JANUARY, 2021 M. Sc in Electronics and Computer Engineering GHULLAM JAILLANI TAKAMUL

HASAN KALYONCU UNIVERSITY

GRADUATE SCHOOL OF

NATURAL & APPLIED SCIENCES

EXPLORATORY VISUALIZATION MODEL FOR MEASURING THE DIGITAL DIVIDE IN ASIAN AND EUROPEAN COUNTRTIES

M. Sc THESIS

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

BY

GHULLAM JAILLANI TAKAMUL

JANUARY 2021

“Exploratory Visualization Model for Measruing the Digital Divide in Asian and European Countries”

M. Sc. Thesis

in

Electrical and Electronics Engineering

Hasan Kalyoncu University

Supervisor

Asst. Prof. Dr. Mohammed K.M. MADI

by

Ghullam Jaillani TAKAMUL

January 2021

	GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES INSTITUTE **M.Sc. ACCEPTANCE AND APPROVAL FROM**

Electronics and Computer Engineering Department, Electronics and Computer Engineering Master programme student **Ghullam Jaillani TAKAMUL** prepared and submitted the thesis titled **"Exploratory Visualization Model for Measuring the Digital Divide in Asian and European Countries"** defended successfully on the date of 10/02/2021 and accepted by the jury as a M. Sc. thesis.

Position	**Title, Name and Surname** **Department/University**	**Signature:**
Supervisor	Asst. Prof. Dr. Mohammed K.M. MADI Computer Engineering Department Hasan Kalyoncu University	
Jury Member	Asst. Prof. Dr. Bülent HAZNEDAR Computer Engineering Department Hasan Kalyoncu University	
Jury Member	Assoc. Prof. Dr. Omar Almomani Department of Information and Network Security The World Islamic Sciences & Education University	

This thesis is accepted by the jury members selected by institute management board and approved by institute management board.

Prof. Dr. Muhammet Fatih HASOĞLU

Director

Document no: ENS.FR.32 Yayın Tarihi: 26.03.2018 Rev no/ Date: 00/

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Ghullam Jaillani TAKAMUL

ABSTRACT

EXPLORATORY VISUALIZATION MODEL FOR MEASURING THE DIGITAL DIVIDE IN ASIAN AND EUROPEAN COUNTRIES.

TAKAMUL, GHULLAM JAILLANI
M. Sc. in Electrical and Electronics Engineering
Supervisor: Asst. Prof. Dr. Mohammed K.M. MADI
February 2021
95 pages

To a large extent, the question of what the digital divide is, including what are its effects on society in general has extensively been studied over the years. Generally, the digital divide is viewed as economic and social inequality to both the access, the use of, or impact of information and communication technologies (ICT) to persons of a given demographic group. The current study proposes an exploratory visualization approach towards the examination of the situation regarding the digital divide in Europe and Asia. In practice, the study uses various visual tools and libraries from python jupyter notebook such as KNN model for imputation for filling missing values using K-Nearest Neighhors, Mann Kendall Trend Test for trend analysis, Ploty for line, box, bar, ,box plots, folium and geopandas for geographical plots to explore distributions on the different aspects of Internet performance in Europe and Asia using data collected based on the PingER methodology that proposes 5 Internet performance metrics i.e., Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packets Lost. Based on the findings obtained after implementing the proposed analytical model for the current study, it is noted that Asia has greater total Packets Lost (848.8074) compared to Europe (562.9666) in the period of examination (2010 – 2018) as well as greater total Out of Order Packets (119.8667 to 91.5), greater Total TCP Throughput (170.909k to 14.80111k), greater total Duplicate Packets (279.0062 to 93), and greater Total Round Trip Time (848.8074 to 562.9666). As such, based on these observations one can argue that there exists a digital divide implying between Asia and Europe with Europe having a better Internet experience compared to Asia in General. However, examining the individual countries it is noted that there are countries such as Pakistani and the United Arab Emirates that show better Internet

performance compared to some countries in Europe. This indicates that apart from the regional digital divide, there potentially exists a country-wise digital divide.

Keywords: PingER, Exploratory Data Visualization, Internet Packets, Pyhton Jupyter.

ACKNOWLEDGEMENTS

I would like to thanks to Assistant Professor Mohammed K. M. MADI for his constant support at all academic and research levels and for his commitment and perseverance with me that enabled me to complete this work.

I would also like to thanks to the Electrical and Electronics Engineering department staffs for their continuous encouragement, excellent guidance, invaluable suggestions, and support at all the stages of my research work. Their interest and confidence in me were the reason for all the success I have made. I have been fortunate to have them as my guides as they have been a great influence on me, both as a person and as a professional.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Pages

LIST OF SYMBOLS AND ABBREVIATIONS

EDA	Exploratory Data Analysis
EDV	Exploratory Data Visualization
TCP	Transmission Control Protocol
RTT	Round-Trip Time
UDP	User Datagram Protocol
ICT	Information and Communication Technology
MPLS	Multiprotocol Label Switching
VoIP	Voice over Internet Protocol
RTP	Real-Time Transport Protocol
ACF	Autocorrelation Function

CHAPTER 1

INTRODUCTION

1.1. Introduction

For several decades, the digital divide has been one of the most celebrated universal phenomena. However, there is no clear definition of the term 'digital divide' owing to the fact that in the mid-1990, an extensive assortment of contentious elucidations and approaches to the digital divide started coming up. The universally accepted technocratic definition of the digital divide is that it is basically the dissimilarity in the provision of access to technological services whereas information sociologist's passive the technological divide as an expression of numerous social, geographical, economical, and informative divides (Van Dijk, 2017). In July 1995, the US department of commerce's National Telecommunications and Information Administration (NTIA) carried out the initial investigation on the Digital Divide in a survey that came to be known as 'Falling Through the Net'. Falling Through the Net is largely credited for introducing the dos and don'ts of what is currently referred to as the Information and Communication Technologies (Kiely and Salazar, 2018). Despite the fact that the digital divide is fuelled by a wide array of elements, the Internet reigns supreme amongst the key drivers of the digital divide at the moment. Internet setups compounded with the quality of Internet performance have gained a lot of significance in all facets of human life whether it is the economic sphere of life, the social aspects, or the political facet of human life.

In recognition of the impact of the digital divide and the role of the Internet performance in widening the digital divides across various societies and regions, the PingER project was brought into existence for the chief purpose of measuring Internet end to end performance in different regions across the globe (White, B. and Cottrell, L., 2016). The formulation of the PingER project was largely informed by the notion that effective measurement of the digital divide in different regions would undoubtedly play a key role in guiding all decision-making processes that are targeted towards tackling the problem of the digital divide (Wenwei and Fang, 2018). The PingER project monitors end to end performance of a wide array of Internet Links globally. Two and a half decades since the PingER project was initiated, the project now boasts of an extensive array of data repositories of Internet performance

measurements sourced from numerous data collection sites around the globe. The PingER project data repository is made up of millions of datasets collected since 1998.

Despite the fact that all pieces of data collected from the PingER project are available online for free, the huge volumes and complexity of the data involved makes it difficult for common individuals to analyse and interpret the data in a manner that will enable them to generate any meaningful insights that can be deemed to be beneficial in the fight against the digital divide. Most individuals tasked with the responsibility of tackling the digital divide lack any meaningful data science knowledge. In recognition of this fact, the administrators of the PingER project sought to employ various data science and information visualization techniques with a goal of making the data understandable to the general populous.

Information visualizations have acquired a huge sense of value in the day to day lives of all human beings. This is because accurate statistical models compounded with information visualization enables humans to come up with a wide array of techniques and decisions. This in turn, eases the burden brought about by various daily challenges by generating accurate predictions and key insight that in one way or another. Therefore, visualization often proves to be quite helpful in human decision-making processes that are targeted towards tackling various problems (Hoeber, 2018). It would be quite impossible for individuals to comprehend the meaning of different forms of data especially when the volumes of the data in question are extremely large. The value of data in human decision-making processes cannot be overlooked but considering the fact raw data cannot be able to relay any meaningful information. Individuals have to rely on general guidelines derived from various statistical models and presented through various information visualization techniques to guide them towards a specific understanding which will enable them to develop a precise action plan. The PingER project is undoubtedly the best practical example of a scenario whereby raw data could be useless especially because extensive volumes of data are involved. Thus, it is definitely impossible to generate any meaningful clues that could guide the next course of action with regard to tackling the problem of the digital divide between Asia and Europe.

In acknowledgment of the extensive negative impact of the digital divide on the general populace, the administrators of the PingER project sought to breakdown the huge volumes of complex data that are continuously generated by the PingER project by employing a wide array of information visualization techniques. According to a

significant number of research studies like (Plaisant and Carpendale, 2011; Balliet and Heimlich, 2016), information visualization is a very robust technique for the exploration of huge volumes of data especially because information visualization amalgamates the superior visual capabilities of human beings with the computational prowess of digital resources. From a general perspective, exploratory visualization is defined as a process whereby a data science practitioner creates different graphics while dealing with complex or relatively unknown data sets. The exploratory visualization process commences with the effective collection of data and culminates with the development of cutting-edge concepts and prepositions (Howe and Heer, 2015). The concepts and prepositions developed from exploratory visualization are used for further analysis and debates. In acknowledgement of the extensive scope and complexity of data, the quantification of the digital divide between Asian countries and European countries employed various exploratory visualization techniques. The utilization of exploratory visualization techniques in the analysis and interpretation of PingER data with regard to Internet performance between Asian countries and European countries provides optimal support for effective monitoring of the digital divide as well as proper decision-making processes with regard to how to tackle the problem of the digital divide between countries situated in both continents.

The most vital element in addition to gathering accurate the data is creation of an accurate supposition that would be helpful in any decision-making process. In most cases, information visualization serves as an instrument for three distinct functions namely, communication, analysis, and exploration (Shneiderman, 2013). When used as instrument for communication, information visualization aims to plainly and accurately relay complex notions to the viewer in a simpler manner that makes comprehension of such complex notions much easier. As for the analytic function, information visualization is employed as a tool for testing different suppositions, with an aim of drawing comparisons or contrasting different elements in order to generate key insights about a specific problem. The exploration function of information visualization is mostly employed as an instrument for generating key concepts that can lead to the formation of meaningful and helpful suppositions on data sets that are not well comprehended by their users. Exploratory visualization models are increasingly gaining relevance as the most important forms of information visualization across various facets of human life. A study on the exploratory visualization model for measuring the digital divide in Asian and European countries

is quite helpful in generating meaningful insights about the exploratory visualization models. Such a study could also shade light on how exploratory visualization models could be effectively employed in generating suppositions and facts that could be useful in tackling a wide variety of problems across various facets of human existence.

1.2.Motivation

Exploratory visualization models are rapidly gaining relevance across a wide variety of professional fields owing to a dramatic increase in the complexity of raw data (Ellis and Mansmann, 2010). Espinosa and Money, 2013) placed lots of emphasis on the development of methodologies and systems that could be used in the visualization of huge volumes of data. Consequently, insignificant interest has been directed towards forming a deeper understanding on the general process of exploratory visualization and the practical effects of exploratory visualization models. There is a strong necessity for the development of a deeper understanding on the process of exploratory visualization. This is because the actual effectiveness of such exploratory visualization models is gaining relevance across different fields and as we look towards the future thus, are most likely to be used by more people across different facets of human life. Information visualization is a potent technique for the effective exploration of complex datasets especially because they amalgamate the loftier visual capabilities of human beings with the computational prowess of different technologies.

If the development of exploratory visualization models as well as the effect of exploratory visualization models were comprehended on the level of the measurement of the digital divide between Asian countries and European countries, it could benefit all types of individuals and organizations. This is because the measurement of the digital divide between Asian countries and European countries is a classic example of the challenges posed by complex and extensive volumes of data. Consequently, exploratory visualization models can be effectively implemented to alleviate the challenges posed by complex extensive volumes of data. The findings of the study would be helpful in developing an extensive comprehension of exploratory visualization models. This is important because a deep understanding of exploratory visualization models can aid data analyst in acquiring key knowledge on how to use exploratory visualization to acquire an overview of the data they are working with. Also, exploratory visualization models provide organizational administrators and

different groups of individuals tasked with the responsibility of making decision with the practical apparatus that enable them to comprehend the issues that they are dealing with and how best to tackle such issues hence the reason why it is quite important to gain an extensive understanding of exploratory visualization models.

1.3. Problem statement

The PingER project data repository is made up of millions of datasets collected since 1998. The huge volumes of complex data have raised several challenges with regard to the analysis and interpretation of different sets of data thereby forcing the administrators of the PingER project to implement various information visualization techniques chief among them with a goal of easing the burden of analysis and interpretation of the extensive and complex data. However, despite noting all these interesting facts about the problem of the digital divide and the role of the PingER project in measuring the digital this study places more emphasis on the Exploratory visualization model used in measuring the digital divide between Asian countries and European countries in PingER.

Exploratory visualization models are rapidly gaining relevance across a wide variety of professional fields owing to a dramatic rise in the complexity and extensiveness of raw data that are gathered by different individuals and entities on a daily basis. However, little interest has been directed towards forming a deeper understanding on the general process of exploratory visualization and the practical effects of exploratory visualization models. Thus, there is a deep needed necessity for further studies on exploratory visualization methods. The measurement of the digital divide between Asian countries and European countries is the best example of a situation where extensive and complex volumes of data are involved (Ordu and Simsek, 2015). Consequently, the measurement of the digital divide between Asian countries and European countries forms a strong basis for developing an extensive and precise understanding of all key concepts of the development and implementation of exploratory visualization models. This study is highly effective in developing a generalised understanding of exploratory visualization models. This is because the measurement of the digital divide between Asian and European countries is not limited to outstanding exploratory visualization algorithms but takes into account the entire process of exploratory visualization modelling.

1.4. Objective of the research

The primary aim of this study is to develop an exploratory visualization model for measuring the digital divide between Asian and European countries with following key objectives in mind:

- To create a clear understanding of what the digital divide using the five network measuring matrices and the importance of exploratory visualization model in measuring those digital divides between Asian countries and European countries.
- To establish an accurate process of exploratory visualization model implementation in the analysis and interpretation of complex huge volumes of data.
- To apply machine learning method on the data to fill the missing data.
- To develop an exploratory visualization model for measuring the digital divide in Asian and European countries.
- To evaluate the developed exploratory visualization model, the step will include some machine learning tests to find the clear image of the output.

1.5. Thesis approach and contributions

The study sought to evaluate exploratory visualization model for measuring the digital divide in Asian countries and European countries. However, the study directed its focus away from the problem of the digital divide towards the exploratory visualization model used in measuring the digital divide between Asian countries and European countries. In seeking to meet its objectives, vital information was sourced from two directions. Firstly, the general process of exploratory visualization was delineated through a comprehensive literature review which will resulted in the generation a fused exploratory visualization process model. Next, the exploratory visualization model was further put to the test through an in-depth analysis of PingER data for European countries and Asian countries. The analysis of the data obtained was done on the python platform and using several another analysis technique including but not limited to regression analysis. The findings of the research were presented using various information visualizations techniques after which the findings were discussed in detail and finally, a clear outline of key recommendations was given. The primary contributions of this thesis include:

i. *The quantification of the digital divide between Asian countries and European countries*- This study will provide an in-depth analysis of key matrices of PingER data with the chief purpose of quantifying the digital divide between Asia and Europe.

ii. *Integration of data visualization in relaying information provided by PingER data*- Considering that all pieces of data in PingER are presented in huge volume and often incomprehensible status the implementation of exploratory visualization processes in the analysis and interpretation of PingER data will generate key insights that will support decision making process aimed at tackling the decision-making processes.

iii. *Visualization in data quality space*- By presenting the different categories of data sourced from the PingER project in a separate view, this project has the potential of making the interpretation of PingER data to be much easier.

iv. *Development of a precise and accurate process model of exploratory visualization*- Basing on relevant models and characterizations in previous literature, this study develops a comprehensive exploratory visualization model using PingER data with a goal of measuring the digital divide between Asian countries and European countries.

1.6. Organization of the thesis

This thesis commences with a clear definition of the digital divide and its relationship with Internet performance and introduction of the concept of exploratory visualization and a clear outline of the reason why information visualization specifically exploratory visualization is an important process in generating key insights that support decision making process especially when the data involved is extensive and complex. In Chapter 2, all literary works that are related to the exploratory visualization model for measuring the digital divide between Asian and European countries are discussed. Chapter three offers in-depth insight into the methodology that was while carrying out the study. On the other hand, chapter provides an in-depth analysis of the available data and a clear presentation of the findings. Chapter five provides conclusions and recommendations for the implementation of the findings of the current research and also the recommendations for future studies.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORKS

2.1.Data Visualization

In the recent past, the universe has observed a dramatic rise in the volumes of data that different organizations collect and process on a daily basis. Owing to the dramatic increase in data volumes, the volumes of data that are currently available across different information platforms including the Internet have skyrocketed in recent times. Despite the fact that a huge percentage of data that are generated on a daily basis are often freely accessible to all interested users, the magnitude of the data often raises a lot of difficulties for interested users in terms of visualization of the data, exploration and ultimate utilization of such data sets (Van Der Aalst, 2016). The realization of the challenges raised by huge amounts of data as well as the acknowledgement of the value of data and appropriate data interpretation to scientific studies prompted different experts have successfully developed different techniques of processing huge volumes of data and presenting them in a manner that can be easily comprehended by any user (Ramakrishnan and Shahabi, 2014). Most notably, the advancement of Information Communication Technology (ICT) as enabled experts in the tech landscape to develop computer software which has the ability to process huge volumes of data and visualize them in an understandable manner.

In the 21st century, each and every individual working within a particular entity whether it is a commercial entity, or a non-profit is increasingly becoming dependent on insights generated from a wide variety of datasets that are collected by their organizations on a daily basis. The insights generated from such datasets help different individuals in making the right decisions, taking the right courses of action in tackling different problems and in overall operational efficiency. Since a huge proportion of the data collected by organizations are huge and complex, it is quite impossible to generate any meaningful insight without enlisting some kind of aid. So, the best possible way of generating meaningful insight from data and avoiding the possibility of missing key correlations entirely rests in in-depth innovative analysis of the available data as well as the use of easy to comprehend data visualizations. From a general context, data visualization entails a wide array of activities including design, advancement and use of graphical representations of processed data sets which are

generated using certain recommended computer software. In most cases, data visualization generates actual representations of the data in question thus enabling the users of the data to realize the data analytics in visual forms that make it simpler for them to comprehend the data (Lewin and Singh, 2018). In summary, data visualization is helpful in the discovery of patterns, process of understanding the information presented by the data and the formation of opinions with regard to the data in question. The concept exploration of data using visualization was brought into the limelight nearly three decades ago by a renowned statistician known as Francis Anscombe. The statistician was able to design a quartet which is famously referred to as the Anscombe quartet. Anscombe's achievements were a clear demonstration that complex datasets can easily be comprehended when presented in a graphical format. Several decades down the line, visual science has undergone tremendous evolution so much that there is currently no doubt regarding the effectiveness of data visualization in relaying information or elucidating complex datasets to a particular audience. However, it is important to note that data visualization can only be effective if the visualizations are attuned in the correct manner such that they can be able to exploit the brain's detection capabilities (Bikakis, 2018). Proper data visualization tends to raise data comprehension speeds as well as the rate at which the information relayed by the data in question is understood especially because visual acuity makes use of the human eye which essentially speaking, has one of the greatest connections to the brain which is the central information processing point.

Owing to the increase in the volume of data available online, the entire globe is currently witnessing a massive gush of attention directed towards data visualization and its ability to relay key pieces of information accurately and efficiently. While taking note of the rising interests directed towards data visualization in recent times, it is important to take note of the fact that a huge proportion of these newly found interests stems from the fact that data visualization is rapidly being acknowledged as a fundamental element in research communication. Despite the fact that the concept of data visualization is a relatively new element in research, it enables several researchers to analyze transform and display complex data sets. Healy (2018) takes note of the fact that the capabilities of data visualization are exceedingly valuable especially in the current information driven environment which is often characterized by massive complex datasets being generated on a daily basis. Lindquist asserts that data visualization is continuously emerging as the most valuable, sense-creating,

analytical and information-relay tool for effectively apprehending and tackling the complexities that are presented by huge volumes of data. The foundation of data visualization rests is rooted in the fact that if the most appropriate data visualization is selected and implemented in the correct manner, it has the potential to accurately reveal the advancement and extensiveness of the underlying issues presented by the datasets in question and possible interventions to such issues while at same time, creating room for further exploration of the datasets in question (Williamson, 2016).

2.2. Standardization of data visualization processes

The creation of accurate data visualizations is a complex process that demands optimum attention and precision. Data visualization is fundamentally a complex form of visual communication and similar to verbal communication, visual communication is largely dependent on semantics and structural accuracy. In this regard, the vitality of the need to fully comprehend the guidelines of proper data visualization development cannot be over-emphasized. It is also important to take note of the fact that there is a huge variation between graphical representation of data and effective visualization of huge complex datasets. Nearly two decades ago, a huge proportion of data scientists who had attempted to use data visualization did so in a poor and absurd manner (Tang and Li, 2018). Despite the fact that huge improvements have been recorded with regard to the use of data visualization in the recent past, renowned experts in the field of data visualization have recently highlighted the value of literacy in all data visualization techniques that have been developed in the recent past. Despite the fact that there is no universal consensus amongst data visualization practitioners regarding the doctrines of proper data visualization, there certain guidelines which are generally recognized as good practice in the field of data visualization. Currently, there are three generally accepted guiding principles related to the design of strong effective data visualizations. These guiding principles include the need comprehend the data, development of a comprehensive understanding what you intend to reveal in the visualization and finally, developing a clear and precise comprehension of your chosen visualization format in terms of its advantages and disadvantages. These guiding principles are elaborated as follows:

- The need to comprehend the data

The most vital dimension of comprehending data is the acquisition of accurate knowledge regarding associations or arrangements within the dataset in question.

From a general perspective dataset are categorized into two distinct groups namely, discrete data and continuous data. Discrete data denotes discrete things that do not have any inherent pattern relative to each other. On the other hand, continuous data is characterized by a particular methodical arrangement. Visualization guidelines dictate that discrete and continuous data should be exhibited in different manners so as to ensure that their correlation can easily be identified. Considering the fact that continuous data is systematically interlinked, visual forms such as line graphs or family trees are most likely helpful in guiding viewers towards the discovery and comprehension of their relationships in a rapid manner (Dur, 2014). As for discrete data, they could be presented using pie charts or any other form of nominal or ordinal scales. It is also important to take not that the two aforementioned groups are the most renowned groups of data, but recent times have seen the development of more distinctions of data types. Despite all the variations that exist in the newly developed data types the common factor that must always be considered when developing data visualizations is the ability to understand the various arrangements of the data as well as the relationships between such pieces of data.

- Development of a comprehensive understanding what you intend to reveal

During any data visualization process, one must always consider the demographics of the target audience as well as the intended purpose of the data visualization process. These two facts must always remain at the top of any data scientist's mind. The aforementioned aspects tend to inform a data scientist's ability to filter the representation of the whole dataset so as not to overwhelm his or her audience. The power of data visualization often stems from its ability to attract the attention and processing capabilities of the viewer as well as their usability traits (Steele and Iliinsky, 2011). However, for data visualization to have strong focus, it is important to put the context of the viewer into consideration as well as his or her inspiration, level of attention and the time that he or she has. As a universal rule, data visualizations have to be kept simple while at same time allowing adequate room for a specific plot or chronological organization of information.

- Developing a clear and precise comprehension of your chosen visualization format in terms of its advantages and disadvantages

In recent times, the universe has observed an intense upsurge in journals seeking to highlight some of the universally accepted best data visualization practices. Such journals have often attempted to provide a step-by-step guide for the creation of highly

effective data visualizations. While there so many rules that have been brought forward by renowned data visualization experts, the most important thing that data visualization designers must always keep in mind is that they must fully comprehend the strengths and weaknesses that different data visualization formats often present (Angus and Wiles, 2015).

2.3. Techniques for the data visualization

There is no doubt that lots of truth lies in the statement that a single picture is worth a thousand words. The substance and reality of this statement is especially evidenced whenever any individual seeks to comprehend a certain dataset or perhaps, attempting to mine certain key discernments from a particular dataset. Considering the fact that in recent times a huge proportion of datasets often occur in large and complex states, visualizations are undoubtedly quite helpful whenever an individual or a group of people are attempting to ascertain the key associations that exist amongst thousands of variables within a particular dataset. In order to come up with meaningful data visualizations, there several basics that must be placed into consideration including the size of the data involved, the type of the data involved and the column composition of the dataset in question (Helfman and Goldberg, 2016). It is also important to take note that in today's rapidly moving society, data visualizations have to be designed in a manner that facilitates quick delivery through different platforms that allow individuals to easily gain access to such visualizations and explore them on their own at their own desired times. Considering the fact that data occur in different forms and magnitude, data visualizations need to be developed based on the configurations of the data in question (Kosara, 2016). In this regard, it is important to take note of the fact that data visualization techniques occur in different forms some of which can only be applied to simple and small datasets and are often referred to as basic data visualizations and those that pertain to the visualization of large and complex data sets. Before embarking on a comprehensive study of data visualization techniques whether it is in the context of small and simple datasets or large and complex datasets, it is important to take note of the fact data visualization is not entirely similar to scientific visualization. Scientific visualization often makes use of animations, simulations and complex computer-generated graphics in creating visual designs of different configurations and procedures that processes that are not entirely visual in nature. On the other hand, data visualization tends to present and exhibit different sets of

information in a manner that is inclined towards encouraging proper interpretations, assortments and associations within the dataset. Unlike scientific visualization, data visualization is known for its ability to utilize human capabilities in the recognition of patterns and the analysis of various trends that exist within the dataset in question while at the same time exploiting the human ability to retrieve significant volumes of information within the shortest time span (Shishkin and Skatkov, 2016).

2.3.1. Basic data visualization techniques

Basic data visualization techniques are those that are often implemented when tacking small and simple datasets. The common basic data visualization techniques include:

- Line charts

Line charts are often used to reveal the association that exist between two or more different variables and are often used to keep an eye on the trajectory of certain variations over a certain period of time. It is also important to note that the usefulness of line charts mostly rests in their ability to draw comparisons between numerous subjects or items over a particular time span (Tu and Chen, 2017).

- Bar charts

In most cases, bar charts are used when drawing comparisons between the quantities of a wide variety of groups. The values of each group are epitomized using a single bar and the bars can be constituted as either vertically or horizontally with the span of each bar acting as a representation of its value. It is also important to note that a simple bar chat often comes in handy when values appear to be quite dissimilar so much that the variations between each bar can be discerned by the ordinary human eye. However, in an event when the bars may appear to be too close or perhaps, the data set is made up of huge numerals meaning that the bars need to be extremely huge, it would be quite difficult to employ the use of bar charts (Xu and Nandi, 2016).

- Scatter plots

Scatter plots are two-dimensional plots that reveal the combined variations that exist between two data objects. In most cases, scatter plots are helpful when seeking to ascertain just how spread out a particular dataset might be.

- Pie charts

In recent times, there have been so many discussions revolving around the usefulness of pie charts which are often used to draw comparisons between different

sections of a whole. The discussions stem from the fact that pie charts are characterized by an elevated level of difficulty in terms of their interpretation especially because it is quite challenging for the human eye to estimate areas and compare visual angles.

- Box Plot:

A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of box to show the range of the data.

- Autocorrelation Plots:

Plots are a commonly used tool for checking randomness in a data set. This randomness is ascertained by computing autocorrelations for data values at varying time lags.

2.3.2. Visualization of Big Data

Essentially speaking, the visualization of enormous datasets is a complex process owing to the fact that a huge proportion of data has to be exhibited on a somewhat tiny space. The measurement of the digital divide between Asia and Europe is an excellent example of the complexity of data visualization whenever huge volumes of data are involved. In recent times a wide array of cutting-edge techniques of big data visualization has been brought into existence.

Big data visualization techniques such as cluttering often deal with sampling and filtering processes. It is also important to take note of the fact that whenever big data visualization is being undertaken, there are some common steps that need to be followed. Such steps are clearly outlined in the below Figure 1.

2.4.Exploratory visualization

Exploratory visualization is the process of creating imagery and graphical utility of statistical components to aid in data presentation. Exploratory visualization is mostly used to showcase the geographical representation of the data analyzed to uncover the underlying relationship among the dataset have a presented. It is not obligatory to have preset statistical model to use the exploratory data visualization techniques since by its definition it has to provide more beyond the formal custom modeling or hypothesis testing task (Li, 2018). The growth of exploratory data

visualization was channeled by the exploratory data analysis (EDA), which was championed by John Turkey in the mid-20th century around the year 1960. John had the aura that data analysis was the backbone of research and that it required to be the most considered part and provided the threshold attention. He envisioned that data had to be analyzed comprehensively with use of adequate technical resources which then led to the invention of open source programming languages; the S, S plus and the R, these developments led to the advancement in the data analytics through dissemination of robust statistics and nonparametric statistics, which were necessitated by the testing of median, mode, mean, standard deviation, deviation and quartiles successes whose findings were orthogonal to the primary analysis task (Cox, 2017).

According John Turkey the main objectives of the EDA were to:

- Advocate for hypothetical causes and their phenomenal parameters
- Evaluate the condition under which the prevailing assumptions will be held.
- Evaluate the decision for selection criteria for appropriate data analytic tool.
- Provide for the basis of data collection and statistical procedures.

The comprehensiveness of the EDA procedures set it at a different class compared to IDA; initial data analysis.

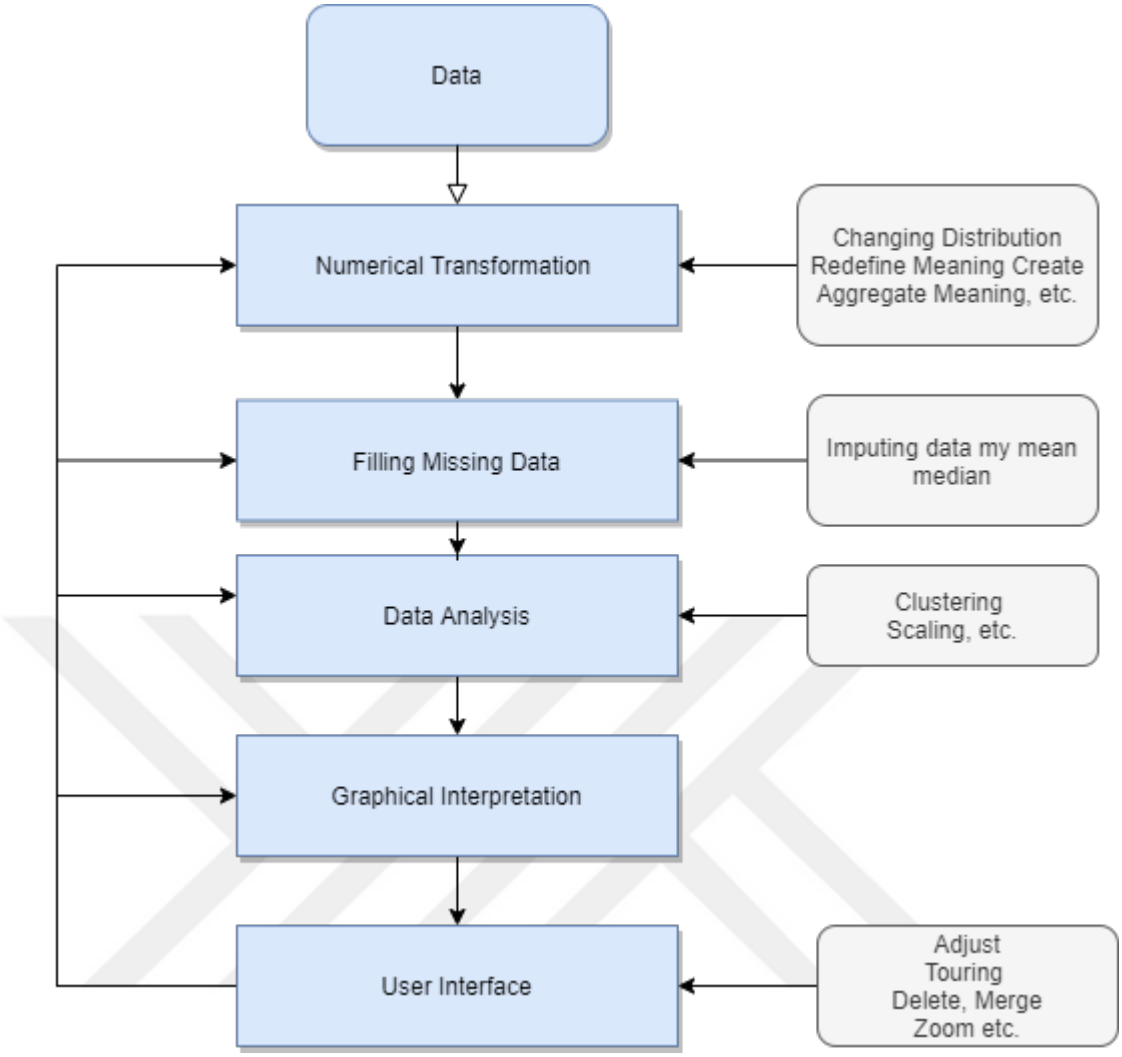


Figure 1: Common steps in Data Visualization

Source: (Nayak and Lenka, 2016)

2.4.1. Extension of Exploratory Data Analysis (EDA)

With the surge of machine language and development of data analytics, there is a void of technical services required to easily link the two. EDA is advancing as the time goes by, year after year since its first invention by its founder Sir John Turkey witnessed by the invention of numerous open sources programming languages championed by Python programming language which have the capacity to handle sophisticated data analytics requirements (Cox, 2017).

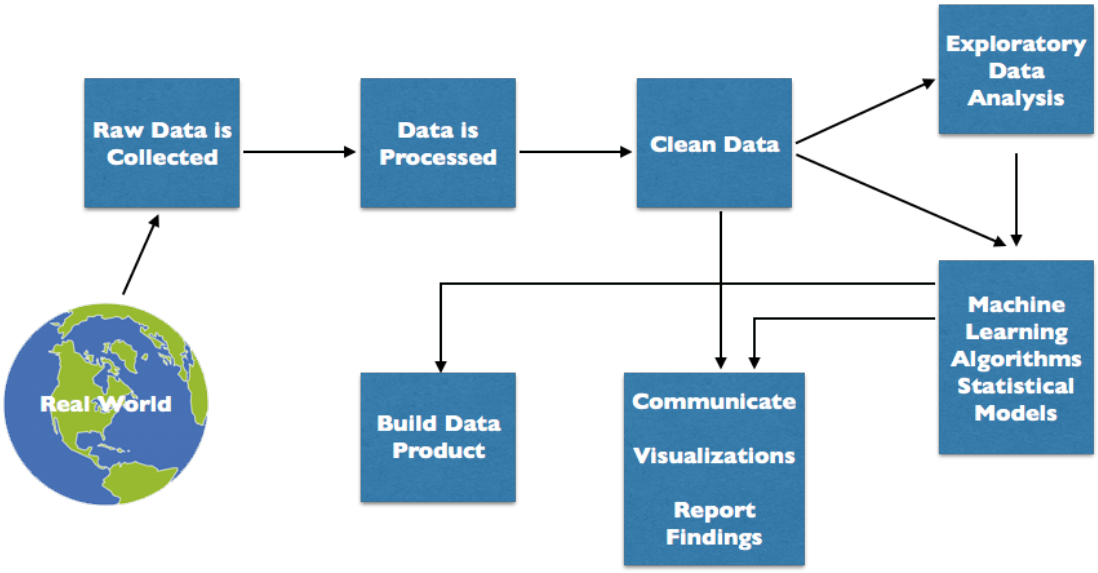


Figure 2: Data Processing

Source: (Trishana , 2020)

From the figure above it is deducible that data visualization is the last step of data handling prior to make decision. This shows how data visualization holds a central position in the data handling and how it is dependent on the data processing, model and algorithm employed in the data processing. Also, from the chart it is deducible that the exploratory data analysis is the engine of the data processing with its fundamental interlinking with the prime parts of the chart; the data processing, data cleansing and lastly the models and algorithms employed.

Data exploration analysis provides the data analyst the opportunity to:

- Verify the data and ascertain the relationship that exist within the data sets.
- To check for unanticipated structures within the data that needs to be altered, removed, or changed accordingly.
- To ensure that the data process is governed by data-driven insights and not in any way motivated by assumptions of stakeholders.
- To provide data-based context relating to the problem sighted for the data science procedures that can place the statistical outputs to its maximum use (Cox, 2017).

2.4.2. EDA Methods

From the definition of the EDA data science and its application on real world problems it can be divided into two major sections. Turkey's work on exploratory data

did not possess clear division within the structuring of the EDA, however with clear precision on the work, it is divided into the two major section; the first method as non-graphical or graphical while the second classification is based on univariate or multivariate (majorly bivariate). The non-graphical method is purely based on computation of numbers and arithmetical mathematics while on the other hand the graphical data method is based on data presentation using appropriate visualization tool to have diagrammatic or pictorial way exhibition (Cox, 2017). The figure 3 displays the EDA methods interfaces and how parts and sub parts are interlinked from the descriptive statistics to their appropriate visualization.

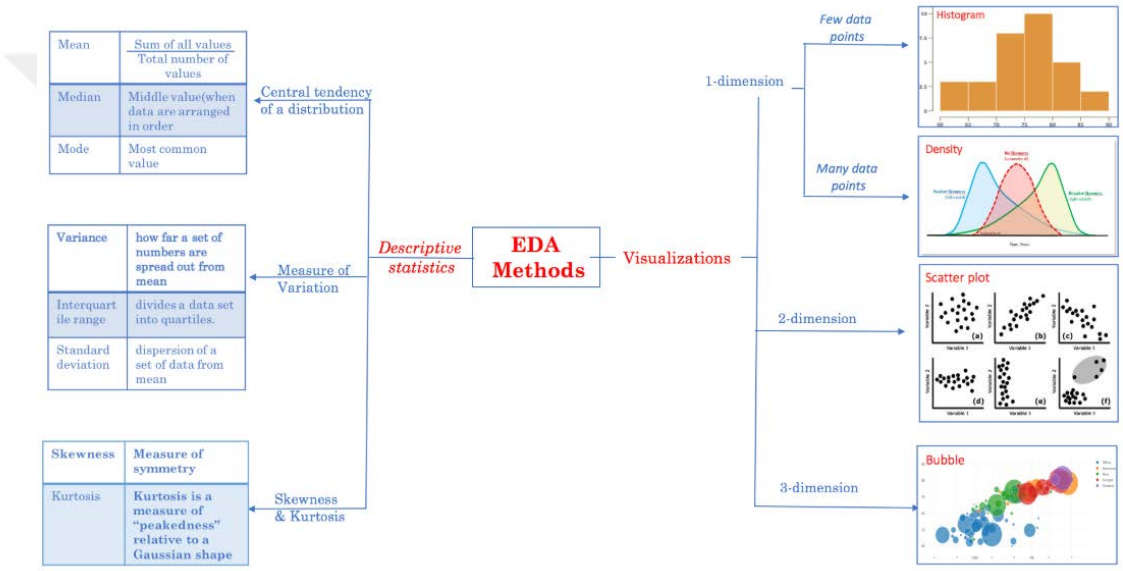


Figure 3: EDA Method

Source: (Sanket Doshi Feb 3, 2019)

2.4.3. Dimensionality reduction and Cluster analysis for EDA

In the last 3 to 4 years more data was produced and recorded compared to data available for equal time periods in the entire world. These huge volumes of data chronicled at single times imply that data is being produced in more dimensions which are increasing day in day out (Reddy and Baker, 2020).

This can be attributed to the surge use of online and digitalization technologies and software such as Facebook, WhatsApp, google among other which allows machines to interact with humans on daily lively routines. For example, the amount of data Facebook collects per minute is baggy considering that it must stores all its

user's personal information, likes, comments and reselects most liked items to repost as favorites.

Considering that slack amounts of data are configured by machines on timely intervals, it is necessary to have a system that sorts the data in such a way that only imperative data are left while superfluous data are wiped out to save storage spaces and computational time. As a result of the mind-boggling numbers of machine data, Dimensionality reduction is developed to assist in analyzing and presenting statistical inference through visualization.

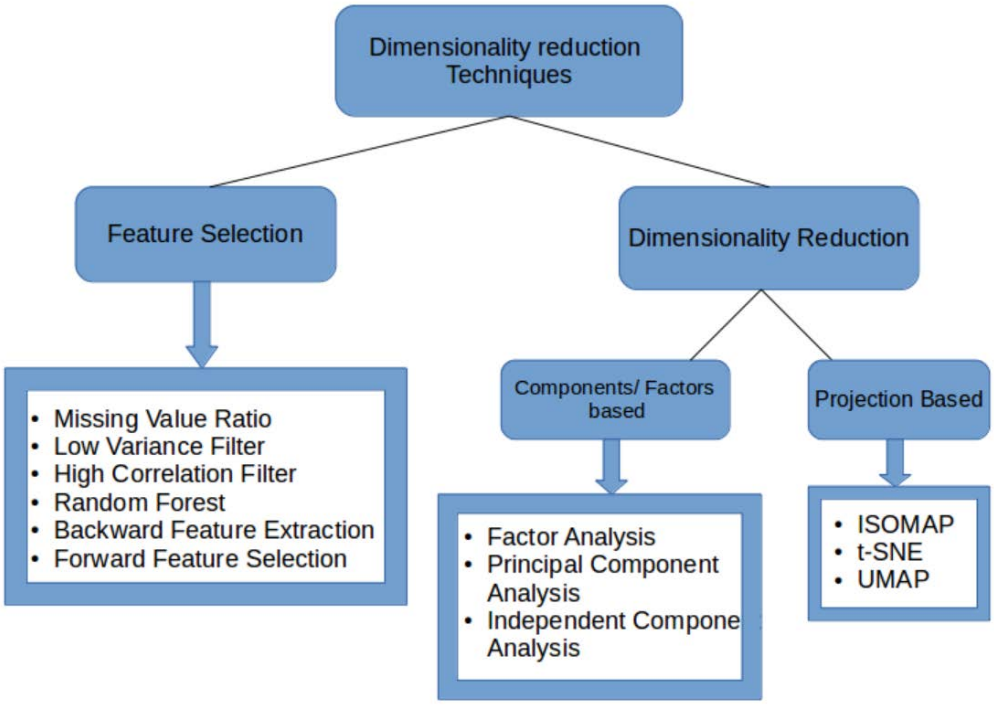


Figure 4: Dimensionality Reduction Techniques

2.4.3.1.Benefit of Dimensionality Reduction

Here are the benefits of applying Dimensionality Reduction on hefty data:

- As mentioned above, Dimensionality Reduction is based on sieving data allowing only pertinent data to be confined by the machine, thus saving storage space. It should be noted that voluminous data have rows and columns that are truly irrelevant considering the threshold of the target analysis objectives (Reddy and Baker, 2020).

- Reducing the scope of data, the machine is dealing with allows the machine to work fast thus also saving the computational time.
- Removing the extraneous data from a group data wipes out multicollinearity. The process of confiscating data is solemn based on deleting columns and rows with redundant information (Reddy and Baker, 2020).

Summary on when to use Dimensionality Reduction.

- Missing Value Ratio: this method is used when the dataset has too many missing values. It is based on dropping variables with most missing values.
- Low Variance filter: this method is used when most variables in the data have constant variables from the dataset. As a result, only variables with low variance are wiped out.
- High Correlation filter; this method is based on selecting variables within the dataset with high multi-correlation and drop them accordingly.
- Random Forest; this is most used most of Dimensionality Reduction and unlike other methods of its kind, it is based on selecting variables based on their importance. The top important variables are maintained while the rest are removed (Becht and Newell, 2019).
- Factor Analysis: this method is used mostly when the variables in the dataset are highly correlated. It works by grouping variables based on their correlation with each other.
- Principal Component Analysis: the method is mostly used with linear data. It works by dividing the data into smaller components groups which are then analyzed accordingly.
- Independent Component Analysis: also used with linear data, but unlike PCA, Independent Component Analysis works by selecting independent variables within the dataset.
- ISOMAP; this techniques bests fits the non-linear data sets
- t-SNE; this method just like ISOMAP best fits the non-linear data sets however it has improved visualization.
- UMAP; the method bests suit the highly dimensional data.

2.4.3.2.Cluster analysis

Clustering is a special type of unsupervised machine learning where variables within a provided data set are classified and grouped base on given similarities to. It is normally used to identify strata of similar objects in a multivariate data set to improve its meaning on generative features on the existing data set. Unsupervised machine learning bank on drawing of statistical reference from dataset comprising of input data with unlabeled responses (Granato and Maggio, 2018). The classification of data divides data into groups containing similar objects and dissimilar objects, thereby making it more statistically viable through data visualization.

2.4.3.3.Need for perform cluster analysis

As mentioned above, clustering is the process of grouping data based on their properties and objects into statistical groups which are based on shared similarities, consequently the intrinsic grouping of dataset will be highly treasured considering unlabeled data. It creates rooms for extracting value from large sets of structured and unstructured data and ease the process of finding the logical patterns existing within the dataset.

Types of clustering analysis

- Partitioning algorithms; this method of clustering is based on dividing the datasets into K groups where K is the number of groups selected by the data analyst.
- Hierarchical clustering; unlike the partitioning algorithm, the data analyst does not have to select the desired number of grouping since it the Hierarchical clustering is grounded on tree modelling system of data basing on dendrogram selection.
- Fuzzy clustering; this type of clustering is more complex than the other types of clustering since it allows a member of a cluster to also be members of other clusters.
- Density-Based Methods; this method of clustering is based on the dense nature of cluster, implying that regions with dense cluster have more similarity compared to region with less dense (Kassambara, 2017).

2.4.4. Importance of Exploratory visualization

Regarding the voluminous amounts of data that are processed daily it is difficult to keep an accurate tract of data without using Exploratory visualization assistance. Human brain is a powerful tool which do have the ability to do computation problems, however it is affected by memory issues and tends to forget easily. Exploratory visualization therefore aids with ample visualization that enables the human brain to understand data trends much more improved than mere presentation of premeditated reports (Kumar and Johnson, 2020). Exploratory visualization matched with exploratory data analysis (EDA) forms a complete fleet of data sciences that comprehensively analyze and present data effectively.

2.5. Measuring the digital divide

As mentioned earlier digital divide is defined as the dissimilarity in the provision of access to technological services whereas information sociologist passive the technological divide as an expression of numerous social, geographical, economical and informative divides. In simple terms it is the technical inequality among nations in the entire universe, with some better placed with Internet services while others have poor Internet related service. Research done earlier which focused on digital divide were all centered on simple technical services surrounding the Internet environment. As a result, these studies were narrowed to number of users capable of accessing Internet, number of computers per given number of people (mostly per hundred), number of persons having computers per given geographical positions (mostly per a square mile), gauging the Internet users basing on age, gender and educational backgrounds (Büchi and Latzer, 2016). Consequently, the results of analysis showed that some nations such as Japan, United states, Singapore, Taiwan and majority of the European counties had an improved accessed to Internet services basing on the above-mentioned technical variables as compared to most third world countries in Africa, southern America and parts of Asia. On the other hand, this methods of measuring digital divide included the cost of accessing Internet on the said countries. Internet service fee is a significant factor in determining the digital divide since the cost of having broadbands in the entire world varies from country to country. Technology divide has brought out as a phenomenon which covers three distinctive aspects namely.

- **The global divide-** This refers to the inequalities in access to Information Communication Technologies in a worldwide context, as in between countries or between different continents.
- **The social divide-** This refers to the inequalities in access to information Communication Technologies in a societal context, as in between different sections of a country's communal organization.
- **The democratic divide-** This refers to the variations between individuals who have access to Information Communication Technologies and those that do not have such privileges with respect to the inability to fully engage in public issues as much as those who have full access to ICTs (Grant and Eynon, 2017).

This study focused on measuring digital divide using exploratory visualization for Asian and European countries. Unlike other previously done studies on digital divide amongst the two continents the study sought to employ techniques were basing on the five Internet packets namely, the **packet loss** which basically happens when all the data packets sent fails to reach the intended destination, **Round-trip time (RTT)** which is the time its takes a browser to receive response from the server, **TCP throughput** which is the ideal trouble shooting time taken by a browser to rectify or to identify problem with the Internet or the browser mostly measured in latency divided into UDP Throughput and TCP Throughput which are classified basing on impacts with latency, **Out of Order Packets** is another major chronic Internet problem which happens when Internet packets received by the browser are different from the one sent by the server, lastly the **Duplicate Packets** which normally happens when identical packets are sent by the server (Abdelsalam and Zampognaro, 2017). The five Internet packets mentioned above are critical in provision of Internet services. Digital divide is formed when other countries of geographical positions receive more improve Internet services than others in other areas. Internet speed is determined by the closeness to the nearest server, which translate to the point that should an Internet server be positioned at Africa then Africans will be experiencing a faster browsing efficiency as compared to other Internet users positioned outside Africa, the longer the distance from Africa the poorer the Internet speed (Kharat and Kulkarni, 2019). Servers have then been positioned on strategic geographical positions which ensure adequate speed for all, while the other Internet packets have continued to affect Internet activities differently.

2.6. Related work

2.6.1. Internet Performance Analysis

According to the studies done under the title of Internet Performance Analysis of South Asian Countries Using End-to-End Internet Performance Measurements, the 2017 article by Saqib Ali, Guojun Wang, Roger Leslie Cottrell and Sara Masood focuses on investigating the relationship between the Internet performance of south Asian countries using End to end Internet performance measurement with key economic development metrics of a region using data from PingER. TCP throughput of the countries was found to correlate with different development indices.

One another article under the Title: Comparison of network performance of India and Pakistan using PingER data, the article was done Prof (Dr.) Bebo White, Akshat Sachan, and Dr. A. Sai Sabitha was conducted in the year 1995 and was oriented on measuring network performance of Pakistan and India for various reasons and was keen on employing pingER using the round-trip time RTT to evaluate the network difference based on network condition that were prevailing at that time. In the current study the main focus is not only on one network measure or not on smaller area the study is done on Asian and European countries level and continent level compare the network measure of five different matrices for 100 pionts data collection, the data source is PingER.

2.6.2. Exploratory Visualization

The article by Shiping Huang is about Exploratory Visualization of Data with Variable Quality in year 2005. According to Huang Data visualization is a process of data mining and analysis using graphical presentation and interpretation. The accuracy and the correctness of the visualization is 100% depend on the quality of the original data collect. Huang emphasises on the high-quality data. The article by Noora Routasuo, is keen on the role of Exploratory visualization in explaining data and data analysis in year 2013. The researcher goes steps further and do suggest that there is so much that exploratory visualization can do apart from visualizing data which includes data collection, processing, interpretation, evaluation, and communication of exploratory results. The current study used pyhton different libraries to perform the different steps of exploratory visualization to obtain the clear understand.

2.6.3. Data Visualization in Exploratory Data Analysis

The work by YINGSEN MAO focuses on the EDA in year 2015. Mao defines EDA as the process of ‘asking questions’ and extracting knowledge from data which slowly gaining momentum in data analytic and being absorbed into business intelligence and other significant business analytics. Mao concludes by acknowledging that EDA compliments the c confirmation data analysis (CDA).

The article by R. Cottrell and Sheryar Khan focuses on Quantifying and Mapping the Digital Divide from an Internet Point. The authors acknowledges that the existing digital divide can be measured using Internet performance measured and quantified using the 700 hosts and 115 countries that have been in the measuring process since the year 1995. The results can be used to improve the conditions of the countries found to be lagging in the digital race.

CHAPTER 3

METHODOLOGY

3.1.PingER Framework

Nearly three decades ago, the world woke up to the realization that Internet performance is vastly linked to various key regional economic growth metrics. In response to the newly found realization, a deep-rooted necessity to develop an elaborate and precise framework for effective monitoring and comprehension of Internet links performances across the globe. Driven by the necessity to monitor and develop a precise understanding of Internet performance across various regions, the PingER project was born. The PingER project was brought into existence with a goal of figuring out various digital infrastructural inadequacies, reduced resource digital distributions and Internet routing problems across various regions in order to generate possible solutions for the future. Various Internet performance metrics across the globe are obtained via the PingER framework which was established by the SLAC National Accelerator Laboratory based in the United States of America. The PingER project as a whole boasts of very extensive deployment levels across the globe. The deployment stats indicate that there are over 700 hosts distributed across more than 115 nations across the globe from hosts in over 30 nations. The 700 hosts correspond to over 2200 remote host pairs across the globes which are known to take an estimated value of 200,000 ping capacities every single day (Mal and Cottrell, 2016).

3.1.1. The PingER framework performance monitoring methodology

The PingER framework is made up of three host types namely, monitoring host, remote hosts and archive hosts. Basically, monitoring host refers to a specific computer which contains software which is referred to as the PingER Monitoring Agent. At the moment the PingER project boast of 50 monitoring hosts spread across 23 nations in different regions across the globe. On the other hand, remote hosts typically refer to a collection of website servers that have stable uptime. Remote hosts are constantly monitored by 50 monitoring agents at designated steady intervals. Unlike monitoring hosts, remote hosts do not require any software, but they must be pingable at all times from the monitoring agents. The PingER project currently boasts of almost 700 remote hosts which are effectively spread across 170 nations in different regions of the globe. Consequently, the PingER framework is made up of 10,000

Monitoring Agent remote hosts pairs spread across the globe and actively monitoring and measuring the performance of Internet links (Pan, and Leslie, 2016). The PingER framework measures Internet performance at a regular thirty-minute interval. The measurement sequence is triggered from each monitoring agent by conveying a set of 100 bytes ping requests and 1 kilobyte ping requests to a particular group of remote hosts.

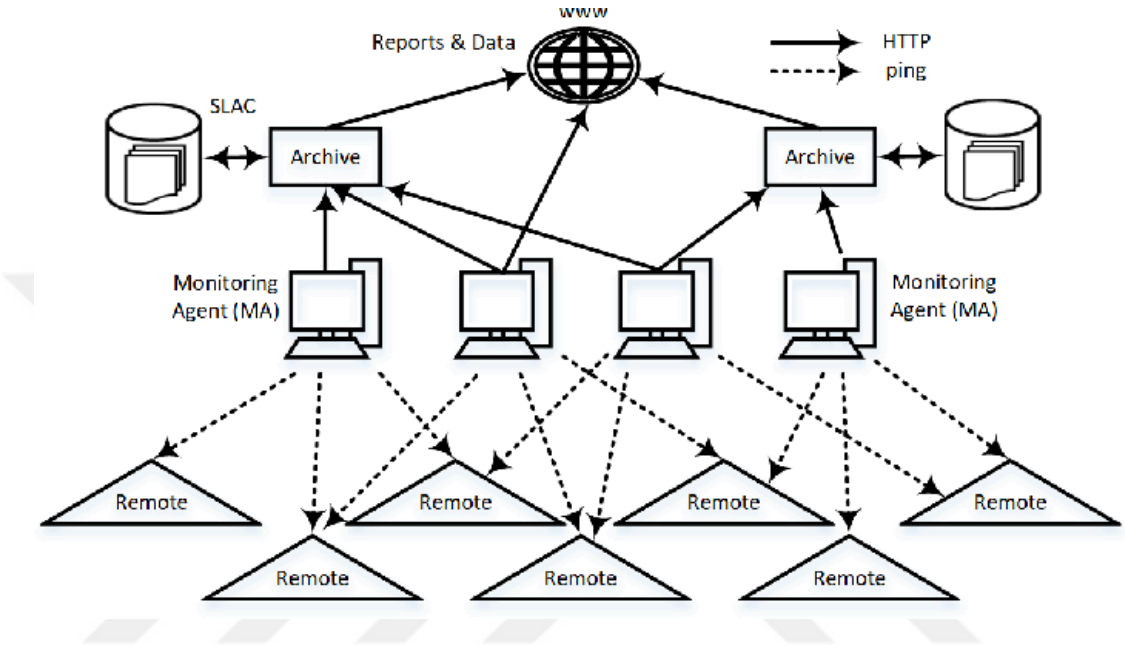


Figure 5: General build of the PingER framework

Source: (S. Ali, G. Wang and R. Cottrell, 2018)

Monitoring agents are designed to halt all processes of sending pings when it collects 10 ping replies from the remote hosts or when the overall ping request conveyed to the remote host hits 30. The PingER framework records data from each set of Ping. From a general context, raw data obtained from the PingER framework is made up of names and IP addresses of the Monitoring Agents and the target remote sites as well as the timestamps, packet sizes, minimum, average and maximum Round Trip Times (RTT) figures of the ping reactions and the individual ping Round Trip Times and their sequence values (Sampson and Cottrell, 2017). The archive host performs the vital role of retrieving all pieces of raw data collected by Monitoring Agents on a day-to-day basis. Fundamentally speaking, the Archive host is a term used to refer to central data storage warehouse located at SLAC headquarters. As things stand at the moment, data obtained by the PingER project through the PingER

framework is the true definition of big and complex data. The volume of compressed PingER datasets is currently estimated to be over 60 GB and is made up of numerous flat files. The raw data obtained in PingER is processed to generate different key performance metrics which can be accessed by the general public through the PingER website.

3.2.Data discovery and selection

Big data is currently generating lots of unparalleled opportunities for commercial entities to accomplish comprehensive, quicker discernments that ultimately inform all their decision-making processes with regard to the enhancement of client experiences and acceleration of their innovation speed. However, due to the extensiveness and complexity of big data, they cannot generate any meaningful value in their original state. Consequently, most organizations dealing with big and complex data have resorted to the utilization of visualization-based data discovery tools. Generally speaking, visualization-based data discovery tools enable users to integrate data from various sources and perform actual analytics that eventually tend to showcase viable results in a persuasive, collaborative and an easily comprehendible format. For analytics and visualization purposes big data is subdivided into three distinct classes of data namely;

- ***Descriptive data***

Descriptive data is the simplest form of data amongst the three-business analytics data. Ideally, basing on its simple raw nature up to 90% of the existing businesses in the world uses the descriptive data to identify the underlying trends in the data (Englander, 2012). The main purpose of the descriptive data is to discover out the ins and outs behind prized success or failure of the organizations in a given time period. It is basically used in answering the question ‘what had happened?’ it is important in deducing the way the corporation will take events focusing on its previous actions and their respective results. It should be noted that if some actions done by the corporation resulted in poor results then the management of the corporation will be keen on establishing and taking different actions that will necessitate better results. In machine leaning descriptive data can be used to speculate the previous action that were undertaken in the training of models in machine learning that effective failed in the

testing model. Such methodologies of training data will not be used again in training data or will be reviewed (Shkedi, 2011).

- ***Predictive data***

Unlike the descriptive data which generally focused on the past the predictive data is centered on predicting what the future holds for such data in business analytics. It is centered on answering the question 'what could happen in the future to data regarding the previous trends?' Analyzing data successfully can enable the data analyzers to effectively and comprehensive predict the future (Crawford and Schultz, 2014). This can enable the business enterprises, corporation and interested organization to set real time realistic and achieve goals, allowing of operative scheduling and confining prospects. Predictive analytics can be used to study the data and ogle into the crystal ball problem solving tool (Hazen and Jones-Farmer, 2014).

- ***Prescriptive data***

The prescriptive data is an extension of the predictive data. It may not an excellent future predictor which can predict soccer and lottery winning numbers but good enough to suggest best decision to the businesses organization when it comes to decision making. The predictive data is centered on answering the question 'what could happen in the future to data regarding the previous trends? while the prescriptive data describes what should the business do, this is the most important aspect has it do contain the action that needs to be done to achieve the set results and goals (Chalamall, and Papotti, 2014). Unlike the previous business analytic data; the descriptive and the predictive data, the prescriptive data is not simple in nature and as a result very few companies and organization spread across globe uses it. Less than 7% of all businesses in the world uses the prescriptive data, since it is expensive and requires complex expertise (Soltanpoor and Sellis, 2016).

3.3.Steps of data preprocessing

Different data require different attention since they all varied combination some may have missing values, others varied formats while others may have duplicate data which spans to unrequired data. This implies that one has to keenly inspect data to realize the kind of treatment it requires (Alasadi and Bhaya, 2017).

The data involved in this study is sourced from PingER containing at least five matrices namely; TCP Throughput, Packet Loss, out of order packets, Duplicate Packets and Average round trip data who is each matrix contains data spanning from the year 2010 to 2020 for the Digital divide existing between the European and the Asian continents.

3.4.Data quality Assessment

It involves the process of accessing the data to ascertain if it contains all the necessary information, while at the same time inspecting for:

- Missing values; from the data collected from the PingER website, each matrix data was inspected for missing values. Fortunately, no rows or columns had missing values, and should one have some, the column or the row would have been deleted or the value of the missing variable estimated using K-Nearest Neighbor approach.
- Inconsistent values; this implies data which are either too large, too small, or contains information that do not match the expected information on a particular row or column. Since data was sourced from a reputable website all values were consistent (Kahn and Liaw, 2016).

Figure 6: Data Quality

The figure above illustrates what composes the data quality. The pinder data was checked for completeness, accuracy and relevance.

3.5.Feature aggression

The data from the Pinger website contains varied information spanning over a long period of time. Feature aggression comes in as a handy tool which enables placing of aggregated values in order to increase data perspective. This eventually reduces data size and consequently save space and data processing time (ZHANG and Lei, 2019).

3.6.Feature Sampling

Almost similar to feature aggression, data from the PingER website is sampled narrowly to contained only the data that contains only required information such as only the data from the year 2010 to 2020, while selecting only significant rows and columns and contains relevant information for the study needs, thus saving space and data processing time.

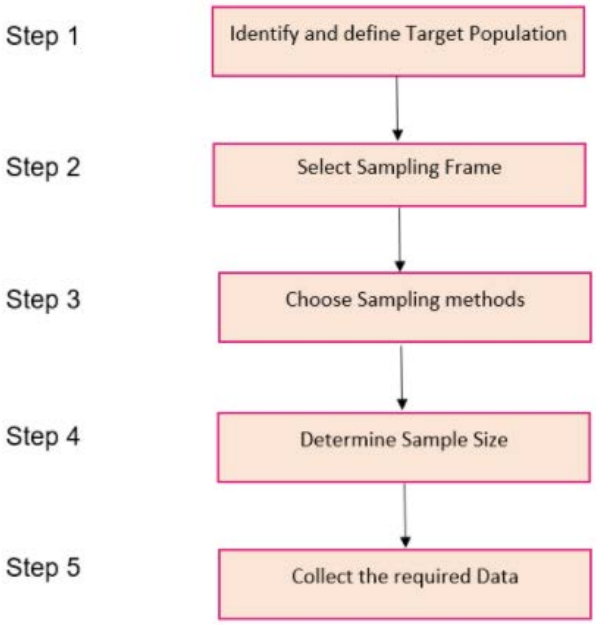


Figure 7: Feature Sampling

Figure 7 shows a stepwise collection of data, beginning with defining of the target population, sampling setting and lastly collecting of the required data.

3.6.1. Dimensionality Reduction

Similar to the above feature reduction and feature aggression, dimensionality reduction is aimed at reducing the space and the time taken for the data pre-processing. But unlike the feature reduction and feature aggression, it does not major on reducing the sizes and numbers of rows and columns involved but reducing the number of features involved in the data by mapping the higher-dimensional feature-space into lower-dimensional feature-space. Note that high dimension takes more planes to map output while lower takes much lesser values making easy to map a 2D image, aided by removal of irrelevant features and noise (Becht and Newell, 2019)

3.6.2. Feature encoding

The whole idea behind data pre-processing is to present correct and accurate data in a manner that is easily understood by machines. The data sourced from the PingER website was robust implying that was suitable for machine consumption. It only required minimal editing and encoding of data which was achieved through ample selection of data for the five Matrices.

3.7.Implementation of visualization model

3.7.1. Visualization

The importance of visualization is that it fastens and eases the process of understanding a given concept by aligning thoughts and ideas to the way the human brain works best.

3.7.2. Visualization model

Visualization is an important aspect in data science and is not only used to create diagrams and images but also used to manipulate data to create varied imageries from same data thus providing more sound meaning of the data analyzed. A model of visualization is used to provide this vivacious data manipulation, through provision of link amongst hypothesis and experiment and link amongst insight and revised hypothesis (Tappini et al 2019).

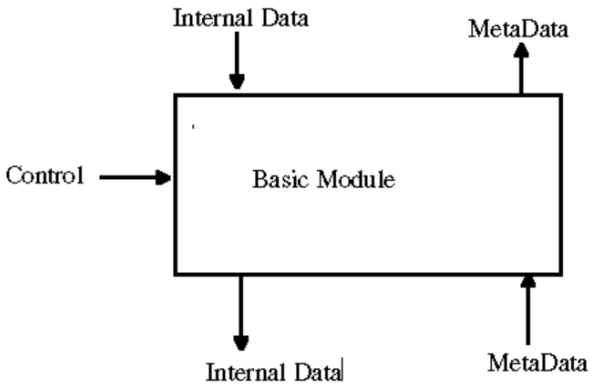


Figure 8: Data Visualization

Source: (Tappini et al 2019).

During this process, several types of data are involved such as:

- **Control Data:** this includes the data that triggers and pedals all the modules in the system. The control data is the Pinger data.
- **User Input/Output:** This includes the system computer common input operations and tools such as the keyboard and the computer output tool such as screen and speakers for sound and sonification, which are then transformed to Metadata for the system modules.
- **Internal Data:** data already existing in the system.
- **External Data:** data can be transferred into the system.
- **Storable Data:** this includes data that can be stored, maintained and accessed within the system.
- **Graphics Data:** includes internal data that can be manipulated into graphical items (2D or 3D)
- **Picture Data:** data which have limited graphical primitives such as the 2D.

A Visualization Technique module takes internal data sourced from the external model and transmutes it into a best matching Base Graphics System module.

3.8.Visualization Process

Visualization of data is a process, which do involve some key steps namely: the data generation phase, data enrichment, mapping of the visualization, the rendering phase and the last step of the process the display phase.

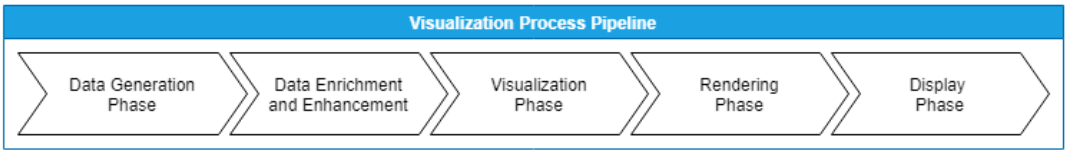


Figure 9: Data Visualization Process Pipeline

From the five steps mentioned above, the first and the last stands outside the process of data visualization.

- The data from PingER is already generated, hence it is uploaded into the system of the data analysis tool.
- The second step; the data enrichment is basically an improvement on data cleansing where operation such as the Domain transformations, interpolation, sampling, and noise filtering are done and enhanced to aid computation.
- The third step is the central part of the visualization process where only the comprehensive and mutual inclusive data from PingER were mapped to visual primitives and attributes.
- The second last step in data visualization; the Rendering phase is the step where the primary images of the data are suited to graphics primitives and attributes, lighting operations and anti-aliasing filtering to create high end 3D images.
- The final step is displaying of the figured visualization with options of copy, storing and renaming.

To suit the needs for the data, clustering analysis was selected for this study since it had the best properties to display the disparities that do exist amongst the European Countries and the Asian Countries based on digital divide.

3.9.Evaluation

Evaluation is the process of determining merit, worth and significance of a project, tool or utility.

3.9.1. Evaluation of the visualization

The selection and the execution of the visualization model and its processes is excellent. The need to exclude noise, duplicate, incorrect and inconsistent data ensures that the only precise data were included in the calculation and computation of the visualization. The data selection and inclusion from the PingER website is exact and

exceptional since only data within the required time range of 2010 to 2020, essential European and Asian counties data on the only five matrices were selected.

To add on the selection of the clustering analysis as the Visualization model is because the data was compact and well separated which fits computation under clustering since clusters are relatively scalable (Kassambara, 2017).

In the coming chapter we will show the visualization using python Jupyter code and covering all the mentioned steps.

CHAPTER 4

RESULTS AND OUTPUTS

The current chapter includes the results of the exploratory data analysis and exploratory data visualization methodology defined earlier. In practice, the results reported herein are aimed at addressing the research objective i.e., to develop an exploratory visualization model for measuring the digital divide between Asian and

Table 1: Network Metrics

Metric	Definition
Duplicate Packet	This includes any packet that is identical to another packet. Duplicate packets tend to generally lower the statistical accuracy of analysis besides increasing the network link saturation as well as its inherent ability to interfere with tools.
Round Trip Time	This generally includes the length of time it takes for a data packet to be sent to a destination in addition to the time it takes for an acknowledgment of that packet to be received back at the origin
TCP Through Put	Throughput measures how many packets arrive at their destinations successfully. Ideally, Packet arrival plays an integral part in leading to high-performance service within a given network. Packets that are lost leads to slower network performance.
Out of Order Packets	Out of Order Packets occur when "the delivery of data packets on a computer network is different from the order in which they were sent." Order Packets is caused by different factors including data streams through multiple sources or through parallel processing paths in a network tool that is not designed to handle the preservation of packet ordering.
Packet Lost	By definition, Packet Loss comes about when single or more packets of data traveling across a computer network do not reach their intended destination.

European countries with the following key objectives in mind:

- To create a clear understanding of what the digital divide is and the importance of an exploratory visualization model in measuring the digital divide between Asian countries and European countries.

- To establish an accurate process of exploratory visualization model implementation in the analysis and interpretation of complex huge volumes of data.
- To develop an exploratory visualization model for measuring the digital divide in Asian and European countries.
- To evaluate the developed exploratory visualization model

In this section, different visuals will be used to provide an in-depth overview of the situation regarding the digital divide between various countries/regions in Europe and Asia. Ideally, the analysis is focused on the five pingER-based metrics i.e., Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Lost.

For a quick reference, Table 1 below provides a brief overview of the description of the metrics provided above.

Having explored the matrices/metrics that will be explored including what their effect is on the performance of Networks across different regions, the sections below provide the results obtained during visualization.

4.1.Descriptive Statistics

4.1.1. Asia

Table 2 shows the mean entries for Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Loss for the Asian region.

From table 2 it is noted that except for Out of Order Packets, which is constant i.e., the Duplicate Packet, TCP Through Put, Out of Order Packets, Round Trip Time, and Packet Lost for Asia are lower in 2020 relative to 2010.

Also, table 3 below provides an overview of the standard deviation of Asian Countries for each of the Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Lost metrics.

Figure 10 is a the boxplot presentation of five metrics for Asia, the different color boxes show different metrices for digital divide and each of them have four quatile (q1, q2, q3, q4) with outliers.

4.1.2. Europe

Table 4 shows the mean entries for Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Loss for the European region.

Table 2: Asia Network Metrics Mean

	Region	Data-For	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	Asia	Duplicate Packet	104.47	95.85	96.69	95.40	89.45	83.56	86.76	80.84	81.03	81.72	83.82
1	Asia	Out of Order Packets	39.96	39.96	39.96	39.96	39.96	39.96	39.96	39.96	39.96	39.96	39.96
2	Asia	Packet Lost	3.34	3.36	2.58	2.17	2.17	2.08	2.30	1.54	1.44	2.18	1.89
3	Asia	Round Trip Time	291.97	281.52	275.32	271.68	263.52	247.45	249.37	237.78	235.42	236.80	243.23
4	Asia	TCP Through Put	22842.27	54788.39	93278.36	75317.44	26502.81	2707.68	4715.78	2534.42	6772.77	6414.98	8963.68

Table 3: Standard Deviation for Asia

	Region	Data-For	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	Asia	Duplicate Packet	164.54	152.42	152.91	146.40	124.81	106.37	114.35	100.87	101.75	103.50	105.90
1	Asia	Out of Order Packets	50.82	50.82	50.82	50.82	50.82	50.82	50.82	50.82	50.82	50.82	50.82
2	Asia	Packet Lost	3.78	7.54	4.27	3.37	3.35	3.41	3.34	3.38	4.78	5.26	4.98
3	Asia	Round Trip Time	131.01	105.59	104.05	94.06	75.83	49.62	53.74	44.05	39.79	38.66	35.65
4	Asia	TCP Through Put	140043.26	255341.14	403429.17	382790.72	139779.40	8462.56	20398.99	4532.99	27199.52	30618.65	41175.13

Similarly, from table 5 above it is evident that except for Out of Order Packets and Duplicate Packet, which are constant, the TCP Through Put, Out of Order Packets, Round Trip Time, and Packet Lost for Asia are lower in 2020 relative to 2010. The claims of the decreasing trend will be explored later in which, a statistical test will be conducted to examine the trend of each of the metrics across the two regions.

Figure 11 is a the boxplot presentation of five metrics for Europe, the different color boxes show different metrices for digital divide and each of them have four quatile (q1, q2, q3, q4) with outliers. Figure 12 is same boxplot but without outliers.

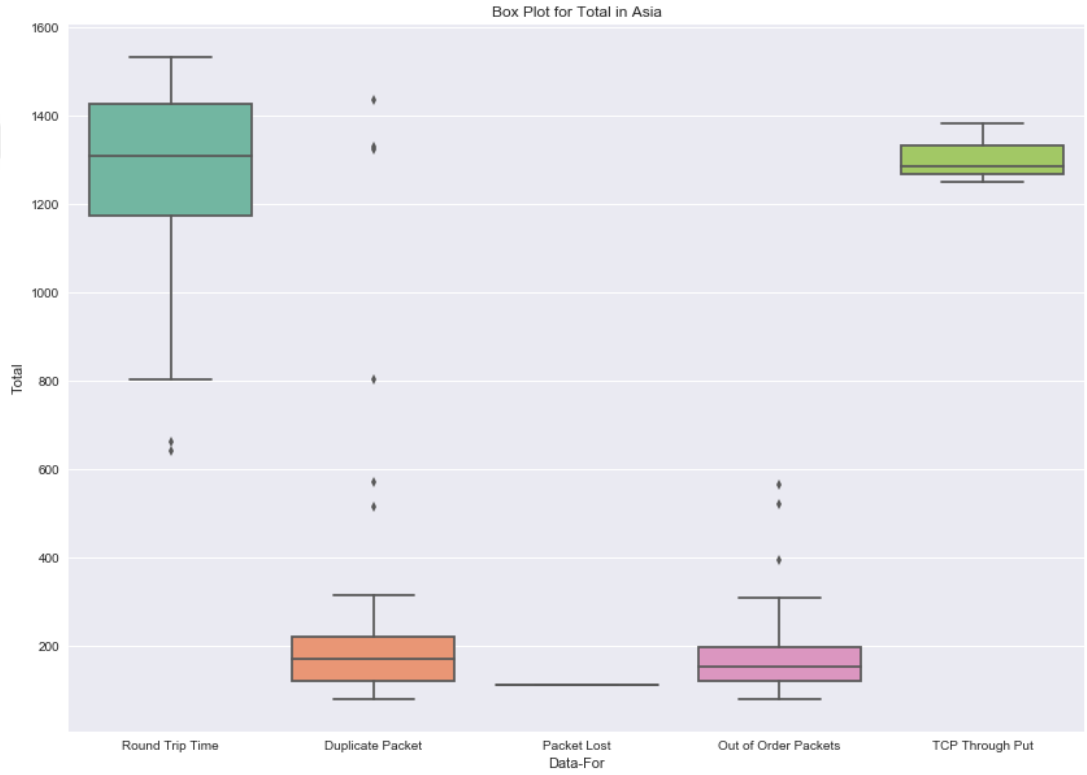


Figure 10: Box plot presentation for Asia

Table 4: Europe Network Metrics Mean

	Region	Data-For	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
5	Europe	Duplicate Packet	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00	31.00
6	Europe	Out of Order Packets	30.50	30.50	30.50	30.50	30.50	30.50	30.50	30.50	30.50	30.50	30.50
7	Europe	Packet Lost	1.88	1.68	1.63	1.44	1.15	1.00	1.41	0.71	2.31	2.24	3.12
8	Europe	Round Trip Time	186.69	189.71	186.57	186.15	185.08	177.54	180.11	182.86	185.38	185.53	170.89
9	Europe	TCP Through Put	4131.29	4641.40	6028.43	19642.05	16430.70	9228.95	5700.47	12055.50	9707.98	9096.28	11030.56

Table 5: Standard Deviation for Europe

	Region	Data-For	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
5	Europe	Duplicate Packet	18.74	18.74	18.74	18.74	18.74	18.74	18.74	18.74	18.74	18.74	18.74
6	Europe	Out of Order Packets	18.26	18.26	18.26	18.26	18.26	18.26	18.26	18.26	18.26	18.26	18.26
7	Europe	Packet Lost	3.02	2.68	3.15	3.15	2.28	1.83	2.02	1.97	7.22	5.33	7.96
8	Europe	Round Trip Time	31.56	33.61	31.28	30.52	30.08	26.62	27.44	29.88	25.60	15.29	12.89
9	Europe	TCP Through Put	4941.74	9819.51	16746.99	90813.01	70613.39	33369.21	11055.35	38792.83	15241.66	14553.83	24295.97

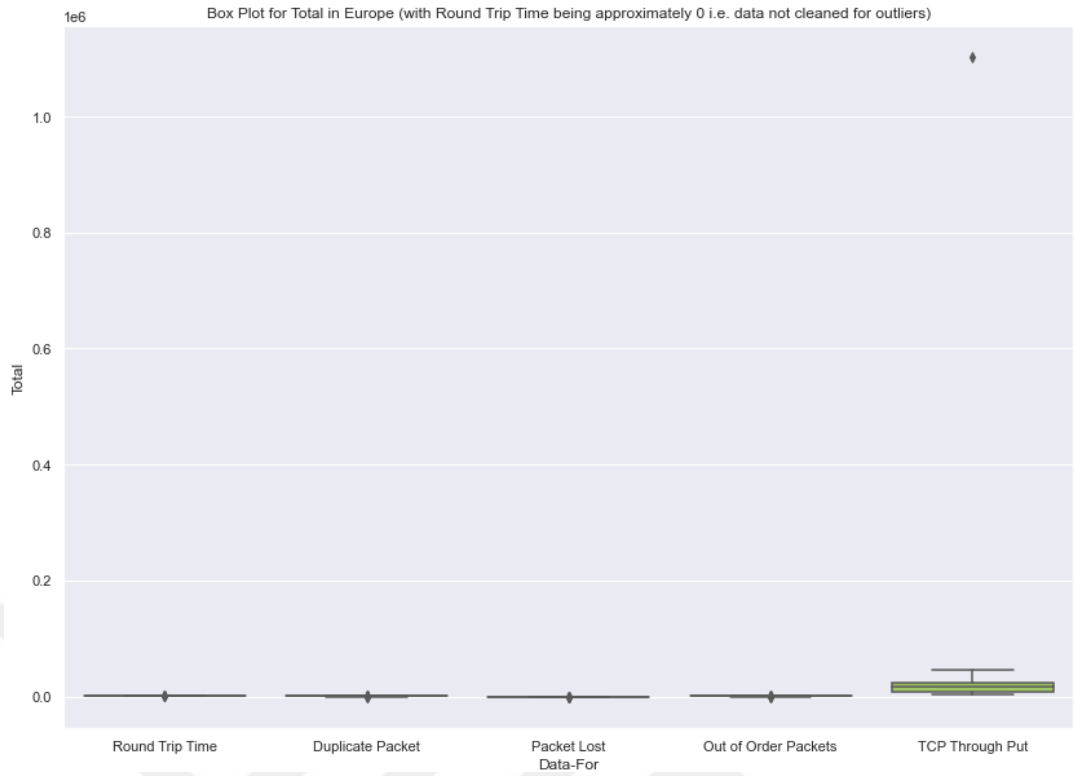


Figure 11: Box plot presentation for Europe with outliers

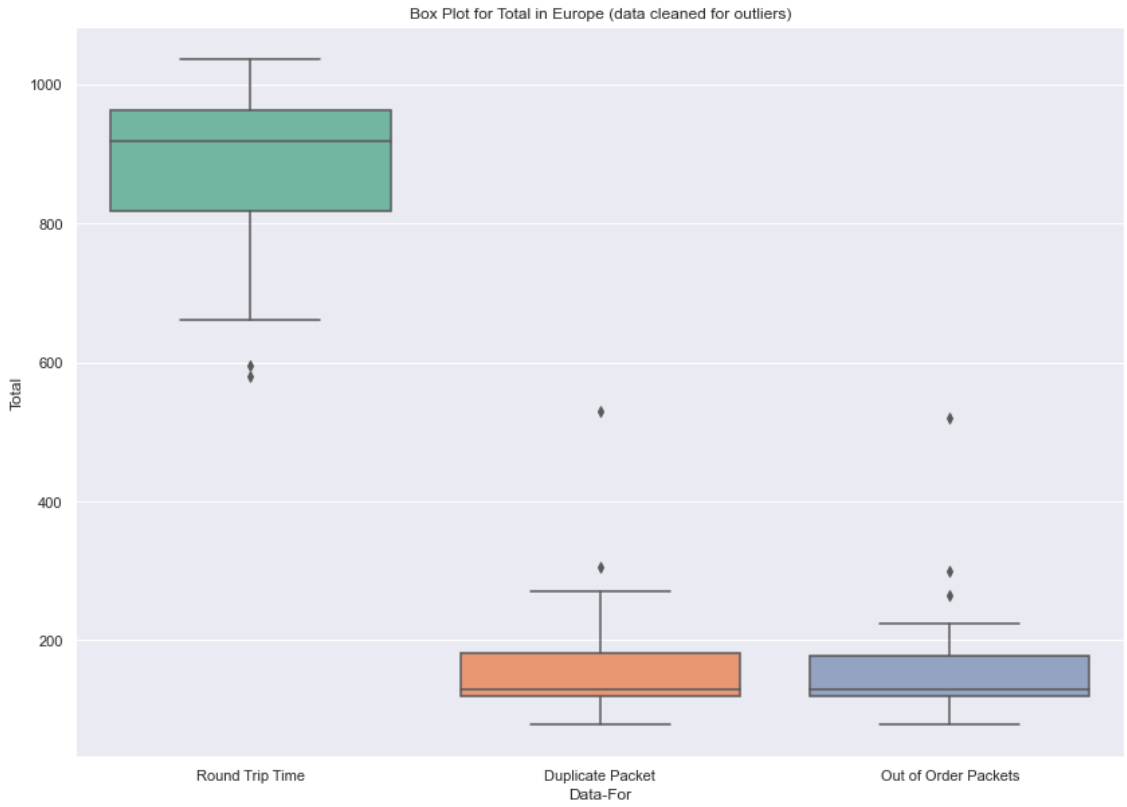


Figure 12: Box plot presentation for Europe with outliers

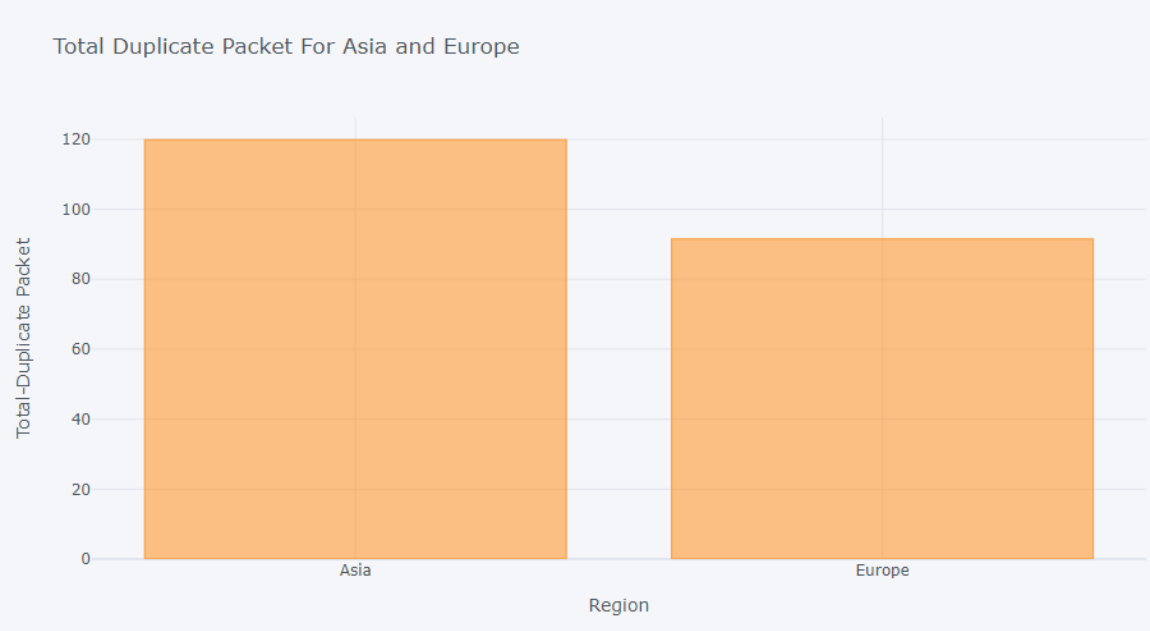


Figure 13: Total Duplicate Packet

4.2.Visual Exploration

General Regions

In concept, the goal of visualization as noted previously is used to enable this vivacious data manipulation, through the provision of link amongst hypothesis and experiment and link amongst insight and revised hypothesis.

4.2.1. Total Duplicate Packet

Figure 13 shows the Total Duplicate Packet distribution while figure 14 shows the Total Duplicate Packet for Asia and Europe over time.

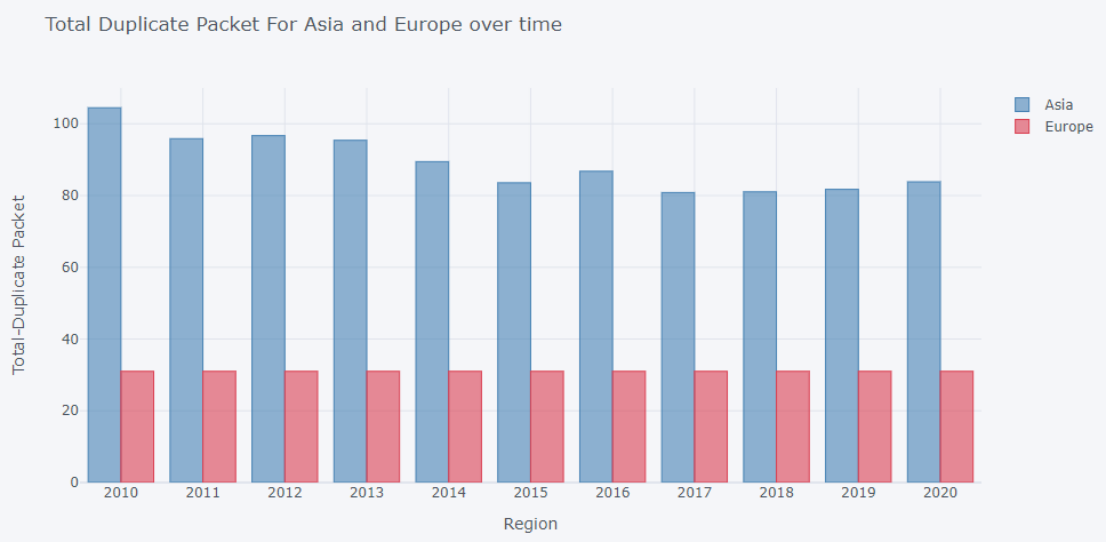
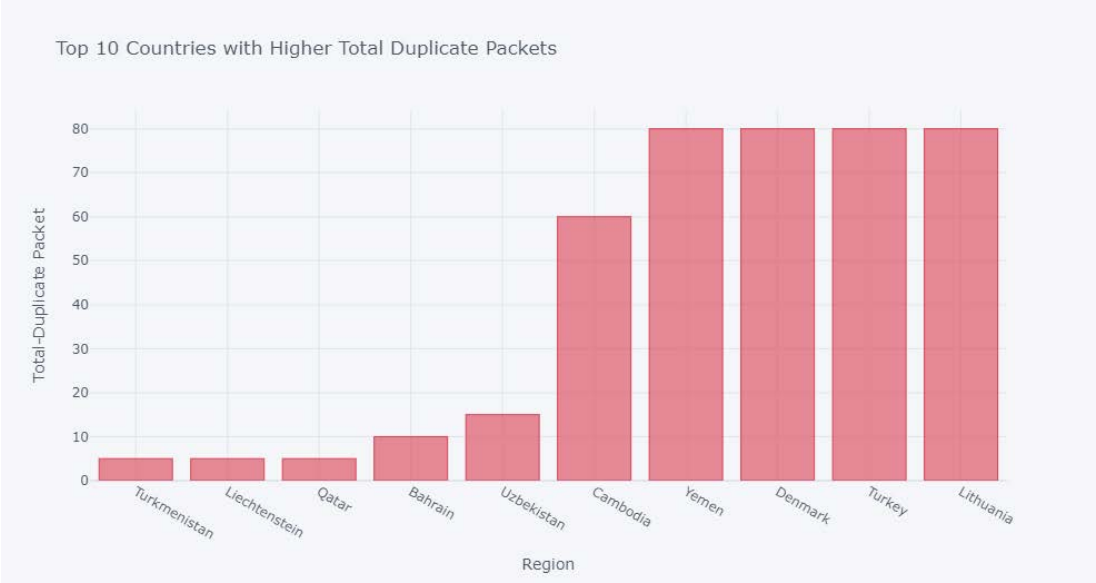


Figure 14: Total Duplicate Packet (2010-2020)



- **Figure 15:** Top 10 Countries with High Total Duplicate Packets-Bar Plot

Asia as noted in figures 13 and 14 has greater Total Duplicate packet scores compared to Europe throughout time. It is noted that Europe has a constant distribution of total duplicate packet scores between the years 2010 and 2020 (see figure 14).

The below figure 15 to figure 17 show the top 10 countries from Asia and Europe with Total Highest Duplicate packets in different visualization persentations.

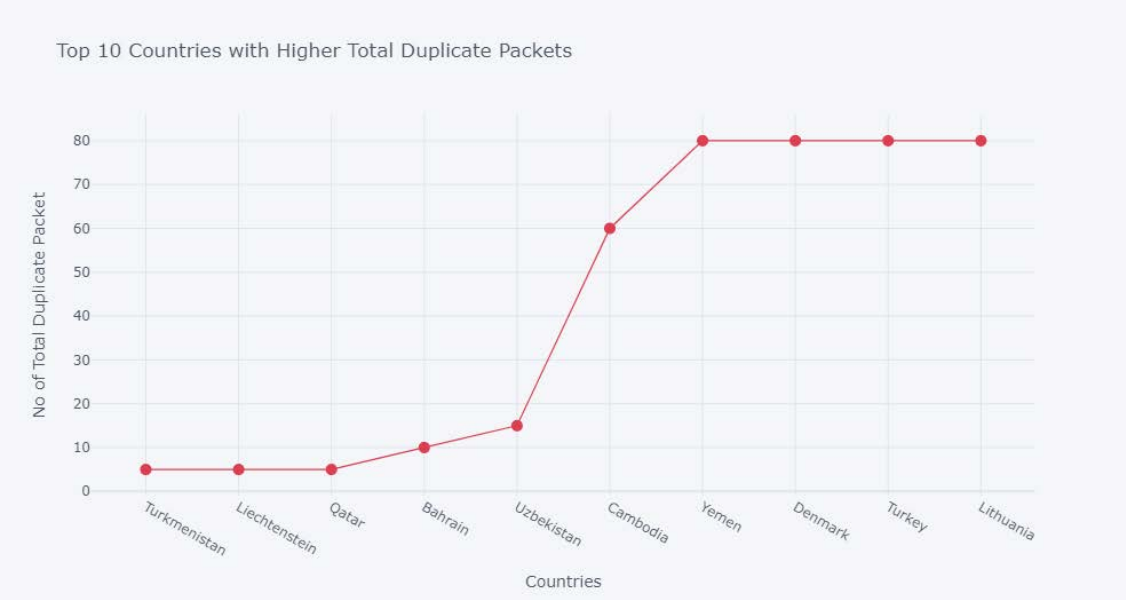


Figure 16: Top 10 Countries with High Total Duplicate Packets- Line Plot

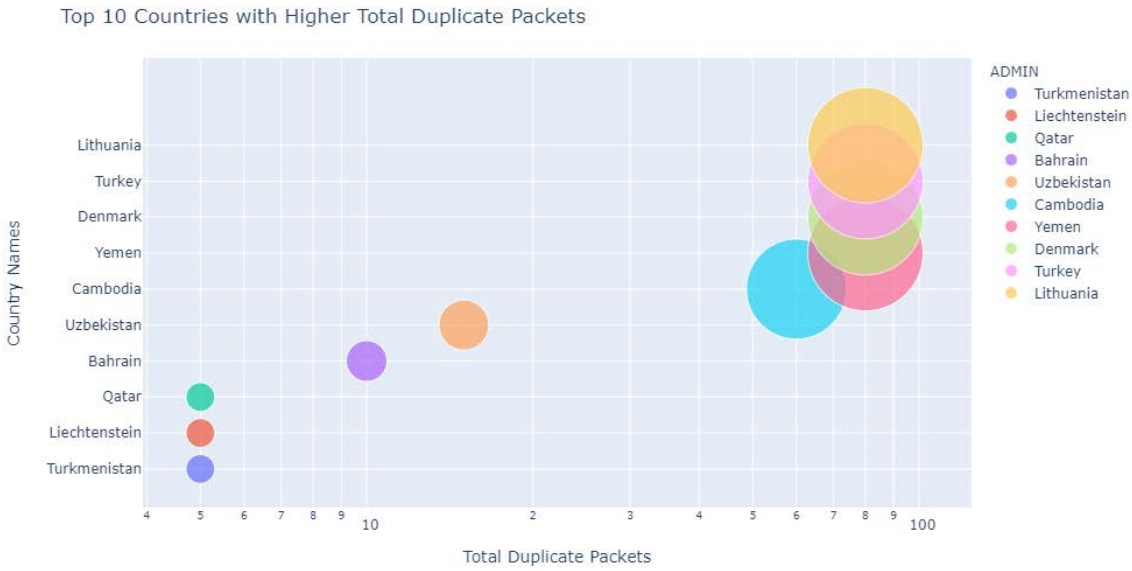


Figure 17: Top 10 Countries with High Total Duplicate Packets- Bubble Plot

The below figure 18 to figure 20 show the top 10 countries from Asia and Europe with Total Lowest Duplicate packets in different visualization persentations.

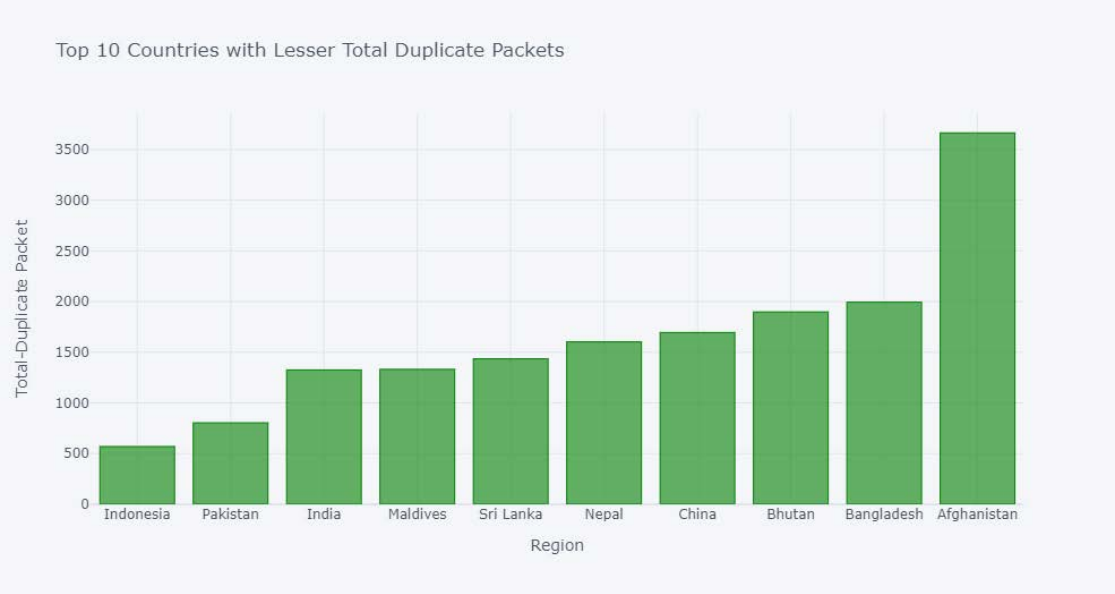


Figure 18: Top 10 Countries with Lesser Total Duplicate Packets-Bar Plot

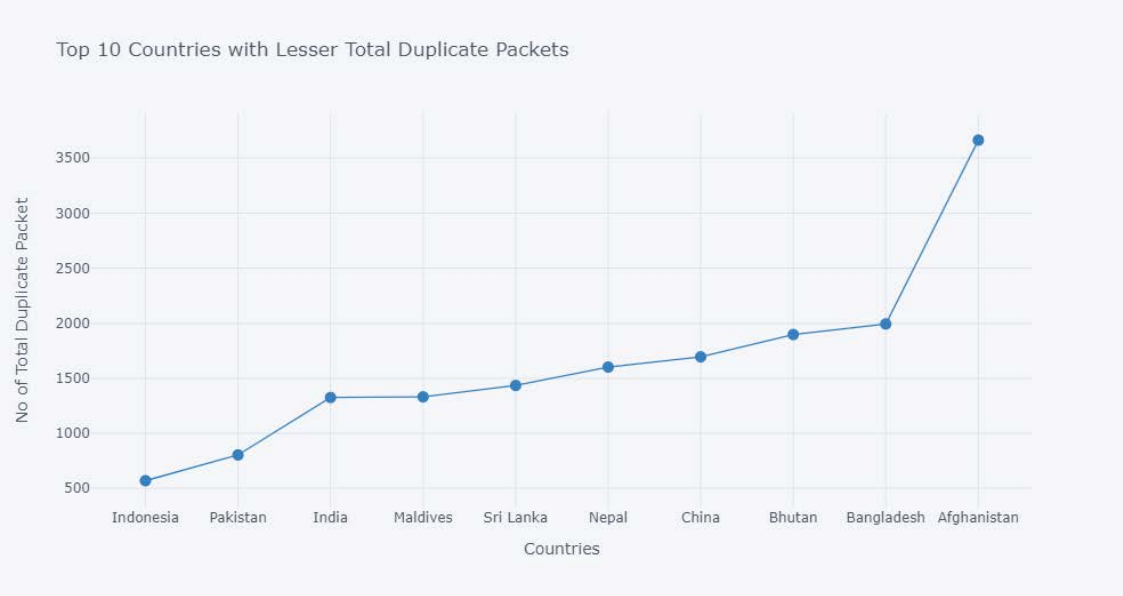


Figure 19: Top 10 Countries with Lesser Total Duplicate Packets- Line Plot

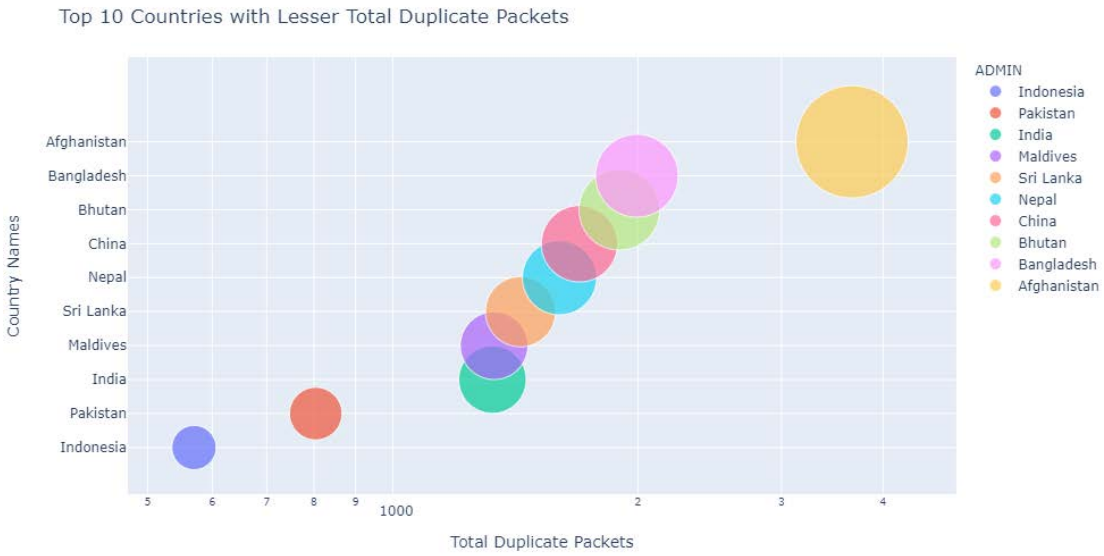


Figure 20: Top 10 Countries with Lesser Total Duplicate Packets- Bubble Plot

Statistical test

Visualizations tend to provide a general distribution overview of the aspects being inspected. To examine the significance of such a distribution, say in the determination of whether a given phenomenon is increasing, decreasing, or constant, it is imperative to use statistical tests. As such, a trend analysis will be conducted whose aim is to determine whether the 5 different matrices are going up (increasing trend), going down (decreasing trend), or staying the same (no trend) using the Mann-Kendall trend test which analyzes the sign of the difference between later-measured data and earlier-measured data.

Mann Kendall Trend Test

Whereas there exist several approaches to analyze for trend, the study will use the Mann Kendall Trend Test. The role of the Mann Kendall Trend Test which is a non-parametric test implying that it is not affected by the distribution of the data being used i.e., doesn't have to meet the assumption of normality. The Mann Kendall Trend Test is mainly to analyze data collected over time for consistently increasing or decreasing trends (monotonic) in Y values (Stephanie, 2016). The main disadvantage of the Mann Kendall Trend Test is that the data used should not have been collected seasonally. Besides, the data should not have any serial correlation since this could affect the significant level (p-value) and could potentially lead to misinterpretation. To address

this problem, the study will adopt a modified Mann-Kendall test i.e., Hamed and Rao Modified MK which is known to account for serial correlation (Blain, 2013). However other modified Mann-Kendall tests include Yue and Wang Modified MK Test, the Modified MK test using the Pre-Whitening method, etc. (Hussain and Mahmud, 2019). The first step towards conducting a Mann Kendall Trend Test is to determine whether the data has any correlation is not, then the original Mann Kendall Trend Test will be used otherwise, the modified Mann-Kendall test i.e., Hamed and Rao Modified MK will be used.

For the total Duplicate Packet in Asia, table 6 below shows the distribution of the Mann Kendall Trend Test.

Trend Test

Autocorrelation Test

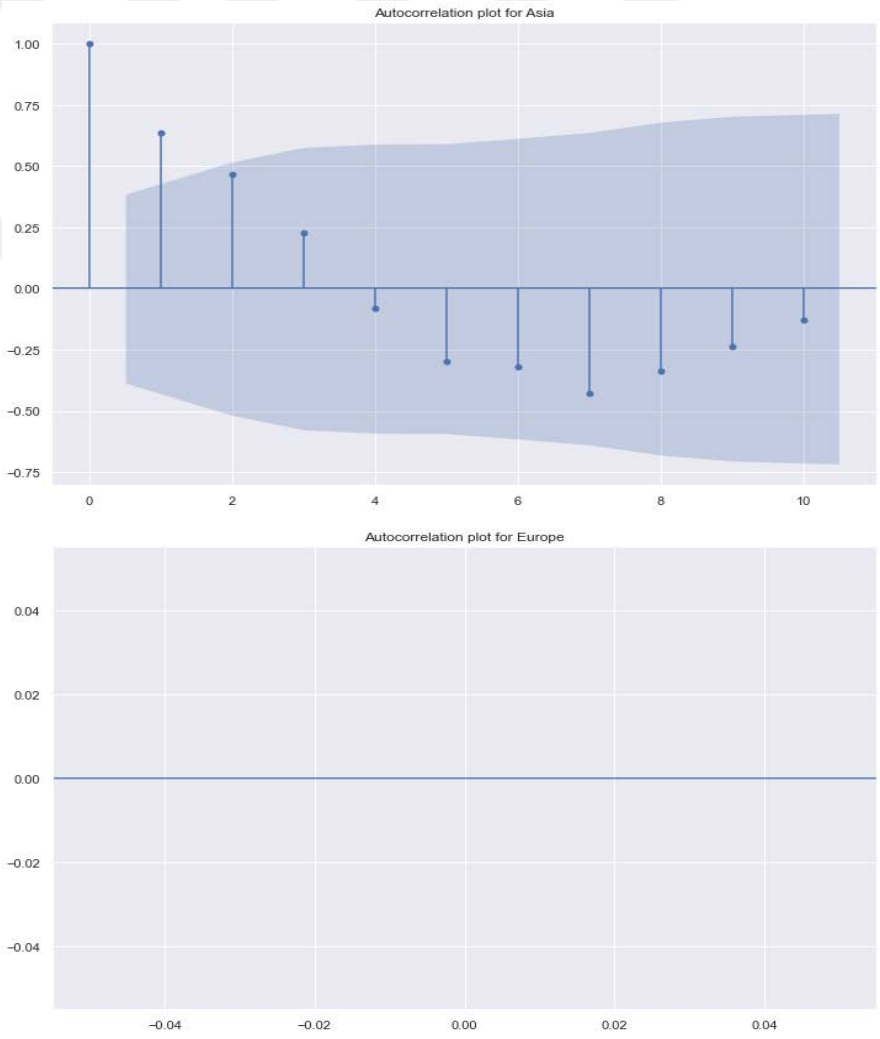


Figure 21: Autocorrelation (ACF) Plot

From the figure 21 (ACF plot) above, it is shown that there is autocorrelation in the first lag for Asia. So, a modified Mann Kendall test was voted to be applied to examine the trend in Asia. There are however no lags for Europe hence, the original Mann Kendall test was used to test for trend in Europe.

Table 6: Total Duplicate Packet Mann Kendall Trend Test Per Regions

	Metric	Asia	Europe
0	Trend	decreasing	no trend
1	h	True	False
2	P-value	0.005069	1.000000
3	Z-Score	-2.802596	0
4	Tau	-0.672727	0.000000
5	S	-37.000000	0.000000
6	Var_s	165.000000	0.000000
7	Slope	-2.133539	0.000000
8	Intercept	97.432655	31.000000

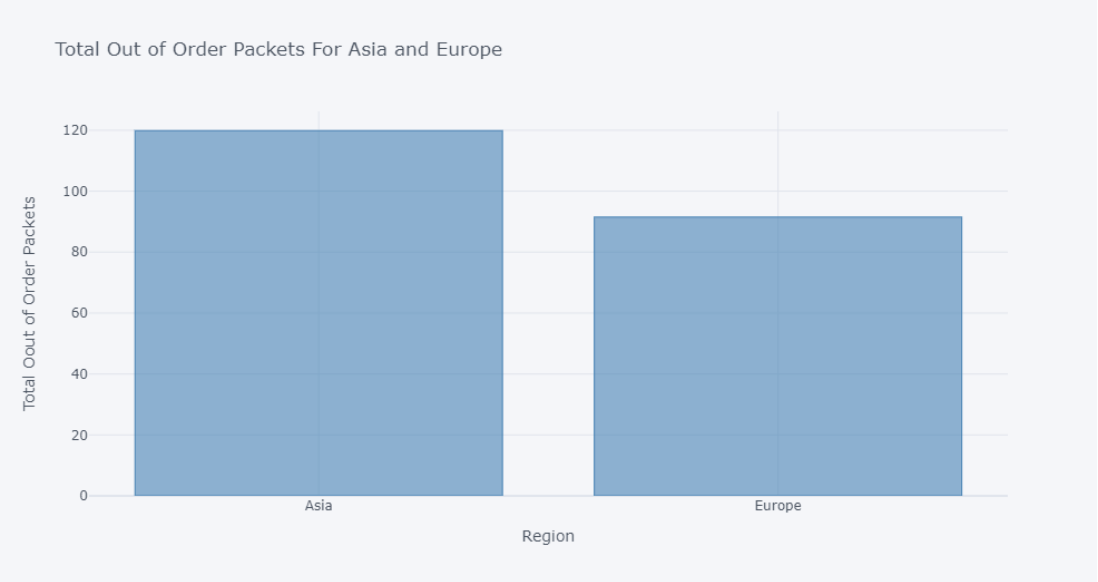


Figure 22: Out of Order Packets

From table 6 above it is evident that the trend for Total Duplicate Packet in Asia is decreasing with p = 0.005069, Z = -2.802596 while there is no statistically significant trend in Europe i.e., no trend with p = 0.005069, Z = -2.802596.

4.2.2. Out of Order Packets

Figure 22 shows the Out of Order Packets distribution while figure 23 shows the Out of Order Packets for Asia and Europe over time.

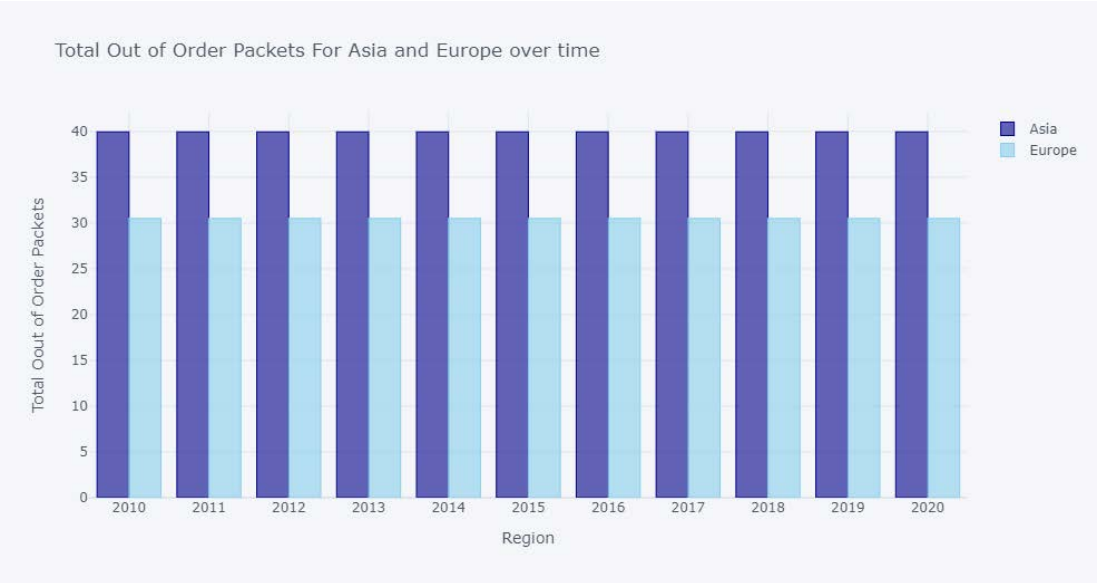


Figure 23: Out of Order Packets for Asia and Europe (2010-2020)

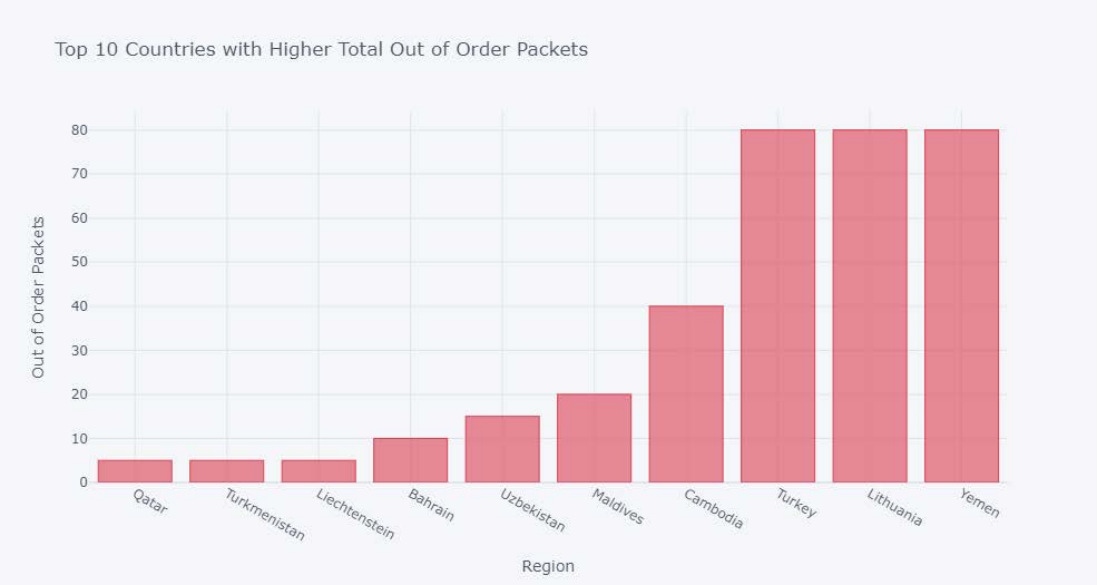


Figure 24: Top 10 Countries with High Total Out of Order Packets-Bar Plot

Examining figures 22 and 23 above it is noted that Asia has greater total Out of Order Packets scores compared to Europe throughout time. It is also noted that both Asia and Europe have a constant distribution of total Out of Order Packets scores between the years 2010 and 2020 (see figure 23). Figure 24 to figure 26 show the top 10 countries from Asia and Europe with Total Highest Out of Order packets in different visualization persentations.

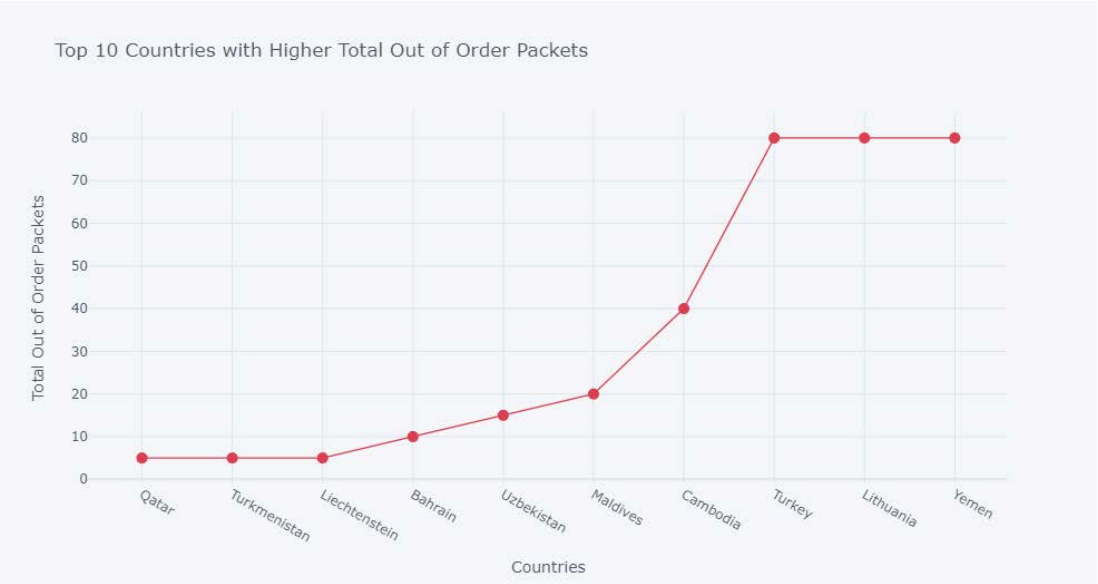


Figure 25: Top 10 Countries with High Total Out of Order Packets - Line Plot

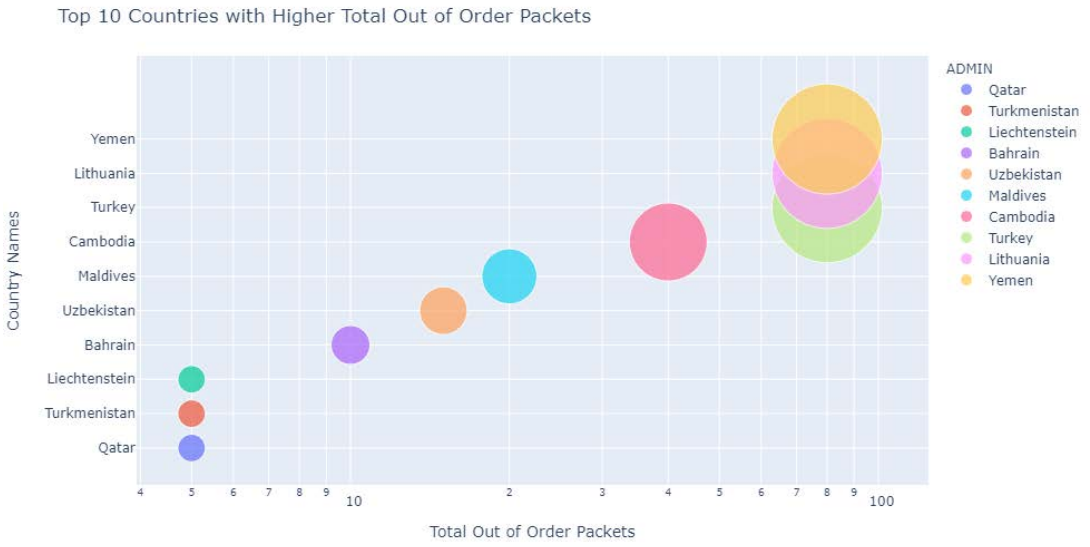


Figure 26: Top 10 Countries with High Total Out of Order Packets - Bubble Plot

The below figure 27 to figure 29 show the top 10 countries from Asia and Europe with Total lowest Out of Order packets in different visualization persentations.

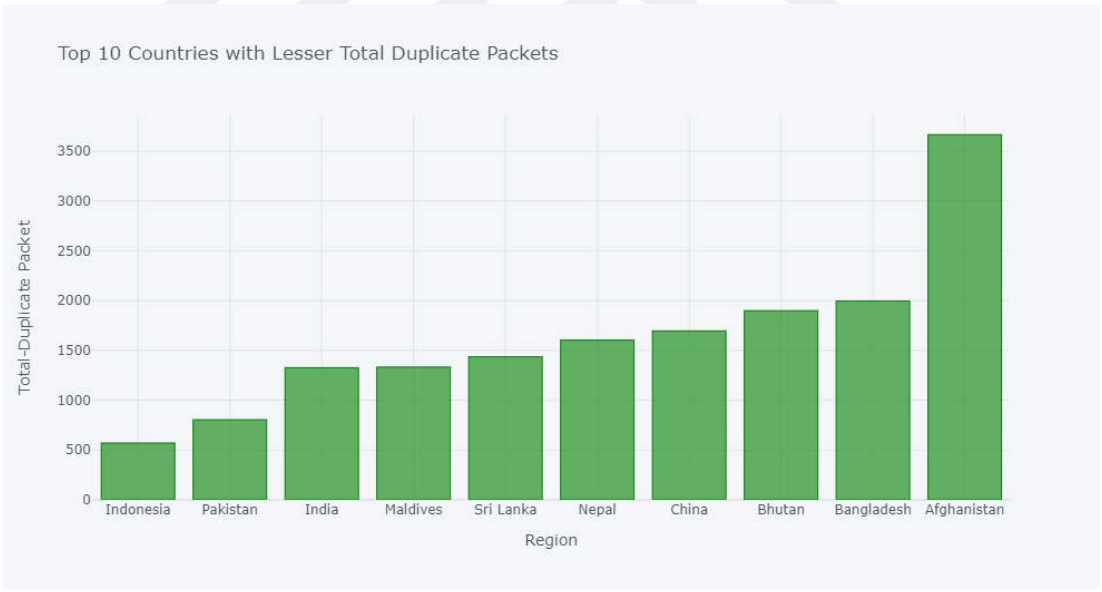


Figure 27: Top 10 Countries with Lesser Total Out of Order Packets -Bar Plot

Figure 28: Top 10 Countries with Lesser Total Out of Order Packets - Line Plot

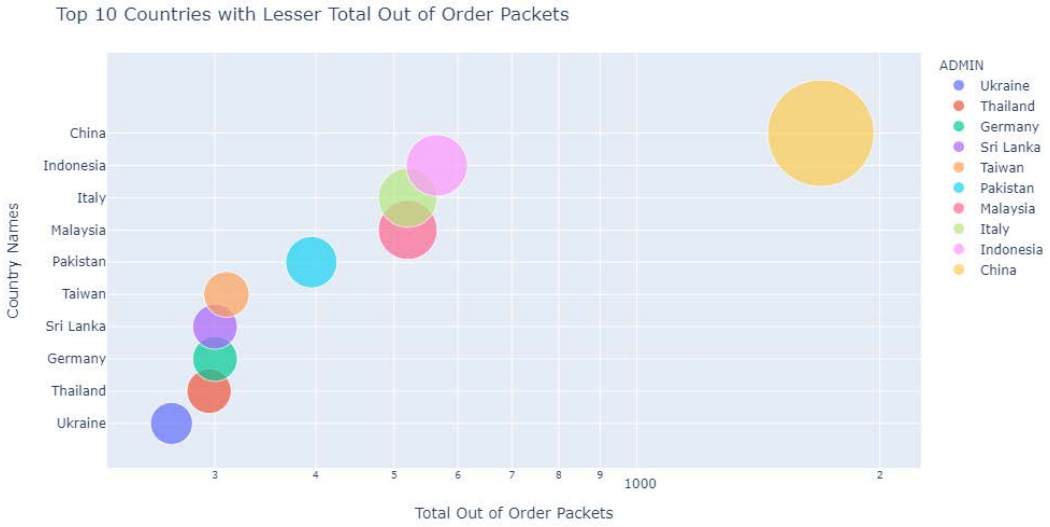


Figure 29: Top 10 Countries with Lesser Total Out of Order Packets - Bubble Plot

Trend Test

Autocorrelation Test

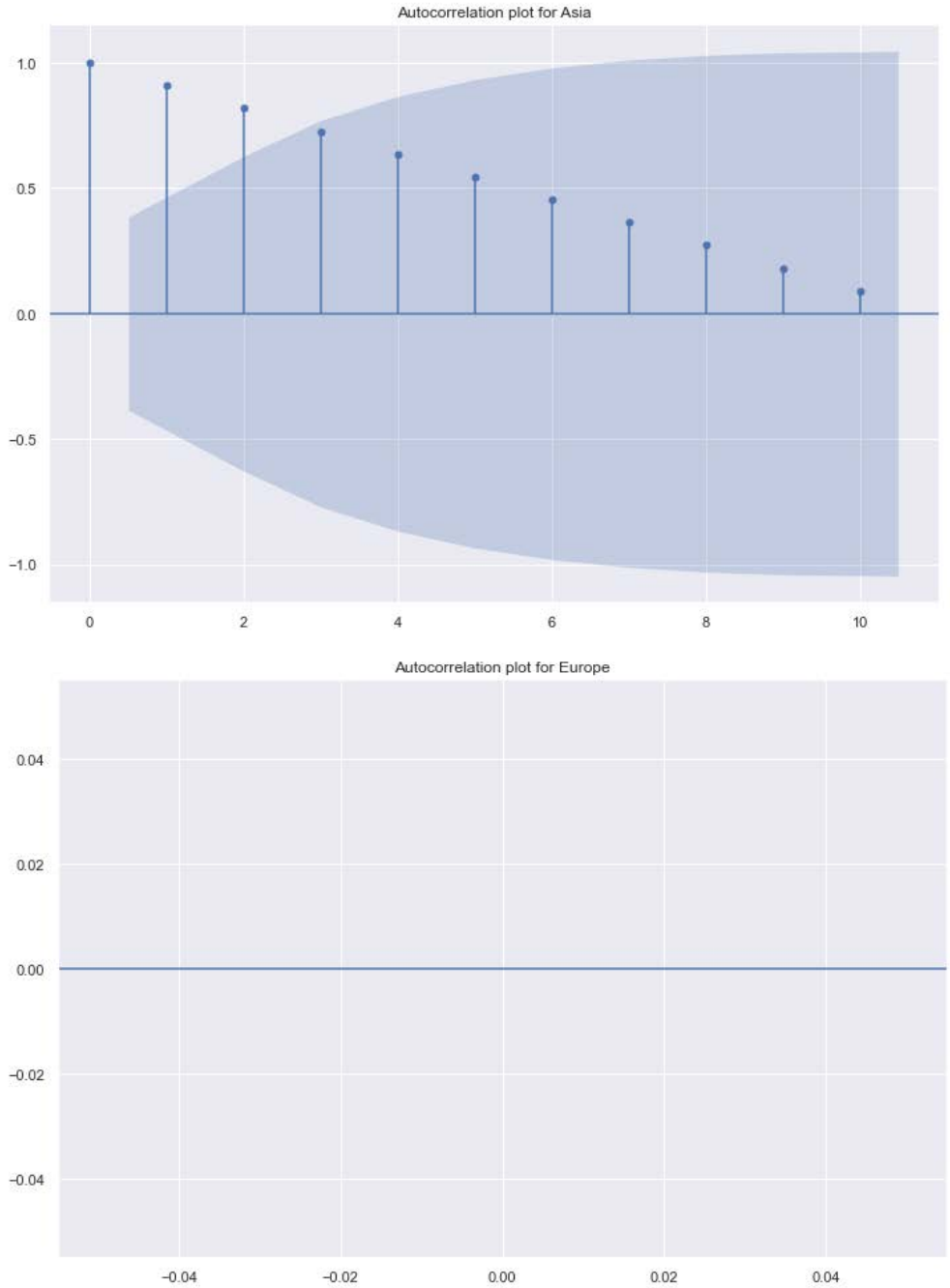


Figure 30: Autocorrelation Plot for Out of Order Packets

From the ACF plot given above, it is noted that there is autocorrelation in the first lag for Asia. As such, the modified Mann Kendall test was used to examine the trend in Asia. It is also evidenced that there is no serial correlation in Europe hence the original Mann Kendall test was used.

Mann Kendall Trend Test

Table 7 shows the Mann Kendall Trend Test for the Out of Order Packets for Asia and Europe over time.

Table 7: Out of Order Packets Mann Kendall Trend Test scores Per Regions

	Metric	Asia	Europe
0	Trend	no trend	no trend
1	h	False	False
2	P-value	1.000000	1.000000
3	Z-Score	0	0
4	Tau	0.000000	0.000000
5	S	0.000000	0.000000
6	Var_s	nan	0.000000
7	Slope	0.000000	0.000000
8	Intercept	39.955556	30.500000

From table 7 above it is shown that the trend for Out of Order Packets in both Asia and Europe is constant with p = 1.000000, Z = 0 and p = 1.000000, Z = 0 respectively.

4.2.3. TCP Throughput

Figure 31 shows the TCP Throughput distribution while figure 32 shows the TCP Throughput for Asia and Europe between the years 2010 to 2020.

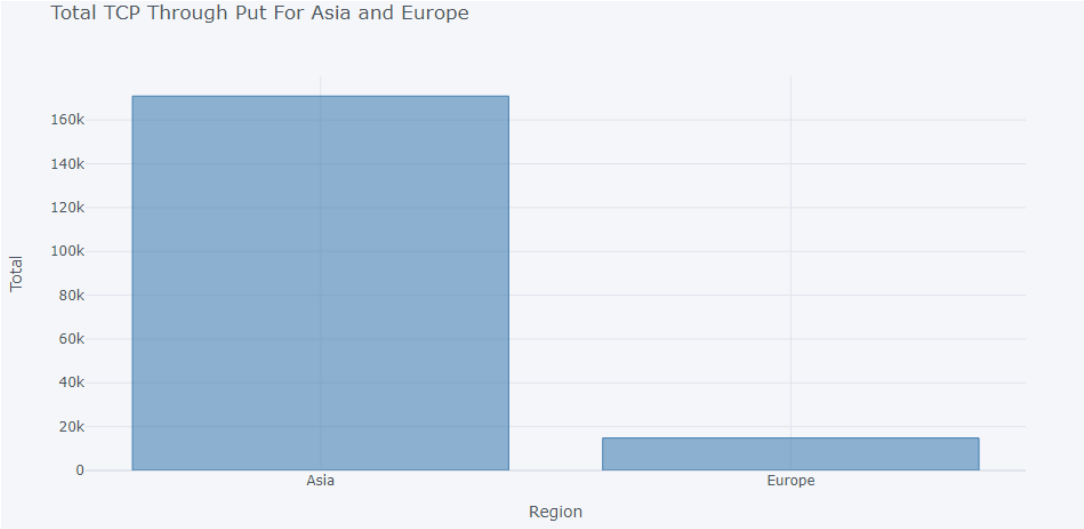


Figure 31: TCP Throughput distribution

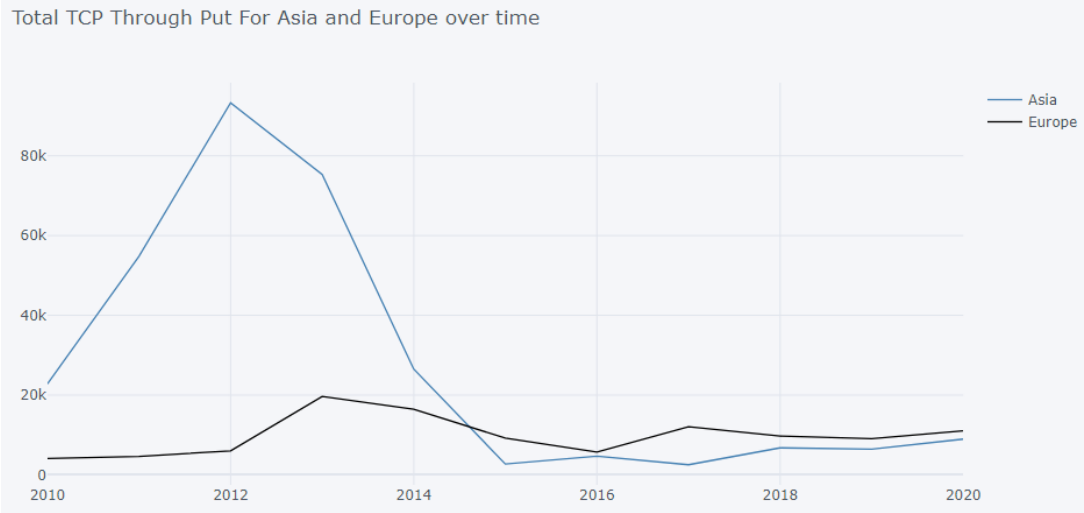


Figure 32: TCP Throughput for Asia and Europe (2010 to 2020)

From figures 31 and 32 above it is evident that Asia has greater total TCP Throughput scores relative to Europe between 2010 and 2020. It is also shown that both Asia and Europe have a varying distribution of total TCP Throughput scores between the underlying years i.e., 2010 and 2020 (see figure 32).

The below figure 33 to figure 35 show the top 10 countries from Asia and Europe with Total highest Duplicate packets in different visualization persentations.

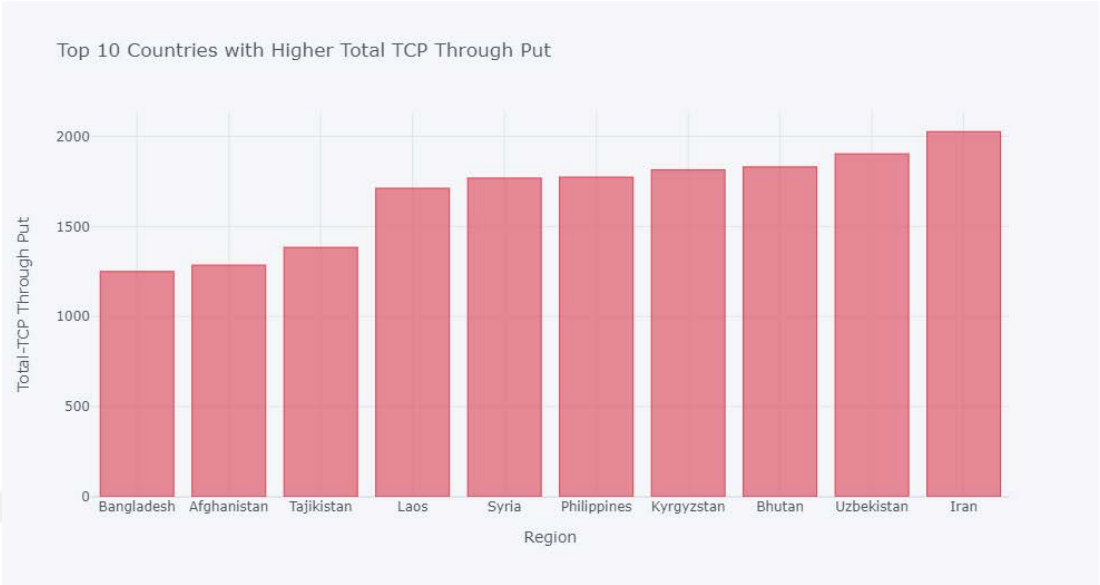


Figure 33: Top 10 Countries with High Total TCP Throughput -Bar Plot

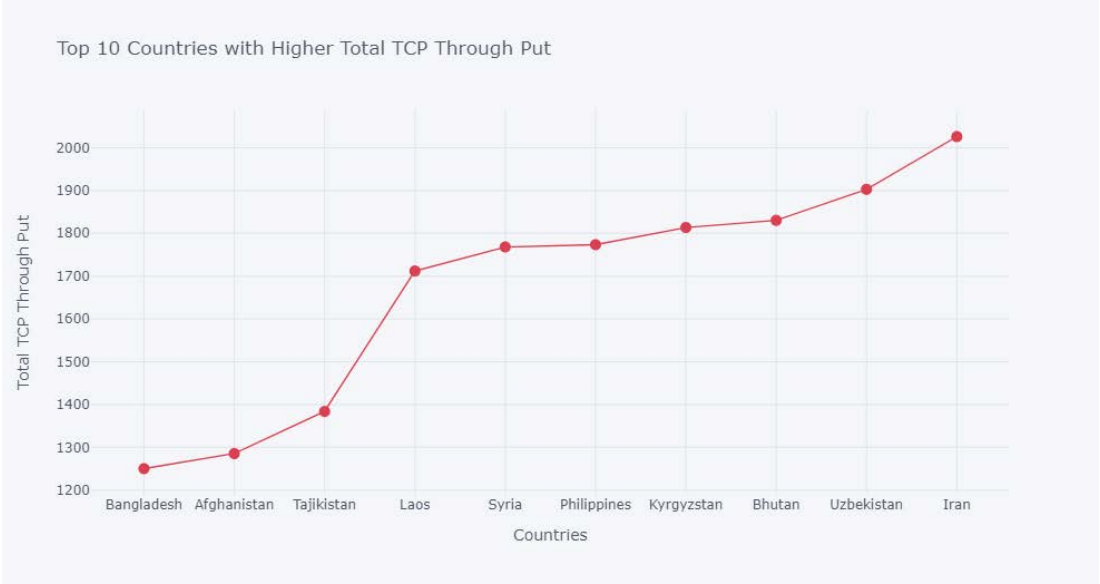


Figure 34: Top 10 Countries with High Total TCP Throughput - Line Plot

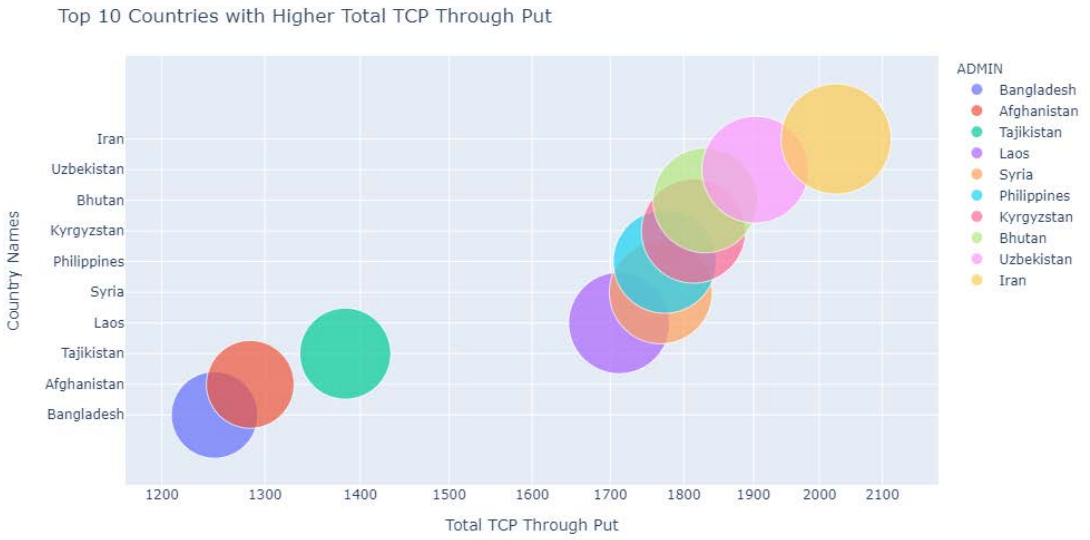


Figure 35: Top 10 Countries with High Total TCP Throughput - Bubble Plot

The below figure 36 to figure 38 show the top 10 countries from Asia and Europe with Total lowest TCP Throughput in different visualization persentations.

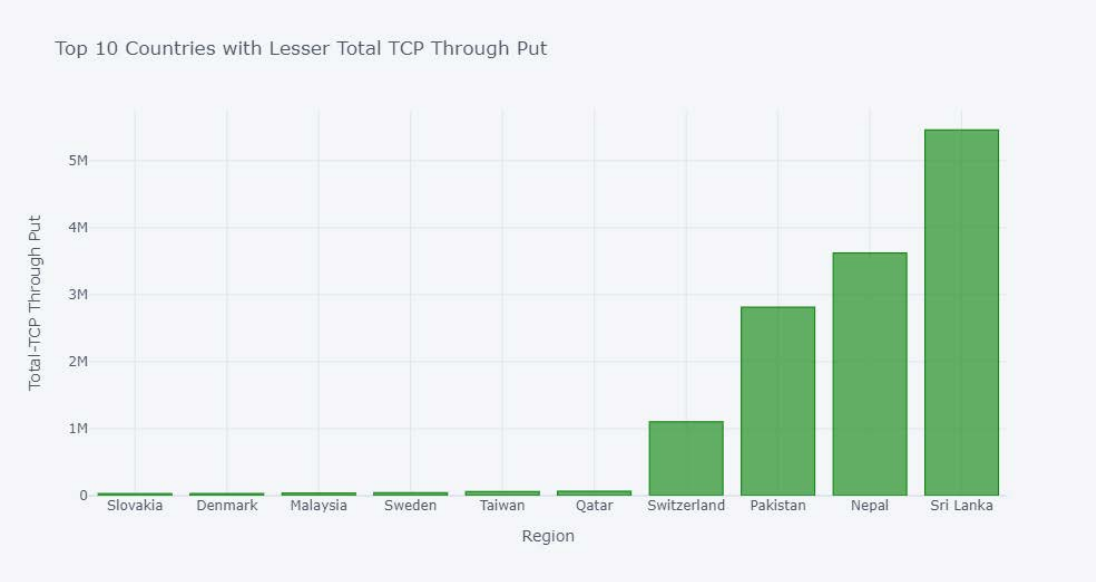


Figure 36: Top 10 Countries with Lesser Total TCP Throughput -Bar Plot

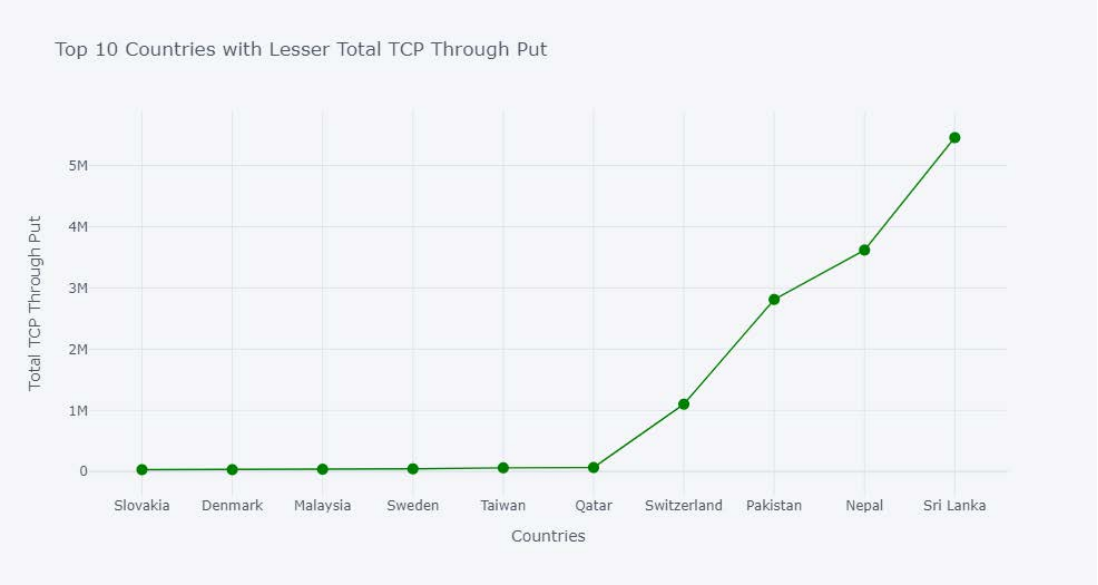


Figure 37: Top 10 Countries with Lesser Total TCP Throughput - Line Plot

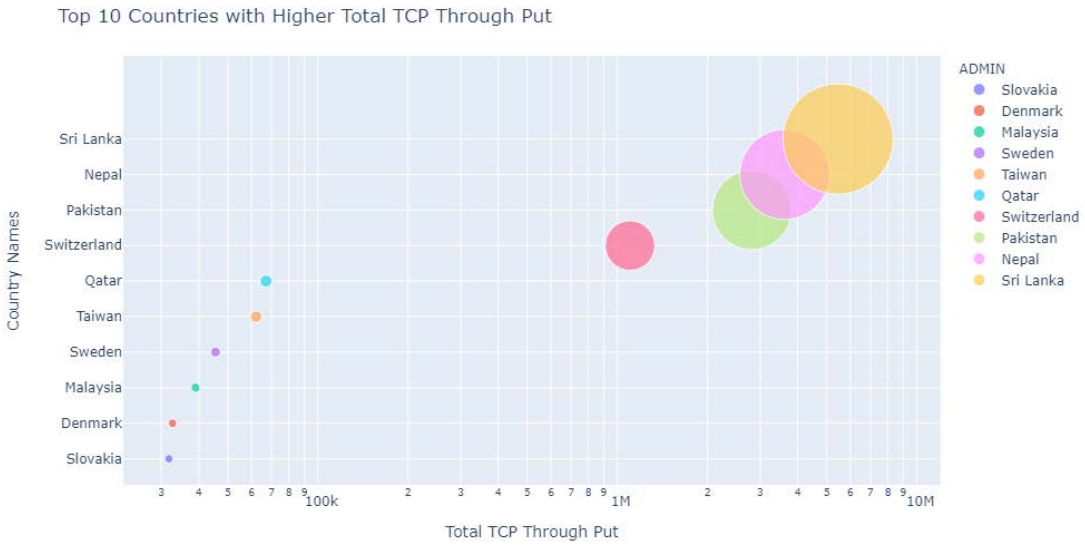


Figure 38: Top 10 Countries with Lesser Total TCP Throughput - Bubble Plot

Trend Test

Autocorrelation Test

From the ACF plot given below, it is noted that there is autocorrelation in the first lag in Asia and Europe. Therefore, the modified Mann Kendall test was used to examine the trend in both Asia and Europe.

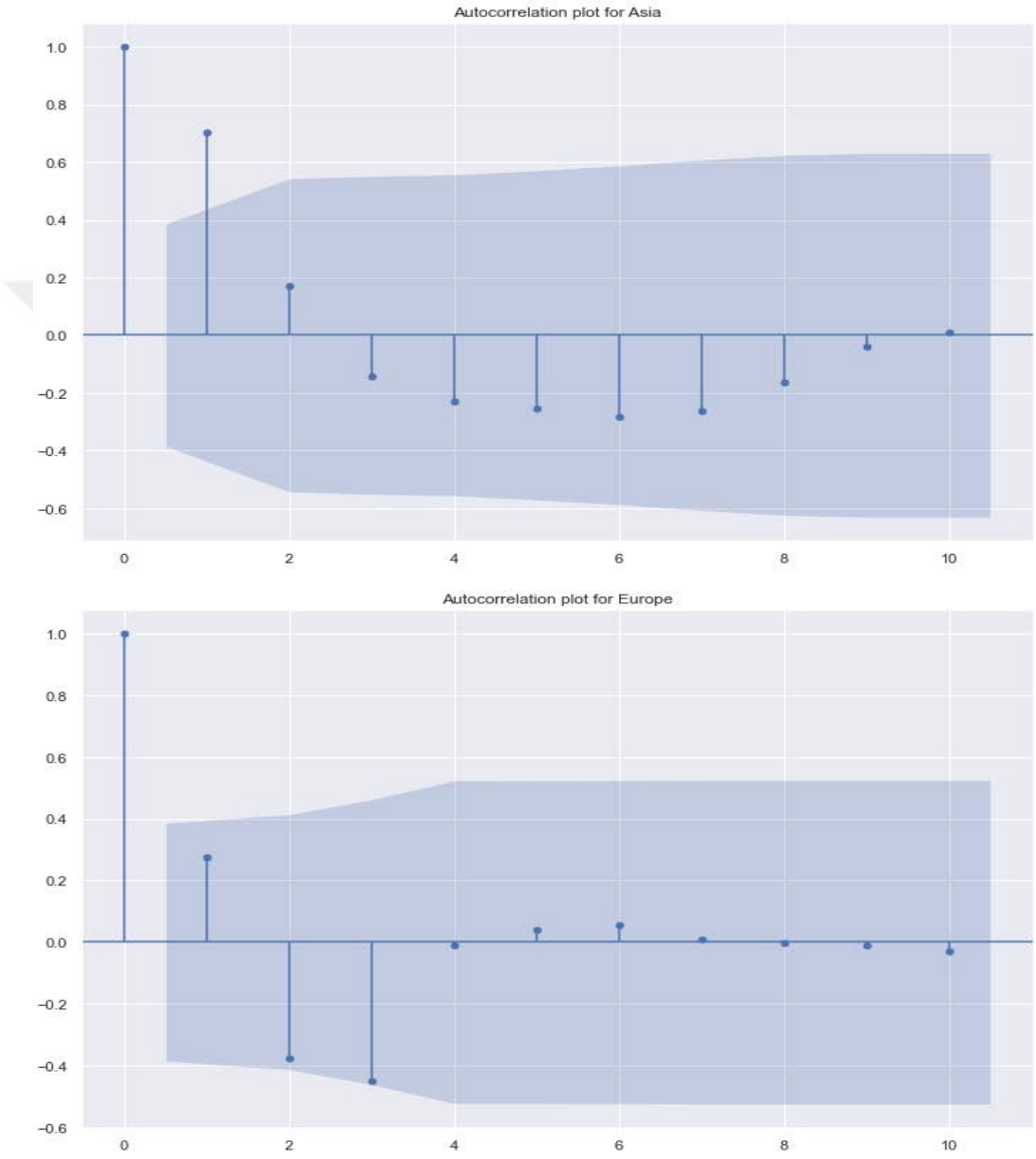


Figure 39: Autocorrelation Plot for total TCP Throughput

Mann Kendall Trend Test

Table 8 shows the Mann Kendall Trend Test for the total TCP Throughput for both Asia and Europe respectively over time.

Table 8: Mann Kendall Trend test for total TCP Throughput

	Metric	Asia	Europe
0	Trend	no trend	no trend
1	h	False	False
2	P-value	0.161125	0.275758
3	Z-Score	-1.401298	1.089899
4	Tau	-0.345455	0.272727
5	S	-19.000000	15.000000
6	Var_s	165.000000	165.000000
7	Slope	-4017.566026	556.860881
8	Intercept	29051.511489	6444.646385

At a 0.05 level of significance, it is shown in table 8 that there is no trend for TCP Throughput in both Asia and Europe is constant with p = 0.161125, Z = -1.401298 and p = 0.275758, Z = 1.089899 respectively.

4.2.4. Packet Loss

Figure 40 shows the total Packet Loss distribution while figure 41 shows the total Packet Loss for Asia and Europe over time (2010 to 2020).

From figures 40 and 41 it is shown that Asia has greater total Packet Loss scores as compared to Europe between the years 2010 and 2020. It is also shown that both

Asia and Europe have a varying distribution of total Packet Loss scores during the defined years i.e., 2010 and 2020 (see figure 41).

The below figure 42 to figure 44 show the top 10 countries from Asia and Europe with Total highest Total Packet Lost in different visualization persentations.

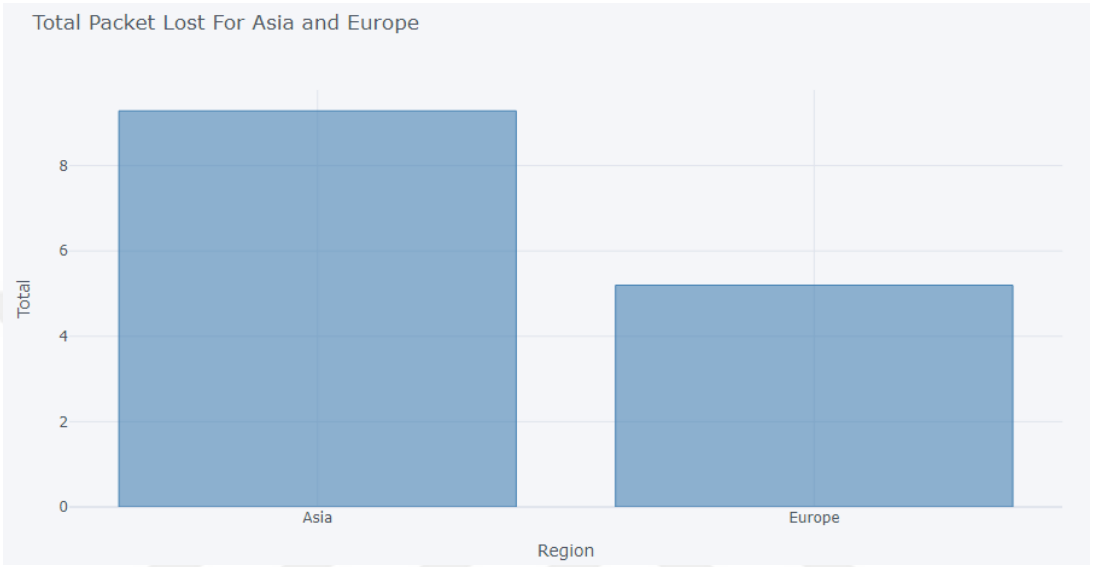


Figure 40: Packet Loss

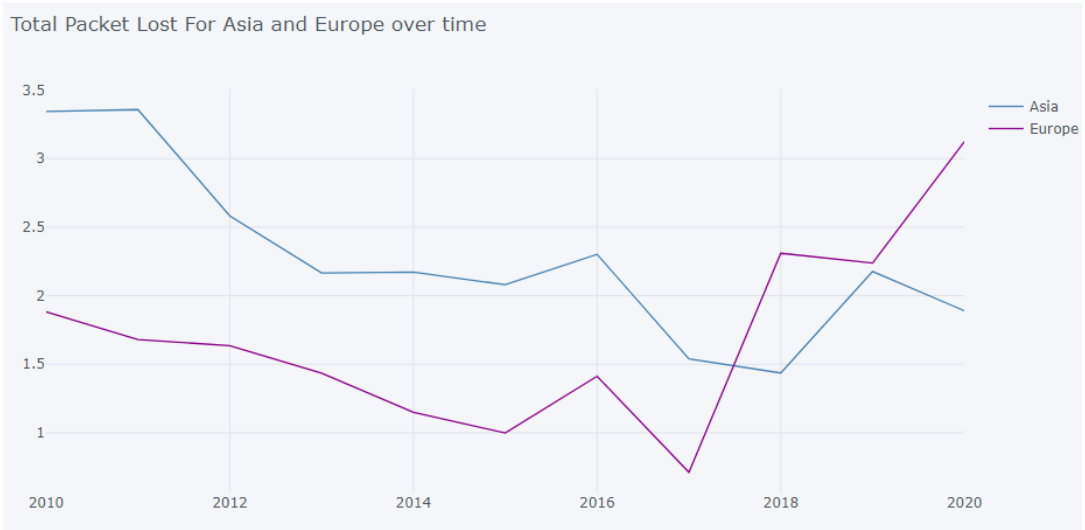


Figure 41: Total Packet Loss for Asia and Europe (2010-2020).

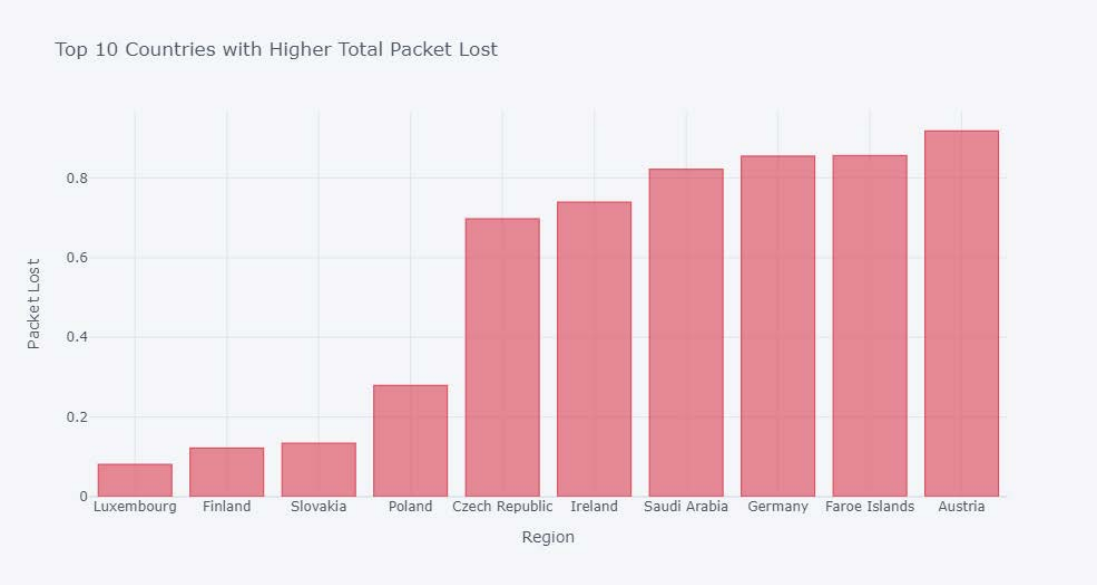


Figure 42: Top 10 Countries with High Total Packet Lost -Bar Plot

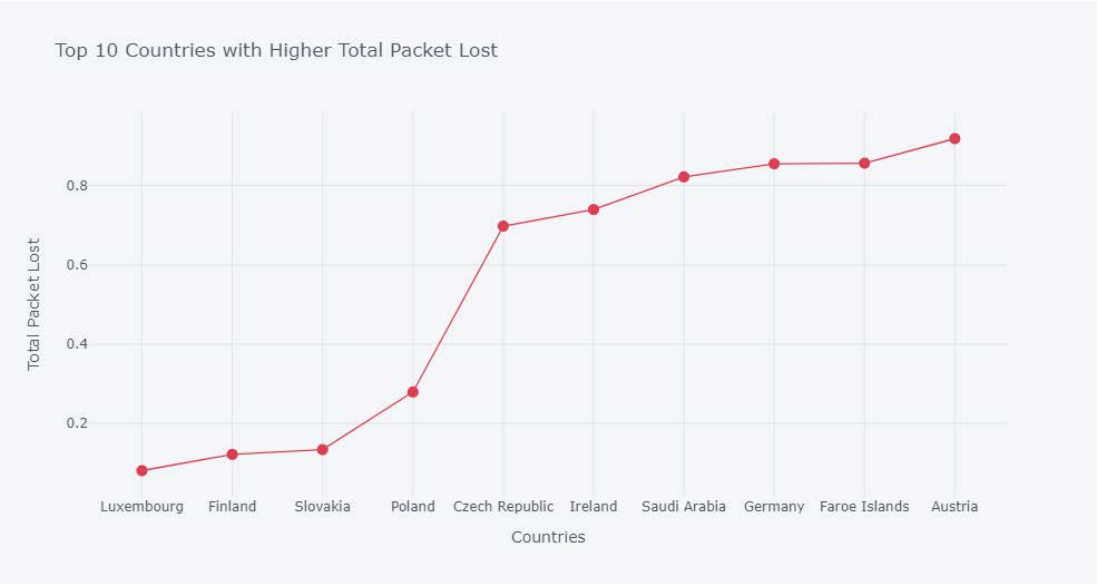


Figure 43: Top 10 Countries with High Total Total Packet Lost - Line Plot

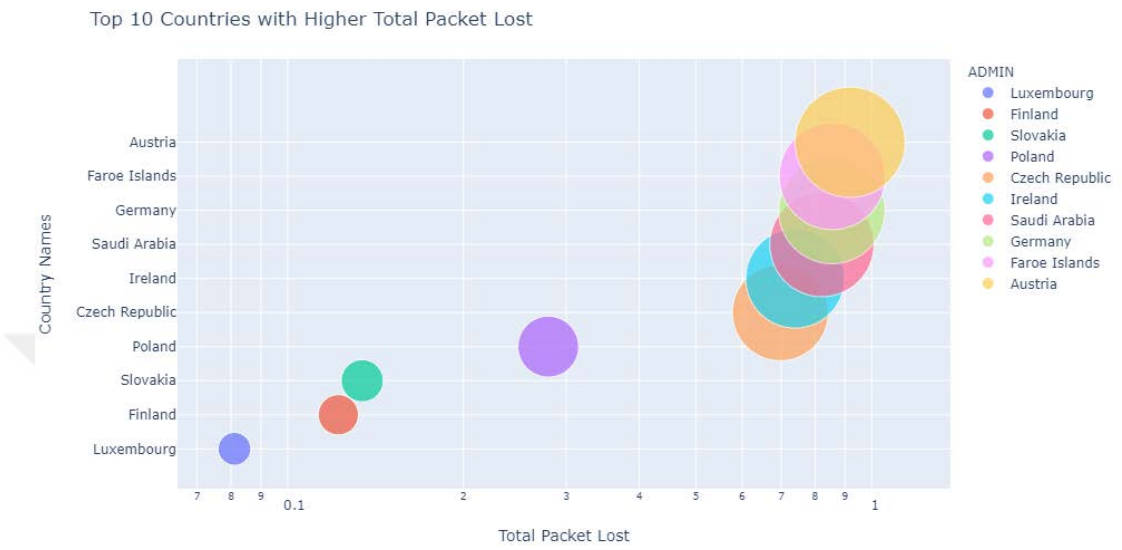


Figure 44: Top 10 Countries with High Total Total Packet Lost - Bubble Plot

The below figure 45 to figure 47 show the top 10 countries from Asia and Europe with Total lowest Total Packet Lost in different visualization persentations.

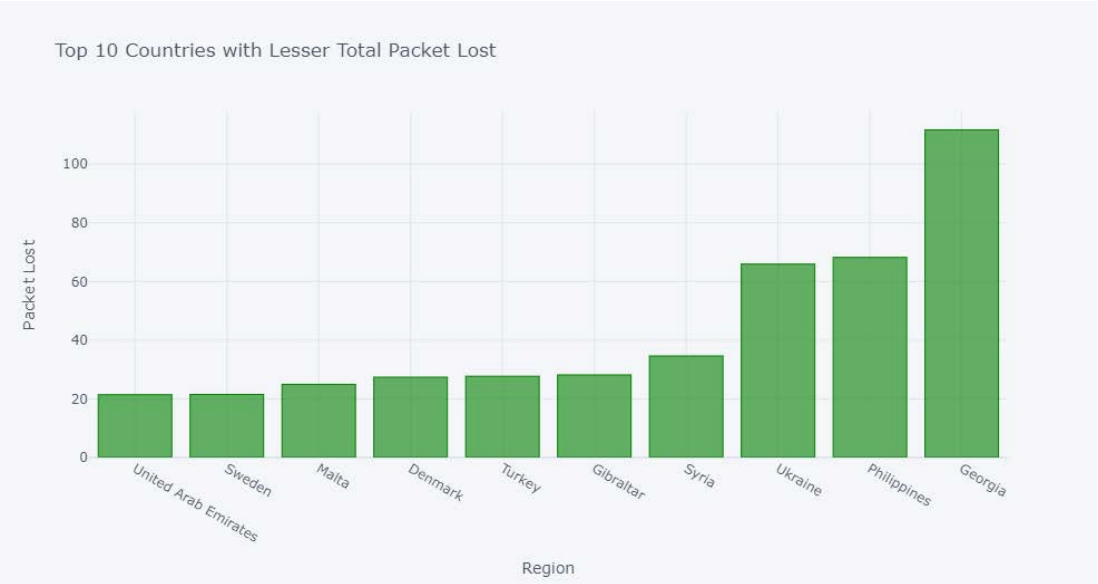


Figure 45: Top 10 Countries with Lesser Total Total Packet Lost -Bar Plot

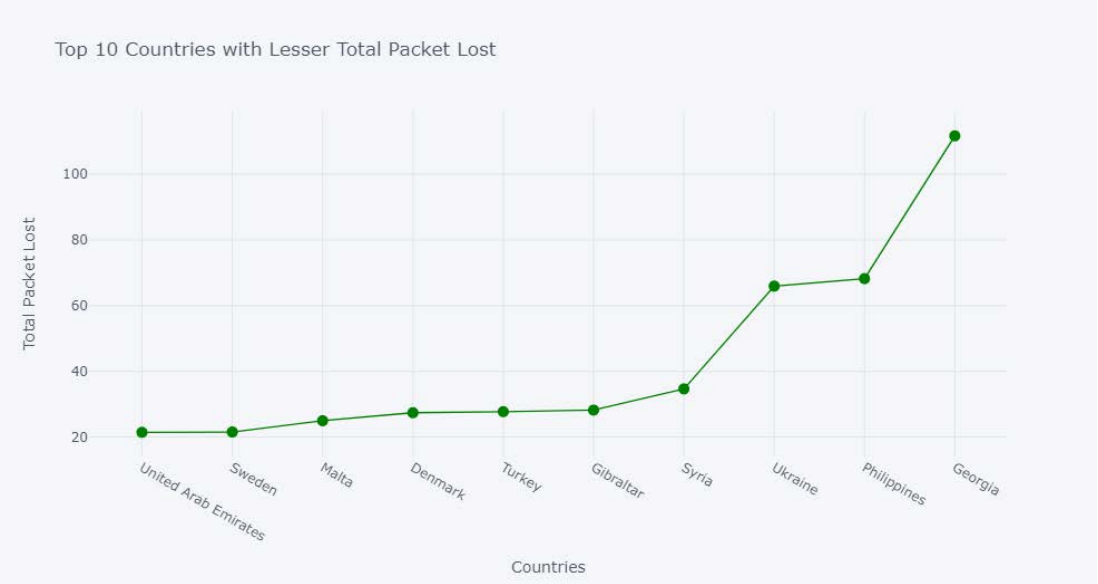


Figure 46: Top 10 Countries with Lesser Total Total Packet Lost - Line Plot

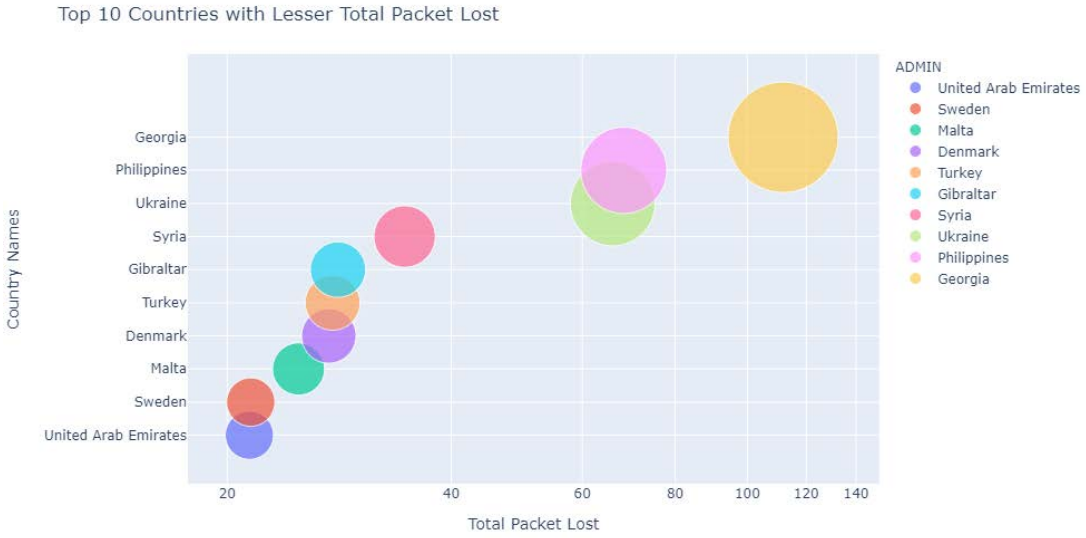


Figure 47: Top 10 Countries with Lesser Total Packet Lost - Bubble Plot

Trend Test

Autocorrelation Test

The plot below is used to test for serial correlation for the Total Packet Loss in both Europe and Asia.

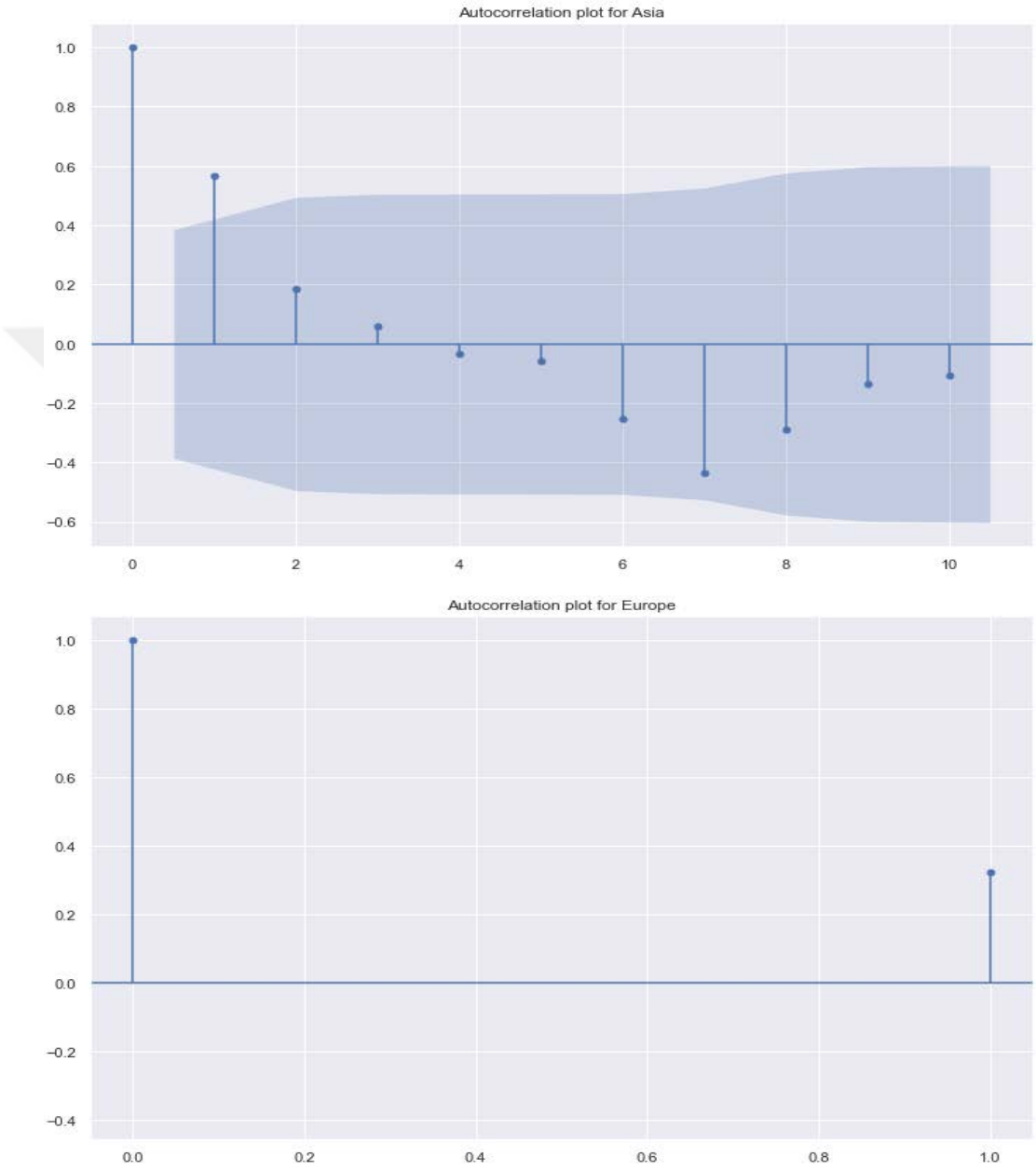


Figure 48: Autocorrelation Plot for total Packets Loss

From the ACF plot given in figure 48 above, it is noted that there is autocorrelation in the first lag for both Asia and Europe. As such, the modified Mann Kendall test was used to examine the trend in the two regions.

Mann Kendall Trend Test

Table 9 shows the Mann Kendall Trend Test for the total Packets lost for Asia and Europe between 2010 and 2020.

Table 9: Mann Kendall Trend test for total Packets lost

	Metric	Asia	Europe
0	Trend	decreasing	no trend
1	h	True	False
2	P-value	0.019517	1.000000
3	Z-Score	-2.335497	0.000000
4	Tau	-0.563636	0.018182
5	S	-31.000000	1.000000
6	Var_s	165.000000	165.000000
7	Slope	-0.156705	0.039501
8	Intercept	2.955400	1.437343

As shown in Table 9 above, at a significance level of 0.05, there is a statistically significant decreasing trend for total packets lost in Asia with p = 0.019517. Z = -2.335497 while in Europe, there is no trend with p = 1.000000, Z = 0.000000.

4.2.5. Round Trip Time

Figures 49 and 50 below show the total Round Trip Time distribution and the total Round Trip Time for Asia and Europe between the years 2010 and 2020 respectively.

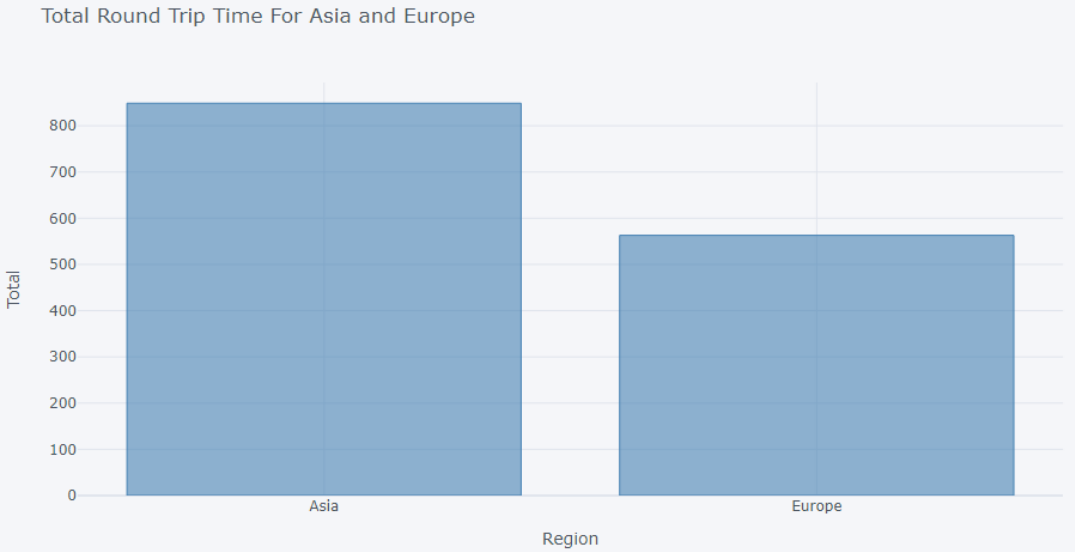


Figure 49: Total Round Trip Time distribution

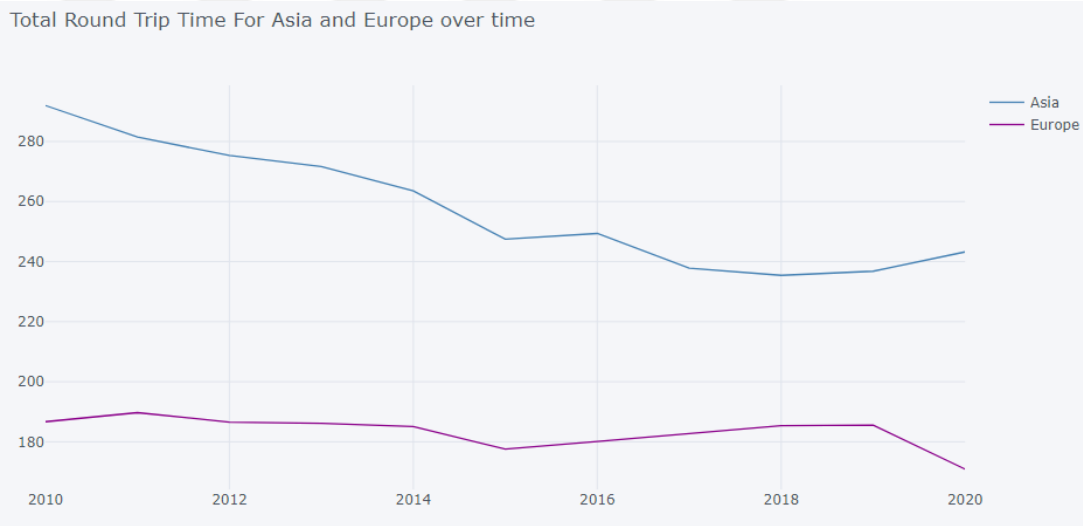


Figure 50: Total Round Trip Time for Asia and Europe (2010-2020)

It is noted from figure 50 that the total Round Trip Time for Asia and Europe is lower in 2020 than in 2010 indicating a decreasing trend despite Asia having a higher total Round Trip Time.

The below figure 51 to figure 53 show the top 10 countries from Asia and Europe with Total highest Total Round Trip Time in different visualization persentations.

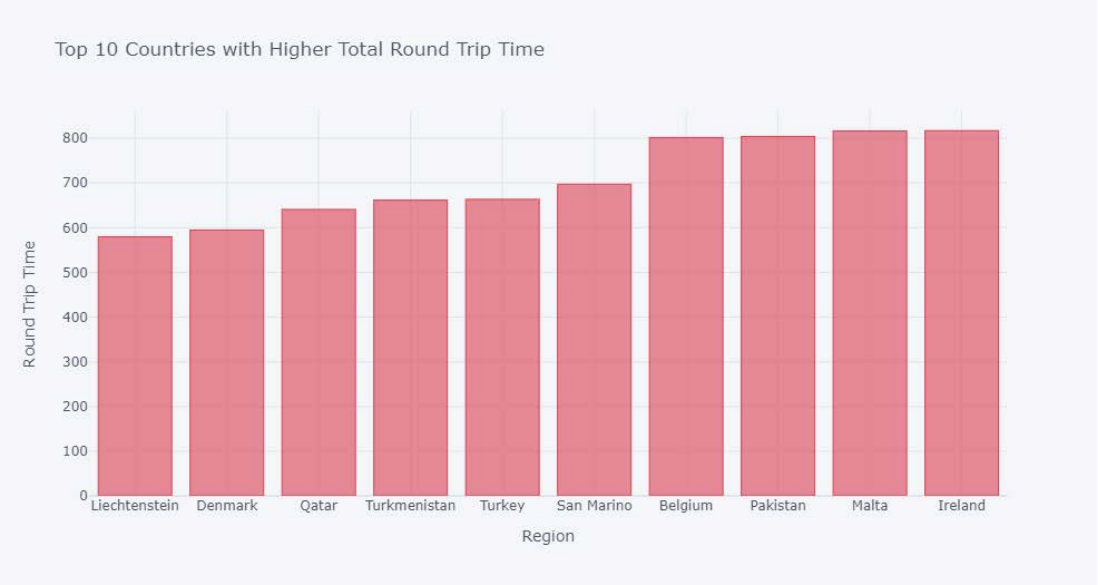


Figure 51: Top 10 Countries with High Total Round Trip Time -Bar Plot

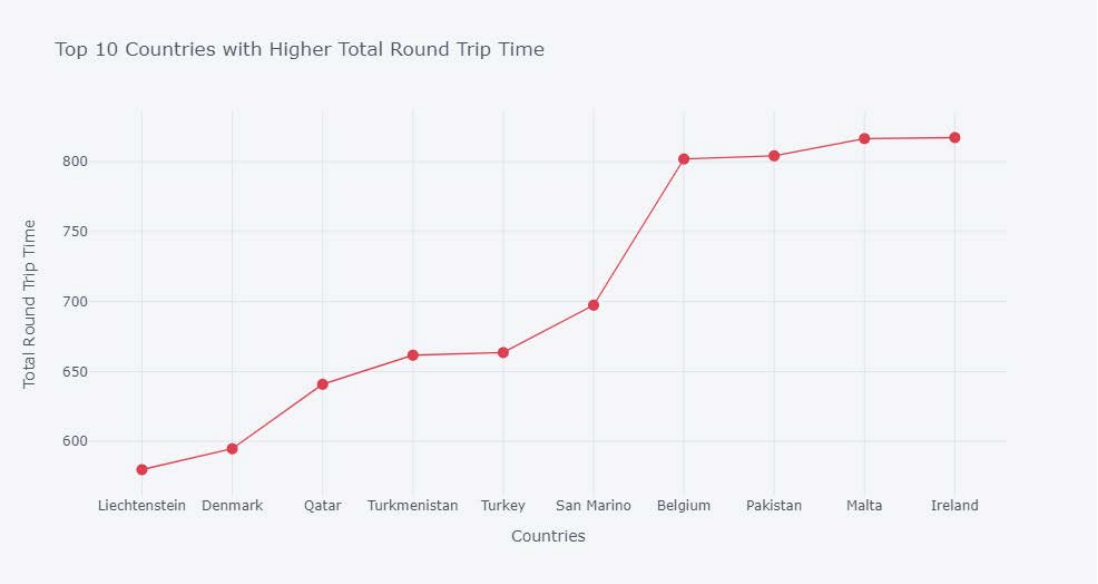


Figure 52: Top 10 Countries with High Total Total Round Trip Time - Line Plot

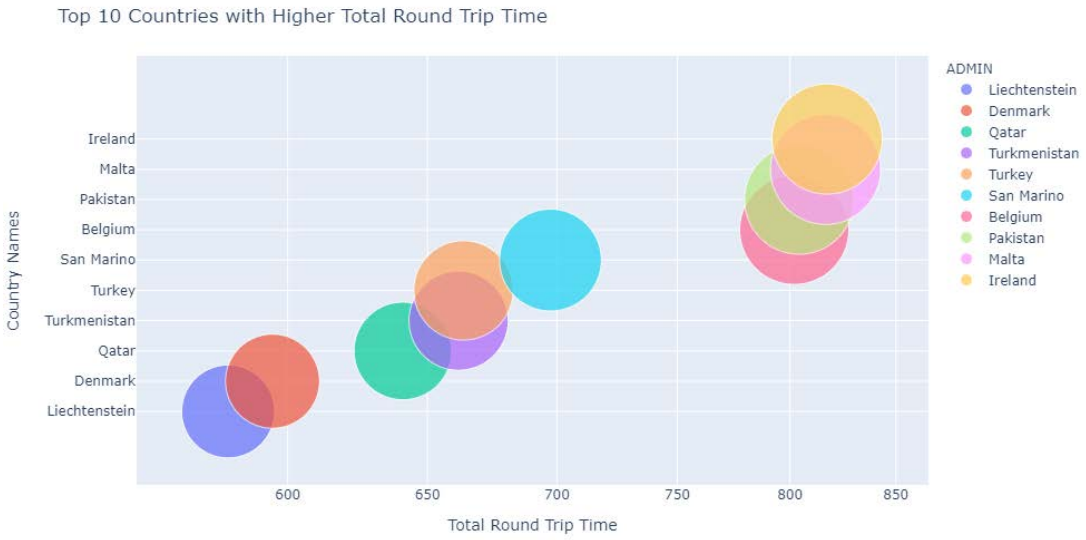


Figure 53: Top 10 Countries with High Total Total Round Trip Time - Bubble Plot

The below figure 54 to figure 56 show the top 10 countries from Asia and Europe with Total lowest Total Round Trip Time in different visualization persentations.

Figure 54: Top 10 Countries with Lesser Total Total Round Trip Time -Bar Plot

Figure 55: Top 10 Countries with Lesser Total Total Round Trip Time - Line Plot

Figure 56: Top 10 Countries with Lesser Total Packet Lost - Bubble Plot

Trend Test

Autocorrelation Test

Figure 57 below is used to show the test for serial correlation for the Total Round Trip Time in both European and Asian regions.

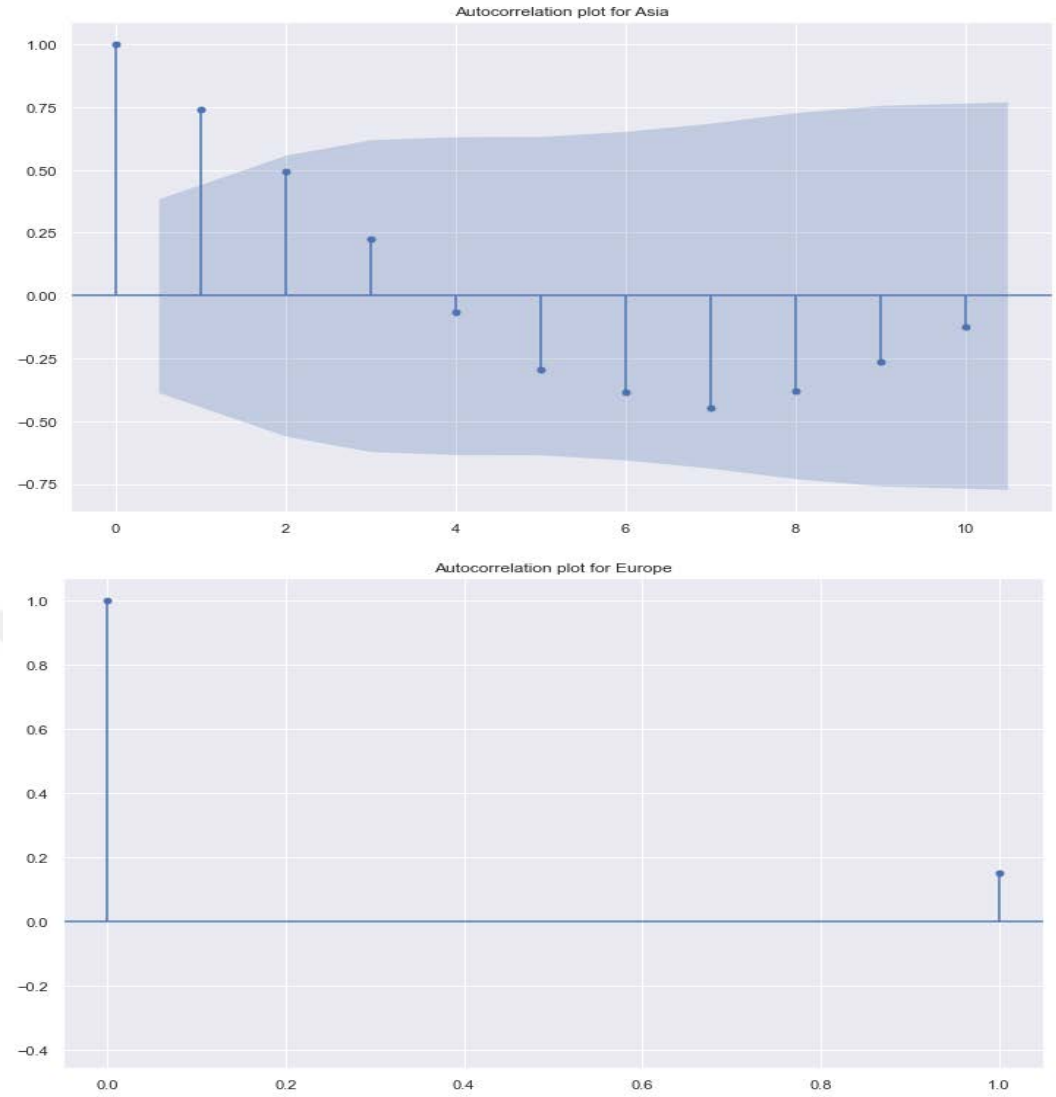


Figure 57: Autocorrelation Plot for the Total Round Trip Time

Examining figure 57 above, it is evident that there is autocorrelation in the first lag for both Asia and Europe. As a result, the modified Mann Kendall test was used to examine the trend in the two regions between 2010 and 2020. The analysis outcome is given in table 10 below.

Table 10: Autocorrelation Plot for the Total Round Trip Time

	Metric	Asia	Europe
0	Trend	decreasing	decreasing
1	h	True	True
2	P-value	0.000614	0.029273
3	Z-Score	-3.425395	-2.179797
4	Tau	-0.818182	-0.527273
5	S	-45.000000	-29.000000
6	Var_s	165.000000	165.000000
7	Slope	-6.197472	-0.741585
8	Intercept	280.360167	189.086046

The results in Table 10 above, show that at a significance level of 0.05, there is a statistically significant decreasing trend for the Total Round Trip Time in Asia with p = 0.000614. Z = -3.425395 as well as in Europe, with p = 0.029273, Z = -2.179797.

4.3.Country Geographical Distribution

4.3.1. Round Trip Time

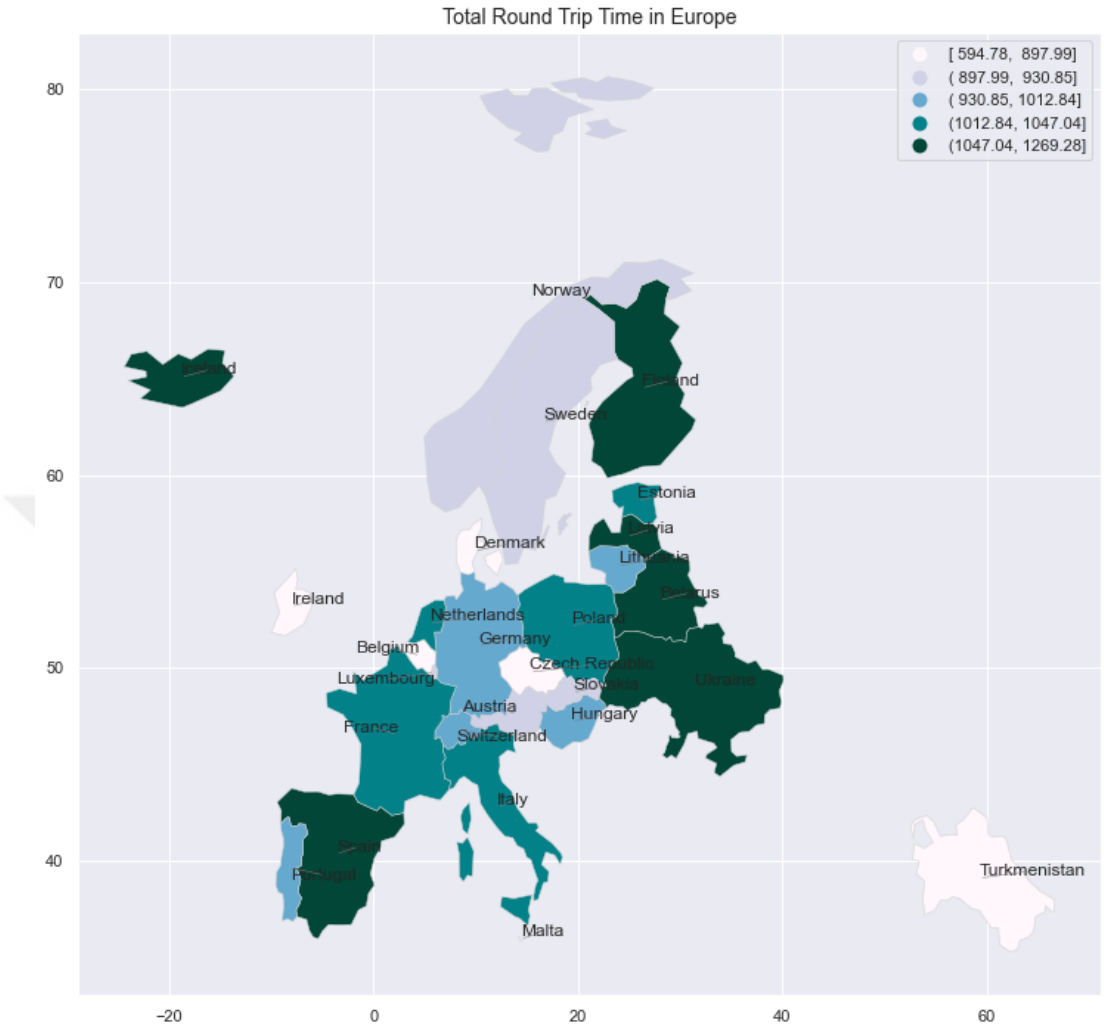


Figure 58: Round Trip Time in Europe

Countries from Europe with the highest Total Round Trip include Belarus, Spain, Finland, Iceland, Latvia, Netherlands, and Ukraine (see figure 58) while those in Asia as shown in figure 59 include Mongolia, Iran, Afghanistan, and Yemen. On the other hand, Belgium, Denmark, Gibraltar, Ireland, Liechtenstein, Malta, San Marino, and Turkmenistan have the lowest Round Trip Time in Europe (see figure 58) while Pakistani, Turkey, Georgia, United Arab Emirates, Cambodia, Malaysia, and Brunei have the lowest Total Round Trip in Asia (see figure 59).

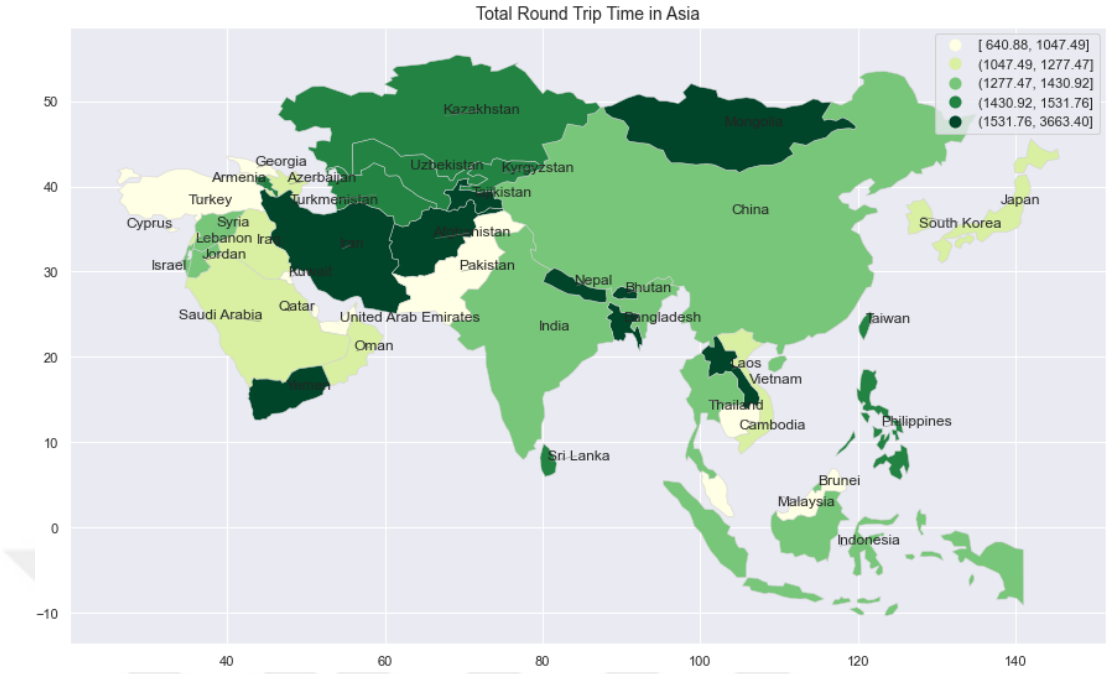


Figure 59: Round Trip Time in Asia

4.3.2. Total Duplicate Packets

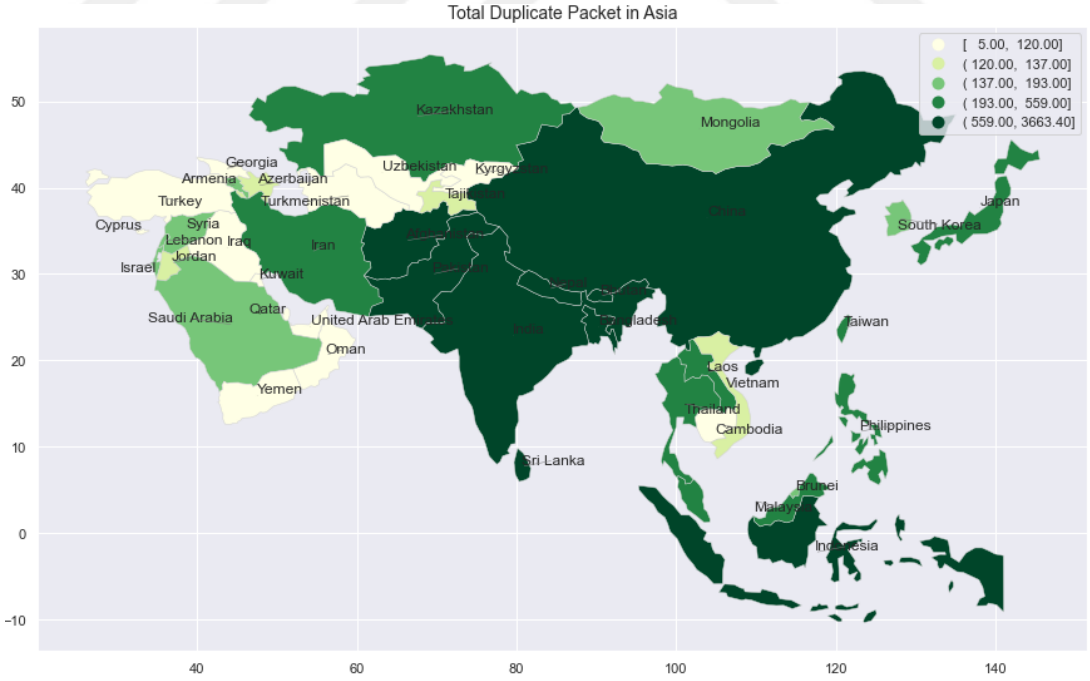


Figure 60: Duplicate packets in Asia

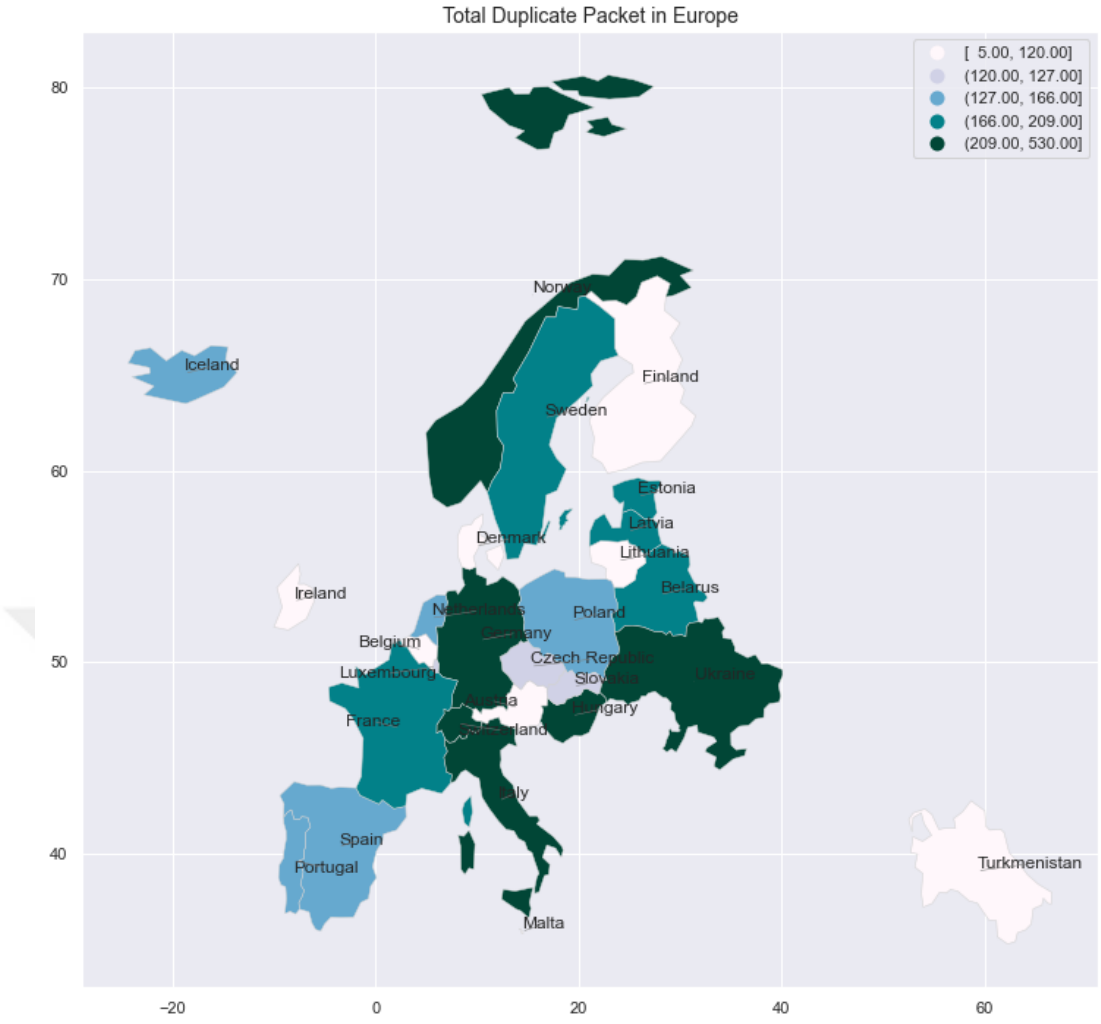


Figure 61: Duplicate packets in Europe

Countries from Europe with the highest Duplicate packets include Switzerland, Germany, Estonia, Hungary, Italy, Norway, and Ukraine (see figure 61) whereas countries in Asia with the highest Duplicate packets as shown in figure 60 include Afghanistan, Bangladesh, Bhutan, China, India, Sri Lanka, Maldives, Nepal, and Pakistan. On the other hand, Andorra, Denmark, Finland, Faroe Islands, Liechtenstein, Lithuania, and Turkmenistan have the lowest Duplicate packets in Europe (see figure 61) while in Asia, Bahrain, Cambodia, Kuwait, Oman, Qatar, Turkmenistan, Turkey, Uzbekistan, and Yemen have the lowest Duplicate packets (see figure 60).

4.3.3. Total Packets Lost

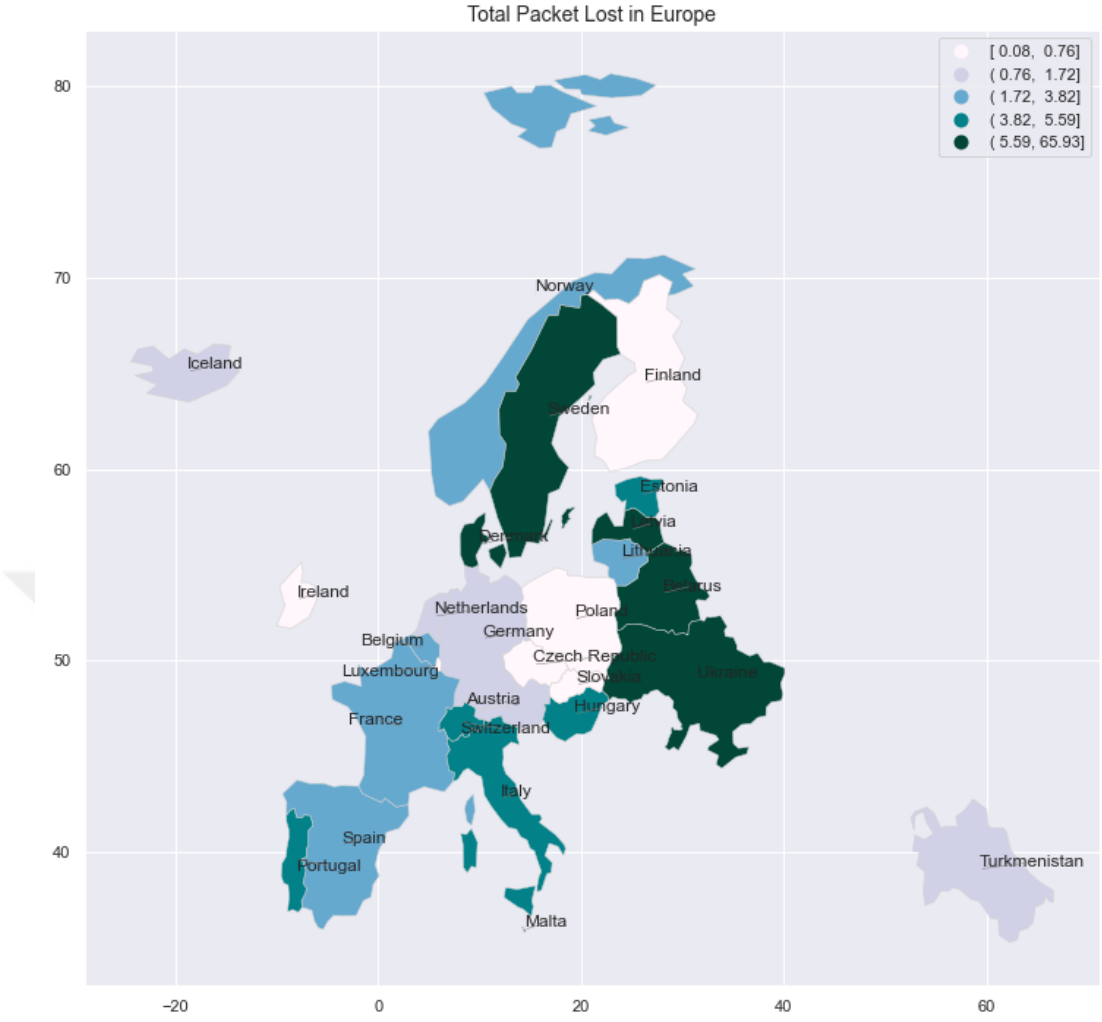


Figure 62: Total Packets Lost in Europe

Examining figures 62 and 63, it is noted that Belarus, Denmark, Gibraltar, Malta, Sweden, San Marino, and Ukraine have the highest packet loss in Europe While Czech Republic, Germany, Finland, Faroe Islands, Ireland, Luxembourg, Poland, and Slovakia have the lowest packet loss in Europe (see figure 62). In Asia, United Arab Emirates, Cyprus, Georgia, Iraq, Kuwait, Laos, Philippines, Syria, and Turkey have the highest packet loss in Asia while Afghanistan, Bahrain, Israel, India, Jordan, Japan, Maldives, Malaysia, Qatar, Saudi Arabia, and Turkmenistan have the lowest packet loss in Asia respectively (see figure 63).

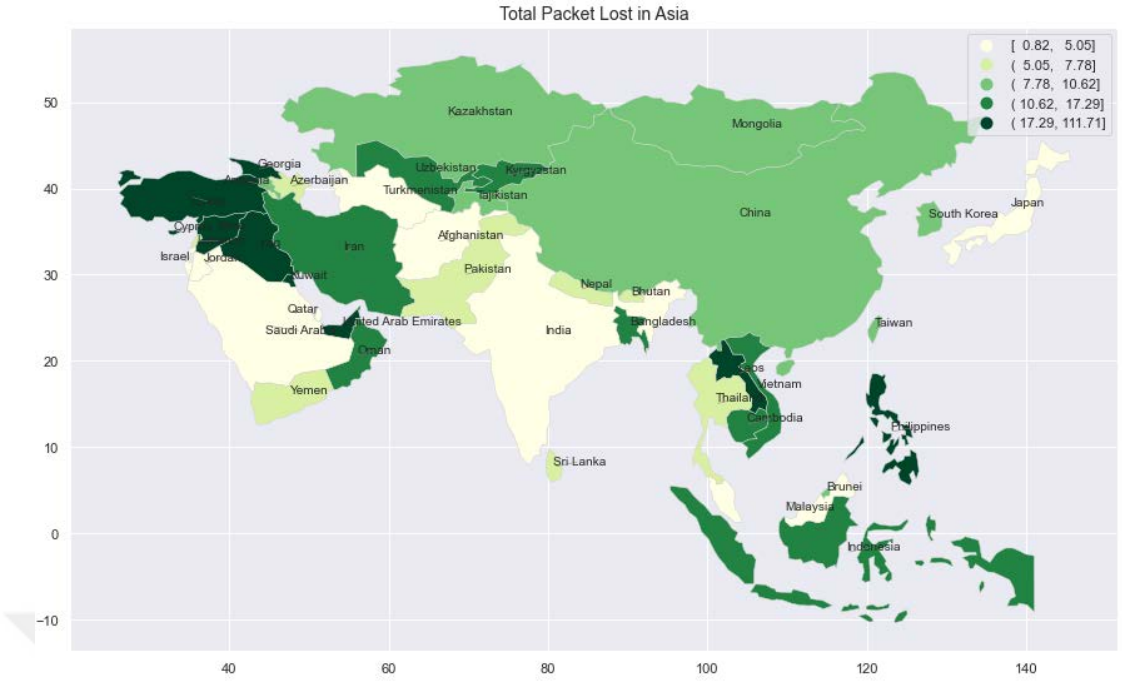


Figure 63: Total Packet Loss in Asia

4.3.4. Out of Order Packets

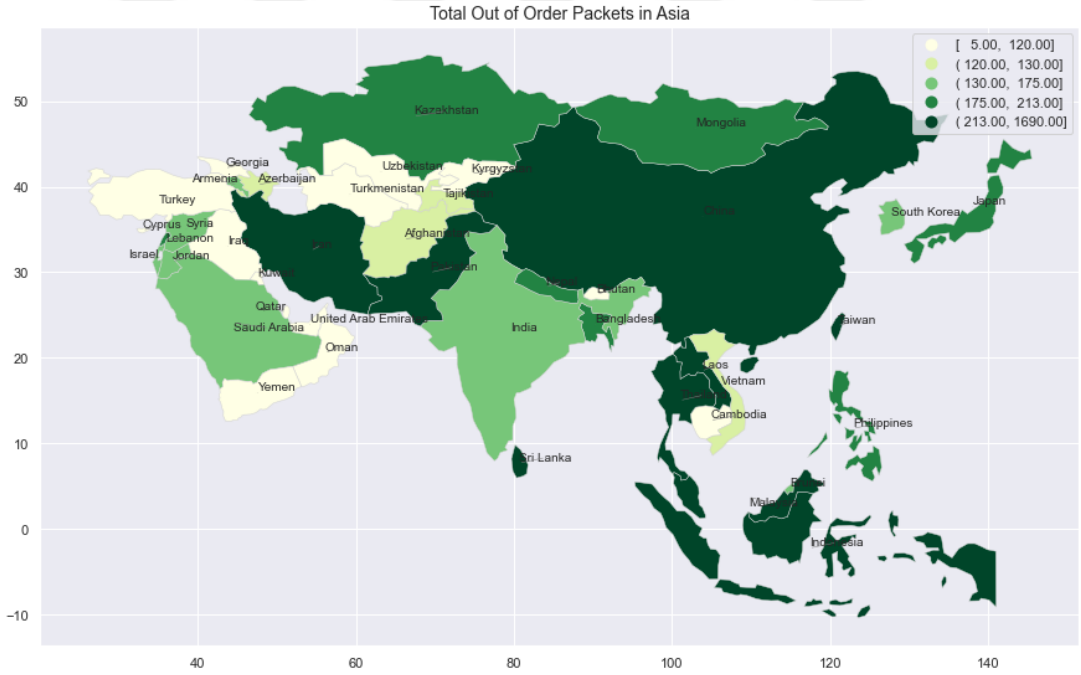


Figure 64: Total Out of Order Packets in Asia

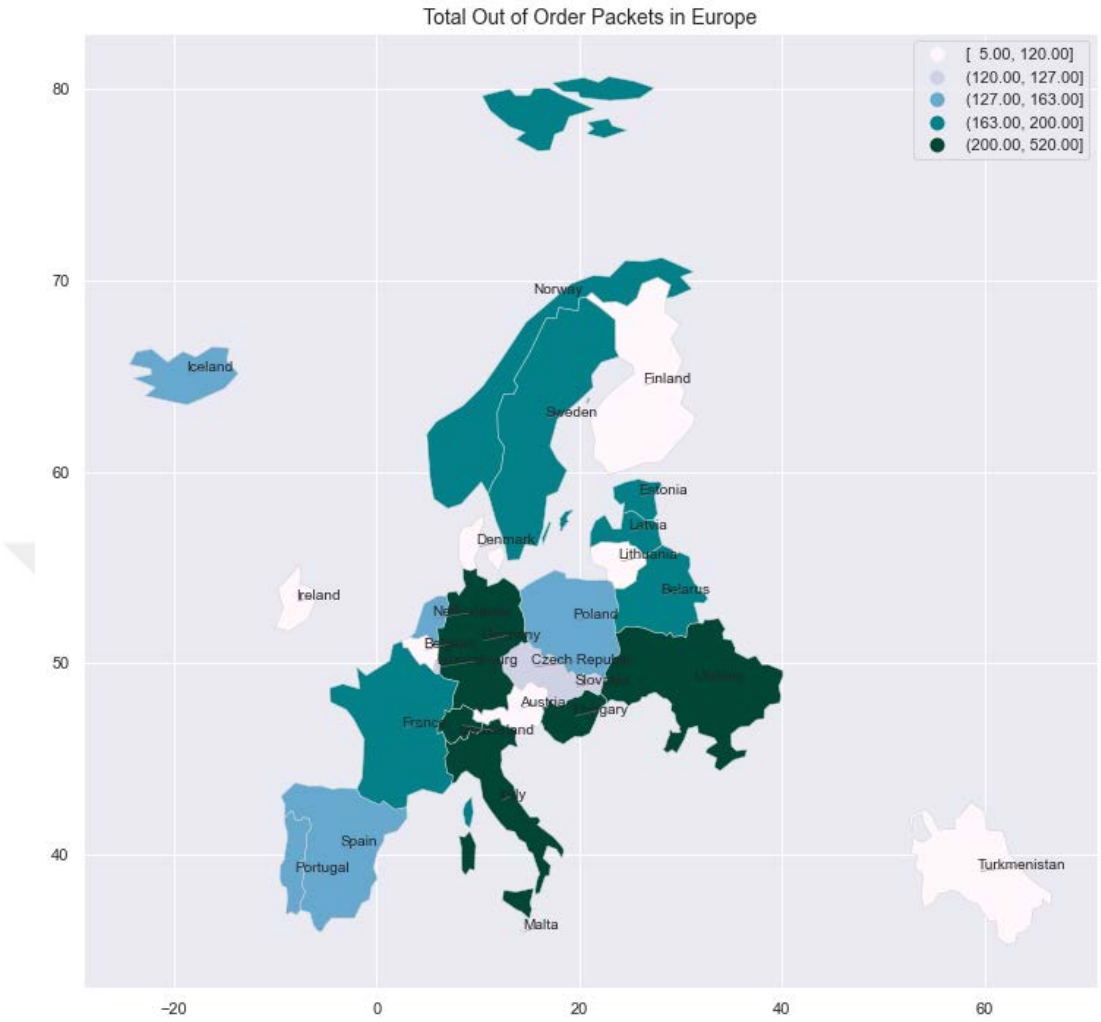


Figure 65: Total Out of Order Packets in Europe

Countries such as Malta, Ukraine, Germany, and Italy from Europe have the highest total out-of-order packets while Finland, Turkmenistan, Ireland, Austria, and Lithuania have the lowest total out-of-order packets (see figure 65) while some of those in Asia as shown in figure 64 include China, Indonesia, Iran, Laos, Sri Lanka, Malaysia, Pakistan, Thailand, Taiwan. While Bahrain, Cambodia, Kuwait, Maldives, Oman, Qatar, Turkmenistan, Turkey, Uzbekistan, and Yemen have the lowest Total Round Trip in Asia (see figure 64).

4.3.5. Total TCP Throughput

Countries such as Switzerland, Denmark, France, Luxembourg, Sweden, Slovakia, and San Marino have the highest Total TCP Throughput in Europe while Belgium,

Belarus, Spain, Lithuania, Latvia, Norway, Portugal, and Ukraine have the lowest Throughput (see figure 66).

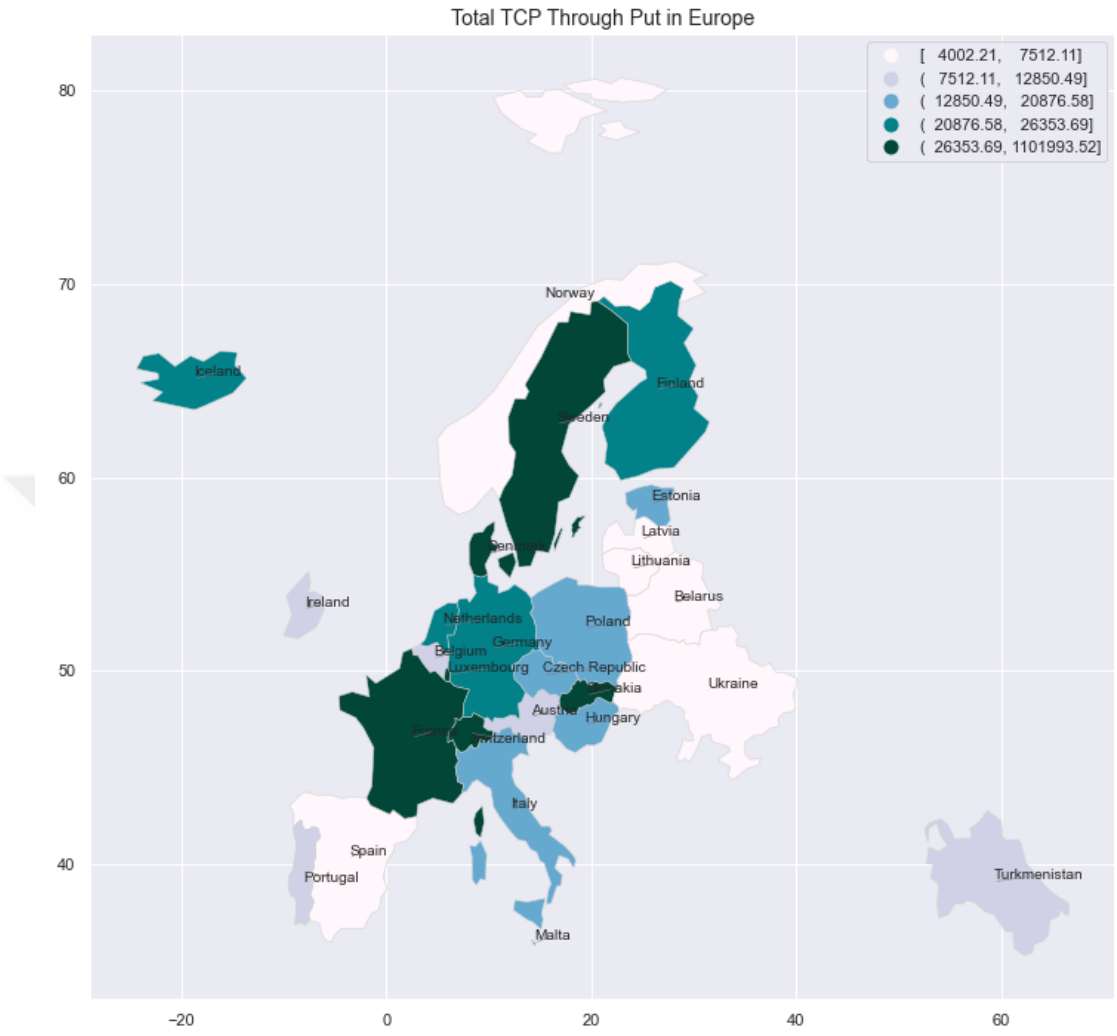


Figure 66: Total TCP Throughput in Europe

In Asia, countries such as Cyprus, Japan, Sri Lanka, Malaysia, Nepal, Pakistan, Qatar, Singapore, and Taiwan have the highest Throughput while Afghanistan, Bangladesh, Bhutan, Iraq, Iran, Kyrgyzstan, Laos, Philippines, Syria, Tajikistan, Uzbekistan have the lowest Throughput (see figure 67).

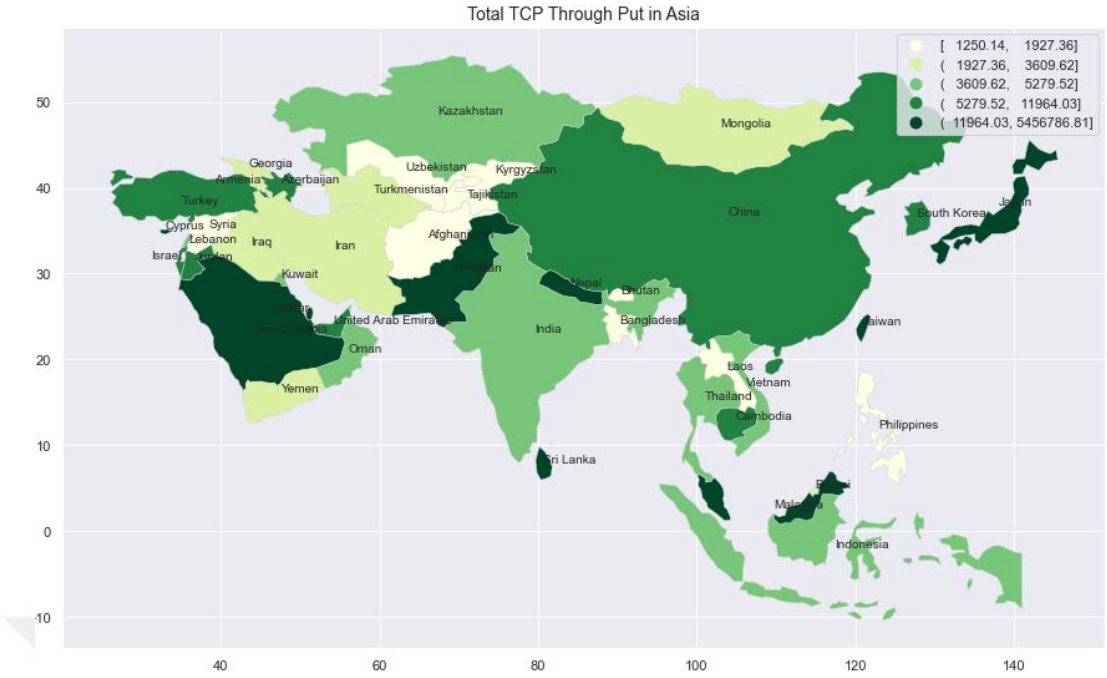


Figure 67: Total TCP Throughput in Europe

4.4.Discussion

The main objective of the current study is to use analytical methods to create a clear understanding of what the digital divide is and the importance of an exploratory visualization model in measuring the digital divide between Asian countries and European countries. As an evaluation metric, the total Duplicate Packet, total Round Trip Time, total TCP Through Put, total Out of Order Packets, and the total Packet Loss will be used.

4.4.1. Round Trip Time

In practice, the round-trip time (RTT) which was defined earlier as "…the amount of time it takes for a signal to be sent plus the amount of time it takes for an acknowledgment of that signal to be received" can be used to evaluate the quality of communications available between two entry points. A high RTT is usually correlated with low-quality user experience hence, the objective is always to ensure the adaptation of measures relevant to lowering the RTT. In this study's analysis, it is reported that Asia has a generally high RTT which translates to a low-quality user experience in Asia compared to Europe. The results show a decreasing trend implying that the RTT of both regions is continuously being optimized. Such might be due to the evolution of the placement of infrastructure which is arguably relevant to

delivering low RTT hence increasing the quality of user-experience. The below table 11 show the RTT breakdown by countries (Asian and Europe) for top 10 higher and top 10 lower for total RTT covering period (2010-2020).

Table 11: Total RTT breakdown by countries

TOTAL ROUND TRIP TIME		
Highest to Low	Afghanistan	Cyprus
	Armenia	Georgia
	Bangladesh	Pakistan
	Bhutan	Qatar
	Iran	Turkey
	Kazakhstan	Andorra
	Laos	Austria
	Sri Lanka	Belgium
	Mongolia	Czech Republic
	Nepal	Denmark
Lowest to High		

4.4.2. Total Packet Lost

Packet Loss is argued to be often encountered when single or more packets of data traveling across a computer network do not reach their intended destination. Studies show that packet loss is caused by among other factors, Faulty or Insufficient Hardware since the hardware plays a significant role in relaying information across the net **Invalid source specified.**. Other causes for packet loss include congestion of the network. Conceptually, “Losses between 5% and 10% of the total packet stream will affect the quality significantly”**Invalid source specified.**. As such, higher losses indicate lower network quality. Asia as shown earlier has a higher Total Packet Lost which is, however, decreasing while Europe has a constant and relatively lower Total Packet Lost between 2010 and 2020 indicating that in general, Europe when using

Total Packet Lost as a metric of evaluating the network performance between the two regions, Europe has relatively better network performance compared to Asia. The below table 12 show the Packet Loss breakdown by countries (Asian and Europe) for top 10 higher and top 10 lower for total Packet Loss covering period (2010-2020).

Table 12: Total Packet Lost breakdown by countries

TOTAL PACKET LOST		
Highest to Low	United Arab Emirates	Bahrain
	Cyprus	Isreal
	Georgia	Jordan
	Iraq	Saudi Arabia
	Iran	Andorra
	Kuwait	Austria
	Laos	Czech Republic
	Philippines	Germany
	Syria	Finland
	Turkey	Faroe
Lowest to High		

4.4.3. TCP Through Put

TCP Through Put was defined earlier on as a metric used to measure how many packets arrive at their destinations successfully. Essentially, the TCP i.e., Transmission Control Protocol is identified as a session-based protocol that makes certain packet delivery and minimization of the packet loss. Low TCP Through Put (quality TCP) is suitable for reliability, quality of images, etcetera **Invalid source specified.**. Asia as shown earlier has a relatively higher TCP Through Put compared to Europe. However, the TCP Through Put has remained fairly stable over time i.e., there is no trend (see table 8 and figure 39). The below table 13 show the TCP Through Put breakdown by

countries (Asian and Europe) for top 10 higher and top 10 lower for total TCP Through Put covering period (2010-2020).

Table 13: Total TCP Through Put breakdown by countries

TOTAL TCP THROUGH PUT			
Highest to Low	Japan	Lowest to High	Afghanistan
	Sri Lanka		Armenia
	Malaysia		Bangladesh
	Nepal		Bahrain
	Pakistan		Brunei
	Qatar		Bhutan
	Taiwan		Georgia
	Switzerland		Iraq
	Germany		Iran
	Denmark		Kyrgyzstan

4.4.4. Out of Order Packets

By definition, Out of Order Packets as noted by studies such as **Invalid source specified.** are frequent in meshed networks MPLS networks. Ideally, out of order packet occurs in cases when the delivery of data packets on a computer network has a variant order from that in which they were sent. Since, if the Out of order packets is high, the TCP will lead to retransmission of packets. Such a case is similar to what happens with dropped packets. Thus, receiving many/high Out of Order Packets affects important systems such as VoIP, video, and RTP**Invalid source specified.**.

Asia was noted earlier on to have relatively higher Out of Order Packets compared to Europe (see figures 22, 23, and table 7) implying the rate at which applications such as VoIP, video, and RTP are affected in Asia is relatively higher compared to Europe.

The below table 14 show the Out of Order Packets breakdown by countries (Asian and Europe) for top 10 higher and top 10 lower for total Out of Order Packets covering period (2010-2020).

Table 14: Total Out of Order Packets breakdown by countries

TOTAL OUT OF ORDER PACKETS			
Highest to Low	China	Lowest to High	Bahrain
	Indonesia		Cambodia
	Iran		Kuwait
	Laos		Maldives
	Sri Lanka		Oman
	Malaysia		Qatar
	Philippines		Turkmenistan
	Pakistan		Turkey
	Thailand		Uzbekistan
	Taiwan		Yemen

4.4.5. Duplicate Packets

Data duplication is one of the most notable generic problems in the field of digital fields and data engineering. In networks, data duplication emerges as Duplicate packets. Data duplication is generally considered as costly to the decisions being made. In networks, data (packet) duplication, affects the traffic between two endpoints by leading to the Pervasion of the volume of information since packet duplication indicates that “packet duplication, which impacts on Service Level Agreement (SLA) planning and threshold-based alerting” which if high, lowers the quality of the delivery hence user-experience**Invalid source specified.**.

Asia as noted earlier has a relatively high by statistically significant decreasing duplicate packets compared to Europe in the years between 2010 and 2020. This in practice implies that Europe has better Service Level Agreement (SLA) planning and threshold-based alerting compared to Asia.

The below table 15 show the Data (packet) Duplication breakdown by countries (Asian and Europe) for top 10 higher and top 10 lower for total Data (packet) Duplication covering period (2010-2020).

Table 15: Total Out of Order Packets breakdown by countries

TOTAL DUPLICATE PACKET		
Highest to Low	Afghanistan	Bahrain
	Bangladesh	Cambodia
	Bhutan	Kuwait
	China	Oman
	Indonesia	Qatar
	India	Turkmenistan
	Sri Lanka	Turkey
	Maldives	Uzbekistan
	Malaysia	Yemen
	Nepal	Andorra
Lowest to High		

CHAPTER 5

CONCLUSIONS

The objective of the current study was:

- To create a clear understanding of what the digital divide is and the importance of an exploratory visualization model in measuring the digital divide between Asian countries and European countries.
- To establish an accurate process of exploratory visualization model implementation in the analysis and interpretation of complex huge volumes of data.
- To develop an exploratory visualization model for measuring the digital divide in Asian and European countries.
- To evaluate the developed exploratory visualization model

Several factors affect the performance of digital networks across different areas with one of the main factors that significantly affect network performance is the Infrastructure on which the network operates. Whereas there is no evidence whatsoever in this paper suggesting infrastructural variation across countries in Europe and those in Asia, it is not entirely illogical to postulate thus i.e. there are infrastructural differences across countries in Europe from those in Asia based on the evidence that Asia has higher Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Loss compared to Europe between the years 2010 and 2020. This supposition is however not entirely true since several other factors influence the performance of the network.

Evidence in this study based on the exploratory visual analysis conducted Suggests that across all the five metrics, Asia performs relatively poorly compared to Europe in general except for country-wise examination where evidence shows that there are countries in Europe that perform poorly that some in Asia. Therefore, regarding the study objective for creating a clear understanding of what the digital divide is and the importance of an exploratory visualization model in measuring the digital divide between Asian countries and European countries, it is clear that there is a large divide. These findings highlight the role of exploratory visualization in drawing insights. For instance, both line graphs and bar graphs played a significant role in illustrating the trend as well as the distribution of the matrices by PingER i.e., Duplicate Packet, Round Trip Time, TCP Through Put, Out of Order Packets, and Packet Loss.

In this paper, the exploratory visualization model followed a methodology that allowed data handling as a preliminary of ensuring clean and accurate data. This includes using prediction models to impute missing observations. The exploratory visualization model adopted in this study included visualization tools such as bar graphs, line graphs, box plots, scatterplots, as well as geographical plots.

Ideally, a digital divide is viewed as the gap between demographics and regions in terms of those that have access to up-to-date information and communications technology, and those that don't or have constrained access. One can argue that countries in Asia based on the preceding observation that Asia has constrained access to information and communications technology since it has been established that there is a digital divide between the two regions.

Evaluation of the performance by the visualization model is not possible statistically. Often, visualizations are judged based on their ability to remain sane while showcasing as much information as possible. In which case the current model performed relatively well in allowing the presentation of the discovered insights.

Therefore, following the findings, the subsequent discussion, and the determination that the model follows the proposed methods, it can be argued that there exists a digital divide between Europe and Asia with Asians having a relatively low-quality user experience compared to Europe.

Recommendation

The important parameters when evaluating the Quality of Experience for networks, Quality of service have a relatively high dependence on network elements. The important attributes include Packet loss rate, jitter, and round-trip time. Therefore, the recommendations are centred on loss rate and round-trip time.

- Network Providing firms as well as other digital companies dealing with connectivity ought to invest in infrastructure that will affect the network performance positively since it is the infrastructure on which the network operates. For instance, an infrastructure that allows faster (shorter) round-trip time will increase the quality of experience.
- Organizations ought to integrate their network systems with real-time analytical tools, methods, and processes which will continually allow organizations to measure the performance of their networks and carry out any necessary adjustments. For instance, in case of packet losses,

organizations can adopt a range of options from restarting hardware to updating the software. Such analytical tools can include BI reporting tools such as dashboards etcetera.

- Digital rights groups and other human rights groups in areas where the loss of packets is due to Internet censorship should push for freedom. This is because censoring methods such as Packet filtering is a method of terminating the TCP in case certain phrases are detected in the network. This affects TCP-type protocols which lead to loss of packets hence low Quality of Service.
- Integrate the analysis and reporting methods with machine learning or data mining methods to allow for the prediction of a drop in the Quality of Experience and automate the system to carry out automatic adjustments to improve the Quality of Experience.

Limitations

One of the main limitations encountered during this study is the problem of data integrity and security. The data used during analysis included some missing observations. Some of the missing observations occurred when merging different files and the determination of how to treat these data problems and which was the optimal problem might slightly affect the accuracy of the findings. Moreover, the use of exploratory analysis alone might have led to the generalization of statistically non-significant observations.

Recommendation for Further Studies

Future studies regarding the effect of a countries social-economic and geopolitical conditions affect the country's digital access. Besides, statistical analysis regarding the hypothesis of digital divide in countries from Asia and Europe should be conducted

References

Abdelsalam, A., and Zampognaro, F. (2017). TCP Wave: A new reliable transport approach for future Internet. Computer Networks, 112, 122-143.

Alasadi, S. A., and Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.

Aldana, C.H.,and Shukla, A.K., Qualcomm Inc, 2015. Methods and systems for enhanced round trip time (RTT) exchange. U.S. Patent 9,154,971.

Angori, L., Didimo, W., Montecchiani, F., Pagliuca, D., and Tappini, A. (2019, September). ChordLink: A new hybrid visualization model. In International Symposium on Graph Drawing and Network Visualization (pp. 276-290). Springer, Cham.

Angus, D. and Wiles, J., 2015. Acquired codes of meaning in data visualization and infographics: Beyond perceptual primitives. IEEE transactions on visualization and computer graphics, 22(1), pp.509-518.

Balliet, R.N. and Heimlich, J., 2016. Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. Information Visualization, 15(3), pp.198-213.

Becht, E., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology, 37(1), 38-44.

Becht, E., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology, 37(1), 38-44.

Bikakis, N., 2018. Big data visualization tools. arXiv preprint arXiv:1801.08336.

Büchi, M., and Latzer, M. (2016). Modeling the second-level digital divide: A five-country study of social differences in Internet use. New media and society, 18(11), 2703-2722.

Chalamalla, A., and Papotti, P. (2014, June). Descriptive and prescriptive data cleaning. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 445-456).

Cox, V. (2017). Exploratory data analysis. In Translating Statistics to Make Decisions (pp. 47-74). Apress, Berkeley, CA.

Crawford, K., and Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. BCL Rev., 55, 93.

Dur, B.I.U., 2014. Data visualization and infographics in visual communication design education at the age of information. Journal of Arts and Humanities, 3(5), pp.39-50.

Ellis, G. and Mansmann, F., 2010. Mastering the information age: solving problems with visual analytics.

Ellsworth, J.L., and Newcombe, C.R., Amazon Technologies Inc, 2018. Forward-based resource delivery network management techniques. U.S. Patent 9,893,957.
Espinosa, J.A. and Money, W., 2013, January. Big data: Issues and challenges moving forward. In 2013 46th Hawaii International Conference on System Sciences (pp. 995-1004). IEEE.

García, S., and Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9.

Granato, D., and Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends in Food Science and Technology, 72, 83-90.

Grant, L., and Eynon, R. (2017). Digital divides and social justice in technology-enhanced learning. In Technology Enhanced Learning (pp. 157-168). Springer, Cham.
Hazen, B. T.,and Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem

and suggestions for research and applications. International Journal of Production Economics, 154, 72-80.

Healy, K., 2018. Data visualization: a practical introduction. Princeton University Press.

Helfman, J. and Goldberg, J., Oracle International Corp, 2016. Filtering for data visualization techniques. U.S. Patent 9,477,732.

Hoeber, O., 2018, March. Information visualization for interactive information retrieval. In Proceedings of the 2018 Conference on Human Information Interaction and Retrieval (pp. 371-374).

Howe, B. and Heer, J., 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. IEEE transactions on visualization and computer graphics, 22(1), pp.649-658.

Kahn, M. G., and Liaw, S. T. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. Egems, 4(1).

Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). Sthda.

Kassambara, A. (2017). Practical guide to cluster analysis in R: Unsupervised machine learning (Vol. 1). Sthda.

Kharat, P., and Kulkarni, M. (2019). Congestion controlling schemes for high-speed data networks: A survey. Journal of High-Speed Networks, 25(1), 41-60.

Kiely, D. and Salazar, S., 2018. Falling Through the Net: The Digital Divide in Western Australia (No. FWA11). Bankwest Curtin Economics Centre (BCEC), Curtin Business School.

Kosara, R., 2016. Presentation-oriented visualization techniques. IEEE computer graphics and applications, 36(1), pp.80-85.

Kumar, A., and Johnson, A. (2020, May). Augmenting Small Data to Classify Contextualized Dialogue Acts for Exploratory Visualization. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 590-599).

Lawrence, J., 2012. Application Performance Monitoring: Out of Order Packets, s.l.: Plixer.

Lewin, B.A. and Singh, A.K., New BIS Safe Luxco SARL, 2018. Methods, apparatus and systems for data visualization and related applications. U.S. Patent 9,870,629.

Li, J. (2018). An Exploratory Analysis of Individual Long-Term Google Search and Browsing History.

Liu, D., and Liu, Y. (2015). Duplicate detectable opportunistic forwarding in duty-cycled wireless sensor networks. IEEE/ACM Transactions on Networking, 24(2), 662-673.

Mal, A., and Cottrell, L. (2016, January). Analysis and clustering of PingER network data. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 268-273). IEEE.

McCraw, C., 2020. How to Fix Packet Loss: Six Common Causes and Solutions [Guide], s.l.: GETVOIP.

Nayak, G.K. and Lenka, R.K., 2016, December. Big data visualization: Tools and challenges. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 656-660). IEEE.

Ordu, M.D. and Simsek, B., 2015. Examining the global digital divide: a cross-country analysis. Communications of the IBIMA, 2015, p.1.

Pan, A.,and Leslie, R. (2016, January). Application for the emulation of PingER on android devices. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 537-541). IEEE.

Plaisant, C. and Carpendale, S., 2011. Empirical studies in information visualization: Seven scenarios. IEEE transactions on visualization and computer graphics, 18(9), pp.1520-1536.

Quevedo, D. E.,and Nesic, D. (2011). Packetized predictive control of stochastic systems over bit-rate limited channels with packet loss. IEEE Transactions on Automatic Control, 56(12), 2854-2868.

Ramakrishnan, R. and Shahabi, C., 2014. Big data and its technical challenges. Communications of the ACM, 57(7), pp.86-94.

Reddy, G. T.,and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access, 8, 54776-54788.

Salyers, D. C., Striegel, A. and Poellabauer, C., 2011. Wireless Reliability: Rethinking 802.11 Packet Loss, Notre Dame: University of Notre Dame.

Sampson, R.,and Cottrell, L. (2017, January). Implementation of pinger on android. In 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence (pp. 306-312). IEEE.

Serral-Gracià, R., Domingo-Pascual, J. and Jakab, L., 2016. Out of order packets analysis on a real network environment. s.l., IEEE.
Shishkin, Y.E. and Skatkov, A.V., 2016. Big Data visualization in decision making. In Science in Progress (pp. 203-205).

Shkedi, R.,2011. Method and stored program for accumulating descriptive profile data along with source information for use in targeting third-party advertisements. U.S. Patent 7,979,307.

Shneiderman, B., 2013. Improving healthcare with interactive visualization. Computer, 46(5), pp.58-66.

Soltanpoor, R., and Sellis, T. (2016, September). Prescriptive analytics for big data. In Australasian Database Conference (pp. 245-256). Springer, Cham.

Spirent, 2016. TCP Network Latency and Throughput: Or 'Why your customer doesn't receive the Throughput they paid for', New York: Spirent.

Steele, J. and Iliinsky, N., 2011. Designing data visualizations. O'Reilly Media, Inc..

Tang, N. and Li, G., 2018, April. Deepeye: Towards automatic data visualization. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 101-112). IEEE.

Taylor, S. and Metzler, J., 2018. The impact of out of order packets on successful transmission. [Online] Available at: https://www.networkworld.com/article/2289506/the-impact-of-out-of-order-packets-on-successful-transmission.html

Tu, C. and Chen, B., 2017. Is there a robust technique for selecting aspect ratios in line charts? IEEE transactions on visualization and computer graphics, 24(12), pp.3096-3110.

Ucar, I., Morato, D., Maga˜na, E. and Izal, M., 2013. Duplicate detection methodology for IP network traffic analysis, Pamplona, Spain: IEEE.

Van Der Aalst, W., 2016. Data science in action. In Process mining (pp. 3-23). Springer, Berlin, Heidelberg.

Van Dijk, J.A., 2017. Digital divide: Impact of access. The international encyclopedia of media effects, pp.1-11.

Wenwei, L. and Fang, C., 2018. Bridging the digital divide: measuring digital literacy. Economics: The Open-Access, Open-Assessment E-Journal, 12(2018-23), pp.1-20.

White, B. and Cottrell, L., 2016, January. Analysis and clustering of PingER network data. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 268-273). IEEE.

Williamson, B., 2016. Digital education governance: data visualization, predictive analytics, and 'real-time'policy instruments. Journal of Education Policy, 31(2), pp.123-141.

Xu, L. and Nandi, A., 2016. Graphical perception in animated bar charts. arXiv preprint arXiv:1604.00080.

ZHANG, X., and Lei, Z. H. U. (2019, June). Fast salient object detection based on multi-scale feature aggression. In 2019 Chinese Control and Decision Conference (CCDC) (pp. 5734-5738). IEEE.