Coursework

Applied Statistics and Data Visualisation

(Principles of Data Science)

MSc Data Science

University of Salford MANCHESTER

Research Titles :

1. **COMPARISON OF ECONOMIC AND HEALTH INDICATORS: AN INTERACTIVE DASHBOARD IMPLEMENTATION**

2. **URBANIZATION AND INTERNET USAGE**

Name : PETER LASISI

Student ID : @00688609

Date : December 3, 2022

Table of Contents

Part One: Interactive Dashboard Design

1. Introduction

Health is a direct contributor to human welfare and a tool for increasing revenue. We talk about a few ways that health can affect income, concentrating on lifetime maternity risk and average life expectancy. Health may have a significant impact on expected lifespans and life cycle behavior, in addition to the average life expenditure. According to studies, a child's health throughout the first few years of life may have a significant impact on the child's physical and cognitive growth as well as later economic success. Changes in health could be just as important as changes in money when it comes to development and human well-being. Healthy living is a goal in and of itself, independent of one's financial situation. Even in the most underdeveloped countries, prioritizing health spending may be desirable because income is a function of health.

A Software called Tableau Public can be used to create dashboards and visualizations based on research and analysis. It is vital to visualize economic and health statistics so that the general public may grasp their link using graphs and plots that even someone who is not particularly tech-inclined can understand and draw the appropriate conclusions from. A story can be told by supplying information and knowledge, but developing visualization dashboards involves a lot of work to establish metrics, comprehend indicators, and portray them graphically. An effective visualization dashboard should also take into account how people perceive information, make it come to life, and encourage action. Lack of consensus over the design and functionality of a visualization dashboard is another issue. A thorough examination of the procedures for transmitting crucial information utilizing a dashboard. Some approaches emphasize the necessity to take the audience's demands into account while others place more emphasis on making the dashboard construction's goals and objectives apparent while always keeping the end goal in mind.

Which continents' (countries') economic and financial healthcare performance is the best? Based on their geographical location, some countries are chosen to represent their respective continents to shed light on how diverse groups of countries are faring. The study aims to evaluate the performance of such countries and make comparisons.

DATASET

The dataset was downloaded from the world databank via https://databank.worldbank.org/source/world-development-indicators, One of the main World Bank and United Nations Data Bank collections of indicators, the World Development Indicators (WDI), is compiled from officially acknowledged international sources. National, regional, and international estimates are provided by both sources. The dataset was organized, cleaned, and filtered. For each continent, two nations were chosen, including:

- Africa – Nigeria and South Africa
- South America – Argentina and Brazil
- North America – United States and Canada
- Asia – China, and Japan
- Europe – United Kingdom and France
- Oceanic - Australia and New Zealand

2. Background Research

Dashboard design is usually interesting. Several factors have to be considered before implementation. Factors like data to be displayed and the targeted audience is worth considering to help in its implementation. Dashboard designs can extract knowledge from a Large Data set from a particular domain to explain concepts, confirm an Hypothesis or to generate knowledge. Technological advancement has also helped in Dashboard designs. Technology has made data readily available for exploration invariably necessitating the need for designs that will help our target audience understand complicated data better. Users must be provided with Tailored designs that will not only help them fulfil their requirements but will also enhance decisions to be taken. Great dashboards communicate information quickly. They display information clearly and efficiently. They show trends and changes in data over time. They are easily customizable. The most important widgets and data components are effectively presented in a limited space. It is therefore important to generate an interactive dashboard that will generate knowledge and that will be user friendly.

In designing Dashboards or visualizations, your end user must be taken in cognizance. When designing single screen dashboards, your must put yourself in the shoes of your targeted

user/audience. If you make the chart too complex, your audience may have difficulty in interpreting your visualization.

Another important point is choosing your metrics or KPI for your dashboard. Your KPI will help shape the direction of your dashboard design. Your visualization should be able to tell a story. A story that will communicate not only to the Technical users but also to the non analytic users. it is also vital not to try to put unnecessary information on your dashboard. This will further complicate issues.

Selecting the right kind of dashboard must also be taken in consideration. You must use the right chart that will help in decision making for the users. In dashboard designs, simplicity is also very important. A complicated dashboard as technical as it may appear may be hard to understand by the non technical users. These days we can play with a lot of effects and charts to make our visualization look more attractive. A rule of thumb is to avoid some of these so that our charts can be as simple as possible. You must also be careful with labeling paying much attention to Font size and colour. Rounding up numbers is also key as you don't want to make your audience uncomfortable with too many decimal numbers to worry about e.g sales figure of $450,244.34 may be rounded up to $450,000.00 for better and easier understanding. Also when choosing colour we must stick to a particular pattern. If for example, we want to show negative figures, we can decide to use RED and must stick to that colour all through our dashboard design. This will help reduce the mental effort it will take for your audience to interpret your visualization.

Single screen dashboard designs should also have elements that will help users drill down into specific clicks to filter and time intervals. They make your visualization neat, and will allow you hide unnecessary details that will overcrowd your design. Recent design paradigm also uses animations to catch the attention of users. It is important to mention that although animation may make your dashboard interactive you must not over use them.

Your dashboard design must also be optimized for different devices (Phones and tablets). This will enable users on the go to have access to your visualization remotely avoiding unnecessary office gatherings.

Graphical integrity must also be taken very important when designing your dashboards. Your graphs must depict the truth about your data and to make them look in a certain way to benefit your analysis.

You must also avoid data visualization mistakes. This include but are not limited to, wrong calculations, wrong choice of visualization and too much data. No matter the dashboard design, whether it is an Executive dashboard or a normal operational dashboard, the rule of simplicity and colour theory will help portray a good and reliable visualization.

Lastly it is important to mention that in dashboard designs, we must not be afriad to evolve. As technological tools get more advanced, the tools for visualization will also follow same trend. It is important we get familiarized with new tools that will help us better visualize data in general.

3. Exploration of Data Set

The dataset was downloaded from the world databank via https://databank.worldbank.org/source/world-development-indicators, One of the main World Bank and United Nations Data Bank collections of indicators, the World Development Indicators (WDI), is compiled from officially acknowledged international sources. National, regional, and international estimates are provided by both sources. The dataset was organized, cleaned, and filtered. For each continent, two nations were chosen, including:

- Africa – Nigeria and South Africa
- South America – Argentina and Brazil
- North America – United States and Canada
- Asia – China, and Japan
- Europe – United Kingdom and France
- Oceania – Australia and New Zealand

Reading and comprehending the instructions was followed by research and data collection. When gathering data, the intended audience was taken into consideration. The choice of data met the criteria for reliable and accurate data collection, which state that the data must be correct, impartial, and ample for proper analysis. On the World Development Indicator website, several series were picked with the project purpose in mind and the years selected, and the countries were chosen based on their effect on the continent. After that, the data was cleaned with the aid of excel software and was also reorganized and reordered. After everything was configured, the data was exported to Tableau for analysis and dash-boarding.

4. Investigation of Data Workflows & Proposal for Design of Dashboard

A dashboard is a communication tool that monitors information quickly and gives the user the most recent details regarding the topic at hand (Few, 2006; Kerzner, 2013). Dashboards are visual displays of the most crucial data required to accomplish one or more objectives that are consolidated and arranged on a single screen for easy monitoring, according to Few (2006). A dashboard should provide data that shows progress and indicates whether any events call for the user's attention. However, it is not required to give all the information that could be required to take action. The primary issue of creating a dashboard is to fit a large quantity of data into a small space while still being able to transmit information effectively and with ease (Few, 2006). Making the best use of the design area is another difficulty. Every piece of information on the dashboard ought to be significant. However, the relevance of the data must be assessed and ranked while constructing the dashboard because certain data may be more relevant than others.

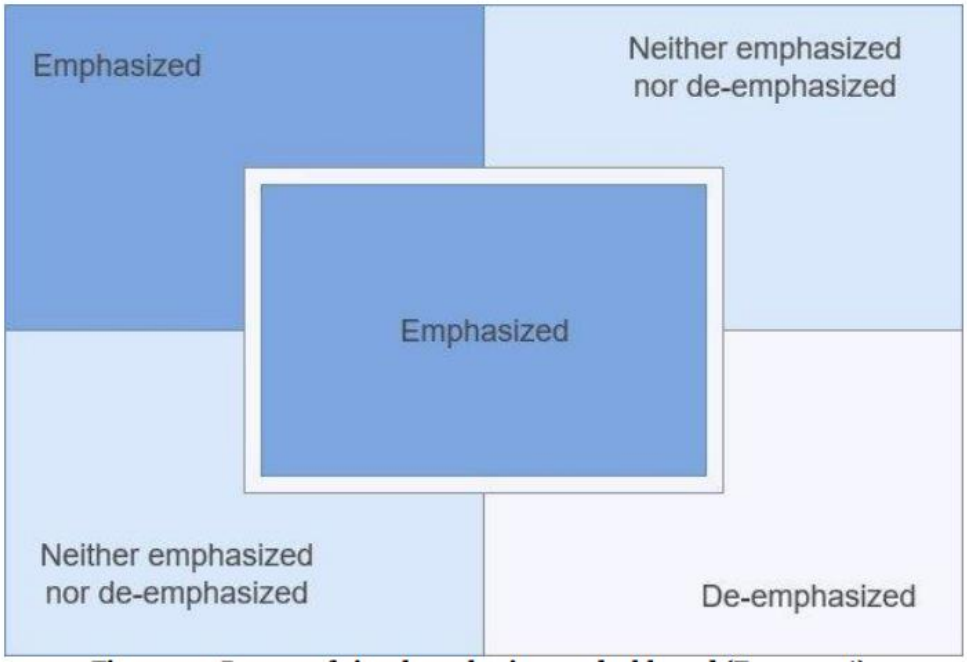

Fig. 1. Degree of visual emphasis on a dashboard (Few, 2006).

INFORMATION VISUALIZATION

The use of graphical representation to display information is the subject of the field of information visualization (Card et al. 1999; Fekete et al. 2008; Mazza, 2014). The ability to help people comprehend data better is one of information visualization's most notable benefits (Fekete et al. 2008). According to Spence (2016), visualization is the process of creating a mental model of something. He is implying that the mental model produced by

visualizations can improve knowledge retention. The visualizations can improve pattern recognition (Fekete et al. 2008). Visuals can operate as short-term memory for human cognitive processes, giving the mind more tools with which to think and analyze. Visual representation can be a useful tool for explaining underlying concepts, ideas, comparisons, and relationships in data when a lot of information needs to be conveyed. Choosing a visual representation that will effectively convey information is difficult. Excellence in statistical graphics, according to Tufte (2001), "consists of complicated ideas expressed with clarity, precision, and efficiency."

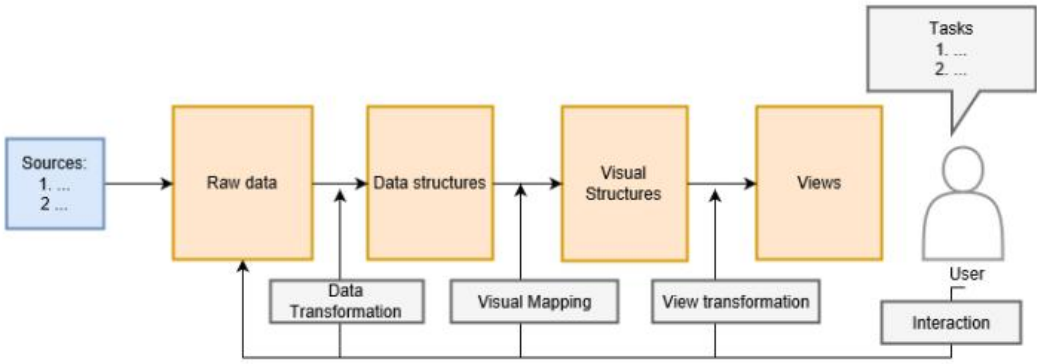

Fig. 2 The process of generating a graphical representation

Four elements are required to create a graphical representation: raw data, data structures, visual structures, and views. The user can interact with the views and modify how the components are transformed (Card et al. 1999; Mazza, 2009; Spence, 2016).

Unstructured data that has been acquired is known as raw data. It may be acquired from a single source or numerous, such as spreadsheets and logs. There are numerous formats for the data, including nominal, numerical, and categorical (Spence, 2014). The name and age of a person are examples of nominal and numerical data. Typically, the acquired data needs pre-processing before being used for visualization because it is not in a format that is suitable for an automatic processing tool (Mazza, 2009).

For instance, it is necessary to fix errors or missing values in raw data before processing it (Card et al. 1999). After that, data structures are created from the raw data. Data transformation is required because raw data frequently has an unsuitable structure or format for processing. The data must be translated into a logical format that the program can understand for it to be processed by the software (Card et al. 1999; Mazza, 2009,). The data structure can be improved by including pertinent details or carrying out some initial

processing. Visual mapping is the process of mapping and translating data into visual structures (Card et al. 1999). Three components must be defined to discover the associated visual structure: the spatial foundation, the graphical elements, and the graphical properties. The view is the outcome that the user sees (Mazza, 2009). View transformation is the process of altering the view through interaction. Graphical parameters like location, scale, and clipping are specified via view transformations (Card et al. 1999). All visual structures might not be supported in the available space when there is a lot of data. Interaction tactics offer a chance to observe specifics and get more understanding of these circumstances.

USER DESIGN DASHBOARD

A user-centered dashboard (UCD) design was employed for this study. In a user-centered design (UCD) process, all design choices are made with the user's needs and requirements in mind (Interaction Design Foundation, 2018). To construct systems tailored to a particular user group, the design approach centers on the user. As a result, the designer can produce an accessible and useable system that improves user happiness. The design process is frequently iterative and begins by determining the user's demands. Investigative and generative methods are then used to comprehend the user's context and wants. The user can then assess how effectively the design is performing in light of the context and needs after the design phase. Following the evaluation, additional iterations are made, starting with specifying the context of use.

Figure 1 depicts the design process. UCD is significant in information visualization since its main objective is to assist the user in discovering patterns and solutions. It's important to identify the user attributes and context early on in the design process. The design of the information visualization can vary depending on the users' objectives and tasks. For instance, the work of an ambulance driver is well-defined and well-known as a result of training. The user may need additional assistance from the offered visualization in finding the information they need quickly. In a different scenario, the design may need to offer additional exploratory tools and visualizations to help the user finish the task because the task was inadequately formulated and speed may not be a requirement (Spence 2014). Romero (2017) argues that to continuously enhance the design, the method for developing information visualization should be user-centered and iterative. When developing the information visualization using a user-centered design method, the following issues should be taken into consideration (Romero 2017):

1. Who is the user, first?

2. What are the assignments?

3. What are the numbers?

4. What visual structures are there?

5. How does the visualization help with the assignments?

6. How could it be made better?

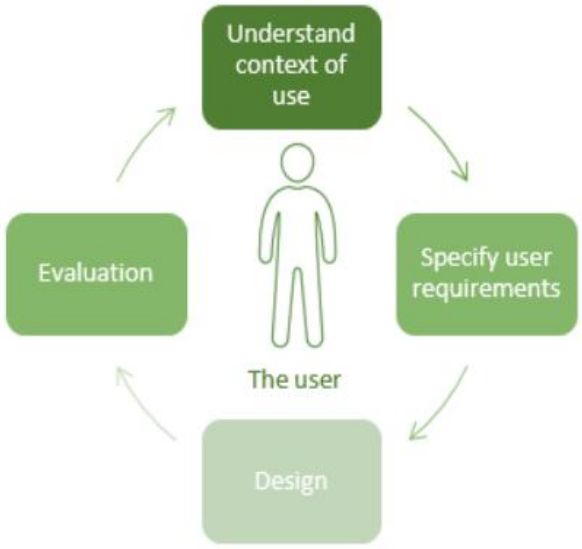

Fig. 3: A user-centered design process

5. Discussion

DASHBOARDING PRINCIPLES

Visualization principles according to www.medium.com include the following: the best visual should be chosen, the design should be balanced, attention should be paid to key areas, the visuals should be kept simple, patterns should be used, parameters should be compared, and interactivity should be added, among other important factors. All of these principles were taken into consideration when the dashboard was designed.

METRICS OF COMPARISONS

1. **Map View of the countries represented**: In this design, it is important to display a map view of the countries being represented for this analysis. As seen in Fig. 4 below, a careful selection of countries have been made so that we can have a Global view of our analysis and

design. Countries were selected in a systematic manner to ensure that all the continents are included in the comparison. The colour legend was introduced here to distinguish each country. We also activated each country label so as to visibly identify the countries under focus.

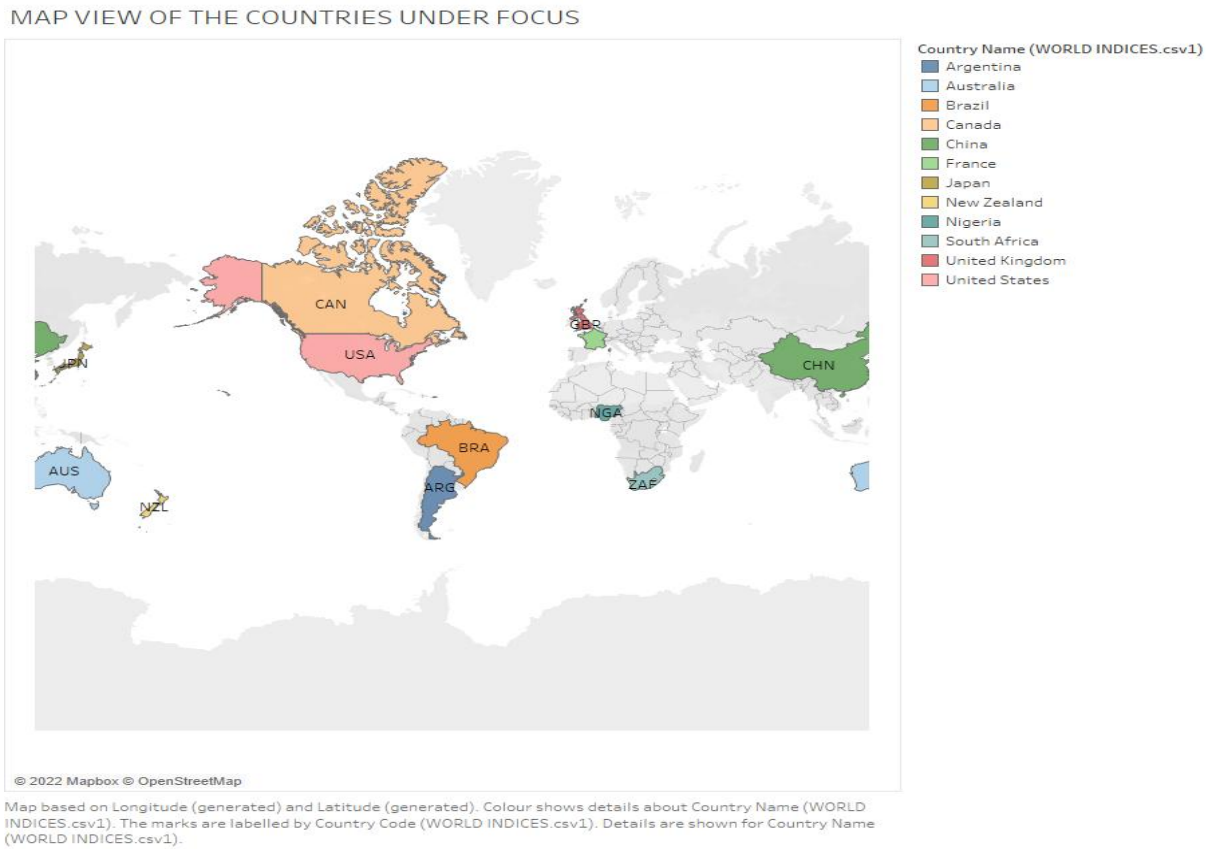

Fig. 4

2. **Average GDP:** The Average GDP is the first economic development indicator (Average Gross Domestic Product). The analysis's findings are displayed in the graph below (Fig. 5). As shown in the diagram, the average GDP displayed represents the average figures for each country from year 2000 to the year 2021, where the US, which represents North America, has the highest average GDP of about $16 trillion dollars. The second and third places are held by two Asian nations due to their strong performances (China and Japan). The European nations, which have maintained a median position in the chart, are listed below the Asian nations. The two biggest African countries in terms of GDP (Nigeria and South Africa) are joined by Argentina and New Zealand to record the lowest average GDP from the year 2000 to 2021 with their average GDP less than a trillion dollars each. It is important to mention that a Bar chart was used to display this visuals as its easy to identify countries

doing well from the height of the Bars. The colour legends, year filters and labels was used in this visuals.

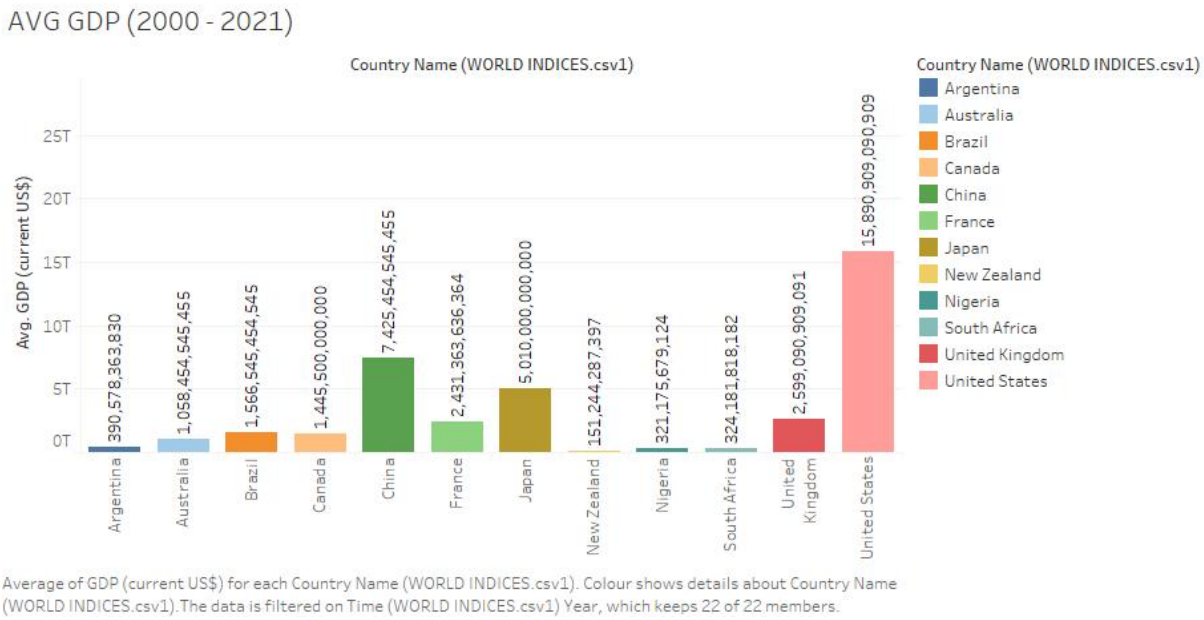

Fig. 5

3. HEALTH INDICATIONS COMBINED:

We have combined the average life time expectancy and the average risk of maternity death in a single chat (fig. 6). It is important to mention that while other countries formed a cluster, Nigeria and South Africa form separate clusters which indicates bad performance.

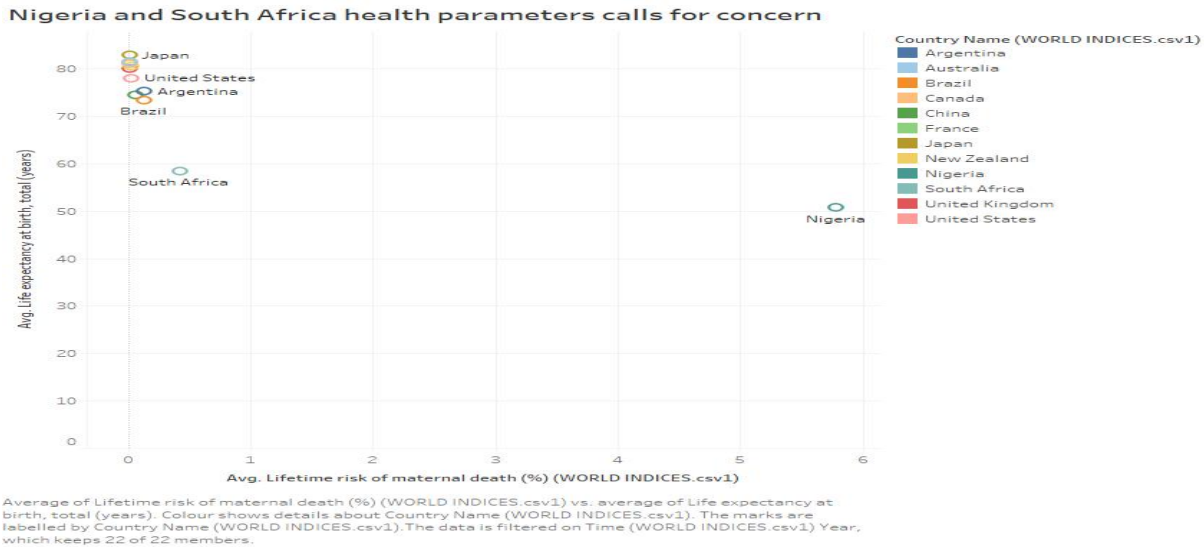

Fig. 6

4. **AVERAGE CURRENT HEALTH EXPENDITURE (% of GDP)**: This is an indicator that shows what percentage of GDP is spent on health related issues. As seen in fig. 7, Nigeria, South Africa and China are least in this category.

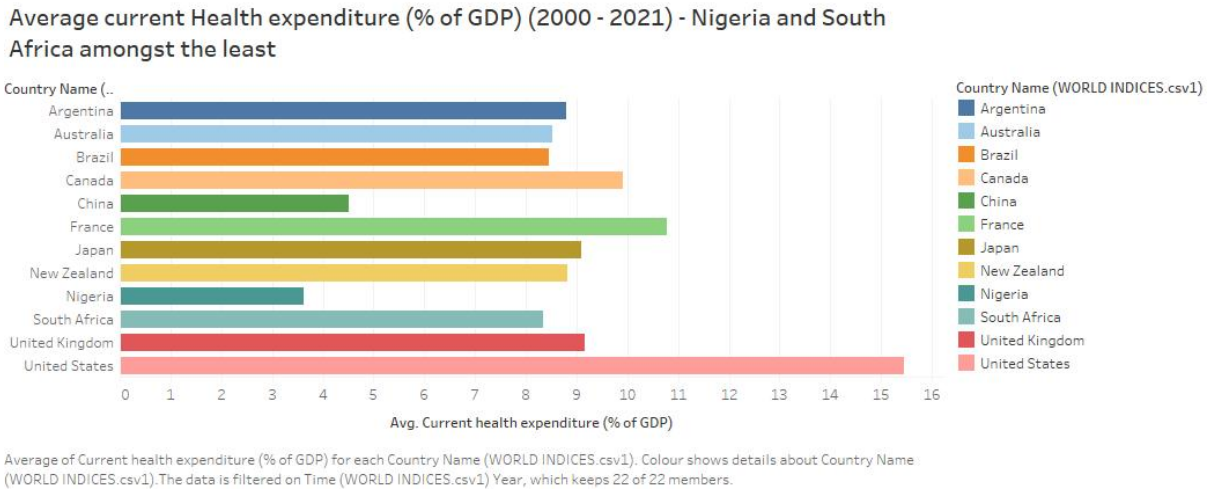

Fig. 7

5. **AVERAGE LIFE-TIME EXPENTANCY AT BIRTH**: Life expectancy at birth is defined as how long, on average, a newborn can expect to live, if current death rates do not change. Looking at Fig. 8 below, every of the countries selected has an average life-time expectancy at birth of 75 years and above except for the African countries. We have introduced colour legends and labels to this visuals. The bar chart is the best fit for this representation for simplicity and less ambiguity.

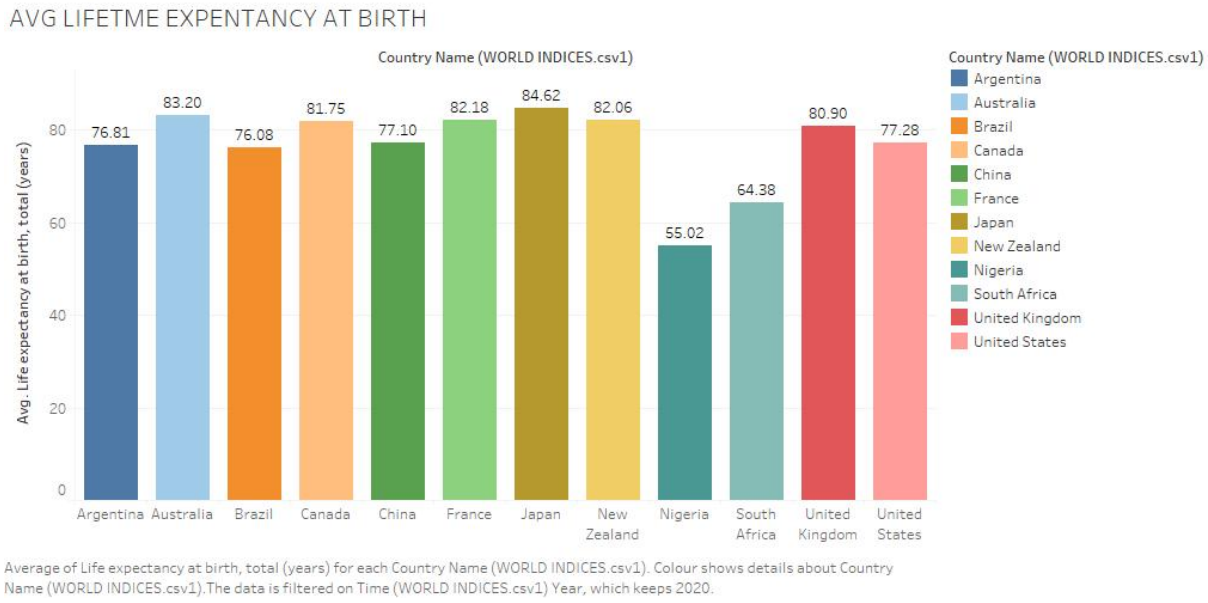

Fig. 8

6. Conclusions

At the end of the analysis, it can be observed that Africa is the least-performed in terms of the GDP and health indices analyzed, coming last in most of the indicators. Asia and North America getting better results in terms of economic capability. The Asian continent is also competing for the best in terms of life expectancy (Fig. 9).

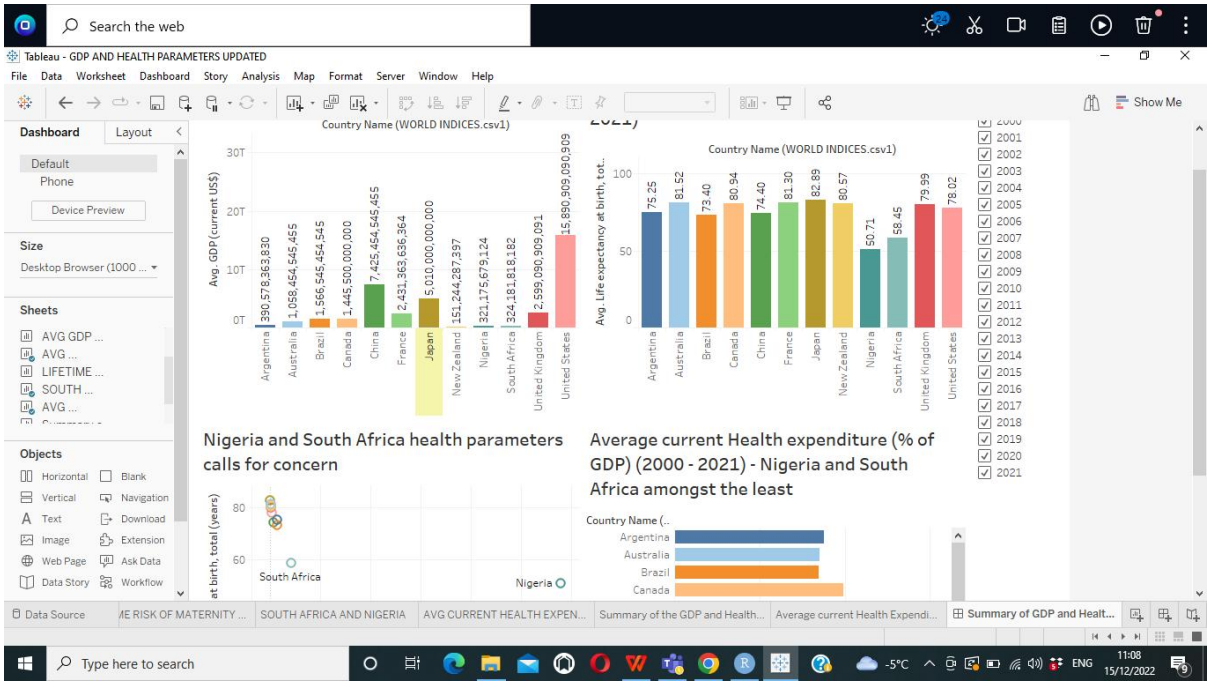

Fig. 9

Part Two: Statistical Analysis

1. Introduction

The research objective is to see how urbanization affects internet usage. Is there a relationship between the urban population and the number of people using the internet?

2. Background Research

Hong & Thakuriah (2018) discovered that although urbanization in developing nations has its drawbacks, it has also sparked technical advancement. As a result of their superior socioeconomic qualities and concentration of high-quality ICT infrastructure, people are more likely to utilize the Internet than those who live elsewhere.

Urban locations also have more Wi-Fi networks overall and better quality data networks. According to Philip et al. (2015), Scotland's urban areas have the highest concentration of 4G mobile Internet coverage.

Hong & Thakuriah (2018) also noted that large urban dwellers typically possess a higher level of computer literacy than those in small towns or rural locations. About 62% of those they observed used their smartphones to access the Internet. The percentage is larger for those who live in metropolitan areas than for those who live in towns or rural regions. This highlights the crucial link between using a smartphone to access the Internet and urban environments. The authors also believed that the ownership of a smartphone affects internet usage.

According to a recent study from the Pew Research Center (2014), 58% of Americans now own smartphones, up from 35% in 2011; age and educational attainment are also substantially connected with smartphone use. Additionally, they discovered that smartphone ownership varied by residential area. Particularly, only 43% of people who live in rural areas have smartphones compared to around 60% of urban and suburban citizens.

The aim of Steven et. al (2014) was to investigate how urban populations affect internet connectivity. They anticipate that demand in non-urban areas will increase if the internet is successful in big cities. Furthermore, demand should be lowest in the major cities since people outside the city would utilize the internet more frequently to replace services that would otherwise be offered by cities.

On the other hand, if the internet complements cities in that it enables individuals to more fully utilize urban services and reduces the costs associated with congestion, then we should observe a larger tendency for internet connectivity in urban regions. Furthermore, the level of connectivity should be highest in the biggest cities. Over a 12-year period, from 2000 to 2011, they used panel data of 50 states. The parameters can be interpreted as elasticities because the variables are in natural logs.

The main result is that high speed internet connections are highly responsive to the level of urbanization within a state. According to the coefficient of 3.37, there are 3.37% more high-speed internet connections for every 1% increase in urbanization. The extremely elastic finding shows that urban residents are far more interested in internet connections than country dwellers are.

The empirical analysis of Steven et. al (2014) reveals that not only do urban populations require more internet access, but that internet access is even more crucial when the urban population of a state is more densely concentrated in a single extremely large metropolis.

Correlation Analysis

Correlation analysis is a statistical technique that determines whether and how strongly two variables or datasets are related. Statistical correlation is typically ranked in three different ways, that is, Spearman, Kendall, and Pearson. In this work, I used the Pearson correlation. The Pearson correlation formula, which assesses the strength of the 'linear' correlations between the raw data from both variables rather than their ranks, is the most used one for correlation analysis. The formula for Pearson correlation is shown below.

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

However, if the relationship between the data is not linear, Spearman's Rank must be used in its place because this particular coefficient will not effectively depict the relationship between the two variables.

Any result between +0.5 and +1 denotes a very strong positive correlation. That is, they both rise at the same time. Any value between -0.5 and -1 denotes a high negative correlation, meaning that as one variable rises, the other falls proportionately. A score of 0, means that there is no correlation between the two variables. The outcome is more accurate with a larger sample size (Emily, n.d.)

Regression Analysis

The estimation of associations between a dependent variable and one or more independent variables is done using a set of statistical techniques called regression analysis. It can be used to model how strongly the relationship between variables will develop in the future and to evaluate the strength of the relationship between variables. Data trends can be discovered using regression analysis.

There are various types of regression analysis, including linear, multiple linear, and nonlinear. Simple linear and multiple linear models are the most prevalent types. For more complex data sets where the connection between the dependent and independent variables is not linear, nonlinear regression analysis is frequently used (CFI Team, 2022).

The relationship between a dependent variable and an independent variable is evaluated using a simple linear regression model. The formula of simple linear regression is shown below.

$$y = b0 + b1X + \epsilon$$

Where:

y – Dependent/target variable

X – Independent variable

b0 – Intercept

b1 – Slope/weight

ϵ – Error

Multiple linear regression is similar to simple linear regression. The only difference is that it uses more than one independent variable. The formula for multiple linear regression is shown below.

$$y = b0 + b1X1 + b2X2 + b3X3 + \epsilon$$

Where:

y – Dependent variable

X1, X2, X3 – Independent variables

b0 – Intercept

b1, b2, b3 – Weights

ϵ – Error

Time Series Analysis

Data collected over a specific time period is referred to as a time-series data. Time-series analysis is a way of analyzing time series data to obtain important statistics and attributes. Predicting future value is one of the study's key objectives (Gupta, 2022).

Some components of time series are trend and seasonality. The trend depicts the time series data's overall direction over a significant amount of time. A trend may be rising, falling, or flat (stationary). A pattern that repeats in terms of timing, direction, and size can be seen in the seasonality component. An example is an increase in the sale of jackets during winter.

A stationary series is one whose values are not affected by the passage of time. The values are therefore not affected by time. As a result, the series' statistical characteristics, such as mean, variance, and autocorrelation, remain consistent across time. The autocorrelation of a series is the correlation between a series and its prior values. If a series has a high degree of autocorrelation, the lags—previous values of the series—might be used to predict the present value.

The majority of statistical forecasting techniques are made to function with stationary time series. A non-stationary series is often transformed to become stationary as the initial stage in the forecasting process. The Augmented Dickey-Fuller (ADF) Test is one of the methods used to check if the series is stationary (Banerjee, 2020).

The ARIMA (Autoregressive Integrated Moving Average) method was used to perform time series analysis in this research. In an ARIMA model, the data is differenced in order to make it stationary.

3. Exploration of Data Set

The dataset has 11 variables and 210 records. It is the data extracted from the world development indicator. The variables are *Time, Time Code, Country Name, Country Code, Individuals using the Internet (% of population) [IT.NET.USER.ZS], Urban population [SP.URB.TOTL], Urban population (% of total population) [SP.URB.TOTL.IN.ZS], Rural population [SP.RUR.TOTL], Rural population (% of total population) [SP.RUR.TOTL.ZS], Secure Internet servers [IT.NET.SECR],* and *Secure Internet servers (per 1 million people) [IT.NET.SECR.P6]*.

The *"Time"* variable shows the years. It ranges from the year 2007 to 2021. It is a numeric variable. The "Time Code" show the year too (e.g., YR2007). The "*Country Name*" variable shows the country. There is data from 14 countries. The countries in the dataset are Germany, the United Kingdom, the United States, Belarus, Canada, Jamaica, Nigeria, Egypt, Burundi, China, Bangladesh, India, Australia, and New Zealand. The "*Country Code*" variable shows the country code (e.g DEU).

The "*Individuals using the Internet (% of the population)*" variable shows data on internet users. These are people who have utilized the Internet in the past three months (from any location). The source of this data is the International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database.

The *"Urban Population"* variable shows the urban population of the country in a particular year. The *"Rural Population"* variable shows the rural population of the country in a particular year. This data is from the World Bank staff estimates based on the United Nations Population Division's World Urbanization Prospects: 2018 Revision.

The *"Secure Internet servers"* shows the number of distinct, publicly-trusted TLS/SSL certificates found in the Netcraft Secure Server Survey. This data is from Netcraft (http://www.netcraft.com/) and World Bank population estimates.

Time <dbl>	Time Code <chr>	Country Name <chr>	Country Code <chr>	Individuals using the Internet (% of population) [IT.NET.USER.ZS] <dbl>	Urban population [SP.URB.TOTL] <dbl>
2007	YR2007	Germany	DEU	75.160000	62833410
2007	YR2007	United Kingdom	GBR	75.090000	49351705
2007	YR2007	United States	USA	75.000000	241795278
2007	YR2007	Belarus	BLR	19.700000	7005597
2007	YR2007	Canada	CAN	73.200000	26441461
2007	YR2007	Jamaica	JAM	21.100000	1472255
2007	YR2007	Nigeria	NGA	6.770000	59734513
2007	YR2007	Egypt, Arab Rep.	EGY	16.030000	33700834
2007	YR2007	Burundi	BDI	0.700000	775530
2007	YR2007	China	CHN	16.000000	595670841

1-10 of 210 rows | 1-6 of 11 columns

Previous 1 2 3 4 5 6 ... 21 Next

Figure 10 Preview of the Data Set

Data Cleaning

I removed the *'Time Code', 'Country Code', 'Secure Internet servers [IT.NET.SECR]'* and *'Secure Internet servers (per 1 million people) [IT.NET.SECR.P6]'* variables. These variables will not be useful for my analysis. I have the *'Time'* variable which shows the year, so the *'Time Code'* variable won't be needed. I also have the *'Country Name'* variable, so the *'Country Code'* variable won't be useful. The last two variables *'Secure Internet servers [IT.NET.SECR]'* and *'Secure Internet servers (per 1 million people) [IT.NET.SECR.P6]'* won't help us with our research objective of seeing how urbanization affects internet usage. I used the names() function in R to filter out the columns I did not need.

There were some missing values in the dataset. I imputed the missing values with the value of the previous year based on the country. For example, for the *"Individuals using the Internet (% of the population) [IT.NET.USER.ZS]"* variable and the country *"Germany"*, if the value for the year 2020 is 89 and the value for the year 2021 is missing, then the missing value will be imputed with 89.

To do this, I created different data frames for each country by using the filter() function to filter the dataset by country. The table below shows the number of missing values in each country's data frame.

Country Dataframe	**Number of missing values**
Germany	1
UK	1
US	1
Belarus	1
Canada	1
Jamaica	3
Nigeria	1
Egypt	1
Burundi	1

China	1
Bangladesh	1
India	1
Australia	1
New Zealand	1

Table 1: Number of missing values in each country data frame

In total, there are 16 missing values. I used the *fill()* function to impute the missing values. I set the ".direction" parameter of the function to "down" to replace missing values with the value before it.

Time <dbl>	Country Name <chr>	Individuals using the Internet (% of population) [IT.NET.USER.ZS] <dbl>	Urban population [SP.URB.TOTL] <dbl>
2017	Germany	84.39415	63861626
2018	Germany	87.03711	64096118
2019	Germany	88.13452	64294010
2020	Germany	89.81294	64410589
2021	Germany	NA	64461773

Figure 11: Preview of the Germany data frame before imputing missing values

Time <dbl>	Country Name <chr>	Individuals using the Internet (% of population) [IT.NET.USER.ZS] <dbl>	Urban population [SP.URB.TOTL] <dbl>
2016	Germany	84.16521	63592936
2017	Germany	84.39415	63861626
2018	Germany	87.03711	64096118
2019	Germany	88.13452	64294010
2020	Germany	89.81294	64410589
2021	Germany	89.81294	64461773

Figure 12: Preview of the Germany data frame after imputing missing values

Exploratory Data Analysis

```
```{r }
# Violin Plots of Internet Usage
x1 <- germany$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x2 <- australia$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x3 <- uk$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x4 <- us$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x5 <- canada$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x6 <- jamaica$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x7 <- nigeria$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x8 <- egypt$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x9 <- burundi$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
x10 <- india$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`

vioplot(x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, names=c("Germany", "Australia", "UK", "US", "Canada",
"Jamaica", "Nigeria", "Egypt", "Burundi", "India"), col="gold", xlab='Countries', ylab='Count')

title("Individual Using Internet (% of population)")
```
```

Figure 13: R code for exploratory data analysis

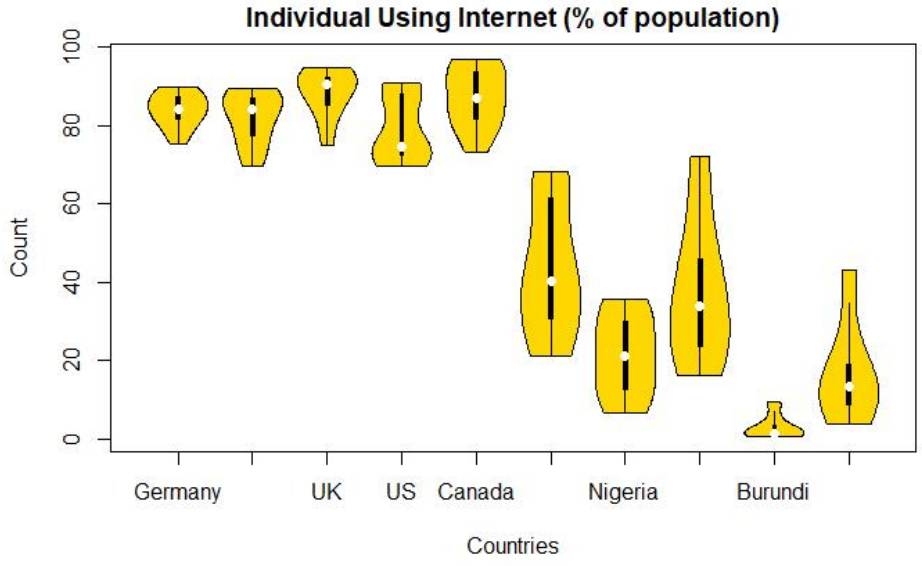

Figure 14: Violin plot showing the percentage of individuals using the internet in developed and developing countries.

While performing exploratory data analysis, I noticed that the percentage of people using the internet in developed countries (like Germany, Australia, the UK, the US and Canada) is higher than those in developing countries (like Jamaica, Nigeria, Egypt, Burundi, India).

4. Analysis

Descriptive Statistical Analysis

I used the *summary()* function in R to get the statistical description of the data. The *summary()* function showed the mean, median, 1st quartile, 3rd quartile, minimum and maximum values of each variable in the dataset. I also created a custom function to calculate the skewness and kurtosis of each variable in the dataset. In the custom function, I made use of the *skewness()* and *kurtosis()* functions from the *moments* package in R to calculate the skewness and kurtosis.

```r
# Create function to calculate the skewness and kurtosis of each variable in each country dataframe

skewness_and_kurtosis <- function(dataframe){
  cat('The skewness of the Individuals using the Internet (% of population) variable is', skewness(dataframe$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`), '\n')
  cat('The kurtosis of the Individuals using the Internet (% of population) variable is', kurtosis(dataframe$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`), '\n')

   cat('The skewness of the Urban population variable is', skewness(dataframe$`Urban population [SP.URB.TOTL]`), '\n')
  cat('The kurtosis of the Urban population variable is', kurtosis(dataframe$`Urban population [SP.URB.TOTL]`), '\n')

  cat('The skewness of the Rural population variable is', skewness(dataframe$`Rural population [SP.RUR.TOTL]`), '\n')
  cat('The kurtosis of the Rural population variable is', kurtosis(dataframe$`Rural population [SP.RUR.TOTL]`), '\n')
}
```

```r
summary(germany)
skewness_and_kurtosis(germany)
```

Figure 15: R code sample of descriptive statistical analysis

Country Dataframe	Minimum Value	1st Quartile	Median	Mean	3rd Quartile	Maximum Value	Skewness	Kurtosis
Germany	75.16	81.64	84.17	83.94	87.31	89.81	-0.4199007	2.313709
UK	75.09	85.19	90.42	88.43	92.26	94.82	-0.9161723	2.92259
US	69.73	72.34	74.70	79.17	87.89	90.90	0.3836022	1.35529
Belarus	19.70	35.72	59.02	56.10	76.78	85.09	-0.216702	1.61648
Canada	73.20	81.65	87.12	87.22	93.67	96.97	-0.2348149	1.891041
Jamaica	21.10	30.73	40.40	43.99	61.64	68.21	0.3325011	1.704211
Nigeria	6.77	12.65	21.00	21.35	29.95	35.50	0.0220789	1.617042
Egypt	16.03	23.60	33.89	37.53	45.94	71.91	0.727593	2.426803

Burundi	0.700	1.055	1.400	2.958	3.380	9.400	1.401331	3.475669
China	16.00	36.30	47.90	46.53	56.75	70.40	-0.2439423	2.200141
Bangladesh	1.80	4.10	11.90	12.34	20.80	24.80	0.2222748	1.393054
India	3.950	8.785	13.500	16.867	19.141	43.000	1.145668	3.226437
Australia	69.45	77.50	84.00	82.02	87.07	89.60	-0.5748551	2.025622
New Zealand	69.76	80.84	84.00	83.55	88.35	91.50	-0.7342846	2.852171

Table 2: Statistical Summary of the Individuals using the Internet (% of population) variable for each country

Country Dataframe	**Minimum Value**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum Value**	**Skewness**	**Kurtosis**
Germany	61940177	62671901	62940432	63204229	63978872	64461773	0.1424339	1.715351
UK	49351705	51315260	53209683	53182308	55174958	56656654	-0.08553573	1.729622
US	241795278	251028926	259430732	259240778	267816372	275050303	-0.07530544	1.792455
Belarus	7005597	7097895	7246350	7246601	7404550	7463845	-0.05625326	1.476367
Canada	26441461	27685179	28781576	28823445	29952302	31229095	0.0814296	1.866847
Jamaica	1472255	1517227	1569365	1572728	1625729	1684526	0.1235373	1.774045
Nigeria	59734513	70573648	82878565	83917375	96564837	111505415	0.153339	1.803146
Egypt	33700834	35975319	38737024	38874943	41608312	44687204	0.1197802	1.773066
Burundi	775530	950534	115926	11941	1416028	1722868	0.2790006	1.870516

			5	53				
China	595670841	668944646	744357517	743092699	819503404	882894483	-0.04579275	1.738455
Bangladesh	40283012	45782482	51817405	52081044	5818775 8	64768559	0.0839205	1.77916
India	353850624	386401610	419567353	421158478	455049784	493169259	0.08390435	1.812309
Australia	17666387	18911562	20095657	20086522	21311046	22228936	-0.05035997	1.760772
New Zealand	3646829	3761594	3896881	3980505	4201504	4445853	0.4475266	1.770094

Table 3: Statistical Summary of the Urban Population variable for each country

Country Dataframe	**Minimum Value**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum Value**	**Skewness**	**Kurtosis**
Germany	18334806	18548328	18755733	18753491	18823081	19432962	0.6124316	2.927956
UK	10669915	11084644	11392615	11374532	11697327	11970758	-0.1914989	1.906289
US	56843442	58163791	58955597	58681030	59405638	59495270	-0.8294279	2.459189
Belarus	1876469	2044338	2202165	2209787	2374845	2555356	0.05777703	1.829968
Canada	6447564	6486930	6655859	6686630	6852858	7017013	0.3819404	1.687992
Jamaica	1288936	1297421	1300903	1300218	1304294	1305789	-0.7480889	2.832596
Nigeria	86605458	90080494	93526366	93415418	96809129	99895289	-0.05895217	1.776674
Egypt	44531290	47669928	51687644	51810501	55824784	59571123	0.06618598	1.698304

Burundi	7086696	7866472	8685036	8737702	9585167	10532561	0.1049699	1.800356
China	529465517	579984096	627502483	626484634	672425354	722214159	-0.01879547	1.836348
Bangladesh	101534935	102319282	102554027	102409427	102655239	102713719	-1.34796	3.738818
India	829358847	855882941	876033415	871789360	890609748	900239774	-0.4586063	1.980958
Australia	3161213	3274324	3380029	3371296	3481228	3535137	-0.2077694	1.760641
New Zealand	576971	605757	619619	627315	655596	677048	0.1553524	1.760843

Table 4: Statistical Summary of the Rural Population variable for each country

Correlation Analysis

I checked the correlation between the "Individuals using the Internet (% of population)" and the "urban population" variables for the different countries. I used the *cor.test()* function from the *ggpubr* package to get the Pearson correlation.

```
```{r}
# Funtion to check the correlation between the Individuals using the Internet (% of population)
# and the Urban population variables for different countries

correlation.urban <- function(dataframe){
  cor.test(dataframe$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`,
           dataframe$`Urban population [SP.URB.TOTL]`, method="pearson")
}
```


```{r}
# correlation analysis for Germany
correlation.urban(germany)
```
```

Figure 16: R code sample for correlation analysis

Country Dataframe	**Correlation**	**95% Confidence interval**	**p-value**
Germany	0.62	0.1560070 - 0.8588322	0.01391
UK	0.91	0.7379278 - 0.9691129	3.043e-06

US	0.86	0.6100960 - 0.9508585	4.848e-05
Belarus	0.99	0.9836650 - 0.9982883	3.122e-14
Canada	0.99	0.9604068 - 0.9958071	1.019e-11
Jamaica	0.97	0.9138836 - 0.9906827	1.701e-09
Nigeria	0.99	0.9843185 - 0.9983573	2.392e-14
Egypt	0.97	0.9121230 - 0.9904844	1.946e-09
Burundi	0.89	0.6981928 - 0.9636977	8.071e-06
China	0.99	0.9639607 - 0.9961897	5.501e-12
Bangladesh	0.98	0.9310375 - 0.9925978	3.917e-10
India	0.92	0.7745001 - 0.9739076	1.086e-06
Australia	0.97	0.9197032 - 0.9913359	1.071e-09
New Zealand	0.91	0.7334804 - 0.9685177	3.417e-06

Table 5: Correlation analysis result for different countries

From the results, there is a very high positive correlation between the “Individuals using the Internet (% of population)” and the “Urban population” variables. It seems that as the urban population increases, the number of people using the internet also increases.

Regression Analysis

I used the plot() function in R to create a scatter plot of the “Individuals using the Internet (% of population)” and the “Urban population” variables.

```
```{r}
plot(uk$`Urban population [SP.URB.TOTL]`,
     uk$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`,
     xlab = 'Urban Population',
     ylab = 'Individual Using the Internet (%)')
```
```

Figure 17: R code sample for the scatter plot of the urban population against the individuals using the internet(%) for the country, UK.

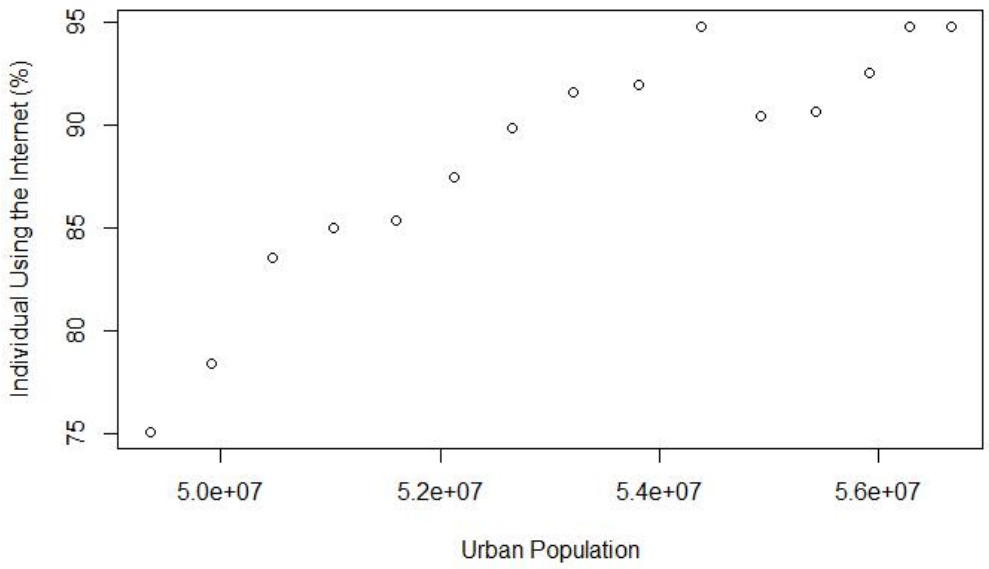

Figure 18: Scatter plot of the urban population against the individuals using the internet(%) for the country, UK.

I fitted the data to a linear regression model with the lm() function in R. I passed the two variables to the *lm()* function to train the linear regression model (Datacamp, 2018).

```
```{r warning=FALSE}
linear_reg.uk <- lm(uk$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]` ~ uk$`Urban population [SP.URB.TOTL]` )

summary(linear_reg.uk)
```
```

Figure 19: R code for fitting data to a linear regression model

```
Call:
lm(formula = uk$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]` ~
    uk$`Urban population [SP.URB.TOTL]`)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6907 -1.9991  0.5243  1.8017  3.6395

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.161e+01  1.545e+01  -2.046   0.0615 .
uk$`Urban population [SP.URB.TOTL]`  2.257e-06  2.902e-07   7.777 3.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.601 on 13 degrees of freedom
Multiple R-squared:  0.8231,    Adjusted R-squared:  0.8095
F-statistic: 60.48 on 1 and 13 DF,  p-value: 3.043e-06
```

Figure 20: Output of the linear regression model

The p-value is 3.043e-06. This is less than 0.05. It means the model is statistically significant. The model can give us a reliable estimate of the number of users using the internet based on the urban population. There is a relationship between the two variables (W3Schools, n.d.).

The Multiple R-squared is 0.8231. The R-squared is high. The linear regression function line fits the data well. This means that the urban population variable can explain 82.31% of the variation in the number of internet users (W3Schools, n.d.).

```
```{r warning=FALSE}
plot(uk$`Urban population [SP.URB.TOTL]`,
     uk$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`,
     xlab = 'Urban Population',
     ylab = 'Individual Using the Internet (%)')
# plot regression line
abline(linear_reg.uk, col='blue')
```
```

Figure 21: R code to plot the regression line on the data

I used the *abline()* function to plot the regression line in the scatter plot. I passed in the linear regression model as an argument to the function and set the *col* parameter to *blue.* The colour of the regression line will be blue.

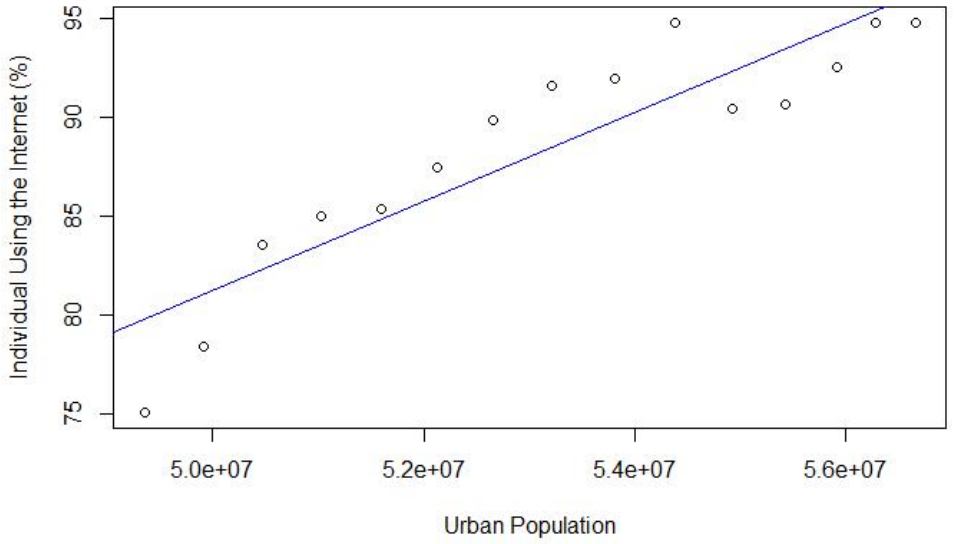

Figure 22: Plotting the regression line on the data

I also trained a ridge regression model on the data. But before that, I pre-processed the data by splitting it into the training and testing set. I used 70% of the data for training and 30% of the data for testing. I used the *sample_frac()* and *anti_join()* functions from the *dplyr* package to do this.

```r
# split data into training set and test set
# use 70% of the data for training and 30% for testing
print("Training Dataset")
training_data  <- uk %>% dplyr::sample_frac(0.7)
print(training_data)
print("Testing Dataset")
testing_data   <- dplyr::anti_join(uk,
                                   training_data, by='Time')
print(testing_data)
```

Figure 23: R code to split dataset to training and testing data

There are 10 records in the training set and 5 records in the testing set. I then separated the features and the target variable. I used the "urban population and the rural population" as the features. I used "Individuals using the Internet" as the target variable.

```r
# Split feature and target variable
# training feature and labels
x.train <- select(training_data, `Urban population [SP.URB.TOTL]`, `Rural population [SP.RUR.TOTL]`)
y.train <- training_data$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`

# testing features and labels
x.test <- select(testing_data, `Urban population [SP.URB.TOTL]`, `Rural population [SP.RUR.TOTL]`)
y.test <- testing_data$`Individuals using the Internet (% of population) [IT.NET.USER.ZS]`
print(x.train)
print(y.train)
print(x.test)
print(y.test)
```

Figure 24: R code to separate the features and target variable

```r
# Fit Ridge model to the training data
ridge_model <- cv.glmnet(as.matrix(x.train),
                         y.train,
                         type.measure='mse',
                         alpha=0,
                         family="gaussian")  # evaluate model with the mean squared error

summary(ridge_model)
```

Figure 25: R code to train a ridge regression model on the training data

I used the *cv.glmnet()* function to train the model. I passed in the training features "x.train" and the training label "y.train" as arguments to the function. The *cv.glmnet()* function need x.train to be a matrix. I used the *as.matrix()* to convert x.train from a

dataframe to a matrix. I set the *type.measure* parameter to *'mse'.* This means I'm evaluating the model with the mean squared error metric. To use ridge regression model I set the *alpha* parameter to 0. I set the *family* parameter to "gaussian" because we want to perform linear regression, not logistic regression (Singh, 2019).

```
```{r}
# use the predict() function to make predictions on the
y_pred <- predict(ridge_model,
                  s=ridge_model$lambda.1se,
                  newx=as.matrix(x.test))
mean((y.test - y_pred)^2)
```
```

```
[1] 19.42906
```

Figure 26: R code to evaluate the model on the test data

I used the *predict()* function to make predictions on the testing set. I passed in the ridge model we trained as the first argument. I also passed in the value of lambda (which is the regularization parameter for the ridge model) as an argument to the *s* parameter. Finally, I passed in the testing data "x.test" as an argument to the *"newx"*. I had to convert the testing data to a matrix first with the *as.matrix()* function. I used the *mean()* function to calculate the mean squared error (Zach, 2020). The mean squared error on the test set is 19.4. This is high, but it is expected since we are working with a small dataset.

Time Series Analysis

I used the ggplot2 library to create a line chart of the urban population of the UK from the year 2007 to 2021.

```
```{r}
# Make a basic graph
ggplot(uk,aes(x=Time))+
  geom_line(aes(y=`Urban population [SP.URB.TOTL]`), color="blue") +
  xlab("Years") +   # add xlabel
  ylab("Urban Population") +   # add ylabel
  labs(title="Time-Series Of the Urban Population in UK") +  # add title
  theme(plot.title = element_text(hjust=0.5))  # centralize the title
```
```

Figure 27: R code to plot the time series graph of the urban population in the UK

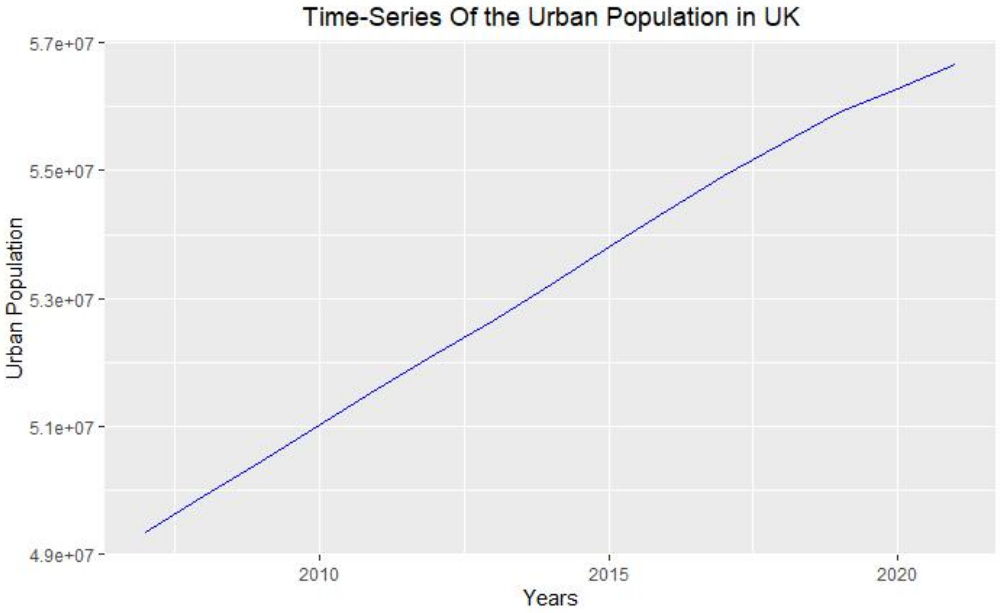

Figure 28: Time series plot of the urban population in the UK

From the line chart above, you'll notice there is an upward trend. As the years go by, the urban population of the UK is increasing.

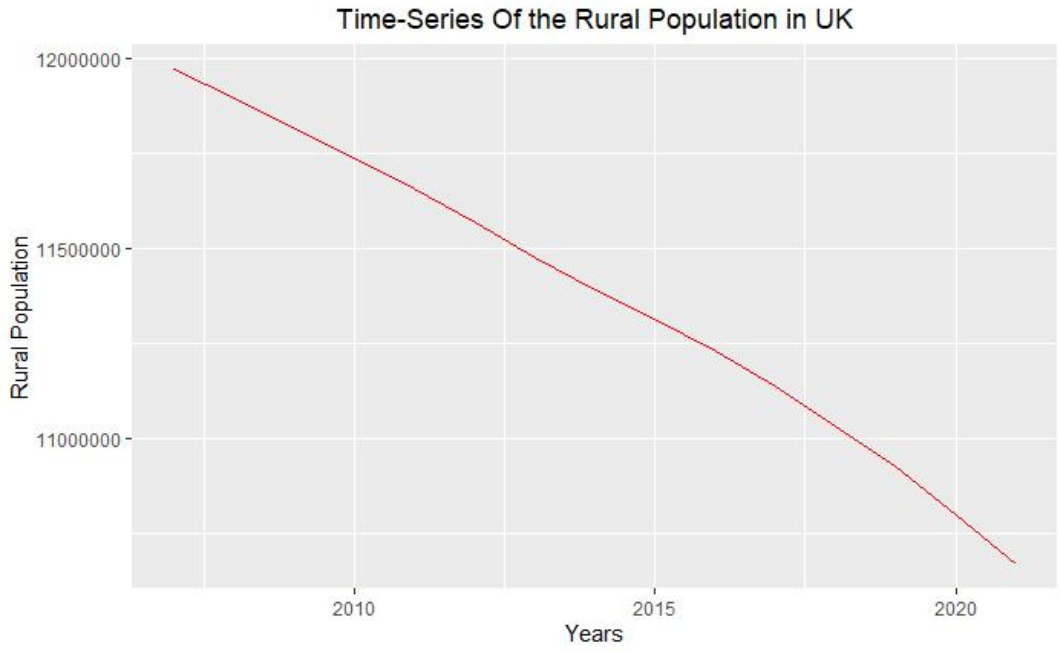

Figure 29: Time series plot of the rural population in the UK

From the line chart above, you'll notice there is a downward trend. As the years go by, the rural population of the UK is decreasing. My goal here is to forecast the UK urban population. I will use the auto ARIMA model for the time series analysis.

Firstly, I convert my dataset to a time series format with the *ts()* function from the *tseries* package. I set the *start* date to the minimum time (2007) and the *end* date to the

maximum date in the dataset (2021). The *frequency* parameter is set to 1 for yearly predictions.

```r
# Convert the dataframe to time-series data
# specify the start time to the earliest date in the time series (2007) and the end to be the latest date in the time series (2021)
# set the frequency to 1 for yearly prediction
uk.ts <- ts(uk$`Urban population [SP.URB.TOTL]`, start=min(uk$Time), end=max(uk$Time), frequency=1)
# confirm that data is a time series data
class(uk.ts)
```

Figure 30: R code to convert the data to time-series format

```r
# check the autocorrelation of the time-series data
acf(uk.ts)
```

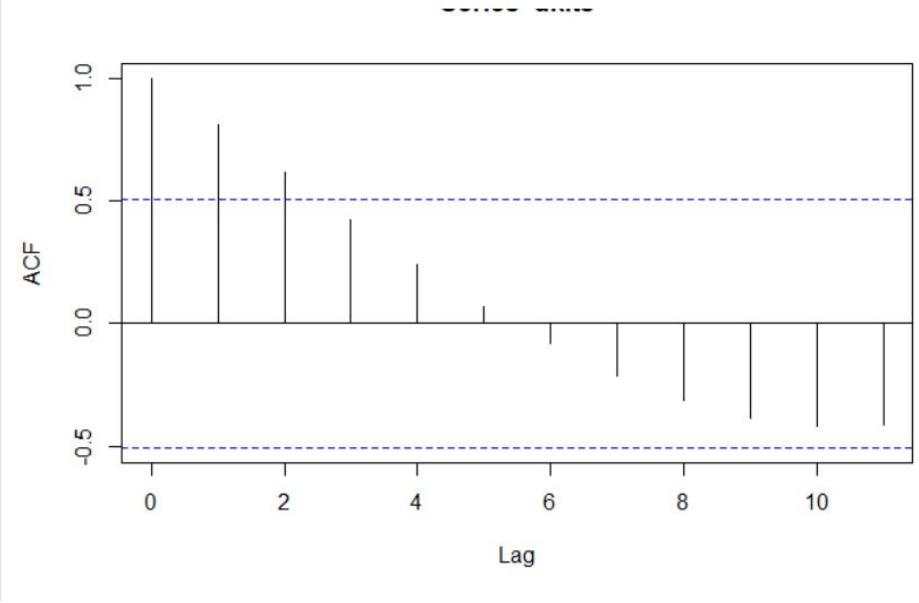

Figure 31: R code to check for auto-correlation in the time-series data.

I used the *acf()* function to check for auto-correlation in the time series data. Some of the vertical lines in the plot above have gone past the blue line. This means there is **auto-correlation** in the data, and the data is **not stationary**. Time series without trends or with seasonality are not stationary since the trend and seasonality will alter the value of the time series at different times. A stationary moment series is one whose attributes do not depend on the time at which the series is examined (Otexts, n.d.).

```r
# use the dickey-fuller test to check if the data is stationery
adf.test(uk.ts)
```

```
        Augmented Dickey-Fuller Test

data:  uk.ts
Dickey-Fuller = -0.46, Lag order = 2, p-value = 0.9767
alternative hypothesis: stationary
```

Figure 32: R code to perform dickey-fuller test

I also performed a dickey-fuller test with the *adf.test()* function to check if the data is stationary. The p-value is 0.9767. It is higher than 0.05. This means it is not stationary. I applied an auto-ARIMA model to the data to convert it from non-stationary to stationary time series data (Chatterjee, 2018).

```r
# use ARIMA
uk.model <- auto.arima(uk.ts, ic="aic", trace=TRUE)
```

```
 ARIMA(2,2,2)                    : Inf
 ARIMA(0,2,0)                    : 314.213
 ARIMA(1,2,0)                    : 315.3133
 ARIMA(0,2,1)                    : 315.2641
 ARIMA(1,2,1)                    : 317.2617

 Best model: ARIMA(0,2,0)
```

Figure 33: R code to fit time series data to auto Arima model

The *ic* parameter is set to *"aic"* and the *trace* parameter is set to TRUE. The time series model will be evaluated using AIC (The Akaike Information Criteria). It is an estimator of prediction error that assesses a statistical model's goodness of fit (Han, 2022).

```r
# check if the data is now stationary
acf(ts(uk.model$residuals))
```

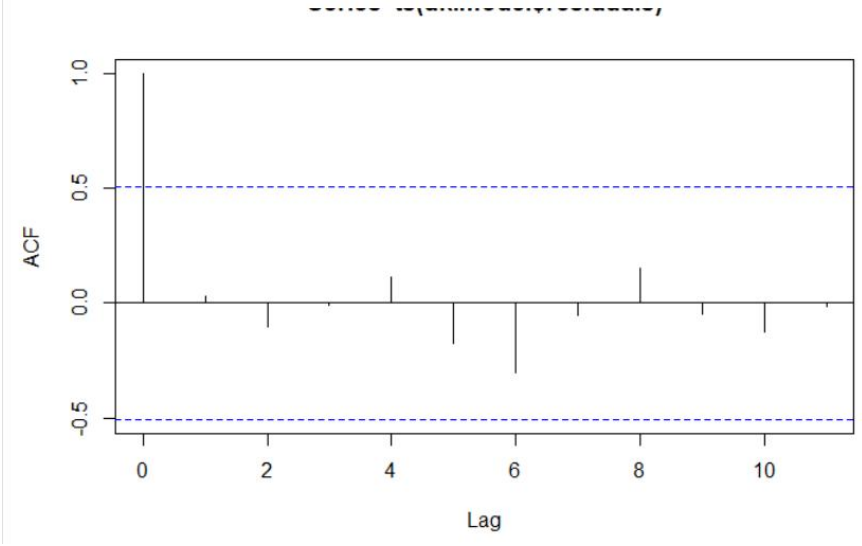

Figure 34: R code to check if the data is stationary

The vertical lines are below the blue line, the data is now stationary. To forecast the urban population for the next 10 years, I used the *forecast()* function. I passed in the auto-arima model as the first argument. I set the level to 95, to get the 95% confidence interval and I set the *h* parameter to 10*1. This will give us the prediction for the next 10 years on a yearly basis.

```r
# set the confidence interval to 95%. We will make predictions for the next 10 years.
uk.forecast <- forecast(uk.model, level=c(95), h=10*1)
uk.forecast
```

	Point Forecast <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2022	57030337	56944041	57116633
2023	57404020	57211056	57596984
2024	57777703	57454813	58100593
2025	58151386	57678724	58624048
2026	58525069	57885081	59165057
2027	58898752	58075541	59721963
2028	59272435	58251367	60293503
2029	59646118	58413565	60878671
2030	60019801	58562958	61476644
2031	60393484	58700235	62086733

Figure 35: R code to forecast the urban population for the next 10 years

To plot the forecast, I used the *plot()* function and passed the *uk.forecast* as the argument.

```r
plot(uk.forecast)
```

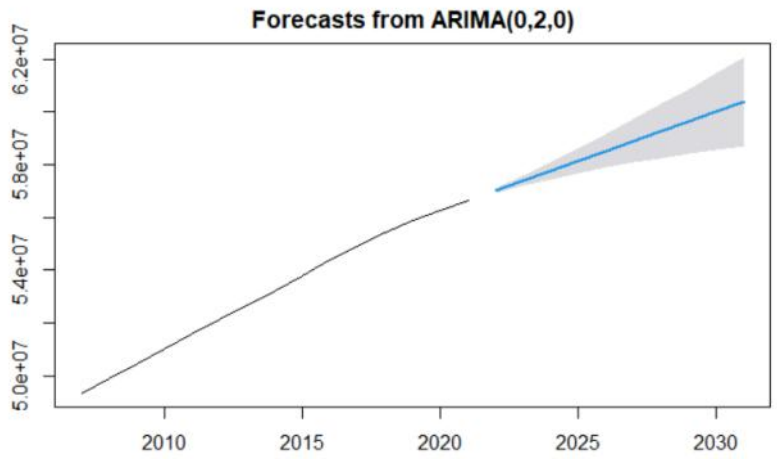

Figure 36: R code to plot the forecast of the urban population for the next 10 years

From the graph, it seems the urban population will keep increasing in the next 10 years. I use the Box.test() function to validate the model. I set the *type* parameter to *Ljung-Box.* The Ljung-Box test determines whether an autocorrelation exists in a time series through a hypothesis test.

```r
Box.test(uk.forecast$resid, lag=5, type="Ljung-Box")
```

```
        Box-Ljung test

data:  uk.forecast$resid
X-squared = 1.299, df = 5, p-value = 0.935
```

Figure 37: R code to validate the forecast model

The chi-squared is 1.299 and the p-value is 0.935. The p-value is greater than 0.05. This means the time series does not have any auto-correlation (KoalaTea, 2021). We cannot reject the null hypothesis that the data is not stationary.

5. Discussion

I started the analysis by checking the correlation between the urban population and the number of internet users. For most of the countries in the dataset, the Pearson correlation was greater than 90%. The correlation for the UK is 91%. This is a high positive correlation.

It means that as the urban population increases, the number of internet users increases as well.

For the rest of the analysis, I focused on the country, the UK. I used a simple linear regression model to see if the urban population can be used to predict the number of internet users in a country. The p-value is 0.000003043. This is under 0.05 which indicates that the model has statistical significance. By using the urban population, the model can provide an accurate estimation of the number of internet users. The two variables are related to one another. The R-squared value is 0.8231, which is very high. It shows that the data are well-fitted by the linear regression function line. This indicates that 82.31% of the variation in the number of internet users can be attributed to the urban population variable. We see that the urban population has an impact on the number of people using the internet.

I also trained a ridge regression model on the data. The data was split into the training set and the test set in the ratio 70:30. I trained the ridge regression model on the training set. The training set had 10 samples, while the test set had 5 samples. I used two features to train the ridge regression model, the urban population and the rural population. My goal was to just use the urban population, but the model needed more than one feature. I used the urban population and the rural population to predict the number of internet users in the UK. I evaluated the model using the mean squared error metric. The mean squared error on the test set is 19.43. The root-mean-squared error is 4.41. This is not bad, considering the model was trained on a small sample of data.

I also performed a time-series analysis to forecast the urban population of the UK. After plotting the data, I noticed an upward trend in the time series data. The urban population was steadily increasing with time. The rural population showed a downward trend. The rural population was steadily decreasing with time. I checked if the time series data was stationary using the auto-correlation function and the dickey-fuller test. The p-value from the dickey-fuller test was 0.9767. It is higher than 0.05, which means the data is not stationary. I performed a time series analysis using auto-ARIMA. I evaluated the time series model with the Akaike Information Criteria (AIC). The best model had an AIC of 314.213.

Through a hypothesis test, the Ljung-Box test establishes whether an autocorrelation exists in a time series. The p-value is 0.935, and the chi-squared is 1.299. The p-value

exceeds 0.05. This indicates that there is no auto-correlation in the time series. The idea that the data are not stationary cannot be ruled out.

6. Conclusions

It is clear from the research that urbanization affects internet usage. There is a correlation between the urban population and the number of internet users. For most of the countries in the dataset, the Pearson correlation was greater than 0.9. This is a high positive correlation. It means that as the urban population increases, the number of internet users increases as well.

Part Three: References and Appendices

[1] Nations Online (2016). The Human Development Index – Going beyond income. Retrieved from http://www.nationsonline.org/oneworld/human_development.html

[2] OECD (2014). Measuring innovation in education: A new perspective, educational research, and innovation. UK: OECD Publishing. http://dx.doi.org/10.1787/9789264215696-en

[3]OECD (2016). Health Inequalities. Retrieved from www.oecd.org

[4] Ololube NP & Kpolovie PJ (2012). Educational Management in Developing Economies: Cases 'n' school effectiveness and quality improvement. Saarbrucken, Germany: LAP LAMBERT Academic Publishing. http://www.amazon.com/Educational-Management-Developing-Economies-Effectiveness/dp/3846589314

[5] Ololube NP; Kpolovie PJ& Makewa LN (2015). Handbook of Research on Enhancing Teacher Education with Advanced Instructional Technology. PA, USA: Information Science Reference (an imprint of IGI Global). ISBN 13: 978146668162; EISBN 13: 9781466681637; DOI: 10.4018/978-1-4666-8162-0. http://www.igi-global.com/book/handbook-research-enhancing-teacher-education/120264

[6] Pew Research Center (2015). World population by income. Retrieved from http://www.pewglobal.org/interactives/global-population-by-income/

[7] Rice, N. and Jones, A.M. Economic analysis of health policies, in S. Glied and P.C. Smith (eds) (2011) The Oxford Handbook of Health Economics. Oxford: Oxford University Press.

[8] Roberts, M.J. et al. (2008) Getting health reform right: A guide to improving performance and equity. Oxford: Oxford University Press.

[9] Roemer, M.I. (1960) Health departments and medical care – a world scanning, American Journal of Public Health, 50: 154–60.

[10] Rosén, M. (2001) Can the WHO Health Report improve the performance of health

systems? Scandinavian Journal of Public Health, 29(1): 76–80.

[11] Agile Alliance. (2018a). What is an Information Radiator? [Online] Available at: https://www.agilealliance.org/glossary/information-radiators/ [Accessed 5 Mar. 2018].

[12] Agile Alliance. (2018b). Daily Meeting. [Online] Available at: https://www.agilealliance.org/glossary/daily-meeting/ [Accessed 9 May 2018].

[13] Card, S.K., Mackinlay, J.D., & Shneiderman, B. (1999). Readings in information visualization - using vision to think.

[14] Chandrasekara, C. (2017). Beginning Build and Release Management with TFS 2017 and VSTS. Berkeley, CA: Apress.

[15] Charters, E. (2003). The use of think-aloud methods in qualitative research - An introduction to think-aloud methods", Brock Education Journal, vol. 12, no. 2, pp. 68-82.

[16] Lehtonen, T., Suonsyrjä, S., Kilamo, T. & Mikkonen, T. (2015). Defining Metrics for Continuous Delivery and Deployment Pipeline. Proceedings of the 14th Symposium on Programming Languages and Software Tools. p. 16-30 15 p. (CEUR Workshop Proceedings; vol. 1525)

[17] Chan, F. (2015). Designing dashboards – visualizing software metrics for Continuous Delivery. KTH Royal Institute Of Technology; Stockham, Sweden.

[18] Roberto T. (2009). Information dashboards and tayloring capabilities - https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8789402 , accessed December 15, 2022

[19] Bernardita C. (2021). 23 Dashboard Design Principles and best practices to enhance your data analysis https://www.datapine.com/blog/dashboard-design-principles-and-best-practices/, accessed December 15, 2022

[20] Banerjee P. (2020) Complete Guide on Time Series Analysis in Python, *Kaggle,* https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python, accessed 30th November, 2022.

[21] CFI Team (November, 2022) Regression Analysis, *Corporate Finance Institute,* https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/, accessed 30th November, 2022.

[22] Chatterjee S., (January, 2018) Time Series Analysis Using ARIMA Model In R, *Data Science Plus*, https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/, accessed 17th November 2022.

[23] Datacamp (July, 2018) Linear Regression in R Tutorial, https://www.datacamp.com/tutorial/linear-regression-R, accessed 14th November 2022.

[24] Emily J. (n.d) What is Correlation Analysis? A Definition and Explanation, *FlexMR Blog*, https://blog.flexmr.net/correlation-analysis-definition-exploration, accessed 14th November 2022.

[25] Gupta, A. (September, 2022) A Complete Guide To Get A Grasp Of Time Series Analysis, *Simplilearn,* https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-time-series-analysis, accessed 30th November 2022.

[26] Han, R (August, 2022) How to evaluate time series models using AIC in R, *Project Pro*, https://www.projectpro.io/recipes/evaluate-time-series-models-aic, accessed 17th November 2022.

[27] Hong, Jinhyun & Thakuriah, Piyushimita. (2018). Examining the relationship between different urbanization settings, smartphone use to access the Internet and trip frequencies. *Journal of Transport Geography*. 69. 11-18. 10.1016/j.jtrangeo.2018.04.006.

[28] KoalaTea (July, 2021) How to Conduct a Ljung-Box Test in R, https://koalatea.io/r-ljung-box-test/, accessed 17th November 2022.

[29] Pew Research Center, (2015) U.S. Smartphone Use in 2015. Retrieved from http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

[30] Philip, L.J., Cottrill, C., Farrington, J., (2015) 'Two-speed' Scotland: Patterns and Implications of the Digital Divide in Contemporary Scotland. https://www.tandfonline.com/doi/full/10.1080/14702541.2015.1067327

[31] OTexts (n.d.) Forecasting: Principles and Practice, https://otexts.com/fpp2/stationarity.html, accessed 17th November 2022.

[32] Singh, D. (November 2019) Linear, Lasso, and Ridge Regression with R, *Pluralsight*, https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r, accessed 15th November 2022.

[33] Steven G., Edward H., Janet E. (October 2014) The Impact of the Internet on Urban Vitality: Does Closeness in Cyber-space Substitute for Urban Space? https://www.uh.edu/~kohlhase/CraigHoangKohlhase_internet_WP_Oct_2014.pdf, accessed 30th November 2022.

[34] W3Schools (n.d.) Data Science - Regression Table: P-Value, https://www.w3schools.com/datascience/ds_linear_regression_pvalue.asp, accessed 17th November 2022.

[35] W3Schools (n.d.) Data Science - Regression Table: R-Squared, https://www.w3schools.com/datascience/ds_linear_regression_rsquared.asp , accessed 17th November 2022.

[36] Zach (November, 2020) Ridge Regression in R (Step-by-Step), *Statology*, https://www.statology.org/ridge-regression-in-r/, accessed 16th November 2022.