POPULATION ESTIMATES & PROJECTIONS

BY:

SRIRAM NARLA
@00646253

Contents

Part One: Dashboard Design

1. Introduction

In the 21st Century, Global concerns of mankind, which encompass not just economic, social, demographic, and ecological issues, are becoming increasingly crucial in the world economy. All of them are intertwined; one leads to various issue. As a result, increase in population leads to resource shortages, a drop in population level and quality of life, and other issues. Simultaneously, why do Western countries pursue policies of demographic stimulation despite continual focus on the problem of global overpopulation, and is it conceivable to link independent demographic concerns to the today's world. In this regard, the world's forecasting of the population is being studied.

In the next 50 years, the total population is probably going to increment from its 7.5 billion to 9-11 billion. The youthful age construction of the populace and the way that in quite a bit of Asia and Africa, ripeness is extremely high making an increment something like another billion practically certain. Practically, all the increment will occur in the creating scene. For the last part of the century, populace adjustment and the beginning of a decrease are reasonable.

In this way, the issue of worldwide overpopulation in the planet makes issues of lack of assets, decline in level. Infant mortality rate is one of the important measures of a society's overall health and is used as significant indicator in UNICEF international comparisons and the Sustainable Development Goals. Infant mortality rates across all the countries have been analysed in the dashboard.

If properly managed, urbanisation can be a powerful growth engine, but it requires attention and solid policies. Few countries achieve high income levels without a substantial amount of urbanisation. Urbanization is a relatively recent phenomenon: only 3% of the total population resided in cities in 18th Century, and less than 15% in 19th Century. More than 50% of world's population now lives in urban areas, and it may increase up to 65% of total population by 2030. Almost all of the urban population growth is now taking place in developing countries, which will be discussed in the dashboard.

The working age population indicates the maximum number of persons who may be expected to be economically productive in a given economy. As a result, the working-age population is defined as the total people in this age group. This indicate whether country's demand for workers (supply) is in balance with the quantity of people available to work (supply). By doing so, it can anticipate how many people will be available to satisfy rising demand if the overall number of available jobs rises or how many will be affected if the total number of jobs falls. The proportion of working-age individuals to more seasoned individuals is probably going to increase generously in all nations, even those that presently have youthful populaces which will be shown in the dashboard.

2. Background Research

Tableau is a data visualisation software that allows you to generate basic oriented graph like data visualisations by taking data from excel sheets, text files, pdf files, JSON files and from servers as well. Tableau, which specialises in attractive graphics, provides insights through simple drag and drop tools, allowing you to quickly evaluate and share essential data. Tableau dashboards can be viewed and operated on a variety of devices, including your laptop, smartphone, or tablet. To make your dashboards mobile-friendly, you do not need to take any further actions. Tableau recognises the device you're using to see the report and adjusts the report accordingly.

Tableau caters to all of your users' demands, regardless of their ability level. With accessible machine learning, statistics, natural language, and smart data prep, augmented analytics solutions allow anyone from data scientists to business users discover insights faster. Tableau users can utilise interactive models and add a variety of formats to it. Data analysts working in many projects at the same time can successfully use all the attributes or measures used in one sheet or workbook to other using an informative, complicated yet easy-to-use dashboard. At a glance, a good dashboard displays actionable and meaningful data. It helps stakeholders understand, analyse, and deliver essential insights by simplifying the visual depiction of complex data. One of the most challenging things to do in a world flooded with data is to provide clear information. On dashboards, displaying only the most relevant data is critical the more information we present, the more difficult it is for users to find what they need.

A dashboard's main goal is to make complex information accessible and digestible. As a result, the data presentation interface should be simple and uncomplicated to reduce users' cognitive burden and time spent searching.

The information architecture should prioritise the most important data while enabling access to additional or supporting measures. It's best to build a progressive drill-down approach that starts with a broad overview and then gets more specific this helps with data prioritising and clarity. Every good dashboard design is built on the foundation of effective communication. Predicting possible scenarios in which users may find themselves can help us gain a better grasp of their situation.

User research aids in the creation of an environment in which relevant, clear, and concise material is provided to users. This allows people to focus on the material and data they need rather than on how to use and obtain it. Some dashboards must function or be easily modifiable for users of various roles seeing the same basic dashboard. User research is crucial since it aids in the identification of the user's goals, mental models, context, and pain areas. These are significant influences on the final dashboard design. A designer must distinguish between the various user categories and determine where their aims are similar and where they differ and what data is most useful to one user type than another. They must decide whether a separate layout is required for each user type or if a solution for a broader use case exists.

With this in mind, it's a good idea to start with basic wireframes and work our way up to prototypes that can be tested with real users throughout the user research phase. A short user

research phase with only five users can yield truly significant knowledge, and it will save you a lot of effort in the long run.

Progressive disclosure is a technique for decreasing clutter and keeping a user's attention. Creating a system of progressive disclosure aids in the creation of a user-centric environment, which aids in the prioritisation of user attention, the avoidance of errors, and the saving of time. It also lets customers to concentrate on the primary aspects that are important to them rather than being compelled to go over all of the options including those they don't need or aren't interested in. When a system is based on feature prioritisation, progressive disclosure is a dashboard design best practise that will significantly reduce mistake rates, improve efficiency, and assist users in better understanding dashboards.

Dashboards are an effective way to communicate data and other information, especially when designed with a user centred, goal-centric approach that adheres to best practises in dashboard design and data visualisation. Despite the fact that each dashboard is unique and has its own set of goals, requirements, and limits, adhering to these basic principles can help you create excellent designs regardless of the details:

- To begin, sympathise with your user types and learn about their goals.
- Use suggestive graphics, labelling, progressive disclosure techniques, and animation to tell a clear story to users.
- By using user research approaches, you can make complicated things simple.
- In a drill-down system, reveal data and information at the proper time.
- To express information in a meaningful way, use data visualisation.

3. Exploration of Data Set

The Dataset "**Population estimates & Projections**" has been taken from the World Bank Data bank https://databank.worldbank.org/source/world-development-indicators . The Data Attributes present in the Dataset are Total population, Male Population, Female Population, Urban & Rural Population growth Rate, Age Dependency ratio, Life Expectancy at Birth, Total Infant Mortality Rate (per 1000 live births), Male & female Infant Mortality Rates (per 1000 live births) for 14 countries for the time frame 2007-2019 (13 years).

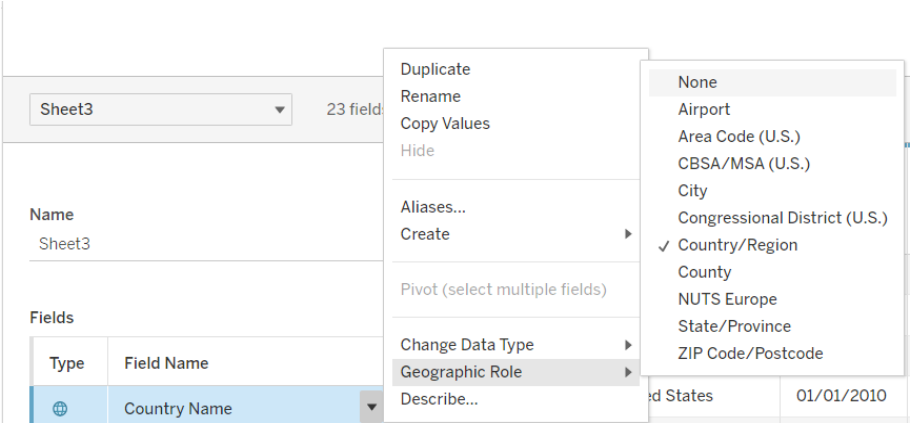
Attributes Population contains the population of the country in the given year. The Age dependence ratio is measured between individuals who are not working to those who are working. It's used to figure out how much strain the productive population is under. A low dependency ratio indicates that enough people are employed to support the dependent population and a high dependency ratio indicates that working people are under more financial hardship and that political instability is a possibility. The mortality rate of a population can be known by attribute life expectancy at birth. It represents the total mortality in all age categories.

The infant mortality rate is a count of infant deaths in children under the age of one year. It is a significant indication of a county’s complete health.

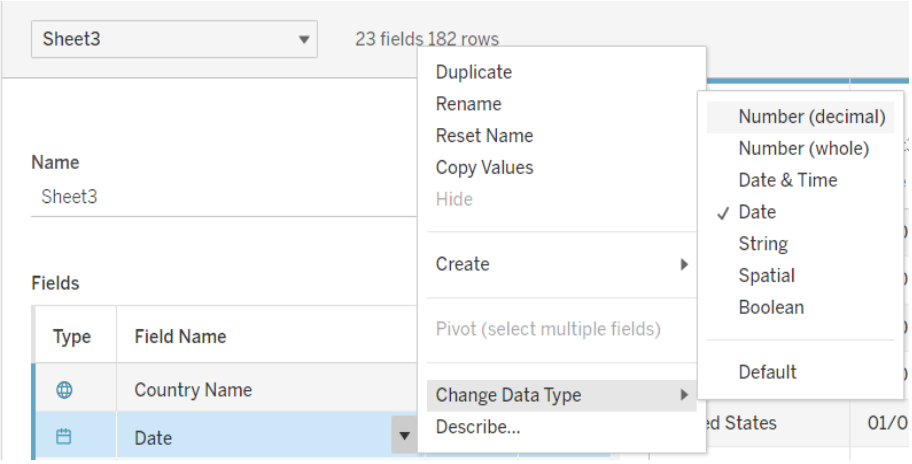
Urban & Rural population growth rates indicate the percentage increase/decrease in the Urban/Rural Population.

Data cleaning has been done on Tableau Public. Steps taken for data preparation are shown below:

- Changed the Field Country Name to the Geographic Role Country/region.

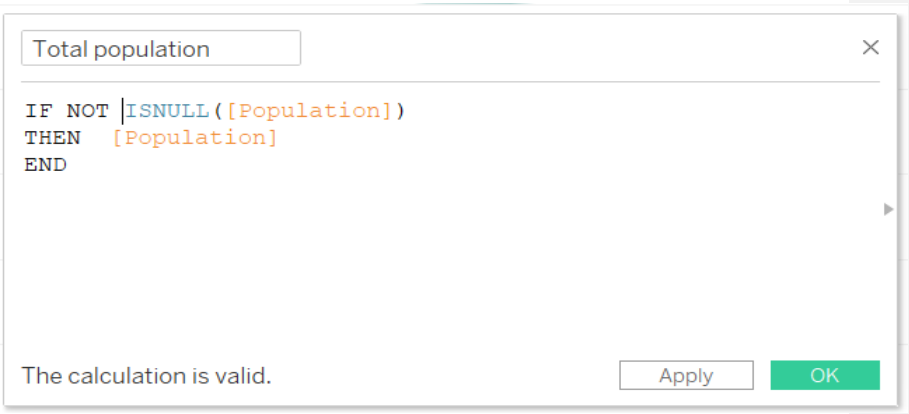


- Changed the Year Field from Data Type String to Date

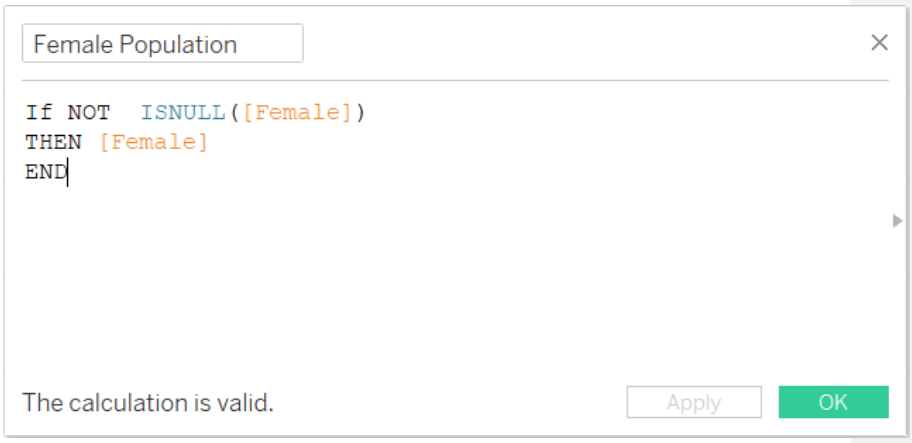


- Created a Calculated Field for each attribute to find out if any Nulls are present in it as shown below for some attributes.

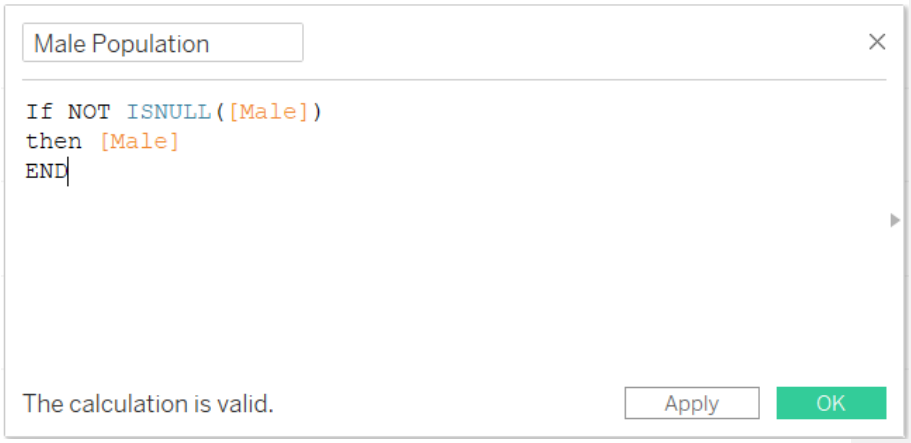
Total Population:



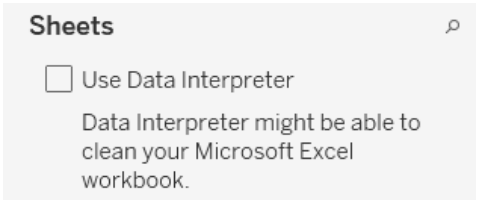
Female Population:



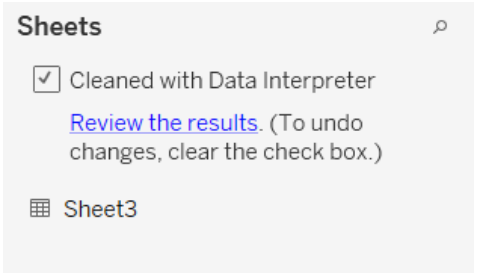
Male Population:



- Used Data Interpreter to clean the Excel Workbook.
  Before:



After:



- Results obtained:

Key for Understanding the Data Interpreter Results

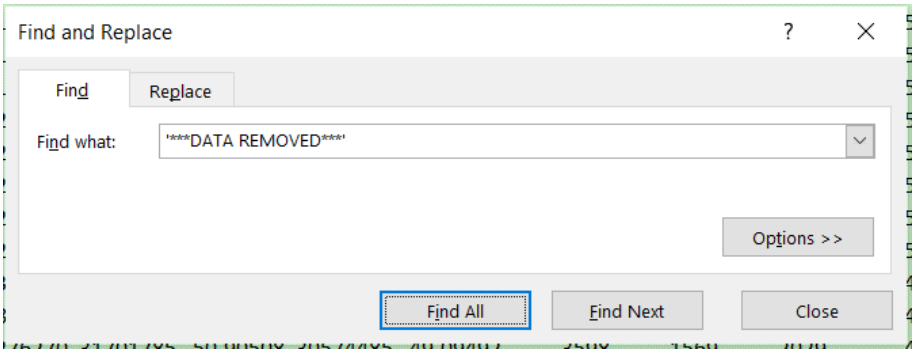
Use the key to understand how your data source has been interpreted.
To view the results, click a worksheet tab.
Note: Tableau never makes changes to your underlying data source.

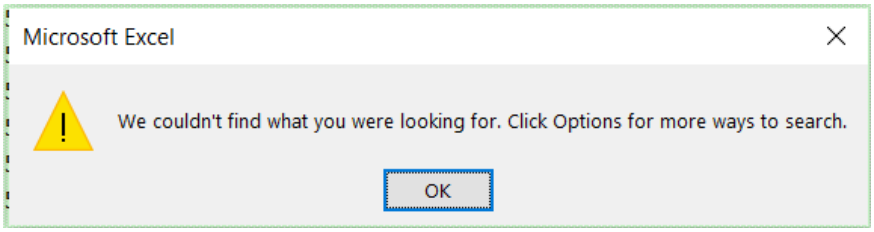
Key:

- **Data is interpreted as column headers (field names).**
- **Data is interpreted as values in your data source.**
- **Data derived from an Excel merged cell is interpreted as value in your data source.**
- **Data is ignored and not included as part of your data source.**
- **Data has been excluded from your data source.**
- **Note: To search for all excluded data, use CRTL +F on Windows or Command F on the Mac, and then type '***DATA REMOVED***'.**

As shown in above screenshot Tableau will be showing "***DATA REMOVED***" for all the values that are removed and a Green coloured box for values in the data source and red coloured box for the Column headers.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Country N	Time	Rural popu	Urban pop	Urban pop	Rural popu	Populatior	female	Populatior	male	Populatior	Number of	Number of	Number of
2	United Sta	2007	0.093157	1.163068	80.269	19.731	3.01E+08	1.53E+08	50.63678	1.49E+08	49.36322	27269	11990	15279
3	United Sta	2008	0.085657	1.156186	80.438	19.562	3.04E+08	1.54E+08	50.63071	1.5E+08	49.36929	26796	11831	14965
4	United Sta	2009	0.014133	1.08529	80.606	19.394	3.07E+08	1.55E+08	50.62335	1.51E+08	49.37665	26195	11541	14654
5	United Sta	2010	-0.03	1.035345	80.772	19.228	3.09E+08	1.57E+08	50.61319	1.53E+08	49.38681	25523	11278	14245
6	United Sta	2011	-0.17177	0.939505	80.944	19.056	3.12E+08	1.58E+08	50.5996	1.54E+08	49.4004	24829	11002	13827
7	United Sta	2012	-0.18899	0.949566	81.119	18.881	3.14E+08	1.59E+08	50.5834	1.55E+08	49.4166	24175	10712	13463
8	United Sta	2013	-0.26505	0.91451	81.299	18.701	3.16E+08	1.6E+08	50.5663	1.56E+08	49.4337	23594	10445	13149
9	United Sta	2014	-0.25542	0.959431	81.483	18.517	3.18E+08	1.61E+08	50.55067	1.57E+08	49.44933	23109	10263	12846
10	United Sta	2015	-0.28426	0.966675	81.671	18.329	3.21E+08	1.62E+08	50.53822	1.59E+08	49.46178	22725	10105	12620
11	United Sta	2016	-0.32286	0.958268	81.862	18.138	3.23E+08	1.63E+08	50.52931	1.6E+08	49.47069	22444	9979	12465
12	United Sta	2017	-0.45384	0.871785	82.058	17.942	3.25E+08	1.64E+08	50.52338	1.61E+08	49.47662	22234	9884	12350
13	United Sta	2018	-0.58325	0.767437	82.256	17.744	3.27E+08	1.65E+08	50.52001	1.62E+08	49.47999	22023	9787	12236
14	United Sta	2019	-0.69526	0.701868	82.459	17.541	3.28E+08	1.66E+08	50.51849	1.62E+08	49.48151	21779	9678	12101
15	United Kin	2007	-0.6455	1.12719	80.479	19.521	61322463	31253394	50.96565	30069069	49.03435	3686	1608	2078
16	United Kin	2008	-0.64731	1.131869	80.757	19.243	61806995	31481382	50.93498	30325613	49.06502	3657	1594	2063
17	United Kin	2009	-0.67773	1.095105	81.031	18.969	62276270	31701785	50.90508	30574485	49.09492	3598	1569	2029
18	United Kin	2010	-0.65506	1.117771	81.302	18.698	62766365	31931743	50.87397	30834622	49.12603	3511	1535	1976
19	United Kin	2011	-0.66217	1.110599	81.57	18.43	63258810	32161519	50.84117	31097291	49.15883	3405	1483	1922
20	United Kin	2012	-0.76397	1.022146	81.837	18.163	63700215	32364469	50.80747	31335746	49.19253	3296	1436	1860
21	United Kin	2013	-0.80002	0.993032	82.102	17.898	64128273	32560344	50.77377	31567929	49.22623	3196	1395	1801
22	United Kin	2014	-0.74388	1.056285	82.365	17.635	64602298	32780144	50.74145	31822154	49.25855	3108	1361	1747
23	United Kin	2015	-0.6987	1.108748	82.626	17.374	65116219	33021393	50.71147	32094826	49.28853	3038	1330	1708
24	United Kin	2016	-0.74993	1.072052	82.886	17.114	65611593	33254646	50.6841	32356947	49.3159	2981	1310	1671
25	United Kin	2017	-0.83371	0.988959	83.143	16.857	66058859	33464674	50.65887	32594185	49.34113	2932	1291	1641
26	United Kin	2018	-0.91836	0.912161	83.398	16.602	66460344	33652378	50.63527	32807966	49.36473	2888	1272	1616
27	United Kin	2019	-0.97762	0.868231	83.652	16.348	66836327	33827559	50.61253	33008768	49.38747	2842	1252	1590
28	India	2007	1.029591	2.642483	29.906	70.094	1.18E+09	5.68E+08	47.96952	6.16E+08	52.03048	1396555	674831	721724

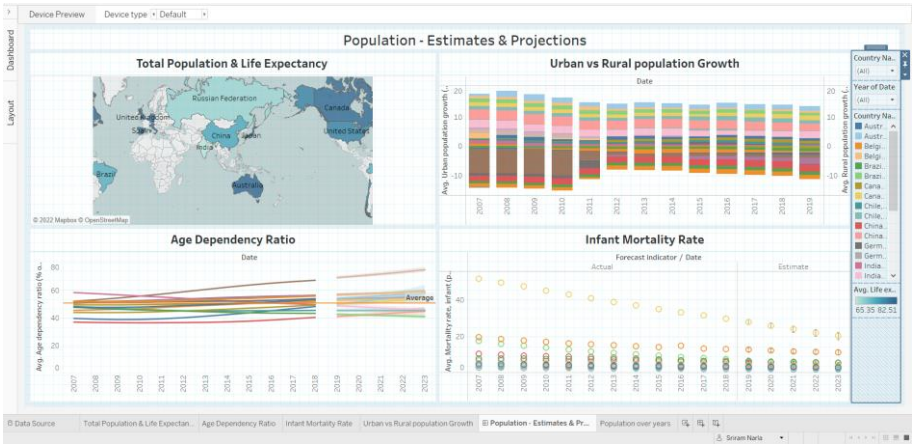




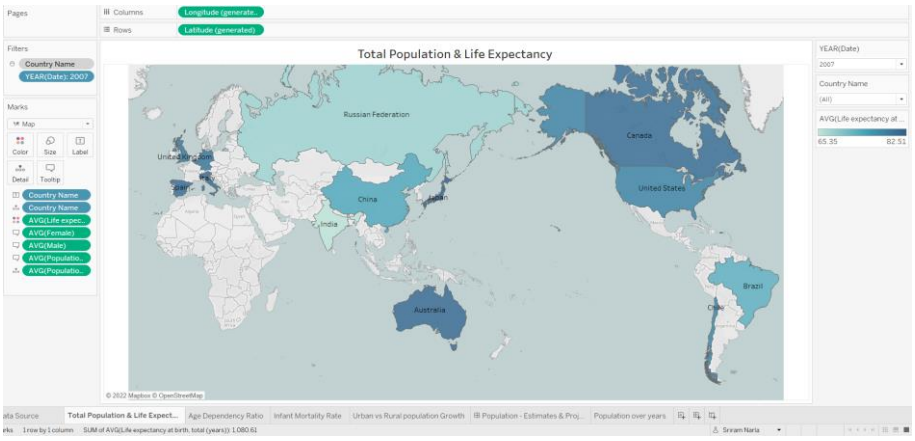
The above screenshot shows the Null Check in dataset. If Nulls are present, we can replace them with either the average/mean of the values. We can also neglect the Nulls by finding the percentage of total Null values to the total number of observations. If the percentage is less, Then Nulls can be neglected and can be deleted. The above screenshot shows that Dataset does not contain the Nulls and there is no excluded data in the excel sheet that Tableau generated after interpreting.

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

Interactive Dashboard:

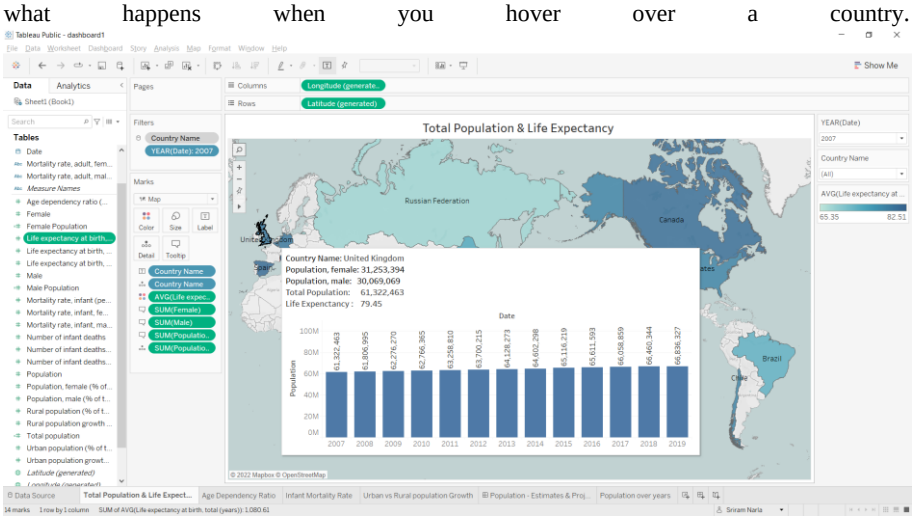


The first sheet in the Dashboard Shows the Total Population of various countries over the years & Life Expectancy over the years.



This is an interactive sheet with filters for Country Name and Years. Life Expectancy at birth measure is mapped to the colour field and countries have been grouped in the colour light blue to the dark blue based on the values included in the dataset. As shown in the legend, it ranges from 65.35 to 82.51.

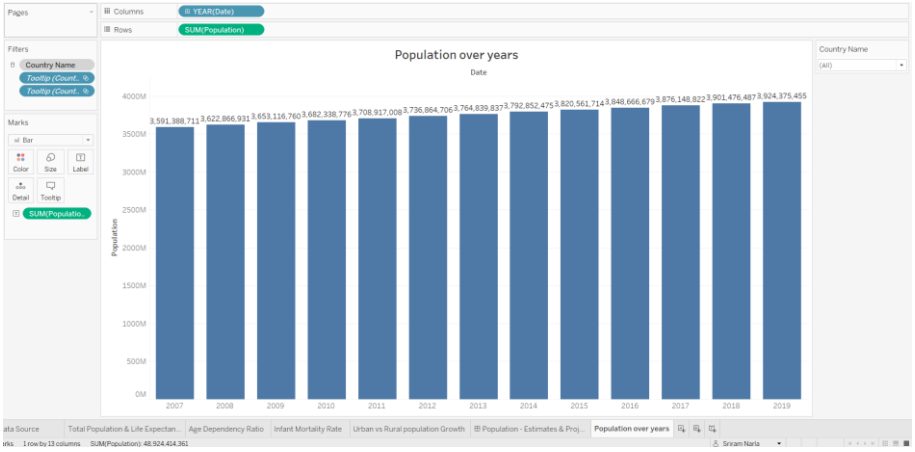
The filter in the above screenshot is for the year 2007, while the screenshot below demonstrates what happens when you hover over a country.



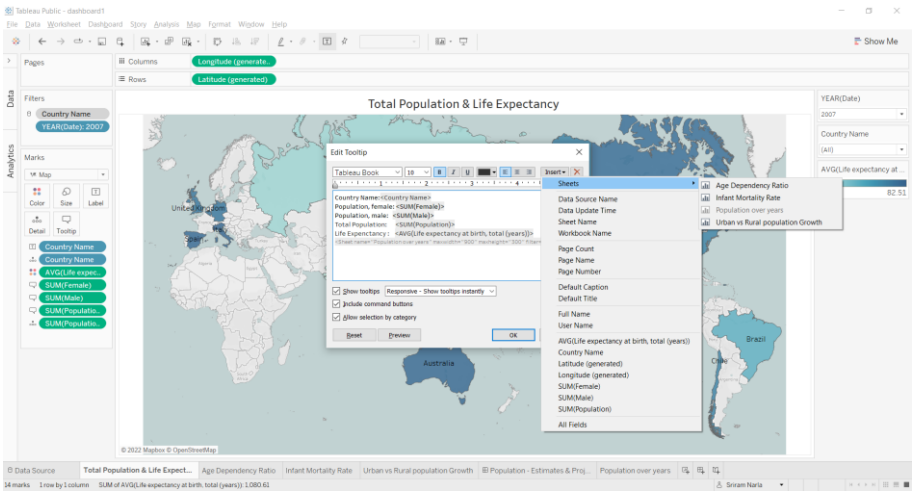
It's an interactive tooltip that shows the population of any country over time when you hover over it. The tool tip was used to import the sheet. For this purpose, I have produced another page which depicts the population over the years. The bar graph depicts the total population over time for all countries. Country has been used as a filter in this sheet. When we choose a country, it displays a bar graph of population changes over time. This population over time sheet has been added to the Total Population & Life Expectancy sheet's tooltip.

Population over the years sheet:

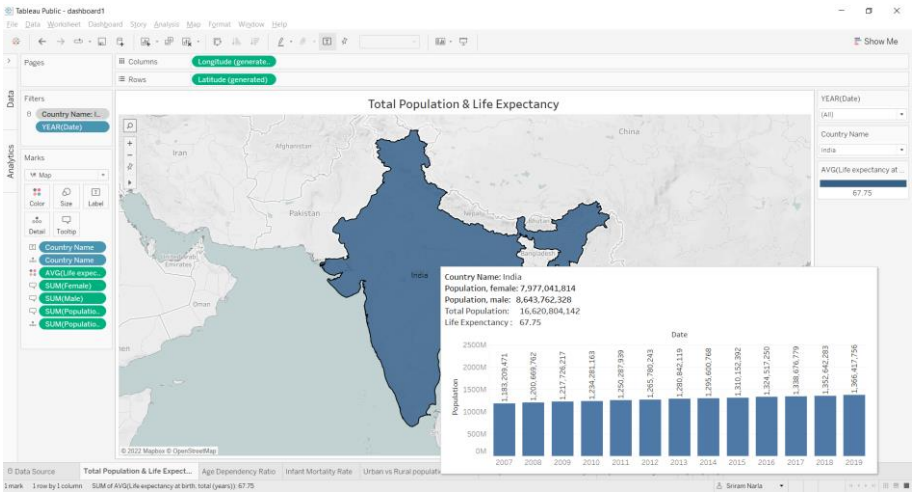
Commented [GV1]:



Sheet Importing has been done as shown below:

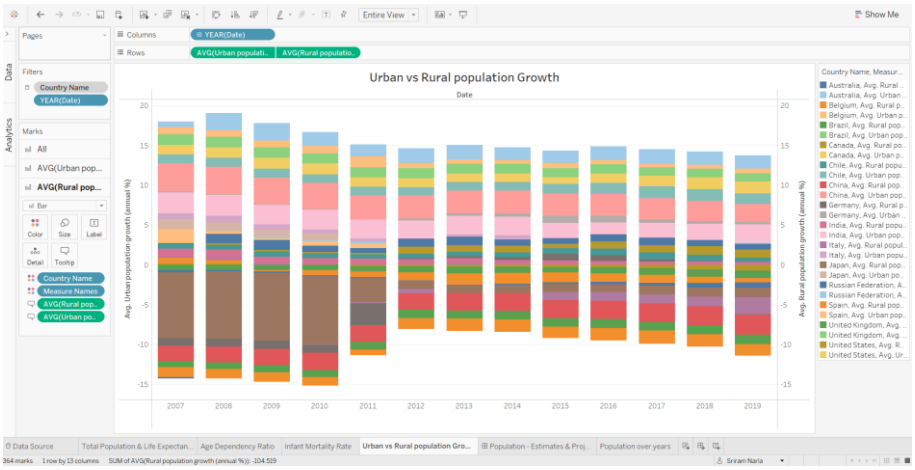


When you select a country in the filter the interactive sheet changes as shown below.

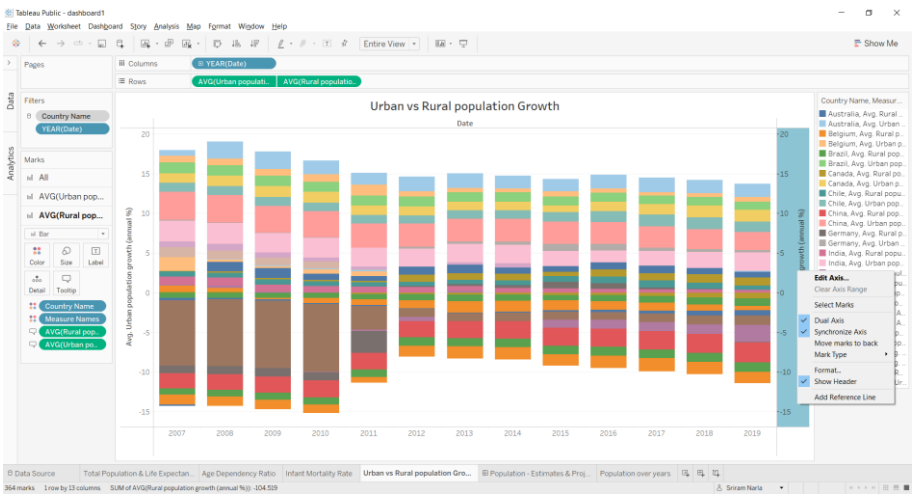


One can easily find the country's population details along with the Life expectancy based on applying the filter.

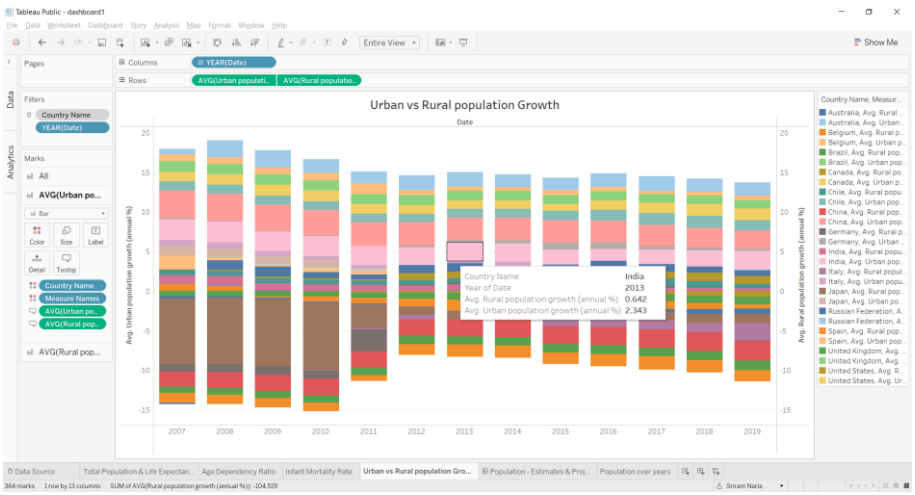
Second sheet in the dashboard shows the Urban and Rural population growth rates over the period of time. Stacked bar chart is used to show the values of the growth rates for each country in a year as shown below.



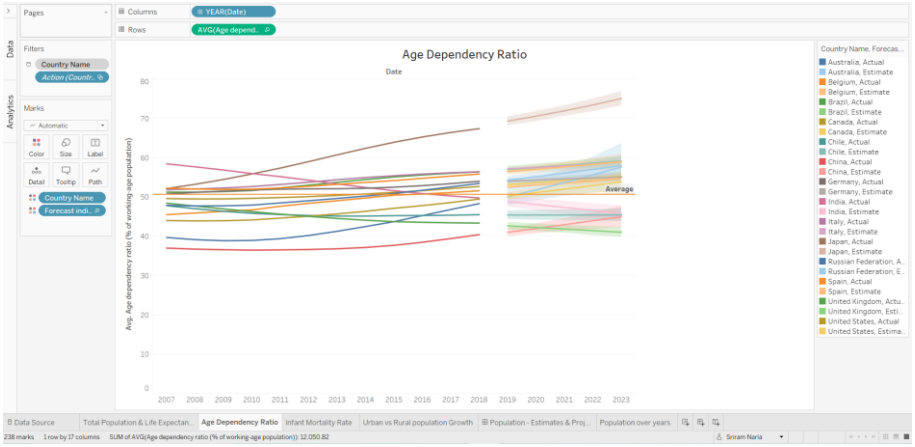
The dual axes approach is employed in this sheet. This sheet was created with Time in the column and Urban and Rural Population Growth Rates in the row. The stacked bar chart is formed when the country name is assigned to the colour. However, the Y axis of the graph will have various values depending on the rate of urban and rural population growth. When we tick the Synchronize Axis, this is synchronised. Following that, the axes will be automatically changed based on both the urban and rural population growth rates.



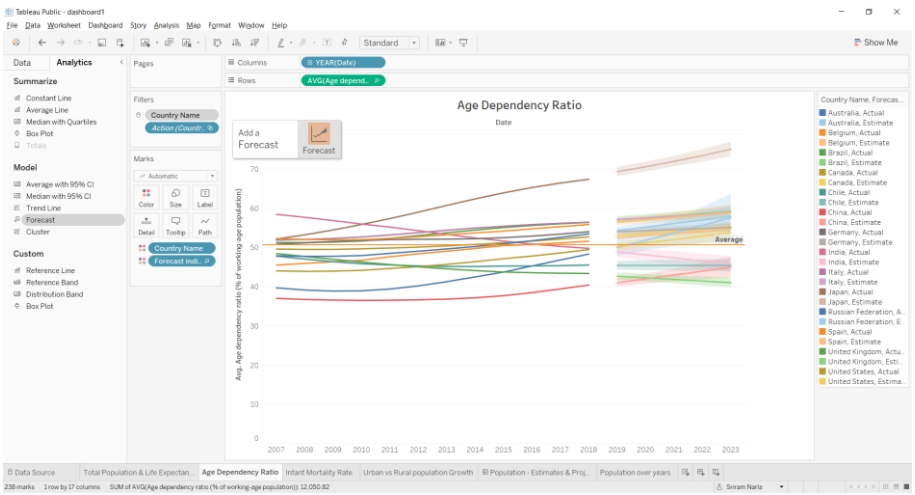
When you hover over any of the bar it shows the country name, year and percentage of growth in Urban population & percentage of growth in Rural population for the year selected as shown below.



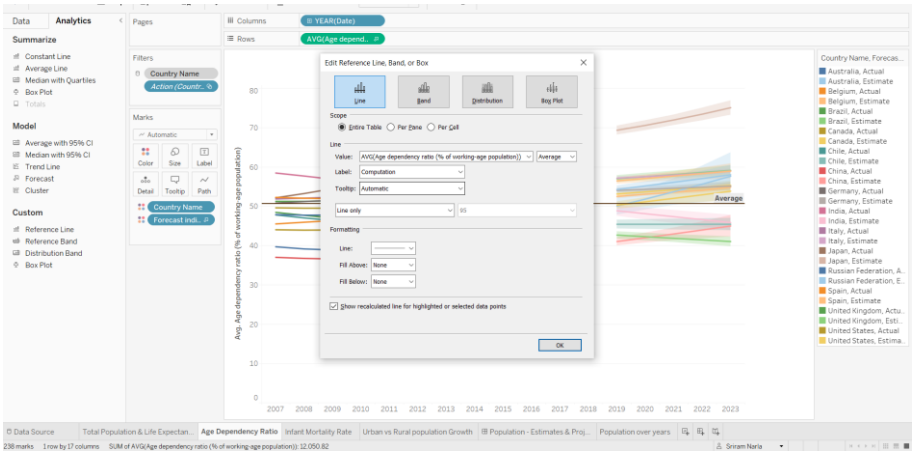
Sheet 3 shows the Age Dependency Ratio for various countries over the years. The continuous line chart has been developed by giving Time (Year) to column and Age dependency Ratio to the Row. Country Name has been given to the colour so that the continuous lines for each country over the period of time has been generated.



Predictive analysis has been used in this sheet. Predictive analysis does the forecasting based on the historical data present in the dataset. This has been done as shown below.



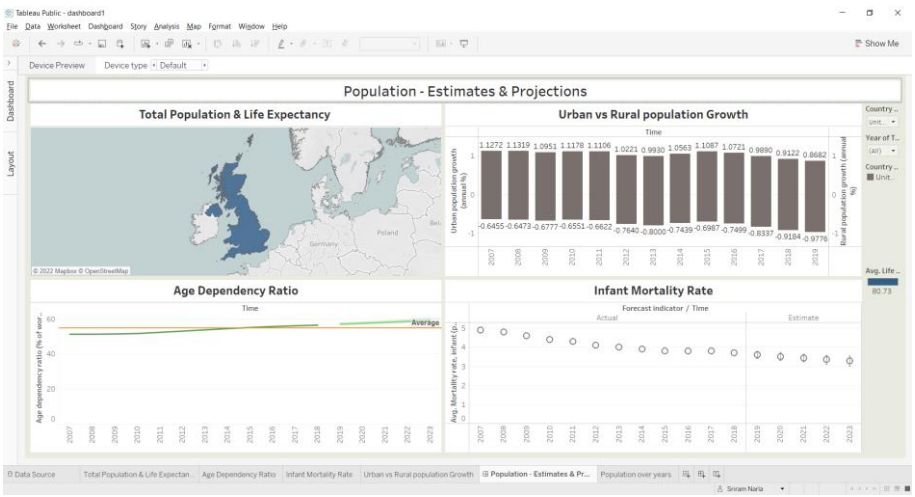
An average Reference line has been drawn in the sheet. Average line has been drawn with the 95% confidence limit. This average reference line has been generated from the analytics tab as shown below.



Infant Mortality rate has been shown in another sheet. Year has been dropped to column & Mortality Rate is in Row Shelf. Measure Mortality rate is based on average. Country name is dropped to colour and in the Marks shape circle is selected to make it a bubble chart.

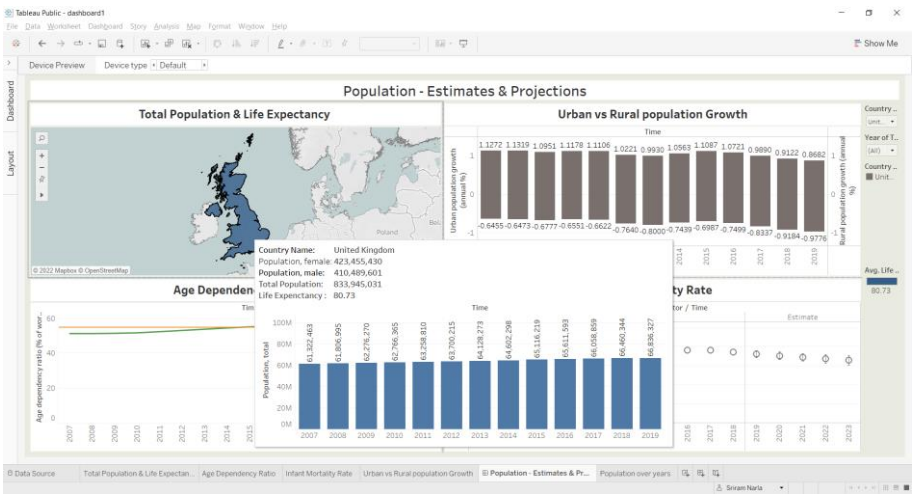
Predictive analysis has been done in this sheet as well for the next five years. It is based on the last 13 years data present in the dataset.

Dashboard when one country is selected in filter, interactive dashboard will be changed as shown below.

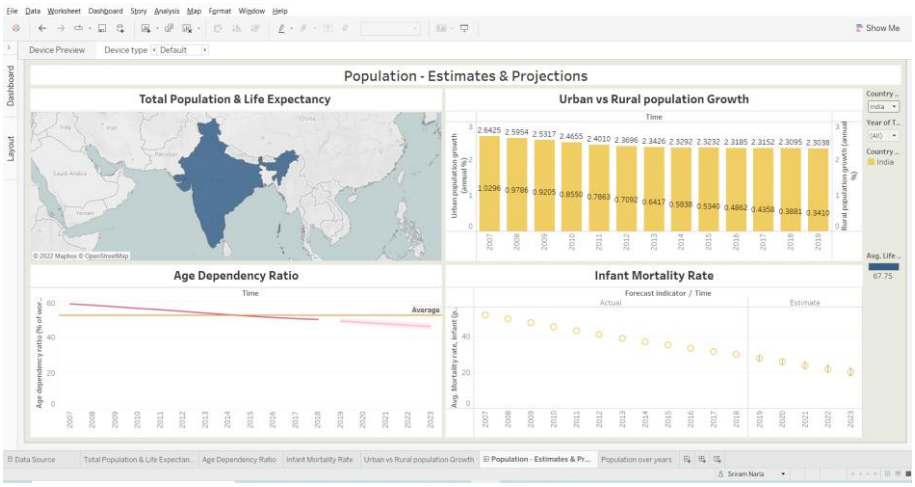


The population estimates and projections for the United Kingdom are shown in the graphic above. The age dependence ratio is increasing over time, while the infant mortality rate is decreasing, which is a positive indicator. Over time, the urban population has constantly increased while the rural population has decreased. This demonstrates that urbanisation is moving at a breakneck speed.

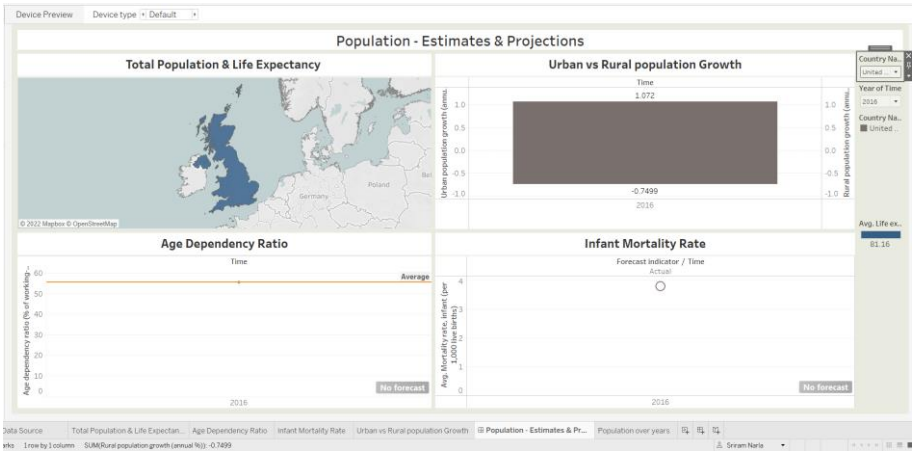
Hovering over the country will show the population of the United Kingdom over time, as illustrated in the graphic below. It plainly demonstrates that the population has been rapidly growing. Over the years, the average life expectancy has been found to be 80.73 years at birth. This figure is based on the average of all of the country's life expectancy over all of its years.



When the country India is selected, the graphic below will appear. Unlike other countries, India's population is rapidly increasing. The population of both urban and rural areas is rapidly growing. The infant mortality rate is falling as sanity and effective hospitalisation services improve. The age dependency ratio is also dropping, indicating that the demand on working individuals is reducing. The average life expectancy has been calculated to be 67.75 years. According to the predictive research, the age dependency rate will be 45.85 percent by 2023, and the infant mortality rate will be 20.08 per 1000 births.



When the year filter is used, all countries' population estimates and projections for that year are displayed. If you apply a country filter, the results will look like this.



Relational modelling can only be done if we have two tables of data. We can do it by linking up the data. Since, I have only one table of data, Relational modelling is not done here.

5. Discussion

During the predicting process, it was discovered that population increase will continue in the first half of the twenty-first century. This expansion will be exceedingly uneven between continents and across countries. As a result of globalisation & urbanisation processes, all countries around the world will be entering the demographic transition stage. In developing countries, population growth will continue which will result in poverty and society imbalance and people starts to migrate for the survival of basic needs.

These findings have significant policy consequences. Environmental, economic, health, governmental, and social challenges can all arise as a result of rapid population growth in these countries. These findings suggest that increased investments in family planning programmes to meet the unmet need for contraception and in girls' education could help to temper the expected high population expansion. In order for faster changes to occur, current investments would need to be increased. It should also be highlighted that the forecasts do not account for any negative feedback from rapid population growth's environmental repercussions. The influx of several billion people into low-income countries could result in significant resource constraints, which could have an impact on population size due to unexpected mortality, migration, or reproductive consequences.

The net effect of urbanisation on households is an increase in average real income. Urban regions provide superior revenue production prospects for people and their households, whether through formal employment or informal sector activities. It is also obvious that in developing countries, the prospect of increased salaries is the primary driver of rural-to-urban

migration. Towns and cities are widely acknowledged as the engines of national economic growth, owing to the agglomeration economies that characterise them. Initially, urbanisation resulted in higher mortality rates in cities than in rural regions, owing to severe overcrowding and poor sanitation in developing countries.

A high dependency ratio is a major issue. Governments must either raise taxes, cut benefits, or increase debt. None of this is a good thing. As a result, it is vital to seek solutions. The following are some examples of solutions to a higher dependency ratio: Increase the retirement age and allow inflation to take its course. Costs of Erode Encourage the immigration of young people. Enhance Economic Growth. A reduction in infant mortality is a positive sign for any country. This has been accomplished by enhancing hospital facilities and implementing government-proposed plans that have delivered positive results.

6. Conclusion

The results of population growth projections and their repercussions presented in this work allow for the formulation of two difficulties that necessarily arise as a result of the forecast. The first issue is to recognise and impose restrictions on the development of goods manufacturing prospects due to natural resource constraints. The second issue is the natural population decrease in those countries where it has already started, which, in current times, may not be regarded as a collapse to the country. More research into all issues will help in determining global population dynamics through the middle of the twenty-first century.

The urban population is rapidly increasing as individuals migrate from rural areas in search of a better quality of life. Population fertility rates and average family sizes are both reduced as a result of urbanisation. This is largely due to the behavioural and lifestyle changes that come with urbanisation, such as improved education, a later age at first marriage, increasing female employment, and higher contraceptive use rates. Furthermore, the cost of caring for children's many requirements, along with a desire for better living conditions and a greater quality of life, tends to deter city dwellers from having big families.

Infant mortality is a key measure of a country's overall health. It is in the downward trend over the years. Basic necessities such as education, purified water, family planning, increase in awareness about the nutrition during pregnancy must all be provided in order to reduce infant death rates. Infant mortality rates may improve as a result of health efforts aimed at preventing preterm birth and improving prenatal care. In industrialised countries, addressing gaps in access to health care must also be a priority.

For all countries, the rise in the age dependence ratio is a cause for concern. Consumption, saving, government spending, taxation, and growth rates will all have substantial ramifications for macroeconomic management. Working age people will see a considerable portion of their labour income reallocated to the elderly in a country with a high rate of age dependency.

Part Two: Statistical Analysis

1. Introduction

Statistics is a science which performs the collection of data, interpreting the data and validating as well. Statistical data analysis is nothing but performing the various statistical operations on the data. Statistics is a discipline that can assist us in comprehending how to use this data to accomplish the following goals: Improve your knowledge of the world around you. Make data-driven judgments. Data may be used to make future predictions.

We need to employ statistical tools to undertake statistical data analysis, something a layperson can't do without knowing statistics. Many software tools, such as Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS), can be used to analyse data in statistics.

The study of population variables includes not only the variables themselves, but also the relationships between them, such as economic, social, biological, political and geographical variables, as well as the interrelationships between them. It takes into account both qualitative and quantitative features of human population.

In the past, study about population were not given much importance. However, in today's world, people from all walks of life demand demographic data, and its relevance is growing by the day. Its research is critical in a variety of fields, including social work, economic development, political work, various administration work, and the creation of laws and regulations. In this regard statistical analysis on the population dataset has been done.

The objective for this statistical analysis is similar to the dashboard design. To know the relation between the variables in the dataset, effect of infant deaths on the population and rural population growth rate analysis is being done.

2. Background Research

SAS Enterprise Guide (EG) is a more GUI-like IDE that includes wizards to help you write code for different procedure. SAS Enterprise Guide guides newcomers through unfamiliar methods and solves some issues for seasoned programmers.

Auto-complete features in the new programme editor include library and table name suggestions, among other things. Wizards aid a programme by abstracting some elements, such as graph colour choices.

SAS Software has the following features:

- A user-friendly, visually appealing, and adaptable interface.
- Tasks for analysis and reporting that are ready to use.
- A facility for altering code.
- Much of SAS's functionality is available to you.
- Exporting data and results to other apps has never been easier.

- Automation and scripting.
- Transparency in data access.

Descriptive Statistics:

Descriptive statistics is a type of data analysis to which is used to describe the data in a meaningful way so that the patterns can be known. We cannot draw the conclusions based on the data we have examined. This is another way of describing the information of the data.

When a large amount of data is present, descriptive statistics is important since it is difficult to see the data if we display it as raw data. Thus, descriptive statistics make us to represent the data in a clear and meaningful way which helps the interpretation of data in a simple way. For instance, if we got the results of 100 pieces of students' assignments, we could be curious about their overall performance. By using the descriptive statistics, the spread or distribution of marks can be known as well.

There are couple of ways to describe the data in the statistics are measures of central tendency and measures of spread. Mean, median& mode are measures of the central tendency. Range, Skewness, standard deviation, quartiles, variance, absolute deviation are the measures of the spread.

To describe the data in SAS programming, we use **PROC UNIVARIATE**. This is used to find the distribution of data, outliers and normality. The UNIVARIATE technique provides a number of descriptive measures, high-resolution graphical presentations, and statistical tools for summarising, visualising, analysing, and modelling numeric variable statistical distributions.

Correlation analysis is to find the relation between two numerical and continuous variables. The correlation coefficient allows researchers to see if two variables recorded on the same subject have a possible linear relationship. When these two variables are of a continuous type, Pearson's correlation coefficient is the most commonly used measure of relationship.

The correlation coefficient, which ranges from 1 to +1, can be used to express this relationship. The sample statistic (correlation coefficient) is r, while the population correlation is commonly written as the rho (r). When shown on an x-y axis, the correlation indicates how well a straight line fits through a dispersion of dots. The **PROC CORR** procedure produces Pearson correlation coefficients of continuous numeric variables. The relation between the continuous variables can be plotted in a scatter plot matrix.

The goal of regression analysis is to simulate the relationship between a response or output variable and a set of input factors. The response is referred to as the target variable, or the variable that one is attempting to predict, while the remaining input variables are referred to as parameters in the algorithm. They are used to calculate the response variable's predicted value. The REG procedure **PROC REG** has extensive capabilities when it comes to fitting linear regression models with single numeric independent variables. We can use various types of regression techniques like Linear Regression, Polynomial Regression.

SAS Hypothesis testing is a procedure in which a user verifies a hypothesis about a population parameter. In SAS Programming Language, this testing is used to deduce the outcome of a hypothesis based on sample data from a broader population. The following SAS methods can be used to do this testing like Chi-Square test, ANOVA and T-test.

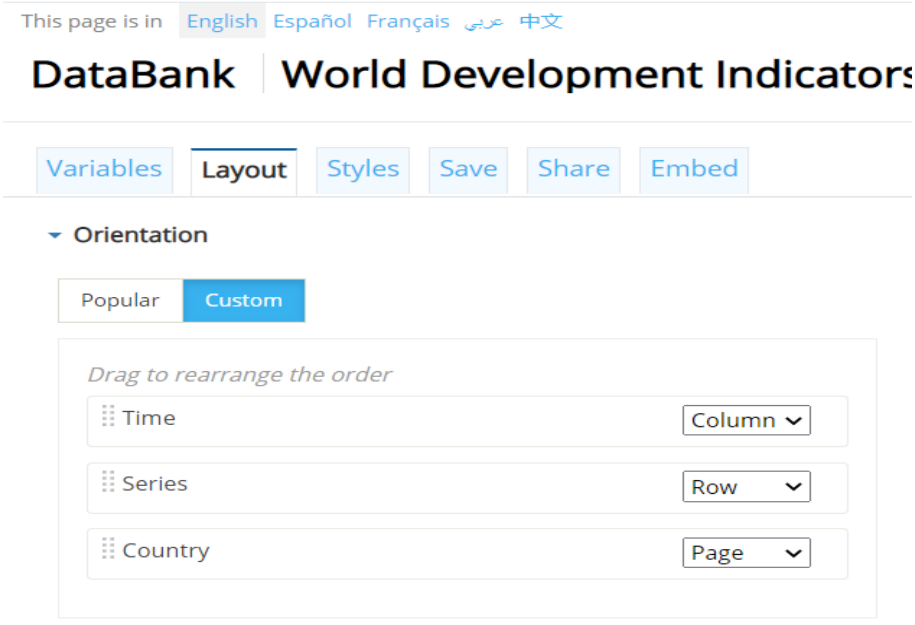
3. Exploration of Data Set

The Dataset "**Population estimates & Projections**" has been taken from the World bank data bank from below link.

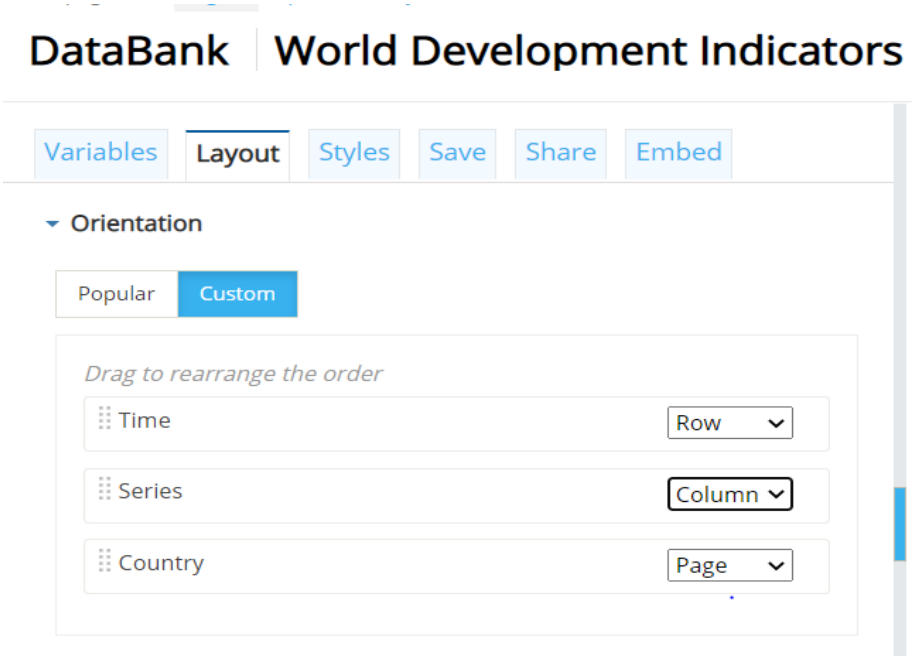
https://databank.worldbank.org/source/world-development-indicators

Data pivoting has been done on the downloaded dataset. It is done in the world bank website only. Data pivoting is just rearranging the rows and columns so that the user can get a different view/perspective of the data for analysis. This is done as shown below.

Initial data layout:



After pivoting, the layout changes to



Total population, Male population, Female population, Urban & Rural Population Growth Rates, Age Dependency Ratio, Life Expectancy at Birth, Total Infant Mortality Rate (per 1000 live births), Male & female Infant Mortality Rates (per 1000 live births), Number of Infant deaths , Male & female Infant deaths for 14 countries (Australia, Brazil, Belgium, China, Chile, Canada, Germany, Italy, India, Japan, Russian Federation, Spain, United Kingdom & United states) for the period 2007-2019 are among the data attributes included in the dataset (13 years).

C	D
Indicator Name	Long definition
Age dependency ratio (% of working-age population)	Age dependency ratio is the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64. Data are shown as the proportion of dependen
Life expectancy at birth, female (years)	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Life expectancy at birth, male (years)	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Life expectancy at birth, total (years)	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Mortality rate, adult, female (per 1,000 female adults)	Adult mortality rate, female, is the probability of dying between the ages of 15 and 60--that is, the probability of a 15-year-old female dying before reaching age 60, if subject to age-specific
Mortality rate, adult, male (per 1,000 male adults)	Adult mortality rate, male, is the probability of dying between the ages of 15 and 60--that is, the probability of a 15-year-old male dying before reaching age 60, if subject to age-specific mort
Mortality rate, infant (per 1,000 live births)	Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.
Mortality rate, infant, male (per 1,000 live births)	Infant mortality rate, male is the number of male infants dying before reaching one year of age, per 1,000 male live births in a given year.
Mortality rate, infant, female (per 1,000 live births)	Infant mortality rate, female is the number of female infants dying before reaching one year of age, per 1,000 female live births in a given year.
Number of infant deaths	Number of infants dying before reaching one year of age.
Population, female	Female population is based on the de facto definition of population, which counts all female residents regardless of legal status or citizenship.
Population, male	Male population is based on the de facto definition of population, which counts all male residents regardless of legal status or citizenship.
Population, total	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
Rural population (% of total population)	Rural population refers to people living in rural areas as defined by national statistical offices. It is calculated as the difference between total population and urban population.
Urban population (% of total population)	Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.

In the data preparation, I have created two columns with names Rural Population & Urban Population using SAS Programming as shown below.

```
data population;
set population;
urban_population =  'Percent of Urban population'n * Population;
rural_population = 'Percent of Rural population'n  * population;
run;
```

I have done the sorting of the data based on the country as shown below.

```
proc sort data=work.population;
by 'Country Name'n;
run;
```

The output of the dataset can be found in the below pdf which id exported from the SAS Results.

population dataset.pdf

This is the dataset that has been saved in the SAS local work folder which can be used to find out the descriptive analysis, Correlation & Regression analysis.

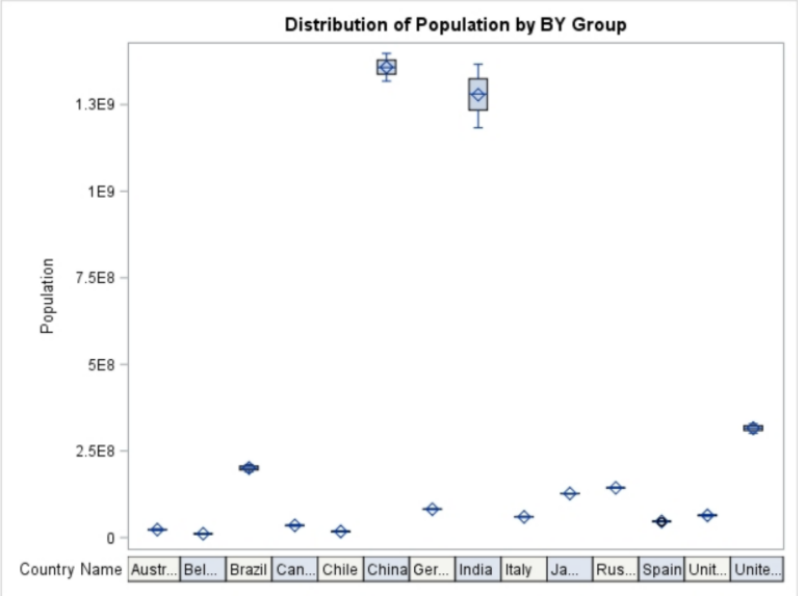
Outliers are very unique values in the dataset that can disturb the statistical analysis. Outliers can occur in the dataset due to data entry errors or during the sampling. To detect the outliers in the dataset, I have executed the below code in the SAS Programming.

```
proc univariate data=work.population plot;
by 'Country Name'n;
Run;
```

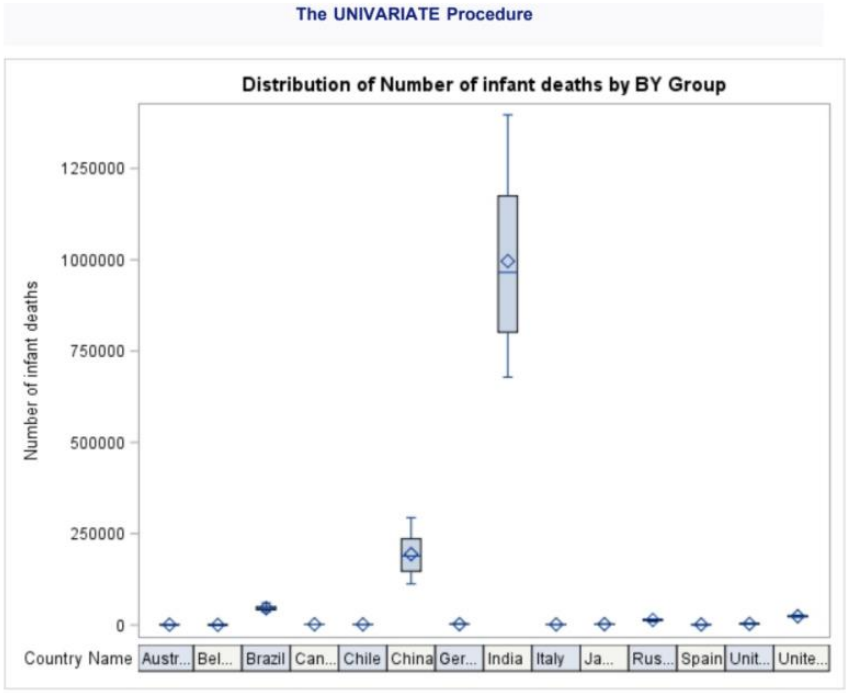
The Distribution plots for the required variables for statistical analysis are as shown below:

Plot for the variable population based on the country:

The UNIVARIATE Procedure



Plot for the variable Number of Infant deaths based on country:



The above graphs shows that the variables do not contain the outliers. The below pdf has plots for all the variables.

Box_plot_for_outliers.pdf

4. Analysis

4.1. Descriptive Statistical Analysis

Descriptive statistical analysis gives easy understanding of the sample data collected by giving the basic statistics of data in a numerical way or graphical way. These help us to give an initial idea about the data during analysing the larger amount of data. The graphical tools or algorithms used to describe the sample data is known as descriptive statistics. They are used to find the central tendency, variance (range of scores) as well.

PROC UNIVARIATE is a procedure in SAS that is primarily used to examine data distributions, including determining normalcy and identifying outliers. PROC UNIVARIATE will produce the following for each numerical variable such as Mean, Median, Mode,

Quantiles, Extreme Observations, Skewness, Kurtosis & Standard Deviation. To separate the analysis based on the variable we can use BY statement. To know the frequency of the observations for a variable FREQ is used. HISTOGRAM, CDFPLOT, PPPPLOT, PROBPLOT, QQPLOT, are used for to create the graphs for the respective variables. CLASS is used to group the data.

The syntax for the UNIVARIATE is as shown below:

Syntax: UNIVARIATE Procedure

```
PROC UNIVARIATE <options>;
    BY variables;
    CDFPLOT <variables> </ options>;
    CLASS variable-1 <(v-options)> <variable-2 <(v-options)>></ KEYLEVEL=value1 | (value1 value2 )>;
    FREQ variable;
    HISTOGRAM <variables> </ options>;
    ID variables;
    INSET keyword-list </ options>;
    OUTPUT <OUT=SAS-data-set> <keyword1=names ...keywordk=names> <percentile-options>;
    PPPLOT <variables> </ options>;
    PROBPLOT <variables> </ options>;
    QQPLOT <variables> </ options>;
    VAR variables;
    WEIGHT variable;
```

The code for the Univariate is as shown below:

```
ods graphics / reset=all imagemap;
proc univariate data= population;
BY 'Country Name'n;
var Population male female urban_population rural_population 'Number of infant deaths'n 'Number of infant deaths male'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
'Age dependency ratio'n;
histogram Population male female urban_population rural_population 'Number of infant deaths'n 'Number of infant deaths male'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
'Age dependency ratio'n /normal(noprint);
run;
```

Procedure Univariate is divided based on the variable country name. This enables us to give the analysis based on the country name. The output for some of the countries is shown below.

Univariate for population variable for country Australia:

The UNIVARIATE Procedure
Variable: Population (Population)

Country Name=Australia

Moments			
N	13	Sum Weights	13
Mean	23110365.3	Sum Observations	300434749
Std Deviation	1449212.43	Variance	2.10022E12
Skewness	0.00169835	Kurtosis	-1.1242628
Uncorrected SS	6.96836E15	Corrected SS	2.52026E13
Coeff Variation	6.27083309	Std Error Mean	401939.211

Basic Statistical Measures			
Location		Variability	
Mean	23110365	Std Deviation	1449212
Median	23128129	Variance	2.10022E12
Mode	.	Range	4538145
		Interquartile Range	2159157

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	57.49717	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0002
Signed Rank	S	45.5	Pr >= |S|	0.0002

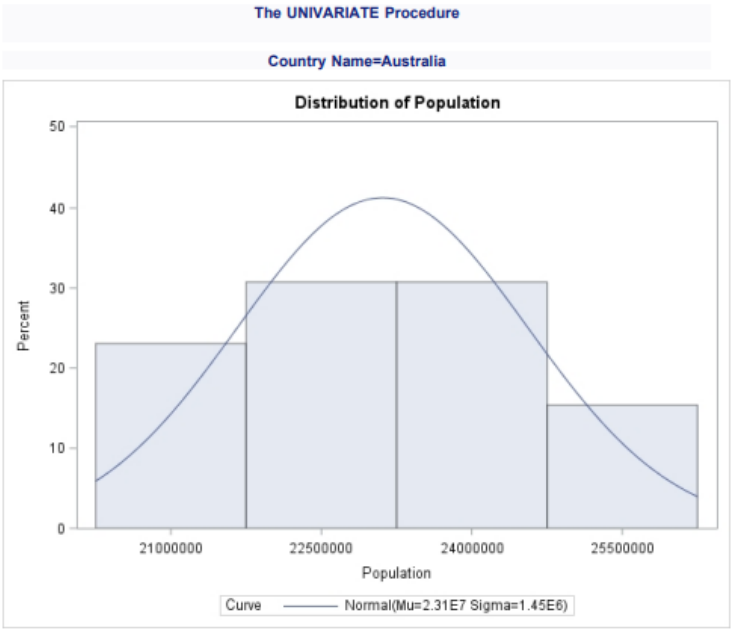
Quantiles (Definition 5)	
Level	Quantile
100% Max	25365745
99%	25365745
95%	25365745
90%	24982688
75% Q3	24190907
50% Median	23128129
25% Q1	22031750
10%	21249200
5%	20827600
1%	20827600
0% Min	20827600

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
20827600	1	23815995	9
21249200	2	24190907	10
21691700	3	24601860	11
22031750	4	24982688	12
22340024	5	25365745	13

Mean for the population variable is 23110365 and median is 23128129. There is no mode value calculated for this variable as there are no repeated values/ observations.

Here mean < median. If mean value is less than the median then we can infer that it is negatively skewed. This can be confirmed using the distribution plot. The skewness for the population variable for country Australia is 0.00169835. Ideal skewness should be near to 0. This skewness very negligible and can be ignored. The kurtosis for the variable population is -1.1242628. Ideal value for kurtosis is to be -3.0 to +3.0. The Kurtosis is in the acceptable range for the population variable.

Distribution curve for the population is as shown below:



The above curve shows that the data is well distributed with Mu ($\bar{x}$) value of 2.31E7(10 to the power 7) and sigma(σ) of 1.45E6(10 to the power 6), it is slightly left skewed (negative skewness)

Univariate for the variable Total Life Expectancy at birth is shown below

The UNIVARIATE Procedure
Variable: Life expectancy at birth total (Life expectancy at birth total)

Country Name=Australia

Moments			
N	13	Sum Weights	13
Mean	82.1011257	Sum Observations	1067.31463
Std Deviation	0.51166118	Variance	0.26179716
Skewness	-0.152542	Kurtosis	-1.0430414
Uncorrected SS	87630.8745	Corrected SS	3.14156592
Coeff Variation	0.62320848	Std Error Mean	0.14190928

Basic Statistical Measures			
Location		Variability	
Mean	82.10113	Std Deviation	0.51166
Median	82.14878	Variance	0.26180
Mode	.	Range	1.60732
		Interquartile Range	0.75366

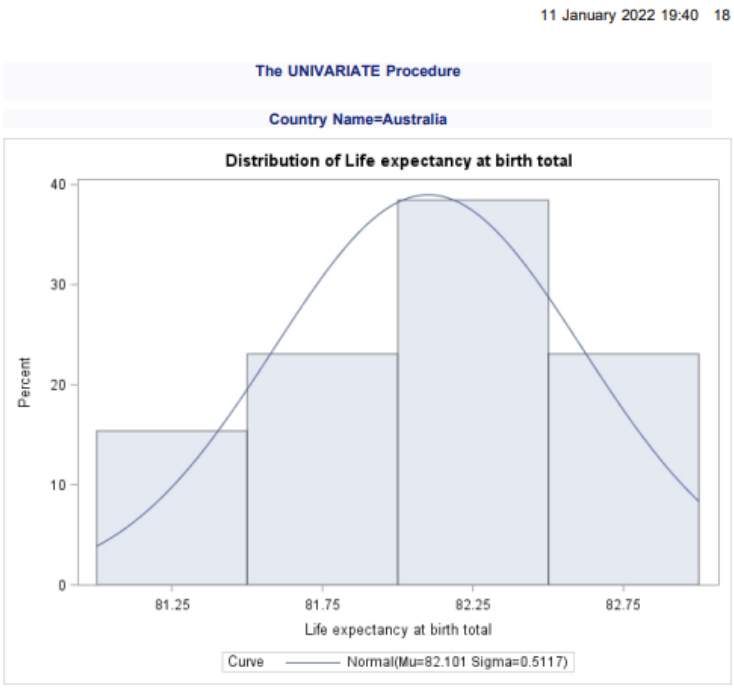
Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	578.5466	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0002
Signed Rank	S	45.5	Pr >= |S|	0.0002

Quantiles (Definition 5)	
Level	Quantile
100% Max	82.9000
99%	82.9000
95%	82.9000
90%	82.7488
75% Q3	82.4488
50% Median	82.1488
25% Q1	81.6951
10%	81.3951
5%	81.2927
1%	81.2927
0% Min	81.2927

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
81.2927	1	82.4000	9
81.3951	2	82.4488	10
81.5439	3	82.5000	11
81.6951	4	82.7488	12
81.8951	5	82.9000	13

Mean for the variable Life Expectancy at birth total is 82.10113 and median is 82.14878. Here mean < median, so we can state that it is negatively skewed which can be verified in the distribution plot below. There is no mode value calculated for this variable as there are no repeated values. Skewness is -0.152542 & Kurtosis is -1.0430414. Skewness & Kurtosis for the above variable is in the acceptable range.

Distribution curve for the Life expectancy at birth for Australia is shown below:



The distribution curve shows that it is slightly left skewed (negative skewness) with skewness of -0.152542. The means Mu ($\bar{x}$) is 82.101 & Sigma (σ) of 0.5117.

The univariate for the variable Total Population for the country INDIA is shown below:

11 January 202

The UNIVARIATE Procedure
Variable: Population (Population)

Country Name=India

Moments			
N	13	Sum Weights	13
Mean	1278523396	Sum Observations	1.66208E10
Std Deviation	59167511	Variance	3.50079E15
Skewness	-0.1121615	Kurtosis	-1.1603474
Uncorrected SS	2.12921E19	Corrected SS	4.20095E16
Coeff Variation	4.62780041	Std Error Mean	16410115

Basic Statistical Measures			
Location		Variability	
Mean	1.2785E9	Std Deviation	59167511
Median	1.2808E9	Variance	3.50079E15
Mode	.	Range	183208285
		Interquartile Range	90236087

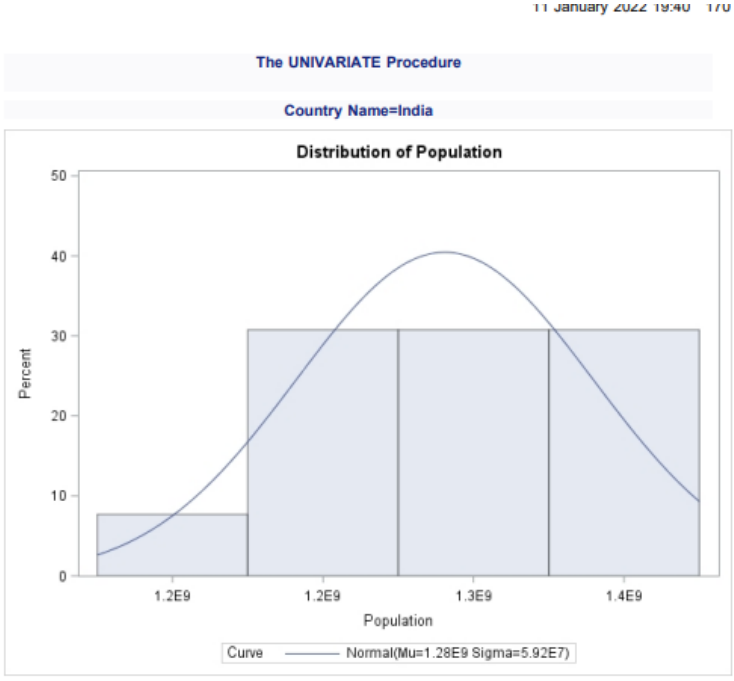
Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	77.91069	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0002
Signed Rank	S	45.5	Pr >= |S|	0.0002

Quantiles (Definition 5)	
Level	Quantile
100% Max	1366417756
99%	1366417756
95%	1366417756
90%	1352642283
75% Q3	1324517250
50% Median	1280842119
25% Q1	1234281163
10%	1200669762
5%	1183209471
1%	1183209471
0% Min	1183209471

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1183209471	92	1310152392	100
1200669762	93	1324517250	101
1217726217	94	1338676779	102
1234281163	95	1352642283	103
1250287939	96	1366417756	104

Mean is 1.2785E9 and median is 1.2808E9. Here mean < median, so we can state that it is negatively skewed which can be verified in the distribution plot below. There is no mode calculated for this variable as the values are not repeated. Here It has skewness of -0.1121615 and kurtosis of -1.1603474 which is in the acceptable range.

Distribution curve is as shown below:



Distribution curve shows that the curve is slightly left skewed (negative skewness) with Mean of 1.28E9(10 to the power 9) and sigma of 5.92E7(10 to the power 7)

The univariate for the variable Number of Infant deaths is as shown below:

The UNIVARIATE Procedure
Variable: Number of infant deaths (Number of infant deaths)

Country Name=India

Moments			
N	13	Sum Weights	13
Mean	996354.308	Sum Observations	12952606
Std Deviation	238420.614	Variance	5.68444E10
Skewness	0.31648893	Kurtosis	-1.1861415
Uncorrected SS	1.35875E13	Corrected SS	6.82133E11
Coeff Variation	23.9293003	Std Error Mean	66125.9807

Basic Statistical Measures			
Location		Variability	
Mean	996354.3	Std Deviation	238421
Median	965381.0	Variance	5.68444E10
Mode	.	Range	717827
		Interquartile Range	373289

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	15.06752	Pr > |t|	<.0001
Sign	M	6.5	Pr >= |M|	0.0002
Signed Rank	S	45.5	Pr >= |S|	0.0002

Quantiles (Definition 5)	
Level	Quantile
100% Max	1396555
99%	1396555
95%	1396555
90%	1324731
75% Q3	1174867
50% Median	965381
25% Q1	801578
10%	715531
5%	678728
1%	678728
0% Min	678728

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
678728	104	1101188	96
715531	103	1174867	95
756546	102	1250379	94
801578	101	1324731	93
851000	100	1396555	92

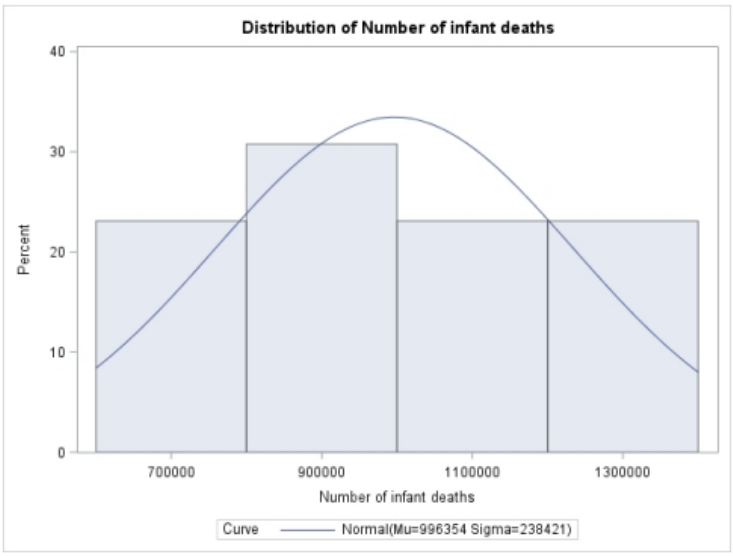
The variable has a mean of 996354.3 and median of 965381. This shows that the curve is positively skewed which can be verified in the below box plot. There is no mode calculated for this variable also. It has a skewness of 0.31648893 and kurtosis of -1.1861415.

Distribution curve for the same variable is as shown below.

11 January 2022 19:40 180

The UNIVARIATE Procedure

Country Name=India



The graph shows that the curve is slightly right skewed (positive skewness) with the Mean of 996354 and sigma of 238421.

The results for each variable in the dataset based on the country can be viewed in the below pdf which is exported from SAS Results.

Univariate Results.pdf

4.2. Correlation Analysis:

Correlation is to find the relationship between the two variables in the dataset which are numerical and continuous as well. This can be measured using the Pearson correlation coefficient. It can range from -1 to +1.

Syntax for the correlation is as shown below:

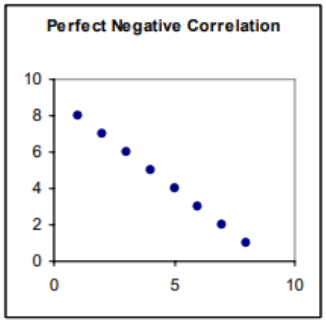
```
proc corr data=dataset;
  by byvars;
  freq freqvar;
  var varlist;
  weight weightvar;
run;
```

We can use BY variable to divide the correlation analysis based on the variable given. If the correlation is non-linear and outliers are present in the data then it will be difficult to find the accurate correlation coefficient.

Types of Correlation are:

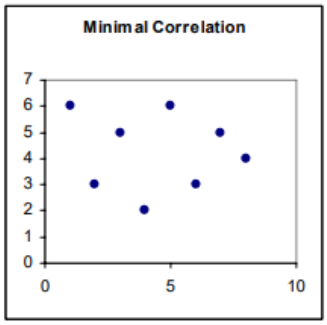
Negative Correlation:

In this type, the variables will be negatively correlated as i.e., they move together but in the negative direction as if one variable is increasing the other is decreasing. The Pearson correlation coefficient will be closer or equal to -1.



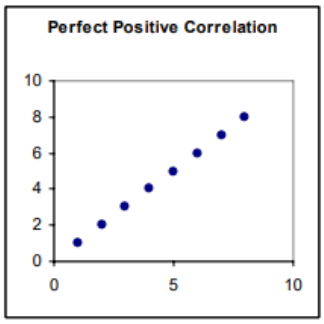
Minimal Correlation:

In this type, the variables are not correlated. It will be difficult to find the relationship between two variables. The Pearson coefficient will be closer or equal to 0. If it is nearer to 0, then the two variables are not correlated.



Positive Correlation:

In this type, the variables are positively correlated i.e., they move together in a positive way. If one variable is increasing then the other variable also increases. The Pearson correlation coefficient will be closer or equal to -1.



The code used for the Correlation Analysis is as shown below:

```
proc corr data=population plots=matrix(hist);
by 'Country Name'n;
var Population 'Number of infant deaths'n 'Number of infant deaths male'n 'Age dependency ratio'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
 male female urban_population rural_population ;
 title "Correlation Analysis for each country";
run;
```

Plots=Matrix(hist) produces the scatter plot for the variables for which correlation analysis is done. SAS has a limitation in producing the scatter plot matrix where only a maximum of five variables can be plotted.

Output for the above code is as shown below for the country Australia:

Correlation Analysis for each country

The CORR Procedure

Country Name=Australia

12 Variables: Population, Number of infant deaths, Number of infant deaths male, Age dependency ratio, Number of infant deaths female, Life expectancy at birth total, Life expectancy at birth female, Life expectancy at birth male, male, urban_population, rural_population

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Population	13	23110365	1449212	300434749	20827600	25365745	Population
Number of infant deaths	13	1103	110.65892	14335	972.00000	1272	Number of infant deaths
Number of infant deaths male	13	616.76923	64.64409	8018	542.00000	715.00000	Number of infant deaths male
Age dependency ratio	13	50.14553	2.29227	651.89187	47.74598	54.31301	Age dependency ratio
Number of infant deaths female	13	485.92308	46.04248	6317	430.00000	557.00000	Number of infant deaths female
Life expectancy at birth total	13	82.10113	0.51166	1067	81.29268	82.90000	Life expectancy at birth total
Life expectancy at birth female	13	84.31538	0.41402	1096	83.70000	85.00000	Life expectancy at birth female
Life expectancy at birth male	13	79.99231	0.60891	1040	79.00000	80.90000	Life expectancy at birth male
male	13	11531248	704717	149906226	10405448	12632259	male
female	13	11579117	744581	150528523	10422152	12733486	female
urban_population	13	1976236650	133460288	2.56911E10	1766638687	2184599422	
rural_population	13	334799681	11466939	4352398456	316121313	351975078	

Pearson Correlation coefficients for the country Australia is:

Pearson Correlation Coefficients, N = 13
Prob > |r| under H0: Rho=0

	Population	Number of infant deaths	Number of infant deaths male	Age dependency ratio	Number of infant deaths female	Life expectancy at birth total	Life expectancy at birth female	Life expectancy at birth male	male	female	urban_population	rural_population
Population	1.00000	-0.97827	-0.97439	0.96879	-0.98312	0.99309	0.98693	0.99915	0.99994	0.99994	1.00000	0.99952
Population		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Number of infant deaths	-0.97827	1.00000	0.99982	-0.92234	0.99965	-0.98491	-0.96483	-0.99104	-0.97747	-0.97891	-0.97804	-0.98050
Number of infant deaths	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Number of infant deaths male	-0.97439	0.99982	1.00000	-0.91688	0.99897	-0.98252	-0.96136	-0.98936	-0.97353	-0.97510	-0.97415	-0.97675
Number of infant deaths male	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Age dependency ratio	0.96879	-0.92234	-0.91688	1.00000	-0.92944	0.95065	0.95394	0.94189	0.96710	0.97027	0.96928	0.96256
Age dependency ratio	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Number of infant deaths female	-0.98312	0.99965	0.99897	-0.92944	1.00000	-0.98768	-0.96912	-0.99281	-0.98241	-0.98368	-0.98290	-0.98517
Number of infant deaths female	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Life expectancy at birth total	0.99309	-0.98491	-0.98252	0.95065	-0.98768	1.00000	0.99345	0.99726	0.99329	0.99279	0.99312	0.99230
Life expectancy at birth total	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Life expectancy at birth female	0.98693	-0.96483	-0.96136	0.95394	-0.96912	0.99345	1.00000	0.98227	0.98757	0.98621	0.98712	0.98421
Life expectancy at birth female	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001
Life expectancy at birth male	0.99915	-0.99104	-0.98936	0.94189	-0.99281	0.99726	0.98227	1.00000	0.99005	0.99012	0.99006	0.99060
Life expectancy at birth male	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001
male	0.99994	-0.97747	-0.97353	0.96710	-0.98241	0.99329	0.98757	0.99005	1.00000	0.99976	0.99993	0.99952
male	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001
female	0.99994	-0.97891	-0.97510	0.97027	-0.98368	0.99279	0.98621	0.99012	0.99976	1.00000	0.99995	0.99941
female	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001
urban_population	1.00000	-0.97804	-0.97415	0.96928	-0.98290	0.99312	0.98712	0.99006	0.99993	0.99995	1.00000	0.99943
	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001		< .0001
rural_population	0.99952	-0.98050	-0.97675	0.96256	-0.98517	0.99230	0.98421	0.99060	0.99952	0.99941	0.99943	1.00000
	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	

For the total population and number of infant deaths, the Pearson correlation coefficient is -0.97827. This indicates that they have a negative correlation.

The Population and Age Dependency Ratio has a Pearson correlation coefficient of 0.96879. This implies that they are related in a positive correlation.

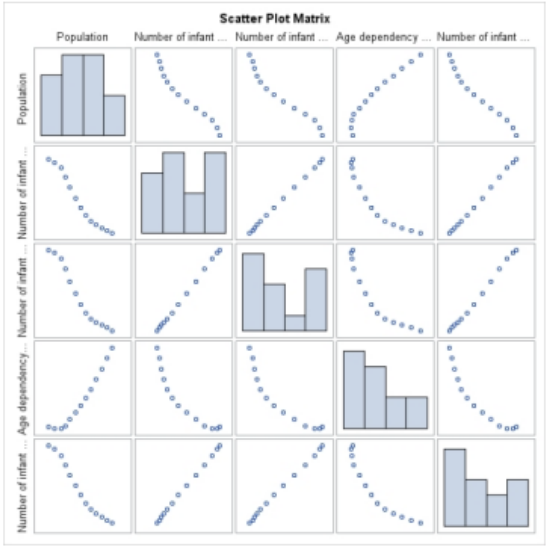
The population and life expectancy at birth total has a Pearson correlation coefficient of 0.99309. This implies that they are related in a positive correlation. Because the value is so close to 1, they are linearly dependant.

This can be seen in the from Scatter plot matrix as shown below:

Correlation Analysis for each country

The CORR Procedure

Country Name=Australia



The above scatter plot matrix shows that negative correlation between the population & Number of infant deaths total. There is a positive correlation between the Population & Age dependency Ratio.

Output for the country United Kingdom is as shown below:

Correlation Analysis for each country

The CORR Procedure

Country Name=United Kingdom

12 Variables: Population Number of infant deaths Number of infant deaths male Age dependency ratio Number of infant deaths female Life expectancy at birth total Life expectancy at birth female Life expectancy at birth male male female urban_population rural_population

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Population	13	64149618	1811517	833945031	61322463	66836327	Population
Number of infant deaths	13	3241	304.02345	42138	2842	3686	Number of infant deaths
Number of infant deaths male	13	1823	176.18416	23702	1590	2078	Number of infant deaths male
Age dependency ratio	13	53.78497	2.08529	699.20459	51.26661	56.75061	Age dependency ratio
Number of infant deaths female	13	1418	127.90025	18436	1252	1608	Number of infant deaths female
Life expectancy at birth total	13	80.73058	0.64445	1049	79.44878	81.30488	Life expectancy at birth total
Life expectancy at birth female	13	82.67692	0.54948	1075	81.60000	83.20000	Life expectancy at birth female
Life expectancy at birth male	13	78.87692	0.73841	1025	77.40000	79.50000	Life expectancy at birth male
male	13	31576123	966990	410489601	30069069	33008768	male
female	13	32573495	844533	423455430	31253394	33827559	female
urban_population	13	5267653840	214643109	6.84795E10	4935170500	5590992426	
rural_population	13	1147307937	33548284	1.4915E10	1092640274	1197075800	

Pearson Correlation coefficients for the country United Kingdom is:

Pearson Correlation Coefficients, N = 13
Prob > |r| under H0: Rho=0

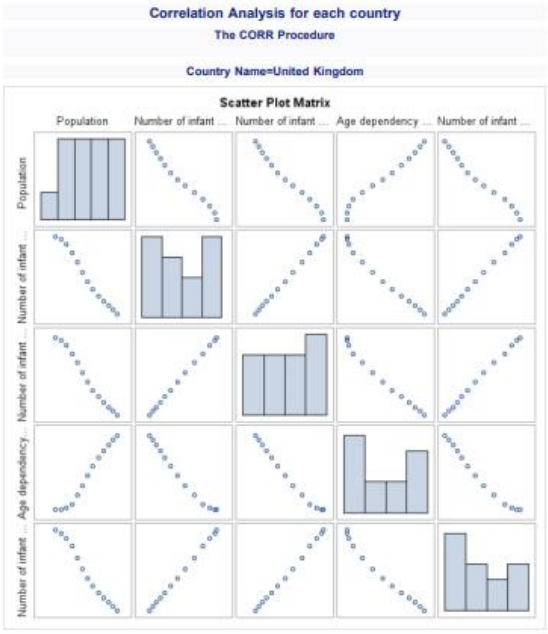
	Population	Number of infant deaths	Number of infant deaths male	Age dependency ratio	Number of infant deaths female	Life expectancy at birth total	Life expectancy at birth female	Life expectancy at birth male	male	female	urban_population	rural_population
Population	1.00000	-0.99243	-0.99340	0.98799	-0.99063	0.87873	0.85388	0.89215	1.00000	1.00000	0.99995	-0.99799
Population		< .0001	< .0001	< .0001	< .0001	< .0001	0.0002	< .0001	< .0001	< .0001	< .0001	< .0001
Number of infant deaths	-0.99243	1.00000	0.99985	-0.99031	0.99972	-0.89674	-0.87002	-0.91141	-0.99270	-0.99212	-0.99191	0.98739
Number of infant deaths	< .0001		< .0001	< .0001	< .0001	< .0001	0.0001	< .0001	< .0001	< .0001	< .0001	< .0001
Number of infant deaths	-0.99340	0.99985	1.00000	-0.99186	0.99918	-0.89196	-0.86496	-0.90684	-0.99365	-0.99310	-0.99294	0.98879
Number of infant deaths mal	< .0001	< .0001		< .0001	< .0001	< .0001	0.0001	< .0001	< .0001	< .0001	< .0001	< .0001
Age dependency ratio	0.98799	-0.99031	-0.99186	1.00000	-0.98771	0.83183	0.80238	0.84875	0.98825	0.98768	0.98827	-0.98814
Age dependency ratio	< .0001	< .0001	< .0001		< .0001	0.0004	0.0010	0.0002	< .0001	< .0001	< .0001	< .0001
Number of infant deaths f	-0.99063	0.99972	0.99918	-0.98771	1.00000	-0.90291	-0.87658	-0.91726	-0.99091	-0.99029	-0.99001	0.98498
Number of infant deaths fe	< .0001	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Life expectancy at birth to	0.87873	-0.89674	-0.89196	0.83183	-0.90291	1.00000	0.99607	0.99803	0.87882	0.87861	0.87568	-0.85773
Life expectancy at birth total	< .0001	< .0001	< .0001	0.0004	< .0001		< .0001	< .0001	< .0001	< .0001	< .0001	0.0002
Life expectancy at birth fe	0.85388	-0.87002	-0.86496	0.80238	-0.87658	0.99607	1.00000	0.98854	0.85390	0.85385	0.85091	-0.83339
Life expectancy at birth fem	0.0002	0.0001	0.0001	0.0010	< .0001	< .0001		< .0001	0.0002	0.0002	0.0002	0.0004
Life expectancy at birth m	0.89215	-0.91141	-0.90684	0.84875	-0.91726	0.99803	0.98854	1.00000	0.89230	0.89198	0.88907	-0.87089
Life expectancy at birth mal	< .0001	< .0001	< .0001	0.0002	< .0001	< .0001	< .0001		< .0001	< .0001	< .0001	0.0001
male	1.00000	-0.99270	-0.99365	0.98825	-0.99091	0.87882	0.85390	0.89230	1.00000	0.99999	0.99995	-0.99798
male	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	0.0002	< .0001		< .0001	< .0001	< .0001
female	1.00000	-0.99212	-0.99310	0.98768	-0.99029	0.87861	0.85385	0.89198	0.99999	1.00000	0.99995	-0.99799
female	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	0.0002	< .0001	< .0001		< .0001	< .0001
urban_population	0.99995	-0.99191	-0.99294	0.98827	-0.99001	0.87568	0.85091	0.88907	0.99995	0.99995	1.00000	-0.99857
	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	0.0002	< .0001	< .0001	< .0001		< .0001
rural_population	-0.99799	0.98739	0.98879	-0.98814	0.98498	-0.85773	-0.83339	-0.87089	-0.99798	-0.99799	-0.99857	1.00000
	< .0001	< .0001	< .0001	< .0001	< .0001	0.0002	0.0004	0.0001	< .0001	< .0001	< .0001	

For the total population and number of infant deaths, the Pearson correlation coefficient is -0.99243. This indicates that they have a strong negative correlation.

The Population and Age Dependency Ratio has a Pearson correlation coefficient of 0.98799. This implies that they are related in a positive correlation.

The population and life expectancy at birth total has a Pearson correlation coefficient of 0.87873. This implies that they are related in a positive correlation.

This can be seen in the from Scatter plot matrix which is plotted as shown below:



The scatter plot matrix above depicts a negative correlation between the population and the overall number of infant deaths. The Population and Age Dependency Ratio have a positive correlation.

The correlation analysis output for all the countries is attached as a pdf which is extracted from the SAS results is as shown below.

Correlation_analysis.pdf

4.3. Regression Analysis:

Regression analysis attempts to replicate the relationship between a response or output variable and a set of input variables. The response is referred to as the target variable, or the variable that is being predicted, while the remaining input variables are referred to as algorithm parameters. They're utilised to calculate the predicted value of the response variable.

I'm trying to do a linear regression analysis on Population with Number of infant deaths as I have found a correlation between them.

Syntax for Regression Analysis:

```
PROC REG <options>;
    <label:> MODEL dependents = <regressors> </ options>;
    BY variables;
    FREQ variable;
    ID variables;
    VAR variables;
    WEIGHT variable;
    ADD variables;
    CODE <options>;
    DELETE variables;
    <label:> MTEST <equation, ..., equation> </ options>;
    OUTPUT <OUT=SAS-data-set> <keyword=names> <...keyword=names>;
    PAINT <condition |ALLOBS> </ options> |<STATUS |UNDO>;
    PLOT <yvariable*xvariable> <=symbol> <...yvariable*xvariable> <=symbol> </ options>;
    PRINT <options> <ANOVA> <MODELDATA>;
    REFIT ;
    RESTRICT equation, ..., equation;
    REWEIGHT <condition |ALLOBS> </ options> |<STATUS |UNDO>;
    STORE <options>;
    <label:> TEST equation, <, ..., equation> </ option>;
```

Code for the Regression Analysis is:

```
proc reg data = population;
by 'Country Name'n;
model population = 'Number of infant deaths'n;
title "Linear Regression Analysis on Population & Number of Infant deaths for each country";
run;
```

Output for the Regression Analysis for the country Australia is as shown below:

Linear Regression Analysis on Population & Number of Infant deaths for each country

The REG Procedure
Model: MODEL1
Dependent Variable: Number of infant deaths Number of infant deaths

Country Name=Australia

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	140628	140628	244.89	<.0001
Error	11	6316.75680	574.25062		
Corrected Total	12	146945			

Root MSE	23.96353	R-Square	0.9570
Dependent Mean	1102.69231	Adj R-Sq	0.9531
Coeff Var	2.17318		

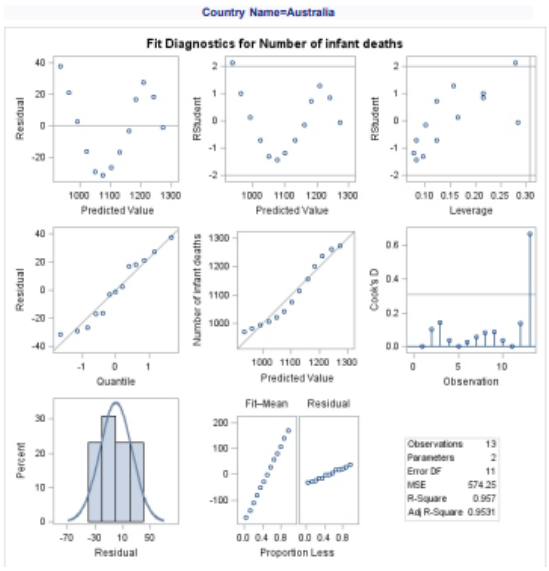
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2829.00735	110.51511	25.60	<.0001
Population	Population	1	-0.00007470	0.00000477	-15.65	<.0001

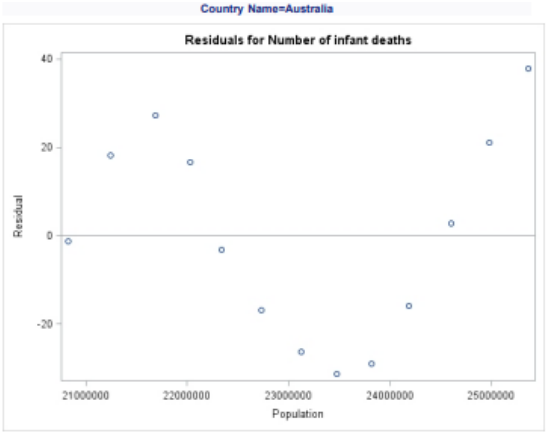
The intercept obtained is 2829.00735 and slope is -0.00007470. By this the fitted plot equation can be calculated as

Number of infant deaths = (-0.00007470)*Population + 2829.00735.

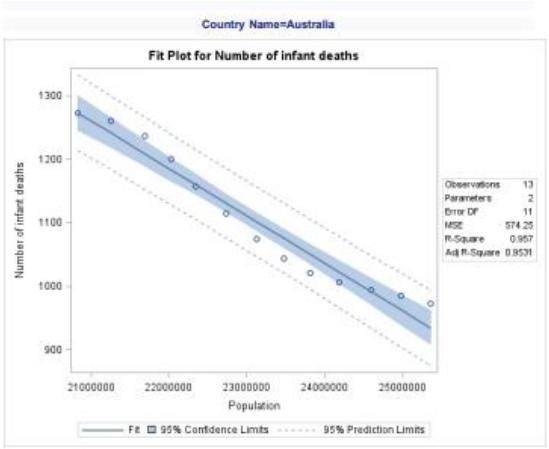
Since the slope the negative the linear curve obtained will be in downward direction which can be verified in the fitplot. R-square value for the graph is 0.957. This means that more data points are fitted to the Regression line will be shown in the fit plot. This also mean that the variability of the response will be around the mean value.



Residuals for the variable population & Number of Infant Deaths:



Fit plot for population & Number of Infant Deaths:



From the above Fit plot, we can see that as the population is increasing the Number of Infant deaths is decreasing which is a good sign. This has been observed in the correlation analysis as well. Above Fit plot is plotted for the 95% of confidence limits & 95% of the Prediction limits.

Output of the Regression Analysis for country United Kingdom is as shown below:

Linear Regression Analysis on Population & Number of Infant deaths for each country

The REG Procedure
Model: MODEL1
Dependent Variable: Number of infant deaths Number of infant deaths

Country Name=United Kingdom

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1092438	1092438	718.48	<.0001
Error	11	16725	1520.48658		
Corrected Total	12	1109163			

Root MSE	38.99342	R-Square	0.9849
Dependent Mean	3241.38462	Adj R-Sq	0.9835
Coeff Var	1.20299		

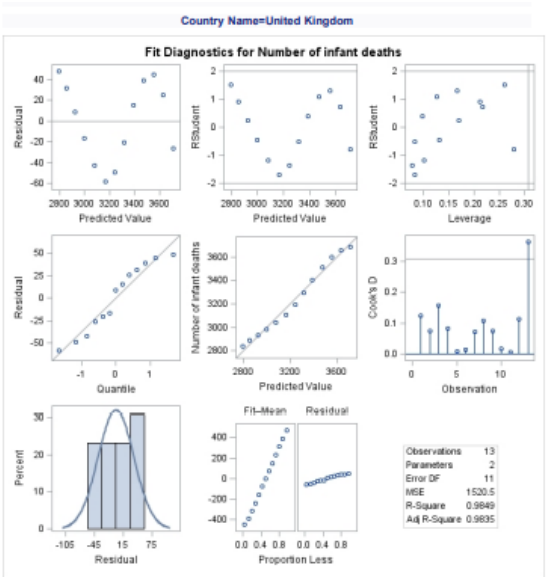
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	13926	398.76043	34.92	<.0001
Population	Population	1	-0.00016656	0.00000621	-26.80	<.0001

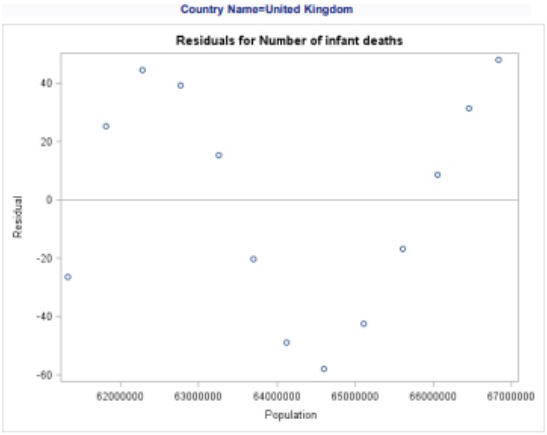
Intercept obtained is 13296 and slope obtained is -0.00016656. The fitted model equation for the above model is:

Number of Infant deaths = (-0.00016656) *Population+13296.

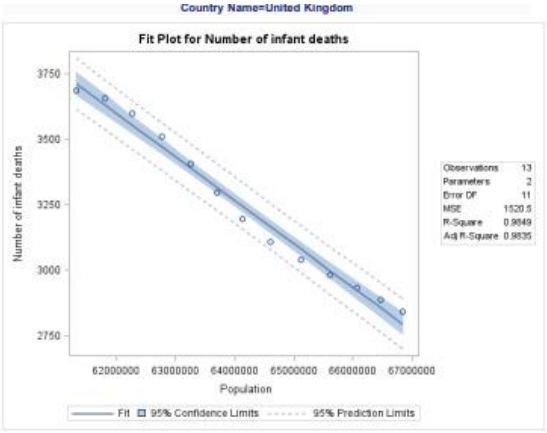
r-square value is 0.9849 which shows that the data points are located around the fitted regression line which can be verified in the fit plot.



Residuals for the variable population & Number of Infant Deaths:



Fit plot for population & Number of Infant Deaths:



The above Fit plot suggest that as the population is increasing the Number of Infant deaths is decreasing. The same pattern has been observed correlation analysis as well.

Regression Analysis output for all the countries is in the below pdf which is exported from the SAS results.

Regression Analysis.pdf

4.4. Hypothesis Testing:

A user verifies a hypothesis regarding a population parameter using the SAS Hypothesis Testing process. This testing is used in SAS Programming Language to determine the outcome of a hypothesis based on sample data from a larger population. This testing can be done using SAS methods such as the Chi-Square test, ANOVA, and TTEST:

TTEST:

This test is used to compare the means of the two individual groups. We require a dependent variable and an independent variable to perform this.

Syntax for the TTEST:

```
PROC TTEST <options>;
    BOOTSTRAP </ options>;
    CLASS variable;
    PAIRED variables;
    BY variables;
    VAR variables </ options>;
    FREQ variable;
    WEIGHT variable;
```

In this procedure, the variables given in the PAIRED statement are to be compared and variable in the CLASS statement groups the dataset based on the variable.

Code for TTEST:

```
proc ttest data = population alpha = 0.05;
     paired 'Number of infant deaths male'n*'Number of infant deaths female'n;
 by 'Country Name'n;
     title "TTEST For Number of Infant Deaths";
    run;
```

I'm trying to test whether there is any difference in the average of the two variables Number of infant deaths, male & Number of infant deaths, female.

Here alpha suggests the confidence level which is 0.05 and null hypothesis by default is 0. The graph has been plotted for 95% confidence.

PROC TTEST's Q-Q plots can be used to see if the observed values of a variable are consistent in comparison with if the variable were actually normally distributed.

Output of the Ttest for the variable Number of Infant Deaths for country Australia is as shown below:

TTEST For Number of Infant Deaths

The TTEST Procedure

Difference: Number of infant deaths male - Number of infant deaths female

Country Name=Australia

N	Mean	Std Dev	Std Err	Minimum	Maximum
13	130.8	18.7654	5.2046	112.0	158.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
130.8	119.5	142.2	18.7654	13.4564	30.9767

DF	t Value	Pr > |t|
12	25.14	<.0001

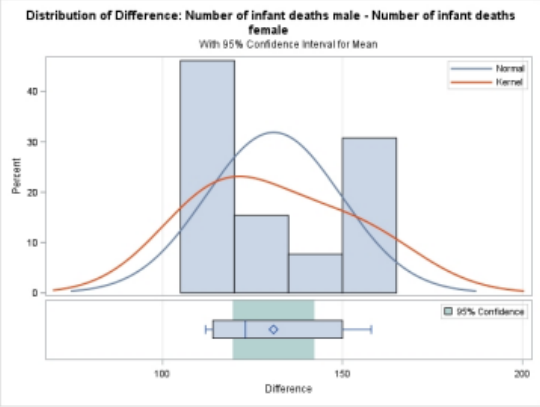
The above graphic shows the descriptive statistics for the difference between the Number of infant deaths male and Number of infant deaths female.

In the first table, N represents the Number of observations. Here N is 13 because we took the data for the 13 years. Mean is the average difference between the two variables here. On an average, 130.8 difference is there between the variables Number of infant deaths male and Number of infant deaths female. Standard deviation for the difference is 18.7564. Minimum difference is 112.0 and maximum difference is 158.0.

The second table in the graphic shows the results of the mean and standard deviation for the 95% confidence limit.

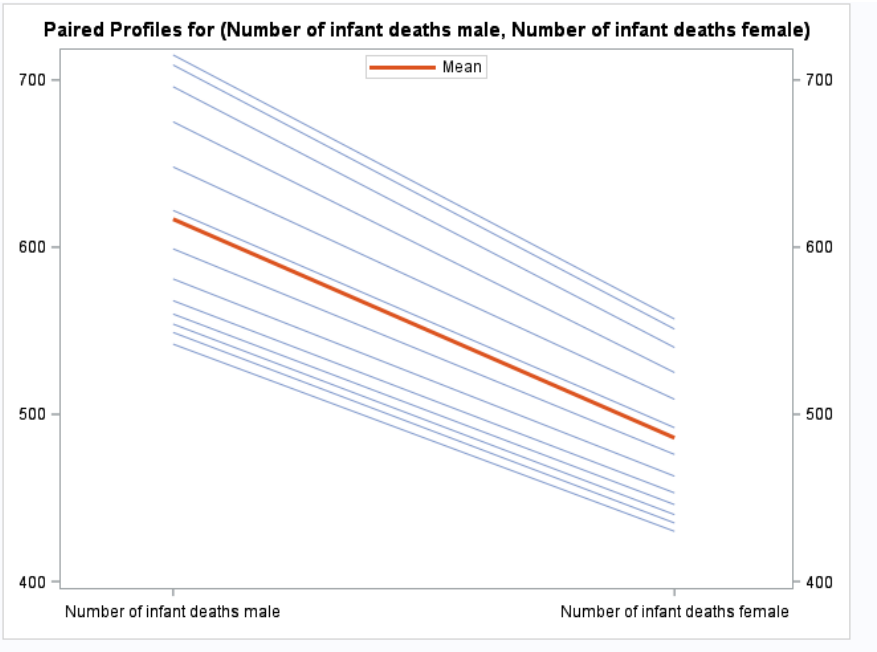
The third table in the graphic shows the results for the paired t test. The P value obtained is very small ($p < 0.0001$). We need to compare this p value with the null hypothesis. By default, the null hypothesis value is 0. But we did the Test for the confidence level of 0.05. Now, by comparing this value 0.05 with p, we can say that the null hypothesis that average of Number of infant deaths male and Number of infant deaths female are same can be rejected. So, we can say that Number of infant deaths male average is considerably different from the average of Number of infant deaths female.

Distribution plot:



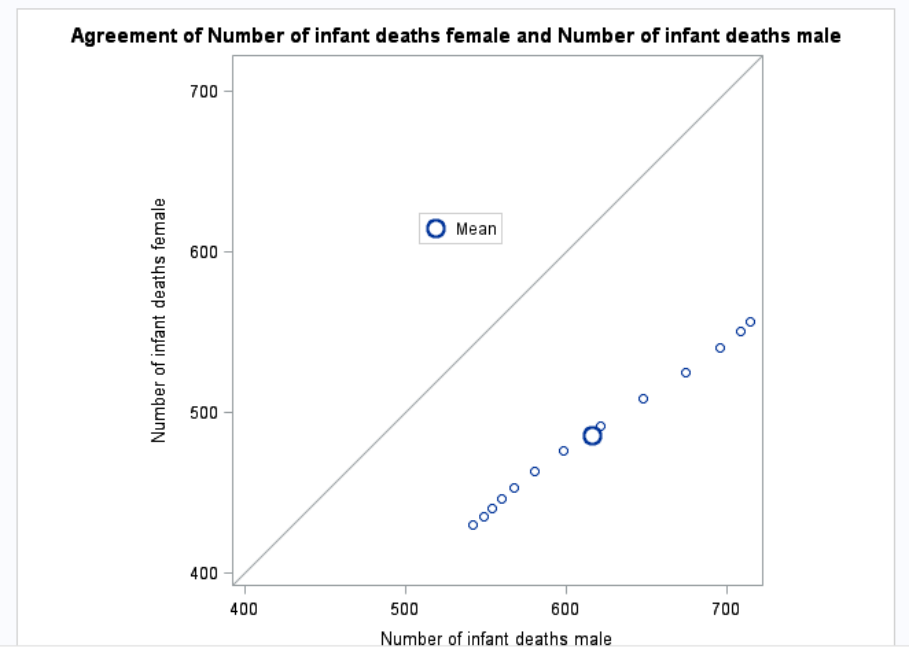
If the centre of the histogram is 0, then we can conclude that there is no difference between the variables Number of infant deaths male and Number of infant deaths female. But the above graphic is catered around the range 125-135. By this also we can conclude that the average of Number of infant deaths male and Number of infant deaths female are not same.

Profile plot:



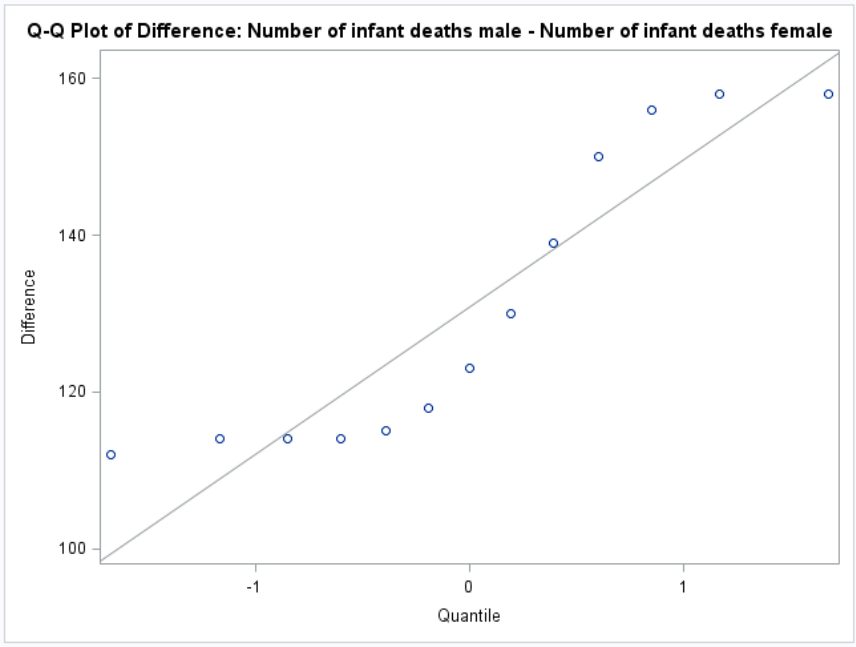
This plot is used to show the behaviour of the variables that are compared. In this case left-side of the graph is Number of infant deaths male and right-side Number of infant deaths female. This graph is in the downward trend. This also shows that the Number of infant deaths male is considerably higher than Number of infant deaths female. The red line represents the average of this trend between the two variables.

Agreement Plot:



This shows the agreement of the of the two variables Number of infant deaths male and Number of infant deaths female. This is a type of a scatter plot. The diagonal is the hypothesis line that values for both the variables are equal. The data point above the line means that Number of infant deaths female are greater in comparison with the Number of infant deaths male and vice versa. We can clearly see that all the data points are below the diagonal line which means that Number of infant deaths male is higher than Number of infant deaths female.

Q-Q plot:



Q-Q plots suggests that the data for the difference between the Number of Infant deaths male and Number of Infant deaths female are around the normally distributed diagonal line. It has been slight deviations from the normality in the heads and tails, but still the normality assumption appears to be satisfied.

Ttest analysis for the country China is shown below:

TTEST For Number of Infant Deaths

The TTEST Procedure

Difference: Number of infant deaths male - Number of infant deaths female

Country Name=China

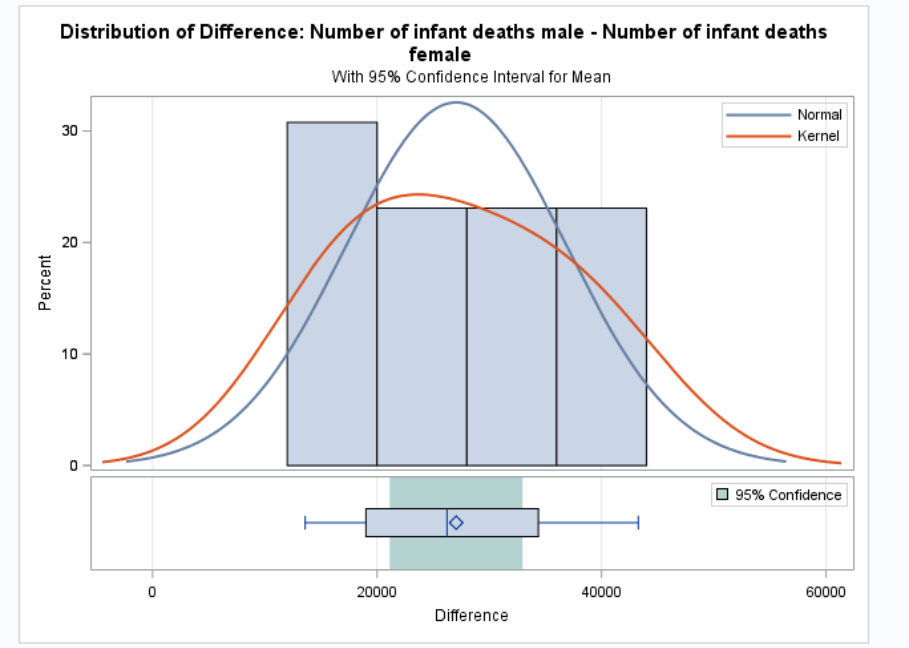
N	Mean	Std Dev	Std Err	Minimum	Maximum
13	27049.6	9800.9	2718.3	13597.0	43273.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
27049.6	21127.0	32972.2	9800.9	7028.1	16178.7

DF	t Value	Pr > |t|
12	9.95	<.0001

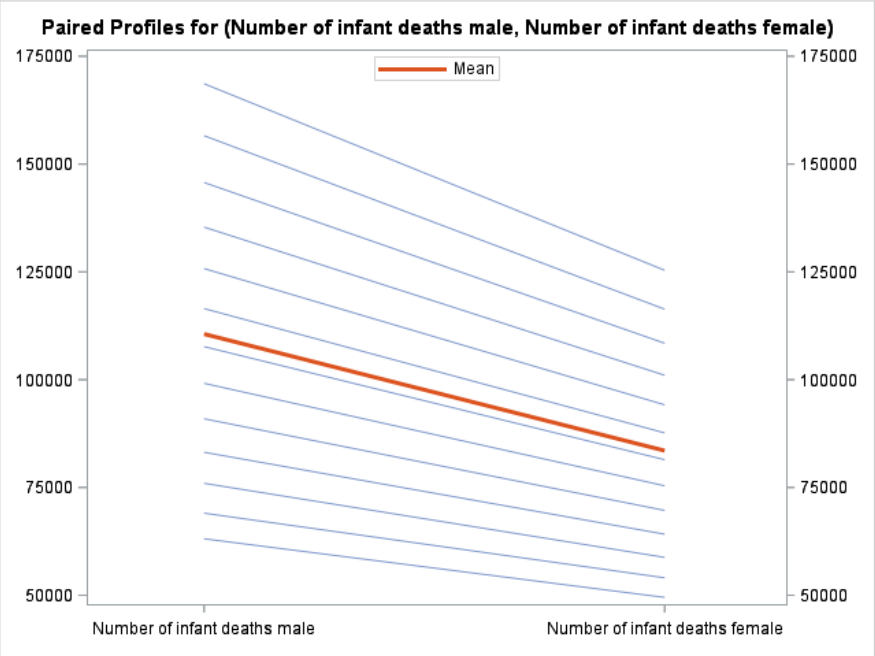
The above graphic shows the mean of 27049.6 for difference of the two variables. Here the P value is very small ($p<0.0001$). Hence, we can reject the null hypothesis.

Distribution Plot:



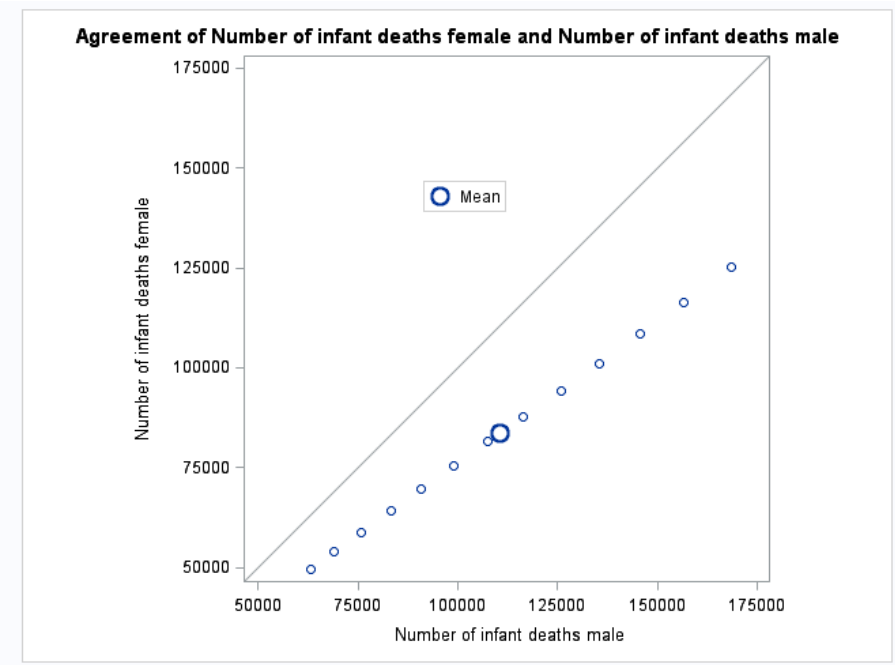
The centre is around the value 26000 – 28000 which is not around 0. This also proves that there is a difference between the average of the two variables Number of Infant deaths male and Number of Infant deaths female.

Paired profile:



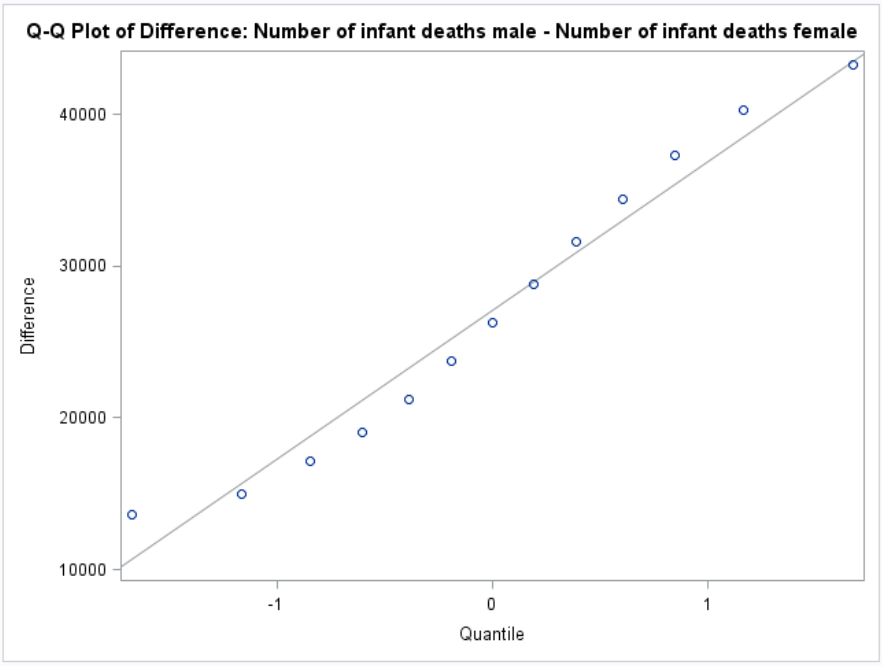
We can clearly see that the graph is in the downward trend which is moving from Number of Infant deaths male to Number of Infant deaths female. This also proves that there is difference between them.

Agreement plot:



The findings from this graph are all the data points are below the diagonal line which shows that the variable Number of Infant deaths male is greater than Number of Infant deaths female.

QQ Plot:



The data for the difference of Number of Infant Deaths male & Number of Infant deaths female appears to be uniformly distributed, according to Q-Q plots. Although there have been minor deviations from normalcy in the heads and tails, the normality assumption looks to be met.

Results for all the countries are in the below pdf:

Ttest_results.pdf

ANOVA Test:

One way analysis of variance (ANOVA) is used to determine if there are any difference between the three or more independent groups. In this test a dependent variable is to be measured with one or more independent variables. This test can be performed only on the balanced data.

Syntax for ANOVA:

The following statements are available in the ANOVA procedure:

PROC ANOVA *<options>*;
CLASS *variables </ option>*;
MODEL *dependents = effects </ options>*;
ABSORB *variables*;
BY *variables*;
FREQ *variable*;
MANOVA *<test-options> </ detail-options>*;
MEANS *effects </ options>*;
REPEATED *factor-specification </ options>*;
TEST <H=*effects*> E=*effect*;

In order to do the ANOVA test, we need a categorical variable in order to fit the model. This categorical variable has been created from the below code:

```
data population;
set work.population;
if 'Rural population growth (annual'n > 0 then
Trend = "Increase";
else Trend = "Decrease";
run;
```

The above code adds a new column Trend to the dataset which has two values Increase if rural population growth rate is positive, decrease if rural population growth rate is negative.

Code for ANOVA test:

```
proc anova data=population;
class Trend;
by'Country Name'n;
model population = Trend;
title "One-Way ANOVA with population as Predictor";
run;
```

Output for the country United States is as shown below:

One-Way ANOVA with population as Predictor

The ANOVA Procedure

Dependent Variable: Population Population

Country Name=United States

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.4030127E14	5.4030127E14	14.72	0.0028
Error	11	4.0363758E14	3.6694325E13		
Corrected Total	12	9.4393885E14			

R-Square	Coeff Var	Root MSE	Population Mean
0.572390	1.918156	6057584	315802484

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Trend	1	5.4030127E14	5.4030127E14	14.72	0.0028

The total number of the levels we used are two (increase, decrease). Analysis results can be known Degrees of Freedom (DF). Model DF can be calculated by number of levels – 1.

Model Df = Number of levels – 1.

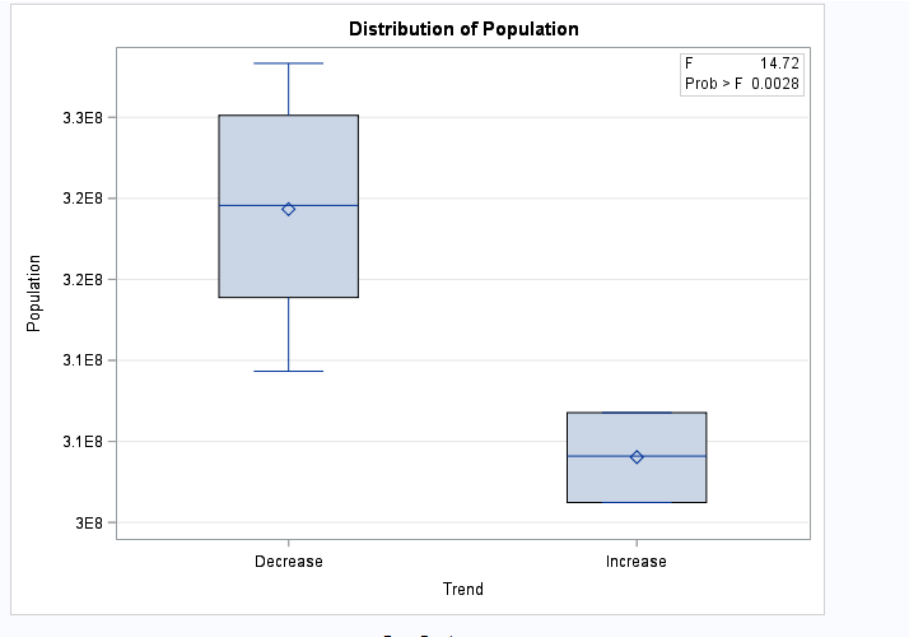
Model Df = 2 -1 = 1 (can be verified from the first table).

The corrected total can be calculated by total observations – 1.

Corrected Total = 13(Total observations) -1 = 12 (can be verified from the first table)

The variability between the groups can be obtained by the F value. If it has the high value then you can say that all the means are not equal and there is a difference between them which shows that null hypothesis can be not accepted.

Distribution plot:



Output for the Country Germany is shown below:

One-Way ANOVA with population as Predictor

The ANOVA Procedure

Dependent Variable: Population Population

Country Name=Germany

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	185118999077	185118999077	0.20	0.6650
Error	11	1.0289275E13	935388648323		
Corrected Total	12	1.0474394E13			

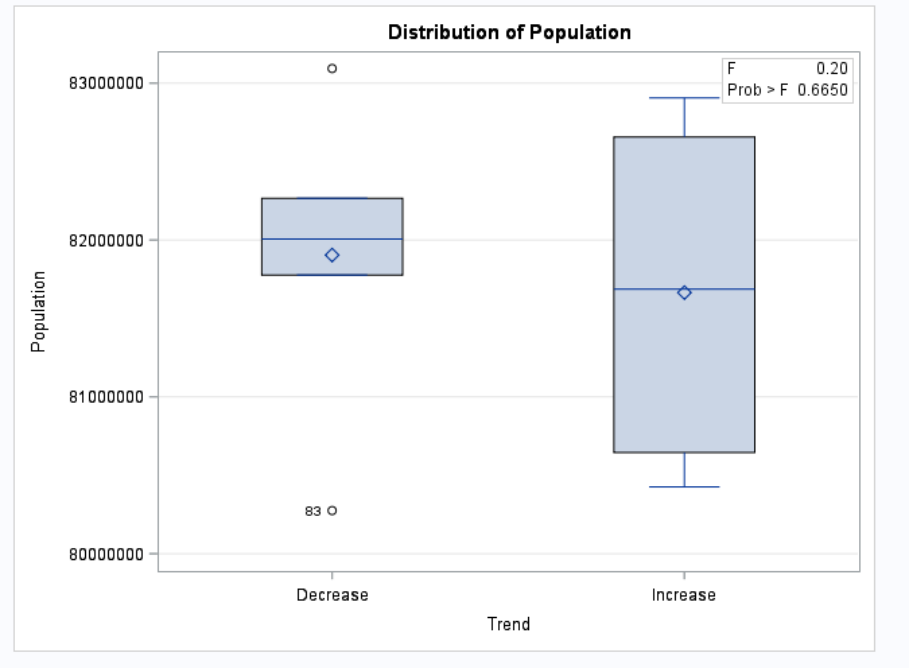
R-Square	Coeff Var	Root MSE	Population Mean
0.017673	1.182702	967154.9	81775049

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Trend	1	185118999077	185118999077	0.20	0.6650

Degrees of Freedom for model is 1 and Corrected total is 12. Here the F value is 0.20. F value is small indicates that there is a slight difference between them. Since F is very small, so we

cannot reject the Null hypothesis here which can be verified in distribution plot where the means are having slight difference.

Distribution plot:



Results for all the countries are in the below pdf.

Anova Results.pdf

5. Discussion

Initially to find the outliers, a box plot has been drawn. Univariate analysis for the all the variables in the dataset for each country has been done. Variables population, Life expectancy at birth for the country Australia are being discussed here. The distribution plot for the same variables is plotted and they are slightly negative skewed with skewness of 0.00169835 and -0.152542. Variables population, Number of Infant deaths for the country India has been negatively & positively skewed respectively with skewness of -0.1121615 and 0.31648893.

Correlation analysis for all the variables in the dataset for each country has been performed. Correlation analysis for the variable's population, Number of infant deaths are negatively correlated which can be verified in the scatter plot for both the countries Australia & United Kingdom.

There is a strong positive correlation for variables population, Age dependency ratio.

Based on the results obtained in the correlation analysis, a linear regression model has been developed for the variables population & Number of Infant deaths. The fit plot obtained is in downward trend with negative slope. Thus, the correlation analysis can be verified as well.

Paired Ttest for the variables Number of infant deaths male and Number of infant deaths female has been performed for each country. The output obtained suggests that there is a significant difference between two of them and Null Hypothesis cannot be accepted.

Anova test for the Categorical variable Trend has been done based on each country. The results obtained that Null hypothesis cannot be accepted for some countries and can be accepted slightly for some countries based on the F value obtained.

6. Conclusion

There has been negative correlation between the population & Number of infant deaths for the countries Australia & United Kingdom. Based on this, the linear regression model has been developed and the results obtained also supports them. They have a negative slope with downward trend as population is increasing the Number of infant deaths is decreasing. This has been observed in the dashboard as well in the sheet of Infant mortality rate. Thus, sheet in the dashboard has the support of the statistical analysis as well. Ttest has been performed for Number of infant deaths male & Number of infant deaths female. The result obtained shows that the Number of infant deaths male is higher than the Number of infant deaths female. Thus, Null hypothesis is not accepted. Anova test for the rural population growth is done for each country by creating a categorical variable.

APPENDIX

SAS programming code:

```
%let path = D:\Sriram\Course work PODS;
libname stat1 "&path";
* write file importing in the document*;
proc import datafile="&path\Population_sas.xlsx" out= population replace;
run;
*adding two colums to the dataset*;
data population;
set population;
urban_population =  'Percent of Urban population'n * Population;
rural_population = 'Percent of Rural population'n  * population;
run;
*sorting the dataset based on country*;
proc sort data=work.population;
by 'Country Name'n;
run;
ods select ssplots;
ods graphics / reset=all imagemap;
*performing the outliers detection using box plot*;
proc univariate data=work.population plot;
by 'Country Name'n;
Run;
*printing the dataset*;
proc print data=population;
run;


*performing the univariate analysis*;
proc univariate data= population;
BY 'Country Name'n;
var Population male female urban_population rural_population 'Number of infant deaths'n 'Number of infant deaths male'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
'Age dependency ratio'n;
histogram Population male female urban_population rural_population 'Number of infant deaths'n 'Number of infant deaths male'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
'Age dependency ratio'n /normal(noprint);
run;

ods select scatterplot;
*performing corelation analysis *;
proc corr data=population plots=matrix(hist);
by 'Country Name'n;
var Population 'Number of infant deaths'n 'Number of infant deaths male'n 'Age dependency ratio'n
'Number of infant deaths female'n 'Life expectancy at birth total'n 'Life expectancy at birth female'n 'Life expectancy at birth male'n
 male female urban_population rural_population ;
 title "Correlation Analysis for each country";
run;
*performing the Regression analysis*;
proc reg data = population;
by 'Country Name'n;
model 'Number of infant deaths'n = population ;
title "Linear Regression Analysis on Population & Number of Infant deaths for each country";
run;
```

```
*performing the paired Ttest*;
proc ttest data = population alpha = 0.05;
     paired 'Number of infant deaths male'n*'Number of infant deaths female'n;
by 'Country Name'n;
     title "TTEST For Number of Infant Deaths";
   run;
*adding a new categorical variable*;
data population;
set work.population;
if 'Rural population growth (annual'n > 0 then
Trend = "Increase";
else Trend = "Decrease";
run;
*performing the one way ANOVA test*;
proc anova data=population;
class Trend;
by'Country Name'n;
model population = Trend;
title "One-Way ANOVA with population as Predictor";
run;
```