National College of Ireland

Bachelor of Science (Honours) in Computing Information

Data Analytics

2020/2021

Luke Sheehan

x17361401

x17361401@student.ncirl.ie

Life Expectancy Project

Technical Report

Contents

Executive Summary

In this report I will be discussing my life expectancy project. I will be covering the background and motivation for this project as well as the key insights and findings that I have taken from it. I will also be explaining the technologies, approach, and data sets that I used to complete my project.

This project focuses on the topic of life expectancy and all the data that surrounds it. This project is built around the fact that there are many different factors that affect life expectancy, and some are not as obvious as others. The aim of the project is to use different data mining and data analytics techniques to help answer the question: what are the factors that affect life expectancy and how does life expectancy differ around the world?

To answer this question, I used the KDD data mining methodology to extract information from the data that I have sourced. Through KDD I discovered insights that help uncover factors that contribute to life expectancy. The data analytics techniques used in this project are Random Forests, clustering, linear and multiple regressions.

Through the KDD methodology and through the data analytics techniques listed, I was able to answer the question that I set for myself (i.e., what are the factors that affect life expectancy?) and produce visualisations of my findings. The key findings from my project are:

- From my Linear regressions I determined that GDP per capita and levels of schooling both correlate to life expectancy.
- Ireland countries has a higher life expectancy than the rest of the world and Ireland's life expectancy is growing at a faster rate as well.
- Through Random Forests I discovered that the most important factors for determining life expectancy are: Income composition of resources, HIV rates and schooling levels.

1.0 Introduction

1.1. Background

Life expectancy at birth is widely accepted to be an indicator of the development of a country. Through this measure alone, major insights can be made on a country's health and social infrastructure. From my research into previous studies on life expectancy I have learned that the list of factors that have been discovered to correlate to life expectancy is growing constantly. Through my research I have learned that there have been many misconceptions on what are the main factors that affect life expectancy. Therefore, I undertook this this project to address these misconceptions and present the real factors of life expectancy through means of data analytics. I also wanted to present through data analytics how life expectancy varies around the world.

I first came up with this idea since I had an interest in statistics about life expectancy and GDP in different parts of the world. I was inspired from a time I watched a news report from England where it was revealed that the life expectancy in England varied drastically depending on where a person was born. I found this to be very interesting and got thinking about what are all the factors that affect a person's life expectancy since it could not just be genetics. It got me thinking about how someone's life can be affected by the differences in their culture and society.

I was also inspired by the book: Man's Body an Owner's Manual. (Diagram Group, 1999) This was a book that provided a complete guide to the mental and physical workings of the male body. It had comprehensive statistics on the male body and health. It allowed readers to chart their own life expectancy based on factors such as where they were born, how often they exercise, do they smoke etc. I found this idea of being able to predict one's own life expectancy to not only be interesting but also educational as one could use this information as an eye opener into how healthy one is and what negative habits one might have.

As I planned on looking at what factors determined life expectancy, looking at correlations and performing linear regressions with the data that I sourced was an obvious choice of approach. Linear regressions are a technique used in data analytics where there is a dependent and independent variable. In this case the dependent variable is life expectancy and the independent variables for my project include factors such as GDP, levels of schooling and happiness rates. Through linear regressions I would be able to predict the value of the dependent variable based on the value of the independent variables. I would also be able to look at how well these two variables correlate to each other.

I was very curious to see if clusters of different countries would form based on their data so performing clustering on the data was an obvious choice of data analysis. For my project I wanted a deeper look at all the factors that affect life expectancy, so I performed Random Forests on my data to see which factors were the most important for determining life expectancy. From my research into other reports and studies on life expectancy I have seen factors that affect life expectancy discussed but I have not seen a true ranking of all the different factors that determine life expectancy.

To compare Ireland to the rest of the world with regards to life expectancy I plotted Ireland against the average life expectancy around the world to see how different Ireland is. I was able to use this data to make predictions on how life expectancy will change for Ireland and the world in the next 10 years.

1.2. Aims

The main objective for my project is to provide an analysis on life expectancy. I aim on providing an analysis on what are the factors that affect life expectancy around the world through data analytics and by creating visualisations. I also aim on using my data to see what are the factors that correlate towards a country's life expectancy e.g., levels of schooling or happiness rates. I aim to display my findings through visualisations that I created through R.

To accomplish these objectives, I have a set of smaller objectives to accomplish. First, I will be gathering relevant data on the following topics:

- Life expectancy around the world
- GDP per country
- Happiness rates around the world
- Smoking rates around the world

With regards to insights into the data sets I have analysed, some examples of insights that I have gained through my research have been:

Is there a correlation between the life expectancy of a particular country and that country's GDP per capita? To test for this, I merged my data sets together and created a linear regression of life expectancy against GDP per capita. I chose to use a linear regression visualisation to visualise the correlation between these two factors. When one factor increases as the other one increases a diagonal linear line shape will be created on the graph stemming from the graph's origin. This means there a strong positive correlation between the two factors.

I also tested to see if there are similarities in life expectancy based on a country's region e.g., Europe, North America, Africa etc. To answer this, I created a scatter plot of the countries in my data set giving a unique colour to each region. From the scatter plot it was easy to see how each region compared to each other.

One of the aims for my project was to create an R Shiny page to host the visualisations and findings of my project. This was an aim of my project as I believed that using a program such as R Shiny is a more user-friendly way of presenting my findings than through using a document alone.

I aimed to benefit the users and readers of my project by clearly presenting the findings and results of my project so that they can gain an understanding on the topic of life expectancy. I hope they will learn what the different factors that affect life expectancy are but also learn

about how countries differ around the world. I also provided a true ranking of the factors that affect life expectancy through Random Forests.

1.3. Technology

To gather the data sets for my project I have been using the Google data set search engine to find data sets on life expectancy. I have been using the website Kaggle to find data sets related to life expectancy as well. On this website I have been able to download data sets in the csv file format.

After I downloaded these data sets, I was able to pre-process them using the language R in R studio. Through R studio I was able to merge multiple data sets and create insights and visualisations from these data sets. To create the visualisations, I used many R packages to create graphs and charts.

The R packages I used in my project are tidyverse, gridExtra, reshape2, ggplot2, ggthemes, scales, dplyr, mice, randomForest and Amelia.

The majority of my coding for this project was done through the language R. I chose to use this language as it is a statistical computing language that is great for data analytics and is also able to create graphics. Through R I was able to process my data, merge data sets and run many different types of algorithms and tests. I was able to test for correlations in my data and produce visualisations of my findings. I was also able to run algorithms on my data.

The algorithms that I used in my project were Random Forests and clustering algorithms. These algorithms helped me to gain key insights into the information in my data set. Through Random Forests I was able to determine the most/least important factors for determining life expectancy. Through clustering I was able to discover clusters of countries that formed in the data set.

To create a web application of my visualisations I used R Shiny. This tool allowed me to build web applications that could display my findings in a user-friendly manner. I used a website called Shinyapps.io to deploy my applications to the web. This website allowed me to run different Shiny applications at the same time and generate usage information on these applications as well.

The model that I followed for this project is KDD. I chose this model as it provides a phase-by-phase method of finding, transforming, and refining meaningful data and patterns from a data set in order to extract information from this data.

For running descriptive statistics and statistical tests on my data I used SPSS. Through SPSS I was able to test for factors such as normality in my data.

1.4. Structure

The remainder of the report is organized as follows. In Section 2 of the document, I will be describing in detail all the data used for my project, this includes all the data sets

used as well as how these data sets were sourced. In this section I will also discuss how the data was compiled in a manner that exploratory analysis could be applied to it.

Section 3 will explain the methodology used for this project. I will be explaining the methodology I chose and why I chose it. In this section I will be going through each step of the methodology and explaining how I completed each step of the methodology to get from my data for my analysis.

In Section 4 of this document, I will be describing the statistical tests that I conducted on my data, these tests were all completed through SPSS. In section 5 I will be presenting the tests that I conducted for my project throughout the project's life cycle. These include validation and functionality tests.

In Section 6 of this document, I will be explaining the analysis that I performed for this project. This section includes the approaches that I used to analyse my data. My reasons for choosing these approaches and the decisions I made with regards to choosing what aspects of the data I will be using for my models and visualisations.

Section 7 presents the results of my analysis. In this section I will be discussing all my results as well as presenting the visualisations of my findings. In section 8 I will be describing how I deployed my Shiny apps to the cloud using Shinyapps.io.

In section 9 I will be concluding my work and presenting the key findings of my project. In this section I will be answering the question of the project that I set for myself (i.e., what are the main factors that affect life expectancy?). I will also be discussing the impact that I aim for my project to have on users/readers of my project. Also, as part of the conclusions of this project I will be analysing my project itself and discussing its strengths, weaknesses, and limitations.

In Section 10 of my project, I will be discussing the possible future development and research I could perform if I were to expand on it. Finally in this project I will be citing my references and I also have included my original project proposal as well as my monthly reports that I completed throughout the academic year to keep track of my progress for the project.

2.0 Data

The two main data sets I have used for this project are the Life Expectancy data set from the world health order and the world happiness report data set. I have sourced these data sets from Kaggle. I was able to download these two data sets in the csv file format. I was then able to start cleaning the data in these data sets through R studio.

The main data set that I used for this project is the Life Expectancy (WHO) data set. This is a data set created by the World Health Organisation using the Global Health Observatory. This data set has data from 193 countries on both social and health factors of each country. This data set has information from each country between the years of 2000 to 2015. This data set has a total of 22 columns for each country. These columns include: Country, year, status (developing/developed), life expectancy, adult mortality, infant deaths, alcohol consumption levels, percentage expenditure, hepatitis B, measles, BMI, under-five deaths, polio infection rates, total expenditure, diphtheria, HIV/AIDS, GDP, population, thinness, thinness 5-9 years, income composition of resources and schooling levels. (Rajarshi, 2021)

I chose this data set to be my primary data set as it came from a reliable organisation (WHO) and because of its extensive information from countries around the world. As my aim for my project is to analyse what affects life expectancy, I knew I had to find a data set which had many different factors that I could compare life expectancy to and hopefully find correlations.

I found this data set on Kaggle by using Google's data set search engine. Through this data set I was able to find more data sets on countries which allowed me to expand my analysis. Another data set that I found through this search engine was the World Happiness Report data set. (Singh, 2021)

The World Happiness Report data set contains information from 149 different countries and has a total of 20 columns of data, each containing useful data on the social factors of different countries such as generosity, freedom levels and corruption levels. Through merging this dataset with my life expectancy data set I was able to discover correlations between life expectancy and these social factors which led to unique insights in the data. Through the merging of data sets I was able to gain insights such as does happiness levels affect life expectancy.

I chose to use this data set as it has unique data on social factors from different countries and because the World Happiness Report is a well-recognised data set that has been ongoing since 2012. This data set was officially released in 2017 by the United Nations as part of International Happiness Day. The data in this report has been used by civil organisations and governments around the world to inform policy making decision.

After I downloaded the data sets, I first explored them for missing or inaccurate values. Through R I was able to identify any missing cells in the data sets which had the NA value. To tackle the missing values, I was able to replace them with the mean value of the column and

for some cases when creating visualisations from the data I replaced the missing values with 0 to allow the visualisations to be made.

For columns in my data set with large amounts of missing values I was able to filter out through R, in certain cases I replaced the columns in my data set with columns for other data sets where the data was more accurate and had no missing values. When analysing the column in my data set on GDP I discovered that a lot of the data was inaccurate when compared to other data sets which had information on GDP. In my GDP column, some countries had their GDP as GDP per capita and other countries had their GDP as their country's overall GDP.

Cleaning my data set

Before I could get started on implementing any methodology or performing any analysis on my data set, I had to make sure that all the data inside the data set was accurate and lacked any missing values. Through performing some statistical analysis on my main data set (life expectancy by the WHO) I was able to see that certain columns had large amounts of missing and inaccurate values. One of the columns in the data set that was very important for my analysis was GDP per capita. Through R I noticed that there were large amounts of inaccurate values that would affect the results of my analysis.

To tackle this issue, I found another data set which had accurate values on GDP for all the countries in my original data set. I found this data set from the website: Our World in Data. Our World in Data is an online scientific publication website that provides data on many different topics around the world such as disease, poverty, and climate change. Through their repositories I was able to find their data set for GDP. As I now had more accurate data on GDP, I was able to merge this data with my current data set replacing the GDP per capita column with the accurate data. This provided me with more accurate data for my analysis. (owid/owid-datasets, 2021)

In my data set there were some columns that had missing values. It was important for me to replace these missing values to draw accurate conclusions from the data. These missing values would also negatively affect any visualisations made from the data. A common and simple approach to this issue is to remove columns with missing values or to replace the missing values with 0. These solutions are quick, but they can also affect the accuracy of any analysis performed on the data.

To tackle this problem, I decided to replace the missing values in each column with the mean values of the column. To do this I ran a function on my data set which allowed me to see how many missing values there were per column. Most columns had none but the four columns; Hepatitis levels, Average BMI, thinness levels age 1-19 and thinness levels age 5-9 all had missing values. To modify these columns, I first calculated the mean vales of each column. I then was then able to replace any cells in the column which had the NA value with the mean value of that column. This allowed me to perform analysis on more accurate data

in the data set. It also allowed me to perform algorithms on my data set such as Random Forests and clustering.

3.0 Methodology

For my project I decided to use the KDD methodology. I decided to use this methodology since it is a very popular methodology used in data analytics projects. KDD is a data mining process with many steps and tasks. The steps begin with data and ends with knowledge and insights. The main benefits of KDD include discovering patterns in data sets, extracting insights from data and being able to automatically summarize data.

The steps/tasks involved in KDD are as follows: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Knowledge Representation.

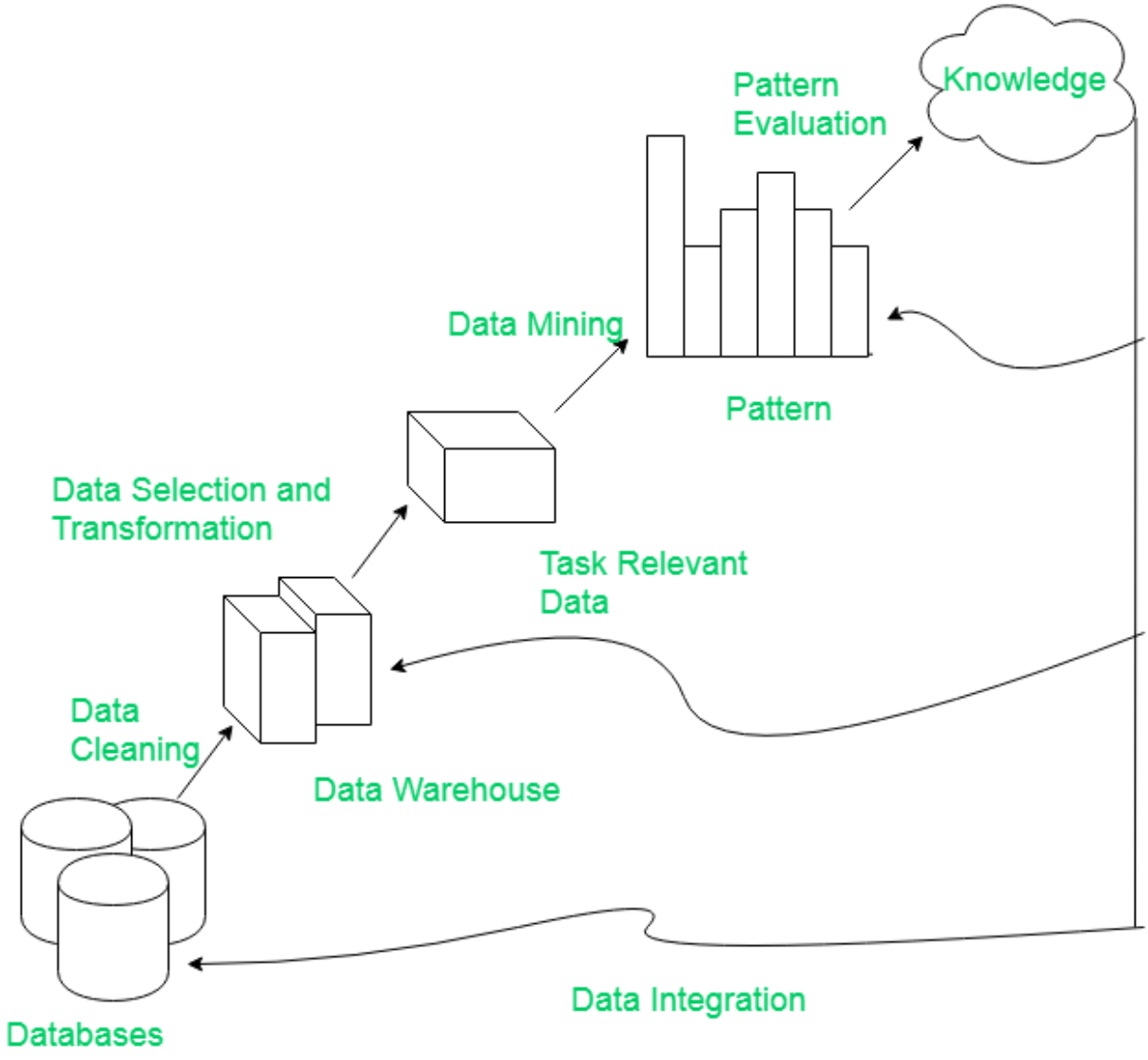


Figure 1 Diagram of the KDD process

I have been using the KDD methodology to create insights into the data sets I have sourced and processed.

Data Cleaning

Before I could begin to start extracting information from my data sets it was important to clean the data of any missing or inaccurate values. To clean the data, I used a combination of data discrepancy detection tools and techniques through R to clean the data. Through R I was able to remove the missing values from the data set and replace them with the mean values of the column they were in. In some cases, whole columns had to be filtered out due to variance errors. For columns with a large number of missing values I was able to replace them with accurate data by merging those columns with columns for other data sets.

Data Integration

Data Integration was a key element of this methodology for my project. As I planned on comparing life expectancy data to data from other data sets it was very important for me to be able to merge data sets together so that I could analyse the data from one common source. I was able to achieve this by merging data sets through the merge command in R and by combining the data based on country name. The data integration step was also important for replacing columns of inaccurate values with columns from other data sets with accurate values.

Data Selection

The data selection phase of KDD is the process where relevant data for the analysis is found. The initial data sets for my project were found through Google's data set search engine and through the website Kaggle. In order to find more relevant data within these data sets I used different approaches such as clustering, regressions and neural networks.

Data Transformation

The data transformation step of KDD is the process of transforming the data collected from its initial form to a more appropriate form for data mining. This includes merging data sets so they can be analysed from the same source, renaming columns and rows to make the data mining process easier for coding and for visualisations and data mapping which is an important part of merging data sets. In the data mapping stage, fields are matched from one data set to the other which enables data integration.

Data Mining

The data mining step of KDD is arguably the most important step in the process. It is in this step that patterns and insights are extracted from the data. This can involve the many different data analytics techniques such as clustering and linear regressions. In this step data is transformed into patterns which can provide useful insights on the data.

Pattern Evaluation

The pattern evaluation step is where the patterns and insights that were discovered in the data mining step are evaluated to determine how useful they are. In this step the

knowledge discovered is represented through various means such as visualisations and statistics about the data. In this step it is important to be able to create presentable visualisations on the data so that the readers and users of my project can understand them clearly. For my project I created a wide set of visualisations which display the insights in the data that I have discovered.

Knowledge Representation

Knowledge representation is the step in KDD where the pattern and trends in the data that have been evaluated are visualised and presented to represent the key insights in the data. This can include, reports, bar plots, line charts and tables. In this step I represented the most useful insights with visualisations to represent my analysis. These visualisations were made using a range of packages in R and were represented together using R Shiny. Some examples of visualisations that I made were, scatter plots and line charts to show correlations in the data.

(KDD Process in Data Mining - GeeksforGeeks, 2021)

4.0 Statistical Tests

Prior to my analysis on my data, I performed a series of statistical tests on my data set to gain a better understanding of the data as well as testing assumptions I had on the data.

4.1. Normality Test

The first statistical test I performed was a normality test through SPSS. I performed this test to see if the data I was using was normally distributed. For this test I looked at the variables, GDP, and life expectancy. I will be determining this through normality tests such as the Shapiro Wilks test. The **null hypothesis** for this test is that the values are sampled from a population that follows a normal distribution. The **alternative hypothesis** is that the values are not sampled from a population that follows a normal distribution. For this test I set the level of significance to 0.05 as that is the standard for statistical tests. Therefore, when I conduct the Shapiro Wilks test, if the Shapiro Wilks test value is over 0.05 then I can accept the null hypothesis that the data is sampled from a normal distribution.

The results of the normality test conducted through SPSS:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
GDP	152	100.0%	0	0.0%	152	100.0%
Life.expectancy	152	100.0%	0	0.0%	152	100.0%

Figure 2 Case Processing Summary of the variables: GDP and Life expectancy

This shows that there are 152 values for each variable with 0 missing values.

Descriptives

			Statistic	Std. Error
GDP	Mean		17526.534867399	1580.0914933267
	95% Confidence Interval for Mean	Lower Bound	14404.591795351	
		Upper Bound	20648.477939447	
	5% Trimmed Mean		15126.021416705	
	Median		11466.985155000	
	Variance		379496747.347	
	Std. Deviation		19480.676254870 3	
	Minimum		566.8460070	
	Maximum		134182.4147000	
	Range		133615.5686930	
	Interquartile Range		20965.0392330	
	Skewness		2.471	.197
	Kurtosis		9.305	.391
Life.expectancy	Mean		71.844	.6529
	95% Confidence Interval for Mean	Lower Bound	70.554	
		Upper Bound	73.134	
	5% Trimmed Mean		72.094	
	Median		73.950	
	Variance		64.793	
	Std. Deviation		8.0494	
	Minimum		51.0	
	Maximum		88.0	
	Range		37.0	
	Interquartile Range		10.9	
	Skewness		-.471	.197
	Kurtosis		-.424	.391

Figure 3 Descriptive Statistics on the GDP and Life expectancy variable

The above table contains detailed descriptive statistics for each of the variables used. For the normality test the main value to look at is the Skewness level. For GDP, the skewness level is 2.471 which suggests a moderate positive skew to this data. For Life expectancy

there is a negative skew of -471. This would suggest that the Life expectancy data is not normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
GDP	.192	152	.000	.764	152	.000
Life.expectancy	.114	152	.000	.967	152	.001

a. Lilliefors Significance Correction

Figure 4 Results of Kolmogorov-Smirnov and Shapiro Wilk test

The next table is the results of the Kolmogorov-Smirnov tests and the Shapiro-Wilks tests. For smaller samples such as the samples I am using for this test, the Shapiro-Wilks test is the relevant test. For both variables the P value/level of significance is below the alpha value of 0.05 which determines that both variables are not normally distributed.

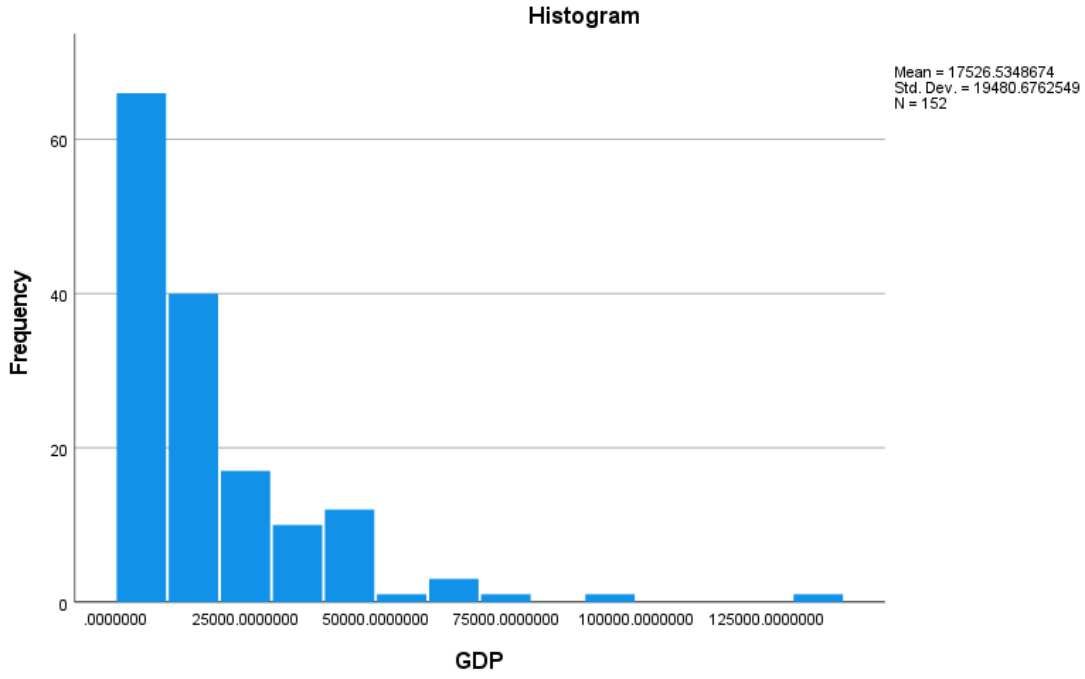


Figure 5 Distribution of GDP

For GDP we can see that the histogram is not bell shaped which is a sign of non-normally distributed data.

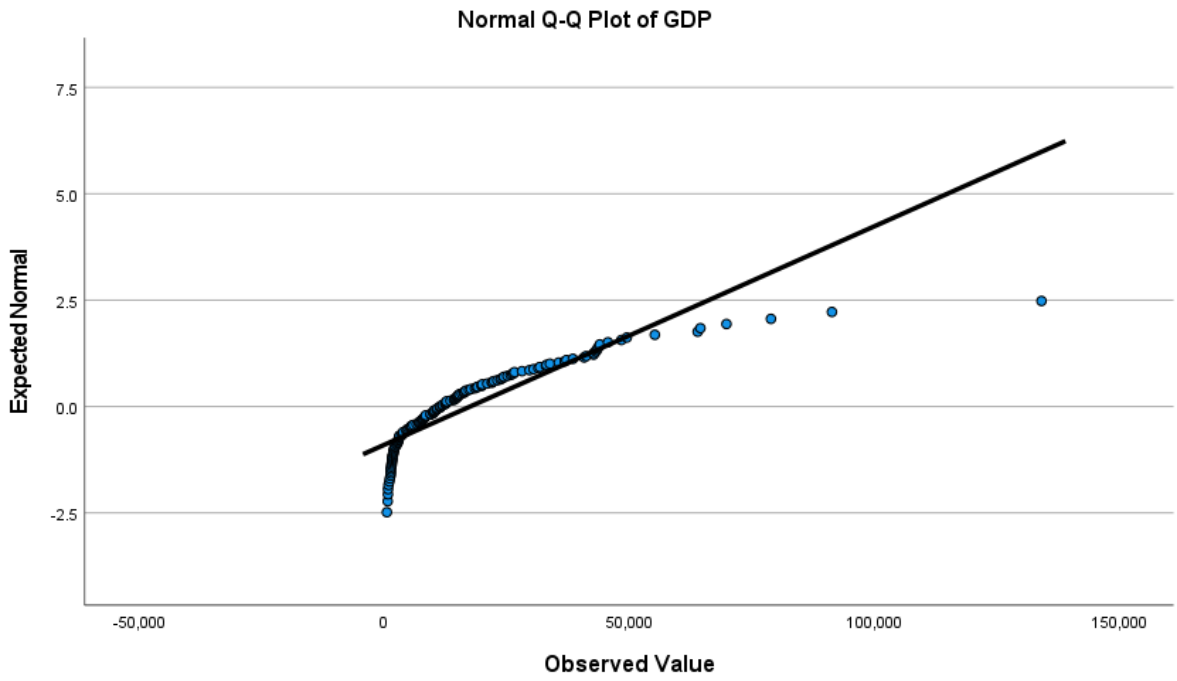


Figure 6 Q-Q plot of GDP

In the Q-Q plot we can see that most of the data does not follow the line. Values along this line suggest normally distributed data so in our case it suggests non normally distributed data.

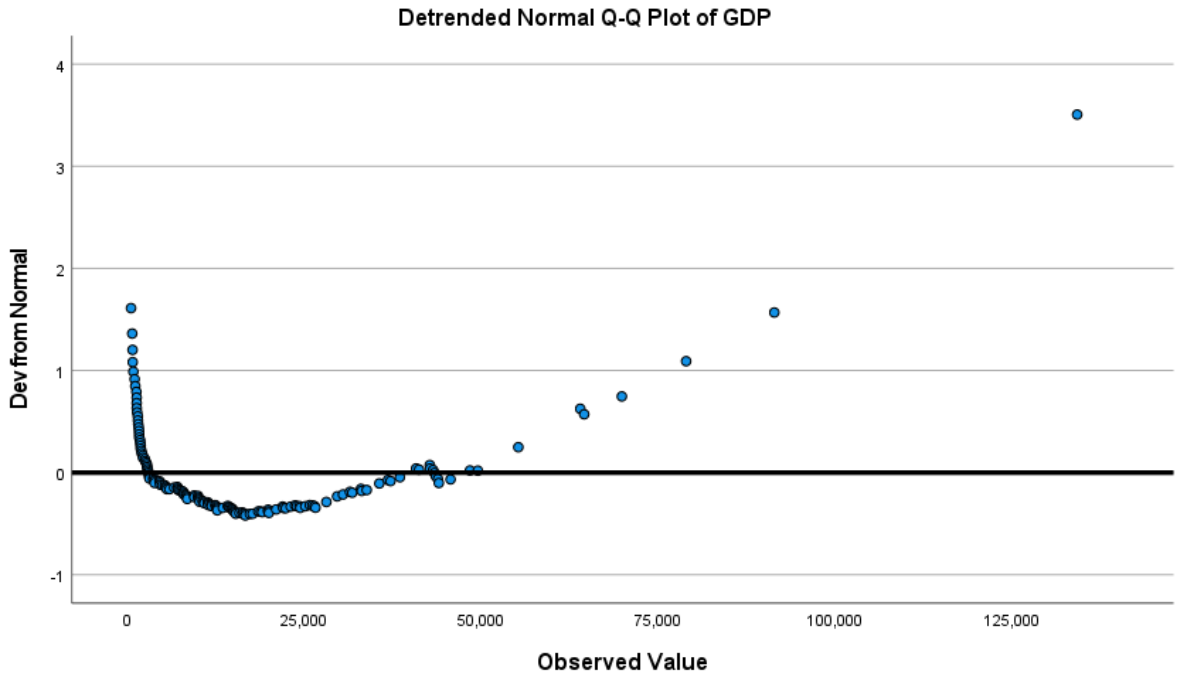


Figure 7 Detrended Plot of GDP

In the detrended plot we can see that most of the data is not close to 0 which suggests non normally distributed data.

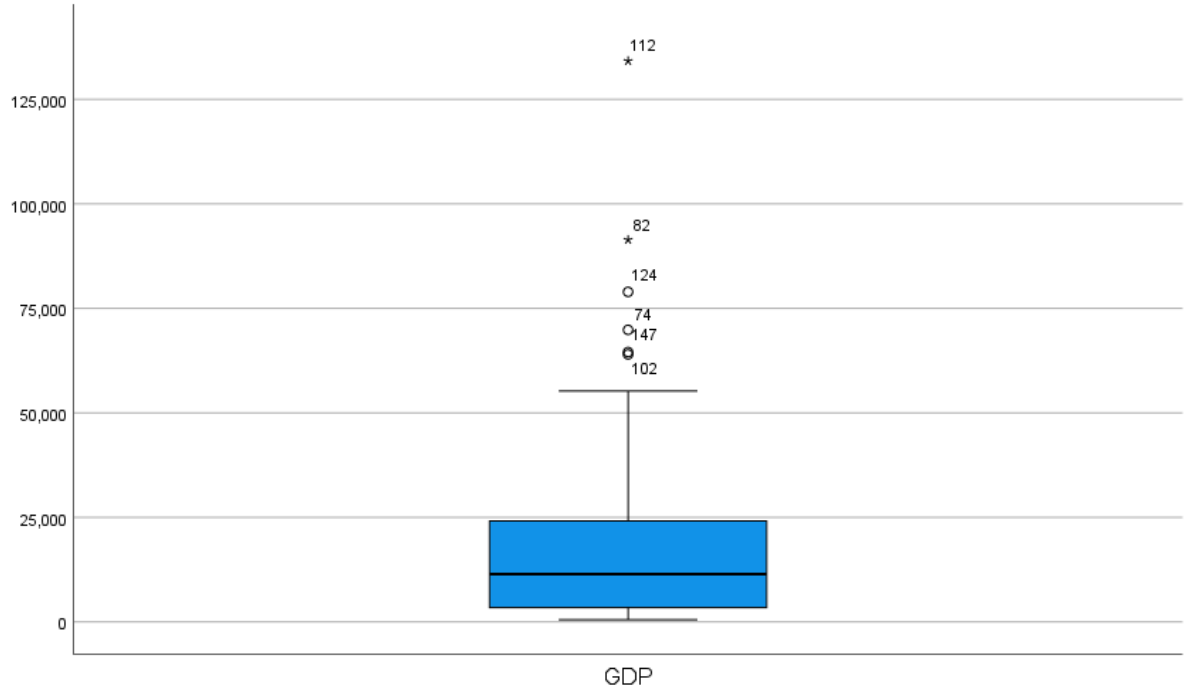


Figure 8 Box and Whisker Plot of GDP

Finally, in the box plot diagram we can see that the box is far from the centre, we can also see outliers in the data far from the mean GDP values. This is another strong indication of a non-normally distributed sample.

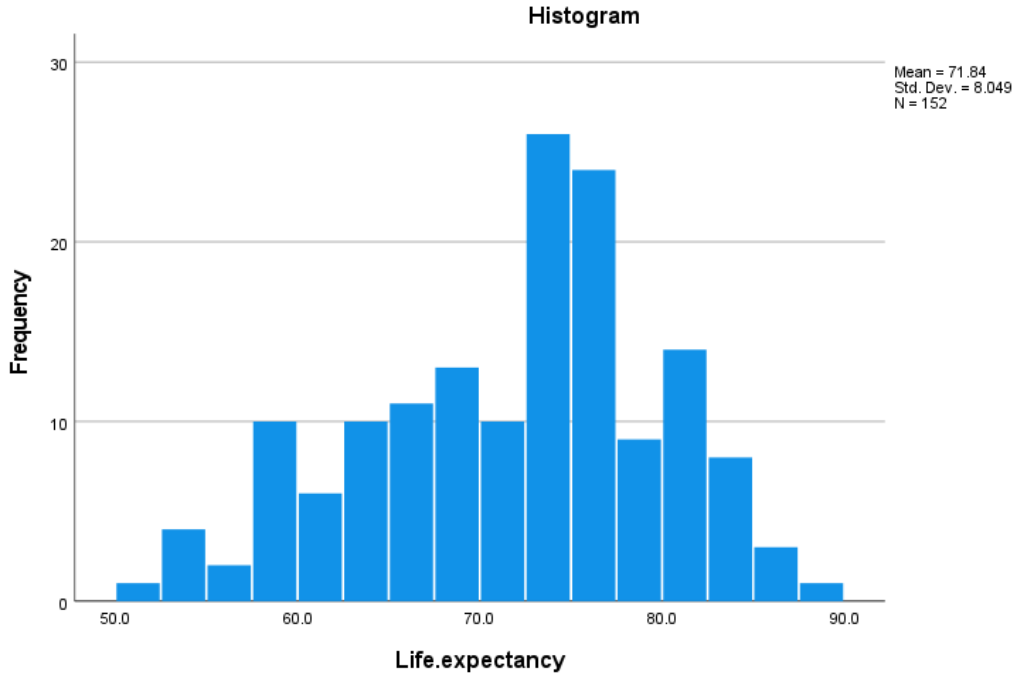


Figure 9 Distribution of Life expectancy

For life expectancy we can see that the data follows more of a bell curve shape but there is a positive skew to the data. We can see that the data sampled for this variable is also not normally distributed but to a lesser extent.

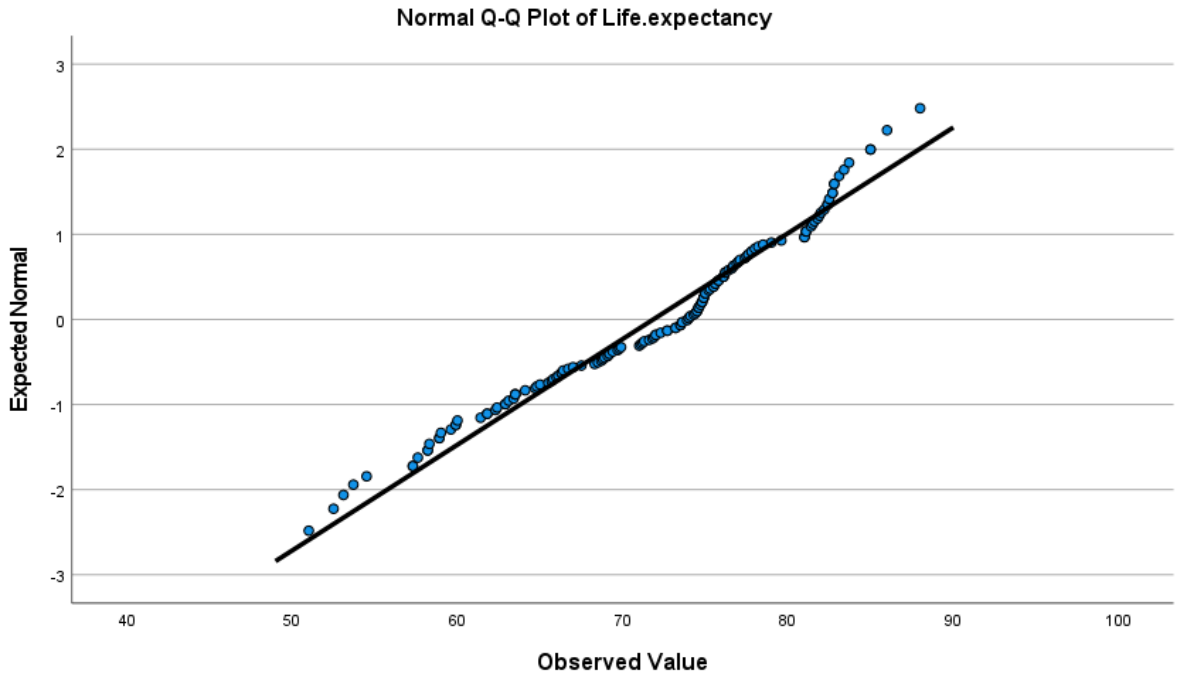


Figure 10 Q-Q plot of life expectancy

In the Q-Q plot we can see that more data is along the line which suggests that it is more normally distributed than GDP but there are still many values off the line.

Figure 11 Detrended Q-Q Plot of Life expectancy

In the detrended plot, the data is a lot closer to 0 but there are plenty of values off the line which suggests this data is closer to being normally distributed than GDP, but it is still not normally distributed.

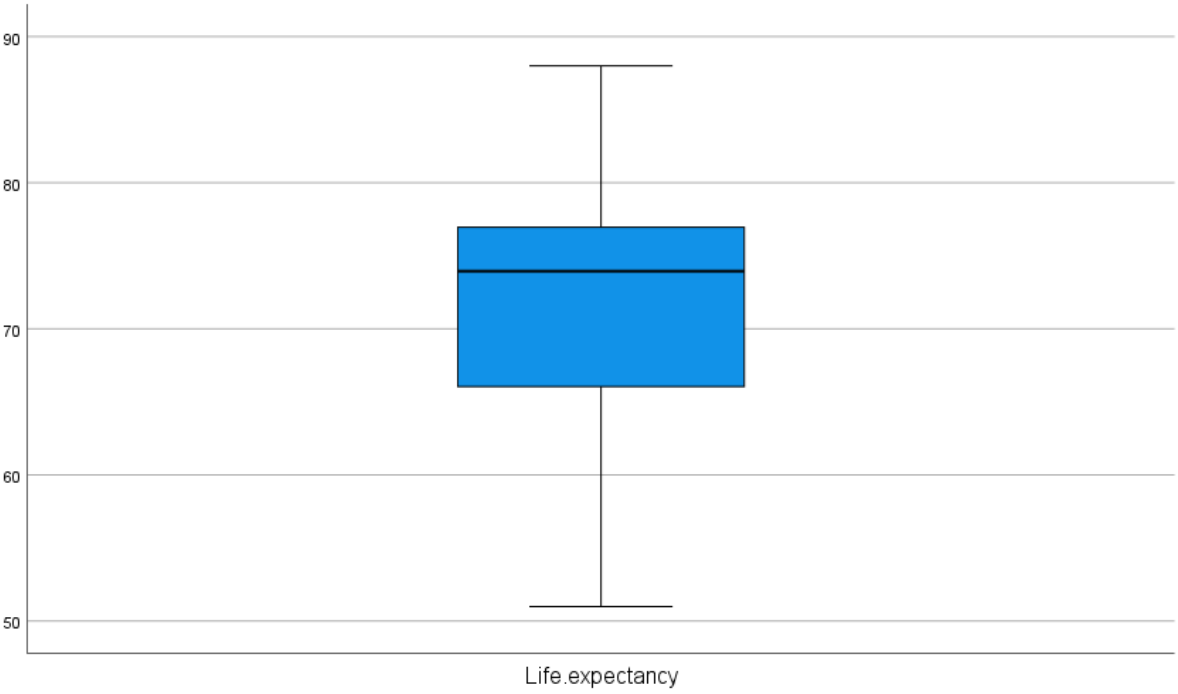


Figure 12 Box and Whisker Plot of Life expectancy

In the box and whisker plot we can see that there are no outliers for the life expectancy variable and median value (the black line through the blue box) is closer to the centre of the graph compared to GDP however it is still moderately skewed from the exact centre.

Conclusions: After completing the normality tests for both the life expectancy and GDP variables I can conclude that the life expectancy data is more normally distributed than the GDP data. However, I must reject the null hypothesis that the data is sampled from a normal distribution for both variables as their P score/Shapiro Wilks score is below 0.05. I must accept the alternative hypothesis that the data is not samples from a normal distribution for both variables.

4.2. Mann- Whitney U Test

The next statistical test I performed was a Mann-Whitney U Test. The purpose of this test is to determine if two samples are likely to derive from the same population. For this test I decided to look at life expectancy and the two variables would be developed and developing countries. In this test I will be determining if on average the status of a country (developed/developing) affects life expectancy. I decided to perform a Mann-Whitney U test as I have proved previously that the data for life expectancy is not normally distributed and for when comparing two variables that come from non-normally distributed samples, the Mann-Whitney U test is recommended. The **Null Hypothesis** for this test is that the status of a country has no effect on life expectancy. The **Alternative Hypothesis** for this test is that the status of a country does affect life expectancy.

The alpha level or **level of significance** will be 0.05. For the **test statistic**, the Mann-Whitney U follows the Z distribution if the sample size is over 20. An important figure to calculate is the Z score. This can be done following a simple two tailed Z table. Since the alpha level is 0.05 this gives me a **Z score** of 1.96. This means if the test statistic is less than -1.96, or greater than 1.96, I can reject the null hypothesis.

	hundredths									
tenths	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	1.00000	0.99202	0.98404	0.97607	0.96809	0.96012	0.95216	0.94419	0.93624	0.92829
0.1	0.92034	0.91241	0.90448	0.89657	0.88866	0.88076	0.87288	0.86501	0.85715	0.84931
0.2	0.84148	0.83367	0.82587	0.81809	0.81033	0.80259	0.79486	0.78716	0.77948	0.77182
0.3	0.76418	0.75656	0.74897	0.74140	0.73386	0.72634	0.71885	0.71138	0.70395	0.69654
0.4	0.68916	0.68181	0.67449	0.66720	0.65994	0.65271	0.64552	0.63836	0.63123	0.62413
0.5	0.61708	0.61005	0.60306	0.59611	0.58920	0.58232	0.57548	0.56868	0.56191	0.55519
0.6	0.54851	0.54186	0.53526	0.52869	0.52217	0.51569	0.50925	0.50286	0.49650	0.49019
0.7	0.48393	0.47770	0.47152	0.46539	0.45930	0.45325	0.44725	0.44130	0.43539	0.42953
0.8	0.42371	0.41794	0.41222	0.40654	0.40091	0.39533	0.38979	0.38430	0.37886	0.37347
0.9	0.36812	0.36282	0.35757	0.35237	0.34722	0.34211	0.33706	0.33205	0.32709	0.32217
1.0	0.31731	0.31250	0.30773	0.30301	0.29834	0.29372	0.28914	0.28462	0.28014	0.27571
1.1	0.27133	0.26700	0.26271	0.25848	0.25429	0.25014	0.24605	0.24200	0.23800	0.23405
1.2	0.23014	0.22628	0.22246	0.21870	0.21498	0.21130	0.20767	0.20408	0.20055	0.19705
1.3	0.19360	0.19020	0.18684	0.18352	0.18025	0.17702	0.17383	0.17069	0.16759	0.16453
1.4	0.16151	0.15854	0.15561	0.15272	0.14987	0.14706	0.14429	0.14156	0.13887	0.13622
1.5	0.13361	0.13104	0.12851	0.12602	0.12356	0.12114	0.11876	0.11642	0.11411	0.11183
1.6	0.10960	0.10740	0.10523	0.10310	0.10101	0.09894	0.09691	0.09492	0.09296	0.09103
1.7	0.08913	0.08727	0.08543	0.08363	0.08186	0.08012	0.07841	0.07673	0.07508	0.07345
1.8	0.07186	0.07030	0.06876	0.06725	0.06577	0.06431	0.06289	0.06148	0.06011	0.05876
1.9	0.05743	0.05613	0.05486	0.05361	0.05238	0.05118	0.05000	0.04884	0.04770	0.04659
2.0	0.04550	0.04443	0.04338	0.04236	0.04135	0.04036	0.03940	0.03845	0.03753	0.03662
2.1	0.03573	0.03486	0.03401	0.03317	0.03235	0.03156	0.03077	0.03001	0.02926	0.02852

Figure 13 two tailed z table

This test was completed through SPSS. The results of the test:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Life.expectancy is the same across categories of Status.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

Figure 14 Mann-Whitney U test summary

In this table we can see what the null hypothesis for this test is and we can also see the decision from SPSS which is to reject the null hypothesis. The P value (significance level) for this test is .000 which is less than .050 which determines that we can reject the null hypothesis.

Independent-Samples Mann-Whitney U Test Summary

Total N	152
Mann-Whitney U	283.500
Wilcoxon W	8033.500
Test Statistic	283.500
Standard Error	210.390
Standardized Test Statistic	-6.904
Asymptotic Sig.(2-sided test)	.000

Figure 15 Mann-Whitney U test summary

This table gives us more statistics generated from the Mann-Whitney U Test, such as the Mann-Whitney U value of 283.5.

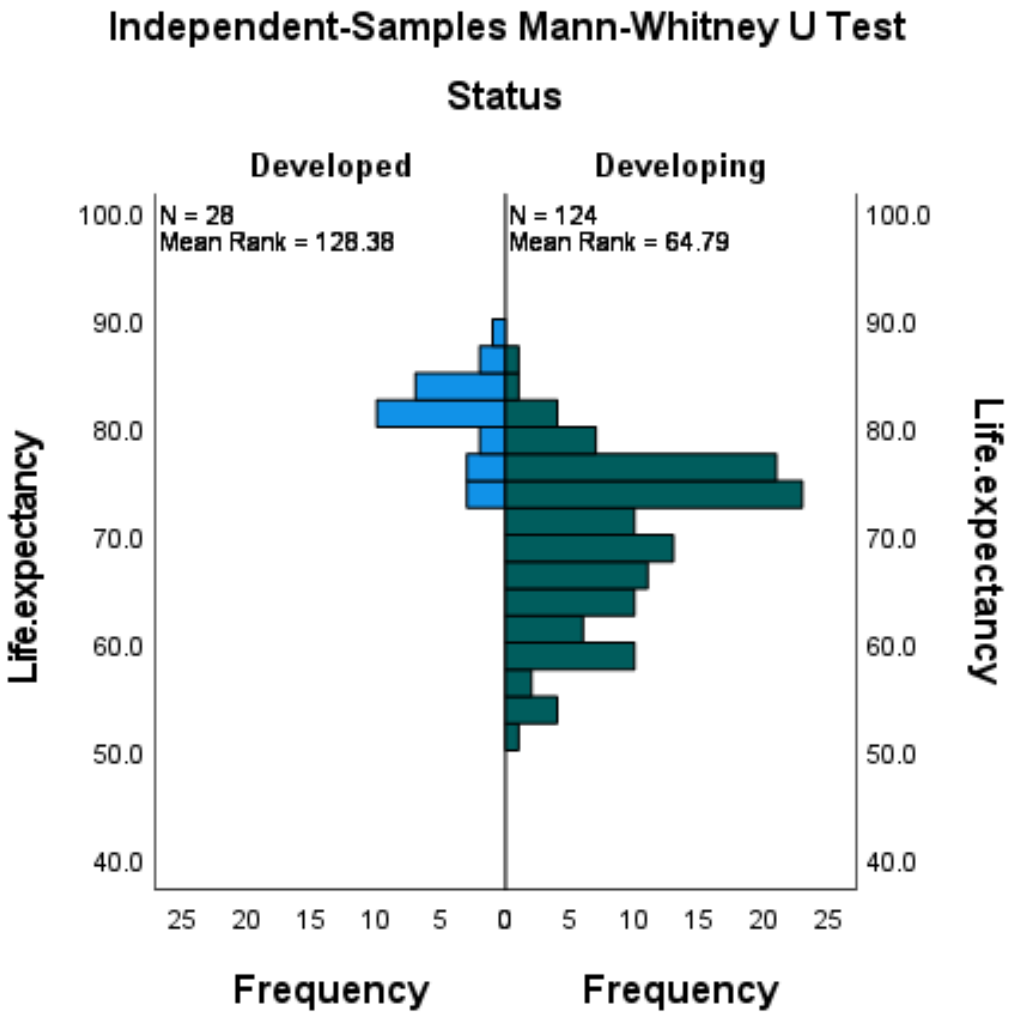


Figure 16 Distribution of Life expectancy for developed and developing countries.

In this visualisation we can see the distribution for both developed and developing countries. As we can see that the mean rank for developed countries is 128.38 which is much greater than the mean rank of the developing countries of 64.79. From this visualisation we can determine that there is a difference between the two variables. In conclusion I can reject the null hypothesis that the status of a country has no effect on life expectancy, and I can accept the null hypothesis that the status of a country does affect a country's life expectancy.

5.0 Testing

As part of my project, I conducted a series of verification and validation tests throughout the different stages of my project. These tests were important as to ensure the accuracy of any data that I used or results that I generated. These tests also tested the functionality of the different technology and algorithms that I used during this project. The objective of the tests is to validate and verify the results of the statistical tests I have performed as well as validating the functionality of the technology that I used such as Shinyapps.io.

Test cases:

Test Case 1			
Name	Shinyapps.io Functionality testing		
Result	Pass	**Date of Test**	1/05/21

Test ID	**1**
Purpose of Test	To Ensure that: It is possible to deploy Shiny Apps on the cloud using Shinyapps.io from my local R studio environment.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, which was connected to a Shinyapps.io cloud account. The data set Life expectancy by WHO was used for creating applications on this cloud environment.
Method	From R Shiny I built an application using the data from my data set. I was able to create an R Shiny application to be deployed by using a ui.R file and a server.R file. I connected to my cloud environment using the function rsconnect in which I specified my account name and token ID so that I could connect to this environment securely.
Expected Result	On completing of the above steps, my application should be deployed to the Shinyapps.io cloud environment where the application will be hosted.

Actual result	The application was successfully deployed to the cloud via Shinyapps.io.
Comments	N/A
Resolution	N/A

Test Case 2			
Name	SPSS Functionality testing		
Result	Pass	**Date of Test**	3/05/21

Test ID	**2**
Purpose of Test	To Ensure that: It is possible to load data sets into SPSS so that statistical tests can be performed on the data.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, which was running IBM SPSS Statistics Data Editor The data set Life expectancy by WHO was loaded into SPSS for the purpose of this test
Method	From SPSS I imported in my data set, by running a statistical test I was able to determine the total number of records in the data set. From this value I was value I was able to determine whether all the data had been successfully imported.
Expected Result	On completing of the above steps, SPSS should import the data set without any missing values.
Actual result	The data set was successfully imported into SPSS Statistics Data Editor without any missing values.
Comments	N/A
Resolution	N/A

Test Case 3			
Name	R Studio Functionality testing		
Result	Pass	**Date of Test**	3/05/21

Test ID	**3**
Purpose of Test	To Ensure that: It is possible to import data sets into RStudio and perform analysis on this data through RStudio using the language R.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, the data set Life expectancy by the WHO was used for this test.
Method	From R Shiny I imported in my data set as a .csv file and saved it as a data frame. From this I was able to run some descriptive statistics on the data set to make sure all the values had been imported correctly and the column and row names of the data set had been unaffected.
Expected Result	On completing of the above steps, my data set should have been imported into R Studio unaffected and should be able to be analysed/processed through the language R.
Actual result	The data set was successfully imported into R Studio with no columns or rows being affected. The data set was able to be analysed through R.
Comments	N/A
Resolution	N/A

Test Case 4			
Name	Missing value identification		
Result	Pass	**Date of Test**	6/03/21

Test ID	**4**
Purpose of Test	To Ensure that: There are no missing values in the main data set
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, which was had my main data set (Life expectancy by the WHO) imported into it.
Method	From R Shiny I was able to analyse my data set for missing values. I used the command: sum(is.na(df$colname)) This command has the purpose of counting all the missing values in a column.
Expected Result	I expect to identify missing values in the data set which I can then remove/replace.
Actual result	The command listed above returned 4 columns which has missing values.
Comments	This was a vital test before running any algorithms on the data set as missing values in the data set prevent these algorithms from running.
Resolution	N/A

Test Case 5			
Name	Running algorithms with replaced missing values		
Result	Pass	**Date of Test**	7/03/21

Test ID	**5**
Purpose of Test	To Ensure that: It is possible to run analysis and algorithms on my data such as Random Forests and clustering. These algorithms do not run if there are missing values in the data.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, which had my main data set (Life expectancy by WHO) imported into its environment.
Method	From R Shiny I had identified the columns in my data set with missing values by calculating the sum of NA values per column in the data set. To tackle this issue, I will be replacing these missing values with the mean values of their columns.
Expected Result	On completing the above steps, my data should have its missing values replaced with the mean values of their columns. After these missing values have been replaced, I should be able to run algorithms such as Clustering on my data set.
Actual result	The missing values were successfully replaced, and I was able to run algorithms such as Clustering on my data set.
Comments	N/A
Resolution	N/A

Test Case 6			
Name	Normalising data prior to clustering		
Result	Pass	**Date of Test**	21/04/21

Test ID	**6**
Purpose of Test	To Ensure that: The data in my data set is normalised/scaled so that I can successfully perform clustering on my data set.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, which had my main data set (Life expectancy by the WHO) imported into it. This dataset had its missing values replaced prior to this analysis.
Method	From R Shiny I imported my data set and used the scale function in R to standardise the data in my data set. This is important as the clustering analysis would be skewed by columns with large values such as GDP.
Expected Result	On completing of the above steps, my clustering algorithm should work normally without any columns skewing the results of this analysis.
Actual result	The data set was successfully normalised, and the clustering analysis worked normally.
Comments	N/A
Resolution	N/A

Test Case 7			
Name	Descriptive statistics results validation		
Result	Pass	**Date of Test**	1/05/21

Test ID	**7**
Purpose of Test	To Ensure that: When performing descriptive statistics on my data set using the software, excel, SPSS and R Studio, the same values such as mean, median and range of values should all be generated in each software.
Test Environment	The test environment is as follows: Client Hardware: HP Laptop 14-ck0xxx running R studio, Excel, and R studio. I imported in my data set (Life expectancy by WHO). I imported the exact same version of the data set in each of the applications.
Method	From R Shiny, Excel and SPSS I imported in my data set. I then performed descriptive statistics on the data set. This generated statistics about the data such as the range of values per column, total records, and mean values of each column. I was able to compare the results from each application.
Expected Result	On completing of the above steps, the three different applications should all return the same descriptive statistics.
Actual result	The applications all produced the same descriptive statistics which validated their results.
Comments	N/A
Resolution	N/A

6.0 Analysis

To analyse my data, I first pre-processed the data so that there were no missing values or inaccuracies in the data sets. I then merged the data sets I had identified so that I could analyse the data from the same source. I explored the data set to find the most relevant data to conduct my analysis on.

Linear regressions

After the data sets had been processed, merged, and structured. I was then able to create some visualisations of the data. The first type of analysis that I performed on the data set was linear regressions. I chose to look at linear regressions as they are a great way to examine the relationship between independent and dependent variables. For my project, the dependent variable was always life expectancy, and the independent variables were the different factors that I was examining such as GDP, schooling and generosity.

Through these plots I was able to create graphs which provided clear insights into what affects life expectancy. I chose to create graphs in my analysis since I felt they were easy to understand and are a great way to see if there is a correlation between two factors. For example, if the scatter plot creates a linear shape stemming from the origin of the graph. This means that the graph has a strong positive correlation and therefore it can be said that there is a correlation between the factors on the x and y axes.

In order to do these linear regressions, I merged my data sets together and plotted different variables against life expectancy. I explored the relationship of life expectancy with both social and political factors. The linear regressions that I performed were:

- Life expectancy vs GDP per capita
- Life expectancy vs Schooling
- Life expectancy vs Happiness levels
- Life expectancy vs Smoking rates
- Life expectancy vs Generosity
- Four variable linear regression

For full results of these regressions see section 7 of the document.

Random Forests

I chose to apply the Random Forests algorithm to my data in order to gain unique insights on my data set and to get a better insight on how each of the variables in my data set affects life expectancy. This is a machine learning algorithm that can perform both classification and prediction tasks from the data.

Random Forests works by creating multiple decision trees. These decision trees have their output combined. Each decision tree will classify data points at each node in the data set and check for information at each node as well. This process is repeated till all the nodes

have been classified and examined. From randomly sampling different data points in the data set Random Forests can eliminate bias from the system.

I performed this algorithm on my data set where I had nineteen variables in total, the explanatory variables were: Country, GDP, Year, Status, Adult Mortality, infant deaths, measles, under five deaths, Polio, Diphtheria, HIV, Population, Income composition of resources, Schooling, Hepatitis B, BMI, thinness age 1-19 and thinness age 5-9. The response variable was Life Expectancy. Through this algorithm I would be able to learn how much each of the explanatory variables affects the response variable.

To perform this, I first loaded in my data set and then replaced the missing values in my data set with the mean values of their column so as to not interfere with the analysis. I filtered the data set so that the values were from 2015. This ensured that the data was as recent as possible as the data in my original data set spans from 2000 to 2015.

I then split my data set into train and validation sets at a ratio of 70:30. I then created a Random Forest model and set the condition to be Life Expectancy. For my first model I set the number of trees to be 500 and the number of variables at each split to be 2. This generated an error rate of 3.64%. For my second model I set the number of trees to also be 500 but set the number of variables at each split to be 6. This generated an error rate of 2.32%. I was now able to predict on both the train data set and the prediction data set. From performing predictions on the data set I was able to view that there was 0 misclassified data points in either data set.

After I built my initial (regression) Random Forest model in R, through the function in R called “importance” I was able to generate two measures for each predictor variable. These two measures are **%IncMSE (Mean Decrease Accuracy)** and **IncNodePurity (Mean Decrease Gini). %IncMSE** is the most robust and informative of the two measures. This measure provides information on the percentage that my model accuracy decreases if I were to leave that variable out. Therefore, the variables with the highest **%IncMSE** score have the most importance.

IncNodePurity or **Mean Decrease Gini** is the measure of variable importance as well. In IncNodePurity the variable importance is based on the Gini impurity index which is used for calculating splits in the decision trees. With IncNodePurity it can also be said that the higher the value, the higher the importance of the variable in my model.

	%IncMSE	IncNodePurity
Country	0.36	38.31
GDP	11.32	618.66
Year	0.00	0.00
Status	6.85	80.023
Adult Mortality	17.4	951.94
Infant Deaths	3.3	60.24

Measles	0.65	48.45
Under five deaths	4.27	112.50
Polio	3.94	86.77
Diphtheria	5.45	68.52
HIV	15.66	1251.99
Population	0.81	40.65
Income composition of resources	20.38	2021.21
Schooling	12.89	1234.62
Hepatitis B	3.66	58.45
BMI	8.31	247.66
Thinness 1-19 years	8.23	228.02
Thinness 5-9 years	11.78	351.59

Figure 17 %IncMSE and IncNodePurity for each variable

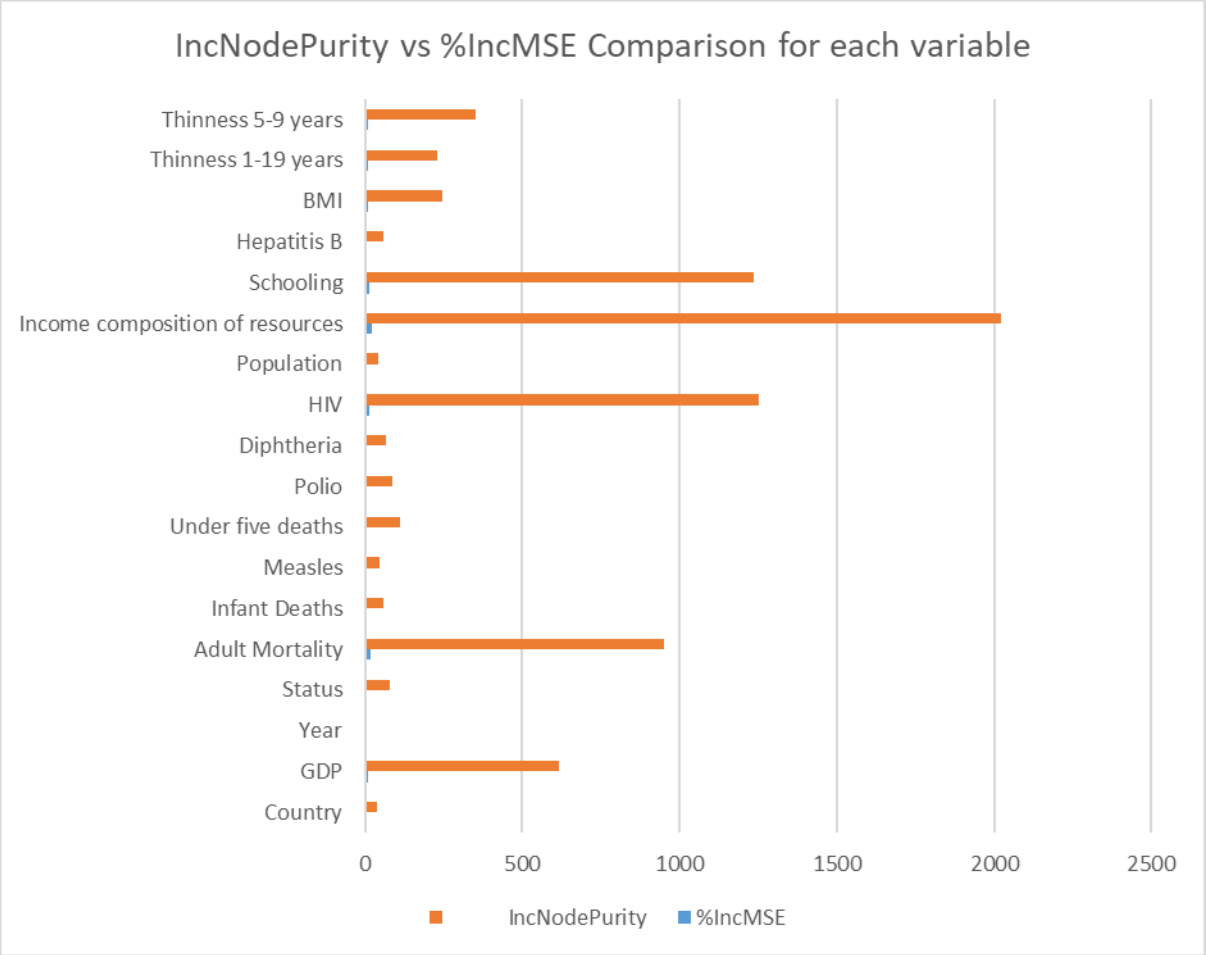


Figure 18 IncNodePurity vs %IncMSE mixed bar plot

As we can see from the table and bar chart above, the explanatory variables have varying levels of importance. Full results of this test are located in Section 7 of the document.

K Means Cluster Analysis

In order to take a broader look at subgroups of information within my data set I performed K Means Cluster Analysis. Through clustering it is possible to find which groups are similar to each other based on their attributes. In this case I will be finding groups of countries which are similar to each other based off the information that I have in my data set such as their populations, GDP and schooling.

To perform K Means Clustering I first filtered out all the columns that had text data as this would interfere with the clustering algorithm. The packages in R that I used for this algorithm are: tidyverse, cluster and factoextra. To prepare my data set for clustering I made sure that the rows were the observations, and the columns were the variables, missing values in a column were replaced by the mean value of their column and finally the data was standardized so that all the variables could be comparable to each other. For example, in my data set the population variable ranges from 1000 all the way to over 1,000,000,000 for populous countries like China and India. I also have variables such as Life Expectancy where all the values range from around 50 to 85 so it was important to scale the data prior to the clustering.

The first stage of clustering was to measure the distance or dissimilarity between each pair of observations (countries) in my data set. The result of this calculation is known as a distance matrix. This step is important for clustering as it allowed me to calculate how similar each element was to each other which defined the size and shape of the clusters. There are many different methods for calculating the distance, the method I chose is Euclidean distance. I chose to use the Euclidean distance method as it is an appropriate method for the size of my data set, and it is commonly used with data sets with different types of numerical values. (Result of distance matrix in results)

To perform K-means clustering on my data set I first specified the number of clusters (K). I decided to set the K-value to 2 as I assumed that countries would fall into two categories naturally based on whether they were developed or developing countries. Through running the algorithm in R, random k objects from the data set were selected to act as the initial cluster centres or mean values. As the centre (centroid) value was calculated, each observation in the data set was assigned to their closest centroid based on the distance between the observation and the centroid value. As new observations were assigned to each cluster the mean (centroid) value was continuously updated. This process is repeated till all the observations are assigned a cluster.

Through the initial clustering step and using the kmeans function I was able to gain information on the clusters that were being formed. This included information such as the number of points assigned to each cluster. In my case the two clusters have sizes of 103 and 49. Through the R command fviz_cluster I was able to create a visualization of the clusters that had formed. I created a range of visualisations for different k values of clusters, I looked at k values from 2-5. These visualisations help provide a clear view of how different countries cluster together. These visualisations also serve as a useful tool for identifying outliers in the dataset.

(See results for visualisations in section 7)

The final analysis I performed with clustering was determining the optimal number of clusters based on the data. I employed the use of three different methods to determine the optimal number of clusters: The Elbow Method and the Silhouette Method. Each of these methods examine the data and determine the optimal number of clusters based on how the similar the data points are to each other.

In each of the methods the k value is varied from 1 to 10 clusters. In the Elbow Method the total within-cluster sum of square (wss) is calculated for each k value. The curve of the total within cluster sum is plotted on a graph against the total number of clusters. In the average silhouette method, the total number of clusters is plotted against the average silhouette value. The average silhouette variable determines how well each object in the data set fits into their cluster. Each method provides a different style of graph to represent this information. Based on the shape of the graphs the optimal number of clusters can be determined (Boehmke, 2021).

(graphs and results in result section)

Growth of Life expectancy, Ireland vs World

As part of my analysis into life expectancy I decided to use my data set to make predictions of life expectancy in the future. In this analysis I decided to compare the life expectancy in Ireland to the life expectancy in the rest of the world to see how Ireland compares. The key parts of this analysis are:

- Comparing the life expectancy in Ireland (2000-2015) to the rest of the world (2000-2015)
- Calculating the rate of change in life expectancy between Ireland and the rest of the world (i.e., calculating how fast life expectancy is growing in Ireland and the rest of the world)
- Analysing the data to predict the life expectancy in Ireland and the rest of the world in the future.

For this analysis I used the R packages readr, gcookbook, ggplot2, tidyverse and Hmisc. The first stage of my analysis was to calculate the average life expectancy of all the 183 countries that were in my data set. I calculated this by calculating the total years of life expectancy per year of all the countries in the data set. I then divided that figure by the total number of countries in the data set to get the average life expectancy per year from 2000 to 2015. With these values I was able to plot this information on a graph with Ireland's life expectancy from 2000 to 2015 so that they could be easily compared. For advanced graphs I saved each of the values I had calculated as a variable and used the variables to create a data frame in R. With the data frame I could make more visualisations of the data and compare Ireland's life expectancy to the rest of the world's life expectancy easily.

The next thing I analysed was the rate of change in the world's life expectancy from year to year and Ireland's life expectancy from year to year. Normally rate of change is calculated in the same way as calculating a slope of a graph, which is the change in the graph's y values over the change in the graph's x values, as seen in the formula below.

$$m = \frac{\text{rise}}{\text{run}} = \frac{y_2 - y_1}{x_2 - x_1}$$

Figure 19 rate of change/slope formula

In the data I had, the change in y and x values were not consistent so I instead had to calculate the average change in y over the average change in x. For this calculation I used the average rate of change formula, as seen below.

$$A(x) = \frac{f(x) - f(a)}{x - a}$$

Figure 20 Average rate of change formula

I calculated the average rate of change for both Ireland and the rest of the world. I calculated the total difference from year to year and then I divided that value by the total number of differences in years I looked at (15).

```
### calculating the sum of the difference in the y values (life expectancy between years)
## this is to calculate the average growth from year to year
TotalIreland = sum((Ireland2001-Ireland2000)+(Ireland2002-Ireland2001)+(Ireland2003-Ireland2002)+(Ireland2004-Ireland2003)+
                     (Ireland2005-Ireland2004)+(Ireland2006-Ireland2005)+(Ireland2007-Ireland2006)+(Ireland2008-Ireland2007)
                   +(Ireland2009-Ireland2008)+(Ireland2010-Ireland2009)+(Ireland2011-Ireland2010)+(Ireland2012-Ireland2011)
                   +(Ireland2013-Ireland2012)+(Ireland2014-Ireland2013)+(Ireland2015-Ireland2014))


##dividing vector by total number of differences  i looked at (15)
AvgrateofchangeIreland = TotalIreland/15
```

Figure 21 Average rate of change in life expectancy for Ireland calculation

```
###same for rest of the world
TotalWorld = sum((AvgWorld2001-AvgWorld2000)+(AvgWorld2002-AvgWorld2001)+(AvgWorld2003-AvgWorld2002)+
                   (AvgWorld2004-AvgWorld2003)+(AvgWorld2005-AvgWorld2004)+(AvgWorld2006-AvgWorld2005)+
                   (AvgWorld2007-AvgWorld2006)+(AvgWorld2008-AvgWorld2007)+(AvgWorld2009-AvgWorld2008)+
                   (AvgWorld2010-AvgWorld2009)+(AvgWorld2011-AvgWorld2010)+(AvgWorld2012-AvgWorld2011)+
                   (AvgWorld2013-AvgWorld2012)+(AvgWorld2014-AvgWorld2013)+(AvgWorld2015-AvgWorld2014))

AvgrateofchangeWorld = TotalWorld/15
```

Figure 22 Average rate of change in life expectancy for the world calculation

From these calculations I was able to calculate the average rate of change (slope) for both Ireland and the rest of the world. From this value alone I was able to determine if Ireland or the rest of the world's life expectancy is growing at a faster rate. (For results of this analysis see section 7).

As I now calculated the average rate of change in years of life expectancy for Ireland and the rest of the world, I could now use this data to make predictions on life expectancy from 2016 to 2030. To make these predictions I continuously added the average rate of change in life expectancy to the current life expectancy to create projections.

Through these calculations I made predictions for life expectancy in Ireland and the rest of the world from 2016-2030 based off the data in my data set. To see how accurate my predictions were, I compared my predicted years of life expectancy to the actual life expectancy in the world and Ireland from 2016 to 2021. This was because the data set that I am using (Life expectancy by the WHO) only goes up to 2015. I was able to compare my predicted values to the actual values in the last five years by using data from the United Nations - World Population Prospects. (World -2021, 2021)

For full results of this analysis please see section 7 of the document.

Descriptive Statistics

As a final analysis on the life expectancy data that I had, I decided to produce some descriptive statistics on this data through SPSS. The descriptive statistics I chose to generate were the mean, median and mode values of the life expectancy values I had. To see how varied my sample was I generated statistics on the standard deviation, variance, range, maximum and minimum values of my sample. To analyse the distribution of the data I generated statistics on skewness and kurtosis levels as well. For full results of this analysis see section 7 of the document.

7.0 Results

Linear Regressions

The linear regressions were created through R using the tidyverse package. The purpose of these regressions is to discover a correlation between Life expectancy and different variables. The first regression I created was Life expectancy plotted against GDP per capita.

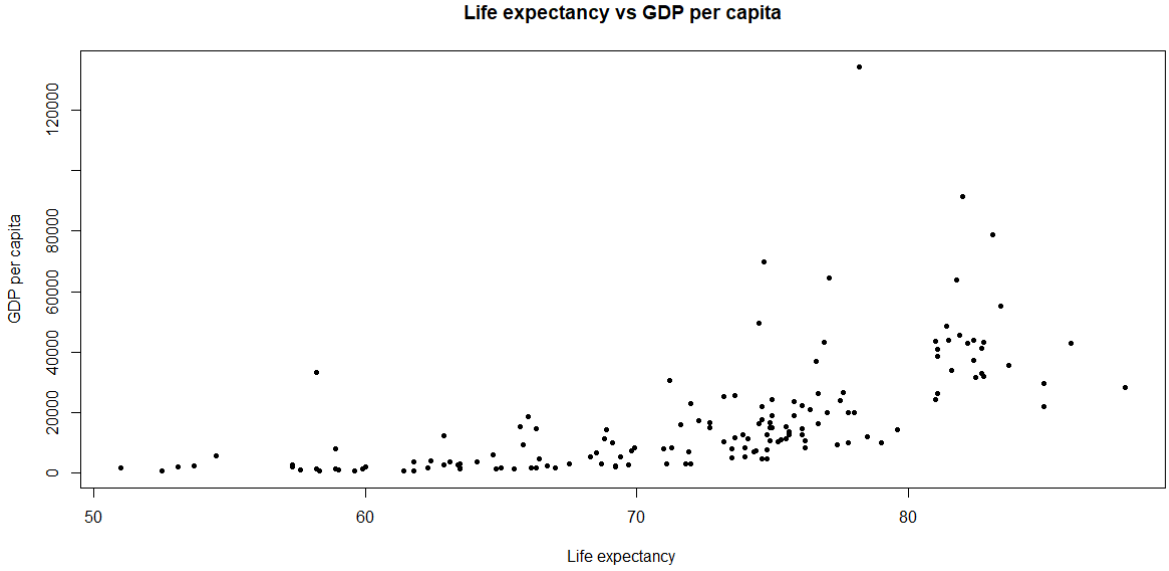


Figure 23 Life expectancy vs GDP linear regression

From this visualisation we can see that there is a clear correlation between these two factors with very few outliers. The next regression I looked at was life expectancy vs smoking levels.

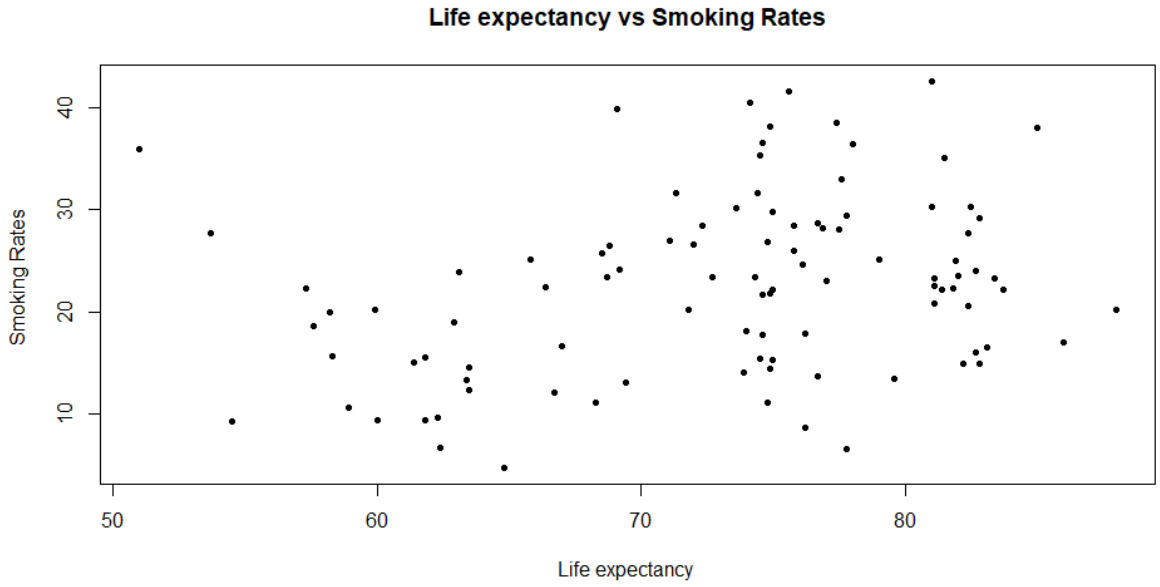


Figure 24 Life expectancy vs Smoking rates Linear regression

As can be seen in this visualisation there seems to be little to no correlation between these factors. I found this to be very surprising, this could be because in countries where more cigarettes are bought, people have more disposable income which could suggest they are from a richer country with better healthcare. I then looked at Life expectancy against Happiness rates.

Figure 25 Life expectancy vs Happiness Score linear regression

The two main data sets that I used for this project were the life expectancy report by the World Health Organisation and the World Happiness Report by the Sustainable Development Solutions Network. As Life expectancy and happiness were the two main subject points of each data set, I decided to plot these two variables against each other after merging the data sets. As can be seen in the graph, there is a strong positive correlation between these two factors. The next regression I performed was life expectancy against schooling. The schooling variable is a value based off how educated each country is on average based on how far each country's citizens goes in their education

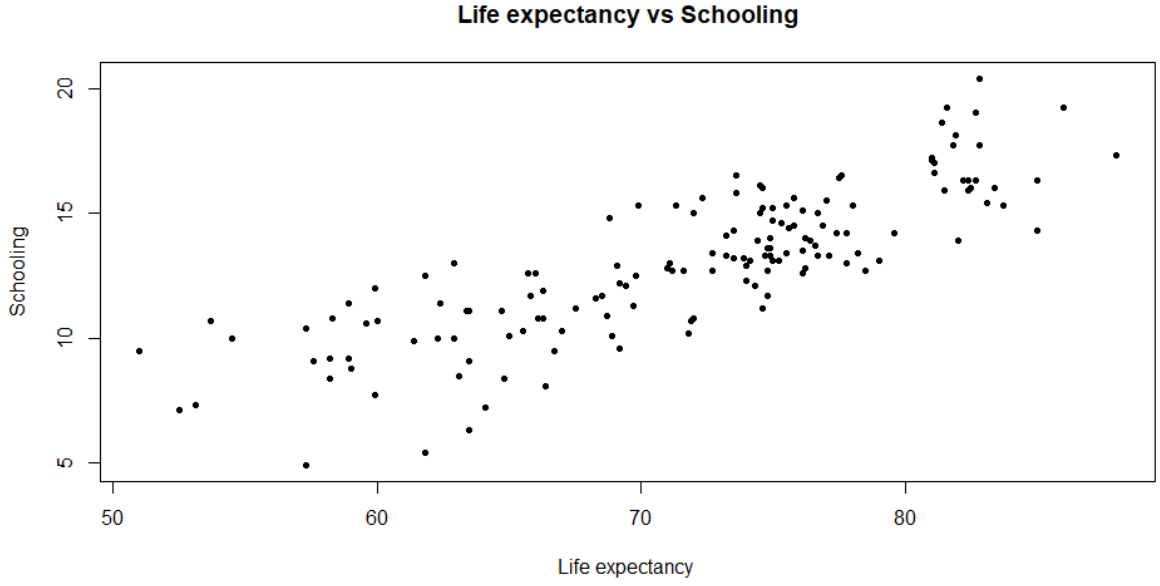


Figure 26 Life expectancy vs Schooling Linear regression

As can be seen in the graph, there is a very strong positive correlation between these two factors. This suggests that countries with better schooling live longer on average. The final linear regression I looked at was Life expectancy against generosity. The generosity value is a score created for the world happiness report, it is a value derived from the average amount of money each country gives to charity per year.

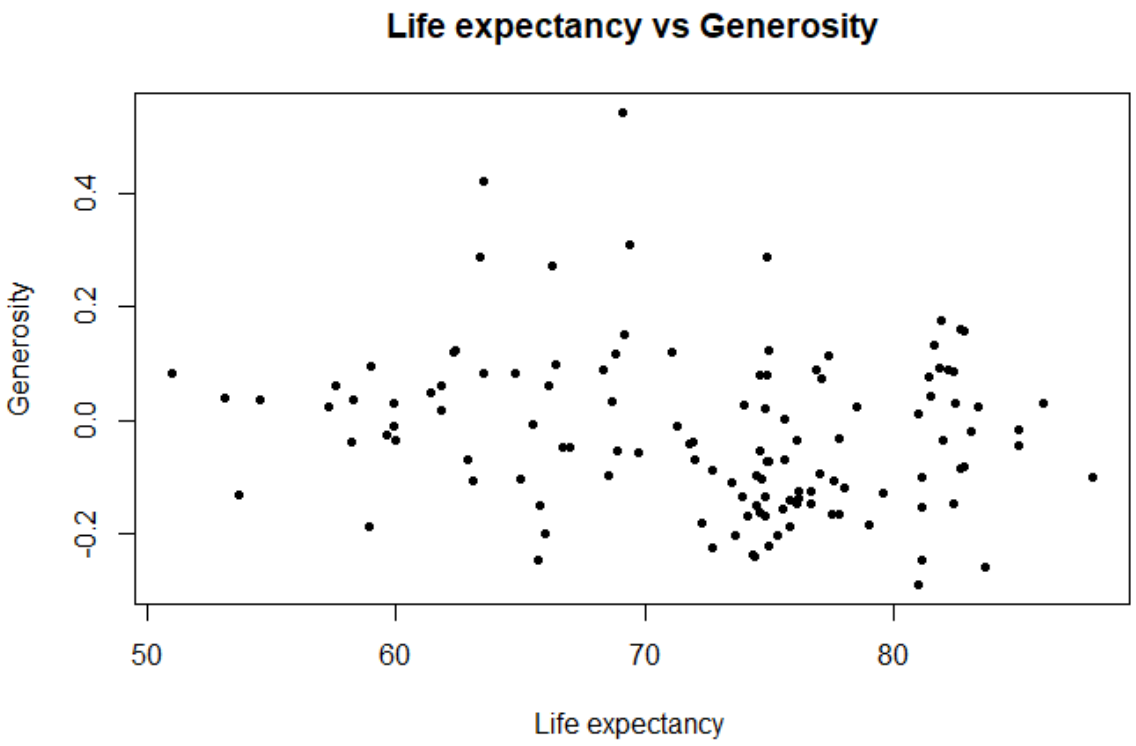


Figure 27 Life expectancy vs Generosity Linear regression

To my surprise there was very little correlation between these two factors. There were seemingly the same levels of generosity regardless of life expectancy. This suggests that being generous or living in a generous country does not affect life expectancy.

The final test I did to compare correlations was comparing the variables, life expectancy, GDP per capita, Schooling and HIV/AIDS rates. I created a 4x4 visualisation which showed how correlated each of the columns were to each other.

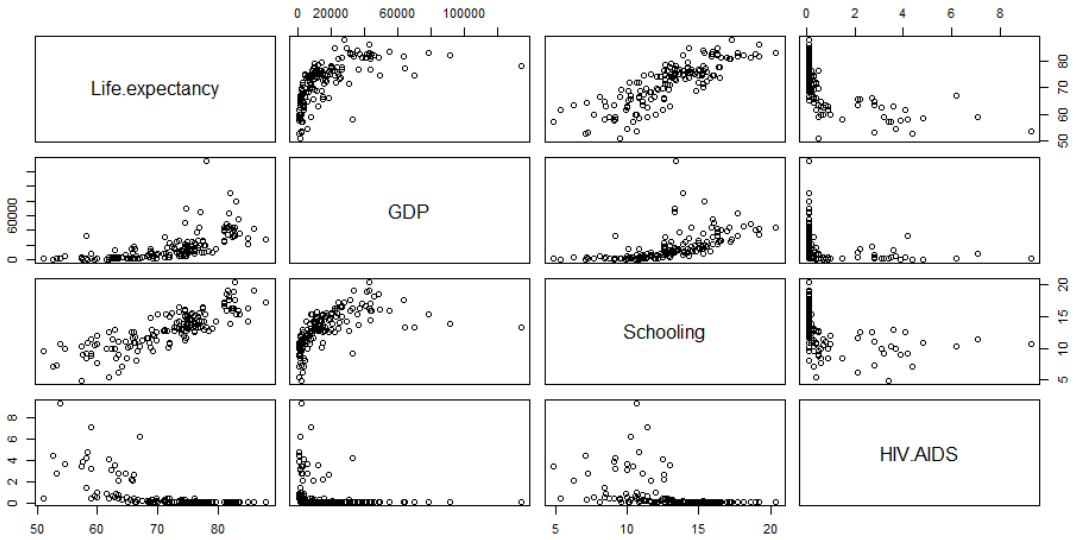


Figure 28 four variable linear regression

The main correlations that can be discovered from this graph are Life expectancy and GDP, Life expectancy and Schooling and Schooling and GDP. There seems to be little to no other correlation between the other variables.

Random Forests

As previously mentioned in the Analysis section (Section 6), I performed a (regression) random forest model through R to calculate the %IncMSE and IncNodePurity scores for each variable. The results are as follows:

	%IncMSE	IncNodePurity
Country	0.36	38.31
GDP	11.32	618.66
Year	0.00	0.00
Status	6.85	80.023
Adult Mortality	17.4	951.94
Infant Deaths	3.3	60.24
Measles	0.65	48.45
Under five deaths	4.27	112.50
Polio	3.94	86.77
Diphtheria	5.45	68.52
HIV	15.66	1251.99
Population	0.81	40.65
Income composition of resources	20.38	2021.21
Schooling	12.89	1234.62
Hepatitis B	3.66	58.45

BMI	8.31	247.66
Thinness 1-19 years	8.23	228.02
Thinness 5-9 years	11.78	351.59

Figure 29 %IncMSE vs IncNodePurity for each variable table

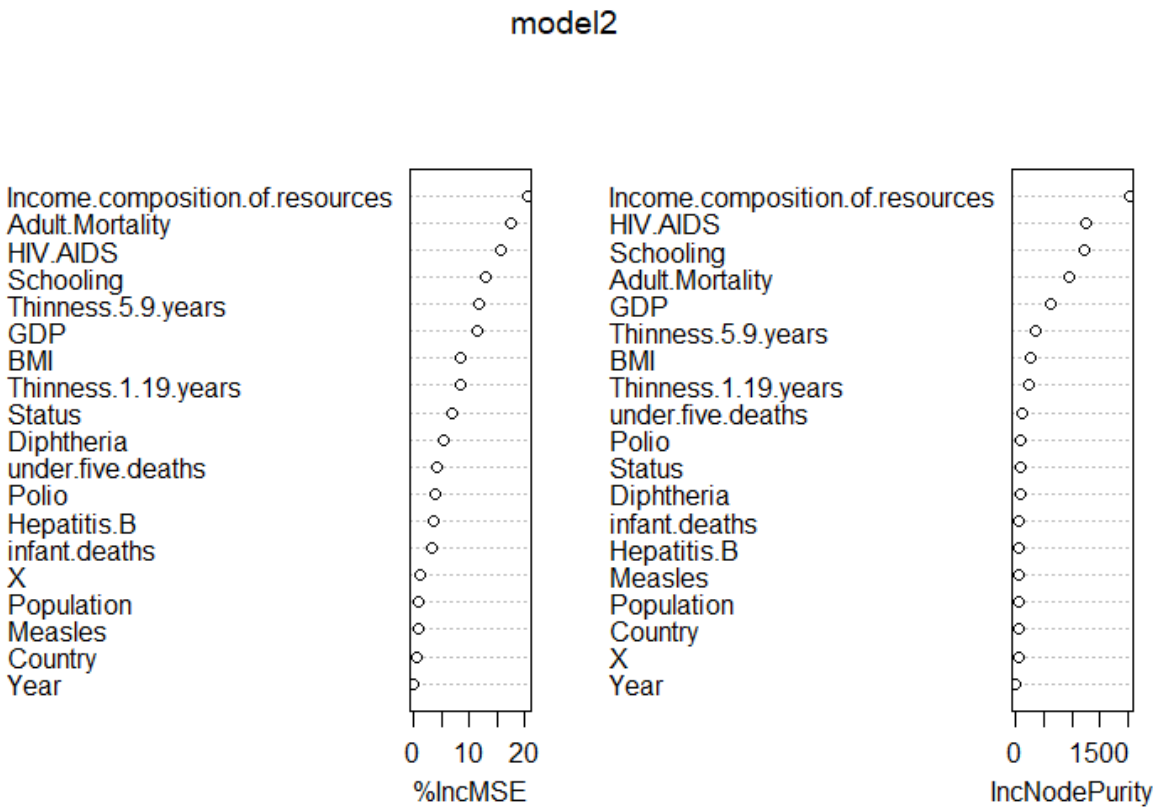


Figure 30 Visualisation of the importance of each variable

From this table and the visualisation that I created in R we can see which variables are the most important with regards to my response variable (life expectancy). Through the visualisation we can see the importance of each of the variables compared to each other. From the table and the visualisation, we can see that the most important variables are: Income Composition of Resources, Adult Mortality and HIV/AIDs levels. The least important variables are Measles, Population, and Infant deaths. For this analysis I set the year to be 2015 so that is why the year is also unimportant. The country name is unique to each variable and the "X" value is for the ID numbers of each row in the data set which explains why they lack importance according to this test.

K Means Cluster Analysis Results

As previously mentioned in the analysis section of the report, the first section of clustering is to calculate the distance or dissimilarity between each pair of observations in the data set. In this case I will be looking at the distance between each pair of countries in the data set. From using the Euclidean distance algorithm, I was able to calculate the distance between all pairs of countries in my data set.

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Figure 31 Euclidean distance algorithm

Clustering Distance Measures Results

From performing this algorithm, I was able to see how similar each of the countries were to each other. According to this test, the two countries that had the furthest distance from each other/least similarity were Pakistan and Austria. I noticed that countries from the same region of the world were more similar to each other. From this test, the two regions that had the least similarity according to the algorithm were central Europe and the middle east. To represent this information, I created a distance matrix through R.

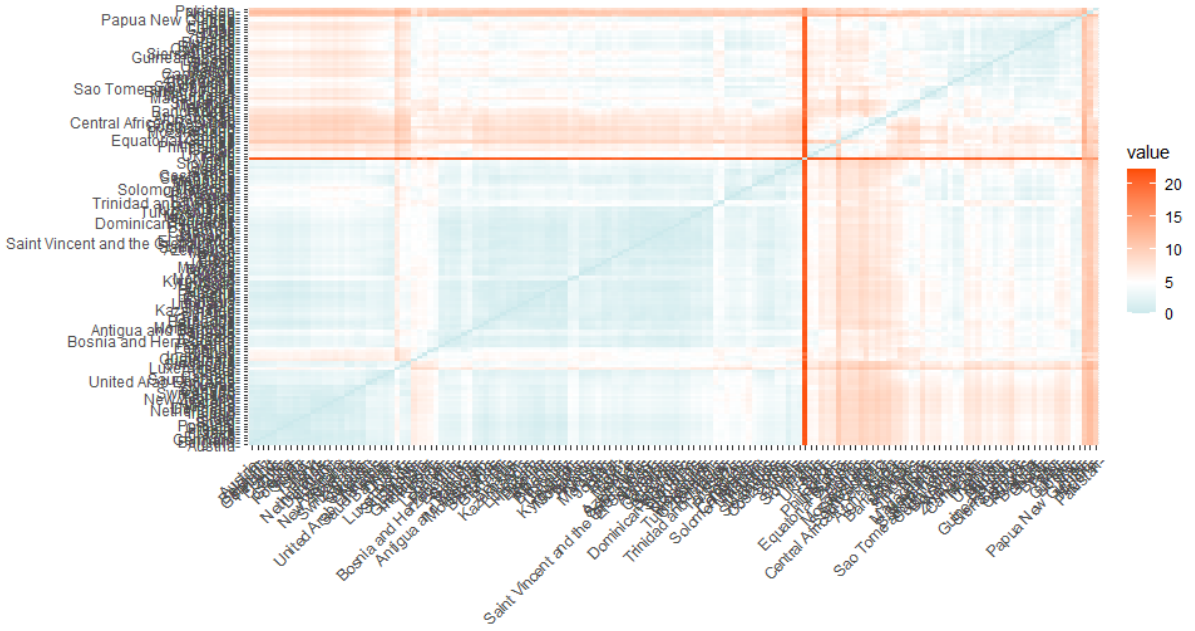


Figure 32 Distance Matrix of all countries in my data set

Due to the number of countries that are in the data set, it is hard to analyse most of the countries on this graph as their names are obscured. To tackle this problem, I filtered down the data set and produced another distance matrix for developed countries alone.

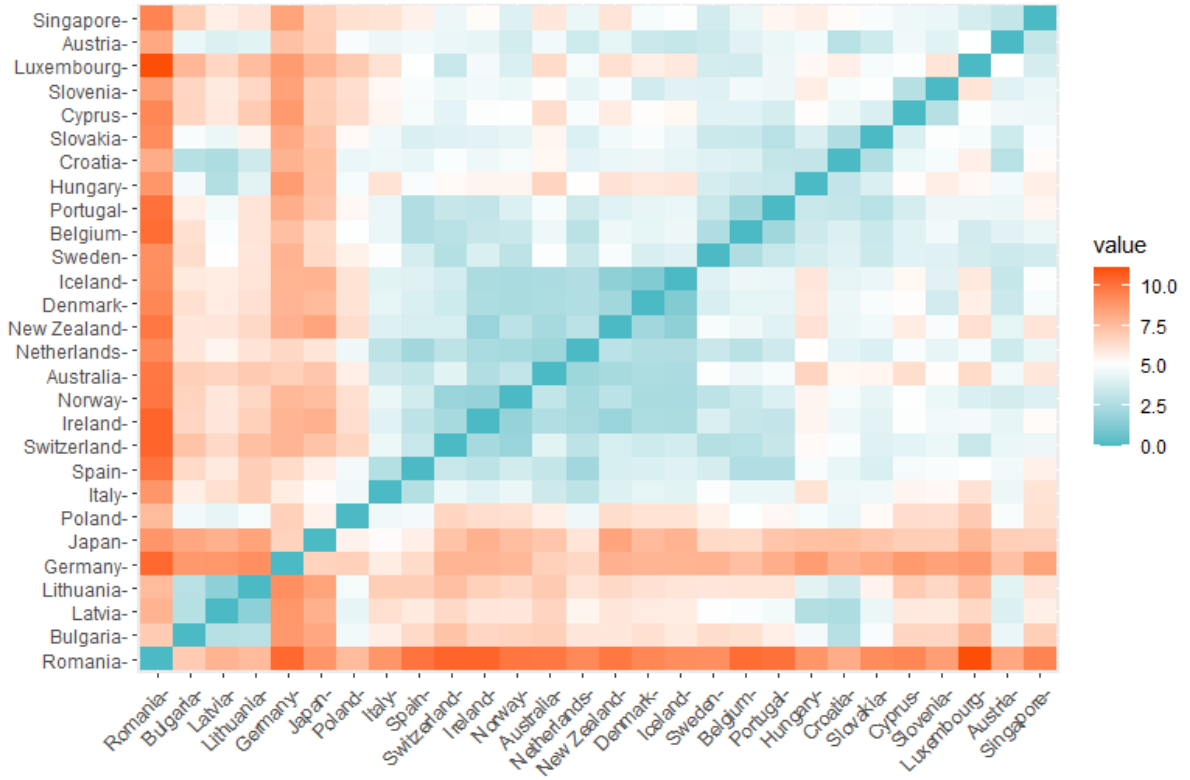


Figure 33 Distance Matrix of all developed countries

Here we can see that Romania and Singapore have the greatest distance from each other. It can also be seen that countries from similar regions e.g., countries from Eastern Europe have a smaller distance from each other.

To begin the main process of clustering I partitioned my data into clusters by using the method K-means clustering. In this method the K value has to be specified before the clustering takes place. This value represents the number of groups/clusters the objects will be separated into. For my first series of clustering, I decided to set the K value to 2 as I wanted to test my theory that the main clusters will be made up for developed and developing countries.

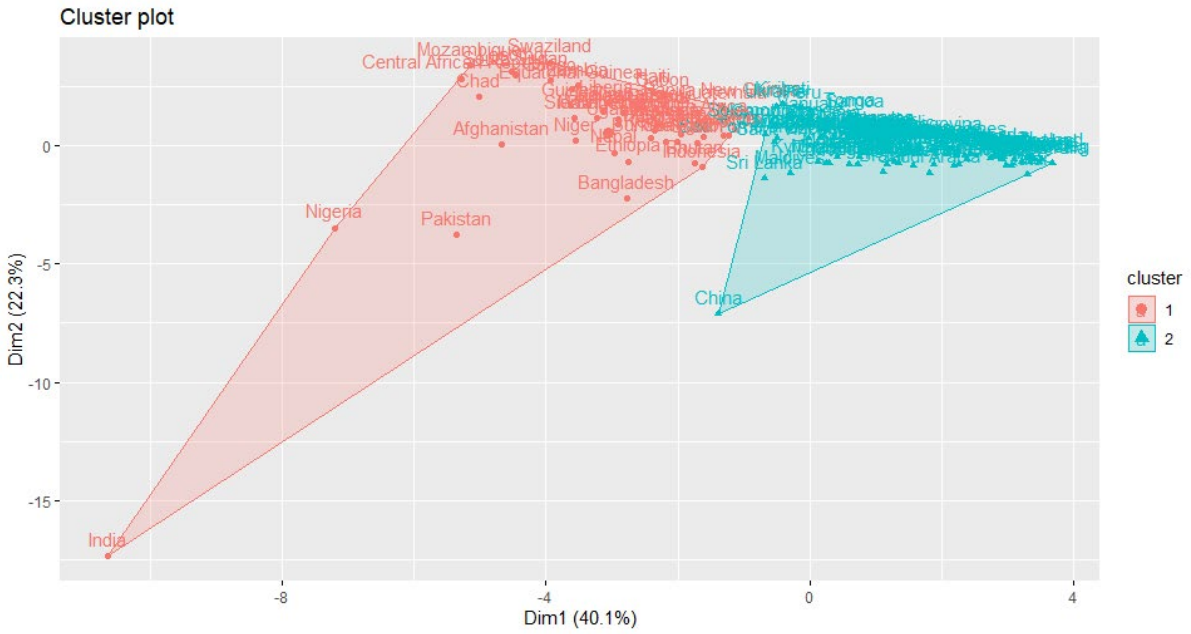


Figure 34 Cluster Plot of all countries in the data set K score = 2

From this visualisation we can see the clusters that were formed. This visualisation serves as a great tool for identifying outliers in the data. In this case the outliers are China and India. For this cluster, the number of centres is 2 and the number of initial configurations was set to 25. From this initial clustering, useful information was generated about these clusters.

totss(the total sum of squares)	2416
Withinss (vector of within-cluster sum of squares)	[1:2] 1047 666
tot.withinss (total within-cluster sum of squares)	1713
Betweenss (the between-cluster sum of squares)	703
Size (the total number of points in each cluster)	49 and 103

Figure 35 Cluster information for K score of 2

I ran the clustering algorithm again, this time I set the k-value to 4 to see if any other observations could be made. As illustrated in figure 36:

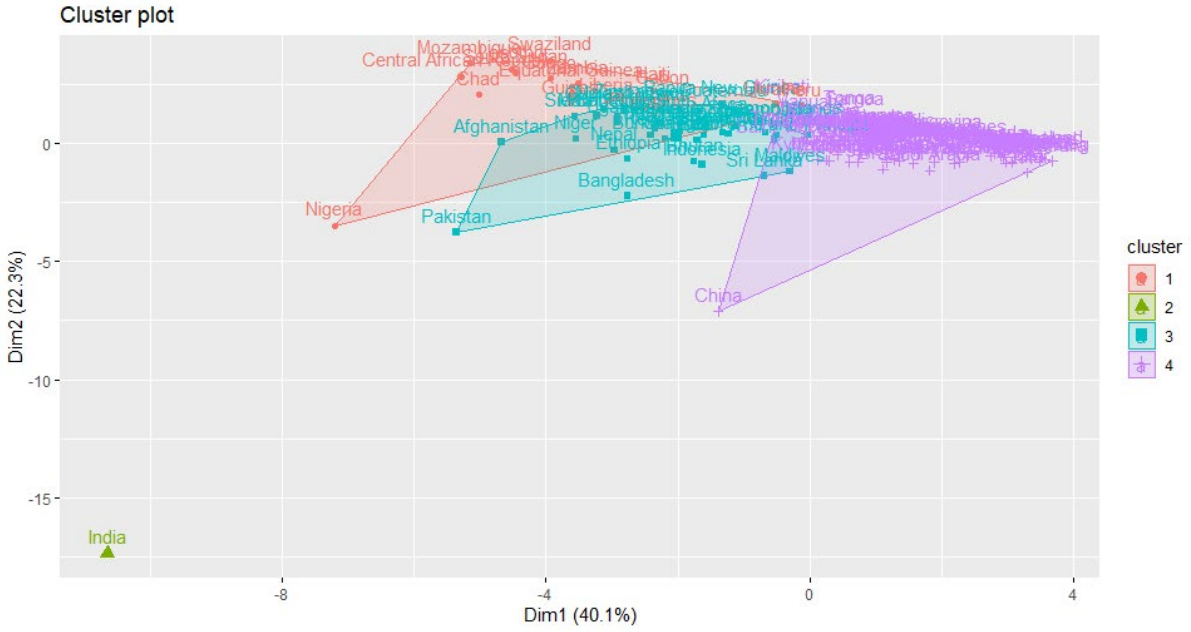


Figure 36 Cluster Plot of all countries in the data set K score = 4

From this visualisation it can be seen that the different regions of the world seem to cluster together. Through this visualisation outliers can be seen, and some countries can be identified. For a closer look at which countries are in each cluster a cluster vector was created with information on which cluster was assigned to each country.

Cluster 1	Central African Republic, Chad, Congo, Equatorial Guinea, Gabon, Guinea, Haiti, Lesotho, Liberia, Mozambique, Nigeria, Peru, Philippines, South Sudan, Swaziland, Ukraine, Zambia
Cluster 2	India
Cluster 3	Afghanistan, Bangladesh, Belize, Benin, Bhutan, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Comoros, Djibouti, Ethiopia, Ghana, Guatemala, Guinea-Bissau, Indonesia, Iraq, Kenya, Madagascar, Malawi, Maldives, Mali, Mauritania, Namibia, Nepal, Niger, Pakistan, Papua New Guinea, Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, Solomon Islands, South Africa, Sri Lanka, Sudan, Tajikistan, Togo, Uganda, Zimbabwe
Cluster 4	Albania, Algeria, Antigua and Barbuda, Australia, Armenia, Austria, Azerbaijan, Bahamas, Bahrain, Barbados, Belarus, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Canada, Chile, China, Colombia,

	Costa Rica, Croatia, Cyprus, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Estonia, Fiji, Finland, France, Georgia, Germany, Greece, Grenada, Honduras, Hungary, Iceland, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kiribati, Kuwait, Kyrgyzstan, Latvia, Lebanon, Libya, Lithuania, Luxembourg, Malaysia, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Netherlands, New Zealand, Nicaragua, Norway, Oman, Panama, Paraguay, Poland, Portugal, Qatar, Romania, Saint Lucia, Saint Vincent and the Grenadines, Samoa, Saudi Arabia, Serbia, Seychelles, Singapore, Slovakia, Slovenia, Spain, Suriname, Sweden, Switzerland, Thailand, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, United Arab Emirates, Uruguay, Uzbekistan, Vanuatu

Figure 37 Categorization of each country, k score = 4

From this table of results, it can be clearly seen that countries from similar regions fall into the same cluster group. Cluster 1 mainly consists of central African countries; Cluster 2 consists of India alone which shows how dissimilar India is to the rest of the countries in the data set. Cluster 3 mainly consists of northern African/middle eastern countries and cluster 4 has countries from the widest range of regions however all 28 of the countries in my data set that were classed as “developed” appeared in this cluster, so it could be said that this cluster has the majority of the wealthier countries in the data set.

To take a further look at how the different clusters of data compare to each other, I created a pairwise scatter plot of the different clusters. I compared the clusters to the variables: GDP and Life Expectancy.

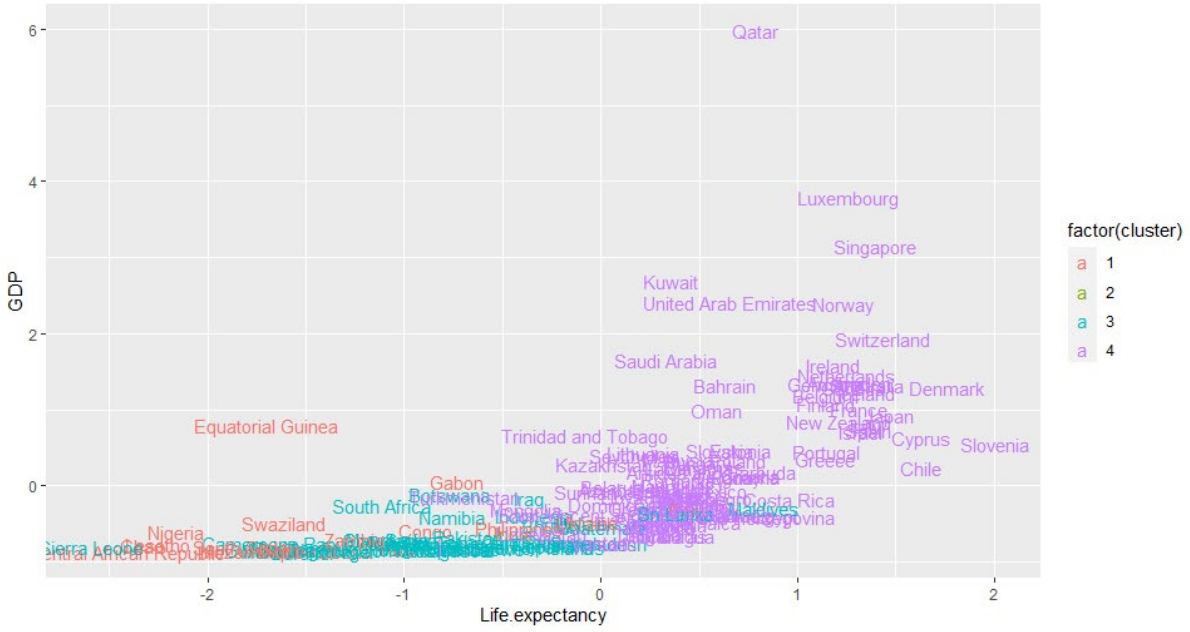


Figure 38 Life expectancy and GDP plot with countries categorised by cluster.

From this graph it can be seen that countries in cluster 4 have the highest GDP and life expectancy on average. Cluster 1 has on average the lowest GDP and life expectancy. From this visualisation alone we can see that Life Expectancy correlates to GDP and we can see that the clusters of countries appear close to each other which proves the clustering algorithm works.

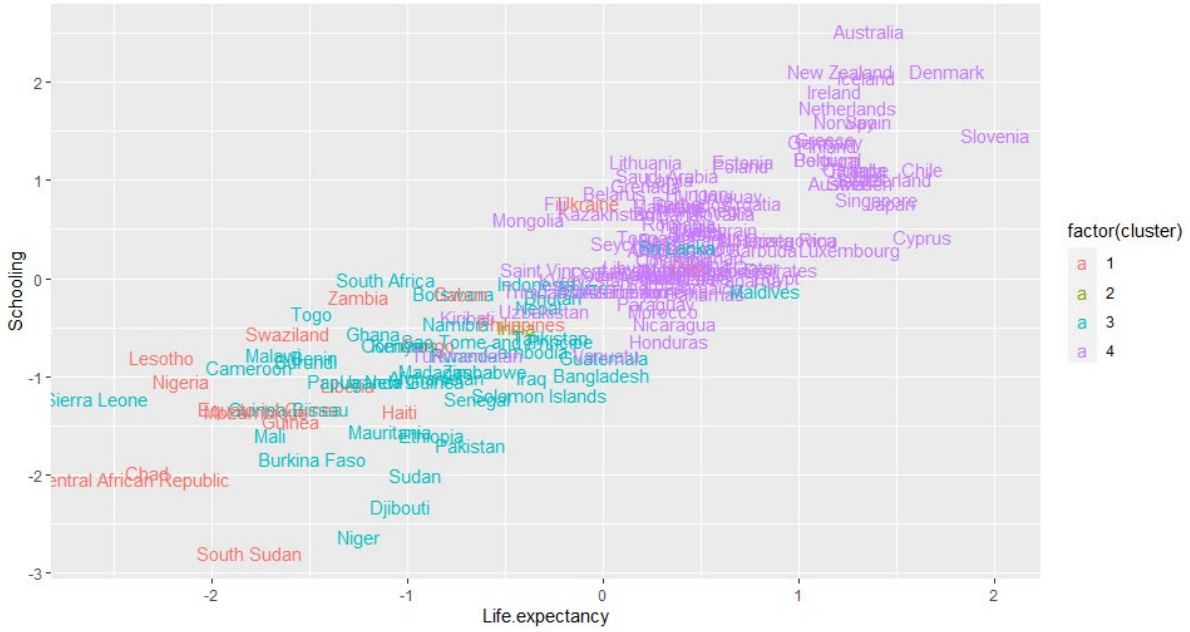


Figure 39 Life expectancy vs Schooling Plot with countries in clusters

To further prove this, I looked at the variables, Life Expectancy and Schooling. In this visualisation as well, the two variables have a correlation between each other, and the countries are showing up on the plot in their clusters.

Since the number of clusters in k-means clustering must be specified before the algorithm, I created a visualisation of the data with 2, 3, 4 and 5 different clusters.

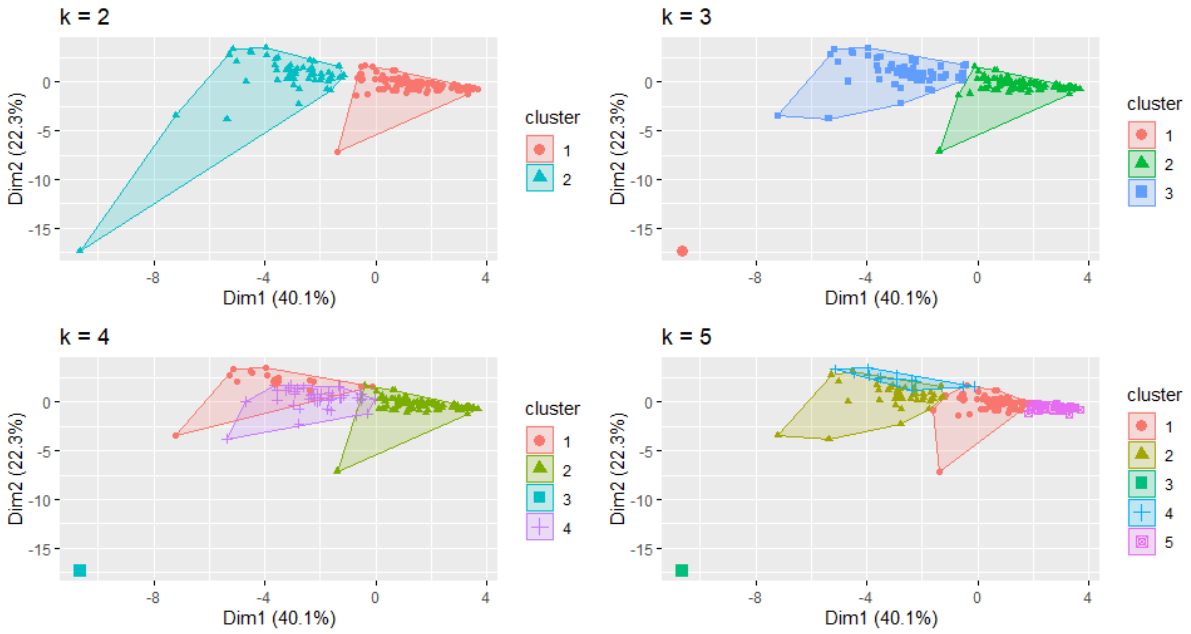


Figure 40 Clusters with K score 2-5

This visualisation provides more information on how the information is clustered together however it does not provide any information on what the optimal number of clusters for the data set is. As previously mentioned in the analysis section of the document, I performed two separate methods to determine the optimal number of clusters.

Elbow Method

In the Elbow Method, several cluster algorithms are performed with varying values of k, for each k value the wss value is calculated. These two values are plotted against each other. In the Elbow Method, the optimal number of clusters is the value that appears at the bend (or elbow) in the graph. In this case the optimal number of clusters is suggested to be from 2-4.

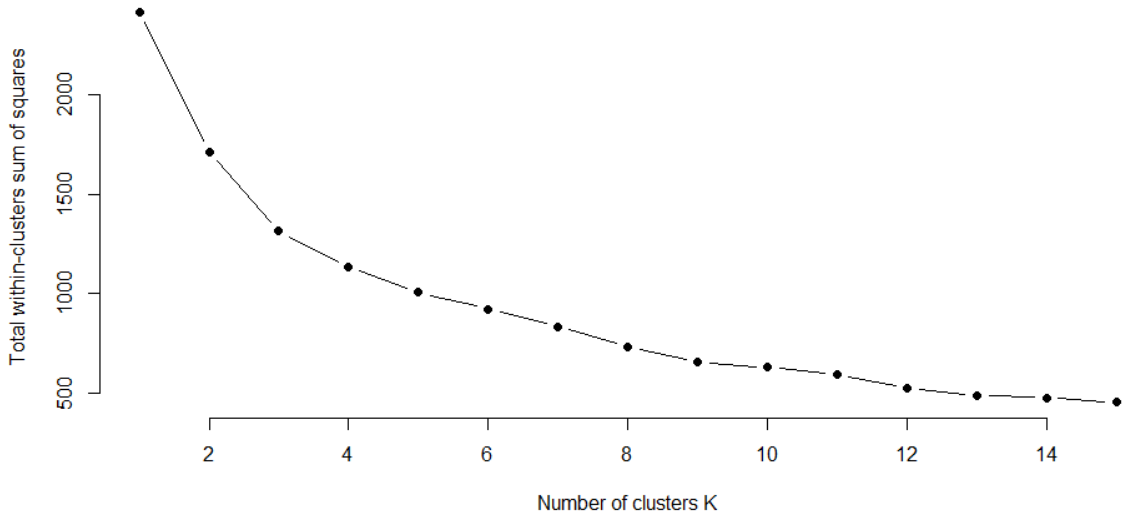


Figure 41 Elbow Method Optimal number of clusters plot

To better visualise the bend or elbow in the graph I used the function in R: fviz_nbclust.

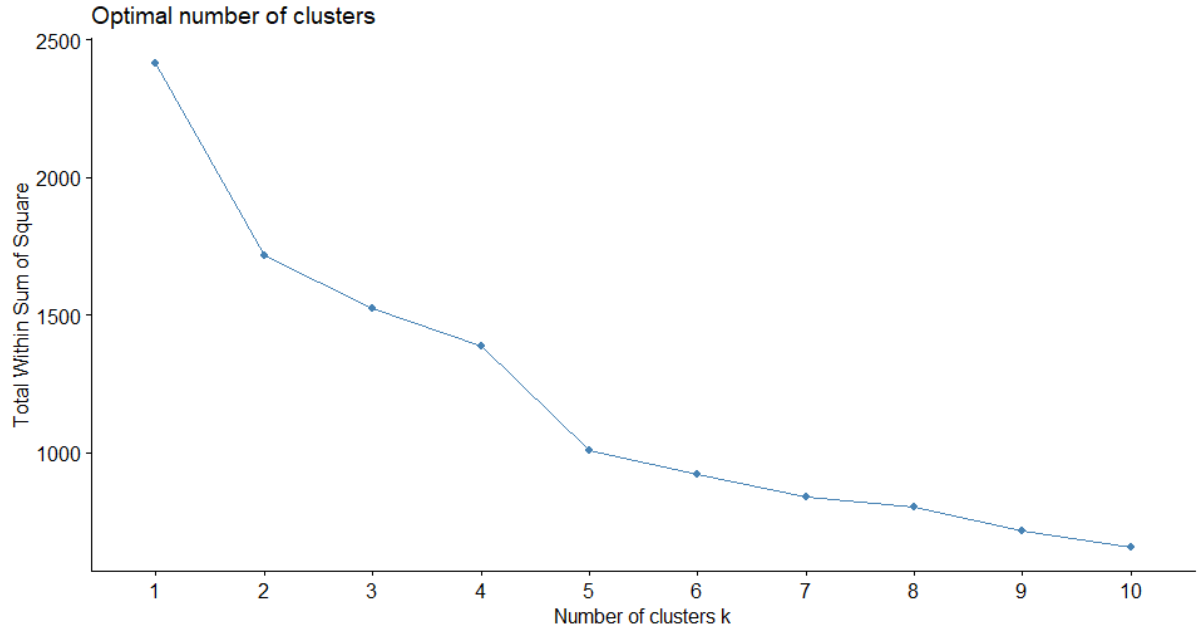


Figure 42 Elbow method optimal number of clusters plot 2

From these visualisations we can see that there are two bends in the graph, one at 2 clusters and another at 4. This suggests that the optimal number of clusters for this data set is between 2 and 4.

Average Silhouette Method

In the Average Silhouette Method, the clustering algorithm is performed for a series of k values like in the Elbow Method. This time the number of clusters is plotted against the

average silhouettes. The average silhouette value determines how well a value fits in the cluster it was assigned to.

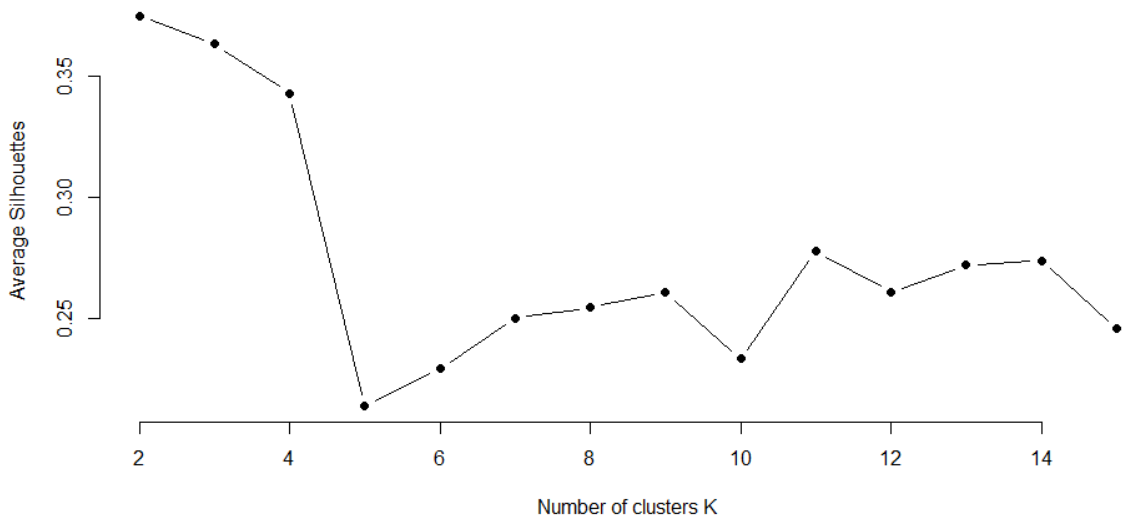


Figure 43 Average Silhouettes Optimal Number of Clusters Plot

From this visualisation we can again see that the optimal number of clusters appears to be from 2 – 4. To better determine this, I used the function: fviz_nbclust.

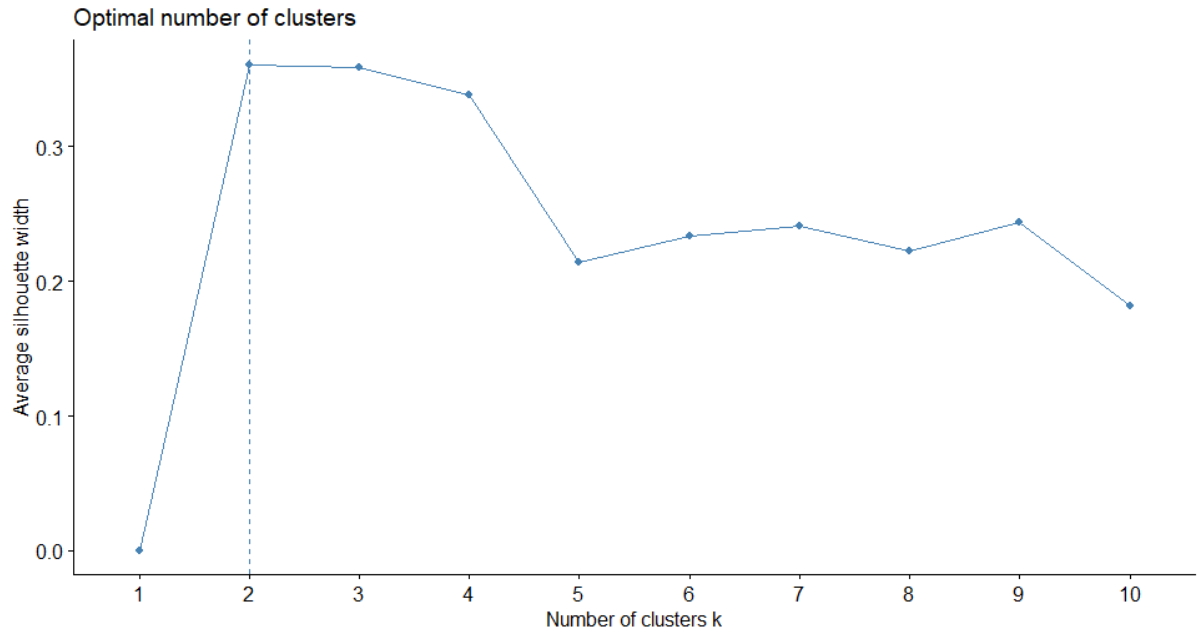


Figure 44 Average Silhouettes Optimal Number of Clusters Plot 2

This function clearly shows that the optimal number of clusters according to the average silhouette method is 2.

In conclusion the optimal number of clusters in the data set was from 2-4. Through analysing the countries that were assigned to each cluster it could be determined that

countries from similar regions geographically cluster together which suggests they share similar social and economic values.

Growth of Life expectancy, Ireland vs World

As previously mentioned in section 6 of the document, as part of this analysis I looked at comparing the life expectancy of Ireland (2000-2015) to the rest of the world (2000-2015). I calculated the rate of change in years of life expectancy for Ireland and the rest of the world. Finally, I used the data to make predictions on life expectancy from 2016-2030.

Ireland vs the World

As previously mentioned in section 6 of the report, I calculated the average life expectancy in Ireland and in the World from 2000-2015 in order to compare their life expectancies over this sixteen-year period. To do this I created a mixed line graph:

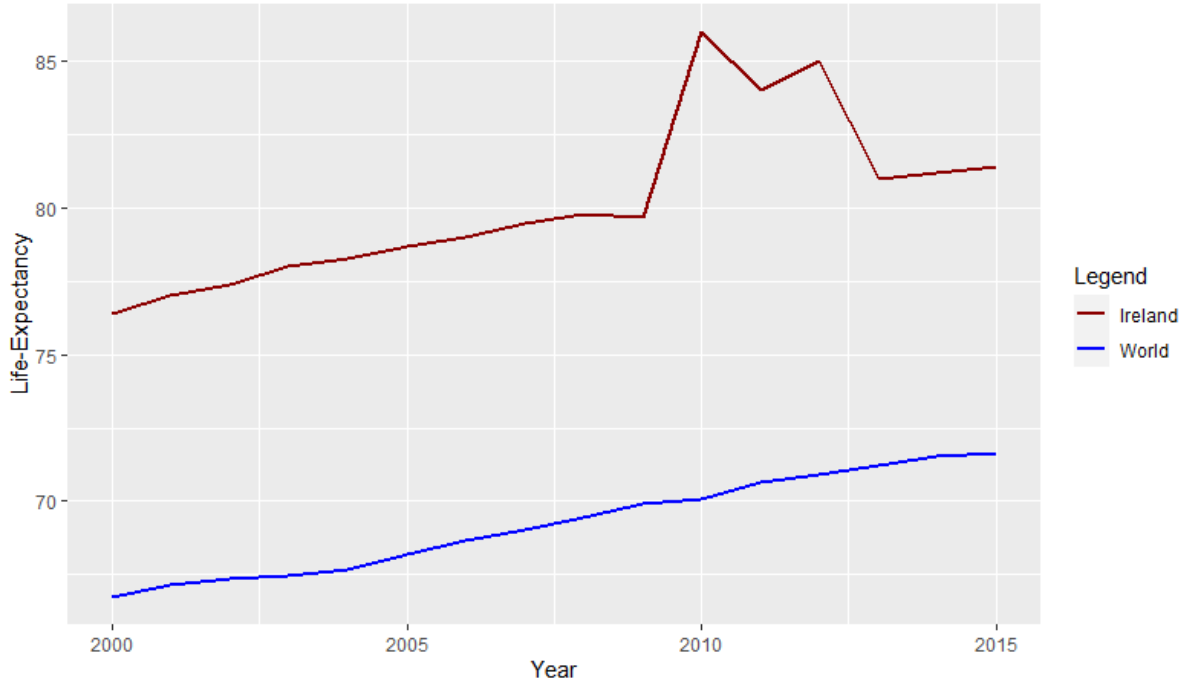


Figure 45 Ireland vs Rest of the world Life expectancy (2000-2015)

From this graph we can see that Ireland's life expectancy is higher than the rest of the world. Through R I calculated Ireland has an average of 10.92507 more years of life expectancy than the rest of the world per year.

Rate of change

As mentioned in the analysis section I calculated the average increase in life expectancy per year for both Ireland and the rest of the world. The results of this test are as follows:

Region	Rate of change (years of life expectancy)
Ireland	0.3333333
Rest of the world	0.3244444

Figure 46 Rate of change (years of life expectancy) Ireland vs the rest of the world

The results of this test are that Ireland's life expectancy is growing at a marginally faster rate than the rest of the world at 0.333 years of life expectancy per year compared to the rest of the world's 0.324 years of life expectancy per year. For a closer look at how the rate of change (slope) of Ireland and the rest of the world compares, I created line plots of their life expectancy from 2000 to 2015 and added a best fit line which had the same slope as the rate of change value that I had calculated.

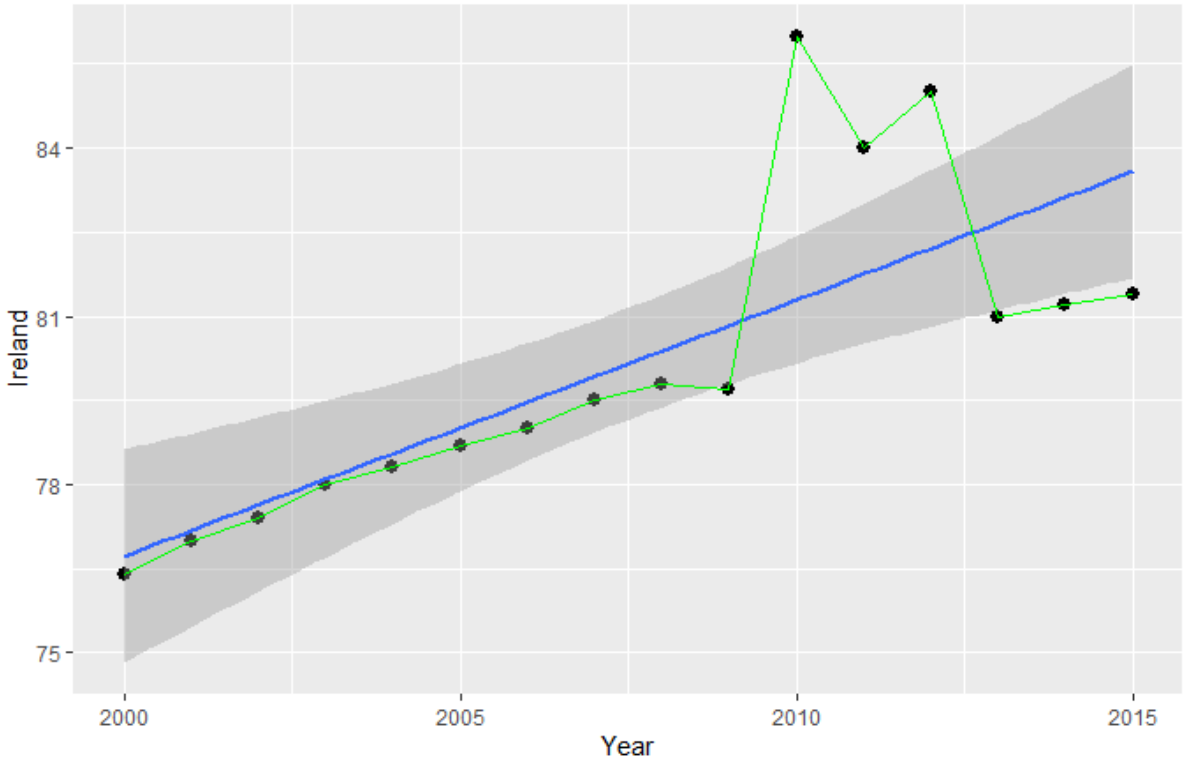


Figure 47 Ireland Life expectancy with best fit line (2000-2015)

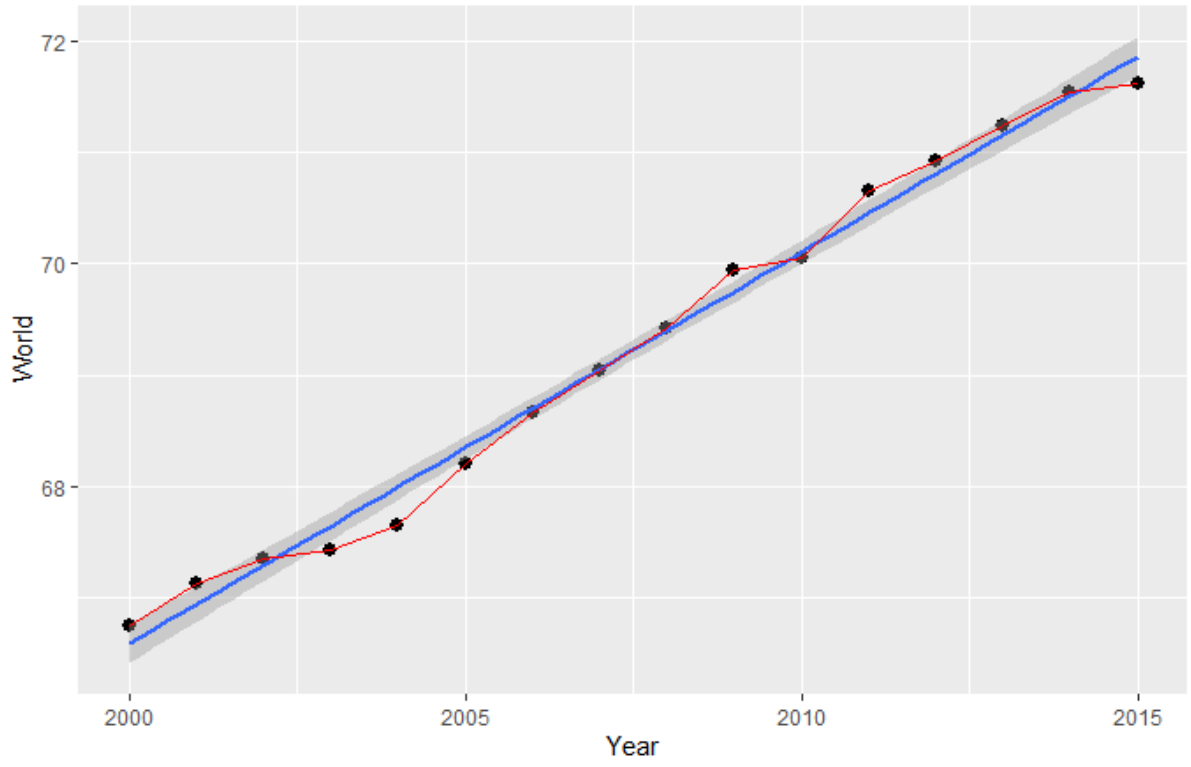


Figure 48 World Life expectancy with best fit line (2000 - 2015)

Future Life expectancy predictions

As I had the rate of change for both Ireland and the rest of the world, I was able to calculate some projections of life expectancy from 2016-2030. The results of calculation are as follows:

Year	Ireland (life expectancy)	World (life expectancy)
2000	76.4	66.75027
2001	77	67.12896
2002	77.4	67.35137
2003	78	67.43333
2004	78.3	67.64645
2005	78.7	68.20929
2006	79	68.66776
2007	79.5	69.03607
2008	79.8	69.42787
2009	79.7	69.93825
2010	86	70.04863
2011	84	70.6541
2012	85	70.91694
2013	81	71.23607
2014	81.2	71.53661
2015	81.4	71.61694

2016	81.73333	71.94138
2017	82.06667	72.26583
2018	82.4	72.59027
2019	82.73333	72.91472
2020	83.06667	73.23916
2021	83.4	73.56361
2022	83.73333	73.88805
2023	84.06667	74.2125
2024	84.4	74.53694
2025	84.73333	74.86138
2026	85.06667	75.18583
2027	85.4	75.51027
2028	85.73333	75.83472
2029	86.06667	76.15916
2030	86.4	76.48361

Figure 49 Prediction table of Life expectancy for Ireland and the world (2000-2030)

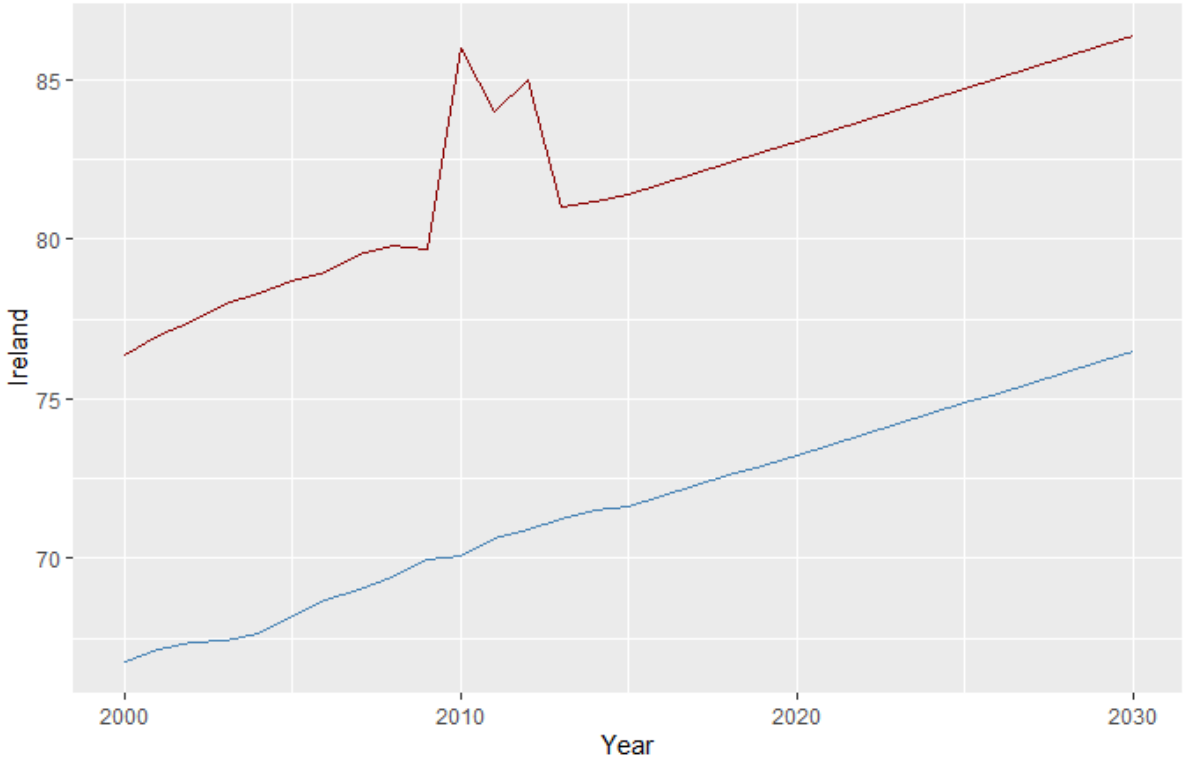


Figure 50 Line graph of predictions for Ireland and the World's life expectancy (2000 - 2030)

The prediction from this analysis is that the life expectancy in 2030 for Ireland will be 86.4 years and the life expectancy in the world will be 76.4 years. As I had predicted the years 2016-2030, to test my predictions I compared my predicted years of life expectancy from 2016-2021 to the real-life life expectancy for both Ireland and the world. The results of that comparison:

Year	Ireland life expectancy (prediction)	Ireland life expectancy (actual data)	Difference (years)
2016	81.73333	81.65	0.08333
2017	82.06667	82.16	0.09333
2018	82.4	82.26	0.14
2019	82.73333	82.2	0.53333
2020	83.06667	82.35	0.71667
2021	83.4	82.51	0.89

Figure 51 Ireland life expectancy predictions vs actual life expectancy (2016-2021)

As can be seen in the table my predictions for Ireland's life expectancy were not far off however the rate of change for Ireland life expectancy decreased by year so my predictions became less accurate. This could suggest that Ireland's life expectancy has started to plateau.

Year	World life expectancy (prediction)	World life expectancy (actual data)	Difference (years)
2016	71.94138	71.72	0.221384
2017	72.26583	72	0.265829
2018	72.59027	72.28	0.310273
2019	72.91472	72.6	0.314718
2020	73.23916	72.63	0.609162
2021	73.56361	72.81	0.753607

Figure 52 World Life expectancy Predictions (2016-2021)

As can be seen from the table, the predicted years in life expectancy get less accurate year by year. This data suggests that life expectancy grew at a faster rate on average from 2000-2015 than from 2016-2021.

Descriptive Statistics

As previously mentioned in the analysis section of the document I generated descriptive statistics on the life expectancy data that I had gathered. I generated these statistics through SPSS.

Statistics

Life.expectancy

N	Valid	152
	Missing	0
Mean		71.844
Median		73.950
Mode		63.5[a]
Std. Deviation		8.0494

Variance	64.793
Skewness	-.471
Std. Error of Skewness	.197
Kurtosis	-.424
Std. Error of Kurtosis	.391
Range	37.0
Minimum	51.0
Maximum	88.0

a. Multiple modes exist. The smallest value is shown

Figure 53 Descriptive Statistics of Life expectancy data

We can see from this table all the statistics that I chose to generate. This shows us that the average life expectancy in the world is 71.844 years, the median life expectancy is 73.95 and smallest most common life expectancy is 63.5 years. According to the table the highest life expectancy in the data set was 88 years and the lowest was 51 years. The skewness and kurtosis values are both under 0 which means the data is not normally distributed and the data has a negative skew.

8.0 Shinyapps.io

In order to present the findings of my research to a wider audience I set myself the goal of deploying my findings to the cloud. To do this I used the website Shinyapp.io which is an online service for hosting Shiny Apps to the cloud. To host the Shiny apps that I had created I needed three files locally. These files were the Shiny Apps connection file where I specified the packages that I would need as well as my Shinyapps.io account details. These details were needed so that the applications I produced would be sent to the cloud from my account.

The server.R file specified the application that I was going to be hosting on the cloud. In this file I specified the data set I was using as well as the Shiny app that I was building e.g., a line plot showing the correlation between two variables. Finally, in the ui.R file I specified the size of the Shiny App I was deploying as well as any other details for the user interface for the Shiny Apps page.

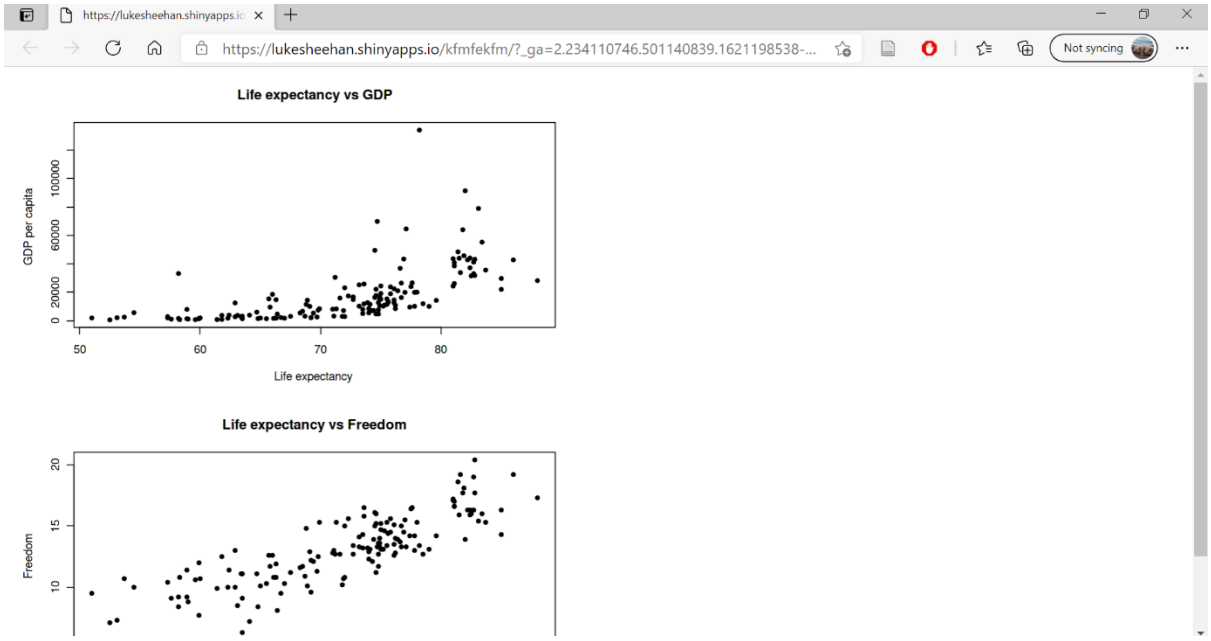


Figure 54 Application deployed on the web.

A very basic Shiny application looks like this. Here I have two R Shiny apps deployed to the cloud from my account.

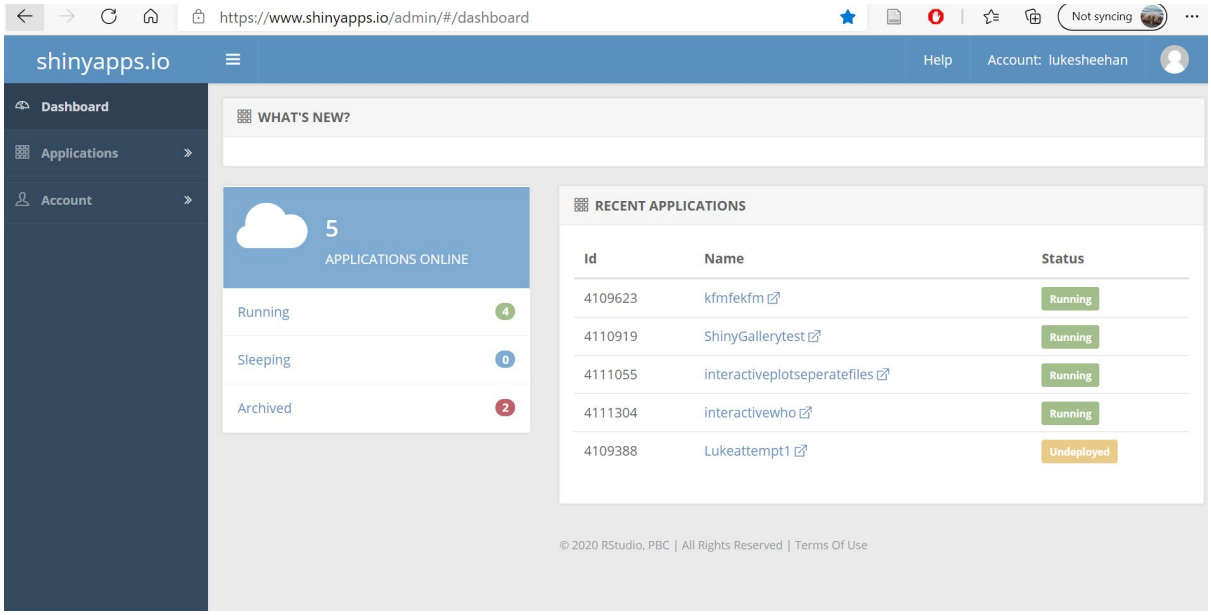


Figure 55 Shinyapps.io dashboard

Through Shinyapps.io I am able to see detailed information on all the apps that I have running or sleeping.

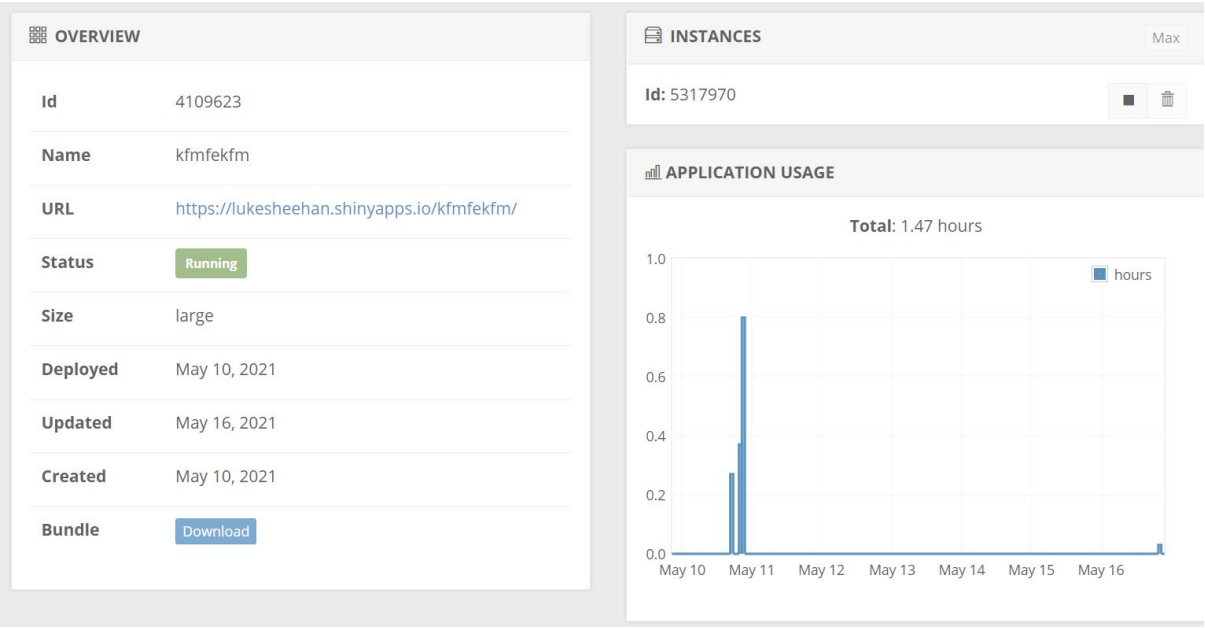


Figure 56 Statistics on usage for R Shiny web application

Here we can see the usage time of one of my Shiny apps. This shows me when I first deployed it as well as its total run time. I chose to use Shinyapps.io as I saw it as a more user-friendly way to present my findings compared to a document alone.

9.0 Conclusions

The question that I set for myself at the start of the project was: what are the main factors that affect life expectancy? To answer this question and to analyse data that surrounds life expectancy in general I used a wide range of techniques and methods. By following the KDD methodology, I collected and organised my data and used data mining tools such as clustering, Random Forests, and linear regressions. Through these methods I believe I have answered the question that I set off to answer.

Through linear regressions I was able to discover what factors correlate/do not correlate to life expectancy around the world. I found this technique to be very useful as it was not too complex, and the findings could be displayed easily through the use of line plots. Due to this technique's low complexity, multiple linear regressions could be run at a time for easy comparison.

With the use of Random Forests, I was able to examine all the variables in my data set that affected life expectancy at the same time. I found this technique to be very useful as it provided me with a clear ranking of the importance of the variables in my data set with regards to affecting life expectancy. Through performing k-means clustering on my data I was able to discover what clusters of countries formed in my data set. Through this I was able to examine which countries had the most similarity to each other. Also, through clustering I was able to determine the optimal number of clusters in my data set. By calculating the rate of change in life expectancy in Ireland and the rest of the world I was able to make predictions of by how much life expectancy would grow in the next decade.

I believe the strengths of this project are how the results of my analysis were presented. Through R I was able to produce visualisations of all my findings in a user-friendly manner. Another strength of this project was the range of data that I could use for further analysis, it was very easy for me to add extra columns to my data set or merge my data set with another data set for further analysis.

In terms of limitations of the project I would say the main limitation is the lack of personal attributes examined that affect life expectancy. Through analysing data that surround life expectancy, I found that the majority of data is per country such as pollution levels and levels of schooling. There was a lack of data per capita such as marital status or hours of exercise per week. Due to this I was not able to explore as many personal factors that affect life expectancy.

The analysis undertaken has demonstrated that data mining and data analytics techniques can be used to analyse complex data sets and solve complex questions such as the factors affecting life expectancy. This is a complex and challenging question that can only realistically be solved with recourse to the powerful data analysis tools deployed in the foregoing study.

Key methodologies such as KDD Data Mining ad Random Forests, Clustering, Linear and Multiple Regressions, when combined, allow powerful data management and data analysis, leading to important findings. In terms of specific findings based on the analysis, I was able to demonstrate the factors affecting life expectancy. These include:

a. From my linear regressions I determined that GDP per capita and levels of schooling both correlate to life expectancy.
b. Ireland's life expectancy is growing at a faster rate than the rest of the world.
c. Through Random Forests I discovered that the most important factors for determining life expectancy are: Income composition of resources, HIV rates and schooling levels.
d. There is an optimal amount of four clusters of countries in the main data set I have looked at.
e. There are similarities in life expectancy based on a country's region (based on scatter plot analysis).

10.0 Further Development or Research

If I were to continue with this project, there are many different things that I would do to build upon my initial research. I would like to look at even more factors that contribute to life expectancy but on more of a personal scale. In my project I mainly researched factors that differ from country to country such as pollution, freedom levels and schooling etc. I would like to look at more personal factors such as exercising and dietary habits of individuals. I would be curious to see how a factor such as number of hours of exercise a week would affect an individual's life expectancy.

In my project I hoped to provide information on the main factors that affect life expectancy to people around the world based on their country. Further research could be conducted into more personal factors that affect life expectancy so that the users of my project would be able to apply this information to their own lives. To accomplish this, I would look at more sources of data to include in my analysis. With additional time I would have begun developing an application for users to chart their own approximate life expectancy based off the data I have collected.

11.0 References

Rajarshi, K., 2021. *Life Expectancy (WHO)*. [online] Kaggle.com. Available at: <https://www.kaggle.com/kumarajarshi/life-expectancy-who> [Accessed 5 May 2021].

Singh, A., 2021. *World Happiness Report 2021*. [online] Kaggle.com. Available at: <https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021> [Accessed 5 May 2021].

GeeksforGeeks. 2021. *KDD Process in Data Mining - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/kdd-process-in-data-mining/> [Accessed 6 May 2021].

Boehmke, B., 2021. *K-means Cluster Analysis · UC Business Analytics R Programming Guide*. [online] Uc-r.github.io. Available at: <https://uc-r.github.io/kmeans_clustering> [Accessed 15 May 2021].

Group, D., 1999. *Man's body*. Chicago, Ill: Contemporary Books.

GitHub. 2021. *owid/owid-datasets*. [online] Available at: <https://github.com/owid/owid-datasets> [Accessed 15 May 2021].

Macrotrends.net. 2021. *World -2021*. [online] Available at: <https://www.macrotrends.net/countries/WLD/world/life-expectancy'>World Life Expectancy 1950-2021> [Accessed 16 May 2021].

12.0 Appendices

12.1. Project Proposal

Contents

1.0 Objectives

The main objective for my project is to provide an analysis on life expectancy. I aim on providing an analysis on what are the factors that affect life expectancy around the world. I also aim on using my data to see what are the factors that correlate towards a countries' life expectancy e.g. pollution levels or happiness rates. I also plan on seeing how a person's life expectancy is affected by their socio-economic group. I aim for my project to provide an insight to how a person's life expectancy is affected.

To accomplish these objectives, I have a set of smaller objectives to accomplish. First, I will be gathering relevant data on the following topics:

- Life expectancy around the world
- GDP per country
- Happiness rates around the world
- Pollution levels per country

I also plan on comparing my research to actuarial models.

2.0 Background

I first came up with this idea since I had an interest in statistics about life expectancy and GDP in different parts of the world. I was inspired from a time I watched a news report from England where it was revealed that the life expectancy in England varied drastically depending on where a person was born. I found this to be very interesting and got thinking about what are all the factors that affected their life expectancy since it couldn't just be genetics. It got me thinking about how someone's life can be affected by how the culture and society they grow up in.

I was also inspired by the book: Man's Body an Owner's Manual. This was a book that provided a complete guide to the mental and physical workings of the male body. It had comprehensive statistics on the male body and health in particular. It allowed readers to chart their own life expectancy based on factors such as where they were born, how often they exercise, do they smoke etc. I found this idea of being able to predict your own life expectancy to not only be interesting but also educational as one could use this information as an eye opener into how healthy they are and what negative habits they have.

I would like to produce my own analysis on life expectancy backed up by data I source which can give readers and insight into their own life expectancy and life expectancy around the world.

3.0 Technical Approach

The approach will be to first gather the required data for my project. I will have to format it and create visualisations of my data. I plan on researching actuarial models to see how they get

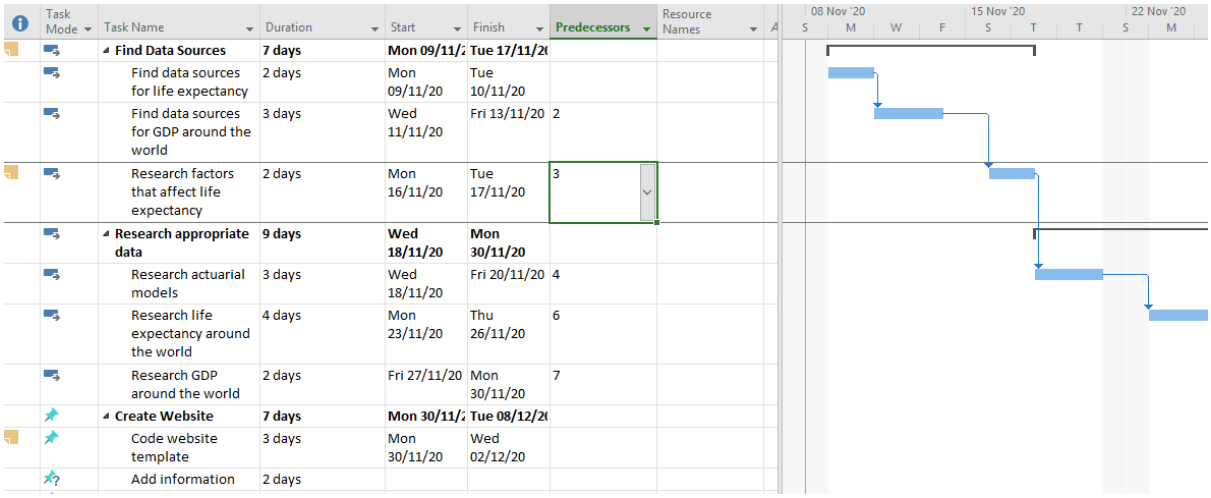
their data and to see what inputs they give that affect life expectancy. I also planning on learning more on how to use languages like python and r for data analysis.

4.0 Special Resources Required

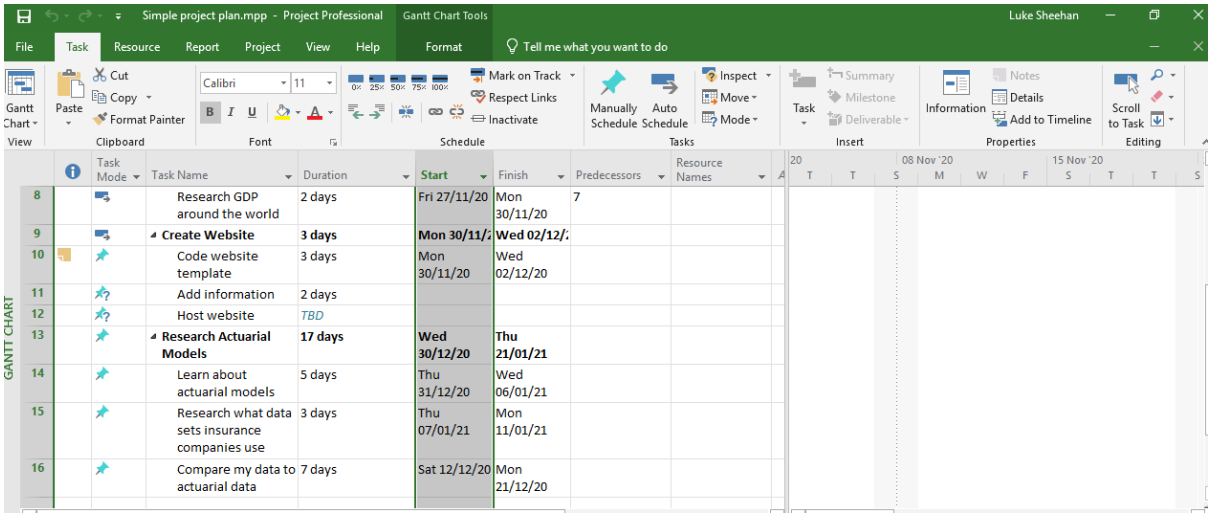
N/A

5.0 Project Plan (in progress, dates may vary)

Below is the initial draft of my project plan which includes the main tasks to be completed. More tasks will be added as the project advances.



	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
		Find Data Sources	**7 days**	**Mon 09/11/**	**Tue 17/11/2(**		
		Find data sources for life expectancy	2 days	Mon 09/11/20	Tue 10/11/20		
		Find data sources for GDP around the world	3 days	Wed 11/11/20	Fri 13/11/20	2	
		Research factors that affect life expectancy	2 days	Mon 16/11/20	Tue 17/11/20	3	
		Research appropriate data	**9 days**	**Wed 18/11/20**	**Mon 30/11/20**		
		Research actuarial models	3 days	Wed 18/11/20	Fri 20/11/20	4	
		Research life expectancy around the world	4 days	Mon 23/11/20	Thu 26/11/20	6	
		Research GDP around the world	2 days	Fri 27/11/20	Mon 30/11/20	7	
		Create Website	**7 days**	**Mon 30/11/**	**Tue 08/12/2(**		
		Code website template	3 days	Mon 30/11/20	Wed 02/12/20		
		Add information	2 days				



	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
8		Research GDP around the world	2 days	Fri 27/11/20	Mon 30/11/20	7	
9		**Create Website**	**3 days**	**Mon 30/11/**	**Wed 02/12/**		
10		Code website template	3 days	Mon 30/11/20	Wed 02/12/20		
11		Add information	2 days				
12		Host website	*TBD*				
13		**Research Actuarial Models**	**17 days**	**Wed 30/12/20**	**Thu 21/01/21**		
14		Learn about actuarial models	5 days	Thu 31/12/20	Wed 06/01/21		
15		Research what data sets insurance companies use	3 days	Thu 07/01/21	Mon 11/01/21		
16		Compare my data to actuarial data	7 days	Sat 12/12/20	Mon 21/12/20		

6.0 Technical Details

For my project I plan on using the languages R and Python as they are the most popular in data analytics and can be used to create visual representations of statistics and data. I also plan on creating a website to host the information that I sourced.

I have been creating graphs and plots using R packages in R studio.

7.0 Evaluation

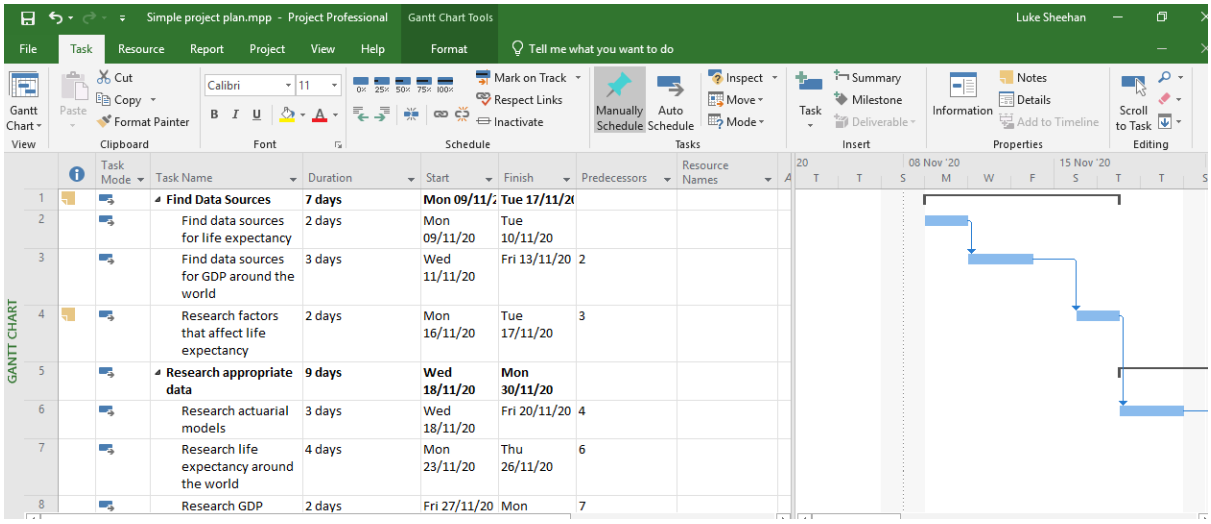
No tests have been decided as of now.

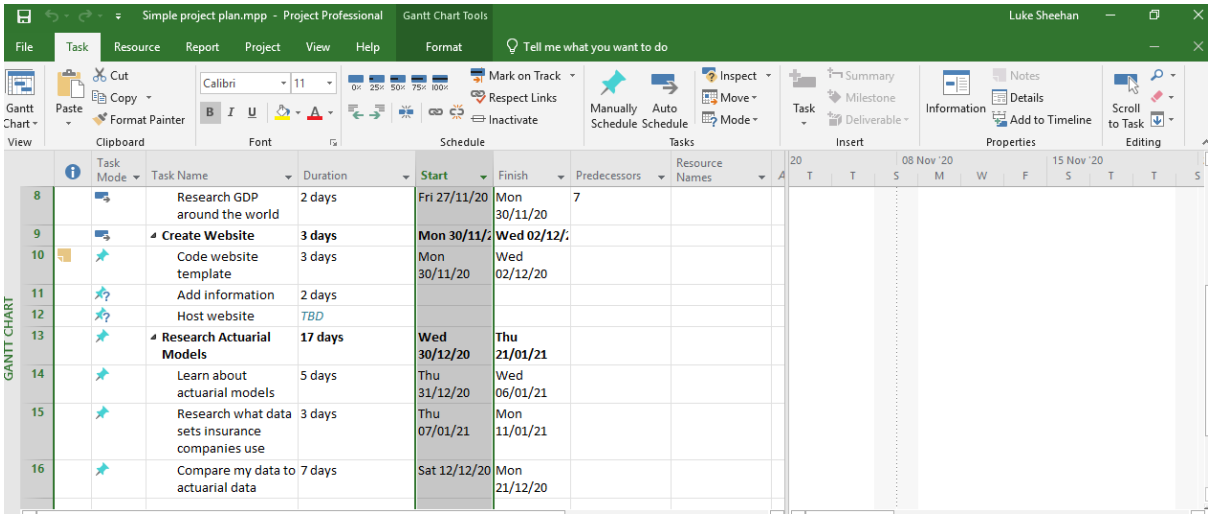
8.0 Invention Disclosure Form (Remove if not filled)

N/A

Signature: 

9.2 Project Plan





9.3 Reflective Journals

Reflective Journal Semester 1 - October

Luke Sheehan – x17361401

Past

This month I developed my project pitch. I did so by researching past data analytics projects from former NCI students as well as thinking about what I am interested in. I wrote up a project pitch and then recorded a video of the pitch. In the video I covered the main questions that needed to be answered. Such as:

- What will the project do?
- Why is it challenging?
- Who is the project for?

- Why should this Project be attempted?
- How is it different than what has been done in this area before i.e. do some preliminary research to ensure there is not an obvious example of the exact same idea?

I recorded the video and submitted it to Moodle.

Present

Recently I met with my supervisor who reviewed my project pitch. Overall, he was happy with it but had some ethical concerns. He believed that some parts of my idea would cost myself a lot of problems in the long run. I was given some advice on how to avoid these problems. He also gave me some tasks to do before our next meeting. We scheduled next weeks meeting and made plans to meet on a bi-weekly basis after that.

Future

For this week I plan on taking my supervisors advice on board. I will start to look at possible data sets I can use for my project. I also will start researching the KDD methodology and actuarial models. I will also start developing a project plan before our next meeting.

Reflective Journal Semester 1 - November
Luke Sheehan – x17361401

Past

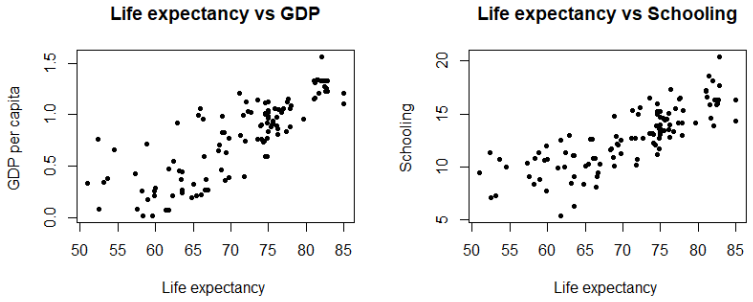
This past month I made progress with my project. I identified two relevant data sets. I was able to explore the data in these data sets using R. I also started cleaning and processing the data through R.

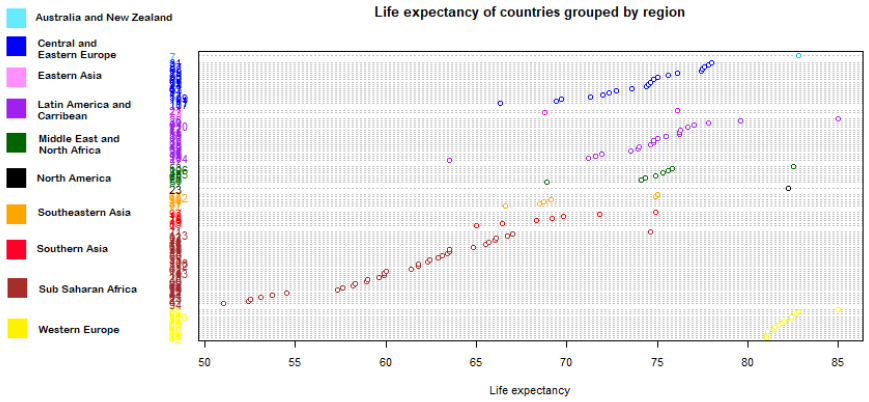
After I had processed my data sets, I merged them together and started to create some visualisations of the trends and patterns that I discovered in my data. I did this through R using the ggplot package.

I also looked into the KDD methodology and explored some of the steps involved in this methodology.

Present

I had a meeting with my student supervisor recently. I caught him up with my progress. We discussed the plan for my next month. He recommended that I look into the lm function in R linear regression. I was also recommended to investigate normalising/standardising numeric attributes methods in data analytics.





Future

For my next month, I plan on looking into the lm function in R linear regression as well as looking into normalising/standardising numeric attributes methods in data analytics. I also plan on looking into actuarial models.

Reflective Journal Semester 1 - December
Luke Sheehan – x17361401

Past

This past month I spoke with my supervisor about my upcoming midpoint presentation. In preparation for this submission I came up with a script for my video and also made a power point slide deck.

Present

This month I had the midpoint submission. The deliverables were: mid-point documentation, a video presentation and a slide show. In the documentation I explained the technologies and

methodologies I had been working with in the last while. I also discussed the data sets I have been using so far in my project.

In my presentation I was given the chance to show some of the insights and observations I have discovered and made with my data analysis so far. As part of my presentation I executed my R code and showed how I generated the visualisations from my data sets. I also touched on my future plans for my project.

Future

For the future of my project I plan on looking at actuarial models and looking at even more data sets. I would like to find more unique insights from my data.

Semester 2 Reflective Journals

Reflective Journal Semester 2 - January
Luke Sheehan – x17361401

Past

The past month I submitted the last of my submissions and finished my exams for semester one. I then had two weeks off where I was able to do some more research on the topic of my project. I researched some more data sets. I also began using the linear regression function with my data sets as well.

Present

Currently I am still looking at more functions for my data sets. I am also looking at more data sets, the purpose of this is to eventually be able to create more insights into what affects life expectancy. I have also been using R more.

Future

For the next month of my project I plan on continuing with what I'm currently doing with regards to data sets. I also plan on creating more visualisations of the insights and trends I found in my data sets.

Reflective Journal Semester 2 – February

Past

The past month I worked more on my project document and worked more on my project plan. I set up my project profiles for the NCI website and I set up a Github repository for my project. I explored other data sets the past month for my project on Kaggle and looked at example of data visualisations that have been made by other people in data analytics.

Present

I am currently looking at R Shiny for my project which I aim to use as a visualisation for my project showcase. I also plan on looking testing for my project as that is a big part of the final project. I'm still working on my project document as there is a lot to it and I want to improve the sections that I've already written.

Future

For the next month of my project, I plan on performing some clustering on the current data I have. I am also planning on looking at self-organising maps, decision trees and Random Forests. I am also planning on using my data to try find certain groups that are of high risk for mortality.

Reflective Journal Semester 2 - March
Luke Sheehan – x17361401

Past

The past month I completed my project profile for the project showcase. I started looking at using SPSS to perform statistical tests on data. Some statistical tests I performed were:

- **Normality tests**
- **Man Whitney test**
- **Descriptive statistics about data**
- **Kruskal-Wallis**
- **R analysis**

Present

Currently I am still learning SPSS and also learning more of R. I have explored more libraries in R and have been able to create new types of visualisations for my data. I have also looked at data analysis tools in excel.

Future

For the next month of my project, I will have more time so there will be more things I plan on looking at. I plan on looking at R Shiny tutorials, forming clusters with normalisation and continuing with my project document.

Some other topics I plan on looking at; One hot encoding, K-medoids, hierarchical clustering, self-organising maps and min-max normalisation prior to clustering.

Reflective Journal Semester 2 - April
Luke Sheehan – x17361401

Past

The past month I have been mainly looking at R Shiny. This is a useful tool for making web apps from R. My aim is to have a user interface to all the visualisations that I have created through R and hopefully have some of the visualisations be dynamic. I have deployed some basic applications on Shinyapps.io which is a platform for hosting Shiny applications.

Present

Currently I am working on my final document for my final submission. I am gathering all the research that I have completed in the last months and adding it into my document. I am also gathering all the code files I have created over the past semesters to be included in my code submission and GitHub repository.

Future

For the final month of my project, I aim to wrap up the analysis side of my project so that I can complete the documentation side of the project for the final submission. I also aim to have a user-friendly R Shiny app deployed to the cloud so that users can view my findings in a user-friendly manner. Finally, I will be wrapping up my documentation and project through writing my conclusions of my project.