National College of Ireland

BSc (Honours) in computing

Data Analytics

2020/2021

Wiktor Wolsza

x17444592

x17444592@student.ncirl.ie

HomePoint

Technical Report

Contents

Executive Summary

Life expectancy is one of the primary factors which indicate to the world how developed a particular place in the world is. In this report I will be focusing on life expectancy however, I will also be looking closely into the correlations between life expectancy and happiness around the world with respect to various factors such as GDP, air quality in order investigate what leads to happiness and ideally long lives around the world.

The goal of this project is to understand the factors which have an effect on life expectancy as well as how life expectancy has changed over time. In this report I will be using various technologies and techniques in order to investigate those key factors around the world.

In order to find the answers to all of the questions I have it was important to firstly start by picking the appropriate methodology. The two main methodologies used in data analytics are KDD and CRISP-DM which meant that my choice was between these two. After carefully analysing the strengths and weaknesses of the methodologies it was clear to me that KDD is the correct methodology for a project of this nature.

Data visualisation was a major component of this project as it was my aim to create something which any user could use and potentially find new and interesting patterns in the rate of change in life expectancy around the world. In order to achieve this, I used tableau to create a user-friendly dashboard. Other technologies such as Rstudio, Excel, SPSS and Jamovi were used to extrapolate as much information from the data as possible.

1.0 Introduction

1.1. Background

Life expectancy is often seen as one of the fundamental metrics for indicating a countries development as well as understanding population growth around the world. Based on my research of this topic I have typically found that not many try to cover the topic with as many aspects in mind as I plan to do. I have decided to undertake this project in order to examine the factors which lead to high and low life expectancies as well as happiness around the world and very importantly the factors which corelate the most with both life expectancy and happiness. (i.e., I want to find out what leads to a long and happy life) As well as investigating through data analytics and data visualisations the trends and changes associated with life expectancy around the world.

For a long time, I have been very intrigued in the statistics behind different countries progress and living conditions. One way in which I have gained an interest in the overall topic was simply by coming across a video on the internet where the creator talked about what each country in the world does best, I found this to be very interesting however not in depth enough which is to some extent how the idea for my project came about.

1.2. Aims

The overall aim and main objective of the HomePoint project is to analyse life expectancy around the world over the course of many years with respect to considerations such as happiness around the world. The key in the project is to carefully examine the various factors which may or may not contribute to life expectancy and happiness around the world. Through statistical tests I aim to even judge how big of an affect each specific factor such as GDP or air quality has on both happiness and life expectancy therefore to an extent answering an age-old question whether or not money buys happiness.

In order to achieve my objectives, it is important to ensure that I have an appropriate collection of data which is sufficient for each of the topics/factors which I will be covering.

In this case it is important to gather data on topics such as:

- Life expectancy around the world
- Historic life expectancy around the world
- Happiness scores around the world
- Air quality around the world

Using the data I find, I plan to develop a user friendly and interactive tableau dashboard which will be used for allowing the user to view a map of the world with colour coded countries where the colour corresponds to the life expectancy of the given country. Using the map, the user can analyse the patterns between different areas and continents around the world.

By utilising the capabilities of tableau, I aim to visualise the changes and trends in terms of life expectancy around the world by creating an animated visualisation of the historic progress the world has made in terms of life expectancy. The user will be able to view the patterns of any of the countries within the data and will be able to highlight several countries trends in order to find patterns between the countries.

After testing several data mining techniques, I have decided that I will use model trees in order to explore and understand my data which can be used to form accurate predictions of future trends.

Another aim is to use technologies like SPSS to analyse the data and specifically the dependencies between the data to carry out tests which will show how big of an impact different factors may have primarily on life expectancy but also happiness around the world. Furthermore Jamovi will be used for certain statistical tests as well as visualisations.

Lastly one of my biggest aims for this project is to present my data and findings in an elegant, aesthetically pleasing, and user-friendly way.

1.3. Technology

In order to make the most out of my project it is important that I choose each of the technologies I use in accordance with each of the technology's strengths and weaknesses as well as my own strengths and weaknesses to ensure that I am being as efficient with my time as possible.

Firstly, to gather my data I have used a wide range of sources such as world bank data, Kaggle as well as google data set search. From most of the sources I was able to download the files in csv format which allowed me to easily examine the data within Excel which is one of the technologies which is used for this project.

After utilising Excel, the next technology on the list is the R language in which I coded within R studio. By utilising R, I was able to prepare my data quickly and efficiently by cleaning it in order to merge the datasets correctly. I have chosen to use R as I have some prior experience with it however one of the key reasons for choosing R is that thanks to its vast numbers of packages R's capabilities are enormous.

R was without a doubt a great choice for this project as it is a statistical computing language which has been around for a very long time meaning that the number of resources as well as packages which are present for the language made it a very powerful tool. I used the R language in order to perform a data mining technique called model trees which is used for the prediction of future values.

To create an interactive and user-friendly dashboard I have used tableau which allowed me to efficiently make a map of the world with countries being colour coded based on their life expectancy.

Furthermore, using tableau, I was also able to create a animation type of visualisations which shows the changes and improvements in life expectancy over a span of many decades. The animation can be paused at any point and allows the user to highlight different countries which can be used to compare and contrast the trends followed by different countries around the world over a chosen period of time.

For further tests including descriptive statistics both SPSS and Jamovi were used in order to calculate the correlation between several factors and the life expectancy as well as happiness around the world. SPSS is a technology made by IBM and is one of the key technologies used in order to perform statistical analysis.

1.4. Structure

My report will be structured as follows. In section 2 will describe the data used for this project. In this section I will be describing the data which used for the project including the source of the data as well as how it was organised to allow me to carry out my analysis.

Section 3 of the document is about the choice I had in terms of methodology for this project. In this section I analyse the methodologies which are most commonly used including their strengths and weaknesses and an explanation as to why I have picked the chosen methodology. The methodology of choice is then described in detail including all of the steps needed to successfully carry out the analysis.

Section 4 of the document focuses on the statistical tests which I have carried out on my data. Using technologies such as SPSS.

Section 5 includes certain tests which have been carried out during the whole project's life.

Section 6 is based on the analysis which I have carried out with details regarding the overall approach and the aims of the analysis as well as the fundamental reasoning behind the chosen analysis backed by visualisations and models.

Section 7 describes the results of my analysis as in this section all of the results are discussed and presented with the aid of visualisations in order to help my findings become easily understandable.

Section 8 focuses on my use of tableau in order to create an interactive map showing life expectancies around the world as well as an animation which shows the changes in life expectancy over a number of years.

Section 9 is the conclusion of my work by once again presenting the key findings within my project. The aim of this section is to answer the question oriented around this project which is to see which countries perform best in various ways such as happiness & air quality as well as very importantly investigating what factors the biggest determining factors are when it comes to happiness and life expectancy around the world.

Lastly in section 10 of the report I am going to discuss what possible additions and developments could be made within my project in order to expand on it. Finally in this report all references will be cited as well as the original project proposal will be attached along with all of the monthly reports which have been completed to aid me in keeping this project on track in terms of progress.

2.0 Data

When beginning this project, I knew that it was important to pick my datasets carefully but also keeping my mind open to add more datasets in order to expand the project further or overall improve the accuracy of my research. One of the places where I looked for data was Kaggle which is where I found my two main datasets right at the beginning of the project which are the World Health Organisation (WHO) life expectancy dataset and also the 2015 world happiness report dataset which were the first two datasets used for this project.

I have decided to use the WHO life expectancy data set as it is from the world health organisation which had used the global health observatory to collect/compile the data. This meant that the data was from a known and reliable organisation which was an added bonus when choosing the dataset. Furthermore, I have chosen this dataset as it covers an extensive amount of not only health but also social factors. The WHO dataset contains a total of 22 columns containing 193 countries with data collected for a total of 16 years from the year 2000 to the year 2015. The columns present in this dataset are as follows 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. This broad range of information present in the dataset allows me to meet my goal of analysing the correlations between various factors and life expectancy and of course happiness around the world. (Rajarshi, 2021)

In order to achieve the goal of also analysing happiness I had to also find a dataset which would be viable to get this information which is how I found the world happiness report dataset. The world happiness report contains data on 149 countries around the world with a total of 20 columns, most importantly including the happiness score each country achieved. This dataset is another dataset from a well known and well-established organisation which is the United Nations. The world happiness report has been growing every year and has been released in 2017 by the United Nations which was done in order to celebrate the international happiness day. (Solutions Network, 2021)

By merging the datasets together, I was able to analyse a wide array of correlations between many factors including the link between happiness and life expectancy.

As it was also my aim to create a visualisation to showcase the change in life expectancy over time, I had to find a dataset which contained data on life expectancy over a span of many decades. In order to find the correct dataset, I had utilised Google's dataset search engine which is how I came across the data which I used from the SuperDataScience website. (SuperDataScience, 2021)

Considering it was my aim to find as many different factors as possible which can have an affect on not only life expectancy but also happiness, I began looking for a dataset which could showcase air quality/pollution levels around the world. I decided to take this factor into consideration simply due to the fact that nowadays there is a big push around the world for greener and more environmentally friendly energy in order to reduce the levels of pollution in the atmosphere which spiked my interest as I wanted to analyse the effects of pollution on life expectancy around the world. When this data was found by utilising Google's dataset search, the data was then merged and processed to include the needed columns for my analysis, this dataset shows the average per year exposure to the fine particulates which demonstrate the quality of air and the level of pollution present in the air per cubic meter. (Weinmeister, 2021)

When analysing this data, I had come across columns which had significant number of missing values, the columns which had a large portion of the data missing I had decided to omit, however in instances where there were not much missing values, I had decided to replace the null values with the mean of the column in question. In order to maintain the highest possible accuracy, I had filtered the data between developed and developing countries which then allowed me to find the average for each of the columns specifically for developed or developing countries. This meant that while replacing missing values for developing countries the average which was substituted in was the average for developing countries and of course vice versa for developed countries.

Lastly, when carrying out my analysis I have found that the GDP column of my data had several values which were somewhat inaccurate. Upon investigating this further I had decided to use a slightly different metric for this particular column, which is GDP PPP, this means that the GDP is based on the Purchasing Power Parity (PPP). I have decided to use this form of GDP calculation as it is seen as a more reliable way of judging each countries domestic market. PPP is in my opinion better than GDP per capita simply due to the fact that PPP takes into consideration more factors, very notably factors such as the relative cost of goods as well as services in each of the countries as opposed to simply taking into account the rates within the global exchange which can to some extent distort the real situation in any given country which is why PPP was my choice. (Nitisha, 2021)

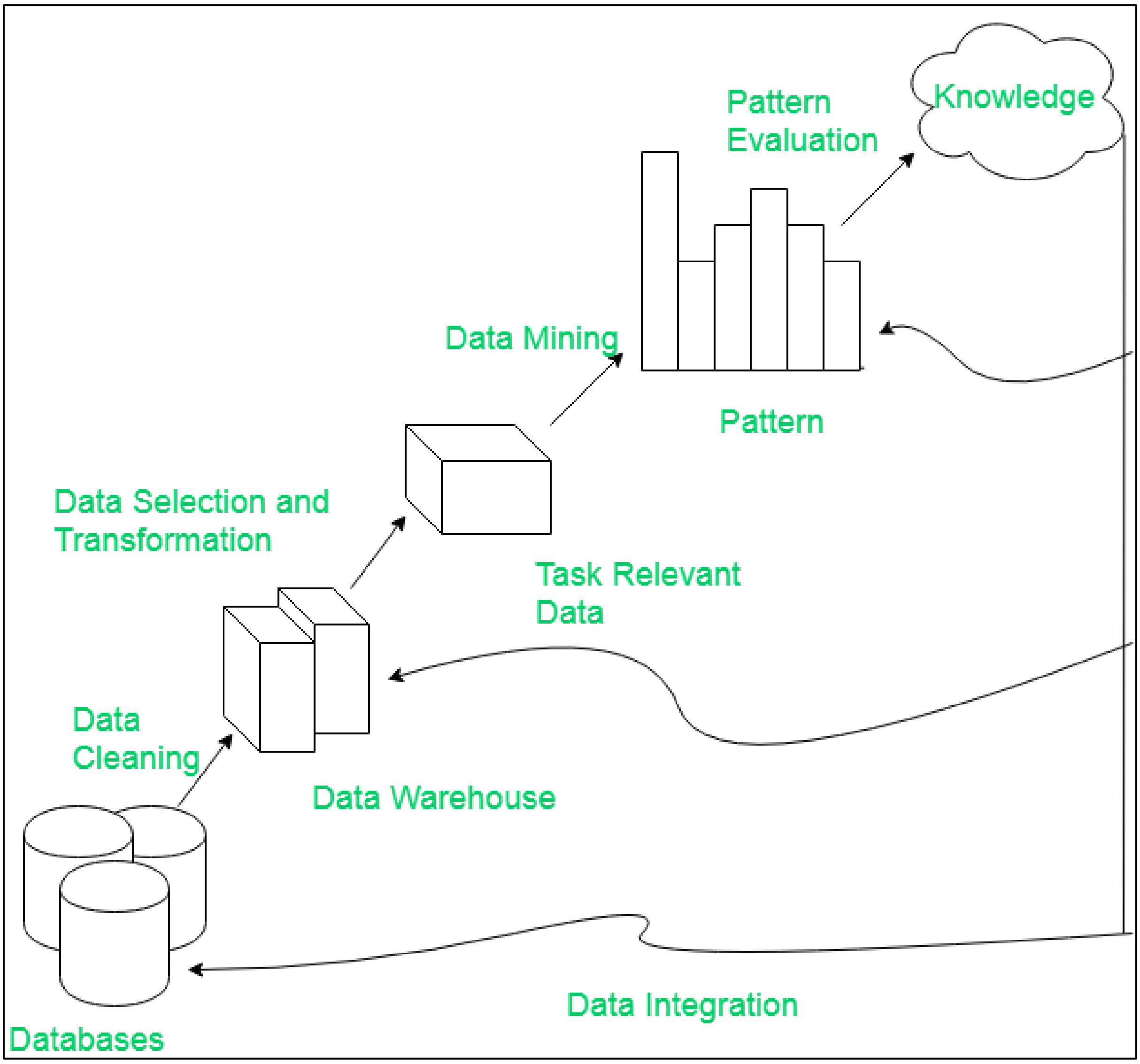
In order to find the data on GDP PPP I have once again used googles dataset search engine to find the dataset on Kaggle, the dataset featured GDP PPP for a total of 195 countries for the years from 1990 to 2018, however as the WHO data is up to 2015, I have decided to of course take the year 2015 for GDP PPP when merging the datasets.

3.0 Methodology

Early on when choosing a methodology, it was important for me to understand what the most popular methodologies within data analytics are and exactly what aspects of each methodology make it so popular. This meant that I had to research the most frequently used methodologies to understand their application as well as their strengths and weaknesses.

The two most popular methodologies which are used for data analytics are CRISP-DM and KDD. Based on my research KDD is more suited to my project as it is the methodology used for scientific research meanwhile CRISP-DM is a lot more business oriented due to the extra steps within CRISP-DM which directly take into consideration the business aspects of the data mining and includes deployment during its cycle which is not needed for this particular project which is why KDD was the only possible option. KDD on the other hand begins its life with data and eventually ends with various insights and knowledge without taking into consideration any business aspects which is why I chose KDD.

There are several steps involved in the KDD methodology which are as follows: Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and lastly knowledge representation.



(Figure 1: KDD methodology)

3.1 Data Cleaning

Data Cleaning was the first step in the methodology and took place right after I have found suitable data in order to carry out my analysis. This stage of the methodology can be known as pre-processing. This stage of the methodology involves aspects like dealing with missing values which can be done by utilising techniques like MICE which is able to take an accurate educated guess on what the missing values could be based on the existing data.

This section also includes the removal of discrepancies or outliers to make the study become a more accurate representation of the scenario. In some cases, the data I have chosen lacked data for several countries that as part of cleaning the data I was forced to omit certain columns as not all countries keep up to date data on the same criteria.

In some scenarios however I was able to find a substitute dataset which allowed me to use more accurate data for the columns which were initially not accurate by performing a merge and delivering the most accurate results possible.

3.2 Data Integration

As my project is based around comparing countries to one another and establishing what factors lead to a high life expectancy as well as happiness the data integration stage of the methodology was one of the most important for my project as various categories of information were needed for example air quality data.

In order to merge the data R was used to combine the data based on country name. This method was also used to substitute columns with more accurate data.

3.3 Data Selection & Transformation

Selecting the most appropriate data for the analysis was also crucial to ensure the highest possible level of precision. Often the data needs to be transformed in order to be best suited for the analysis which can extent beyond merging of data.

This section of the methodology can include Principal Component Analysis (PCA) which can effectively condense the data while retaining the vast majority of the information needed which can make certain algorithms run a lot more efficiently as the data had been transformed into a more appropriate state.

3.4 Data Mining

This stage of the methodology is important as an appropriate algorithm is applied on the data in order to extrapolate the necessary information. Crucially this step includes the tuning of parameters from the dataset to be applied in the algorithm. The techniques used in this project were model trees for the prediction of life expectancy as well as regression. It is at this stage where particular patterns and trends can emerge, and interesting insights can be observed.

3.5 Pattern Evaluation & Knowledge Representation

Pattern evaluation is where the patterns found in the previous stage are examined and carefully analysed in order to determine how accurate as well as useful the patterns are. The knowledge gained can be understood by utilising visualisations such as bar charts and other plots.

In my project I have put a big emphasis on creating clear and easy to understand visualisations even including animated visualisations within tableau. In this part it is crucial to ensure that all knowledge and insights gained from the study are appropriately presented in a way that is user friendly and easy to understand.

(KDD Process in Data Mining - GeeksforGeeks, 2021)

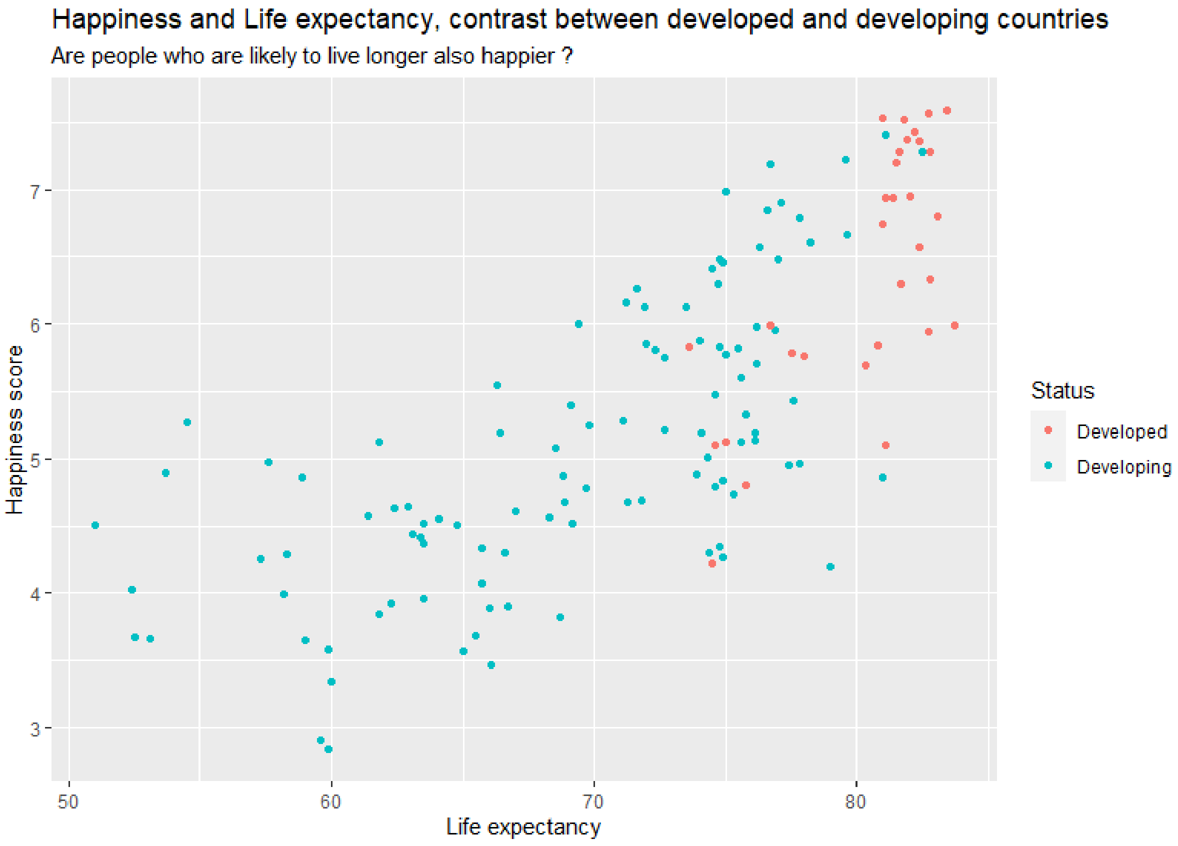
4.0 Analysis

4.1. Difference between developed and developing countries

In order to begin my analysis, I had firstly revised the methodology chosen which is why straight after choosing my datasets I began cleaning the data. The first part of cleaning the data included filtering the data in order to understand what columns and values are of most interest to me, this is why I began by removing columns which were either not needed or simply contained many null values. Developing countries had a tendency of missing values in certain columns, in instances where I felt like there was so many data missing, I had decided that it was not worthwhile carrying out techniques to fill in the missing data as in that case it would mean that even more than half of the data would be generated by a technique such as MICE/AMELIA or using the average to substitute for null values.

The next step which I had undertaken as part of following the methodology was the integration of data. This meant that I began to merge the WHO dataset with the 2015 happiness report.

The key values which were of interest to me were of course the life expectancy values and the happiness scores which I could now compare as I have merged the data successfully. I began by utilising R packages to help me code clear and easy to understand visualisations, in this case I made use of the ggplot2 technology/package within RStudio. To begin I had decided to categorise and colour code the countries based on whether or not the country is developing or developed.



(Figure 2: Happiness and Life expectancy)

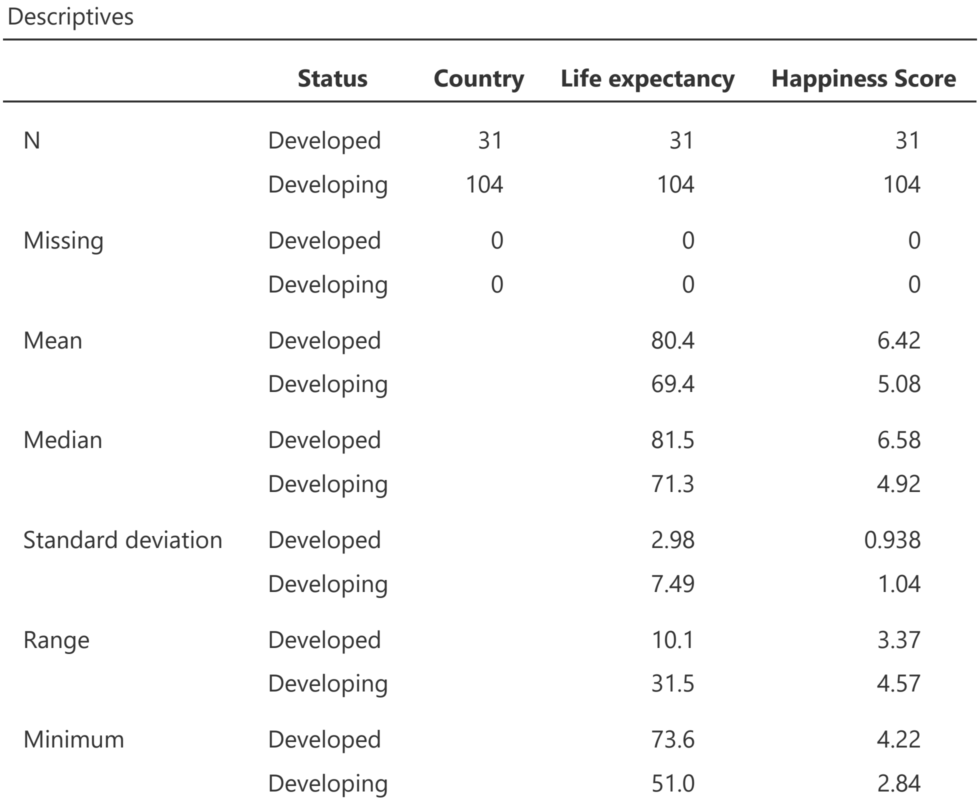
The visualisation above shoes quite a strong correlation between people being happy also living longer. The other clear correlation which was to be expected is that of course people in developing countries tend to live shorter lives and less happy lives however this is certainly not always the case.

Chile was one of the countries that peaked my interest due to their very high life expectancy in 2015 which was at very close to 80 years despite being a developing country. To add to that Chile had also scored nearly 6.7 in terms of happiness score which is an excellent result.

In analysing my data Qatar had also intrigued me and in particular its status as a developing country according to the UN in 2015 despite being such as wealthy country already at this point.

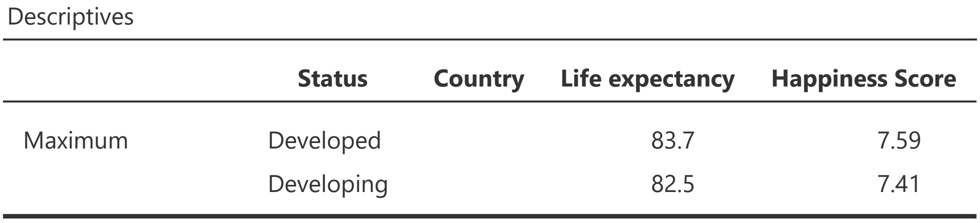
The reason why Qatar was at the time still considered a developing country was due to the fact that the actual development which had been done in the country was quite low in proportion to factors such as GDP for instance which is why it was considered a developing country in 2015. As this country quite clearly is a very big outlier in terms of developing countries, I had decided to investigate whether or not this was correct, and of course it was indeed still a developing country.

To analyse the data which I had already cleaned and processed I had decided to export it as a CSV in order to carry out additional analysis using the Jamovi software as it is my aim to use a wide variety of technologies for this project.



Descriptives

	Status	Country	Life expectancy	Happiness Score
N	Developed	31	31	31
	Developing	104	104	104
Missing	Developed	0	0	0
	Developing	0	0	0
Mean	Developed		80.4	6.42
	Developing		69.4	5.08
Median	Developed		81.5	6.58
	Developing		71.3	4.92
Standard deviation	Developed		2.98	0.938
	Developing		7.49	1.04
Range	Developed		10.1	3.37
	Developing		31.5	4.57
Minimum	Developed		73.6	4.22
	Developing		51.0	2.84



Descriptives

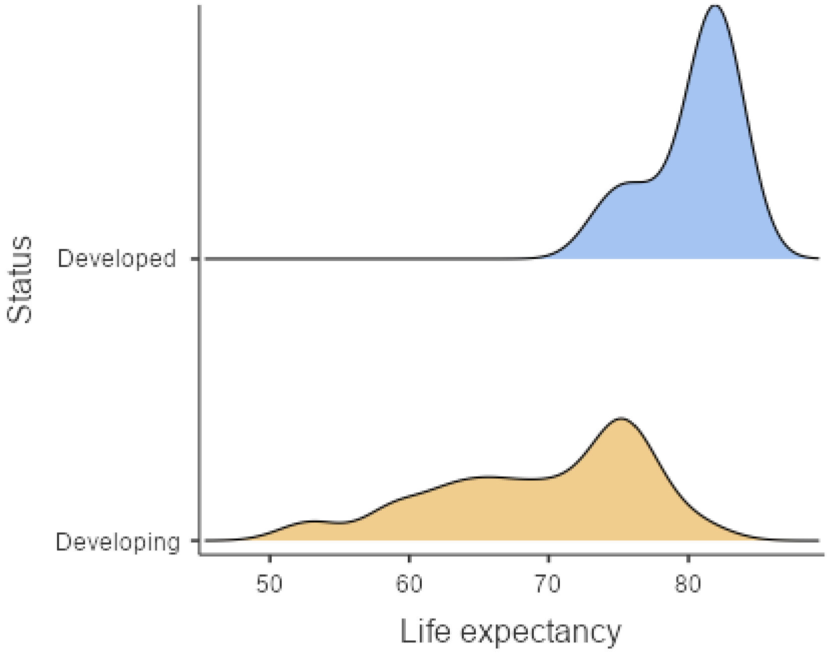
	Status	Country	Life expectancy	Happiness Score
Maximum	Developed		83.7	7.59
	Developing		82.5	7.41

(Figure 3: Happiness and Life expectancy, Descriptive Statistics)

As can be observed in Figure 3 there as expected there is quite a drastic difference between what life is like in developing and developed countries however during my analysis to some extent, I was surprised how life is different to the vast majority of people depending on whether or not they were born in a developed or developing country.

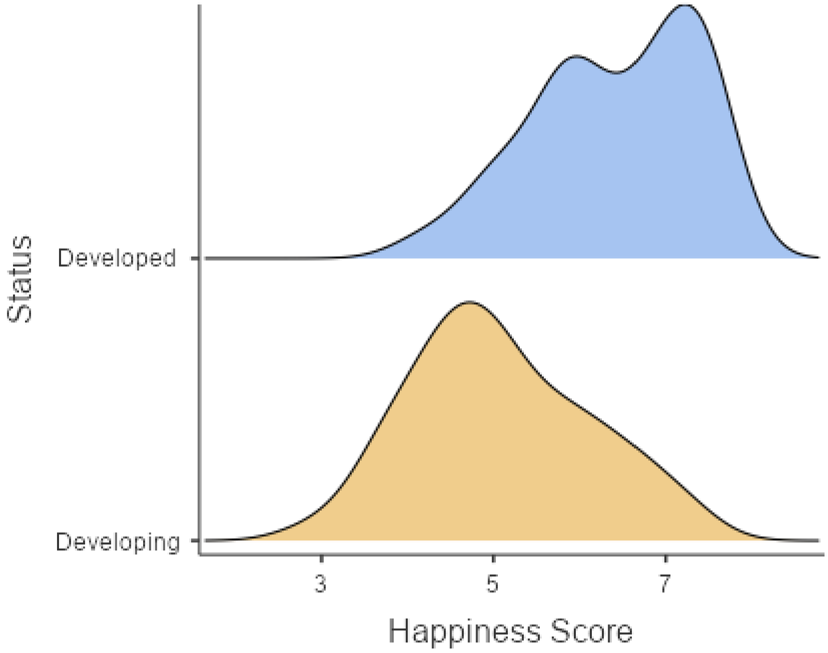
Based on figure 3 it can be determined that there are a total of 0 missing values for both developing and developed countries which indicates that cleaning/processing of the data was done correctly. It is also important to note that as mentioned above in terms of developing countries there are some outliers such as Qatar which despite being a rich country with a good life expectancy was technically not yet a developed country in 2015, the inclusion of countries like Qatar within the data means that to some extent this is the best-case scenario for developing countries.

Across developed and developing countries there is a difference of 11 years in terms of life expectancy on average. This difference from 80.4 to just 69.4 years is close to a 15% difference in life expectancy despite some of the best-case scenarios which can be found within the data such as Qatar. The difference observed between happiness is even more pronounced when looking at the median difference which is at 6.58 and 4.92 which means that the percentage difference is at a very meaningful close to 29%.



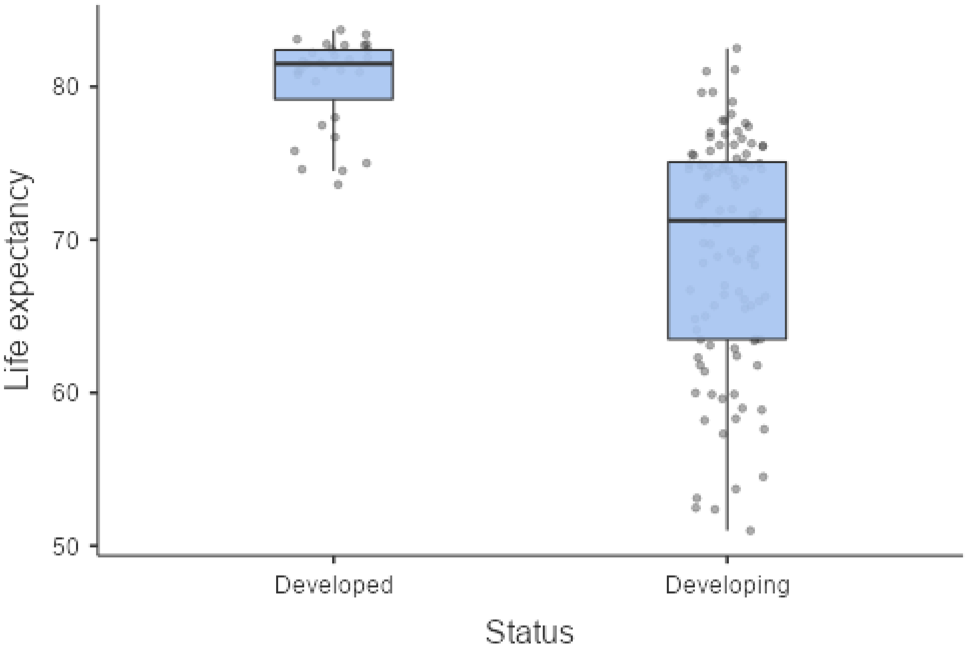
(Figure 4: Life Expectancy Density Difference Between Developed and Developing countries)

As can be seen very clearly via the visualisation in figure 4 there is an extreme difference in terms of range between developed and developing countries. In Figure 3 it can be seen that the range for developing countries is 10.1 years however the range for developing countries is at a staggering 31.5 years showing that there are without a doubt country's which are still developing who are getting close to developed countries in terms of life expectancy however there are also countries which will need decades to catch up.



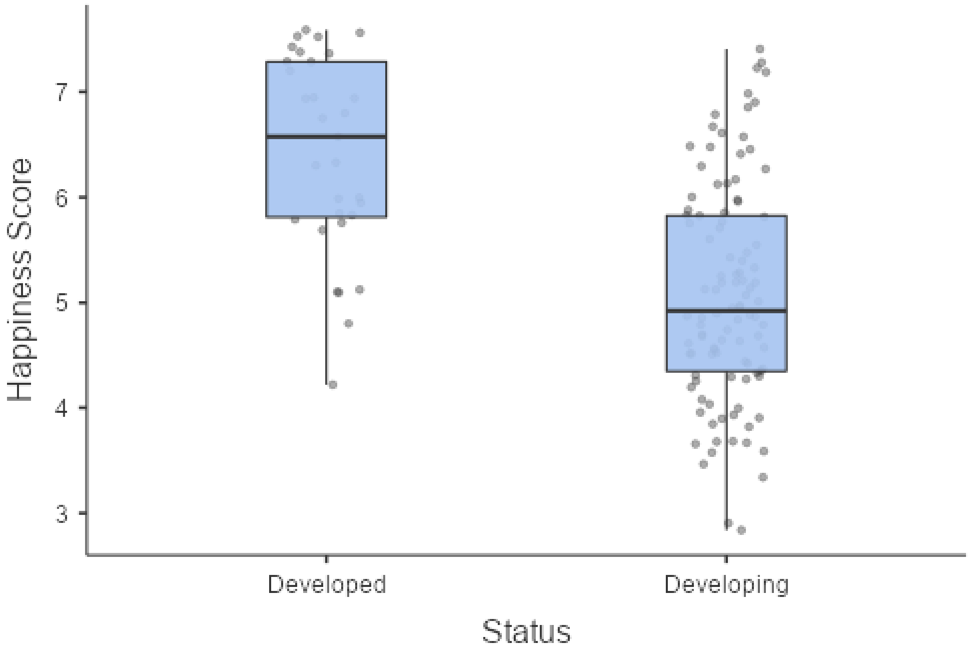
(Figure 5: Happiness Density Difference Between Developed and Developing countries)

As can be seen above in figure 5 the trends are overall quite similar with the peak for developing countries being under 5 while developed countries have a peak of density beyond 7 in terms of happiness.



(Figure 6: Life Expectancy Box Plot)

The boxplot above further demonstrates the differences between life expectancy in developed and developing countries where again it can be seen that the average differences are very substantial but especially the ranges in data are the most drastic.



(Figure 7: Happiness Box Plot)

Figure 7 above shows quite a similar pattern to the life expectancy however interestingly it shows that happiness scores in the case of developed countries tend to vary quite a lot more than they do for life expectancy which shows that interestingly enough clearly not in all cases long lives mean happy lives in these particular countries. This is quite insightful as it clearly shows that simply living long lives in certain scenarios does not guarantee being happy. However, in contrast living in developed countries the range in life expectancy is quite narrow which can be observed in figure 6.

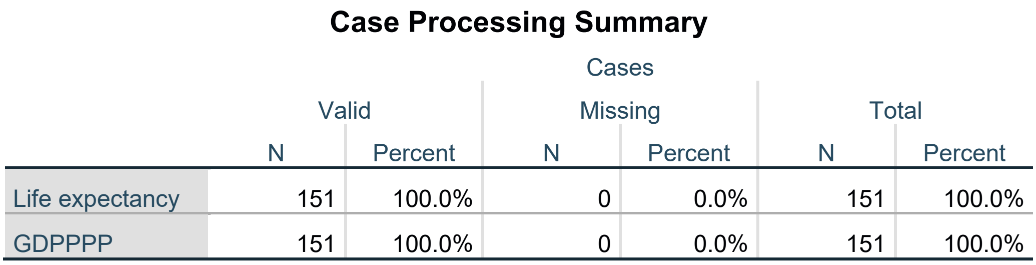
4.2.Statistical Tests for GDP (PPP) & Life Expectancy

After cleaning, processing and of course merging my data in order to understand my data and most importantly the correlations present within the data I have decided to carry out a range of statistical tests for the variables of GDP (PPP) and life expectancy.

The first test I decided to carry out was the normality test which was done through SPSS after data processing and filtering was done via RStudio which was then exported to Excel. The test was carried out in order to investigate whether or not the data which I am dealing with is normally distributed. This will be determined via tests such as the Shapiro Wilks test.

In this case the **null hypothesis** is that the values are sampled from a population which follows a normal distribution while the **alternative hypothesis** is that the data is does not follow a normal distribution. As it is standard practice for this type of test to set a significance level to 0.05 this means that if the Shapiro Wilks test result is above 0.05 in that case, I will accept the null hypothesis meaning that the data is in fact sampled from a normal distribution.

The results of the tests carried out within SPSS are as follows:

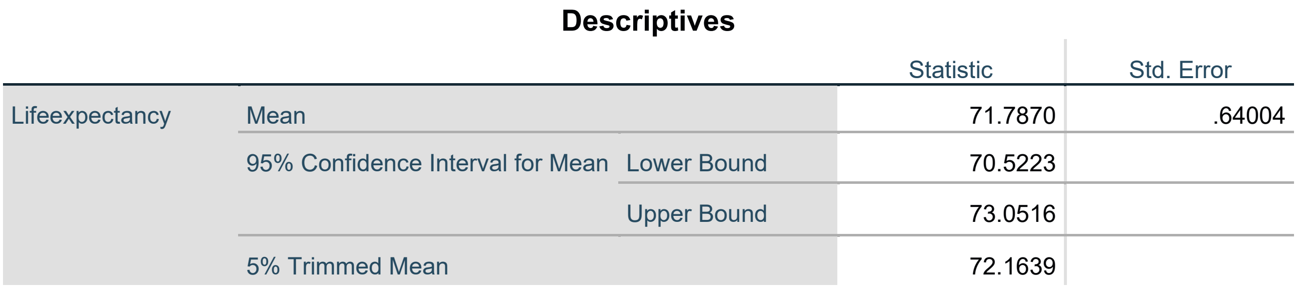


Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Life expectancy	151	100.0%	0	0.0%	151	100.0%
GDPPPP	151	100.0%	0	0.0%	151	100.0%

(Figure 8: Case Processing Summary for GDP (PPP) & Life Expectancy)

As can be seen figure 8 shows that there are no missing values which proves that the processing of the data was done correctly, there are a total of 151 rows present for the variables of GDP and life expectancy.



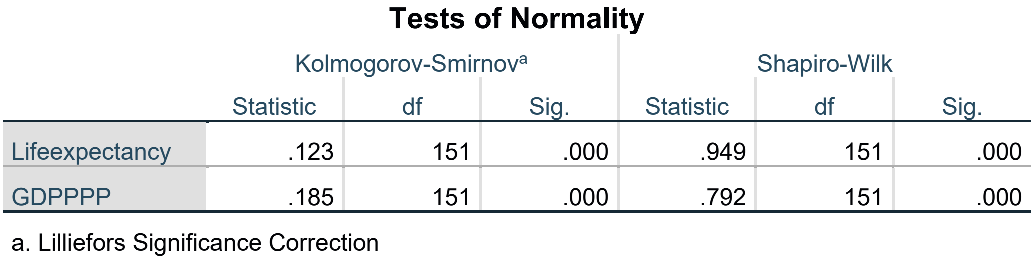
Descriptives

			Statistic	Std. Error
Lifeexpectancy	Mean		71.7870	.64004
	95% Confidence Interval for Mean	Lower Bound	70.5223	
		Upper Bound	73.0516	
	5% Trimmed Mean		72.1639	

	Median		74.0000	
	Variance		61.858	
	Std. Deviation		7.86500	
	Minimum		51.00	
	Maximum		83.70	
	Range		32.70	
	Interquartile Range		10.80	
	Skewness		-.623	.197
	Kurtosis		-.301	.392
GDPPPP	Mean		19908.481338256	1737.7321915227
	95% Confidence Interval for Mean	Lower Bound	16474.887037357	
		Upper Bound	23342.075639156	
	5% Trimmed Mean		17301.766071910	
	Median		13185.253280000	
	Variance		455976688.588	
	Std. Deviation		21353.6106686345	
	Minimum		744.7345426	
	Maximum		123822.0833000	
	Range		123077.3487574	
	Interquartile Range		22821.2270730	
	Skewness		2.003	.197
	Kurtosis		5.060	.392

(Figure 9: Descriptive Statistics for GDP (PPP) & Life Expectancy)

The table above (Figure 9) shows the descriptive statistics for both of the variables which are being examined. As the focus of this analysis is the normality in this case one of the key values to focus on is the skewness. In the case of GDP (PPP) it is at a 2.003 which shows a relatively positive skew for this variable. In the case of life expectancy, a negative skew is present of -623 which is an indication towards data being non normally distributed.



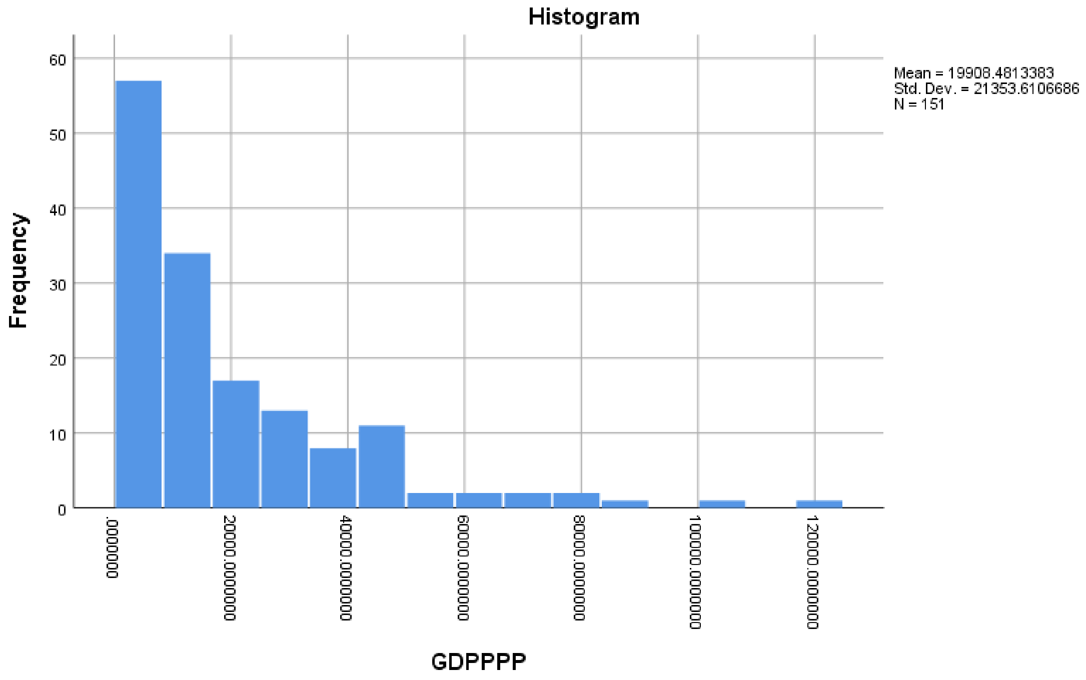
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Lifeexpectancy	.123	151	.000	.949	151	.000
GDPPPP	.185	151	.000	.792	151	.000

a. Lilliefors Significance Correction

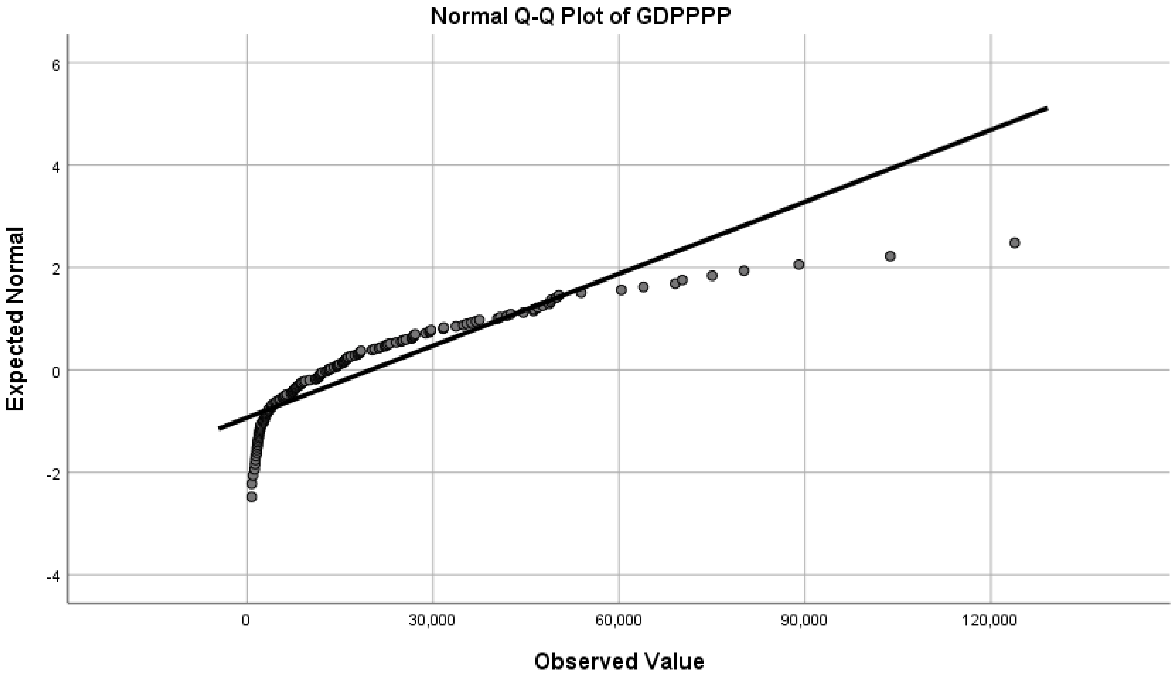
(Figure 10: Test of Normality for GDP (PPP) & Life Expectancy)

As typically the Kolmogrov-Smirinov test is more suited to very large data samples in this case it is more appropriate for me to focus on the results of the Shapiro-Wilks test. As can be seen in figure 10, in the case of both GDP and life expectancy variables the level of significance is below the alpha value of 0.05 which is a clear indication that both variables are not normally distributed.



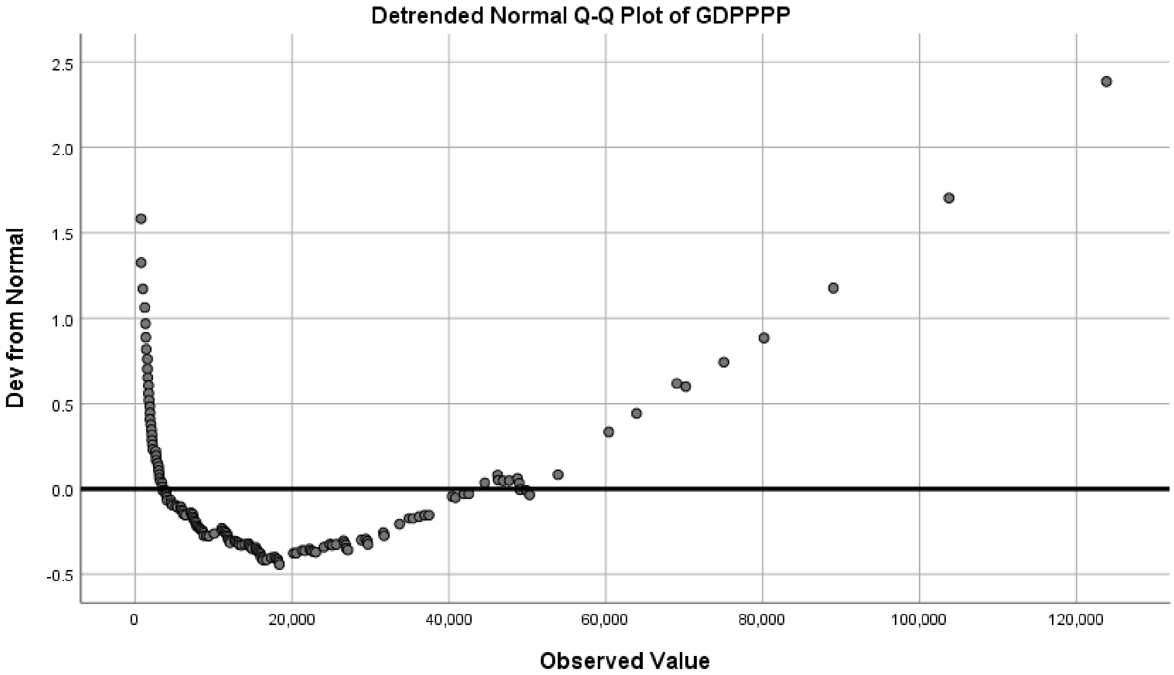
(Figure 11: Histogram showing the distribution of GDP PPP)

Figure 11 is a very clear indicator that the data is in fact not even close to being normally distributed. This can be seen clearly based on the histogram which is not even close to resembling a standard bell curve which would suggest normally distributed data.



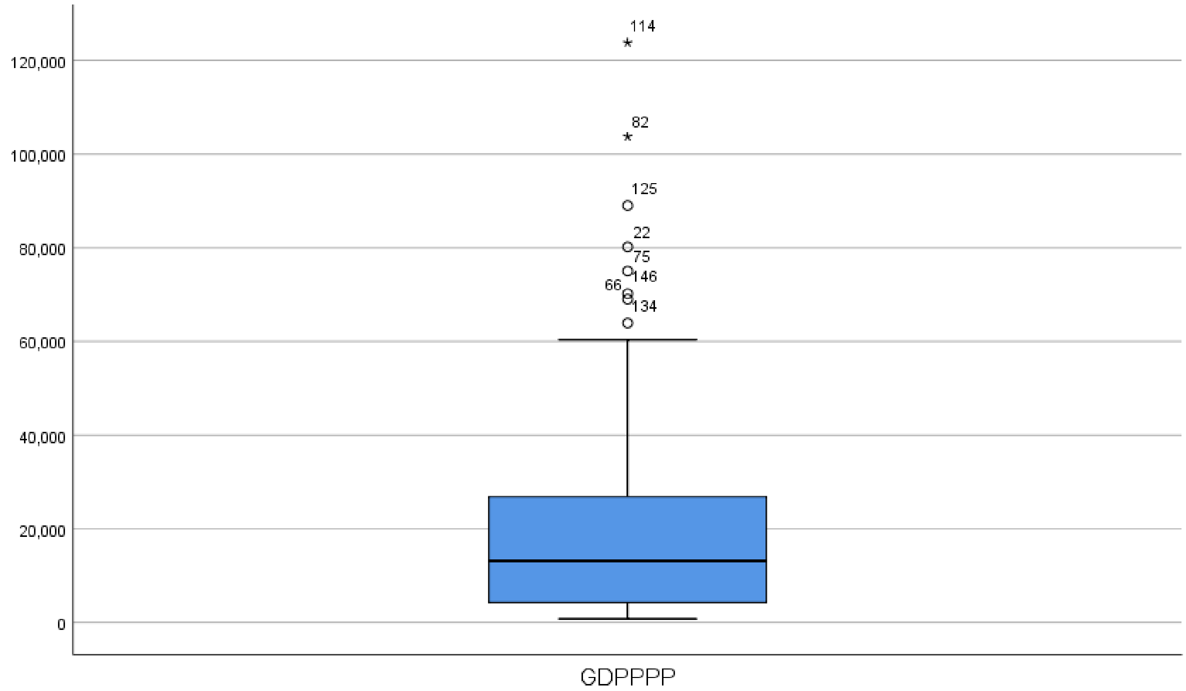
(Figure 12: Q-Q plot of GDP PPP)

The above plot is another visual representation of the data being non normally distributed, this is so as not many of the values follow the line. If the values did follow the line, it would show that the data is normally distributed, in this case however it is clear that the data is not normally distributed.



(Figure 13: Detrended Q-Q plot of GDP PPP)

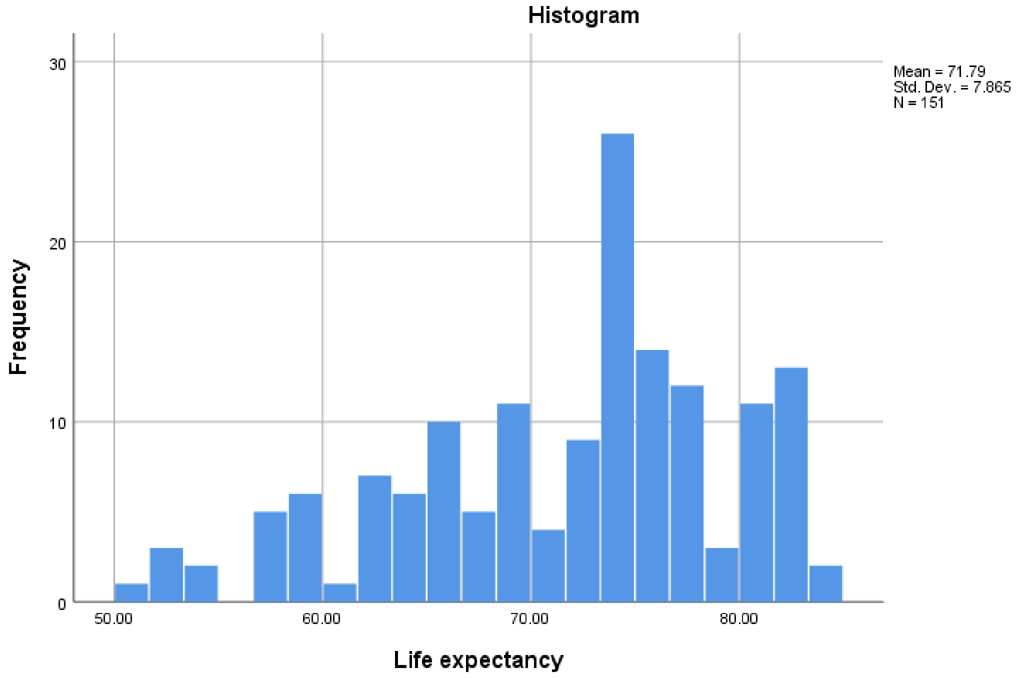
The detrended plot very much like the previous tests also confirms that the data is not normally distributed as there is great deviation from the norms within the data.



(Figure 14: Box and Whisker Plot of GDP PPP)

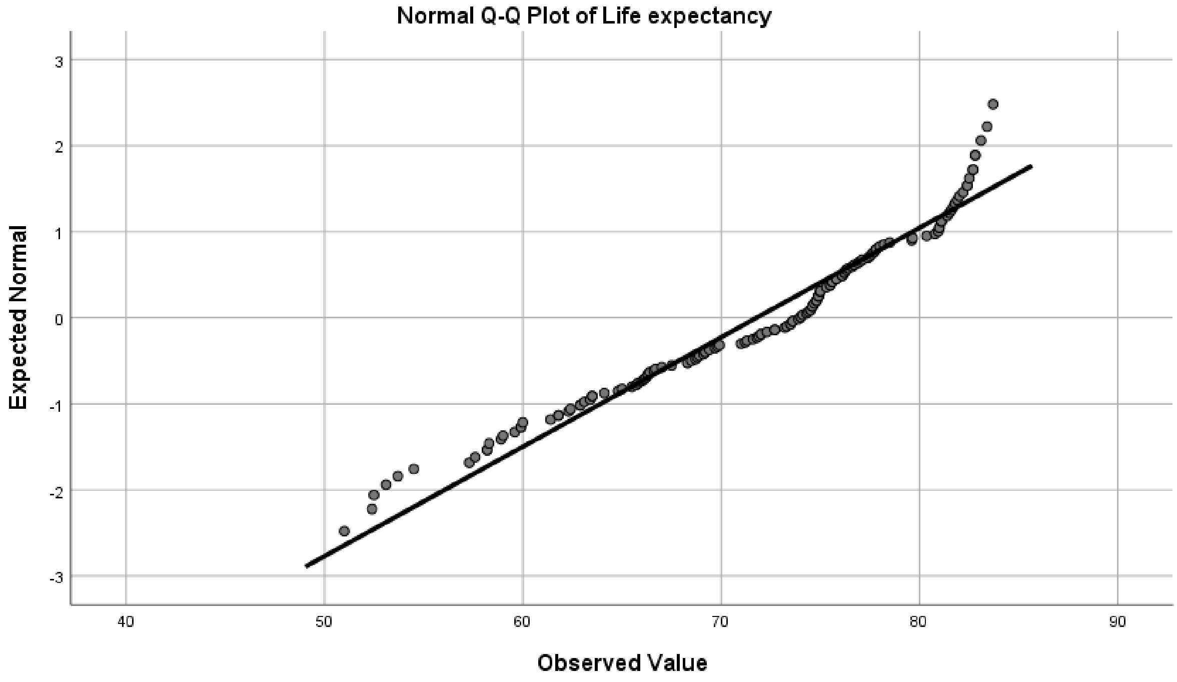
As can be seen above the box plot is very far from the centre with a vast number of outliers present in the data which are quite far from the mean which is yet another and the final indication towards data which is far from being normally distributed.

However, when looking at the plots for Life Expectancy it becomes clear that the data is closer to being normally distributed however it still does not meet the requirements to be considered normally distributed data.



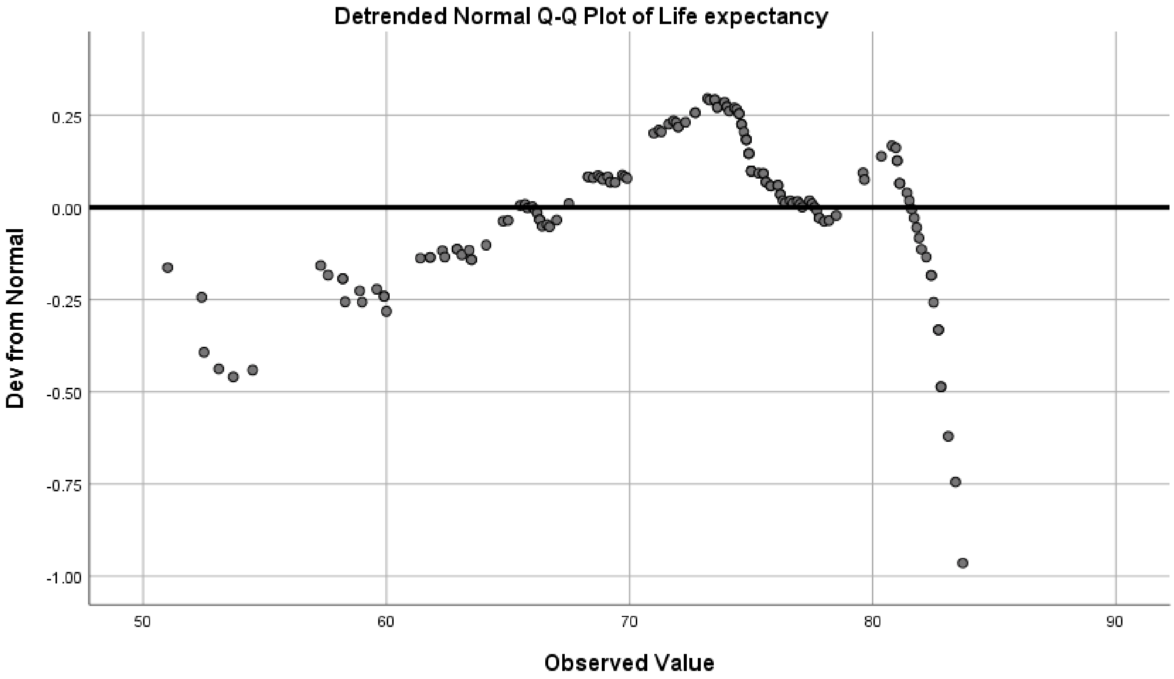
(Figure 15: Distribution of Life Expectancy)

In the case of life expectancy, it is clear that the data is closer to match the bell curve than the data for GDP however it is still not close enough to be considered normally distributed data.



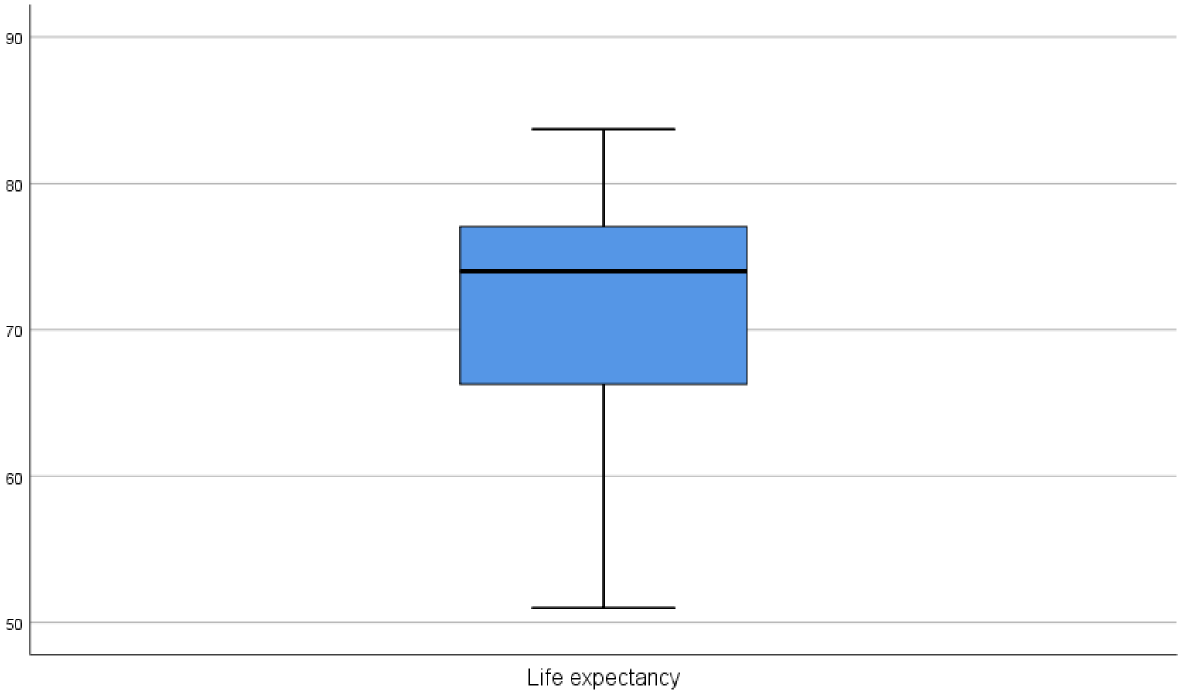
(Figure 16: Q-Q plot of Life Expectancy)

In figure 16 the trend is set by once again the visualisation showing data which is considerably closer to being normally distributed than the data for GDP however it is still evident that life expectancy is also not normally distributed.



(Figure 17: Detrended Q-Q plot of Life Expectancy)

As can be seen above in the detrended Q-Q plot the of Life Expectancy figure 17 the variations are more pronounced than in Figure 16 which shows that there is in fact a significant amount of variation showing clearly that the data is definitely not normally distributed.



(Figure 18: Box Plot of Life Expectancy)

The box plot in figure 18 is very meaningful to show the lack of extreme variables which were present in the case of GDP in figure 14 however as the line signifying the median which is the black line going through the blue box is relatively close to the middle of the graph especially when compared to figure 14 however it is not in the middle further showing that the life expectancy data is not normally distributed.

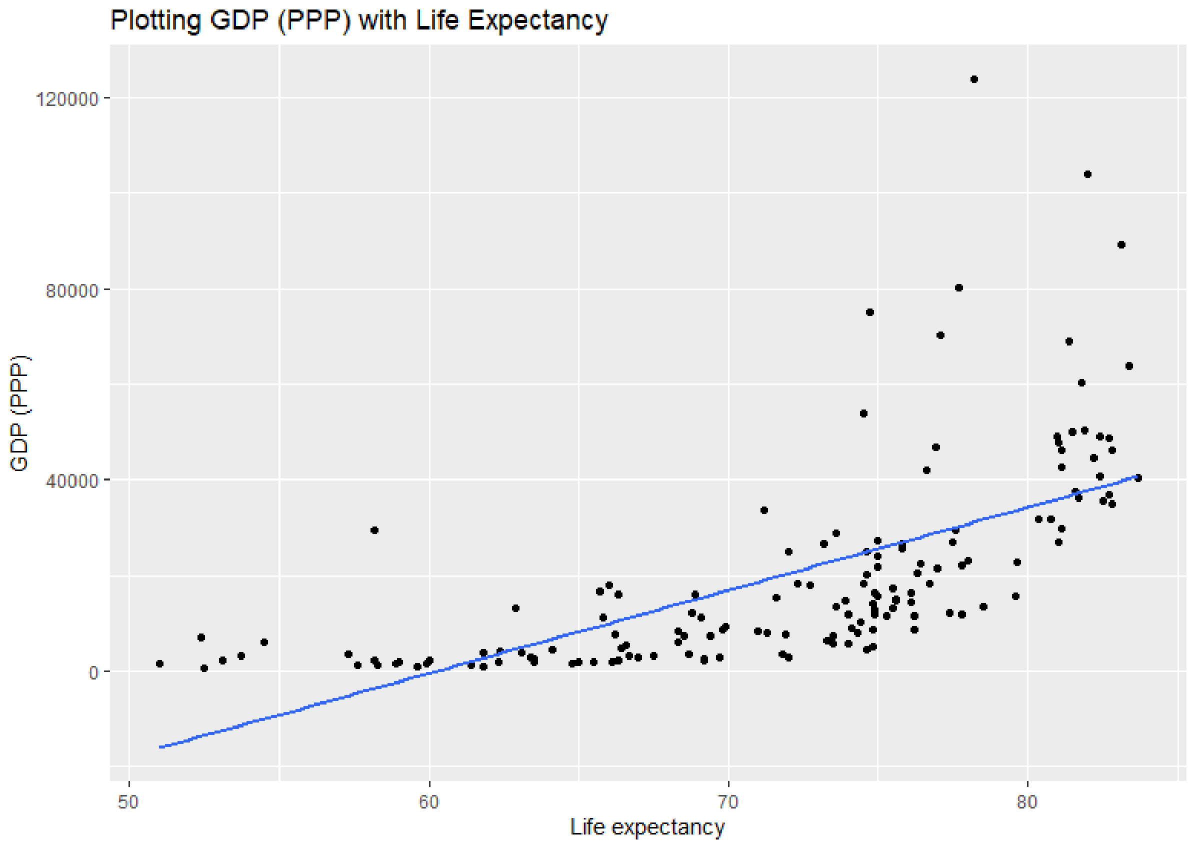
4.3. Linear regressions

In order to analyse in more detail, the effects of each of the variables in my data after performing multiple merges, followed by data cleaning and transformations I have decided to carry out linear regressions to show the correlations between various different factors and life expectancy as well as happiness levels around the world.

As the linear regression is the best way to investigate the relationship between variables, this is one of the best ways to find insights as to what kind of factors play the biggest role in deciding life expectancy and happiness. By utilising R studios ggplot package I was able to code various different graphs to visually represent the correlations between all of the various factors examined below.

4.4. GDP PPP vs Life Expectancy

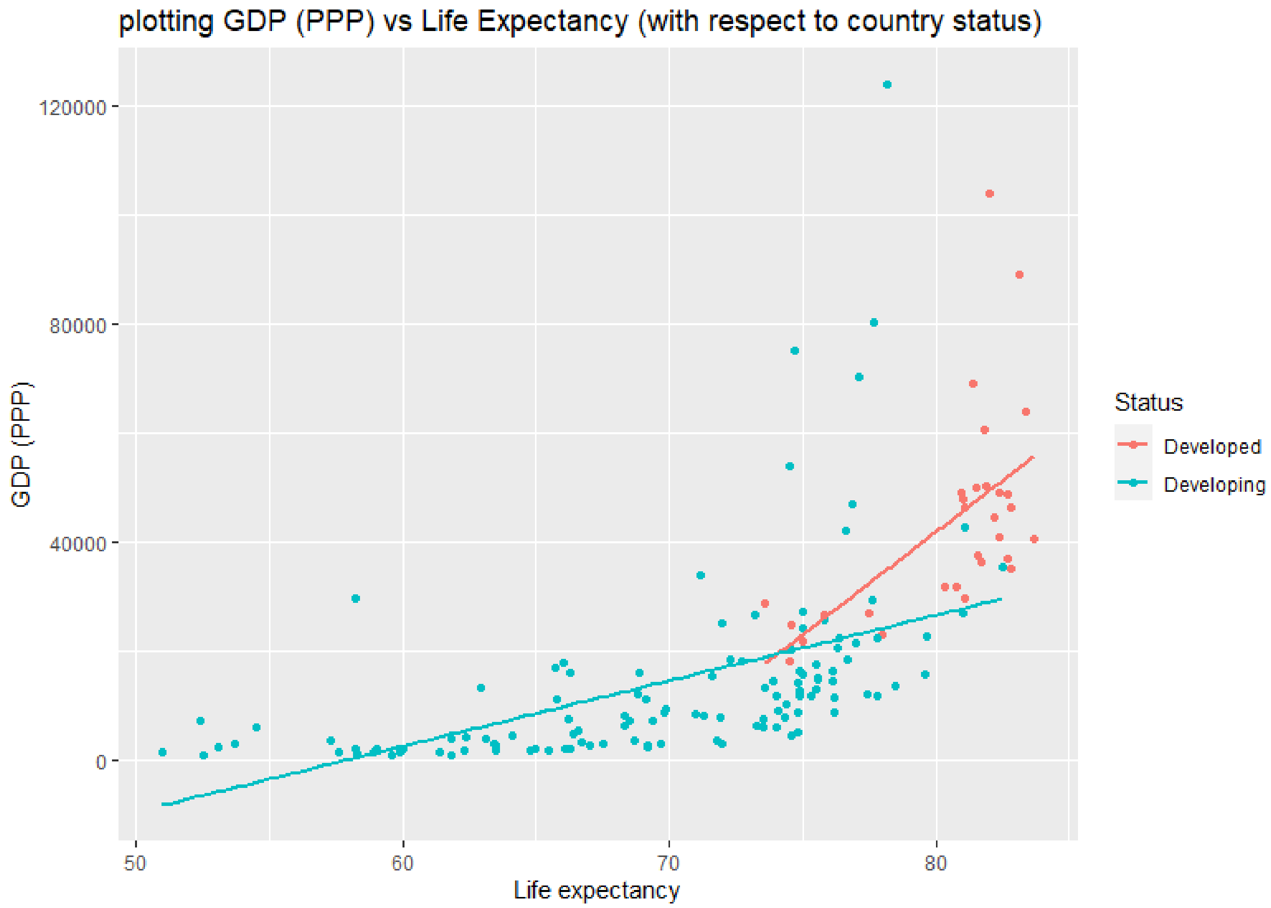
As money and wealth has a huge impact on peoples lives, I wanted to examine the correlation between GDP and Life Expectancy.



(Figure19: Life Expectancy vs GDP)

As can be seen above there is a very clear correlation present between life expectancy and GDP. It can be seen that money/wealth definitely does buy a longer life however it can be seen quite clearly that there are some exceptions where some countries scored below 3000 in terms of GDP PPP while almost achieving a life expectancy of close to 80 years which is hugely impressive and will be looked into in more detail in the results section part 6.

Based on figure 19 it is clear that GDP may in fact be the biggest determining factor for a longer life expectancy however there may be some factors even more determining which can be seen below in figure 20.

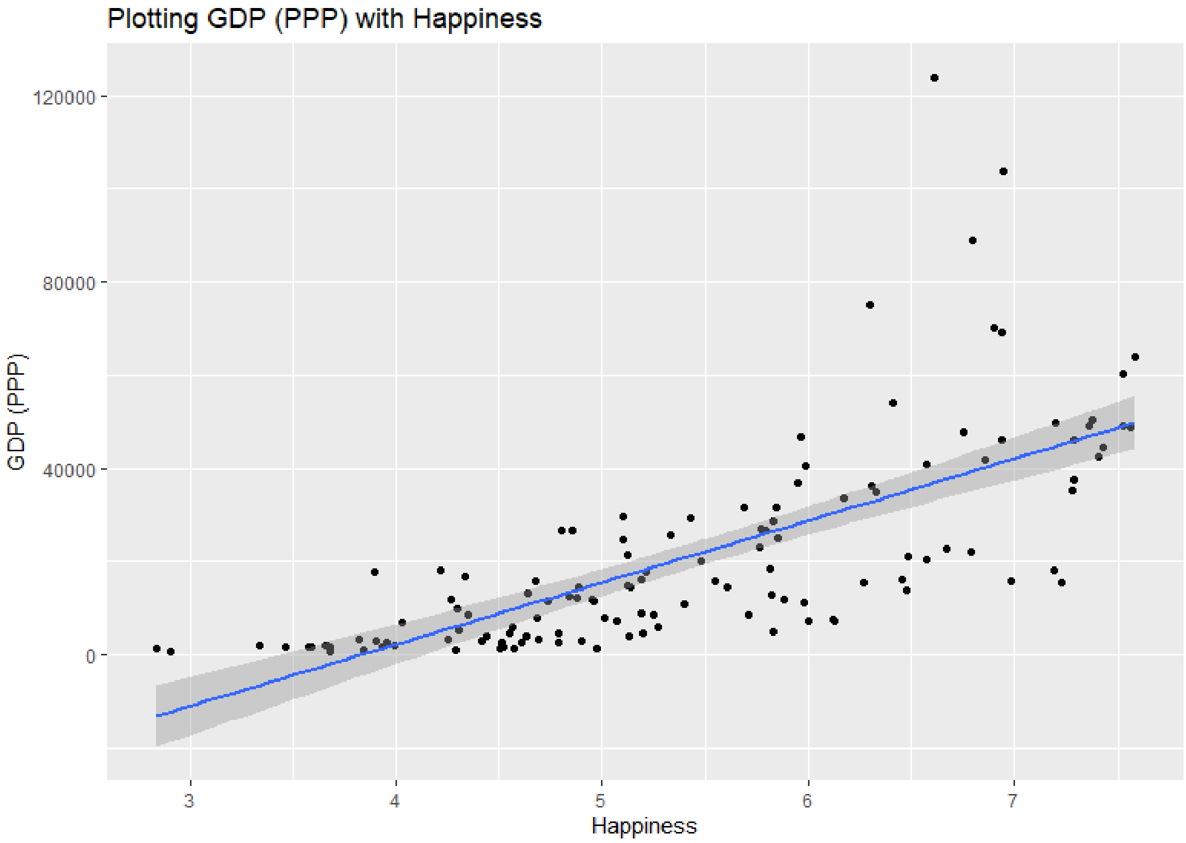


(Figure 20: Life Expectancy vs GDP (with respect to status))

Based on the above graph it can be seen that the line difference between developing and developed countries is very differently skewed which shows the huge extent to which life expectancy depends on the status of a given country especially considering that GDP has a major effect on whether a country is developing or developed.

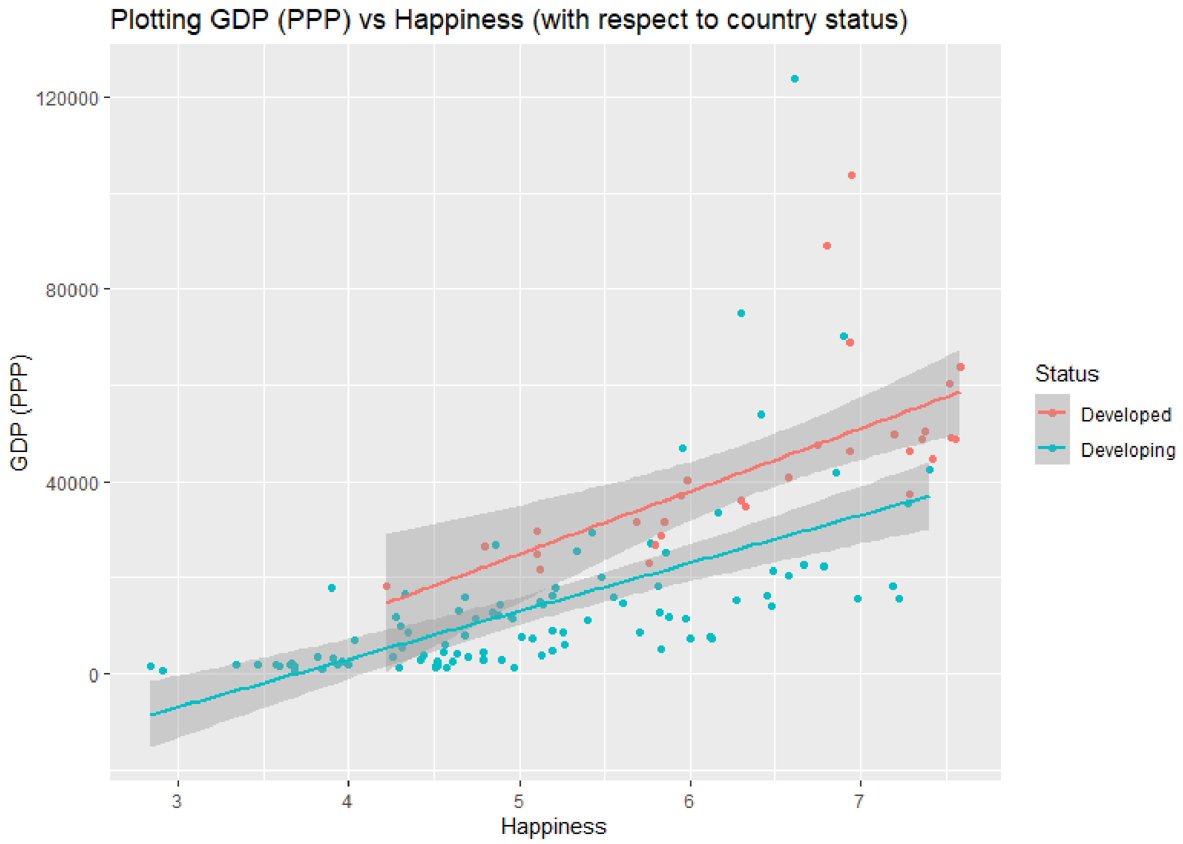
As mentioned above however it is very important to remember that simply having a very high GDP does not make a country be automatically a developed country as could be seen in the example of Qatar explained in section 4.1. However, it is clear that the majority of developed countries are very close to the right side on the x axis of figure 20 showing a very high life expectancy.

4.5. Happiness vs GDP PPP



(Figure 21: Happiness with GDP PPP)

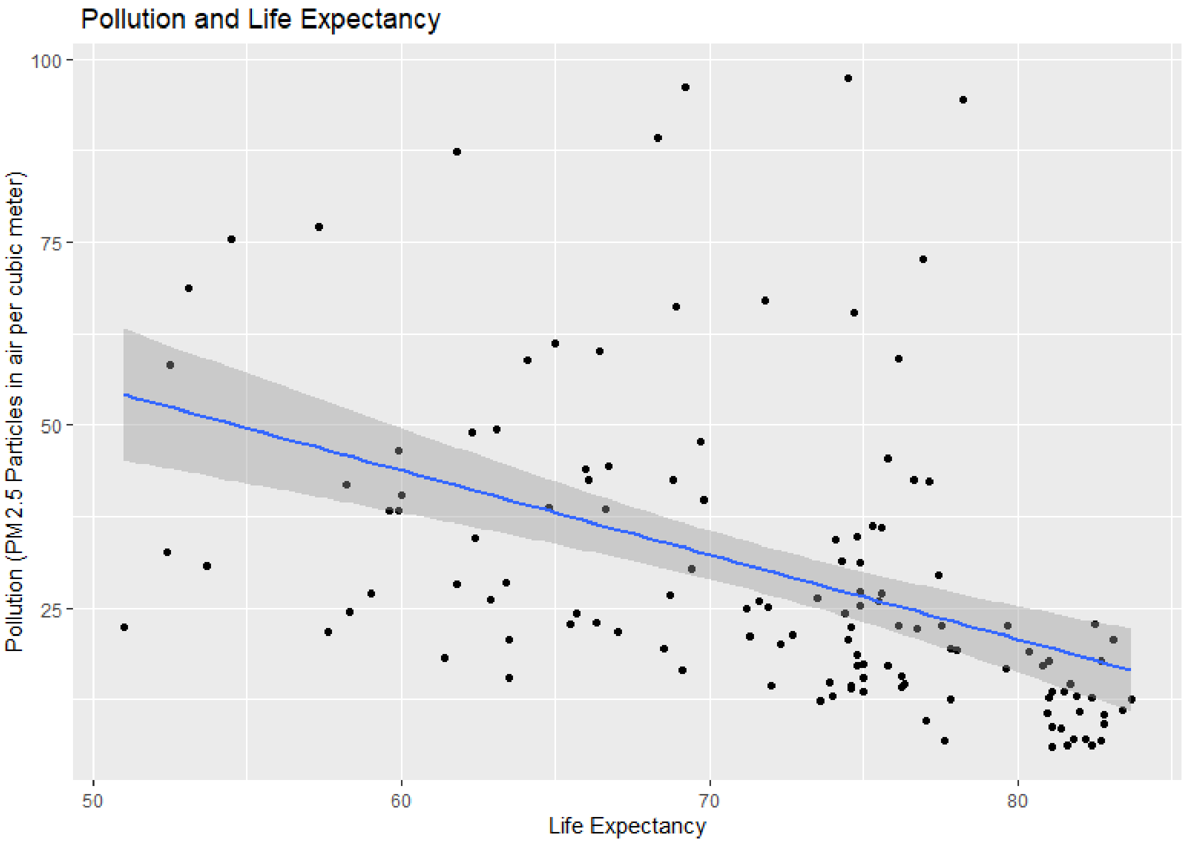
The graph above shows me that quite surprisingly the deviations in terms of happiness are quite significant in some instances. It can be seen in figure 21 that some countries even scoring above 6000 in terms of GDP PPP were less happy than a large number of countries with a GDP PPP of less than around 5500 which is very interesting to me. It is however clear that the deviations become less pronounced the closer we look to the left on the x axis meaning that when the GDP PPP is quite low the deviations are smaller as happiness scores are also lower. However, when looking on the right it is very clear that deviations (from the grey shadow around the blue line which represents the 95% confidence interval) are far more pronounced suggesting that money does not buy happiness to the extent I believed before carrying out the analysis which will be expanded upon in the results section.



(Figure 22: Happiness with GDP PPP with respect to status)

As can be seen in the above graph there are several countries which are still developing yet seemingly fully capable of housing happy citizens even despite being below the 95% confidence interval in terms of GDP which once again shoes that to an extent money does not buy happiness beyond a certain point which will need to be discussed in the results section.

4.6. Air pollution's effect on life expectancy

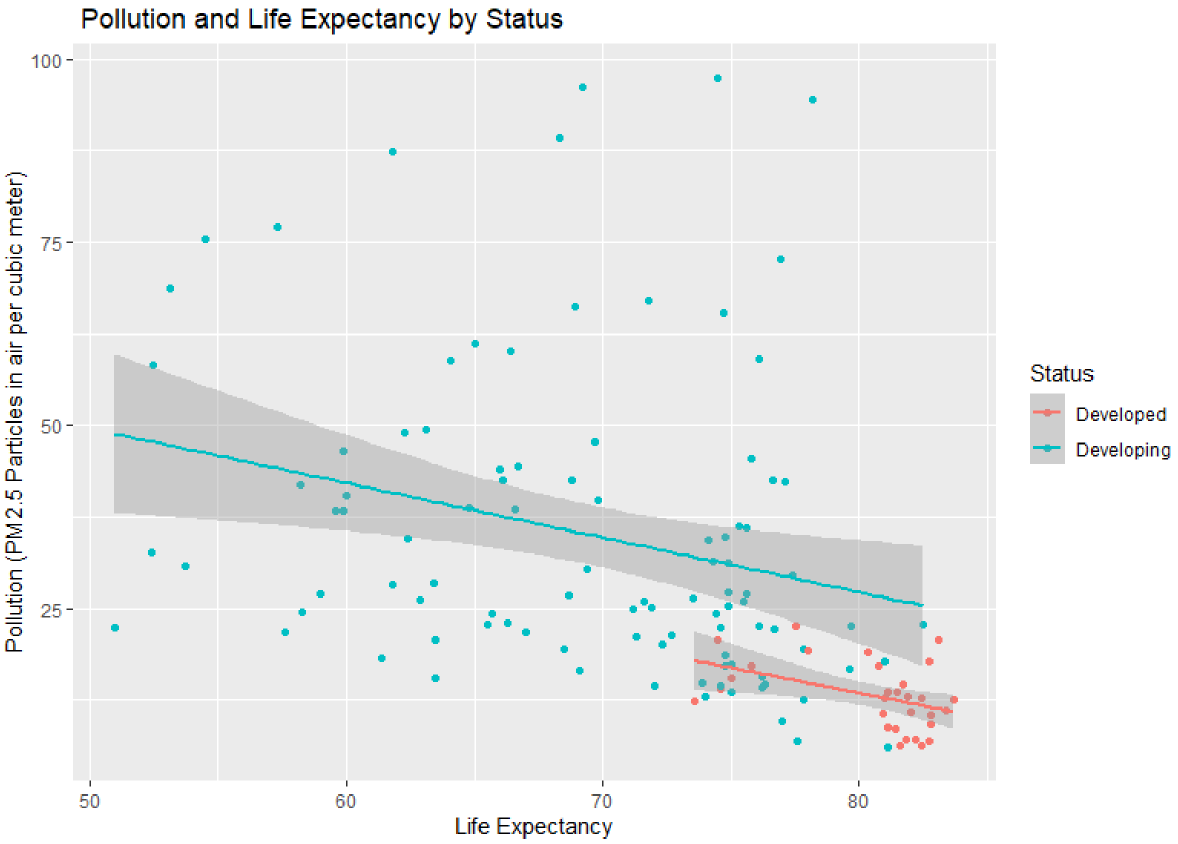


(Figure 23: Pollution levels with Life Expectancy graph)

The above figure 23 which visualises the correlation between the life expectancy and air pollution around the world is perhaps the most interesting insight so far. It is clear that many countries are doing their best in terms of moving towards green and eco friendly solutions for needs such as electricity which based on figure 23 is an absolute must if the country aspires to reach a life expectancy of above 80 years old.

Based on the visualisation from figure 23 it is very interesting to see that there is no country in the world with a life expectancy of 80 or over which contains more than 25 PM 2.5 particles in the air for every cubic meter in average per annum. Very interestingly there appears to be somewhat of a barrier to getting a life expectancy over 80 without taking care of the air people breath. This is clear as every country with a life expectancy of 80 years or more falls within or below the 95% confidence interval which shows that a score of below 25 PM 2.5 particles per cubic meter is a must in order to obtain a very high life expectancy.

It is also however very interesting to observe that for life expectancies just below 80 there is a very wide range of variation in terms of air quality which suggests that air quality is an extremely meaningful factor in influencing life expectancy however this is so only for life expectancies over 80. In the case of expectancies below 80 it can be seen that air quality plays less of a factor which is something which will also be discussed in the results section.



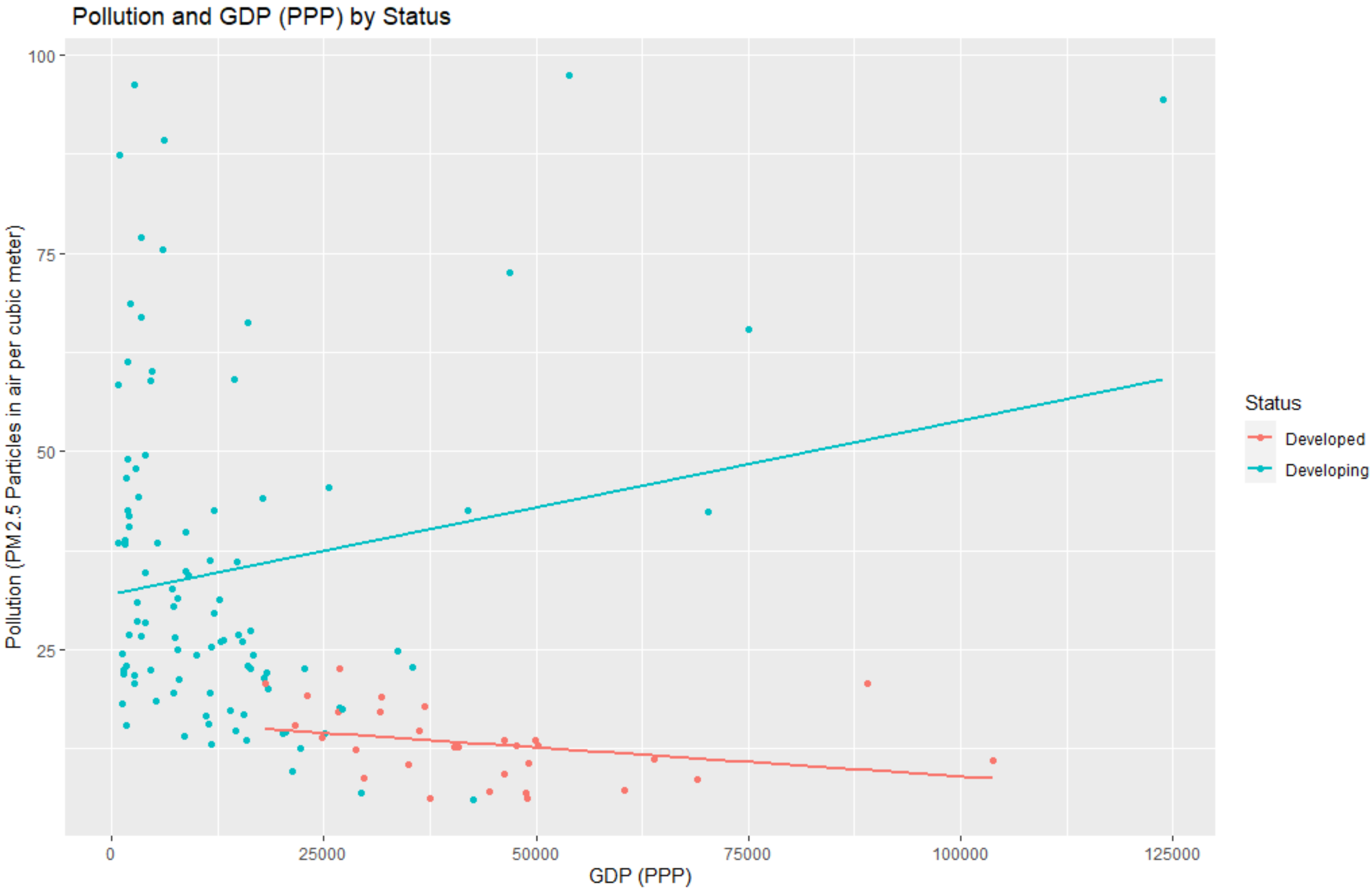
(Figure 24: Pollution levels with Life Expectancy by Status)

The visualisation above shows an extra layer on the air quality and life expectancy correlation around the world. In Figure 24 it can be observed that developed countries all fall below that threshold of 25 PM 2.5 particles per cubic meter with only developing countries exceeding the threshold established from figure 23. It is however important to notice that some developing countries above the threshold of 25 PM 2.5 particles per cubic meter are able to compete against the developed countries which have the lowest life expectancy but of course as mentioned previously none have gone beyond the threshold of 25 while maintaining a life expectancy above 80 which is crucial.

It is also worth taking note of the fact that developing countries contain much more variation in their air quality than developing countries.

4.7. GDP PPP and Air Pollution

After being so intrigued by the insights gathered from figures 24 and 23 in section 4.6, I had to investigate further using data to find answers and explanations to what I had found. In order to look for more insights I decided to do another linear regression, this time based on GDP PPP and air pollution to understand the relationship between these key factors.



(Figure 25: Pollution levels with GDP PPP by Status)

The visualisation above answers many questions I have had after the analysis of part 4.6. The visualisation (figure 25) shows that the trend between developed and developing countries is in fact reversed which can be seen by the blue and red lines going in the opposite direction.

Figure 25 shows clearly that in the scenario where a developed country has a higher GDP (PPP) the countries air pollution is being brought down and down as scene by the red line in figure 25. This indicates that countries with higher GDP's are capable of reducing harmful emissions which also extends life expectancy.

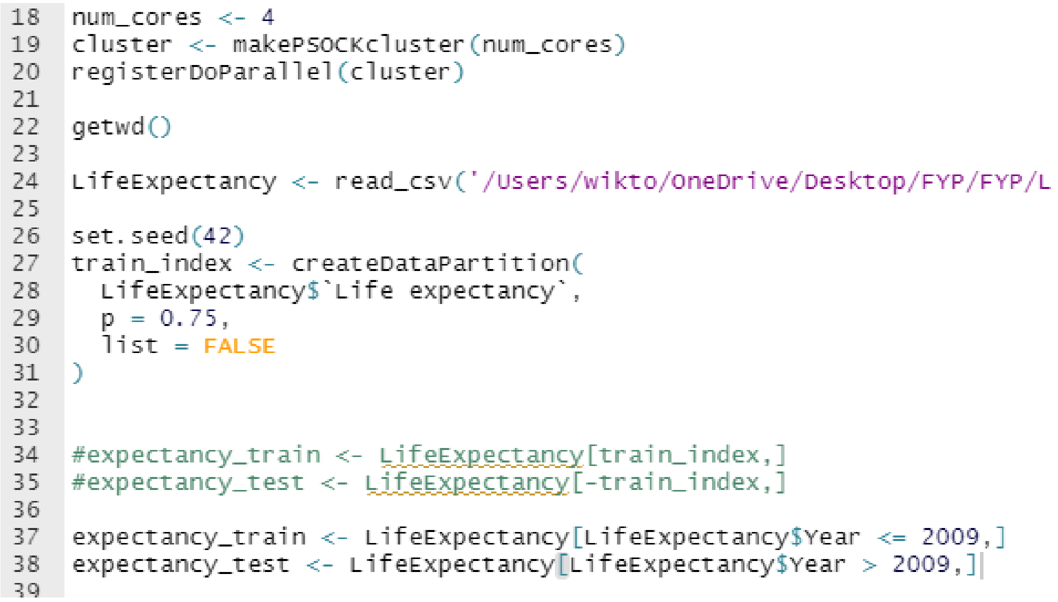
As it could be seen in figure 24 the range between developing countries in terms of air pollution are very broad which even suggests that until a country is reaching a life expectancy of 80 the air quality doesn't play too much of a role. However, when trying to exceed a life expectancy of 80 there is seemingly a barrier present in the form of air quality as a threshold of 25 PM 2.5 particles per cubic meter can not be crossed. There is however a clear correlation between developing countries with a relatively high GDP having poor air quality which will be further discussed in the results section.

4.8. Model trees and SVR

As one of my goals was to study the change in life expectancy over time, I had decided that it was important create a model based around prediction of life expectancy. As I have made an animation which visually shows the change if life expectancy over time using tableau (which will be discussed in more detail in the results section) I believe that the natural next step was to create models which can potentially predict the future. The models do this by taking a fragment on the dataset and purely learning the correlations from that training segment which can then be compared with the existing data which was not trained on.

The results from the training can then be compared to the actual pre-existing data which is in the dataset. The results from the training data can be then compared to the data which is present therefore showing the accuracy of the model. It is important to note here that the in order to validate the result I can not go beyond the year I have which is 2015 therefore I can only predict up to 2015 as only up to the year 2015 can be verified in terms of accuracy. This is of course the same in the case of SVR which is Support Vector Regression.

As the models can take quite some time to be built it was important to also analyse how long each model takes to finish. It is also important to note that this is hugely dependant on the quality of the CPU. In my case I was able to conduct the analysis using an AMD Ryzen 5 5600x which is a very modern CPU with 6 cores and 12 threads which allowed for a relatively quick completion of the models.

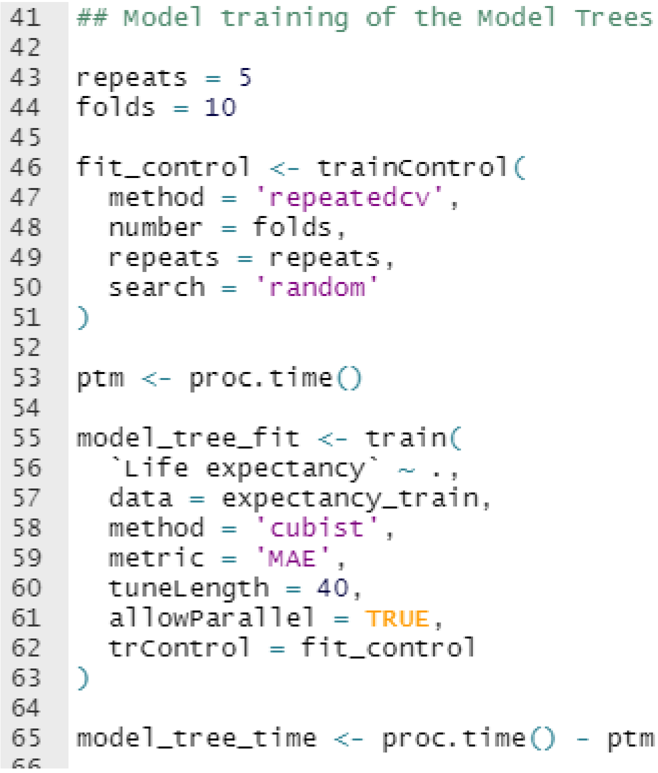


```
num_cores <- 4
cluster <- makePSOCKcluster(num_cores)
registerDoParallel(cluster)

getwd()

LifeExpectancy <- read_csv('/Users/wikto/OneDrive/Desktop/FYP/FYP/L

set.seed(42)
train_index <- createDataPartition(
  LifeExpectancy$`Life expectancy`,
  p = 0.75,
  list = FALSE
)


#expectancy_train <- LifeExpectancy[train_index,]
#expectancy_test <- LifeExpectancy[-train_index,]

expectancy_train <- LifeExpectancy[LifeExpectancy$Year <= 2009,]
expectancy_test <- LifeExpectancy[LifeExpectancy$Year > 2009,]
```

(Figure 26: Model trees code snippet)

As can be seen in figure 26 it was very important on line 18 to set the number of cores. This is so as allowing R studio to utilise all 6 cores at 100% could even result in a system crash or at least potentially a major slow down of the entire system as important background processes may not work accordingly.

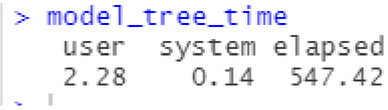
Line 27 is also vital as it is the line of code which creates the training and testing partitions which is used to compare the results to the actual available data and the prediction based on the training data. Setting the seed is also important however the number which is chosen does not matter however it is vital to keep it consistent when trying to recreate the results.



```r
## Model training of the Model Trees

repeats = 5
folds = 10

fit_control <- trainControl(
  method = 'repeatedcv',
  number = folds,
  repeats = repeats,
  search = 'random'
)

ptm <- proc.time()

model_tree_fit <- train(
  `Life expectancy` ~ .,
  data = expectancy_train,
  method = 'cubist',
  metric = 'MAE',
  tuneLength = 40,
  allowParallel = TRUE,
  trControl = fit_control
)

model_tree_time <- proc.time() - ptm
```

(Figure 27: Model trees code folds and repeats)

The number of folds and repeats is also very important for models as they can be a big determining factor in the accuracy of the prediction. It is however important to note that the higher the repeats and folds the longer it will take to complete the model which can happen in an exponential fashion in terms of time. It is also important to note that even if the number of repeats and folds were to be doubled it would not result in double the accuracy however it would most likely result in more than double the time to complete therefore diminishing returns is present.



```
> model_tree_time
   user  system elapsed 
   2.28    0.14  547.42 
```

(Figure 28: Model trees code snippet time)

As can be seen in figure 28 the total elapsed time for the model to be completed was 547 seconds which is just over 9 minutes.

The code for SVR was structured in a very similar fashion in order to ensure that the comparison is fair. Upon running both of the models it was important to understand which is more accurate and to analyse that model in more detail which is what will be covered in the results section.

5.0 Testing

Test Case 1			
Name	Tableau Testing		
Result	Pass	**Date of Test**	3/07/21

Test ID	**1**
Purpose of Test	To Ensure that: It is possible to use the data for performing a animation style visualisation showing life expectancy change over several decades
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. Tableau public 2020.4
Method	After creating the animation in tableau, it was important to verify that several randomly chosen countries had the correct life expectancy for the correct year.
Expected Result	Upon completing the animation, the verification of the results for countries life expectancy and year matches the original dataset correctly.
Actual result	The dataset matched the visual output as expected and no adjustments were needed.
Comments	N/A
Resolution	N/A

Test Case 2			
Name	R Studio Functionality Testing		
Result	Pass	**Date of Test**	3/07/21

Test ID	**2**
Purpose of Test	To Ensure that: R studio is functioning as expected with the ability to set the working directory and read the required data correctly and of course to ensure that analysis can be performed.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. R Studio.
Method	Checks performed to compare the data which was loaded in with the original data which will be used for the analysis. Also running basic trial tests to ensure everything is setup correctly.
Expected Result	Data is loaded in correctly and the trial analysis yields expected results.
Actual result	The data loads in correctly without any issues and R Studio functions as required.
Comments	N/A
Resolution	N/A

Test Case 3			
Name	SPSS Functionality Testing		
Result	Pass	**Date of Test**	3/07/21

Test ID	**3**
Purpose of Test	To Ensure that: Loading of processed data extracted from R into SPSS to ensure formatting of data is correct therefore analysis within SPSS is accurate.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. IBM SPSS Statistics 26
Method	After processing the data using R including merging the WHO dataset with the happiness scores dataset and cleaning the data it was important to ensure that the data can be extracted correctly for further analysis in other software (in this case SPSS). By importing the dataset into SPSS and validating the data was structured correctly I was able to arrive at a result.
Expected Result	Upon importing the dataset into SPSS the data is structured appropriately and ready for further analysis.
Actual result	The data loads in correctly without any issues.
Comments	N/A
Resolution	N/A

Test Case 4			
Name	R Studio missing values		
Result	Pass	**Date of Test**	3/07/21

Test ID	**4**
Purpose of Test	To Ensure that: There are no null values present in the dataset.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. R studio. WHO and Happiness datasets.
Method	After processing the data using R including merging the WHO dataset with Happiness dataset it was important to carryout analysis to view the number of missing values.
Expected Result	The missing values are displayed and can be dealt with accordingly.
Actual result	All missing values are presented in a clear way and dealt with by either omitting them and or replacing the null values with the average of that particular column.
Comments	N/A
Resolution	N/A

Test Case 5			
Name	Validating results across different software		
Result	Pass	**Date of Test**	3/07/21

Test ID	**5**
Purpose of Test	To Ensure that: When performing analysis such as descriptive statistics the results are consistent throughout various software such as Excel, RStudio, Jamovi and SPSS.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. R studio, Excel, Jamovi 1.6.21 and SPSS 26
Method	After performing analysis using R studio I then exported the data to the above mentioned software's. The same tests were then carried out across all of the software to validate the results.
Expected Result	The results of the tests are fully consistent between all software used.
Actual result	All outputs are exactly consistent with one another meaning that the analysis is correct across all sofrware used.
Comments	N/A
Resolution	N/A

Test Case 6			
Name	Testing model trees and SVR.		
Result	Pass	**Date of Test**	3/07/21

Test ID	**6**
Purpose of Test	To Ensure that: The models which have been created (model trees and SVR) are performing well and crucially comparing them to each other to decide which performs better in terms of accuracy.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. R studio.
Method	By analysing the outputs from the model trees and VSR it was possible to decide which model performs better.
Expected Result	Ideally both models perform well however one will be better than the other.
Actual result	Model trees performs slightly better than VSR meaning that it is more accurate.
Comments	N/A
Resolution	N/A

Test Case 7			
Name	Tableau hosted online (Animation)		
Result	Pass	**Date of Test**	3/07/21

Test ID	**7**
Purpose of Test	To Ensure that: The visualisation created have the capability to be hosted online and are accessible to users.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. Tableau 2020.4
Method	conduct a check to ensure that the visualisation was present and accurate when hosted online.
Expected Result	The animated visualisation works smoothly without issues as it does offline.
Actual result	Animation works as intended online.
Comments	N/A
Resolution	N/A

Test Case 8			
Name	Tableau hosted online (Map)		
Result	Pass	**Date of Test**	3/07/21

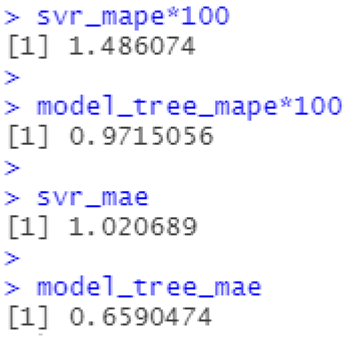
Test ID	**8**
Purpose of Test	To Ensure that: The map visualisation created is accessible online.
Test Environment	The test environment was as follows: Client Hardware: Ryzen 5 5600x, RTX 3070, 16 gb RAM. Tableau 2020.4
Method	conduct a check to ensure that the map visualisation is accessible online.
Expected Result	The map is accessible online and functions as it does offline.
Actual result	The map appears the same it does offline without any discrepancies.
Comments	N/A
Resolution	N/A

6.0 Results

Model Trees and SVR

I am beginning the results section with the two models which can be used in order to predict the data based on the training model. As the two models were used for the same purpose testing was important to understand which model performed better.

One of the best ways to analyse which model performed better is to analyse the errors by utilising mape and mae which can be used to show the number of errors found between the prediction and the actual data in the dataset.

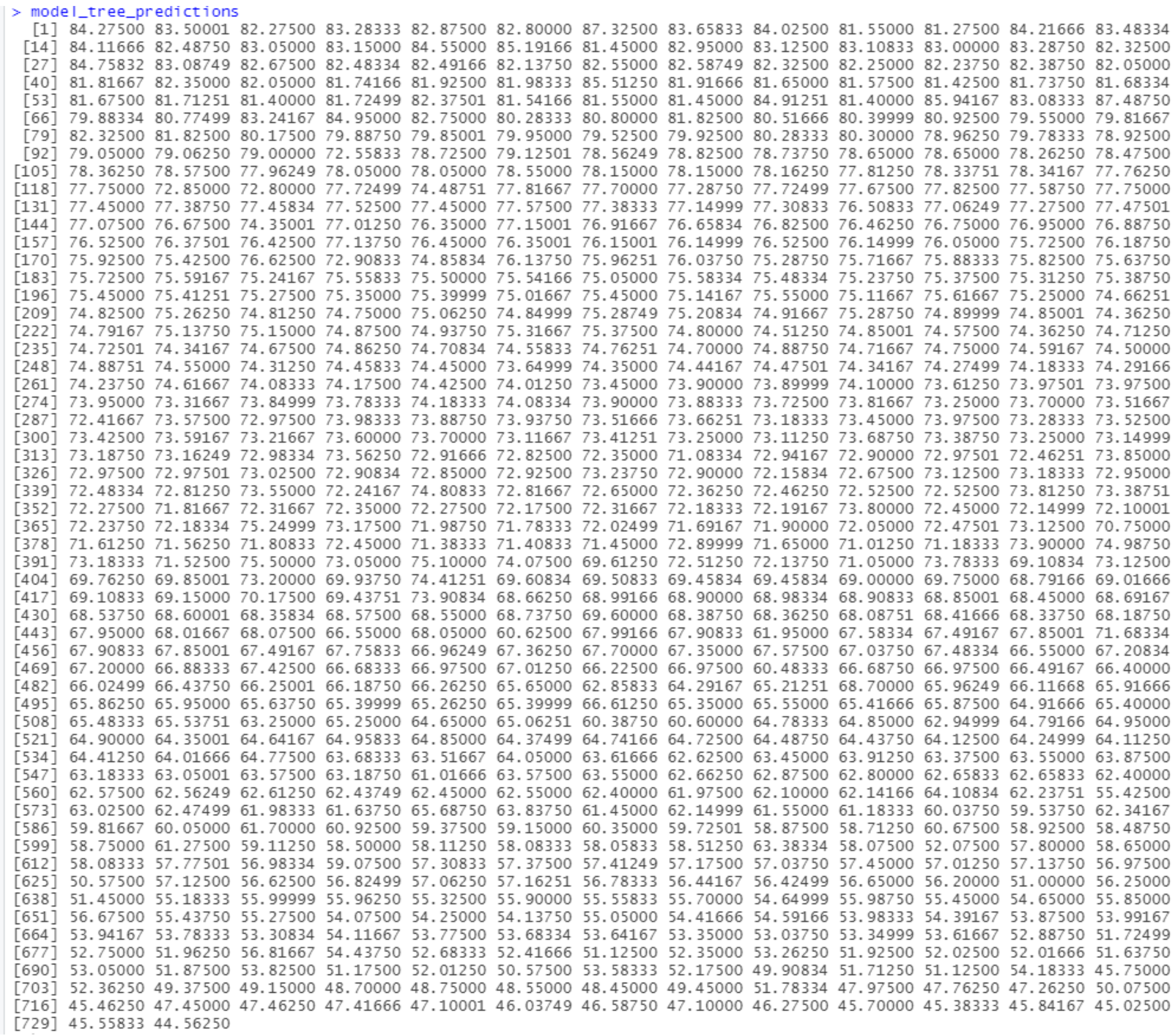


```
> svr_mape*100
[1] 1.486074
>
> model_tree_mape*100
[1] 0.9715056
>
> svr_mae
[1] 1.020689
>
> model_tree_mae
[1] 0.6590474
```

(Figure 29: Model trees and SVR comparison code snippet)

As can be seen in figure 29 model trees outperformed SVR quite significantly as model trees had scored lower in terms of errors as SVR scored 1.48 while model trees scored 0.97 in terms of mape while mae scores were 1.02 and 0.65 which is another clear lead for the model trees, based on these results it was clear that focus needs to be put towards the model trees for this analysis.

Based on the findings via figure 29 I was able to make an informed decision regarding looking into model trees which presented the following results for prediction.

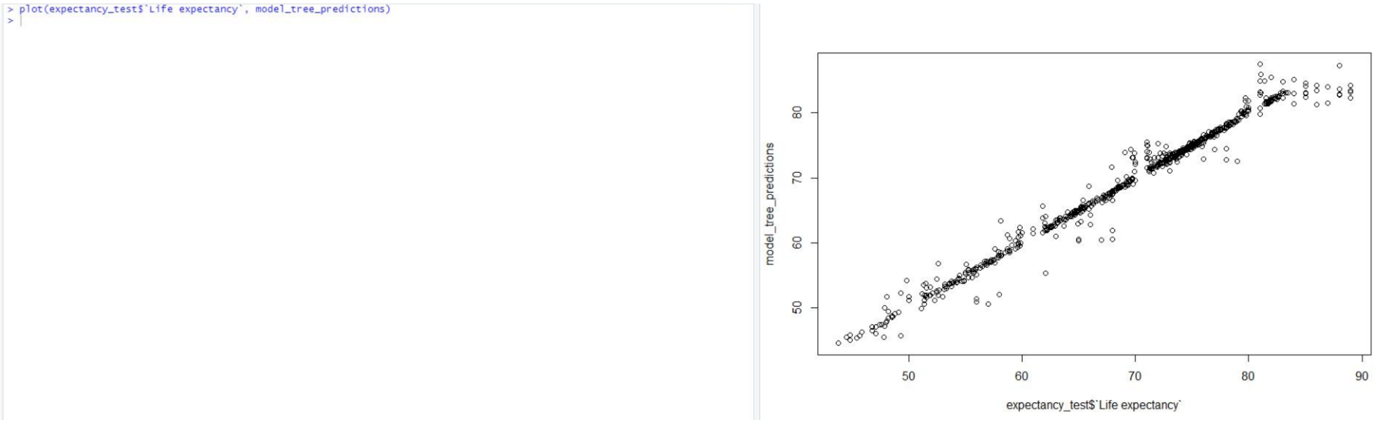


```
> model_tree_predictions
  [1] 84.27500 83.50001 82.27500 83.28333 82.87500 82.80000 87.32500 83.65833 84.02500 81.55000 81.27500 84.21666 83.48334
 [14] 84.11666 82.48750 83.05000 83.15000 84.55000 85.19166 81.45000 82.95000 83.12500 83.10833 83.00000 83.28750 82.32500
 [27] 84.75832 83.08749 82.67500 82.48334 82.49166 82.13750 82.55000 82.58749 82.32500 82.25000 82.23750 82.38750 82.05000
 [40] 81.81667 82.35000 82.05000 81.74166 81.92500 81.98333 85.51250 81.91666 81.65000 81.57500 81.42500 81.73750 81.68334
 [53] 81.67500 81.71251 81.40000 81.72499 82.37501 81.54166 81.55000 81.45000 84.91251 81.40000 85.94167 83.08333 87.48750
 [66] 79.88334 80.77499 83.24167 84.95000 82.75000 80.28333 80.80000 81.82500 80.51666 80.39999 80.92500 79.55000 79.81667
 [79] 82.32500 81.82500 80.17500 79.88750 79.85001 79.95000 79.52500 79.92500 80.28333 80.30000 78.96250 79.78333 78.92500
 [92] 79.05000 79.06250 79.00000 72.55833 78.72500 79.12501 78.56249 78.82500 78.73750 78.65000 78.65000 78.26250 78.47500
[105] 78.36250 78.57500 77.96249 78.05000 78.05000 78.55000 78.15000 78.15000 78.16250 77.81250 78.33751 78.34167 77.76250
[118] 77.75000 72.85000 72.80000 77.72499 74.48751 77.81667 77.70000 77.28750 77.72499 77.67500 77.82500 77.58750 77.75000
[131] 77.45000 77.38750 77.45834 77.52500 77.45000 77.57500 77.38333 77.14999 77.30833 76.50833 77.06249 77.27500 77.47501
[144] 77.07500 76.67500 74.35001 77.01250 76.35000 77.15001 76.91667 76.65834 76.82500 76.46250 76.75000 76.95000 76.88750
[157] 76.52500 76.37501 76.42500 77.13750 76.45000 76.35001 76.15001 76.14999 76.52500 76.14999 76.05000 75.72500 76.18750
[170] 75.92500 75.42500 76.62500 72.90833 74.85834 76.13750 75.96251 76.03750 75.28750 75.71667 75.88333 75.82500 75.63750
[183] 75.72500 75.59167 75.24167 75.55833 75.50000 75.54166 75.05000 75.58334 75.48334 75.23750 75.37500 75.31250 75.38750
[196] 75.45000 75.41251 75.27500 75.35000 75.39999 75.01667 75.45000 75.14167 75.55000 75.11667 75.61667 75.25000 74.66251
[209] 74.82500 75.26250 74.81250 74.75000 75.06250 74.84999 75.28749 75.20834 74.91667 75.28750 74.89999 74.85001 74.36250
[222] 74.79167 75.13750 75.15000 74.87500 74.93750 75.31667 75.37500 74.80000 74.51250 74.85001 74.57500 74.36250 74.71250
[235] 74.72501 74.34167 74.67500 74.86250 74.70834 74.55833 74.76251 74.70000 74.88750 74.71667 74.75000 74.59167 74.50000
[248] 74.88751 74.55000 74.31250 74.45833 74.45000 73.64999 74.35000 74.44167 74.47501 74.34167 74.27499 74.18333 74.29166
[261] 74.23750 74.61667 74.08333 74.17500 74.42500 74.01250 73.45000 73.90000 73.89999 74.10000 73.61250 73.97501 73.97500
[274] 73.95000 73.31667 73.84999 73.78333 74.18333 74.08334 73.90000 73.88333 73.72500 73.81667 73.25000 73.70000 73.51667
[287] 72.41667 73.57500 72.97500 73.98333 73.88750 73.93750 73.51666 73.66251 73.18333 73.45000 73.97500 73.28333 73.52500
[300] 73.42500 73.59167 73.21667 73.60000 73.70000 73.11667 73.41251 73.25000 73.11250 73.68750 73.38750 73.25000 73.14999
[313] 73.18750 73.16249 72.98334 73.56250 72.91666 72.82500 72.35000 71.08334 72.94167 72.90000 72.97501 72.46251 73.85000
[326] 72.97500 72.97501 73.02500 72.90834 72.85000 72.92500 73.23750 72.90000 72.15834 72.67500 73.12500 73.18333 72.95000
[339] 72.48334 72.81250 73.55000 72.24167 74.80833 72.81667 72.65000 72.36250 72.46250 72.52500 72.52500 73.81250 73.38751
[352] 72.27500 71.81667 72.31667 72.35000 72.27500 72.17500 72.31667 72.18333 72.19167 73.80000 72.45000 72.14999 72.10001
[365] 72.23750 72.18334 75.24999 73.17500 71.98750 71.78333 72.02499 71.69167 71.90000 72.05000 72.47501 73.12500 70.75000
[378] 71.61250 71.56250 71.80833 72.45000 71.38333 71.40833 71.45000 72.89999 71.65000 71.01250 71.18333 73.90000 74.98750
[391] 73.18333 71.52500 75.50000 73.05000 75.10000 74.07500 69.61250 72.51250 72.13750 71.05000 73.78333 69.10834 73.12500
[404] 69.76250 69.85001 73.20000 69.93750 74.41251 69.60834 69.50833 69.45834 69.45834 69.00000 69.75000 68.79166 69.01666
[417] 69.10833 69.15000 70.17500 69.43751 73.90834 68.66250 68.99166 68.90000 68.98334 68.90833 68.85001 68.45000 68.69167
[430] 68.53750 68.60001 68.35834 68.57500 68.55000 68.73750 69.60000 68.38750 68.36250 68.08751 68.41666 68.33750 68.18750
[443] 67.95000 68.01667 68.07500 66.55000 68.05000 60.62500 67.99166 67.90833 61.95000 67.58334 67.49167 67.85001 71.68334
[456] 67.90833 67.85001 67.49167 67.75833 66.96249 67.36250 67.70000 67.35000 67.57500 67.03750 67.48334 66.55000 67.20834
[469] 67.20000 66.88333 67.42500 66.68333 66.97500 67.01250 66.22500 66.97500 60.48333 66.68750 66.97500 66.49167 66.40000
[482] 66.02499 66.43750 66.25001 66.18750 66.26250 65.65000 62.85833 64.29167 65.21251 68.70000 65.96249 66.11668 65.91666
[495] 65.86250 65.95000 65.63750 65.39999 65.26250 65.39999 66.61250 65.35000 65.55000 65.41666 65.87500 64.91666 65.40000
[508] 65.48333 65.53751 63.25000 65.25000 64.65000 65.06251 60.38750 60.60000 64.78333 64.85000 62.94999 64.79166 64.95000
[521] 64.90000 64.35001 64.64167 64.95833 64.85000 64.37499 64.74166 64.72500 64.48750 64.43750 64.12500 64.24999 64.11250
[534] 64.41250 64.01666 64.77500 63.68333 63.51667 64.05000 63.61666 62.62500 63.45000 63.91250 63.37500 63.55000 63.87500
[547] 63.18333 63.05001 63.57500 63.18750 61.01666 63.57500 63.55000 62.66250 62.87500 62.80000 62.65833 62.65833 62.40000
[560] 62.57500 62.56249 62.61250 62.43749 62.45000 62.55000 62.40000 61.97500 62.10000 62.14166 64.10834 62.23751 55.42500
[573] 63.02500 62.47499 61.98333 61.63750 65.68750 63.83750 61.45000 62.14999 61.55000 61.18333 60.03750 59.53750 62.34167
[586] 59.81667 60.05000 61.70000 60.92500 59.37500 59.15000 60.35000 59.72501 58.87500 58.71250 60.67500 58.92500 58.48750
[599] 58.75000 61.27500 59.11250 58.50000 58.11250 58.08333 58.05833 58.51250 63.38334 58.07500 52.07500 57.80000 58.65000
[612] 58.08333 57.77501 56.98334 59.07500 57.30833 57.37500 57.41249 57.17500 57.03750 57.45000 57.01250 57.13750 56.97500
[625] 50.57500 57.12500 56.62500 56.82499 57.06250 57.16251 56.78333 56.44167 56.42499 56.65000 56.20000 51.00000 56.25000
[638] 51.45000 55.18333 55.99999 55.96250 55.32500 55.90000 55.55833 55.70000 54.64999 55.98750 55.45000 54.65000 55.85000
[651] 56.67500 55.43750 55.27500 54.07500 54.25000 54.13750 55.05000 54.41666 54.59166 53.98333 54.39167 53.87500 53.99167
[664] 53.94167 53.78333 53.30834 54.11667 53.77500 53.68334 53.64167 53.35000 53.03750 53.34999 53.61667 52.88750 51.72499
[677] 52.75000 51.96250 56.81667 54.43750 52.68333 52.41666 51.12500 52.35000 53.26250 51.92500 52.02500 52.01666 51.63750
[690] 53.05000 51.87500 53.82500 51.17500 52.01250 50.57500 53.58333 52.17500 49.90834 51.71250 51.12500 54.18333 45.75000
[703] 52.36250 49.37500 49.15000 48.70000 48.75000 48.55000 48.45000 49.45000 51.78334 47.97500 47.76250 47.26250 50.07500
[716] 45.46250 47.45000 47.46250 47.41666 47.10001 46.03749 46.58750 47.10000 46.27500 45.70000 45.38333 45.84167 45.02500
[729] 45.55833 44.56250
```

(Figure 30: Model trees prediction code snippet and ouput)

Figure 30 shows all of the predictions gathered from the model however it is important to use visualisations to understand the accuracy of the prediction.

In order to visualise and interpret the accuracy of the data I had made a plot to show the actuals vs the residuals.



```
> plot(expectancy_test$`Life expectancy`, model_tree_predictions)
> 
```

(Figure 31: actuals vs residuals plot and code snippet)

As I was predicating the output is very close to a straight line with some errors which contain some errors of a larger magnitude however there is not many of them which means that the model is working very well and predicted the life expectancy using the training data quite accurately which is seen as it is compared to the actual data. As described above in the analysis section for models of this nature it was advised to me not to go beyond the 2015-year time frame as there is no way to validate the results beyond that point. Currently due to the global pandemic it is extremely difficult to build a model which can show life expectancy in the year 2030 for example as the sheer scale of the pandemic will not be know for many years to come.

To conclude this results section having tested for serial correlation, the null hpotheiss is that there is no serial correlation while the alternate hypothesis is that there isa correlation. Based on the analysis it is clear that correlation is present which means that as P is very low which means that the null hypothesis can be rejected.

Developed and Developing countries

The key take insights from the analysis on developed and developing countries I believe is that the difference between living in a developed country and a developing country can be extremely drastic. The most shocking insight in my opinion presented itself in figure 6 in part 4.1 which shows the box plots for life expectancy for developed and developing countries. What was most staggering to see was the magnitude of variation found in the life expectancies among developing countries.

As mentioned above there are some countries which despite being developing are doing extraordinary well such as Chile, but it can also be seen that there are a very large number of countries below a life expectancy of 60.

Figure 7 brought another interesting insight as it had showed that yes there is a correlation between longer lives being happier, but it did also present the fact that there are many countries which are developing who managed to live happier lives despite typically living shorter lives which is absolutely evident in figure 4 as the graph for life expectancy amongst developing countries is extremely broad in comparison to developed countries.

GDP (PPP) and life expectancy

Figure’s 10 and 11 show very clearly how non normally distributed the data is between life expectancy and especially GDP (PPP). In these insights it can be seen very clearly how a small percentage of the world has the vast majority of the wealth in the world which keeps the rich countries progressing while the developing countries continue to struggle. Figure 14 shows the box plot for GDP(PPP) which shows how far the average GDP is from the rich countries which are extremely rich in comparison.

Figure 19 however shines a light which shows that there are countries which despite a relatively low GDP (PPP) of below even 2000 are able to exceed a life expectancy of 75 years. There are however countries with a very similar GDP (PPP) which have a drastically lower life expectancy at even lower than 75 years. This is an important insight as purely having a higher GDP (PPP) does not guarantee a longer life expectancy, but it suggests that the governments approach to factors like health care carries a lot of weight too especially in the case of poorer countries.

Happiness vs GDP PPP

Figure 21 in part 4.5 of the report was aimed to answer a question which so many people ask which is, does money buy happiness? There is no direct answer to this question but based on the analysis I believe that it does not buy happiness to the extent I originally believed. I believe so as there are a lot of countries present which have a GDP PPP of barely over 2000 or lower with a happiness score which can easily exceed 6.0 and in some cases even 7.0 is exceeded.

What is however very intriguing to me is that out of all of the nations present not a single nation which scored above 6500 GDP PPP was able to score above 7.0 in terms of happiness leading me to believe that money isn’t a deciding factor at all when it comes to scoring over 7.0 in terms of happiness. In fact, there are countries which scored less than 2000 in terms of

GDP PPP (which is over 3 times less than the richest countries) which were still able to achieve 7 in terms of happiness score which the richest countries were clearly unable to do.

The huge variations in terms of GDP PPP being so high in certain instances did not translate to the happiest lives which shows that although there is a correlation between happiness and GDP in order to be extraordinarily happy an extraordinary amount of wealth within a country is definitely not needed. Which shows that money doesn’t necessarily by happiness at that stage of wealth.

On the other hand, we see that there is far less extremes in terms of wealth and happiness the lower down the Y axis (GDP PPP axis) I look which would suggest that there is a bigger correlation with poor countries being unhappy than there is with very rich countries being very happy.

To conclude this analysis, I am led to believe that although until a certain point money doesn’t buy happiness. However unfortunately it can also be seen that on the other hand a lack of money does buy unhappiness to a stronger level than money buying happiness.

Air pollution effect’s effects on life expectancy

The insights gathered surrounding air pollution have been some of the most intriguing. Figure 23 in part 4.6 has been particularly interesting. Figure 23 shows the correlation between life expectancy and air quality in terms of PM 2.5 particles in air per cubic meter as an annual average. In this visualisation it can be seen that there is huge variation in terms of air pollution for all ages below 80. In fact, there are many countries with relatively good air scoring less than 25 but with life expectancies below even 55.

On the other end of the spectrum there are countries which scored over 60 particles per cubic meter but maintained a life expectancy of over 70 years, crucially however not a single country with over 25 particles per cubic meter was able to achieve a life expectancy of 80 years or more. This hints at a somewhat of a threshold where a country can not achieve a life expectancy of over 80 without reducing the pollution below 25.

The findings from figure 23 had intrigued me to much that I had to analyse further.

GDP PPP and Air Pollution

I believe that figure 25 in section 4.7 is perhaps the most intriguing of all. In this visualisation I aim to understand the correlation between wealth and air quality. The first thing aspect of figure 25 which stands out tremendously is the fact that the trend between developed and developing countries is completely reversed.

Developed countries with a higher GDP PPP tend to have much better air quality, this is presumably due to the fact that those countries are able to invest in new technologies which are more eco friendly such as electric cars and of course renewable energy. This reduction of pollution in the air clearly helps in achieving an impressive life expectancy of over 80 years.

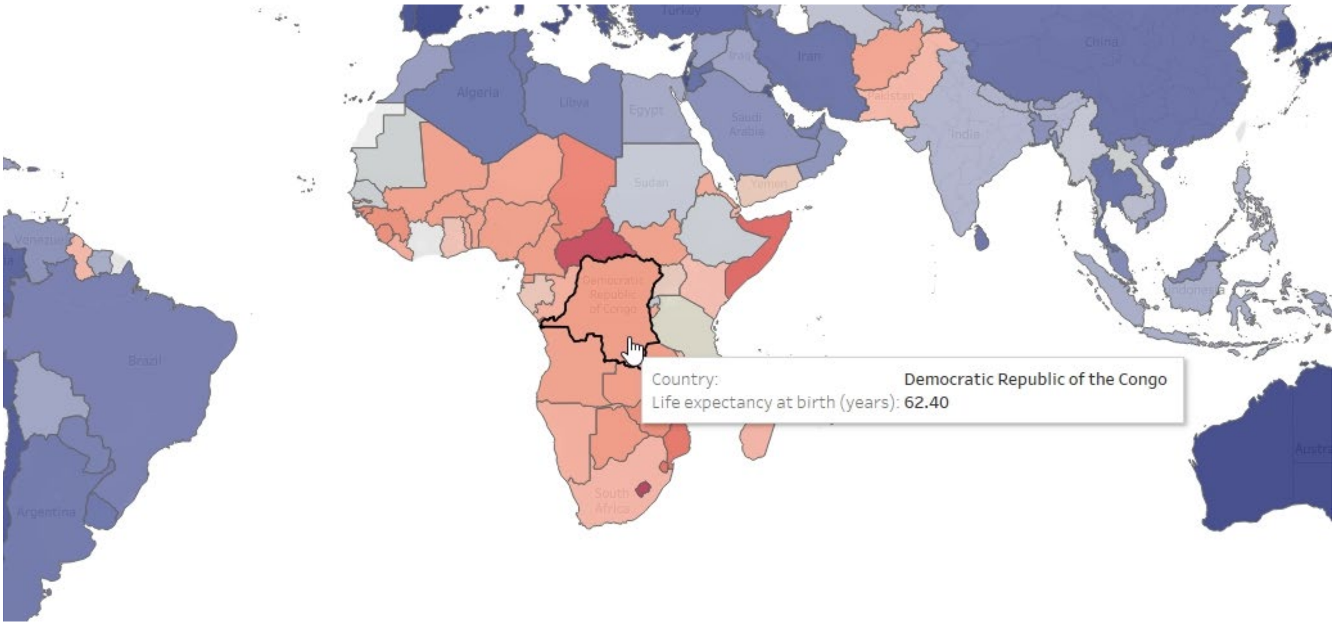
In the case of developing countries these countries do not have the wealth and technological advancements to focus on renewable energy and therefore have poor air quality which seemingly caps those countries at a life expectancy below 80.

To conclude, as it can be seen that developing countries with a higher GDP PPP tend to have a lower quality of air showing that these countries are seemingly sacrificing their air quality for the sake of higher wealth. I believe these countries are rapidly expanding their infrastructure and as big factories in theses countries are in operation without adequate use of technology to combat their air pollution. I also believe that the GDP PPP is higher more people use cars for their daily commuting which also in their case contributes to poor air quality as most cars in developing countries would be older cars with engines which are far worse for the environment than those cars found in developed countries.

Tableau Animation and Map

The goal of the tableau component of this project was to deliver something which is very insightful but also looks aesthetically pleasing and is user friendly.

In order to understand the life expectancy around the world by regions I have created an interactive colour coded map of the world which displays life expectancy when the user hovers their mouse over it the life expectancy of the given country is displayed.

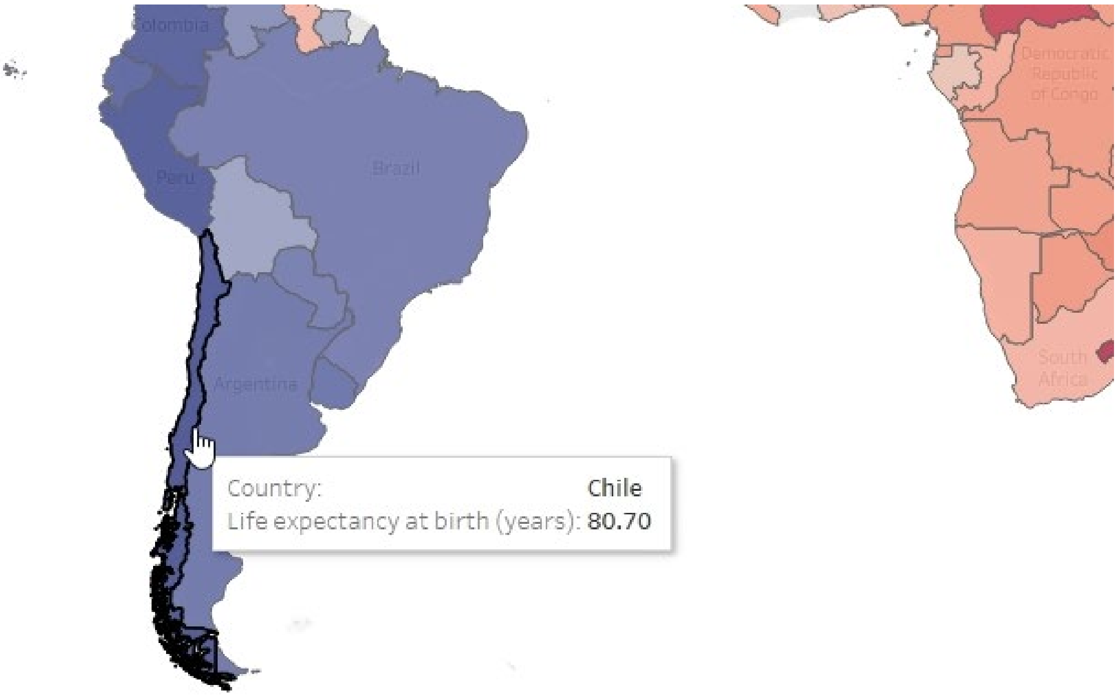


Country: Democratic Republic of the Congo
Life expectancy at birth (years): 62.40

(Figure 32: Tableau interactive map)

Figure 32 shows the functionality of the tableau map which shows countries being colour coded based on their life expectancy. In figure 32 the user is examining the democratic republic of Congo.

During my analysis Chile was a country which caught my attention for being a developing country with a high life expectancy.

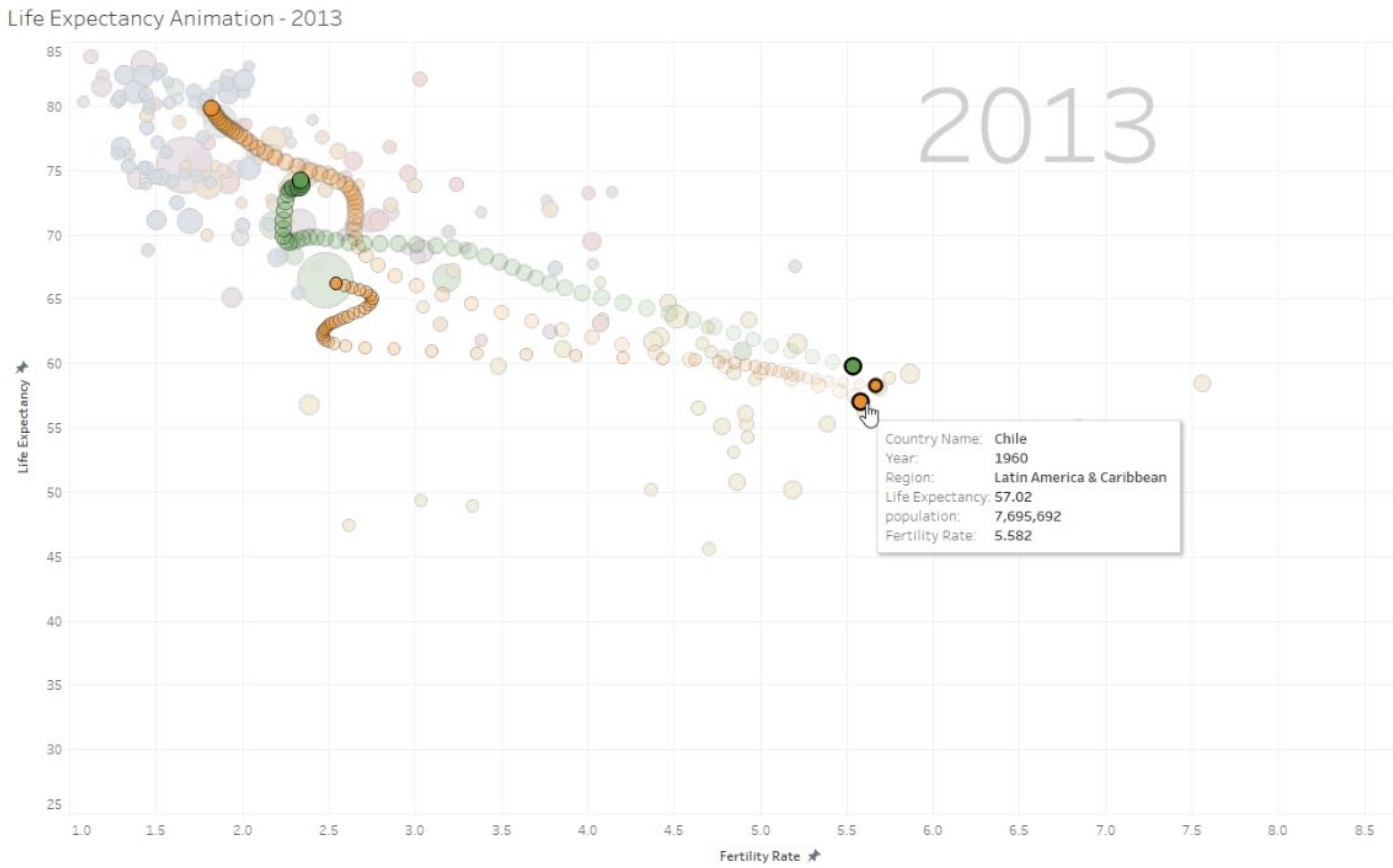


Country: Chile
Life expectancy at birth (years): 80.70

(Figure 33: tableau map showing Chile)

As can be seen in Figure 33 Chile has a very impressive life expectancy at 80.7 years.

The small country of Chile had impressed me for the fact that despite being a country with only the 43rd highest GDP PPP in my dataset while being a developing country for the year 2015 according to the WHO dataset had a life expectancy very close to 80 years. As it can be seen in figure 33 only a few years later in the year 2019 Chile achieved a life expectancy of 80.70 which is very impressive. In the merged dataset with GDP PPP, happiness and of course WHO Chile had the 27th highest life expectancy of any country which is hugely impressive.



Country Name: Chile
Year: 1960
Region: Latin America & Caribbean
Life Expectancy: 57.02
population: 7,695,692
Fertility Rate: 5.582

(Figure 34: tableau life expectancy animation from 1960 to 2013)

Figure 34 shows exactly how well Chile has done to improve their life expectancy from 1960 to 2013. When looking at the countries historic pattern it can be seen that countries which had higher life expectancy in 1960 than Chile were clearly left behind as Chile continued to improve in comparison to Sri Lanka (green dot) and Guyana (orange dot).

This analysis once again shows just how well Chile has done to achieve their life expectancy.

7.0 Conclusions

During the countless meetings with my supervisor Keith, there was always one major thing which my supervisor Keith kept reminding me of in terms of the approach needed in a data analytics project's. This one approach that we kept referring to was that a data analytics project in his eyes should be journey in exploring the data going from one insight to the next. Typically, the researcher is not meant to know exactly where this journey will go and with the next piece of analysis will be.

Thanks to this approach I am very happy with how my project turned out in the end especially considering how much I enjoyed working on it. The case where this journey approach was most prevalent to me during my project was when analysing the correlation between air quality and life expectancy. After this piece of analysis, I was very intrigued and asked questions regarding why the data presented itself the way it did. This then led me to analysing the correlations between GDP PPP and air quality which gave many answers, and this is what I loved most about this project.

The questions which I had asked myself were: what are the correlations between long lives and happy lives? does money actually buy happiness? What are the effects of bad air quality on our life expectancy and crucially what are the correlations between each of those factors?

In order to answer these questions, I used a wide array of technologies as well as techniques by following the KDD methodology. By finding my data and cleaning it followed by processing it and merging I believe that with the use of linear, regressions, tableau visualisations, model trees, descriptive statistics and of course hypothesis testing I am happy to have been able to answer all of the questions which I had.

By analysing the differences between developed and developing countries by using the Jamovi software I was able to understand the scope of how much life can be different depending on being from a developing or developed country with respect to life expectancy as well as visually understanding the difference in life expectancy density in figure 4 for instance.

Via statistical analysis using SPSS to test GDP PPP and life expectancy I was made aware of how non normally distributed the wealth is around the world and even life in figure 11 and 15.

The linear regressions then made it easy for me to understand and answer many questions in terms of what factors effect life expectancy and happiness. One of the most interesting parts was to find that in fact money does buy happiness to an extent but can not make a nation be the happiest in the world as the countries with the highest GDP PPP actually scored lower than many countries with as low as only 30% of the GDP PPP which is incredible. In this part of the analysis, I had come to the conclusion that it is more so that a lack of money buys unhappiness as opposed to money buying happiness because having all of the money in the world does not make people be the happiest in the world, but lack of money has a bigger correlation with being unhappy as seen in figure 21.

Perhaps the most interesting finding was the effect of air pollution on life expectancy where I had found that no country with more than 25 PM 2.5 particles per cubic meter was able to achieve a higher life expectancy than 80 years as seen in figure 23. However, figure 25 answered even more questions in demonstrating that developing countries with a higher GDP PPP tend to sacrifice their air quality.

When model tress showed how it is possible to accurately predict the future of life expectancy, tableau on the other hand figure 34 had shown that historically speaking Chile needs to be an example for many countries in terms of achieving a good life expectancy. Which are all great strengths of the project.

I believe that this type of project has one key limitation that it very difficult to get around, that limitation is that life is very complex for everyone, and an infinite number of factors may have an effect on an infinite number of different factors. As a result, I think that for a project of this nature there is never enough factors being taken into consideration which is its key weakness.

8.0 Further Development or Research

Without a doubt if I were to continue with this project, I would certainly keep adding to the list of factors which influence life and happiness. One of those factors would be safety and crime rates for example and how this influences happiness in particular.

With enough time given in order to expand upon this project I would be very interested in investigating the effect that covid has had on countries around the world. Unfortunately for this type of analysis to be performed to the highest accuracy possible it would be important to wait several years before analysing how each country dealt with the pandemic and of course how the life expectancy has been affected.

Covid has also had a huge effect on people’s mental health and wellbeing which is also something I would like to analyse in the context of world happiness.

9.0 References

Rajarshi, K., 2021. Life Expectancy (WHO). [online] Kaggle.com. Available at: <https://www.kaggle.com/kumarajarshi/life-expectancy-who> [Accessed 2 August 2021].

Solutions Network, S., 2021. *World Happiness Report*. [online] Kaggle.com. Available at: <https://www.kaggle.com/unsdsn/world-happiness> [Accessed 2 August 2021].

Weinmeister, K., 2021. *PM2.5 Global Air Pollution 2010-2017*. [online] Kaggle.com. Available at: <https://www.kaggle.com/kweinmeister/pm25-global-air-pollution-20102017> [Accessed 2 August 2021].

Superdatascience.com. 2021. *SuperDataScience*. [online] Available at: <https://www.superdatascience.com/pages/tableau-advanced> [Accessed 2 August 2021].

Nitisha, 2021. *GDP per capita all countries*. [online] Kaggle.com. Available at: <https://www.kaggle.com/nitishabharathi/gdp-per-capita-all-countries> [Accessed 2 August 2021].

GeeksforGeeks. 2021. *KDD Process in Data Mining - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/kdd-process-in-data-mining/> [Accessed 2 August 2021].

10.0 Appendices, Proposal

10.1. Objectives

(Max 1 Page)

This project will centre around data which will help to compare different countries in various different aspects such as life expectancy, wages, house prices, happiness, stability and so on. The project also aims to dive deeper and analyse what it really is that makes a country a happy place to live in for example is it safety or high wages and so on. The project also aims to discover what has the biggest effect on life expectancy for example is it the countries GDP or is it that people need access to free health care etc.

The objective is to show the users which country may be the best for them to live in and also highlight what people really value when it comes to choosing where to live and what it is that makes a country happy to live in i.e. is it wages, good health care or simply safety/low crime rates.

The objective is to also show what each country in the analysis is good at but also bad at, this will hopefully allow us to view what each country does well and what each country could improve upon in order to aim to become the perfect place to live in.

10.2. Background

(Max 2 Pages)

As this project may help to point someone towards their perfect home, I have come up with the name Home Point for this project.

Due to this project being quite broad in scope by nature of it there is a lot of potential data sets which will need to be covered and considered to be included in the project.

As a side effect of the project being so broad in scope there are also however positive aspects to it. One of the main positives is that since the data used will be only secondary data the project will rely very heavily on current data and its accessibility. By nature of this type of project not all of the various categories which are in mind will be possible due to a lack of sufficient data however thanks to the projects broad scope there is quite a lot of potential for finding new and interesting categories to add to this project during its research and even creation stages.

As someone who has been interested in research as to what each country does well in comparison to others I have been looking for quite a while for a tool which would allow me to see all of this information in an easily accessible and clear format however unfortunately I am yet to find a service which ticks those boxes.

It is clear that there is an interest in this kind of topic as a small and brief YouTube series on this kind of topic was able to gather 9 million views in total. The problem is that this was very brief typically only covering what each country is best at and moving on without and real comparison to other countries. If a user wanted to dive deeper into this topic the user would need to spend a lot of time reading through typically lengthy articles to find out what the user is interested in which is why the goal of this project is to deliver as much information as possible including all of the interesting finings in a format which is very accessible, clear and understandable.

Later into the development of the project there is also potential to look into what the future may bring for certain countries using previously gathered data.

By using the data which has been analysed there may be an opportunity to explore the future developments of some particular countries. For example if there will be a category regarding ranking countries educational systems there may be a country found which has a much better educational system when compared to countries with a higher GDP as inevitably wealthier countries should be doing better in a lot of aspects. When a country like this is spotted (a country which seems to over achieve considering its low GDP) based on the countries current trends it may be possible to predict by how much this countries education system can improve in later years.

Along with learning about what the future may bring for certain countries it will also be possible to analyse what are the key factors in delivering good systems in countries, in this case for example in order to have a good educational system is a high GDP vital or perhaps can good leadership and maybe a good sense of direction really allow a country to excel in various ways despite having less potential due to its GDP.

10.3. Technical Approach

Brief description of the approach to be followed (Max. 1 Page), Research, literature review, requirements capture, implementation etc…

The research aspect of this project will most likely be the most important part of the entire project as this element will dictate the whole direction which will be taken in terms of what categories will be covered and at what level of detail the categories can be covered.

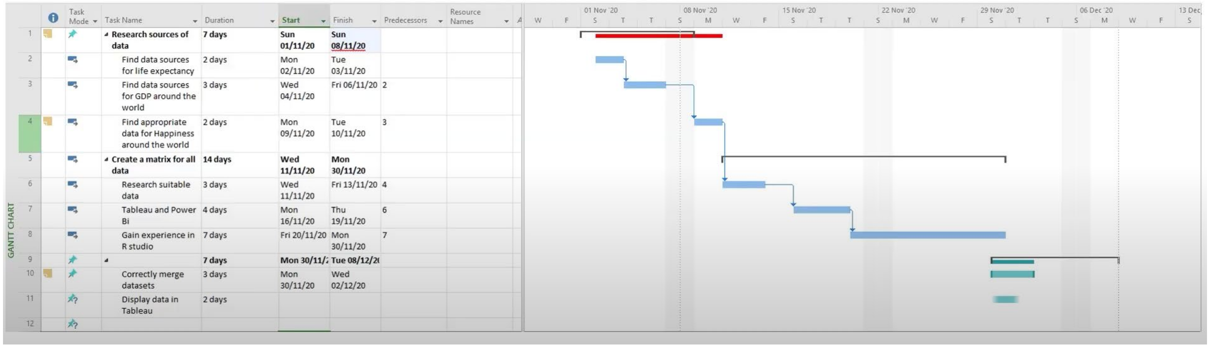
This is so as of course all of the data which will be used will be secondary data meaning that although I may have a great idea for a section/category which would tell the user a lot about certain countries it may not necessarily be possible to do that category due to insufficient data.

One of the key requirements I have for this project is too make all of my findings and data as clear and visible to the users as possible, in order to achieve that I will most likely have to use a tool such as tableau and utilise many of its powerful tools and make the service as user friendly as possible.

This means that one of the key features of the project should be that it is aesthetically pleasing and offer a clear and easy to use GUI with a lot of data to display but to be displayed in a way which is not overwhelming for the user which means that a clear layout will be very important.

10.4. Project Plan

Gantt chart using Microsoft Project with details on implementation steps and timelines



	Task Mode	Task Name	Duration	Start	Finish	Predecessors	Resource Names
1		**Research sources of data**	**7 days**	**Sun 01/11/20**	**Sun 08/11/20**		
2		Find data sources for life expectancy	2 days	Mon 02/11/20	Tue 03/11/20		
3		Find data sources for GDP around the world	3 days	Wed 04/11/20	Fri 06/11/20	2	
4		Find appropriate data for Happiness around the world	2 days	Mon 09/11/20	Tue 10/11/20	3	
5		**Create a matrix for all data**	**14 days**	**Wed 11/11/20**	**Mon 30/11/20**		
6		Research suitable data	3 days	Wed 11/11/20	Fri 13/11/20	4	
7		Tableau and Power Bi	4 days	Mon 16/11/20	Thu 19/11/20	6	
8		Gain experience in R studio	7 days	Fri 20/11/20	Mon 30/11/20	7	
9			**7 days**	**Mon 30/11/2**	**Tue 08/12/2**		
10		Correctly merge datasets	3 days	Mon 30/11/20	Wed 02/12/20		
11		Display data in Tableau	2 days				
12							

10.5. Technical Details

Implementation language and principal libraries

One of the languages which I will use in this project will the R language as this is so far the language which we have covered in class which is most suitable for data related projects.

The R language will be used in conjunction with technologies such as tableau in order to show my findings as clearly as possible in the most aesthetically pleasing way.

10.6. Evaluation

Describe how you will evaluate the system with real technical data using system tests, integration tests etc. In addition, where possible describe how you will evaluate the system with an **end user. (be careful here re Ethics etc)**

To ensure that my project is good for the end user I will ask friends and family to use it and take their feedback regarding clarity design and potentially even what categories they think would be a good fit for my project.

10.7. Reflective Journals

Reflective journal October: Wiktor Wolsza

What?:

For the purpose of the project and the video I have done research regarding the various different aspects which I can cover in my project such as comparing countries by their life expectancy, and discovering what are the main factors which contribute to a longer life expectancy i.e. do countries with high GDP tend to have much higher life expectancies than countries with lower GDP and if so how big is that difference or maybe is it access to free health care which has a bigger effect on life expectancy instead of it being purely dependent on GDP.

Other potential categories include things like rating a countries safety and stability etc. All of the sections are yet to be confirmed based on my research of what data is accessible and how accessible this data is.

After having the opportunity to discuss my project with my supervisor earlier this week I gained quite a lot of valuable information and a sense of direction in terms of where I should be going next in terms of the project.

So what?:

One of the main points taken from my meeting with the supervisor was that this project is very broad in scope which can lead to some problems down the line however it is also a good thing that it is broad in scope as it means that if certain sections of my project are not doable due to insufficient data those sections can be replaced by various different sections for which there is data.

The next thing we discussed was the presentation of this data, here we briefly discussed the use of tableau to show my findings as clearly as possible and I derived that Tableau is certainly an option which I will consider.

Now what?:

Now the main things to focus on in the development of the project is surrounding the data which I can use. At the moment I am researching different suitable data sources and as the supervisor recommended and listing them all in a matrix on excel for easy access later on.

Reflective journal November: Wiktor Wolsza

What?

From the last reflective journal progress has been made in identifying datasets which will be suitable. One of those key datasets is the World Health Organisation dataset which covers a wide array of aspects such as GDP per capita, HIV rates, life expectancy and so on. Another data set which can be merged with the WHO data to bring new and interesting insights is the happiness rank data which covers several years from 2015 allowing for the merging of the data sets to be as smooth as possible.

Research was also done into methodologies such as KDD which is the methodology I will be using in this project.

I have also done merging and practice data filtration and processing in the R language in R studio along with making clear and easy to understand visualisation of the data and information which was gathered.

Research was also done on Tableau to gain an understanding of how to use this tool as well as the Power Bi software.

So what?

The meetings with my supervisor have helped me build a clearer and clearer picture of how I want my project to look and also how I should be approaching tackling going about some of the issues which have arisen for example an issue with some of my data having seemingly incorrect values.

Now what?

Now the main thing is to keep developing a better understanding of using the tools which will be utilised in the project and continue to look for new data and potential interesting categories which can be made for the project.

Reflective journal December: Wiktor Wolsza

What?

By following online tutorials, I was able to construct an animation type of visualisation which covers the years from 1960 to 2013. The visualisation can be used to explore the trends which have occurred over all of those years across 216 countries in terms of life expectancy at birth as well as fertility rates. The user can inspect any of the countries in particular by highlighting any of the countries which will show the historic trend lines for that particular country, the user can also select more than one country in order to compare and contrast the trend between a number of countries.

An interactive map was also made with life expectancy which allows the user to hover over any of the countries on the map in order to inspect their data. The map is also colour coded which means that the user can spot certain trends across continents in terms of life expectancy as certain countries stand out more than others due to their life expectancy.

More experience was also gained in R studio particularly in data merging as well as other tasks such as writing the new data to new CSV's in order to be prepared for visualisations to be made within Tableau.

So what?

The experience gained in R studio and Tableau are so far showing good results as I am particularly happy with the visualisations which are very intuitive clear and interactive which allow me and the users to explore the data and the trends with ease.

Now what?

As I keep expanding the project I aim to continue to research new potential categories as well as to continue to develop the life expectancy category in order to give that category as much depth as possible. I also plan to apply new techniques which I have not used before such as data mining in order to gain experience in this area and extract as much as possible from the data.

Reflective journal January: Wiktor Wolsza

What?

After editing the mid point presentation video and uploading the presentation along with all of its documents the exam period came about which meant that the focus on the project had to be moved over to focus on terminal based assessments for the rest of the modules. This meant that for the most part this month was the slowest month there has been so far in terms of progress as the majority of my attention had to be put towards other assessments.

However, this time was used wisely to reflect on the progress so far and also on what has to be done. Overall, I was happy with my grade for the mid point which was certainly very reassuring as it meant that I was on the right track to achieve my desired grade.

So what?

Understanding that I am on the right track in terms of what I need to be doing for the project was a big confidence boost going forward as this meant that I know knew the formula which I can follow for the rest of the project.

Now what?

The next step in the project is to wait for a meeting with my supervisor in order to gain valuable feedback regarding my grades in order to understand where there is most space for improvement in order to "collect the low laying fruit" in terms of marks to maximise my final grade. After that I will implement whatever changes are recommended and verify them with my supervisor.

Reflective journal February: Wiktor Wolsza

What?

As soon as the opportunity was there to meet with my supervisor, I was able to get a detailed breakdown of my marks which was very important to me in order to point me in the direction in which there is most potential for improvement. Upon finding out that the weakest component of the marking rubric was the proposal document.

At this stage I was also beginning to narrow down some other parts of the marking rubric such as the testing phase of the project with which I have very little experience as testing for a software project which I have done before is a lot different to testing on a Data Analytics project.

So what?

One of the initial testing strategies I thought of was to verify and cross reference some of my calculations with the same test carried out on a different software in order to make sure that the primary software of my choice is consistent and also accurate. After talking about this with my supervisor we had established that this is not a bad idea however we both think a better solution is possible which will require research to find.

As part of other modules, I am also learning very interesting data analytics techniques especially in the area of data mining which means that I can transfer my experience from this module into my main software project, some of the techniques which are of most interest to me are techniques such as clustering, decision trees, logistic regression and so on.

Now what?

At the moment I am gaining familiarity in those new techniques which can be used to great effect in my project. I am also researching the possibility of reusing some of those techniques on existing data sets as well as examining what types of data sets would be applicable for the new techniques which were introduced to me.

Reflective journal March: Wiktor Wolsza

What?

After getting marking feedback from my supervisor, I had found out that it was my proposal document which is where I had lost the most marks. As this part of the submission was not the biggest component the loss of marks was not too dramatic however it was still important for me to understand where I could have improved in order not to make the same mistakes in the future.

With the help of my supervisor, I had come to the conclusion that the main reason for my lower marks in the proposal was due to my initial approach which was overly colloquial and not detailed enough which was partially due to the fact that the proposal was done very early in the development of the project. However, the key here was that I learned more about the types of language I should use in this type of documents going forward.

As the vast majority of the data and web mining module has been covered, I have also gained a lot more knowledge in a lot of techniques which will be particularly useful to me and especially for the data which I am using particularly the Random forest which is applicable to my data.

As in the later parts of this month I have been focusing more on CA's the progress on the final year project has been quite limited however knowledge gained during those CA's will be a big help in progress of the final year project.

So what?

This month the Project showcase has also been finalised and completed.

Now what?

Now the next objectives are to begin the final project documentation to allow enough time for my supervisor to read through it before the final submission allowing me to make vital changes before the final marking. As I am going to be developing new categories I am also going to use different techniques to deal with new obstacles such as missing data for which I have found new solutions such as Mice and Amelia which are both better than any of the previous methods I had known.

Reflective journal April: Wiktor Wolsza

What?

In this month I have looked a lot more carefully in the models I can apply to my data. After doing some research I had decided to apply the model trees algorithm on my dataset in order to analyse the data. Using model trees graphs and predictions were made to get further insights into my project.

Based on my supervisors' comments from the mid point marking I have decided to make broad changes in my document so far starting with the executive summary. This month due to my situation I had began asking my supervisor questions regarding possibly getting a deferral for the project to allow me extra time to deliver the best result I can.

So What?

This month I had made progress in my analysis via model trees as well as very importantly making major adjustments to my report including rewriting the majority of the report so far.

Now what?

Refinements need to be made to the model trees in order to produce easy to understand graphs and derive useful information from it.

Reflective journal May: Wiktor Wolsza

In this month after a very hectic time and a stressful period after careful consideration as well as close feedback with my supervisor I had decided to defer the project to ensure that I achieve the grade I was aiming for.

After completing the rest of the projects and submissions I had taken some time off to take a much-needed break from college in order to come back refreshed and work at my best.

Reflective journal June: Wiktor Wolsza

What?

In this month I began to write more of the report including a detailed description of why I have picked the chosen methodology as well as its pros and cons amongst the methodologies competitors. I have also written the technologies section carefully outlining each of the technologies I have used to complete my project. I have also made a poster including elegant screen shots of my project as well as a HomePoint logo.

So What?

As I know have a very good foundation set for my project, I feel very confident going forward regarding reporting all of my analysis.

Now what?

Since the nature of my project makes it quite broad in scope, I have been looking at adding new datasets for extra analysis such as air quality datasets. This dataset proved to have been very useful and delivered on many great insights.

Reflective journal July: Wiktor Wolsza

This month the focus has been on finalising the findings as well as visualisations and graphs which I have completed. This was a very important step as it was crucial to ensure an elegant completion of my report. After I have polished my findings and analysis which I have carried out using SPSS, R, Jamovi and Excel it was time to describe it all in detail and present my findings in an elegant and easily readable way.

Signature: Wiktor WOLSZa