University of Strathclyde

MSc Artificial Intelligence and Applications

CS982 : Big Data Technologies - Assignment

Analysis of World Development Indicators

Barry Smart

Registration Number: 201962939

Monday 11th November 2019

Contents

Introduction 1
Approach 1
1 - Problem Definition 3
2 - Data Acquisition 3
3 - Initial Exploration 4
4 - Data Wrangling 5
5 – Addressing Gaps in the Data 6
6 – Analysis of Individual Variables 7
Life Expectancy 7
Gross Domestic Product (GDP) Per Capita 8
Population Growth 10
7 –Analysis of Relationships Between Variables 11
8 – Application of Unsupervised Methods 15
Rationale 15
Application 16
Findings 18
9 – Application of Supervised Methods 19
Rationale 19
Application 20
Findings 21
10 - Conclusions 23
Assessment Against Problem Definition 23
Reflection 23
References 25
Figures 26
Tables 27
Appendix 28
A1 - Design Decisions 28
A2 – Development Environment 29
A3 – Process to Acquire Data From World Bank 31
A4 – Definition World Development Indicators (WDIs) 34
A5 – Results of Data Interpolation 37
A6 – Comparison of Clustering Models 38
Agglomerative Clustering 38
K-Means Clustering 39
A7 – Dendrogram 40

Note - word count for main body of document (sections “Introduction” to “10 – Conclusions” inclusive) is 3,242 excluding captions for Figures and Tables.

Introduction

The choice of data set and the objective of exploring global wealth and health trends was inspired by the BBC Four programme “The Joy of Stats - 200 Countries, 200 Years, 4 Minutes” in which Hans Rosling uses animated data visualisation to tell the story of world development from 1810 to 2010 (BBC Four, 2010):

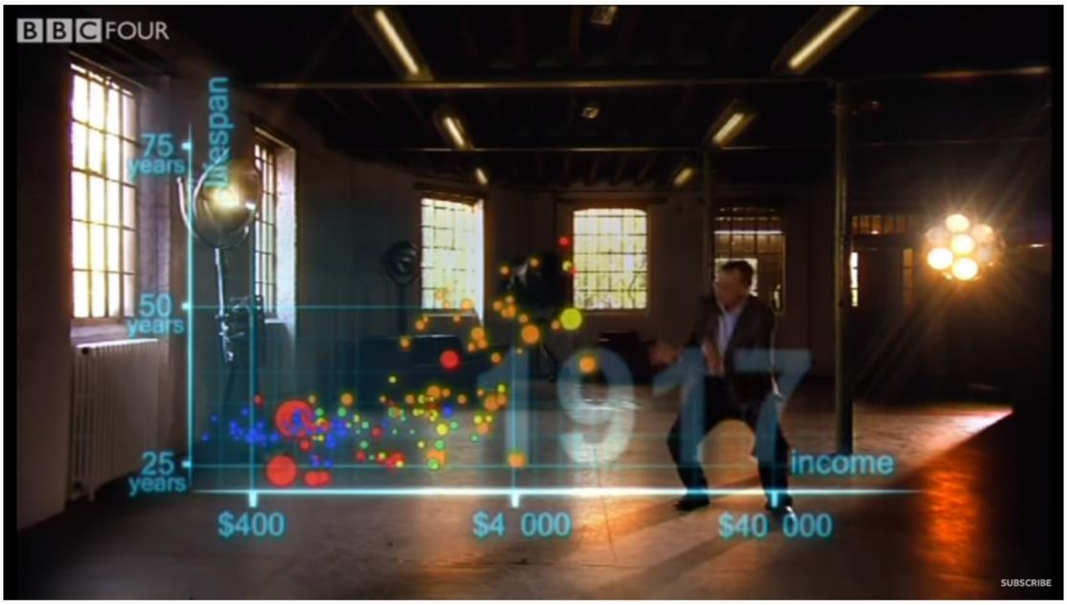


Figure 1 - frame captured from the YouTube video.

https://www.youtube.com/watch?v=jbkSRLYSojo

Approach

To complete the assignment I followed the series of steps illustrated in Figure 2 below. The process was not entirely linear as the illustration suggests, exhibiting many iterations or “loops within loops” (Zumel & Mount, 2014) in line with normal data science practice.

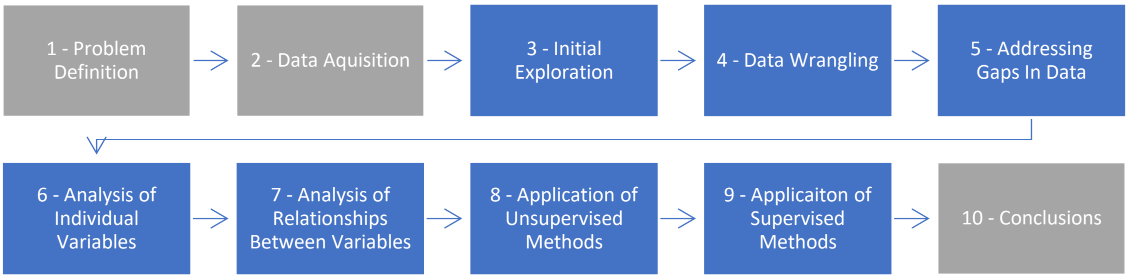

1 - Problem Definition
2 - Data Aquisition
3 - Initial Exploration
4 - Data Wrangling
5 - Addressing Gaps In Data
6 - Analysis of Individual Variables
7 - Analysis of Relationships Between Variables
8 - Application of Unsupervised Methods
9 - Applicaiton of Supervised Methods
10 - Conclusions

Figure 2 - illustration of major stages of work undertaken to complete the assignment. Stages highlighted in blue correspond to a discrete Python files.

A discrete Jupyter notebook was created for each of the stages highlighted in blue Figure 2 above.

Figure 3 below shows how the data flowed through these notebooks:

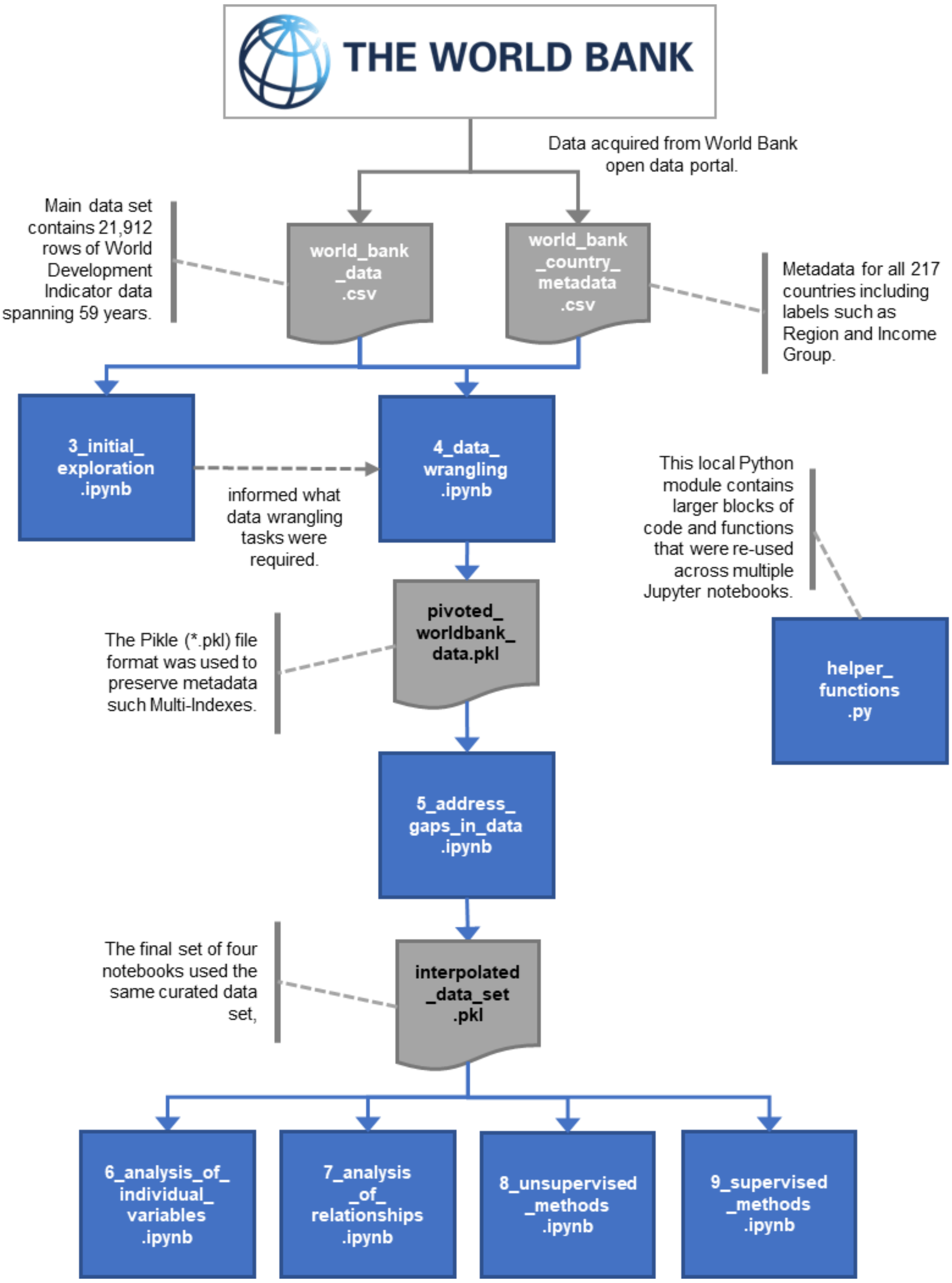

THE WORLD BANK
Data acquired from World Bank open data portal.
Main data set contains 21,912 rows of World Development Indicator data spanning 59 years.
world_bank_data.csv
world_bank_country_metadata.csv
Metadata for all 217 countries including labels such as Region and Income Group.
3_initial_exploration.ipynb
informed what data wrangling tasks were required.
4_data_wrangling.ipynb
This local Python module contains larger blocks of code and functions that were re-used across multiple Jupyter notebooks.
helper_functions.py
The Pikle (*.pkl) file format was used to preserve metadata such Multi-Indexes.
pivoted_worldbank_data.pkl
5_address_gaps_in_data.ipynb
The final set of four notebooks used the same curated data set,
interpolated_data_set.pkl
6_analysis_of_individual_variables.ipynb
7_analysis_of_relationships.ipynb
8_unsupervised_methods.ipynb
9_supervised_methods.ipynb

Figure 3 - overview of end to end data flow.

A number of design principles underpinned this approach – these are set out below in Appendix "A1 - Design Decisions" on page 28

See Appendix "A2 – Development Environment" below on page 29 for full details of the development tools used for the assignment.

1 - Problem Definition

The overarching objective was to apply the tools and methods covered by CS982 whilst exploring the following 5 questions in relation to World Development Indicators (WDIs):

#	Question
1	How has the gap in economic prosperity between the poorest and richest countries developed over time?
2	Does the economic prosperity of a country have an impact on the health of its citizens?
3	Could visualisation techniques be used to enable people to "acquire an evidence based world view" (Rosling, et al., 2005)?
4	Could unsupervised methods be used to cluster countries according to their relative economic standing?
5	Could a regression model be trained to enable future life expectancy to be reliably predicted based on economic growth assumptions?

Table 1 - the five questions set at the start of the assignment to frame the problem.

2 - Data Acquisition

Suitable open data was sourced from the World Bank (World Bank, The, 2019): https://databank.worldbank.org/

See Appendix "A3 – Process to Acquire Data From World Bank" (below on page 31) for the process followed to download the data.

This is provided under Creative Commons license by the World Bank and International Energy Agency.

The databank spans 3 core dimensions as follows:

1. **Country** - data for 217 countries;
2. **Time** - annual data from 1960 to 2018 inclusive (59 years);
3. **World Development Indicators (WDIs)** - an extensive range of 1,432 WDIs. 85 were downloaded, but ultimately these were reduced to 20 following analysis of data quality and correlations. See Appendix "A4 – Definition World Development Indicators (WDIs)" (below on page 34) for a detailed description of the 20 WDIs ultimately chosen for use in the assignment.

The three WDIs featured most extensively in this assignment were:

- **Life Expectancy at Birth, total (years)** – the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life;
- **Gross Domestic Product (GDP) per Capita (current US$)** – "is a measure of the size and health of a country's economy" (Bank of England, 2019) in this case normalised as a "per capita" amount and standardised in US dollars.
- **Population Growth (annual %)** – the rate at which population has grown (or declined) in that year as percentage change from the prior year.

3 - Initial Exploration

The World Bank data was provided in structured CSV files, so ingestion into a Pandas data frame was straightforward. However, during initial exploration, a range of issues became apparent that would need to be addressed, these included:

1. Pivoting the variables (WDIs) into discrete columns;
2. Trailing rows at the end of the dataset;
3. Unpivoting the Year columns and converting them from string format (eg "1968 [1968]") into an integer (eg 1968);
4. Translating variables into floating point numbers. Pandas had failed to do this automatically due to the presence of double full stops in the data (see next issue);
5. Replacing double full stops ".." with a null (np.nan);
6. Integrating categorical information about each country such as "Region" and "Income Group" from a separate meta data file.

4 - Data Wrangling

The data wrangling stage was used to address all six issues identified in the exploration stage above. The most significant task being to transform the data into a better shape for downstream analysis and modelling. This was achieved in two steps illustrated below:

Initial data structure – Years as columns (59 in total), combinations of Country, and Series defining each row as follows:

Index	Country	Series	1960	1961	…	2017	2018
0	Afghanistan	Population, total	8996973	9169410	…	36296400	37172386
1	Afghanistan	Life expectancy at birth, total (years)	32.45	57.37	…	64.13	64.13
…	…	…	…	…	…		…
21911	Zimbabwe	GDP per capita (current US$)	580.32	832.86	…	510.74	1430.10

Table 2 - initial data structure with years as columns.

Step 1 – the first step was to transform this data into a "tall and skinny" structure through use of Pandas **melt** function so that Year columns were now transformed into single column:

Index	Country	Series	Year	Value
0	Afghanistan	Population, total	1960	8996973
1	Afghanistan	Population, total	1961	9169410
…	…	…	…	…
311519	Zimbabwe	GDP per capita (current US$)	2018	1430.10

Table 3 – outcome from Step 1 was a "tall and skinny" structure.

Step 2 – the final step was to pivot the data into a "short and fat" structure through use of Pandas **pivot_table** function so that Series columns now become value columns:

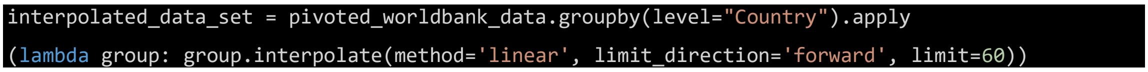
Index	Country	Year	Population, total	Life expectancy at birth, total (years)	…	GDP per capita (current US$)
0	Afghanistan	1960	8996973	32.45	…	581.13
1	Afghanistan	1961	9169410	57.37	…	209.67
…	…	…	…	…	…	…
12725	Zimbabwe	2018	13573890	57.25	…	1430.10

Table 4 – the outcome from Step 2 was a "short and fat" structure.

The data set was placed into a multi-dimensional Pandas dataframe with 20 continuous variables in columns and 12,726 rows of data. A multi-index was estbalished that encoded categorical variables such as Region and Income group as well as a hierarchical time dimension (decade and year.)

5 – Addressing Gaps in the Data

For some of the countries, there were significant gaps in the data. Rather than resort to discarding this data, I chose to apply forward linear interpolation to fill these gaps. Given that multi-indexing was in place, this was achieved through a single line of code as follows:



```
interpolated_data_set = pivoted_worldbank_data.groupby(level="Country").apply
(lambda group: group.interpolate(method='linear', limit_direction='forward', limit=60))
```

The outcome of this approach is illustrated below:

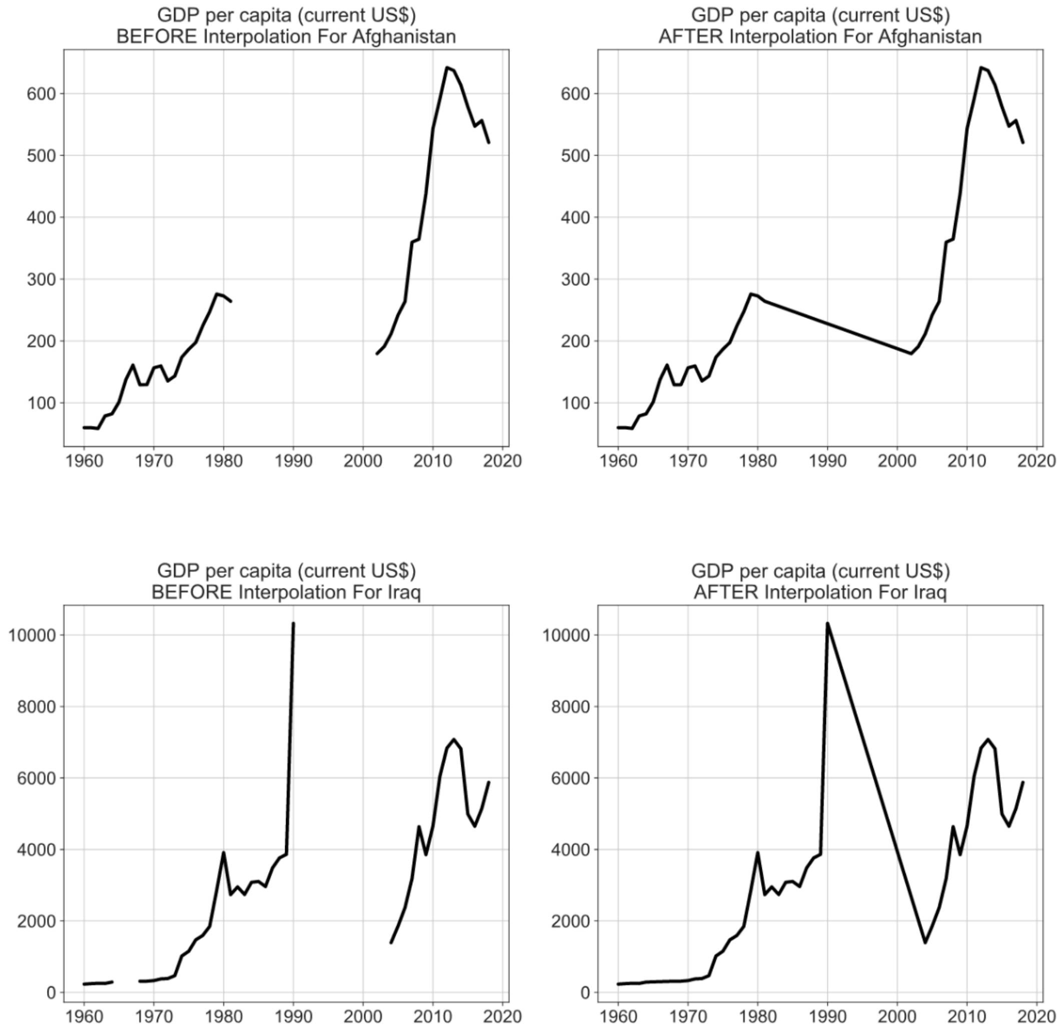


Figure 4 - pairs of line plots for Afghanistan, Iran and Iraq showing how the forward linear interpolation technique filled gaps in data. This approach was applied across all data series.

This approach enabled 13,059 empty data points to be interpolated – reducing the percentage of NaNs in the data set from 42% to 37%. See Appendix A5 – Results of Data Interpolation below on page 37 for more details.

6 – Analysis of Individual Variables

In order to get a feel for the data and to highlight interesting trends in the data, extensive analysis was carried out of the three primary continuous variables: Life Expectancy, Gross Domestic Product (GDP) per Capita and Population Growth.

A few highlights are captured in this section of the document, please refer to the associated Jupyter notebook for a full set of analytics.

Life Expectancy

Figure 5 below illustrates how Life Expectancy has developed over time. Given that only 217 data points (ie Countries) existed, the swarm plot was an effective way of visualising this data:

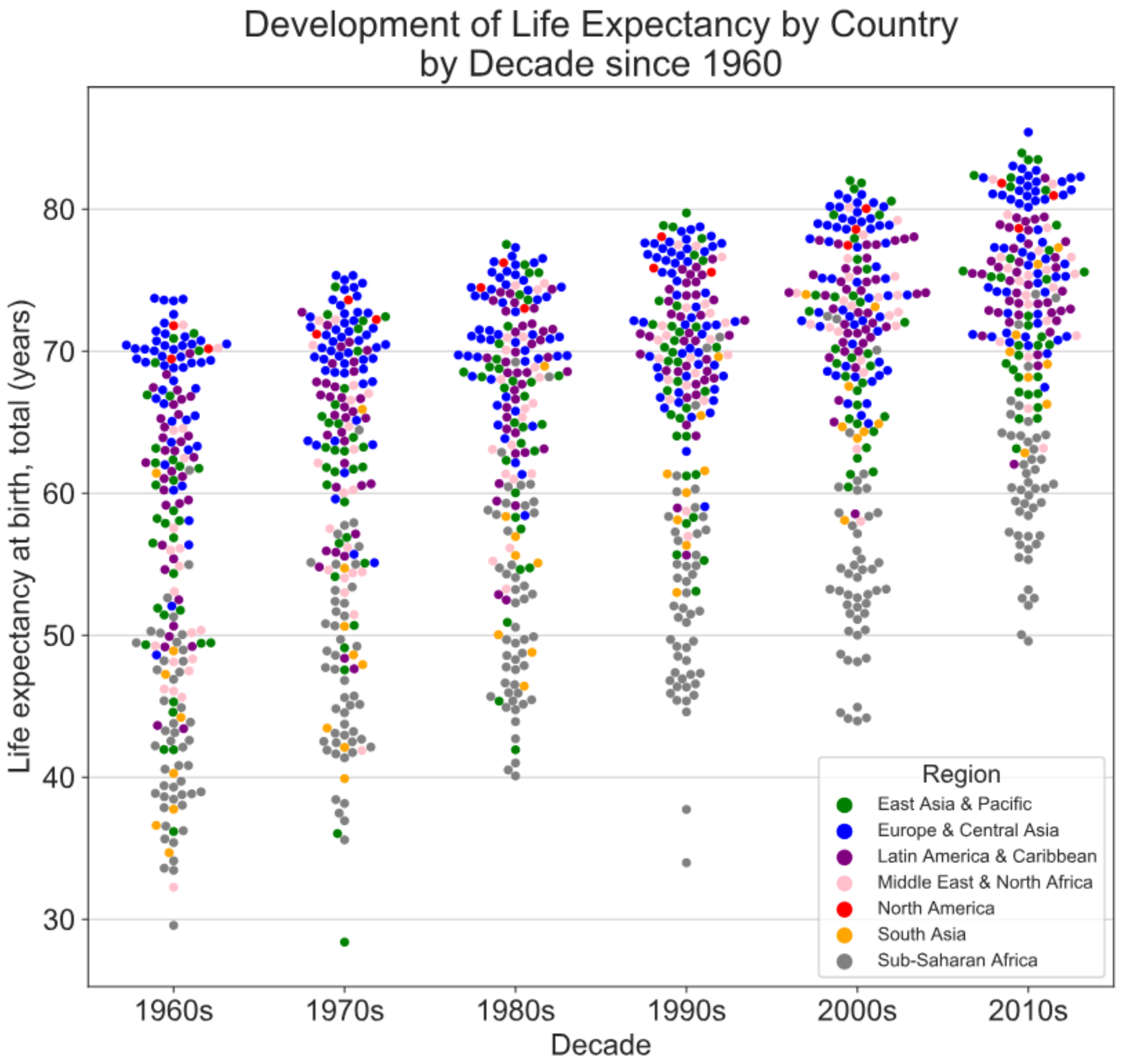


Figure 5 - swarm plot showing the average life expectancy for each country across each decade since 1960. Each point on the chart is a country, colour coded by region.

Figure 5 above shows that life expectancy has generally improved over the last 59 years and that during this time, the gap between the most healthy and least healthy countries has closed from 45 years in 1960 to 33 years in 2018.

Figure 6 below now focuses on Life Expectancy in 2018:

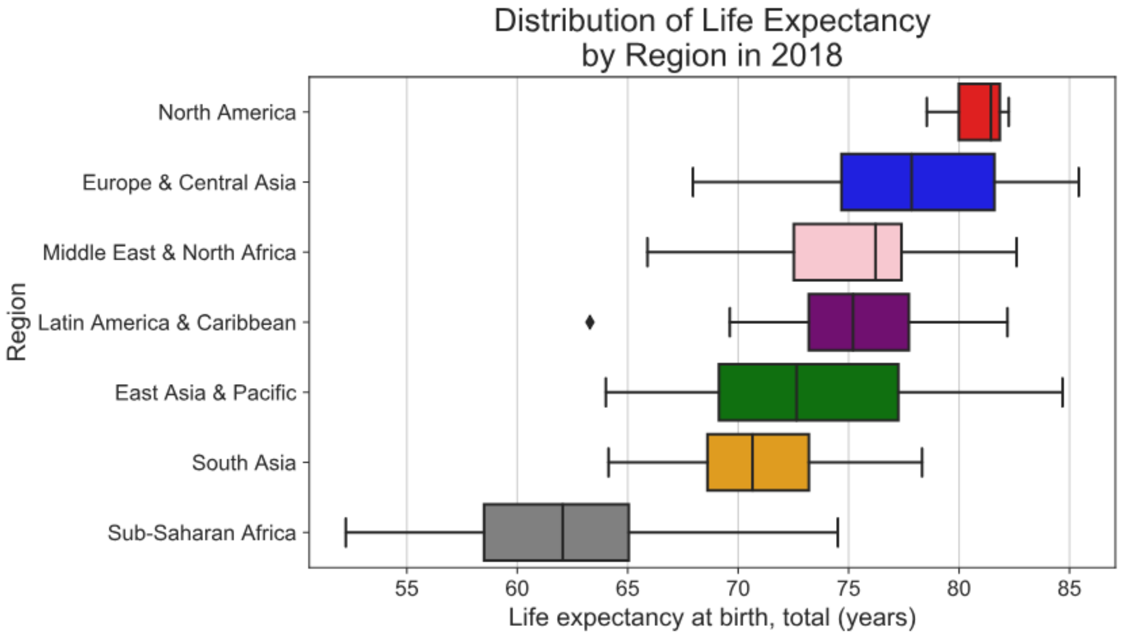


Figure 6 - boxplot showing distribution of life expectancy by Region in 2018.

One significant outlier exists in Figure 6 above for the Latin America & Caribbean region : this is **Haiti**. Research confirms that Haiti is suffering from “a domino-effect of massive natural disasters, fragile health care infrastructure and low access to preventative care in a country where half of the population lives in extreme poverty” (Borgen Project, The, 2019).

Gross Domestic Product (GDP) Per Capita

Figure 7 below illustrates how the economic prosperity of each Region has developed over the last 59 years.

One significant feature in Figure 7 below on page 9 is the dip of GDP Per Capita in 2009. Research confirms that the impact of the **2008 global financial crisis** caused “the year 2009 [to become] the first on record where global GDP contracted in real terms” and “many of the direct effects of the crisis still remain active concerns” (Chatham House, 2018).

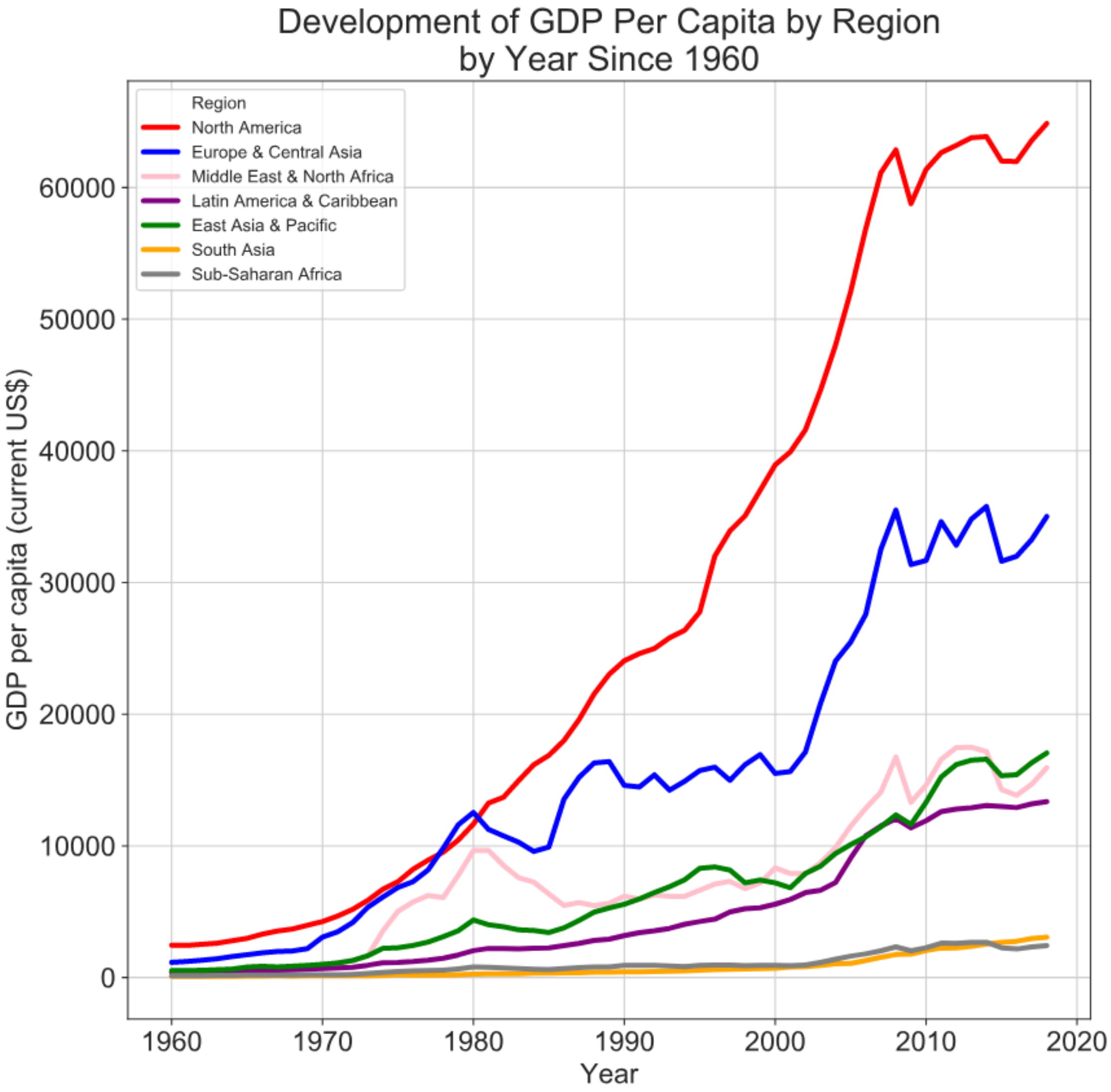


Figure 7 - development of average GDP Per Capita by region for each year since 1960. Illustrating how the gap between the richest an poorest countries has opened up exponentially over that period.

As illustrated in Figure 7 above, GDP per Capita spans 4 orders of magnitude across the entire data set. As a result, **log base 10 of GDP per Capita** was used as it enabled more meaningful visualisations and analysis.

Population Growth

The initial analysis of Population Growth is summarised in Figure 8 below:

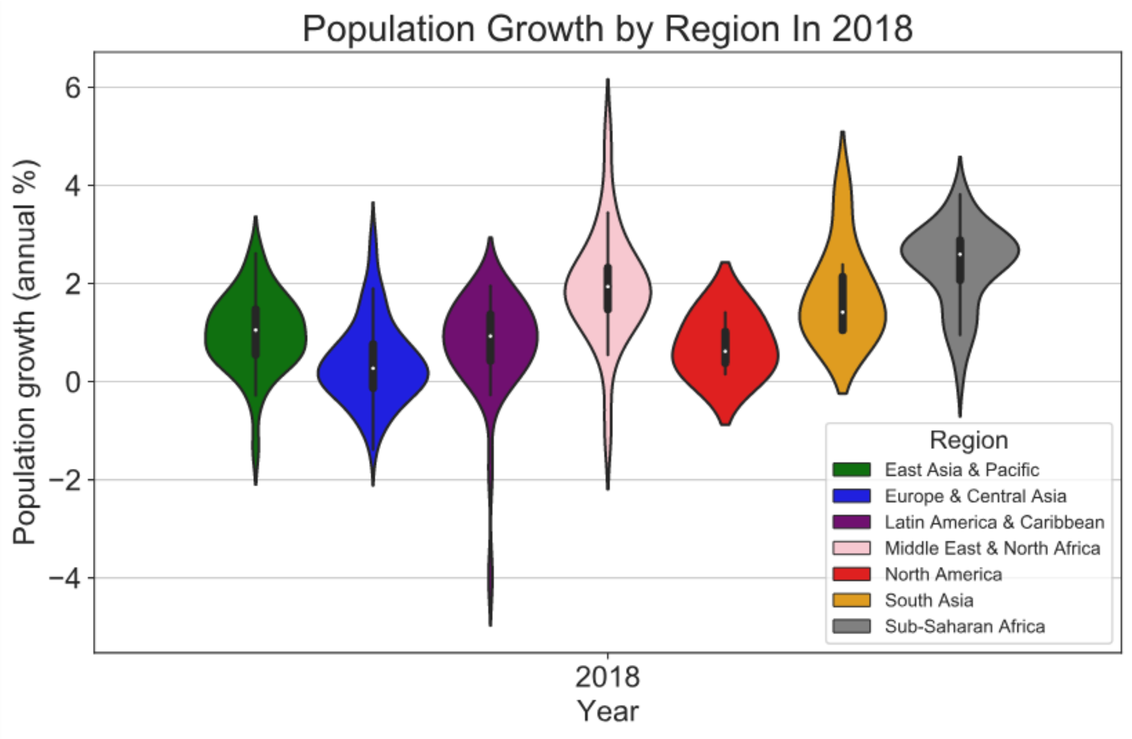


Figure 8 - distribution of population growth by region in 2018.

Note the Y axis scale in Figure 8 above : many countries are exhibiting negative population growth. **Puerto Rico** is the outlier in Latin America & Caribbean with a population growth of -3.9%. The country is facing outmigration due to “the effects of a decade-long economic recession, Puerto Ricans – who are U.S. citizens at birth – have increasingly moved to the U.S. mainland” ” (Pew Research Center, 2016). This has been further compounded when in September 2017 “hurricanes Maria and Irma hit the island” (Pew Research Centre, 2019). These trends are reflected in the World Bank WDI dataset and illustrated in Figure 9 below:

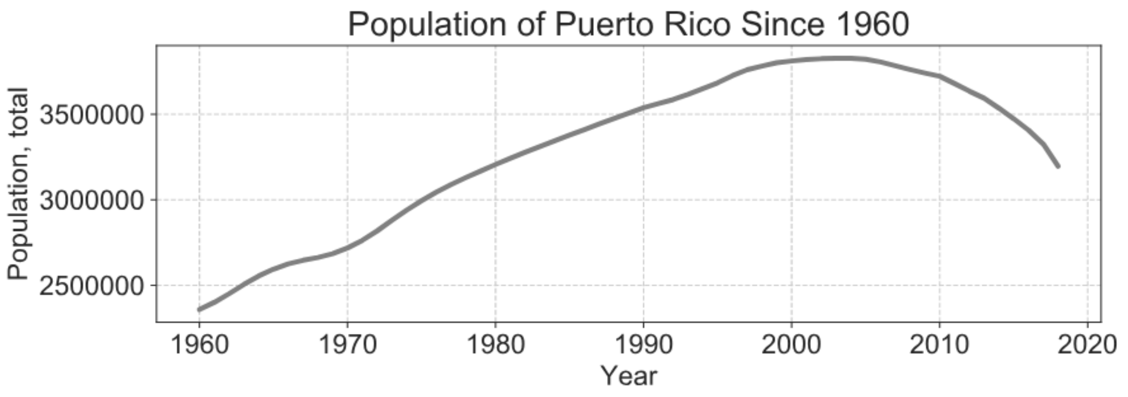


Figure 9 - this line plot clearly shows the decline in Puerto Rico's population since the mid 2000's.

7 –Analysis of Relationships Between Variables

The purpose of this stage of the process was to explore potential relationships between the range of 20 World Development Indicators (WDIs) selected from the World Banks open data set – see Appendix A3 below on page 34 for detailed descriptions of each WDI. Figure 10 below shows the initial analysis of correlations across the 20 variables:

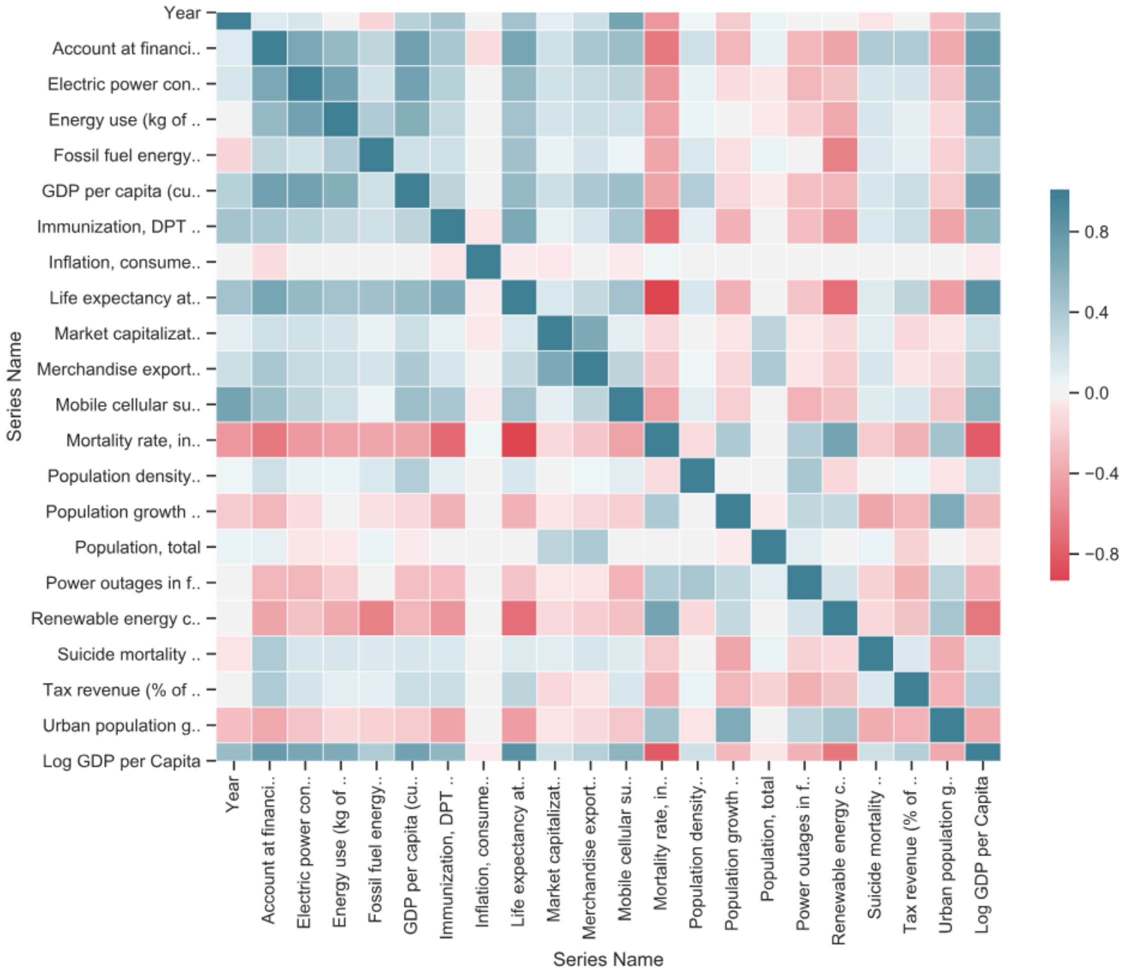


Figure 10 - heat map showing correlations between variables in the data set. Shades of green indicate positive correlation, whilst shades of red indicate negative correlation. The darker the shading, the stronger the correlation.

Figure 10 above was analysed with a focus on Life Expectancy and selecting variables that exhibited strong positive or negative correlations with it. The following three variables were selected accordingly for further analysis:

- Log GDP Per Capita;
- Mortality Rate, Infants;
- Renewable Energy Consumption.

This detailed analysis was completed using the “pair plot” in Figure 11 below, given the scale of the data set and the desire not to distort analysis by looking at multiple observations across the time dimension, this focused specifically on data for the year 2018:

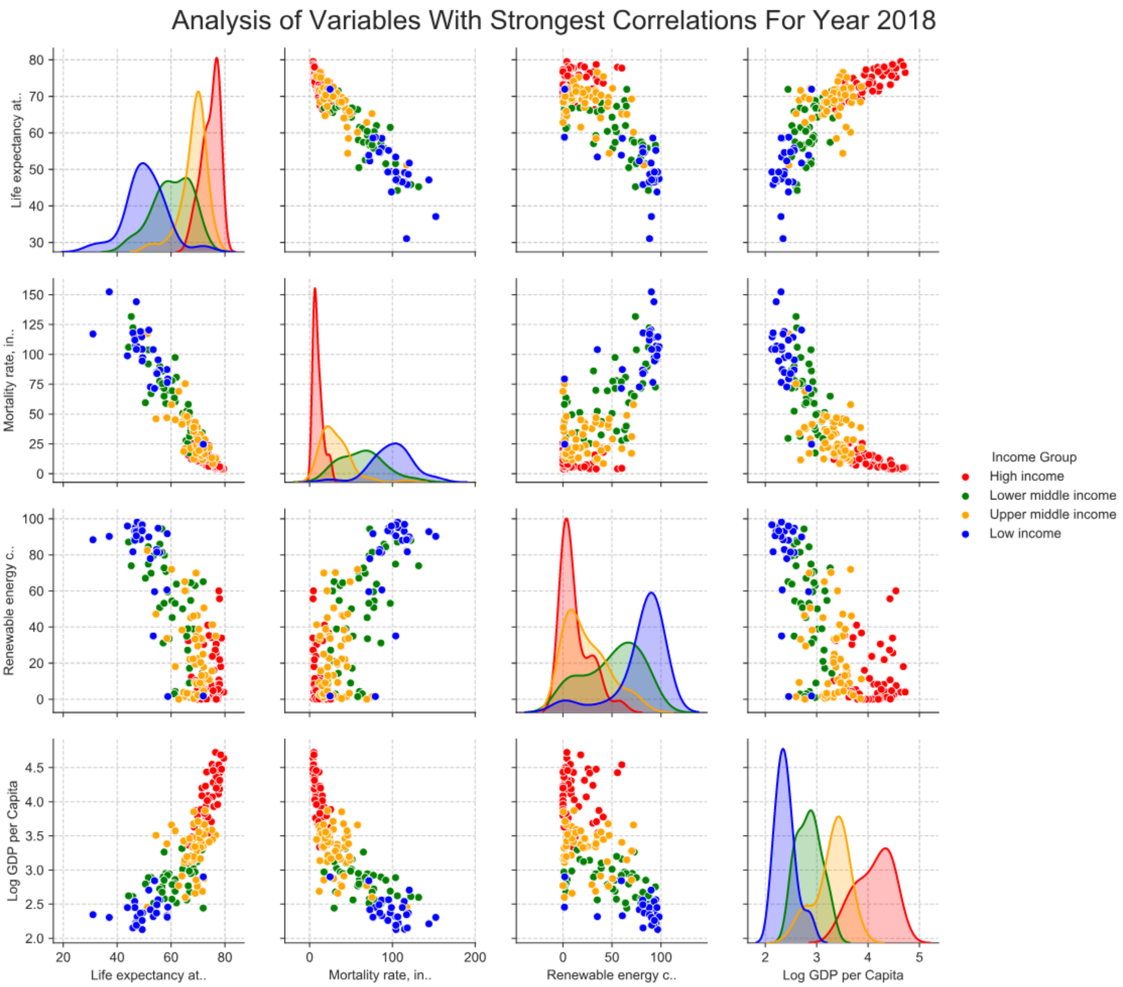


Figure 11 – pair plot showing scatter plots illustrating relationships between each pair of variables, each data point represents a country in the year 2018. Down the leading diagonal, density plots are also provided for the individual variables. Colour is used to segment data by Income Group.

The scatter chart in the top right corner of Figure 11 provides further insight into the strong positive correlation between Life Expectancy and Log GDP Per Capita.

Deeper investigation of this relationship provided an opportunity to apply Hans Rosling’s visualisation technique for this data (BBC Four, 2010). This is presented below in Figure 12 on page 13, and Figure 13 on page 14.

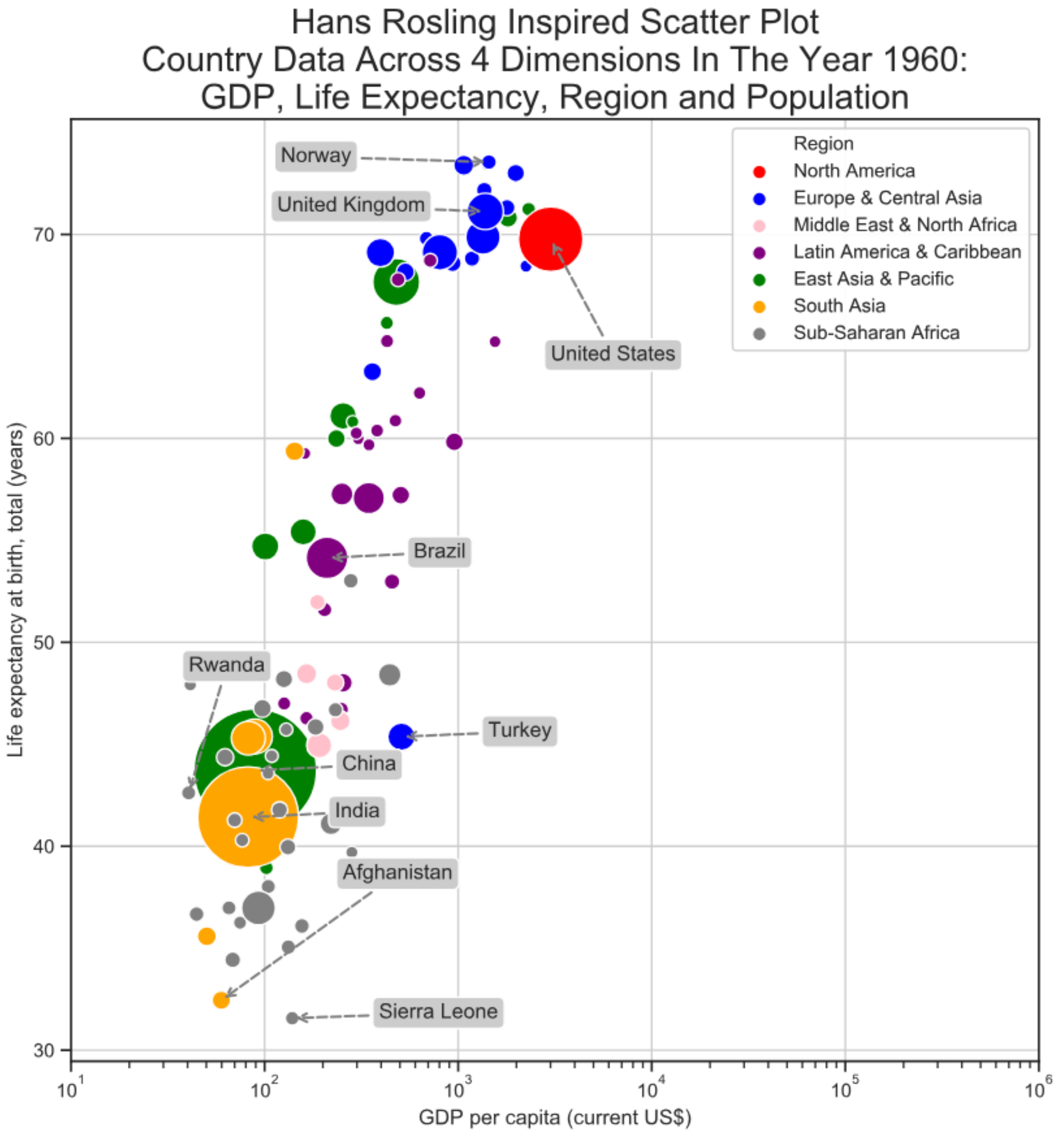


Figure 12 - scatter plot of country data from **1960** showing information from 4 dimensions: GDP (note log scale), life expectancy, population and region.

Figure 12 above captures a snapshot for the year 1960. There is clear separation between the poorest / least healthy regions of South Asia and Sub-Saharan Africa and the richest / most healthy regions of Europe & Central Asia and North America.

The spread in GDP Per Capita and Life Expectancy is line with previous analysis showing a sizeable gap between richest and poorest.

Fast forward to the year 2018, as shown in Figure 13 below on page 14, and the picture evolves.

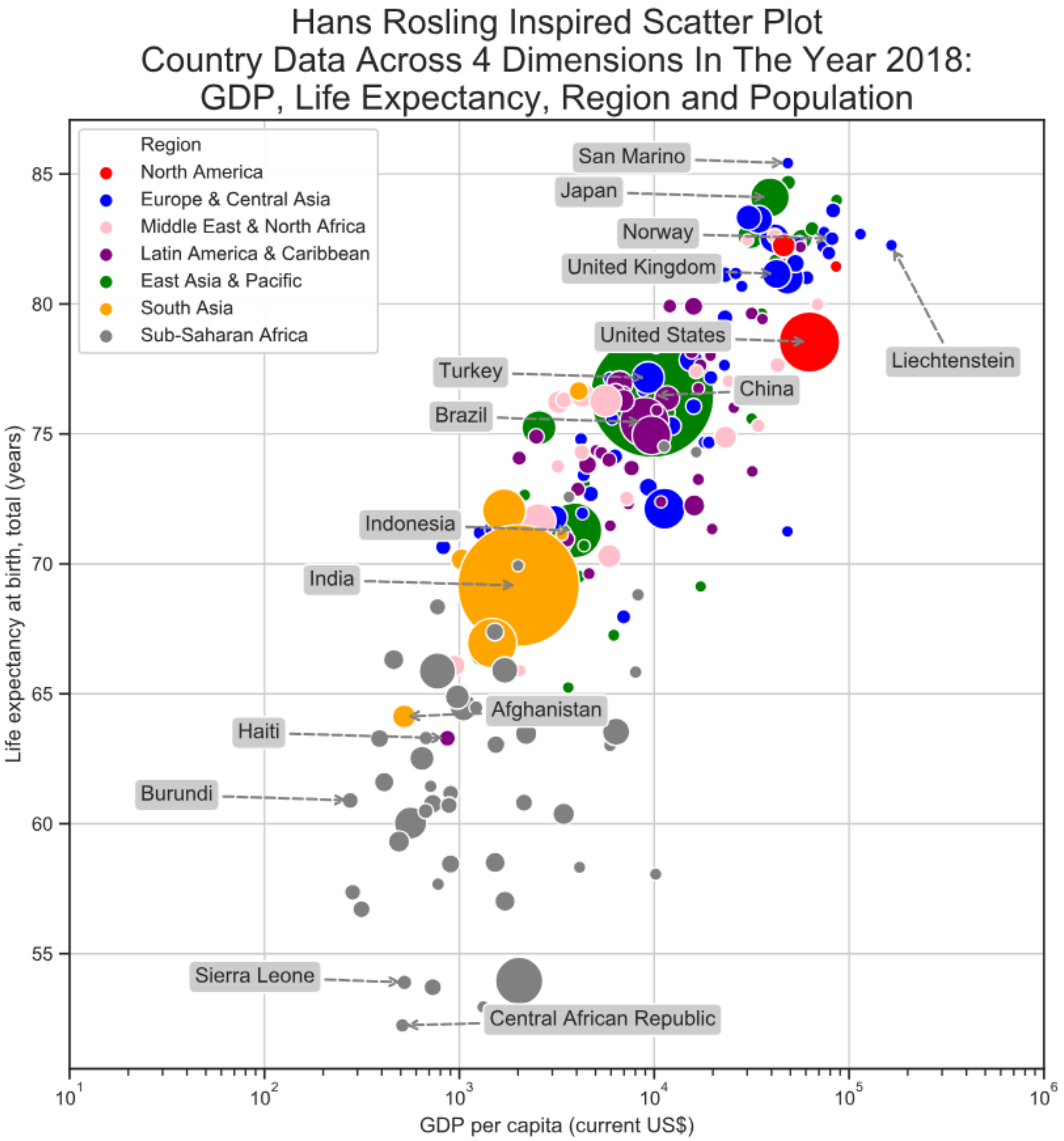


Figure 13 - scatter plot of country data from **2018** showing information from 4 dimensions: GDP (note log scale), life expectancy, population and region.

Figure 13 captures a snapshot for the year 2018 and there are some significant changes from 1960 (see Figure 12 above on page 13):

- China has closed the gap significantly with the European and North American regions;
- The United States has fallen back from its leading position in 1960;
- Afghanistan has climbed significantly from its low position in 1960;
- As we identified earlier, Haiti stands out as an outlier for the Latin American & Caribbean region.

8 – Application of Unsupervised Methods

The objective at this stage of the process was to evaluate how well clustering algorithms could replicate the "ground truth" Income Group label.

Rationale

Both agglomerative (bottom up) and K-means (top down) clustering models were evaluated - see Appendix "A6 – Comparison of Clustering Models" below on page 38 for more details. The Agglomerative model was chosen with a configuration of 4 clusters, Euclidian affinity and Ward linkage. The rationale being:

- The target of 4 clusters aligned with the "ground truth" Income Group label: ie "High Income", "Upper Middle Income", "Lower Middle Income" and "Low Income";
- This configuration of the model exhibited the best trade off in performance across Silhouette, Completeness and Homogeneity scores.

Agglomerative clustering adopts a "bottom-up" approach to clustering "beginning with every observation representing a singleton cluster. As each of the N – 1 steps (where N is the number of observations) the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level" (Hastie, 2001).

- Affinity - Euclidian – the affinity parameter defines the method through which distances between clusters are computed. In this case, Euclidean being the "straight line" distance between the points in Euclidean space;
- Linkage - Ward – linkage defines the method used to determine the closeness (or in some cases dissimilarity) between clusters. In this case "minimizes the sum of squared differences within all clusters" (Sci-kit Learn, 2019).

This is bottom up approach is illustrated in Appendix "A7 – Dendrogram" below on page 40.

Application

Before applying the clustering model to the data, the “ground truth” labelling of the data was visualised – see Figure 14 below:

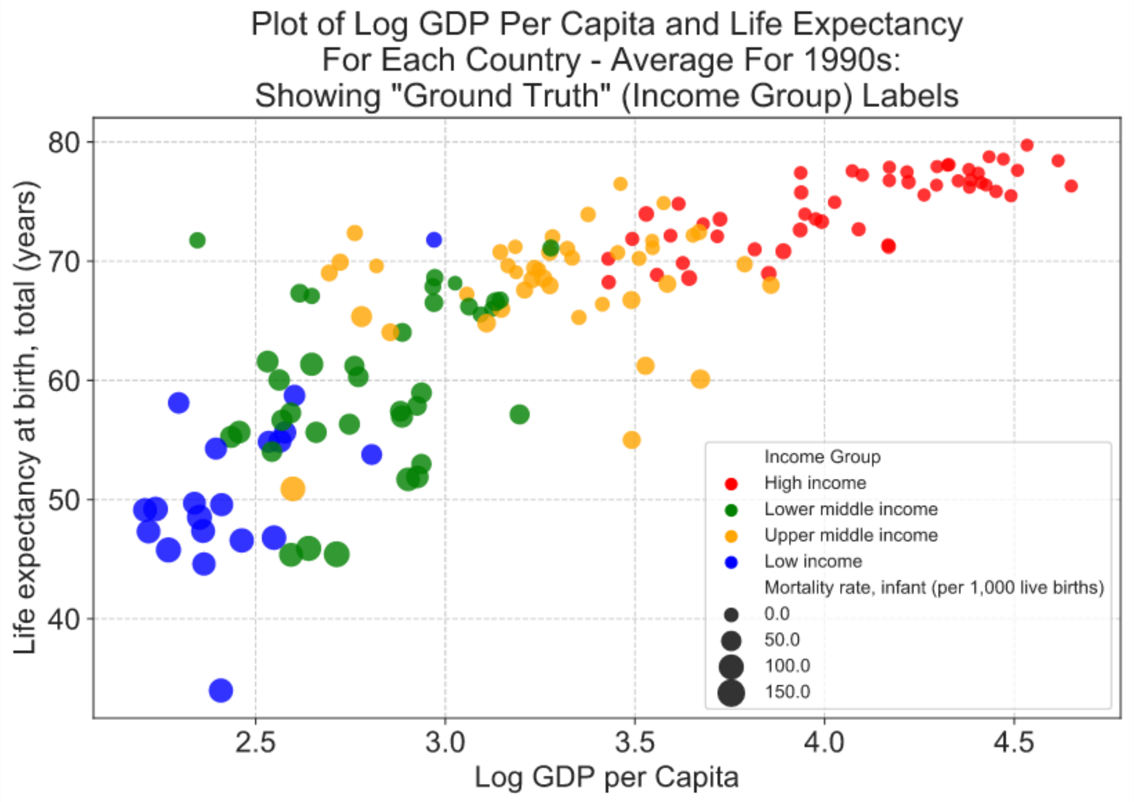


Figure 14 - "ground truth" labelling of the data.

The Agglomerative clustering model was then applied to the data and the labels assigned by the model were visualised - see Figure 15 below:

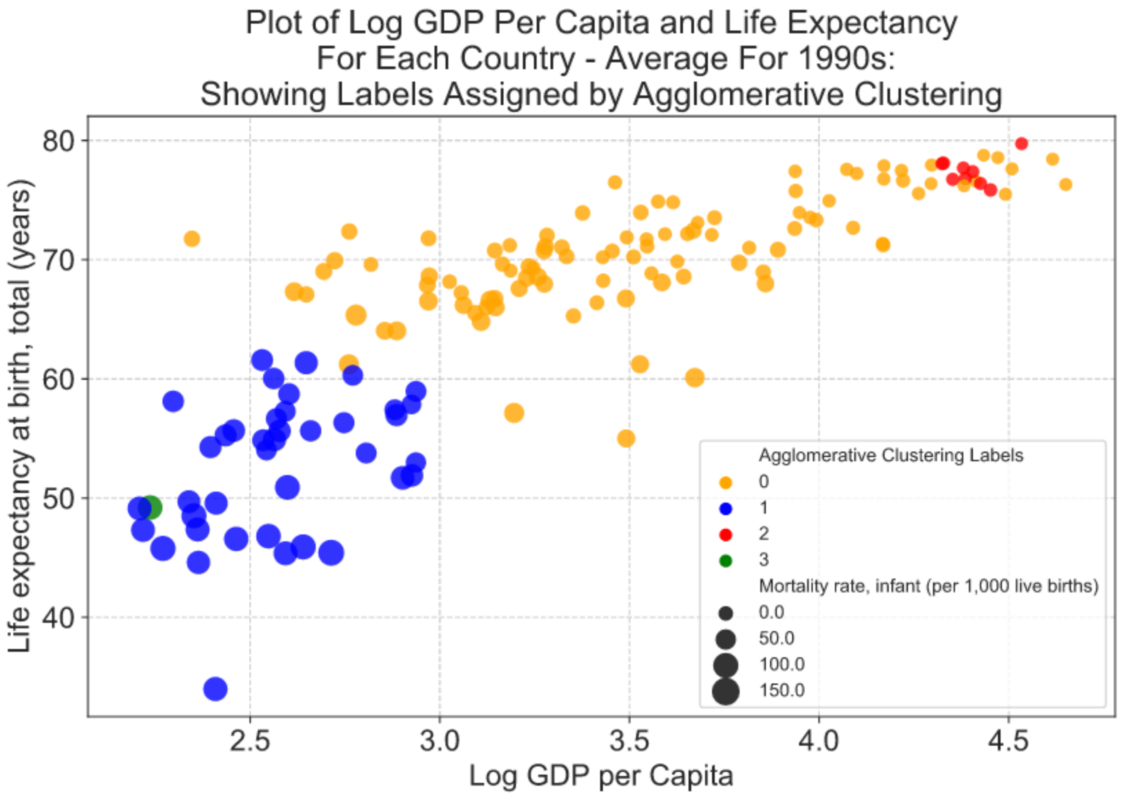


Figure 15 - visualisation of the labelling applied by the Agglomerative.

The plot above in Figure 15 illustrates reflects the fact that “Agglomerative clustering has a ‘rich get richer’ behaviour that leads to uneven cluster sizes.” (Sci-kit Learn, 2019). The outcome being that two of the four clusters (labels 0 and 1) dominate.

As a result, the model struggled to find four distinct clusters. Whilst it showed reasonable success at identifying clusters at extremes (High Income and Low Income) it struggle to achieve separation – in particular in the areas where Upper and Lower Middle Incomes applied,

This was reflected in scores achieved by the model (as set out in Table 5 below). All three metrics would need to be closer to 1 to indicate well defined dense clusters had been located in the data:

Silhouette Score	0.435
Completeness Score	0.501
Homogeneity Score	0.309

Table 5 - scores achieved for agglomerative clustering, configured to seek 4 clusters using Euclidian affinity and Ward linkage.

The next step was to visualise the performance of the model by highlighting points where the clustering agreed with the “ground truth” and those where a fit was not found - see Figure 16 below.

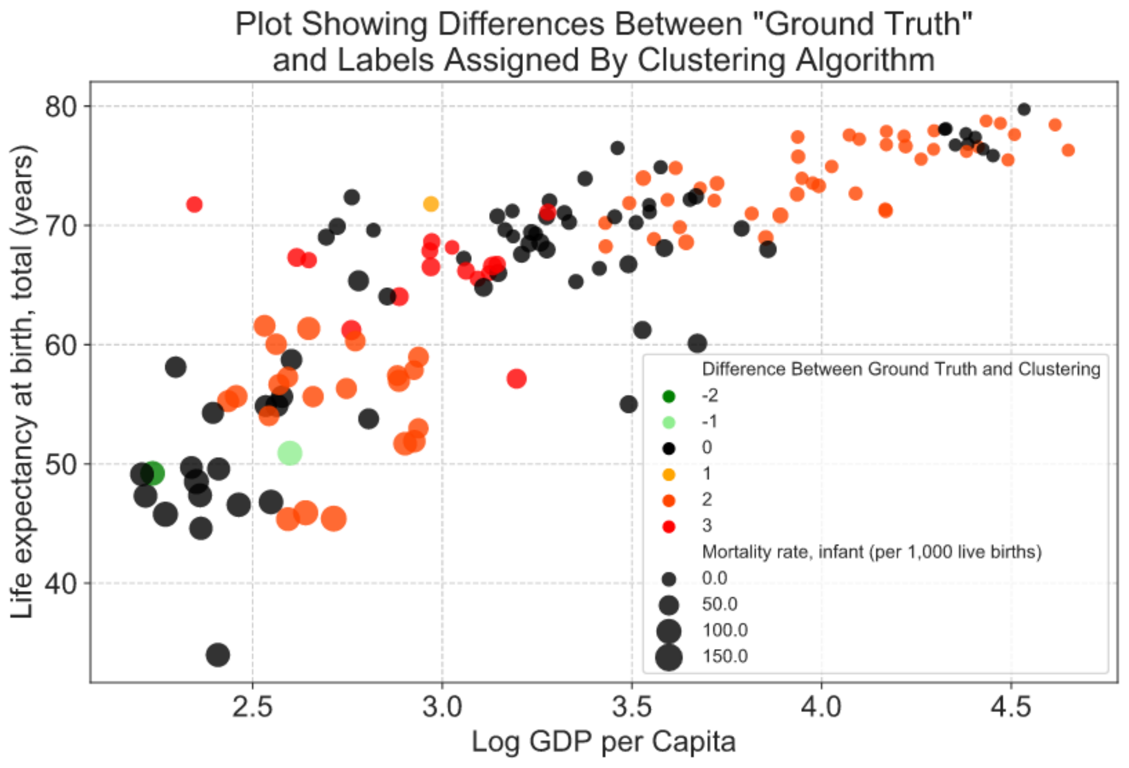


Figure 16 - black points show where model and "ground truth" were aligned. Coloured points show mis-matched labels.

Finally a Dendrogram was generated to visualise how the agglomerative clustering algorithm had constructed the clusters - see Appendix “A7 – Dendrogram” below on page 40.

Findings

Figure 16 above led to the conclusion that the application of clustering is problematic because the data does not exhibit clusters that are compact and well separated from others. This is reflected in the Silhouette Scores which failed to exceed 0.5 across all models and model parameter configurations evaluated.

9 – Application of Supervised Methods

The objective at this stage of the process was to evaluate Linear Regression as a method of predicting future Life Expectancy. “Linear regression is a popular regression lerning algorithm that learns a model which is linear combination of features of the input example” (Burkov, 2019)

Rationale

The Linear Regression model was chosen because Life Expectancy is a continuous (as opposed to categorical) variable. Furthermore, previous analysis indicated strong correlation between Life Expectancy and other continuous variables that could be exploited to train this type of model.

Application

The approach taken is illustrated below in Figure 18 below on page 21. The year 1990 was selected to train and test the model. This achieved a mean squared error of 8.35 and an R^2 error[1] of 0.878. The model found a fit with an intercept of 69.61 and coefficients as set out below in Table 6:

Feature	Coefficient
Immunization, DPT (% of children ages 12-23 months)	-0.013511
Mortality rate, infant (per 1,000 live births)	-0.195201
Renewable energy consumption (% of total final energy consumption)	-0.028483
Log GDP per Capita	2.145506

Table 6 - coefficients achieved by training linear regression model on data from 1990.

The trained model was then applied to all future years, and the resulting mean squared errors were calculated. The output from this is captured below in Figure 17.

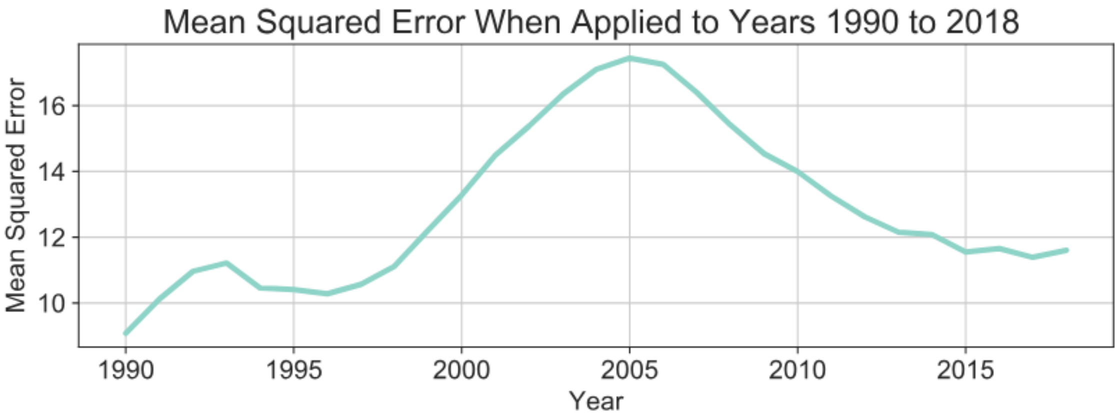


Figure 17 - chart showing mean squared effor for predicted versus actual Life Expectnacy. Model was trained using data for 1990 and then applied to all future years in the data set.

[1] "the coefficient of determination, denoted R^2 or r^2 and pronounced 'R squared'. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model." (Wikipedia, 2019)

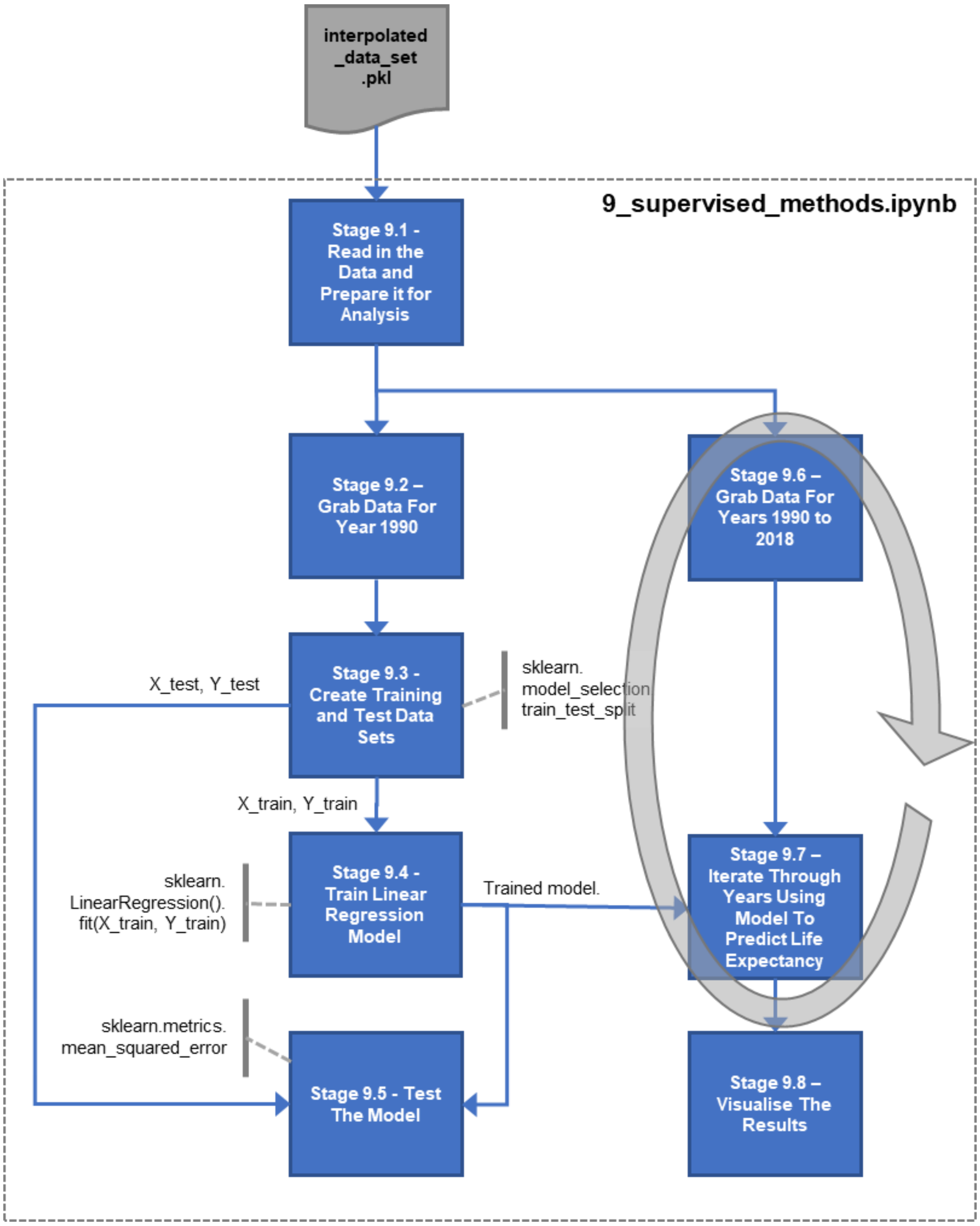

interpolated _data_set .pkl
9_supervised_methods.ipynb
Stage 9.1 - Read in the Data and Prepare it for Analysis
Stage 9.2 – Grab Data For Year 1990
Stage 9.3 - Create Training and Test Data Sets
sklearn. model_selection. train_test_split
X_test, Y_test
X_train, Y_train
Stage 9.4 - Train Linear Regression Model
sklearn. LinearRegression(). fit(X_train, Y_train)
Trained model.
sklearn.metrics. mean_squared_error
Stage 9.5 - Test The Model
Stage 9.6 – Grab Data For Years 1990 to 2018
Stage 9.7 – Iterate Through Years Using Model To Predict Life Expectancy
Stage 9.8 – Visualise The Results

Figure 18 - process followed to train, test and run the linear regression model to predict Life Expectancy.

Findings

Figure 19 below on page 22 indicates that the linear regression model appears to be reasonably successful at predicting Life Expectancy across the full range of future years in the data set. However, closer inspection of Figure 17 above on page 20 shows that the mean squared error increases significantly after 1996. This would indicate that the reliability of this model reduces after a horizon of 5 or 6 years.

In practical terms, if this type of model was put into production, it would be recalibrated on an annual basis (based on availability of new data each year) to sustain the most accurate forward predictions. A further benefit of adopting a linear regression model is that is highly explainable. This would enable shifts in model fit intercept and coefficients to be examined year on year, providing further insights into the development of the relationship between the chosen variables and Life Exectancy.

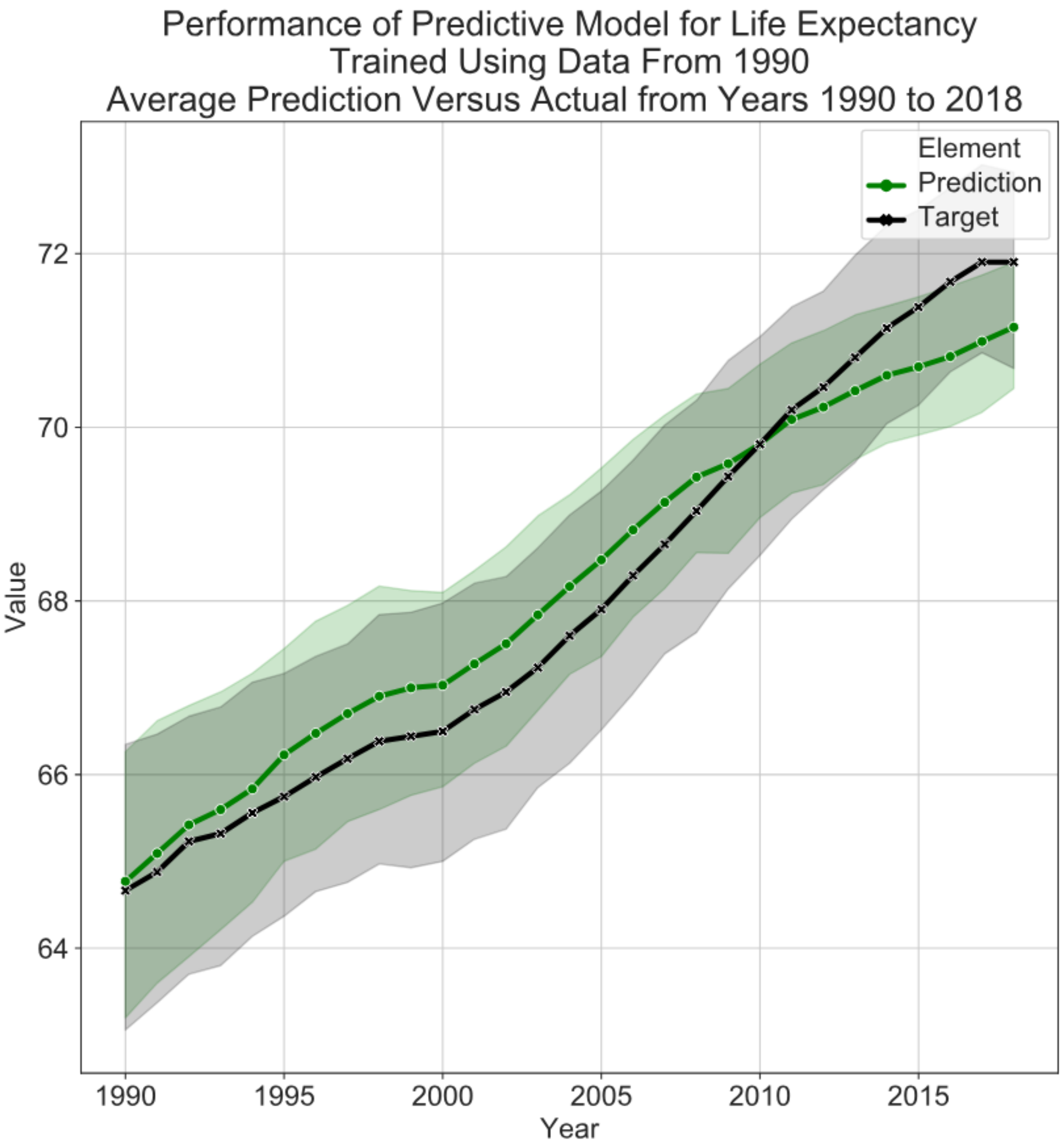


Figure 19 - chart showing how well the model performed when compared to actual results. Solid lines show the mean, the shaded bands indicate the distribution of data in each year - one data point was available per country in each year.

10 - Conclusions

Assessment Against Problem Definition

Table 7 below summarises the findings against the Problem Definition:

#	Question	Finding
1	How has the gap in economic prosperity between the poorest and richest countries developed over time?	Figure 7 (above on page 9) provides evidence that the gap between the poorest and richest countries has opened up significantly over the last 59 years : by 3 orders of magnitude from circa $3,000 in 1960 to circa $166,000 in 2018.
2	Does the economic prosperity of a country have an impact on the health of its citizens?	The chart in the top right corner of Figure 11 (above on page 12) and the subsequent Hans Rosling inspired plots (see next question) provide strong evidence that richer countries have healthier citizens.
3	Could visualisation techniques be used to enable people to "acquire an evidence based world view" (Rosling, et al., 2005)?	The Hans Rosling inspired scatter plots (Figure 12 on page 13, and Figure 13 on page 14) were shown to friends and family – all became engaged in the content, and agreed that they had gained deeper insights into the historic development of global wealth and health.
4	Could unsupervised methods be used to cluster countries according to their relative economic standing?	As reinforced by Figure 16 (above on page 18) and the associated scores, the application of a clustering algorithm was problematic given that the data does not inherently exhibit well defined, dense clusters.
5	Could a regression model be trained to enable future life expectancy to be reliably predicted based on economic growth assumptions?	The application of a simple linear regression model demonstrated that predictive models could be created – as illustrated in Figure 19 above on page 22.

Table 7 - summary of conclusions against original questions raised.

Reflection

In addition to the findings above, the points below capture observations made during the end to end process:

1. This was a rich data set. I only managed to scratch the surface of the insights that could be generated from it. Further research could be conducted into global health and wealth, as well as other important topics such as global energy usage trends, and projecting future global population growth;

2. It was surprising how much time (~25%) was required to "wrangle" the data into the right shape. But the effort paid off – for example, the implementation of a

multi-index on the Pandas data frames made the task of slicing, interpolating and charting data significantly easier;

3. The implementation of functions in key areas of the assignment helped to keep the code and Jupyter notebooks condense, as well as making analysis consistent and reliable;
4. Wrapping key stages of analysis in loops, stepping through different configurations, collecting the data at each and then applying data analysis to that provided useful insights that allowed optimum models and model parameters be identified;
5. The use of modern IDE[2] and source control significantly improved productivity;
6. A more interactive experience would be beneficial for presenting the data : for example the ability to animate data using the time dimension, or to enable users to interact with the data. This would be an interesting follow up, and could be enabled by tools such as Python widgets or Microsoft's PowerBI.

[2] IDE – integrated development environment. A software application that is designed to optimise the productivity and quality of work generated by software engineers.

References

Bank of England, 2019. *What is GDP?.* [Online]
Available at: https://www.bankofengland.co.uk/knowledgebank/what-is-gdp
[Accessed 10 11 2019].

BBC Four, 2010. *Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four.* [Online]
Available at: https://www.youtube.com/watch?v=jbkSRLYSojo
[Accessed 10 11 2019].

Borgen Project, The, 2019. *Top 10 Facts About Life Expectancy In Haiti.* [Online]
Available at: https://borgenproject.org/top-10-facts-about-life-expectancy-in-haiti/
[Accessed 10 11 2019].

Burkov, A., 2019. Linear Regression. In: *The Hundred-Page Machine Learning Book.* s.l.:Andriy Burkov, pp. 21-25.

Chatham House, 2018. *The Lasting Effects of the Financial Crisis Have Yet to Be Felt.* [Online]
Available at: https://www.chathamhouse.org/expert/comment/lasting-effects-financial-crisis-have-yet-be-felt
[Accessed 10 11 2019].

Hastie, T. &. F., 2001. *The Elements of Statistical Learning.* s.l.:Springer.

Investopedia, 2019. *https://www.investopedia.com/.* [Online]
Available at: https://www.investopedia.com/terms/g/gdp.asp
[Accessed 10 11 2019].

Markdown Guides, 2019. *Free and open-source reference guide that explains how to use Markdown.* [Online]
Available at: https://www.markdownguide.org/
[Accessed 10 11 2019].

Pew Research Center, 2016. *Historic population losses continue across Puerto Rico.* [Online]
Available at: https://www.pewresearch.org/fact-tank/2016/03/24/historic-population-losses-continue-across-puerto-rico/
[Accessed 10 11 2019].

Pew Research Centre, 2019. *Puerto Rico's population declined sharply after hurricanes Maria and Irma.* [Online]
Available at: https://www.pewresearch.org/fact-tank/2019/07/26/puerto-rico-population-2018/
[Accessed 10 11 2019].

Rosling, H., Rosling, A. & Rosling, O., 2005. New Software Brings Statistics Beyond the Eye. *"Statistics, Knowledge and Policy", the first OECD World Forum on Key Indicators,* p. 522 to 530.

Sci-kit Learn, 2019. *Clustering User Guide.* [Online]
Available at: https://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness
[Accessed 10 11 2019].

Wikipedia, 2019. *Coefficient of determination.* [Online]
Available at: https://en.wikipedia.org/wiki/Coefficient_of_determination
[Accessed 10 11 2019].

World Bank, The, 2019. *World Development Indicators.* [Online]
Available at: https://data.worldbank.org/
[Accessed 10 11 2019].

Zumel, N. & Mount, J., 2014. *Practical data science with R.* s.l.:Shelter Island, NY : Manning Publications Co..

Figures

Figure 1 - frame captured from the YouTube video. 1

Figure 2 - illustration of major stages of work undertaken to complete the assignment. Stages highlighted in blue correspond to a discrete Python files. 1

Figure 3 - overview of end to end data flow. 2

Figure 4 - pairs of line plots for Afghanistan, Iran and Iraq showing how the forward linear interpolation technique filled gaps in data. This approach was applied across all data series....................6

Figure 5 - swarm plot showing the average life expectancy for each country across each decade since 1960. Each point on the chart is a country, colour coded by region....................7

Figure 6 - boxplot showing distribution of life expectancy by Region in 2018. 8

Figure 7 - development of average GDP Per Capita by region for each year since 1960. Illustrating how the gap between the richest an poorest countries has opened up exponentially over that period. 9

Figure 8 - distribution of population growth by region in 2018....................10

Figure 9 - this line plot clearly shows the decline in Puerto Rico's population since the mid 2000's. 10

Figure 10 - heat map showing correlations between variables in the data set. Shades of green indicate positive correlation, whilst shades of red indicate negative correlation. The darker the shading, the stronger the correlation. 11

Figure 11 – pair plot showing scatter plots illustrating relationships between each pair of variables, each data point represents a country in the year 2018. Down the leading diagonal, density plots are also provided for the individual variables. Colour is used to segment data by Income Group. 12

Figure 12 - scatter plot of country data from **1960** showing information from 4 dimensions: GDP (note log scale), life expectancy, population and region. 13

Figure 13 - scatter plot of country data from **2018** showing information from 4 dimensions: GDP (note log scale), life expectancy, population and region. 14

Figure 14 - "ground truth" labelling of the data....................16

Figure 15 - visualisation of the labelling applied by the Agglomerative. 17

Figure 16 - black points show where model and "ground truth" were aligned. Coloured points show mis-matched labels....................18

Figure 17 - chart showing mean squared effor for predicted versus actual Life Expectnacy. Model was trained using data for 1990 and then applied to all future years in the data set. 20

Figure 18 - process followed to train, test and run the linear regression model to predict Life Expectancy.21

Figure 19 - chart showing how well the model performed when compared to actual results. Solid lines show the mean, the shaded bands indicate the distribution of data in each year - one data point was available per country in each year. 22

Figure 20 – illustration of high level components used to develop the Python code and Jupyter notebooks for this assignment....................30

Figure 21 - screen shot of World Bank data portal....................31

Figure 22 - screen shot showing panel on left used to choose data points to download....................32

Figure 23 - screen shot of "advanced download options" used to generate CSV files. 33

Figure 24 - visualisation of the performance of the K-means algorithm as clusters are stepped up from 2 to 10. There is a dramatic step change in both Silhouette and Homogeneity Score as between 4 and 5 clusters. 39

Tables

Table 1 - the five questions set at the start of the assignment to frame the problem. 3

Table 2 - initial data structure with years as columns...5

Table 3 – outcome from Step 1 was a "tall and skinny" structure. 5

Table 4 – the outcome from Step 2 was a “short and fat” structure. 5

Table 5 - scores achieved for agglomerative clustering, configured to seek 4 clusters using Euclidian affinity and Ward linkage..17

Table 6 - coefficients achieved by training linear regression model on data from 1990. 20

Table 7 - summary of conclusions against original questions raised. 23

Table 8 - overview of the sub-set of World Bank “Economic Development Indicator” open data selected for this assignment. 34

Table 9 - definitions for all 20 World Development Indicators (WDIs) analysed in this assignment. 36

Table 10 – summary of the impact of forward linear interpolation on the data set. 37

Table 11 - table showing relative performance of Agglomerative model with target number of clusters fixed at 4. All possible combinations of "Affinity" and "Linkage" parameters were explored. 38

Table 12 - table showing performance of K-means clustering as number of target clusters is stepped up from 2 to 10. The row of specific interest is that with 4 clusters as this matches the number of clusters in the “ground truth” label Income Group. 39

Appendix

A1 - Design Decisions

The following principles were adopted:

- **Loosely coupled logical architecture** - each stage of the process illustrated in blue in Figure 2 (above on page 1) was captured as a discrete **Jupyter** notebook. The output from one notebook was used as the input for the next through use of the Pandas **pikle** file format in order to preserve metadata;

- **Strictly no use of spreadsheets** – no pre-processing of data was performed before data was ingested into a Python. Furthermore, some options could have been chosen in the World Bank web site to make the data wrangling simpler, but I wanted to prove that Python was capable of addressing these issues itself;

- **Documentation** – extensive use was made of **Markdown** (Markdown Guides, 2019) in each notebook to provide description of steps and to capture observations;

- **Flexible code** – the data structures implemented and code were designed to enable new cuts of the data to be ingested from source without breaking downstream analysis

- **Multi-index** – use of a hierarchical index was used to index the core Pandas dataframe. This made actions such as interpolating, slicing, averaging and charting data strightforward;

- **Promoting Re-use** - blocks of re-usable code were written as functions. Some of which were held in a separate "helper_functions.py" Python file so that they could imported into each notebook;

- **Modern development tools –** a suite of modern tools were used to build the project. This included the open source Visual Studio Code IDE to maximise productivity and GitHub for source control.

A2 – Development Environment

The following software tools, libraries and Cloud services were used to complete the assignment:

Python version 3.7.4. **Python libraries** used:

- Arrays, dataframes and general data wrangling:
    - Numpy
    - Pandas
- Data visualisation:
    - Matplotlib
    - Seaborn
    - Scipy (Dendrogram)
- Data preparation and machine learning:
    - Sci-Kit Learn – including: scaling, label encoding, clustering algorithms, generation of metrics, linear regression and naïve Bayes.

Visual Studio Code used as IDE[3] - 1.39.2 with the following extensions:

- Python – enables code to be run in Python Interactive window.

Extensive use of **Markdown** cells within Jupyter to describe process and capture thoughts and actions.

GitHub used as source control – enabled code to be secured on the cloud, code to be easily copied to PCs in the lab and for an audit trail of changes to be maintained:

https://github.com/Vesperpiano/CS982/tree/master/assignment1

Jupyter notebooks were generated from Visual Studio Code Python Interactive Window and hosted on Azure Notebooks:

My backlog of "to do" tasks were maintained on a **Trello** Kanban board.

I used **Slack** to provide an integrated view across GitHub and Trello and to provide a place for capturing further notes and thoughts.

[3] Integrated Development Environment

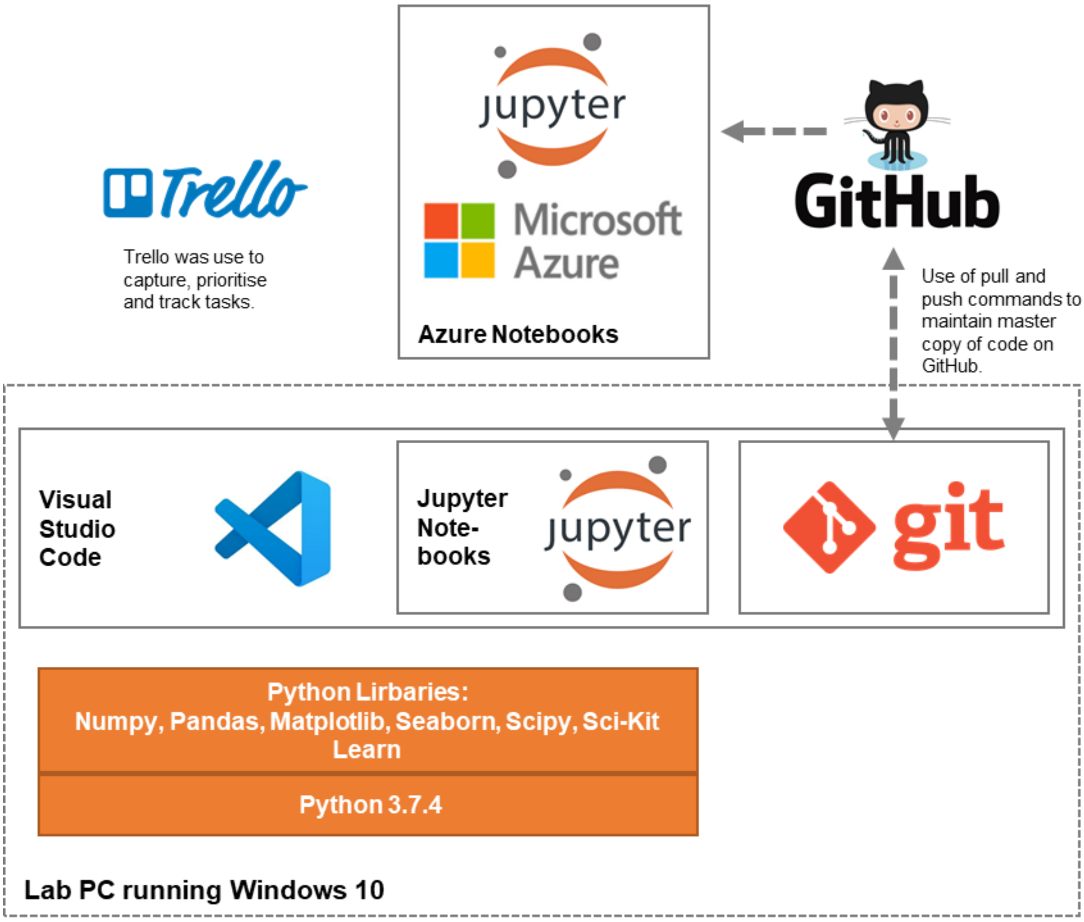
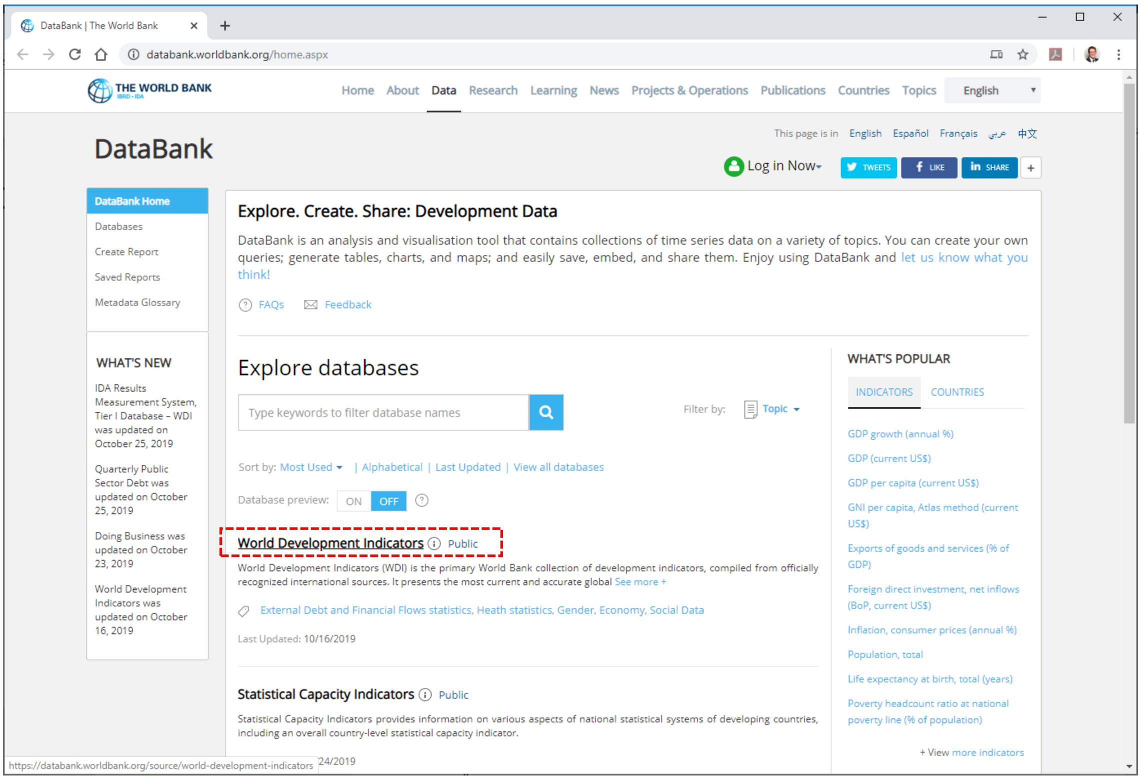

Trello
Trello was use to capture, prioritise and track tasks.
jupyter
Microsoft Azure
Azure Notebooks
GitHub
Use of pull and push commands to maintain master copy of code on GitHub.
Visual Studio Code
Jupyter Note-books
jupyter
git
Python Lirbaries:
Numpy, Pandas, Matplotlib, Seaborn, Scipy, Sci-Kit Learn
Python 3.7.4
Lab PC running Windows 10

Figure 20 – illustration of high level components used to develop the Python code and Jupyter notebooks for this assignment.

A3 – Process to Acquire Data From World Bank

Step 1 – open the World Bank's databank web site: https://databank.worldbank.org/

Step 2 – select the "World Development Indicators" database:



DataBank

Explore. Create. Share: Development Data

DataBank is an analysis and visualisation tool that contains collections of time series data on a variety of topics. You can create your own queries; generate tables, charts, and maps; and easily save, embed, and share them. Enjoy using DataBank and let us know what you think!

WHAT'S NEW

IDA Results Measurement System, Tier I Database – WDI was updated on October 25, 2019

Quarterly Public Sector Debt was updated on October 25, 2019

Doing Business was updated on October 23, 2019

World Development Indicators was updated on October 16, 2019

Explore databases

World Development Indicators Public

World Development Indicators (WDI) is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global See more +

External Debt and Financial Flows statistics, Heath statistics, Gender, Economy, Social Data

Last Updated: 10/16/2019

Statistical Capacity Indicators Public

Statistical Capacity Indicators provides information on various aspects of national statistical systems of developing countries, including an overall country-level statistical capacity indicator.

24/2019

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

WHAT'S POPULAR

- GDP growth (annual %)
- GDP (current US$)
- GDP per capita (current US$)
- GNI per capita, Atlas method (current US$)
- Exports of goods and services (% of GDP)
- Foreign direct investment, net inflows (BoP, current US$)
- Inflation, consumer prices (annual %)
- Population, total
- Life expectancy at birth, total (years)
- Poverty headcount ratio at national poverty line (% of population)

Figure 21 - screen shot of World Bank data portal.

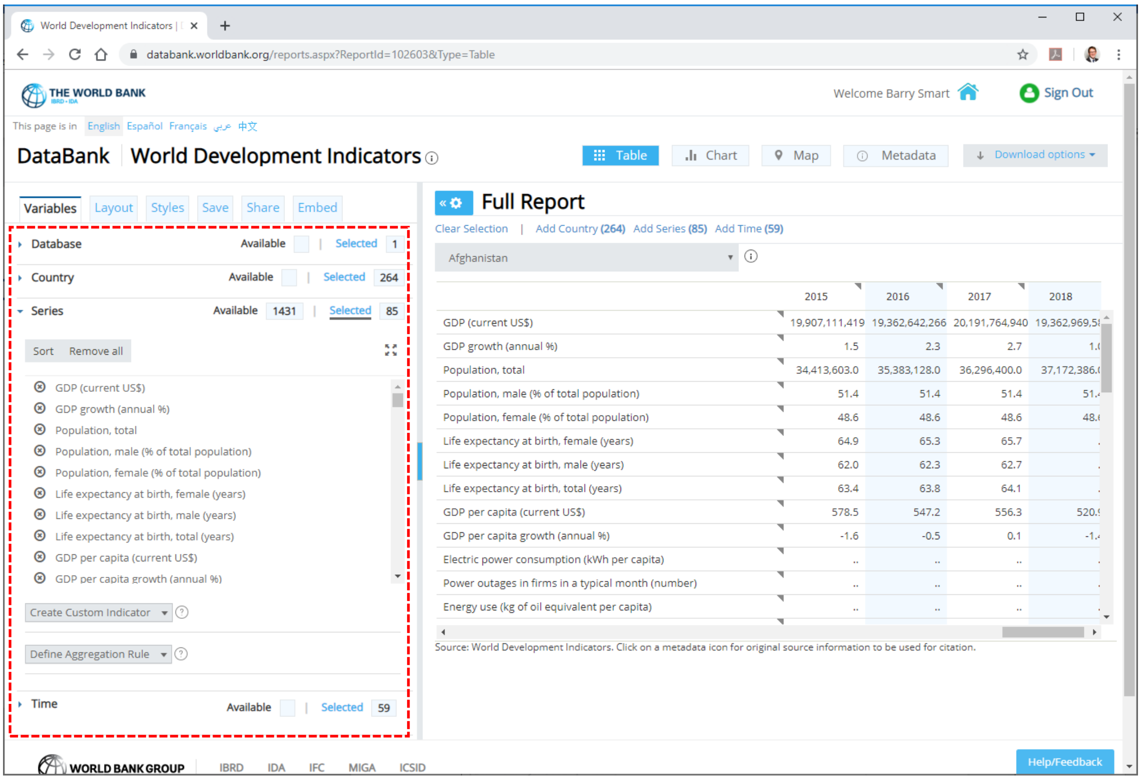
Step 3 – use the panel on the left had side to select:

- Country – all 264 were selected;
- Time – all 59 years were selected;
- Series – a subset of 85 were selected – this included:
    - **Gross Domestic Product (GDP) per capita** – "Gross Domestic Product (GDP) is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. As a broad measure of overall domestic production, it functions as a comprehensive scorecard of the country's economic health." (Investopedia, 2019);
    - **Life Expectancy at birth**– "Life expectancy at birth used here is the average number of years a newborn is expected to live if mortality

> patterns at the time of its birth remain constant in the future." (World Bank, The, 2019).

Note – a user account can be created to enable report configurations to be saved, thus avoiding the need to perform Step 3 each time.

Step 4 – once all options have been chosen, generate the report:



World Development Indicators | D

databank.worldbank.org/reports.aspx?ReportId=102603&Type=Table

Welcome Barry Smart

Sign Out

This page is in English Español Français عربي 中文

DataBank | World Development Indicators

Table | Chart | Map | Metadata | Download options

Variables | Layout | Styles | Save | Share | Embed

	Available	Selected
Database		1
Country		264
Series	1431	85
Time		59

Sort Remove all

- GDP (current US$)
- GDP growth (annual %)
- Population, total
- Population, male (% of total population)
- Population, female (% of total population)
- Life expectancy at birth, female (years)
- Life expectancy at birth, male (years)
- Life expectancy at birth, total (years)
- GDP per capita (current US$)
- GDP per capita growth (annual %)

Create Custom Indicator

Define Aggregation Rule

Full Report

Clear Selection | Add Country (264) Add Series (85) Add Time (59)

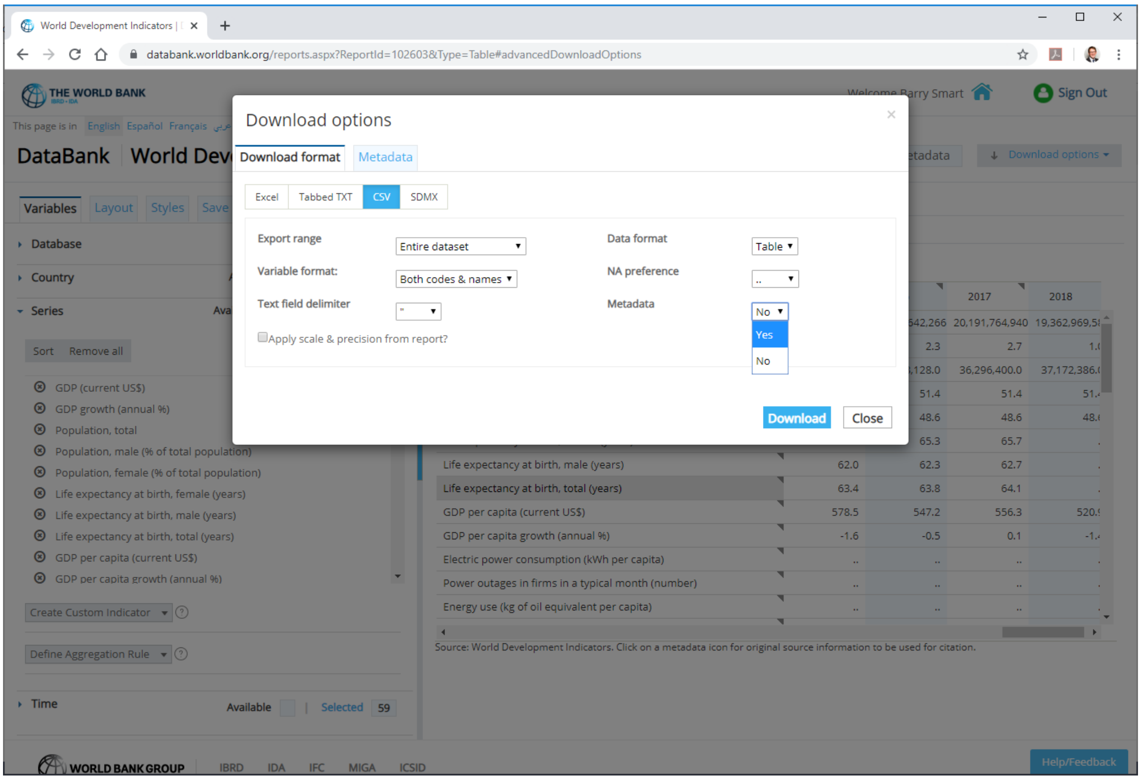
Afghanistan

	2015	2016	2017	2018
GDP (current US$)	19,907,111,419	19,362,642,266	20,191,764,940	19,362,969,5
GDP growth (annual %)	1.5	2.3	2.7	1.
Population, total	34,413,603.0	35,383,128.0	36,296,400.0	37,172,386.
Population, male (% of total population)	51.4	51.4	51.4	51.
Population, female (% of total population)	48.6	48.6	48.6	48.
Life expectancy at birth, female (years)	64.9	65.3	65.7	.
Life expectancy at birth, male (years)	62.0	62.3	62.7	.
Life expectancy at birth, total (years)	63.4	63.8	64.1	.
GDP per capita (current US$)	578.5	547.2	556.3	520.
GDP per capita growth (annual %)	-1.6	-0.5	0.1	-1.
Electric power consumption (kWh per capita)	..	..	..	.
Power outages in firms in a typical month (number)	..	..	..	.
Energy use (kg of oil equivalent per capita)	..	..	..	.

Source: World Development Indicators. Click on a metadata icon for original source information to be used for citation.

Figure 22 - screen shot showing panel on left used to choose data points to download.

Step 5 – use the advanced download options to download the report and associated metadata in CSV form:



Download options

Download format | Metadata

Excel | Tabbed TXT | CSV | SDMX

Export range: Entire dataset

Variable format: Both codes & names

Text field delimiter: "

□ Apply scale & precision from report?

Data format: Table

NA preference: ..

Metadata: No (Yes / No)

Download | Close

World Development Indicators
databank.worldbank.org/reports.aspx?ReportId=102603&Type=Table#advancedDownloadOptions
THE WORLD BANK
Welcome Barry Smart
Sign Out
This page is in English Español Français عربي
DataBank World Dev
Variables
Layout
Styles
Save
Database
Country
Series
Sort
Remove all
GDP (current US$)
GDP growth (annual %)
Population, total
Population, male (% of total population)
Population, female (% of total population)
Life expectancy at birth, female (years)
Life expectancy at birth, male (years)
Life expectancy at birth, total (years)
GDP per capita (current US$)
GDP per capita growth (annual %)
Create Custom Indicator
Define Aggregation Rule
Time
Available
Selected 59
Metadata
Download options
2017
2018
Life expectancy at birth, male (years)
Life expectancy at birth, total (years)
GDP per capita (current US$)
GDP per capita growth (annual %)
Electric power consumption (kWh per capita)
Power outages in firms in a typical month (number)
Energy use (kg of oil equivalent per capita)
Source: World Development Indicators. Click on a metadata icon for original source information to be used for citation.
WORLD BANK GROUP
IBRD
IDA
IFC
MIGA
ICSID
Help/Feedback

Figure 23 - screen shot of "advanced download options" used to generate CSV files.

A4 – Definition World Development Indicators (WDIs)

A summary of the data selected is set out below in Table 8:

Country	Data is available for 264 countries, all were selected. Meta data was also downloaded for each country as it provided useful labels including: Region and Income Group.
Year	Annual data is available from 1960 to 2018. All 59 years were selected.
Data Series	Each data series captures a specific World Development Indicator (WDI). 1,432 are available. A subset of 85 were chosen initially and then subsequently reduced based on analysis.

Table 8 - overview of the sub-set of World Bank "Economic Development Indicator" open data selected for this assignment.

Therefore, a theoretical total of 1,323,960[4] data points were available for analysis, but as discovered in subsequent phases:

- A number of countries were "dummy" countries created by the World Bank to enable aggregated data to be captured – for example by geographic region or economic group. These were removed during the data wrangling phase;
- There were some significant gaps in the data – where it made sense to do so, interpolation techniques were used to address these gaps;
- Application of data analysis enabled the number of WDIs required to be reduced significantly to the 20 variables presented below in Table 9 below.

Data Series	Definition	Topic
Account ownership at a financial institution or with a mobile-money-service provider, young adults (% of population ages 15-24)	Account denotes the percentage of respondents who report having an account (by themselves or together with someone else) at a bank or another type of financial institution or report personally using a mobile money service in the past 12 months (young adults, % of population ages 15-24).	Financial Sector: Access
Electric power consumption (kWh per capita)	Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants.	Environment: Energy production & use
Energy use (kg of oil equivalent per capita)	Energy use refers to use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports and fuels supplied to ships and aircraft engaged in international transport.	Environment: Energy production & use

[4] The product of the size of each dimension in the data set: 264 Countries, 85 Data Series, 59 Years.

Fossil fuel energy consumption (% of total)	Fossil fuel comprises coal, oil, petroleum, and natural gas products.	Environment: Energy production & use
GDP per capita (current US$)	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.	Economic Policy & Debt: National accounts: US$ at current prices: Aggregate indicators
Immunization, DPT (% of children ages 12-23 months)	Child immunization, DPT, measures the percentage of children ages 12-23 months who received DPT vaccinations before 12 months or at any time before the survey. A child is considered adequately immunized against diphtheria, pertussis (or whooping cough), and tetanus (DPT) after receiving three doses of vaccine.	Health: Disease prevention
Inflation, consumer prices (annual %)	Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used.	Financial Sector: Exchange rates & prices
Life expectancy at birth, total (years)	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	Health: Mortality
Market capitalization of listed domestic companies (current US$)	Market capitalization (also known as market value) is the share price times the number of shares outstanding (including their several classes) for listed domestic companies. Investment funds, unit trusts, and companies whose only business goal is to hold shares of other listed companies are excluded. Data are end of year values converted to U.S. dollars using corresponding year-end foreign exchange rates.	Financial Sector: Capital markets
Merchandise exports (current US$)	Merchandise exports show the f.o.b. value of goods provided to the rest of the world valued in current U.S. dollars.	Private Sector & Trade: Exports
Mobile cellular subscriptions (per 100 people)	Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts (i.e. that have been used during the last three months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging and telemetry services.	Infrastructure: Communication s
Mortality rate, infant (per 1,000 live births)	Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year.	Health: Mortality

Population density (people per sq. km of land area)	Population density is midyear population divided by land area in square kilometers. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship--except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin. Land area is a country's total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones. In most cases the definition of inland water bodies includes major rivers and lakes.	Environment: Density & urbanization
Population growth (annual %)	Annual population growth rate for year t is the exponential rate of growth of midyear population from year t-1 to t, expressed as a percentage . Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.	Health: Population: Dynamics
Population, total	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.	Health: Population: Structure
Power outages in firms in a typical month (number)	Power outages are the average number of power outages that establishments experience in a typical month.	Private Sector & Trade: Business environment
Renewable energy consumption (% of total final energy consumption)	Renewable energy consumption is the share of renewables energy in total final energy consumption.	Environment: Energy production & use
Suicide mortality rate (per 100,000 population)	Suicide mortality rate is the number of suicide deaths in a year per 100,000 population. Crude suicide rate (not age-adjusted).	Health: Mortality
Tax revenue (% of GDP)	Tax revenue refers to compulsory transfers to the central government for public purposes. Certain compulsory transfers such as fines, penalties, and most social security contributions are excluded. Refunds and corrections of erroneously collected tax revenue are treated as negative revenue.	Public Sector: Government finance: Revenue
Urban population growth (annual %)	Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.	Environment: Density & urbanization

Table 9 - definitions for all 20 World Development Indicators (WDIs) analysed in this assignment.

A5 – Results of Data Interpolation

The following table captures the impact that linear interpolation had on the different variables in the data set

Series Name	Data Points in Original Data	NaNs in Original Data	Data Points in Interpolated Data	NaNs in Interpolated Data	Number of NaNs Interpolated
Account at financial institution (% of population ages 15-24)	427	12,299	1,207	11,519	780
Electric power consumption (kWh per capita)	5,907	6,819	6,477	6,249	570
Energy use (kg of oil equivalent per capita)	6,082	6,644	7,320	5,406	1,238
Fossil fuel energy consumption (% of total)	5,860	6,866	7,274	5,452	1,414
GDP per capita (current US$)	9,675	3,051	9,879	2,847	204
Immunization, DPT (% of children ages 12-23 months)	6,871	5,855	6,871	5,855	-
Inflation, consumer prices (annual %)	7,671	5,055	7,842	4,884	171
Life expectancy at birth, total (years)	11,329	1,397	11,747	979	418
Market capitalization of listed domestic companies (current US$)	2,199	10,527	2,623	10,103	424
Merchandise exports (current US$)	10,380	2,346	10,416	2,310	36
Mobile cellular subscriptions (per 100 people)	9,418	3,308	12,098	628	2,680
Mortality rate, infant (per 1,000 live births)	10,097	2,629	10,097	2,629	-
Population density (people per sq. km of land area)	12,162	564	12,177	549	15
Population growth (annual %)	12,690	36	12,702	24	12
Population, total	12,695	31	12,705	21	10
Power outages in firms in a typical month (number)	259	12,467	1,447	11,279	1,188
Renewable energy consumption (% of total final energy consumption)	5,392	7,334	6,031	6,695	639
Suicide mortality rate (per 100,000 population)	915	11,811	3,477	9,249	2,562
Tax revenue (% of GDP)	3,843	8,883	4,530	8,196	687
Urban population growth (annual %)	12,573	153	12,584	142	11
Sum	146,445	108,075	159,504	95,016	13,059
Percentage Of Data Points That are NaNs		42%		37%	

Table 10 – summary of the impact of forward linear interpolation on the data set.

A6 – Comparison of Clustering Models

Clustering algorithm configurations were evaluated based on the following scoring mechanisms (Sci-kit Learn, 2019):

- **Silhouette Score –** scores around zero indicate overlapping clusters, when the score is higher when clusters are dense and well separated;
- **Completeness Score –** a clustering result satisfies completeness if all the data points that are members of a given class (ie our “ground truth” Income Group label) are elements of the same cluster.
- **Homogeneity Score** - a clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.

Agglomerative Clustering

Table 11 below shows the results for Agglomerative clustering based on a target number of clusters of 4 (ie the number of discrete labels in the “ground truth” Income Group label).

The parameters control the behaviour of the model as follows:

- **Affinity** - method used to compute the distance between clusters;
- **Linkage** - the algorithm will merge the pairs of clusters that minimize this criterion.

Affinity	Linkage	Silhouette Score	Completeness Score	Homogeneity Score
cosine	complete	0.117966	0.456784	0.338985
cosine	average	0.123303	0.452613	0.298469
euclidean	**ward**	**0.435368**	**0.50151**	**0.30942**
euclidean	complete	0.469286	0.478634	0.268287
euclidean	average	0.389241	0.251034	0.033384
l1	complete	0.432461	0.447195	0.294452
l1	average	0.446828	0.532109	0.268474
l2	complete	0.469286	0.478634	0.268287
l2	average	0.389241	0.251034	0.033384
manhattan	complete	0.432461	0.447195	0.294452
manhattan	average	0.446828	0.532109	0.268474

Table 11 - table showing relative performance of Agglomerative model with target number of clusters fixed at 4. All possible combinations of "Affinity" and "Linkage" parameters were explored.

K-Means Clustering

Table 12 and Figure 24 below show the performance of the K-means clustering model as the number of clusters was stepped up from 2 to 10. Specific focus was given to the results when the number of clusters was set to 4 as this was where the number of clusters matched the number of labels in the “ground truth” Income Group label.

Number Of Clusters	Silhouette Score	Completeness Score	Homogeneity Score
2	0.46729	0.558602	0.259738
3	0.439871	0.484364	0.295449
4	**0.296581**	**0.551667**	**0.454572**
5	0.309472	0.463645	0.442522
6	0.311227	0.444627	0.495534
7	0.281263	0.395941	0.493447
8	0.256917	0.349959	0.484001
9	0.253352	0.361373	0.524386
10	0.265022	0.343203	0.523603

Table 12 - table showing performance of K-means clustering as number of target clusters is stepped up from 2 to 10. The row of specific interest is that with 4 clusters as this matches the number of clusters in the “ground truth” label Income Group.

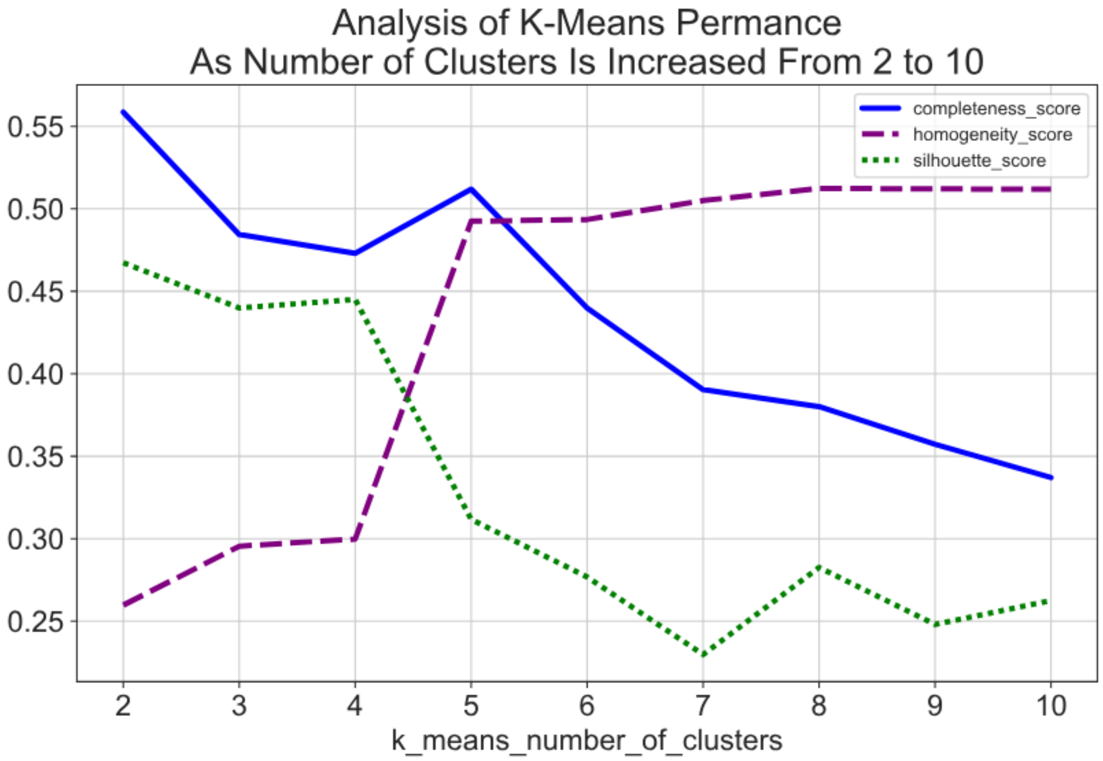
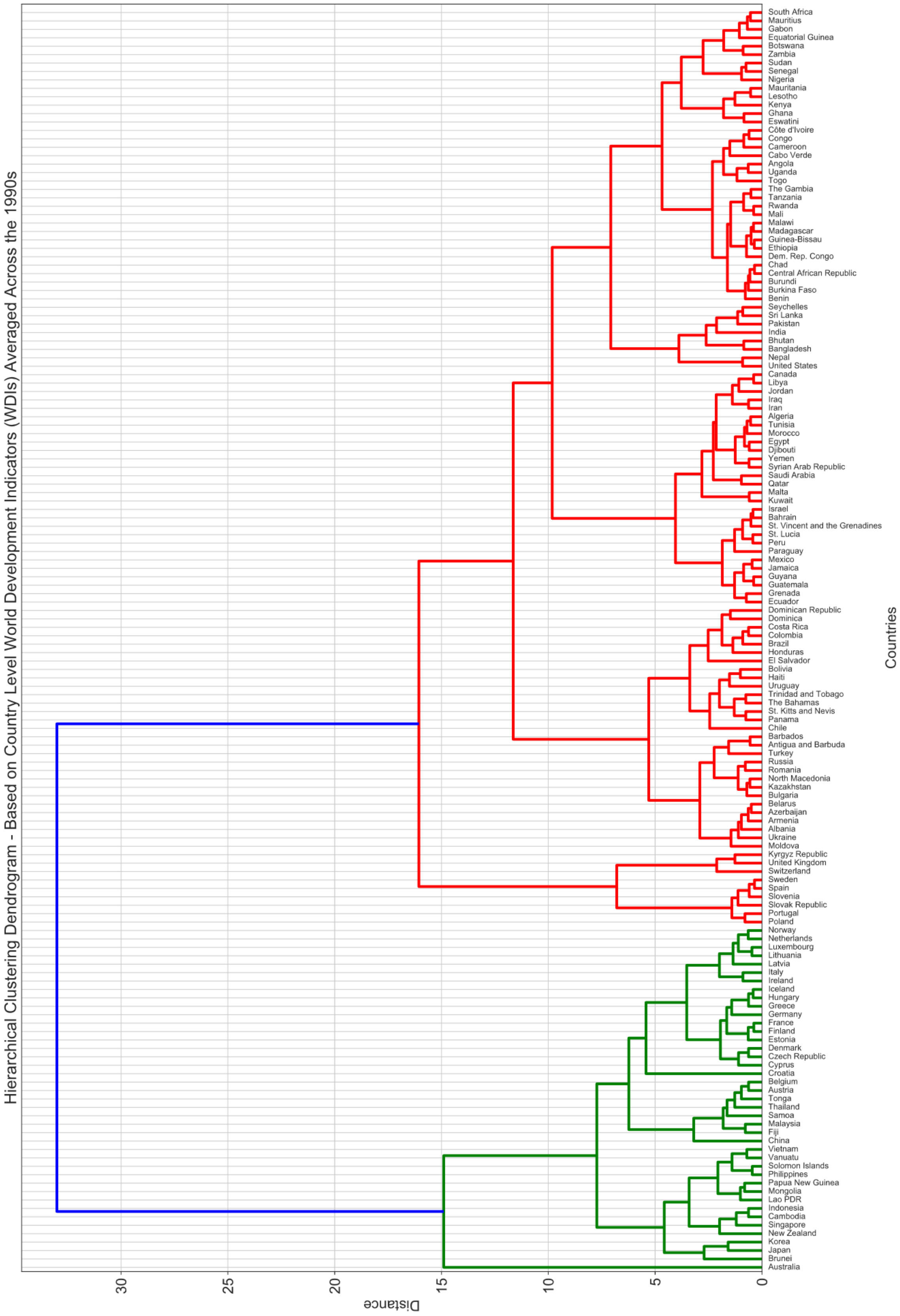


Figure 24 - visualisation of the performance of the K-means algorithm as clusters are stepped up from 2 to 10. There is a dramatic step change in both Silhouette and Homogeneity Score as between 4 and 5 clusters.

A7 – Dendrogram

The dendrogram below illustrates how the agglomerative clustering model has clustered the data:


Hierarchical Clustering Dendrogram - Based on Country Level World Development Indicators (WDIs) Averaged Across the 1990s
South Africa
Mauritius
Gabon
Equatorial Guinea
Botswana
Zambia
Sudan
Senegal
Nigeria
Mauritania
Lesotho
Kenya
Ghana
Eswatini
Côte d'Ivoire
Congo
Cameroon
Cabo Verde
Angola
Uganda
Togo
The Gambia
Tanzania
Rwanda
Mali
Malawi
Madagascar
Guinea-Bissau
Ethiopia
Dem. Rep. Congo
Chad
Central African Republic
Burundi
Burkina Faso
Benin
Seychelles
Sri Lanka
Pakistan
India
Bhutan
Bangladesh
Nepal
United States
Canada
Libya
Jordan
Iraq
Iran
Algeria
Tunisia
Morocco
Egypt
Djibouti
Yemen
Syrian Arab Republic
Saudi Arabia
Qatar
Malta
Kuwait
Israel
Bahrain
St. Vincent and the Grenadines
St. Lucia
Peru
Paraguay
Mexico
Jamaica
Guyana
Guatemala
Grenada
Ecuador
Dominican Republic
Dominica
Costa Rica
Colombia
Brazil
Honduras
El Salvador
Bolivia
Haiti
Uruguay
Trinidad and Tobago
The Bahamas
St. Kitts and Nevis
Panama
Chile
Barbados
Antigua and Barbuda
Turkey
Russia
Romania
North Macedonia
Kazakhstan
Bulgaria
Belarus
Azerbaijan
Armenia
Albania
Ukraine
Moldova
Kyrgyz Republic
United Kingdom
Switzerland
Sweden
Spain
Slovenia
Slovak Republic
Portugal
Poland
Norway
Netherlands
Luxembourg
Lithuania
Latvia
Italy
Ireland
Iceland
Hungary
Greece
Germany
France
Finland
Estonia
Denmark
Czech Republic
Cyprus
Croatia
Belgium
Austria
Tonga
Thailand
Samoa
Malaysia
Fiji
China
Vietnam
Vanuatu
Solomon Islands
Philippines
Papua New Guinea
Mongolia
Lao PDR
Indonesia
Cambodia
Singapore
New Zealand
Korea
Japan
Brunei
Australia
Countries
Distance