Exploring socio-economic properties of countries and life expectancy

Yousef Gharib

Abstract— This paper uses data sourced from the World Health Organization to classify countries as developed,developing and underdeveloped allowing us to determine which countries are most suitable for foreign aid, as well as predict the life expectancy of a country based on socio-economic features. First, clustering is used to identify a countrys' status as well as provide insight on the relationship between socio-economic features of a country and their geographic location. The last method used in this approach is linear regression, features in the dataset are used to in a multiple linear regression model to predict the life expectancy of a country. Our main findings show which countries are most suitable for financial aid by identifying underdeveloped countries in out cluster. The clusters also give us a spatial idea of the distribution of country status, and we see that western countries tend to be more developed. Testing our linear regression model, we find that life expectancy can be predicted signitificantly well, visualizations help us validate our clusters and linear regression models as well as spark ideas for new research questions.

1 Problem Statement

Countries can be classified as developed, developing, and in some cases, underdeveloped. This classification allows for underdeveloped countries to be identified and provided financial aid for economic and humanitarian development, especially during difficult situations. In 2017, the United States spent a total of $49.87 billion in aid, $14.77 billion as military assistance, and $35.10 billion as economic assistance [1]. Foreign aid is usually provided to the countries that need it most, deciding which countries need it can be a tough task since there are many external factors, however, socio-economic data of countries can provide insight on possible countries.

Life expectancy is a very popular measure used to assess the health of a country or population and informs us how long a certain population is expected to live for. Socio-economic status of a population such as income as well as factors like disease play major roles in life expectancy. This paper aims to answer the following research questions:

- How has life expectancy changed over time? What are factors affecting life expectancy?
- Is there a relationship between socio-economic properties of countries and their geographic location?
- Which countries are most suitable for financial aid?
- Can we accurately predict life expectancy of a country?

The data used contains socio-economic and health data as well as life expectancy of countries from the years 2000 to 2015 and will help conduct the analysis and build predictive models to answer the research questions.

2 State of the Art

Hennig and Liao (2013) talk about partitioning a population into social classes as well as clustering mixed-type data. They use data surveys of consumer finances carried out in 2007. The main approach taken was clustering, comparing latent class clustering and k-medoids. They suggested the best data to use for socio-economic stratification is a balance of variables measuring reward, achievement, and wealth. Clustering philosophy is mentioned along with advice on choosing clusters, stating there is a 'natural human intuition' of what true clusters are given a dataset.

The main visual tool used are scatter plots, to envision clustering theory but also to demonstrate how transforming variables for higher variance can prove to be beneficial. Bar charts and line graphs are used to compare clustering methods as well validate clusters, visualizing frequency distributions based on cluster and cluster method, and contrast silhouette width depending on the number of clusters.

Ketchen Jr. and Shook (1996) focused more on the application of clustering and aimed to investigate issues regarding how clustering is used and provide suggestions for more reliable clusters. They analysed 45 published papers and journals researching strategic management using clustering methods. They state fundamental questions to ask before clustering: "how to select variables; whether or not to standardize variables; and how to address multicollinearity among variables" and suggest approaches and solutions for each question. Exploratory analysis was conducted on the journals summarizing clustering decisions such as method and number of clusters. Results from the analysis show that validation of clusters is commonly neglected declaring the lack of reliability among clusters used.

The importance of validating clusters is emphasized heavily especially since it is often looked over. The main suggestions for more reliable clusters are to perform multiple cluster analyses, testing different clustering methods among other things to address the questions mentioned, and to try splitting data to analyse multiple independent clusters as well as carrying out significance tests to justify variable choices.

Life expectancy tells us a lot about a country's socio-economic status. Roser et al (2013) uses data collected from as early as 1543 regarding life expectancy as well as socio-economic features. With the aim to answer how life expectancy has changed, what improves life expectancy as well as explore mortality and life expectancy by age, they use a lot of diverse visualization methods to assist the analysis. Time series graphs are used to show the change in life expectancy, distinguishing countries, and continents with the biggest change. Animated choropleth maps are used as a spatio-temporal demonstration of how life expectancy progressively develops.

These papers provide a good balance of theory, analytical and visual techniques and relate to the domain of this analysis, allowing for methods and suggestions to be adopted and tailored to the datasets used.

3 Properties of the Data

The two datasets used in the analysis were collected from the World health organisation (WHO) as well as other sources and were found on Kaggle. The first dataset contains information on socio-economic properties of 167 countries while the second dataset provides more information on populations of countries as well as the life expectancy of countries from the year 2000 to 2015. These datasets complement each other as they can be linked by country and fit the research questions. The first dataset is used as a base for the analysis and the second dataset is used to provide extra insight since it contains life expectancy over a 15-year period.

The most notable feature used in this analysis is life expectancy, merging the two datasets allows for rich analysis to be conducted through time series and spatial visualizations as well as constructing predictive models using other variables.

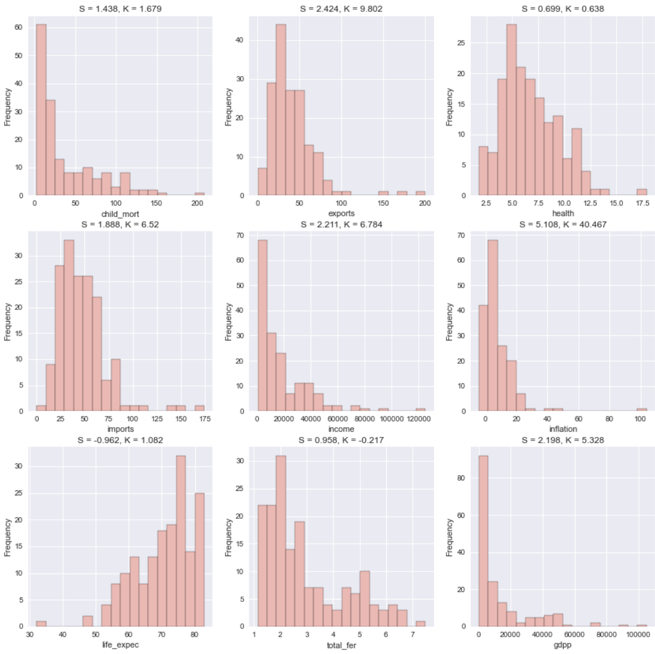


Figure 1: Histogram of features

Histograms and boxplots are used to investigate the data for outliers and visualize properties of the data. The features in the first dataset contain no null values, however, variables in the second dataset do, but since they are not used the null values were not dropped. Computational tools such as skewness and kurtosis along with visual tools are used to assess properties of data and provide insight on how issues with variables can be dealt with.

A significant number of variables are heavily skewed with relatively large values for skewness and kurtosis, suggesting transforming variables is suitable and will aid the analysis.

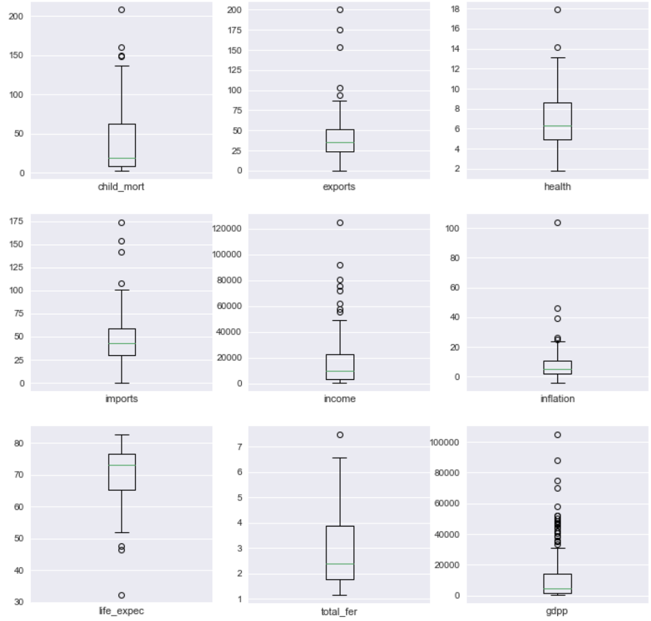


Figure 2: Boxplot of features

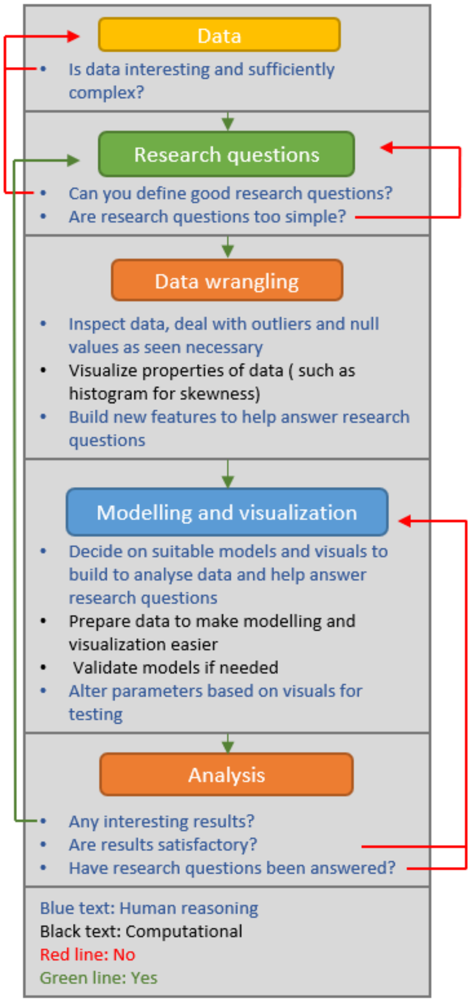
Figure 2 visualizes the distribution of outliers among variables, **gdpp** contains the largest number of outliers. Outliers can contain valuable information and are generally not dealt with just by normalizing data. During the exploratory analysis, outliers are kept as they are expected to provide useful insight, however, during the construction of predictive models, tests are conducted with and without outliers to compare and are chosen as seen necessary. Data transformation techniques such as Box-cox and log transformations are tested to increase the variance of features and standardise them.

While the data chosen allows for rich analysis and is considered satisfactory as it provides a balance of variables as mentioned by Hennig and Liao (2013), some drawbacks of the first dataset is that it does not mention the year this data was collected, meaning data could be outdated. In addition to that, it does not contain all the countries, as well as lack of information which may be considered crucial such as population of the country the year the data was collected. With these in mind, the analysis is still expected to provide worthwhile results and the analytical approaches used allow for methods to be recycled with new data to provide up to date results.

4 Analysis

4.1 Approach

This section discusses the workflow analysis displayed in figure 3, going into more detail on the approach taken as well as the human reasoning and computational tools practiced and how they connect. The workflow plan was designed to be vague but sufficiently complicated, describing the process to take for any dataset, this allows datasets to be replaced or added and adapted easier.



Data
Is data interesting and sufficiently complex?
Research questions
Can you define good research questions?
Are research questions too simple?
Data wrangling
Inspect data, deal with outliers and null values as seen necessary
Visualize properties of data (such as histogram for skewness)
Build new features to help answer research questions
Modelling and visualization
Decide on suitable models and visuals to build to analyse data and help answer research questions
Prepare data to make modelling and visualization easier
Validate models if needed
Alter parameters based on visuals for testing
Analysis
Any interesting results?
Are results satisfactory?
Have research questions been answered?
Blue text: Human reasoning
Black text: Computational
Red line: No
Green line: Yes

Figure 3: Workflow plan

The first step requiring human reasoning and intuition is selecting an interesting and sufficiently complex dataset that will help answer research questions, taking into consideration advice and suggestions mentioned in previous sections such as having a good mix of variables describing different features. The main steps in the process after selecting the dataset are data wrangling, modelling and visualization with a large emphasis on validation and testing, and finally the analysis. With an aim to involve human reasoning and intuition to explore and test, each section uses visualizations to support human cognition, resulting in further development of the analytical process.

Visualizing properties of data is crucial step during the early analysis. Inspecting and handling null values as well as outliers are necessary before proceeding with the analysis. Histograms are used to display skewness and whether the data resembles a distribution, and boxplots are used to inspect outlier's data for outliers. Human reason is required in deciding how to deal with null values and outliers, deriving new features based on existing ones, as well as interpreting visualizations and whether data transformation is required, applying these methods require computational assistance.

Based on the research questions, computational models are used to test methods and are visualized. The methods used to answer the research questions are clustering, time-series graphs as well as predictive linear regression models. While these methods are carried out computationally, significant human reasoning is required to interpret visualizations and test results as well as validate techniques and test different models. For clustering, the validation techniques used are elbow plots, silhouette plots and visual models to decipher cluster quality. For linear regression, mean squared error and R^2 value are used to validate the model, as well as a comparison of models, Q-Q plots are also used to visualize the distribution of our residuals.

Analysis of final results based on chosen methods is the final step in this approach. The main analysis is the interpretation of computational methods and visualizations built as well as the validation of models. Findings allow for potentially new, interesting research questions, taking us back to modelling and visualization.

4.2 Process

1. Data wrangling

The initial process before model building is analysing qualities of variables to see if any improvements can be made to support the analysis, such as deriving new variables and scaling features.

Since we are using 2 datasets, they can be merged on mutual columns, however, this is not necessary as we are using the second dataset only to build a time series. Using boxplots, we visualise outliers of features, 57 datapoints are considered outliers with 25 being part of the **gdpp** column which represents GDP per capita, this contributing to about 3% of the data with. The outliers suggest that there are certain countries which have significantly higher or lower values corresponding to that column, from figure 2 we see that 25 countries have a GDP per capita significantly higher than others, therefore outliers are not removed as they are expected to provide useful insight in the analysis.

The histogram in figure 1 informs us that majority of the variables are heavily skewed, this is also proven with computational values of skewness and kurtosis. The values of skewness and kurtosis are significantly over the general range of values, suggesting a transformation of variables is suitable. A log transformation as well as a box-cox transformation are applied to the features independently of each other and the transform better suited was used. Once the transformations were applied, new histograms are plotted to visualize skewness of data as well as new calculations of skewness and kurtosis. Both transformations methods provided much better results and increase the variance between variables, in the end, a log transformation was used since it is easier to understand and apply computationally.

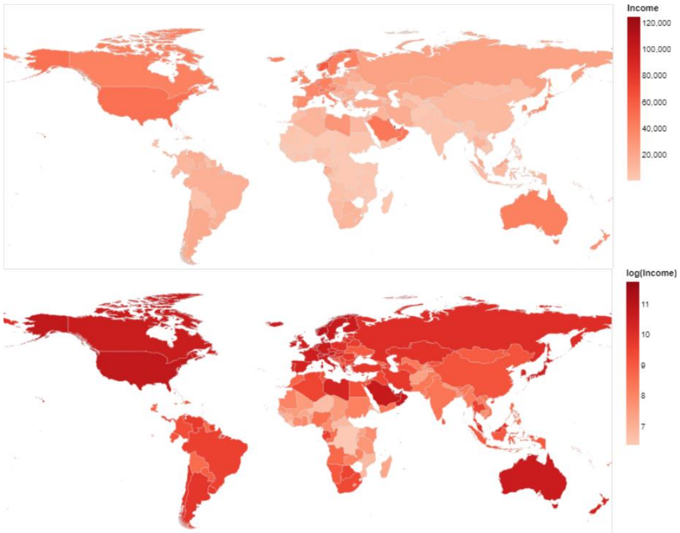


Figure 4: Comparison of income each country (top) and a comparison of income with a log transformation (bottom)

A visual benefit of transforming variables is shown by figure 4, the transformation allows for easier display of differences between countries but does not take care of null values.

New variables can be derived from existing ones to allow for easier analysis, a new column called **binned_life_expec** was created which is a binned version of the life expectancy column with bin sizes of 5 years since it results in better visualizations compared to 10, creating a categotical column for life expectancy as well as a continuous one, this also allows for easier grouping of countries based on binned life expectancy.

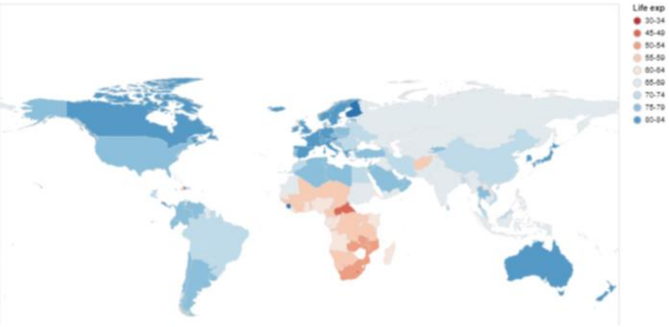


Figure 5: Binned life expectancy

The second dataset is then analysed before the time series line graphs are built. The dataset contains life expectancy for 193 unique countries from the years 2000 to 2015. Using the **groupby** function in python, we can group variables based on features and aggregate variables to allow for easier plots to be produced as well as for easier analysis of the data. After grouping the data by year and counting the number of unique countries, we found that each year, there is data for only 183 countries except for the year 2014 which has 193 countries. For a fair comparison in the time series, the intersection of countries is used. A function is created to validate that the same 183 countries contain a value for life expectancy each year, once this is done, the intersecting countries are used in the time series.

The analysis for skewness and outliers of life expectancy based on year was carried out the same way as it was for the first dataset. From the boxplots we see a significant number of outliers especially in the year 2005, however removing them does not make a significant difference in the time series and the trends can still be seen. The histograms are not heavily skewed and do not have significant values of skewness and kurtosis, therefore the data is not transformed.

2. Modelling and Visualization

The temporal analysis was carried out using intersecting countries from the years 2000 to 2015. Visualizations of how the world life expectancy has changed as well as how they have changed for developed and developing countries.

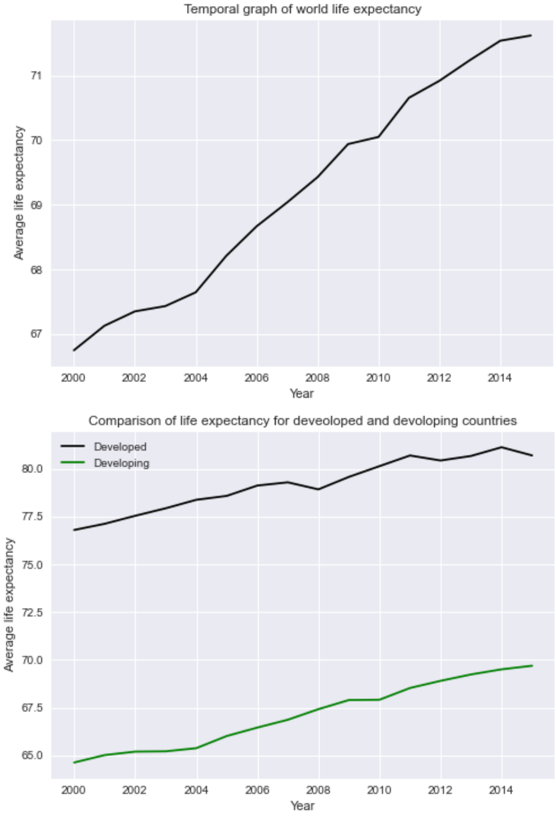


Figure 6: Time series of life expectancy

The results of the time series seem reasonable except that the life expectancy of developed countries decreased from 2014 to 2015, this could be for many factors ranging from disease to issues with the dataset, however, it is not significant enough to consider seriously.

To try and group countries based on socio-economic properties and classify them into groups, clustering is the next approach applied. After investigating the data and testing clustering methods, we found the best clustering algorithm to use was K-means clustering. While other methods such as hierarchical clustering works well, visuals produced testing both methods found that K-means helps answer the research questions better. Hierarchical clustering is mainly interpreted through a dendrogram whereas K-means clustering can be interpreted and validated using multiple methods.

The initial testing of clusters was done using all the variables in the dataset, an elbow plot is used to give us an idea of the best number of clusters to use. The elbow plot in figure 7 shows the optimal number of clusters is 2 or 3 since they are the points with the largest increase in gradient when using all variables as well as specific ones. Using a silhouette plot, we visualize the consistency of clusters and are able to validate them. With 2 clusters, an average silhouette score if around 0.4 is achieved, while with 3 clusters, and average score of approximately 0.3 is achieved. This suggests that 2 clusters provide more consistent data within clusters. We visualize the clusters spatially to allow for geographic interpretation.

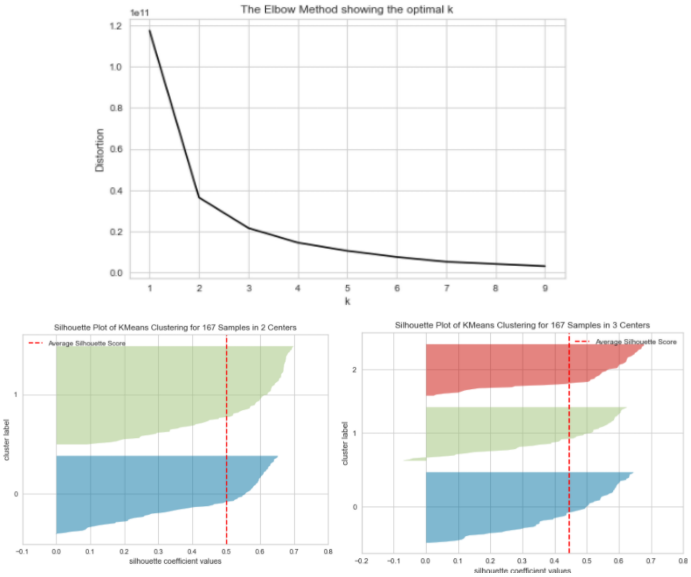


Figure 7: Elbow plot (top) and silhouette plots for 2 clusters (bottom left) and 3 clusters (bottom right) using income, GDP per capita, life expectancy and exports

Further cluster testing is carried out by trying out different combinations of variables to cluster on. Features are ranked based on interest in and relation to the research questions, the features used are income, GDP per capita, life expectancy and export of goods and services. The silhouette scores for 2 clusters have an average of 0.5 and around 0.45 for 3 clusters as shown in figure 7. The scores are significantly higher compared to when all the features are used, however, some of the scores are negative when using 3 clusters suggesting certain countries can be considered misclassified. Multiple independent clusters are created and built by splitting the dataset based on features, the next independent clusters are modelled based on the features not used in the first. The features used for the second independent cluster are child mortality rate, import of goods and services, inflation of currency and birth rate. Silhouette scores for the second independent cluster are not as high compared to the first, but both have proved to cluster countries better compared to clustering all features. Spatially visualizing these clusters shows they appear to group countries better. Figure 8 visualizes how countries are clustered based on socio-economic features in the dataset. We see that for both 2 and 3 clusters, majority of countries surrounding each other have been categorized into the same cluster, this makes for visually satisfying and satisfactory clusters except for the gaps in the map due to the missing countries in the dataset.

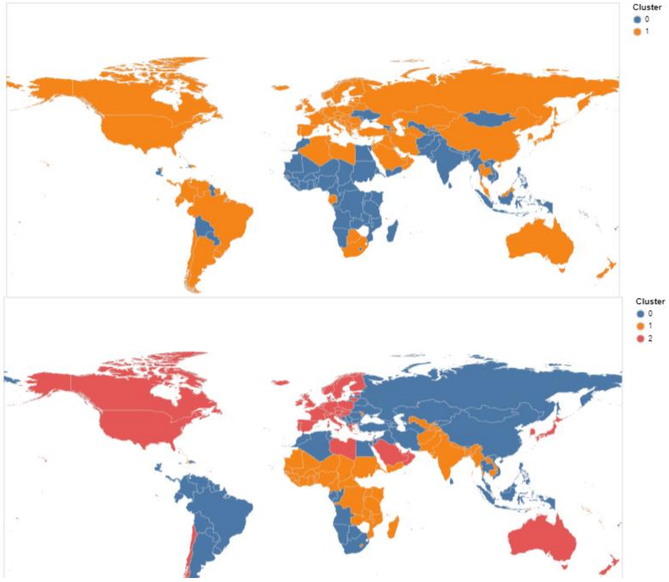


Figure 8: Spatial visualization of 2 and 3 clusters using income, GDP per capita, life expectancy and exports

The countries can be classified as developed, developing, or underdeveloped based on cluster as well as number of clusters used. Since we are interested in finding countries which may be suitable for financial aid, using 3 clusters may be more appropriate as we can identify underdeveloped countries. While some countries are missing, clustering still gives a good representation of socio-economic features and geographic location and their relationship.

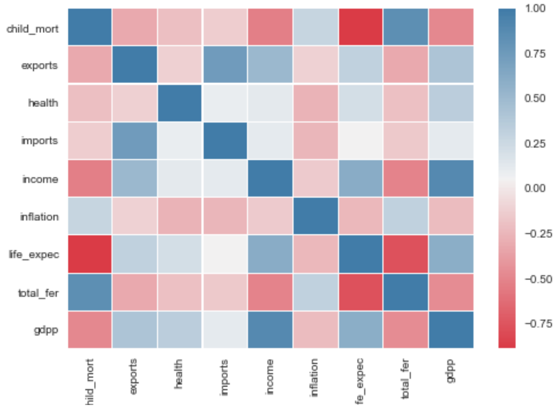


Figure 9: Correlation heatmap

Linear regression is the final approach used in this analysis as we are interested in predicting life expectancy. The initial step is deciding which variables to use as predictors.

We use figure 9 to understand how variables are correlated to help us decide on which variables to use to avoid multicollinearity. The predictors chosen are child mortality, income, GDP per capita and health, these variables have a relatively high correlation to life expectancy and relatively low correlation to each other. Before building out linear regression model, our data is divided into 75%/25% train test split to prevent overfitting, if we consider which cluster the countries were assigned to, we will want our training data to be a good balance of countries from each country to avoid bias, another model is built using cross validation instead of a train/test split and the results are compared.

Models are validated by calculating the mean square error, R^2 value and by visualizing the residuals using a Q-Q plot. The first linear regression model uses a train/test split, the mean sum square achieved by this model is 0.0104 with an R^2 value of 0.484, while the residual sum is low, the R^2 value suggests that only 48% of the variance of life expectancy is explained through the predictors. The second model uses cross validation which is expected to give significantly better and more reliable results. The mean squared error achieved by this model is approximately 0.00549 with a R^2 value of 0.72, suggesting it is a much better model compared to the first.

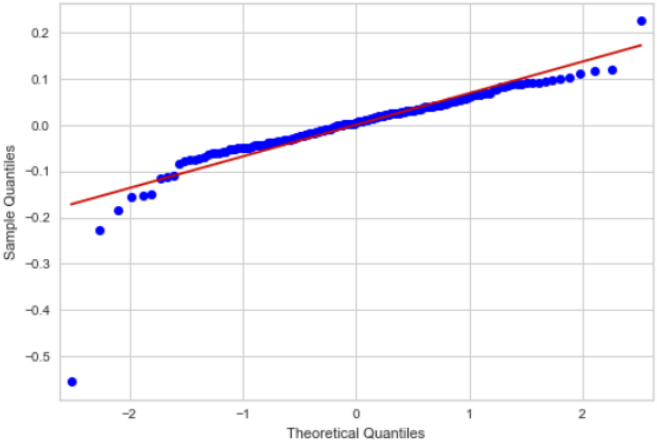


Figure 10: Q-Q plot of residuals

Using a Q-Q plot, we compare how the residuals are distributed in the second model. Figure 10 shows that the residuals resemble a normal distribution which is an indication that our model is valid.

4.3 Results

It is shown through the time series, that life expectancy steadily increases each year and that developed countries have an average life expectancy approximately 10 years more than developing countries. Life expectancy is determined by numerous features, the features in the dataset with the most effect are child mortality, income, GDP per capita and health.

Using 3 clusters provides reliable insight on how countries are distributed geographically as well as classify them based on whether they are developed, developing or underdeveloped. Analysing the clusters, we determine cluster 0 represents underdeveloped countries, 2 represents developing and 1 represents developed. Western countries appear to be more developed compared to eastern countries, Africa as well as Middle east Asia have the highest concentration of underdeveloped countries. Countries such Afghanistan, Yemen are most suitable for financial aid as they are classified as underdeveloped.

The linear regression model performed significantly better when using cross validation compared to using a train/test split. The cross-validation model produced a mean square error of 0.00549 and an R^2 value of 0.72 suggesting the model is reliable for predicting life expectancy based on the predictors used.

5 Critical reflection

Plotting a time series helped inform us how life expectancy has changed. Grouping the dataset by year and counting the number of unique countries we found that 2014 included 10 countries which are absent in the other years. Plotting the time series with these values showed no significant difference in the average life expectancy compared to plotting without, however, for a fair comparison of a time series, only the same countries should be included each year, and so those 10 countries were not included in the final time series. While the dataset contained variables with a clear relationship to life expectancy, further experimentation with different variables such as education should be tested with for more insight on life expectancy.

The clusters produced in this analysis are satisfactory and have supported the analysis and helped answer the research questions, while there is no solid evidence there is a relationship between socio-economic properties of countries and their geographic location, the clusters allow for certain trends to be seen and interpreted, suggesting they are good clusters. Splitting the data based on feature interest and producing multiple independent clusters proved to construct significantly more reliable clusters as shown by the silhouette and spatial plots, which was surprising, and is most likely due to features chosen. Testing out more combinations of features may prove to be useful and provide even more reliable clusters. Interpreting the clusters is done through the analysis of their centroids for each feature, from this we determined cluster 0, 2 and 1 represent underdeveloped, developing and developed countries respectively. The cluster centroids are also visualized by plotting features on a graph with points classified by cluster and centroids, we see that there is overlap of points in clusters suggesting some countries may also be considered part of a different cluster, however, the clusters provide a reliable overall representation how the countries are classified. Further work would involve narrowing the selection of countries suitable for foreign aid.

The linear regression model built proves to be reliable since the mean square error is relatively low and the R^2 value is high. Further testing for the predictive model would include testing other combinations of variables, as well as testing out boosting methods such as gradient boosting or adaptive boosting to see if a better model can be constructed.

The dataset used for this analysis is good but does have some drawbacks. The biggest drawback is that it does not provide data on all the countries, giving us white spaces in the spatial visualizations showing that country is not classified, also, it only provides socio-economic factors, not factors such as disease or war, which play major roles in the effect of life expectancy of a country.

Table of word counts

Problem statement	236/250
State of the art	456/500
Properties of the data	451/500
Analysis: Approach	412/500
Analysis: Process	1498/1500
Analysis: Results	186/200
Critical reflection	451/500

REFERENCES

The list below provides examples of formatting references.

[1] Wikipedia Contributors (2021). *United States foreign aid*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/United_States_foreign_aid#:~:text=In%20fiscal%20year%202017%20.

[2] Hennig, C. and Liao, T.F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics), [online] 62(3), pp.309–369

[3] KETCHEN Jr., D.J. and SHOOK, C.L. (1996). THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE. Strategic Management Journal, [online] 17(6), pp.441–458.

[4] Roser, M., Ortiz-Ospina, E. and Ritchie, H. (2013). Life Expectancy. [online] Our World in Data.

[5] Nist.gov. (2021). 1.3.5.11. Measures of Skewness and Kurtosis.

[6] Torri, T. and Vaupel, J.W. (2012). Forecasting life expectancy in an international context. International Journal of Forecasting, [online] 28(2), pp.519–531.

[7] Nguyen, D., Smith, N. and Rosé, C. (2011). Author Age Prediction from Text using Linear Regression. [online] Association for Computational Linguistics, pp.115–123.