ILLINOIS

Predicting Life Expectancy
STAT448 Final Project Report

Jiangyan Feng
NetID: jf8 (Group 4)
Chemical & Biomolecular Engineering
University of Illinois at Urbana-Champaign
April 29, 2020

1 Introduction

Problem statement: Life expectancy is the average age of death in a population, which indicates population health. There are many factors affecting life expectancy, such as economic, health, and social factors. The question is do countries with different economic, health, and social factors tend to have different life expectancies? To answer this question, this work employs clustering analysis and principal component analysis in order to explore how does life expectancy differ among different clusters.

Description of data: The dataset used in this work was collected from the World Health Organization (WHO) data repository website and the corresponding economic data was collected from the United Nation website. This dataset is composed of life expectancy and 21 related factors for 193 countries from 2000 to 2015. In total, there are 22 variables and 2938 instances. Among these 22 variables, 3 variables are categorical variables (country, status, year) and the others are all numerical variables.

Data source: `https://www.kaggle.com/kumarajarshi/life-expectancy-who`

Description of variables:

1. country (Nominal) - the country name (193 unique values)
2. year (Ordinal) - the calendar year (ranging from 2000 to 2015)
3. status (Nominal) - “Developing" or “Developed" country
4. life expectancy (Ratio) - the average age of death for a particular country and year
5. adult mortality (Ratio) - probability of dying between 15 and 60 years per 1000 population
6. infant deaths (Ratio) - number of infant deaths per 1000 population
7. alcohol (Ratio) - alcohol consumption rate measured as liters of pure alcohol consumption per capita
8. percentage expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (GDP) per capita (%)
9. hepatitis b (Ratio) - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
10. measles (Ratio) - number of reported Measles cases per 1000 population
11. bmi (Interval/Ordinal) - average Body Mass Index (BMI) of a country’s total population
12. under-five deaths (Ratio) - number of under the age of five deaths per 1000 population
13. polio (Ratio) - Polio (Pol3) immunization coverage among 1-year-olds (%)
14. total expenditure (Ratio) - government expenditure on health as a percentage of total government expenditure (%)
15. diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
16. hiv/aids (Ratio) - deaths per 1000 live births caused by HIV/AIDS (0-4 years)
17. gdp (Ratio) - Gross Domestic Product per capita (in USD)
18. population (Ratio) - population of a country
19. thinness 1-19 years (Ratio) - prevalence of thinness among children and adolescents for Age 10 to 19 (%)
20. thinness 5-9 years (Ratio) - prevalence of thinness among children for Age 5 to 9(%)
21. income composition of resources (Ratio) - Human Development Index in terms of income composition of resources (ranging from 0 to 1)
22. schooling (Ratio) - average number of years of schooling of a population

2 Methods and Results

2.1 Initial exploratory analysis

2.1.1 Description of the data

To get a quick description of the data, we use info() method from Pandas to extract the total number of row, each attribute's type and number of non-null values. As shown in Figure 1, this dataset contains 2938 instances and 22 variables as previously described. The data type for Country and Status is object and the data type for other variables are either integer and float. According to the non-null numbers, there are missing data in many variables, except Country, Year, Status, infant deaths, percentage expenditure, Measles, under-five deaths, HIV/AIDS. This indicates that dealing with missing data is required for the further analysis.

```
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
Country                          2938 non-null object
Year                             2938 non-null int64
Status                           2938 non-null object
Life expectancy                  2928 non-null float64
Adult Mortality                  2928 non-null float64
infant deaths                    2938 non-null int64
Alcohol                          2744 non-null float64
percentage expenditure           2938 non-null float64
Hepatitis B                      2385 non-null float64
Measles                          2938 non-null int64
 BMI                             2904 non-null float64
under-five deaths                2938 non-null int64
Polio                            2919 non-null float64
Total expenditure                2712 non-null float64
Diphtheria                       2919 non-null float64
 HIV/AIDS                        2938 non-null float64
GDP                              2490 non-null float64
Population                       2286 non-null float64
 thinness  1-19 years            2904 non-null float64
 thinness 5-9 years              2904 non-null float64
Income composition of resources  2771 non-null float64
Schooling                        2775 non-null float64
dtypes: float64(16), int64(4), object(2)
```

Figure 1: Data info.

2.1.2 Descriptive statistics of the data

We use describe() method from Pandas to obtain the basic statistics of all the numerical attributes (Figure 2). This allows us to understand the mean, standard deviation, min, max, 25%, 50%, and 75% of the attributes. There are some attributes do not make sense. For example, the mean of percentage expenditure is 738 which is much larger than 100. In addition, Measles is the number of reported Measles cases per 1000 population. However, the mean of Measles is 2419 which is

far larger than 1000. This could be a result of missing values and outliers. This indicates that we need to understand the percentage of outliers and investigate these two variables in the further analysis. From the basic statistics, we can clearly see the value range of different attributes varies a lot. For example, Polio ranges from 3 to 99 whereas under-five deaths ranges from 0 to 2500. This indicates that feature scaling is required in the further analysis.

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.000000	2919.000000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.035739	82.550188
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.445548	23.428046
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.000000	78.000000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.000000	93.000000
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.000000	97.000000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000

Figure 2: Summary of each numerical attribute.

To understand the distribution of each attribute, we plot the histogram of all the numerical attributes (Figure 3). From the distribution plots, life expectancy, total expenditure histogram, income composition of resources, and schooling seem to be normally distributed.

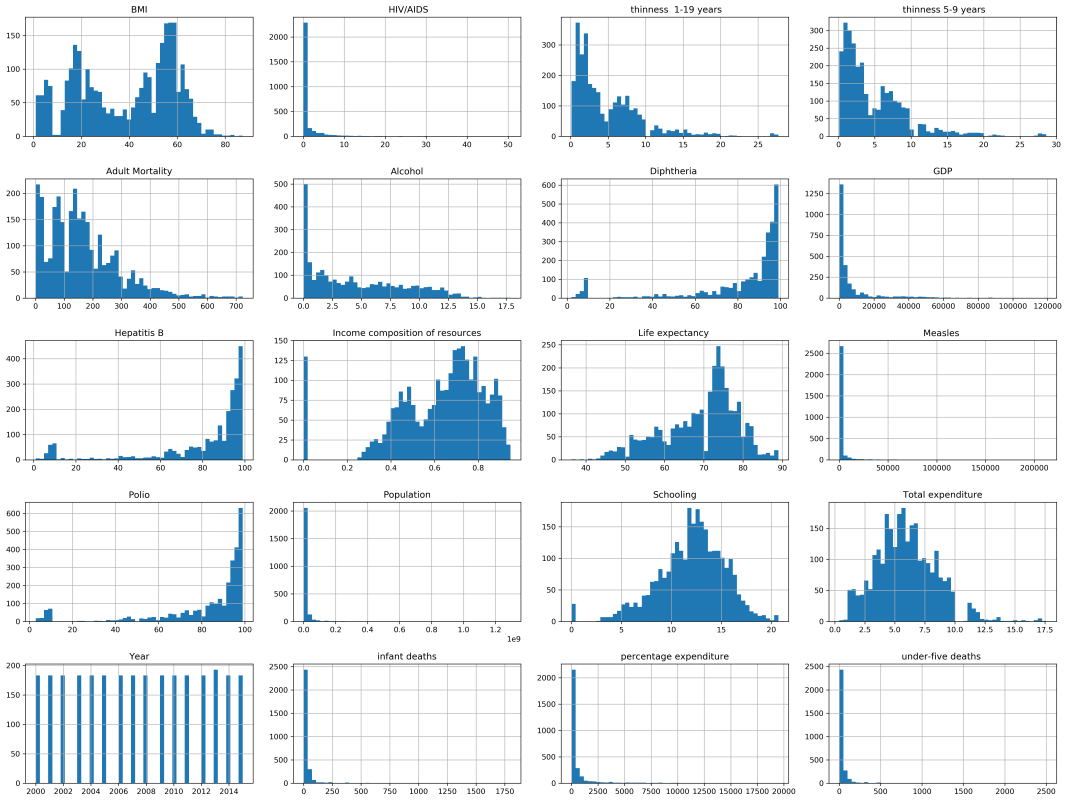

Figure 3: Histogram of each numerical attribute.

2.2 Diagnostic checking and data cleaning

2.2.1 Filling in missing data

From previous analysis, we notice that there are lots of missing values. To understand how much data is missing, we computed the percentage of missing values in each attribute (Figure 4). Overall, population has the maximum percentage of missing data (around 22%). On average, around 0.13% data is missing, which is relatively small.

Country	0.000000
Year	0.000000
Status	0.000000
Life expectancy	0.340368
Adult Mortality	0.340368
infant deaths	0.000000
Alcohol	6.603131
percentage expenditure	0.000000
Hepatitis B	18.822328
Measles	0.000000
BMI	1.157250
under-five deaths	0.000000
Polio	0.646698
Total expenditure	7.692308
Diphtheria	0.646698
HIV/AIDS	0.000000
GDP	15.248468
Population	22.191967
thinness 1-19 years	1.157250
thinness 5-9 years	1.157250
Income composition of resources	5.684139
Schooling	5.547992

Figure 4: Missing data in each attribute.

In order to deal with the missing data, there are few choices. The first one is either removing instances with missing value or removing attributes with missing values. However, this will lead to a large loss of information. The second choice is filling in the missing data. Since only numerical attributes have missing data. In this work, we decide to fill in the missing data by the median value of each year. This is better than using the median value among all the years since the values of attributes may change dramatically with years. After data imputing, there is no missing data and all attributes have 2938 instances (Figure 5).

```
Data columns (total 22 columns):
Country                          2938 non-null object
Year                             2938 non-null int64
Status                           2938 non-null object
Life expectancy                  2938 non-null float64
Adult Mortality                  2938 non-null float64
infant deaths                    2938 non-null int64
Alcohol                          2938 non-null float64
percentage expenditure           2938 non-null float64
Hepatitis B                      2938 non-null float64
Measles                          2938 non-null int64
 BMI                             2938 non-null float64
under-five deaths                2938 non-null int64
Polio                            2938 non-null float64
Total expenditure                2938 non-null float64
Diphtheria                       2938 non-null float64
 HIV/AIDS                        2938 non-null float64
GDP                              2938 non-null float64
Population                       2938 non-null float64
 thinness  1-19 years            2938 non-null float64
 thinness 5-9 years              2938 non-null float64
Income composition of resources  2938 non-null float64
Schooling                        2938 non-null float64
```

Figure 5: No missing data after data imputing.

2.2.2 Winsorizing outliers

To detect the outliers, we first visualize the box plots of each attribute (Figure 6). According to the box plots, there are lots of outliers in most of the attributes including the target attribute, life expectancy. To quantitatively measure the outliers of each attribute, we computed the percentage of outliers for each attribute (Figure 7). According to Figure 7, Measles and HIV/AIDS have the highest outlier percentage, which is around 18%. Overall, the percent of outliers is relatively low.

To minimize the influence of outliers, there are several approaches. Approach 1 is deleting all the outliers which may cause the loss of information. Approach 2 is using Winsorization technique, which modifying the outlier value so that it is close to other values. In this work, we employed Winsorization technique in order to keep the information intact. We used the winsorize function from scipy.stats.mstats package. Considering the outlier percentage is different for each attribute, we fine tune the winsorization parameters for each attribute (Figure 8).

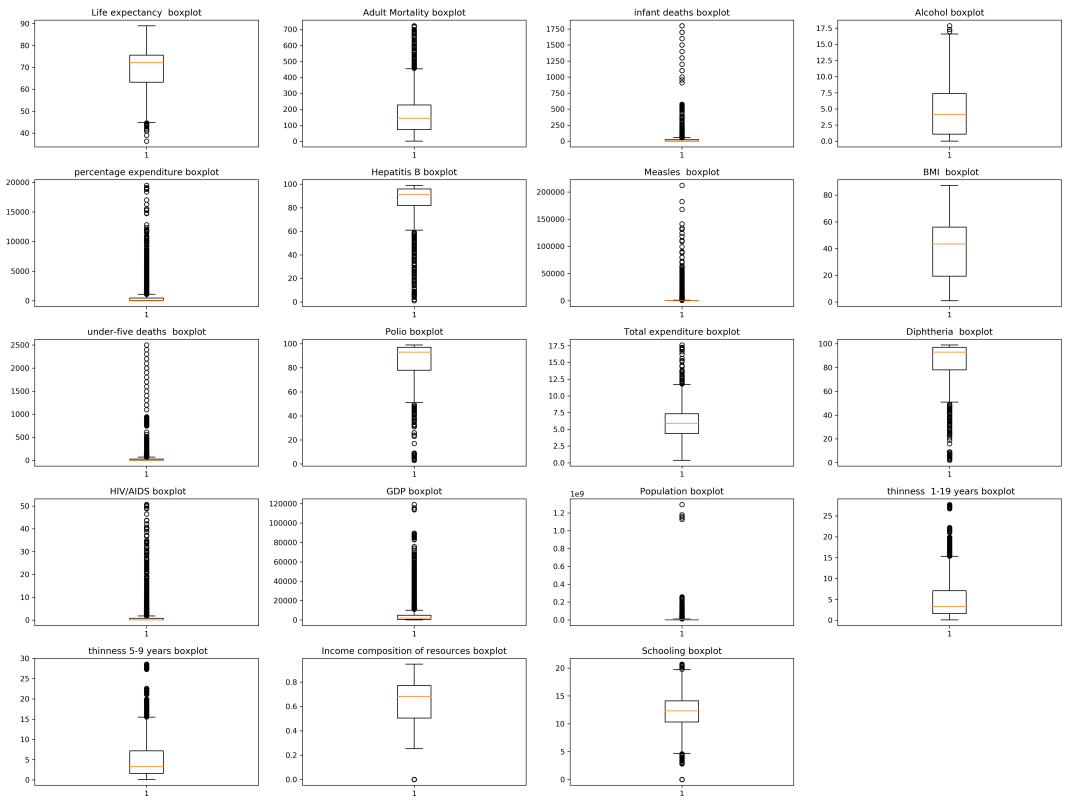

Figure 6: Box plot of each attribute.

```
--------------Life expectancy --------------
Number of outliers: 17
Percent of outliers: 0.58%
--------------Adult Mortality--------------
Number of outliers: 86
Percent of outliers: 2.93%
--------------infant deaths--------------
Number of outliers: 315
Percent of outliers: 10.72%
--------------Alcohol--------------
Number of outliers: 3
Percent of outliers: 0.1%
--------------percentage expenditure--------------
Number of outliers: 389
Percent of outliers: 13.24%
--------------Hepatitis B--------------
Number of outliers: 322
Percent of outliers: 10.96%
--------------Measles --------------
Number of outliers: 542
Percent of outliers: 18.45%
-------------- BMI --------------
Number of outliers: 0
Percent of outliers: 0.0%
--------------under-five deaths --------------
Number of outliers: 394
Percent of outliers: 13.41%
--------------Polio--------------
Number of outliers: 279
Percent of outliers: 9.5%
--------------Total expenditure--------------
Number of outliers: 51
Percent of outliers: 1.74%
--------------Diphtheria --------------
Number of outliers: 298
Percent of outliers: 10.14%
-------------- HIV/AIDS--------------
Number of outliers: 542
Percent of outliers: 18.45%
--------------GDP--------------
Number of outliers: 445
Percent of outliers: 15.15%
--------------Population--------------
Number of outliers: 452
Percent of outliers: 15.38%
-------------- thinness  1-19 years--------------
Number of outliers: 100
Percent of outliers: 3.4%
-------------- thinness 5-9 years--------------
Number of outliers: 99
Percent of outliers: 3.37%
--------------Income composition of resources--------------
Number of outliers: 130
Percent of outliers: 4.42%
--------------Schooling--------------
Number of outliers: 77
Percent of outliers: 2.62%
```

Figure 7: Percentage of outliers in each attribute.

```
wins_dict = {}
wins_outliers(num_vars[0], lower_limit=0.006)
wins_outliers(num_vars[1], upper_limit=0.03)
wins_outliers(num_vars[2], upper_limit=0.11)
wins_outliers(num_vars[3], upper_limit=0.01)
wins_outliers(num_vars[4], upper_limit=0.14)
wins_outliers(num_vars[5], lower_limit=0.11)
wins_outliers(num_vars[6], upper_limit=0.19)
wins_outliers(num_vars[7])
wins_outliers(num_vars[8], upper_limit=0.14)
wins_outliers(num_vars[9], lower_limit=0.10)
wins_outliers(num_vars[10], upper_limit=0.018)
wins_outliers(num_vars[11], lower_limit=0.11)
wins_outliers(num_vars[12], upper_limit=0.185)
wins_outliers(num_vars[13], upper_limit=0.152)
wins_outliers(num_vars[14], upper_limit=0.154)
wins_outliers(num_vars[15], upper_limit=0.035)
wins_outliers(num_vars[16], upper_limit=0.034)
wins_outliers(num_vars[17], lower_limit=0.05)
wins_outliers(num_vars[18], lower_limit=0.025, upper_limit=0.01)
```

Figure 8: Winsorization parameters for each attribute.

2.2.3 Drop Measles and percentage expenditure

Previous analysis shows that the value range of Measles and percentage expenditure do not make sense. The mean of percentage expenditure is 738, which is much larger than 100. In addition, Measles is the number of reported Measles cases per 1000 population. However, the mean of Measles is 2419 which is far larger than 1000. Previously, we suggest this could be a result of missing values and outliers. However, after filling in the missing values and dealing with the outliers, the value range of these two variables still do not make sense. Therefore, we decide to drop these two variables for further analysis.

2.2.4 Visualize the final data distribution and outliers

After all the procedures of data cleaning, we plot the box plots and histograms of all the numeric variables to ensure all the concerning problems are resolved (Figure 9). As shown in Figure 9, there are no outliers and the Winsorized life expectancy, Winsorized total expenditure histogram, Winsorized income composition of resources, and Winsorized schooling seem to be normally distributed.

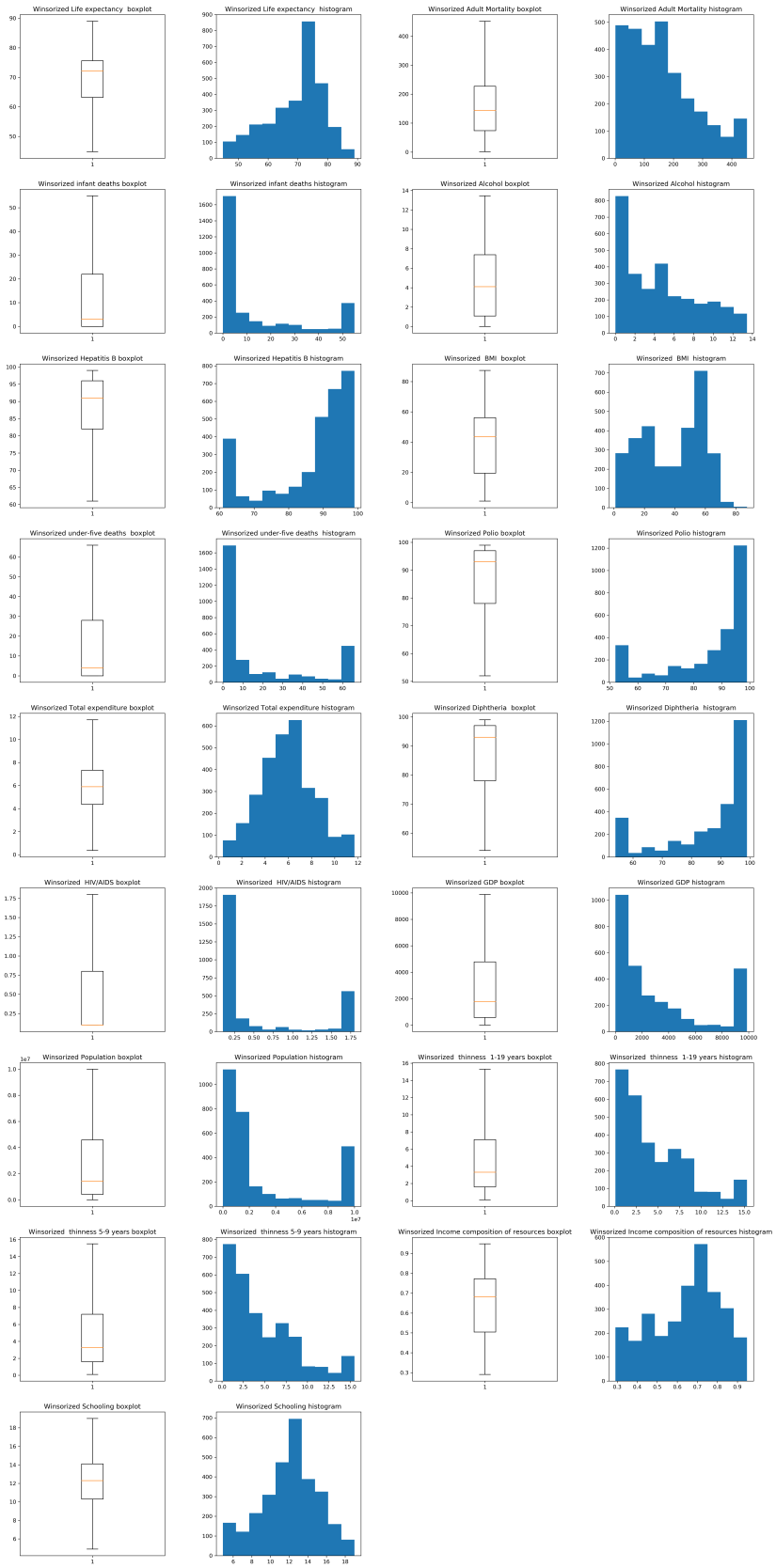

Figure 9: Box plots and histograms of Winsorization attributes.

2.3 Clustering analysis

2.3.1 Split the data into numeric and categorical parts

We first split the data into two parts: numeric (Figure 10) and categorical (Figure 11). We drop the life expectancy variable in the numeric parts and use the remaining 16 numeric variables to perform clustering analysis.

```
Data columns (total 17 columns):
Life expectancy                  2938 non-null float64
Adult Mortality                  2938 non-null float64
infant deaths                    2938 non-null int64
Alcohol                          2938 non-null float64
Hepatitis B                      2938 non-null float64
 BMI                             2938 non-null float64
under-five deaths                2938 non-null int64
Polio                            2938 non-null float64
Total expenditure                2938 non-null float64
Diphtheria                       2938 non-null float64
 HIV/AIDS                        2938 non-null float64
GDP                              2938 non-null float64
Population                       2938 non-null float64
 thinness  1-19 years            2938 non-null float64
 thinness 5-9 years              2938 non-null float64
Income composition of resources  2938 non-null float64
Schooling                        2938 non-null float64
```

Figure 10: Numeric variables information.

	Country	Status
count	2938	2938
unique	193	2
top	Sweden	Developing
freq	16	2426

Figure 11: Categorical variables information.

2.3.2 Feature scaling

As noticed previously, the value range of different attributes varies a lot. For example, Polio ranges from 3 to 99 whereas under-five deaths ranges from 0 to 2500. This indicates that feature scaling is required in the further analysis. To ensure all the features have similar scales, we employed StandardScaler() method from sklearn.preprocessing package. This method will compute the Z-score: $z = \frac{(x-\mu)}{s}$ where μ is the mean and s is the standard deviation.

2.3.3 Using Elbow method to find the optimal number of clusters

In this work, we focus on KMeans clustering method, which requires us to define the number of clusters: k. In order to find out the optimal number of clusters, we plotted the inertia versus the increasing number of numbers (ranging from 1 to 50) (Figure 12). Inertia measures the sum of squared distances of samples to their closest cluster center. Therefore, the lower the inertia, the better the model. From Figure 12, the upper elbow is around 5. Therefore, we choose number of clusters as 5 for further analysis.

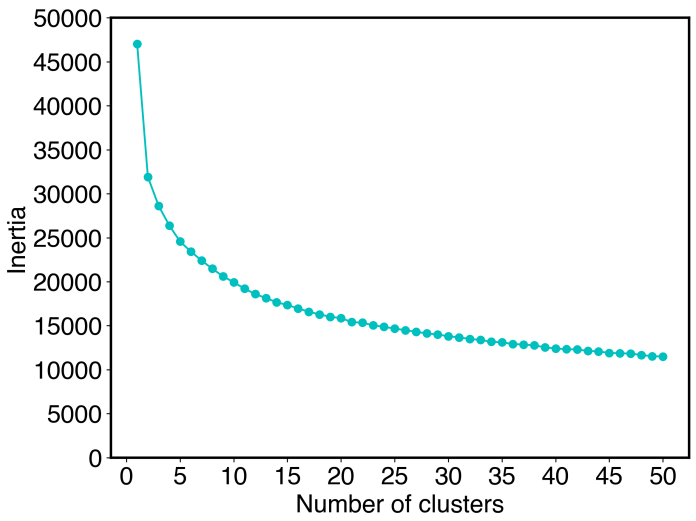

Figure 12: Inertia versus the increasing number of numbers.

2.3.4 Build the kMeans model with optimal number of clusters

We use number of clusters = 5 to build kMeans model to fit the scaled numerical data (Figure 13).

```
k = 5
kmeans = KMeans(n_clusters = k)
y_pred = kmeans.fit_predict(data_num_drop_scaled)
```

Figure 13: Parameters for the final kMeans model.

2.3.5 Compare the life expectancy among different clusters

To understand the difference of life expectancy among different clusters, we computed the basic statistics of life expectancy for each cluster (Figure 14) and plotted the box plots for each cluster (Figure 15). The results indicate that cluster 1 and 5 tend to have the highest and lowest life expectancy respectively.

	Life expectancy							
	count	mean	std	min	25%	50%	75%	max
Label								
0	735.0	78.20	4.31	64.6	75.0	78.5	81.2	89.0
1	401.0	61.73	7.72	44.8	56.6	61.7	66.6	79.0
2	349.0	68.45	6.46	44.8	64.5	69.7	73.2	79.9
3	892.0	72.96	3.95	62.4	71.0	73.3	75.2	88.0
4	561.0	57.50	6.87	44.8	52.2	57.2	63.0	77.0

Figure 14: Basic statistics of life expectancy for each cluster.

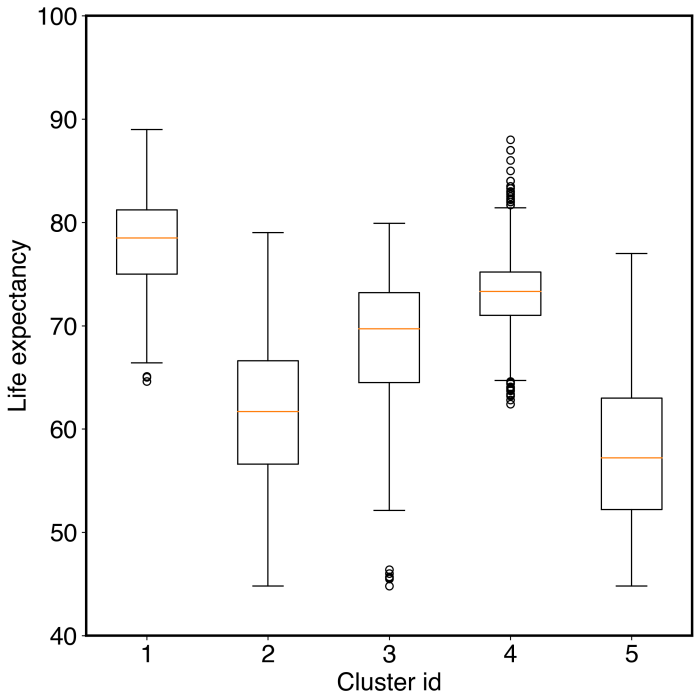

Figure 15: Box plot of life expectancy for five clusters.

Next, to understand whether the difference is statistically significant, we used ANOVA test to compare each pair of clusters (Figure 16) and all the clusters (Figure 17). The pvalue is smaller than 0.05, suggesting that there is significant difference of life expectancy among different clusters.

```
Statistics=2130.122, p=0.000
Statistics=863.789, p=0.000
Statistics=651.335, p=0.000
Statistics=4399.499, p=0.000
Statistics=164.530, p=0.000
Statistics=1196.075, p=0.000
Statistics=79.936, p=0.000
Statistics=223.048, p=0.000
Statistics=572.076, p=0.000
Statistics=2964.061, p=0.000
```

Figure 16: ANOVA test of each pair of clusters.

```
Statistics=1362.620, p=0.000
```

Figure 17: ANOVA test of all the clusters.

2.3.6 General characteristics of different clusters

To understand the general characteristics of these five clusters, we computed the mean value of each attribute for different clusters (Figure 18). In previous section, we notice cluster 1 and 5 tend to have the highest and lowest life expectancy respectively. Here, we focus on the characteristics of cluster 1 and 5. According to Figure 18, cluster 1 tends to have the lowest adult mortality, infant deaths, under-five deaths, thinness 1-19 years and thinness 5-9 years. In addition, cluster 1 tends to have the highest total expenditure, GDP, income composition of resources and schooling. Cluster 5 is the opposite of cluster 1. These results suggest that these attributes distinguish cluster 1 from 5 and contribute the most to the life expectancy.

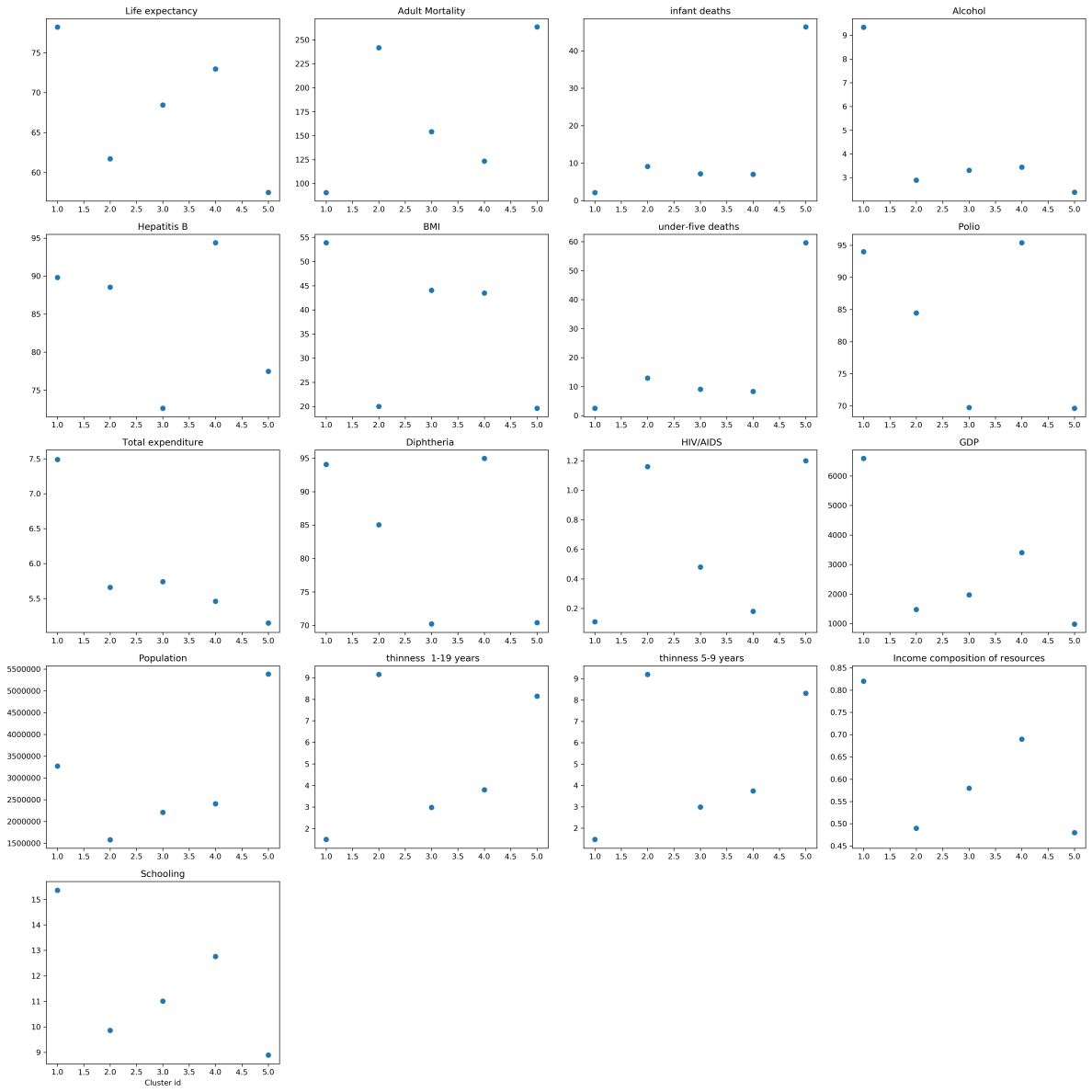

Figure 18: Mean value of each attribute for different clusters.

2.3.7 Principal component analysis

In this work, a total of 16 numeric attributes are used. In order to check whether these features can be decomposed, we employed principal component analysis. In order to find the optimal number of components, we computed the eigenvalues and explained variance ratio (Figure 19). As shown below, elbow is at components = 3, top 7 components are required to achieve explained variance > 80%, and top 4 components have eigenvalues > 1. Therefore, 7 components should be maintained. This suggests that 7 linear combinations of these 16 numeric attributes can be used to explain the dataset.

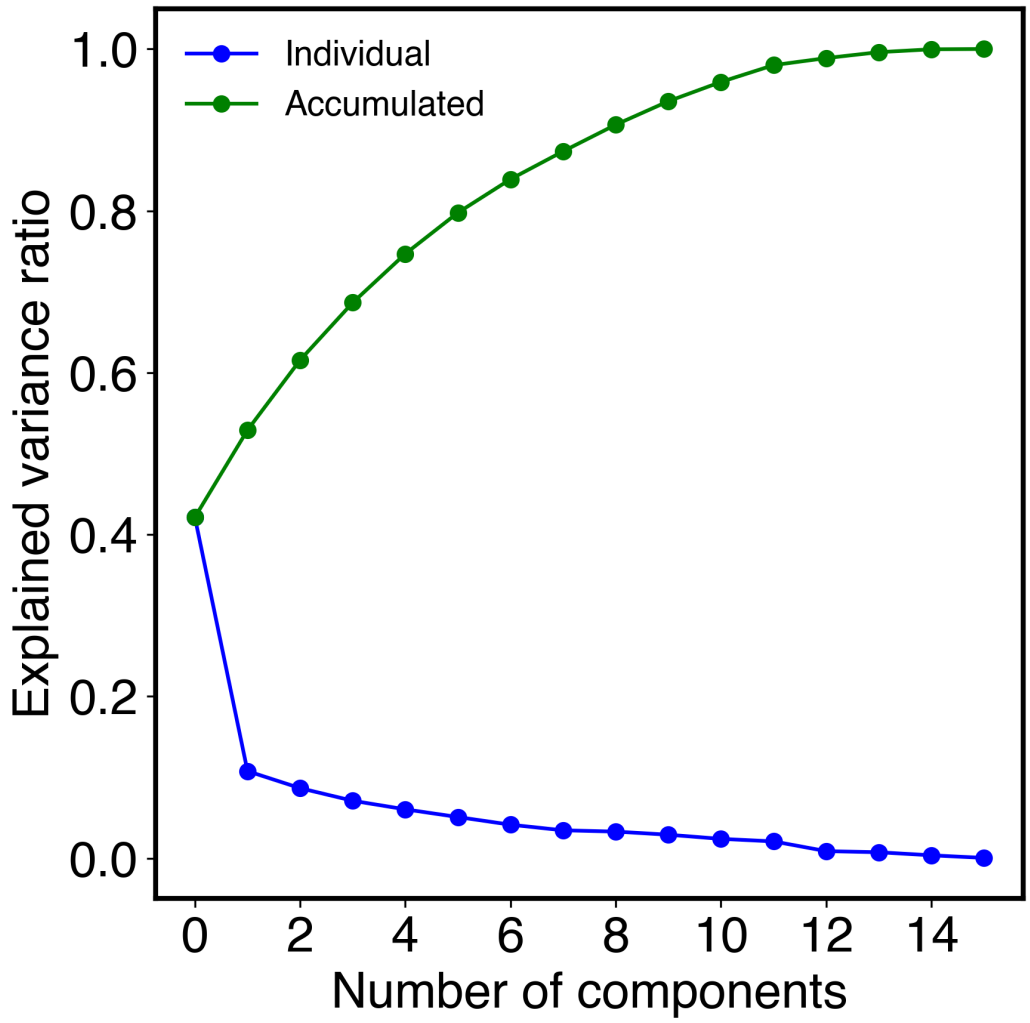

Figure 19: Explained variance ratio versus number of components.

3 Conclusion

To sum up, this dataset started with 21 unclean variables (including the target: life expectancy). There are many missing values, outliers and two variables do not make sense. The first step is to clean the data, which included detecting and dealing with missing values, outliers, and unreasonable variables. After obtaining the cleaned data, the next step is scaling the data using standardization to achieve good analysis. In this work, we used kMeans clustering method and performed Elbow method to select the optimal number of clusters. Finally, five clusters were constructed. According to the ANOVA test of both pair-wise clusters and all the clusters, there is significant difference of life expectancy among different clusters. Therefore, we can answer the stated question: countries with different economic, health, and social factors do have different life expectancies. To understand which cluster tends to have the lowest and highest life expectancy and the corresponding characteristics, we computed the mean value of each attribute for different clusters. The results indicate that countries with higher life expectancy share the following features: lower adult mortality, infant deaths, under-five deaths, thinness 1-19 years and thinness

5-9 years, and higher total expenditure, GDP, income composition of resources and schooling. In other words, countries with better economics and health conditions tend to have the higher life expectancy. Finally, we performed principal component analysis and shows that 7 components will be sufficient to explain 80% of variance in the data. Therefore, seven linear combinations of the 16 numeric attributes may be used to further compress the data.

4 Appendices

```
life.head(5)
```

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	G
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537

5 rows × 22 columns

Figure 20: Structure of the dataset. Each row is an instance, and each column is an attribute.

```
# percent of total data is missing
total_cells = np.product(life.shape)
total_missing = missing_values_count.sum()
(total_missing / total_cells) * 100
```

```
0.1349653677707799
```

Figure 21: Percent of total missing data. On average, around 0.13% data is missing.

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio
count	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000
mean	2007.518720	69.238462	164.695031	30.303948	4.637600	738.251295	82.644656	2419.592240	38.386555	42.035739	82.605344
std	4.613841	9.510459	124.092441	117.926501	3.921306	1987.914858	22.881890	11467.272489	19.939693	160.445548	23.362728
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000
25%	2004.000000	63.200000	74.000000	0.000000	1.082500	4.685343	82.000000	0.000000	19.400000	0.000000	78.000000
50%	2008.000000	72.100000	144.000000	3.000000	4.100000	64.912906	91.000000	17.000000	43.450000	4.000000	93.000000
75%	2012.000000	75.600000	227.000000	22.000000	7.390000	441.534144	96.000000	360.250000	56.100000	28.000000	97.000000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000

Figure 22: Basic statistics of the imputed data.

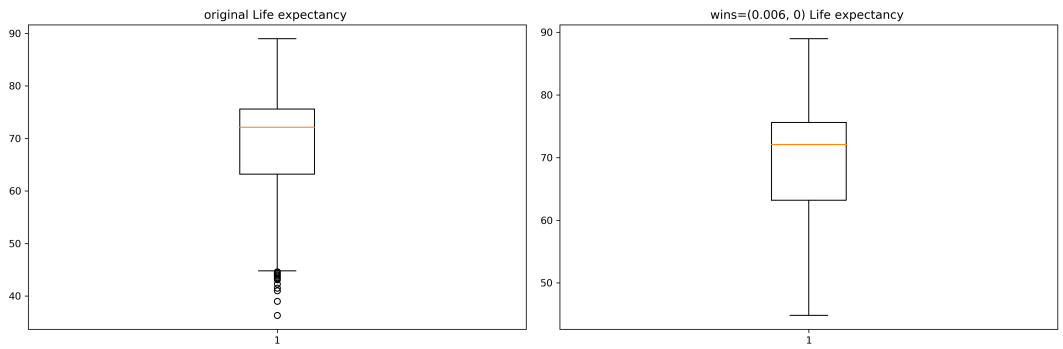

Figure 23: An example of comparison of box plots before and after winsorization.

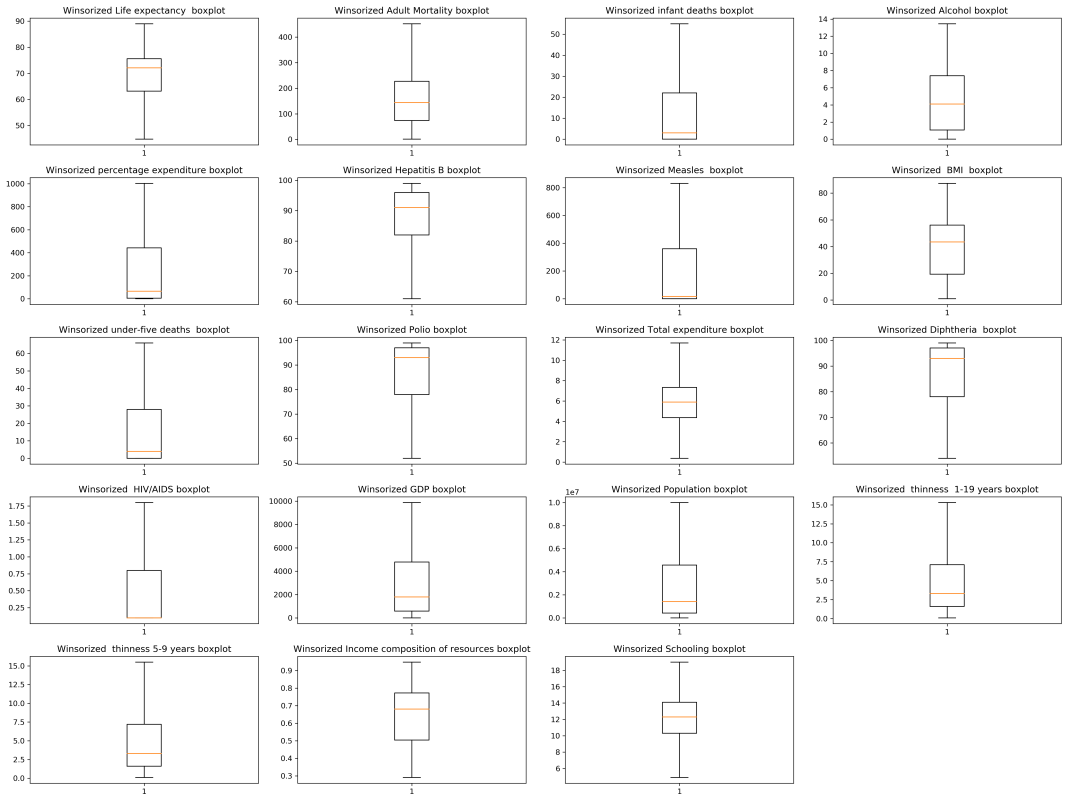

Figure 24: Box plots of winsorized data.

```
Explained variance by principal components:
PC1 0.4213845349701042
PC2 0.5289984318135044
PC3 0.6155498766354988
PC4 0.6865226061909595
PC5 0.7468892596389503
PC6 0.7977317748758083
PC7 0.8391130663068973
PC8 0.8735670263231305
PC9 0.9064241734679248
PC10 0.9354206627384677
PC11 0.9592827515609149
PC12 0.9800539174481899
PC13 0.9887557648556828
PC14 0.9960897852812163
PC15 0.9996328324285343
PC16 0.9999999999999999
```

Figure 25: Explained variance ratio of each principal component.

```
Eigenvalues:
[6.74444815e+00 1.72240860e+00 1.38529463e+00 1.13595031e+00
 9.66195317e-01 8.13757220e-01 6.62326097e-01 5.51451056e-01
 5.25893351e-01 4.64101794e-01 3.81923416e-01 3.32451810e-01
 1.39276964e-01 1.17384281e-01 5.67080559e-02 5.87668138e-03]
```

Figure 26: Eigenvalues of all the eigenvectors.