Lessons from countries with high life expectancy

`Exploratory Data Analysis`

Applied Multivariate Statistics	Rohith Krishna
Assignment 01 *25 November 2020*	PGDM in Research & Business Analytics Madras School of Economics `pgdm19rohith@mse.ac.in`

In this exercise, exploratory data analysis and visualization is carried out on **the Life Expectancy dataset.** The variables in this dataset have been collected from various sources such as WHO, UNO, UNICEF, IMF etc. The analysis has been carried out on `python` , `R` & `Tableau` and the corresponding plots and insights are stated here.

Purpose of EDA

- detection of mistakes
- checking violations of statistical assumptions like CLRM
- direction and size of relationships between variables.
- observing patterns
- preliminary model selection
- generate appropriate hypothesis
- pre-emptive tests preferred over alternatives that cost heavily later.

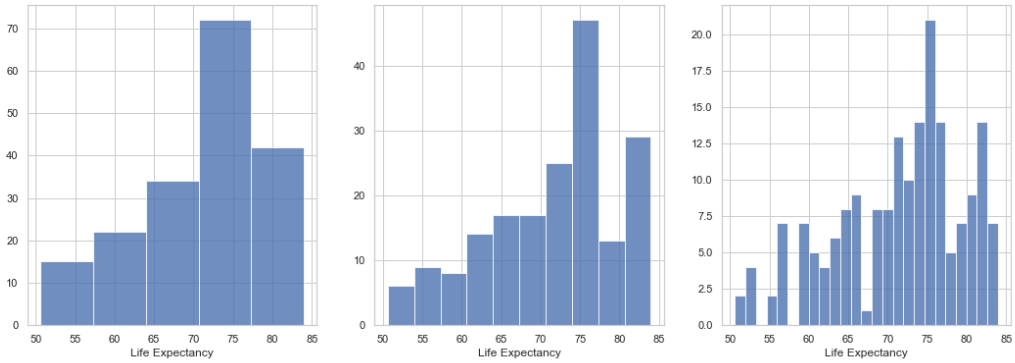
Data description

```
 #   Column   Description
---  ------   -----------------------------------------
 1   country  name of the country
 2   code     country codes
 3   lex      life expectancy at time of birth (UN Population Division)
 4   gdp      per capita GDP per annum in current USD (World Bank)
 5   lit      literacy rate for people aged 15 and above (UNESCO statistics)
 6   hex      health expense per capita in current USD (WHO)
 7   urb      number of people living in urban areas (UN Population Division)
 8   unt      undernourished population (Food & Agriculture Organization)
 9   phy      number of physicians for every 1000 people (WHO)
 10  san      % of people using basic sanitation services (UNICEF JMP)
 11  dri      % of people using basic water services (UNICEF JMP)
 12  fer      fertility rate (UN Population Division)
 13  smo      smoking prevalance [% of population] (WHO)
 14  alc      per capita alcohol consumption (WHO)
 15  dev      development category [high, upper-mid, lower-mid, low incomes]
```

Histograms

- Numeric (or continuous) data is analyzed using a histogram. The first step in a histogram is to **bin the continuous distribution into a certain number of quantiles.**
- For instance say 10 deciles, each entry in the `lex` distribution is put into one of the 10 bins and their corresponding frequency of occurrence is computed. This frequency-ratio can be interpreted as the probability distribution of the chosen `lex` variable.
- Note that the bins must be equal sized in terms of the variable range. They need not be equal sized in terms of number of observations in each bin.

Distribution of Life Expectancy `lex` - Histogram



Inference. We see that selection of number of bins essential for appropriate interpretation. If for instance, the first histogram tells us that about 70% of the observations (countries) have an average life expectancy between 70 and 77. If we choose a more granular binning, we find that the modal life expectancy is 75-76, further, the mean life expectancy across countries is 71 and the median is 74.

```
Summary Statistics for lex (life expectancy):
count    185.000000        25%       65.975000
mean      71.389138        50%       73.181000
std        8.133747        75%       76.917000
min       50.621000        max       83.980488
```

Sturges' Formula - optimal number of bins - a heuristic

In order to choose an appropriate number of bins for the histogram, a popular method is to use the Sturges' formula. Let k be the number of bins and n be the number of observations. Then, the optimal number of bins is:

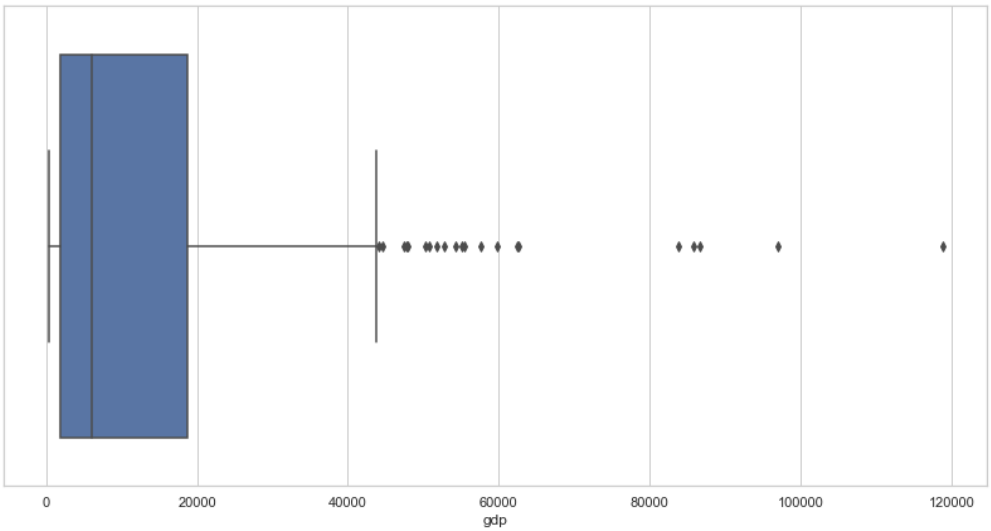
$$k = 1 + 3.3322 \log_{10}(n)$$

```
from math import log10
def get_num_bins(n):
    k = 1 +3.3322*log10(n)
    return round(k)
get_num_bins(len(df))

Out [1]: 9
```

Distribution of GDP per capita - Box Plots

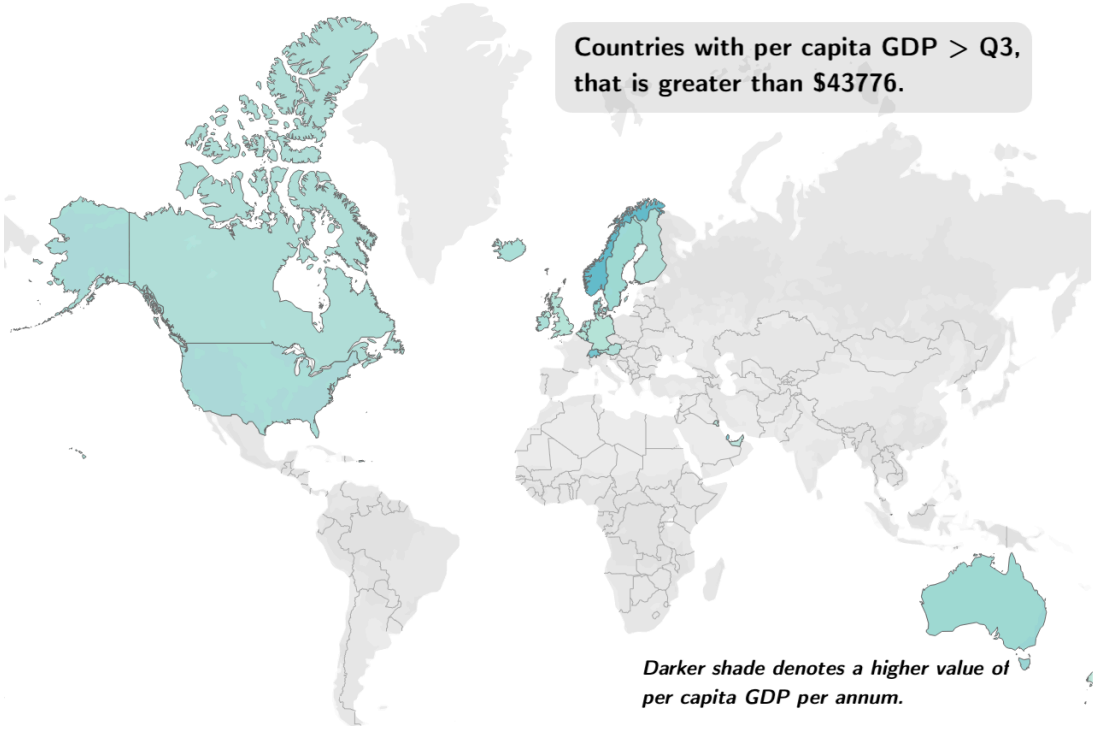
GDP per capita, much link any income distribution is highly skewed to the right. Most countries have a GDP per capita value less than 5000 USD per annum. Further, there are several outlier countries with a very high GDP per capita.



The box plot below suggests that all countries with GDP percapita above 43000 USD per annum, are outliers in the `gdp` distribution. These countries can now be isolated to further inspect any patterns in this data.

```
q1 = df['gdp'].quantile(q=0.25)
q3 = df['gdp'].quantile(q=0.75)
IQR = q3 - q1
upper_whisker = q3 + 1.5*IQR
list = df[df['gdp']>upper_whisker][['country','gdp']]
print(upper_whisker, list)

Out [1]: 43776.239976
Out [2]: Australia, Austria, Belgium, Bermuda, Canada, Switzerland, Germany,
Denmark, Finland, United Kingdom, Ireland, Iceland, Kuwait, Luxembourg,
Netherlands, Norway, New Zealand, Qatar, Singapore, Sweden, United States,
```



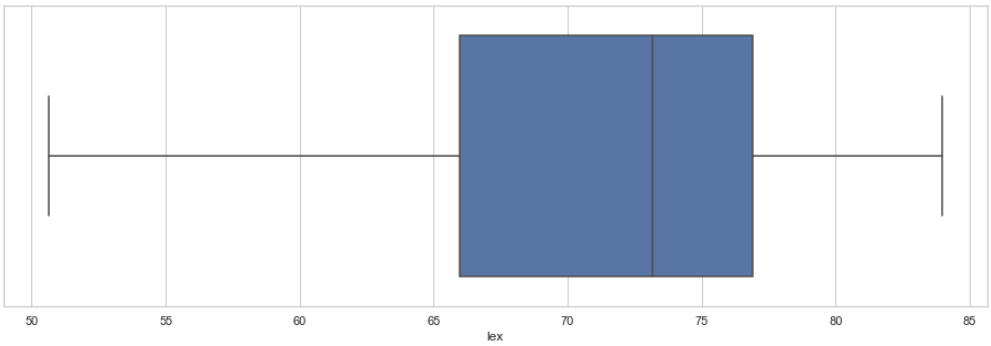
Pivot Table

In order to understand clearly, the consequences of GDP on life expectancy in these outlier countries, we shall inspect the data further. Simple pivot tables prove very useful in this scenario.

Life Expectancy for outlier countries wrt GDP

Entity	Lex	Gdp	Hex	hex/gdp	Lit	Urb	Fer	Phy
Switzerland	83	86,606	10,015	11.56%	99	74	1.54	4.11
Iceland	83	54,242	4,515	8.32%	99	94	1.93	3.63
Singapore	82	57,563	2,215	3.85%	96	100	1.25	2.04
Australia	82	62,511	5,638	9.02%	99	86	1.83	3.46
Sweden	82	59,844	6,628	11.08%	99	86	1.88	4.20
Luxembourg	82	118,824	7,757	6.53%	99	90	1.50	2.86
Norway	82	97,019	9,118	9.40%	99	81	1.75	4.43
Canada	82	50,836	5,038	9.91%	99	81	1.58	2.50
Netherlands	82	52,830	5,676	10.74%	99	90	1.71	3.42
Austria	81	51,717	5,386	10.42%	100	58	1.46	5.00
New Zealand	81	44,534	4,145	9.31%	99	86	1.92	2.81
Ireland	81	55,493	5,391	9.71%	99	62	1.89	2.77
United Kingdom	81	47,418	4,568	9.63%	99	82	1.81	2.78
Belgium	81	47,701	4,909	10.29%	99	98	1.74	2.97
Finland	81	50,260	4,733	9.42%	99	85	1.71	3.21
Germany	81	47,960	5,278	11.01%	99	77	1.47	4.08
Bermuda	81	85,748	10,830	12.63%	98	100	1.63	0.09
Denmark	81	62,549	6,381	10.20%	99	87	1.69	3.64
United States	79	55,033	9,053	16.45%	99	81	1.86	2.58
Qatar	78	83,859	2,104	2.51%	91	99	1.95	1.80
Kuwait	74	44,062	1,375	3.12%	96	100	2.09	2.59
Grand Total	**81**	**62,696**	**5,750**	**9.29%**	**98**	**86**	**1.72**	**3.09**

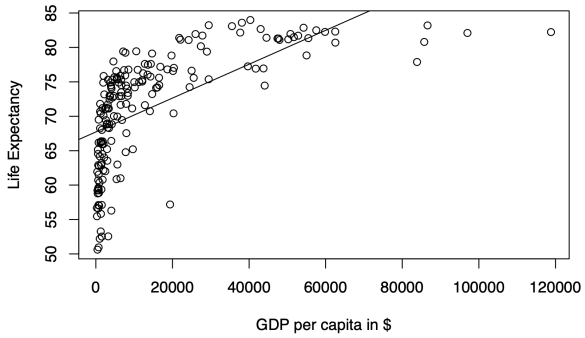
Distribution of life expectancy `lex`

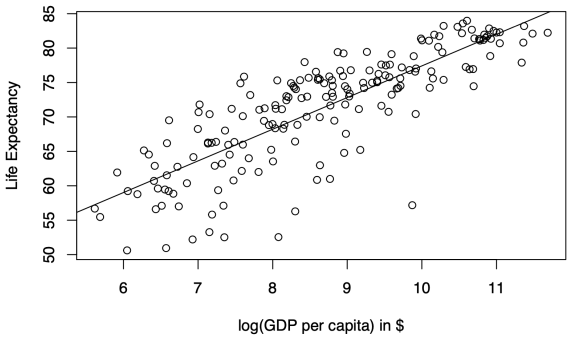


We observe from the above table and the box plot that for high GDP countries, the life expectancy is also found to be quite high. Literacy rate is a high 98% and the number of physicians is an average 3 per 1000 population, which is again a high figure. Likewise, we find that these countries with exceedingly high ratio of health expenditure to GDP per capita. On an average these countries spend about 9.3% of their national income on health and health related factors. While we cannot prove any relation conclusively from this, we still understand that there appears to be some relation between GDP per capita, Health Expenditure and life expectancy.

GDP per capita and Life Expectancy - Scatter plots

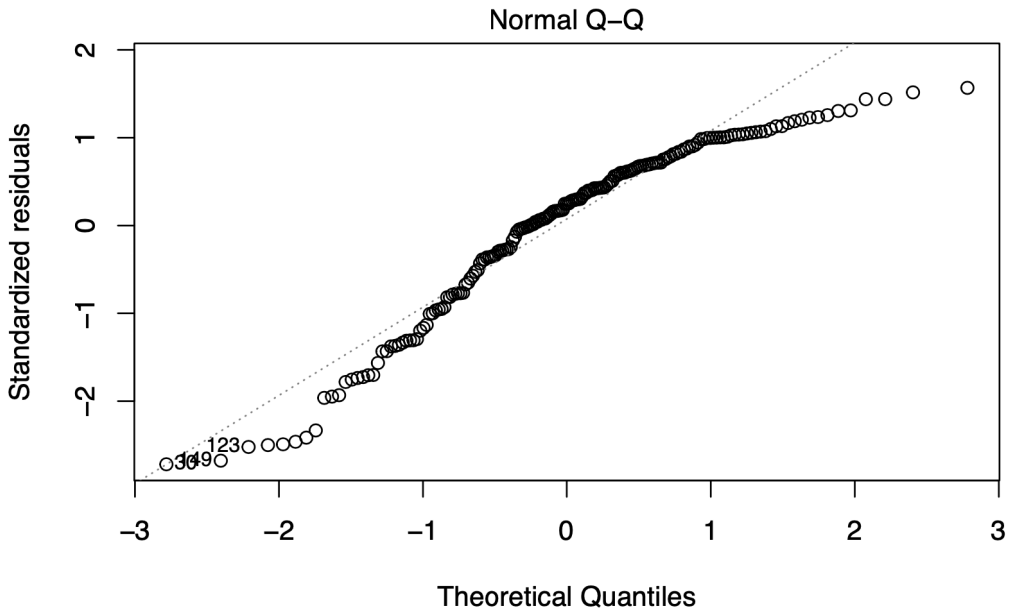
Scatter plot are used in finding correlations or comovements between two numerical variables. Here, we plot life expectancy as a function of GDP per capita. We immediately find out that a linear relationship between these variables does not exist. In fact the empirical **Preston curve** suggests that this relationship is infact non-linear and logarithmic. If we plot the relationship between `log(GDP)` and `lex` we observe a linear trend.



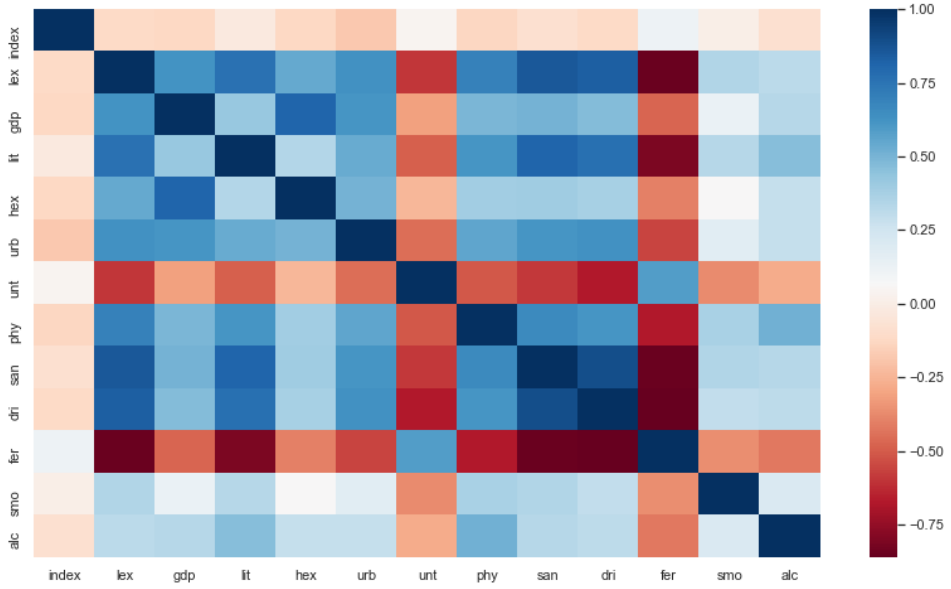


Violation of statistical assumptions

Heteroscedasticity. From the Q-Q plot we make out whether the errors in the data are normally distributed and thereby have constant variance. Any deviation from the straight line in the Q-Q plot tells us non-constant variance of errors - also known as heteroscedasticity.

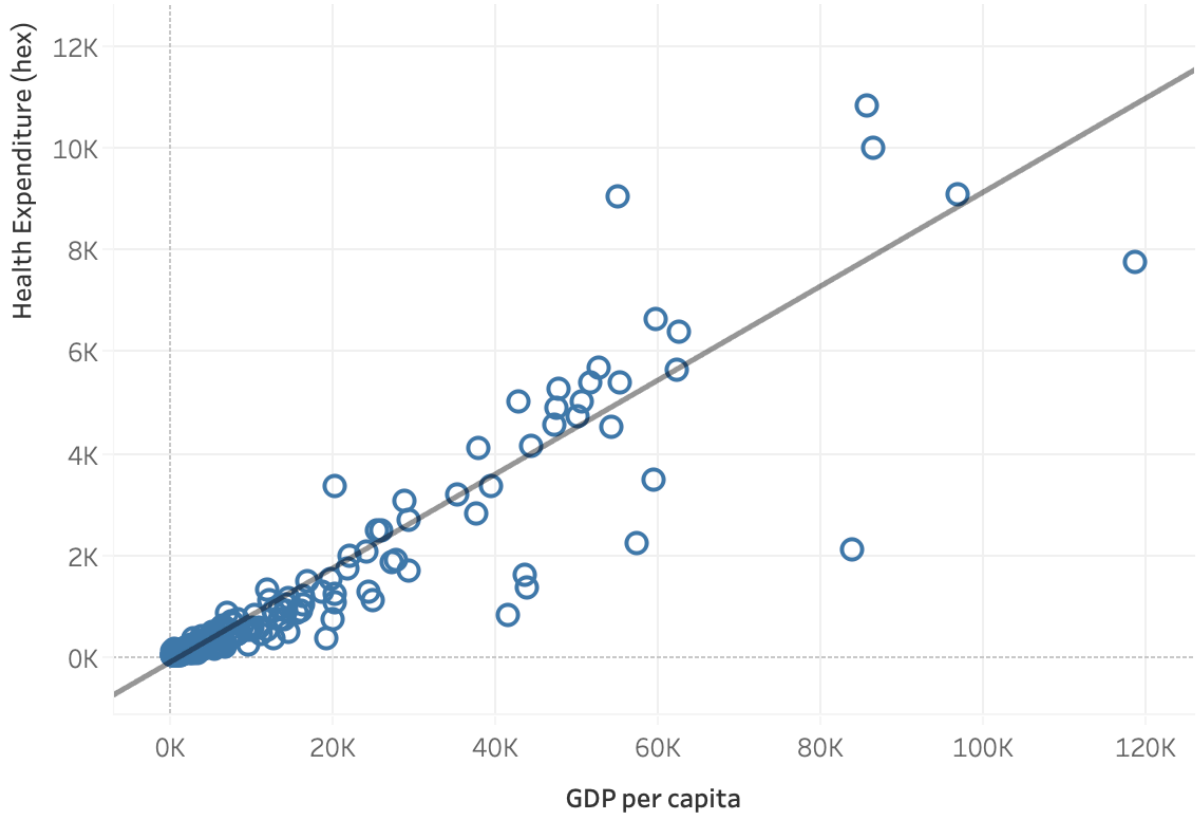


Multicollinearity. It might be the case that rich countries also spend higher in health expenditure. One can eother check for multicollinearity using correlation plots or through statistical tests. From the correlation plot below, qw observe several correlations between variables to different extents.



Scatter Plots between `gdp` and `hex`

As expected, the plot between `gdp` and `hex` is a linear one, with very few outlier datapoints. Thus in order to properly model the linear relationship between `gdp` and `lex` one must ensure that multicollinearity is absent.



Categorical Variable

Before concluding we make a final remark on our categorical variable. Giving testament to our hunch that income and life expectancy are related, we plot the income-group-wise distribution of life expectancy. Certainly, it turned out that countries with a higher life expectancy also belonged to the high-income category. This is seen the following plot: