Coursework 1: Using Big Data Tools

Big Data Engineering

(Report)

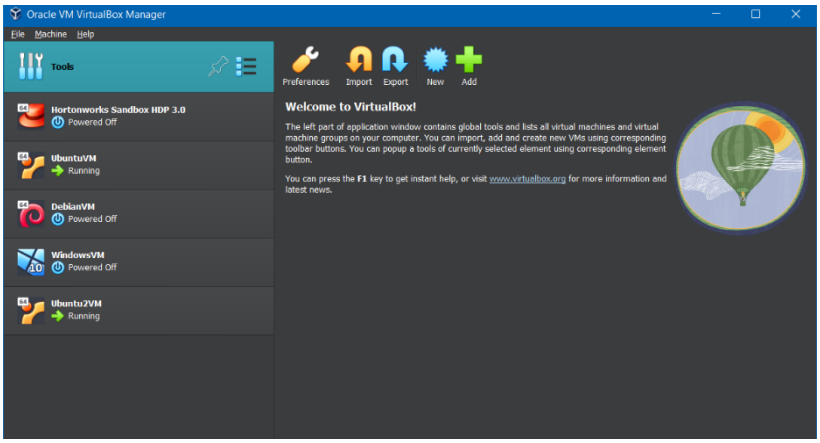
Stefan Velev, 0MI3400521

Introduction

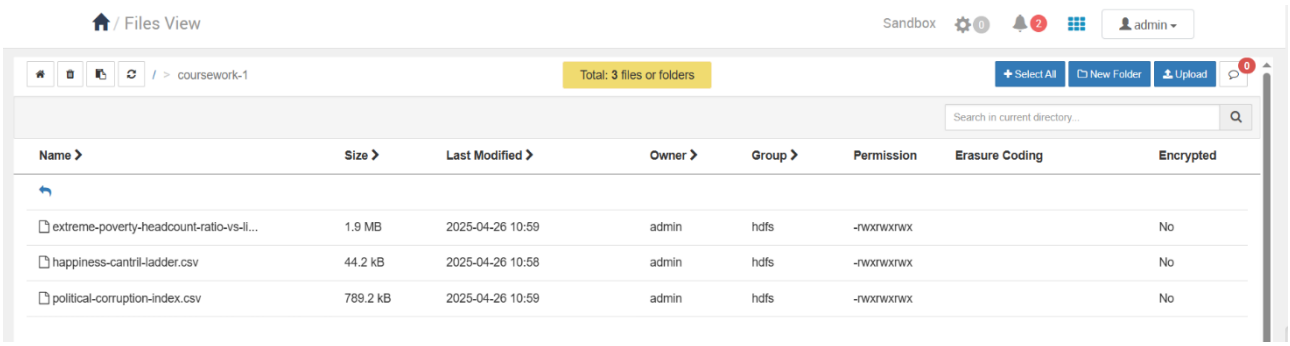
This coursework assignment covers the typical **data lifecycle** from the data source to the end user who is looking into the data and the results from its analysis using statistical and data analytics methods. It is primarily focused on the tools from the **Hadoop ecosystem** as the most popular platform for processing **Big Data**.

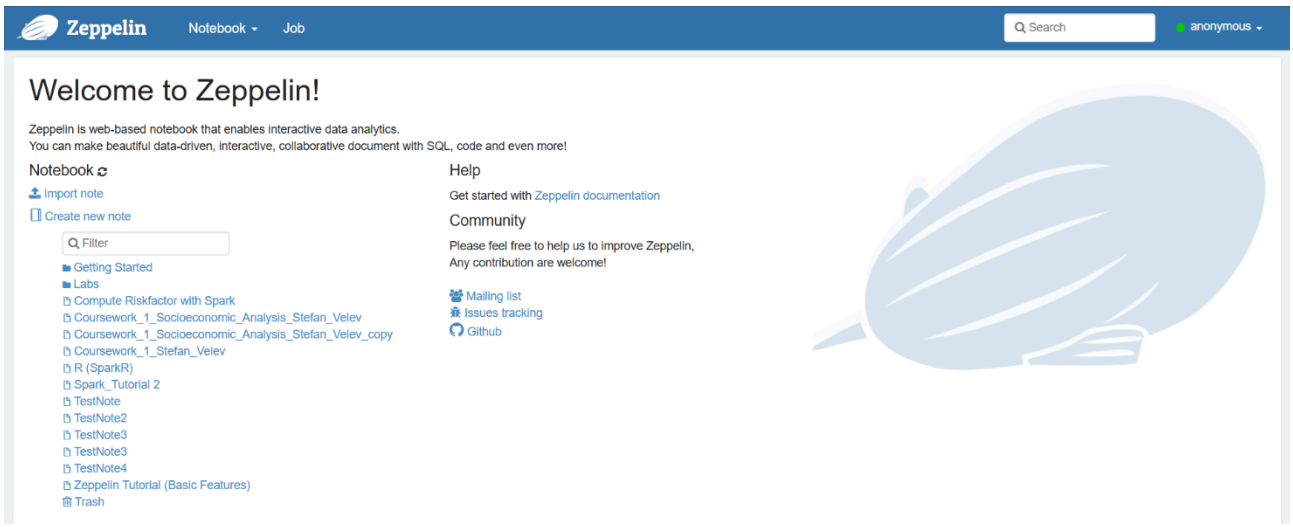
Software Tools Used

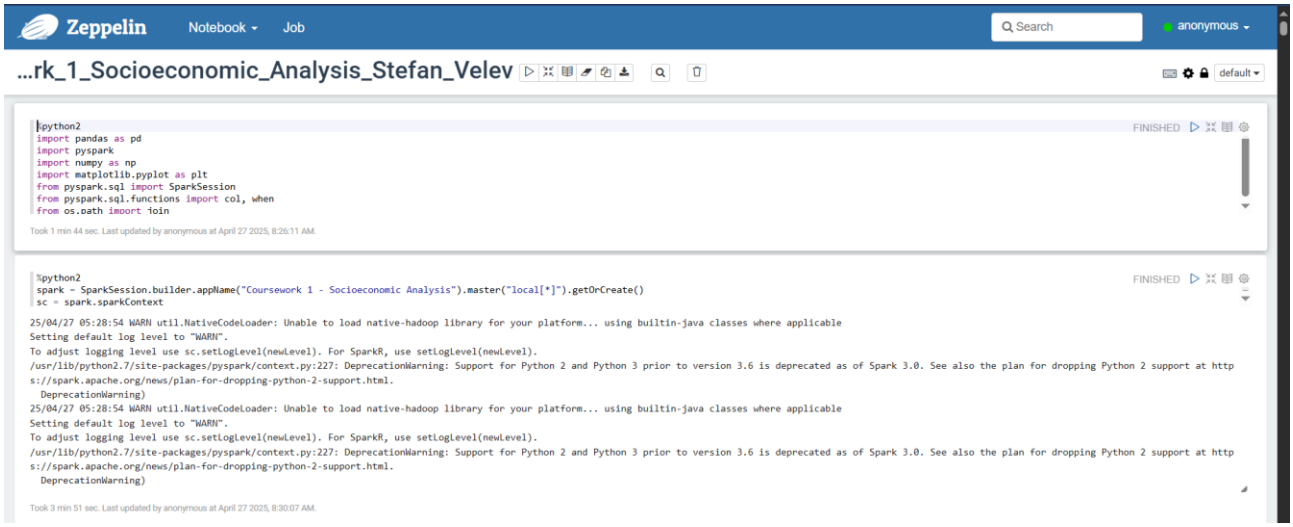
The software tools for this coursework assignment are from the Hadoop ecosystem. They are entailed in **Hortonworks Sandbox HDP 3.0** that I have entirely relied on. I have loaded it in **Oracle VM VirtualBox**.



The tools from the **Hadoop ecosystem** that I have used are: **Ambari** for accessing the UI and loading locally the CSV files that I have used for performing different kinds of analyses, **HDFS** is the distributed file system that I have used for storing the CSV files, and **Apache Spark** with **Python** (**PySpark**) in **Zeppelin** environment. I have used the **Shell In a Box** (localhost:4200) for working with **HDFS**, and installing some **Python** packages such as ***Pip***, ***Pandas***, ***PySpark***, ***Matplotlib*** that I have relied on in the analysis part.

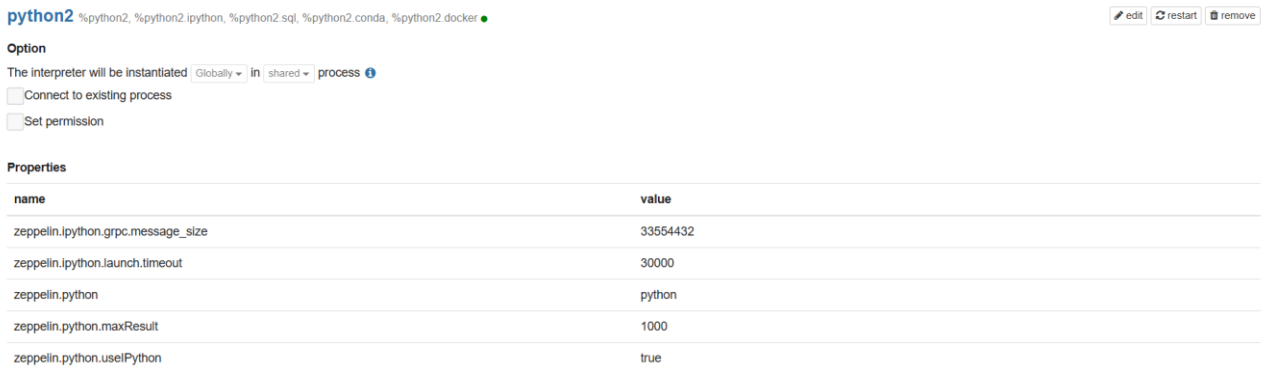






Data Modelling and Analysis – Apache Spark with Python (PySpark)

The packages I've used in this assignment are: ***PySpark*** (for working with Big Data efficiently in Spark Data Frames & SQL), ***Pandas*** (for reading and storing data), ***NumPy*** (for making statistics), and ***Matplotlib*** (for visualisation). In order to use **Python** environment in **Zeppelin** I had to create an interpreter for that purpose:



It was a challenge to configure ***PySpark***, ***Pandas*** and ***Matplotlib*** on the ***Python*** version that is already installed in the sandbox because it was an older one (*Python 2.7.5*) and I could not update it because

the services from the Hadoop ecosystem in the sandbox are already configured to work with it. Therefore, the packages that I downloaded had to be older versions as well. Installing ***Matplotlib*** was the greatest challenge because it tried to replace internally a package that is fixed in the sandbox and could not be deleted. Therefore, I had to install it with the option *--ignore-installed* that means "do not uninstall old stuff, just install what I need locally". Another problem was with the backend that ***Matplotlib*** relied on. Since **HDP Sandbox** is a server, it has no GUI, so I needed to tell the package to use a non-GUI backend like **Agg**, which only creates images. I did that with the help of hidden configuration folder with a file that ***Matplotlib*** used for changing the backend from **Tkinter** to **Agg**.

The **CSV** files are stored in the **HDFS** file system and read from there with the help of ***PySpark***:

```
%python2
path_data_1 = "hdfs:///coursework-1/happiness-cantril-ladder.csv"
df_life_satisfaction = spark.read.format("csv") \
  .option("sep", ",") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .load(path_data_1)
```

They are stored in **Spark Data Frames** that are optimised for working with distributed data. For the analysis part I have relied entirely on them. For the visualisations with ***Matplotlib*** I have converted the data in **Pandas Data Frames** since ***Matplotlib*** cannot work with **Spark Data Frames** and such kind of detailed plots cannot be achieved only with the visualisation properties of the **Zeppelin** environment.

The **Python** code for the **Zeppelin** notebook will be attached to the submission of the coursework assignment. One of the most challenging parts with the processing of data was with the syntax of ***PySpark*** which is different from the way we work with ***Pandas Data Frames***. However, it is still intuitive when additional literature such as books and tutorials for reference.

```
%python2
df_life_satisfaction = (
    df_life_satisfaction
    .select('Entity', 'Year', 'Cantril ladder score')
    .withColumnRenamed('Entity', 'Country')
    .filter(col('Year') == 2021)
    .na.drop()
)
```

For the **data modelling** part I have relied on the ***PySpark Data Frames*** mechanism for inferring the **data schema** of the CSV files. In fact, the results turned out to be satisfactory for all the three datasets.

```
happiness-cantril-ladder.csv
|-- Entity: string (nullable = true)
|-- Code: string (nullable = true)
|-- Year: integer (nullable = true)
|-- Cantril ladder score: double (nullable = true)

extreme-poverty-headcount-ratio-vs-life-expectancy-at-birth.csv
|-- Entity: string (nullable = true)
|-- Code: string (nullable = true)
|-- Year: integer (nullable = true)
|-- Life expectancy - Sex: all - Age: 0 - Variant: estimates: double (nullable = true)
|-- $2.15 a day - Share of population in poverty: double (nullable = true)
|-- 990305-annotations: string (nullable = true)
|-- Population (historical): long (nullable = true)
|-- World regions according to OWID: string (nullable = true)

political-corruption-index.csv
|-- Entity: string (nullable = true)
|-- Code: string (nullable = true)
|-- Year: integer (nullable = true)
|-- Political corruption index (best estimate, aggregate: average): double (nullable = true)
```

I've downloaded the three data sets from ***Our World in Data***© which I will analyse later.

1. Self-reported life satisfaction

Average of responses to the 'Cantril Ladder' question in the Gallup World Poll. The survey asks respondents to think of their current place on a ladder, with the best possible life for them being a 10, and the worst possible life being a 0.

Data URL: *https://ourworldindata.org/grapher/happiness-cantril-ladder?time=latest*

Data sources: *World Happiness Report (2012-2024); Wellbeing Research Centre (2024); Population based on various sources (2023)*

The data is stored in a *csv file*. The name of the separate columns can be found in the *Zeppelin notebook* accompanying this report. The *csv file* contains 1 787 entries. However, most of them will not be used for our purposes. Here comes the **data cleaning** part. First, I decided to remove the columns which are not required for the task. In that case it is only one column – 'Code'. After that, I renamed one of the left columns with more concise names:

'Entity' → 'Country'

The next step was to choose the year which I wanted to examine. Since I'd like up-to-date statistics, my focus was on 2021. The reason not to choose more contemporary year is that not all entries are updated for later period which would disrupt the analyses later. Then I dropped the rows with missing values (if there are any). After examining the data, I decided to remove the following aggregated entries with country field: 'High-income countries', 'Low-income countries', 'Lower-middle-income countries', 'Upper-middle-income countries', 'World'. The reason is that they are not necessary for the statistics about individual countries.

Since I wanted to analyse data for the separate continents (such data is also provided in the data set) as well, I decided to do **data segregation** and extract the entries for the continents into another data set. After that, I removed these rows from the original data set (without the entry for 'Australia' since it is a country and a continent). Finally, I had a *DataFrame* with 147 entries which is a satisfactory result since the total number of countries is 195 and I assume entries for some of them are probably missing or incomplete.

2. Share in extreme poverty vs. life expectancy

The period life expectancy at birth, in a given year. Extreme poverty is defined as living below the International Poverty Line of $2.15 per day.

Data URL: *https://ourworldindata.org/grapher/extreme-poverty-headcount-ratio-vs-life-expectancy-at-birth*

Data sources: *UN, World Population Prospects (2024); World Bank Poverty and Inequality Platform (2024); HYDE (2023); Gapminder - Population v7 (2022); Gapminder - Systema Globalis (2022)*

The data is stored in a *csv file*. The name of the separate columns can be found in the *Zeppelin notebook* accompanying this report. The *csv file* contains 60 100 entries. However, most of them will not be used for our purposes. Here comes the **data cleaning** part. First, I decided to remove the columns which are not required for the task – 'Code', '990305-annotations', 'World regions according to OWID'. After that, I renamed some of the left columns with more concise names:

'Entity' → 'Country'

'Life expectancy – Sex: all – Age: 0 – Variant: estimates' → 'Life expectancy'

'$2.15 a day – Share of population in poverty' → 'Share in extreme poverty'

'Population (historical)' → 'Population'

The next step was to choose the year which I wanted to examine. Since I'd like up-to-date statistics, my focus was on 2021. The reason not to choose more contemporary year is that not all entries are updated for later period which would disrupt the analyses later. Then I dropped the rows with missing values (if there are any). After examining the data, I decided to remove the following aggregated entry with country field: 'World'. The reason is that it is not necessary for the statistics about individual countries.

While inspecting the data, I noticed that there was an outlier. That is for the life expectancy of the Central African Republic for 2021 – according to the *World Health Organization Data* for 2021[1] it is 52.31 years, not 40.279 years. For the purposes of the task and for better visualisation, I decided to correct it.

3. Political corruption index

Based on the expert estimates and index by V-Dem. It captures the extent to which the executive, legislative, judiciary, and bureaucracy engage in bribery and theft, and the making and implementing of laws are susceptible to corruption.

Data URL: https://ourworldindata.org/grapher/extreme-poverty-headcount-ratio-vs-life-expectancy-at-birth

Data sources: V-Dem (2024)

The data is stored in a *csv file*. The name of the separate columns can be found in the *Zeppelin notebook* accompanying this report. The *csv file* contains 33 090 entries. However, most of them will not be used for our purposes. Here comes the **data cleaning** part. First, I decided to remove the columns which are not required for the task. In that case it is only one column – 'Code'. After that, I renamed some of the left columns with more concise names:

'Entity' → 'Country'

'Political corruption index (best estimate, aggregate: average)' → 'Political corruption index'

The next step was to choose the year which I wanted to examine. Since I'd like up-to-date statistics, my focus was on 2021. The reason not to choose more contemporary year is that not all entries are updated for later period which would disrupt the analyses later. Then I dropped the rows with missing values (if there are any). After examining the data, I decided to remove the following aggregated entry with country field: 'World'. The reason is that it is not necessary for the statistics about individual countries.

Since I wanted to analyse data for the separate continents (such data is also provided in the data set) as well, I decided to do **data segregation** and extract the entries for the continents into another data set. After that, I removed these rows from the original data set (without the entry for 'Australia' since it is a country and a continent). Finally, I had a *DataFrame* with 180 entries which is a satisfactory result since the total number of countries is 195 and I assume entries for some of them are probably missing or incomplete.

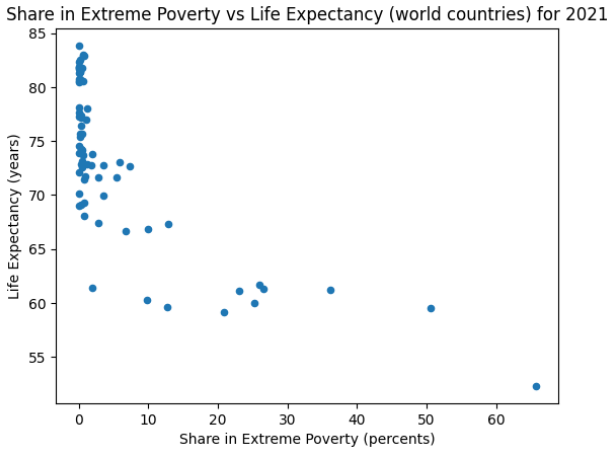
[1] https://data.who.int/countries/140

Data Visualisation – Matplotlib

For the visualisation tasks I've made several plots using different approaches. They would help me answer some meaningful questions which is always the point when we make analyses. The rationale behind choosing these data sets is connected to the type of questions to which I wanted to receive an answer. We will start looking at the separate plots and to the questions that they answer. We will state any assumptions and motivate any decision when selecting data to be plotted, and in combining data. We will discuss any observations or insights obtained from the **data visualisations**.

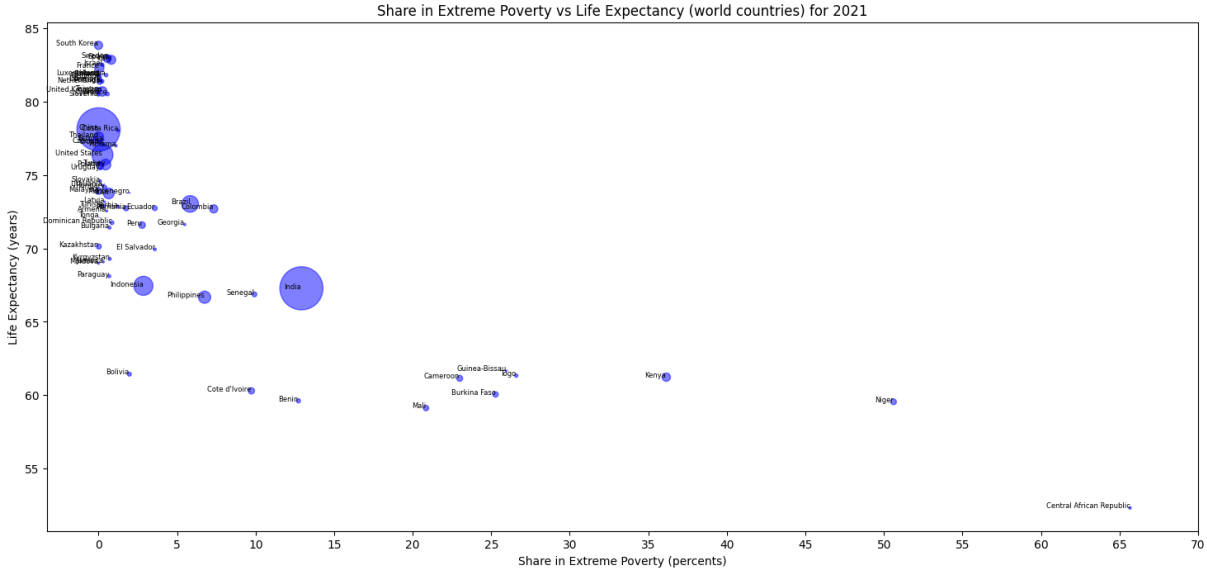
I. A scatter plot of Share of Population in Extreme Poverty vs Life Expectancy for 2021

The first scatter plot I've presented is one made with the help of ***Pandas***. This diagram can only be used for an overall picture of how data is distributed. That is why the name of the countries is not shown on the scatter plot.

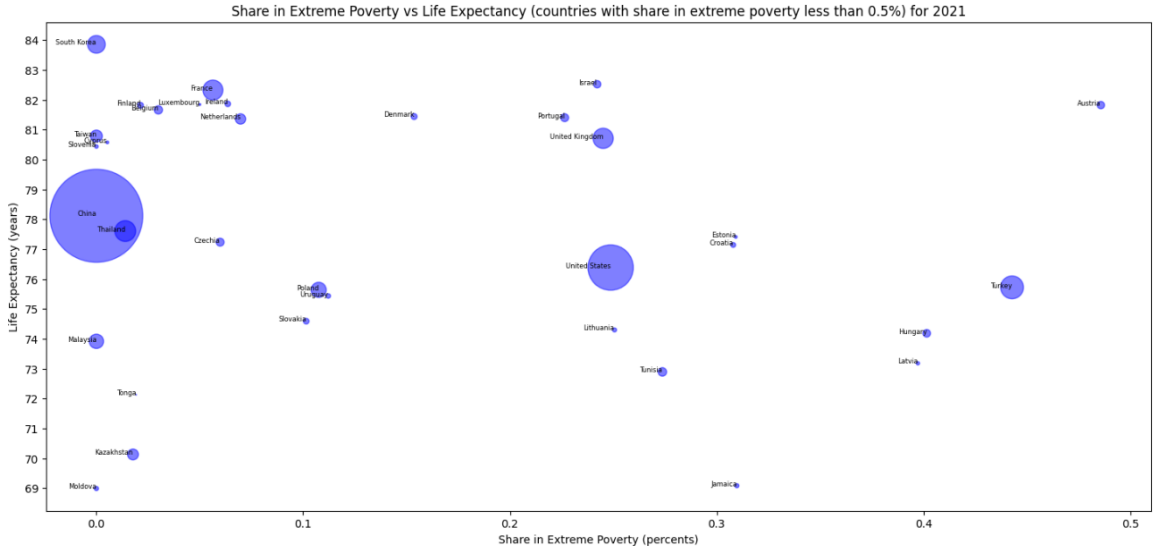


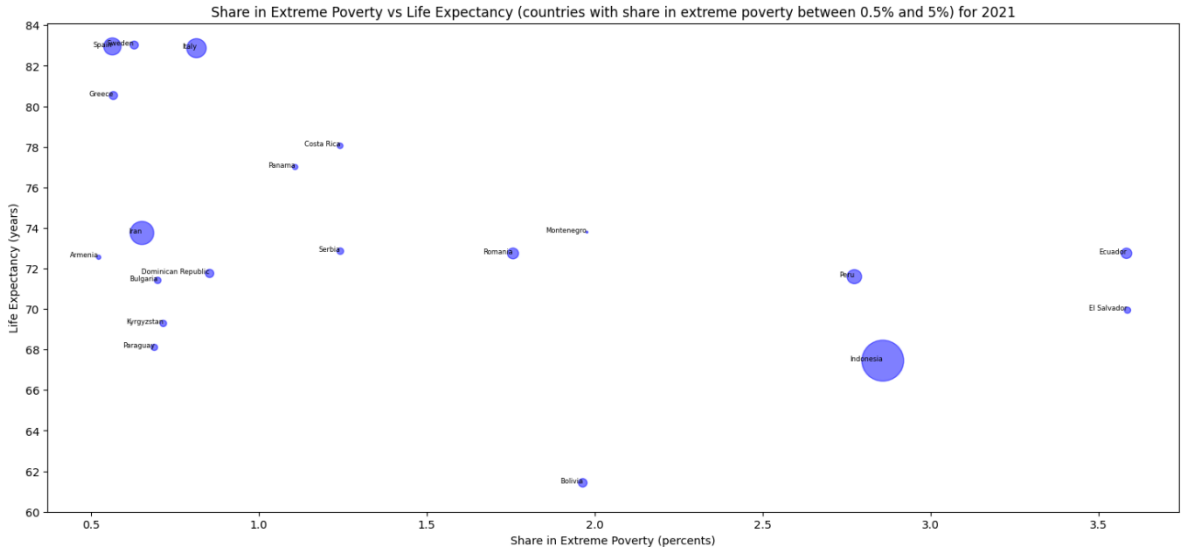
The most detailed scatter plot is the second one for which I've used ***Matplotlib***. For that, some considerations were taken into account:

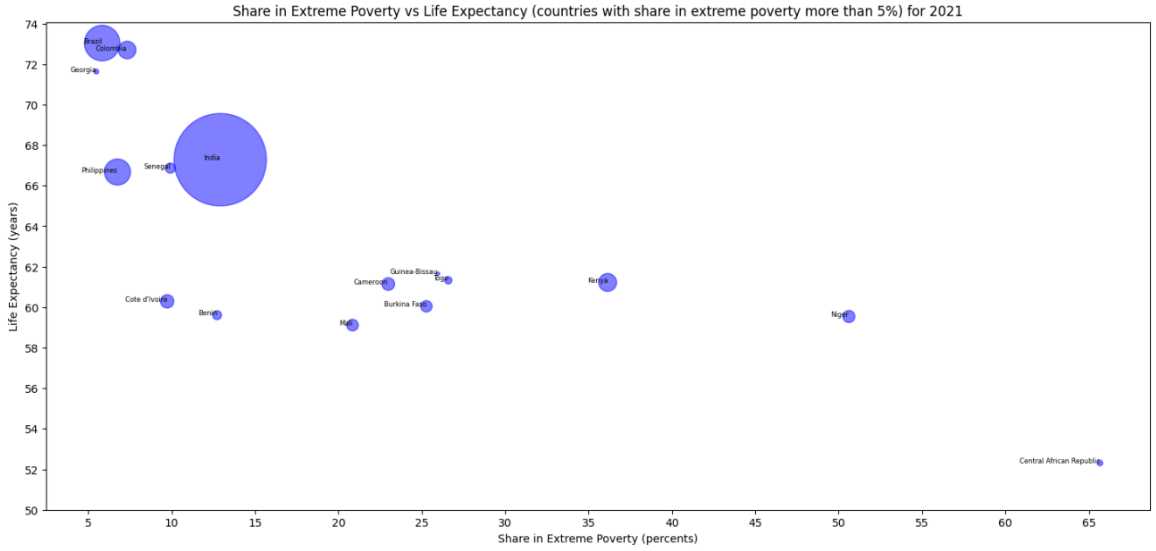
- the scatter plot is much larger as the number of entries is 70 (most of them settled in one area)
- the circles of the scatter plot are proportional to the population of each country for 2021
- the name of each country can be found next to its circle – for better clarity it would be good if it is visualised only on hoovering over it; however this cannot be achieved with ***Matplotlib***



However large the scatter plot is, so many names of countries cannot be presented on just one diagram. That's why I've taken the decision to divide this scatter plot into three separate ones. The criteria used is 'Share in extreme poverty' – less than 0.5% in the first scatter plot, between 0.5% and 5% in the second scatter plot, above 5% in the third scatter plot. In that way, the names of the countries become clearer.





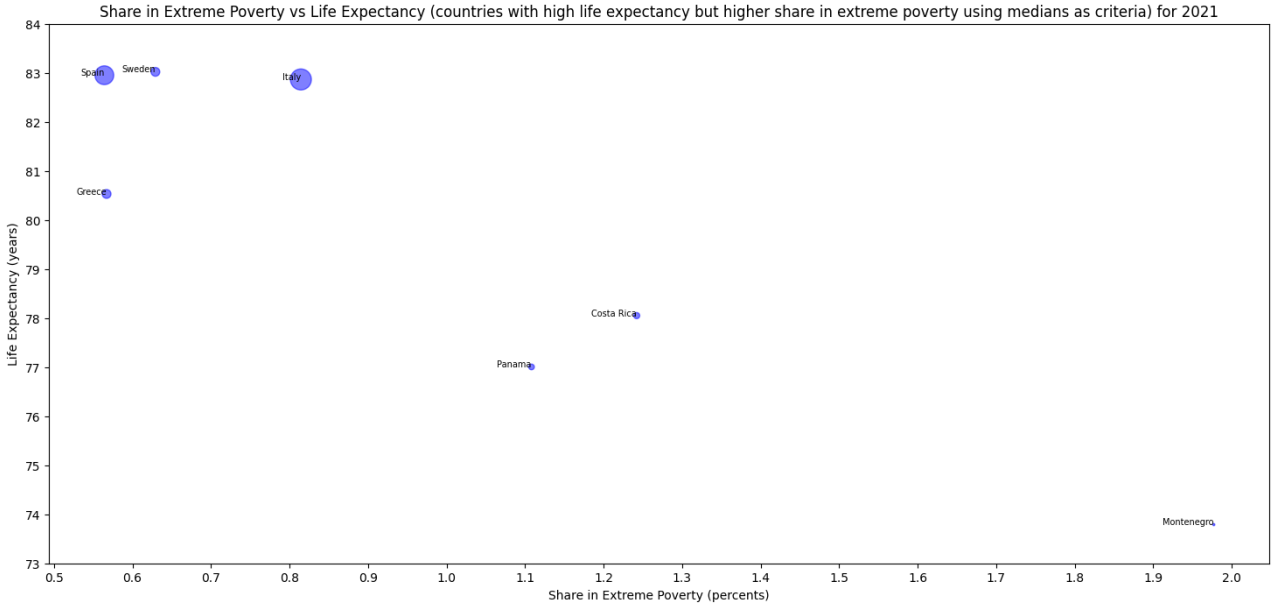


Questions to be answered with the help of the scatter plot:

1. Check which countries have high life expectancy but have higher share in extreme poverty for 2021?

With ***NumPy*** I've determined that the median life expectancy of the sample is 73.77 and the median share in extreme poverty is 0.54. So, using ***Pandas*** I've concluded that the number of the countries for 2021 which have life expectancy above the found median and above the median share in extreme poverty is 7 and they are as follows:

Costa Rica, Greece, Italy, Montenegro, Panama, Spain, Sweden

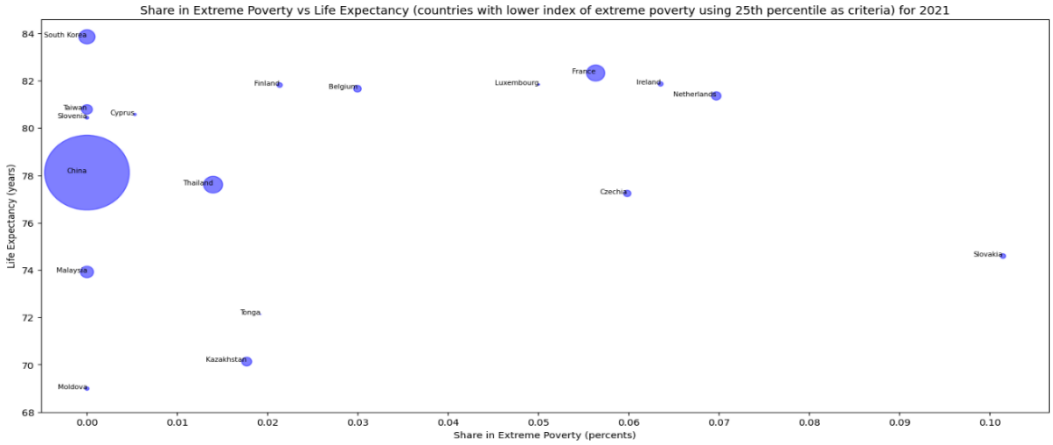


If we use the mean as criteria for high life expectancy and high share in extreme poverty, we will find that there are no such countries. This means we can draw a conclusion that the countries with high life expectancy tend to have a smaller share in extreme poverty. That is natural since high life expectancy means that local people maintain a healthier lifestyle which contrasts with extreme poverty. We can notice the great difference in the mean and the median for the column 'Share in extreme poverty'. The mean is 5.41 and the median is 0.54. By definition, the mean is the number we get by dividing the sum of a set of values by the number of values in the set. In contrast, the median is the middle number in a set of values when those values are arranged from smallest to largest. The big difference is due to the skewness of the data.

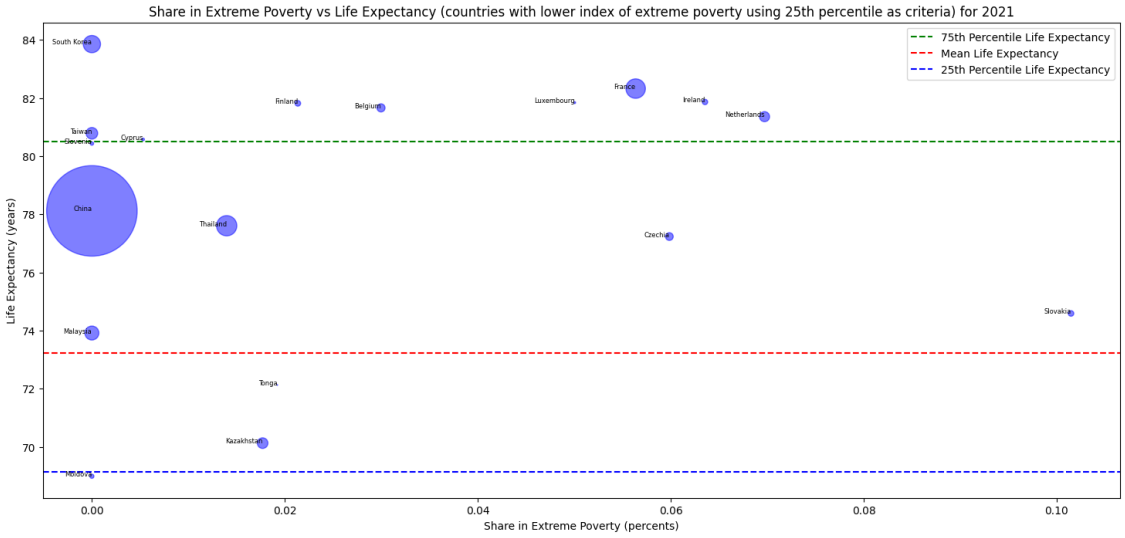
2. Find whether each country with lower share in extreme poverty have high life expectancy?

For defining lower share in extreme poverty, I use the 25th percentile of the column 'Share in extreme poverty'. This will be my upper boundary for the searched countries. The exact share is 0.1029%. With that consideration the searched number of countries is 18 and they are:

Belgium, China, Cyprus, Czechia, Finland, France, Ireland, Kazakhstan, Luxembourg, Malaysia, Moldova, Netherlands, Slovakia, Slovenia, South Korea, Taiwan, Thailand, Tonga



For defining higher life expectancy, I use the 75th percentile of the column 'Life expectancy'. This will be my lower boundary for high life expectancy. In addition to that, I find the mean life expectancy and the 25th percentile representing the upper boundary for low life expectancy. I will provide the answer to the question by plotting and observing the following scatter plot:



The circles represent each country with lower share in extreme poverty. Now we have to see if all of the above countries have high life expectancy. Apparently, 9 countries are above the 75th percentile of life expectancy which is enough to call them countries with higher level of life expectancy. 6 countries are within the mean and the 75th percentile of life expectancy. As they are above the average value, we can call them countries with high level of life expectancy. 2 of the countries are under the mean life expectancy but still above the 25th percentile which classifies them as countries with low life expectancy. There is one country (Moldova) that is below the 25th percentile of life expectancy which means that it is a country with one of the lowest life expectancies. This can answer the question that there are countries with lower share of extreme poverty that are still further away from high life expectancy standard.

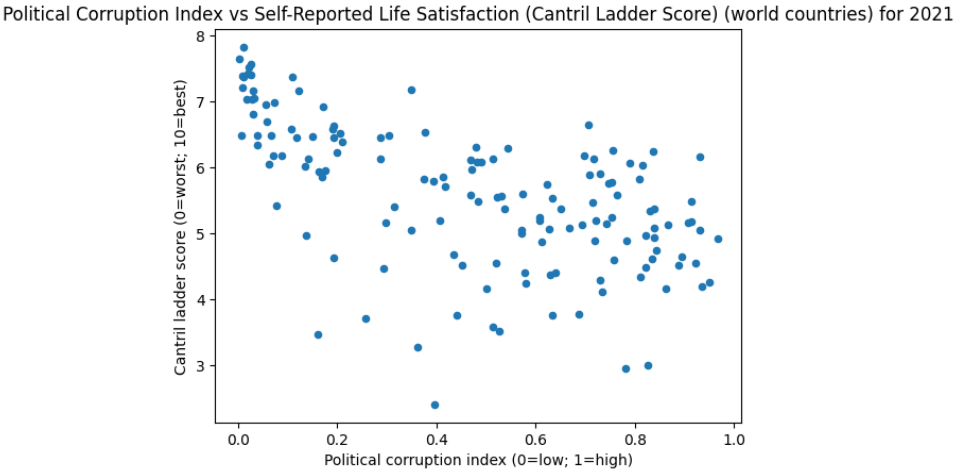
From the obtained results for the 18th countries we can notice something odd. Countries like China, Malaysia, Moldova, Slovenia, South Korea, Taiwan have 0% share in extreme poverty. Most of these entries (e.g. Moldova) have missing values for that column. That is why it is important to verify the results for outliers.

II. A scatter plot of Self-Reported Life Satisfaction (Cantril Ladder Score) vs Political Corruption Index for 2021

I want to make a scatter plot merging two separate data sets. For that purpose, I used ***PySpark*** and its *join* function to perform **inner join** on the 'Country' column.

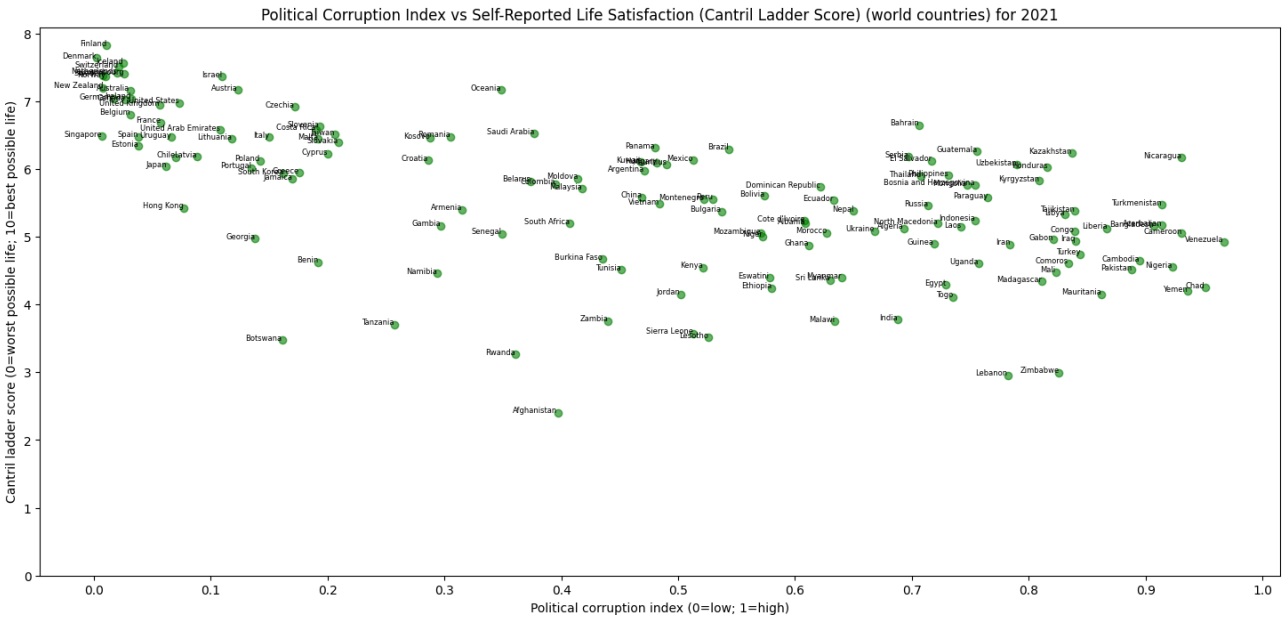
```
df_life_satisfaction_political_corruption = df_life_satisfaction.join(df_political_corruption,
df_life_satisfaction.Country == df_political_corruption.Country, "inner")
```

As a result, I get a data frame with 145 rows. On the following scatter plot I've done with the help of ***Pandas***, we can see the distribution of the data. This diagram can only be used for an overall picture. That is why the name of the countries is not shown on the scatter plot.

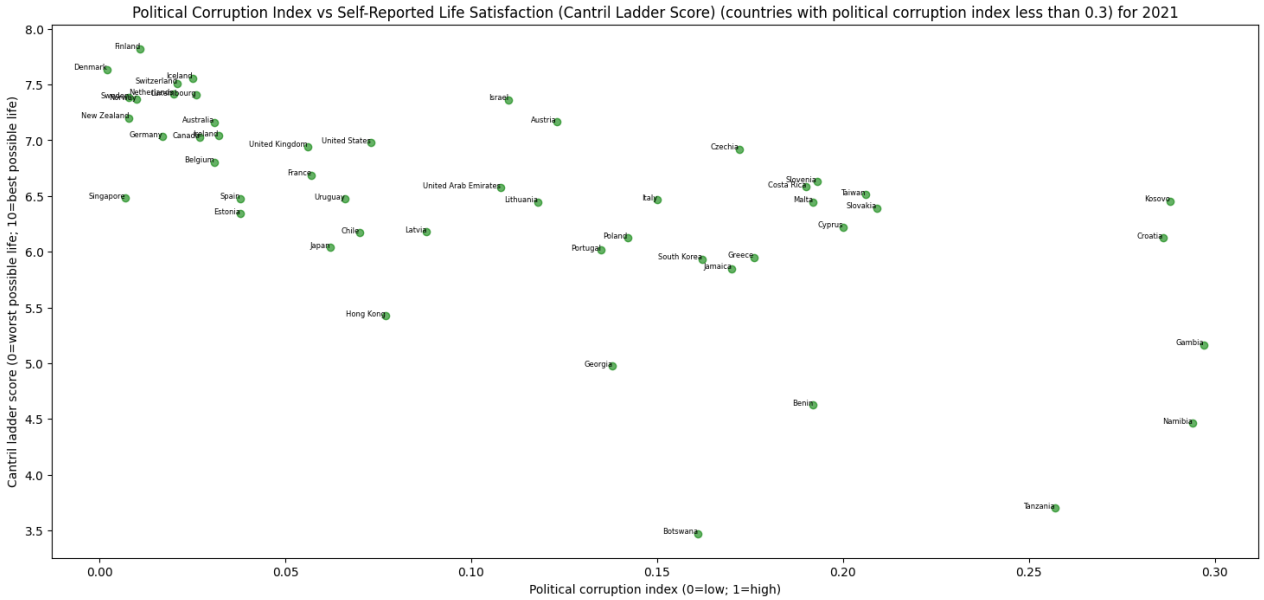


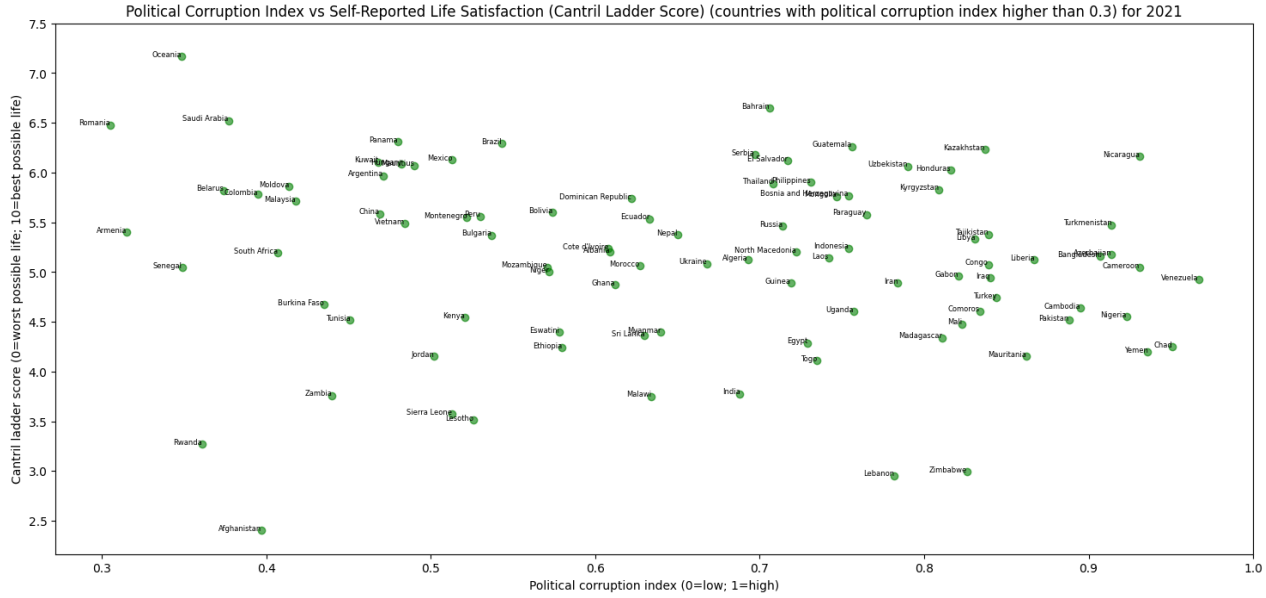
The most detailed scatter plot is the second one for which I've used ***Matplotlib***. For that, some considerations were taken into account:

- the scatter plot is much larger as the number of entries is 145 (most of them settled in one area)
- the name of each country can be found next to its circle – for better clarity it would be good if it is visualised only on hoovering over it; however this cannot be achieved with ***Matplotlib***



However large the scatter plot is, so many names of countries cannot be presented on just one diagram. That's why I've taken the decision to divide this scatter plot into two separate ones. The criteria used is 'Political corruption index' – less than 0.3% in the first scatter plot, above 0.3% in the second scatter plot. In that way, the names of the countries become clearer.

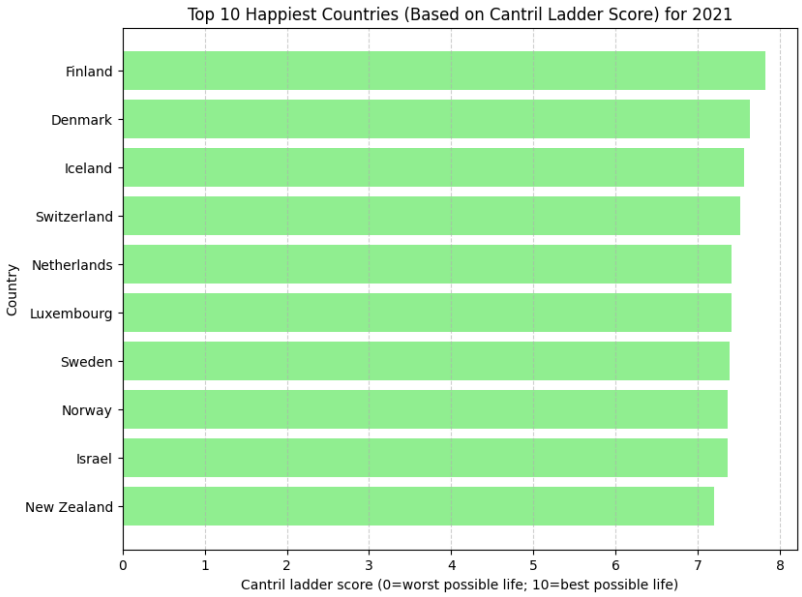




Questions to be answered with the help of the scatter plot:

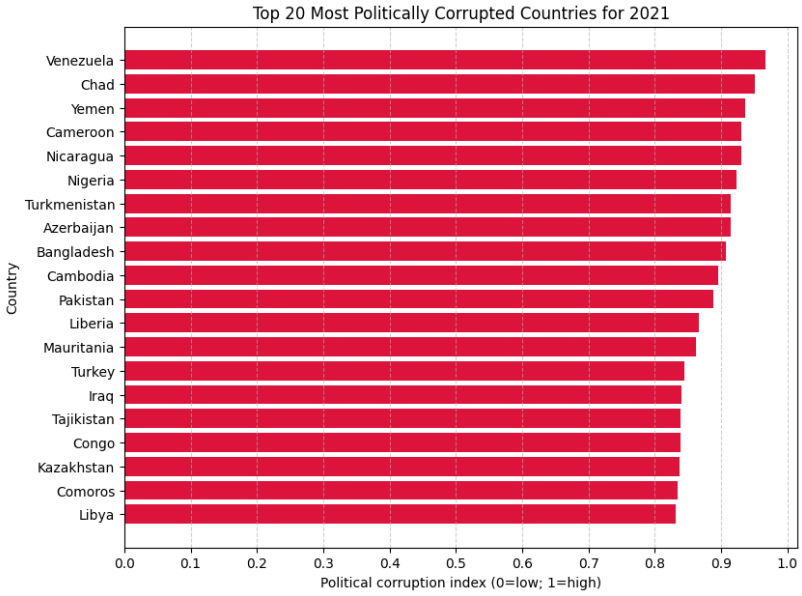
1. Which are the top 10 happiest world countries (using the Cantril ladder score)?

With ***Pandas*** and its method *sort_values* performed on the merged data frame, I sorted the data frame on the 'Cantril ladder score' column in descending order and took the first 10 rows with the *head* method. The results I visualised using horizontal bar chart with the help of ***Matplotlib***:



2. Which are the top 20 most politically corrupted world countries?

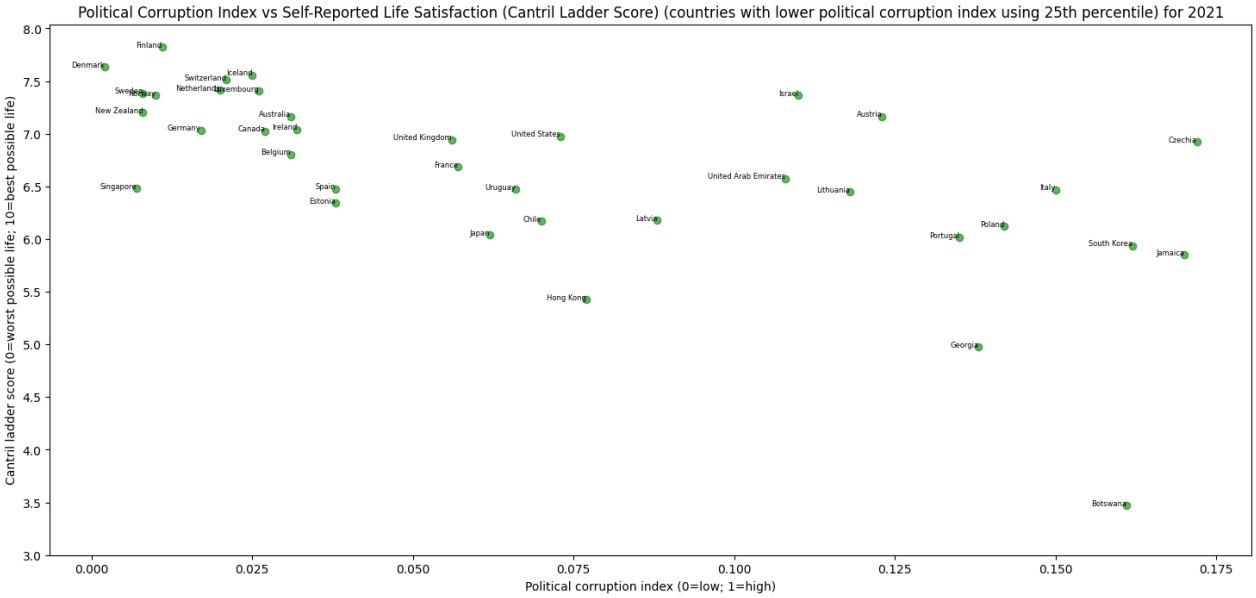
With ***Pandas*** and its method *sort_values* performed on the merged data frame, I sorted the data frame on the 'Political corruption index' column in descending order and took the first 20 rows with the *head* method. The results I visualised using horizontal bar chart with the help of ***Matplotlib***:



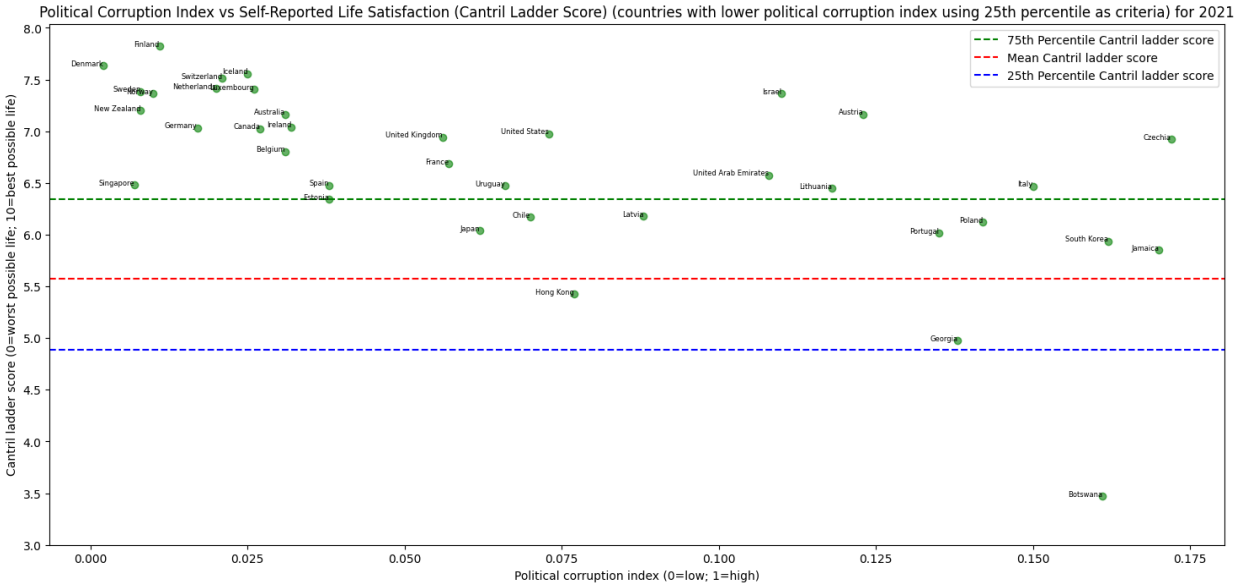
3. Find whether each country with low political corruption index have high Cantril ladder score?

For defining low political corruption index, I use the 25th percentile of the column 'Political corruption index'. This will be my upper boundary for the searched countries. The exact value is 0.172. With that consideration the searched number of countries is 37 and they are:

Australia, Austria, Belgium, Botswana, Canada, Chile, Czechia, Denmark, Estonia, Finland, France, Georgia, Germany, Hong Kong, Iceland, Ireland, Israel, Italy, Jamaica, Japan, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Singapore, South Korea, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States, Uruguay



For defining higher Cantril ladder score (better possible life), I use the 75th percentile of the column 'Cantril ladder score'. This will be my lower boundary for high life satisfaction. In addition to that, I find the mean Cantril ladder score and the 25th percentile representing the upper boundary for low life satisfaction. I will provide the answer to the question by plotting and observing the following scatter plot:



We can see that there are 3 countries below the mean Cantril ladder score that are with low index of political corruption. Most of the corrupted-free countries are happy which strengthens the trend. Botswana is the only African country that is with low level of political corruption but is still with low level of life satisfaction. That must be related to the poor quality of life of the local population.

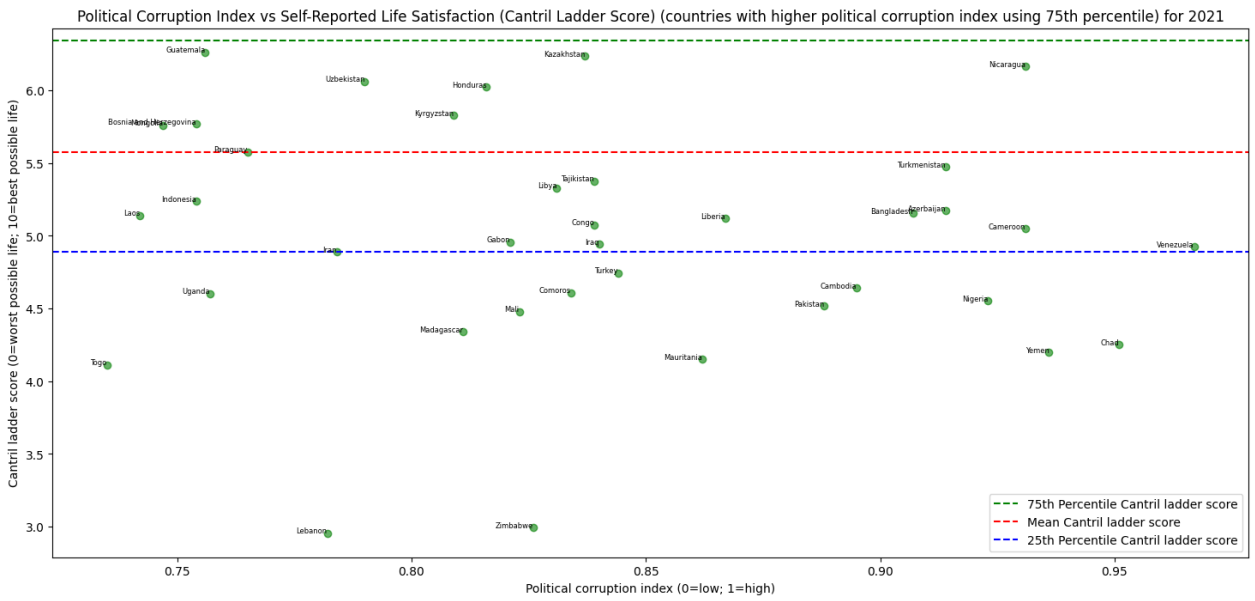
Generally, we can conclude that the lower the political corruption is, the higher level of life satisfaction a country will have (most countries on the above scatter plot prove that).

4. Check if there are highly politically corrupted countries that have high Cantril ladder score?

For defining high political corruption index, I use the 75th percentile of the column 'Political corruption index'. This will be my lower boundary for the searched countries. The exact value is 0.735. With that consideration the searched number of countries is 37 and they are:

Azerbaijan, Bangladesh, Bosnia and Herzegovina, Cambodia, Cameroon, Chad, Comoros, Congo, Gabon, Guatemala, Honduras, Indonesia, Iran, Iraq, Kazakhstan, Kyrgyzstan, Laos, Lebanon, Liberia, Libya, Madagascar, Mali, Mauritania, Mongolia, Nicaragua, Nigeria, Pakistan, Paraguay, Tajikistan, Togo, Turkey, Turkmenistan, Uganda, Uzbekistan, Venezuela, Yemen, Zimbabwe

For defining higher Cantril ladder score (better possible life), I use the 75th percentile of the column 'Cantril ladder score'. This will be my lower boundary for high life satisfaction. In addition to that, I find the mean Cantril ladder score and the 25th percentile representing the upper boundary for low life satisfaction. I will provide the answer to the question by plotting and observing the following scatter plot:

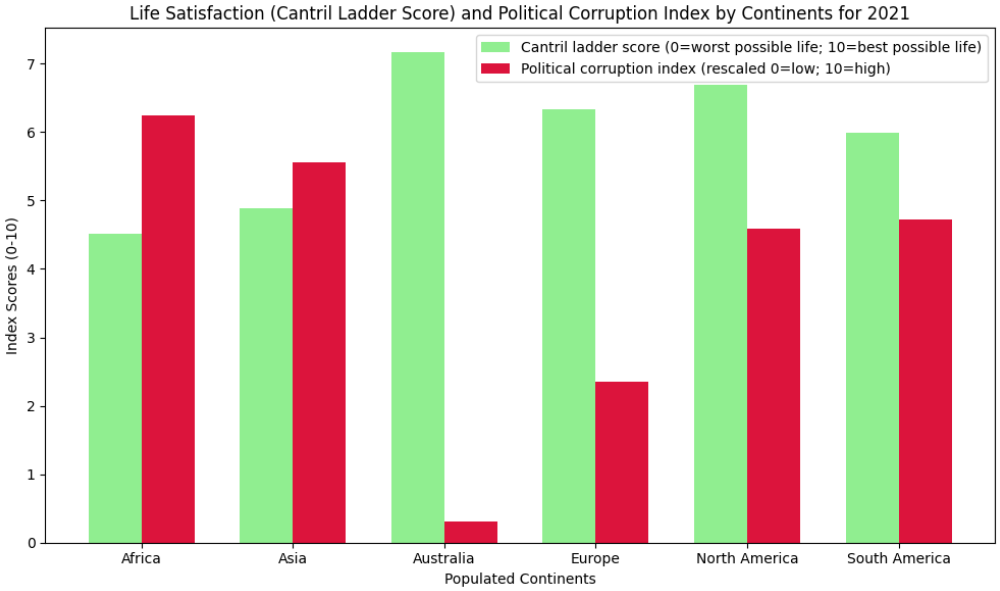


We can see that the plenty of the most politically corrupted countries tend to be unhappy (with low level of life satisfaction). There are still 9 countries which are above the mean Cantril ladder score although being highly corrupted. Most of these countries are poor countries and as such it is possible that they can be easily manipulated to think they are happy although living in a corrupted country. As a conclusion, we have to fight for eradicating political corruption if we want to live in a better, fairer world.

III. A grouped bar plot of Life Satisfaction (Cantril Ladder Score) and Political Corruption Index by Continents for 2021

We have already extracted the necessary data for the continents. As we have to compare two indices for the same entities (the six populated continents) I decided to use a grouped bar plot. For those purposes, it is important that the scale is the same for the two features – Cantril ladder score and political corruption index. For that reason, I transformed the political corruption index from

boundaries 0-1 to 0-10 by multiplying the values for all the continents by 10. In this way, I can ensure equal index scale.



From the grouped bar plot we can see that Australia is the continent with highest Cantril ladder score, Africa – with the lowest one. Africa is the continent with highest index of political corruption whereas Australia is the continent with the lowest. Therefore, Australia seems to be the winner among the other continents. Europe is the second-best continents for living – with high Cantril ladder score and comparatively low level of political corruption.

Conclusion

In this project, we integrated and analysed data from three different **CSV datasets**: life expectancy and extreme poverty indicators, the Cantril Ladder (subjective life satisfaction), and the Political Corruption Index for the year 2021. The data were uploaded into the **Hadoop Sandbox** environment using the **Ambari** interface and processed with **PySpark** in **Apache Zeppelin**. We inferred the data schemas programmatically and performed visual exploration by creating scatter plots using **Matplotlib**, highlighting relationships between poverty levels, life expectancy, subjective well-being, and perceived corruption across world countries.

Through this work, we demonstrated the ability to perform distributed data analysis, schema extraction, and visual data storytelling within a big data environment, emphasizing the importance of socioeconomic indicators for understanding global well-being.