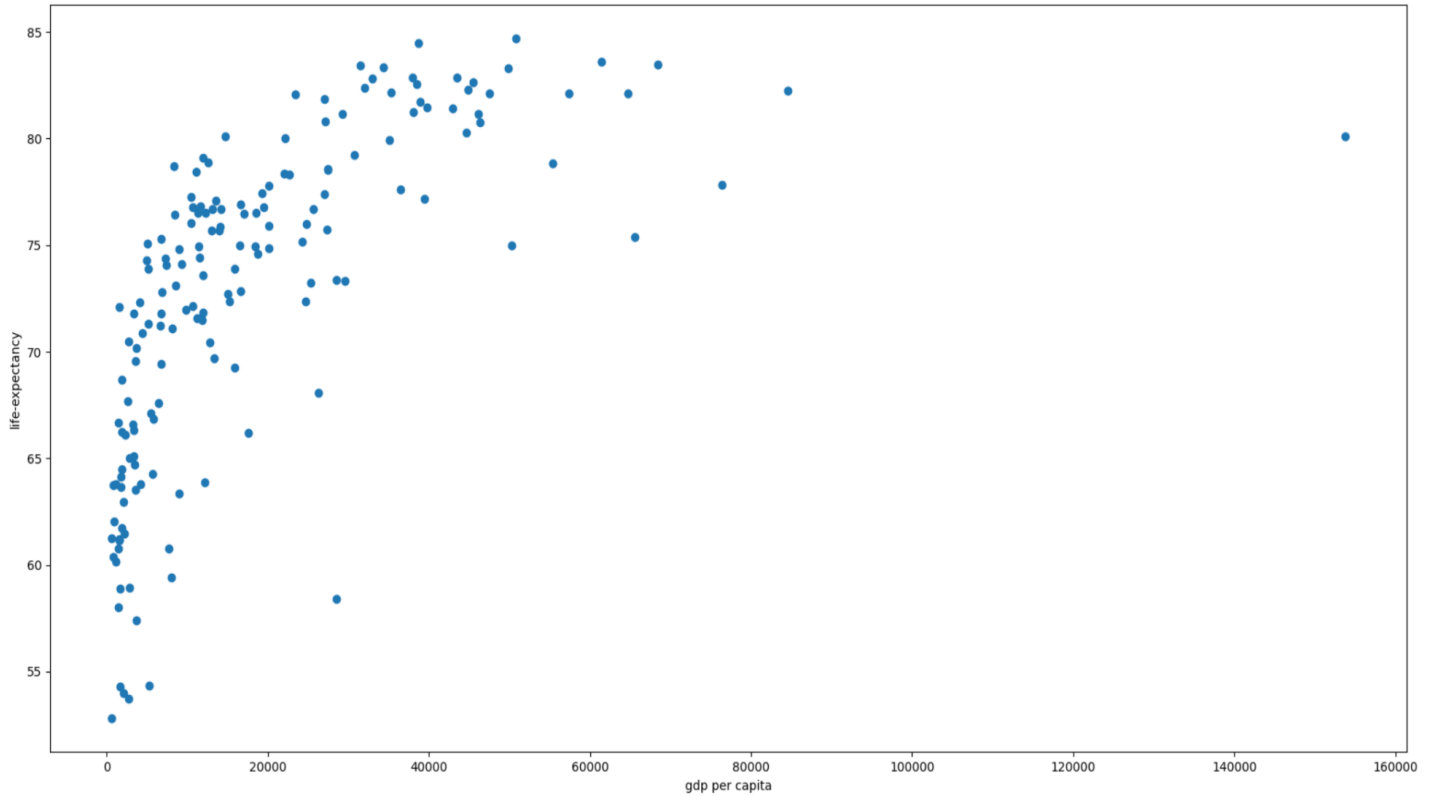


Lab 1 World In Data

1. Below you can find our scatterplot with life expectancy on the y-axis and GDP per capita on the x-axis.



By:
Group: 59
Johan Östling
Malte Carlstedt

1 a). We made the assumption that the most recent data was the most valuable data in this case. Since we were most interested in the present, this data best shows how our world looks today. We can mention that if one wanted to reassure that there is a correlation between life expectancy and GDP per capita it could be interesting to look at all years. In our case we sorted out all data that we downloaded from the "Our World In Data" website that are not from the year 2018. Even though Our World in Data had data from the year 2019 in regards to life expectancy. GDP per capita did not. Thus we chose the data that was most recent in both categories. The two data sets did not have the same number of countries, therefore we had to combine the data only where the same country existed in both sets.

1 b). The plot seems reasonable since the countries with low GDP per capita have for the most part lower life expectancy than countries with high GDP per capita. It seems reasonable because richer countries can afford to improve their public sector while poorer countries do not have the same possibility. We can also see from the plot that the life expectancy for countries with low GDP per capita can differ and the same with countries with higher GDP per capita. This seems reasonable because there are certainly more factors that predict life expectancy than GDP per capita such as war and state.

When it comes to the pattern of the scatter plot it also seems reasonable that it looks like a logarithmic function, and not a linear. This is because GDP per capita can grow fast and there is not any limit for how high GDP per capita a country can have. On the other hand life expectancy has a limit, a human can not live forever. And the more years a person has lived, the likelihood to live another decreases. So a country can not increase its life expectancy above a certain number, for example 150 years, no matter how much they increase their GDP per capita.

1 c). We did not remove any entries other than the data that was not from the year 2018 and countries that did not exist in both the GDP per capita data and in the life expectancy data. We also removed the column "Code" that had the abbreviation for each country. We removed this since we didn't need this information when creating our plots. There was no need to fill in missing values because there weren't any, nor was there any misspellings in the data so this wasn't edited either.

1 d). We calculated that the mean life expectancy was: 72.29468263473056 and the standard deviation: 7.654279505385729. We used our function "CountriesAboveStandard" to conclude that the following 34 countries have at least one standard deviation above mean: 'Australia', 'Austria', 'Belgium', 'Canada', 'Chile', 'Costa Rica', 'Cyprus', 'Denmark', 'Finland', 'France', 'Germany', 'Greece', 'Hong Kong', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Japan', 'Luxembourg', 'Malta', 'Netherlands', 'New Zealand', 'Norway', 'Portugal', 'Puerto Rico', 'Qatar', 'Singapore', 'Slovenia', 'South Korea', 'Spain', 'Sweden', 'Switzerland', 'Taiwan', 'United Kingdom'.

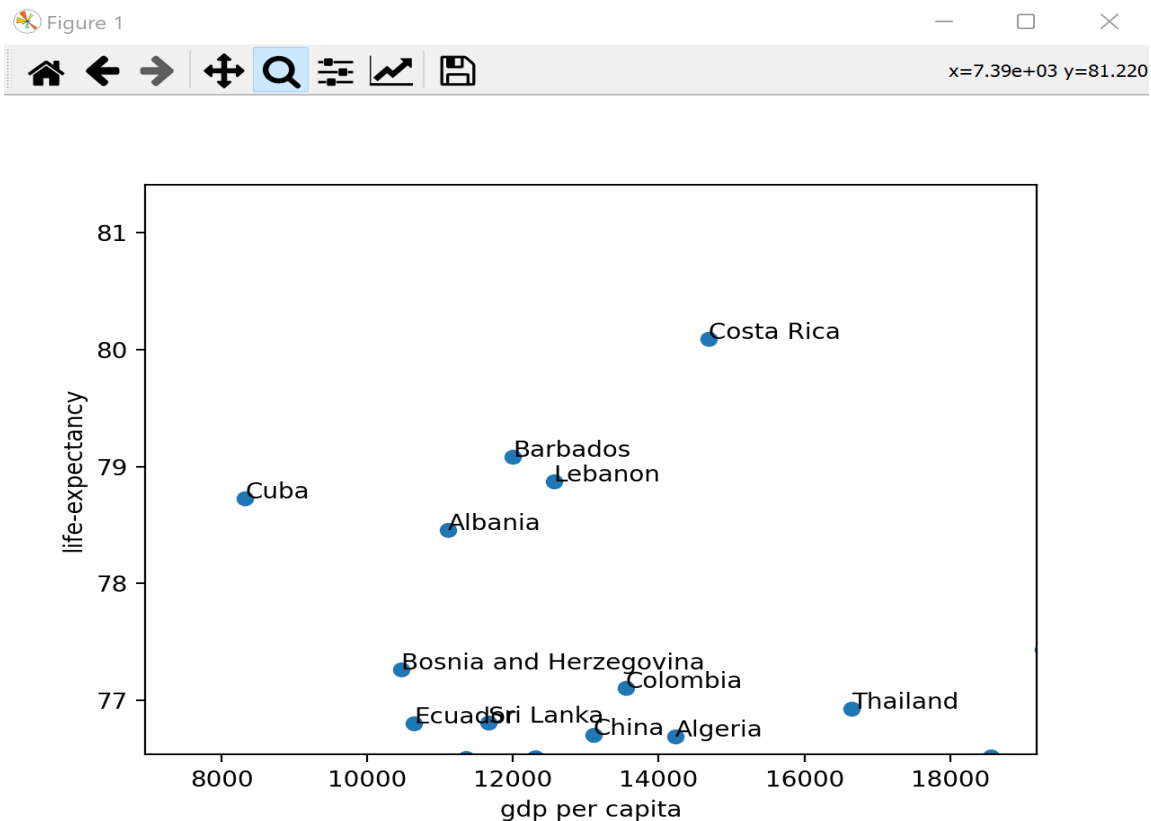
1 e). We chose to define high life expectancy as **the mean plus one standard deviation**. The countries with high life expectancy are shown in answer 1d. For defining what low gdp is we thought that the mean of GDP per capita of all countries minus one standard deviation was appropriate. But this was not the case since the standard deviation is greater than the mean in some cases. This is a possibility if the value of GDP per capita has a big spread. So instead we defined low GDP per capita as **being lower than the mean**. The following countries has GDP per capita lower than the mean:

['Afghanistan', 'Albania', 'Algeria', 'Angola', 'Argentina', 'Armenia', 'Azerbaijan', 'Bangladesh', 'Barbados', 'Belarus', 'Benin', 'Bolivia', 'Bosnia and Herzegovina', 'Botswana', 'Brazil', 'Bulgaria', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon', 'Cape Verde', 'Central African Republic', 'Chad', 'China', 'Colombia', 'Comoros', 'Congo', 'Costa Rica', 'Cote d'Ivoire', 'Cuba', 'Democratic Republic of Congo', 'Djibouti', 'Dominica', 'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Eswatini', 'Ethiopia', 'Gabon', 'Gambia', 'Georgia', 'Ghana', 'Guatemala', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Honduras', 'India', 'Indonesia', 'Iran', 'Iraq', 'Jamaica', 'Jordan', 'Kenya', 'Kyrgyzstan', 'Laos', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mexico', 'Moldova', 'Mongolia', 'Morocco', 'Mozambique', 'Myanmar', 'Namibia', 'Nepal', 'Nicaragua', 'Niger', 'Nigeria', 'North Korea', 'North Macedonia', 'Pakistan', 'Palestine', 'Paraguay', 'Peru', 'Philippines', 'Rwanda', 'Saint Lucia', 'Sao Tome and Principe', 'Senegal', 'Serbia', 'Sierra Leone', 'South Africa', 'Sri Lanka', 'Sudan', 'Syria', 'Tajikistan', 'Tanzania', 'Thailand', 'Togo', 'Tunisia', 'Uganda', 'Ukraine', 'Uzbekistan', 'Venezuela', 'Vietnam', 'World', 'Yemen', 'Zambia', 'Zimbabwe']

To then find out which countries have high life expectancy but low GDP per capita we need to take the intersection between the list of countries with high life expectancy and the list of countries with low GDP per capita. The result was the following countries:

There was only one country which was in both lists and it was **Costa Rica**.

Now if we choose a less structured method we can look at the scatterplot and by ocular inspection determine which countries have low GDP per capita but a high life expectancy We chose Cuba, Albania, Barbados, Lebanon and Costa Rica. See picture below.



1 f). To find out if all countries with a high GDP per capita have a high life expectancy we try to find at least one country that has high GDP per capita but low life expectancy. If we found a country like this then the statement given in question 1f is therefore false. If we don't find any, then we can conclude that every country with a high GDP per capita also has a high life expectancy. We can also mention that high GDP per capita does not necessarily mean a strong economy. But since GDP per capita is the only variable we are plotting we assume for this assignment that it is.

We define *high GDP per capita* as countries that are above the mean, and *low life expectancy* as the mean minus one standard deviation. We got the follow list for countries with high GDP per capita:

['Australia', 'Austria', 'Bahrain', 'Belgium', 'Canada', 'Chile', 'Croatia', 'Cyprus', 'Czechia', 'Denmark', 'Equatorial Guinea', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hong Kong', 'Hungary', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Japan', 'Kazakhstan', 'Kuwait', 'Latvia', 'Lithuania', 'Luxembourg', 'Malaysia', 'Malta', 'Mauritius', 'Montenegro', 'Netherlands', 'New Zealand', 'Norway', 'Oman', 'Panama', 'Poland', 'Portugal', 'Puerto Rico', 'Qatar', 'Romania', 'Russia', 'Saudi Arabia', 'Seychelles', 'Singapore', 'Slovakia', 'Slovenia', 'South Korea', 'Spain', 'Sweden', 'Switzerland', 'Taiwan', 'Trinidad and Tobago', 'Turkey', 'Turkmenistan', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay']

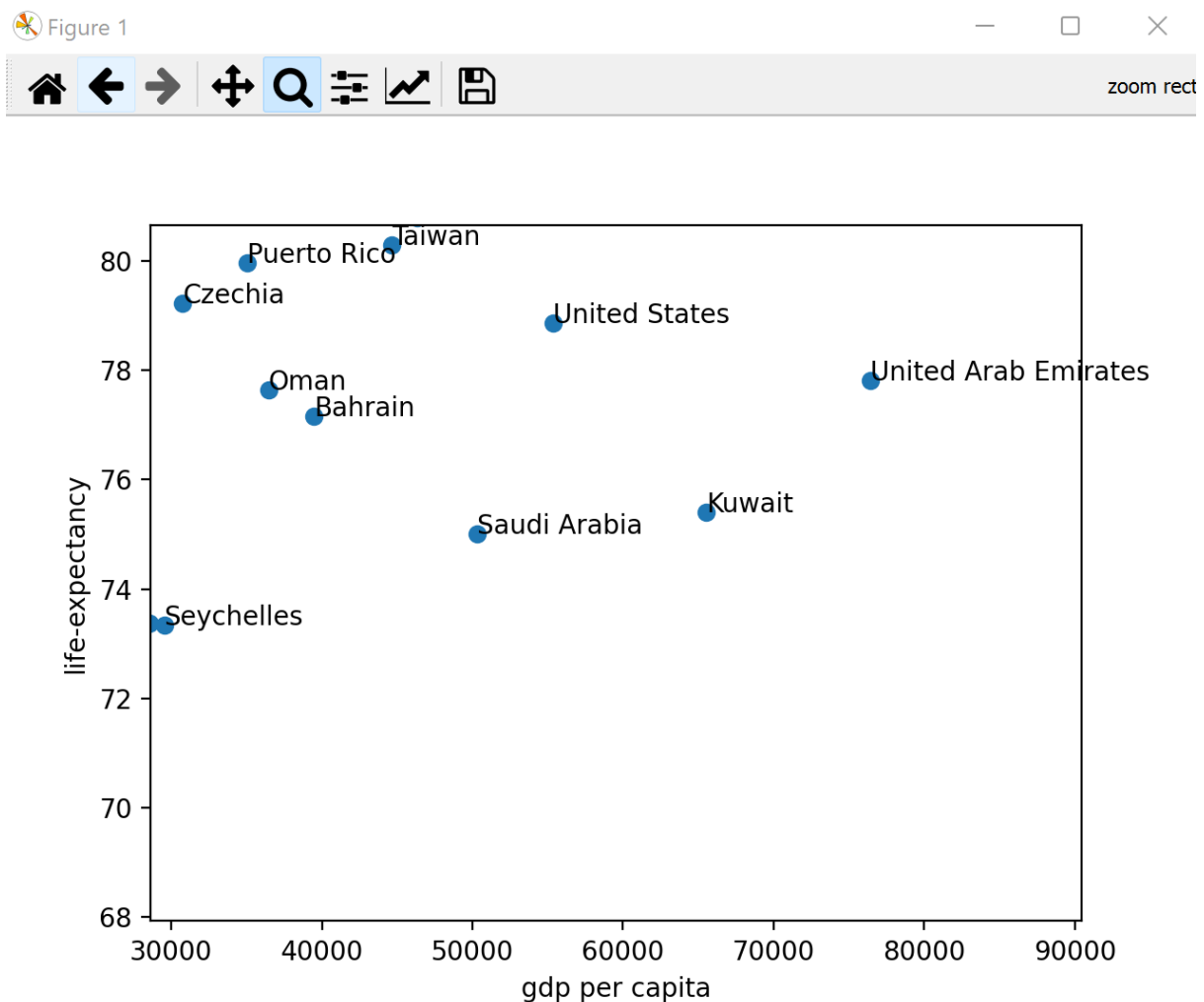
And the following list for countries with low life expectancy:

['Afghanistan', 'Angola', 'Benin', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo', 'Cote d'Ivoire', 'Democratic Republic of Congo',

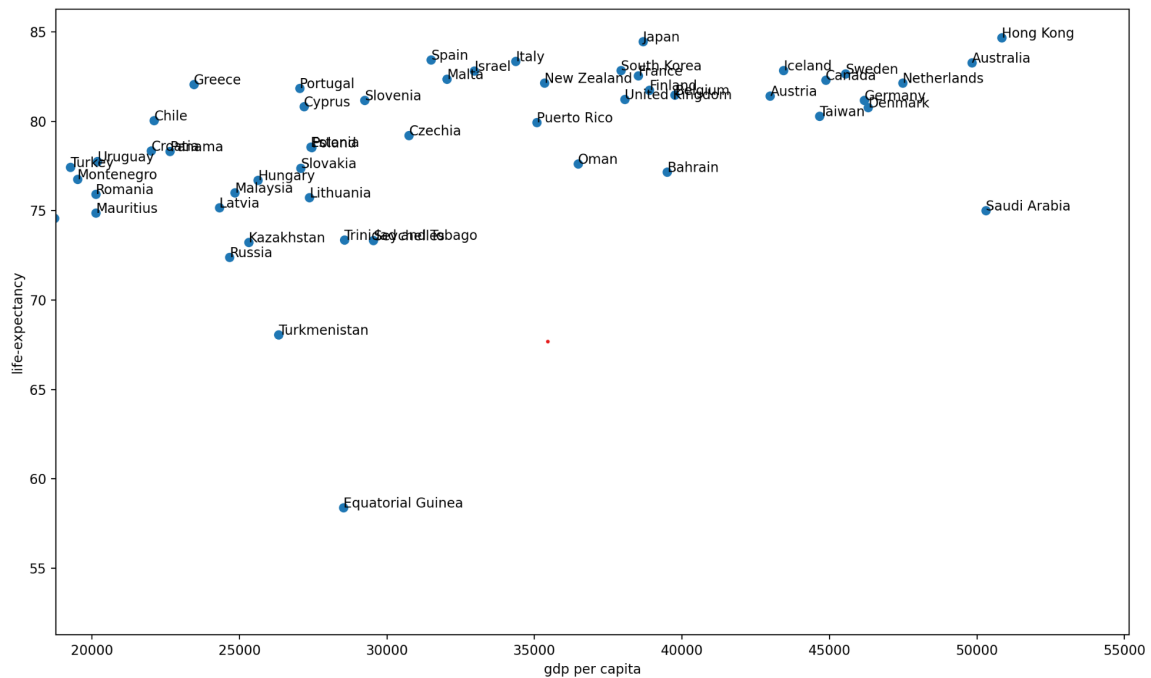
'Equatorial Guinea', 'Eswatini', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Lesotho', 'Liberia', 'Malawi', 'Mali', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Sierra Leone', 'South Africa', 'Togo', 'Uganda', 'Zambia', 'Zimbabwe']

The intersection between the two lists shows the country **Equatorial Guinea**. From this we conclude that not every strong economy has a high life expectancy. Since we could find a country that disapproves of the original statement.

If we instead use the method to inspect the scatter plot we can see that the countries like **Saudi Arabia** and **Kuwait** have low life expectancy compared with countries with equal GDP per capita. From this we could also get the same approximate conclusion as above.



1 g). If we assumed that a strong economy was represented by GDP per capita then we would only get a clear insight of how much money that the country has. It however does not show where and how that money is spent. This is a great example for e.g. Equatorial Guinea where their GDP per capita is relatively high but their life expectancy is very low. Other countries with the approximate same gdp have much higher life expectancy. As shown in the picture below.



Countries like Slovenia and Spain have ~20 year higher life expectancy but around the same GDP per capita as Equatorial Guinea. This is the danger by using only GDP per capita to measure if a country has a strong economy since it doesn't measure where the money is being used in the country. Neither does it show how much of the money is distributed over its citizens and how much of it is used for e.g. corruption among politicians. Neither does it show the rift between poor and rich in a country. Some may be very poor and some may be very rich.

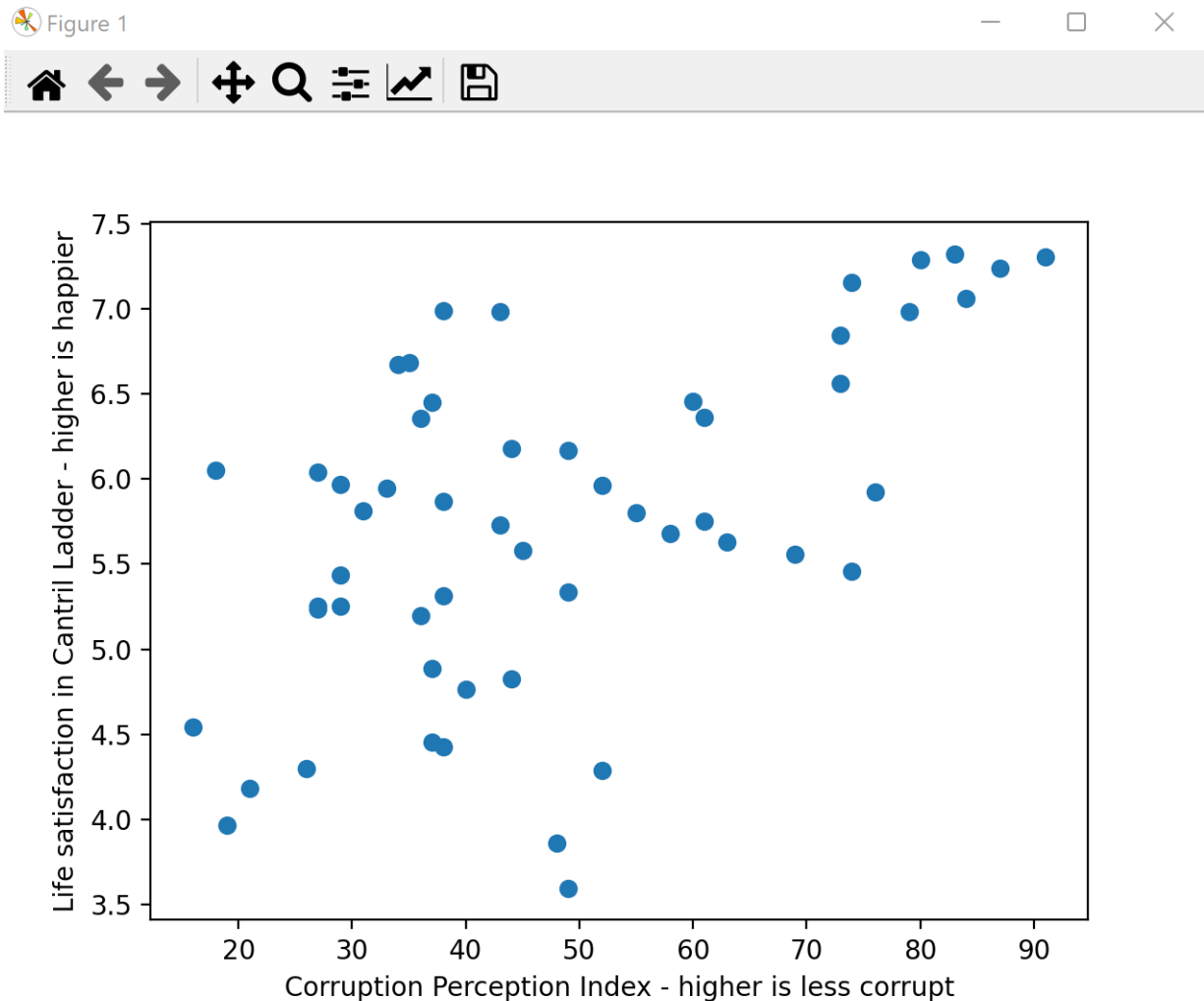
2. a) We chose to visualize the following data sets: Corruption, Happiness and Life Satisfaction, and trust.

Our meaningful questions are:

1. Does higher levels of corruption in a country affect people's perceived level of happiness?
2. Does higher levels of corruption mean that people don't trust each other?
3. Is there a correlation between happiness and trust?
4. Can there be a connection between corruption, happiness and life satisfaction, and trust?

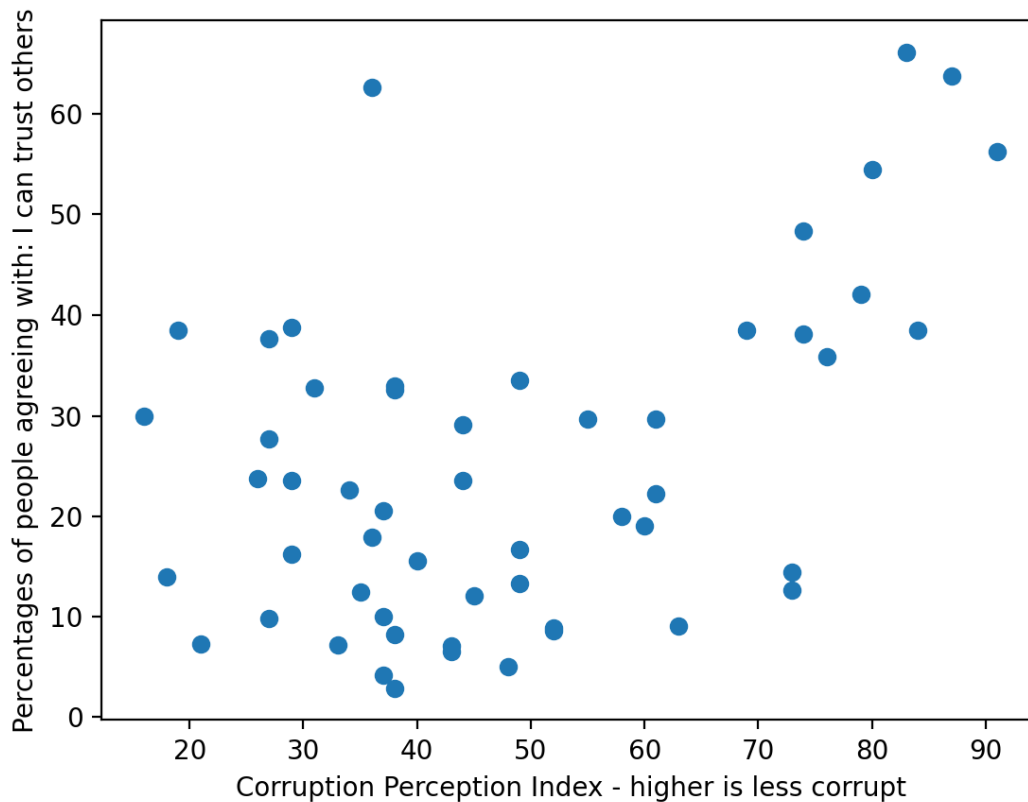
For the first question we found data about corruption from the dataset 'Corruption' that shows on a scale from 0-100 how corrupt a country is perceived. Where 100 is not corrupt and 0 is very corrupt. The number of corruption is generated by standardized surveys, in which only experts on the subject were allowed to take part. For the happiness level we used the dataset 'Happiness and life satisfaction' in which they got data by asking people to write a number from 0-10 on how happy they are. 0 represents the worst possible life for you and 10 represent the best possible life for you. We choose to plot for each country the corruption level and happiness in a scatter plot, where each point represents a country. From this we can see if, in general, countries with low corruption also are happy, or the other way around.

We chose to visualize the data in scatter plots since it is a very good way to see if two variables are correlated. By inspecting the points in a scatter plot you can see if there is any pattern or curve that fits. And by that way see if there are any connections between the two variables. Because of these reasons we only used scatter plots to visualize the information for the “the meaningful questions” that we listed earlier. We choose to use the most recent data that all data sets have in common, this was only true for the year 2014. We also had to filter out the countries that were not present in both data sets. We got the following plot:



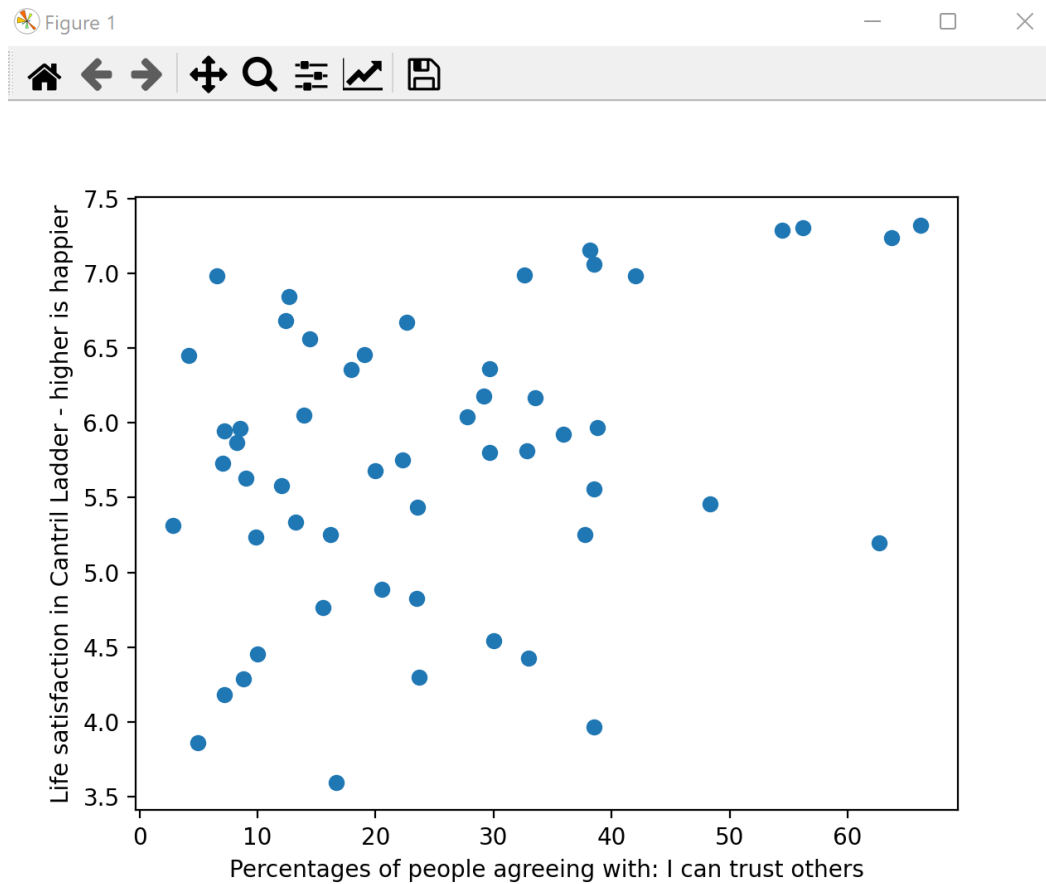
For the second question we compared corruption against trust. The general idea behind this question was to see if corruption in a country created disbelief among the citizens of the country. The trust dataset is based on how many percent of the population in a country agree with the statement “I can trust others”. We got the following scatter plot:

Figure 1



From this scatter plot we concluded that in general people with less corruption in their respective country tend to trust other people more. However, as can be seen in the plot there is a big cluster in the bottom left that we believe has to do with where and the type of state the country has. After lab 2 it would be interesting to see how we could define these countries within the cluster and then try to find a correlation regarding where in the world the country is and what type of state it has. E.g. democracy or dictatorship.

For the third question we did as before and got the following scatter plot:



This plot was the most difficult to read and explain. The plot is very scattered and no direct curve or connection can be seen instantly. Our general conclusion is that trust and happiness don't have a clear connection.

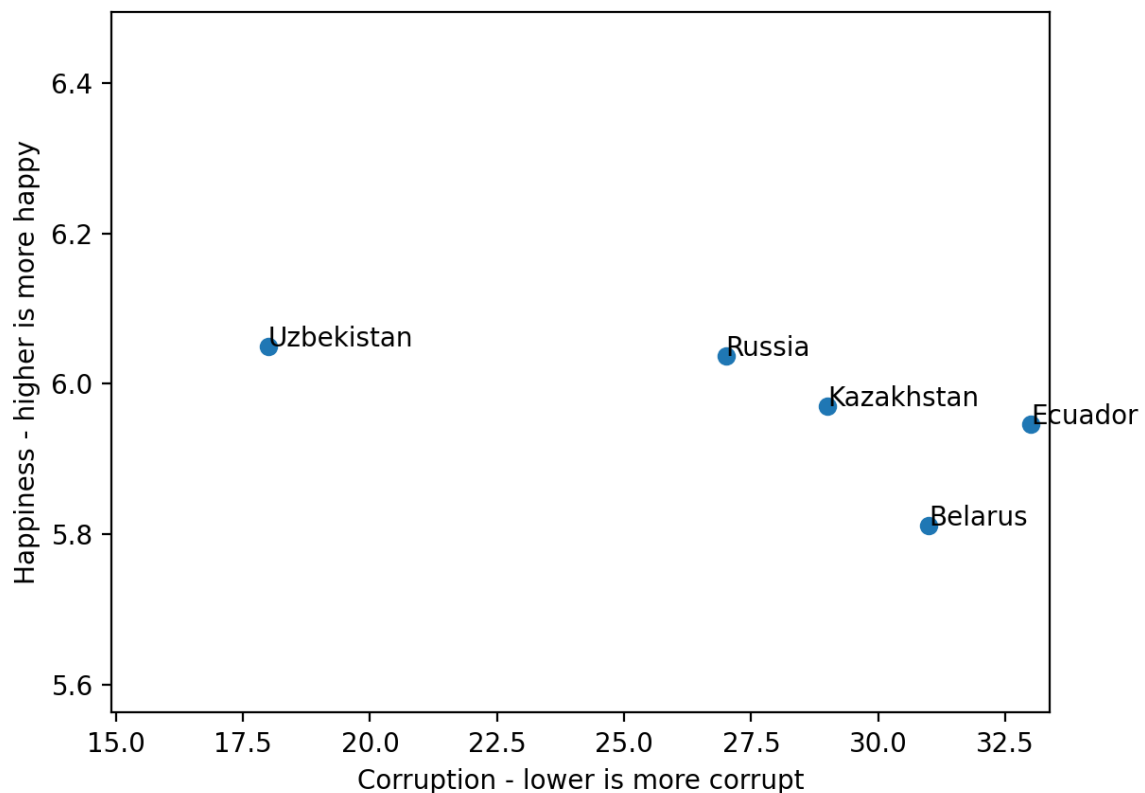
The fourth question can be answered by analyzing all three scatter plots made for answering the three first questions.

2. b) From the data collected we came to the following conclusions:

- People who are happy do not necessarily trust others more.
- People living in a corrupt country tend to be distrusting of others.
- People living in a corrupt country are less happy.

The third point however is controversial since some of the people who are living in a corrupt country may be under the influence of their leaders to tell the world that they are happy when they are not. A good example of this is Uzbekistan which you can see in the picture below,

has very high corruption in the country but still a quite high happiness level. It might be that the population actually is happy but looking at the tendencies of the other countries where high corruption leads to lower levels of happiness this is probably not true. Our conclusion is therefore that many of the countries with high corruption and high level of happiness are also under high influence from their leaders and therefore not trustworthy.



Another insight we got from these data sets is that the vast majority of countries have high corruption. In this day and age you could have hoped for that the world had come further along with democracy etc but as can be seen in our scatterplots, the majority of data points are to the further left on the x-axis. Which means that most countries in the world are still in fact corrupt.