```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Assignment 1: Introduction to Data Science and Python

DAT405 Introduction to Data Science and AI

By Pauline Nässlander and Albin Ekström

Hours spent on the assignment:

- Pauline Nässlander: 8.5 hours
- Albin Ekström: 8.5 hours

1a)

The data collected from https://ourworldindata.org on life expectancy and GDP per capita included data from several years. Since we only wanted to have one data point per country, one year in common and the most recent data, we chose to use the data from 2019.

We made sure that the countries plotted had data from both life expectancy and GDP per capita. This was made by first creating a dictionary with countries as keys and life expectancy as values for the year 2019. When we later iterated through the GDP per capita file we made sure that the country was in the dictionary and that the data was from 2019, see code below.

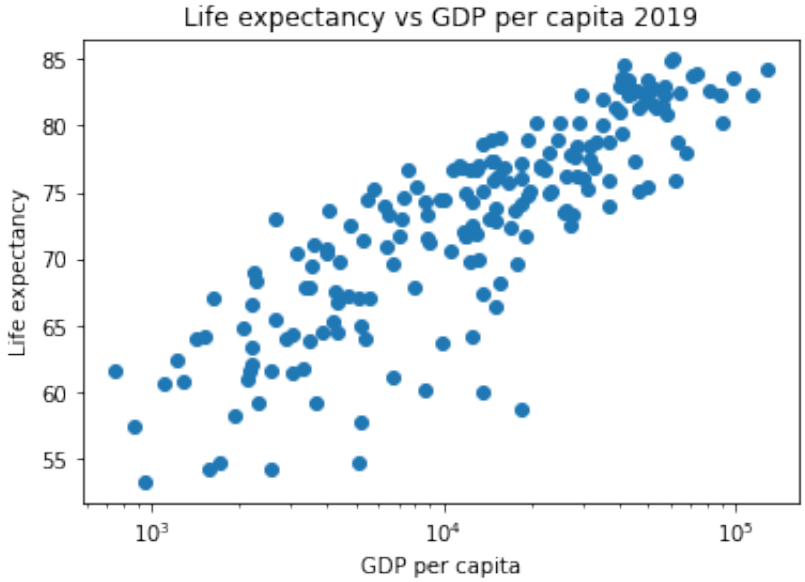
1b)

From the graph we obtained (see graph "*Life expectancy vs GDP per capita 2019*") we can draw the conclusion that life expectancy increases linearly with increasing GDP per capita. Note how the x-axis increases logarithmic.

The conclusion above is not unreasonable as increased GDP per capita implies that the country and its citizens have a higher life standard which leads to longer life in general. We refer to Hans Rosling's comment on the life expectancy vs GDP per capita graph; 'this is because basic modernities have reached most people and drastically improved their lives. They have plastic bags to defend and transport food. They have plastic buckets to carry water and soap to get rid of bacteria. Most of their children are vaccinated.' (Rosling 2018, p.76). When the inhabitants' economy has reached a certain level where they can afford all the necessities of life, such as health care etc. Increased economy does not give any major increase in life expectancy, hence the logarithmic appearance of the graph.

code for assignment 1a and 1b:

```
import matplotlib.pyplot as plt
import pandas as pd


life = {}
df = pd.DataFrame(columns=['Country', 'Age', 'GDP'])

life_exp = pd.read_csv(r'/content/drive/MyDrive/DAT405/Ass_1/life-expectancy.csv
gdp_capita = pd.read_csv('/content/drive/MyDrive/DAT405/Ass_1/gdp-per-capita-wor

for index, row in life_exp.iterrows():
  if row["Year"] == 2019:
    life[row["Entity"]] = row["Life expectancy"] # Create a dict with the specif

for index, row in gdp_capita.iterrows():
  if row["Entity"] in life.keys() and row["Year"] == 2019 and not row["Entity"]
    df = df.append({'Country': row["Entity"], 'Age': life.get(row["Entity"]), 'G

x_values = df['GDP']
y_values = df['Age']
plt.scatter(x_values, y_values)
ax = plt.gca()
ax.set_xscale('log')
plt.xlabel('GDP per capita')
plt.ylabel('Life expectancy')
plt.title('Life expectancy vs GDP per capita 2019')
plt.show()
```



1c)

Firstly when executing assignment a and b we only picked out the data from 2019 since we thought that data from earlier years was not relevant and we also assessed that there was not enough data from later years than 2019. In addition to this, we only wanted data from different countries, not regions built up of several countries since we did not think regions would supply relevant information. After skimming the data, we saw that the “world” was included in both the GDP per capita file and the life expectancy file and therefore we had to manually erase it by including the line “and not row["Entity"] == "World"” in the if statement before adding entities to the dataframe.

Secondly, when we did assignment f we imported new data of GDP per country. This file did not include enough data after the year of 2018 and therefore we chose to use the data from 2018 in this assignment. Also in this task we had to manually erase the entity “World”.

1d)

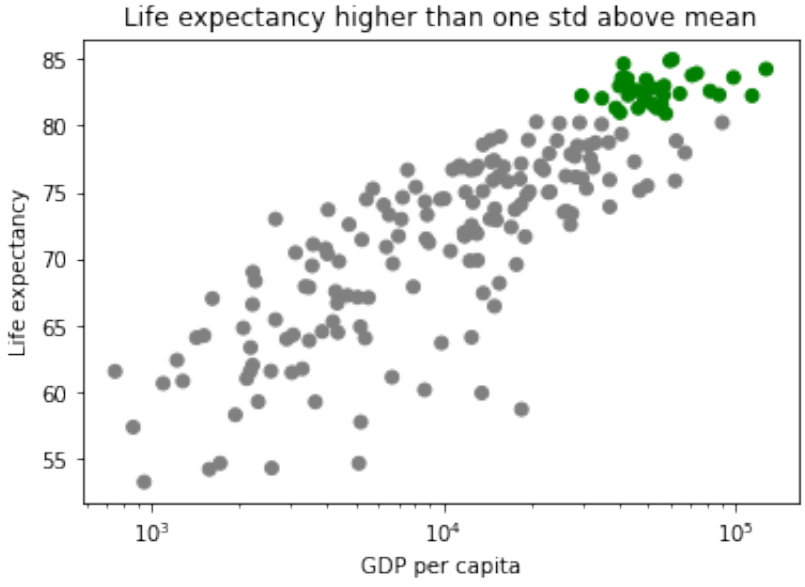
The countries that have a life expectancy one standard deviation above the mean is listed below, keeping in mind that this only includes countries with available data on both life expectancy and GDP per capita from 2019.

code for assignment 1d:

```
import statistics as st

# Calculating one std and mean for life expectancy
y_mean = st.mean(y_values)
y_std = st.stdev(y_values)

countries = []
color = []

# Categorize countries if life expectancy higher than one std above mean
for index, rows in df.iterrows():
  if rows["Age"] > (y_mean + y_std): # If it is higher color is green
    countries.append(rows["Country"])
    color.append('green')
  else: # If it is less color is grey
    color.append('gray')
```

```
print("Countries with very high life expectancy:")
display(countries)

plt.scatter(x_values, y_values,c=color)
ax = plt.gca()
ax.set_xscale('log')
plt.xlabel('GDP per capita')
plt.ylabel('Life expectancy')
plt.title('Life expectancy higher than one std above mean')
plt.show()
```

```
Countries with very high life expectancy:
['Australia',
 'Austria',
 'Belgium',
 'Bermuda',
 'Canada',
 'Cayman Islands',
 'Cyprus',
 'Denmark',
 'Finland',
 'France',
 'Germany',
 'Greece',
 'Hong Kong',
 'Iceland',
 'Ireland',
 'Israel',
 'Italy',
 'Japan',
 'Luxembourg',
 'Macao',
 'Malta',
 'Netherlands',
 'New Zealand',
 'Norway',
 'Portugal',
 'San Marino',
 'Singapore',
 'Slovenia',
 'South Korea',
 'Spain',
 'Sweden',
 'Switzerland',
 'United Kingdom']
```



1e)

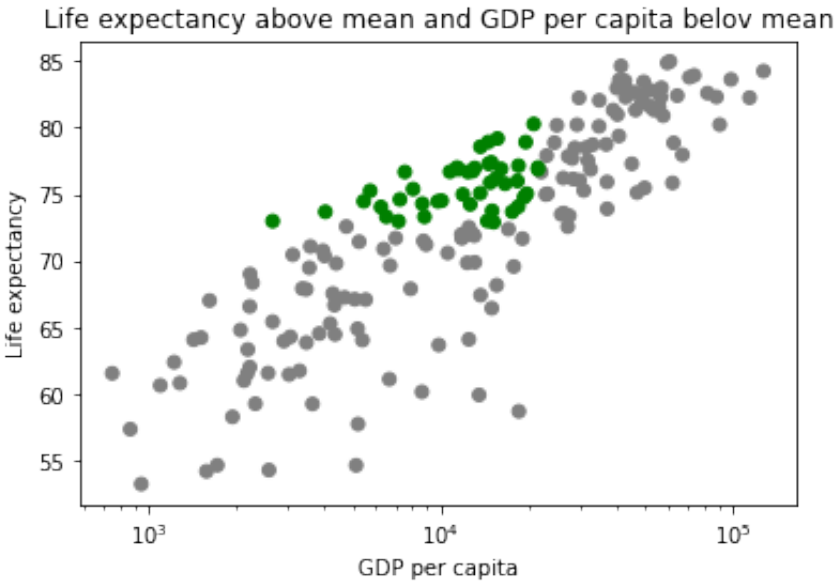
Assuming high life expectancy is above mean and low GDP per capita is belove mean we obtain the result below. One could argue that high life expectancy is one standard deviation above mean, but there's no country which has both life expectancy above mean plus one standard deviation and GDP per capita belove mean and therefore the previous assumption was made instead.

code for assignment 1e:

```
x_mean = st.mean(x_values)

countries = []
color = []

for index, rows in df.iterrows():
  if rows["Age"] > (y_mean) and rows["GDP"] < (x_mean):
    countries.append(rows["Country"])
    color.append('green')
  else:
    color.append('gray')

print("Countries with high life expectancy and low GDP per capita:")
display(countries)

plt.scatter(xValues, yValues,c=color)
ax = plt.gca()
ax.set_xscale('log')
plt.xlabel('GDP per capita')
plt.ylabel('Life expectancy')
plt.title('Life expectancy above mean and GDP per capita belov mean')
plt.show()
```

```
Countries with high life expectancy and low GDP per capita:
['Albania',
 'Algeria',
 'Antigua and Barbuda',
 'Armenia',
 'Azerbaijan',
 'Barbados',
 'Belarus',
 'Belize',
 'Bosnia and Herzegovina',
 'Brazil',
 'Cape Verde',
 'China',
```

```
'Colombia',
'Costa Rica',
'Dominica',
'Dominican Republic',
'Ecuador',
'El Salvador',
'Georgia',
'Guatemala',
'Honduras',
'Iran',
'Jamaica',
'Jordan',
'Lebanon',
'Libya',
'Maldives',
'Marshall Islands',
'Mexico',
'Montenegro',
'Morocco',
'Nicaragua',
'North Macedonia',
'Palau',
'Palestine',
'Paraguay',
'Peru',
'Saint Lucia',
'Samoa',
'Serbia',
'Solomon Islands',
'Sri Lanka',
'Thailand',
'Tunisia',
'Vietnam']
```



1f)

Assuming a strong economy is above mean GDP and high life expectancy is above mean.

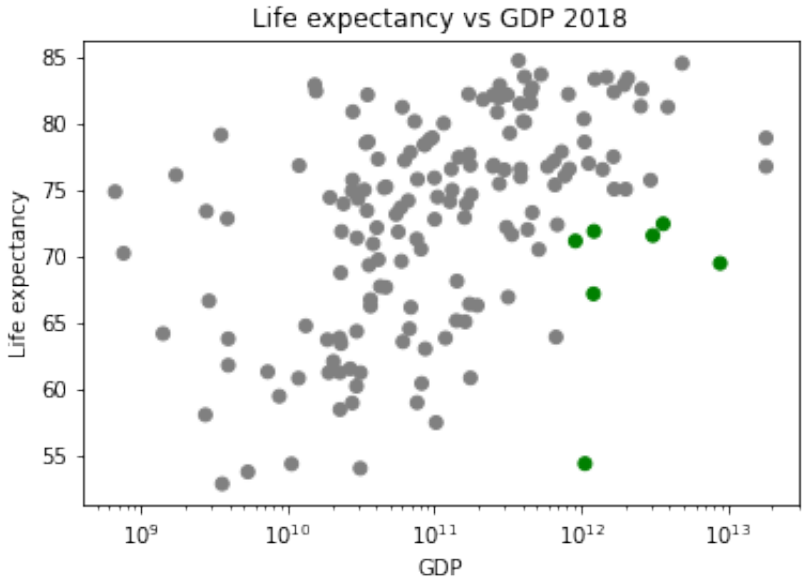
We can conclude that most countries with a strong economy (high GDP) have a high life expectancy, but not all (see list belove). Those countries have a GDP higher than the mean but a life expectancy belove mean.

code for assignment 1f:

```
countries_2018 = []
color_2018 = []

life_2018 = {}
df_2018 = pd.DataFrame(columns=['Country', 'Age', 'GDP'])

gdp = pd.read_csv('/content/drive/MyDrive/DAT405/Ass_1/gdp-world-regions-stacked

for index, row in life_exp.iterrows():
  if row["Year"] == 2018:
    life_2018[row["Entity"]] = row["Life expectancy"]

for index, row in gdp.iterrows():
  if row["Entity"] in life_2018.keys() and row["Year"] == 2018 and not row["Enti
    df_2018 = df_2018.append({'Country': row["Entity"], 'Age': life_2018.get(row

x_values_2018 = df_2018['GDP']
y_values_2018 = df_2018['Age']

x_mean_2018 = st.mean(x_values_2018)
y_mean_2018 = st.mean(y_values_2018)


for index, rows in df_2018.iterrows():
  if rows["Age"] < (y_mean_2018) and rows["GDP"] > (x_mean_2018):
    countries_2018.append(rows["Country"])
    color_2018.append('green')
  else:
    color_2018.append('gray')

print("Countries with life expectancy belove mean and GDP above mean:")
display(countries_2018)

plt.scatter(x_values_2018, y_values_2018, c=color_2018)
ax = plt.gca()
```

```
ax.set_xscale('log')
plt.xlabel('GDP')
plt.ylabel('Life expectancy')
plt.title('Life expectancy vs GDP 2018')
plt.show()
```

```
Countries with life expectancy belove mean and GDP above mean:
['Egypt', 'India', 'Indonesia', 'Nigeria', 'Pakistan', 'Philippines', 'Russ
```



1g)

First and foremost we do the same comparison as in f), but we now compare with GDP per capita instead of only GDP and we use the data from 2019. The difference in data points between 2019 and 2018 shouldn't be that big, hence it won't affect the result.

The result was that Russia is the only country which has a strong economy (in this case GDP per capita), but a low life expectancy.

From the results in assignment f and g we see that there are some strong economies in the world that have a low life expectancy, but the conclusions to be drawn from this differ if we compare the expected life length of the population with the GDP of the entire country or the GDP per capita. It is also important to take into account how populated the country is. If the country has a high population the GDP per capita can be low even if the GDP is very high this implies that even though the country has a lot of resources, the resources per citizen is low. On the other hand if the country has a low population, a strong economy (high GDP per capita) and low life expectancy indicates that it also has high segregation between social classes since the money is accumulated at the richest few in the country.

The conclusion we can draw from this is that it is important to know what parameters are used and what they mean. It's also important to analyze which parameters is compared with which since small variations in parameters (e.g. GDP vs GDP per capita) can lead to very different results and conclusions.

code for assignment 1g:

```
countries = []
color = []

for index, rows in df.iterrows():
  if rows["Age"] < (y_mean) and rows["GDP"] > (x_mean):
    countries.append(rows["Country"])

print("Countries with low life expectancy and high GDP per capita:")
display(countries)
```

```
Countries with low life expectancy and high GDP per capita:
['Russia']
```

2a)

The first dataset we chose to compare was CO2 emissions versus GDP per capita. The most recent data on CO2 emissions was from 2012 and therefore we chose to use the 2012 data on GDP per capita as well for a fair comparison. With this data we will answer the following two questions:

Does a country's CO2 emissions increase with increased GDP per capita?

As can be seen in the graph below ("*CO2 emissions per capita vs GDP per capita 2012*") there is a pattern where countries with higher GDP per capita tend to also have a higher CO2 emission rate. This together with the general knowledge that with higher GDP per capita the people in general have a better economy and can spend their money on things that aren't mandatory for their survival. Hence, they can spend money on fuel for their car instead of cycling/walking to their work/school, buy a flight ticket instead of taking the train (even if the train in most cases are more expensive in richer countries then the flight), they can afford to buy clothes and furniture when they already have perfectly working ones. All these actions can contribute to a higher CO2 per capita and this makes for the conclusion that there is a relationship between increased CO2 emissions and increasing GDP.

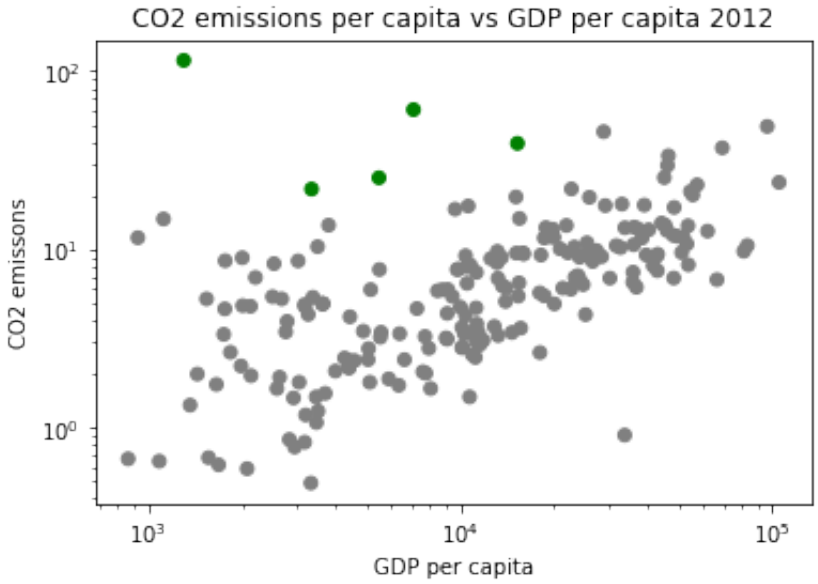
Are there any countries with very high CO2 emissions but low GDP per capita?

Assuming that very high CO2 emissions is defined by emissions higher than the mean plus one standard deviation and that low GDP per capita is less than the mean, the answer to this question is yes. The counties with high CO2 emissions despite of low GDP per capita is marked with a green color in the graph "*CO2 emissions per capita vs GDP per capita 2012*". This may be a result of the fact that companies in richer countries often plant the environmentally hazardous part of their production in poorer countries, which causes higher emissions in these places even when they are not caused by the country itself.

```
CO2 = {}
df = pd.DataFrame(columns=['Country', 'CO2', 'GDP'])

countries = []
color = []

CO2_capita = pd.read_csv(r'/content/drive/MyDrive/DAT405/Ass_1/ghg-emissions-per
```

```
for index, row in CO2_capita.iterrows():
  if row["Year"] == 2012:
    CO2[row["Entity"]] = row["GHG"]

for index, row in gdp_capita.iterrows():
  if row["Entity"] in CO2.keys() and row["Year"] == 2012 and not row["Entity"] =
    df = df.append({'Country': row["Entity"], 'CO2': CO2.get(row["Entity"]), 'GD

x_values = df['GDP']
y_values = df['CO2']

x_mean = st.mean(x_values)
x_std = st.stdev(x_values)
y_mean = st.mean(y_values)
y_std = st.stdev(y_values)

for index, rows in df.iterrows():
  if rows["CO2"] > (y_mean + y_std) and rows["GDP"] < (x_mean):
    countries.append(rows["Country"])
    color.append('green')
  else:
    color.append('gray')

print("Countries with very high CO2 emissions and low GDP per cpaita:")
display(countries)

plt.scatter(x_values, y_values, c=color)
plt.xlabel('GDP per capita')
plt.ylabel('CO2 emissons')
ay = plt.gca()
ay.set_xscale('log')
ay.set_yscale('log')
plt.title('CO2 emissions per capita vs GDP per capita 2012')
plt.show()
```

```
Countries with very high CO2 emissions and low GDP per cpaita:
['Bolivia', 'Botswana', 'Central African Republic', 'Laos', 'Zambia']
```



The second dataset that we compared against each other was experienced happiness among citizens of different countries and GDP per capita. The most recent data in the happiness file was from 2014 and therefore we chose to use the 2014 data from the GDP per capita file as well. We also made sure to only add countries with data from 2014 in both files. The questions that we will answer with this data is:

Does citizens experienced happiness increase with increased GDP per capita?

Looking at the graph below ("*Experienced happiness vs GDP per capita*") we can state that it is not necessarily a connection between increased wealth and increased happiness. Keeping this in mind we also note that countries with very high GDP per capita have a high happiness index as well. This being said, there is nothing in this graph that suggests that a low GDP per capita automatically means that the inhabitants are unhappy.

Which countries have a high happiness index but low GDP per capita?

Assuming that a "high" is defined as above mean and "low" is defined as below mean the answer is found in the first list of countries below and these are also marked with green in the graph, "*Experienced happiness vs GDP per capita*".

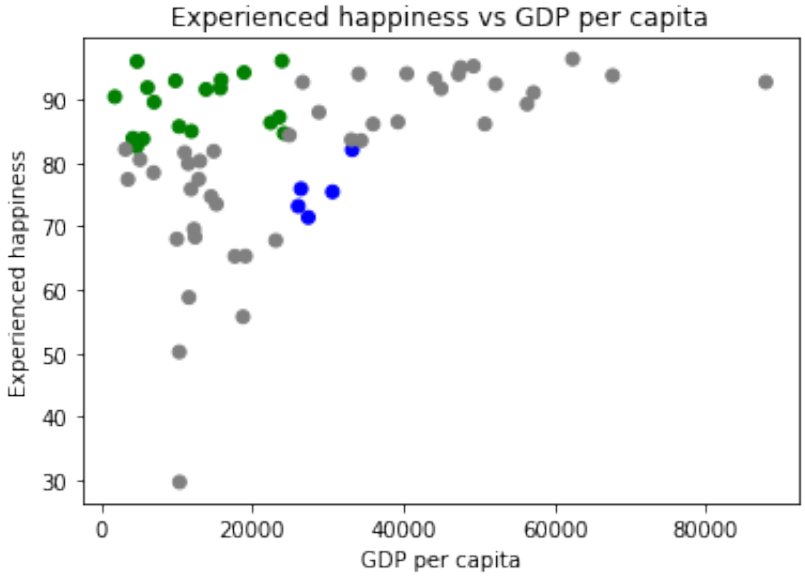
Which countries have a high GDP per capita but citizens with low experienced happiness?

Making the same assumptions as made in the previous question the answer is the countries in the second list below and these are marked with blue in the graph below.

```
happiness = {}
df = pd.DataFrame(columns=['Country', 'Happiness', 'GDP'])

countries_l = []
countries_h = []
color = []

exp_happiness = pd.read_csv(r'/content/drive/MyDrive/DAT405/Ass_1/share-who-say-

for index, row in exp_happiness.iterrows():
    happiness[row["Entity"]] = row["Share of people who are happy (World Value S

for index, row in gdp_capita.iterrows():
  if row["Entity"] in happiness.keys() and row["Year"] == 2014 and not row["Enti
```

```
    df = df.append({'Country': row["Entity"], 'Happiness': happiness.get(row["En

x_values = df['GDP']
y_values = df['Happiness']

x_mean = st.mean(x_values)
y_mean = st.mean(y_values)

for index, rows in df.iterrows():
  if rows["Happiness"] > (y_mean) and rows["GDP"] < (x_mean):
    countries_h.append(rows["Country"])
    color.append('green')
  elif rows["Happiness"] < (y_mean) and rows["GDP"] > (x_mean):
    color.append('blue')
    countries_l.append(rows["Country"])
  else:
    color.append('gray')

print("Countries with high experience and low GDP per capita")
print(countries_h)
print("")
print("Countries with low experience and high GDP per capita")
print(countries_l)

plt.scatter(x_values, y_values, c=color)

plt.title('Experienced happiness vs GDP per capita')
plt.xlabel('GDP per capita')
plt.ylabel('Experienced happiness')
plt.show()
```

```
Countries with high experience and low GDP per capita
['Argentina', 'Brazil', 'Chile', 'China', 'Colombia', 'Ghana', 'Indonesia',

Countries with low experience and high GDP per capita
['Cyprus', 'Estonia', 'Hungary', 'Russia', 'Slovakia']
```



Lastly, we chose to compare corruption levels against average years of schooling in different countries. When cleaning the data we chose to use the data from 2017 as this was the most recent for most countries. As in previous tasks we also made sure to only add countries with data from 2017 in both files (corruption and average years of schooling). With this data we will answer the following question.

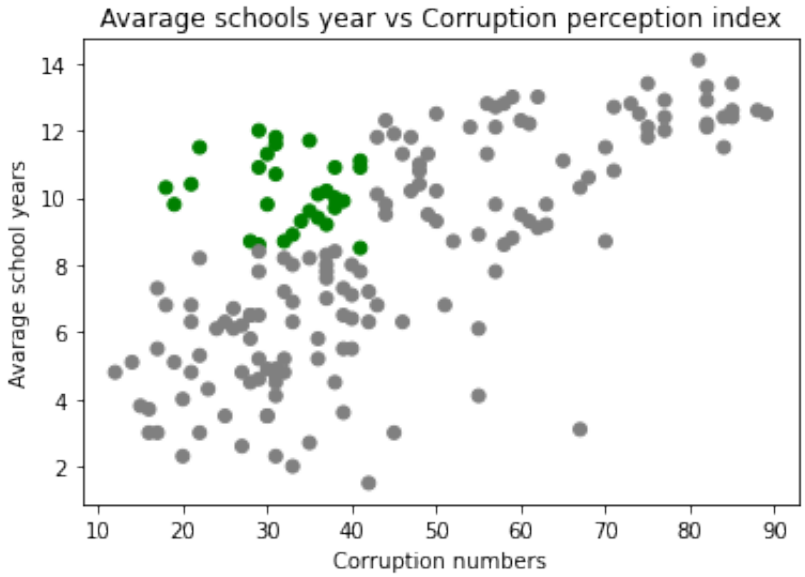
Can high levels of corruption be traced back to low average education.

Looking at the graph below called "*Average schools year vs Corruption perception index*" we can note that there seems to be some correlation between education and corruption since as the corruption perception index increases (meaning that the corruption levels decrease) the average schooling years of those countries tend to get higher. An interesting observation to be made though is that assuming high is above mean and low is below mean there are quite many countries with low corruption perception index (meaning much corruption) and high average schooling years. These countries are listed below and marked with green in the graph. This indicates that high levels of corruption can not only be traced to low levels of education.

```
import numpy as np

school = {}
```

```
df = pd.DataFrame(columns=['Country', 'School', 'Corruption'])

countries = []
color = []

avg_school = pd.read_csv(r'/content/drive/MyDrive/DAT405/Ass_1/average-schoolir
corruption = pd.read_csv(r'/content/drive/MyDrive/DAT405/Ass_1/TI-corruption-pe

for index, row in avg_school.iterrows():
  if row["Year"] == 2017:
    school[row["Entity"]] = row["Avarage"]

for index, row in corruption.iterrows():
  if row["Entity"] in school.keys() and row["Year"] == 2017 and not row["Entity
    df = df.append({'Country': row["Entity"], 'School': school.get(row["Entity"

x_values = df['Corruption']
y_values = df['School']

x_mean = st.mean(x_values)
x_std = st.stdev(x_values)

y_mean = np.nanmean(y_values)

for index, rows in df.iterrows():
  if rows["School"] > (y_mean) and rows["Corruption"] < (x_mean):
    countries.append(rows["Country"])
    color.append('green')
  else:
    color.append('gray')

print("Countries with average school years above mean and corruption belove mea
display(countries)

plt.scatter(x_values, y_values, c=color)
plt.title('Avarage schools year vs Corruption perception index')
plt.xlabel('Corruption numbers')
plt.ylabel('Avarage school years')
plt.show()
```

```
Countries with average school years above mean and corruption belove mean
['Albania',
 'Argentina',
 'Armenia',
 'Azerbaijan',
 'Bahrain',
 'Bolivia',
 'Bosnia and Herzegovina',
 'Ecuador',
 'Iran',
 'Kazakhstan',
 'Kyrgyzstan',
 'Lebanon',
 'Mexico',
 'Moldova',
 'Mongolia',
 'North Macedonia',
 'Panama',
 'Peru',
 'Philippines',
 'Russia',
 'Serbia',
 'Sri Lanka',
 'Suriname',
 'Tajikistan',
 'Trinidad and Tobago',
 'Turkmenistan',
 'Ukraine',
 'Uzbekistan',
 'Venezuela']
```



2b)

In general there are always data points (in this case countries) which are not following the pattern. This can make it harder to predict a single country's data point in a graph. On the other hand almost every graph has some pattern that all data points follow, which makes it easy to draw conclusions for a group of data points.

Another important aspect of drawing conclusions from a graph is to always have some background knowledge about what's going on. If you only draw a conclusion from how the data points follow a pattern in the graph, they can in reality have no correlation. For example if we look at cars being stolen and cars being bought over a couple of years. Maybe the number of cars being stolen increases at the same rate that the number of cars being bought increases. From this we could draw a hasty conclution that if there are more car thefts, more cars will be sold, but in reality maybe the increase in car theft is caused by a gang and the increase of cars being sold is a result of more families having the money to buy cars.

References

Rosling, H, Rosling A, Rosling O 2018, *Factfulness*, Natur & Kultur, Stockholm