DAT405/DIT406 - Introduction to data science and AI

Assignment 1: Introduction to Data Science and Python

Meng Yuan & Tomas Ekholm
Group 71

November 9, 2021

*Statement: We devote 20 hours each person working on this assignment.

1 Download some data related to GDP per capita and life expectancy.

a. Write a Python program that draws a scatter plot of GDP per capita vs life expectancy. State any assumptions and motivate decisions that you make when selecting data to be plotted, and in combining data. [1p]

Firstly, we download one original dataset related to GDP per capita and life expectancy [1] from Our World in Data. We extract only the columns named *'Entity'(e.g. Country names), 'Year','Life expectancy','GDP per capita', 'Total population (Gapminder, HYDE UN)'*. Then we exclude all non-country names (in this case, they are *'Africa', 'Asia', 'Europe', 'North America', 'Oceania', 'South America'* and *'World'*.) and weed out all NaN data for further analysis. At this step we have a new dataframe called *df_new* with 12403 rows × 5 columns.
Secondly, we draw a scatter plot based on non-duplicate countries only at Year 2018, at the same time, we scale size of dot with population as we use GDP per capita to indicate economy while it also relates to population. In some countries the overall GDP may be high but the GDP per capita is not relatively high due to the large population. This does not mean that the country's economy is not good, and vice versa. We hope that this situation will not affect our objectivity when discussing relevance. Finally, we draw the scatter and add a trend line as shown in Fig.1.
It is apparent that there is a logarithmic relationship between GDP per capita and Life expectancy. However, there are also some countries that do not conform to this trend, which we will discuss later.

b. Consider whether the results obtained seem reasonable and discuss what might be the explanation for the results you obtained. [1p]

It seems reasonable as first of all, there is no doubt that with the increase in GDP per capita, people's lives will be better, and life expectancy will definitely increase. However, life expectancy is also limited by some other factors, such as the level of medical development, so life expectancy will not keep increasing with GDP per capita infinitely. It will gradually tend to a constant value. The data on the diagram also conforms to the logarithmic performance.

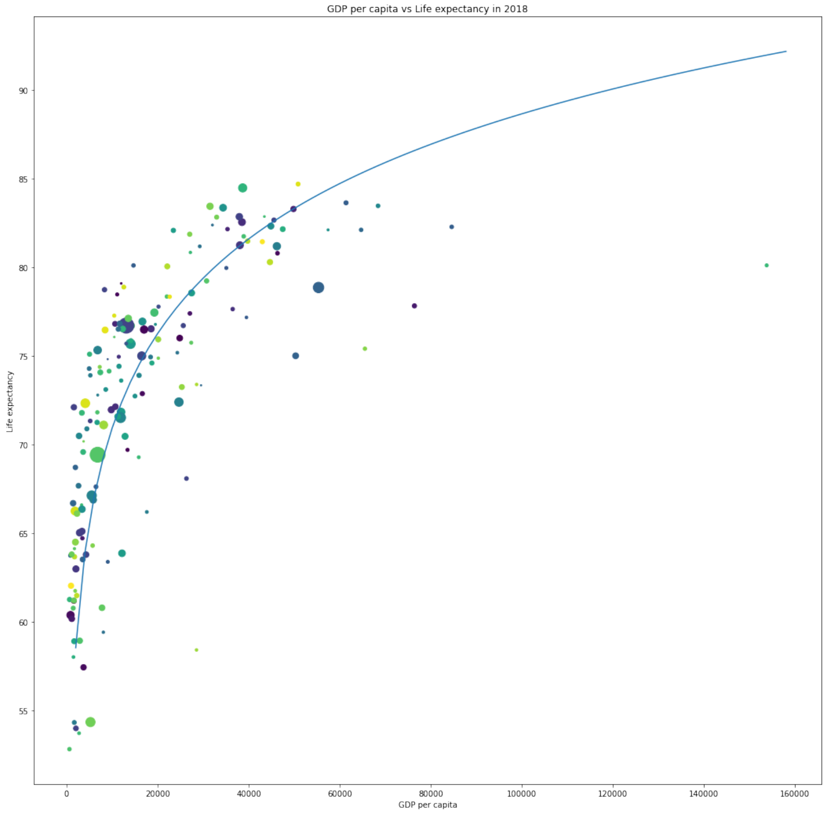

Figure 1: GDP per capita vs Life expectancy in 2018

c. Did you do any data cleaning (e.g., by removing entries that you think are not useful) for the task of drawing scatter plot(s) and the task of answering the questions d, e, f, and g? If so, explain what kind of entries that you chose to remove and why. [0.5p]

Yes we did. We only use *'GDP per capita'*, *'Life expectancy'* and *'Population'* data for countries in 2018 and None of them is NaN.
Reason: We only want to see relationship between *'GDP per capita'* and *'Life expectancy'* if there is no entry in any of these columns the data doesn't fit our purpose.

d. Which countries have a life expectancy higher than one standard deviation above the mean? [0.5p]

Australia, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany, Greece, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Kingdom
Related python code is attached in Appendix a. (This goes for all questions as well)

e. Which countries have high life expectancy but have low GDP? [0.5p]

In this question we define "high life expectancy" as being higher than 0.3(times the standard deviation) above the mean and "low GDP" as lower than 0.3(times the standard deviation) below the mean. To get our result we extract High_LE_list and Low_GDP_list to see if the data intersects. The resulting countries that fit the criteria are *Barbados, North Macedonia, Montenegro, Saint Lucia, Malta.* and *Iceland* Worth nothing is that they are all countries with a small population.

f. Does every strong economy (normally indicated by GDP) have high life expectancy? [1p]

In this question we define "strong economy" as having a GDP per capita higher than one standard deviation above the mean and correspondingly "high life expectancy" as one standard deviation above the mean, same as in question d. In this case, we extract the countries in High_GDP_list and compare against those not in the High_LE_list. This gives us a list of strong economies without high life expectancy. In this case, they are *China, India, Russia, Indonesia, United States*, and *Brazil*. This is 6 of the worlds top 9 most populous countries. In essence these countries are opposite to the ones we found in in question h.

g. Related to question f, what would happen if you use GDP per capita as an indicator of strong economy? Explain the results you obtained, and discuss any insights you get from comparing the results of g and f. [1p]

The countries that would have been found if using GDP per capita instead of GDP are the following. *Qatar, Bahrain, United States, Kuwait, United Arab Emirates, Saudi Arabia* and *Taiwan*. This is an entirely different set of countries. The only overlap is *The United States.* So how do they differ? Well the common denominator in the first set is that they are very populous, why does this matter? Because it drives up the GDP while not necessarily doing anything for the GDP per capita. In the second set of countries 6 out of the 8 countries are gulf-states with huge oil reserves and a low population. Is any of the measurements better than the other? Well that would depend on the purpose. Both of them highlight two different kinds of "wealthy" countries, so is there any commonality between these groups? Without looking at data one speculation is that intra-country inequality in the distribution of wealth is the reason both these different groups of countries makes it onto this dubious list. Something well worth looking into.

2 Download some other data sets, e.g. related to happiness and life satisfaction, trust, corruption, etc.

In this section, we download two datasets. The first one regarding Life satisfaction/Happiness vs Life Expectancy [2]. The second is on Trust in government [3] which we decided to compare against our existing data on GDP per capita. Below we have visualised the extracted data.

a. Think of several meaningful questions that can be answered with these data, make several informative visualisations to answer those questions. State any assumptions and motivate decisions that you make when selecting data to be plotted, and in combining data. [2.5p]

One possible question to investigate is: Do life expectancy affect how satisfied we are with life? One might think it is obvious that we would be happier the longer we live but consider the effects of various ailments as old age takes its toll on both body and mind. On the opposite extreme one might find it hard to be happy in a society where death might suddenly strike you.
The other question we want to try and answer is: Do people trust their government more if their country does well financially? It definitely seems plausible that GDP per capita is a decent measure of how good a government is doing.
To plot GDP per capita vs Share of people who trust their national government, since we only obtained trust data in 2018, we first have to merge it with the "dataset" in 1 into a new dataset named 'df_new3', de-duplicate and draw a scatter plot based on it. Since the data does not have a very strong correlation, we don't add a trend line in figure 3.

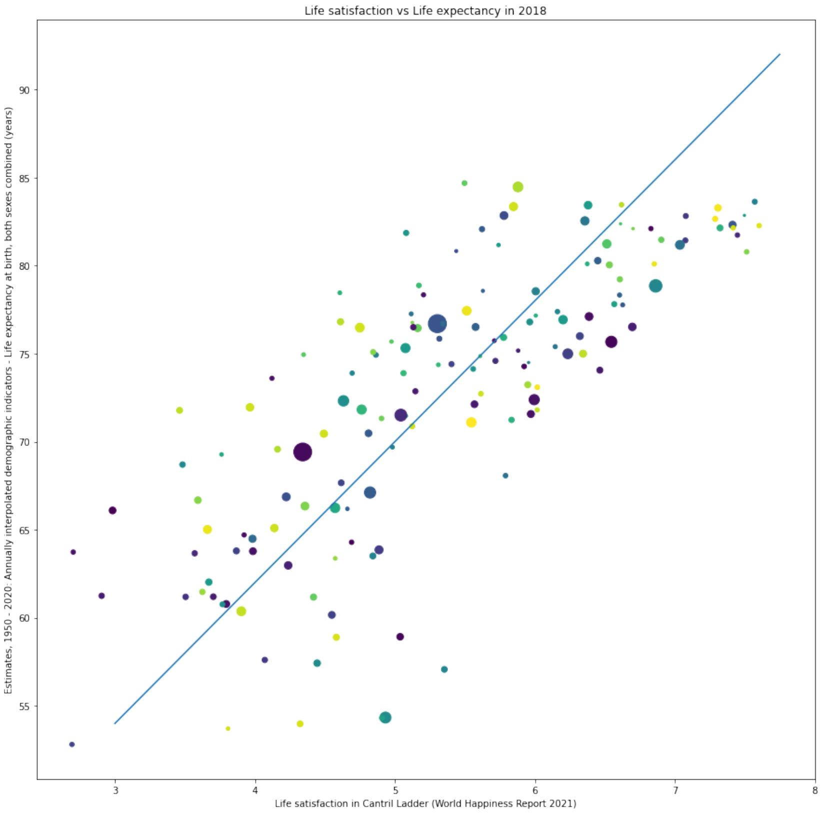

Figure 2: Life satisfaction vs Life expectancy in 2018

b. Discuss any observations that you make, or insights obtained, from the data visualisations. [2p]

Life satisfaction and Life expectancy clearly have a strong linear relationship. So it is definitely plausible to conclude that either we are more satisfied from the prospect of a long life or that we simply live longer when we are happy. Which causes which can be hard to say. It would also be wise to investigate if possible other factors could impact both. For example a wealthy society might be able offer its citizens both amenities and entertainment that increase Life satisfaction as well as Good and well funded health-care systems that increase Life expectancy. Further analysis can be done here if given time.

Life satisfaction vs Life expectancy, we extract data on life expectancy, happiness and population in 2018, have de-duplication on dataset as before, and draw a trend line.

There is no strong correlation between GDP per capita and Share of people who trust their national government. It seems that whether people trust their government or not does it is not correlated with the macroeconomic situation

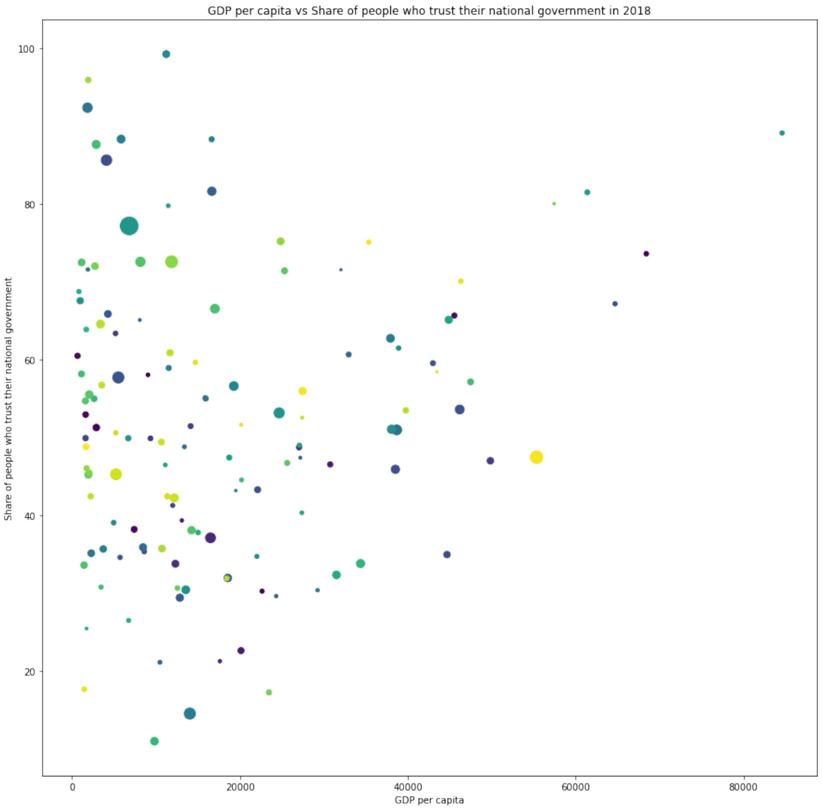

Figure 3: GDP per capita vs Share of people who trust their national government in 2018

and vice versa. Contrary to our previous predictions.

References

[1] O. W. in Data, "Life expectancy vs. gdp per capita, 2018," https://ourworldindata.org/grapher/life-expectancy-vs-gdp-per-capita?minPopulationFilter=1000000.

[2] ——, "Life satisfaction vs life expectancy, 1950 to 2020," https://ourworldindata.org/grapher/life-satisfaction-vs-life-expectancy?time=1950..2020.

[3] ——, "Share of people that trust their national government, 2018," https://ourworldindata.org/grapher/share-who-trust-government.

A Appendix a

```
#!/usr/bin/env python
# coding: utf-8

# ## DAT405 - Assignment 1
#
# ### Name1: Meng Yuan
#
# ### Name2: Tomas Ekholm
#
# ### Group: 71
#
# **Statement: We devote 20 hours each person working on this assignment.**

# In[1]:


# imports part
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from scipy.optimize import leastsq


# ### 1. Download some data related to GDP per capita and life expectancy.

# In[2]:


# import data set
df = pd.read_csv (r'/Users/mengyuan/Desktop/master_materials/intro/ass1/life-exp
df


# In[3]:


#extract data with country Name, Year, Life expectancy and GDP per capita and po
df_new = pd.DataFrame(df, columns=['Entity','Year','Life␣expectancy','GDP␣per␣ca

#Weed out/filter NaN data
df_new = df_new.dropna()

# exclude all continents names as they are not countries
```

```
df_new = df_new[(df_new['Entity']!='Africa')&(df_new['Entity']!='Asia')&(df_new[

df_new


# **a. Write a Python program that draws a scatter plot of GDP per capita vs lif
#
#

# In[4]:


# draw a scatter plot only based on Year 2018
dataset = pd.DataFrame(df_new.loc[(df_new['Year']==2018)], columns=['Entity','Li

country_names = np.unique(dataset['Entity']) #in case any duplicate countries
colors = np.random.rand(len(country_names))

plt.figure(figsize=(18,18))

plt.scatter(dataset['GDP per capita'],
            dataset['Life expectancy'],
            c=colors, s=(0.0001*dataset['Total population (Gapminder, HYDE & UN)

x1 = np.arange(0,160000,2000)
y1 = 7.7*np.log(x1)
plt.plot(x1,y1,label='Fitting function: y = 7.7log(x)')

plt.xlabel("GDP per capita")
plt.ylabel("Life expectancy")
plt.title('GDP per capita vs Life expectancy in 2018')

plt.show()


# **b. Consider whether the results obtained seem reasonable and discuss what m
#
# **Answer:**
# It seems reasonable as first of all, there is no doubt that with the increase

# ### Answer these questions:
#
# **c. Did you do any data cleaning (e.g., by removing entries that you think ar
#
# **Answer:**
#
```

```
# Yes we did. We only use 'GDP per capita', 'Life expectancy' and
'Population' data for countries in 2018 and None of them is NaN.
#
# Reason: We only want to see relationship between 'GDP per capita' and 'Life ex

# **d. Which countries have a life expectancy higher than one standard deviation
#
# **Answer:**
# Australia, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany

# In[5]:


country_name_list = dataset['Entity'].values.tolist()
Life_expectancy_list = dataset['Life␣expectancy'].values.tolist()
GDP_list = dataset['GDP␣per␣capita'].values.tolist()
population_list = dataset['Total␣population␣(Gapminder,␣HYDE␣&␣UN)'].values.toli


#mean of Life expectancy
mean_of_LE = np.mean(Life_expectancy_list)
#standard deviation of Life expectancy
std_of_LE = np.std(Life_expectancy_list)
# print(std_of_LE)

#calculate overall GDP
all_GDP_list = []
for i in range(len(Life_expectancy_list)):
    all_GDP_list.append(GDP_list[i]*population_list[i])

#mean of GDP per capita
mean_of_GDP = np.mean(GDP_list)
#standard deviation of GDP per capita
std_of_GDP = np.std(GDP_list)

#mean of overall GDP
mean_of_overall_GDP = np.mean(all_GDP_list)
#standard deviation of overall GDP
std_of_overall_GDP = np.std(all_GDP_list)

Answer1 = []
for i in range(len(Life_expectancy_list)):
    if Life_expectancy_list[i] - mean_of_LE >= std_of_LE:
        Answer1.append(country_name_list[i])

print(Answer1)
```

```
# **e. Which countries have high life expectancy but have low GDP? [0.5p]**
#
# **Answer:**
#
# In this question we define ”high life expectancy” as being higher than 0.3(tim
criteria are **Barbados, North Macedonia, Montenegro, Saint Lucia, Malta.** and

# In[6]:


High_LE_list = []
for i in range(len(Life_expectancy_list)):
    if Life_expectancy_list[i] - mean_of_LE >= 0.3*std_of_LE:
        High_LE_list.append(country_name_list[i])

# print(High_LE_list)

Low_GDP_list = []
for i in range(len(all_GDP_list)):
    if mean_of_overall_GDP - all_GDP_list[i] >= 0.3*std_of_overall_GDP:
        Low_GDP_list.append(country_name_list[i])

# print(Low_GDP_list)

Answer2 = list(set(High_LE_list).intersection(set(Low_GDP_list)))
print(Answer2)


# **f. Does every strong economy (normally indicated by GDP) have high life expe
#
# **Answer:**
#
# In this question we define ”strong ecomomy” as GDP per capita higher than one

# In[7]:


High_overall_GDP_list = []

for i in range(len(all_GDP_list)):
    if all_GDP_list[i] - mean_of_overall_GDP >= std_of_overall_GDP:
        High_overall_GDP_list.append(country_name_list[i])

Answer3 = list(set(High_overall_GDP_list).difference(set(Answer1)))
```

```
print(Answer3)


# **g. Related to question f, what would happen if you use GDP per capita as an
#
# **Answer:**
#
# The countries that would have been found if using GDP per capita instead of GD
# United Arab Emirates, Saudi Arabia and Taiwan**. This is an entirely different

# In[8]:


High_GDP_list = []

for i in range(len(GDP_list)):
    if GDP_list[i] - mean_of_GDP >= std_of_GDP:
        High_GDP_list.append(country_name_list[i])

Answer4 = list(set(High_GDP_list).difference(set(Answer1)))
print(Answer4)


# ### 2. Download some other data sets, e.g. related to happiness and life satis
#
# **a. Think of several meaningful questions that can be answered with these dat
#
# **Answer:**
#

# In[9]:


#In first section we compare Life satisfaction/Hapiness and Life Expectancy

# import dataset
df2 = pd.read_csv (r'/Users/mengyuan/Desktop/master_materials/intro/ass1/life-sa
#extract data with country Name, Year, Life expectancy and Life satisfaction and
df_new2 = pd.DataFrame(df2, columns=['Entity','Year','Estimates,␣1950␣-␣2020:␣Ar

#Weed out/filter NaN data
df_new2 = df_new2.dropna()

# exclude all continents names as they are not countries
df_new2 = df_new2[(df_new2['Entity']!='Africa')&(df_new2['Entity']!='Asia')&(df_
df_new2


# In[10]:


# draw a scatter plot only based on Year 2018
dataset2 = pd.DataFrame(df_new2.loc[(df_new2['Year']==2018.0)], columns=['Entity

country_names2 = np.unique(dataset2['Entity']) #in case any duplicate countries
colors2 = np.random.rand(len(country_names2))

plt.figure(figsize=(15,15))

plt.scatter(dataset2['Life␣satisfaction␣in␣Cantril␣Ladder␣(World␣Happiness␣Repor
            dataset2['Estimates,␣1950␣-␣2020:␣Annually␣interpolated␣demographic␣
            c=colors2, s=(0.0001*dataset2['Total␣population␣(Gapminder,␣HYDE␣&␣U

x2 = np.arange(3,8,0.25)
y2 = 8*x2+30
plt.plot(x2,y2,label='Fitting␣function:␣y␣=␣8x+30')

plt.xlabel("Life␣satisfaction␣in␣Cantril␣Ladder␣(World␣Happiness␣Report␣2021)")
plt.ylabel("Estimates,␣1950␣-␣2020:␣Annually␣interpolated␣demographic␣indicators
plt.title('Life␣satisfaction␣vs␣Life␣expectancy␣in␣2018')

plt.show()


# In[11]:


#In second section we compare GDP per capita and trust in government
# import dataset and keep trust only
df3 = pd.read_csv (r'/Users/mengyuan/Desktop/master_materials/intro/ass1/share-w
df3


# In[12]:


#As we only have year(2018) and trust data in this dataset, we have to merge it
df_new3 = pd.merge(dataset, pd.DataFrame(df3, columns=['Entity','Share␣of␣people

#Weed out/filter NaN data
df_new3 = df_new3.dropna()
```

```
# exclude all continents names as they are not countries
df_new3 = df_new3[(df_new3['Entity']!='Africa')&(df_new3['Entity']!='Asia')&(df_n
df_new3


# In[13]:


country_names3 = np.unique(df_new3['Entity']) #in case any duplicate countries
colors3 = np.random.rand(len(country_names3))

plt.figure(figsize=(15,15))

plt.scatter(df_new3['GDP␣per␣capita'],
            df_new3['Share␣of␣people␣who␣trust␣their␣national␣government'],
            c=colors3, s=(0.0001*df_new3['Total␣population␣(Gapminder,␣HYDE␣&␣UN

# x1 = np.arange(0,160000,2000)
# y1 = 7.7*np.log(x1)
# plt.plot(x1,y1)

plt.xlabel("GDP␣per␣capita")
plt.ylabel("Share␣of␣people␣who␣trust␣their␣national␣government")
plt.title('GDP␣per␣capita␣vs␣Share␣of␣people␣who␣trust␣their␣national␣government

plt.show()


# **b. Discuss any observations that you make, or insights obtained, from the da
#
# **Answer:**
#
# Life satisfaction and Life expectancy clearly have a strong linear relationshi
#
# Life satisfaction vs Life expectancy, we extract data on life expectancy, happ
# draw a trend line.
#
# There is no strong correlation between GDP per capita and Share of people who

# In[ ]:
```