DAT405 Assignment 1 – Group 53

Venkata Sai Dinesh Uddagiri - (14 hrs)
Madhumitha Venkatesan - (14 hrs)

November 15, 2022

Problem 1

1(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

```
import pandas as pnd
import matplotlib.pyplot as plt
#Reading csv file in to data frame
le_Gdp_DF_Raw=pnd.read_csv("life-expectancy-vs-gdp-per-capita.csv")
#Data cleaning
le_gdp_DF_clean = le_Gdp_DF_Raw.dropna(subset = ['Life expectancy', 'GDP per capita','
    Population (historical estimates)'])
del le_gdp_DF_clean["417485-annotations"], le_gdp_DF_clean["Continent"]
le_gdp_DF_clean = (le_Gdp_DF_Raw[(le_Gdp_DF_Raw['Year'] >= 2011) & (le_Gdp_DF_Raw['
    Year'] <= 2019)]
                .dropna(subset = ['Life expectancy', 'GDP per capita','Population (
    historical estimates)']))
del le_gdp_DF_clean["417485-annotations"], le_gdp_DF_clean["Continent"]
#Obtaining mean 'Life expectancy','GDP per capita' and 'Population (historical
    estimates)'of based on 'Entity'
le_gdp_DF_clean =le_gdp_DF_clean.groupby(['Entity'])[['Life expectancy','GDP per
    capita','Population (historical estimates)'
                                                    ]].agg('mean')
#Drawing scatter plot with corresponding axis Labels and title
plt.scatter(le_gdp_DF_clean['GDP per capita'], le_gdp_DF_clean['Life expectancy'])
plt.title("GDP per capita versus Life-expectancy ", fontsize = 18)
plt.xlabel("GDP per capita", fontsize = 18)
plt.ylabel("Life Expectancy", fontsize = 18)
plt.xscale("log")
plt.show()
```

Listing 1: GDP per capita versus Life-expectancy Python code – scatter plot.

We have used the data provided by Ourworldindata.org. The original data provided, has the data of multiple years and for more than 150 countries. But We have considered data from years 2011 to 2019.Since these years are the most recent in our data set and which also includes information on life expectancy and GDP per capita. We have obtained mean of Life expectancy and GDP per capita for each entity, from years 2011 to 2019 to limit amount of data to correlate. We select GDP per capita for the X axis and life expectancy for the Y axis. By observing the graph we believe that GDP per capita can have a significant impact on life expectancy, but not the other way around.

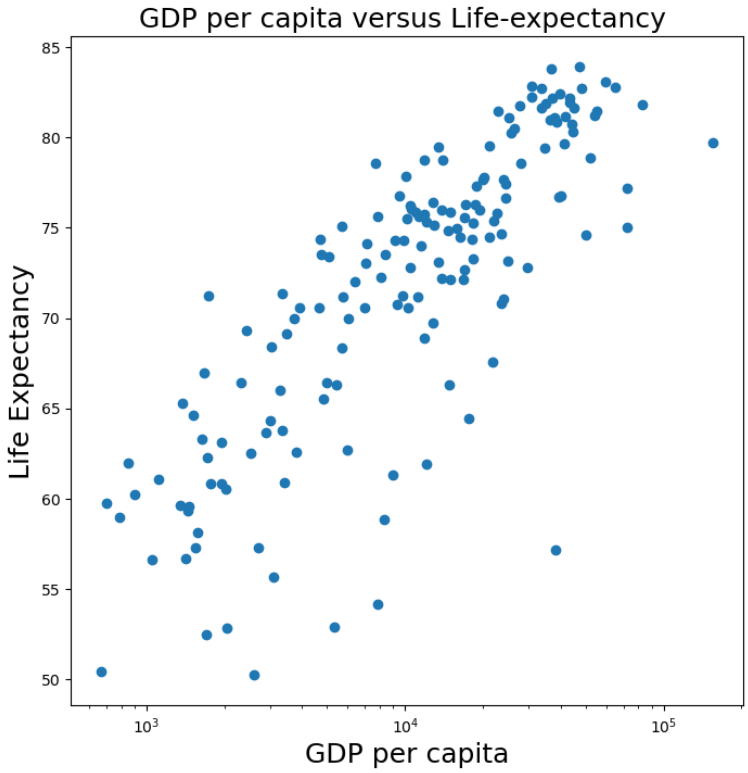

Figure 1: Scatter plot for GDP per capita Vs Life expectancy

1(b) Consider whether the results obtained seem reasonable and discuss what might be the explanation for the results you obtained.

According to the scatter plot we produced above, there is a correlation between GDP per capita and life expectancy.This finding suggests that a country with a high GDP per capita has a prosperous economy because life expectancy rises as GDP per capita rises. Which seems reasonable, based on real world knowledge. Which means country with high GDP capita has the rich economy. The rich economy in turn has best facilities to its citizens, one of the facility is healthcare. The citizens having better health care access in result have high life expectancy. One more thing we observed in scatter plot is life expectancy become constant after certain GDP point. which is acceptable fact according to human nature. By this we conclude that results obtained looks reasonable.

1(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.

Yes, we have cleaned up the data. The major clean-up was removing all the data other than the years 2011 to 2019.

Due to the fact that we are establishing a relationship between GDP per capita and life expectancy, we have dropped the rows with NaN values from the GDP per capita and life expectancy columns.

We are also deleted two columns 417485-annotations and Continent. Continent data might be useful but, looking at the raw data above one notices that it mostly contains NaN values.

For 1.e: Because we are calculating the GDP, we have deleted the entries with NaN values in the

Population (historical estimates) column. This is a combination of population and GDP per capita in the country.

1(d) Which countries have a life expectancy higher than one standard deviation above the mean?

```
#Calculation of standard deviation for Life expectancy
StandardDeviation = le_gdp_DF_clean.describe().loc['std', 'Life expectancy']
#Calculation of mean for Life expectancy
mean = le_gdp_DF_clean.describe().loc['mean', 'Life expectancy']
#Printing mean and StandardDeviation of Life expectancy
print("StandardDeviation of Life expectancy is", StandardDeviation)
print("mean of Life expectancy is", StandardDeviation)
#calculation of life expectancy higher than one standard deviation above the mean
le_gdp_DF_clean[le_gdp_DF_clean['Life expectancy'] > mean + StandardDeviation].
    sort_values(by= 'Life expectancy', ascending = False)
```

Listing 2: Countries with One standard deviation above the mean Python code.

Standard Deviation of Life expectancy is 8.235309996831452
mean of Life expectancy is 8.235309996831452
one standard deviation above the mean = mean + StandardDeviation

One standard deviation above the mean is sum of the mean and Standard Deviation. The Countries with One standard deviation above the mean are Hong Kong, Japan, Switzerland, Spain, Singapore, Italy, Australia, Iceland, Israel, Sweden, France, Canada, South Korea, Norway, Malta,New Zealand, Netherlands, Luxembourg, Greece, Ireland, Austria, Portugal, Finland, United Kingdom, Belgium, Germany, Slovenia, Denmark, Cyprus.

1(e) Which countries have high life expectancy but have low GDP?

We have used the median value as reference to obtain "countries have high life expectancy but have low GDP". which is, countries with life expectancy greater than the median and countries with GDP lower than the median GDP. List of countries have high life expectancy but have low GDP using median are Iceland, Malta, Luxembourg, Slovenia, Cyprus, Costa Rica, Barbados, Lebanon, Cuba, Albania, Croatia, Panama, Estonia, Uruguay, Bosnia and Herzegovina, Bahrain, Montenegro, Saint Lucia, North Macedonia, Lithuania, Latvia, Mauritius, Honduras, Armenia, Dominica, Jamaica

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

Considering countries with GDP greater than the median as highest and Life expectancy greater than the median as highest.

We used GDP that was above the median and life expectancy that was below the median to perform this check. If the statement in the question is accurate, we ought to receive no output. However, results indicate that few countries have high GDP which means, above the median has low life expectancy's which means below median. By this we can conclude that every stronger economy doesn't have high life expectancy.

1(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.

According to the statistics, GDP per capita correlates with life expectancy significantly more strongly than GDP does, making it a much better measure of personal well-being as even countries with high

GDPs may nevertheless have low GDP per capita due to their big populations.

Furthermore, a high GDP is not always accompanied by a high GDP per person. In general, countries with large populations will have greater GDP's, but when that GDP is divided up among the population, or GDP per capita, it might be much lower. Therefore, unless GDP is converted to a percentage of a country's population, it truly doesn't mean anything.

Problem 2

2(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.

We decided to select two data sets from https://ourworldindata.org/ - life-expectancy-vs-health-expenditure.csv and average-years-of-schooling-vs-gdp-per-capita.csv for problem 2. Visualisations were made between Health Expenditure vs Life expectancy and Average Total Years of Schooling vs GDP per capita. The goal in selecting this dataset was to determine correlation between Health Expenditure vs Life expectancy and Average Total Years of Schooling vs GDP per capita .

Health expenditure versus life-expectancy

```
import pandas as pnd
import matplotlib.pyplot as plt
#Reading csv file in to data frame
import seaborn as sns
import numpy as np
le_he_DF_Raw=pnd.read_csv("life-expectancy-vs-health-expenditure.csv")
#Data cleaning
le_he_DF_clean = (le_he_DF_Raw[(le_he_DF_Raw['Year'] >= 2011) & (le_he_DF_Raw['Year']
    <= 2020)]
                  .dropna(subset = ['Life expectancy at birth, total (years)', 'Health
     Expenditure and Financing (per capita) (OECDstat (2017))']))
Countries_list= ['Australia','Canada','France','Denmark','Finland','Japan','Sweden','
    Germany']
#plot size
plt.figure(figsize=(8,8))
for i in range(len(Countries_list)):
    df_Entity = le_he_DF_clean.loc[le_he_DF_clean['Entity'] == Countries_list[i]]
    df_Entity_le = df_Entity['Life expectancy at birth, total (years)']
    df_Entity_helExp = df_Entity['Health Expenditure and Financing (per capita) (
    OECDstat (2017))']
    #generating random clour for each entity
    clr = (np.random.random(), np.random.random(), np.random.random())
    #Instance of plot
    plt.plot(df_Entity_helExp,df_Entity_le, color = clr)
#Assigning plot with corresponding axis Labels and title
plt.xticks(rotation ='vertical')
plt.legend(Countries_list, loc = "lower left")
fs=18
plt.title('health-expenditure-vs-life-expectancy',fontsize = fs)
plt.xlabel('Health Expenditure',fontsize = fs)
plt.ylabel('Life expectancy',fontsize = fs)
plt.show()
```

Listing 3: Health expenditure versus life-expectancy

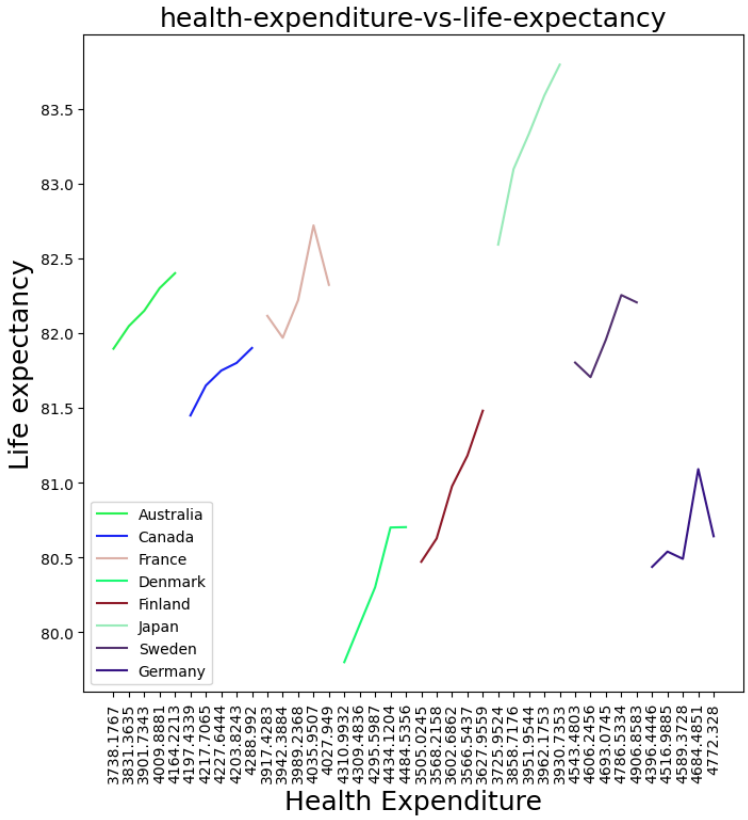

Figure 2: Health expenditure versus life-expectancy

Average years of schooling versus Gdp per capita

```
import pandas as pnd
import matplotlib.pyplot as plt
import seaborn as sns
#Reading csv file in to data frame
ayos_gdp_DF_Raw=pnd.read_csv("average-years-of-schooling-vs-gdp-per-capita.csv")
#Data cleaning
ayos_gdp_DF_clean = (ayos_gdp_DF_Raw[(ayos_gdp_DF_Raw['Year'] == 2017)]
                  .dropna(subset = ['Average Total Years of Schooling for Adult
    Population (Lee-Lee (2016), Barro-Lee (2018) and UNDP (2018))', 'GDP per capita,
    PPP (constant 2017 international $)']))
#Plot size
plt.figure(figsize=(8,8))
#Scatter plot with corresponding axis Labels and title
plt.scatter(ayos_gdp_DF_clean['GDP per capita, PPP (constant 2017 international $)'],
    ayos_gdp_DF_clean['Average Total Years of Schooling for Adult Population (Lee-Lee
    (2016), Barro-Lee (2018) and UNDP (2018))'])
fs=18
plt.title("average-years-of-schooling-vs-gdp-per-capita ", fontsize = fs)
plt.xlabel("GDP per capita", fontsize = 18)
plt.ylabel("Average Total Years of Schooling", fontsize = fs)
plt.xticks(rotation ='vertical')
plt.xscale('log')
plt.show()
```

Listing 4: Code for Average years of schooling versus Gdp per capita

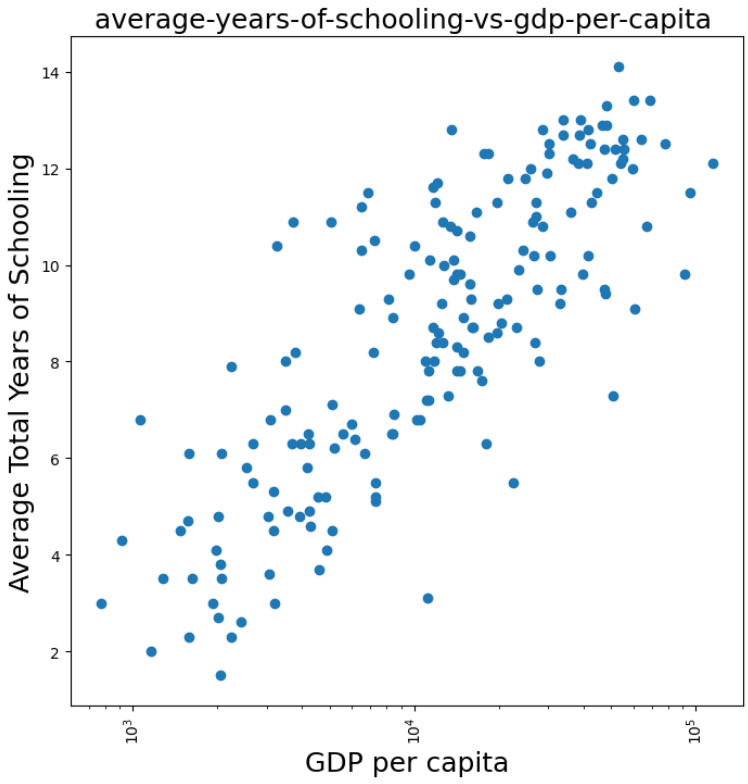

Figure 3: Average years of schooling versus GDP per capita

Questions

i. What is technique used to visualise the data?
For Health Expenditure vs Life expectancy, we have used the standard plot to visualize the data of certain countries, picked randomly, which places Life Expectancy on the y axis and Health Expenditure on the x axis. For each entity, a random color has been assigned in order to distinguish the countries.

For Average Total Years of Schooling vs GDP per capita , we have used the scatter plot to visualize the data of all the countries for year 2017, with GDP per Capita on the x axis and Average Total Years of Schooling on the y axis. Although we utilized the same method for problem 1, we think a correlation between Average Total Years of Schooling and GDP per Capita is optimal.

ii. Is any data cleaning preformed?
Data cleaning is performed in both the data frames.

In Health Expenditure vs Life expectancy, we have considered data of the recent years (2011 to 2019), removed irrelevant data and removed the data which has null values in Health Expenditure and Life expectancy columns.

In Average Total Years of Schooling vs GDP per capita, we have considered only the data of year 2017 as this the latest year of data available in our data set and which have information for most of countries and removed the data which has null values in Average Total Years of Schooling and GDP per capita columns.

iii(a). What is relation between Health Expenditure and Life expectancy?
The countries spend more on health expenditure, which leads to more advanced and well infrastructure health care systems and can provide the best treatment to their people. Which in result increases average Life expectancy of people.

iii(b). What is relation between Average Total Years of Schooling and GDP per capita?
People in nations with long average school years are more educated and knowledgeable. As a result, there will be an increase in the capacity to produce new innovations, an increase in personal income , increase in the income of the organizations for which they work, an increase in labor force size and decrease in unemployment rates. Which in total increase the GDP per capita of the country.

2(b) Discuss any observations that you make, or insights obtained, from the data visualisations.

Health Expenditure vs Life expectancy

In the plot above, we can see that when a nation's health expenditure rises, the average life expectancy of the nation generally rises as well. Although health expenditure costs have increased, life expectancy has decreased for the countries of France, Sweden, and Germany in recent years. We can conclude that there is a relationship between expenditures on health care and life expectancy based on the first observation made above. However, the second observation leads us to believe that it might not always be the case. There may be many additional factors that influence life expectancy outside healthcare expenditures, such as the prevalence of diseases for which there are no effective therapies, higher mortality rates, and more. Another finding that can be seen is that certain nations have higher healthcare expenditure than others, but having shorter life expectancy. The cost of producing medicine, paying doctors' salaries, and other factors may also affect healthcare expenditure.

Average Total Years of Schooling vs GDP per capita

We can observe a strong link between GDP per capita and Average Total Years of Education in the scatter plot above, the nation with high average years of schooling has high GDP's. However, the majority of the graph appears to be expanding linearly. According to 2a, a high level of education and knowledge among the population leads to high income for each individual. which country has a high GDP per capita as a result. The more GDP a country has, the more likely it is that it would provide free education to its citizens. As a result, more people may consider pursuing higher education to gain competence in their desired fields, which raises the average number of years spent in school. Here, we can see the two-way relationship between GDP per capita and the average total number of years spent in school.

References

[1] Donald E. Knuth (1986) *The TEX Book*, Addison-Wesley Professional.

[2] Leslie Lamport (1994) *LATEX: a document preparation system*, Addison Wesley, Massachusetts, 2nd ed.

[3] Max Roser- "Economic Growth". Published online at ourworldindata.org Retrieved from:' https://ourworldindata.org/economic-growth' [Online Resource]

[4] by Max Roser(May 26, 2017)- "Link between health spending and life expectancy". Published online at ourworldindata.org Retrieved from:' https://ourworldindata.org/the-link-between-life-expectancy-and-health-spending-us-focus' [Online Resource]

[5] by LEE-LEE (2016), BARRO-LEE (2018) - "Average years of schooling vs. GDP per capita, 2017". Published online at ourworldindata.org Retrieved from:' https://ourworldindata.org/grapher/average-years-of-schooling-vs-gdp-per-capita' [Online Resource]