Assignment 1

Students

- Total number of hours spent on assignment per person: 7 hours.
- Davíð Freyr Björnsson
- Eric Guldbrand

1

Write a Python program that draws a scatter plot of GDP per capita vs life expectancy. State any assumptions an be plotted, and in combining data. Answer these questions:

1. Which countries have a life expectancy higher than one standard deviation above the mean?
2. Which countries have high life expectancy but have low GDP?
3. Does every strong economy have high life expectancy?

Description of the data

- Data for GDP per capita comes from the World Bank and contains the period 1990 to 2017.
- Data for life expectancy comes from OWID. OWID has combined two data sources: Clio-Infra's dataset for Population Division for data from 1950 to 2015.

```
# Load Data

import pandas as pd

gdpPerCapitaURL = 'https://raw.githubusercontent.com/ecen/data-science-intro/master/g
gdpPerCapitaDF = pd.read_csv(gdpPerCapitaURL)

lifeExpectancyURL = 'https://raw.githubusercontent.com/ecen/data-science-intro/master
lifeExpectancyDF = pd.read_csv(lifeExpectancyURL)
```

```
# Look at the structure of the data
gdpPerCapitaDF.describe()
```

```
# Rename the GDP per capita
gdpPerCapitaDF.rename(columns={"GDP per capita (int.-$) (constant 2011 international
```

```
lifeExpectancyDF.rename(columns={"Life expectancy (Clio-Infra up to 1949; UN Populati
```

```
# Select and combine data

gdpPerCapita2015 = gdpPerCapitaDF.loc[gdpPerCapitaDF['Year'] == 2015]
lifeExpectancy2015 = lifeExpectancyDF.loc[lifeExpectancyDF['Year'] == 2015]

gdpVsLife = pd.merge(gdpPerCapita2015, lifeExpectancy2015)
```

It was decided to use data from 2015, the most recent year that existed in both the selected datasets.

The 2015 data from both datasets was then merged, retaining only countries that existed in both.

```
# Create a scatterplot, with GDP per capita in 2015 on the x-axis
# and life expectancy in 2015 on the y-axis
import matplotlib.pyplot as plt
import numpy as np

plt.scatter(gdpVsLife['gdpPerCapita'], gdpVsLife['lifeExp'])
plt.xlabel("GDP per capita (2015)")
plt.ylabel("life expectancy in 2015")
plt.show()
```

a. Which countries have a life expectancy higher than one standard deviat

```
# Selected countries with a life expectancy of one standard deviation or more over th
lifeStd = gdpVsLife['lifeExp'].std()
lifeMean = gdpVsLife['lifeExp'].mean()

lifeAboveOneStd = gdpVsLife.loc[gdpVsLife['lifeExp'] > lifeMean + lifeStd]

lifeAboveOneStd.Entity
```

b. Which countries have high life expectancy but have low GDP?

```
# Select countries with high life expectancy and low gdp
# We define high and low as the 75th respective 25th percentile

highGdp75 = gdpVsLife.gdpPerCapita.quantile(0.75)
highGdp50 = gdpVsLife.gdpPerCapita.quantile(0.5)
lowGdp50 = gdpVsLife.gdpPerCapita.quantile(0.5)
lowGdp25 = gdpVsLife.gdpPerCapita.quantile(0.25)
highLife75 = gdpVsLife.lifeExp.quantile(0.75)
highLife50 = gdpVsLife.lifeExp.quantile(0.5)

lifeHighGdpLow = gdpVsLife.loc[(gdpVsLife['lifeExp'] > highLife75) & (gdpVsLife['gdpP

lifeHighGdpLow.Entity
```

Given a more strict definition of high and low, 75th and 25th percentiles respectively, there are no countries with h

```
# Select countries with high life expectancy and low gdp
# We define high and low as the 50th respective 50th percentile

lifeHighGdpLow = gdpVsLife.loc[(gdpVsLife['lifeExp'] > highLife50) & (gdpVsLife['gdpP

lifeHighGdpLow.Entity
```

Given a less strict definition, with high and low being considered above and below the median (50th percentile) re
expectancy but low GDP.

c. Does every strong economy have high life expectancy?

```
gdpHigh = gdpVsLife.loc[gdpVsLife['gdpPerCapita'] > highGdp75]
lifeHighGdpHigh = gdpVsLife.loc[(gdpVsLife['lifeExp'] > highLife75) & (gdpVsLife['gdp

gdpHigh.shape[0] - lifeHighGdpHigh.shape[0]
```

```
plt.scatter(gdpVsLife['gdpPerCapita'], gdpVsLife['lifeExp'])
plt.xlabel("GDP per capita (2015)")
plt.ylabel("life expectancy in 2015")
plt.axvline(highGdp75, color='orange')
plt.axhline(highLife75, color='orange')
plt.text(highGdp75 - 10000, 87, '75th percentile')
plt.text(130000, highLife75, '75th percentile')
plt.show()
```

```
## Does the result change much if we alter our definition of low and high?
## Set high and low both as the 50th percentile

gdpHigh = gdpVsLife.loc[gdpVsLife['gdpPerCapita'] > highGdp50]

lifeHighGdpHigh = gdpVsLife.loc[(gdpVsLife['lifeExp'] > highLife50) & (gdpVsLife['gdp

gdpHigh.shape[0] - lifeHighGdpHigh.shape[0]
```

```
plt.scatter(gdpVsLife['gdpPerCapita'], gdpVsLife['lifeExp'])
plt.xlabel("GDP per capita (2015)")
plt.ylabel("life expectancy in 2015")
plt.axvline(highGdp50, color='orange')
```

```
plt.axhline(highLife50, color='orange')
plt.text(highGdp50 - 10000, 87, '50th percentile')
plt.text(130000, highLife50, '50th percentile')
plt.show()
```

Not every strong economy has high life expectancy. If we define high as above 75th percentile and low as below
strong economies and not high life expectancy. If we define high and low to be above and below the median then

2. Economic growth compared to tertiary education enr

```
# Annual growth of GDP per capita, adjusted for inflation, 1961 to 2014: https://ourw
gdpPerCapitaGrowth = pd.read_csv('https://raw.githubusercontent.com/ecen/data-science
gdpPerCapitaGrowth.rename(columns={"GDP per capita growth (annual %) (%)":"growth"},

# Tertiary enrollment: Total enrollment in tertiary education (ISCED 5 to 8), regardl
# expressed as a percentage of the total population of the five-year age group follow
# Years: 1970 to 2015. Source: https://data.worldbank.org/data-catalog/ed-stats, UNES
# expressed as a percentage of the total population of the five-year age group follow
enrollment = pd.read_csv('https://raw.githubusercontent.com/ecen/data-science-intro/m
enrollment.rename(columns={"Gross enrolment ratio, tertiary, both sexes (%) (%)":"enr
```

```
def plotYears(startYear, endYear):
  growthData = gdpPerCapitaGrowth[(gdpPerCapitaGrowth['Year'] >= startYear) & (gdpPer
  enrollmentData = enrollment[(enrollment['Year'] >= startYear) & (enrollment['Year']
  data = growthData.merge(enrollmentData, on='Entity')

  # Remove outliers to make it easier to compare the distribution of the majority
  axes = plt.gca()
  axes.set_xlim([-15,15])
  axes.set_ylim([0,80])
  plt.scatter(data.growth, data.enrollment)

  plt.xlabel("Average growth in GDP per Capita (%)")
  plt.ylabel("Average growth in tertiary enrollment (%)")
  plt.title("Average growth in GDP per capita" + "\n" + "and tertiary enrollment for
            + str(startYear) + " - " + str(endYear))
  plt.show()

plotYears(1970, 1980)
plotYears(1980, 1990)
plotYears(1990, 2000)
plotYears(2000, 2010)
```

Figures show the 10-year average growth in GDP per capita in percent versus tertiary enrollment in percent in the school leaving.

A few outliers have been excluded from these figures to ease comparison of the distributions between the 10-yea

Over time we see an increase in tertiary education enrollment (again, of those who have recently finished second

The first 10 year period (1970-1980) seems to indicate two groups. Countries with low tertiary enrollment have a countries with higher tertiary enrollment all have a more moderate spread. In later 10 year periods, this split seem