DAT405 Assignment 1 Shivneshwar Velayutham

November 27, 2022

1 Problem 1

1.1 Problem 1a)

Data source: Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2013) - “Life Expectancy”. Published online at OurWorldInData.org. Retrieved from: ‘https://ourworldindata.org/life-expectancy’ [Online Resource]

The above source has data for many years but for the purpose of this scatter plot the year **2018** has been chosen. The motivation for selecting 2018 is because a year needs to selected for us to draw a 2D scatter plot. The specific year 2018 was chosen randomly as just one of the years available in the dataset. There are no assumptions made.

Below is the python script to draw a scatter plot of life expectancy vs gdp per capita.

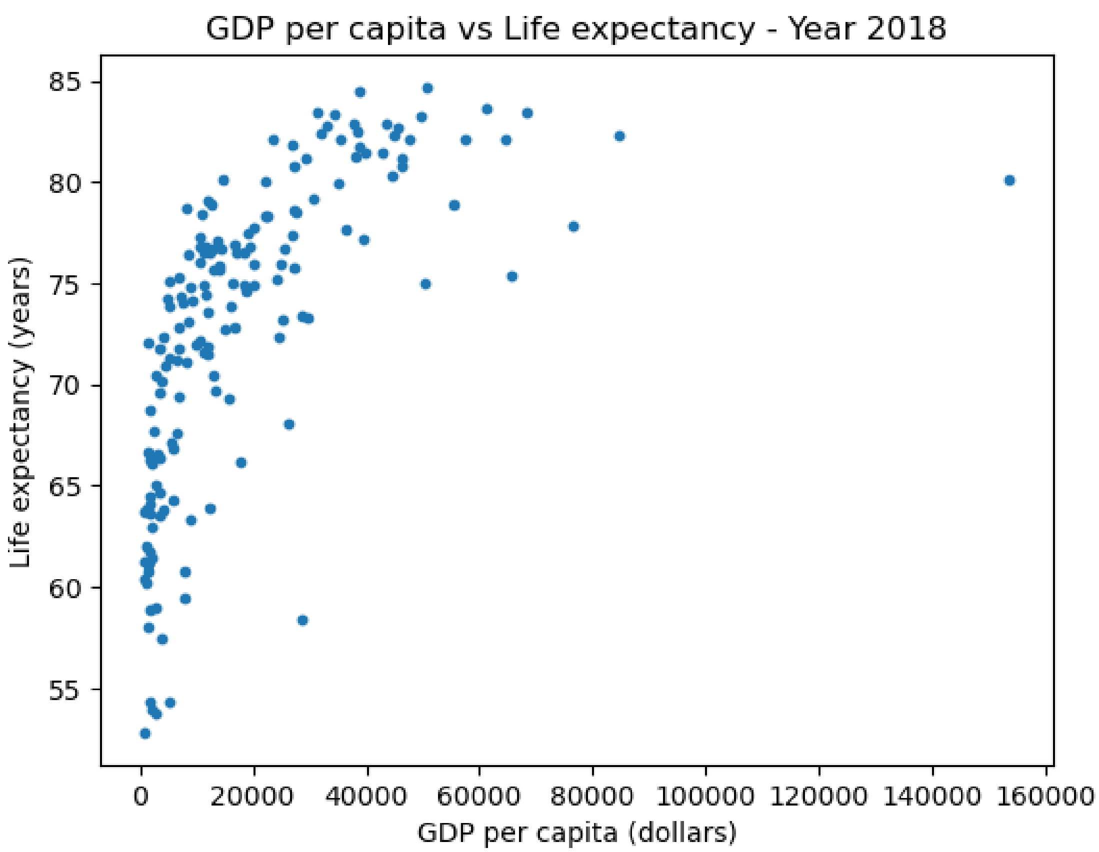
[118]:
```
import pandas
import matplotlib.pyplot as plt
```

[119]:
```
# Converting input csv to dataframe
df = pandas.read_csv("life-expectancy-vs-gdp-per-capita.csv")

# Removing rows from table where we have null values for the cols that need to␣
↪plotted
df.dropna(subset=['Life expectancy', 'GDP per capita'], inplace=True)

# Removing unused columns
df.drop(columns=['Code', '417485-annotations', 'Population (historical␣
↪estimates)', 'Continent'], inplace=True)

# Selecting the year 2018 for the scatter plot
df18 = df[df['Year'] == 2018].drop(columns=['Year'])

print("Checking for illegal countries")
entities_to_be_dropped = ['America', 'Antartica', 'Africa', 'Asia', 'Oceania',␣
↪'Europe', 'Zealandia', 'USSR', 'World']
for entity in entities_to_be_dropped:
    if df18['Entity'].str.contains(entity, case=False).any():
```

```
        print("Found while checking for " + entity + ": " +␣
↪str(df18[df18['Entity'].str.contains(entity, case=False)]['Entity'].
↪tolist()))

print("Only World is an illegal country so is dropped")
df18 = df18[df18['Entity'].str.contains('World', case=False) == False]
```

```
Checking for illegal countries
Found while checking for Africa: ['Central African Republic', 'South Africa']
Found while checking for World: ['World']
Only World is an illegal country so is dropped
```

[120]:
```
# Display graph
plt.scatter(df18['GDP per capita'], df18['Life expectancy'], s=10)
plt.title('GDP per capita vs Life expectancy - Year 2018')
plt.xlabel('GDP per capita (dollars)')
plt.ylabel('Life expectancy (years)')
plt.show()
```



1.2 Problem 1b)

Yes, it's reasonable that as GDP increases, the life expectancy increases. This is because access to good healthcare and medicines increases as the country's economy improves. A better economy often results in better social infrastructure (such as health and sanitary infrastructure). More access to money means better nourishment which is especially important for young children. It is obvious to see that there are some outliers that have high GDP but low life expectancy and vice versa as well. These will be discussed below.

1.3 Problem 1c)

Yes, I've cleaned the rows that had insufficient information to draw the scatter plot ie. when a row didn't have the GDP value or life expectancy value then that row is discarded. Also unused columns have been removed. As per comment, I have searched for plausible illegal countries, continents and the row for World have been removed.

[121]:
```
df18['GDP per capita'].describe()
```

[121]:
```
count       165.000000
mean      19053.786628
std       20346.341909
min         623.488892
25%        4440.381836
50%       12165.794922
75%       27370.554688
max      153764.171875
Name: GDP per capita, dtype: float64
```

[122]:
```
df18['Life expectancy'].describe()
```

[122]:
```
count    165.000000
mean      72.708642
std        7.748274
min       52.805000
25%       66.867000
50%       74.405000
75%       78.458000
max       84.687000
Name: Life expectancy, dtype: float64
```

1.4 Problem 1d)

The mean and the standard deviation for life expectancy can be seen above. The below are the countries whose life expectancy is higher than one standard deviation over the mean.

[123]:
```
df18[(df18['Life expectancy'] > 80.456916)]
```

[123]:
```
                Entity  Life expectancy  GDP per capita
3117         Australia           83.281    49830.800781
```

3354	Austria	81.434	42988.070312
5260	Belgium	81.468	39756.203125
9041	Canada	82.315	44868.742188
12649	Cyprus	80.828	27184.416016
13632	Denmark	80.784	46312.343750
17487	Finland	81.736	38896.699219
17870	France	82.541	38515.917969
19476	Germany	81.180	46177.617188
20073	Greece	82.072	23450.765625
22766	Hong Kong	84.687	50839.371094
23396	Iceland	82.855	43438.542969
24590	Ireland	82.103	64684.300781
24915	Israel	82.819	32954.769531
25253	Italy	83.352	34364.167969
26183	Japan	84.470	38673.808594
30819	Luxembourg	82.102	57427.500000
32359	Malta	82.376	32028.912109
36935	Netherlands	82.143	47474.109375
37722	New Zealand	82.145	35336.136719
39831	Norway	82.271	84580.132812
43346	Portugal	81.857	27035.599609
48185	Singapore	83.458	68402.343750
48775	Slovenia	81.172	29244.919922
50164	South Korea	82.846	37927.609375
50702	Spain	83.433	31496.519531
52014	Sweden	82.654	45541.890625
52634	Switzerland	83.630	61372.730469
57387	United Kingdom	81.236	38058.085938

1.5 Problem 1e)

The below filtering of data is used to figure out the countries that have high life expectancy but low GDP. As seen from the output below we can see **Cuba and Barbados** both have high life expectancy (greater than 78.458) but have low GDP (lesser than 12000 dollars per capita).

Motivation for choosing 78.458 as high life expectancy is since that is 75th percentile of the data and only 25% of data has a greater value than that.

Motivation for 12000 dollars per capita as low is because it's much lesser than the mean (19053.786628). Note: Please find the mean and percentile information above from the describe output.

[124]:
```
df18[(df18['Life expectancy'] > 78.458000)].sort_values(by=['GDP per capita']).
↪head(2)
```

[124]:

	Entity	Life expectancy	GDP per capita
12320	Cuba	78.726	8325.630859
4641	Barbados	79.081	11995.186523

1.6 Problem 1f)

The below filtering of data is used to figure out if there are countries that have high GDP but low life expectancy. As seen from the output below we can see **Equatorial Guinea and Turkmenistan** both have high GDP per capita (greater than 26000) but have low life expectancy (lesser than 68.1). So the answer is **No**, not every strong economy has high life expectancy.

Motivation for choosing 68.1 as low life expectancy is since it's much lesser than the mean (72.708642) and close to the 25th percentile which is 66.867000.

Motivation for 26000 dollars per capita as high is because it's much higher than the mean (19053.786628) and close to the 75th percentile which is 27370.554688.

Note: Please find the mean and percentile information above from the describe output.

[125]:
```
df18[(df18['Life expectancy'] < 68.1)].sort_values(by=['GDP per capita'],
    ascending=False).head(2)
```

[125]:
```
                  Entity  Life expectancy  GDP per capita
15373  Equatorial Guinea           58.402    28528.953125
55724       Turkmenistan           68.073    26318.365234
```

1.7 Problem 1g)

If we use GDP as a strong indicator of economy then this means that a country with a good economy need not necessarily result in it's citizens having a high life expectancy. Sometimes improvement of economy through industrialization might cause pollution in a variety of ways and actually end up hurting the overall health of the population. To improve the life expectancy of its citizens, a country must not just look to improve it's economy but to try to understand the other variables that might be related to life expectancy.

The main difference between GDP and GDP per capita is that per capita means per person. So it's possible for there to be countries where wealth disparity is quite high and there's a small subset of people who have a lot of money but majority still live in poverty. So GDP per capita might not be a appropriate metric to measure overall wealth/economy of a country.

2 Problem 2

2.1 Problem 2a)

Hannah Ritchie and Max Roser (2013) - "Indoor Air Pollution". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/indoor-air-pollution' [Online Resource]

Hannah Ritchie and Max Roser (2019) - "Outdoor Air Pollution". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/outdoor-air-pollution' [Online Resource]

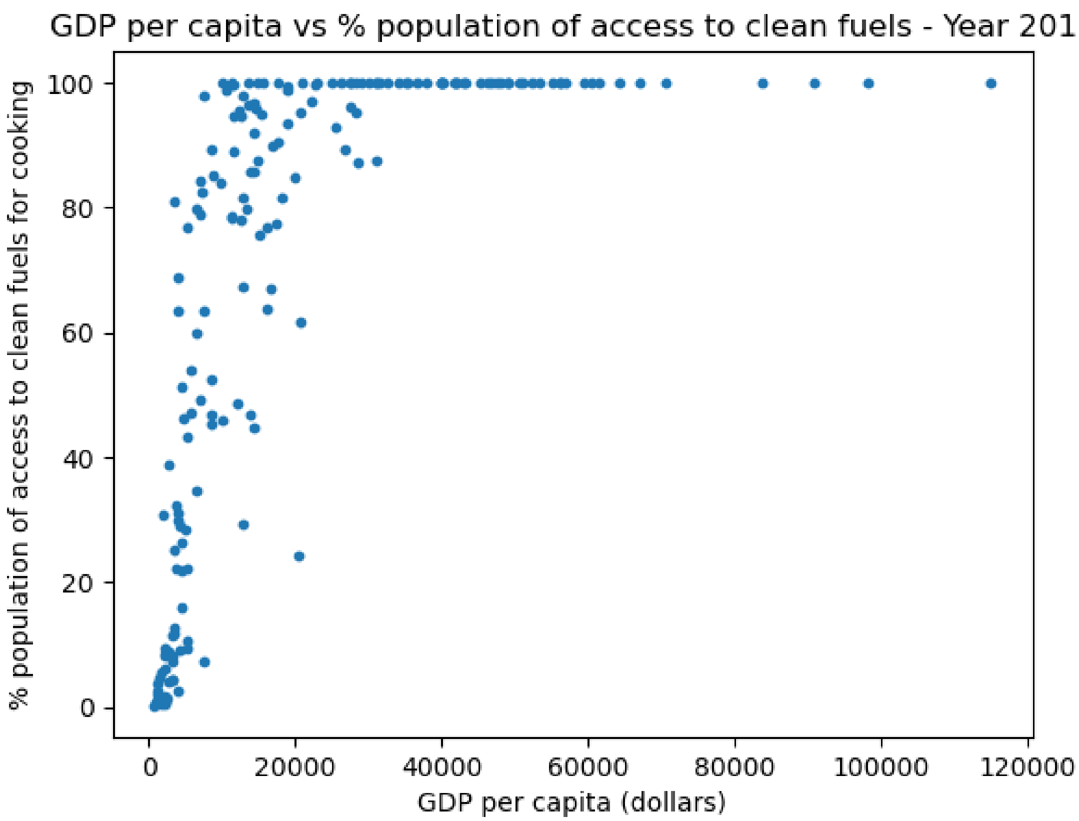
Questions: 1. How does GDP per capita relate to access to clean fuels? 2. How does access to clean fuels relate to the death rate due to indoor pollution? 3. How does GDP per capita relate to death rate due to indoor pollution? 4. How does GDP per capita relate to death rate due to outdoor pollution?

Below is the python script used to visualize the same.

No assumptions made. The motivation for selecting 2018 since a year needs to selected for us to draw a 2D scatter plot. The year 2018 has been chosen for both scatter plots since it was the year used in the previous scatter plots to maintain consistency.

[126]:
```
# Converting input csv to dataframe
df = pandas.read_csv("access-to-clean-fuels-for-cooking-vs-gdp-per-capita.csv")

# Renaming columns for ease of use
df.rename(columns={'Indicator:Proportion of population with primary reliance on␣
↪clean fuels and technologies for cooking (%) - Residence Area Type:Total':
                   'Access to clean fuels for cooking',
           'GDP per capita, PPP (constant 2017 international $)': 'GDP per␣
↪capita'}, inplace=True)

# Removing rows from table where we have null values for the cols that need to␣
↪plotted
df.dropna(subset=['Access to clean fuels for cooking', 'GDP per capita'],␣
↪inplace=True)

# Removing unused columns
df.drop(columns=['Code', 'Population (historical estimates)', 'Continent'],␣
↪inplace=True)

# Selecting the year
df18 = df[df['Year'] == 2018].drop(columns=['Year'])
```

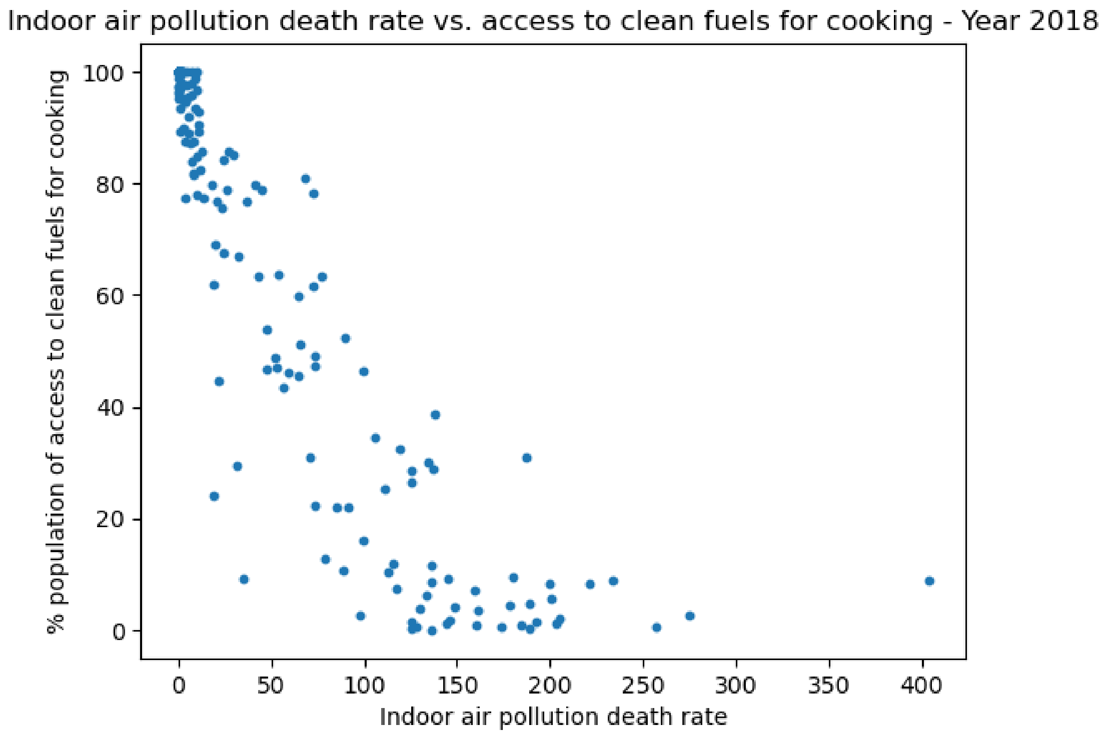
[127]:
```
# Display graph
plt.scatter(df18['GDP per capita'], df18['Access to clean fuels for cooking'],␣
↪s=10)
plt.title('GDP per capita vs % population of access to clean fuels - Year 2018')
plt.xlabel('GDP per capita (dollars)')
plt.ylabel('% population of access to clean fuels for cooking')
plt.show()
```



[128]:
```
# Converting input csv to dataframe
df = pandas.read_csv("indoor-pollution-death-rates-clean-fuels.csv")

# Renaming columns for ease of use
df.rename(columns={'Indicator:Proportion of population with primary reliance on␣
↪clean fuels and technologies for cooking (%) - Residence Area Type:Total':
                 'Access to clean fuels for cooking',
           'Deaths - Cause: All causes - Risk: Household air pollution from␣
↪solid fuels - Sex: Both - Age: Age-standardized (Rate)': 'Deaths'},␣
↪inplace=True)

# Removing rows from table where we have null values for the cols that need to␣
↪plotted
df.dropna(subset=['Access to clean fuels for cooking', 'Deaths'], inplace=True)

# Removing unused columns
df.drop(columns=['Code', 'Population (historical estimates)', 'Continent'],␣
↪inplace=True)

# Selecting the year
```

```
df18 = df[df['Year'] == 2018].drop(columns=['Year'])
```

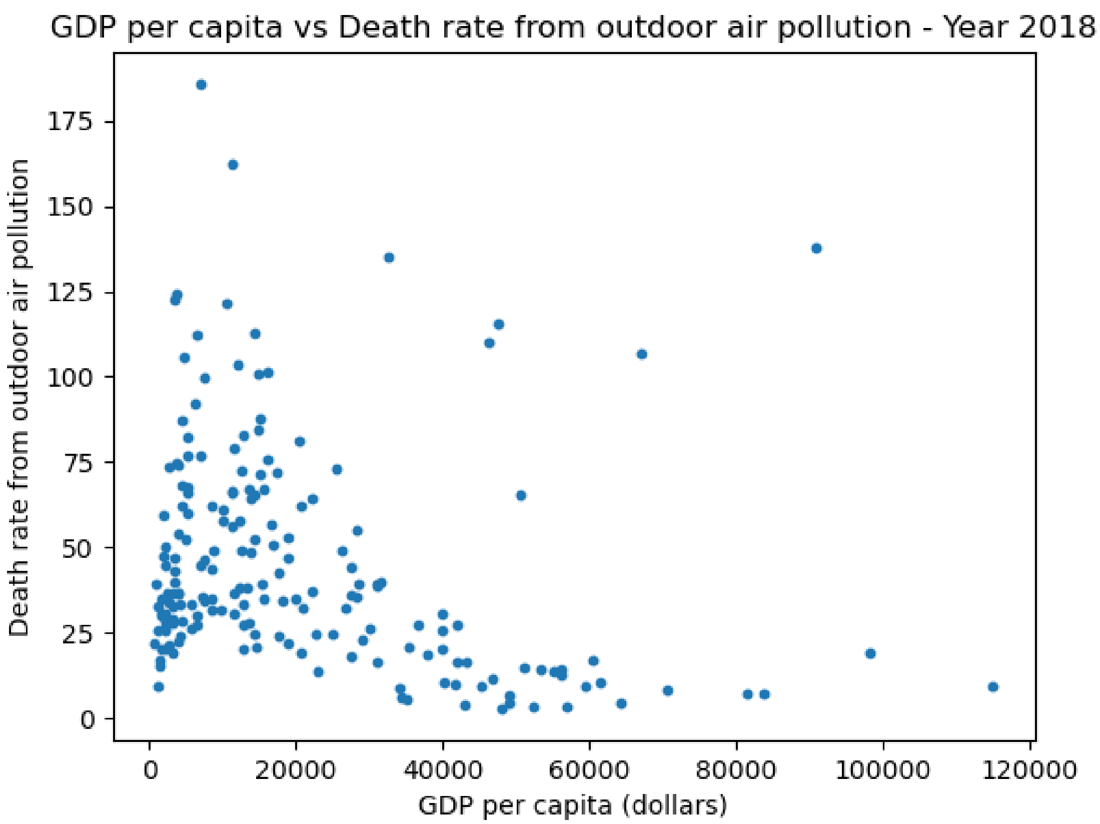
[129]:
```
# Display graph
plt.scatter(df18['Deaths'], df18['Access to clean fuels for cooking'], s=10)
plt.title('Indoor air pollution death rate vs. access to clean fuels for␣
 ↪cooking - Year 2018')
plt.xlabel('Indoor air pollution death rate')
plt.ylabel('% population of access to clean fuels for cooking')
plt.show()
```



[130]:
```
# Converting input csv to dataframe
df = pandas.read_csv("outdoor-pollution-rate-vs-gdp.csv")

# Renaming columns for ease of use
df.rename(columns={'GDP per capita, PPP (constant 2017 international $)': 'GDP␣
 ↪per capita',
           'Deaths - Cause: All causes - Risk: Outdoor air pollution - OWID -␣
 ↪Sex: Both - Age: Age-standardized (Rate)': 'Deaths'}, inplace=True)

# Removing rows from table where we have null values for the cols that need to␣
 ↪plotted
df.dropna(subset=['GDP per capita', 'Deaths'], inplace=True)
```

```
# Removing unused columns
df.drop(columns=['Code', 'Population (historical estimates)', 'Continent'],␣
 ↪inplace=True)

# Selecting the year
df18 = df[df['Year'] == 2018].drop(columns=['Year'])
```

[131]:
```
# Display graph
plt.scatter(df18['GDP per capita'], df18['Deaths'], s=10)
plt.title('GDP per capita vs Death rate from outdoor air pollution - Year 2018')
plt.xlabel('GDP per capita (dollars)')
plt.ylabel('Death rate from outdoor air pollution')
plt.show()
```



2.2 Problem 2b)

It's clear to see that as the GDP per capita increases, access to clean fuels for cooking improves in it's country. There are of course some outliers where some countries that don't have a small GDP per capita have a small percentage of population who have acess to clean fuels. It's also clear that as access to clean fuels improves, the death rate due to indoor air pollution reduces. Therefore we can say that as the GDP per capita of a country increases, the death rate due to indoor air pollution reduces.

The same cannot be said about outdoor pollution, I can see an upside down U shape curve where the countries that medium GDP per capita have high number of deaths due to outdoor pollution and the low and high GDP countries have lower number of deaths due to outdoor air pollution.