BANA 680 Final Exam

Tayo Obafaiye

Introduction

In this exam project, we are trying to establish a relationship between Cancer Incidence Rate and Death Rate. Intuitively, one would expect that the higher the rate of incidence, the higher the death rate and vice versa. The challenge in this project is that we would be making use of data from three different datasets.

We would be analyzing this data at state level thus we would be able to answer questions such as which State is most/least prone to cancer. We can also determine the State with the most/least death rate due to cancer.

Analysis of Data

The first dataset (NCHS_-_Leading_Causes_of_Death__United_States.csv) contains data about the number of deaths caused by different ailments for each year between the years 2000 to 2016. It also contains the corresponding Age-adjusted Death Rate.

The second dataset (nst-est2018-01.xlsx) contains population information between the years 2010 and 2018. The population estimates are given in this dataset at national, regional and state level.

The additional data (cancerdeaths.csv) used in this project contains county-level information of metrics relating to cancer deaths for almost all counties in the U.S. The metric of concern for this project is the Cancer Incidence Rate. Although the data is gathered at county level, the row entries are sorted by zip-codes and because there are usually multiple zip-codes within a county, there arises an issue of duplicate values for the variables in this dataset including Incidence Rate. The population estimate used in this dataset is from 2015. However, in this project we would use the 2015 population estimate obtained from the 'nst-est2018-01.xlsx' file.

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

We subset our first dataframe to create a new dataframe for deaths caused by cancer in the year 2015. Because the original data is sorted by states, no further cleaning is necessary. However we drop the '113 Cause Name' & 'Age-adjusted Death Rate' columns and rename 'State' as 'State_Name' ahead of merging the datasets.

In [2]:

```
df1 = pd.read_csv("C:/Users/Tayo Obafaiye/Desktop/BANA680/BANA 680 Data/NCHS_-_Leading_
Causes_of_Death__United_States.csv")
df1 = df1[(df1['Cause Name'] == 'Cancer') & (df1['Year'] == 2015)]
df1 = df1.rename({'State': 'State_Name', 'Cause Name': 'Cause_Name'}, axis=1)
df1 = df1[['Year','Cause_Name', 'State_Name', 'Deaths']]
df1.head(1)
```

Out[2]:

	Year	Cause_Name	State_Name	Deaths
7498	2015	Cancer	Utah	3091

From the second dataset, we obtain the 2015 population estimate for each state in the U.S. We create a new dataframe with one column made up of US States and the other having the 2015 population estimate for each corresponding state. This new dataframe has the population estimate for all 50 states, DC and Puerto Rico.

In [3]:

```
df2 = pd.read_excel("C:/Users/Tayo Obafaiye/Desktop/BANA680/BANA 680 Data/nst-est2018-0
1.xlsx", usecols = [0, 8], skiprows = [0, 1, 2, 4, 5, 6, 7, 8])
df2 = df2.dropna()
df2 = df2.rename({'Unnamed: 0': 'State_Name', 2015: '2015_Population_Estimate'}, axis=1
)
df2['State_Name'] = df2['State_Name'].str.strip('.')
df2.head(1)
```

Out[3]:

	State_Name	2015_Population_Estimate
0	Alabama	4853160.0

The cancer incidence rate like other variables in this dataset is obtained at county-level however the dataset is sorted by zip codes. Hence there is the issue of duplicate incidence rates. We drop duplicate incidence rates by using the drop_duplicates function keeping the the first row entry for each unique county code.

This dataset does contain the 2015 population estimate for each county in the United States. However rather than getting the state population from sums of the county population, we would make use of the population estimate derived from the second dataset.

This dataset throws up an interesting challenge in that the States are listed by their two letter codes. I went about solving this by importing a US State abbreviation dictionary from https://gist.github.com/rogerallen/1583593 (https://gist.github.com/rogerallen/1583593) and creating a new 'State_Name' column by mapping the 'State' column using the imported dictionary.

We use the aggregate function to calculate the sum and average incidence rate of cancer, grouping our results by State.

In [4]:

```
df3 = pd.read_csv("C:/Users/Tayo Obafaiye/Desktop/BANA680/BANA 680 Data/cancerdeaths.cs
v")
df3.head(2)
```

Out[4]:

	zipCode	countyCode	studyCount	State	PovertyEst	povertyPercent	medIncome	Nar
0	1001	25013	0	MA	80178	17.7	49072	Hampd Cour
1	1008	25013	0	MA	80178	17.7	49072	Hampd Cour

In [5]:

```
df3 = df3.drop_duplicates(subset=['countyCode'], keep='first')
us_state_abbrev = {'Alabama': 'AL','Alaska': 'AK','American Samoa': 'AS','Arizona': 'A
Z','Arkansas': 'AR','California': 'CA','Colorado': 'CO','Connecticut': 'CT','Delaware':
'DE','District of Columbia': 'DC','Florida': 'FL','Georgia': 'GA','Guam': 'GU','Hawaii'
: 'HI','Idaho': 'ID','Illinois': 'IL','Indiana': 'IN','Iowa': 'IA','Kansas': 'KS','Kent
ucky': 'KY','Louisiana': 'LA','Maine': 'ME','Maryland': 'MD','Massachusetts': 'MA','Mic
higan': 'MI','Minnesota': 'MN','Mississippi': 'MS','Missouri': 'MO','Montana': 'MT','Ne
braska': 'NE','Nevada': 'NV','New Hampshire': 'NH','New Jersey': 'NJ','New Mexico': 'N
M','New York': 'NY','North Carolina': 'NC','North Dakota': 'ND','Northern Mariana Islan
ds':'MP','Ohio': 'OH','Oklahoma': 'OK','Oregon': 'OR','Pennsylvania': 'PA','Puerto Ric
o': 'PR','Rhode Island': 'RI','South Carolina': 'SC','South Dakota': 'SD','Tennessee':
'TN','Texas': 'TX','Utah': 'UT','Vermont': 'VT','Virgin Islands': 'VI','Virginia': 'VA'
,'Washington': 'WA','West Virginia': 'WV','Wisconsin': 'WI','Wyoming': 'WY'}
abbrev_us_state = dict(map(reversed, us_state_abbrev.items()))
df3['State_Name'] = df3['State'].map(abbrev_us_state).fillna(df3['State'])
df4 = df3.groupby('State_Name')[['incidenceRate']].agg(['sum', 'mean'])
df4.head(1)
```

Out[5]:

	incidenceRate	
	sum	mean
State_Name		
Alabama	30952.6	461.979104

Merging datasets

All three dataframes are merged on the unique key 'State_Name' in an inner join. The first two dataframes are merged and in turn the new combined dataframe (dfA) is merged with the third dataframe (df4).

In [6]:

```
dfA = pd.merge(df1, df2, on='State_Name', how='inner')
finaldf = pd.merge(dfA, df4, on='State_Name', how='inner')
finaldf['Cancer_Deaths_Per_Population'] = finaldf.Deaths/finaldf['2015_Population_Estimate']
finaldf.columns = ['Year', 'Cause_Name', 'State_Name', 'Deaths', '2015_Population_Estimate', 'Total_Incidence_Rate', 'Avg_Incidence_Rate', 'Cancer_Deaths_Per_Population']
finaldf.head(2)
```

```
C:\Users\Tayo Obafaiye\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:618: UserWarning: merging between different levels can give an unintended result (1 levels on the left, 2 on the right)
  warnings.warn(msg, UserWarning)
```

Out[6]:

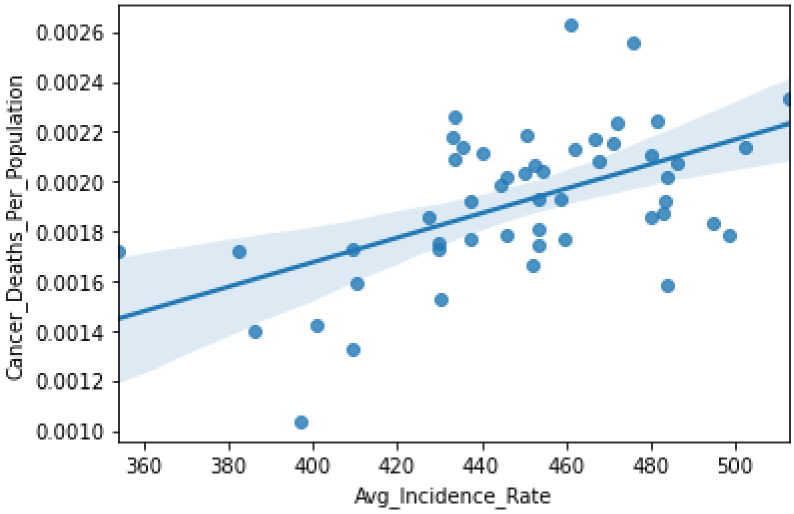
	Year	Cause_Name	State_Name	Deaths	2015_Population_Estimate	Total_Incidence_Rate
0	2015	Cancer	Utah	3091	2982497.0	10713.9
1	2015	Cancer	Colorado	7604	5452107.0	23154.7

Visualization

The plot below is a scatter plot of the mean incidence rate vs. death rate with each individual point representing the intersecting value of each of the 50 states and District of Columbia (DC). The plot includes a trend line or regression line. The trend line has a positive slope suggesting that as the average incidnce rate increases, the deaths due to cancer per population increases as well.

In [7]:

```
fig, ax = plt.subplots()
sns.regplot(x='Avg_Incidence_Rate',y='Cancer_Deaths_Per_Population', data=finaldf, ax=ax)
plt.show()
```



Follow-up Questions

1. Which State is most/least prone to cancer? Kentucky has the highest Average Incidence Rate while Arizona has the least.
2. Which State has the most/least deaths due to cancer? West Virginia has the highest rate of deaths due to cancer, while Utah has the least.

In [8]:

```
print(finaldf['State_Name'][finaldf['Avg_Incidence_Rate'] == finaldf['Avg_Incidence_Rate'].max()])
print(finaldf['State_Name'][finaldf['Avg_Incidence_Rate'] == finaldf['Avg_Incidence_Rate'].min()])
print(finaldf["State_Name"][finaldf['Cancer_Deaths_Per_Population'] == finaldf['Cancer_Deaths_Per_Population'].max()])
print(finaldf["State_Name"][finaldf['Cancer_Deaths_Per_Population'] == finaldf['Cancer_Deaths_Per_Population'].min()])
```

```
50    Kentucky
Name: State_Name, dtype: object
4    Arizona
Name: State_Name, dtype: object
49    West Virginia
Name: State_Name, dtype: object
0    Utah
Name: State_Name, dtype: object
```

Conclusion

It is important to note that we normalized our Incidence Rate and Cancer Deaths variables before comparing them among states. We used the mean of cancer incidence rates and the cancer deaths per population respectively. Also to note that the data used is based on or around the year 2015, it might not reflect the true picture of things today. I suspect due to the covid-19 pandemic, other medical services like cancer screenings might have been forced to take a back seat. In all we can conclude that the data supports our earlier intuition that the higher the average cancer incidence rate, the higher the cancer death rate and vice versa.