BANA680.01 Final Exam

By Dana DiVincenzo

Introduction

In this assignment, I was tasked with using data sets from a previous assignment, along with a new dataset of my choosing, to create and answer a new and significant question that I create. The purpose of finding the data and creating the question myself is to demonstrate my understanding in what can be accomplished with the right datasets, and to show my Pandas, Matplotlib, and Jupyter Notebook skills. The data sets from the previous assignment held information about population data in the United States from 2010 to 2018, and about causes of death for Americans from 1998 to 2016. I used those to answer four questions about death trends in the US and the leading causes of death in the country and in individual states.

To find a new dataset to help me create and answer a new question, I went to the United States Census Bureau website and found a dataset by American Community Survey (TableID: B02001). The dataset showed the population of different races in each state in the year 2016, which is the most recent year shared by the two previous datasets. With this dataset, I wanted to see if I could find a relationship between race and death rate. I hypothesized that states with a greater population percentage of white people would have a lower death rate, because people of color may face systematic challenges to proper health care access that white people do not.

My Question: Do states with a higher percentage of white residents have higher, lower, or unrelated death rates in 2016?

Step 1: Cleaning the data

In the following section, I read in all three data frames. I then remove any unnecessary information from each dataframe and make adjustments such as transposing a dataframe or sorting by state so they can be properly merged together.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [2]:
```
# read in files
file = 'C:/Users/Owner/Downloads/USPop.csv'
popdf_full = pd.read_csv(file)

file = 'C:/Users/Owner/Downloads/Deaths.csv'
deathsdf = pd.read_csv(file)

# this is a file I downloaded from the US census website
# it shows the population per race for each state
file = 'C:/Users/Owner/Downloads/CensusRace.csv'
racedf = pd.read_csv(file)
```

In [3]:
```
# drops an empty column and renames the first column to ""
updatedf = racedf.drop(10)
updatedf.rename(columns={ updatedf.columns[0]: "" }, inplace = True)
```

In [4]:
```
# transposes the df
racedf = updatedf.transpose()

# moves the first row to the header
new_header = racedf.iloc[0]
racedf = racedf[1:]
racedf.columns = new_header

# re-formats the df and fixes the column names. Then sorts by State.
modified = racedf.reset_index()
modified.rename(columns={ modified.columns[0]: "State" }, inplace = True)
modified.rename(columns={ modified.columns[1]: "Total" }, inplace = True)
modified.rename(columns={ modified.columns[2]: "White" }, inplace = True)
modified = modified.sort_values("State")

# drops Puerto Rico and resets the index
modified = modified[modified["State"] != "Puerto Rico"]
modified.reset_index(drop=True, inplace=True)

modified = modified[["State", "White"]]
# I dropped population because another data frame already had population and I
can't use both
# the values were very similar - both from US Census website from 2016
```

In [5]:
```
# gets data frame where state is not "United States", cause name is "All cause
s",
# this data frame only displays the year, state, and deaths columns
totaldeathsdf = deathsdf.loc[(deathsdf["State"] != "United States") & \
                             (deathsdf["Cause Name"] == "All causes"), \
                            ["Year", "State", "Deaths"]]

# takes same data frame where year is only 2016
totaldeathsdf = totaldeathsdf.loc[(totaldeathsdf["Year"] == 2016), \
                            ["State", "Deaths"]]

# sorts data frame by year and resets the index
totaldeathsdf = totaldeathsdf.sort_values("State")
totaldeathsdf.reset_index(drop=True, inplace=True)
```

In [6]:
```
# removes the geographic areas from the dataframe which are not states
popdf = popdf_full.copy(deep=True)
bad_areas = ["United States", "Northeast", "Midwest", "South", "West", "Puerto
Rico"]
popdf = popdf[~popdf['Geographic Area'].isin(bad_areas)]
popdf.reset_index(drop=True, inplace=True)

# correctly formats the state names
popdf["Geographic Area"] = popdf["Geographic Area"].str.replace('[^\w\s]','')

# takes only the geographic area and 2016 columns and renames them
popdf = popdf[["Geographic Area", "2016"]]
popdf.rename(columns={ "Geographic Area" : "State", "2016" : "2016_Pop" }, inp
lace = True)
```

Step 2: Merging the data frames and adding additional columns

In the following section, I merge all three data frames together. I then use the columns containing the number of deaths per state, the number of white people per state, and the population per state, to create new columns which tell me the death rate and percent white per state.

In [13]:
```
# merges all three data frames
combine = pd.merge(popdf, totaldeathsdf, on=['State'], how='outer')
combinedf = pd.merge(combine, modified, on=['State'], how='outer')

# removed commas from the numbers in the data frame
combinedf = combinedf.replace(",", "", regex=True)

# changes types of 2016_Pop and of White to int
combinedf["2016_Pop"]=combinedf["2016_Pop"].astype(int)
combinedf["White"]=combinedf["White"].astype(int)
```

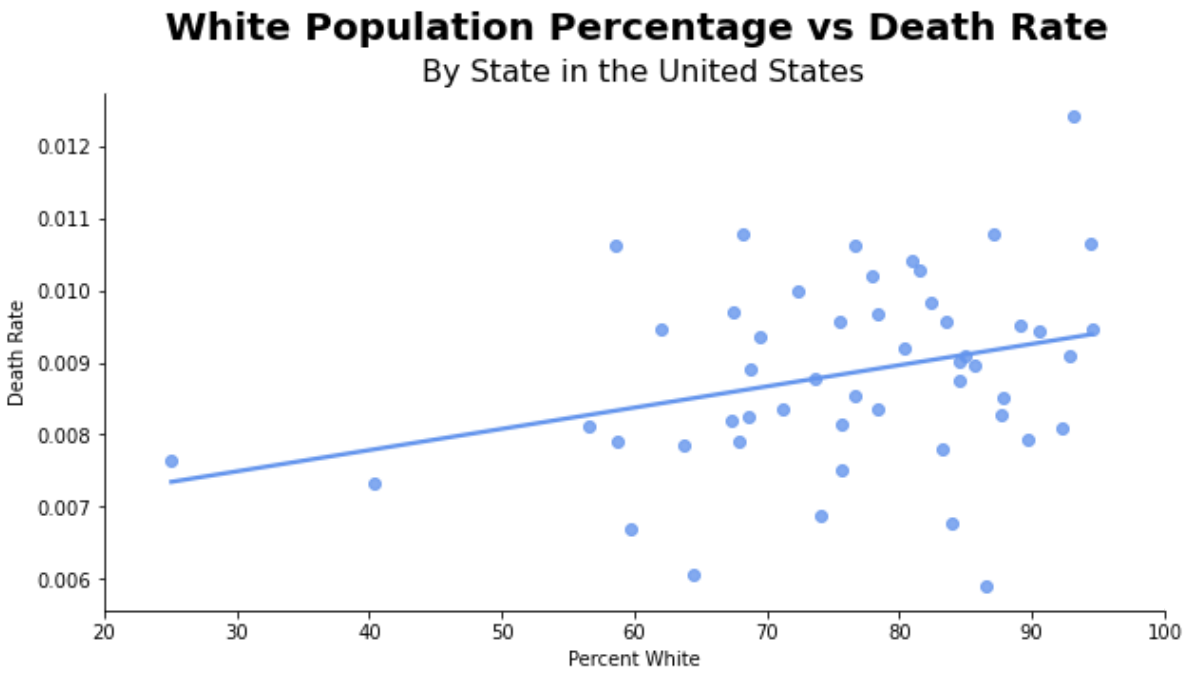
In [14]:
```
# calculates the death rate by dividing the number of deaths in 2016 by the po
pulation in 2016
combinedf["Death_Rate"] = (combinedf["Deaths"] / combinedf["2016_Pop"])
```

In [15]:
```
# calculates the percent white of each state in 2016
# by dividing the number of white people with the total population
combinedf["Percent_White"] = (combinedf["White"] / combinedf["2016_Pop"]) * 10
0
```

Step 3: Creating a graph

In the following section, I create a scatterplot which shows the white population percentage vs the death rate by state in 2016. The line of best fit shows whether the trend is positive, negative, or unrelated.

In [16]:
```
# creates a plot
plt.figure(figsize=(10,5))
ax = plt.subplot()

# adds title and subtitle
plt.suptitle("White Population Percentage vs Death Rate", fontweight = "bold",
fontsize = 20, ha = "left", x = 0.17, y = 1)
plt.title("By State in the United States", fontsize = 16, ha = "left", x = .3)

# adds line of best fit
sns.regplot(x= "Percent_White",
            y= "Death_Rate",
            ci=None,
            data=combinedf,
            color = "cornflowerblue");

# adds x and y labels
plt.xlabel('Percent White', fontsize=10)
plt.ylabel('Death Rate', fontsize=10)

# shows x axis from 20 to 100
plt.xlim([20, 100])

# removes top and right spines of graph
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)
plt.show()
```



Step 4: Intepreting the graph

The graph above shows us that as the white population percentage increases, the death rate also increases. Therefore, they have a positive relationship. This contradicts the original hypothesis that assumed states with higher percentages of white people would have lower death rates.

Conclusions

My visualization showed that the relationship between white population percentage and death rate is positive. I was surprised by this, as white people do not have the lowest life expectancy of all races in the United States. In 2014, the CDC reported that White Americans had the second highest life expectency (78.8), when compared to Hispanic Americans (81.8) and African Americans (75.2). Additionally, differences in quality of medical care betewen white Americans and non-white Americans is well documented. According to Williams and Rucker in an article in Health Care Financing Review, "Compared with white persons, black persons and other minorities have lower levels of access to medical care in the United States due to their higher rates of unemployment and under-representation in good-paying jobs that include health insurance as part of the benefit package."

Although the results seem counter-intuitive, the dataset I found does not tell us how these people are dying or at what age, so it is not complete. To get a better picture of how race effects death rate these factors should be included. Other interesting things to include would be health insurance, proximity to medical care, and salary.

Overall, I feel that this assignment was a success and I was able to pose and answer a significant question while demonstrating my Pandas, Matplotlib, and Jupyter Notebook skills.

Works Cited

Arias E. Changes in life expectancy by race and Hispanic origin in the United States, 2013–2014. NCHS data brief, no 244. Hyattsville, MD: National Center for Health Statistics. 2016.

Williams, D R, and T D Rucker. “Understanding and addressing racial disparities in health care.” Health care financing review vol. 21,4 (2000): 75-90.