US Heart Disease Analysis for 2014

Mary Alice Salazar

Kaushik Sanaddar

Prasanna Jaiswal

DG OMG BRO

Motivation & Summary

We looked at heart disease because it is the number one cause of death in the United States.

According to the CDC:

- About **610,000 people** die of heart disease in the United States every year- that's **1 in every 4 deaths**
- Heart disease is the leading cause of death for both men and women. **More than half** of the deaths due to heart disease in 2009 were in men

https://www.cdc.gov/heartdisease/facts.htm

We found that there is a high number of occurrence of heart disease in every state. We drilled down on gender, poverty, obesity and risk factors such as smoking, alcohol consumption, high school obesity, and adult obesity. We also found a strong correlation between the risk factors and data values.

Data Sources and Questions

- Data sources:
  - Data.gov: Heart_Disease_Mortality_Data 2014
  - CDC: Heart Disease Death Rates (2014)
  - US census 2014

  - Questions :
  - What is the rate of heart disease among the US population?
  - Is there a difference in occurrence for gender, poverty, and other groups?
  - What are possible causes? i.e. smoking, alcohol, adult obesity, high school obesity?

These data sources contained data for each of these questions.

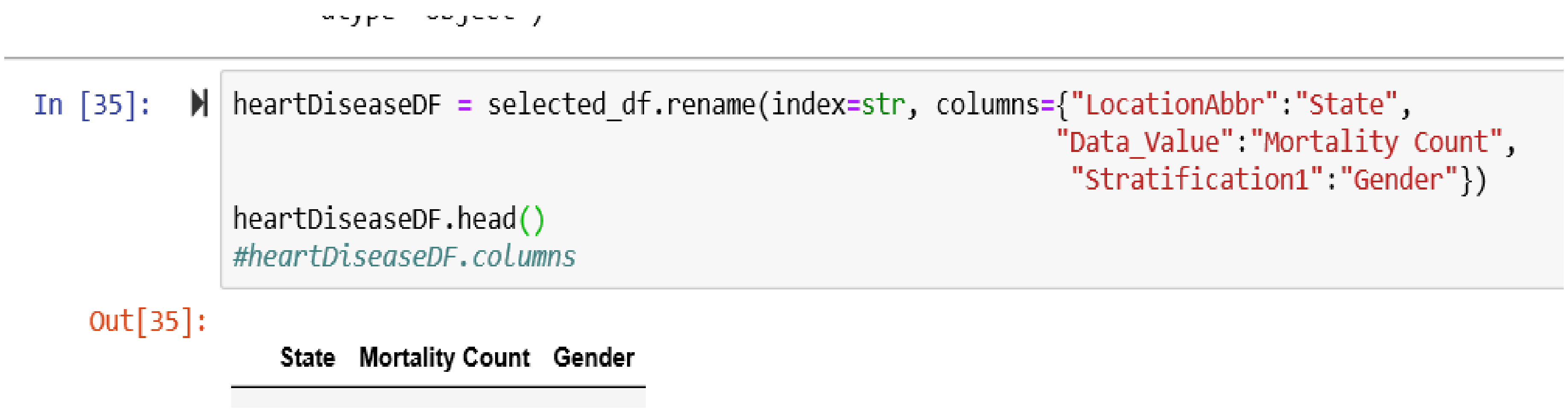
From the first, we gained information about gender

From the second data source, we gained data for mortality and risk factors.

From the census, we found poverty data.

Data cleanup and exploration

We did minor cleanup on the data. There were some renamed columns, and we sliced the csv columns. For example, different data sources had different nomenclature for the states. We picked good sources.



```
dtype='object')
```

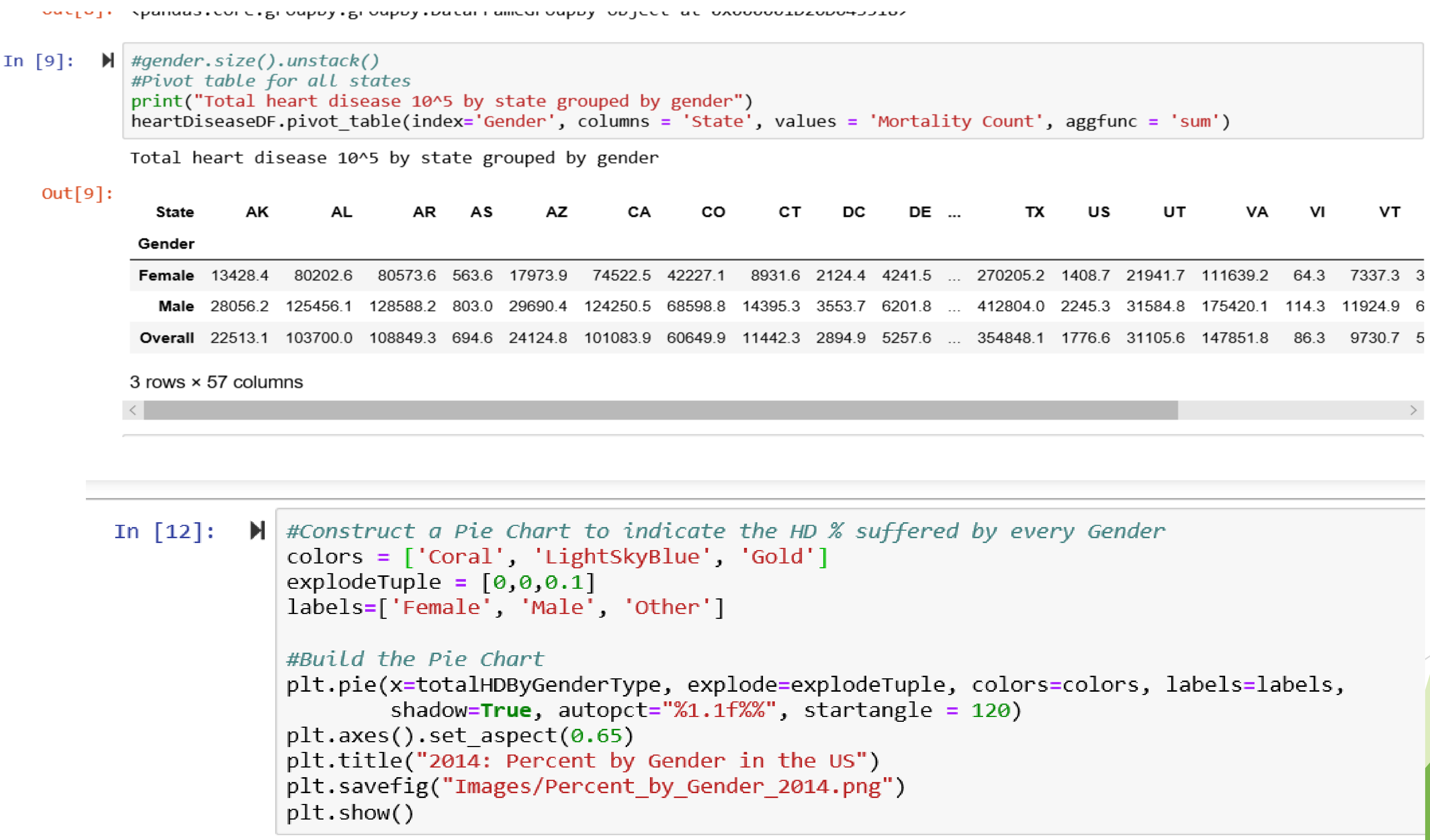
```
In [35]:  heartDiseaseDF = selected_df.rename(index=str, columns={"LocationAbbr":"State",
                                                             "Data_Value":"Mortality Count",
                                                              "Stratification1":"Gender"})

          heartDiseaseDF.head()
          #heartDiseaseDF.columns
```

Out[35]:

State	Mortality Count	Gender

Pivot table and pie: gender



```
Out[8]:  <pandas.core.groupby.groupby.DataFrameGroupBy object at 0x000001D26D645918>
```

```
In [9]:  #gender.size().unstack()
         #Pivot table for all states
         print("Total heart disease 10^5 by state grouped by gender")
         heartDiseaseDF.pivot_table(index='Gender', columns = 'State', values = 'Mortality Count', aggfunc = 'sum')
```

```
Total heart disease 10^5 by state grouped by gender
```

Out[9]:

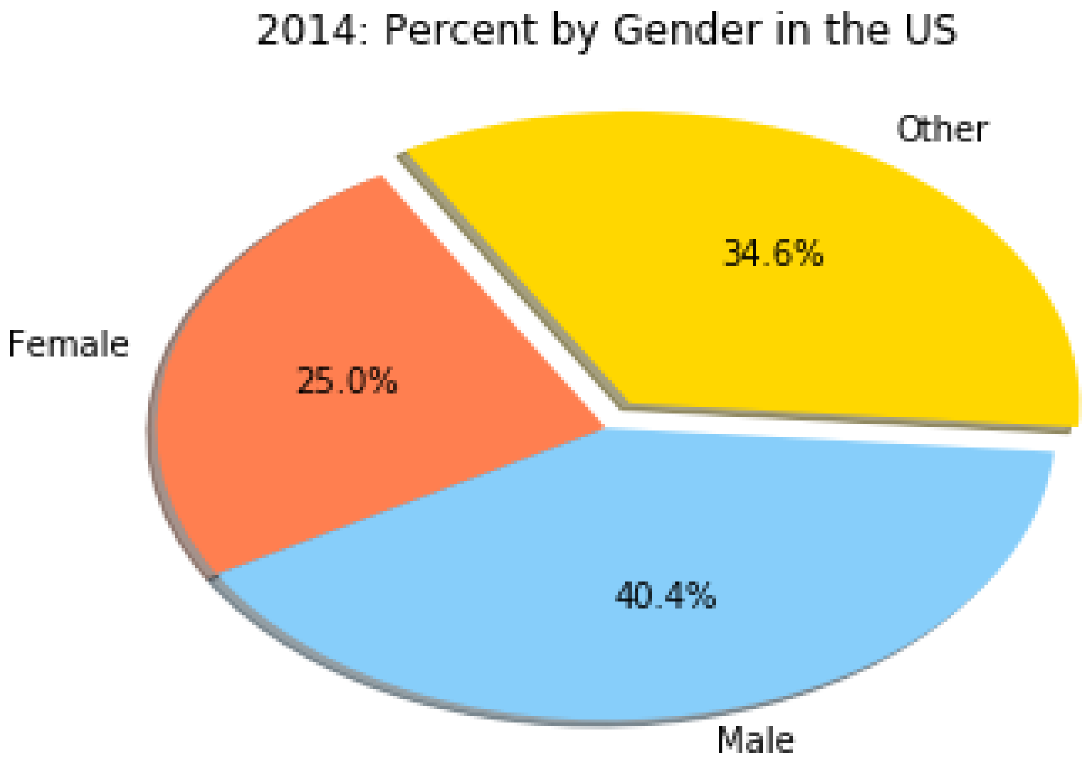
State	AK	AL	AR	AS	AZ	CA	CO	CT	DC	DE	...	TX	US	UT	VA	VI	VT	
Gender																		
Female	13428.4	80202.6	80573.6	563.6	17973.9	74522.5	42227.1	8931.6	2124.4	4241.5	...	270205.2	1408.7	21941.7	111639.2	64.3	7337.3	3
Male	28056.2	125456.1	128588.2	803.0	29690.4	124250.5	68598.8	14395.3	3553.7	6201.8	...	412804.0	2245.3	31584.8	175420.1	114.3	11924.9	6
Overall	22513.1	103700.0	108849.3	694.6	24124.8	101083.9	60649.9	11442.3	2894.9	5257.6	...	354848.1	1776.6	31105.6	147851.8	86.3	9730.7	5

3 rows × 57 columns

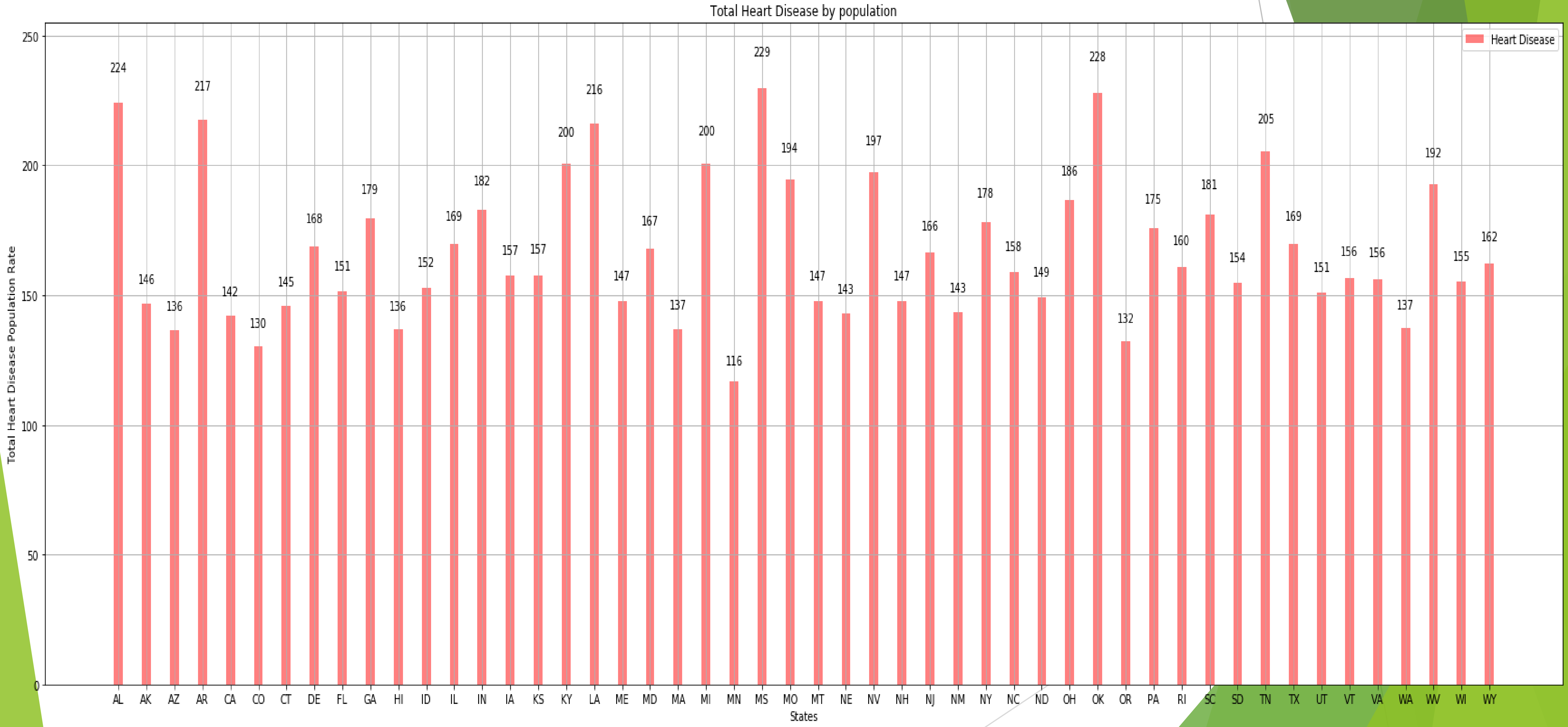
```
In [12]:  #Construct a Pie Chart to indicate the HD % suffered by every Gender
          colors = ['Coral', 'LightSkyBlue', 'Gold']
          explodeTuple = [0,0,0.1]
          labels=['Female', 'Male', 'Other']

          #Build the Pie Chart
          plt.pie(x=totalHDByGenderType, explode=explodeTuple, colors=colors, labels=labels,
                  shadow=True, autopct="%1.1f%%", startangle = 120)
          plt.axes().set_aspect(0.65)
          plt.title("2014: Percent by Gender in the US")
          plt.savefig("Images/Percent_by_Gender_2014.png")
          plt.show()
```

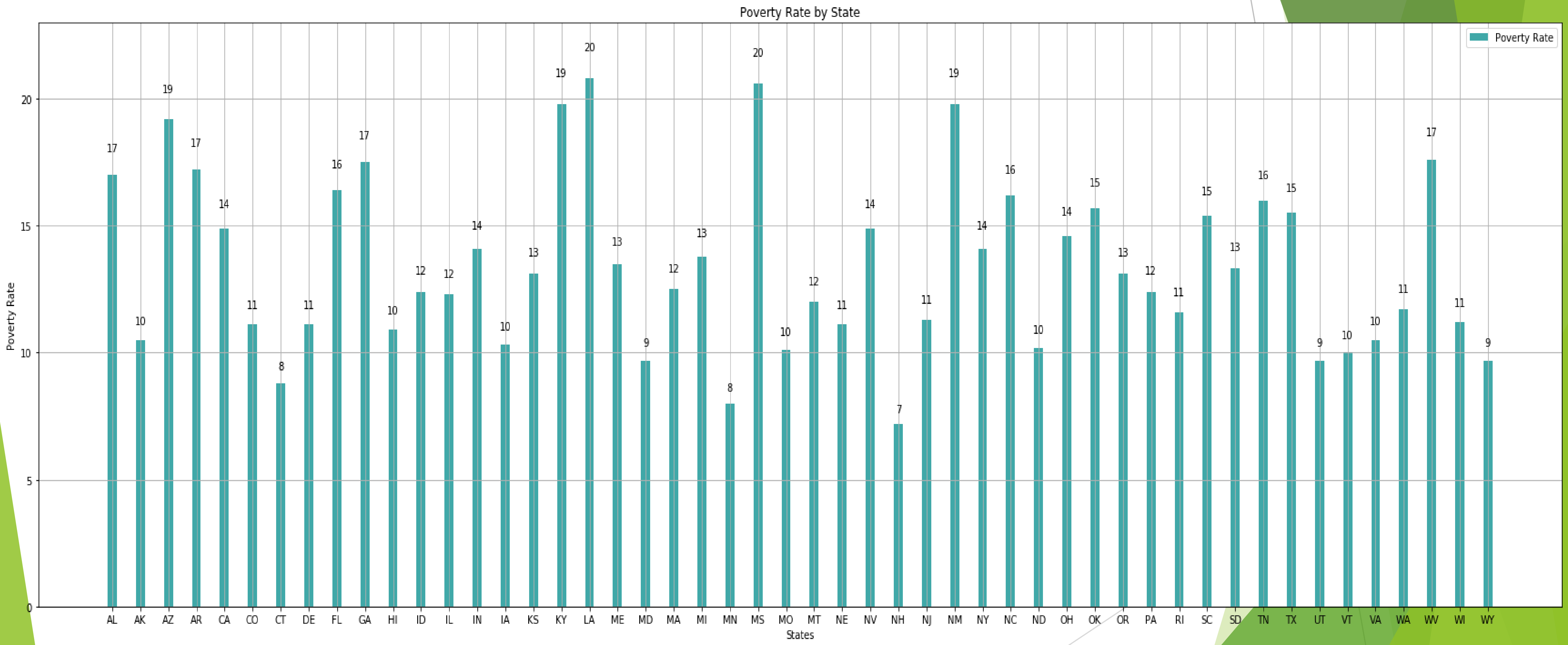
Males had a higher occurrence of heart disease in US



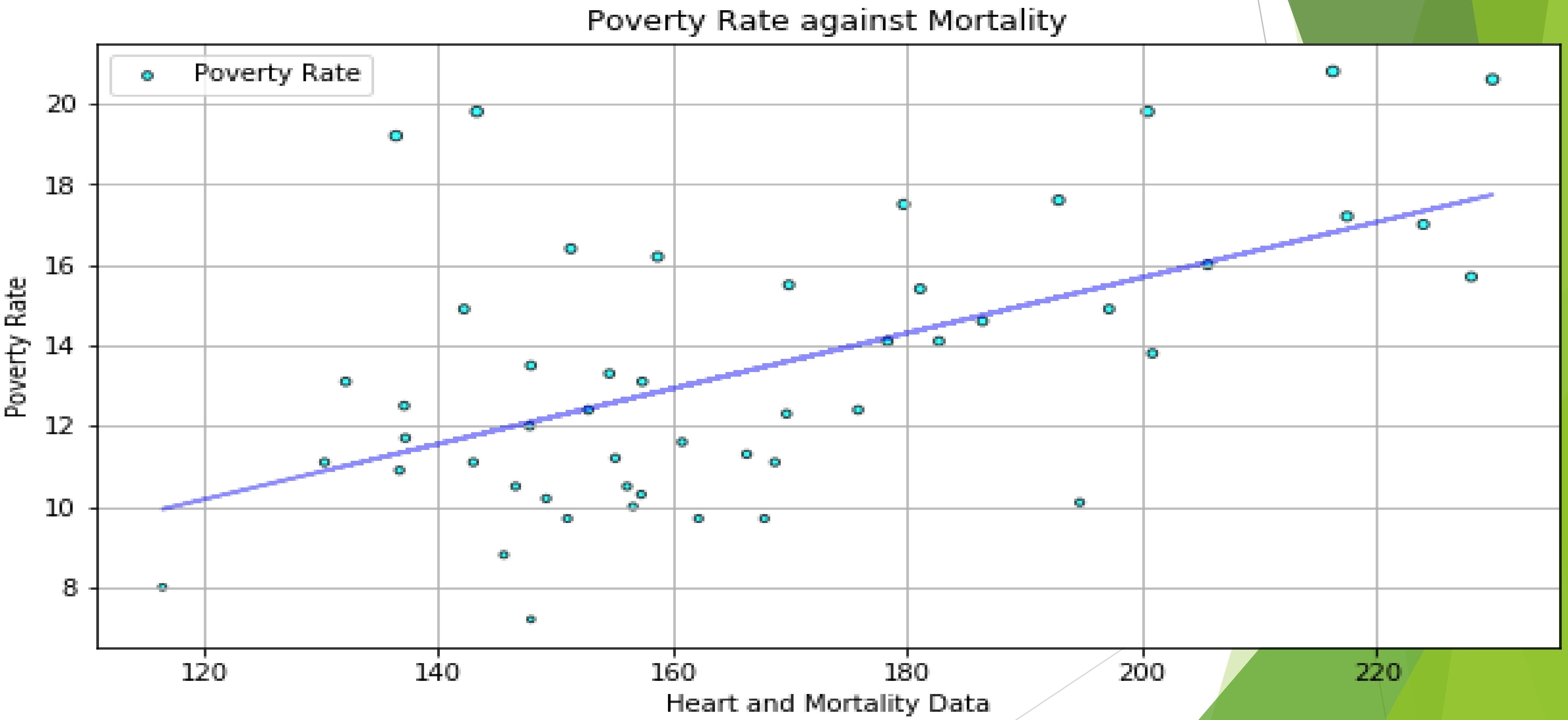
Heart Disease Analysis by States



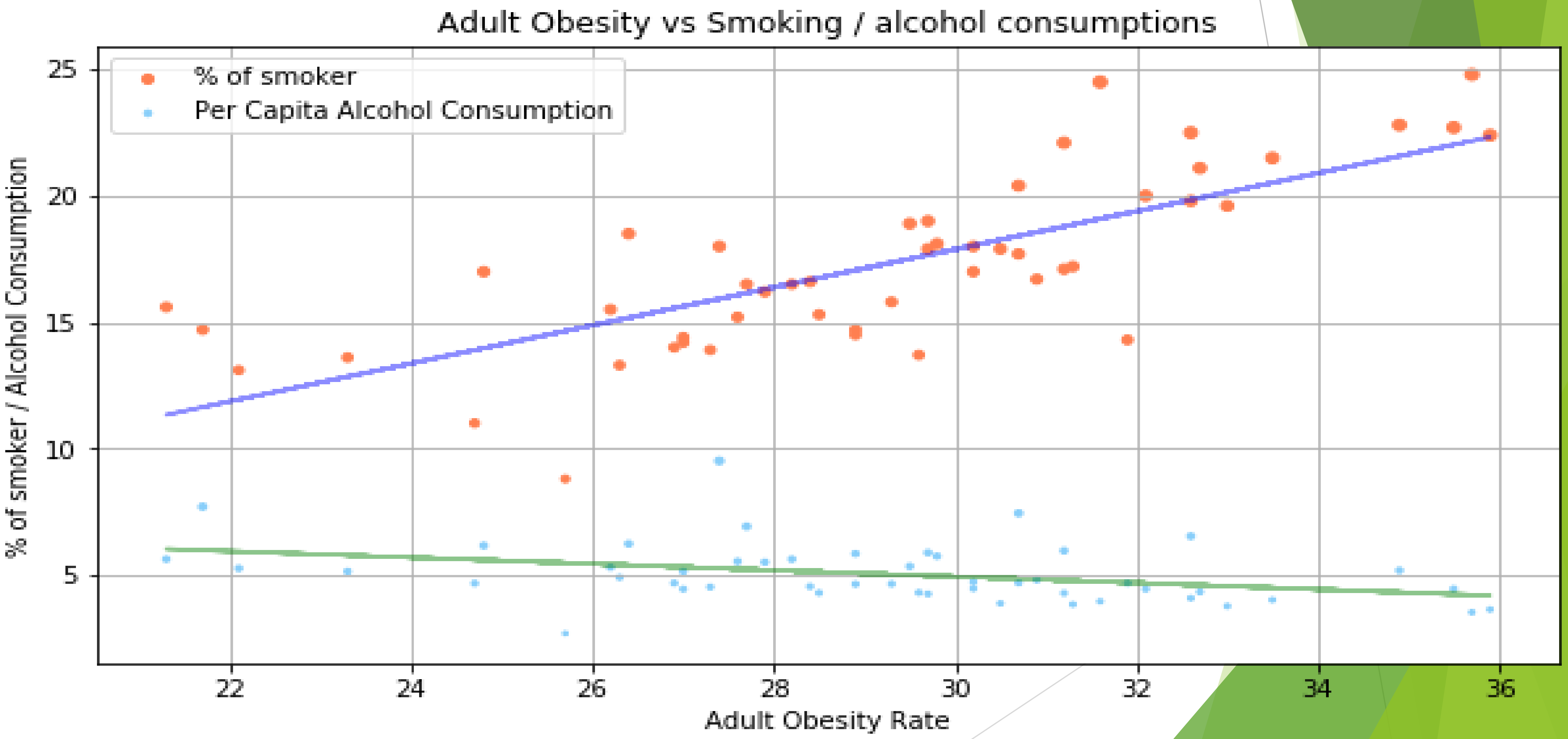
Poverty Rate by State



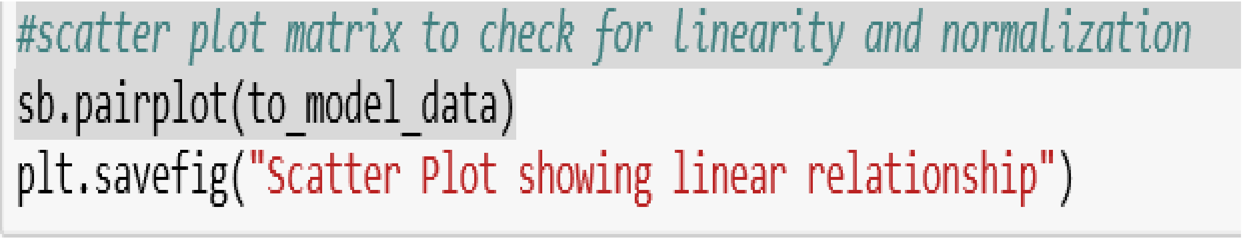
Poverty Rate by Mortality



Adult Obesity Vs Smoking / Alcohol

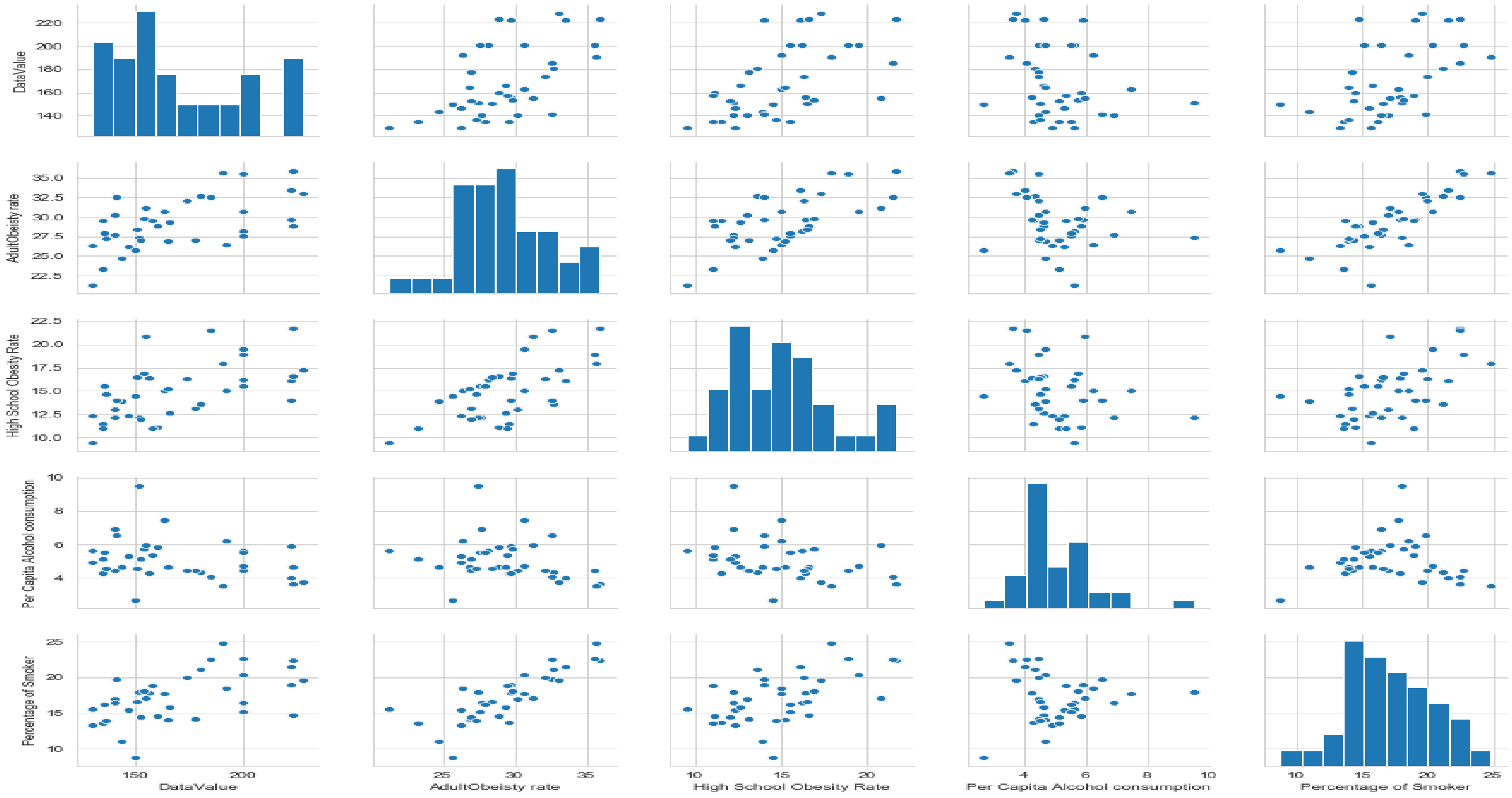


Analyzing Diabetic Cause Relation



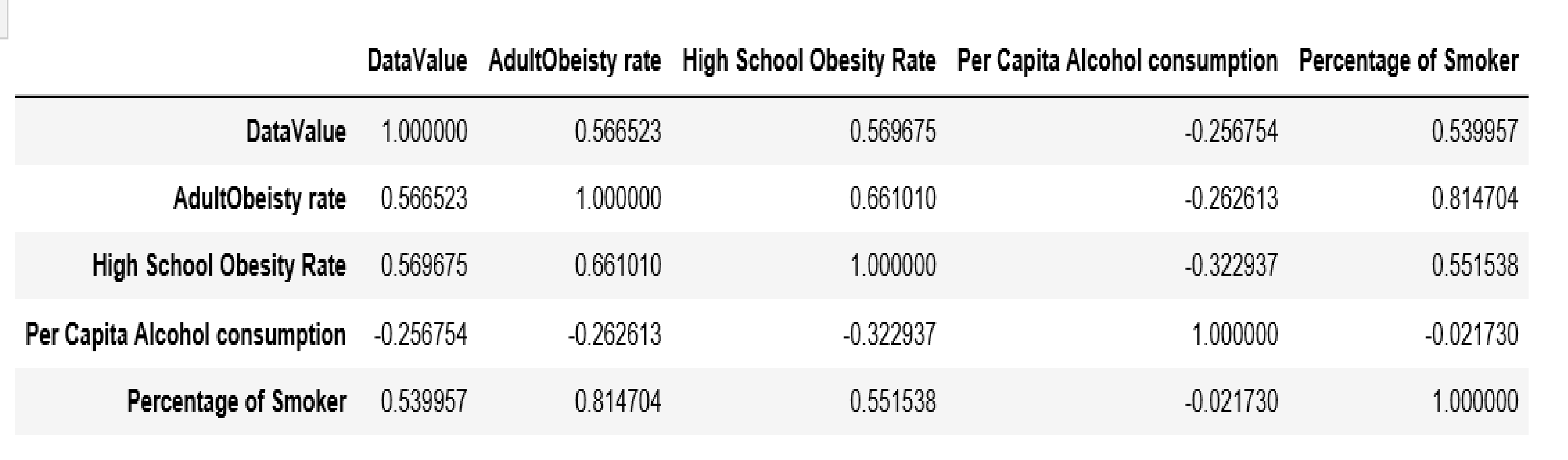
```
#scatter plot matrix to check for linearity and normalization
sb.pairplot(to_model_data)
plt.savefig("Scatter Plot showing linear relationship")
```

Checking for Normalization and Linearity



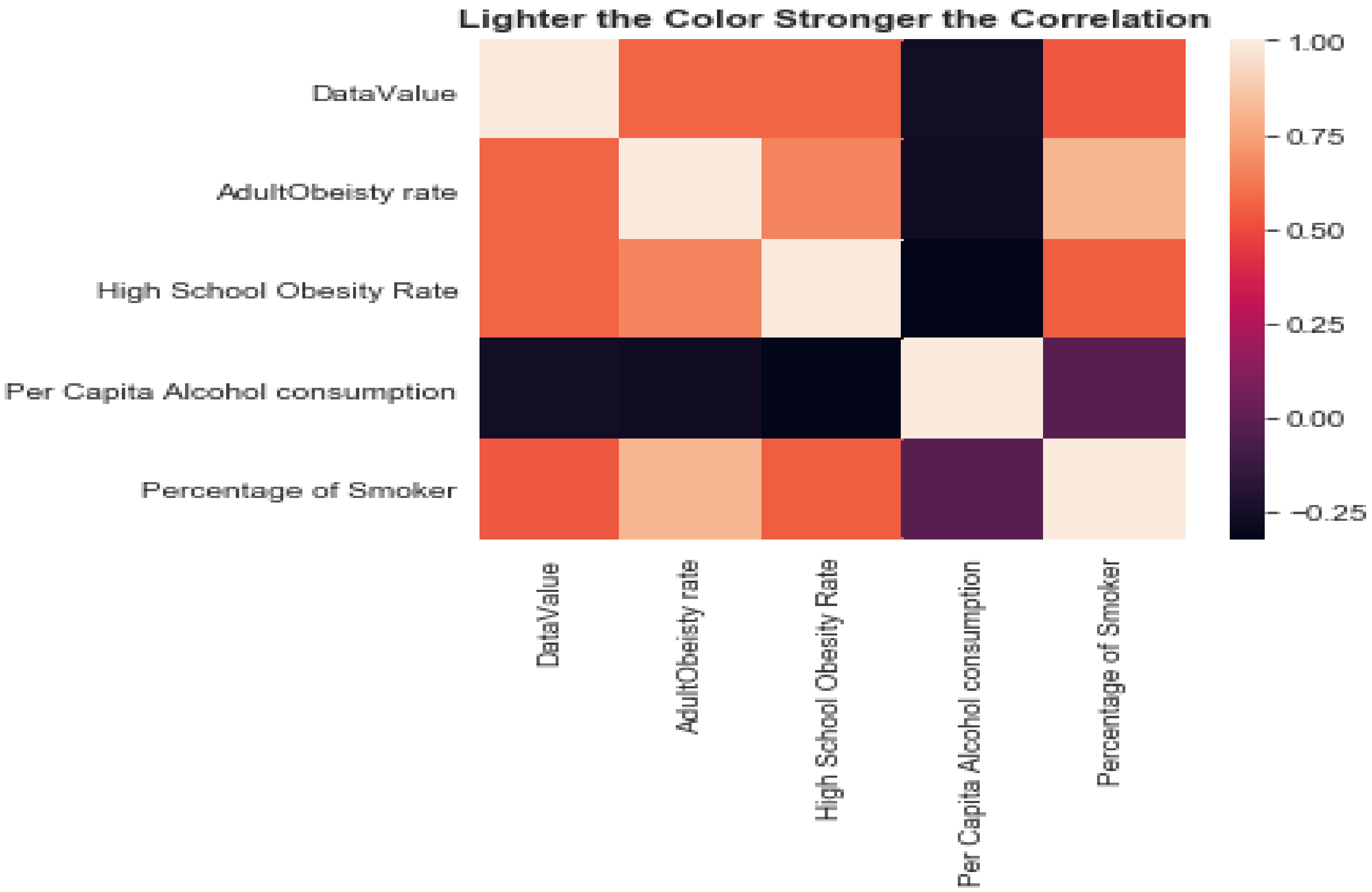
Correlation Coefficient Value (P Value)

- #Calculate Pearson correlation Coefficent
- corr = test_new.corr()
- corr.to_csv('Correlation Values.csv')



	DataValue	AdultObeisty rate	High School Obesity Rate	Per Capita Alcohol consumption	Percentage of Smoker
DataValue	1.000000	0.566523	0.569675	-0.256754	0.539957
AdultObeisty rate	0.566523	1.000000	0.661010	-0.262613	0.814704
High School Obesity Rate	0.569675	0.661010	1.000000	-0.322937	0.551538
Per Capita Alcohol consumption	-0.256754	-0.262613	-0.322937	1.000000	-0.021730
Percentage of Smoker	0.539957	0.814704	0.551538	-0.021730	1.000000

Correlation Coefficient



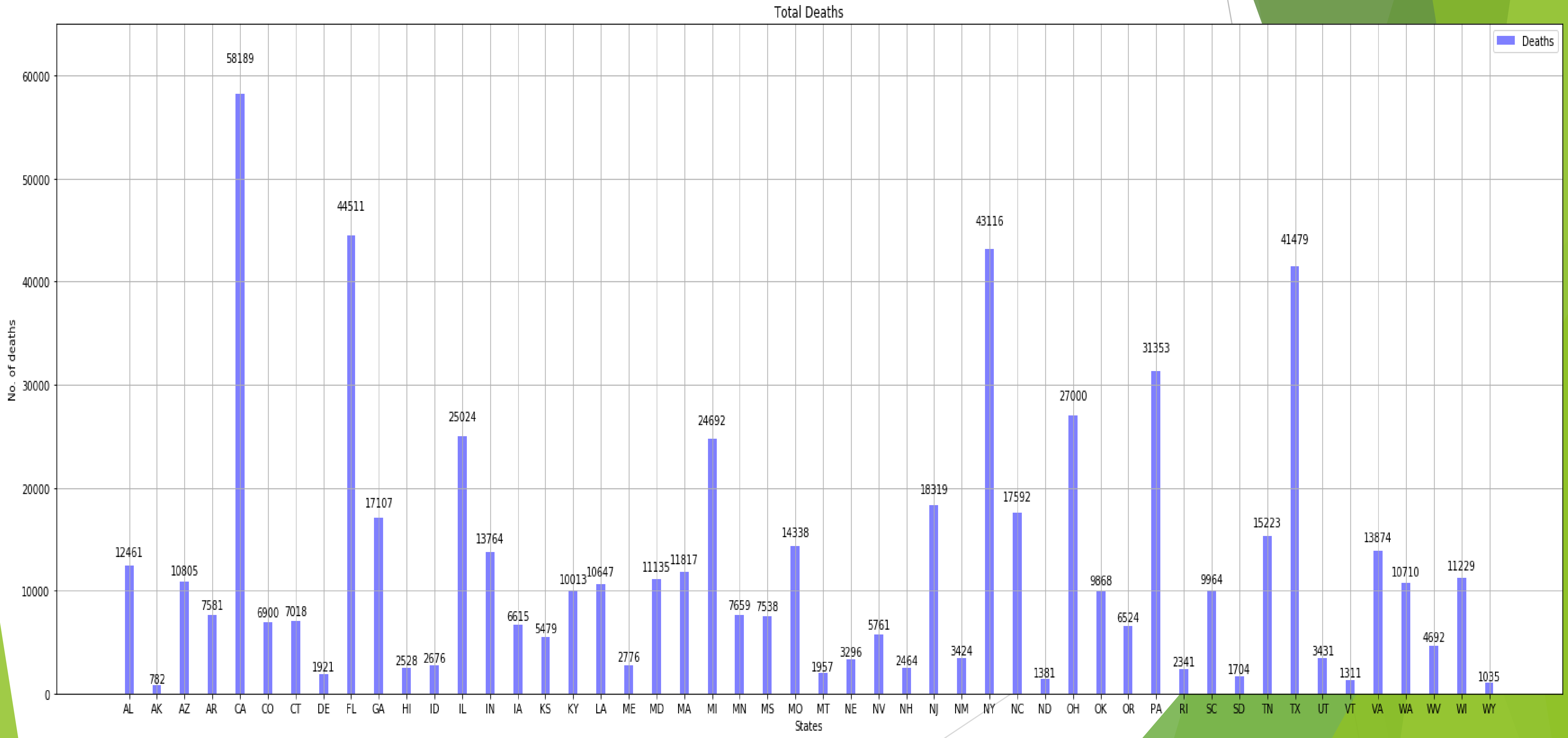
Observations

- Males have a greater % of heart disease rate than the females.
- The top 3 states with Heart Disease are Mississippi, Oklahoma and Alabama.
- The bottom 3 states with Heart Disease are Minnesota, Oregon and Colorado.
- There is a linear relationship between poverty rate and Heart / Mortality rate.
- The P values indicate people who smoke more tend to be more Obese.
- The P values also indicate that High School Students who are obese tend to more likely to be Obese if they continue with the lifestyle.

Further Research

- Considered additional years to understand trend over time for all the factors.
- Consider additional factors such as Nutrition, Education, % of grocery stores etc.
- Heatmap visualization of the data.
- Created Models to demonstrate top 5 causes.

Death by State


Total Deaths
Deaths
No. of deaths
AL
AK
AZ
AR
CA
CO
CT
DE
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MA
MI
MN
MS
MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY
States