

Data Exploration: Car Accidents

Looking over the dataset, a few questions came to my mind. Do states which a higher urban population percentage in respect to total population have the most traffic fatalities? If the state has a higher urban population, does that correlate to more drunk driving accidents as well? My hypothesis is that the higher the urban population, the more traffic fatalities occur with a great percentage being drunk driving. My reasoning for my hypothesis is that most urban areas are populated with clubs, bars, and even concerts and most people that live in urban areas live an urban lifestyle which includes drinking. An individual cannot even go to a baseball game and not see someone drinking a beer. With that being said, I decided to analysis the data provided to test my hypothesis.

My code that provided me with my graphs: Python

```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt

##Set up to import the excel file with the data
file = 'P1Dataset.xls'
worksheet = pd.ExcelFile(file)
df = worksheet.parse('dataset')
urban_percentage = df['Pct Urban Population']
urban = df['Urban Population']
total = df['Total Population']
traffic_fatalities = df['Traffic fatalities']
drunk_driving = df['Fatalities involving high blood alcohol']

##Graph showing the slight difference between total and urban population
total.plot('line', color='red', label='Total')
urban.plot('line', color='blue', label='Urban')
plt.title('Urban Population Vs State')
plt.xlabel('States')
plt.ylabel('Population (per 100k)')
plt.legend()
plt.show()

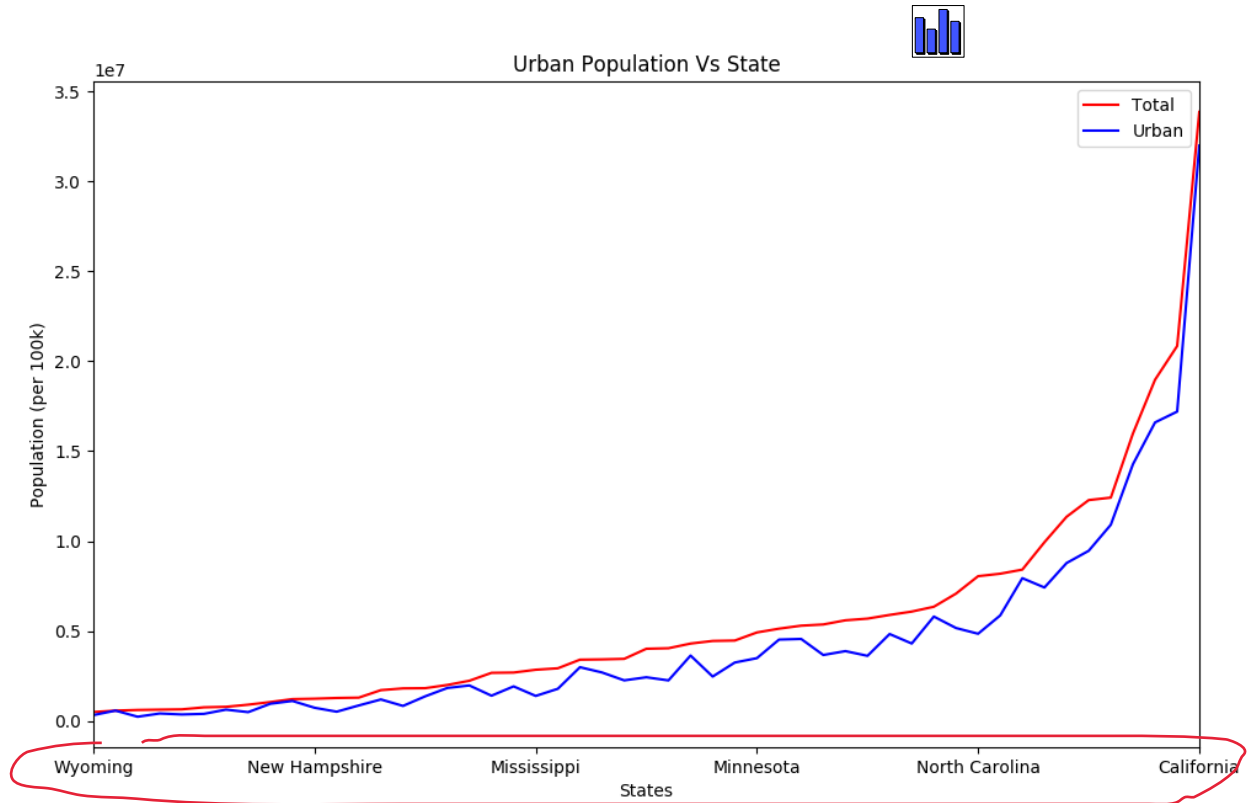
## Traffic Fatalities Graph
traffic_fatalities.plot('bar', color='black')
plt.title('Traffic Fatalities Vs State')
plt.xlabel('States')
plt.ylabel('# of Fatalities')
plt.xticks(rotation=90)
plt.show()

##Drunk Driving graph
drunk_driving.plot('bar', color='red')
plt.title('Drunk Driving Incidents')
plt.xlabel('States')
plt.ylabel('# of Fatalities')
plt.xticks(rotation=90)
plt.show()

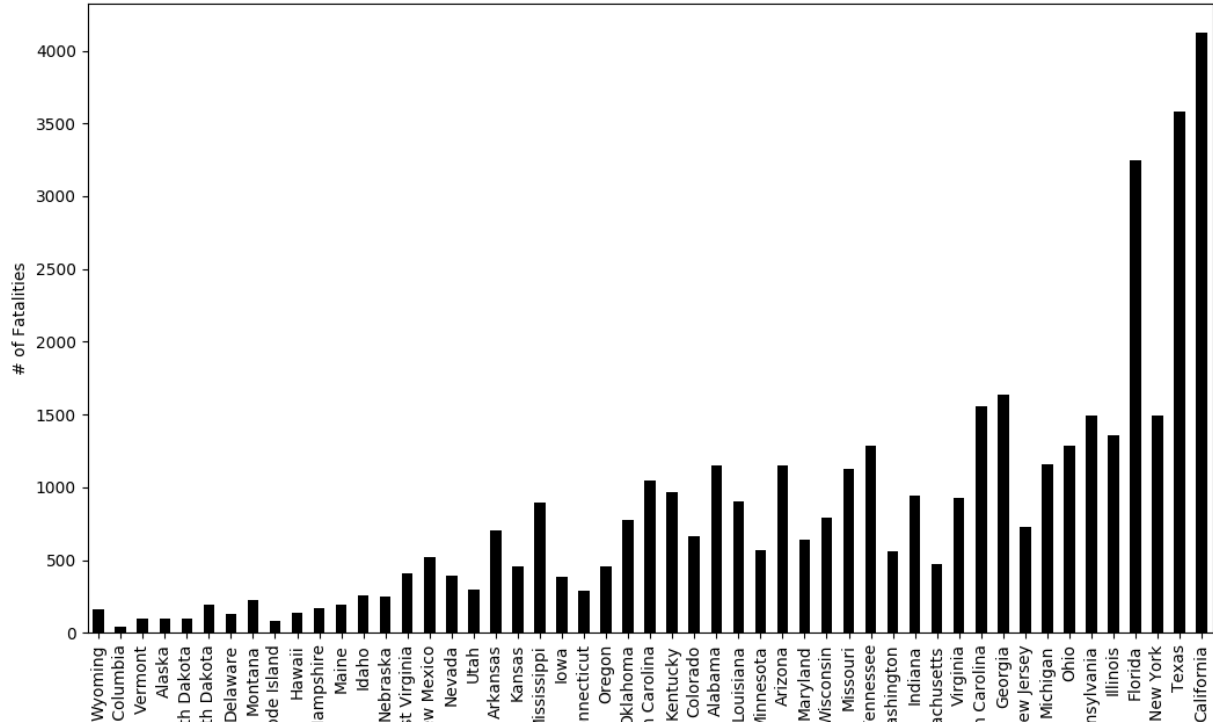
##Graphs of Traffic Deaths and Drunk Driving Incidents
traffic_fatalities.plot('line', color='black', label='Traffic Deaths')
drunk_driving.plot('line', color='green', label='Drunk Driving')
```

```
plt.title('Incidents')
plt.xlabel('States')
plt.ylabel('Traffic Deaths')
plt.legend()
plt.show()
```

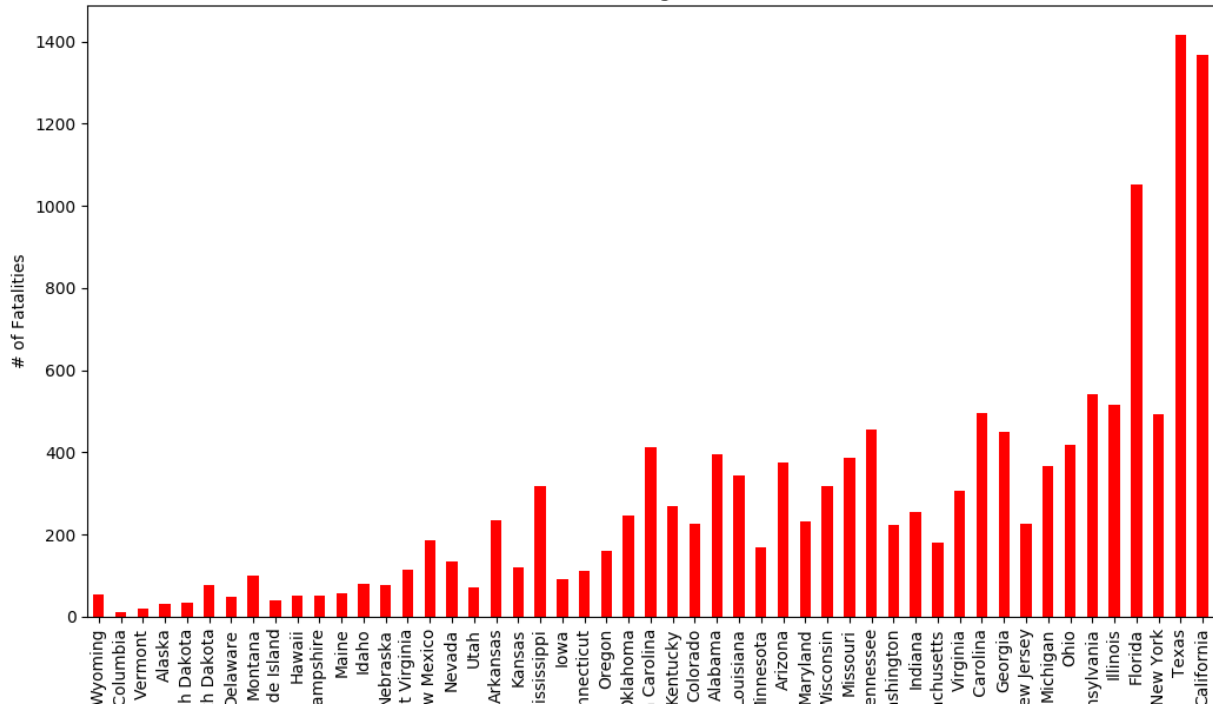
The output of the graphs was rather interesting. Listed below are a few graphs from my data explorations:

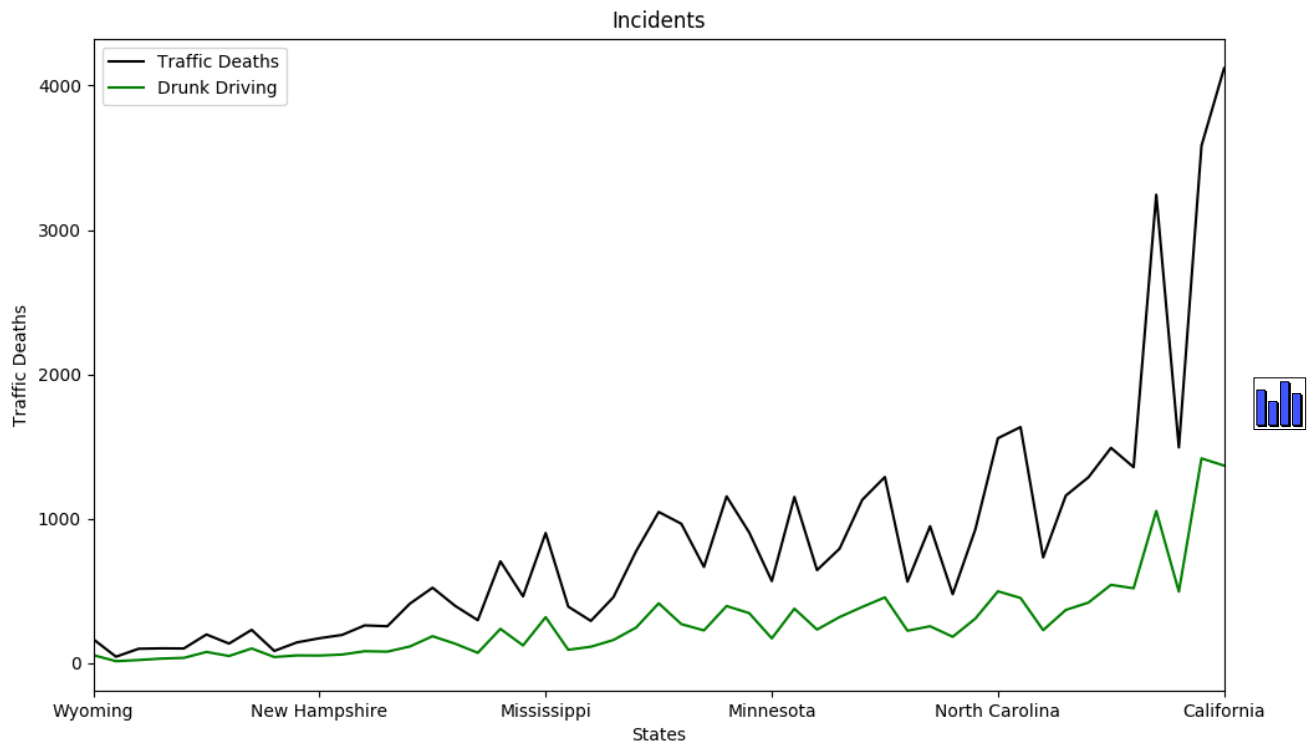


Traffic Fatalities Vs State



Drunk Driving Incidents





From the graphs above, my hypothesis seems to be partially incorrect simply because it seems that the urban population makes up majority of the total population from the first graph above. So instead of focusing on urban population, it looks like states with a higher population as a whole has the most car incidents. The traffic fatalities graph and the drunk driving graph looked very similar, so similar that I decided to combine the two on the incident graph above. Drunk driving accidents almost exactly corresponds with the trend of the total traffic deaths by state. It seems that of the traffic deaths, about half of them where drunk driving incidents. Overall, both that traffic deaths and the drunk driving graph stats both show that it does indeed correlate with population. Wyoming, for example, have very low population along with very low traffic deaths versus California with an enormous population with the greatest amount of traffic deaths. Surprisingly, California drunk driving incidents does not seem so high in comparison to the total traffic deaths. (Research for another day). Other than California, drunk driving is the cause for about half of the traffic incidents.

A few ideas to lower the statistics of car accidents would be to focus on the drunk driving incidents. We cannot lower population but the least we could do is take action on things we can change. Uber could create a secondary app for drunk individuals to where if a friend or intoxicated person tap a button that says "I need a ride", the uber will take them where they need to go and the app will reward the user by providing a free ride after "x" amount of drunk calls. There are a few ways to try to create initiatives to stop the reckless drunk driving and that would possibly lower all traffic incidents by 25% possibly.

What could be done with this data in order to generate new data or predictions is collect all of the information in that dataset daily. Since my main focus was population and traffic incidents, I would create a model that will learn from the data being brought

in year after year to predict the occurrences in accidents by state and the population correlation. Hopefully that would help with predicting the number of incidents by state which would allow users like traffic control officers to lower the rates by placing more officers on duty in those main locations. I would also add data such as time of incidents by state which would calculate the average time of day the most accidents occur. That could help narrow down the time in which to have more officers on duty by state. There is so much that could be done with this data and with the right algorithm, models could be created that would help produce solutions in the long run.