Data Exploration

2023-10-22

1) My problem statement for this project is, is Unemployment and Labor Force Participation a predictor of future crimes in a state? For my project I'm examining 3 different states and 9 total data sets.

Maryland Unemployment Rate - MDUR.csv Source: https://fred.stlouisfed.org/series/MDUR

Maryland Labor Force Participation - maryland_labor.csv Source: https://fred.stlouisfed.org/series/LBSSA24

Maryland Crime Data - Maryland crime 1975-2022.csv Source: https://catalog.data.gov/dataset/violent-crime-property-crime-by-county-1975-to-present

New York Unemployment Rate - NYUR.csv Source: https://fred.stlouisfed.org/series/NYUR

New York Labor Force Participation - newyork_labor.csv Source: https://fred.stlouisfed.org/series/LBSSA36

New York Crime Data - New York State Crime 1990-2022 Source: https://catalog.data.gov/dataset/index-violent-property-and-firearm-rates-by-county-beginning-1990

Washington Unemployment Rate - WAUR.csv Source: https://fred.stlouisfed.org/series/WAUR

Washington Labor Force Participation - washington_labor.csv Source: https://fred.stlouisfed.org/series/LBSSA53

Washington Crime Data - Washington state crime.csv Source: https://catalog.data.gov/dataset/washington-state-criminal-justice-data-book-6c019

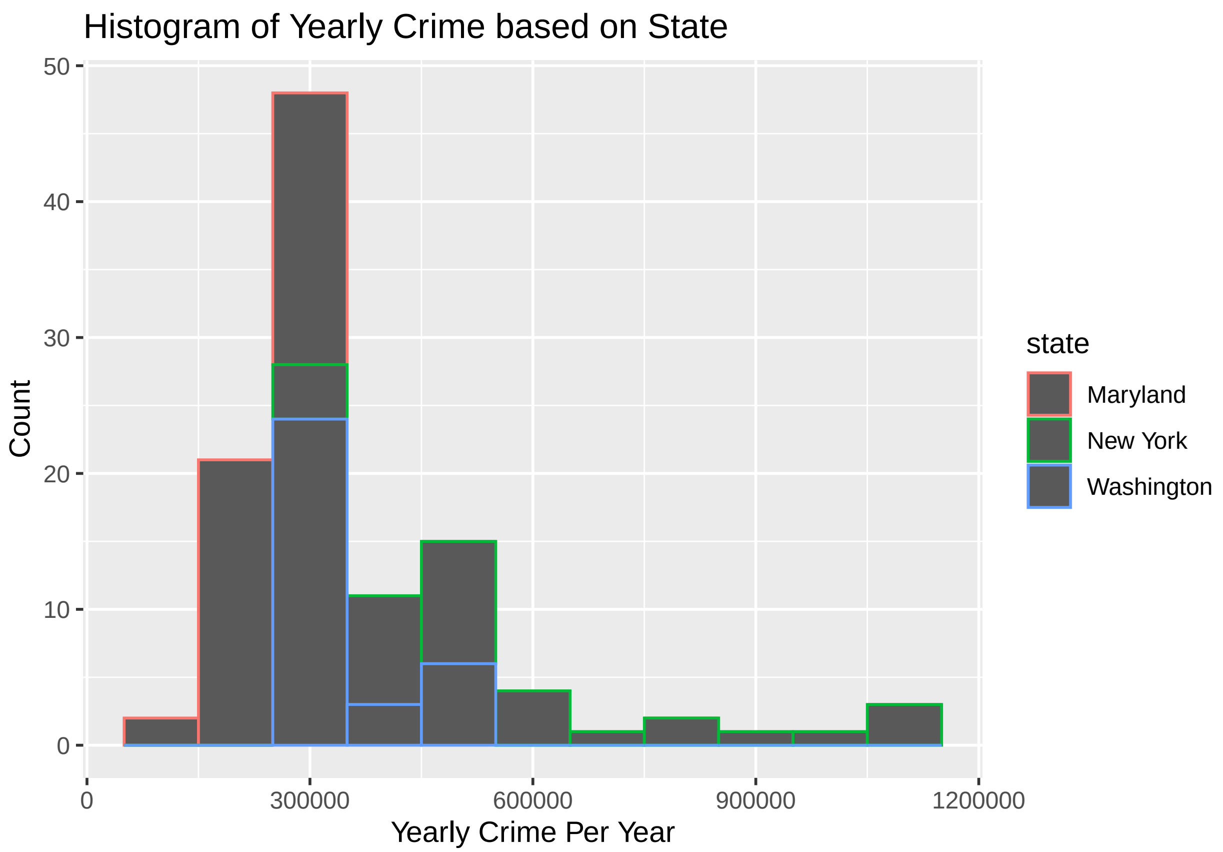
2) For all of the Unemployment data sets for each state the data engineering process was fairly simple. I first brought in the data by reading the csv file. In the same block I also converted the Date column to only include the year. Then I renamed all of the date columns to Year. After that I renamed the column that contained unemployment data to unemployment. After that I converted the Unemployment columns to numeric and rounded them off to the second decimal place.

All of the Labor force data sets had the same structure as well. I had to rename some columns and convert the labor force columns to numeric. I also rounded the percentage to the second decimal place.

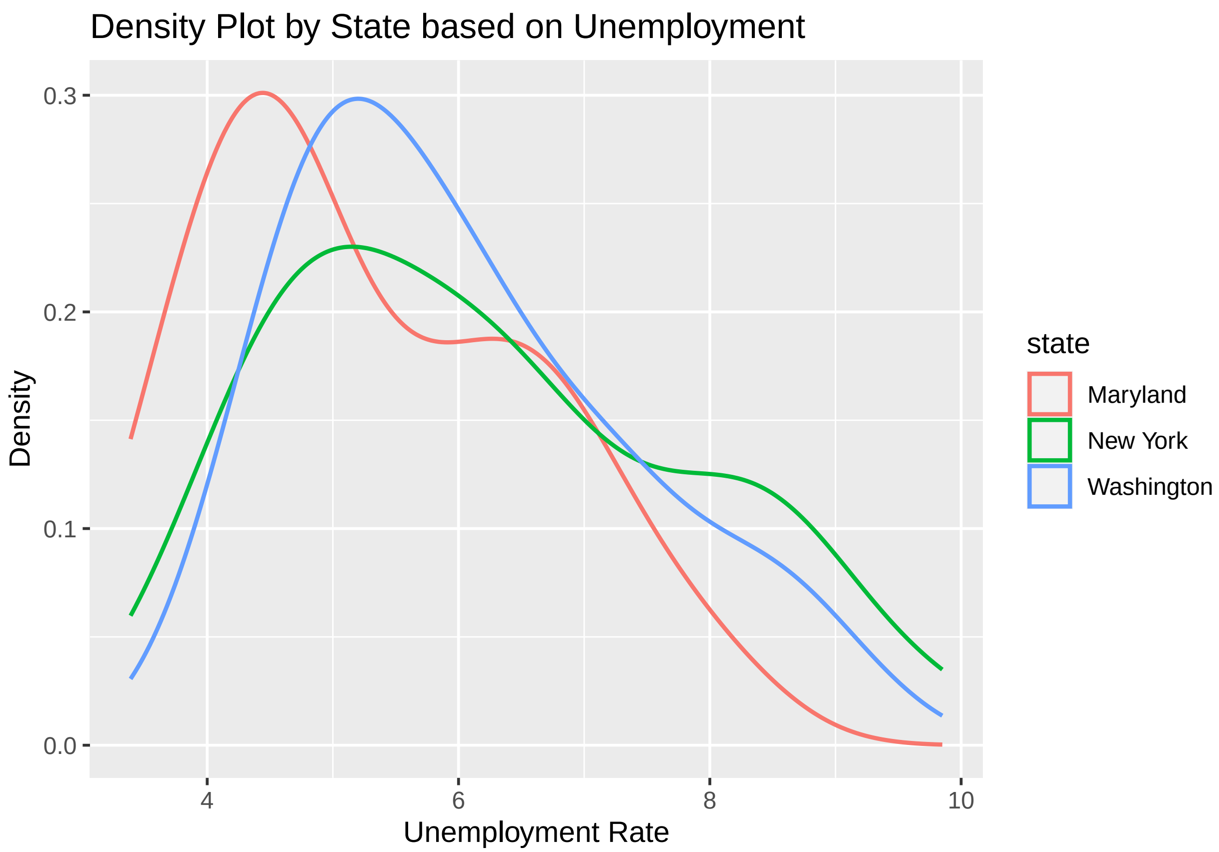
For all the crime data sets I had to rename the columns that contained information to the year. Then I had to select the relevant columns from the data set. The Washington data set had over 50 columns that I had to sort through. New York and Maryland were closer to 20 rows. I kept the original datasets so that I could go back and conduct more data exploration later.

After this I combined that Unemployment, Labor Force percentage, and crime data into one data set for each state. I joined them all on the year. Finally, I combined all the datasets for all the states into one data set so I could compare them against each other.

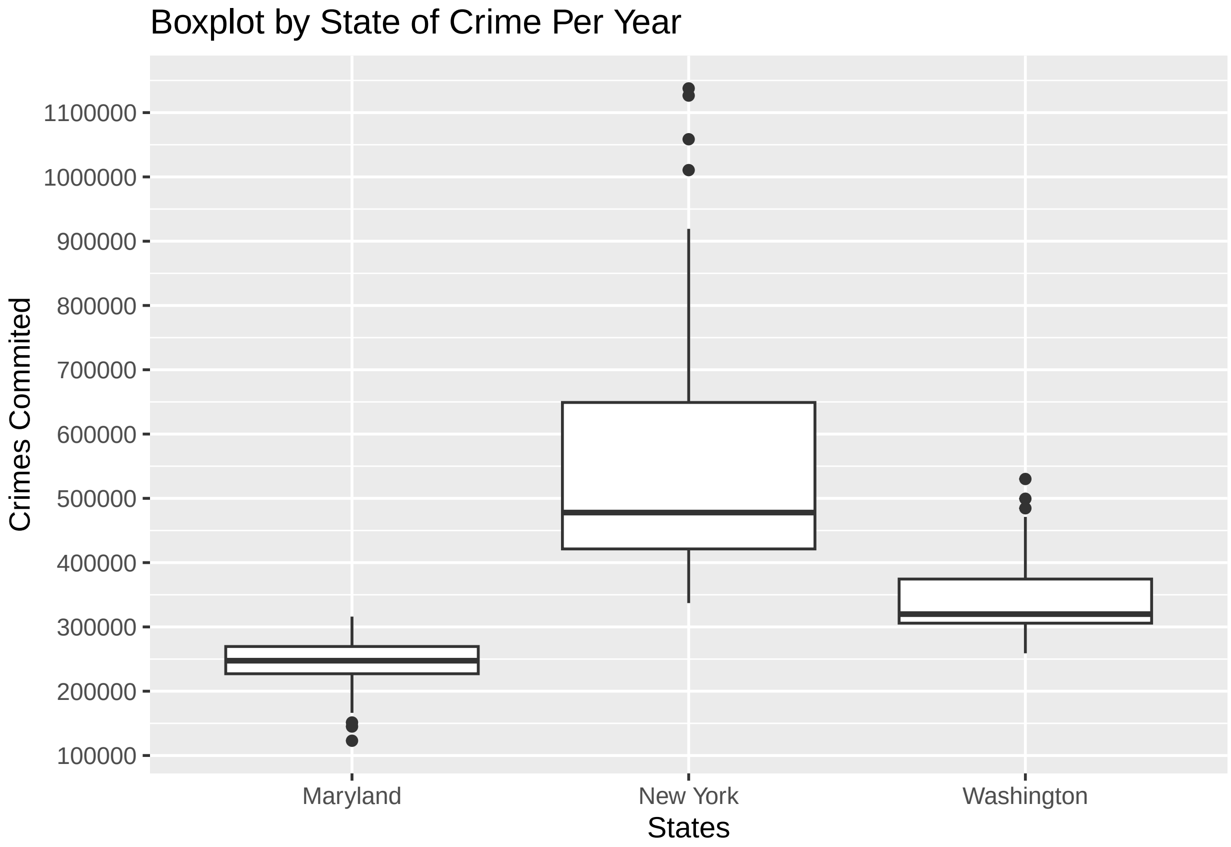
3) For my previously submitted Data Visualizations the only feedback I got was to explain how analyzing the relationships will be useful. Other than that all my feedback was positive towards my data visualizations.



Analyzing the distribution of crimes for each state is helpful because it helps me visualize the trends in Crime. For example Maryland tends to have less crime each year than New York and Washington



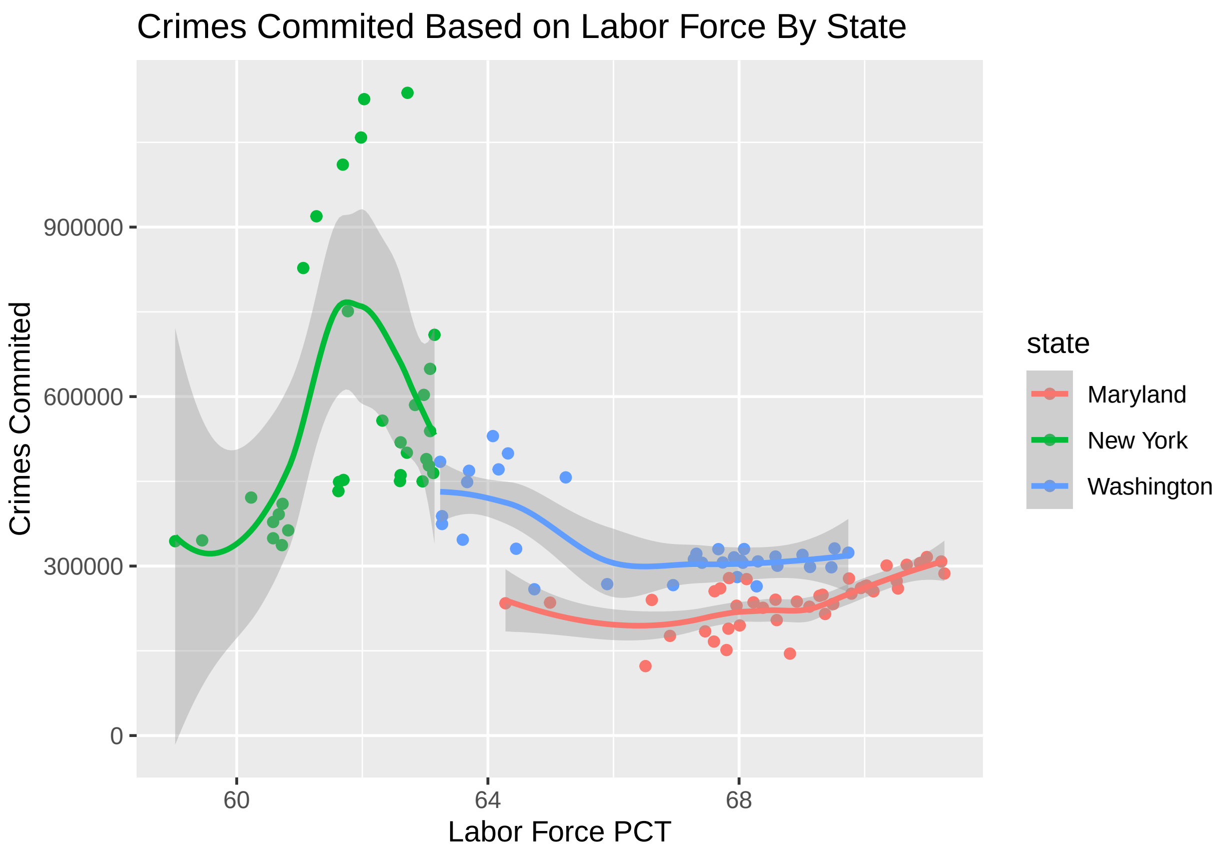
This density plot is important because it tells me where most of the data for unemployment sits. With this I can examine if crime is consistent when the unemployment is at the same percent. This will help explain the crime and unemployment relationship



This box plot is important because it shows me outliers in the crime data. With this I go about figuring out how to address these data points and how they skew my results

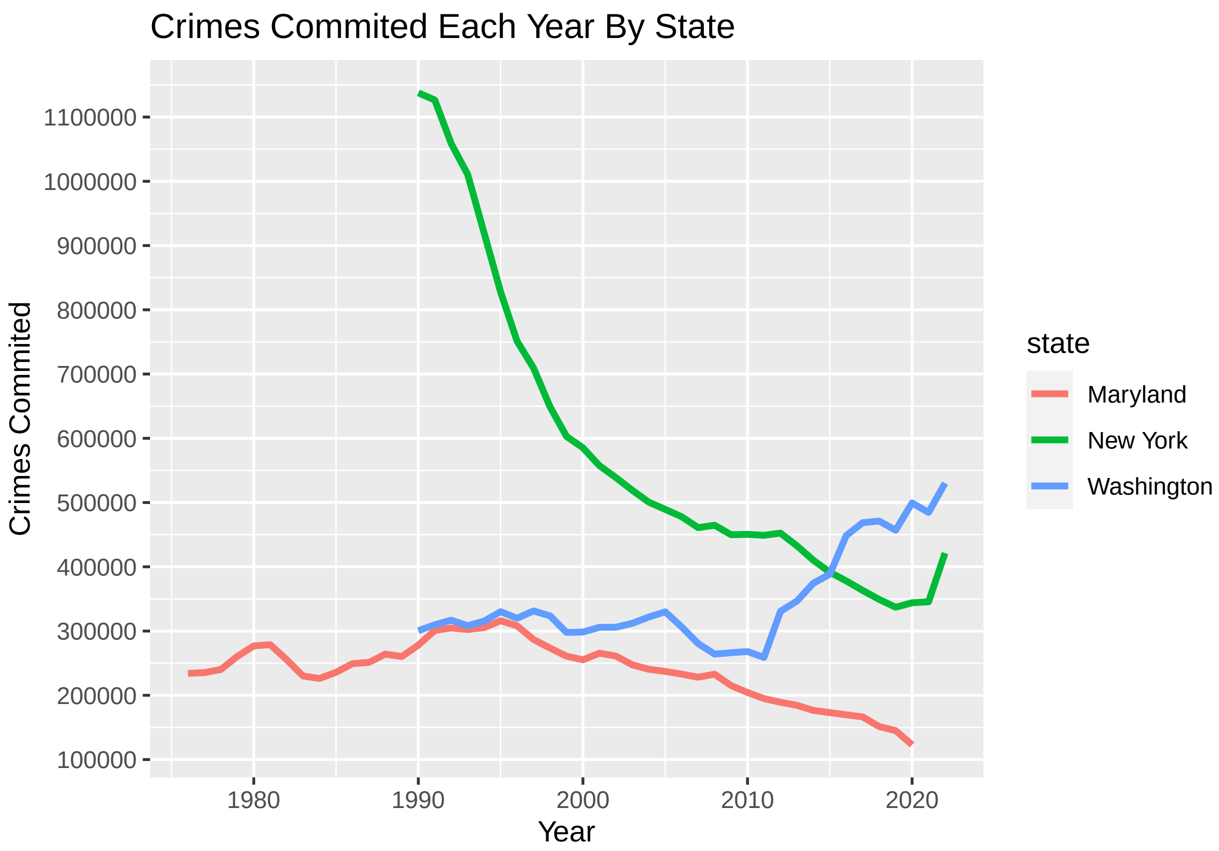
This graphic is important because it helps me visualize the relationship betwen unemployment and crime which is what my problem statement is about.

```
## ‘geom_smooth()‘ using method = ’loess’ and formula = ’y ~ x’
```

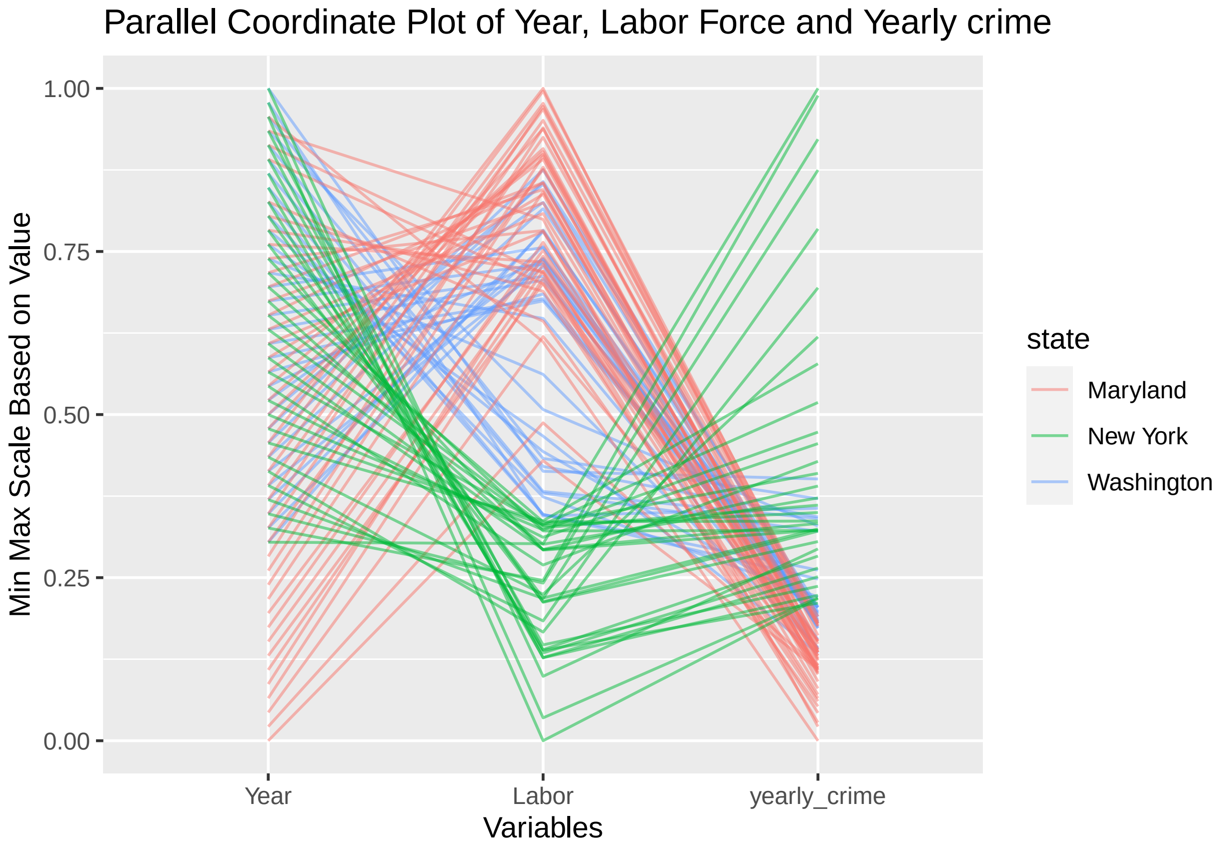


```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```

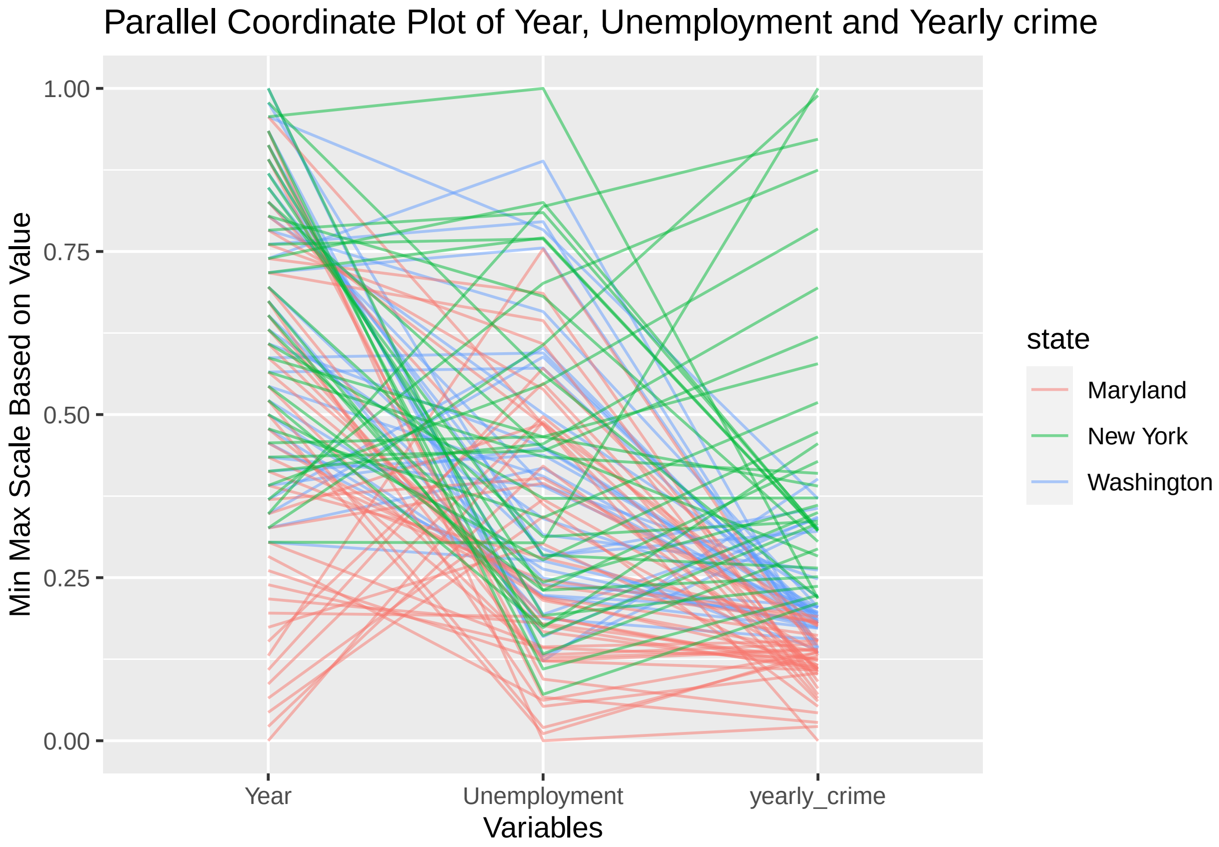
This graphic is important because it visualizes the realationship between labor force percentage and crime whic is what my problem statement is about.



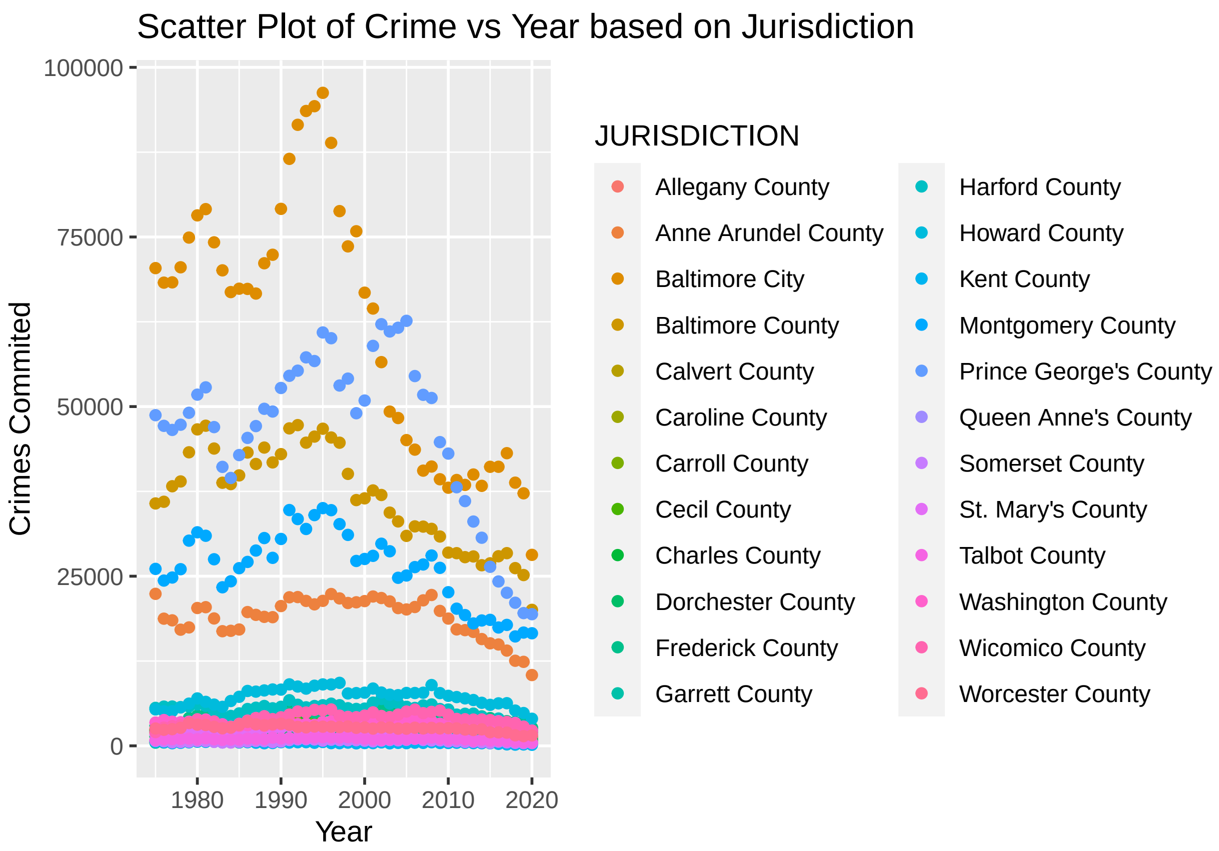
This visualization is important because it helps me understand crime trends over time. If there is a relationship between crime and labor force percentage or unemployment then they should look similar



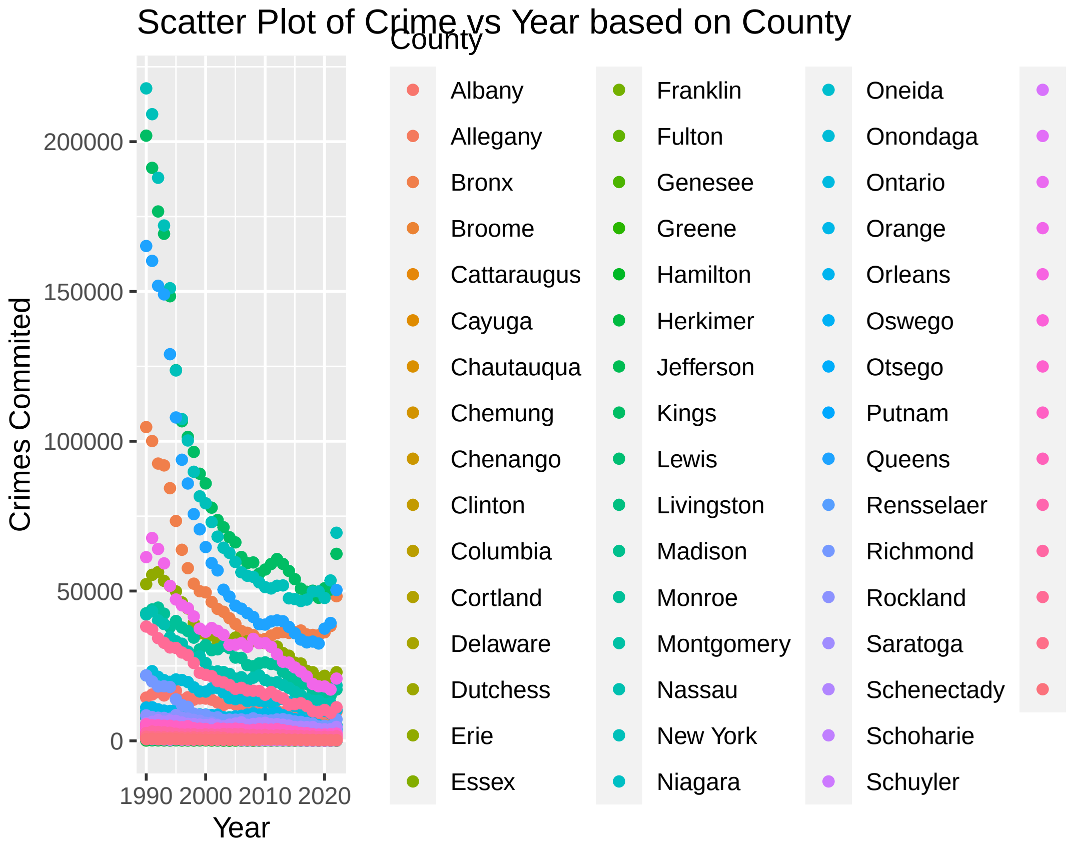
This parallel coordinate plot is important because it helps me understand how crime relates to labor and overall trends on a state by state basis.



This parallel coordinate plot is important because it helps me understand how crime relates to unemployment on a state by state basis.



This scatter plot is important because it helps me understand that most counties experience significantly lower crime. It also helps me pick out counties that I might want examine futher.



Similar to the last scatter plot this helps me pick which counties I need to examine further. It also important because it reminds me to examine how population might effect crime rates for each county.

This visualization is important because it shows me the relationship between population and crime for counties with less than 40,000 people basis in Maryland

4) The three data exploration techniques I decided to go with were Summarizing by Group, Outlier Detection, and correlation.

I picked summarizing by group rather than other exploration techniques like spatial analysis because I was more consider with the data from the counties itself rather than the actual location of the counties. With this technique I was able to see that most counties reported less than 25,000 crimes in New York and Maryland. I also found that most counties had a population under 250,000.

I picked Outlier Detection rather than other exploration techniques because my data has so many points. I need to make sure that my visualization and conclusions are not skewed by outliers. This will improve the accuracy of my data exploration with correlation. With Outlier Section I found that Kings county in New York had four outliers in the yearly crime columns. I was also able to find more outliers in other counties.

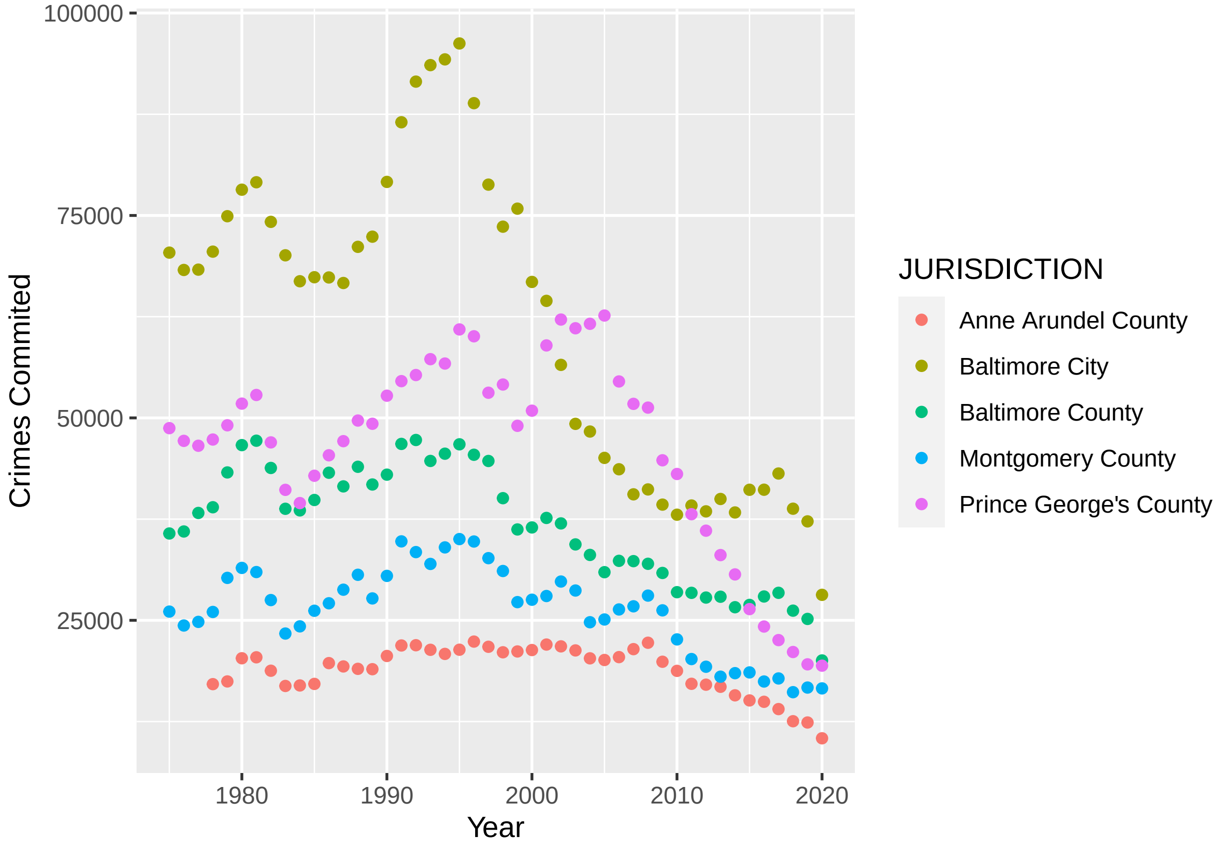
And the final data exploration technique is correlations as I briefly mentioned before. I select correlation over other data exploration techniques because my project centers around the relationship between variables. And correlation is an excellent way to show that. When examining the correlation for New York between crime and unemployment as well as labor the correlation coefficient was quite low. For Washington there was a low negative correlation between crime and unemployment but for labor and crime there was a moderately high positive correlation. Maryland had a negligible correlation between unemployment and crime. But when it came to labor and crime the correlation it was moderately high and positive.

5)Summarizing by Group I can integrate these results into my research by separating the the counties that experience less crime from counties that experience more crime. I can also separate the counties with more than 250,000 people from the smaller ones to get more informative visualizations. Summarizing by group provided me with the insight that I might need to answer my problem statement separately for these two groups to provide an accurate answer.

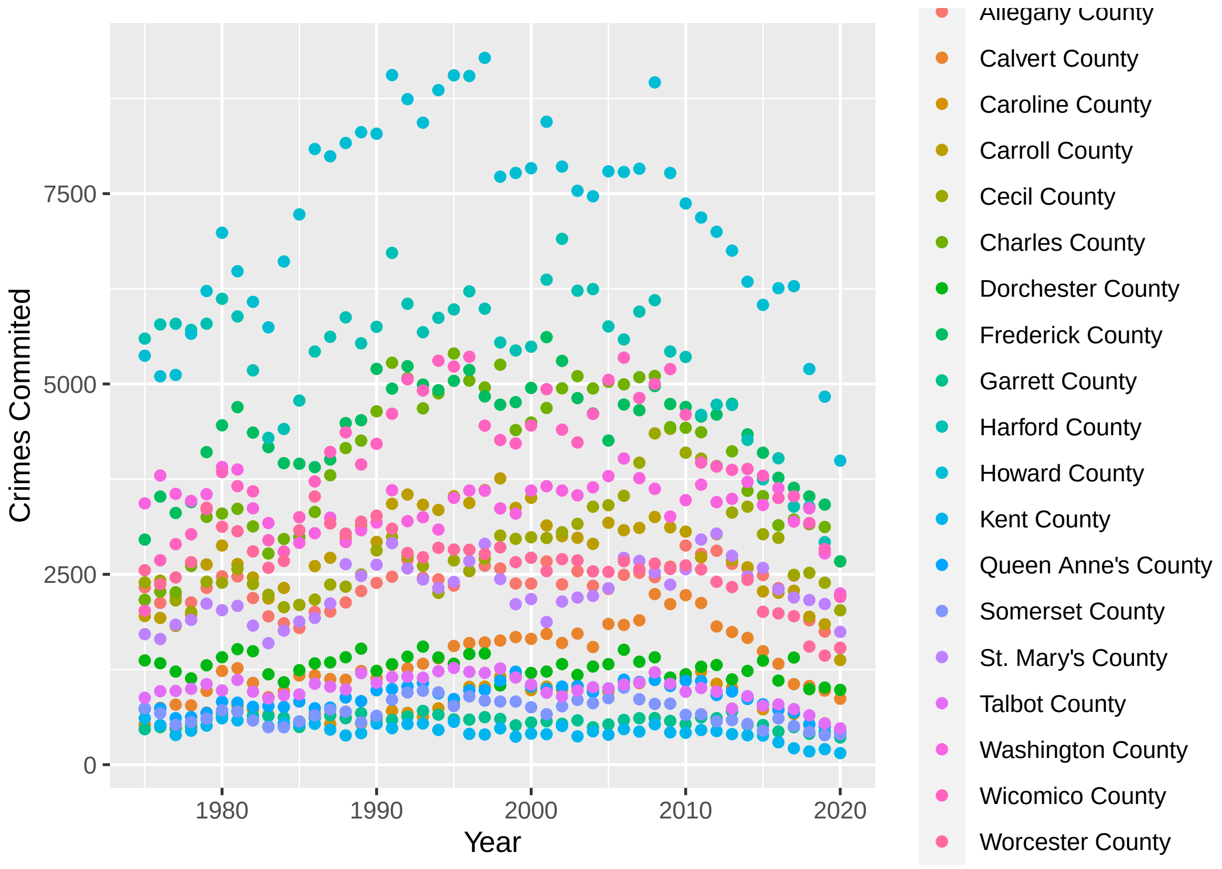
Outlier Detection I will integrate these results by trimming outliers for these counties. This technique provided me with the insight that I need to deal with outliers on a county by county basis. It also provided me with the insight that I need to compare check the correlation between these outliers and their respective unemployment or labor force percentage. This could show that if there's a significant spike in crime unemployment responded similarly or differently giving insight on the relationship.

Correlation I can integrate these results by using them to answer my problem statement. These results provided me with the insight I might need to compare unemployment and labor force on a county by county basis to see if these correlation stand up so that I can answer my problem statement more accurately.

```r
ggplot(yearly_crime_m, aes(x = Year, y = GRAND.TOTAL, color = JURISDICTION)) +

  geom_point(data = subset(yearly_crime_m, POPULATION < 2000000 & 350000 < POPULATION))+

  labs(x = "Year", y = "Crimes Commited")
```



```r
ggplot(yearly_crime_m, aes(x = Year, y = GRAND.TOTAL, color = JURISDICTION)) +

  geom_point(data = subset(yearly_crime_m, POPULATION < 330000))+

  labs(x = "Year", y = "Crimes Commited")
```



```
png("larceny vs state maryland.png", width = 800, height = 600)
ggplot(yearly_crime_m, aes(LARCENY.THEFT, JURISDICTION)) + geom_boxplot()
```