DATA 490 FINAL PRESENTATION

PHILIP BAILEY

PROBLEM STATEMENT

Can Unemployment and Labor Force Participation predict future crimes on a statewide scale?

DATA SETS

	Maryland	New York	Washington
Unemployment	1976 - 2022	1976 - 2022	1976 - 2022
Labor Force Participation	1976 - 2022	1976 - 2022	1976 - 2022
Crime Statistics	1975 - 2020	1990 - 2022	2000 - 2022

DATA STRUCTURES

Unemployment Data Sets

Date	Unemployment
1976-01-01	8.78333333333333
Numeric	Numeric

Labor Force Data Sets

Date	Labor Force
1976-01-01	61.3916666666667
Numeric	Numeric

Washington Crime Data Set

Year	County	Pop_Total	SRS_TOTAL	NIB Total	-
1990	STATE	4866692	300546	*	-
Numeric	Character	Numeric	Numeric	Numeric	-

*Washington Police switched crime reporting systems in 2011 from SRS to NIB

-This data set has over 50 Columns. Taking this into account I only included relevant columns

Maryland Crime Data Set

Jurisdiction	Year	Population	Grand.Total	-
Allegany C ounty	1975	79655	300546	-
Character	Numeric	Numeric	Numeric	-

-This data set has over 30 Columns. Taking this into account I only included relevant columns

New York Crime Data Set

County	Year	Population	Yearly_Crime	-
Albany	1990	292594	1137689	-
Character	Numeric	Numeric	Numeric	-

-Only relevant columns were included

DATA STRUCTURES CONT.

DATA JOINS EXAMPLE

Maryland State Data

Year (PK)	Numeric
Total Crime	Numeric
Unemployment	Numeric
Labor Force	Numeric
State	Character

New York State Data

Year (PK)	Numeric
Total Crime	Numeric
Unemployment	Numeric
Labor Force	Numeric
State	Character

Washington State Data

Year (PK)	Numeric
Total Crime	Numeric
Unemployment	Numeric
Labor Force	Numeric
State	Numeric

Combined State Data

Year (PK)	Numeric
Total Crime	Numeric
Unemployme nt	Numeric
Labor Force	Numeric
State	Character

DATA JOINS EXAMPLE

Maryland Crime Clean (Source)

Year (PK)	Numeric
Total Crime	Numeric

Maryland Unemployment (Source)

Year (PK)	Numeric
Unemployment	Numeric

Maryland Labor Force (Source)

Year (PK)	Numeric
Labor Force	Numeric

Maryland State Data

Year (PK)	Numeric
Total Crime	Numeric
Unemployment	Numeric
Labor Force	Numeric
State	Character

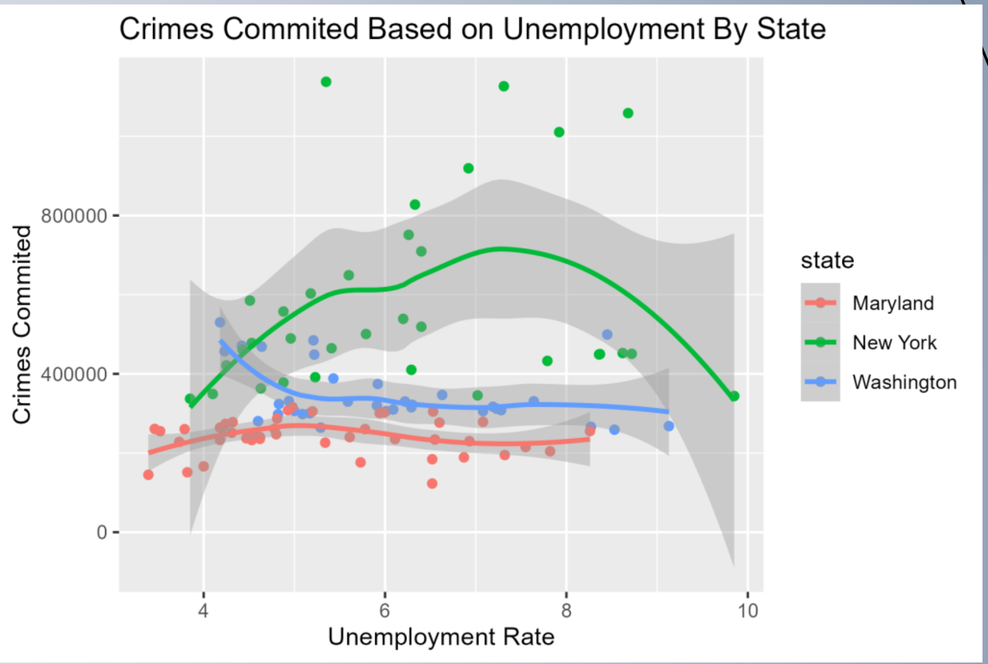
Maryland State Data

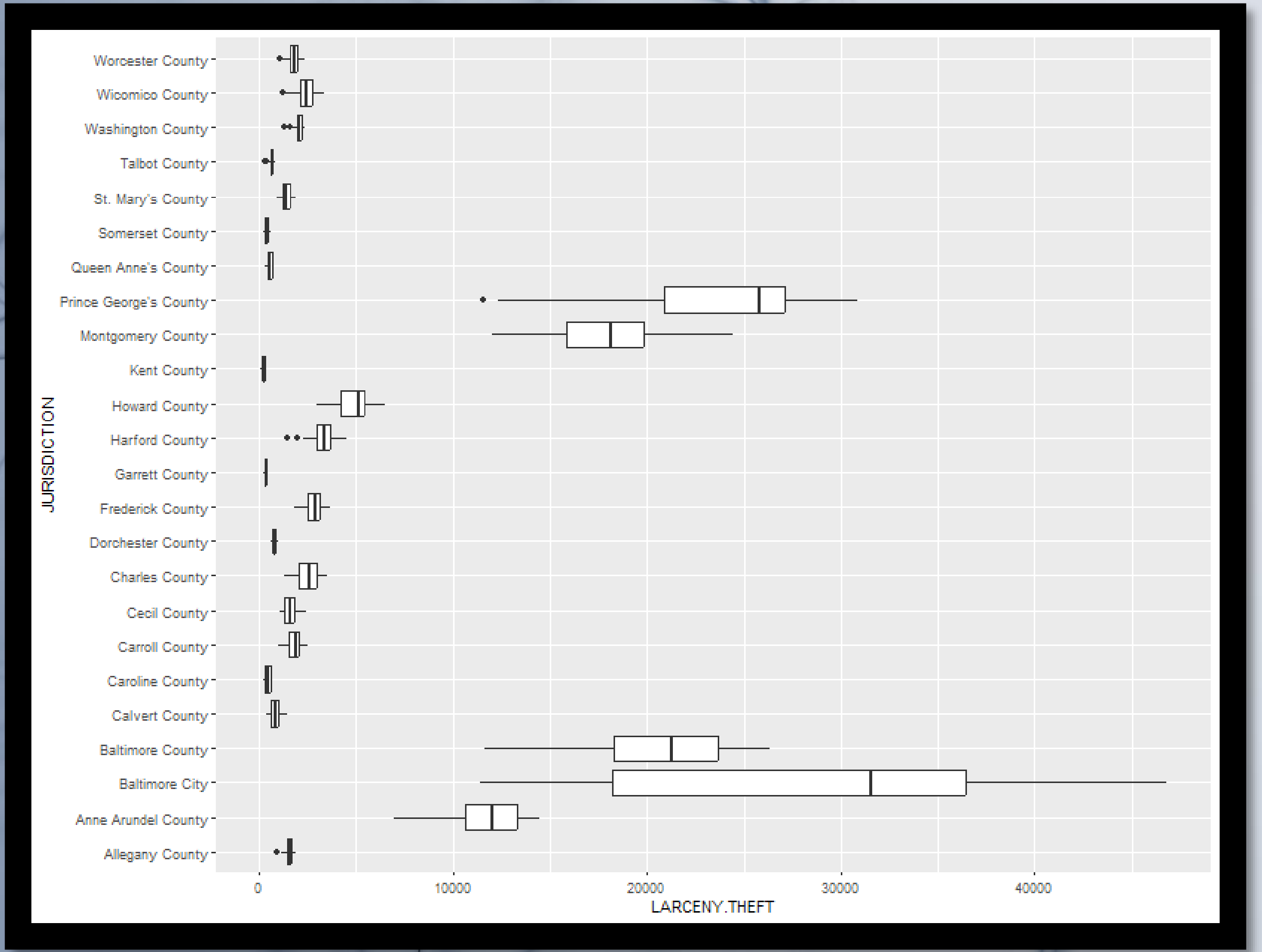
- Columns containing the date were all renamed Year
- Total Crime columns for all three states needed to be renamed
- All states data were combined into their own subset
- State Column added to each state data set to identify them when combined

DATA EXPLORATION

Unemployment vs Crime State by State

Multivariate





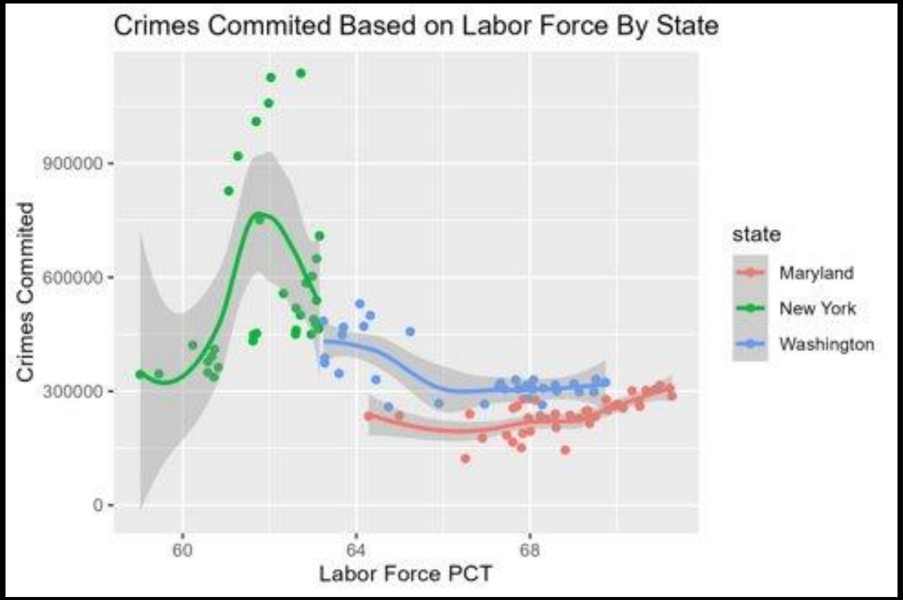
Boxplot of Larceny Theft of Maryland Counties

Bivariate

DATA EXPLORATION CONT.

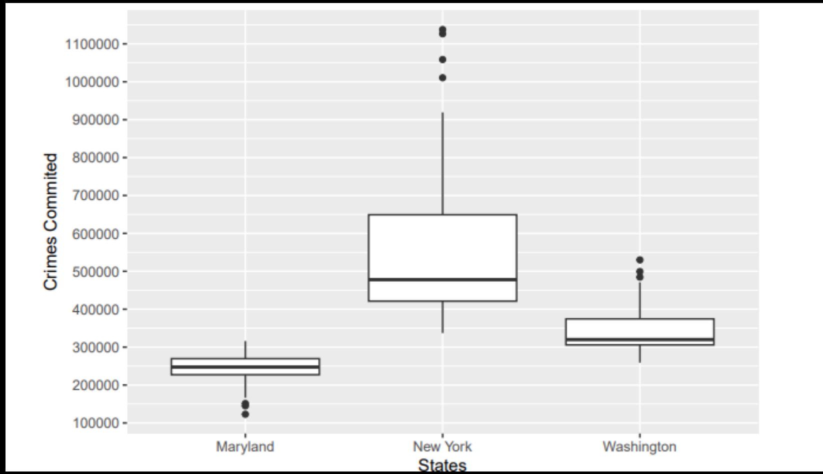
Scatterplot of Labor Force vs Crime State by Sate

Multivariate



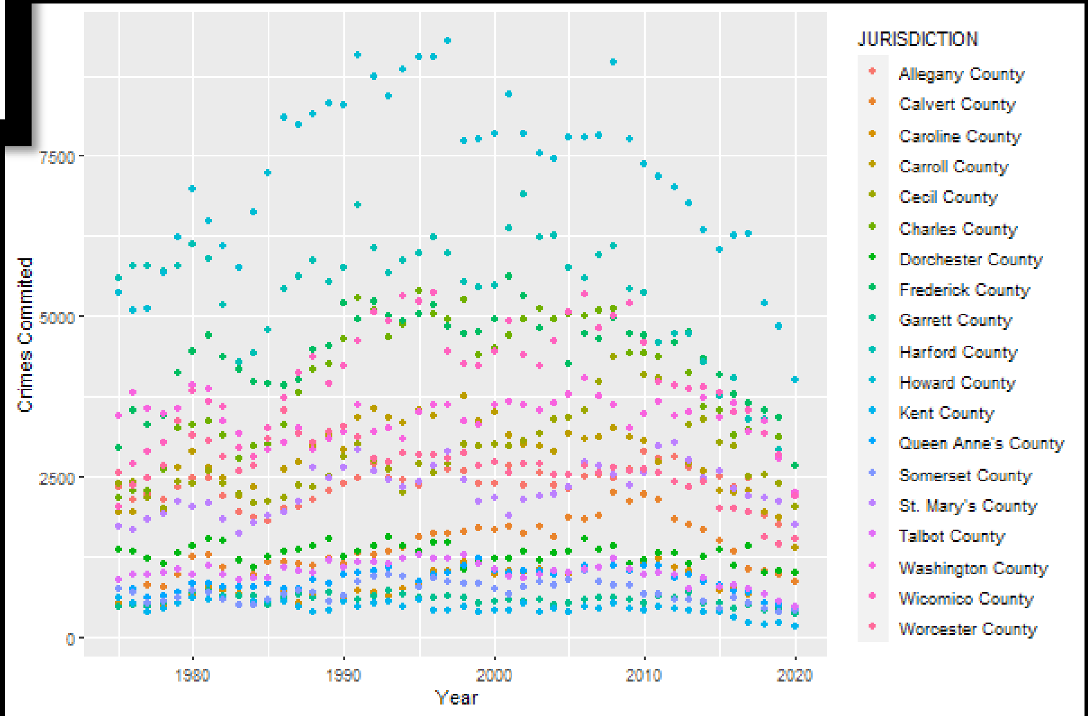
Boxplot of Crime Distribution by State

Univariate



Scatterplot Larceny Crime Maryland W/ Pop Cap

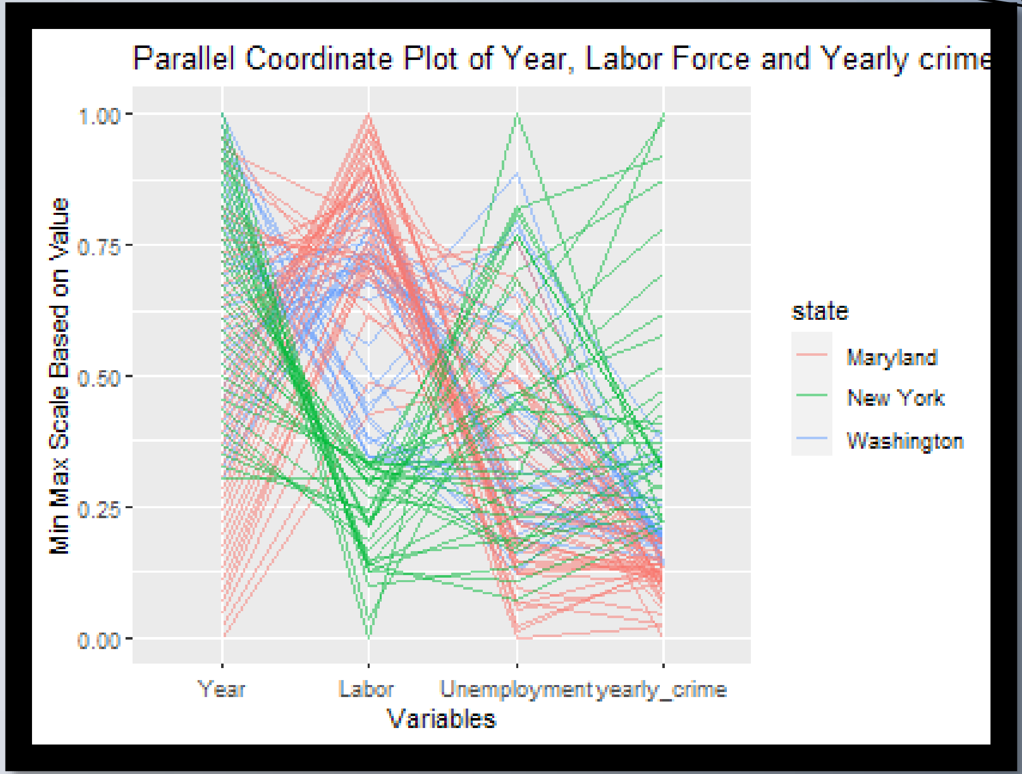
Multivariate

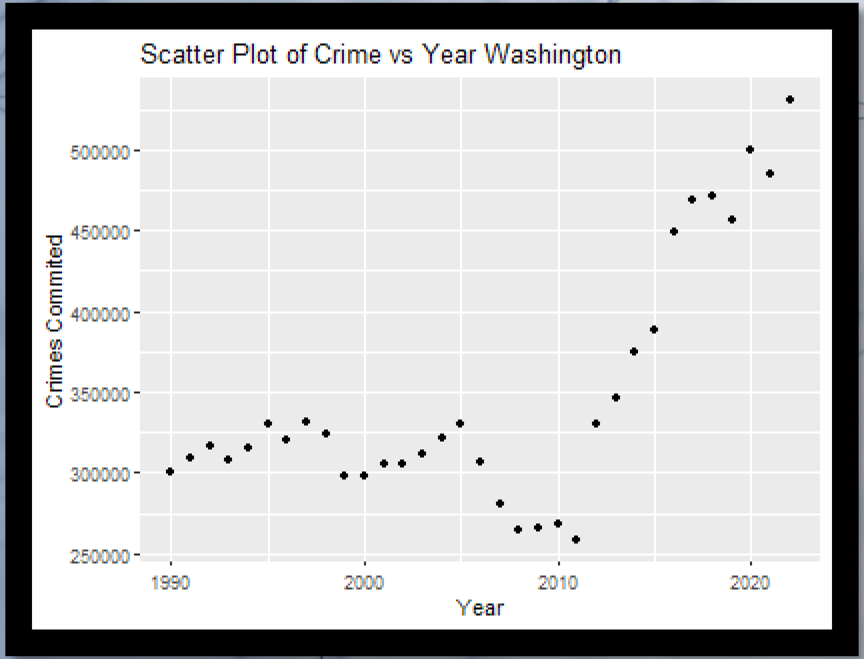


DATA EXPLORATION

Plot of Year, Labor, Unemployment and Yearly Crime

Multivariate





Scatterplot of Crime vs Year in Washington

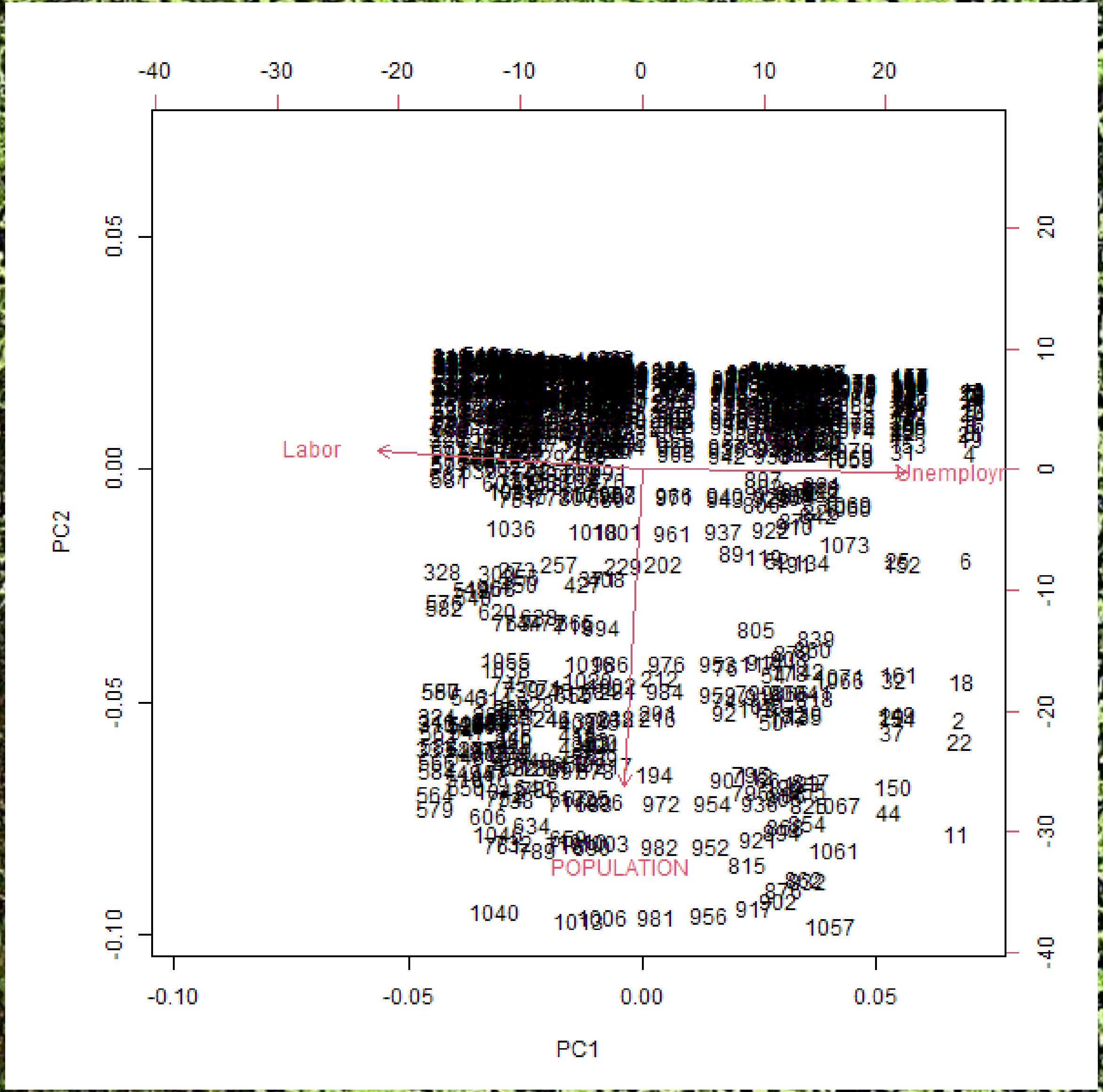
Bivariate

MODELS USED

- PCA
- Negative Binomial

Model 1:PCA

In this machine learning model, I tested to see if certain crimes could be predicted better than others. I used unemployment, labor force percentage and population as my independent variables. To the right is a biplot of the PCA with my independent variables



Model 1:PCA

```
[1] "Larceny Theft"

Call:
lm(formula = grand.total ~ ., data = regression_crime)

Residuals:
     Min       1Q   Median       3Q      Max 
-16996.1   -966.3    -17.8    757.0  25928.9 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  5558.87     111.92  49.670 < 0.0000000000000002 ***
PC1          -576.66      95.64  -6.029   0.0000000022596247 ***
PC2         -7682.35     111.99 -68.599 < 0.0000000000000002 ***
PC3          1102.37     141.08   7.814   0.0000000000000132 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3678 on 1076 degrees of freedom
Multiple R-squared:  0.817,	Adjusted R-squared:  0.8165 
F-statistic:  1601 on 3 and 1076 DF,  p-value: < 0.00000000000000022
```

```
[1] "Overall Crime"

Call:
lm(formula = grand.total ~ ., data = regression_crime)

Residuals:
   Min     1Q Median     3Q    Max 
-37070  -2195    -81   1481  57990 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   9931.2      266.4  37.278 < 0.0000000000000002 ***
PC1          -1138.3      227.7  -5.000          0.000000670085 ***
PC2         -14178.6      266.6 -53.187 < 0.0000000000000002 ***
PC3           2108.7      335.8   6.279         0.000000000494 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8755 on 1076 degrees of freedom
Multiple R-squared:  0.7289,	Adjusted R-squared:  0.7282 
F-statistic: 964.4 on 3 and 1076 DF,  p-value: < 0.00000000000000022
```

My dependents were MURDER, ROBBERY, AGG..ASSAULT, B...E, LARCENY.THEFT, M.V.THEFT, GRAND.TOTAL, VIOLENT.CRIME.TOTAL. My control was overall crime. I then fit a linear model to each type of crime and compared the r scores. This is a measure for variability explained by a model. The higher the score the better the model can have an accurate prediction. In the end the only dependent to have a higher r score than the control was Larceny theft.

Model 2: Negative Binomial

```
Summary for 1 :

Call:
glm.nb(formula = Index_Count ~ Unemployment + Labor + Population,
    data = subset_data, init.theta = 8.974178909, link = log)

Coefficients:
                Estimate  Std. Error z value                Pr(>|z|)
(Intercept)  -5.473723187  0.822162781  -6.658      0.0000000000278 ***
Unemployment  0.056990798  0.009332048   6.107      0.0000000010152 ***
Labor         0.155498251  0.013153770  11.822 < 0.0000000000000002 ***
Population    0.000050008  0.000001133  44.120 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(8.9742) family taken to be 1)

    Null deviance: 2149.13  on 527  degrees of freedom
Residual deviance:  543.34  on 524  degrees of freedom
AIC: 7073.9

Number of Fisher Scoring iterations: 1


              Theta:  8.974
          Std. Err.:  0.560

 2 x log-likelihood:  -7063.886
```

```
Summary for 4 :

Call:
glm.nb(formula = Index_Count ~ Unemployment + Labor + Population,
    data = subset_data, init.theta = 3.593792842, link = log)

Coefficients:
                   Estimate     Std. Error z value               Pr(>|z|)
(Intercept)  -2.00031187889  1.28650821502  -1.555                  0.12
Unemployment  0.07654103919  0.01459819719   5.243 0.000000157828696834 ***
Labor         0.16783529488  0.02054972257   8.167 0.000000000000000315 ***
Population    0.00000121380  0.00000003351  36.221 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.5938) family taken to be 1)

    Null deviance: 1729.35  on 527  degrees of freedom
Residual deviance:  552.26  on 524  degrees of freedom
AIC: 11305

Number of Fisher Scoring iterations: 1


              Theta:  3.594
          Std. Err.:  0.212

 2 x log-likelihood:  -11295.297
```

For my second model I investigated if population would have an effect on a model's ability to predict crime. In this model I used data from New York State and broke it into four population groups. I settled on a Negative Binomial Regression because of the high variability in the data. When I attempted a Poisson Regression the deviance and AIC were way too high to draw accurate conclusions. In the end I discovered that as population increases it decreases the accuracy of a model to predict crime. I measured this with the AIC which is a measure of goodness of fit for the model.

Model 3: Negative Binomial

```
Call:
glm.nb(formula = Index_Count ~ Unemployment + Labor + Population,
    data = newyork_noagg, init.theta = 1.610919759, link = log)

Coefficients:
                   Estimate      Std. Error  z value                Pr(>|z|)    
(Intercept)   -2.39546887242   0.97532708750  -2.456                   0.014 *  
Unemployment   0.05798159067   0.01107848971   5.234             0.000000166 ***
Labor          0.15302227399   0.01559176446   9.814 < 0.0000000000000002 ***
Population     0.00000253563   0.00000003349  75.722 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.6109) family taken to be 1)

    Null deviance: 8371.3  on 2045  degrees of freedom
Residual deviance: 2250.3  on 2042  degrees of freedom
AIC: 37436

Number of Fisher Scoring iterations: 1


              Theta:  1.6109 
          Std. Err.:  0.0462 

 2 x log-likelihood:  -37425.6080 
```

```
Call:
glm.nb(formula = yearly_crime ~ Unemployment + Labor + yearly_pop,
    data = newyork, init.theta = 56.86523968, link = log)

Coefficients:
                   Estimate      Std. Error  z value               Pr(>|z|)    
(Intercept)   21.23907739861   1.85795505327  11.431 <0.0000000000000002 ***
Unemployment   0.02123500997   0.01494943088   1.420                0.155    
Labor          0.02274002705   0.02256394563   1.008                0.314    
yearly_pop    -0.00000050257   0.00000003969 -12.663 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(56.8652) family taken to be 1)

    Null deviance: 261.075  on 32  degrees of freedom
Residual deviance:  33.096  on 29  degrees of freedom
AIC: 841.17

Number of Fisher Scoring iterations: 1


              Theta:  56.9 
          Std. Err.:  14.0 

 2 x log-likelihood:  -831.167 
```

In my third model I went with another Negative Binomial Regression to explore if aggregating crime data would influence a model's ability to predict crime. For this model I used the data from New York State. I found that aggregated crime data was more accurate because its AIC was lower, and its deviance was close to the degrees of freedom.

PROJECT CONCLUSION

After all the data exploration, model development and hypothesis testing I came to a couple conclusions. First, I do believe that you can predict crime with Unemployment and Labor Force Prediction. While they cannot predict crime with precision, I do believe they explain a significant portion and rather paint with a wide brush the general trends of crime. Second, I believe that these economic factors can be used to predict certain crimes better than others. This becomes evident when looking back at my first model where Larceny theft has the highest accuracy and murder with the lowest. Explaining why someone stole is a lot more predictable than explaining why someone commits murder. There are clear economic factors that would lead someone to steal like if they lost their job or can't find one. Third and last, the population of county drastically effects the ability for Unemployment and Labor to project future crimes. This is makes sense when you think about how smaller counties have less people and thus are less impacted by economic conditions.

THANK YOU

SOMEONE@EXAMPLE.COM