New York City Crime Statistics

Description of the Data

The dataset we will use for this project is the “NYPD Complaint Data Historic,” created by the Police Department of New York. The Police Department of New York recorded the data from 2006 to the end of the year 2019. The Police Department of New York is responsible for New York City’s five central boroughs, including Brooklyn, Staten Island, Bronx, Queens, and Manhattan. The dataset contains over six million observations with thirty-five variables. Some of the variables include the victim’s race, sex, and age group, even the borough that the crime took place. Some of the other variables include a description of the offense with its severity and the type of location it took place in. The dataset was obtained by NYC Open Data, a website that holds free public data published by New York City’s agencies.

Research Objective

New York City is one of the largest cities in America by population, economy, and cultural influence. It is home to roughly eight million people with diverse backgrounds. It is a city that attracts people to move and tourists to visit every day. As one of the largest economic and cultural capitals of America, NYC strongly influences America. It is important to understand the factors that influence the city.

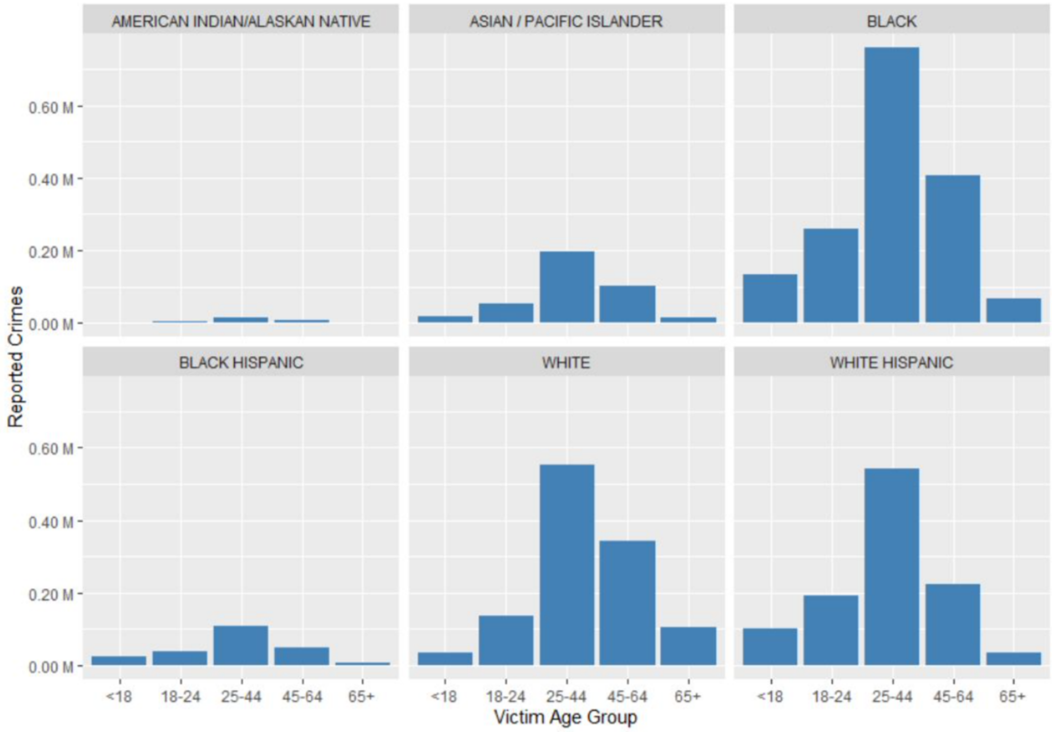
We will be looking at one aspect of the city, its notorious levels of crime. Many stories have emerged of the dangers of big cities. We want to analyze this data to better understand not just the city but the complexities of crime and the several factors that influence it. With curiosity among us, we are using this project to analyze just that with our data. We will look at the statistics of crimes through several perspectives, Geography, Demographics, Time, Societal.

Discussion of Methods

The methods used in the project are Logistic Regression, Time-Series, and PCA. Logistic Regression usage will allow us to see any relationships between the dependent and independent variables and their odds. Time-Series will enable us to see the time frame of crimes and reporting and also forecasting the crime. At the same time, PCA allows to see if there was any correlation between crimes.

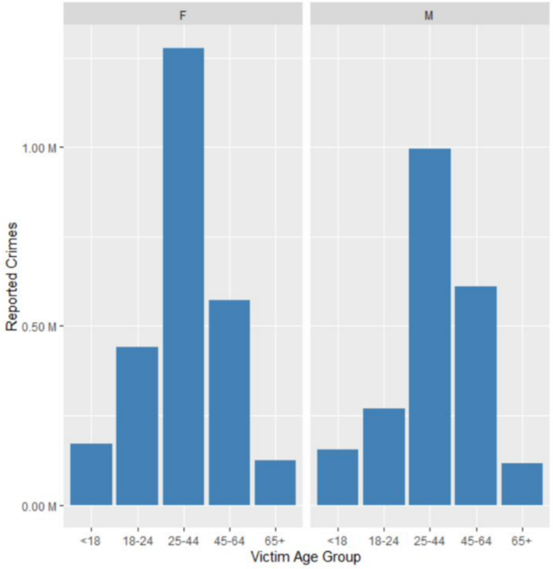
Demographics

Victim Ethnicity



In our data we had over six million reports, we then divided the data by age group based on ethnicity. Most of the crimes reported were mainly with white, white Hispanics and black, but within those groups, the ages of 25-44 had the highest, while ages 45 and plus the crime started to decrease. American Indian/ Alaskan Native had the least number of crimes per age group. The second least happened to be Asian / Pacific Islander, followed by black Hispanic.

Victim Sex



Now looking at it by the gender of a person, it was notably high for females in the age group of 25-44 to be a victim of a crime. Ages lower than 18 was the least, but it starts to go up from 18-24, but it goes back down when it hits age 45 and plus. When it came to males, 25-44 was the highest, second highest was 45-64, then coming down at age 65 plus. Even though ages 18 and less was low, it was higher than 65 and plus.

Logistic Regression

When it came to our data, in the factor of offense, there was 55 different factors, we then decided to reduce it to only two factors. Offense was our dependent variable.

```
crime1<- crime1[crime1$OFNS_DESC==c("HARRASSMENT 2","ASSAULT 3 & RELATED
OFFENSES"),]
```

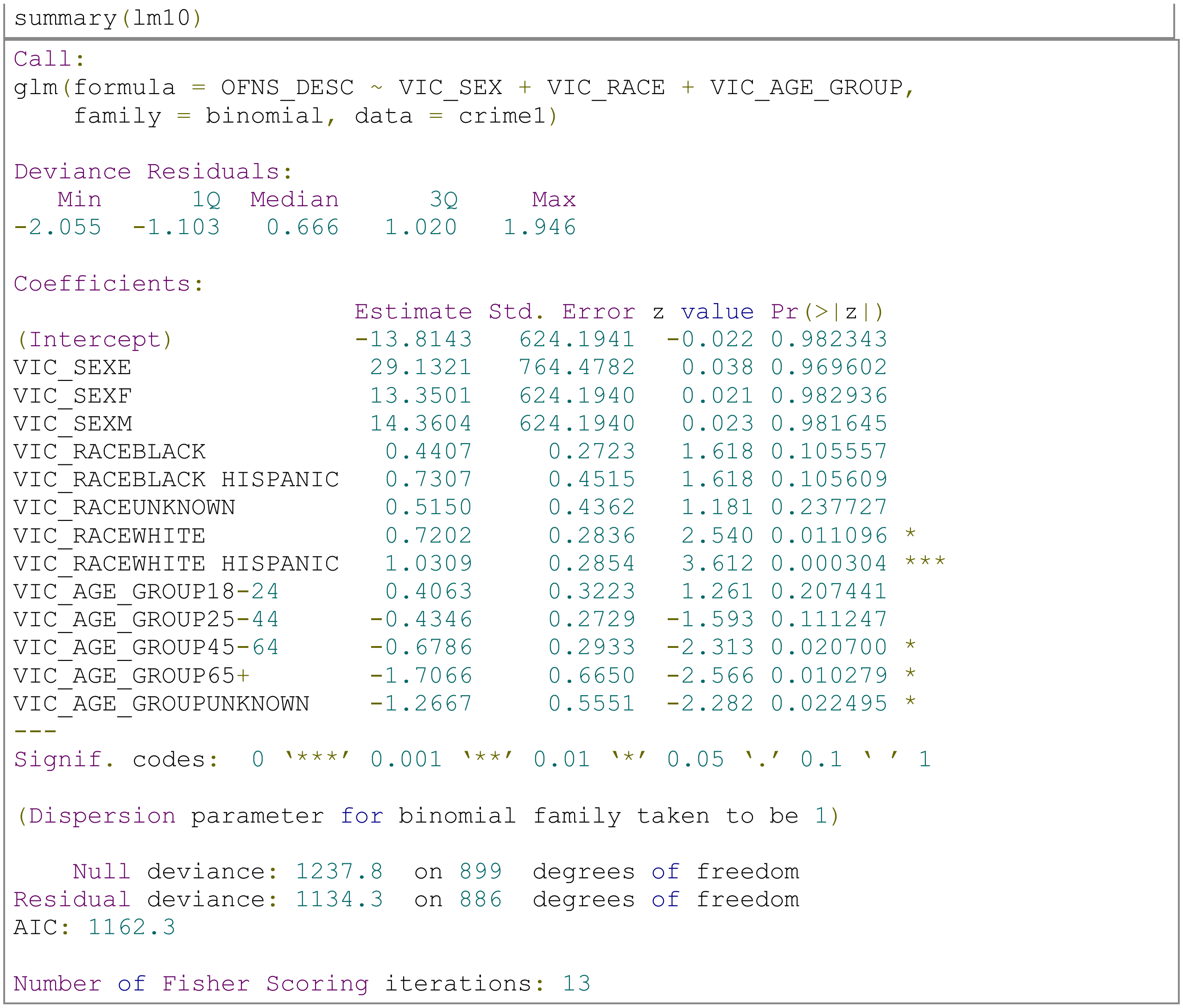
Harassment and Assault were the two factors, since there was a lot of reporting of it. To start the regression, we called and ifelse statement to replace the factors with 0s and 1s.

```
crime1$OFNS_DESC=ifelse(crime1$OFNS_DESC=="HARRASSMENT 2",0,1)
crime1$OFNS_DESC=factor(crime1$OFNS_DESC)
```

Our independent variables were victim sex, victim race and victim age group. We picked these, because we want to see these factors and the odds of them being a victim of a crime.

H0: There is no association between Crime Offense and Victim's demographic; α = 0.05
Ha: There is an association between Crime Offense and Victim's demographic; α = 0.05

```
lm10=glm(OFNS_DESC ~ VIC_SEX +VIC_RACE+VIC_AGE_GROUP,
family=binomial,data=crime1)
```



```
summary(lm10)
```

```
Call:
glm(formula = OFNS_DESC ~ VIC_SEX + VIC_RACE + VIC_AGE_GROUP,
    family = binomial, data = crime1)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.055   -1.103   0.666    1.020    1.946

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -13.8143   624.1941  -0.022 0.982343
VIC_SEXE                29.1321   764.4782   0.038 0.969602
VIC_SEXF                13.3501   624.1940   0.021 0.982936
VIC_SEXM                14.3604   624.1940   0.023 0.981645
VIC_RACEBLACK            0.4407     0.2723   1.618 0.105557
VIC_RACEBLACK HISPANIC   0.7307     0.4515   1.618 0.105609
VIC_RACEUNKNOWN          0.5150     0.4362   1.181 0.237727
VIC_RACEWHITE            0.7202     0.2836   2.540 0.011096 *
VIC_RACEWHITE HISPANIC   1.0309     0.2854   3.612 0.000304 ***
VIC_AGE_GROUP18-24       0.4063     0.3223   1.261 0.207441
VIC_AGE_GROUP25-44      -0.4346     0.2729  -1.593 0.111247
VIC_AGE_GROUP45-64      -0.6786     0.2933  -2.313 0.020700 *
VIC_AGE_GROUP65+        -1.7066     0.6650  -2.566 0.010279 *
VIC_AGE_GROUPUNKNOWN    -1.2667     0.5551  -2.282 0.022495 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1237.8  on 899  degrees of freedom
Residual deviance: 1134.3  on 886  degrees of freedom
AIC: 1162.3

Number of Fisher Scoring iterations: 13
```

Based on the results we got, we concluded that white and white Hispanic are significant, as to also the age groups of 45 and plus and to even unknown. While alpha is 0.05, the p-value for white was less, at 0.011096, and for white Hispanic at 0.000304, while VIC_AGE_GROUP45-64 had a p-value of 0.020700 and VIC_AGE_GROUP65+ had a p-value of 0.010279, lastly VIC_AGE_GROUPUNKNOWN had 0.022495. Meaning that these factors ended up rejecting the null hypothesis, concluding that there is an association between crime offense and the victim’s demographic. The other factors such as VIC_SEXE, VIC_SEXF, VIC_SEXM, VIC_RACEBLACK, VIC_RACEBLACK HISPANIC, VIC_RACEUNKNOWN, VIC_AGE_GROUP18-24, and VIC_AGE_GROUP25-44 had p-value higher than alpha, meaning that they failed to reject the null hypothesis.

Our Logistic Regression Model

```
Model1 = logit(p/1-p)=13.8143245+ 29.1321+ 13.3500614+ 14.3603813+ 0.4406906+
0.7306793 +0.5150016+0.7202155+1.0308666+0.4063325-0.4346253-0.6786154-
1.7066153-1.2667448
```

Coefficients

(Intercept)	**VIC_SEXE**	**VIC_SEXF**	**VIC_SEXM**	**VIC_RACEBLACK**
-13.8143245	29.1321355	13.3500614	14.3603813	0.4406906
VIC_RACEBLACK HISPANIC	**VIC_RACEUNKNOWN**	**VIC_RACEWHITE**	**VIC_RACEWHITE HISPANIC**	**VIC_AGE_GROUP18-24**
0.7306793	0.5150016	0.7202155	1.0308666	0.4063325
VIC_AGE_GROUP25-44	**VIC_AGE_GROUP45-64**	**VIC_AGE_GROUP65+**	**VIC_AGE_GROUPUNKNOWN**	
-0.4346253	-0.6786154	-1.7066153	-1.2667448	

On average the crime offense increases by 29.13 for a unit increase in VIC_SEXE. While for VIC_SEXF, crime offense increases by 13.35 and for VIC_SEXM, crime offense increases by 14.36. When it comes to race VIC_RACEBLACK increases by 0.44, VIC_RACEBLACK HISPANIC, 0.73, VIC_RACEUNKNOWN, 0.51, VIC_RACEWHITE 0.72, VIC_RACEWHITE HISPANIC 1.03. When it comes to age groups, VIC_AGE_GROUP18-24 increases by 0.40, but for VIC_AGE_GROUP25-44 the average changes by -0.43, while VIC_AGE_GROUP45-64 -0.67, VIC_AGE_GROUP65+ - 1.70, and lastly VIC_AGE_GROUPUNKNOWN -1.26

```
glm.probs1 = predict(lm10, type = "response")
glm.pred1 = character()
glm.pred1[glm.probs1 < 0.5] = "HARRASSMENT 2"
glm.pred1[glm.probs1 > 0.5] = "ASSAULT 3 & RELATED OFFENSES"
```

```
glm.pred1                         ASSAULT 3 & RELATED OFFENSES HARRASSMENT 2
  ASSAULT 3 & RELATED OFFENSES                             340           176
  HARRASSMENT 2                                            157           227
```

We created a confusion matrix, to look into at the true negative and true positives when it came into prediction. True positive was 227 and true negative was 340, this gave us a 63% accuracy rate while having a 37% of error rate. We also wanted to see the multicollinearity to check whether or not any of the factors would have given us any issues. By calling the VIF function, we were able to look just that. None of the factors reached nor went over 10, meaning none of the factors would cause an issue.

	VIC_SEXE	VIC_SEXF	VIC_SEXM	VIC_RACEBLACK
	2.327318e+06	8.666665e+07	8.640171e+07	1.497314e+01
VIC_RACEBLACK HISPANIC	VIC_RACEUNKNOWN	VIC_RACEWHITE	VIC_RACEWHITE HISPANIC	VIC_AGE_GROUP18-24

5.912346e+00	9.657591e+00	1.324390e+01	1.349959e+01	1.170249e+01
VIC_AGE_GROUP25-44	VIC_AGE_GROUP45-64	VIC_AGE_GROUP65+	VIC_AGE_GROUPUNKNOWN	
1.675157e+01	1.376155e+01	6.095080e+00	9.796566e+00	

```
glm.pred1                         D   E   F   M
  ASSAULT 3 & RELATED OFFENSES    0   4 151 361
  HARRASSMENT 2                   2   0 347  35
```

We also did some predictions, in the data there was unknown gender filing such as D and E, in which is why we got them during the prediction process. But we saw that females had 151 when it came to assault and 347 when it came to harassment. For males, there was 361 for assault, while 35 for harassment. We noticed that females had a higher odd when it came for harassment while for males, they had a higher odd for assault.

```
glm.pred1                        <18 18-24 25-44 45-64 65+ UNKNOWN
  ASSAULT 3 & RELATED OFFENSES    46   124   246    93   0       7
  HARRASSMENT 2                   25     8   196   115  14      26
```

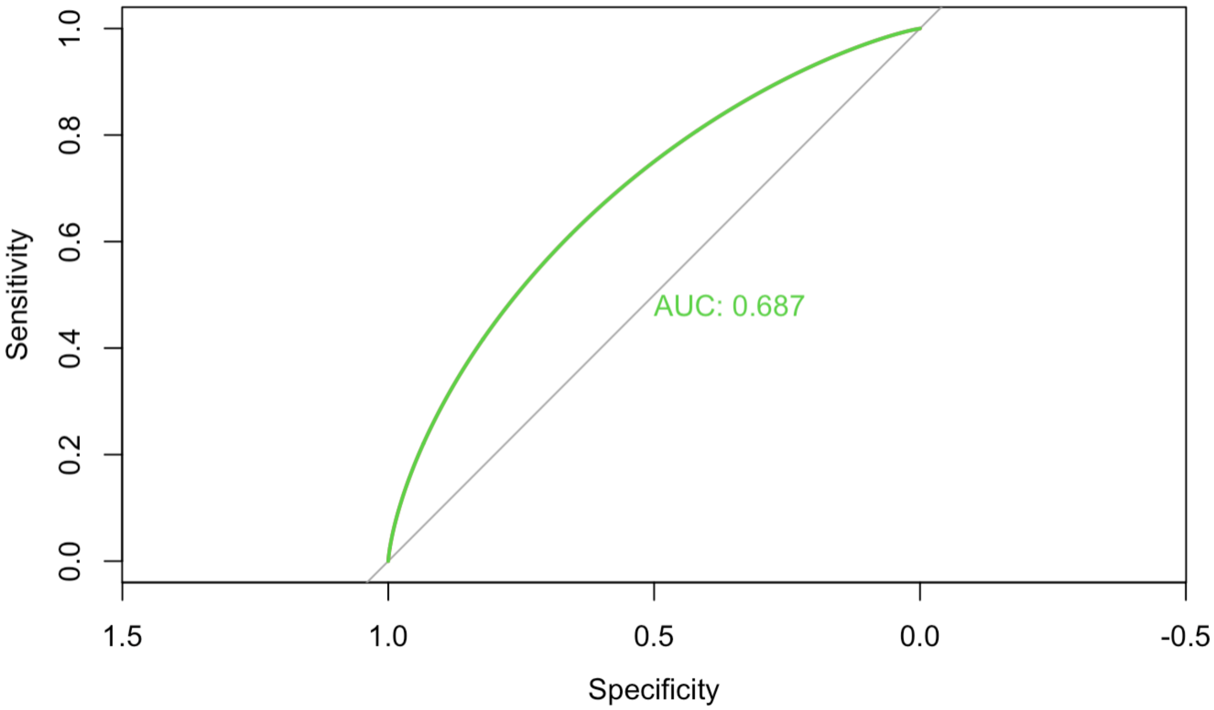
Prediction for age groups were also done, when it came to less than 18, 46 of them went to assault and 25 for harassment. For age group 18 through 24, the number for assault picked up to 124, but 8 for harassment. The next age group of 25-44, there was 246 for assault, the highest for any group. While 196 for harassment. In the age group of 45 through 64, we had 93 go to assault and 115 go to harassment. For 65 and over it was 0 for assault and 14 for harassment. In the data we also got an unknown group due to failure of reporting for the victim, there we got 7 for assault and 26 for harassment.

	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC
ASSAULT 3 & RELATED OFFENSES	25	137	19	17	117	201
HARRASSMENT 2	49	169	11	37	100	18

We also did prediction for ethnicity, for Asian / Pacific Islander, we got 25 for assault and 49 for harassment. While for black we got 137 for assault and 169 for harassment, but for black Hispanic we got 19 for assault and 11 for harassment. For the unknown group we got 17 assault and 37 harassment. White group gave us 117 for assault and 100 for harassment while lastly the white Hispanic group, we got 201 for assault and 18 for harassment.

Lastly, we did an AUC, the area under the curve is 0.687, the model has 68.7% chance of being able to tell the difference between positive and negative class. Not the best model but not that bad either since there is a 68.7% chance.

Area under the Curve



PCA

Principal component analysis (PCA) is to find the subset of variables that best explain the data's variation.PCA is used for analyzing a table of variables with the main idea of converting the observed variables into a set of new variables. The principal components are uncorrelated and explain the variation in the data. For this reason, PCA reduces a complex data set to a lower dimension to reveal the structures of the dominant types of variations in both the observations and the variables.PCA is an unsupervised approach meaning that it is performed on a set of variables X1, X2, …, Xp with no associated response Y. Slashing the number of input variables for a predictive model and allowing most of the variability to be explained using fewer variables are referred to as dimensionality reduction. In the data set, PCA reduces the dimensionality and is commonly used as one step in a series of analyses. PCA can reduce the number of variables and avoid multicollinearity or numerous predictors relative to the number of observations. Some of the significant benefits of PCA are

Dimensionality reduction

We can efficiently perform dimensionality reduction on a high dimensional dataset and then fit a linear regression model to a reduced set of variables by using PCR. Simultaneously, keep most of the variability of the original predictors. The use of only a few of the principal components reduces the number of variables in the model. As a result, this can reduce the model complexity. PCR manages to perform well when the first principal components explain most of the predictors' variation.

Mitigation Overfitting

If all the assumptions underlying PCR hold, then fitting a least-squares model to the principal components will lead to better outcomes than fitting a least-squares model to the original data since most of the variation and information related to the dependent variable is condensed in the principal components and by estimating fewer coefficients, you can reduce the risk of overfitting.

Avoiding multicollinearity

Suppose there is some degree of multicollinearity between the variables in the dataset. In that case, PCR will avoid this obstacle because performing PCA on the raw data produces linear combinations of the uncorrelated predictors.

Potential drawbacks and warnings

A classic mistake is to think of PCR as a feature selection method. It is not a feature selection method because each of the calculated principal components is a linear combination of the original variables. Using principal components instead of the basic features can make it complex to explain what is influencing others.

Another major disadvantage of PCR is that the directions that best represent each predictor are obtained unsupervised. The dependent variable is not used to distinguish each principal component direction. This means that it is not sure that the directions found will be the best possible directions to use when getting predictions on the dependent variable.

Disadvantages of Principal Component Analysis

1. Information Loss:

Even Though Principal Components try to include the highest variance among the elements in a dataset. Still, if we don't choose the number of Principal Components with care, it may lose certain information as related to the original list of features.

2. Independent variables become less interpretable:

After implementing PCA on the dataset, our original features will turn into Principal Components. The principal components are the linear arrangement of our original features. Principal Components are not as readable and interpretable as original elements.

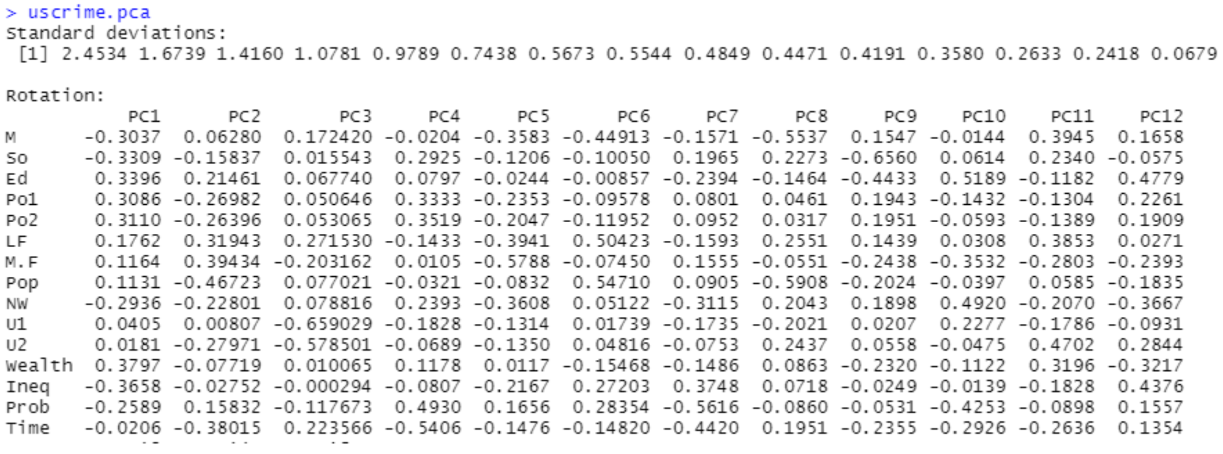
3. Data standardization is essential before PCA:

PCA will not obtain the optimal Principal Components until we standardize our data before PCA implementation. For example, if a feature set has data expressed in Light years, Kilograms, or Millions, the training set's variance scale is enormous. If PCA is operated on such a feature set, the resultant loadings for high variance features will also be significant. Consequently, principal components will be biased towards features with high variance, leading to false results. Also, for standardization, all the categorical features must be converted into numerical features before PCA can be applied. PCA is affected by scale, so we need to scale our data features before using PCA.

Converting to PCA:

For PCA analysis, I have taken two different datasets. Besides the NYC crimes dataset, I have taken one additional dataset for analysis from http://www.statsci.org/. With 79 variables and around 7 Million records, 1st dataset is a good candidate for dimensionality reduction. With 10 Principle vomponent 2nd dataset is a good candidate for Principle Component reduction. When applying prcomp R function scaling, the scree plot is used to determine the number of factors to retain in exploratory factor analysis or principal components to keep a principal component analysis centering set to TRUE. A Scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. Scaling is necessary to ensure that higher denomination data points don't overly influence the model, so we scale to provide all the data attribute's data points are in the same range. Summary of PCAs for 1st dataset in Fig 2B indicates that PC1 explains 75.5 % of the variance, PC2 17.4% of the variance, PC3 7.09 of the variance etc. The variance proportion decreases with the PC values and after PC3 there is no substantial marginal increase. Some of the significant findings on data correlation are Kinapping and Felony sex crime are positively corelated. Loitering for drug , prostitution and Alcohol control law violation are positively correlated.Summary of PCAs for 2nd dataset in Fig 2Aindicates that PC1 explains 40% of the variance, PC2 18%, etc. The variance proportion decreases with the PC values and after PC7 there is no substantial marginal increase. This will be used to pick the PCAs to use and be more clearly displayed by plotting a graph as shown in the following analysis.

Studying the output of the PCA we can find various exciting factors. Consider 2nd dataset PC1, for example. The most important predictors in PC1 are Wealth which has a correlation coefficient of 0.3797, and the lowest correlation is Ineq which correlates -0.3658. This indicates that when PC1 increases, Wealth increases with and Ineq decreases as it is negatively correlated. We also find out that Po1 and Po2 have positive correlation coefficients that are very close, which indicates a close correlation in the positive direction for these two variables. In Wealth and Ineq the correlation coefficient absolute values are relative, showing a correlation, but as the signs are opposite, the correlation is negative in this pair. PCA will ensure that such correlation effects are nullified when creating the models. For PC2 we can see the number of Males per 100 females M.F has a very high positive correlation of 0.39, and state population of 1960 Pop has a negative correlation of -0.47.



```
> uscrime.pca
Standard deviations:
 [1] 2.4534 1.6739 1.4160 1.0781 0.9789 0.7438 0.5673 0.5544 0.4849 0.4471 0.4191 0.3580 0.2633 0.2418 0.0679

Rotation:
           PC1      PC2       PC3     PC4     PC5      PC6     PC7     PC8     PC9    PC10    PC11    PC12
M      -0.3037  0.06280  0.172420 -0.0204 -0.3583 -0.44913 -0.1571 -0.5537  0.1547 -0.0144  0.3945  0.1658
So     -0.3309 -0.15837  0.015543  0.2925 -0.1206 -0.10050  0.1965  0.2273 -0.6560  0.0614  0.2340 -0.0575
Ed      0.3396  0.21461  0.067740  0.0797 -0.0244 -0.00857 -0.2394 -0.1464 -0.4433  0.5189 -0.1182  0.4779
Po1     0.3086 -0.26982  0.050646  0.3333 -0.2353 -0.09578  0.0801  0.0461  0.1943 -0.1432 -0.1304  0.2261
Po2     0.3110 -0.26396  0.053065  0.3519 -0.2047 -0.11952  0.0952  0.0317  0.1951 -0.0593 -0.1389  0.1909
LF      0.1762  0.31943  0.271530 -0.1433 -0.3941  0.50423 -0.1593  0.2551  0.1439  0.0308  0.3853  0.0271
M.F     0.1164  0.39434 -0.203162  0.0105 -0.5788 -0.07450  0.1555 -0.0551 -0.2438 -0.3532 -0.2803 -0.2393
Pop     0.1131 -0.46723  0.077021 -0.0321 -0.0832  0.54710  0.0905 -0.5908 -0.2024 -0.0397  0.0585 -0.1835
NW     -0.2936 -0.22801  0.078816  0.2393 -0.3608  0.05122 -0.3115  0.2043  0.1898  0.4920 -0.2070 -0.3667
U1      0.0405  0.00807 -0.659029 -0.1828 -0.1314  0.01739 -0.1735 -0.2021  0.0207  0.2277 -0.1786 -0.0931
U2      0.0181 -0.27971 -0.578501 -0.0689 -0.1350  0.04816 -0.0753  0.2437  0.0558 -0.0475  0.4702  0.2844
Wealth  0.3797 -0.07719  0.010065  0.1178  0.0117 -0.15468 -0.1486  0.0863 -0.2320 -0.1122  0.3196 -0.3217
Ineq   -0.3658 -0.02752 -0.000294 -0.0807 -0.2167  0.27203  0.3748  0.0718 -0.0249 -0.0139 -0.1828  0.4376
Prob   -0.2589  0.15832 -0.117673  0.4930  0.1656  0.28354 -0.5616 -0.0860 -0.0531 -0.4253 -0.0898  0.1557
Time   -0.0206 -0.38015  0.223566 -0.5406 -0.1476 -0.14820 -0.4420  0.1951 -0.2355 -0.2926 -0.2636  0.1354
```

Fig: 1A.

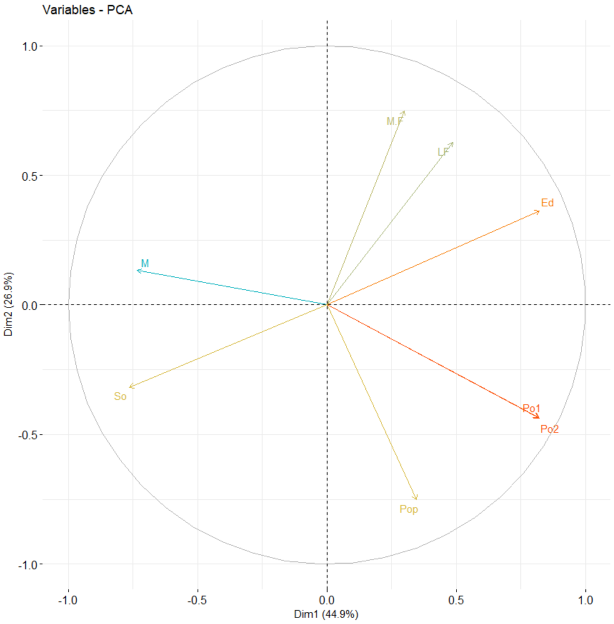


Fig: 2A.

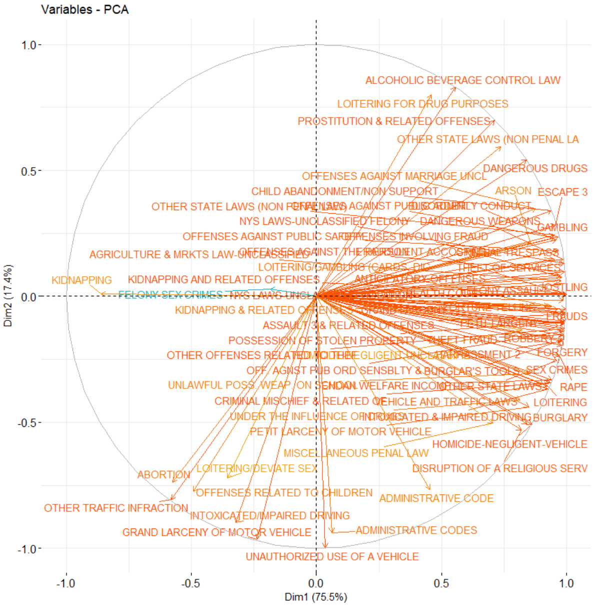


Fig: 2B.

Picking the PCAs:

We want to pick the PCAs which explain the variance in the output of the model results. We want to maximize the amount of variance dictated by the PCAs. . Based on the PCA data's scree plot, we see that the variance prediction's marginal increase drops off after 3 for the 1st dataset in Fig 3B. On the otherhand, based on the PCA data's scree plot, we see that the variance prediction's marginal increase drops off after 7 for the 2nd dataset in Fig 3A. So an ideal value to pick would be between 4 and 7. We would prefer 5 as the value in the median range between 4 and 7 and see that it explains about 6% of the output variance. 6% is slightly better than the 5% threshold we usually chose when filtering out statistically significant variances. Plotting the cumulative variance vs. the principal component also indicates that the proportion of variance explained by PCA diminishes after 6.

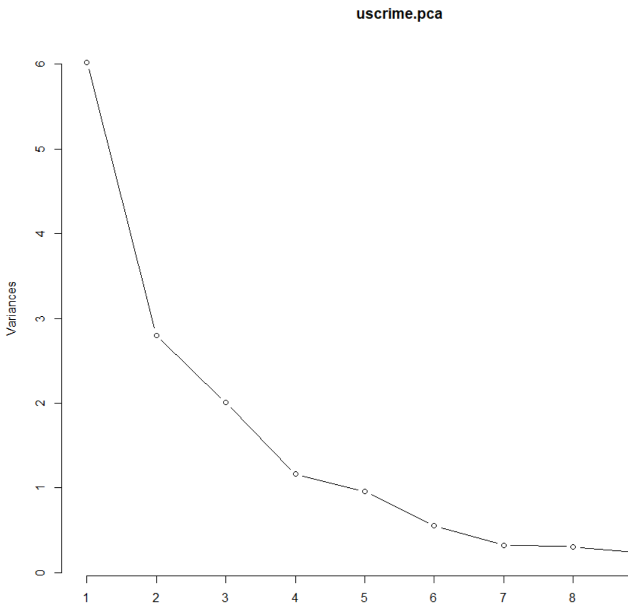


Fig:3A

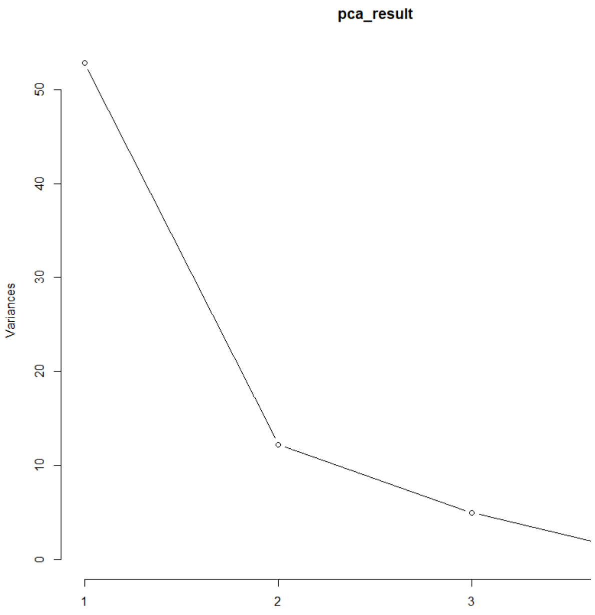


Fig:3B

Cross-validation:

We are using the cv.lm R function in the DAAG library, we can perform cross-validation on the models to verify their quality and can compare the R squared and Adjusted R Square values to confirm which model has better quality. The overall ms value for model1 is 75069, and the overall ms value for model2 is 66756, which indicates that model1 is doing better. In both cases, fold2 has the best outcomes for cross-validation.

model1_cv <- cv.lm(PCcrime1,model1,m=5) Analysis of Variance Table Response: Crime Df Sum Sq Mean Sq F value Pr(>F)	> model_cv <- cv.lm(PCcrime2,model2,m=5) Analysis of Variance Table Response: Crime

PC1 1 1177568 1177568 19.78 6.5e-05 ***	Df Sum Sq Mean Sq F value Pr(>F)
PC2 1 633037 633037 10.63 0.0022 **	PC1 1 1177568 1177568 19.78 6.3e-05 ***
PC3 1 58541 58541 0.98 0.3272	PC2 1 633037 633037 10.64 0.0022 **
PC4 1 257832 257832 4.33 0.0437 *	PC4 1 257832 257832 4.33 0.0435 *
PC5 1 2312556 2312556 38.84 2.0e-07 ***	PC5 1 2312556 2312556 38.85 1.8e-07 ***
Residuals 41 2441394 59546	Residuals 42 2499935 59522
---	---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
fold 1	fold 1
Observations in test set: 9	Observations in test set: 9
1 4 8 9 18 20 23 32 47	1 4 8 9 18 20 23 32 47
Predicted 714 1745 1158 862.7 1098 1238.8 768 970 1139	Predicted 741.9 1739 1143 830 1108 1310.0 750 984 1179
cvpred 686 1698 1114 807.9 1089 1245.7 732 945 1206	cvpred 696.5 1700 1110 795 1096 1276.8 725 953 1225
Crime 791 1969 1555 856.0 929 1225.0 1216 754 849	Crime 791.0 1969 1555 856 929 1225.0 1216 754 849
CV residual 105 271 441 48.1 -160 -20.7 484 -191 -357	CV residual 94.5 269 445 61 -167 -51.8 491 -199 -376

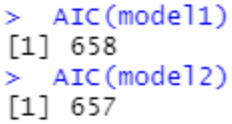
Sum of squares = 706357 Mean square = 78484 n = 9	Sum of squares = 735120 Mean square = 81680 n = 9
fold 2	fold 2
Observations in test set: 10	Observations in test set: 10
5 13 15 17 25 34 39 40 42 46	5 13 15 17 25 34 39 40 42 46
Predicted 1004 669 663 468 604.2 842 628 1069.9 272 927	Predicted 972 622.5 689 502.5 558.08 860.8 623 1049 314 902
cvpred 1020 691 607 406 609.3 815 588 1074.2 185 927	cvpred 963 607.6 653 466.3 530.44 846.2 584 1037 263 886
Crime 1234 511 798 539 523.0 923 826 1151.0 542 508	Crime 1234 511.0 798 539.0 523.00 923.0 826 1151 542 508
CV residual 214 -180 191 133 -86.3 108 238 76.8 357 -419	CV residual 271 -96.6 145 72.7 -7.44 76.8 242 114 279 -378
Sum of squares = 517100 Mean square = 51710 n = 10	Sum of squares = 407907 Mean square = 40791 n = 10
fold 3	fold 3
Observations in test set: 10	Observations in test set: 10
2 3 11 14 16 22 28 31 33 38	2 3 11 14 16 22 28 31 33 38
Predicted 1196 506.4 1310 653.8 933.8 770 1015 688 723 604.3	Predicted 1197 515.7 1271 627.9 979.9 767 1012 761 687 613.5
cvpred 1161 560.1 1205 618.9 994.2 815 994 765 697 622.2	cvpred 1161 558.9 1209 620.8 990.4 814 994 759 700 621.1

Crime 1635 578.0 1674 664.0 946.0 439 1216 373 1072 566.0 CV residual 474 17.9 469 45.1 -48.2 -376 222 -392 375 -56.2 Sum of squares = 936438 Mean square = 93644 n = 10 fold 4 Observations in test set: 9 19 21 26 27 29 30 36 44 45 Predicted 975 806 1846 480 1464 802 978 1126.3 425.5 cvpred 1135 820 1764 573 1664 818 1045 1113.6 418.8 Crime 750 742 1993 342 1043 696 1272 1030.0 455.0 CV residual -385 -78 229 -231 -621 -122 227 -83.6 36.2 Sum of squares = 719985 Mean square = 79998 n = 9 fold 5 Observations in test set: 9 6 7 10 12 24 35 37 41 43	Crime 1635 578.0 1674 664.0 946.0 439 1216 373 1072 566.0 CV residual 474 19.1 465 43.2 -44.4 -375 222 -386 372 -55.1 Sum of squares = 926214 Mean square = 92621 n = 10 fold 4 Observations in test set: 9 19 21 26 27 29 30 36 44 45 Predicted 951 799 1866 448 1445 765.5 958 1143 510 cvpred 1078 812 1801 514 1615 763.1 1012 1141 556 Crime 750 742 1993 342 1043 696.0 1272 1030 455 CV residual -328 -70 192 -172 -572 -67.1 260 -111 -101 Sum of squares = 600715 Mean square = 66746 n = 9 fold 5 Observations in test set: 9 6 7 10 12 24 35 37 41 43

Predicted 901 818 906 831.7 929.0 915 1212 841.5 1043 cvpred 856 785 960 886.8 911.7 898 1422 913.8 1150 Crime 682 963 705 849.0 968.0 653 831 880.0 823 CV residual -174 178 -255 -37.8 56.3 -245 -591 -33.8 -327 Sum of squares = 648385 Mean square = 72043 n = 9 Overall (Sum over all 9 folds) ms 75069	Predicted 888 841 897 801.8 1004.2 959 1152 795.7 1007 cvpred 855 835 929 832.8 1042.7 995 1264 823.9 1063 Crime 682 963 705 849.0 968.0 653 831 880.0 823 CV residual -173 128 -224 16.2 -74.7 -342 -433 56.1 -240 Sum of squares = 467577 Mean square = 51953 n = 9 Overall (Sum overall 9 folds) ms 66756

AIC comparison

Model 1 has an AIC value of 658, and Model 2 has an AIC value of 657. A comparatively lower value of AIC indicates a better model as AIC punishes the models for having too many variables. So there is an opportunity to reduce the number of PCAs used to build the model.



```
>  AIC(model1)
[1] 658
>  AIC(model2)
[1] 657
```

Conclusion

My project's primary goal is to maximize the amount of variance dictated by the PCAs. Based on the PCA data's scree plot, we see that the variance prediction's marginal increase drops off after 7. Therefore an ideal value to pick would be between 4 and 7. On top of it, the model comparison indicates that there is still an chance to lower the amount of PCAs used to build the model.

```
R code:
install.packages("dplyr")
install.packages("factoextra", dependencies=TRUE)
install.packages("mlbench", dependencies=TRUE)
install.packages("tidyverse", dependencies=TRUE)
install.packages("GGally", dependencies=TRUE)
library(dplyr)
library(tidyverse)
library(mlbench)
library(tidyverse)  # data manipulation and visualization
library(gridExtra)  # plot arrangement
library(magrittr)
library(factoextra)
library(GGally)
###################################LoadDataset##############################
## 2 different dataset
NYPD_Complaint_Data_Historic <-
read.csv("C:/Users/sasoy/Downloads/NYPD_Complaint_Data_Historic.csv")
usacrime <- read.delim("http://www.statsci.org/data/general/uscrime.txt")
##View(usacrime)
##??usacrime
#############################################################################
####
###1st sataset processing and cleanup
#############################################################################
####
crimedata<- NYPD_Complaint_Data_Historic
crimedata<-crimedata %>% select(c(BORO_NM, OFNS_DESC))
crimedata<-crimedata %>%
group_by(BORO_NM,OFNS_DESC)%>% summarise(Count = n())
crimedata[is.na(crimedata)] =0
crimedata$Count <- as.numeric(crimedata$Count)
crimedata<-subset(crimedata, crimedata$BORO_NM != "")   ## Remove unknown
location
crimedata<-subset(crimedata, crimedata$OFNS_DESC != "")   ## Remove unknown
location
crimedata <-
union(
union(
union(
union(
  crimedata %>%
    filter(BORO_NM =='BRONX') %>%
  select(BORO_NM,OFNS_DESC,Count)%>%
  mutate( Count=(Count/2648403)*10000 ), ## Population 2648403,
  crimedata %>%
    filter(BORO_NM =='BROOKLYN') %>%
    select(BORO_NM,OFNS_DESC,Count)%>%
    mutate( Count=(Count/2589970)*10000))  ,
crimedata %>%
  filter(BORO_NM =='MANHATTAN') %>%
  select(BORO_NM,OFNS_DESC,Count)%>%
  mutate( Count=(Count/1631990)*10000 )),

crimedata %>%
  filter(BORO_NM =='QUEENS') %>%
  select(BORO_NM,OFNS_DESC,Count)%>%
  mutate( Count=(Count/2287390)*10000)),
crimedata %>%
  filter(BORO_NM =='MANHATTAN') %>%
  select(BORO_NM,OFNS_DESC,Count)%>%
  mutate( Count=(Count/1631990)*10000))
View(crimedata)
##Manhattan Population 2021 1,631,990
##QUEENS 2021 2,287,390
#BRONX 2021 1,435,070
##BROOKLYN 2021 2,589,970
##STATEN ISLAND 2021 474,893
crimedata<-crimedata %>%
spread(OFNS_DESC,Count)
str(crimedata)
crimedata[is.na(crimedata)] =0
crimedata[is.na(crimedata)] =0
crimedata1<-crimedata
str(crimedata1)
crimedata1$BORO_NM<-factor(crimedata1$BORO_NM)
crimedata<-crimedata[,-1]
####################################1stDatasetProcessingCompleted###########
###########
##############################################################################
##############

#pairs(crimedata) ## figure margins too large
#ggpairs(crimedata) ##figure margins too large
#pca for 1st dataset
pca_result <- prcomp(crimedata, center=TRUE,scale = TRUE)  ##
## princomp
pca_result
#pca for 2nd dataset
usacrime.pca <- prcomp(usacrime[,1:15],scale.=TRUE,center=TRUE)
## PCA without response veritable
usacrime.pca
# With connected lines - useful for looking for the "elbow"
plot(pca_result, type = "l")
abline(h=10,col="red")
summary(pca_result)
## decending order of how much component is the reflection of varion of data
## Component 1 is responsible for 66.29% variation of data
## Component 2 is responsible for 26.83% variation of data
## Component 1 , 2 and 3 gives us 100% variation of data

plot(usacrime.pca, type = "l")
abline(h=1,col="red")
summary(usacrime.pca)
```

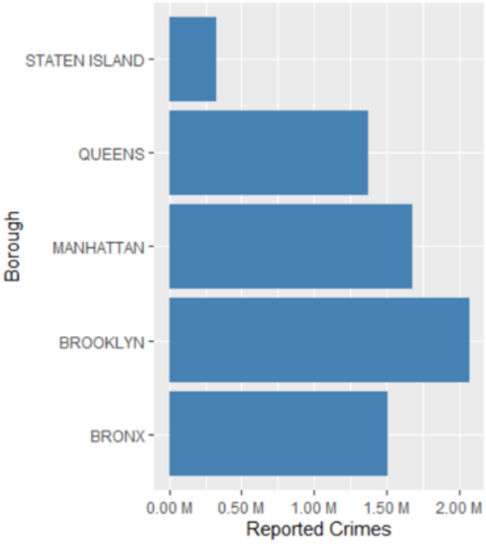
```
#1st Component 40%
#2nd Components 18%
#3rd Components 13%
fviz_pca_var(pca_result,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
pca_result
## Cosign between the vectors are corellation between the variables
## Each component is ploted for PC1 and PC2
# each arrow represent igon vector in our dataframe
## Important Findings:
## This plot help us to interpret the importance of relationship between the
different variables and Pricile component and a good way looking for patters
in data
## Kinapping and Felony sex crime are positively corelated.
## Loitering for drug , prostitution and Alcohol control law violation are
positively correlated.
fviz_pca_var(usacrime.pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
usacrime.pca
##Po1:police expenditure in 1960.Po2 police expenditure in 1959. <>Prob:
probability of imprisonment.
##Ed: mean years of schooling<>NW number of non-whites per 1000 people.
#Ed: mean years of schooling<>Ineq: income inequality.(Schooling decrease
disparity)
#The most important predictors in PC1 are Wealth which has a correlation
#coefficient of 0.3797 and the lowest correlation is Ineq
#which has a correlation of -0.3658. This indicates that
#when PC1 increases Wealth increases with and Ineq (income
inequality)decreases
#as it is negatively correlated. We also see that both Po1 and Po2 have
#correlation coefficients which are very close, that indicates a close
correlation in the positive direction for these two variables. In the case of
Wealth and Ineq the correlation coefficient absolute values are close
indicating a correlation, but as the signs are opposite the correlation is
negative in the case of this pair. PCA will ensure that such correlation
effects are nullified when creating the models.
#In the case of PC2 we can observe that number of Males per 100 females M.F
#has a very high positive correlation of 0.39 and state population of 1960
Pop (state population)
#has a negative correlation of -0.47.
##############################################################################
###
###############LinearRegression###############################################
##
##############################################################################
###
  lm.crime <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob,data=usacrime)
  summary(lm.crime)
##only 3 of them is statistically significant .
#That do not mean that all 4 of them is significant. There might be multi-
colinearity. We deal with it we will use PCA.
```

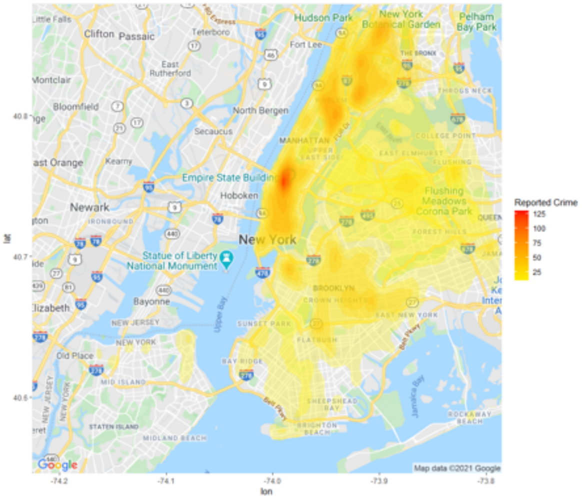
```
 ############################################################
 #linear regression models
  #regression on first 5 PCs
  PCs <- usacrime.pca$x[,1:5]
  PCcrime1 <- cbind(PCs,usacrime[,16])
  PCcrime1 <- as.data.frame(PCcrime1)
  summary(PCcrime1)
  PCs1 <- usacrime.pca$x[,c(1,2,4,5)]
  PCcrime2 <- cbind(PCs1,usacrime[,16])
  PCcrime2 <- as.data.frame(PCcrime2)
  summary(PCcrime2)
  #linear regression models
  Crime <- PCcrime1$V6
    model1 <- lm(Crime~PC1+PC2+PC3+PC4+PC5,data = PCcrime1)
    summary(model1)
    model1
 ## PC3 is not significant.This is very easy to interpte the are all
orthogonal to each other.There is no multicolinearity. I simple can ignore
PC2, PC3,PC4 because there is no multicolinerity.
 #dropped PC3
 model2 <- lm(Crime~PC1+PC2+PC4+PC5,data = PCcrime2)
 summary(model2)
 ##Invistigate can reduce the number of PCAs used to build the model
 #compare two model
 anova(model1,model2)
 ## P value of 0.33 indicate there is no statistically significant different
between model 1 and model 2
 #cross validation : Why? Because :training and test set validation are
fragile outlier can change outcome
 #1. compare the R squared and Adjusted R Square values to verify which model
has better quality.
 library(DAAG)
# cv.lm R function in DAAG library we can perform cross validation on the
#This following function gives internal and cross-validation measures of
predictive accuracy    #for multiple linear regression.
 model1_cv <- cv.lm(PCcrime1,model1,m=5)
 model_cv <- cv.lm(PCcrime2,model2,m=5)
 #The overall ms value for model1 is 75069
 #and the overall ms value for model2 is 66756
 #which suggests that model1 is doing better.
 #In both cases fold2 has the best results for cross validation.
 ##2. AIC comparison
 AIC(model1)
 AIC(model2)
 #Summary :
 #Model1 has an AIC value of 658 and Model 2 has 657. We’re being penalized
for using additional predictors.
```

Geographic

It is helpful to analyze the data not just via demographics but also geography. Brooklyn may have higher crime, but the Brooklyn neighborhood is much larger. It is clear to see that some areas experience very little crime while other areas experience a plethora of crime.

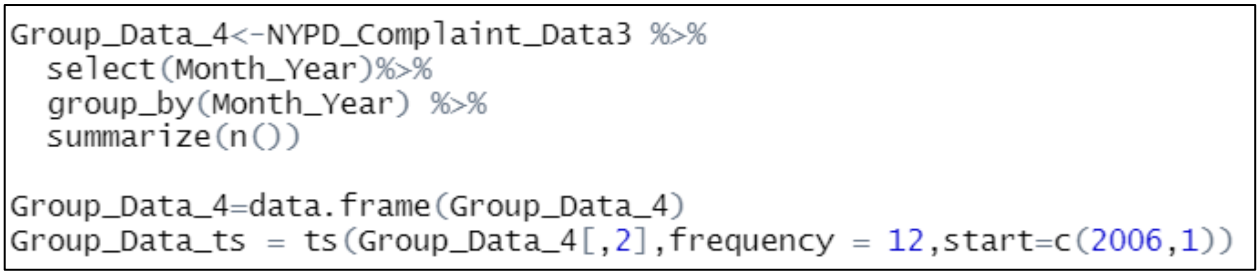
This data could be useful in determining which areas could use additional police presence in order to bring fight crime more effectively.





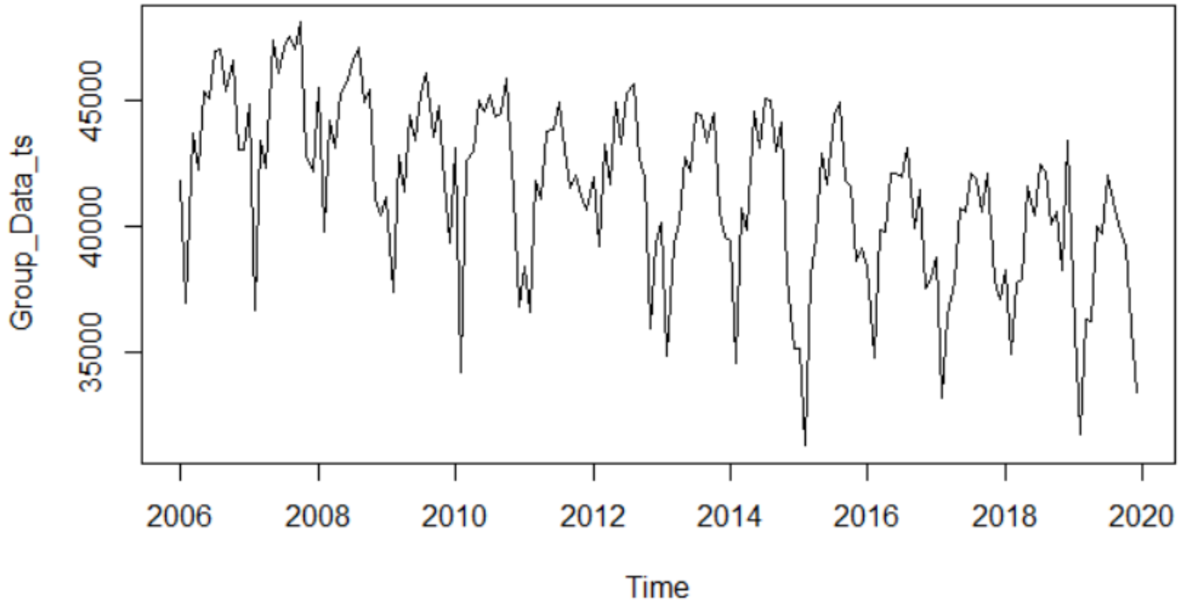
Time

In order to analyze the data, it must first be turned into timeseries data. Monthly frequency was chosen in order to interpret the effect of seasonality easier.

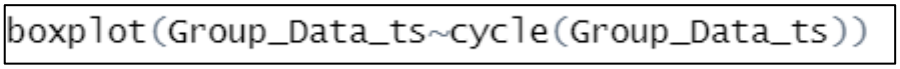


```
Group_Data_4<-NYPD_Complaint_Data3 %>%
  select(Month_Year)%>%
  group_by(Month_Year) %>%
  summarize(n())

Group_Data_4=data.frame(Group_Data_4)
Group_Data_ts = ts(Group_Data_4[,2],frequency = 12,start=c(2006,1))
```

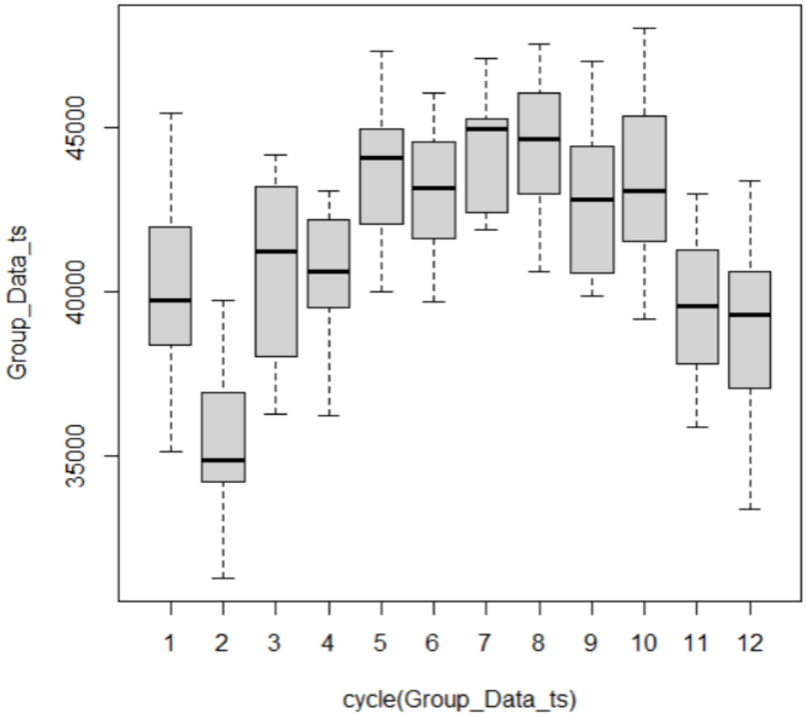
Monthly Crime in NYC



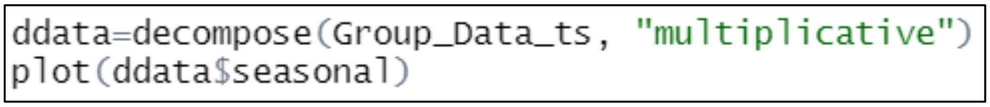
Components



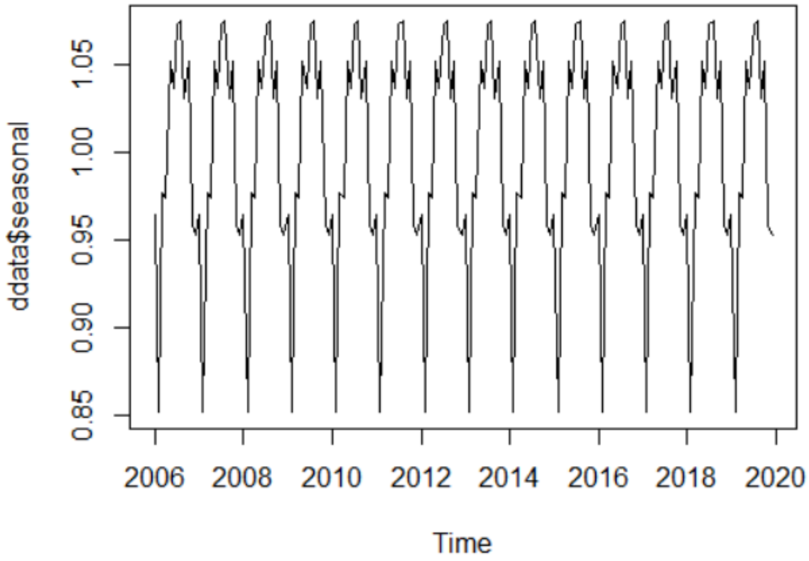
```
boxplot(Group_Data_ts~cycle(Group_Data_ts))
```



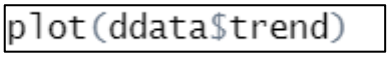
- There is clearly seasonality variation in the data. Visually one can see that crime is lower during winter months and higher during warmer months.
- It is mysterious as to why February has such low amount of crime over the years. Perhaps it is due to being less days in February 28 compared to the other months



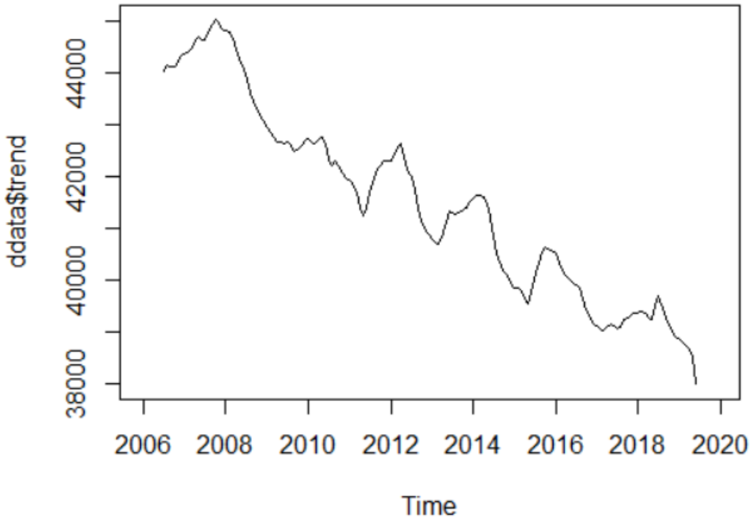
```
ddata=decompose(Group_Data_ts, "multiplicative")
plot(ddata$seasonal)
```



- Further support that the data experiences strong seasonal variation.

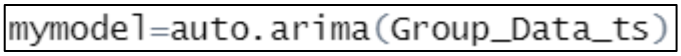


```
plot(ddata$trend)
```

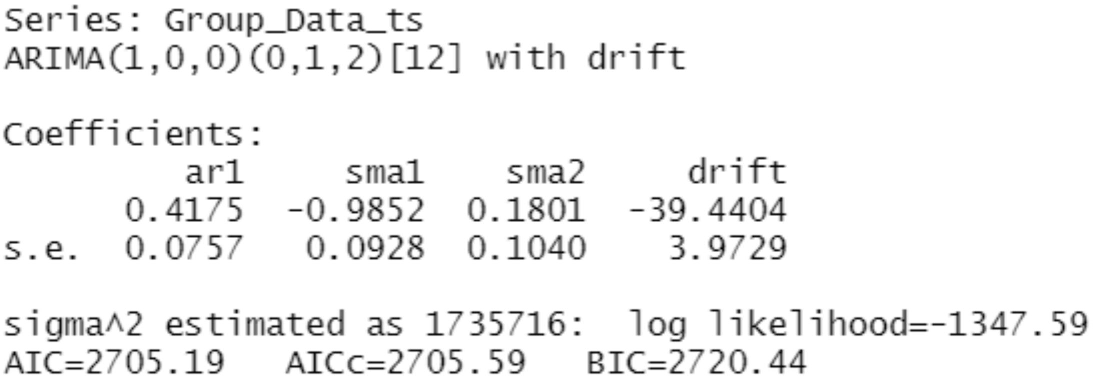


- There appears to be a trend over the last 15 years

Model

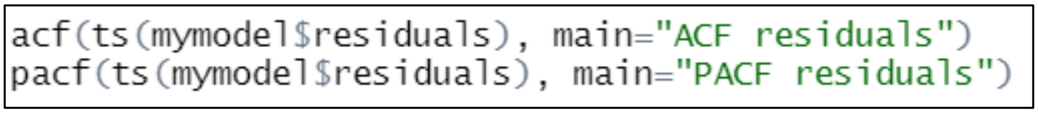


```
mymodel=auto.arima(Group_Data_ts)
```

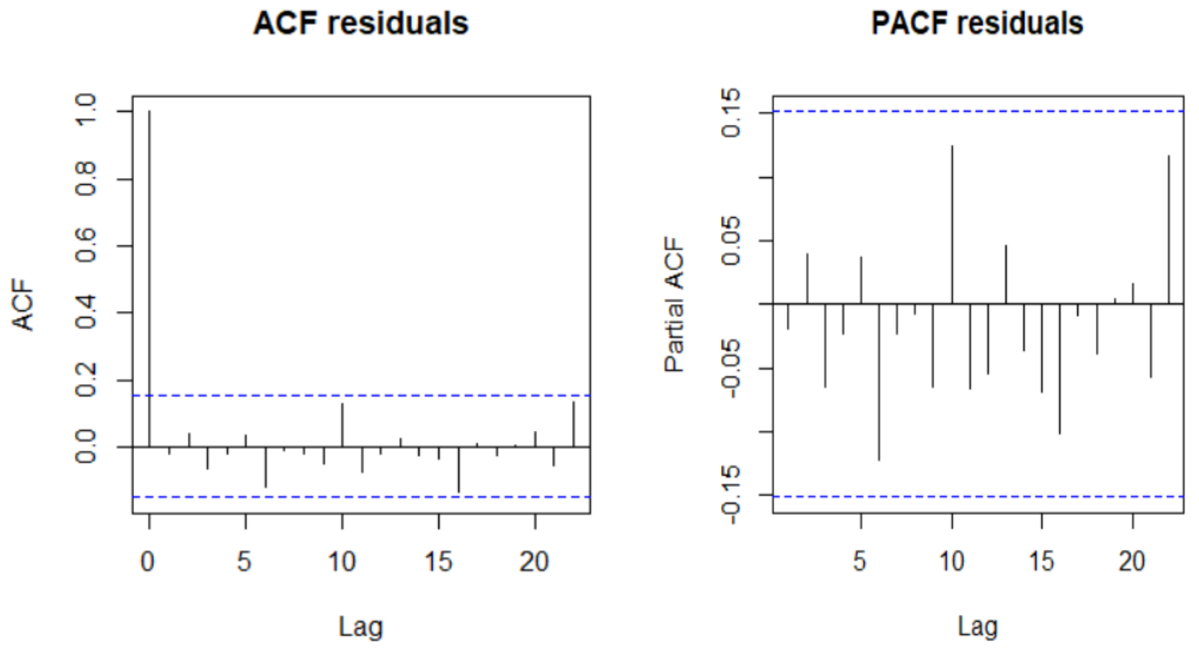


```
Series: Group_Data_ts
ARIMA(1,0,0)(0,1,2)[12] with drift

Coefficients:
         ar1     sma1    sma2     drift
      0.4175  -0.9852  0.1801  -39.4404
s.e.  0.0757   0.0928  0.1040    3.9729

sigma^2 estimated as 1735716:  log likelihood=-1347.59
AIC=2705.19   AICc=2705.59   BIC=2720.44
```

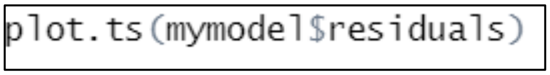
- Arima model is an Autoregressive Model as p=1, d=0, q=0
- There is 1 lagged observation



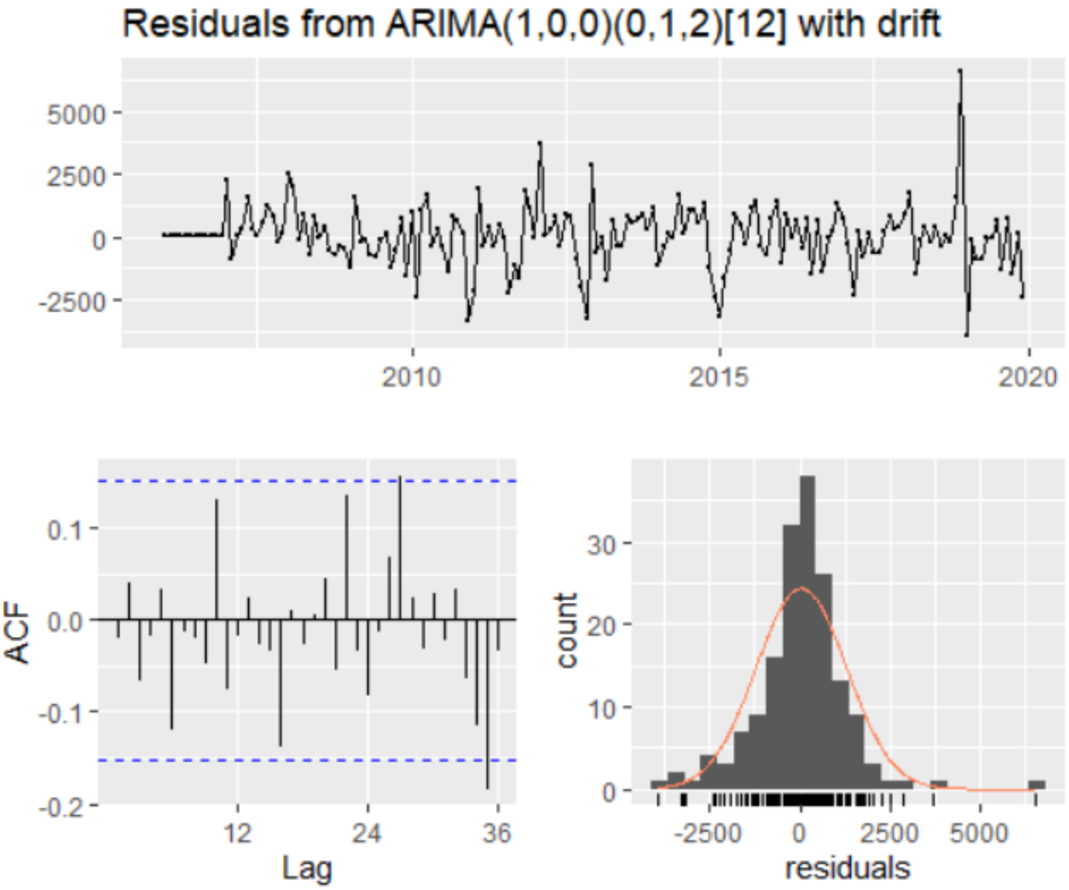
```
acf(ts(mymodel$residuals), main="ACF residuals")
pacf(ts(mymodel$residuals), main="PACF residuals")
```



Assumptions

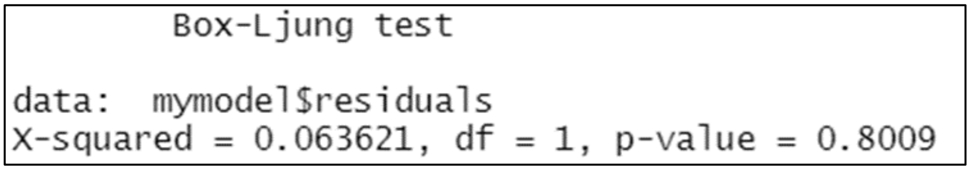


```
plot.ts(mymodel$residuals)
```



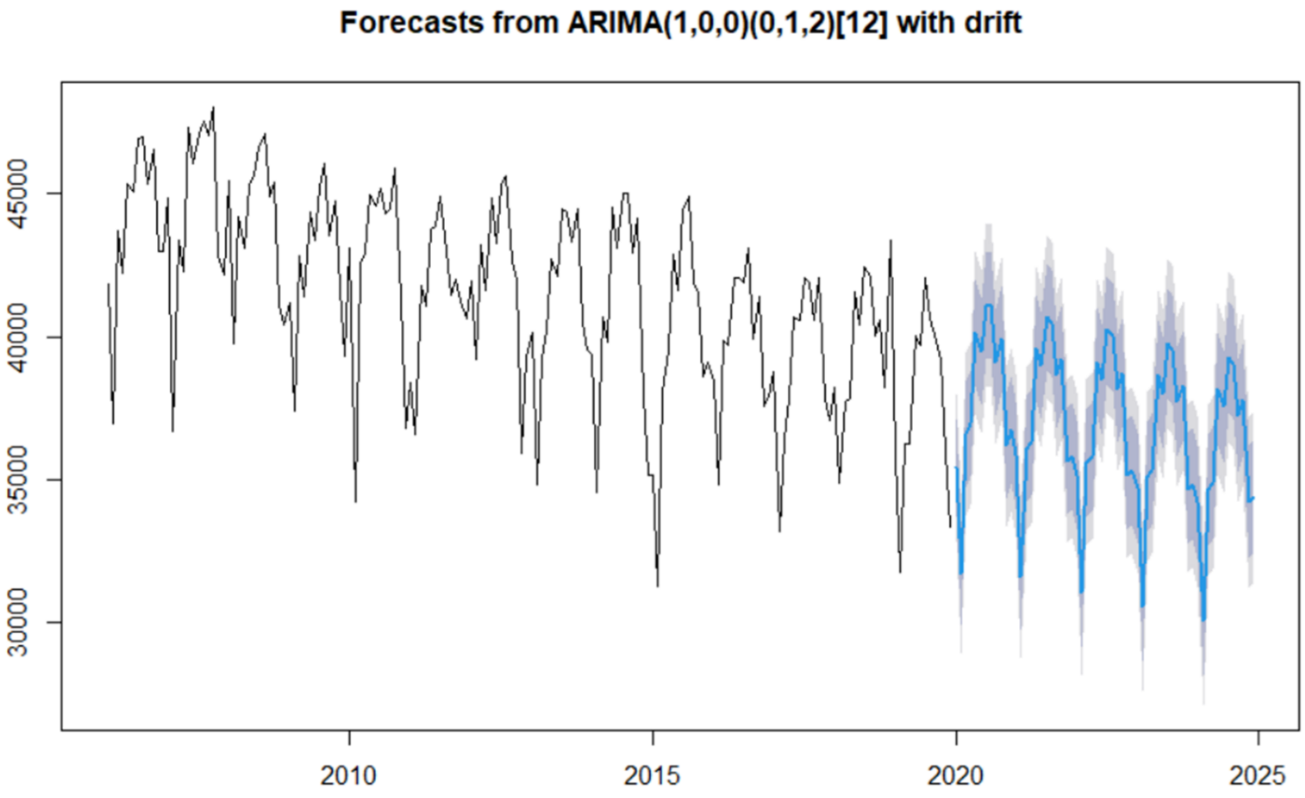
Residuals from ARIMA(1,0,0)(0,1,2)[12] with drift

- The data seems to be stationary as there does not seem to be a trend or momentum in the residuals.
- Residuals appear to resemble normal distribution.
- There's does not seem to be an issue with autocorrelation.
- According to the Box-Ljung the autocorrelations are not significant. We can conclude that our ARIMA (1,0,0) has shown to adequately fit the data



```
        Box-Ljung test

data:  mymodel$residuals
X-squared = 0.063621, df = 1, p-value = 0.8009
```

- After concluding that our data meets the assumptions needed for an ARIMA model we forecasted out five years from our last data point.
- As can be expected for data with strong seasonal variation the ARIMA forecast includes peaks and troughs.
- The confidence interval is widest near those peaks and troughs
- The Model forecast that crime will continue to decrease gradually over the years.
- The ARIMA model has combined the Trend and the Seasonal Variation in order to make the most a proper forecast.



Data:

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

http://www.statsci.org/