Introduction

The purpose of this document is for the NYC Council Data Science Interview. It is done by Caffrey Lee solely. This document only contains a brief description of the data, analysis, and the answers for the 4 questions. For detailed analysis and the code, please see the corresponding files that I submit it along with the file.

Data

The given data has 4798339 obs. of 18 variables. And, it has some issues:

- arrest_date is not in a date format;
- missing values;
- non-descriptive factors in categorical variables;
- too many categories for a single feature;

After cleaning the data, the final data that I utilized for the analysis has 4789450 of observations with 20 columns (two additional date-related columns) with a missing rate of 0%. In short, the following analysis is based on the 99.81% of the given data.

Data Preparation

- *Missing values*: the missing rows were filled up with NA for all columns, leading me to simply drop them. Also, the missing rates of the given data were not critical, 0.19%.

- *Cleaning:*

1. AREREST_BORO has 6 unique factors: B, K, M, Q, S, and " ". And, there were only 8 cases. So, I simply dropped them.
2. PERP_RACE has 8 categories with UNKNOWN AND OTHERS. I prefer not to have unknown, especially when there is a category called "Other". So, I put them together.
3. (Note: one may think that this may not be a good idea, I agree)
4. AGE_GROUP has 2000, 378, and other numbers that are not descriptive at all. Those non-descriptive numbers were recategorized into UNKNOWN
5. yr and yr_mon were introduced for the analysis

Assumption

The analysis is based on the following assumptions:

- I assume that the population does not fluctuate across time.
- For simplicity, I ignore the traits of the neighbors such as the proportion of each ethnicity, social environment, economic status, and such.
- I presume that every case is independent. This also excludes multiple crime, arrest, cases of a single person.
- Consecutive observations are equally spaced
- Apply a discrete-time index, yearly, (even though this may only hold approximately)

Analysis & Answer

Q1. Has the arrest rate been decreasing from 2015-2018? Describe the trend and defend any statistical tests used to support this conclusion.

Note: Generally speaking, arrest rate is calculated as Numb. Of Arrest (Types) / (Total) Population. In the question, the definition of the arrest rate is a bit abstract. The rate could be based on the population(uniformed) or other variables such as race, age, borough, sex, arrest types, and such. Hence, I utilized the above-mentioned variables to calculate the arrest rate of each case by year as well as the total population to compute the overall arrest rate.

The overall arrest rate throughout the time shows the decreasing trend; however, the arrest rate of race, gender, age group, borough, and the top 5 arrest types shows the different results arrest rates of race, borough, and gender do not change much, while the top 5 arrest types(i.e. Felony Assault is listed as Top 5 in 2018 when Other Offenses Related is no longer in Top 5) and age group(i.e. in age group: 18-24 and 25-44, 5% decrease and increase for each group) display distinctive changes.

The result of the multivariate linear regression also states that the arrest rates are decreasing across the time, holding the above-mentioned variable fixed. Controlling for other variables, one year increase in a year would decrease the arrest rate by -8.014e-07 on average.

Q2. What are the top 5 most frequent arrests as described in the column 'pd_desc' in 2018? Compare & describe the overall trends of these arrests across time.

Note: Here, the question asked about the frequency, which I take it as the number of occurrences, instead of rate. That is, the following analysis is based on the absolute number rather than the rate, calculated by frequency/population

The most frequent pd_desc types in 2018 are Assult 3, Larceny Petit From Open Areas, Traffic, Assault 2&1, and Controlled Substance.

Regarding the overall trend of the top 5 arrests, one could say that the trend is moving from increasing to decreasing starting from 2014 and that there is a high seasonality.

Yet, looking at the trend for each type, one could see that the arrest frequency varies across the year: the arrest frequency of Assault 3 has been varied between 29% and 43%, even though it accounts for the biggest portion of the arrests across the year, from 2009 to 2018. The frequency of another type of Assault, 2&1, does not show much fluctuation. Still, one could see that since 2012, it is not listed on the top 3 anymore. Lastly, December has the lowest arrest rates throughout the years, when May has the highest.

Still, one could say that the frequency of those top 5 types is increased due to the fact that the mean of the overall number of arrests of those types is increased compared to that of the beginning year. This could be supported by the Linear Regression. (P-val of yr is 1.629e-01)

Q3. If we think of arrests as a sample of total crime, is there more crime in precinct 19 (Upper East Side) than precinct 73 (Brownsville)? Describe the trend, variability and justify any statistical tests used to support this conclusion.

Note: The question mentioned that the data, sample, could represent the total crime. That is, it is a representative sample that allows me to draw a conclusion of the population based on the analysis of the sample with confidence.

Say that, number of arrests can represent the crime rates. With that in mind, the analysis is conducted. Also, I calculated the monthly mean of arrests of each precinct across the year, then take the difference between them. Here, I took the mean difference to alleviate the difference between the two districts as well as the difference across the time.

Based on the monthly mean of arrests, one can say that Precinct 73 has a greater average amount of crime than precinct 19 across the time on average. Yet, to answer this question, one should be careful since the analysis disregard the demographic difference between the two communities. To confirm, I ran a linear regression. And, the P-val of the precinct 73 says that location does not have a statistical impact on the mean of crimes, controlling the year fixed. In contrast, the year has a statistical impact on the monthly mean number of arrests.

Regarding the trend, one could say that it is moving from increasing to decreasing trend from 2011 with seasonality. This happens mostly because the crime rates in precinct 73 are decreasing. For instance, after reaching the peak rate, 43.87%, in 2011-07, it gradually decreases. And, the rate remains around 13% since 2018-08.

Further research on the characteristics of each neighborhood such as ethnicity, environment, economic status, and such could help one to have a better understanding of the crime rates between the two communities.

Q4. What model would you build to predict crime to better allocate NYPD resources? What challenges do you foresee? What variables would be included? How would you evaluate the model? Discuss in no more than 150 words.

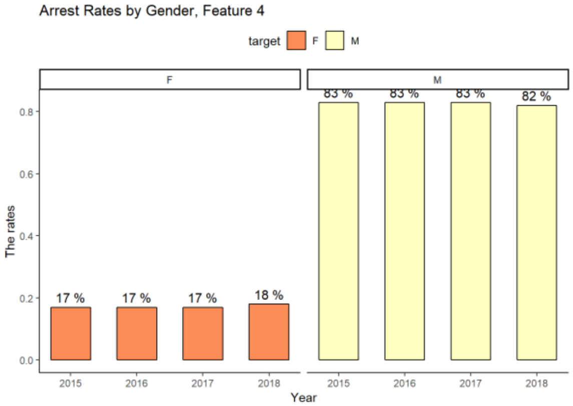
Forecasting, in general, is difficult, especially when the outcome is expected to be influenced by unexpected variables. Hence, to select a model, I generated various models. And, based on the AIC, RMSE, ME, and MPE of the multiple models, I would utilize the ARIMA(0,1,1)(0,1,1)[12] model to forecast crime.

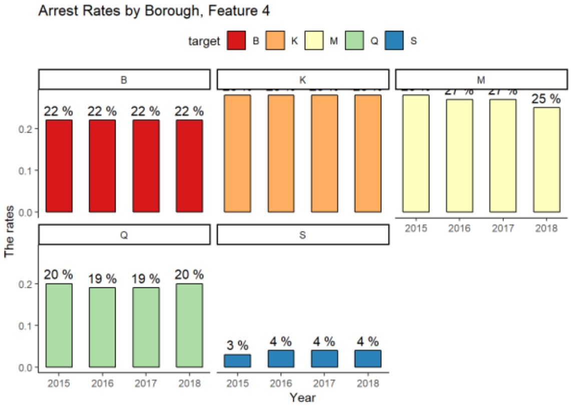
The analysis is based on the arrest rate based on uniformed total population across the year. As the above analysis shows, this tends to ignore the difference among the districts, race, age group, borough, crime types and such. Incorporating the variables that could explain the characteristics of each district - race, age group, borough, and crime types - would be helpful. Still, this is vulnerable to bias data.

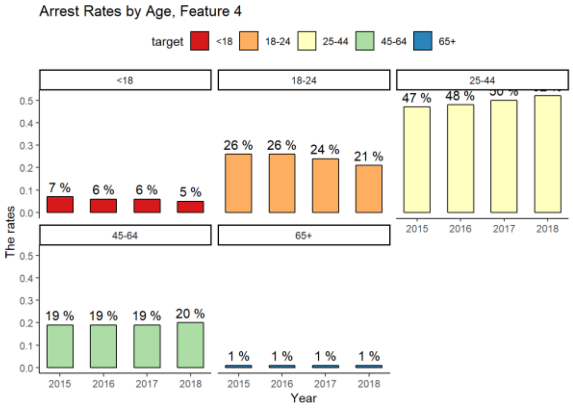
As an evaluation, I would pay attention to RMSE, ME, and MPE more because large errors are undesirable in terms of forecasting, and ME and MPE allow for the bias check.

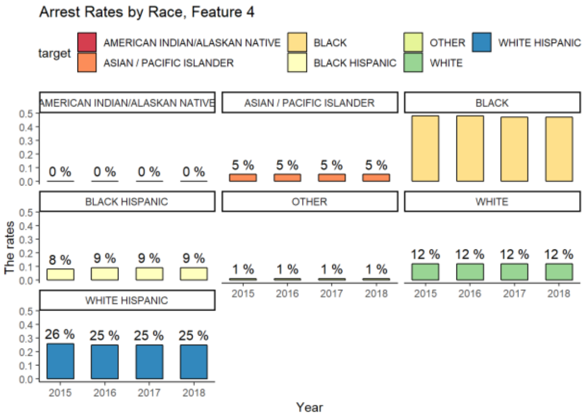
Appendix

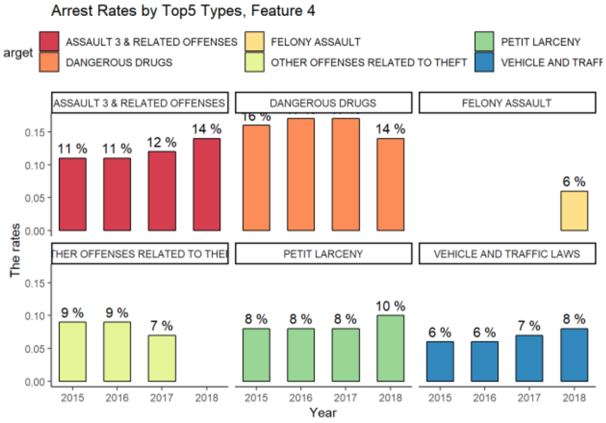
Q1 Appendix

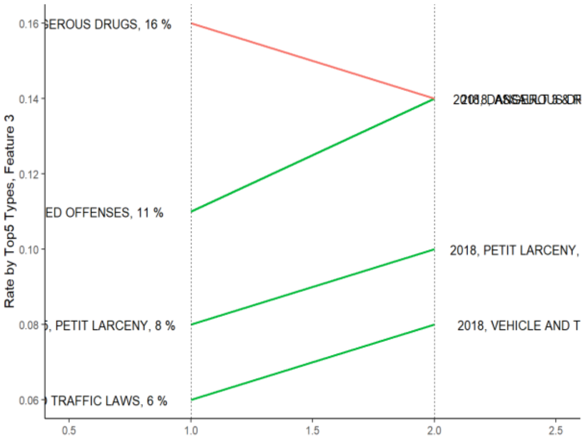

Arrest Rates by Gender, Feature 4


Arrest Rates by Borough, Feature 4

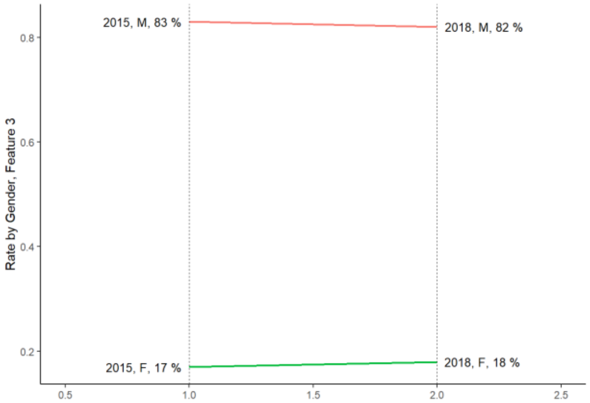

Arrest Rates by Age, Feature 4

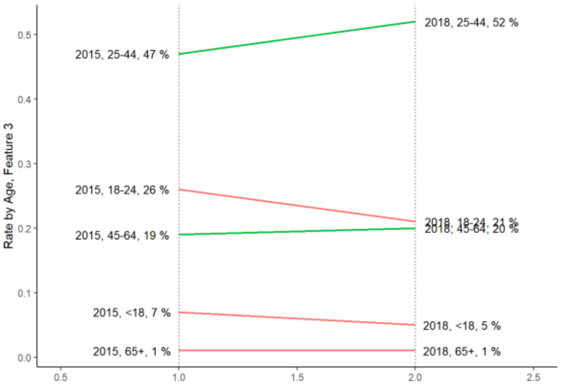

Arrest Rates by Race, Feature 4

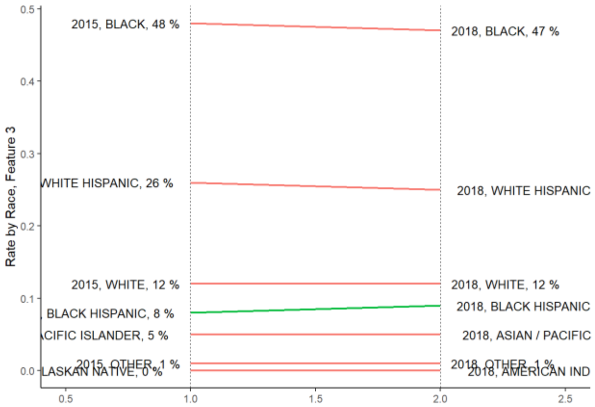

Arrest Rates by Top5 Types, Feature 4

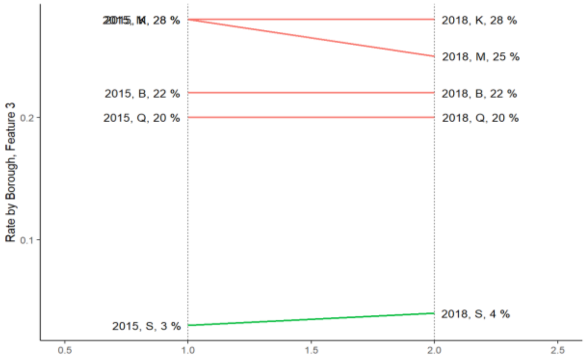


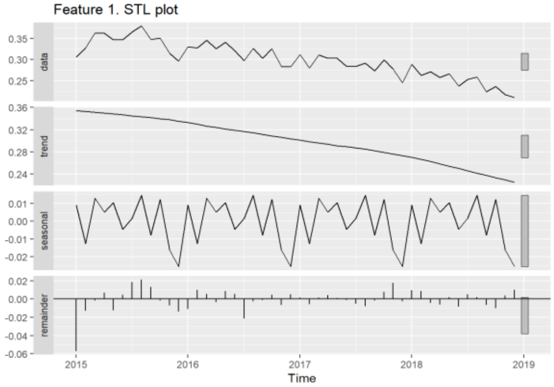
Caffrey Lee – NYC Council Data Science Interview

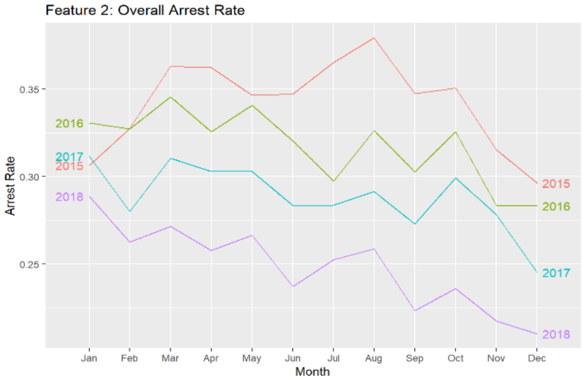


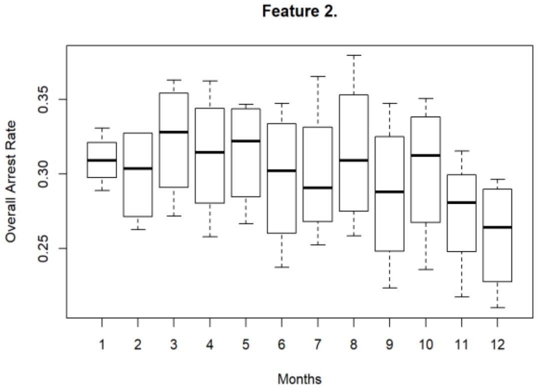




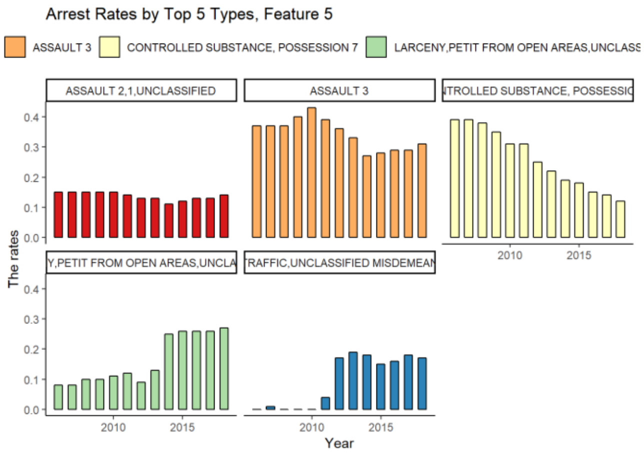


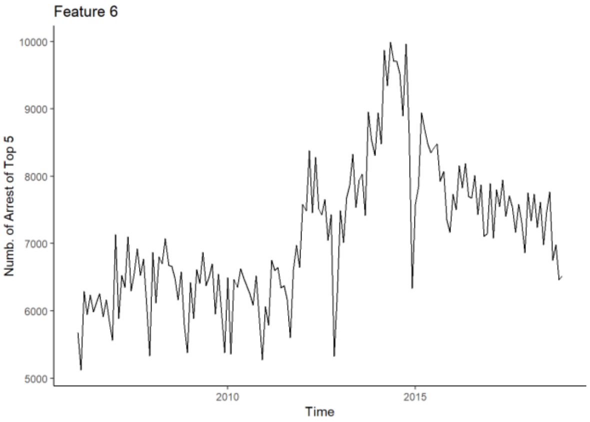


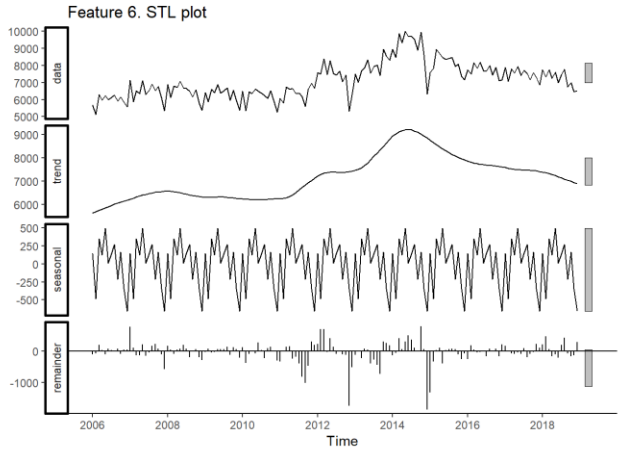


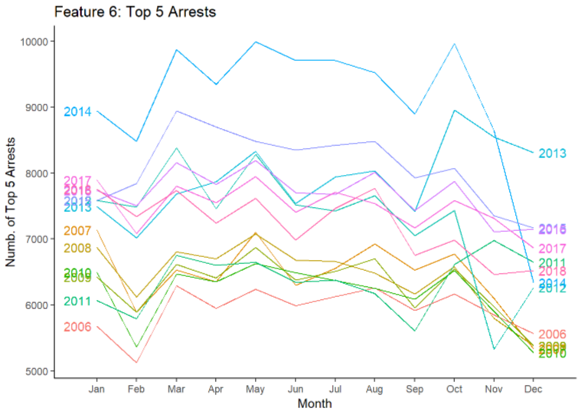


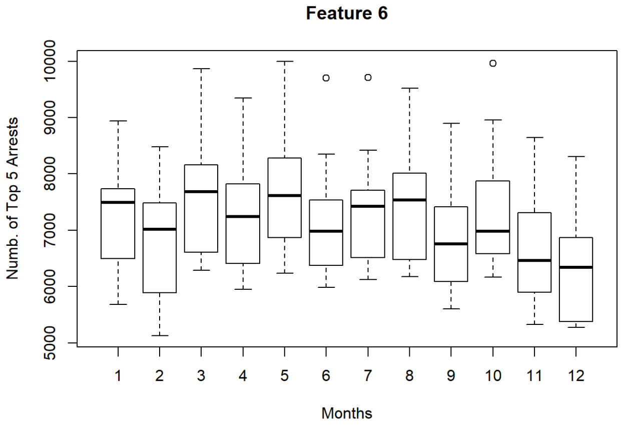
Q2 Appendix

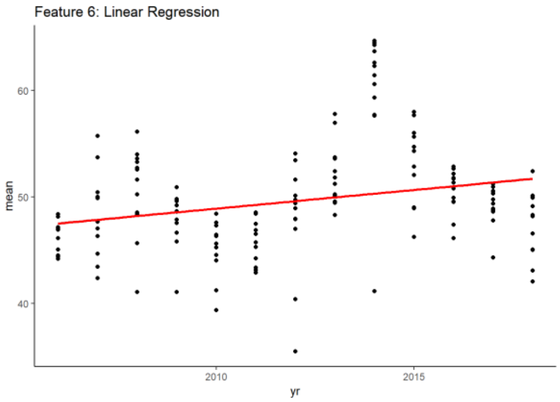

Arrest Rates by Top 5 Types, Feature 5


Feature 6


Feature 6. STL plot

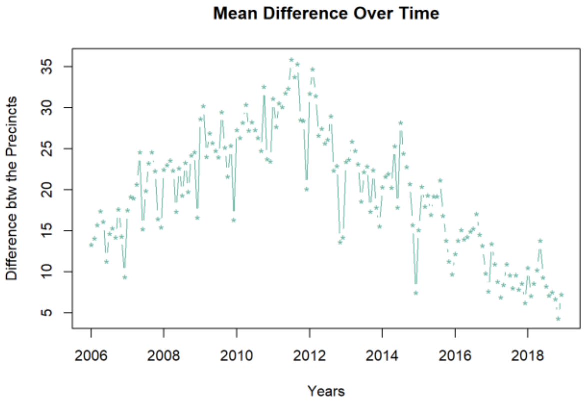

Feature 6: Top 5 Arrests

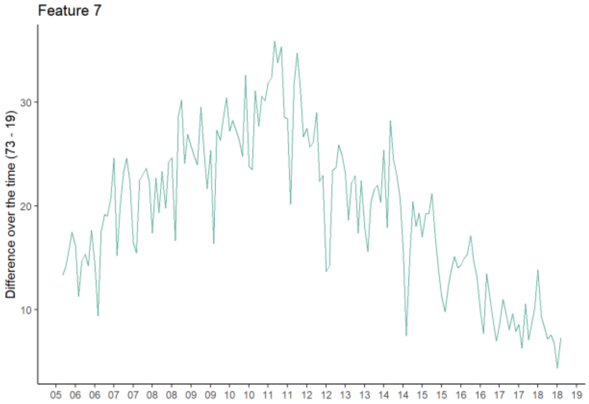

Feature 6

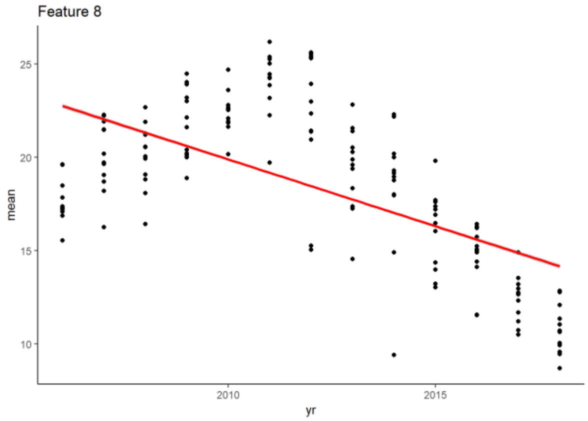

Feature 6: Linear Regression

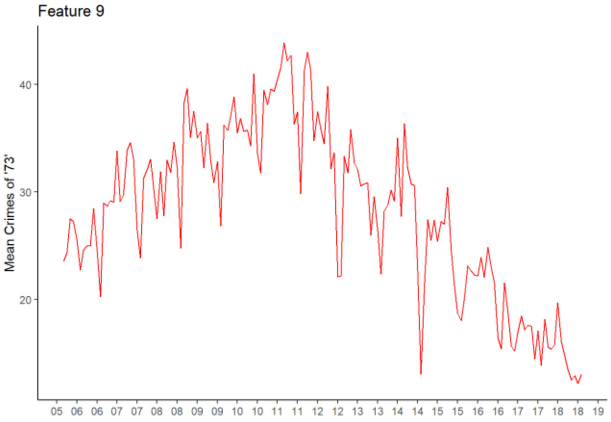
Caffrey Lee – NYC Council Data Science Interview

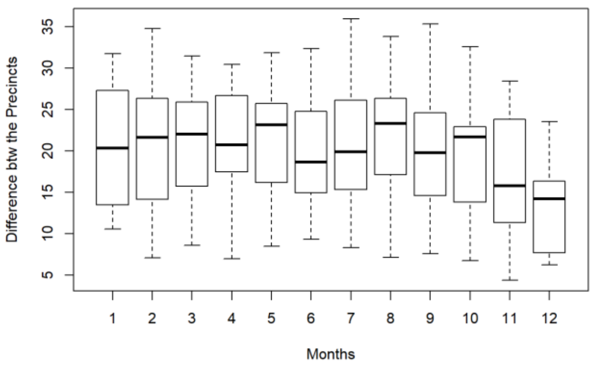
Q3 Appendix

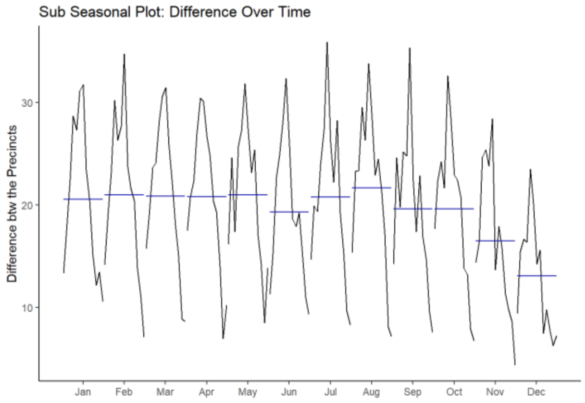


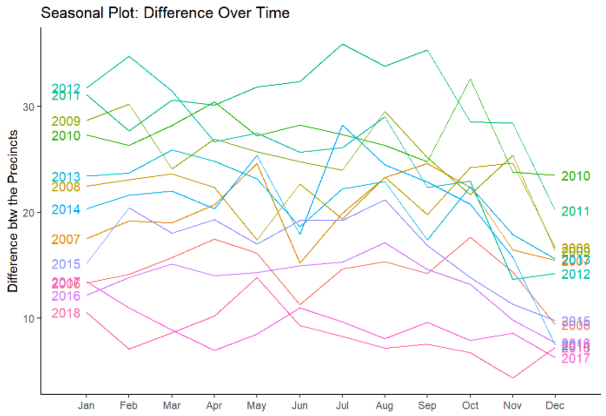


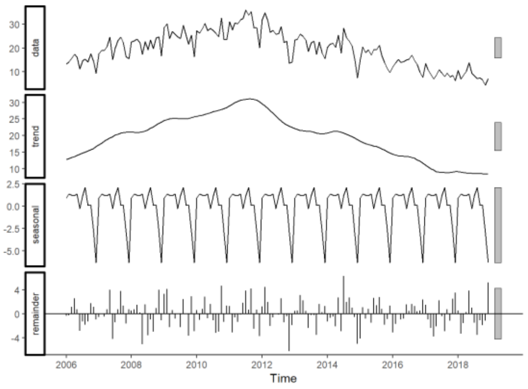




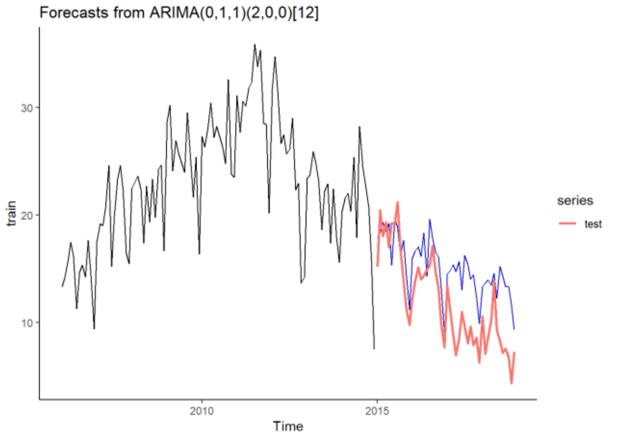


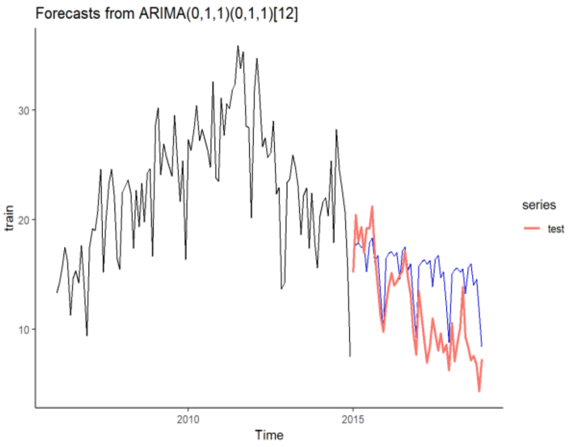

Sub Seasonal Plot: Difference Over Time

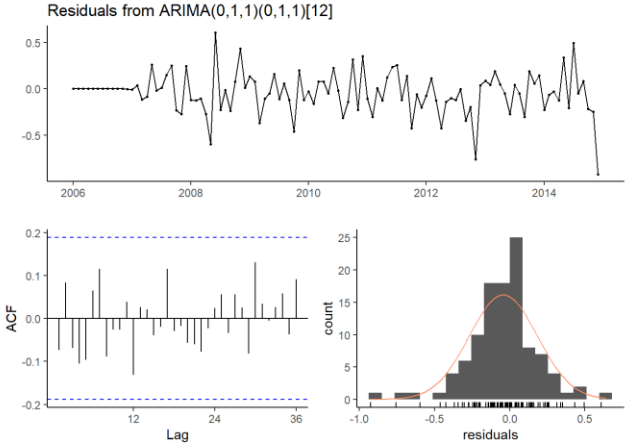

Seasonal Plot: Difference Over Time

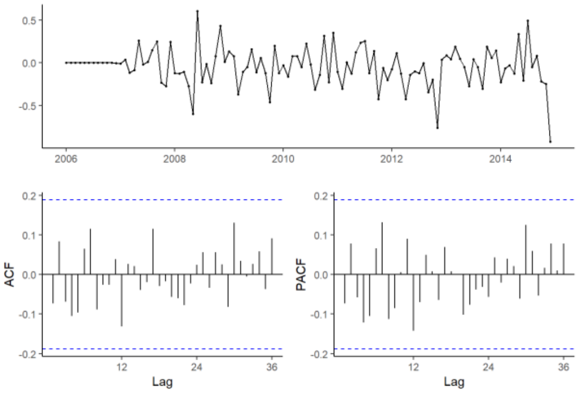


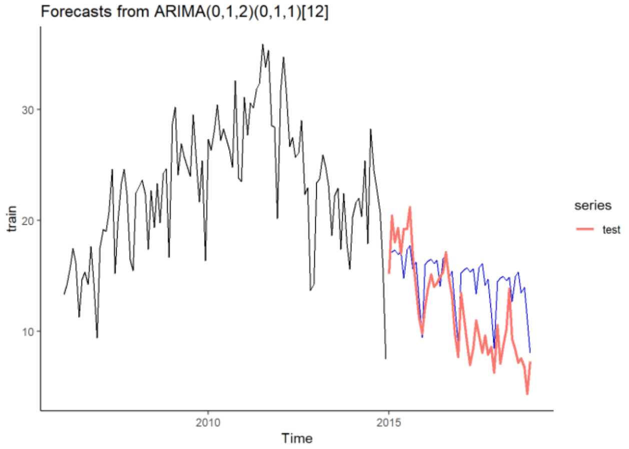
Q4 Appendix

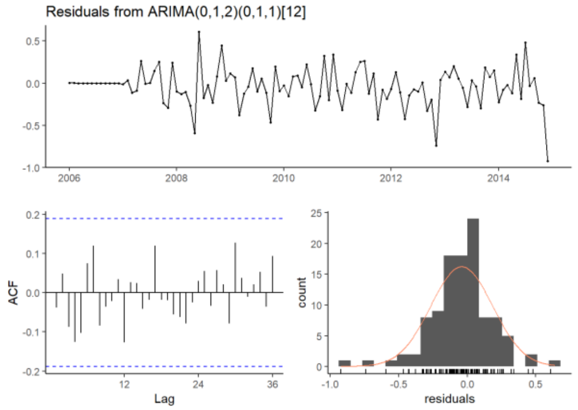

Forecasts from ARIMA(0,1,1)(2,0,0)[12]

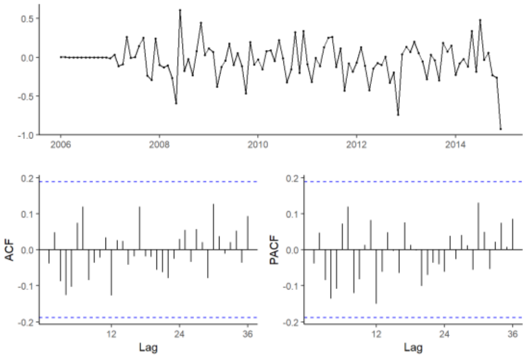

Forecasts from ARIMA(0,1,1)(0,1,1)[12]

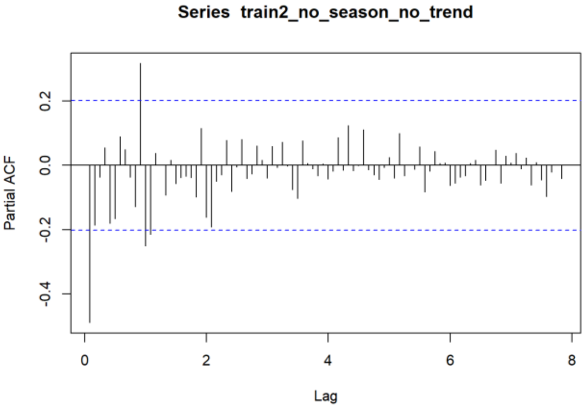

Residuals from ARIMA(0,1,1)(0,1,1)[12]

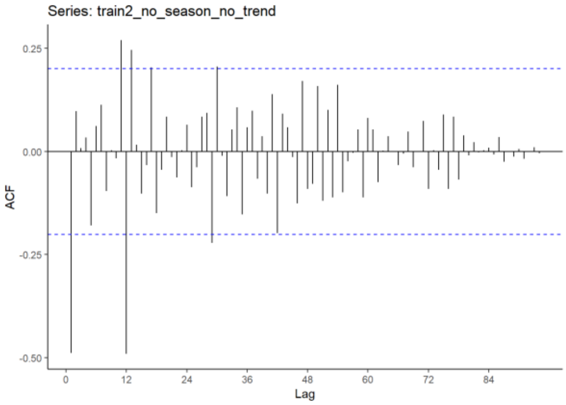


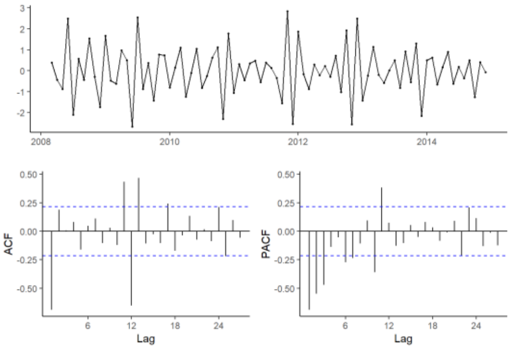

Forecasts from ARIMA(0,1,2)(0,1,1)[12]

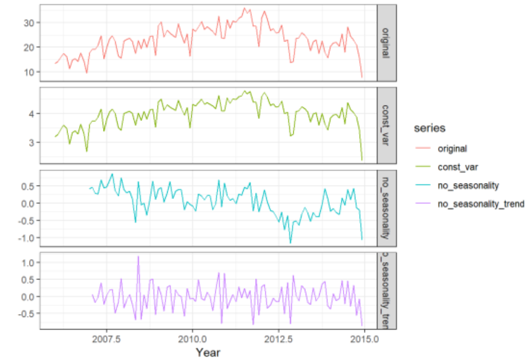

Residuals from ARIMA(0,1,2)(0,1,1)[12]

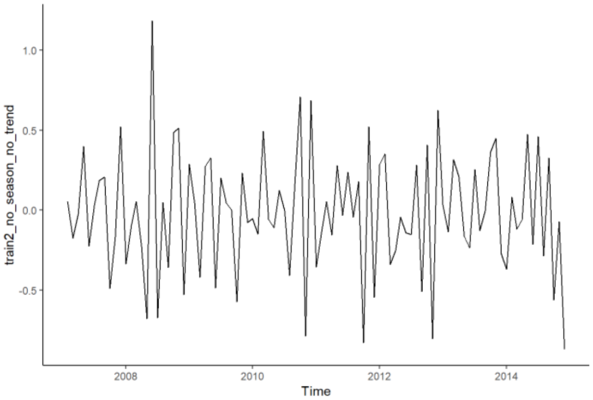

0.5
0.0
-0.5
-1.0
2006
2008
2010
2012
2014
ACF
PACF
0.2
0.1
0.0
-0.1
-0.2
12
24
36
Lag
Lag

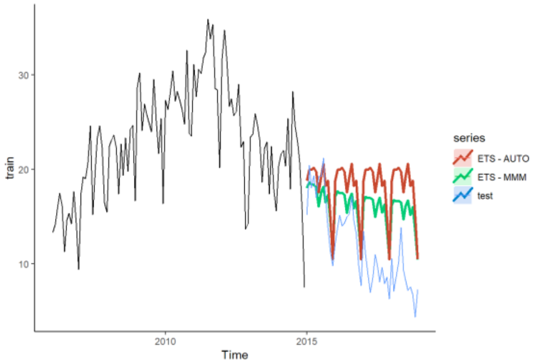

Series train2_no_season_no_trend
Partial ACF
0.2
0.0
-0.2
-0.4
0
2
4
6
8
Lag

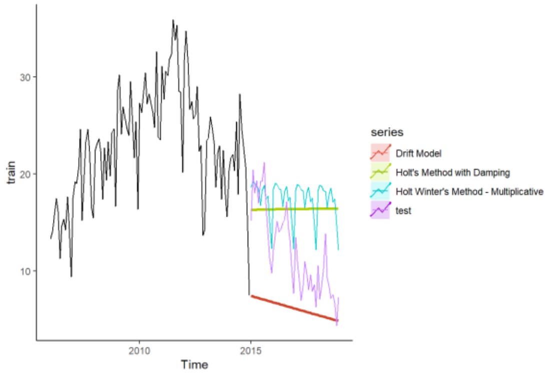

Series: train2_no_season_no_trend
ACF
0.25
0.00
-0.25
-0.50
0
12
24
36
48
60
72
84
Lag

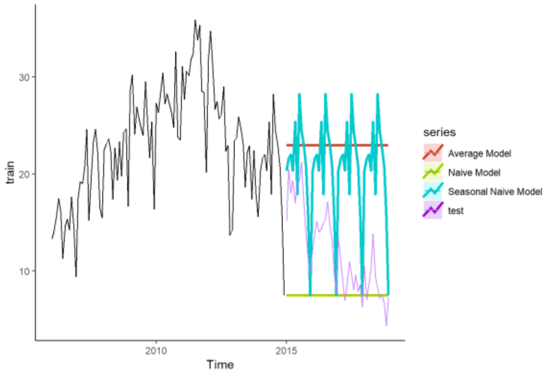

3
2
1
0
-1
-2
2008
2010
2012
2014
ACF
PACF
0.50
0.25
0.00
-0.25
-0.50
6
12
18
24
Lag
Lag


original
const_var
no_seasonality
o_seasonality_tren
series
original
const_var
no_seasonality
no_seasonality_trend
2007.5
2010.0
2012.5
2015.0
Year


train2_no_season_no_trend
1.0
0.5
0.0
-0.5
2008
2010
2012
2014
Time


train
30
20
10
2010
2015
Time
series
ETS - AUTO
ETS - MMM
test


train
30
20
10
2010
2015
Time
series
Drift Model
Holt's Method with Damping
Holt Winter's Method - Multiplicative
test



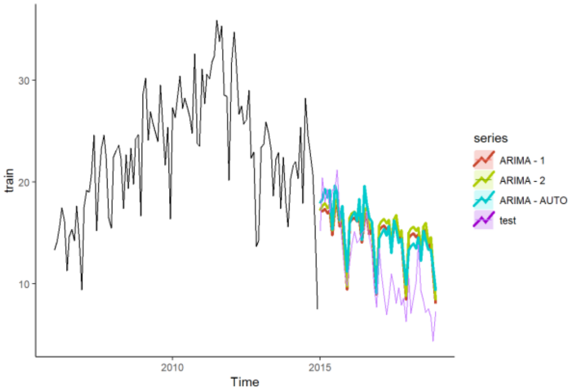
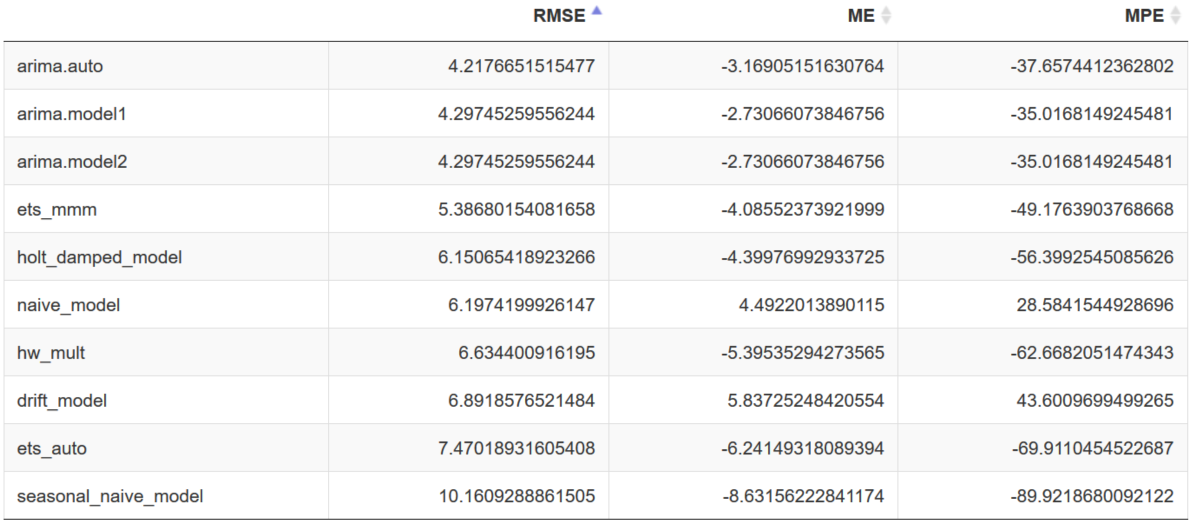


Table 11.



	RMSE	ME	MPE
arima.auto	4.2176651515477	-3.16905151630764	-37.6574412362802
arima.model1	4.29745259556244	-2.73066073846756	-35.0168149245481
arima.model2	4.29745259556244	-2.73066073846756	-35.0168149245481
ets_mmm	5.38680154081658	-4.08552373921999	-49.1763903768668
holt_damped_model	6.15065418923266	-4.39976992933725	-56.3992545085626
naive_model	6.1974199926147	4.4922013890115	28.5841544928696
hw_mult	6.634400916195	-5.39535294273565	-62.6682051474343
drift_model	6.8918576521484	5.83725248420554	43.6009699499265
ets_auto	7.47018931605408	-6.24149318089394	-69.9110454522687
seasonal_naive_model	10.1609288861505	-8.63156222841174	-89.9218680092122