Forecasting NYC Restaurant Animal-related Violations

Chavanie Joseph, Tiffany Ramkaran, Samantha Wang

Fall 2020

Introduction

Every restaurant in New York City (NYC) is randomly inspected at least a year.[1] With each inspection, the inspector notes violations against each restaurant. In our project, we analyzed weekly Animal-related Violations as our dependent variable. We modeled it against our independent variables, Cuisine Types and 311 complaints, along with seasonality to understand the correlation between them. Our data is from 1/1/2017-12/28/2019. We omitted 2020 data due to the COVID-19 pandemic which led to restaurant closure between the months of March and June. However, we forecasted the number of Animal-related Violations in 2020 as if COVID-19 did not exist. To forecast, we used the 4-period moving average of the dependent variable to mitigate for any anomalies in the weekly data set.

We gathered daily data from the NYC Open Data source – `data.cityofnewyork.us` – and transformed it into weekly data. Our first dataset, DOHMH New York City Restaurant Inspection Results, includes our dependent variable, Animal-related Violations, and independent variables, Cuisine Types and location (Community Districts). We filtered the violation codes that has an animal-related description. The table below displays all the animal-relation violation code and their descriptions.

Code	Violation Description
04L	Evidence of mice or live mice present in facility's food and/or non-food areas.
08A	Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.
04N	Filth flies or food/refuse/sewage-associated (FRSA) flies present in facility's food and/or non-food areas. Filth flies include house flies, little house flies, blow flies, bottle flies and flesh flies. Food/refuse/sewage-associated flies include fruit flies, drain flies and Phorid flies.
04M	Live roaches present in facility's food and/or non-food areas.
04K	Evidence of rats or live rats present in facility's food and/or non-food areas.
04O	Live animals other than fish in tank or service animal present in facility's food and/or non-food areas.

The Cuisine Type column originally contained 80+ types of cuisines, but we noticed that many of the cuisines can be grouped. We tided the data to present Cuisine Type variables. The Community District represents a location variable. Community Districts that begin with a '1' signifies that the district is in Manhattan, '2' is the Bronx, '3' is Brooklyn, '4' is Queens, and '5' is Staten Island. Our second dataset – 311 Service Requests from 2010 to Present – was consolidated to display only complaints of interest before importing due to its large size. The independent variables included in this dataset are complaints about Rodents, Litter Baskets/Garbage, Sanitation, and Sewage. After gathering our independent variables, we tested seasonality in the dependent variable. The number of Animal-related Violations turned out to be seasonal; therefore, we modeled seasonality as well.

Using our tidied data, we explored and analyzed the effect the independent variables had on the count of Animal-related Violations using visualizations and modeling.

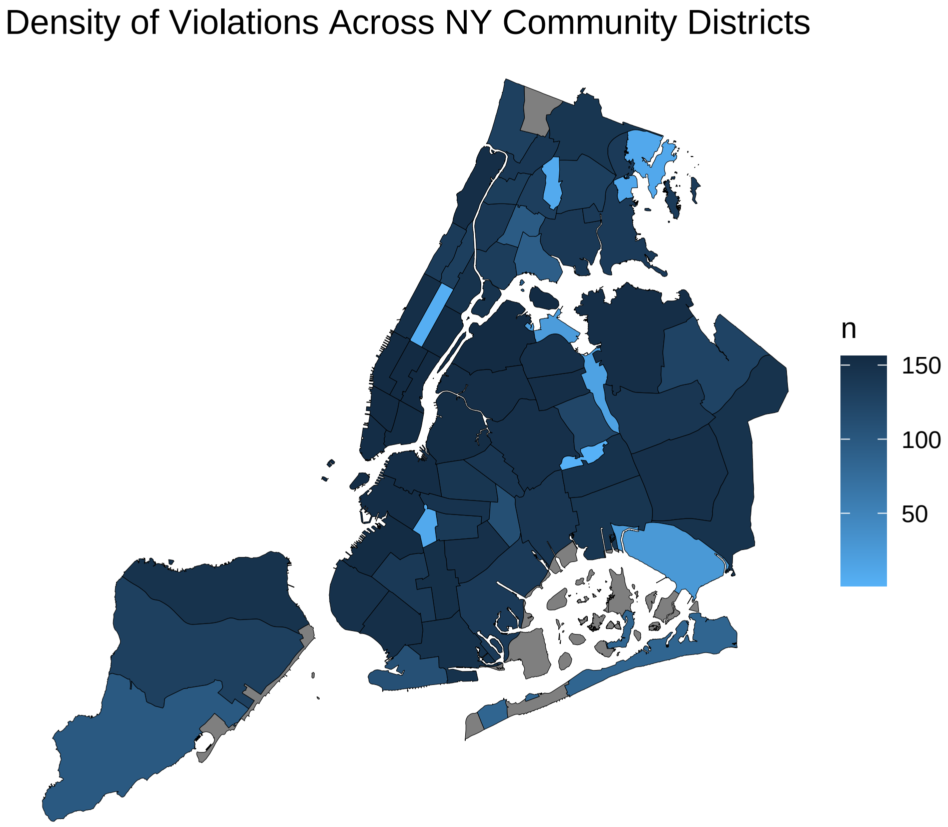
[1] https://www1.nyc.gov/site/doh/business/food-operators/the-inspection-process.page#

Exploration of Data

We explored the relationship between number of Animal-related Violations in restaurants and our independent variables – Cuisine Type, Location (Community District), Complaints about Rodents, Litter Baskets/Garbage, Sanitation, and Sewage. We also observed the timeline for any significant increases during specific times of the year (i.e. Summer months).

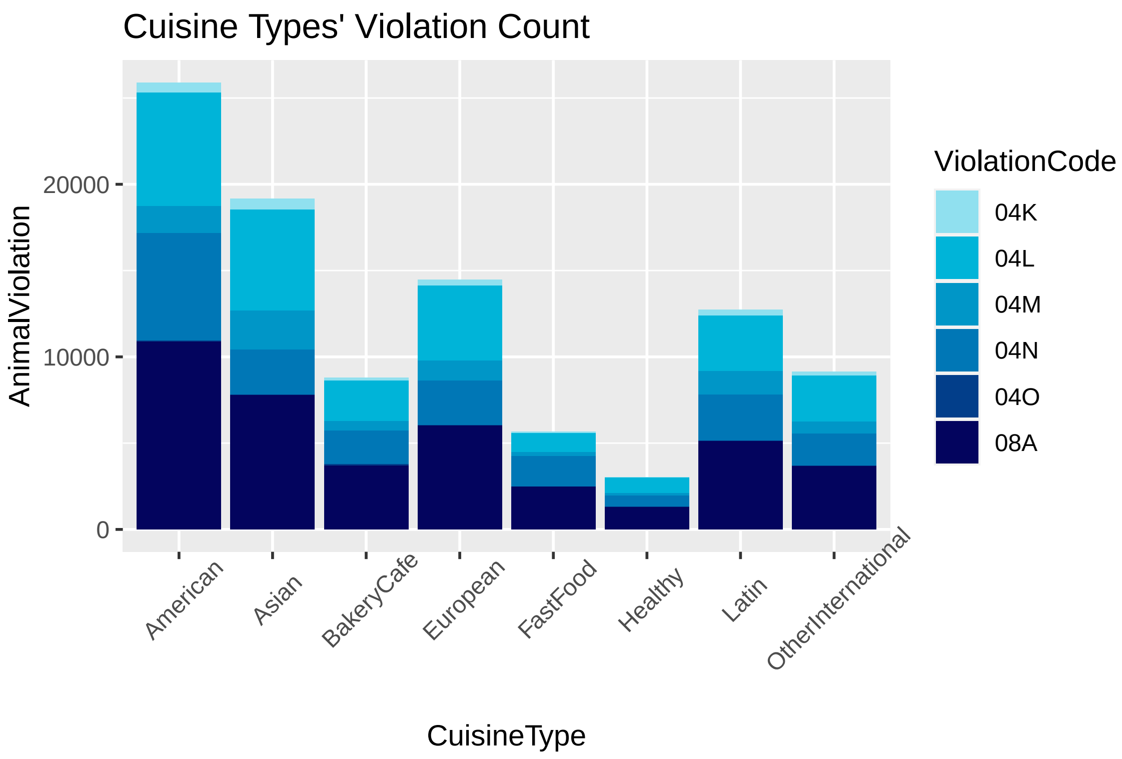
Relationship between Community Districts and Violations

We first looked at the relationship between Location and Violations. Community Districts represent the different locations throughout NYC. In the density map below, we can see which areas in NYC are more susceptible to Animal-related Violations and concluded majority of the Community Districts have about the same number of violations (130-150 violations total). We presumed the areas with few violations may be wealthier or have less restaurants. We determined Location was not a strong variable to further analyze due to the findings in the density map.



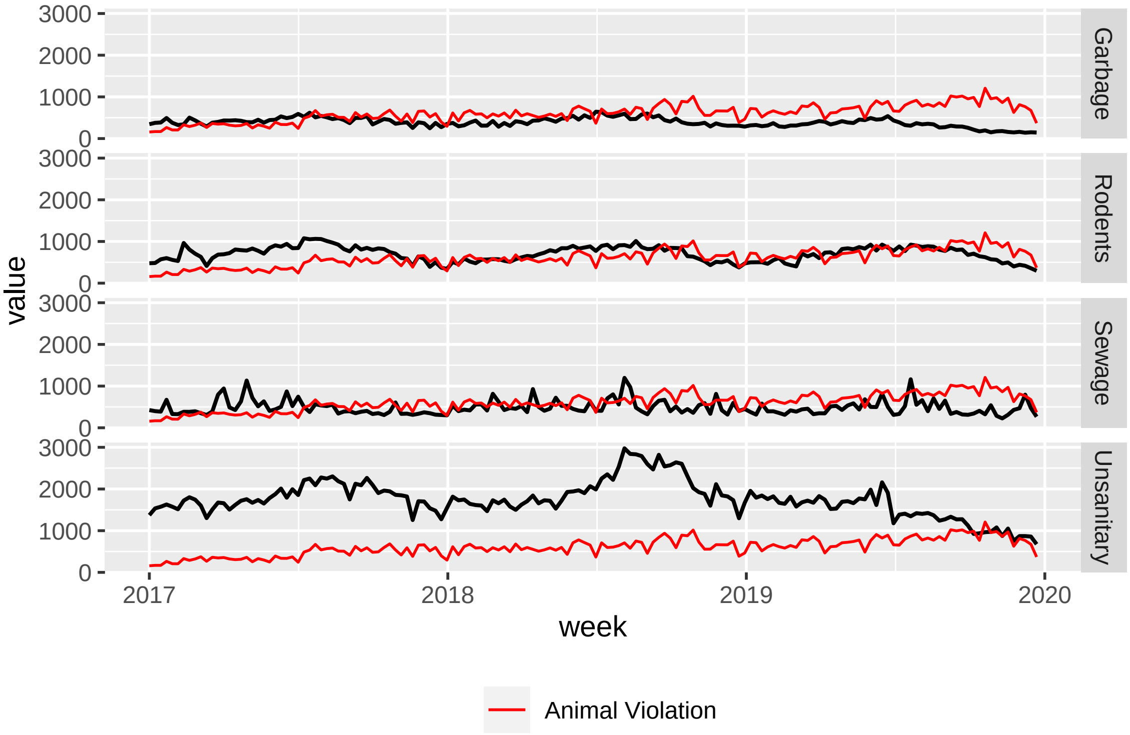
Relationship between Cuisine Type and Violations

Cuisine Types are grouped into eight categories – American, Asian, Bakery/Cafe, European, Fast Food, Healthy, Latin, and Other International. They were grouped based on their Cuisine Description likeliness. The bar graph below displays the number of Violations in each Cuisine Type. According to the graph below, the most present violation codes are 08A, 04N, and 04L. American Cuisine Type has the most violations against it.



Relationship between Complaints and Violations

Complaints were grouped into four categories – Rodents, Sewage, Garbage, and Unsanitary. In the figures below, we plotted each Complaint against the number of Animal-related Violations over three years. We noticed the trends in the depedent variable were similar to the trends of the complaint type variables. There are notable spikes in each of the figures. We noted in the Rodent Complaint figure, the spikes appear to be more prominent in the summer months. We explored and analyzed this further using a seasonality model in our analysis.



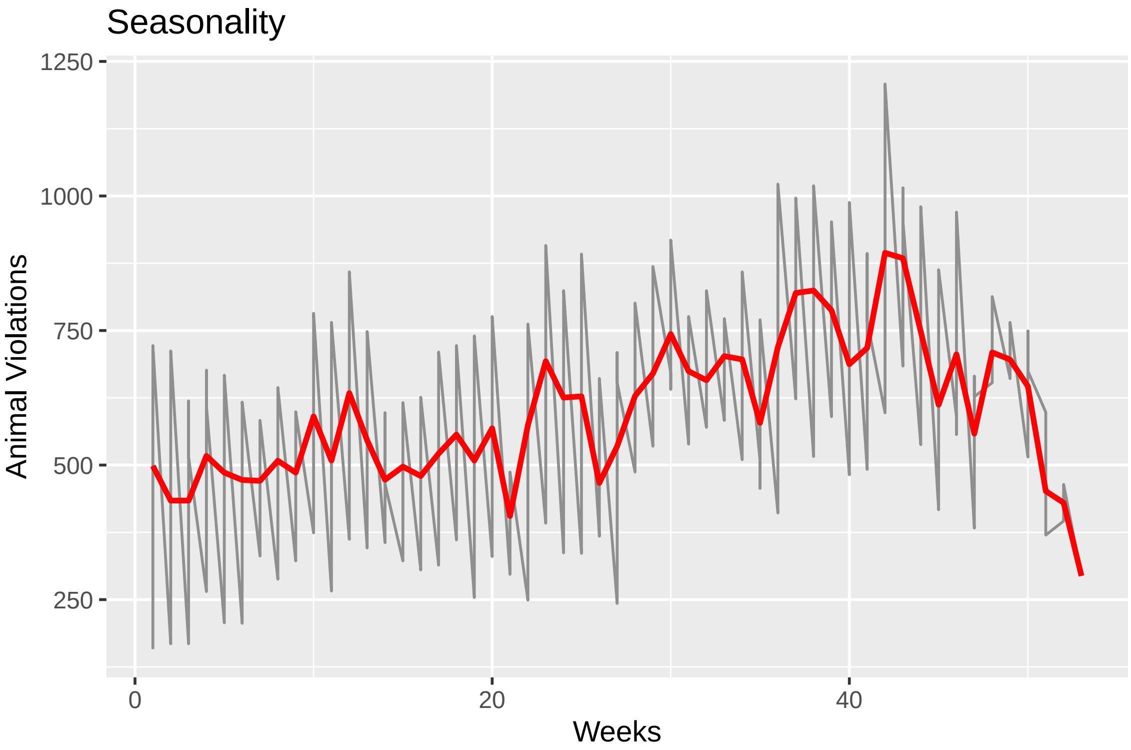
Time Series, Seasonality, and Moving Average

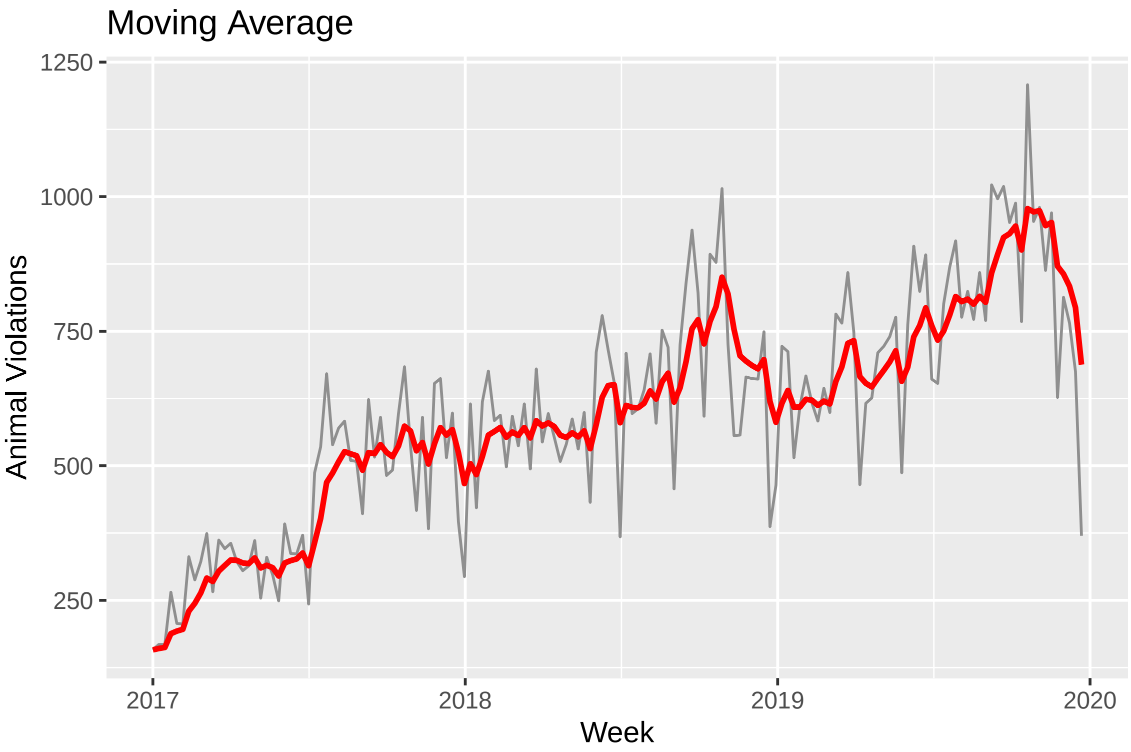
When analyzing our dependent and independent variables, we noticed that there were consistent spikes in the data. There were spikes in Animal-related Violations and Rodent Complaints in the summer, leading us to believe that the data is highly seasonal.

Therefore, we created a time series and used the Augmented Dickey Fuller (ADF) Test to test if the time series is stationary. We discovered that it is non- stationary, due to its p-value being above 0.05 and confirmed that seasonality exists.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_seasonality
## Dickey-Fuller = -2.4844, Lag order = 5, p-value = 0.3744
## alternative hypothesis: stationary
```

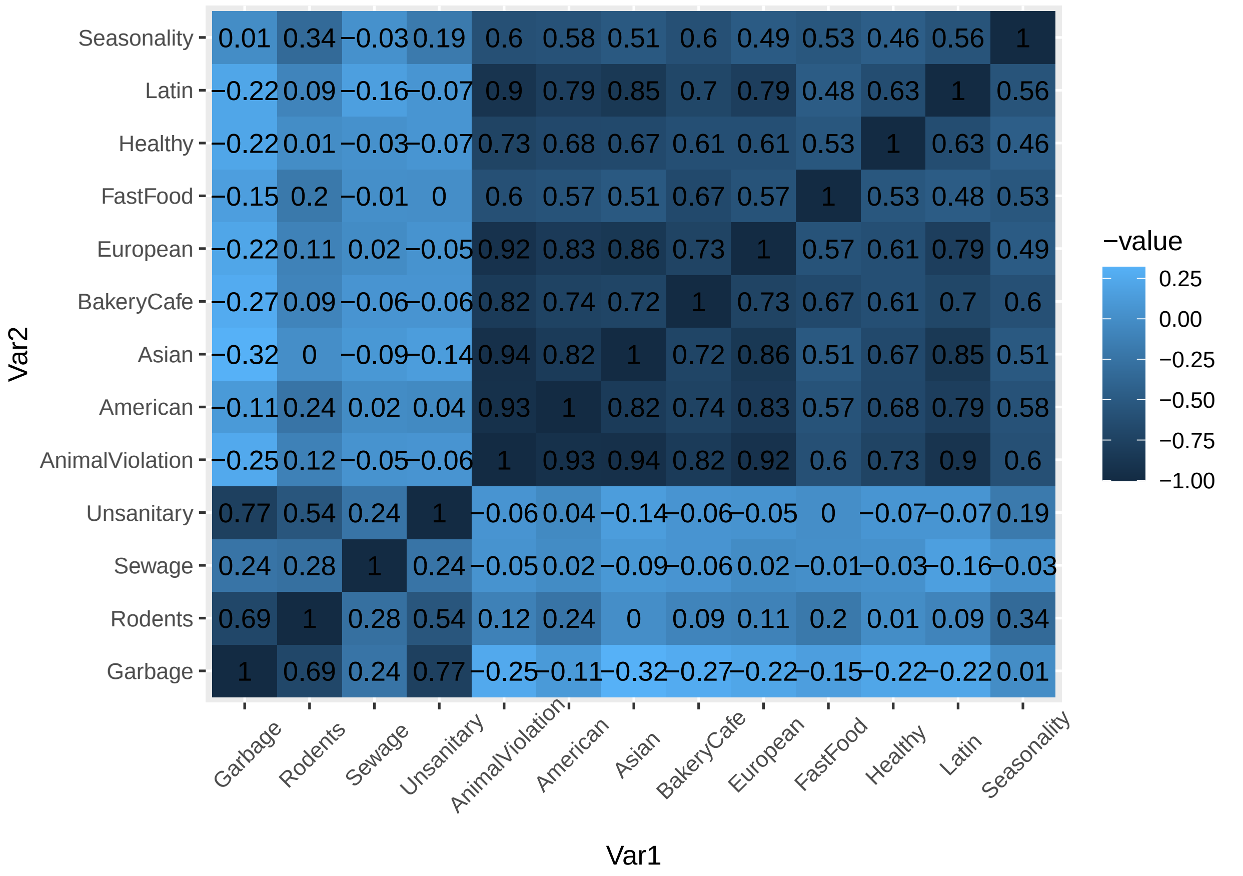
Furthermore, below displays a graph of Seasonality values against the sum of the Animal-related Violations over three years. To stabilize the Animal-related Violations, we also graphed the Moving Average.


Seasonality
Animal Violations
1250
1000
750
500
250
0
20
40
Weeks


Moving Average
Animal Violations
1250
1000
750
500
250
2017
2018
2019
2020
Week

Heatmap

Before we modeled, we created a heatmap to depict the correlation between the independent variables in order to make hypothesizes about the model. Animal-related Violations had a strong correlation with Asian (0.94), American (.93), European (0.92) and Latin (.90) Cuisine Types. From that, we hypothesized that Asian, American, European and Latin Cuisine Types will have the highest coefficients in the model.



Fitted Models/Simulations

In total, we ran 10 models. Each of those 10 models were statistically significant, with p-values under 0.05.

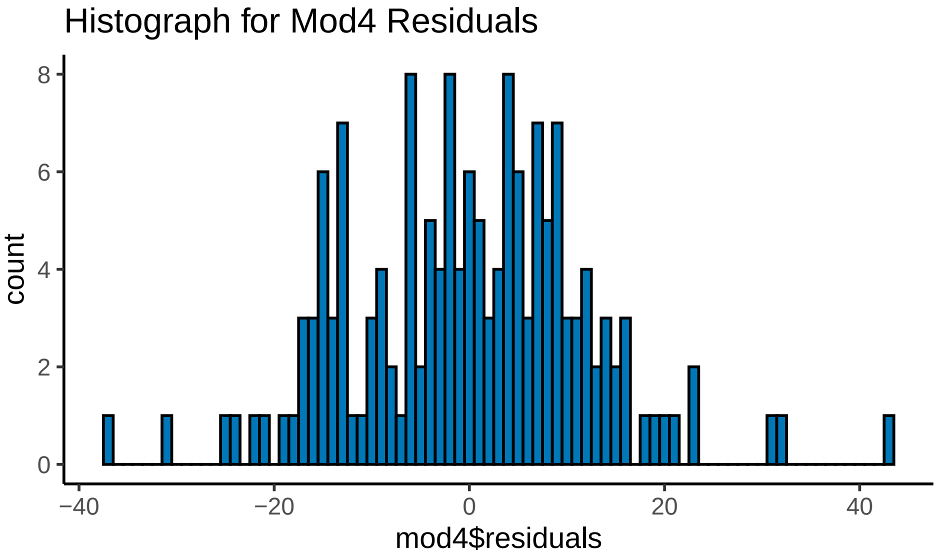
The first model we ran, `mod1`, modeled the complaint type variables against the dependent variable. This resulted in a low R-Square of .26.

In `mod2`, we modeled the cuisine variables against the dependent variable. `mod2` resulted in a much higher R-Square than mod1 at .99. The p-value was extremely low at 2.2e-16. While having a low p-value and a high R-Squared is usually the best-case scenario, we knew there had to be an error in the model. We hoped that adding the Cuisine Type variables to the complaint type variables would result in a better model.

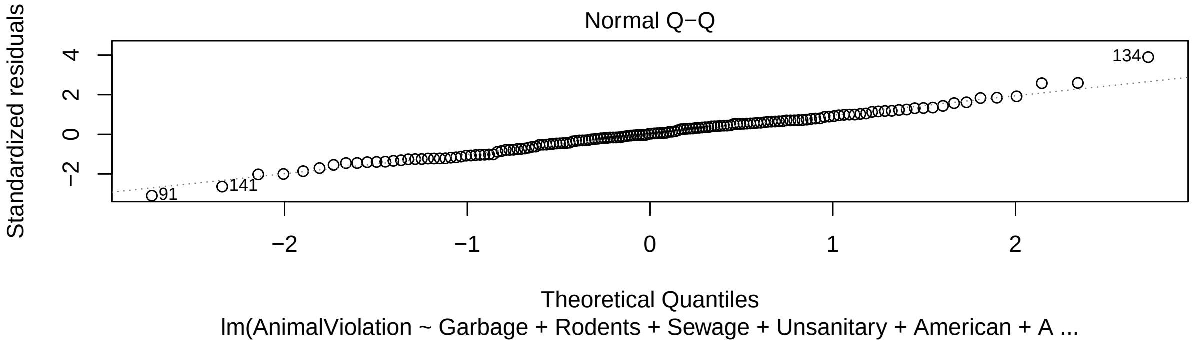
In `mod3`, we modeled all the variables together, apart from seasonality. `mod3` resulted in similar R-Squared and p-value as `mod2`. From this, we assumed that the Cuisine Type variables have greater impact on Animal-related Violations than the complaint type variables. We confirmed by checking the coefficients of each variable. The Cuisine Type variables had the highest coefficients, apart from Fast Food, which garnered a negative coefficient that was not statistically significant.

In `mod4`, we tested all the variables in `mod3` and added in seasonality, in hopes that seasonality will take some weight off the Cuisine Type variables. However, when we added in seasonality, it had a low coefficient of 0.001, and those results were statistically insignificant. R-Squared and p-value remained unchanged from `mod2` and `mod3`.

The normal distribution bell curve reflects the residuals of `mod4`. Most of the residuals are close to 0, indicating that the model was over-fitting.



In the graph below, you can see that the model is overfitting. After `mod4`, we decided to test different combinations of variables.



We tested different combinations of variables in models 5-10 in hopes of achieving a more realistic R-Squared. We dropped variables from models and combined them. We tested different combinations of the independent variables. With each model run, it became clearer that the models must be inaccurate due to the multicollinearity of the variables.

```
## # A tibble: 10 x 3
##    Model  ‘R Square‘ ‘P Value‘
##    <chr>       <dbl>     <dbl>
##  1 Mod 1       0.269 1.18e-  9
##  2 Mod 2       0.996 3.71e-178
##  3 Mod 3       0.997 4.56e-172
##  4 Mod 4       0.997 2.88e-170
##  5 Mod 5       0.997 6.51e-172
##  6 Mod 6       0.997 2.80e-171
##  7 Mod 7       0.997 4.05e-171
##  8 Mod 8       0.997 4.86e-173
##  9 Mod 9       0.997 8.67e-173
## 10 Mod 10      0.996 2.44e-176
```

To validate our models, we checked for multicollinearity by running Variance Inflation Factor (VIF) tests. VIF is used to "quantify the extent of correlation between one predictor and the other predictors in a model." If the VIF is higher than 4 or 5, multicollinearity is considered "moderate to high." If VIF is greater than 10, it is regarded as high.[2]

In `mod1`, each independent variable had VIF values lower than 4 which meant that the model results were not heavily impacted by multicollinearity. This was surprising to us, as the Garbage, Rodents and Unsanitary variables are closely correlated.

```
vif(mod1)
```

```
##    Garbage    Rodents     Sewage Unsanitary
##   3.326217   1.967423   1.098098   2.475453
```

In `mod2`, the American and Latin variables had VIF values of 4 while the Asian European variables had VIF values of 6 and 5, respectively. `mod2` has results that were impacted by the multicollinearity of the Cuisine Type variables.

```
vif(mod2)
```

```
##   American      Asian BakeryCafe   European    Healthy      Latin   FastFood
##   4.475033   6.145342   3.148983   5.141116   2.141671   4.182517   1.992073
```

From `mod2`, we recognized that each model ran with those Cuisine Type variables would end up with unreasonably high R-Squares given its multicollinearity between the independent variables. We tested VIF for each model there-after and confirmed that the Asian, American, Latin, and European variables consistently tested for high VIF values.

```
vif(mod3)
```

```
##    Garbage    Rodents     Sewage Unsanitary   American      Asian BakeryCafe
##   4.709568   2.763973   1.215878   2.682536   5.203434   6.746873   3.335305
##   European    Healthy      Latin   FastFood
##   5.298694   2.215921   4.475144   2.106930
```

```
vif(mod4)
```

```
##     Garbage     Rodents      Sewage  Unsanitary    American       Asian
##    4.840393    2.961419    1.221406    2.830097    5.230379    6.773482
##  BakeryCafe    European     Healthy       Latin    FastFood Seasonality
##    3.456457    5.443676    2.223826    4.588416    2.159398    2.135145
```

[2] https://www.displayr.com/variance-inflation-factors-vifs/

```
vif(mod5_rmUnsanitary)
```

```
##     Garbage     Rodents      Sewage    American       Asian  BakeryCafe
##    2.601834    2.876365    1.203765    5.230295    6.771181    3.419369
##    European     Healthy       Latin    FastFood Seasonality
##    5.437485    2.222793    4.585333    2.159396    2.023819
```

```
vif(mod6_rmRodents)
```

```
##     Garbage      Sewage  Unsanitary    American       Asian  BakeryCafe
##    3.035262    1.174288    2.748815    5.027215    6.710654    3.454448
##    European     Healthy       Latin    FastFood Seasonality
##    5.443411    2.156466    4.577382    2.088748    1.992788
```

```
vif(mod7_rmGarbage)
```

```
##     Rodents      Sewage  Unsanitary    American       Asian  BakeryCafe
##    1.857015    1.215072    1.521249    5.226954    6.724685    3.342819
##    European     Healthy       Latin    FastFood Seasonality
##    5.442337    2.215561    4.586567    2.135939    2.077437
```

```
vif(mod8_rmUnsanRodents)
```

```
##     Garbage      Sewage    American       Asian  BakeryCafe    European
##    1.378620    1.164793    5.019679    6.710623    3.419217    5.437476
##     Healthy       Latin    FastFood Seasonality
##    2.156322    4.571846    2.086774    1.917892
```

```
vif(mod9)
```

```
##                    Sewage Rodent_Garbage_Unsanitary                  American
##                  1.192622                  1.388593                  5.000049
##                     Asian                BakeryCafe                  European
##                  6.658276                  3.334418                  5.440854
##                   Healthy                     Latin                  FastFood
##                  2.164042                  4.565963                  2.085791
##               Seasonality
##                  2.032416
```

```
vif(mod10)
```

```
##                    Sewage Rodent_Garbage_Unsanitary                BakeryCafe
##                  1.174504                  1.246518                  3.321982
##                   Healthy                     Latin                  FastFood
##                  2.085064                  6.517879                  2.010208
##               Seasonality          Combined_Cuisine
##                  1.972186                  8.761324
```

Random Forest Model

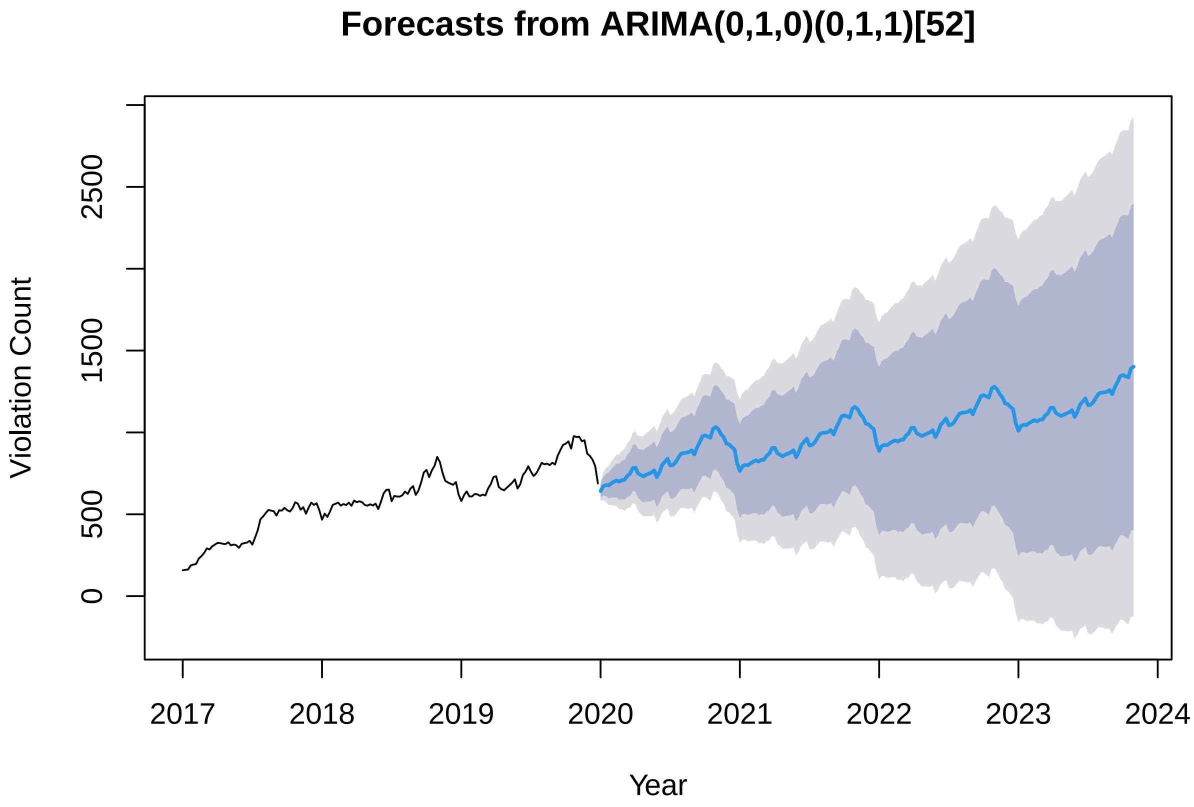
Although our regression analysis did not result in positive results, we tried to run a random forest model to see if we could gain any learnings from it. Going into the random forest model, we hypothesized that the Cuisine Types will be the most important given its high volume and strong coefficients. As expected, the Cuisine Type variables were the most important. American, Asian, European and Latin were the most important variables, as demonstrated in their high %IncMSE and IncNodePurity.

```
##              %IncMSE IncNodePurity
## American   14.129348    1376170.34
## European   12.926635    1414463.10
## Asian      12.580623    1388102.81
## Latin      11.495517    1207082.14
## BakeryCafe  8.629004     585745.63
## Healthy     6.691382     264479.78
## FastFood    4.418948     203314.54
## Unsanitary  3.928387     104683.51
## Garbage     3.213911      53432.50
## Sewage      2.971882      49798.40
## Rodents     2.132967      33946.38
```

Forecast Model

Due to the COVID-19 Pandemic, there was very few data for restaurant inspections, since NYC restaurants closed their doors in March 2020. Based on our earlier models, we noticed that the Animal-related Violations were significant throughout the three years and believed that the violations would've continued if restuarants were opened in 2020. We used a ARIMA model to fit and forecast the Animal-related Violations for the year 2020 and beyond.

ARIMA models require the data to be stationary. We saw earlier in our analysis, that our data is non-stationary, due to its seasonality. Therefore, we used the moving average component to make up a non-seasonal ARIMA model and mitigate for any anomalies in the data.



Based on the low MAPE in our training set (shown in the summary below), we can determine that our forecast model is accurate.

```
##                     ME     RMSE      MAE        MPE     MAPE       MASE
## Training set -1.902248 24.38827 15.81725 -0.3311785 2.291026 0.08610417
##                     ACF1
## Training set -0.03984505
```

Conclusion

Our goal in this project was to understand which variables influenced Animal-related Violations the most, to then forecast beyond our timeframe. Our linear model and random forest model results were inconclusive, given the multicollinearity present within the independent variables. Therefore, when we decided to forecast the rest of 2020 and beyond, we did not use those model results. Instead, we used the weekly timeseries forecast model to get a more accurate result. The timeseries forecast was able to forecast Animal-related Violations with high accuracy.