The Data Open

Improving 311 Service through Geospatial and Time Series Analysis

Neil Liu, Ankai Jie, Sai Praneeth, Priyank Jaini

May 12, 2018

1 Topic question

New York city's 311 non emergency call system sees an incredible amount of traffic every year. In 2017, the number of 311 contacts set its 4th consecutive annual record with nearly 40 million requests, surpassing 2016 by 11% (nyc.gov). With such a large volume of traffic, it is important for the government to be able to not only address requests efficiently, but allocate appropriate resources as well. With this in mind, we pose the following topic question in analyzing the dataset:

Topic Question: What suggestions can we make to improve the resolution efficiency of 311 service requests in New York?

Improved government services can do a lot for improving the daily life of New Yorkers. In this report we show that the structure of the 311 service complaints/ requests is a criterion develop a unique profile of the local city, thereby serving as an efficient and low-cost decision system for relevant city stakeholders. We will be analyzing two factors in particular to assist 311 service.

1. How can we accommodate the needs of people from different zip codes?
2. Can we use time series analysis to make predictions on call volume?

By answering these questions, we hope that 311-services can optimize their services to improve the health and welfare of NYC residents.

2 Executive summary

Through our analysis of the data, we discovered the following key insights.

Firstly, the distribution of 311 complaints varied with geographic area. We successfully clustered New York city to distinct regions based on the distributions of 311 complaints at different zip codes. Furthermore, we used exploratory factor analysis to discover that there are certain key qualities in a geographic area that correlate strongly with complaint types.This insight leads to the conclusion that any policy and action changes to 311 resolution handling (e.g. determining where to place specialized employees) should take into account the qualities of the region.

We also strongly suspected that 311 regions could act as a predictor or at least indicator for socio-economic performance on a regional level, but we were unable to complete a detailed analysis on time.

The second insight relates to how service requests activity varies over time. We discovered both long term and short term trends in seasonality for 311 service requests, scoped to geographic locations. We found that for regions in New York, the busiest days for 311 requests range from June - August. We used these trends to forecast 311 request volumes, conditioned on areas.

311 workers can easily use these trends and forecasts to anticipate periods of high and low activity, leading to efficiency in addressing caller needs. This represents a low-cost and quick way to make the lives of service workers more predicable and accurately equalize resolution activity with request activity.

3 Technical Exposition

Beginning our analysis, it is important to note the geographic heterogeneity of the data. In particular, the 311 data is limited to New York City while all other data is county-level on a state basis. Hence, the other data were not useful pertaining to the 311 data.

3.1 Initial Exploration

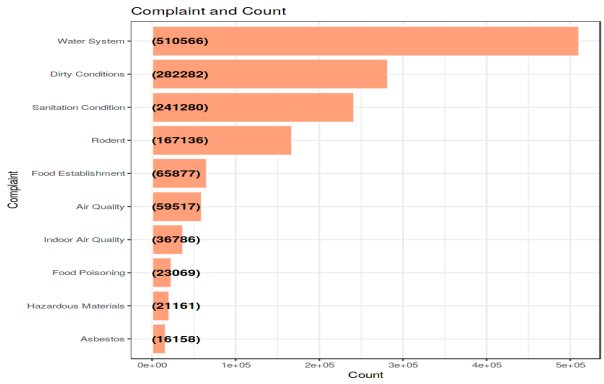

Figure 1: Frequency of 311 Complaints

New York is a big state and depending on the demographics, 311 requests types can vary greatly in different areas. We wanted to extract the differences between 311 requests in different communities, so that we could make tailored suggestions on how to improve the 311 system depending on the community. We began be conducting some exploratory classifications on the data. Fig. 1 gives of a summary of the top complaint types.

For different complaint types, we analyzed how important they were geographically by creating a density heatmap over the map of New York. For example, Fig. 2 depicts the geographic significance of 4 top complaint types in New York.

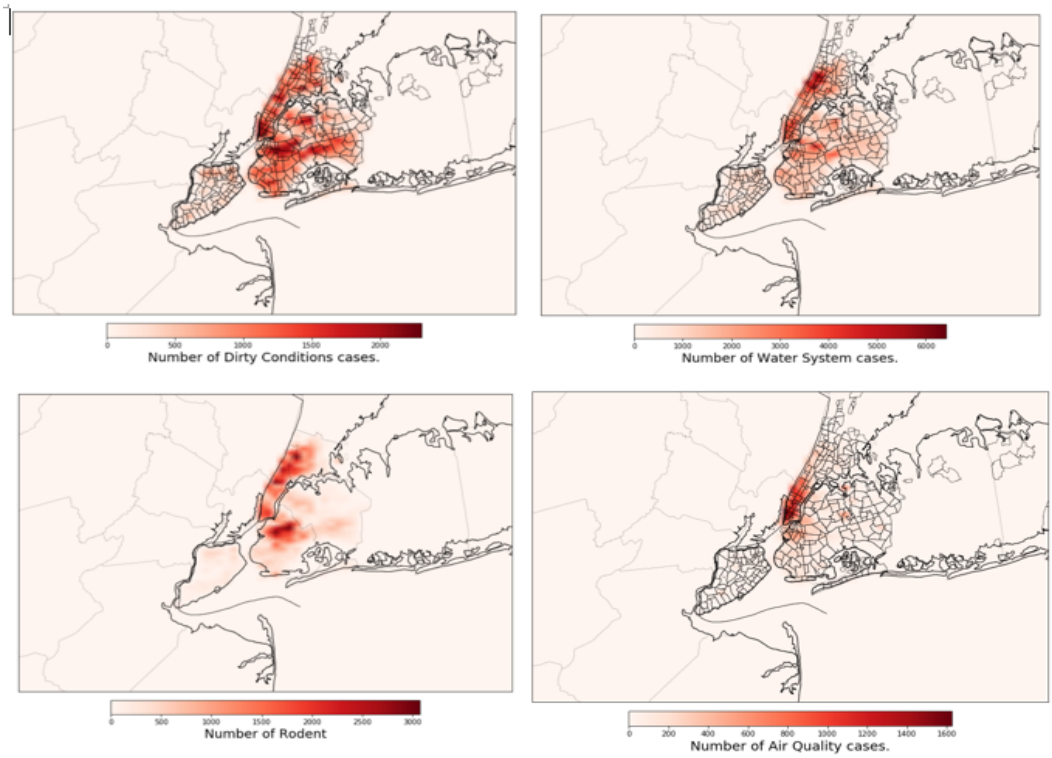

Figure 2: Four different Heatmaps of top 311 Case Complaints

What we noticed was that the profile of 311 complaints could differ drastically depending on the geographic region. Air quality complaints, for example, are only especially prominent in the Manhattan area whereas Rodents tend to appear prominently in both Manhattan and Brooklyn.

We decided that any suggestions to improve the 311 resolution system should be tailored to different geographic profiles. As such, we decided to cluster the 311 complaints based on location data.

On the time series side, we wanted to know if there were seasonality to such requests that can be modelled with an Arima or Holt-Winters model.

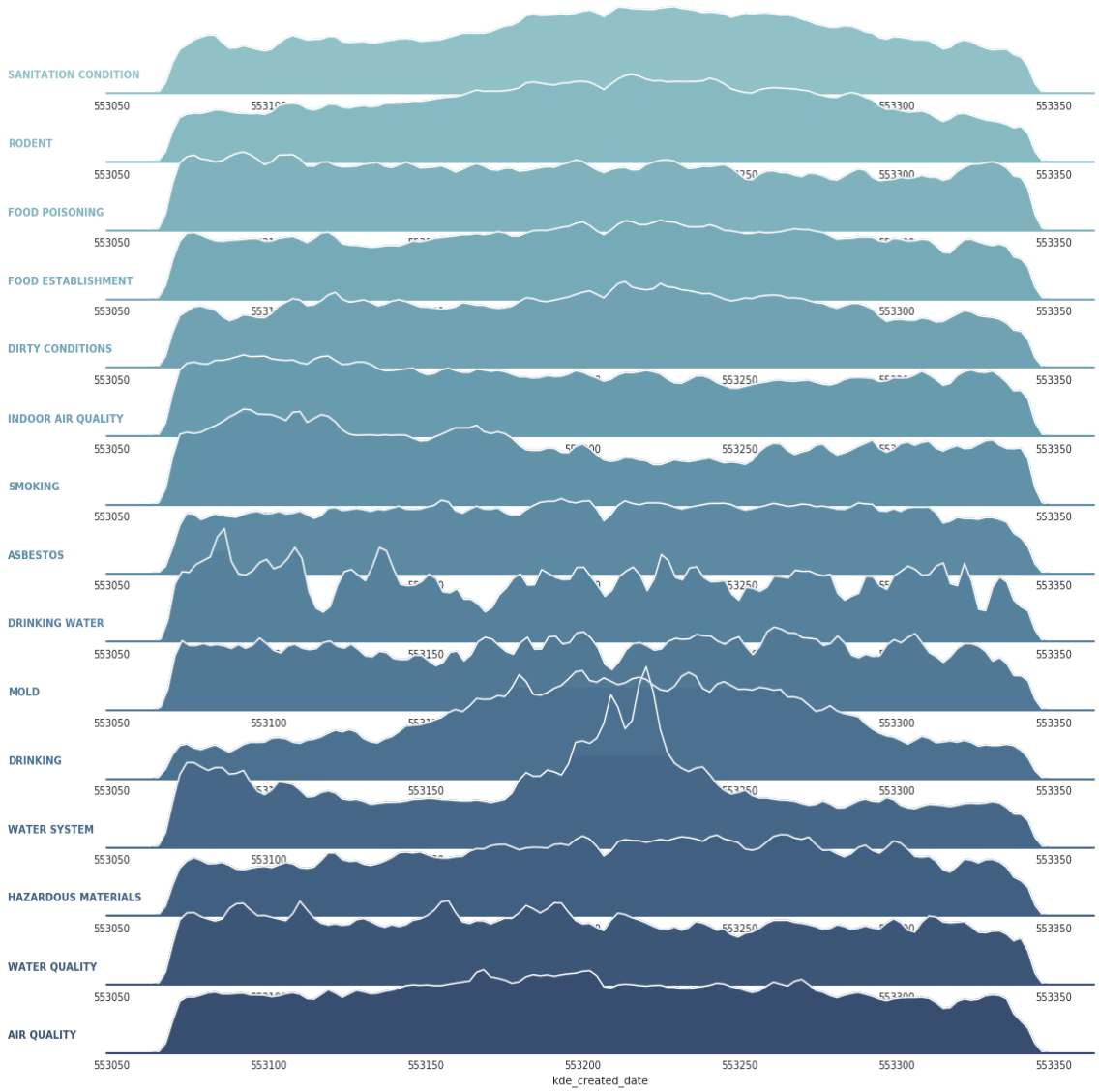

Figure 3: Time density plots of the top 15 complaint types across a year.

We discovered that there exists seasonality of complaint types. Things related to drinking and water systems evidently spike up in summer season which is expectedk. On the other hand, smoking and indoor air quality seems to be more important within the winter seasons. Hence, we know that there exists a possibility to model the concentration of 311 calls with a time series model.

3.2 Classification of 311 service categories

Our goal was to separate New York into distinct 311 sections. We did this by classifying zip codes of the complains to different clusters. There was a total of 245 unique zip codes in the data to classify.

To begin this analysis, we decided to cluster the 311 requests based on the relative frequencies of requests types, as taken from the "complaint type" column of the 311 services requests dataset. The top 15 most frequent complaint types were selected to be the features. We define a 311 pattern signature which is a vector of relative frequencies for each of the top 15 distinct features. We calculated each relative frequency as follows:

$$\mathbf{X}_{ij} \leftarrow \frac{\mathbf{X}_{ij}}{\sum_{j=1}^{1} 5\mathbf{X}_{ij}} \qquad \mathbf{X} \in R^{245\times 15} \tag{1}$$

However, specific complaint types such as water systems and dirty conditions dominated others because of the sheer number of reports. Hence, the data was further normalized against the average of each complaint for all districts.

$$\mathbf{X}_{ij} \leftarrow \frac{\mathbf{X}_{ij}}{\frac{1}{245}\sum_{i=1}^{245} X_{ij}} \tag{2}$$

This helped better create understandable plots.

So now, each datapoint has a zip code and a relative frequency vector. Aggregating by zip code, we obtain a relative frequency matrix, each row corresponding to the calculated vector of a data point for that zip code. To classify each of the zip codes, we then applied a k-means clustering algorithm, running the k-means 100 times with different centroid seeds to obtain the best clustering.

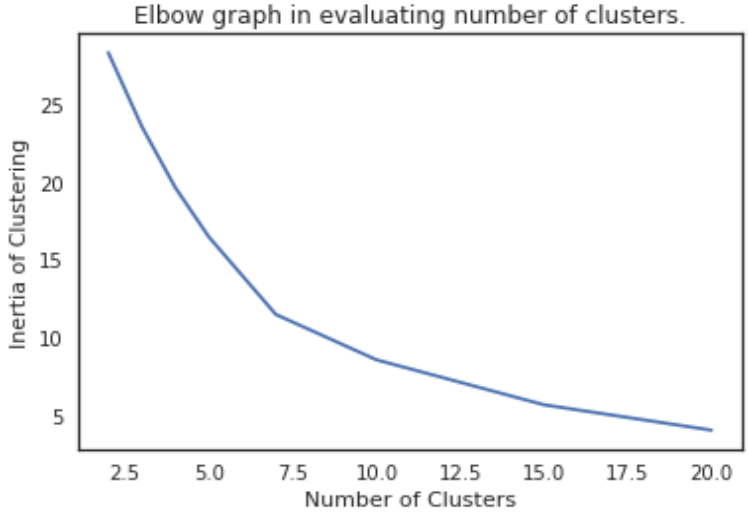

Figure 4: Figure for elbow method

An important step in our approach is to pick the number of appropriate clusters. We evaluated our clustering model using the Elbow method and found that a total 10 clusters defined a maximum proportion of variation in the data (see Fig.3). We could have chosen lesser number of clusters as well but for this report we chose 10 clusters. Part of Queens falls in clusters 1, 4 and 5 respectively. Cluster 2 comprises of parts of Bronx and Queens. The TSNE embeddings as shown in Fig. 5 show that the clusters are well formed thereby reinforcing the number of clusters we used in our method.

In the first figure in Fig. 6 we divide the NYC map into these 10 clusters. The maps reveals that mostly 4 clusters appear often according to 311 service requests. Cluster 9 is the largest and includes Staten Island, Eastern and Southern Brooklyn, Flushing and some eastern parts of Bronx. Cluster 0 includes most parts of Manhattan.

In order to evaluate how different each cluster is with respect to the nature of 311 service requests, in the second figure of Fig.6 we present the distribution of top service requests for each cluster. The figures clearly show variation in the distribution. Cluster 9 mostly gets service requests related to Water System, Dirty Conditions

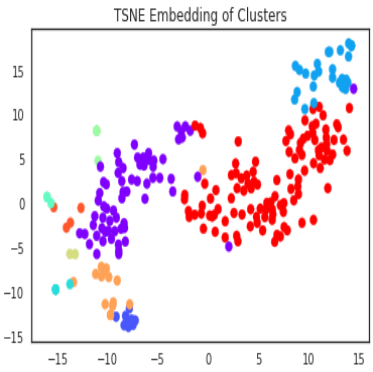

Figure 5: TSNE Embedding of 311 Clusters

and Sanitation Conditions. This is expected given the areas cluster 9 covers in the NYC map. Cluster 0 mostly results in complaints related to rodents and air quality. This is also expected given that cluster 0 covers most of Manhattan with its congested traffic conditions. Service requests with complaints about rodents are most prevalent in cluster 6 and 7 around parts of Queens area. Similarly, Food Establishment related service requests are also mostly from areas in clusters 1 and 7 which forms part of Queens and Bronx. Smoking related requests are mostly from clusters 6 and 8 which is mostly parts of Queens area. Due to time constraints we were not able to collate this with socio-economic data to predict how these complaints vary with average income distribution and health coverage. We conjecture that socio-economic indicators can serve as proxy for the type of 311 service requests generated from a particular area.

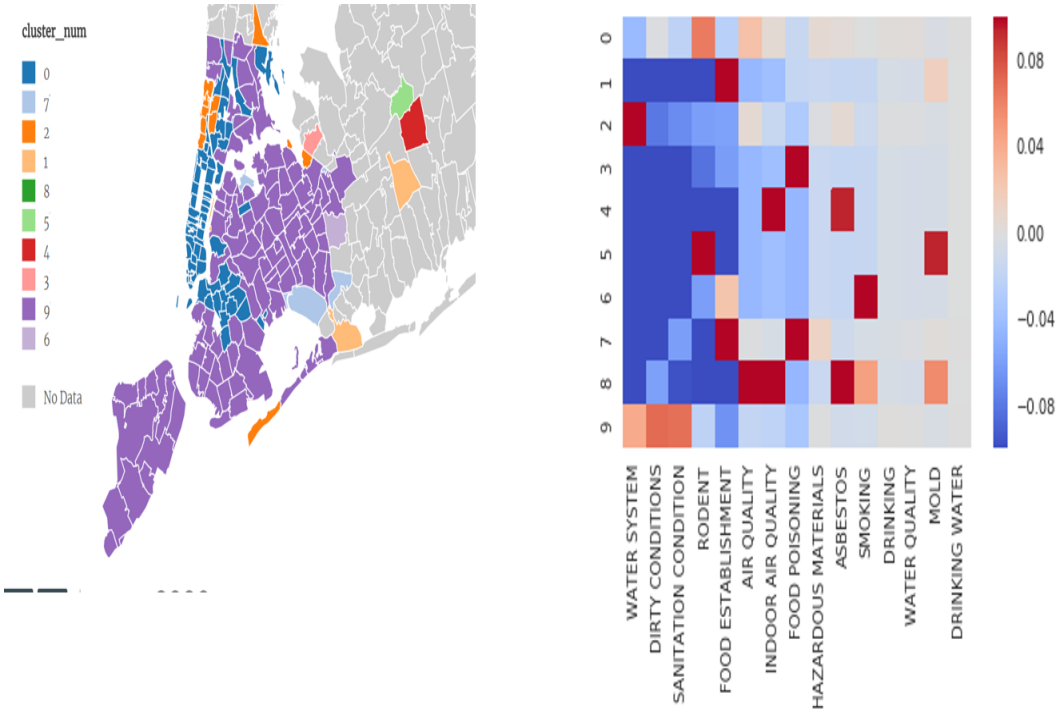

Figure 6: (a) Classification of urban location based on categorical structure of 311 service requests. (b) Patterns of 311 activity within clusters for top 15 service requests and their relative frequency in the distribution. A high positive value or high negative value implies high frequency

3.3 Exploratory Factor Analysis

We also performed Exploratory Factor Analysis (EFA) on the cleaned data to explore the underlying structure of the data set to understand if any latent variables might explain the variance seen in multiple predictors and show the results in Fig. 7. The results show seven factors have multiple variable loadings ≥ 0.8 indicating there are

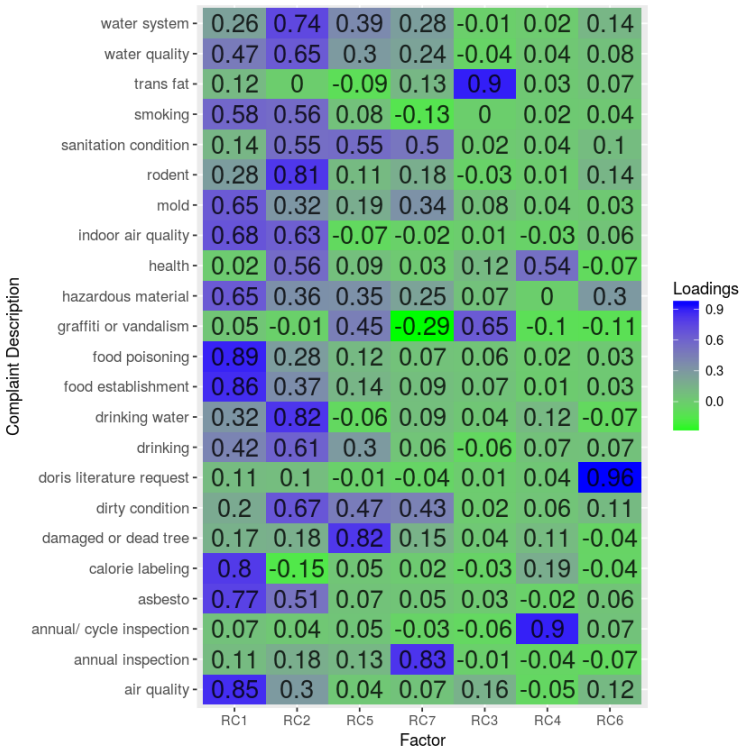

Figure 7: Exploratory Factor Analysis for Complaint Description

four latent variables which cause residents to make similar complaints. This helps us in precisely stating the first of our recommendations : these latent variables can help to predict the most likely complaint associated with an area. This would help in making the response and decision support for 311 support requests fast and efficient.

3.4 Time Series Analysis To Predict Future 311 requests volume

We wanted to analyze changes in 311 data over time in order to both model and forecast activity. We started this analysis with seasonality. Fig. 8 depicts the total 311 call activity on a per month basis. We can see that request activity goes up consistently in the middle of the year. This indicates a consistent seasonal shift in activity. It is beneficial to government workers to have a predictable time to prepare for higher activity.

Some of the findings, shown in Fig.10(A) and (B) are :

- July , June , August are the busiest months
- 22, 19, 21 (the third week) are the busiest days in the month
- 4, 1, 19 are the least busy days in the month
- Tuesday , Monday and Wednesday are the most busy days in the week
- 12 , 9 , 10 are the busiest hours in the day

For the monthly and weekly forecast (see Fig.9) we divide the dataset into 2 parts, the train dataset having entries having the first 40 entries and the test dataset having entries having the next 5 entries. We perform a ARIMA forecasting and evaluate the Root Mean Square Error for the test set. The predictions are 13539.58 14012.89 14086.07 14011.03 13934.79, and the RMSE is 5014.285. and for the short term weekly forecast same as before we divide the dataset again into 2 parts, the train dataset having entries having the first 188 entries and the test dataset having entries having the next 5 entries.

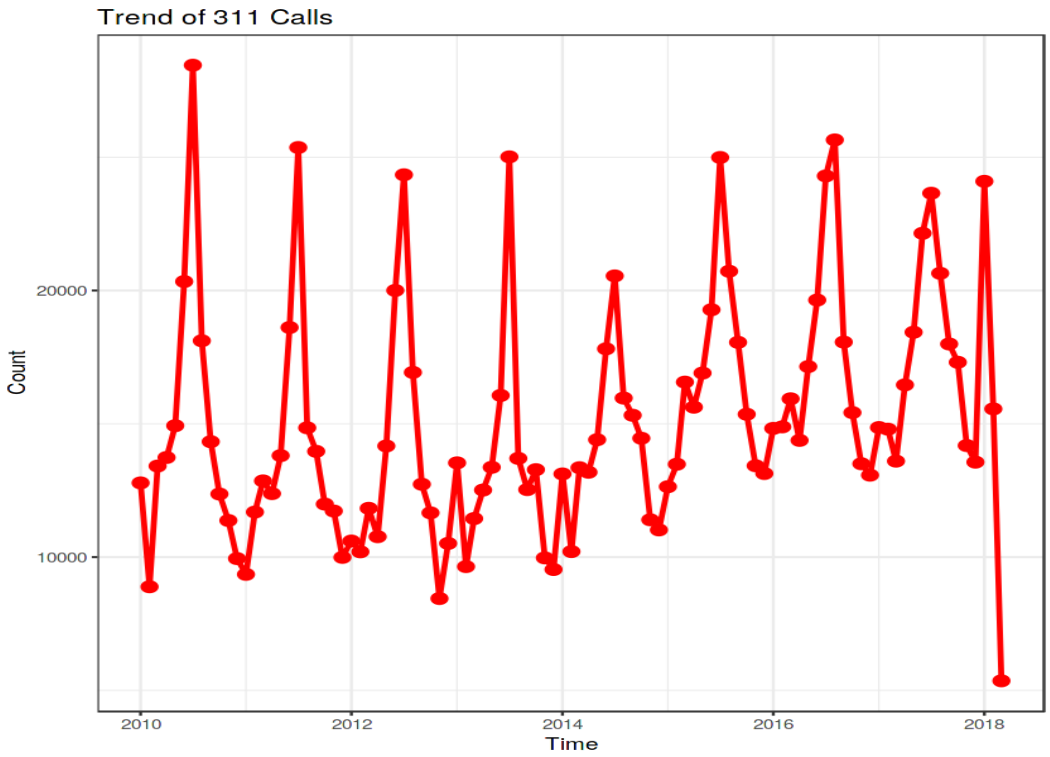

Figure 8: Frequency of 311 Requests by Year

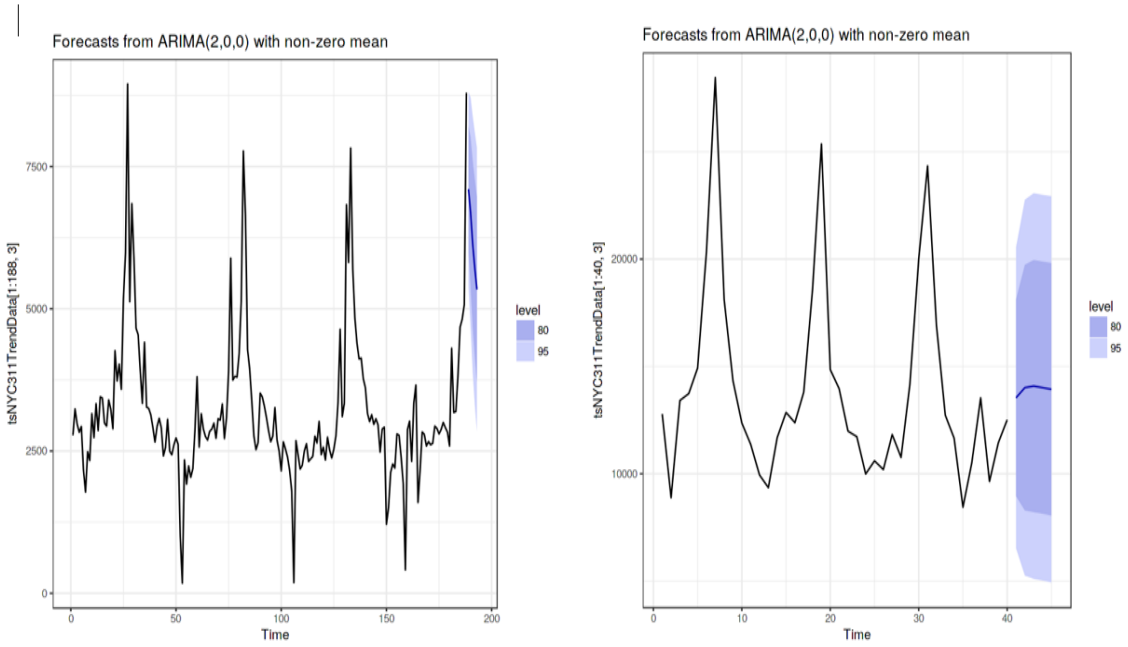

Figure 9: (a) Weekly forecasts for complaints shows clear seasonality. (b) Monthly forecast for complaints

3.5 Conclusion

In conclusion, we have identified the needs of different zip codes across NYC using unsupervised learning. With time series analysis, we present actionable insights that can help NYC services to better optimize the call centre schedule. We hope that these findings can help improve 311 services in NYC.

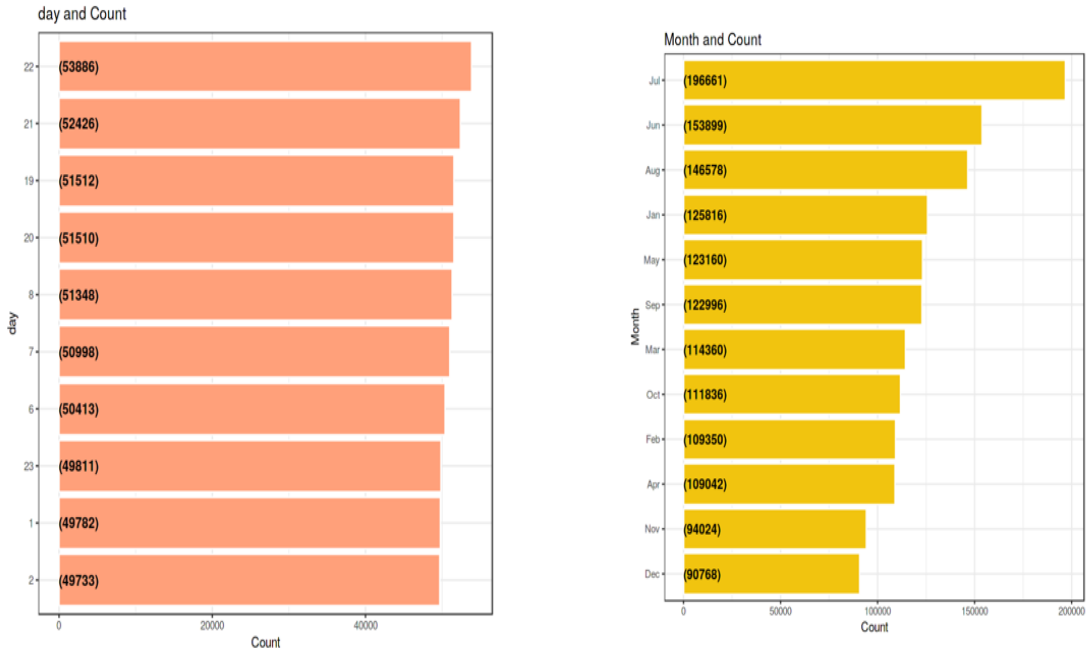

Figure 10: (a) Frequency of calls on (A) Day Basis and (B) Month Basis