311 for your help

TEAM *Z*

Saurabh V Gagpalliwar
Ganesh N Prasad
Naveen Kalaga
Shriram Rangarajan

You are in IBM headquarters doing predictive modeling on data given by your client. You feel happy about it. You found a very accurate model and are going to showcase it to the client tomorrow. With a sense of pride, you see out of the glass window only to see your car being moved. It is snowing heavily and your car was parked in the snow emergency route. How do you get home? You need some help?

Here comes 311 for your help. You call 311 and they send you an escort vehicle to drop you home. It's 3AM and they eventually make it. You are home!

311 helped bring the IBMer back home.

What if the vehicle reaches you faster? We want quick solutions to household problems like gas, fire, sanitation etc.

How do we optimize the efficiency of the 311 office? We need to allocate sufficient resources seeing measures such as snowfall, temperature, precipitation and other factors. We also need to see what specific type of complaints come from a particular area. So there is a lot we can do, let us get going.

Let's collect data before we start:

So what data could be related to 311 calls?

New York Public Works Department releases the 311 call center call log to the public. From that source, we used features such as
-> complaint type (311 classifies the complaint into a type like 'snowfall', 'parking issue'),
-> agency name (agency that will handle the problem eg. Police, Public Works),
-> location coordinates (latitude and longitude)

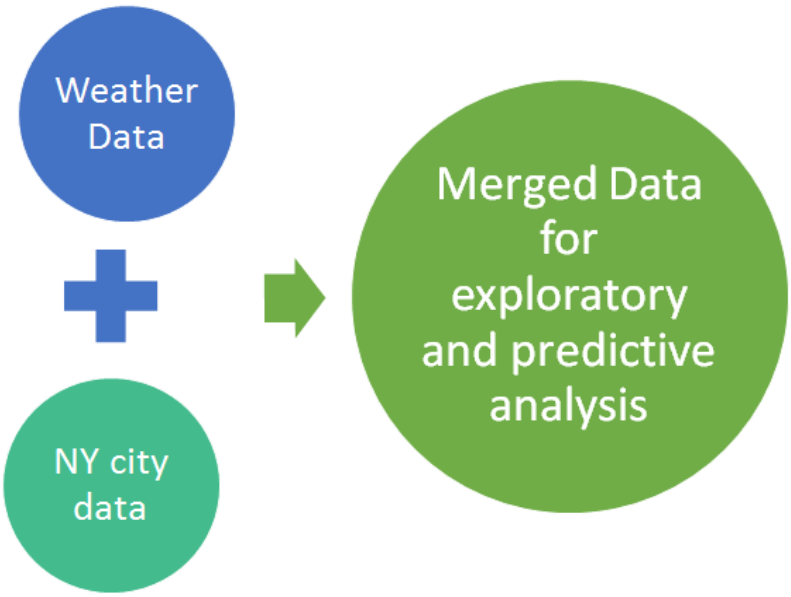
We also got the weather data from National Climatic Data Center to understand how complaints correlate with climatic changes. This is based on an assumption that climate is a statistically good predictor of the complaints. We conducted our initial analysis by merging all complaints with climate.

We started our initial analysis with core climatic features such as
-> Snowfall
-> Maximum Temperature
-> Minimum Temperature
-> Precipitation

How can these measures be indicative and accurate at the same time?
We calculated the nearest measuring station based on the complaint location to understand climatic diversity based on location.

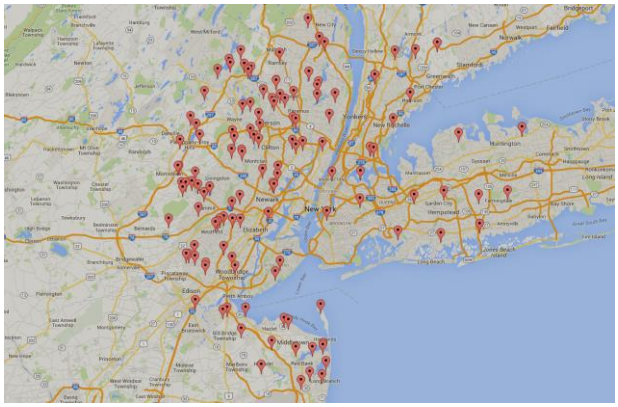
Another data source is the population density data. This is will give an information about the population of each borough in every year and we can then estimate the number of complaints based on the number of people. This will help in better allocate resources by understanding complaint diversity geographically.

How do we merge these data sources?



Weather data has location coordinates of the station that recorded those weather readings. We calculated the geodesic distance (distance of two points given latitude and longitude) from the location of the station to the location of the complaint coming from the boroughs.

Then we took the nearest station to understand how the climate would be when we got a complaint.



Map of Climate Measuring Stations in NYC

We also aggregated the data sets based on the 5 boroughs, to match the population of each borough to the number of complaints coming from that borough.

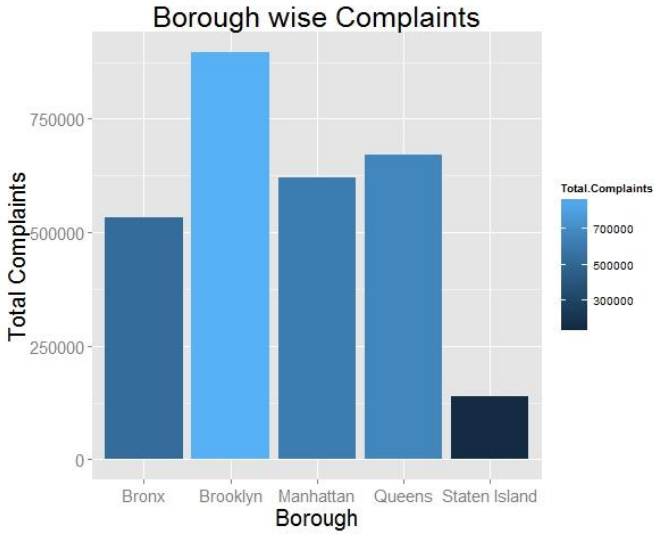
Jumping into Analysis:

Lets define some important components which we need before our Analysis -

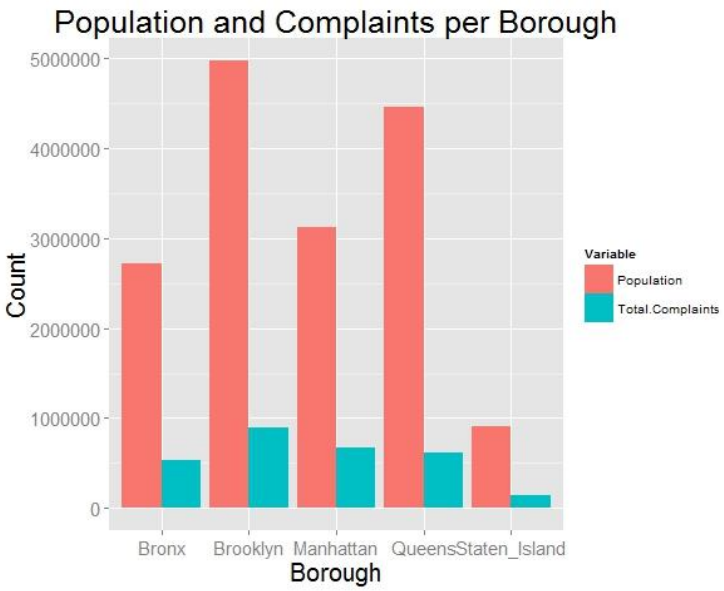
Exploratory Data Analysis:

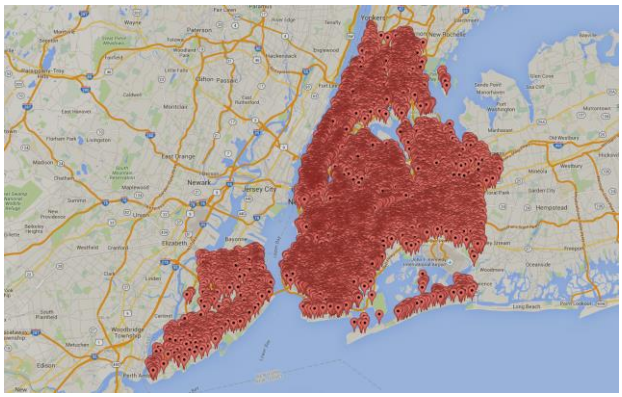
Before starting our analysis, we wanted to understand our data better and did some basic exploration.

The below graph shows how No. of complaints vary across different boroughs



We wanted to understand if population directly corresponds to No. of complaints. The graph below tells us that the population does not directly correspond to No. of complaints.





The original dataset has location coordinates

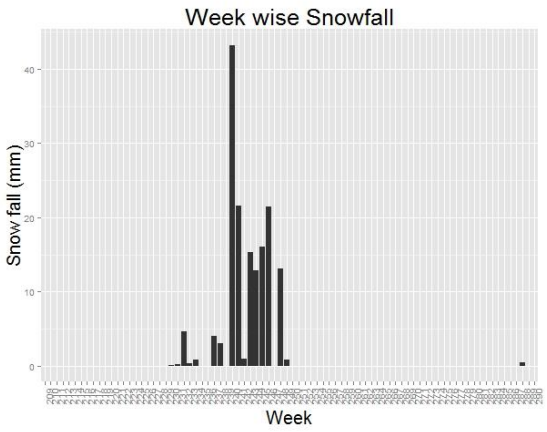
Map of Sample complaint density

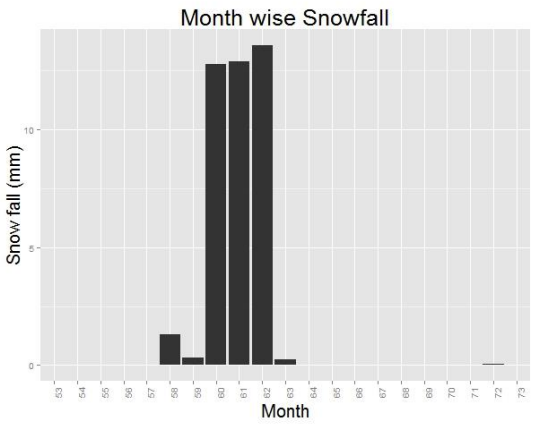
The above graph suggests we need a level of granularity to understand the data and correlation better.

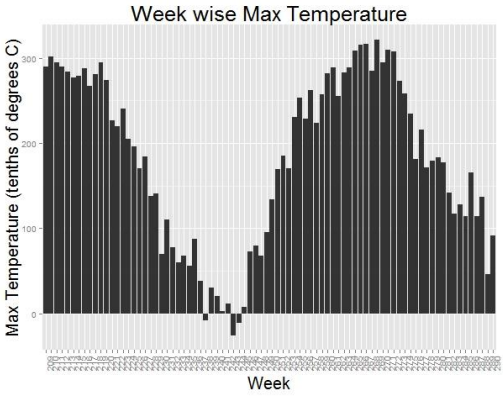
Level of Granularity:

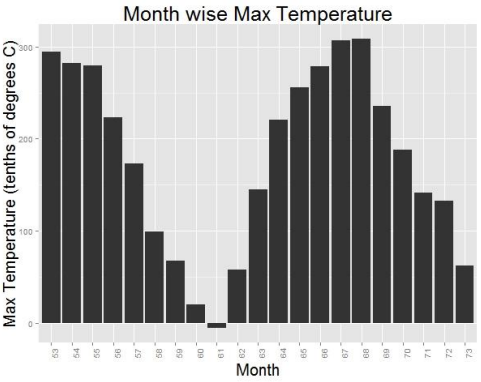
We aim to determine an optimum level of granularity which is statistically significant enough with climate changes. This, we aim to find by doing exploratory analysis on weather data and how the climate changes for various levels of time viz., daily, weekly, monthly.

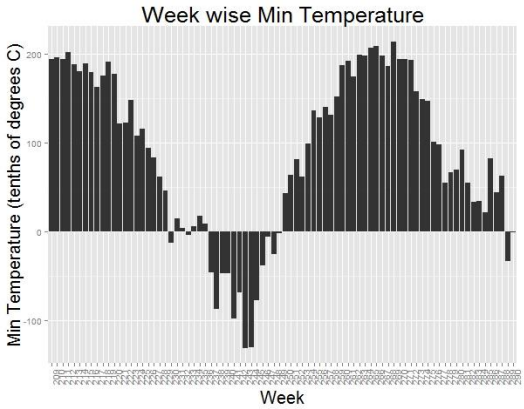
The graphs below show sample analysis on how the core climate features temperature and precipitation levels change with various levels of granularity. Ideally, we would see sharp spike in climate for a given optimum level of granularity.

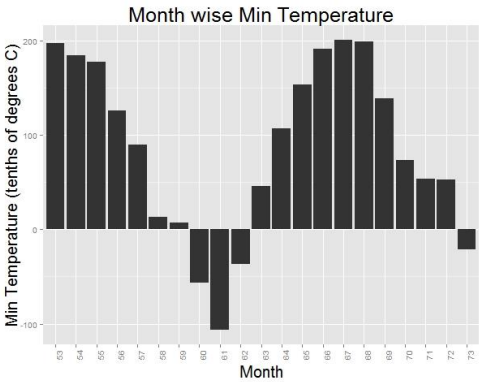


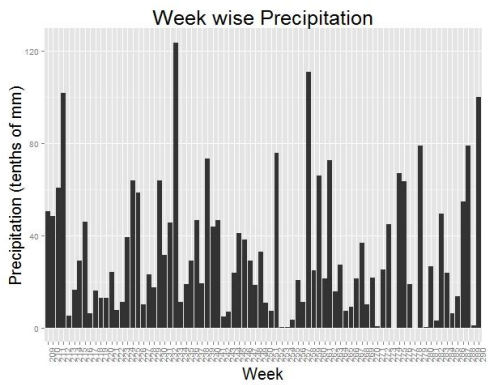


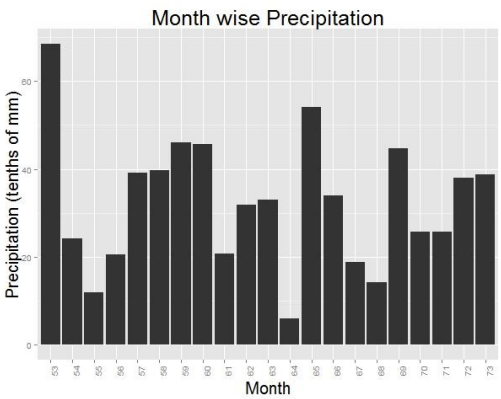












This shows sample analysis. We aim to look at all climatic ranges for the final report

Feature Selection:

Not everything is important for analyzing and predicting the number of complaints. How do we eliminate the redundant and not contributing features?

First, we removed the columns that were redundant for example -a column named agency was just a short code from the actual agency name column. So it was not providing any extra information to predict the number of complaints in a particular area.

Then we checked the correlation coefficient of the number of complaints against independent variables like complaint type, snowfall rate, precipitation etc.

Prediction:

Our first analysis was an aggregation of the number of complaints on the month and the complaint type. So we had say 30 complaints in the month of March for 'Illegal Parking'. Then what to do with 2 factor variables as the input and the output was the number of complaints. We ran the anova test which is used when the input variables are factors.

Here is the screenshot of the results

```
> summary(aov(Total.Complaints ~ Month + Agency.Name, data = nyca_bronx_agg))
             Df     Sum Sq  Mean Sq F value              Pr(>F)
Month        11    9070422   824584   0.772               0.669
Agency.Name 237 4396259272 18549617  17.360 <0.0000000000000002 ***
Residuals   762  814196230  1068499
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(Total.Complaints ~ Month + Complaint.Type, data = nyca_bronx_agg))
                Df    Sum Sq Mean Sq F value              Pr(>F)
Month           11   3817449  347041   1.335               0.198
Complaint.Type 171 852395331 4984768  19.175 <0.0000000000000002 ***
Residuals      2349 610640655  259958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How can we use this?

If we know the month and the complaint type, we can run this model to predict the number of complaints.

Linear Regression Model:

Once we had the merged information from 311 call logs dataset and the weather dataset we aggregated the different climate parameters like snowfall, maximum temperature, minimum temperature, precipitation weekly to get the number of complaints. We fitted a linear regression model using these parameters to predict the number of complaints for a particular week given the weather conditions.

```
Call:
lm(formula = as.formula(formula), data = merged_agg)

Residuals:
       Min         1Q     Median         3Q        Max
-4.800e-15 -5.349e-16 -1.056e-16  4.507e-16  8.543e-15

Coefficients:
                Estimate Std. Error    t value Pr(>|t|)
(Intercept)   -1.736e-14  2.022e-15 -8.584e+00 1.24e-12 ***
Weeks          4.057e-17  7.175e-18  5.655e+00 2.95e-07 ***
TMAX          -7.514e-18  1.137e-17 -6.610e-01 0.510802
tmin           2.103e-17  1.212e-17  1.734e+00 0.087110 .
prcp          -9.604e-18  3.757e-18 -2.556e+00 0.012687 *
snow           5.000e-19  2.119e-17  2.400e-02 0.981241
snowd          2.691e-17  7.353e-18  3.660e+00 0.000477 ***
no_complaints  1.000e+00  1.476e-17  6.774e+16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.447e-15 on 72 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.096e+32 on 7 and 72 DF,  p-value: < 2.2e-16
```

Achievements till date:

- Data Sources Finding
- Data Cleaning
- Data Merging
- Exploratory Analysis
- Feature Selection
- Basic Anova and Linear Regression

Future Scope:

- Exploring Feature Selection techniques
- Predictive Modeling for the entire data set
- Text Analytics of Complaint Description
- Clustering to find weather related tags
- Visualizing call spikes based on natural calamities