Final Project: What do the 311 calls tell us?

Reshma Elizabeth Roy Kurian

2018-03-15

Introduction

This report analysis the NYC311 call data. The NYC 311 call center provides the residents of New York city 24/7 help with more than 3600 non-emergency government services.The call center logs the service requests and complaints and routes it to the concerned government agencies. When the problem is resolved, the call log is closed.

The objective of this project is to understand the NYC 311 data and explore possible solutions to predict/anticipate complaints. This report will provide some insights which could help the agencies reduce the volume of different complaints in the future. R has been used to conduct visual analytics on the data and to generate this report. NYC population census data and NYC Weather data are used in this project to analyse of the 311 call data from different perspectives.

Preparing the datasets

1. Initialization & Reading of the 311 data file

At the beginning of the rmd, the tidyverse packages & the `data.table` package are loaded. Then the NYC311 data set is loaded and its column names are fixed to remove spaces.

Description of nyc311 dataset

The Data in the 311 csv file has 52 columns and over 9Million rows(9124937 to be exact). Each row in the file refers to a 311 complaint call and logs the nature, schedules and actions taken.

2. Cleaning the nyc311 dataset

Removing duplicate rows

The unique key column is removed and the rows are checked for duplication using distinct() and all_equal().

```
## [1] "Different number of rows"
```

853538 rows are identified as duplicates and these are removed.

The resulting dataset has 8271399 rows.

Columns that are redundant/doesnt add value (26 columns in all) are dropped

A description of the columns that are dropped is given in the appendix#2. This leaves nyc311_1 with 25 columns.

Removing rows given as Unspecified in Borough

Borough column is cleaned for the entry "Unspecified" to retain only the rows with calls recieved from one of the 5 NewYork.

This removes 859,858 rows from the dataset and results in the nyc311 data with 7411541 rows.

Formatting & modifying some relevant columns

1. Working on Date and time columns:

There are 3 the date-time columns in the dataset namely, Closed.Date, Created.Date and Due.Date. All the 3 columns are of the datatype "chr". In order to use these columns for data & time operations, their format are changed to POSIXct using Lubridate package. In addition the following columns are added to the data frame:

RTime : This indicates the total time taken to resolve the complaint and is calculated as the difference between Closed Date and Created Date. All erroneous rows with negative RTime are filtered out in this chunk.

Weekday : This column indicates the day of the week the complaint call was made.

Hr : This column indicates the hour the complaint call was made

Mt:This column indicates the month the call was created

Yr : This column indicates the Year of creation of the call service request

Created.Date & Created.Time : The column Create.Date of the dataframe has both Date and Time which is separated into 2 columns with only Date in mdy format and only Time in hms format.

This leaves 30 columns in the dataset.

2. Modification of Street columns

Cross streets and Intersection streets are joined for better readability.The resulting data frame is nyc311_1. This data frame and its subsets will be used for exploration and analysis of the 311 data. It has 26 columns and 8065672 rows of 311 calls.

Creating a dataframe from a subset of nyc311 only for noise complaints

Creating a dataframe from a subset of nyc311 only for heat related complaints

Sampling 100000 rows for testing code chunks

A sample of 100000 rows are taken from nyc311 to use with maps.

The NYC311 data table

Find below a table with some relevant columns from NYC311.

Table 1: Table continues below

Agency	Complaint.Type	Status	Borough	Hr	Mt	Yr	RTime
NYPD	Vending	Closed	BRONX	2	4	2015	48.7
NYPD	Noise - Street/Sidewalk	Closed	MANHATTAN	2	4	2015	47.48
NYPD	Noise - Street/Sidewalk	Closed	BROOKLYN	1	4	2015	18.92
NYPD	Derelict Vehicle	Closed	QUEENS	1	4	2015	137.7
NYPD	Noise - Street/Sidewalk	Closed	BROOKLYN	1	4	2015	37.47
NYPD	Illegal Parking	Closed	QUEENS	1	4	2015	38.4

Created.Date	Location.Type
2015-04-14	Street/Sidewalk
2015-04-14	Street/Sidewalk
2015-04-14	Street/Sidewalk
2015-04-14	Street/Sidewalk
2015-04-14	Street/Sidewalk
2015-04-14	Street/Sidewalk

Reading the New York Census population file for all counties of New York.

This file has population data from 1970 to 2018 for New York. It will be read with all spaces in column names replaced with a dot. For example, Program Type becomes Program.Type

Description

The population dataset contains census data of New York counties. Census data is collected once every decade and population estimations are made there after and corrections made to previous estimations. The data in this dataset is based on the 2010 census. The data includes “intercensal estimates” for 1970-2009 and postcensal estimates for 2010-2018 and the decennial Census counts for 1970-2010. This contains 3339 observations and has 5 columns as listed in the data dictionary in the appendix.

Cleaning the population dataset

Selecting relevant rows from nycpop.

The selection criteria used are: keep only those rows for the years 2009 to 2015 & keep only those rows corresponding to the 5 Boroughs. The FIPS codes corresponding to the NY boroughs are 36005 for Bronx, 36047 for Brooklyn/Kings, 36061 for Manhattan/New York, 36081 for Queens & 36085 for Staten Island/Richmond

Preparing the population dataset for joining with nyc311

2 columns are prepared for the join with nyc311:Year & Geography

1. The Year column is formatted to numeric so it can match with Yr from nyc311
2. Entries in the column "Geography: of nycpop will be renamed to their corresponding Borough names so that they can match with entries of the column Borough in NYC311. This column, thus, can be made the common column for the join with nyc311. The following replacement will be done for the values in Geography

BRONX will replace Bronx County

BROOKLYN will replace Kings County

MANHATTAN will replace New York County

QUEENS will replace Queens County

STATEN ISLAND will replace Richmond County

Rows for 2010 with Postcensal Population Estimate are removed, so that the 2010 rows for each borough has the Censal Base Population numbers. After removing these 5 rows, the Program.Type column is dropped. FIPS Code is also dropped as it is redundant with Geography

The head of the population data table

Here we produce a header table with some rows of data.

Geography	Year	Population
BRONX	2015	1460412
BROOKLYN	2015	2643546
MANHATTAN	2015	1657183
QUEENS	2015	2346005
STATEN ISLAND	2015	475313
BRONX	2014	1445800

Reading file with New York daily weather data

Description

This dataset gives the daily weather information of New York from 2009 to 2015 and was obtained from the National Oceanic and Atmospheric Administration (NOAA).
All the data in this dataset were measured and collected from the weather station at JFK International Airport. It contains 2556 rows and has 30 columns as listed in the data dictionary.

Cleaning the weather dataframe

Removing irrelevant columns

Dropping columns STATION and NAME :all data are measured from the same weather station (USW000094789) located in JFK International Airport.

Populating the Daily Average temperature TAVG for all rows.

The average temperature, TAVG is not NA only for rows from April 2013. So, TAVG is calculated from TMAX and TMIN and populated for all rows.

Creating an Extreme weather data frame (nycweatherextr) from the weather dataframe.

All columns with Weather Type data i.e., columns 10 to 28 labelled WT01 and so on. All these different WT columns are gathered into a single column.for a dataframe of values with extreme weather condition

The head of the weather data table

Here we produce a header table with some rows of data.

DATE	AWND	PRCP	SNOW	TAVG	TMAX	TMIN	WESD
2009-01-01	18.57	0	0	22	28	16	0
2009-01-02	12.75	0	0	30	37	23	0
2009-01-03	16.55	0	0	35.5	41	30	0
2009-01-04	11.63	0	0	35	44	26	0
2009-01-05	10.29	0.01	0	41	45	37	0
2009-01-06	7.83	0.03	0	35.5	39	32	0

Joining datasets

A. Joining 311 dataframe and Population dataframe

nyc311 is joined with the population dataframe nycpop on the columns with Borough and Year data. The join is a left join and gives 7411541 rows and 7 columns.

Table of the population and nyc311 join data

Here the columns of the joined dataframe and a few rows are displayed

Geography	Year	Population
BRONX	2015	1460412
BROOKLYN	2015	2643546
MANHATTAN	2015	1657183
QUEENS	2015	2346005
STATEN ISLAND	2015	475313
BRONX	2014	1445800

Joining 311 noise dataframe and Population dataframe

nyc311 is also joined to the noise dataframe for analysis of noise complaints with population.

Table with the population and nyc311 join data only for noise complaints

This table displays noise data information with the population. It diplays a few rows with the columns Complaint.Type, Borough, Hr, Yr and Population.

Complaint.Type	Borough	Hr	Yr	Population
Noise - Street/Sidewalk	MANHATTAN	2	2015	1657183
Noise - Street/Sidewalk	BROOKLYN	1	2015	2643546
Noise - Street/Sidewalk	BROOKLYN	1	2015	2643546
Noise - Commercial	MANHATTAN	1	2015	1657183
Noise - Commercial	BROOKLYN	1	2015	2643546
Noise - Commercial	BROOKLYN	1	2015	2643546

B. Joining 311 dataframe and the Weather dataframe

The head of the weather data table

Here we produce a header table with some rows of data.

Table 7: Table continues below

Created.Date	Complaint.Type	Borough	RTime	Hr
2015-04-14	Vending	BRONX	48.7	2
2015-04-14	Noise - Street/Sidewalk	MANHATTAN	47.48	2
2015-04-14	Noise - Street/Sidewalk	BROOKLYN	18.92	1
2015-04-14	Derelict Vehicle	QUEENS	137.7	1
2015-04-14	Noise - Street/Sidewalk	BROOKLYN	37.47	1
2015-04-14	Illegal Parking	QUEENS	38.4	1

Location.Type	AWND	PRCP	SNOW	SNWD	TAVG	TMAX	TMIN	WESD
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA
Street/Sidewalk	5.82	0.04	0	0	58	66	50	NA

C. Joining 311 dataframe and the Extreme Weather dataframe

The head of the extreme weather data table

Here we produce a header table with some rows of data.

Complaint.Type	Borough	Hr	Location.Type	WEATHER
Consumer Complaint	BROOKLYN	19		FOG
Consumer Complaint	BROOKLYN	18		FOG
Consumer Complaint	MANHATTAN	16		FOG
Consumer Complaint	QUEENS	16		FOG
Consumer Complaint	MANHATTAN	12		FOG
Consumer Complaint	BRONX	11		FOG

D. Joining 311 dataframe with Heat complaints and the Weather dataframe

Table with data on heat complaints and weather

Here we produce a header table with some rows of data.

Complaint.Type	Borough	Hr	Location.Type	TAVG	WEATHER
HEAT/HOT WATER	BRONX	0	RESIDENTIAL BUILDING	NA	NA
HEAT/HOT WATER	BRONX	0	RESIDENTIAL BUILDING	NA	NA
HEAT/HOT WATER	BRONX	0	RESIDENTIAL BUILDING	NA	NA
HEAT/HOT WATER	BRONX	0	RESIDENTIAL BUILDING	NA	NA
HEAT/HOT WATER	BROOKLYN	0	RESIDENTIAL BUILDING	NA	NA
HEAT/HOT WATER	BRONX	0	RESIDENTIAL BUILDING	NA	NA

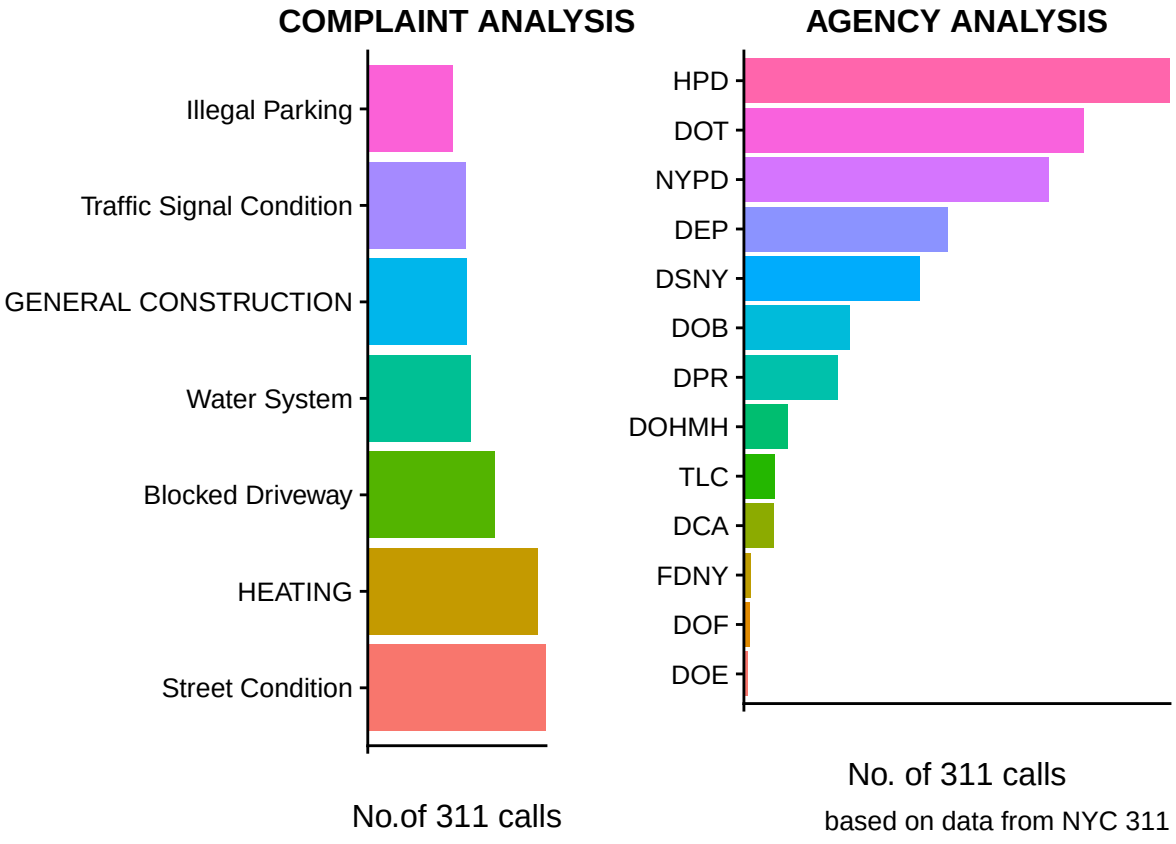
Exploration of the NYC311

The data available include the primary data from nyc311, population data and weather data. Every 311 call is associated with 3 things: WHAT:Complaint Type, WHEN: the date-time of call & when it gets resolved, WHERE: the location of call. Exploratory analytics was done on these 3 aspects to look for patterns and insights.

Complaint Types & Agencies

Complaints and agencies are analysed to find the most common complaints and the agencies that deal with most calls.

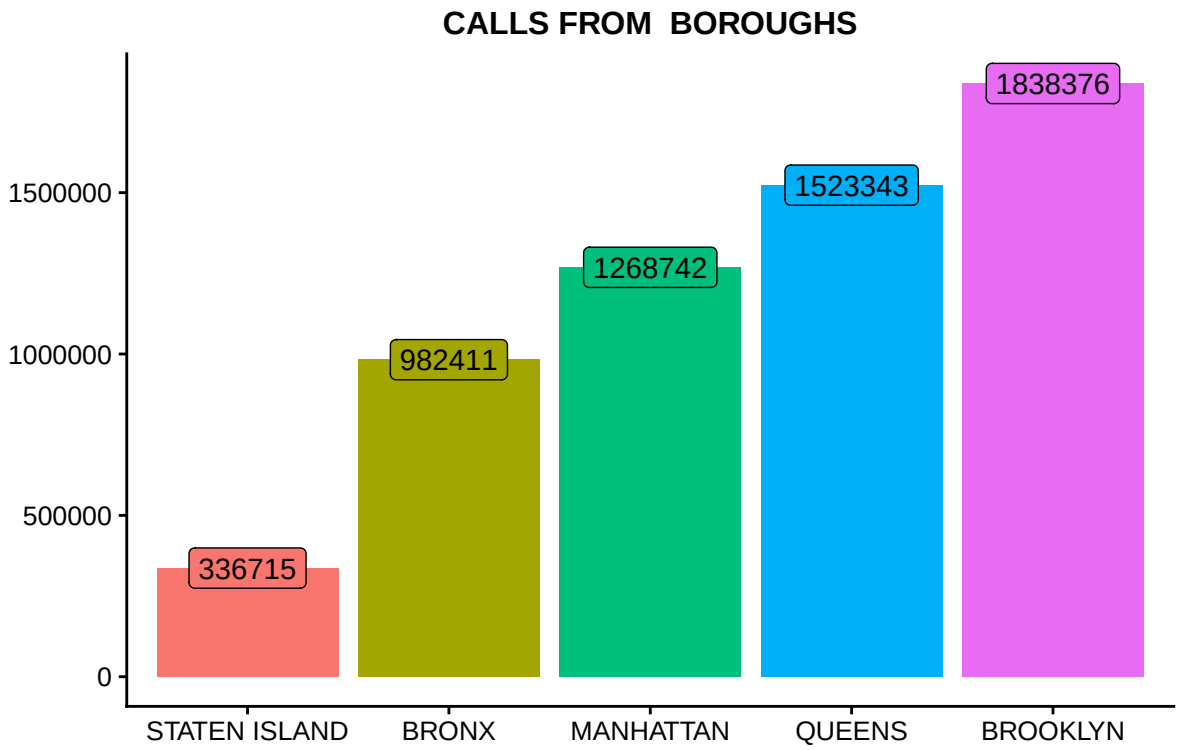
```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 70
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



Complaint Analysis: nyc311 data is sorted by Complaint.Type and top 7 Complaint Types are found as shown in the figure on the left. Agency Analysis: Rows are grouped by Agency and those with incidence > 10000 are displayed in the figure on the right.

Where are the 311 calls made from?

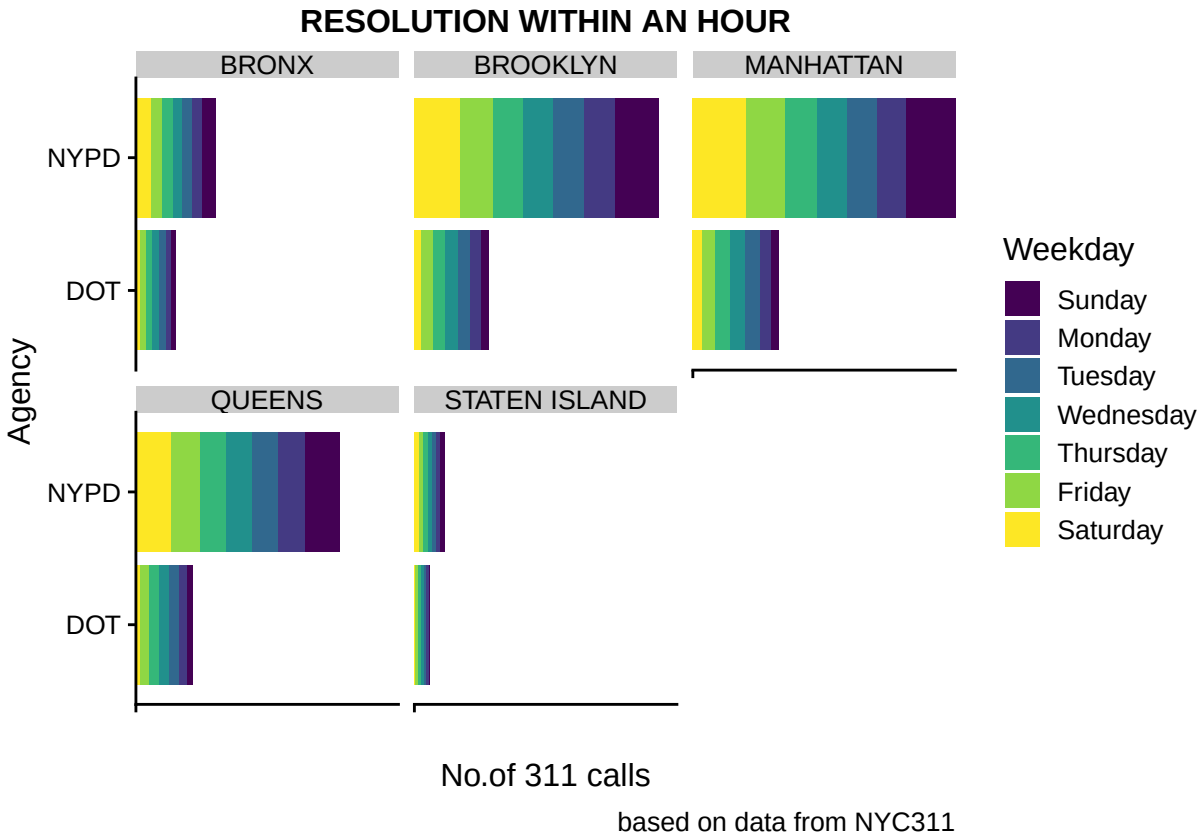
The plot illustrates the number of calls received by borough.



Brooklyn & Queens receive the most 311 calls whereas Staten Island receives the fewest number of calls.

Resolution within an hour in the 3 big agencies

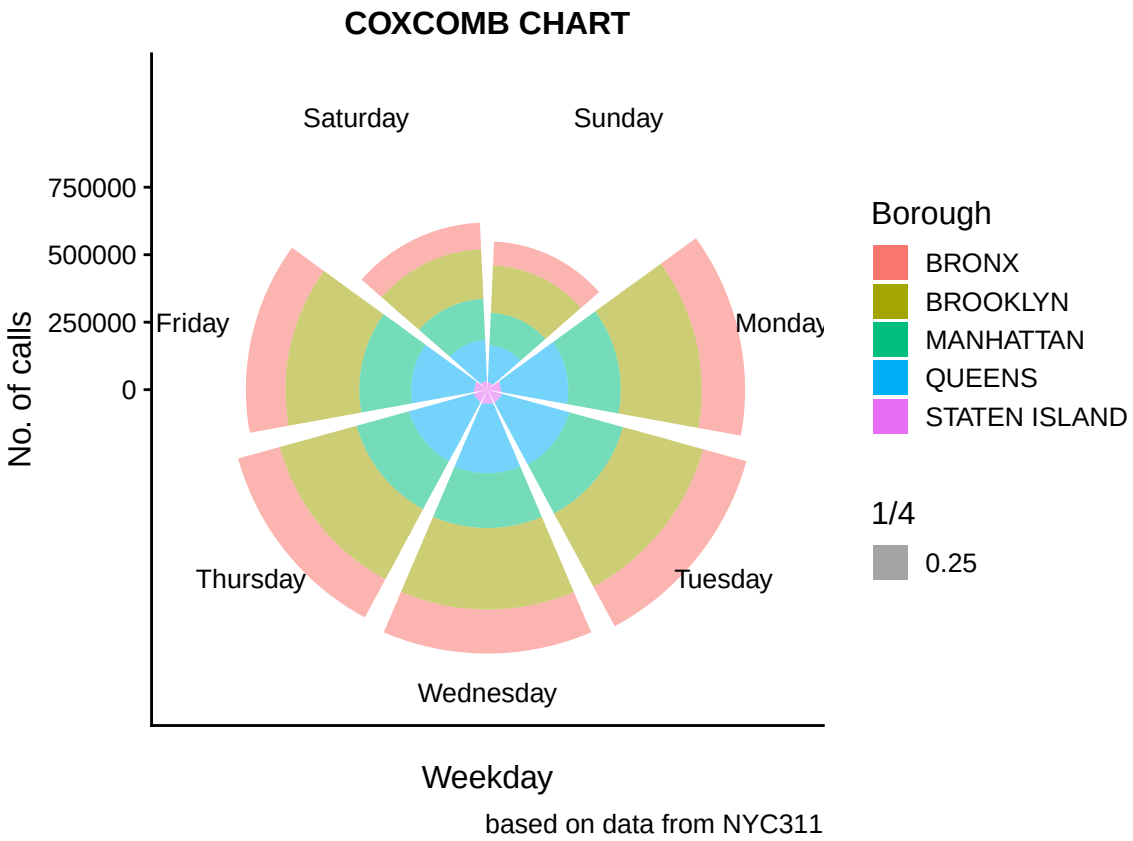
This plot explores the correlation between the number of cases in each Borough resolved within an hour by HPD, NYPD and DOT and day of the week. Insights: HPD takes more than an hour to resolve most cases.The resolution is slower on weekends at HPD and DOT. NYPD shows similar resolution numbers through out the week.



Comments: Confirms earlier findings from the plot for agencies.Most Heating Complaints arent solved within an hour. Resoliution time is longer on Weekends. Street Light and Street Condition related complaints are resolved quicker in Queens than in Brooklyn.

Which days of the week receives the most calls in each borough?

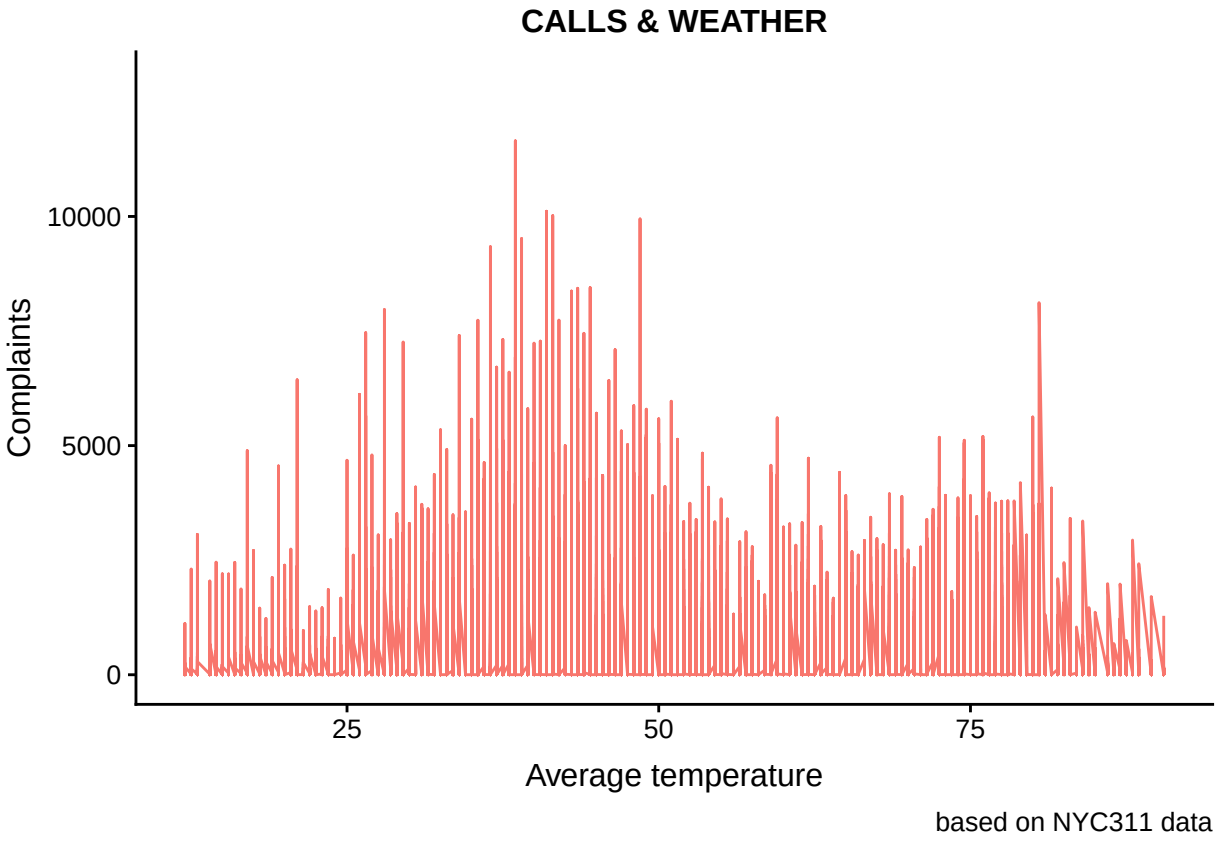
A coxcomb chart is used to dislay the relationship between number of calls, Weekday & Borough.



Comments: The chart shows the distribution of calls in Boroughs on weekdays is very similar. Tuesday, Wednesday and Thursdays are the day on which the call center is the most busiest. Friday and Mondays are less busier days. Calls are fewier on weekends, the least on Sundays.

Exploring correlation between weather & 311 calls

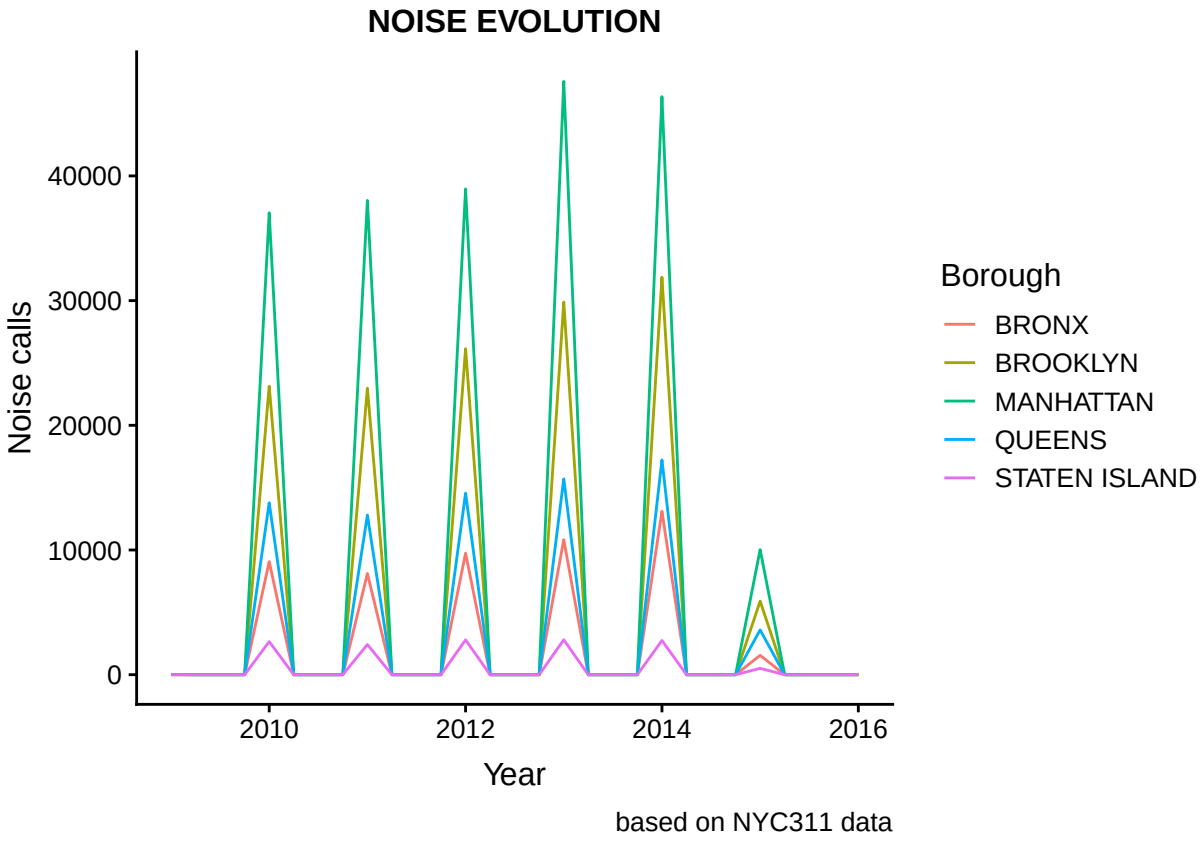
```
## Warning: Removed 43 rows containing missing values (geom_path).
```



The figure shows that more complaints are made on colder days.

Noise evolution over the years across boroughs

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).
## Warning: Removed 10 rows containing missing values (geom_path).
```



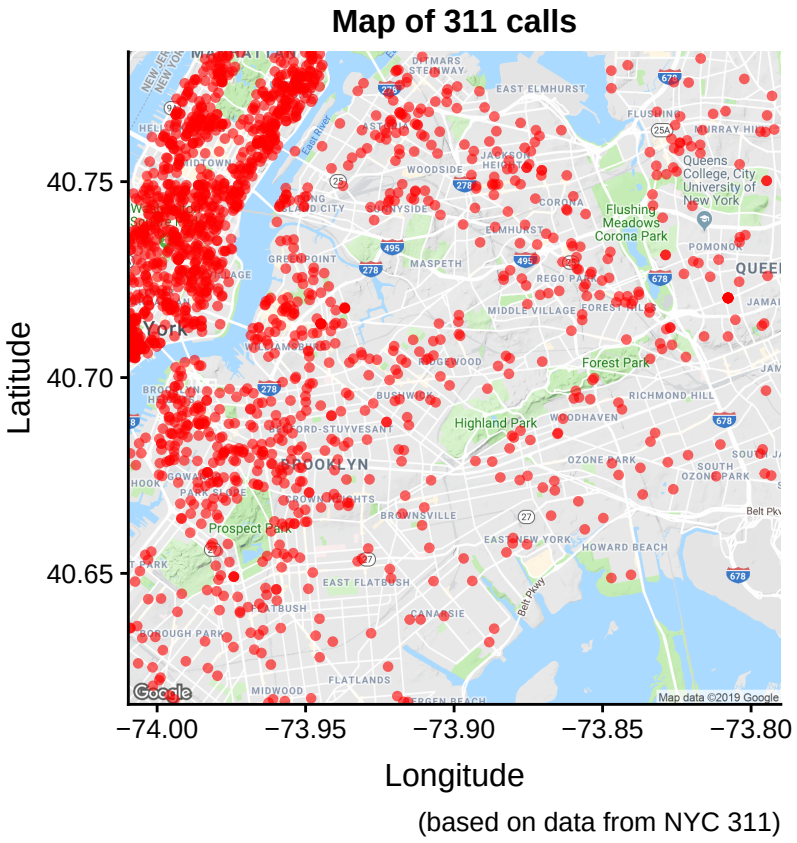
This frequency polygon shows the evolution of Noise complaint over the years across Boroughs. Manhattan shows the most Noise complaints. 2013 saw a spike in Noise complaints in Brooklyn and Manhattan. Highly effective corrective measures seem to have been employed in Manhattan & Brooklyn as the numbers dropped to less than a quarter in 2015.

Mapping noise complaint calls to their location

```
## Skipping install of 'ggmap' from a github remote, the SHA1 (2d756e5e) has not changed since last install.
##   Use `force = TRUE` to force installation
```

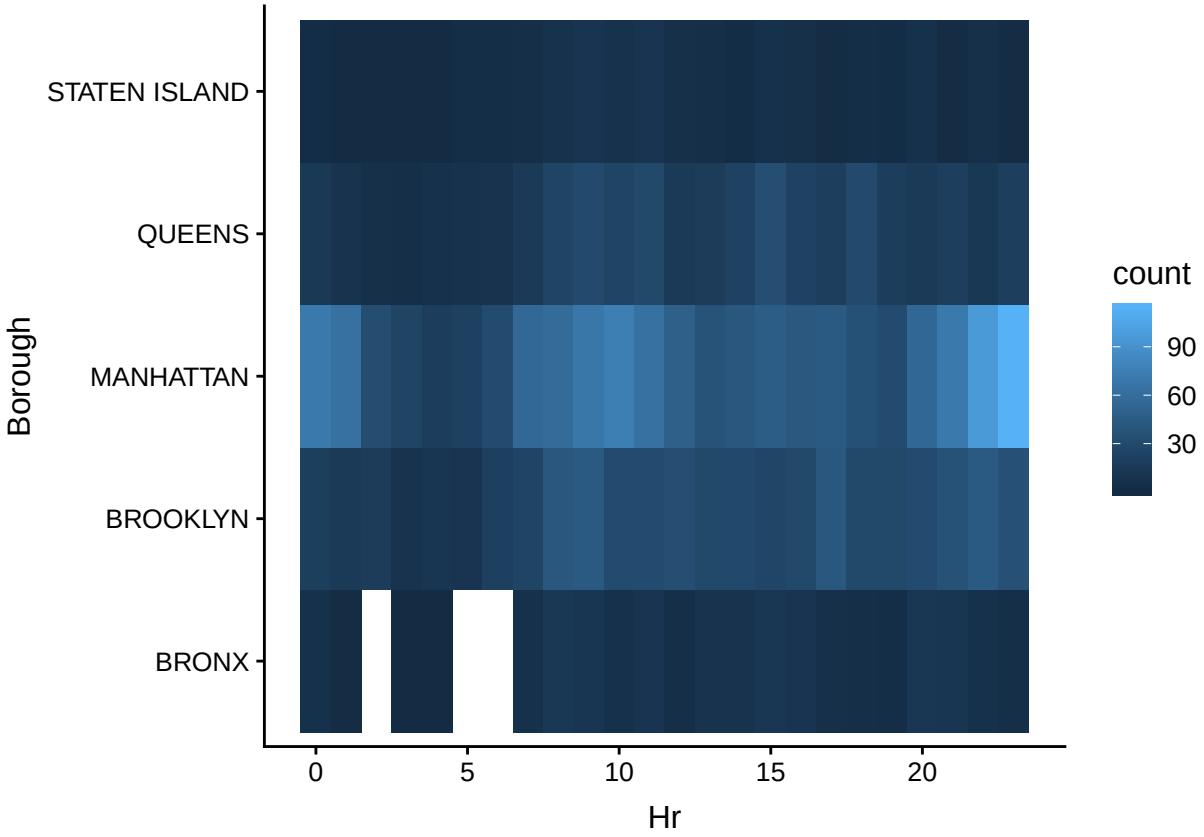
```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=40.7,-73.9&zoom=12&size=640x640&scale=2&maptyp
```

```
## Warning: Removed 866 rows containing missing values (geom_point).
```



The map shows the distribution of calls over New York for Noise complaints. Manhattan is evidently the borough with most Noise complaints from the sample used.

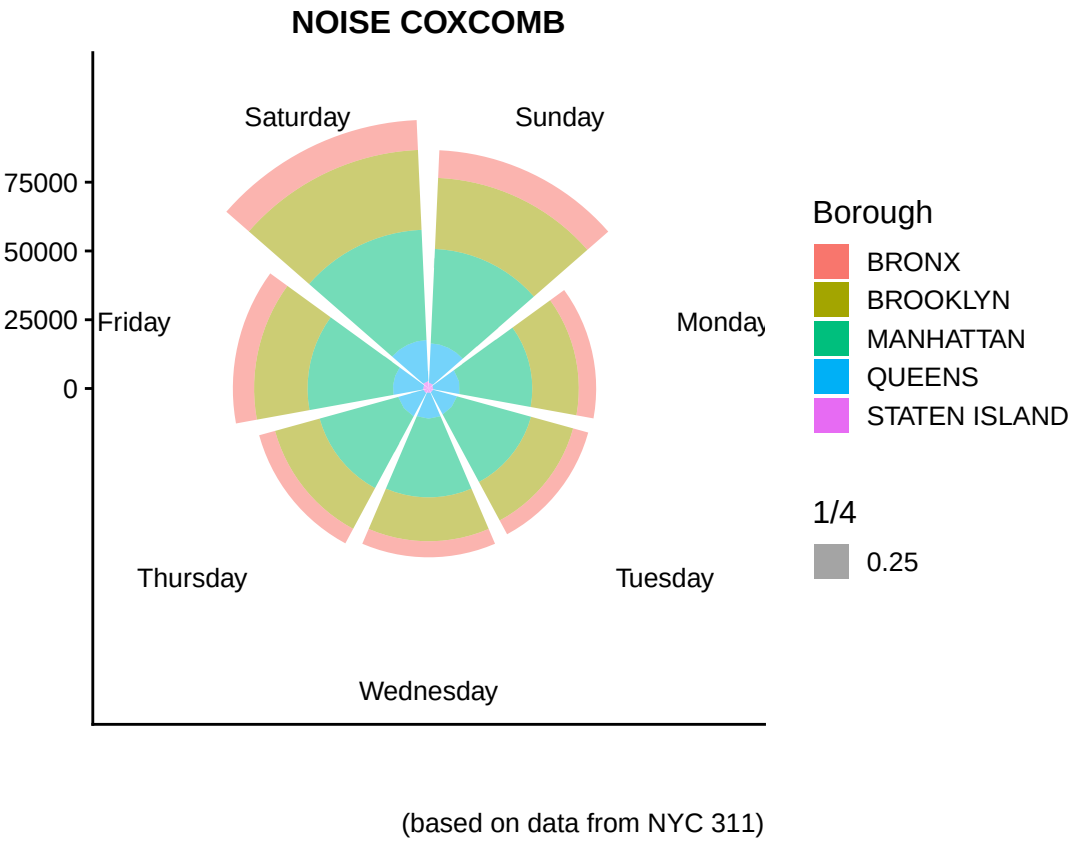
Exploring Noise Complaints by hour



This tile plot confirms that Manhattan has the most number of noise complaints among the boroughs. There are more noise

complaint calls made between 9 in the evening and 2 in the morning from Manhattan. This could indicate that there are noisier evening activities in Manhattan.

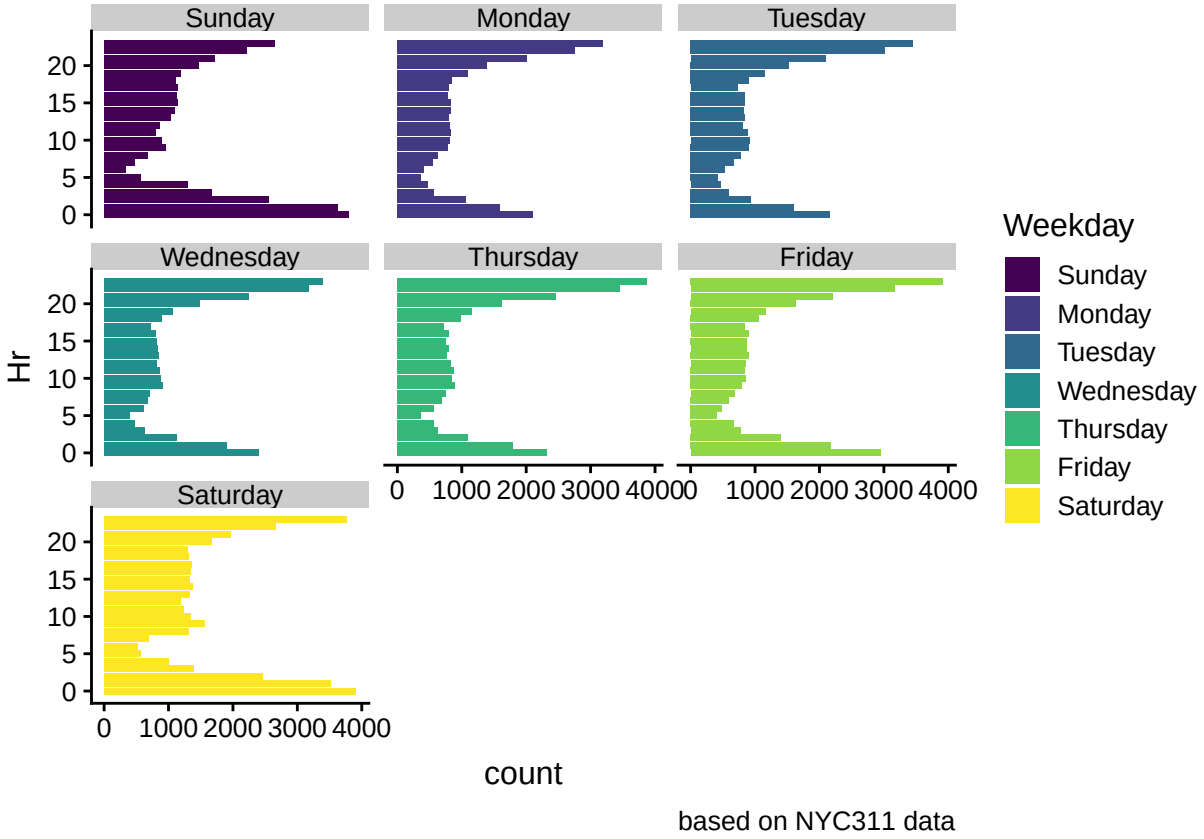
Weekday noise complaint distribution



(based on data from NYC 311)

The Combchart shows that most noise complaints are made on Saturdays. Surprisingly Fridays and Sundays have fewer noise complaint calls.

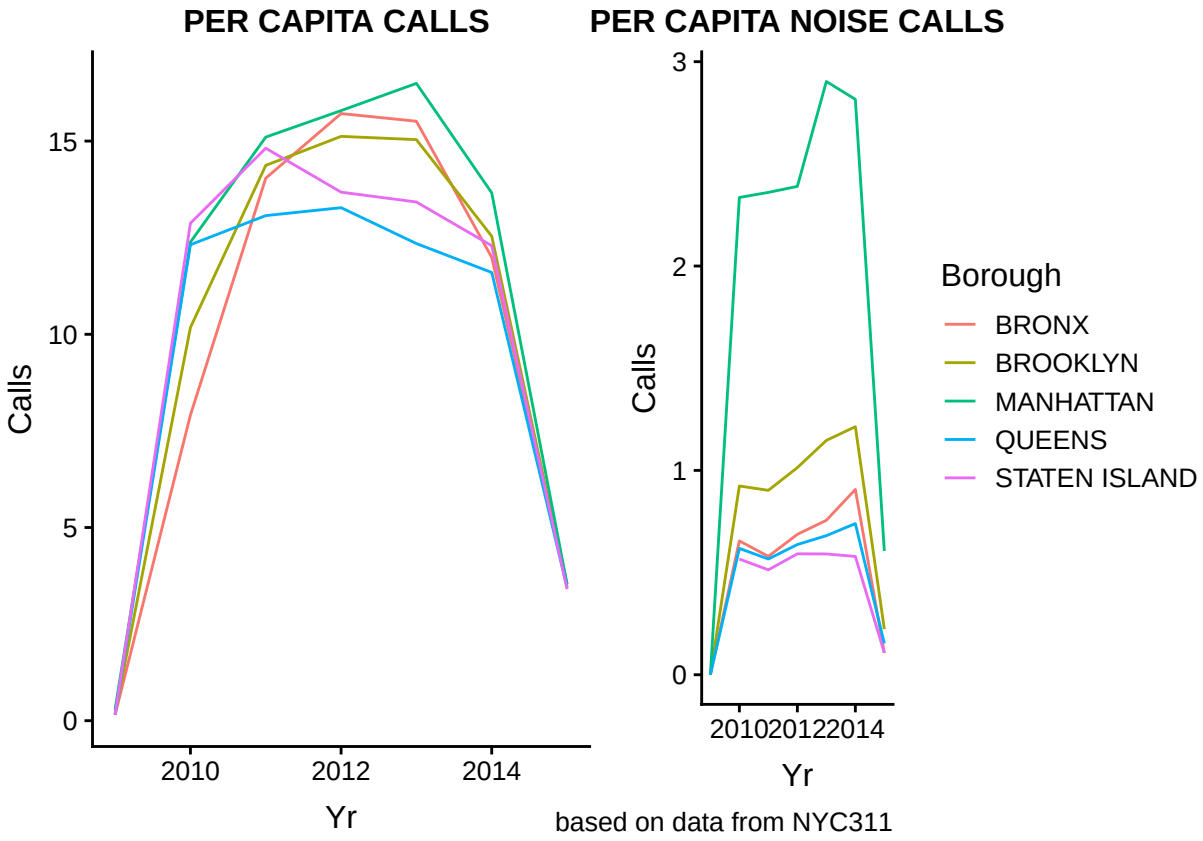
Noise complaints weekdays and Hr



Here, the noise complaints from Manhattan is analysed for the days and the time of complaints. As confirmed in earlier analysis, Saturday shows a peak in average number of noise complaints received per hour. However, the most calls are recieved in the day time with a peak around 9am. For the remaining days, most complaints follow the similar pattern for noise complaints. Most complaints are lodged in the between 9pm and 2am.

##Trend of complaints per resident and noise complaints per resident over the years across boroughs

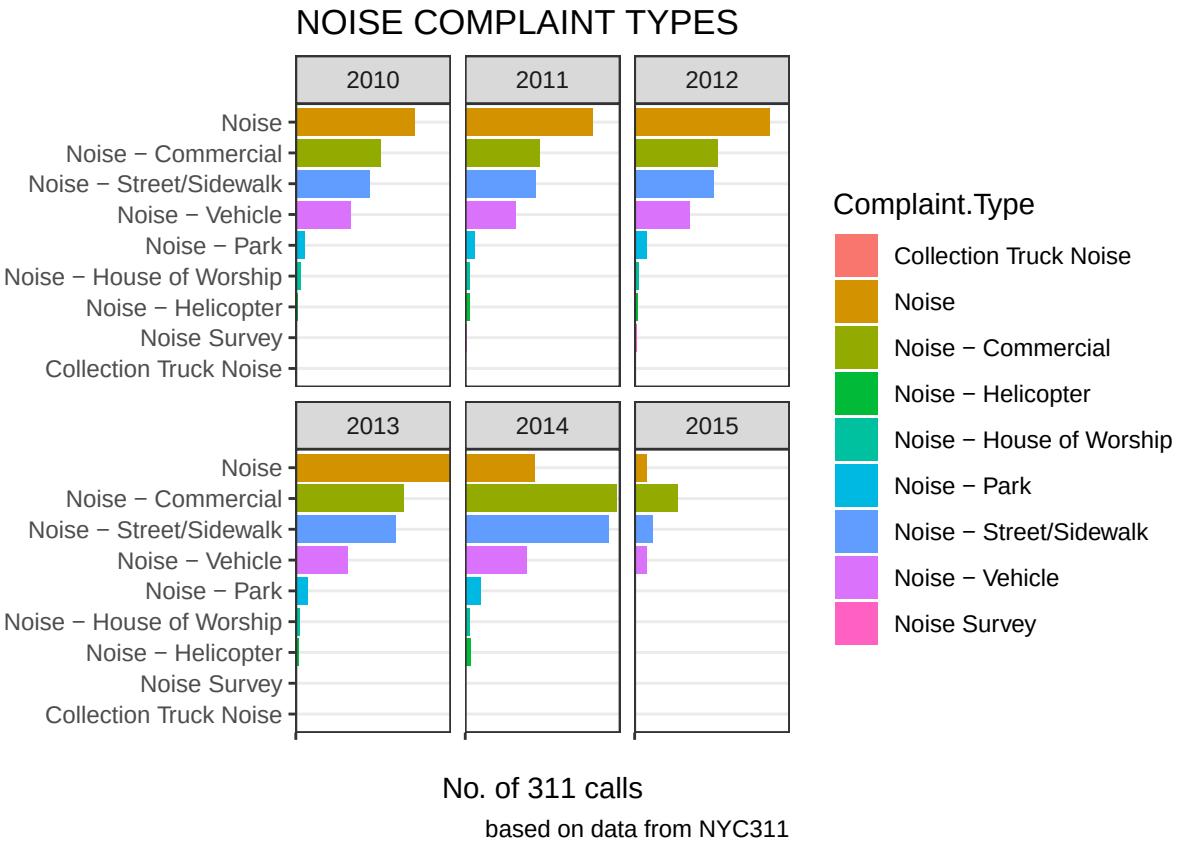
```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 70
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



This figure on the left shows that 311 call rate is highest for Manhattan followed by Bronx , Brooklyn, Staten Island and then by Queens. It is interesting to note that the number of calls per resident of Bronx was the least among New York Boroughs until the 2nd half of 2010, after which it Bronx residents have made the most 311 calls. This could be caused by the close to 2% increase in population from 2009 to 2011 in Bronx. Also, Queens has the least 311 calls per capita indicating either low awareness or better facilities.

The visualization above shows the evolution of noise complaints per person in each Borough over the years 2009 to 2015. The rate of calls increased in 2014 across New York in 2014 and dipped in 2015.

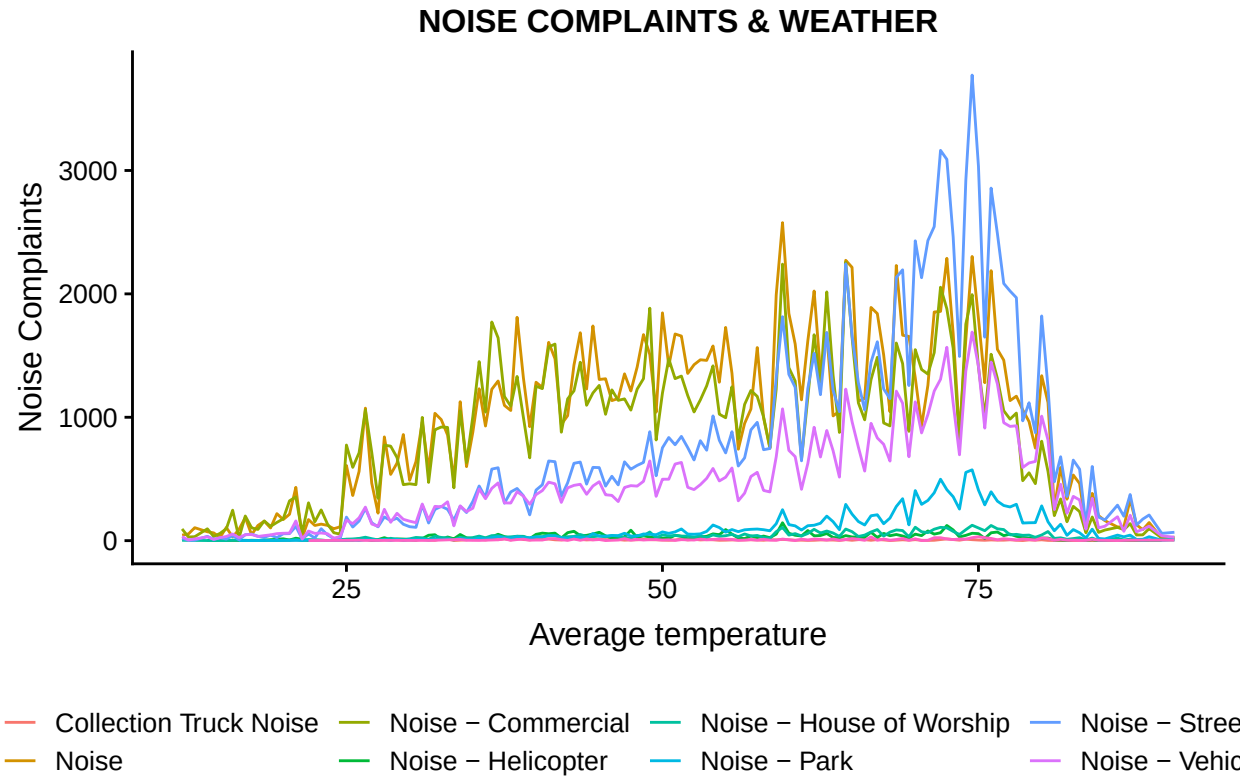
Exploring types of noise complaints



This figure shows that the Commercial noise complaints have increased over the years until it dropped in 2014-2015. Noise compaints from Vehicles had been constant until it dropped in 2015. The measures taken to drop calls in 2015 can be continued to furthur reduce noise complaints.

Exploring correlation between weather & 311 noise calls

```
## Warning: Removed 2 rows containing missing values (geom_path).
```



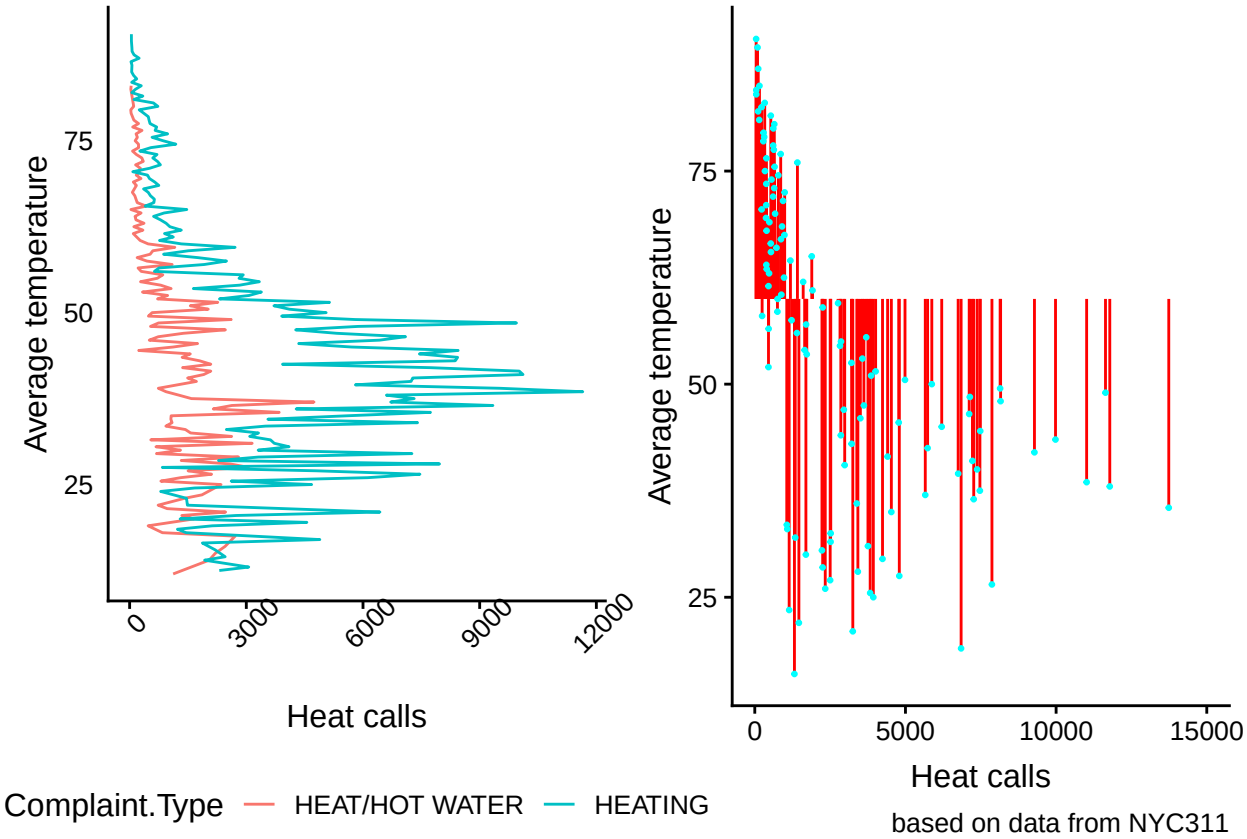
This plot indicates that noise complaints have quite the opposite trend. There are more noise complaints on warmer days.This could be because people engage in more "noisy" activities such as park noise, party noise etc on warmer days.

Exploring correlation between heat complaints & average temperature related to heating

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The figure on the left shows that the number of heat related complaints, mainly related to HEATING are higher during lower average tempertures. This is confirmed in the figure on the right which indicates that heat related complaints can be expected when the mercury drops below 60 degrees. More agents could be deployed at HPD when forecasts of colder temperatures are announced.

Conclusion

From the analysis done, the following insights can be used for predicting the possibility of a complaint and taking proactive actions:

Heat related complaints- Heat related complaints are the most common complaints of New Yorkers across boroughs. The resolution time taken by HPD is much longer than other agencies(Not one complaint is solved in under an hour). The number of Heat complaints escalate as the mercury drops. =>HPD could anticipate more calls when colder days are announced and deploy more personnels.

Noise Complaints- Noise complaints are one of the other complaint that has been analysed. Manhattan has the most noise related complaints. The call density is highest between 9pm-2am on warmer days and more on the weekends than on weekdays. =>Evening activities such as parties in Manhattan could be responsible for this. Improving patrolling in the night to curb noisy activities could reduce these calls.There has been a very significant drop in noise complaints in 2015. The measures taken can be reproduced to furthur bring down noise complaints.

Queens- It has the 2nd largest proportion of calls and has the 2nd largest population. But it has the least per capita calls despite its large population. Queens has a good responsive team who solve Street Light and Street Condition related complaints within an hour. =>The good records in Queens could be either due to lack of awareness or better response time from agencies. This need to be investigated and best practices can be shared to improve the call rates in other Boroughs. Brooklyn in particular could benefit in high volumes as it receives the maximum number of 311 complaint calls with Complaints related to Heating and Street Condition being the most common.

Appendix

1. Data Dictionary of nyc311

1. Unique Key: Unique Identifier of service request (SR) call in the open data set

2. Created Date: Date the call was made and SR was created
3. Closed Date: Date SR was closed by responding agency
4. Agency:Acronym of responding city government agency
5. Agency Name: Full agency name of responding city government agency
6. Complaint Type: This is the first level of a hierarchy identifying the topic of the incident or condition.
7. Descriptor: This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.
8. Location Type: Describes the type of location used in the address information
9. Incident Type: 10479 zip code depending on the area
10. Incident Address: House number of incident address provided by submitter.
11. Street Name: Street name of incident address provided by the submitter
12. Cross Street 1: First Cross street based on the geo validated incident location
13. Cross Street 2: Second Cross Street based on the geo validated incident location
14. Intersection Street 1: First intersecting street based on geo validated incident location
15. Intersection Street 2: Second intersecting street based on geo validated incident location
16. Address Type: Type of incident location information available.
17. City: City where the Incident occurred
18. Landmark: If the incident location is identified as a Landmark the name of the landmark will display here
19. Facility Type: If available, this field describes the type of city facility associated to the SR
20. Status: Status of SR
21. Due Date: Date when responding agency is expected to update the SR
22. Resolution.Action.Updated.Date: Date when responding agency last updated the SR.
23. Community Board: This indicated the community district where the incident occurred
24. Borough: Bronx, Brooklyn, Manhattan, Queens, Staten Island Borough Provided by the submitter and confirmed by geo validation.
25. X coordinate: Geo validated, X coordinate of the incident location.
26. Y coordinate: Geo validated, Y coordinate of the incident location. Open Data channel type: Indicates how the SR was submitted to 311. i.e. By Phone, Online, Mobile, Other or Unknown.
27. Park Facility Name: If the incident location is a Parks Dept facility, the Name of the facility will appear here
28. Park Borough: The borough of incident if it is a Parks Dept facility
29. School Number: Schools registered number
30. School Name: Name of school
31. School Region: Region of the school
32. School Code: School'scode
33. School Phone: School phone number
34. School Address: School complete address
35. School City :City of the schools location
36. School State: Which state the school is based in
37. School Zip :School zip code
38. School not found : Has the value Y when School isnt found . Else has the value N or " ".
39. School or city wide complaint : Complaint type
40. Vehicle Type: If the incident is a taxi, this field describes the type of TLC vehicle.
41. Taxi Company Borough: If the incident is identified as a taxi, this field will display the borough of the taxi company.
42. Taxi Pick Up Location :If the incident is identified as a taxi, this field displays the taxi pick up location
43. Bridge Highway Name :If the incident is identified as a Bridge/Highway, the name will be displayed here.
44. Bridge Highway Direction : Direction of the highway
45. Road Ramp : If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.
46. Bridge Highway Segment : Additional information on the section of the Bridge/Highway were the incident took place.
47. Garage.Lot.Name: Related to DOT Parking Meter SR, this field shows what garage lot the meter is located in

48. Ferry.Direction: Used when the incident location is within a Ferry, this field indicates the direction of ferry

49. Ferry.Terminal.Name: Used for ferry incidents. This field indicates the ferry terminal where the incident took place.
50. Latitude: Geo based Lat of the incident location
51. Longitude: Geo based Long of the incident location
52. Location : Combination of the geo based lat & long of the incident location

2. Columns dropped from NYC311

The following columns identified to be removed are:

1. Col#10 Street.Name - is redundant with the column Incident Address
2. Col#18 Facility.Type - same & more specific information is given in column Location.Type
3. Col#21 Resolution.Action.Update.Date - all date info is got from Created Date, Closed Date and Due Date
4. Col#24 & 25 X coordinate and Y coordinate - Longitude and Latitude will be used for mapping
5. Col#27 &Col#22 Park.Borough and Community Board- is redundant with the column Borough
6. Col#28 & 48 School.Name and Ferry.Terminal.Name - are redundant with the column Park.Facility.Name
7. Col#29 School.Number - is a unique number for each school, but so is School.Code
8. Col#30 School.Region - no value add
9. Col#32 School.Phone.Number - no value add
10. Col#33 School Address - is redudant with Incident Address
11. Col#34&16 School.City & City - has many mispellings and odd entries
12. Col#35 School.State - has a single value fr all rows “NY”
13. Col#36 School.Zip - is the same as the column Incident.Zip
14. Col#37 School.Not.Found - no value add. Has “Y”" for senior citizen complaints (deducable from Complaint.Type) and N or "" for others
15. Col#38 School.or.Citywide.Complaint - is “School” or "". Location.Type will indicate School for all rows concerning schools.
16. Col#51 Location - is redundant with columns Longitude and Latitude
17. Col#44 Road.Ramp - is either Ramp or Roadway and doesnt add value to data
18. Col#17 Landmark - the whole dataset has only 15 rows where this column in filled. So, it will not add much value to the analysis
19. Col#47 Ferry.Direction -only 7 rows have values, hence wont add value to the analysis
20. col#39, 40,41 Vehicle.Type, Taxi.Company.Borough, Taxi.Pick.Up.Location -10, 13 and 117 rows only. Hence can be removed

3. Data Dictionary of nycpop

1.FIPS Code - Text data- Federal Information Processing Standards (FIPS) codes which identifies the differnt counties of New York. 2.Geography - Text data- Gives the name of the county 3.Year - Numeric- Gives the year 4.Program Type -Text data - Could be “Postcensal Population estimate”/“Census Base population”/“Intercensal Population” 5.Population - Numeric - Gives the number of residents

4. Data Dictionary of nycweather

1. STATION - (17 characters) is the station identification code.
2. NAME - (max 50 characters) is the name of the station (usually city/airport name)
3. DATE - Date of weather measurement
4. AWND - Average Wind Speed
5. PRCP - Precipitation
6. SNOW - Snowfall
7. SNWD - Snow depth
8. TAVG - Average temperature
9. TMAX - Maximum temperature
10. TMIN - Minimum temperature
11. WESD - Water equivalent of snow on the ground
12. WT01- Fog, ice fog, or freezing fog (may include heavy fog)
13. WT02- Heavy fog or heaving freezing fog (not always distinguished from fog)
14. WT03 - Thunder
15. WT04 - Ice pellets, sleet, snow pellets, or small hail"
16. WT05 - Hail (may include small hail)
17. WT06 - Glaze or rime
18. WT07 - Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction
19. WT08 - Smoke or haze
20. WT09 - Blowing or drifting snow
21. WT11 - High or damaging winds
22. WT13 - Mist
23. WT14 - Drizzle

24. WT15 - Freezing drizzle
25. WT16 - Rain (may include freezing rain, drizzle, and freezing drizzle)"
26. WT17 - Freezing rain
27. WT18 - Snow, snow pellets, snow grains, or ice crystals
28. WT19 - Unknown source of precipitation
29. WT21 - Ground fog
30. WT22 - Ice fog or freezing fog