homework ii

Khavya Seshadri

2019-09-12

Introduction

311 is a telephone number similar to 911, where people call to access non-emergency government services. The dataset consists of about 9 million records which indicates the service call requests reported in the New York city from the year 2010 to the present year.

Initialization

Here we load the tidyverse packages and the `data.table` package and load the nyc311 data set. Then we fix the column names of the nyc311 data so that they have no spaces.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------------------- t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.2
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv")
names(nyc311)<-names(nyc311) %>%
  stringr::str_replace_all("\\s", ".")
```

Data pre-processing

Here we perform data pre-processing steps, by dropping irrelevant columns and removing duplicate rows from the dataset.

```
nyc311 <- nyc311[,c(-1,-10:-19,-23, -25:-49)]
names(nyc311)
```

```
##  [1] "Created.Date"                   "Closed.Date"
##  [3] "Agency"                         "Agency.Name"
##  [5] "Complaint.Type"                 "Descriptor"
##  [7] "Location.Type"                  "Incident.Zip"
##  [9] "Status"                         "Due.Date"
## [11] "Resolution.Action.Updated.Date" "Borough"
## [13] "Latitude"                       "Longitude"
## [15] "Location"
```

```
nyc311nodups <- distinct(nyc311)
dim(nyc311nodups)
```

```
## [1] 8250329      15
```

Description

Here we describe the data, showing both a sample and a data dictionary.

The head of the table

Here we produce a table of just some relevant columns of data.

```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.5.2
```

```
options(xtable.comment=FALSE)
options(xtable.booktabs=TRUE)
narrow<-nyc311nodups %>%
  select(Agency,
     Complaint.Type,
     Descriptor,
     Incident.Zip,
     Status,
     Borough)
xtable(head(narrow))
```

	Agency	Complaint.Type	Descriptor	Incident.Zip	Status	Borough
1	NYPD	Vending	In Prohibited Area	10465	Closed	BRONX
2	NYPD	Blocked Driveway	No Access	11234	Open	BROOKLYN
3	NYPD	Noise - Street/Sidewalk	Loud Music/Party	11204	Open	BROOKLYN
4	NYPD	Noise - Street/Sidewalk	Loud Talking	11211	Assigned	BROOKLYN
5	NYPD	Noise - Street/Sidewalk	Loud Talking	10025	Closed	MANHATTAN
6	NYPD	Noise - Street/Sidewalk	Loud Talking	11205	Closed	BROOKLYN

Data Dictionary

• Created.Date – The date when the service request was created. (Type: timestamp (mm/dd/yyyy hh:mm:ss)) • Closed.Date – The date when the service request was closed by the responding agency. (Type: timestamp) • Agency – The responding City Government agency (For example: NYPD, DPR,etc.) • Agency.Name – The full agency name of responding city government agency. (Type: text) • Complaint.Type – The type of complaint reported (For example: vending, illegal parking, blocked driveway). • Descriptor - Detailed description of the corresponding complaint type. (Type: text) • Location.Type – The type of location based on the address. (For example: Street/Sidewalk, Park, etc.) • Incident.Zip – Zip code of the incident location. (For example: 14623 or (5-digit integer / 9 digits with dash between fifth and sixth digit)) • Status – The status of the service request submitted. (Allowed values: Open, Started, Assigned, Unassigned, Email sent, Pending, Email sent, Closed.) • Due.Date – The date, during when the responding agency is expected to update the service request. (Type: Date/timestamp) • Resolution.Action.Updated.Date – Date when the responding agency last updated the service request. • Borough – town/ district of the NYC provided by submitter. (Values: BRONX, BROOKLYN, MANHATTAN, QUEENS, STATEN ISLAND) • Latitude – Geo-based latitude of the incident location. (Type: degrees) • Longitude – Geo-based longitude of the incident location. (Type: degrees) • Location – Combination of the geo-based latitude and longitude of the incident location. (Type: location)

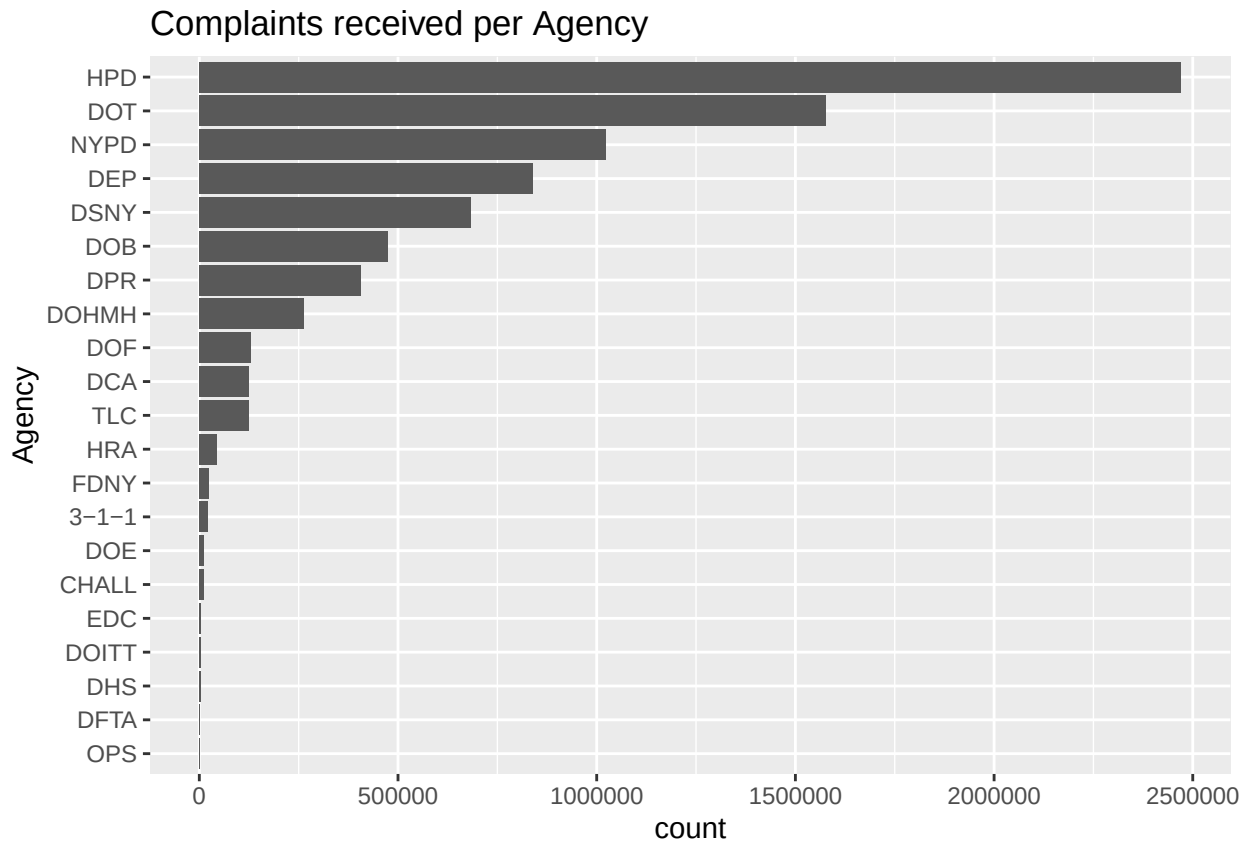
Exploration

Here we explore the columns in the data set.

The following plot shows a horizontal bar chart showing the top agencies that received service call requests along with the count of service call requests for each agency.

```
bigAgency <- narrow %>%
  group_by(Agency) %>%
  summarize(count=n()) %>%
  filter(count>1000)
```

```
bigAgency$Agency<-factor(bigAgency$Agency,
  levels=bigAgency$Agency[order(bigAgency$count)])
p<-ggplot(bigAgency,aes(x=Agency,y=count)) +
   geom_bar(stat="identity") +
   coord_flip() +
   ggtitle("Complaints received per Agency")
p
```



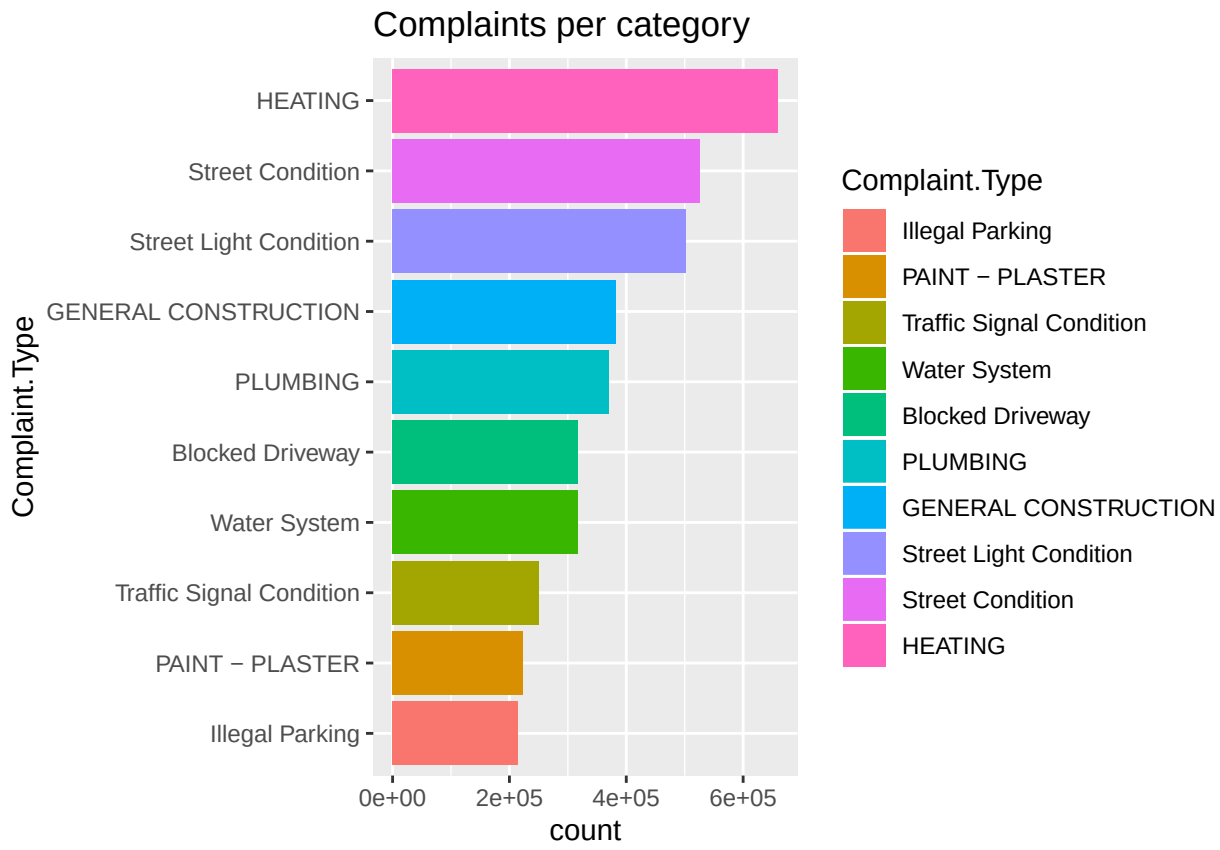
The following horizontal bar chart shows the top 10 complaint types received, with the color specified for each complaint type.

```
topComplaints <- narrow %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>%
  filter(count>100000) %>%
  top_n(10)
```

```
## Selecting by count
```

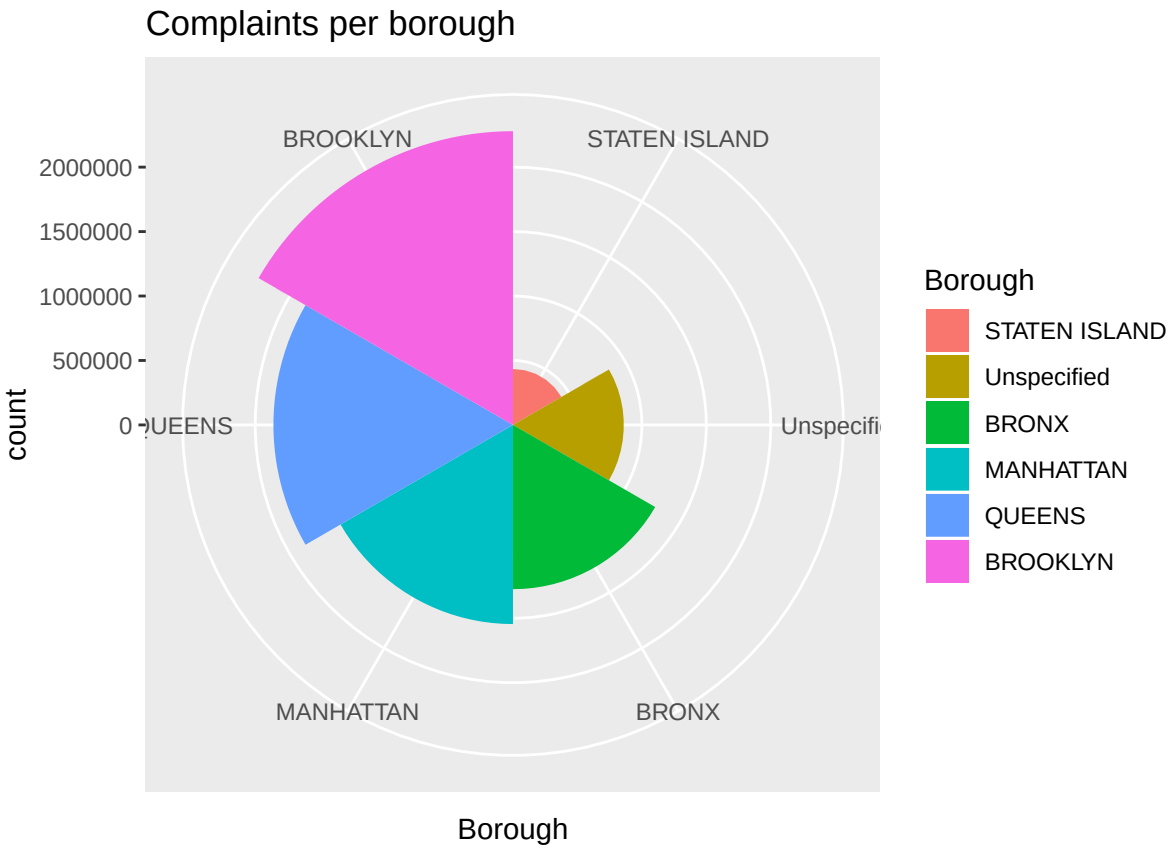
```
topComplaints$Complaint.Type<-factor(topComplaints$Complaint.Type,
  levels=topComplaints$Complaint.Type[order(topComplaints$count)])
plotA<-ggplot(topComplaints,aes(x=Complaint.Type,y=count, fill=Complaint.Type)) +
   geom_bar(stat="identity") +
   coord_flip() +
```

```
   ggtitle("Complaints per category")
plotA
```



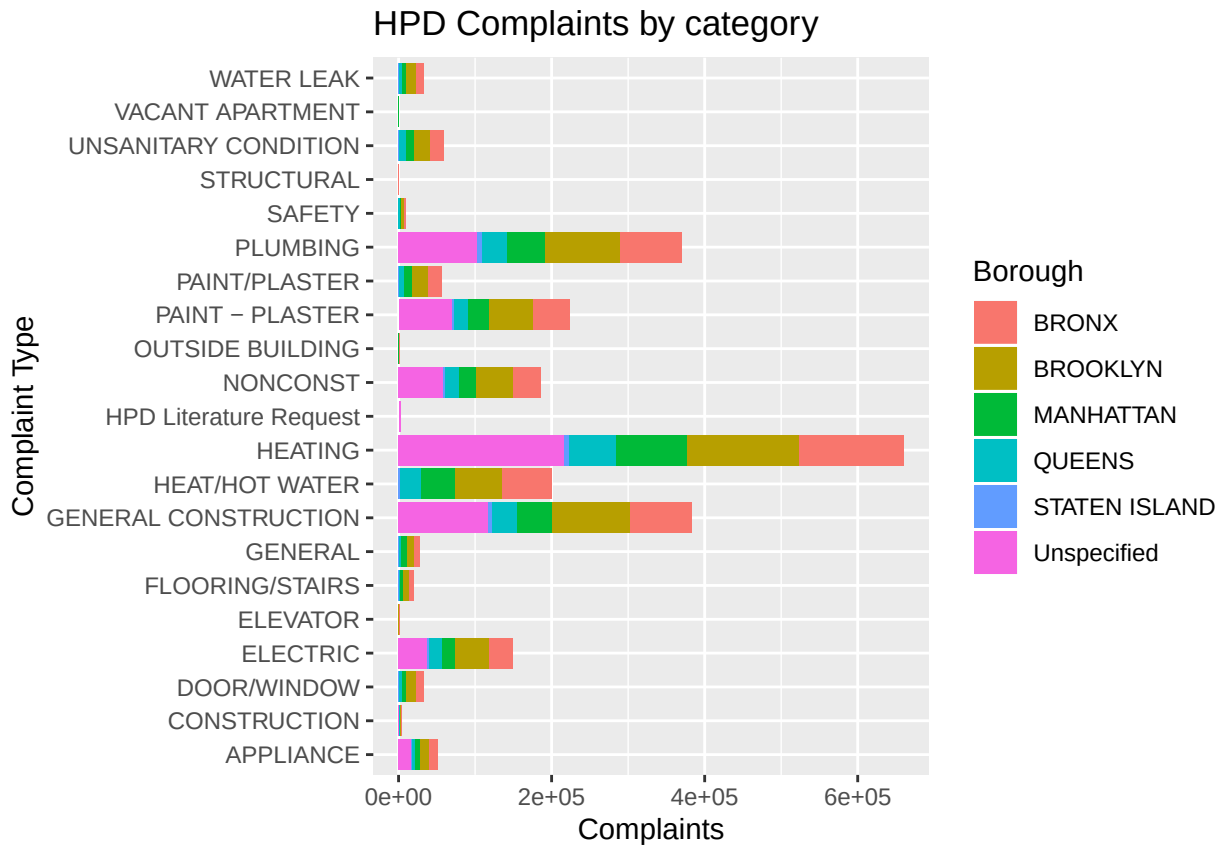
The following shows a coxcomb plot showing the boroughs that received the most service call requests depicted in the form of coxComb.

```
boroughs <- narrow %>%
  group_by(Borough) %>%
  summarize(count=n())
boroughs$Borough<-factor(boroughs$Borough,
  levels=boroughs$Borough[order(boroughs$count)])
plotB<-ggplot(boroughs,aes(x=Borough,y=count, fill=Borough)) +
   geom_bar(stat="identity", width=1) +
   theme(aspect.ratio = 1) +
   coord_polar() +
   ggtitle("Complaints per borough")
plotB
```



Considering the HPD agency alone, below is a plot depicting the HPD complaints by type across each borough.

```
hpdComplaints <-dplyr::filter(narrow, Agency=='HPD')
hpdComp <- hpdComplaints %>%
  group_by(Complaint.Type,Borough) %>%
  summarize(Complaints = length(Complaint.Type))
ggplot(hpdComp, aes(x=Complaint.Type,y=Complaints, fill=Borough)) +
  xlab("Complaint Type") +
  geom_bar(stat ="identity") +
  coord_flip() +
  ggtitle("HPD Complaints by category")
```

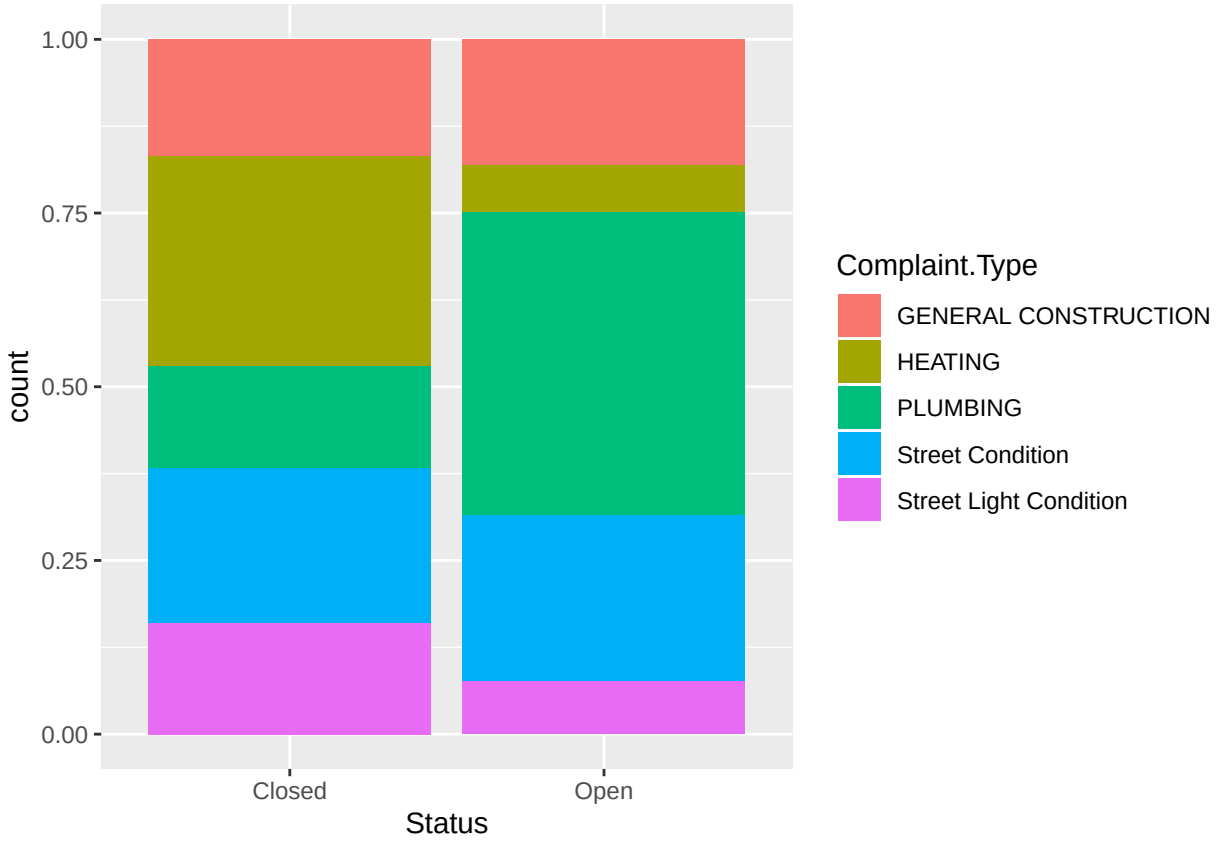


The table below shows information about the number of open and closed service call requests.

```
statusFrequency <- narrow %>%
  group_by(Status) %>%
  summarize(count=n()) %>%
  filter(Status=="Open" | Status=="Closed")
statusFrequency$Status<-factor(statusFrequency$Status,
  levels=statusFrequency$Status[order(statusFrequency$count)])
statusFrequency
```

```
## # A tibble: 2 x 2
##   Status   count
##   <fct>    <int>
## 1 Closed 7043557
## 2 Open    772819
```

```
filteredData <- dplyr::filter(narrow, (Complaint.Type=='HEATING' | Complaint.Type=='GENERAL CONSTRUCTIO
complaintStatus <- filteredData %>%
  group_by(Status,Complaint.Type) %>%
  summarize(count=n())
plotC<-ggplot(complaintStatus,aes(x=Status,y=count, fill=Complaint.Type)) +
   geom_bar(stat="identity", position = "fill")
plotC
```



The above plot shows the percentage of each of the top 5 complaint types(using color) in the open and closed status of the report. This shows that majority of the open service requests are of complaint type "Plumbing" and the Heating complaint requests have a good record in the closed status.

Next we include a crosstabulation.

```
xtabA<-dplyr::filter(narrow,
  Complaint.Type=='HEATING' |
  Complaint.Type=='GENERAL CONSTRUCTION' |
  Complaint.Type=='PLUMBING'
)
xtabB<-select(xtabA,Borough,"Complaint.Type")
library(gmodels)
CrossTable(xtabB$Borough,xtabB$'Complaint.Type')
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1413129
##
##
##                | xtabB$Complaint.Type
## xtabB$Borough | GENERAL CONSTRUCTION |             HEATING |            PLUMBING |           Row Total |
```

```
## --------------|----------------------|----------------------|----------------------|----------------------|
##         BRONX |                80645 |               137147 |                79858 |               297650 |
##               |                0.001 |               25.733 |               45.489 |                      |
##               |                0.271 |                0.461 |                0.268 |                0.211 |
##               |                0.211 |                0.208 |                0.216 |                      |
##               |                0.057 |                0.097 |                0.057 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##      BROOKLYN |               101025 |               145473 |                98696 |               345194 |
##               |              602.771 |             1543.150 |              755.645 |                      |
##               |                0.293 |                0.421 |                0.286 |                0.244 |
##               |                0.264 |                0.220 |                0.267 |                      |
##               |                0.071 |                0.103 |                0.070 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##     MANHATTAN |                46839 |               93851 |                49068 |               189758 |
##               |              406.019 |              306.356 |                8.303 |                      |
##               |                0.247 |                0.495 |                0.259 |                0.134 |
##               |                0.122 |                0.142 |                0.133 |                      |
##               |                0.033 |                0.066 |                0.035 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##        QUEENS |                31671 |               61224 |                33427 |               126322 |
##               |              190.172 |               83.248 |                3.386 |                      |
##               |                0.251 |                0.485 |                0.265 |                0.089 |
##               |                0.083 |                0.093 |                0.090 |                      |
##               |                0.022 |                0.043 |                0.024 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
## STATEN ISLAND |                 6275 |                5299 |                5801 |                17375 |
##               |              522.265 |              977.887 |              342.902 |                      |
##               |                0.361 |                0.305 |                0.334 |                0.012 |
##               |                0.016 |                0.008 |                0.016 |                      |
##               |                0.004 |                0.004 |                0.004 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##   Unspecified |               116378 |               217108 |               103344 |               436830 |
##               |               32.604 |              835.319 |             1074.992 |                      |
##               |                0.266 |                0.497 |                0.237 |                0.309 |
##               |                0.304 |                0.329 |                0.279 |                      |
##               |                0.082 |                0.154 |                0.073 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##  Column Total |               382833 |               660102 |               370194 |              1413129 |
##               |                0.271 |                0.467 |                0.262 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##
##
```

The above crosstab shows tabulation of every borough with respect to the complaint types: heating, general construction and plumbing, that is it shows the number of complaints received in every borrough for the three specific complaint types and along with chi-square contribution, the percentage of complaints in every borough(N/row total), percentage of each complaint type(N/column total) and percentage of complaints for a specific complaint type and at a specific borough.(N/table total).

```
xtabA1<-dplyr::filter(narrow, ( Agency=='HPD' | Agency=='DOT'))
xtabB1<-select(xtabA1,Borough, Agency)
library(gmodels)
CrossTable(xtabB1$Borough,xtabB1$Agency)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
```

```
##
## Total Observations in Table:  4048906
##
##
##                | xtabB1$Agency
## xtabB1$Borough |       DOT |       HPD | Row Total |
## ---------------|-----------|-----------|-----------|
##          BRONX |    216610 |    561050 |    777660 |
##                | 24651.418 | 15739.804 |           |
##                |     0.279 |     0.721 |     0.192 |
##                |     0.137 |     0.227 |           |
##                |     0.053 |     0.139 |           |
## ---------------|-----------|-----------|-----------|
##       BROOKLYN |    445952 |    656034 |   1101986 |
##                |   635.980 |   406.070 |           |
##                |     0.405 |     0.595 |     0.272 |
##                |     0.283 |     0.265 |           |
##                |     0.110 |     0.162 |           |
## ---------------|-----------|-----------|-----------|
##      MANHATTAN |    325233 |    357578 |    682811 |
##                | 13150.287 |  8396.391 |           |
##                |     0.476 |     0.524 |     0.169 |
##                |     0.206 |     0.145 |           |
##                |     0.080 |     0.088 |           |
## ---------------|-----------|-----------|-----------|
##         QUEENS |    443440 |    239431 |    682871 |
##                | 118180.010 | 75457.332 |           |
##                |     0.649 |     0.351 |     0.169 |
##                |     0.281 |     0.097 |           |
##                |     0.110 |     0.059 |           |
## ---------------|-----------|-----------|-----------|
##  STATEN ISLAND |    121484 |     33479 |    154963 |
##                | 61816.665 | 39469.624 |           |
##                |     0.784 |     0.216 |     0.038 |
##                |     0.077 |     0.014 |           |
##                |     0.030 |     0.008 |           |
## ---------------|-----------|-----------|-----------|
##    Unspecified |     25074 |    623541 |    648615 |
##                | 205094.159 | 130951.571 |           |
##                |     0.039 |     0.961 |     0.160 |
##                |     0.016 |     0.252 |           |
##                |     0.006 |     0.154 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |   1577793 |   2471113 |   4048906 |
##                |     0.390 |     0.610 |           |
## ---------------|-----------|-----------|-----------|
##
##
```

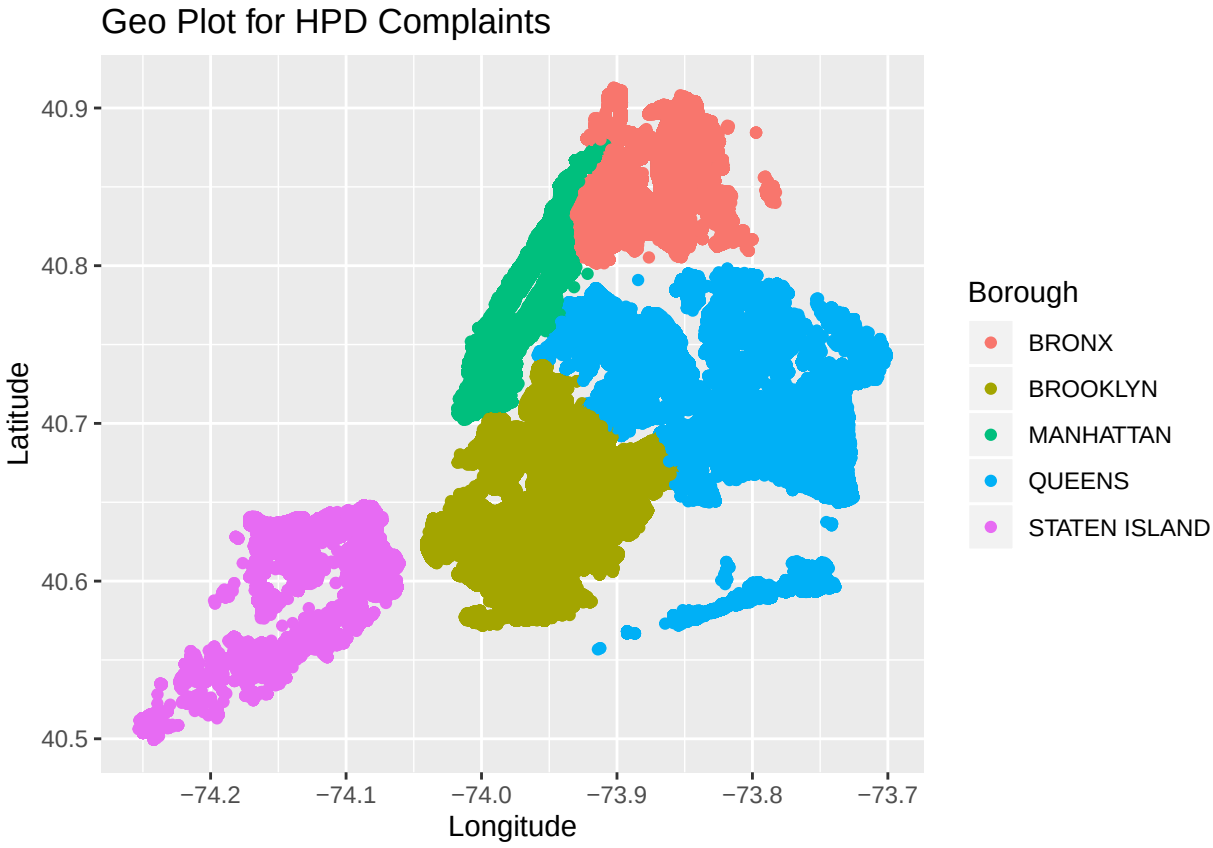
The above crosstab shows the amount of service requests received by HPD and DOT agencies with respect to each borough.

The following is where the latitude and longitude is plotted for HPD complaints representing boroughs with colors.

```
locationData <-nyc311nodups %>%
  select(Agency,
     Complaint.Type,
     Latitude,
   Longitude,
     Borough) %>%
  filter(Agency=="HPD" & Borough!="Unspecified")
ggplot(data = locationData) +
      geom_point(mapping = aes(x = Longitude, y = Latitude,  color=Borough)) +
```

```
      ggtitle("Geo Plot for HPD Complaints")
```

```
## Warning: Removed 4656 rows containing missing values (geom_point).
```



Conclusion

In this document, I have gained a good understanding of the 311 NYC service call requests dataset. I have performed data pre-processing steps i.e. ignoring irrelevant features for better analysis and removing duplicates, included a data dictionary which I will be working on and explored the various relevant features of the service call requests data and depicted my findings by visualizing them with plots and tabulations.