homework ii

Sachin Mohan Sujir

2020-09-11

Introduction

311 is a telephone number that is used for non-emergency government services. The dataset consists of more than 9 million records which has data of 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)

library(data.table)
# fast for when you are starting out:
# nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv",nrow=10000)
# after you get going:
nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv")
names(nyc311)<-names(nyc311) %>%
  stringr::str_replace_all("\\s", ".")
```

Data pre-processing

We perform data pre-processing by dropping irrelevant columns and also removing duplicate rows.

```
names(nyc311)
```

```
##  [1] "Unique.Key"                    "Created.Date"
##  [3] "Closed.Date"                   "Agency"
##  [5] "Agency.Name"                   "Complaint.Type"
##  [7] "Descriptor"                    "Location.Type"
##  [9] "Incident.Zip"                  "Incident.Address"
## [11] "Street.Name"                   "Cross.Street.1"
## [13] "Cross.Street.2"                "Intersection.Street.1"
## [15] "Intersection.Street.2"         "Address.Type"
## [17] "City"                          "Landmark"
## [19] "Facility.Type"                 "Status"
## [21] "Due.Date"                      "Resolution.Action.Updated.Date"
## [23] "Community.Board"               "Borough"
## [25] "X.Coordinate.(State.Plane)"    "Y.Coordinate.(State.Plane)"
## [27] "Park.Facility.Name"            "Park.Borough"
## [29] "School.Name"                   "School.Number"
## [31] "School.Region"                 "School.Code"
## [33] "School.Phone.Number"           "School.Address"
## [35] "School.City"                   "School.State"
## [37] "School.Zip"                    "School.Not.Found"
```

```
## [39] "School.or.Citywide.Complaint"   "Vehicle.Type"
## [41] "Taxi.Company.Borough"           "Taxi.Pick.Up.Location"
## [43] "Bridge.Highway.Name"            "Bridge.Highway.Direction"
## [45] "Road.Ramp"                      "Bridge.Highway.Segment"
## [47] "Garage.Lot.Name"                "Ferry.Direction"
## [49] "Ferry.Terminal.Name"            "Latitude"
## [51] "Longitude"                      "Location"
```

```
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"
```

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

```
## [1] 8250344      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)
options(xtable.comment=FALSE)
options(xtable.booktabs=TRUE)
narrow<-nyc311 %>%
  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: Date and Time) • Closed.Date – The date when the service request was closed by the responding agency. (Type: Data and Time) • Agency – Acronym of responding agency. (Plain Text) • Agency.Name – Full Agency name of responding City Government Agency. (Type: Plain Text) • Complaint.Type – The type of complaint reported (For example: vending, illegal parking, blocked driveway). • Descriptor - Detailed description of the corresponding complaint type. (Type: Plain Text) • Location.Type – The type of location based on the address information. (Plain Text) • Incident.Zip – Zip code of the incident location. (Type: Plain Text) • Status – The status of the service request submitted. (Type: Plain Text) • Due.Date – The date, during when the responding agency is expected to update the service request. (Type: Date and Time) • Resolution.Action.Updated.Date – Date when the responding agency last updated the service request. (Type: Date and Time) • Borough – town/ district of the NYC provided by submitter. (Values: BRONX, BROOKLYN, MANHATTAN, QUEENS, STATEN ISLAND) (Type: Plain Text) • Latitude – Geo-based latitude of the incident location. (Type: Number) • Longitude – Geo-based longitude of the incident location. (Type: Number) • Location – Combination of the geo-based latitude and longitude of the incident location. (Type: location)

```
names(nyc311)
```

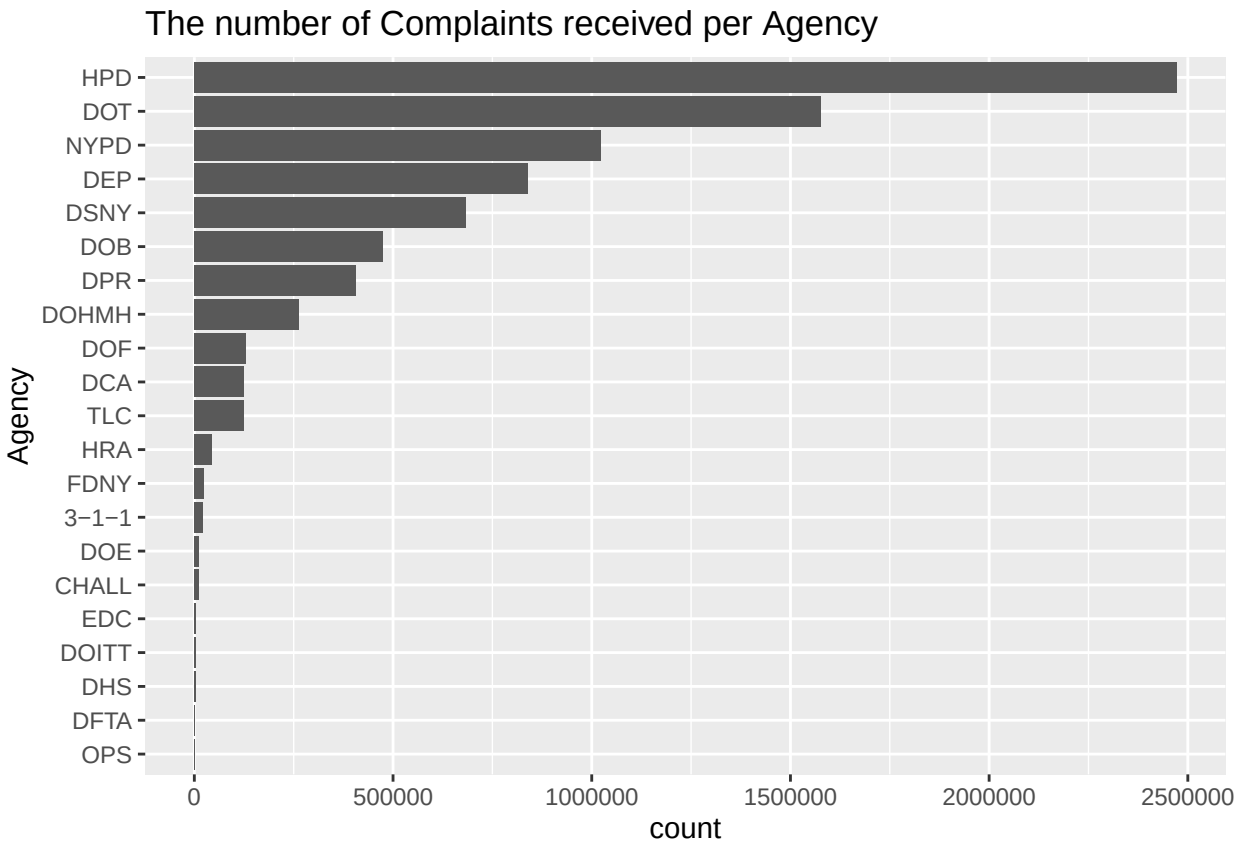
```
##  [1] "Created.Date"                    "Closed.Date"
##  [3] "Agency"                          "Agency.Name"
##  [5] "Complaint.Type"                  "Descriptor"
##  [7] "Location.Type"                   "Incident.Zip"
##  [9] "Status"                          "Due.Date"
## [11] "Resolution.Action.Updated.Date"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

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("The number of Complaints received per Agency")
p
```

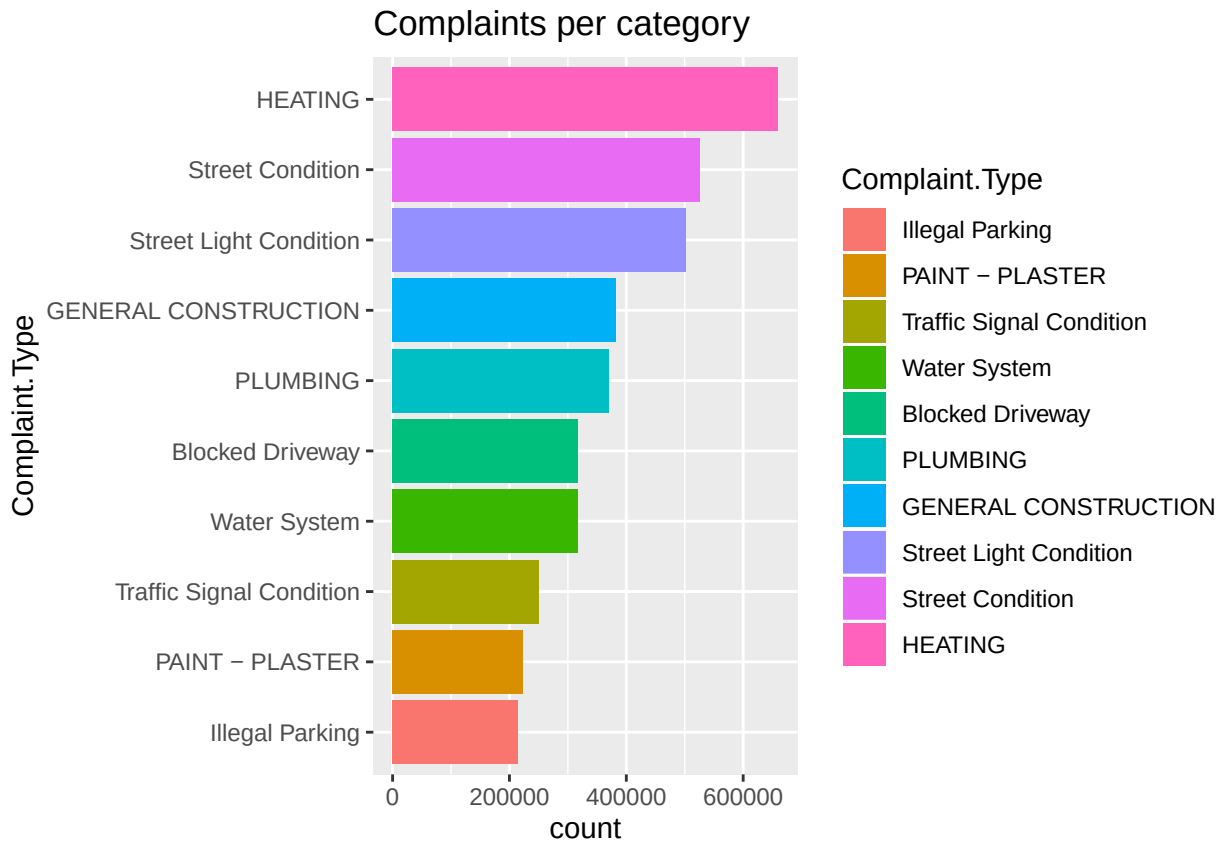


The following bar chart shows the top 10 complaint types received, with the color specified for each complaint type.

```
options(scipen = 999)
topComplaints <- narrow %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>%
  filter(count>100000) %>%
  top_n(10)
```

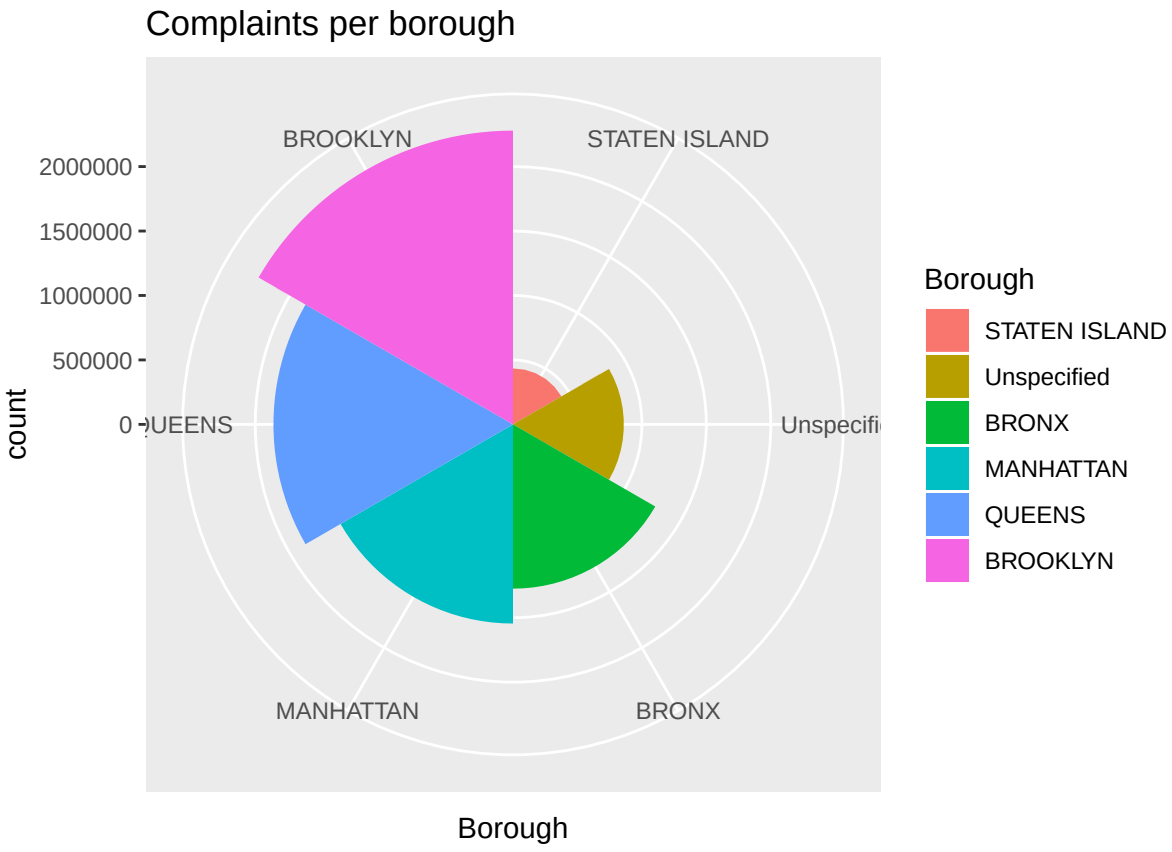
```
## Selecting by count
```

```
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
```



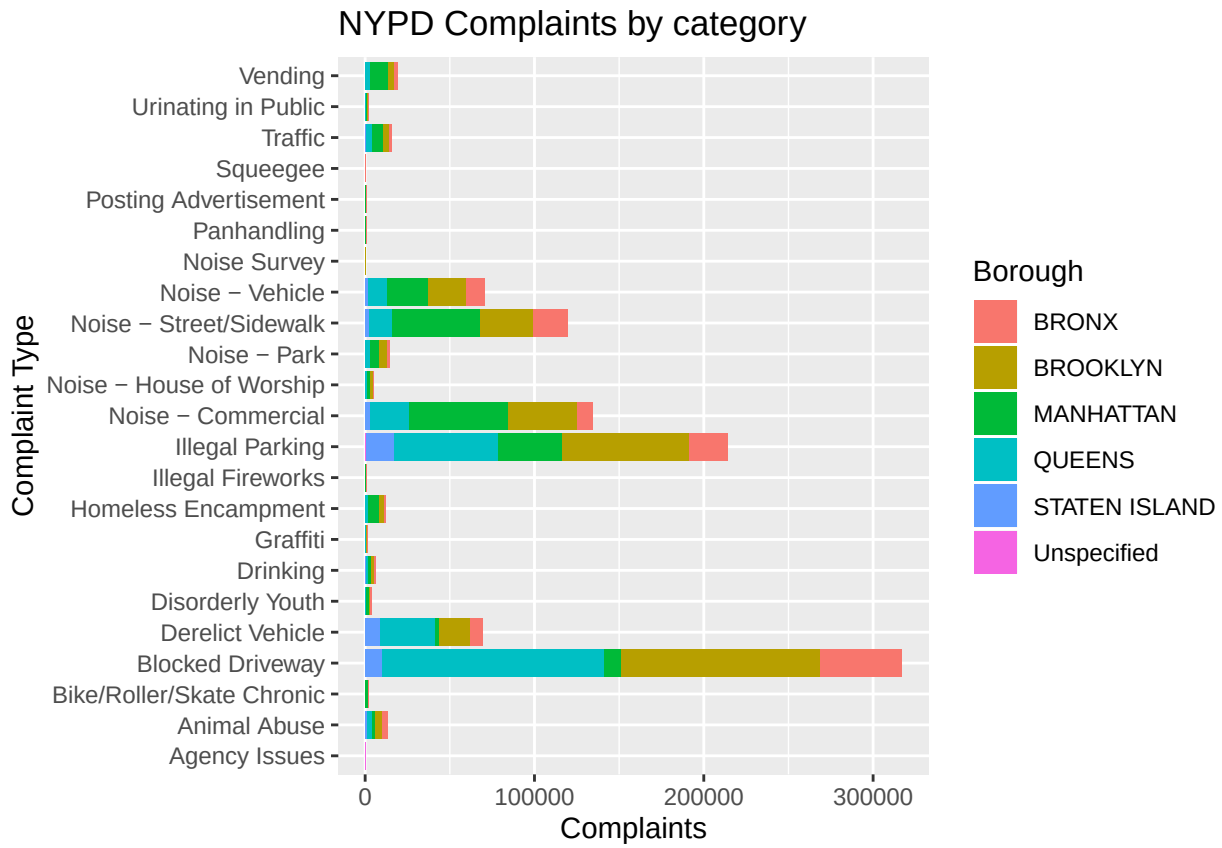
The following shows a coxcomb plot showing the boroughs that received the most service call requests depicted in the form of coxComb.

```
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
```



Compaints That NYPD received across each borough

```
hpdComplaints <-dplyr::filter(narrow, Agency=='NYPD')
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("NYPD Complaints by category")
```



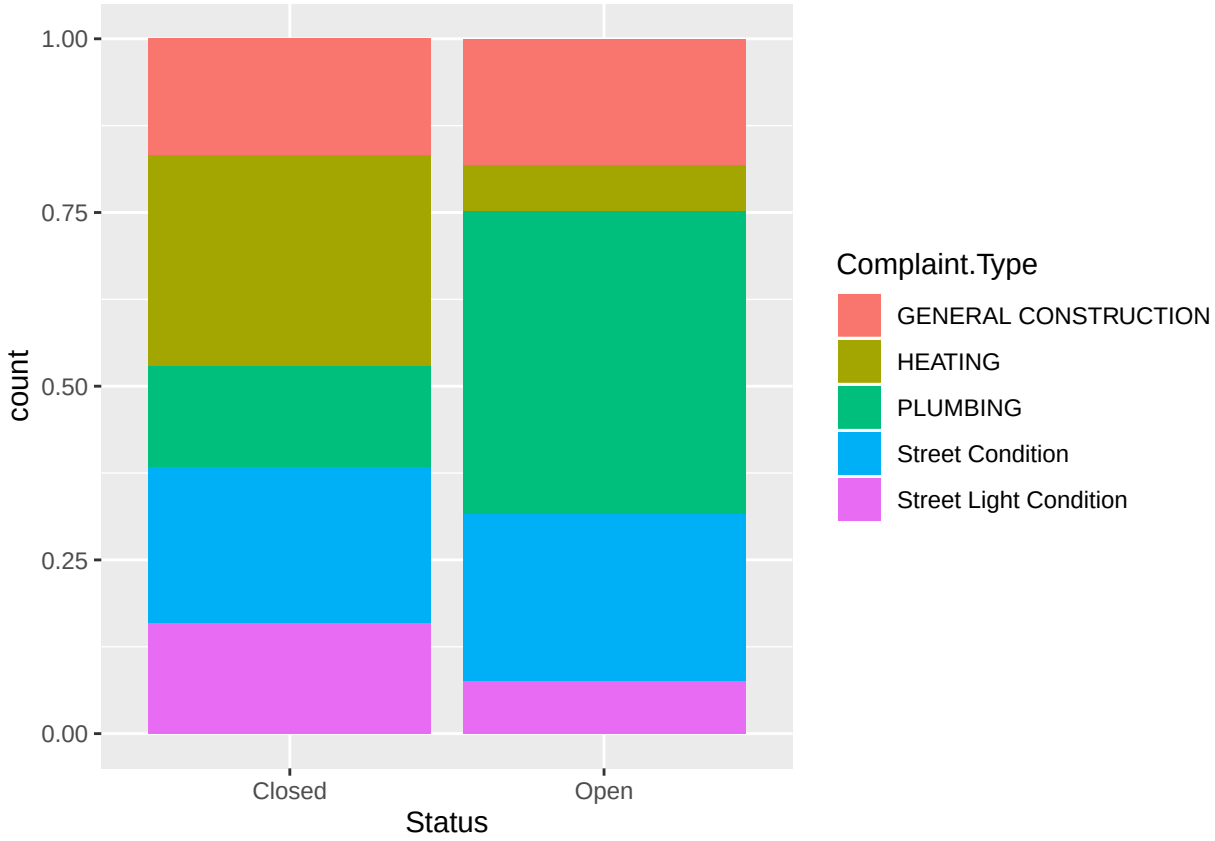
The table below shows information about the number of open and closed service call requests.

```
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 7043568
## 2 Open    772823
```

The bar below shows the percentage complaint type that are open and closed service requests of top 5 complaints. It seems like “Plumbing” has too many open service requests abd “Heating” requets have a good record.

```
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
```



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:  1413138
##
##
##               | xtabB$Complaint.Type
## xtabB$Borough | GENERAL CONSTRUCTION |              HEATING |             PLUMBING |            Row Total |
## --------------|----------------------|----------------------|----------------------|----------------------|
##         BRONX |                80647 |               137147 |                79859 |               297653 |
##               |                0.001 |               25.769 |               45.513 |                      |
##               |                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 |
##               |              602.707 |             1543.141 |              755.708 |                      |
##               |                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 |
##               |              406.054 |              306.360 |                8.298 |                      |
##               |                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 |
##               |              190.192 |               83.249 |                3.388 |                      |
##               |                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 |
##               |              522.250 |              977.886 |              342.913 |                      |
##               |                0.361 |                0.305 |                0.334 |                0.012 |
##               |                0.016 |                0.008 |                0.016 |                      |
##               |                0.004 |                0.004 |                0.004 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##   Unspecified |               116380 |               217112 |               103344 |               436836 |
##               |               32.607 |              835.469 |             1075.204 |                      |
##               |                0.266 |                0.497 |                0.237 |                0.309 |
##               |                0.304 |                0.329 |                0.279 |                      |
##               |                0.082 |                0.154 |                0.073 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##  Column Total |               382837 |               660106 |               370195 |              1413138 |
##               |                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=='NYPD'))
xtabB1<-select(xtabA1,Borough, Agency)
library(gmodels)
CrossTable(xtabB1$Borough,xtabB1$Agency)
```

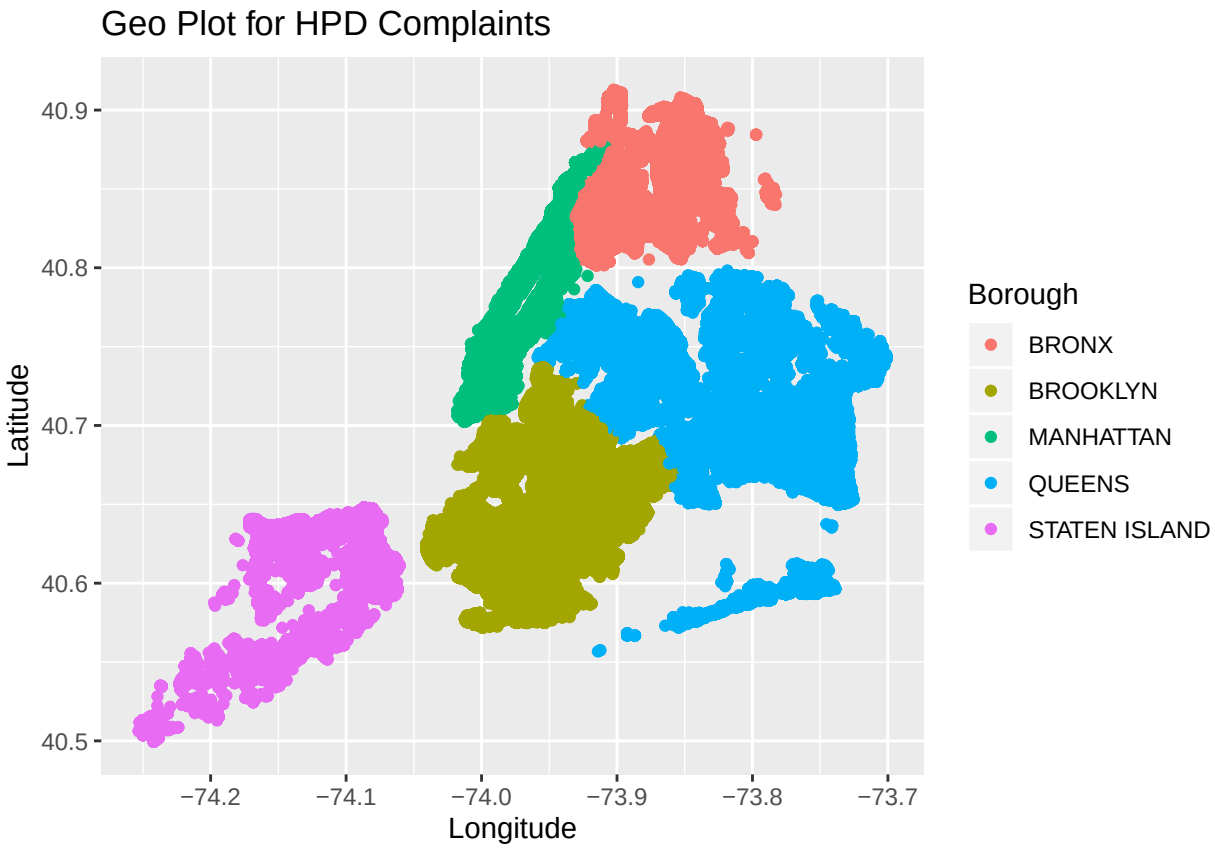
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  3494282
##
##
##                | xtabB1$Agency
## xtabB1$Borough |       HPD |      NYPD | Row Total |
## ---------------|-----------|-----------|-----------|
##          BRONX |    561056 |    130312 |    691368 |
```

```
##                | 10639.930 | 25697.626 |           |
##                |     0.812 |     0.188 |     0.198 |
##                |     0.227 |     0.127 |           |
##                |     0.161 |     0.037 |           |
## ---------------|-----------|-----------|-----------|
##       BROOKLYN |    656036 |    330496 |    986532 |
##                |  2484.247 |  5999.970 |           |
##                |     0.665 |     0.335 |     0.282 |
##                |     0.265 |     0.323 |           |
##                |     0.188 |     0.095 |           |
## ---------------|-----------|-----------|-----------|
##      MANHATTAN |    357578 |    223584 |    581162 |
##                |  6942.136 | 16766.691 |           |
##                |     0.615 |     0.385 |     0.166 |
##                |     0.145 |     0.219 |           |
##                |     0.102 |     0.064 |           |
## ---------------|-----------|-----------|-----------|
##         QUEENS |    239431 |    291643 |    531074 |
##                | 49349.267 | 119188.661 |           |
##                |     0.451 |     0.549 |     0.152 |
##                |     0.097 |     0.285 |           |
##                |     0.069 |     0.083 |           |
## ---------------|-----------|-----------|-----------|
##  STATEN ISLAND |     33479 |     46386 |     79865 |
##                |  9366.886 | 22622.962 |           |
##                |     0.419 |     0.581 |     0.023 |
##                |     0.014 |     0.045 |           |
##                |     0.010 |     0.013 |           |
## ---------------|-----------|-----------|-----------|
##    Unspecified |    623548 |       733 |    624281 |
##                | 75079.077 | 181331.460 |           |
##                |     0.999 |     0.001 |     0.179 |
##                |     0.252 |     0.001 |           |
##                |     0.178 |     0.000 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |   2471128 |   1023154 |   3494282 |
##                |     0.707 |     0.293 |           |
## ---------------|-----------|-----------|-----------|
##
##
```

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

```
locationData <-nyc311 %>%
  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 homework, I have gained a good understanding of the 311 NYC service call requests dataset. I have performed data pre-processing steps like removing irrelevant features for easier 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.