homework ii

Vivek Panchal, Rohit Kunjilikattil

2019-09-12

Introduction

NYC311 is an open data initiative.The main purpose of this non emergency line being generated is to filtrate calls from the emergency phone line 911.This dataset talks about all the complaints that are received in the five Boroughs of NYC i.e Bronx, Queens, Staten Island, Brooklyn and Manhattan.There are several complaints registered with each passing day.Few complaints from the many complaints that are reported are Illegal parked cars, noise complaints, taxi complaints, vending , plumbing and many more. This is a huge dataset.NYC311 receives these complaints and forward them to agencies operating in that area. Agencies are namely NYPD, HPD,TLC,DOT,DPR. The requests are addressed by the agencies and once the request is sorted they then close it.

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  2.1.3     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
```

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

```
install.packages('tidyverse', repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/xz/_k5p6kjn1bg2wc2f5cg88qym0000gn/T//Rtmpo3YlOn/downloaded_packages
```

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

```
##
## 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", ".")
all_complaints <-  nyc311 %>% select(2, 6, 24)
```

Removing two columns

```
df2 <- nyc311[,c("Unique.Key","City"):=NULL]
```

##DUPLICATES

```
if (!require(dplyr)) {
install.packages("dplyr",dependencies=TRUE)
library(dplyr)
}
nyc311nodups<-distinct(df2)
isTRUE(all.equal(nyc311nodups,df2))
```

```
## [1] FALSE
```

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

For our analysis we are working with following columns from the dataset.There were a total of 52 Columns in our dataset. A detailed description is given below: Agency - It has acronym of responding agency in the New York city. Agency Name - It has full agency name. Borough - It has the names of five boroughs in NYC i.e Bronx, Manhattan, Brooklyn, Staten Island and Queens. Complaint.Type - It tells us about the complaint that was registered for example Plumbing, Vending, Noise Complaints, Taxi complaint and many more. Descriptor - It is dependent on Complaint type and provides more information about the incident/complaint. Status - It shows the status of the complaint that was registered. The statuses are as follows assigned,open, and closed.
Incident.zip - It gives zip code of the incident location.

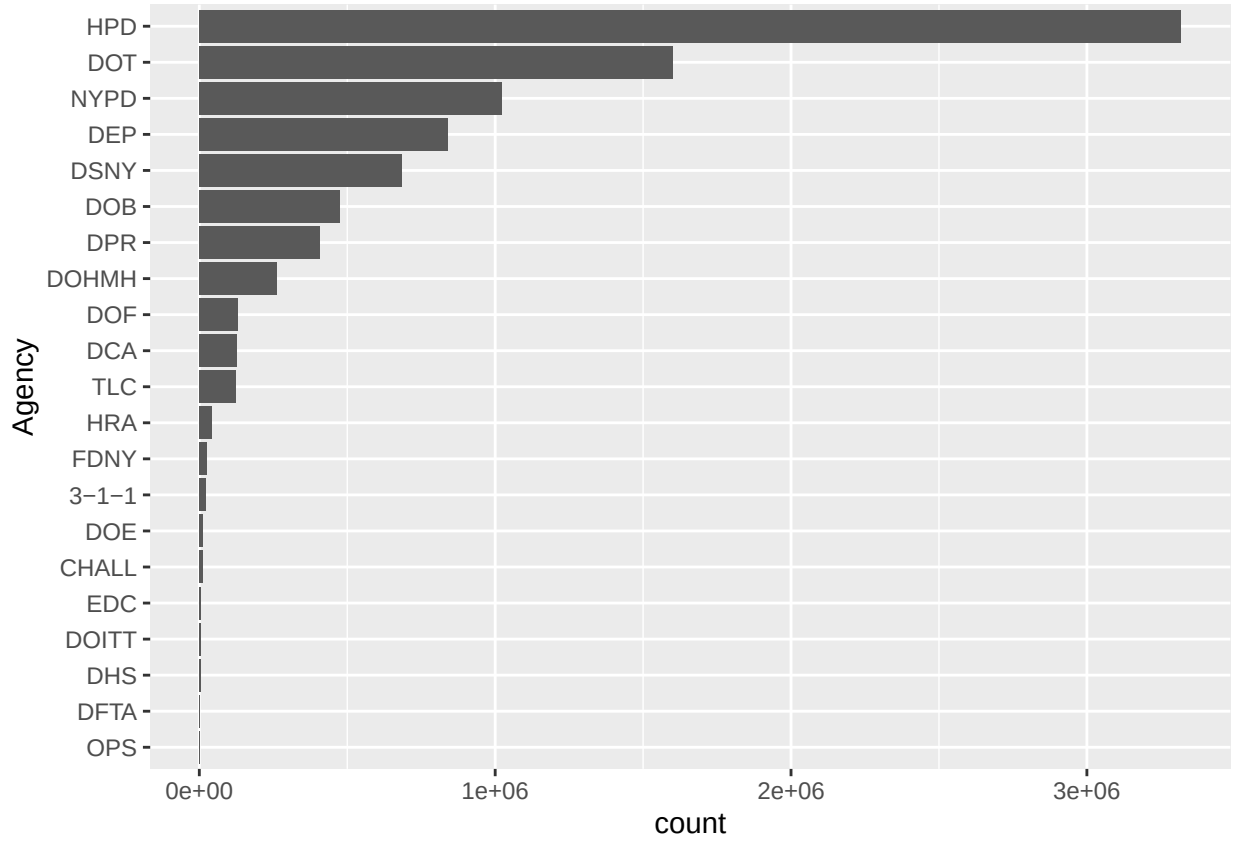
Other Columns in the dataset are as follows: Latitude- Latitude of the location. Longitude - Longitude of the location. Location.type - It tells the type of location it was for example Street, Sidewalk or Park.

Exploration

Here we explore the columns in the data set.

The following cross tabulation is done in order to visulaize the relationship between the agency and the number of complaints. One of the Xtab is a list of different agencies and the other xtab is a count of complaints registered in each agency. Crosstabing these two xtabs will give a chart displaying the number of complaints registered in 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()
p
```



(More plots should follow here.)

#CORRPLOT

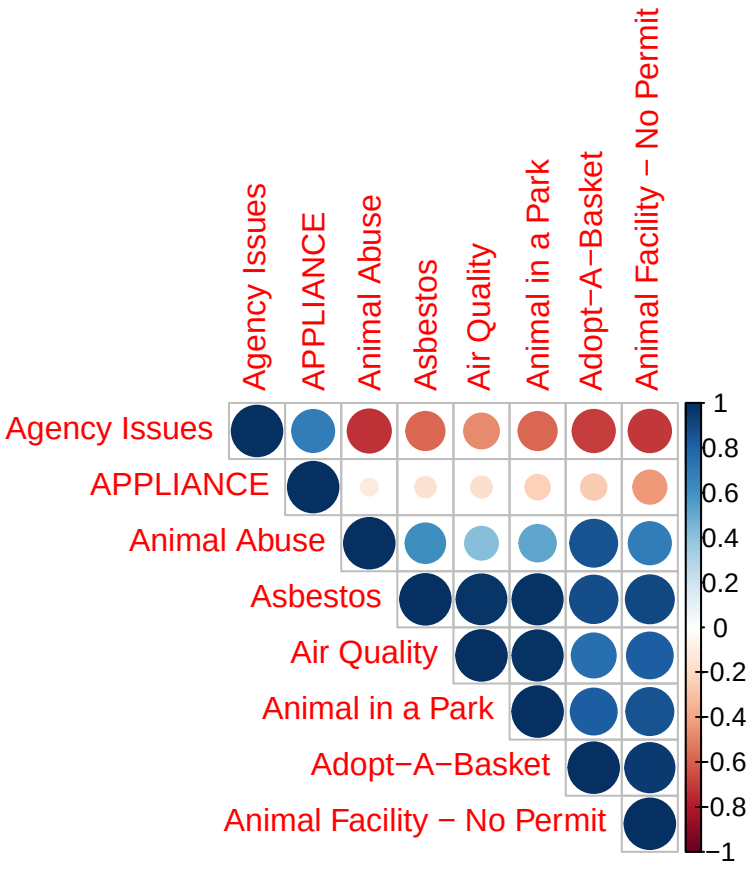
```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following objects are masked from 'package:data.table':
##
##     dcast, melt

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
nyc311_corr <- nyc311 %>% select( 5, 22)
new_table2 <- table(melt(nyc311_corr, id.var="Complaint.Type"))
new_table2<- as.data.table(new_table2)
nyc311_corr2<-new_table2 %>% select( 1, 4)
wide_table <- spread(new_table2,'Complaint.Type',N)
wide_table <- wide_table[,3:10]
resultfinal <- cor(wide_table, use = "complete.obs")
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
p <- corrplot(resultfinal, type="upper", order="hclust")
```




```
p
```

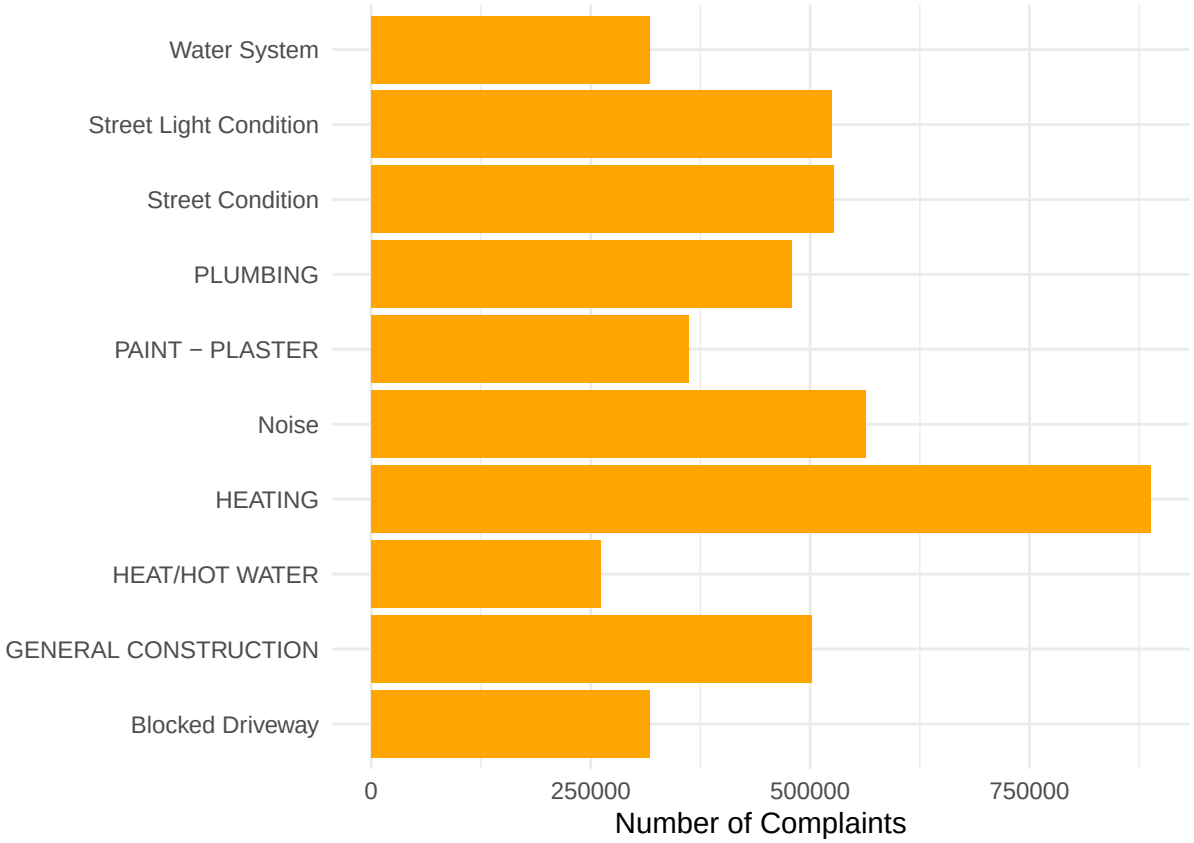
```
##                              Agency Issues  APPLIANCE Animal Abuse
## Agency Issues                    1.0000000  0.6928202   -0.7293635
## APPLIANCE                        0.6928202  1.0000000   -0.1128827
## Animal Abuse                    -0.7293635 -0.1128827    1.0000000
## Asbestos                        -0.5724859 -0.1646990    0.6137686
## Air Quality                     -0.4764309 -0.1708998    0.4242370
## Animal in a Park                -0.5706198 -0.2364906    0.5215976
## Adopt-A-Basket                  -0.6999158 -0.2557213    0.8610525
## Animal Facility - No Permit     -0.7059419 -0.4329925    0.6981924
##                                 Asbestos Air Quality Animal in a Park
## Agency Issues                  -0.5724859  -0.4764309       -0.5706198
## APPLIANCE                      -0.1646990  -0.1708998       -0.2364906
## Animal Abuse                    0.6137686   0.4242370        0.5215976
## Asbestos                        1.0000000   0.9728805        0.9843040
## Air Quality                     0.9728805   1.0000000        0.9891640
## Animal in a Park                0.9843040   0.9891640        1.0000000
## Adopt-A-Basket                  0.8820955   0.7597247        0.8271594
## Animal Facility - No Permit     0.9016892   0.8222111        0.8677723
##                                 Adopt-A-Basket Animal Facility - No Permit
```

```
## Agency Issues                  -0.6999158                -0.7059419
## APPLIANCE                      -0.2557213                -0.4329925
## Animal Abuse                    0.8610525                 0.6981924
## Asbestos                        0.8820955                 0.9016892
## Air Quality                     0.7597247                 0.8222111
## Animal in a Park                0.8271594                 0.8677723
## Adopt-A-Basket                  1.0000000                 0.9503262
## Animal Facility - No Permit     0.9503262                 1.0000000
```

As it is not possible to create a correlation matrix of non-numeric data, we first created a table of the frequency of different complaint types against different boroughs. Then we selected the complaint type and the frequency column from this table and calculated its correlation matrix which was then used to plot the corrplot. This corrplot shows how the different complaint types are related to each other. There are a total of 182 complaint types but we are only displaying the first few for clarity in the corrplot

#BAR PLOT

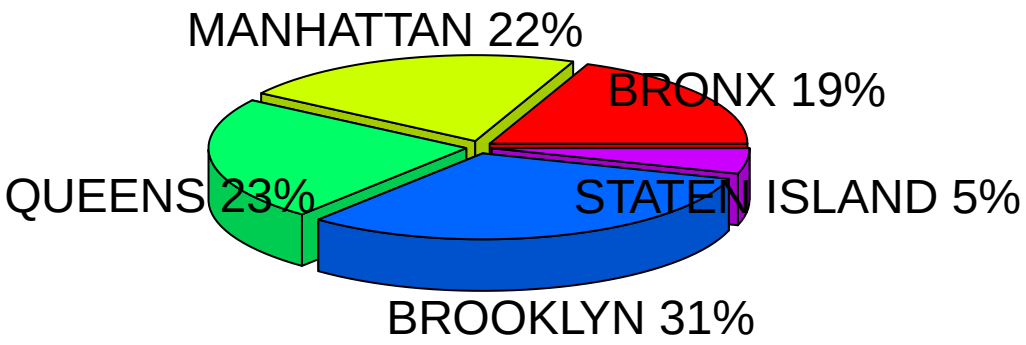
```
all_complaints$Complaint.Type[grepl("^Noise.*", all_complaints$Complaint.Type)] <- "Noise"

all_complaints_temp <- all_complaints %>%
    group_by(Complaint.Type) %>%
    summarise(count=n()) %>%
    arrange(desc(count))
top10_complaints <- top_n(all_complaints_temp, 10, count)
ggplot(top10_complaints) + geom_bar(aes(x=top10_complaints$Complaint.Type , y = top10_complaints$count
                                        ), fill = "Orange",
             stat = "identity") + theme_minimal() +
    xlab("") + ylab("Number of Complaints") + coord_flip()
```



This plot shows that against which complaint type maximum number of complaints were reported.We have made use of ggplot for analysis.It helps in creating graph that can be both univariate or multivariate categorical or numerical data.

#PLOTRIX

```
library(plotrix)
slices <- c(18.8, 21.8, 23.4, 31.2, 4.8)
lbl <- c("BRONX", "MANHATTAN", "QUEENS", "BROOKLYN", "STATEN ISLAND")
pct <- round(slices/sum(slices)*100)
lbl <- paste(lbl, pct)
lbl <- paste(lbl,"%",sep="")
pie3D(slices,labels=lbl,explode=0.05,
   main="Complaints for each Borough ")
```



This plot shows the % percentage complaints that were reported for each Borough. We made use of the library plotrix in this analysis.We have made use of "pie3D" function which displays pie chart in 3D manner.The maximum number of complaints were registered for Brooklyn Borough. This was calculated by selecting the columns 'Complaint.Type' and Borough.Grouping of this data was done by using Borough and summarization was done by finding the total length of(Complaint.Type) and then arranging them in Descending order.Hence,the following analysis was obtained.

(Next we include a crosstabulation.)

#CROSSTAB-1

```
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:  1868064
##
##
##               | xtabB$Complaint.Type
## xtabB$Borough | GENERAL CONSTRUCTION |              HEATING |             PLUMBING |            Row Total |
## --------------|----------------------|----------------------|----------------------|----------------------|
##         BRONX |               107626 |               195246 |               103964 |               406836 |
##               |               23.326 |               19.145 |                1.030 |                      |
##               |                0.265 |                0.480 |                0.256 |                0.218 |
##               |                0.215 |                0.220 |                0.217 |                      |
##               |                0.058 |                0.105 |                0.056 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##      BROOKLYN |               132552 |               190268 |               128383 |               451203 |
##               |             1076.405 |             2717.190 |             1398.387 |                      |
##               |                0.294 |                0.422 |                0.285 |                0.242 |
##               |                0.264 |                0.214 |                0.268 |                      |
##               |                0.071 |                0.102 |                0.069 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##     MANHATTAN |                61453 |               137458 |                63103 |               262014 |
##               |             1123.330 |             1347.582 |              245.877 |                      |
##               |                0.235 |                0.525 |                0.241 |                0.140 |
##               |                0.123 |                0.155 |                0.132 |                      |
##               |                0.033 |                0.074 |                0.034 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##        QUEENS |                41277 |                75776 |                43604 |               160657 |
##               |               79.707 |                4.192 |              142.183 |                      |
##               |                0.257 |                0.472 |                0.271 |                0.086 |
##               |                0.082 |                0.085 |                0.091 |                      |
##               |                0.022 |                0.041 |                0.023 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
## STATEN ISLAND |                 8329 |                 6011 |                 7525 |                21865 |
##               |             1030.062 |             1845.525 |              657.654 |                      |
##               |                0.381 |                0.275 |                0.344 |                0.012 |
##               |                0.017 |                0.007 |                0.016 |                      |
##               |                0.004 |                0.003 |                0.004 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##   Unspecified |               150277 |               282916 |               132296 |               565489 |
##               |               15.587 |              750.862 |             1106.709 |                      |
##               |                0.266 |                0.500 |                0.234 |                0.303 |
##               |                0.300 |                0.319 |                0.276 |                      |
##               |                0.080 |                0.151 |                0.071 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##  Column Total |               501514 |               887675 |               478875 |              1868064 |
##               |                0.268 |                0.475 |                0.256 |                      |
## --------------|----------------------|----------------------|----------------------|----------------------|
##
##
```

Cross-tabulations are used to represent the relationship between two or more variables in a dataset analytically. The axes of the crosstable are the variables whose relationship is to be represented. Considering the above cross-tab, we can see that it shows the number of 'PLUMBING' complaints against the different boroughs in New York City. From the crosstab, we can see in detail how many complaints were reported in a particular borough. As we can see, Brookyln has the largest contribution. The crosstab also sheds light on table proportions, showing that Brooklyn contributes to 0.406 % of the total 'Plumbing' complaints, leading to a total of 99 complaints.

#CROSSTAB -2

```
xtabA<-dplyr::filter(narrow,
  Complaint.Type == 'Noise' | Complaint.Type == 'Illegal Parking' | Complaint.Type == 'Blocked Driveway' | Complaint.Type == 'HE
xtabB<-select(xtabA,Borough,"Borough")
library(gmodels)
CrossTable(xtabB$Borough,xtabA$Complaint.Type)
```

```
##
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1517833
##
##
##               | xtabA$Complaint.Type
## xtabB$Borough | Blocked Driveway |   HEAT/HOT WATER |  Illegal Parking |            Noise | Street Condition |        Row Tot
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##         BRONX |            48247 |            87391 |            22796 |            12101 |            58515 |           2290
##               |            3.101 |        58546.276 |         2803.546 |        10679.602 |         5537.778 |
##               |            0.211 |            0.382 |            0.100 |            0.053 |            0.255 |           0.1
##               |            0.152 |            0.335 |            0.106 |            0.061 |            0.111 |
##               |            0.032 |            0.058 |            0.015 |            0.008 |            0.039 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##      BROOKLYN |           117895 |            78269 |            74929 |            48476 |           147547 |           4671
##               |         4216.703 |           51.547 |         1237.026 |         2638.725 |         1310.448 |
##               |            0.252 |            0.168 |            0.160 |            0.104 |            0.316 |           0.3
##               |            0.372 |            0.300 |            0.350 |            0.244 |            0.280 |
##               |            0.078 |            0.052 |            0.049 |            0.032 |            0.097 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##     MANHATTAN |             9894 |            59292 |            37752 |            99038 |           101264 |           3072
##               |        45936.967 |          793.346 |          721.699 |        85905.857 |          270.461 |
##               |            0.032 |            0.193 |            0.123 |            0.322 |            0.330 |           0.2
##               |            0.031 |            0.227 |            0.176 |            0.498 |            0.192 |
##               |            0.007 |            0.039 |            0.025 |            0.065 |            0.067 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##        QUEENS |           130899 |            33487 |            61451 |            31876 |           150515 |           4082
##               |        24372.684 |        19184.544 |          258.534 |         8720.123 |          550.372 |
##               |            0.321 |            0.082 |            0.151 |            0.078 |            0.369 |           0.2
##               |            0.413 |            0.128 |            0.287 |            0.160 |            0.286 |
##               |            0.086 |            0.022 |            0.040 |            0.021 |            0.099 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
## STATEN ISLAND |            10139 |             2497 |            16838 |             7087 |            68456 |           1050
##               |         6350.714 |        13405.198 |          276.047 |         3232.393 |        28107.765 |
##               |            0.097 |            0.024 |            0.160 |            0.067 |            0.652 |           0.0
##               |            0.032 |            0.010 |            0.079 |            0.036 |            0.130 |
##               |            0.007 |            0.002 |            0.011 |            0.005 |            0.045 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##   Unspecified |               89 |                0 |              368 |              225 |              500 |             11
##               |          101.058 |          203.202 |          242.868 |           31.817 |           19.640 |
##               |            0.075 |            0.000 |            0.311 |            0.190 |            0.423 |           0.0
##               |            0.000 |            0.000 |            0.002 |            0.001 |            0.001 |
##               |            0.000 |            0.000 |            0.000 |            0.000 |            0.000 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##  Column Total |           317163 |           260936 |           214134 |           198803 |           526797 |          15178
##               |            0.209 |            0.172 |            0.141 |            0.131 |            0.347 |
## --------------|------------------|------------------|------------------|------------------|------------------|---------------
##
##
```

In this crosstab, we look at analysing how the complaints types are spread with repect to the different boroughs. For this we take one of the xtabs to be the differnt types of complaints and list the borough names in a series of OR operations in the other xtab. By cross tabulating, these two tabs we get to see the division of complaint types in each borough. As calculating all types of complaints yielded a large result which was difficult to go through we decided to filter it to show only the top 5 most registered complaints using the information from the top 10 complaints bar chart.

#CROSSTAB- 3

```r
xtabA<-dplyr::filter(nyc311,
  Complaint.Type == 'Noise'|Complaint.Type == 'Street Condition'|Complaint.Type == 'HEAT/HOT WATER')
xtabB<-select(xtabA,Status,"Complaint.Type")
library(gmodels)
CrossTable(xtabA$Status,xtabB$Complaint.Type)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  986536
##
##
##              | xtabB$Complaint.Type
## xtabA$Status |   HEAT/HOT WATER |            Noise | Street Condition |        Row Total |
## -------------|------------------|------------------|------------------|------------------|
##     Assigned |                0 |                1 |             2310 |             2311 |
##              |          611.253 |          463.706 |          938.122 |                  |
##              |            0.000 |            0.000 |            1.000 |            0.002 |
##              |            0.000 |            0.000 |            0.004 |                  |
##              |            0.000 |            0.000 |            0.002 |                  |
## -------------|------------------|------------------|------------------|------------------|
##       Closed |           179006 |           152332 |           480827 |           812165 |
##              |         5969.360 |          784.677 |         5124.332 |                  |
##              |            0.220 |            0.188 |            0.592 |            0.823 |
##              |            0.686 |            0.766 |            0.913 |                  |
##              |            0.181 |            0.154 |            0.487 |                  |
## -------------|------------------|------------------|------------------|------------------|
##         Open |            81930 |            45777 |            31912 |           159619 |
##              |        37352.603 |         5759.665 |        33358.348 |                  |
##              |            0.513 |            0.287 |            0.200 |            0.162 |
##              |            0.314 |            0.230 |            0.061 |                  |
##              |            0.083 |            0.046 |            0.032 |                  |
## -------------|------------------|------------------|------------------|------------------|
##      Pending |                0 |                0 |            11725 |            11725 |
##              |         3101.230 |         2362.778 |         4768.473 |                  |
##              |            0.000 |            0.000 |            1.000 |            0.012 |
##              |            0.000 |            0.000 |            0.022 |                  |
##              |            0.000 |            0.000 |            0.012 |                  |
## -------------|------------------|------------------|------------------|------------------|
##      Started |                0 |              693 |                0 |              693 |
##              |          183.297 |         2192.580 |          370.053 |                  |
##              |            0.000 |            1.000 |            0.000 |            0.001 |
##              |            0.000 |            0.003 |            0.000 |                  |
##              |            0.000 |            0.001 |            0.000 |                  |
## -------------|------------------|------------------|------------------|------------------|
##   Unassigned |                0 |                0 |                2 |                2 |
##              |            0.529 |            0.403 |            0.813 |                  |
##              |            0.000 |            0.000 |            1.000 |            0.000 |
##              |            0.000 |            0.000 |            0.000 |                  |
##              |            0.000 |            0.000 |            0.000 |                  |
## -------------|------------------|------------------|------------------|------------------|
##  Unspecified |                0 |                0 |               21 |               21 |
##              |            5.554 |            4.232 |            8.541 |                  |
##              |            0.000 |            0.000 |            1.000 |            0.000 |
##              |            0.000 |            0.000 |            0.000 |                  |
##              |            0.000 |            0.000 |            0.000 |                  |
## -------------|------------------|------------------|------------------|------------------|
## Column Total |           260936 |           198803 |           526797 |           986536 |
```

```
##              |           0.264 |           0.202 |           0.534 |                 |
## -------------|-----------------|-----------------|-----------------|-----------------|
##
##
```

Using the information we got from the top 10 complaints graph, we create a cross tabulation of the top 3 complaints with their complaint statuses. This information is useful to understand how responsive the government is about the types of complaints that are registered the most. Here the first xtable is a series of OR conditions in order to select the top 3 noise complaints and then the other xtab is the statuses. The cross tablulation then shows different statuses of each noise complaint along with their proper proportions.

Installing TinyTex

```
install.packages('tinytex', repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/xz/_k5p6kjn1bg2wc2f5cg88qym0000gn/T//Rtmpo3YlOn/downloaded_packages
```

```
tinytex::install_tinytex()
```

```
## Warning: Detected an existing tlmgr at /Users/Student/Library/TinyTeX/
## bin/x86_64-darwin/tlmgr. It seems TeX Live has been installed (check
## tinytex::tinytex_root()). You are recommended to uninstall it, although
## TinyTeX should work well alongside another LaTeX distribution if a LaTeX
## document is compiled through tinytex::latexmk().

## The directory /usr/local/bin is not writable. I recommend that you make it writable. See https://gitl

## TinyTeX installed to /Users/Student/Library/TinyTeX
```

Conclusion

We undertook following steps for analysis and came up with a conclusion.Firstly, we removed two columns from our dataset i.e 'Unique.Key'(to identify duplicate values) and 'City' because it is always going to be New York.Secondly, we checked for duplicates in our dataset and found out those values. We checked for distinct values and stored them in a variable called as 'nyc311nodups' and later compared it with the duplicate values that we had obtained and the result was False.We made use of several libraries for the purpose of analysis.In corrplots,we represented how each complaint types affect all oter complaint types.A bar plot created showed that the maximum number of complaints were registered against Noise in all the Boroughs.The maximum number of complaints were reported to HPD(Department of Housing Preservation and Development).We made use of Plotrix which showed that Brooklyn was the Borough where maximum number of complaints were made followed by Staten Island where the least % of complaints were made.Then,crosstabulation was performed on dataset which showed that maximum number of complaints for Plumbing was registered in Brooklyn.CrossTabulation 2 showed the number of top five complaints in each borough. CrossTabulation 3 showed the statuses for complaint type Noise in each Borough.The results of these analysis and visualizations revealed about the volume of complaints filed at New York city.This can surely help all the government agencies to take necessary steps in the future to overcome such incidents.It was also help in resolving various issues in a dedicated time.It will also help them in recruiting more people where there is need of an hour in this case for Brooklyn where the maximum number of complaints were lodged.