R Project

Swati Mehla

10/12/2018

Contents

Loading Libraries

```
library(tidyverse)
if(!require(kableExtra)){
    install.packages("kableExtra")
    library(kableExtra)
}
if(!require(data.table)){
    install.packages("data.table")
    library(data.table)
}
if(!require(lubridate)){
    install.packages("lubridate")
    library(lubridate)
}
if(!require(dplyr)){
    install.packages("dplyr")
```

```
    library(dplyr)
}
if(!require(ggplot2)){
    install.packages("ggplot2")
    library(ggplot2)
}
library(stringr)
if(!require(stringr)){
    install.packages("stringr")
    library(stringr)
}
if(!require(DT)){
    install.packages("DT")
    library(DT)
}
if(!require(leaflet)){
    install.packages('leaflet')
    library(leaflet)
}
if(!require(grid)){
    install.packages('grid')
    library(grid)
}
library(gridExtra)
if(!require(knitr)){
    install.packages("knitr")
    library(knitr)
}

fillColor = "#FFA07A"
fillColor2 = "#F1C40F"
```

Introduction:

About NYC 311 Data:

In this project, we are visualizing New York City 311 incident call data. 311 is a phone number supported in United States that provides access to non-emergency municipal services. The phone number 311 is intended in part to divert routine inquiries and non-urgent community concerns from the 911 number which is reserved for emergency service.NYC311 receives thousands of requests related to several complaint types like noise complaints, street conditions, plumbing issues etc.These requests are forwarded to the relevant agencies, such as the Police, Buildings or Transportation.

About NYC_Crimes:

NYC_Crimes data set is a data set about all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD). We care about public safety and wanted to visualize what kind of non-urgent and urgent requests are reported to NYPD. We have joined NYC_Crimes with NYC311 dataset to create a new dataset named as Merged_Data which is a joined dataset about all the complaints made to NYPD.

Approach:

Using these datasets, we did some exploratory visualization to gain insights into the type of complaints, number of complaints and area of complaints reported through 311 and how it varies through the year. We have created a data dictionary for all the important columns which can be seen in APPENDIX A.

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

Data Preparation:

1) NYC311 is a huge data set with over 1 million records, 52 columns and over 240 different types of Complaints.
   Most of our visualization is concentrated around Borough and Complaint Type of the service request. We removed all the rows which had "NA", "Unspecified" or "blank" Borough. Values of column such as Cross/Intersection Street were mutually Exclusive. We united such columns.

We used regular expressions to determine which all columns hold similar properties of data. For example, there were 7 columns which included Information about School such as School's phone number, address, district etc. We felt that some of these attributes are not useful for our analysis. Hence, we removed those columns. We used regular expressions on following names: School Bridge Park Ferry

We used following lines of code to identify redundant columns NumberOfColumn_forschool = grep('^School', colnames(nyc311)) NumberOfColumn_forbridge = grep('^Bridge', colnames(nyc311))

Moreover, we removed columns which had similar data like Park Borough and Borough. A list of all the columns deleted can be seen the code below.

```
nyc311 <- unite(nyc311,Cross_Street1/Intersection_Street1,Cross.Street.1,
                Intersection.Street.1, sep = "")
nyc311 <- unite(nyc311,Cross_Street2/Intersection_Street2,Cross.Street.2,
                Intersection.Street.2, sep = "")
nyc311 <- na.omit(nyc311, cols = "Complaint.Type")
nyc311 <- na.omit(nyc311, cols = "Descriptor")
nyc311 <- na.omit(nyc311, cols = "Borough")
nyc311 <- nyc311[- grep("Unspecified", nyc311$Borough),]
nyc311 <- (data.frame(nyc311[nyc311$Created.Date != "" & nyc311$Created.Date != " " &
                               nyc311$Created.Date != "N/A"  & nyc311$Created.Date!=
                               "Unspecified" & nyc311$Created.Date != "NA" ,]))

nyc311.Created.Date.Time <- data.frame(nyc311$Created.Date)
colnames(nyc311.Created.Date.Time) <- c("Created.Date")
# Removing unwanted columns
nyc311 <- subset(nyc311, select=-c(Agency.Name, Incident.Address, Landmark, School.Number,
                                   School.Not.Found, School.Code,School.Address,
                                   School.Phone.Number,Bridge.Highway.Segment,
                                   Park.Borough, Created.Date,
                                  Community.Board))
```

2) Time stamp:- We generated additional features from the timestamps to be able to extract year, month, and hour. We are separating the Created.Date column on the basis of a format. Then we are separating the columns using different seperators like "-" and " " to extract Created year and Created hour

from the Created Date column. Then we use cbind to bind these new columns "Created.Year" and "Created.Hour" to our current data set. Later on we will use these columns to find yearly and hourly trends of the data.

```
nyc311.Created.Date.Time <- strptime(nyc311.Created.Date.Time$Created.Date,
                                     format="%m/%d/%Y %I:%M:%S %p")
nyc311.Created.Date.Time <- data.frame(nyc311.Created.Date.Time)
colnames(nyc311.Created.Date.Time ) <- c("Created.Date")
nyc311.Created.Date.Time <- separate(nyc311.Created.Date.Time,
                                     Created.Date,c('Created.Date','Created.Time'),
                                     sep = " ")
nyc311.Created.Date.Time <- separate(nyc311.Created.Date.Time,
                                  Created.Date,c( 'Created.Year', 'Created.Month',
                                                  'Created.Day'), sep = "-")
nyc311.Created.Date.Time <- separate( nyc311.Created.Date.Time, Created.Time,
                                      c('Created.Hour','Ignore1','Ignore2'))
nyc311.Created.Date.Time <- select( nyc311.Created.Date.Time, Created.Year, Created.Month,
                                    Created.Day, Created.Hour)
nyc311<-cbind(nyc311, nyc311.Created.Date.Time)
nyc311.Created.Date.Time <- NULL
nyc311 <- (data.frame(nyc311[nyc311$Unique.Key != "" & nyc311$Unique.Key != " " &
                               nyc311$Unique.Key != "N/A"  & nyc311$Unique.Key !=
                               "Unspecified" & nyc311$Unique.Key != "NA" ,]))

nyc311 <- (data.frame(nyc311[nyc311$Created.Year != "" & nyc311$Created.Year != " " &
                               nyc311$Created.Year != "N/A"  & nyc311$Created.Year !=
                               "Unspecified" & nyc311$Created.Year != "NA" ,]))
```

3) We considered a new dataset NYC_Crimes to get more insight into the type of complaints reported to NYPD per Borough. We cleaned this data set and changed the name of the columns to meaningful names.

```
colnames(NYC_Crimes)[colnames(NYC_Crimes)=="CMPLNT_FR_DT"] <- "COMPLAINT.FROM.DATE"
colnames(NYC_Crimes)[colnames(NYC_Crimes)=="CMPLNT_FR_TM"] <- "COMPLAINT.FROM.TIME"
colnames(NYC_Crimes)[colnames(NYC_Crimes)=="CMPLNT_TO_DT"] <- "COMPLAINT.TO.DATE"
colnames(NYC_Crimes)[colnames(NYC_Crimes)=="CMPLNT_TO_TM"] <- "COMPLAINT.TO.TIME"
colnames(NYC_Crimes)[colnames(NYC_Crimes)=="BORO_NM"] <- "Borough"

NYC_Crimes <- subset(NYC_Crimes, select=-c(HADEVELOPT,Lat_Lon,X_COORD_CD,Y_COORD_CD))
NYC_Crimes <- na.omit(NYC_Crimes,cols ="Borough")

nyc311_NYPD <- (data.frame(nyc311[nyc311$Agency == "NYPD",]))
nyc_Crimes_NYPD <- NYC_Crimes
nyc_Crimes_NYPD <- nyc_Crimes_NYPD[nyc_Crimes_NYPD$Latitude != "NA" &
                                     nyc_Crimes_NYPD$Longitude != "NA" &
                                     !is.na(nyc_Crimes_NYPD$Latitude) &
                                     !is.na(nyc_Crimes_NYPD$Longitude) ]
nyc_Crimes_NYPD <- (data.frame(nyc_Crimes_NYPD[nyc_Crimes_NYPD$JURIS_DESC ==
                                                 "N.Y. POLICE DEPT"]))
```

4) Lastly, we filtered the NYC_311 dataset for Agency name = "NYPD" and merged it with NYPD_Crimes data set on Borough, Latitude and longitude to compare NYC311 and NYPD_Crimes complaints with respect to Borough.

```
nyc311_NYPD$Latitude <-formatC(nyc311_NYPD$Latitude, digits = 4, format = "f")
nyc311_NYPD$Longitude <-formatC(nyc311_NYPD$Longitude, digits = 4, format = "f")
```

```
nyc_Crimes_NYPD$Latitude <-formatC(nyc_Crimes_NYPD$Latitude, digits = 4, format = "f")
nyc_Crimes_NYPD$Longitude <-formatC(nyc_Crimes_NYPD$Longitude, digits = 4, format = "f")
Merged_Data <- left_join(nyc311_NYPD, nyc_Crimes_NYPD, by = c("Borough" = "Borough",
                                                              "Latitude" = "Latitude",
                                                              "Longitude" = "Longitude"))
nyc311_NYPD$Latitude <- as.numeric(nyc311_NYPD$Latitude)
nyc311_NYPD$Longitude <- as.numeric(nyc311_NYPD$Longitude)
nyc_Crimes_NYPD$Latitude <- as.numeric(nyc_Crimes_NYPD$Latitude)
nyc_Crimes_NYPD$Longitude <- as.numeric(nyc_Crimes_NYPD$Longitude)
nyc311$Latitude <-as.numeric(nyc311$Latitude)
nyc311$Longitude <- as.numeric(nyc311$Longitude)
NYC_Crimes <- NULL
```

Visualization:

1) Complaint Per Boroughs:

We wanted to visualize the frequencies of different types of complaints in every Borough.

To Visualize this:

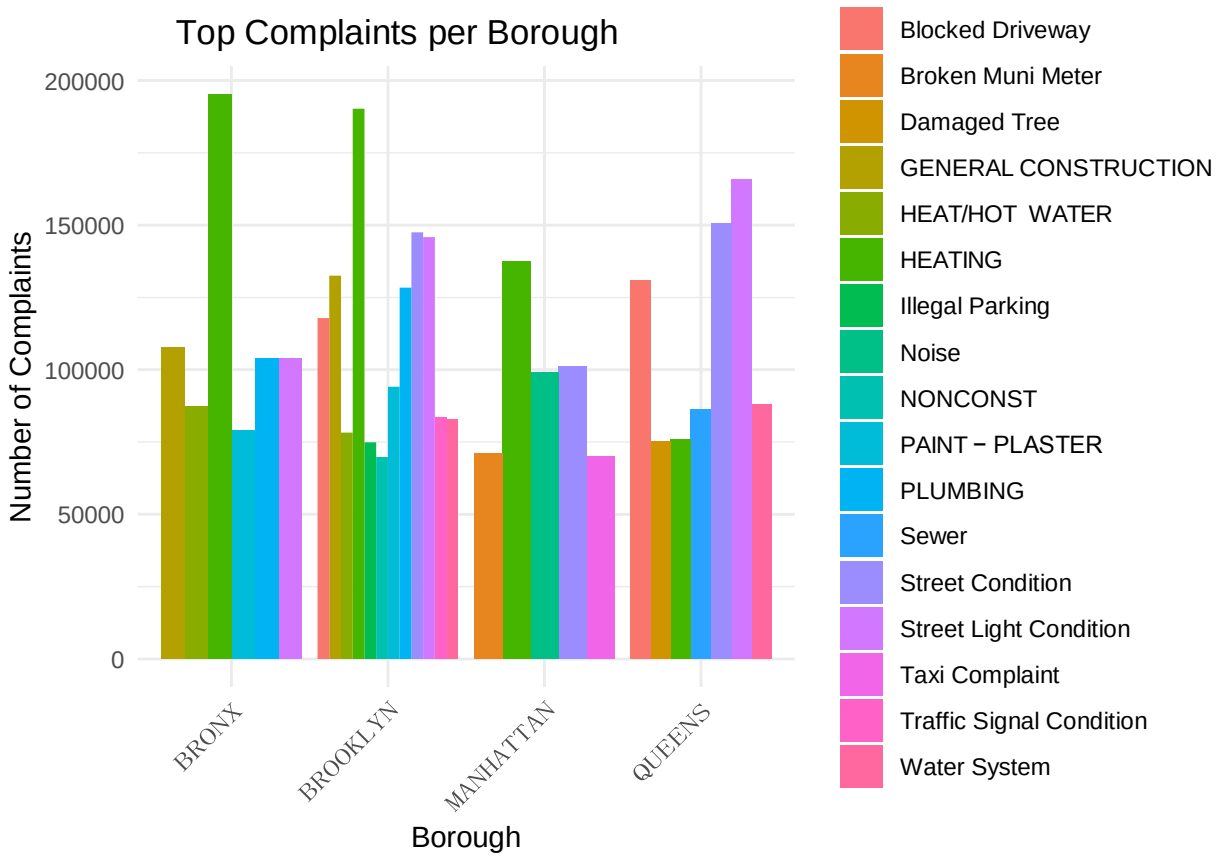
We drew a histogram by grouping the attributes “Borough” and “Complaint Type”. We displayed 30 such obsevations.

Observation:

It can be observed that Brooklyn received the highest number of Complaints followed by Bronx. The most frequent complaint Type in Brooklyn is Heating and the second highest is Street Conditions.

Moreover, we found that the highest number of complaints in Bronx were for heating, whereas the Queens dominated the Complaints about street conditions.

```
nyc311 %>% group_by(Borough, Complaint.Type) %>%
  filter(!is.na(Borough)) %>%
  summarize(num_call = n())>%%
  ungroup() %>%
  arrange(desc(num_call)) %>%
  head(30)%>%
  ggplot(aes(y=num_call, x=Borough, fill = Complaint.Type, color = Complaint.Type)) +
  geom_histogram(stat="identity",position = 'dodge') +
  ggtitle("    Top Complaints per Borough") +
  xlab("Borough") + ylab("Number of Complaints") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



2) Status of Request:

We wanted to see the status of service requests for every borough. That is the frequency of service resolutions. Moreover, what kind of location type has more service request still open or pending.

To Visualize this:
We grouped the data on the basis of status. Then for Status = "pending", "Closed" and "Open" we are filtering the data set Using the filtered data set for every status, we are checking the distribution for every Borough.

Observation:
Brooklyn had the highest number of service requests which are still open. Number of requests closed in brooklyn are high as well. Number of Closed Requests are more in Queen than Bronx whereas, number of Pending requests are more in Bronx Then Queens.

Location Types with highest complaints: It can also be seen that there are lot of issues created at residential Locations. Number of complaints with "Open" status are more for the residential locations. This might be because Street Conditions issues are easily fixable than heating problems.

```
p <- (data.frame(nyc311$Status[nyc311$Status != "" & nyc311$Status!= " " &
                               nyc311$Status != "N/A" & nyc311$Status != "Unknown" ]))
a <- as.data.frame(table(p))
newdata <- head(a[order(-a$Freq),],5 )

p1 <- (data.frame(nyc311$Borough[nyc311$Status == toString(newdata[1, "p"])]))
a1 <- as.data.frame(table(p1))
newdata1 <- head(a1[order(-a1$Freq),],5 )
titlestr = paste("Closed Service Requests")
```

```
p1 <- ggplot(data=newdata1, aes(x = reorder( newdata1$p1, -newdata1$Freq) ,
                                y=newdata1$Freq)) +
  geom_bar(stat="identity",color='steelblue',fill='steelblue') +
  xlab(toString(titlestr)) +
  ylab("Service Requests") +
  theme( axis.text.x = element_text(angle=15), legend.text = element_text(size=5),
         legend.position="bottom", legend.key.size = unit(0.4, "cm"))

p3 <- (data.frame(nyc311$Borough[nyc311$Status == toString(newdata[2, "p"])]))
a1 <- as.data.frame(table(p3))
newdata3 <- head(a1[order(-a1$Freq),],5 )
titlestr = paste("Open Service Requests")
p3<-ggplot(data=newdata3, aes(x=reorder(newdata3$p3, -newdata3$Freq), y=newdata3$Freq)) +
  geom_bar(stat="identity",color='green',fill = 'green') +
  xlab(toString(titlestr)) +
  theme(axis.text.x=element_text(angle=20)) +
  ylab("Service Requests") +
  theme(axis.text.x=element_text(angle=15), legend.text = element_text(size=5),
        legend.position="bottom",legend.key.size = unit(0.4, "cm"))

##We wanted to see, what kind of complaints are their in Bronx and Queens,
p2 <- (data.frame(nyc311$Borough[nyc311$Status == toString(newdata[3, "p"])]))
a1 <- as.data.frame(table(p2))
newdata2 <- head(a1[order(-a1$Freq),],5 )
titlestr = paste("Pending Service Requests")
p2<-ggplot(data=newdata2, aes(x=reorder(newdata2$p2, -newdata2$Freq), y=newdata2$Freq)) +
  geom_bar(stat="identity",color='orange',fill = 'orange') +
  xlab(toString(titlestr)) +
  ylab("Service Requests") +
  theme(axis.text.x=element_text(angle=15),legend.text =  element_text(size=5),
        legend.position="bottom", legend.key.size = unit(0.4, "cm"))


locations <- (data.frame(nyc311$Location.Type[nyc311$Location.Type != "" &
                                              nyc311$Location.Type != " " &
                                              nyc311$Location.Type != "N/A"  &
                                              nyc311$Location.Type != "Unspecified" ]))
colnames(locations) <- "Location"
a <- as.data.frame(table(locations))
newdata <- head(a[order(-a$Freq),],4 )

locationSR <- nyc311[ nyc311$Status != "" & nyc311$Status != " " &
                      nyc311$Status != "NA"  & nyc311$Status != "Closed - Testing" &
                      nyc311$Status != "Unspecified" & nyc311$Status != "Unassigned" &
                      nyc311$Location.Type != "NA" &
                      ((nyc311$Location.Type == toString(newdata[1,1])) |
                         (nyc311$Location.Type == toString(newdata[2,1])) |
                         (nyc311$Location.Type == toString(newdata[3,1])) |
                         (nyc311$Location.Type == toString(newdata[4,1]))),]
locationSR <- na.omit(locationSR, cols = "Location.Type")
bar_plt <- ggplot(locationSR, aes(x = Location.Type, fill = Status))
bar_plt <- bar_plt + geom_bar() +
  xlab("Location Types") +
```

```
  ylab("Service Requests") +
  theme(axis.text.x=element_text(angle=15), legend.text = element_text(size=4),
        legend.position="bottom", legend.key.size = unit(0.3, "cm"))

grid.arrange(p1,p3,p2,bar_plt,ncol=2,layout_matrix=cbind(c(1,2,3),c(4,4,4)))
```

3) Agency Vs Complaints.

We wanted to visualize the number of complaints made to agencies. For that we Made several bar plots. Which can be seen in the grid below.

To visualize this:
We constructed bar plots for most frequent complaint types. Then we constructed a bar plot for the Agencies which received most complaint types. From there we targeted the Agencies, which interests us, HPD(most frequent) and NYPD. We then filtered the data by a given agency to see complaint trends in the agency. Then we used the grid package to arrange all the plots in 2*2 grid.
Observation:
Top 5 Complaint Types: These are most frequent complaint types of the service requests. Largest service requests are for heating, followed by street light conditions, street conditions, general constructions and plumbing Top 5 Agencies Receiving most requests: According to the visualization, the most complaint requests are made to HPD, which stands for "Housing Preservation and development", followed by "department of transportation"(DOT), NYPD (New York Police department), DEP(Department of Environmental Protection Police) and DSNY (Department Of Sanitation).

Then we considered each agency separately to see, what types of complaints are registered for every agency. HPD has most complaints about: Heating, General Construction, Plumbing, Heat and paint. NYPD has

most complaints for: Blocked Driveway, Illegal Parking, Noise Construction, Noise Street, and Noise vehicle. Note: The plot shows the abbreviated names of the complaint types, so that it is easily readable by the audience.

```
p <- (data.frame(nyc311$Complaint.Type[nyc311$Complaint.Type != "" &
                                       nyc311$Complaint.Type != " " &
                                       nyc311$Complaint.Type != "N/A" ]))
a <- as.data.frame(table(p))
newdata <- head(a[order(-a$Freq),],5 )
vec <- newdata$p
vec <- substring(vec, first = 1, last = 10)
vec <- toupper(vec)
p<-ggplot(data=newdata, aes(x=reorder(vec, -newdata$Freq), y=newdata$Freq)) +
  geom_bar(stat="identity",color='steelblue',fill = 'steelblue') +
  xlab("Complaint Types of 311 Data") +theme(axis.text.x=element_text(angle=20))+
  ylab("Service Requests")

p2 <- (data.frame(nyc311$Agency[nyc311$Agency != "" & nyc311$Agency!= " " &
                                nyc311$Agency != "N/A" ]))
a <- as.data.frame(table(p2))
newdata2 <- head(a[order(-a$Freq),],5 )
vec2 <- newdata2$p
vec2 <- substring(vec2, first = 1, last = 10)
p2<-ggplot(data=newdata2, aes(x=reorder(vec2, -newdata2$Freq), y=newdata2$Freq)) +
  geom_bar(stat="identity",color='steelblue',fill = 'steelblue') +
  theme(axis.text.x=element_text(angle=20))+
  xlab("Agencies Receiving Most Requests") +
  ylab("Service Requests")

p1 <- (data.frame(nyc311$Complaint.Type[nyc311$Agency == "HPD"]))
a1 <- as.data.frame(table(p1))
newdata3 <- head(a1[order(-a1$Freq),],5 )
vec3 <- newdata3$p
vec3 <- substring(vec3, first = 1, last = 10)
vec3 <- toupper(vec3)
titlestr = paste("Complaints Types made to \"HPD\"")
p1<-ggplot(data=newdata3, aes(x=reorder(vec3, -newdata3$Freq), y=newdata3$Freq)) +
  geom_bar(stat="identity",color='steelblue',fill = 'steelblue') +
  theme(axis.text.x=element_text(angle=20))+
  xlab(toString(titlestr)) +
  ylab("Service Requests")

p3 <- (data.frame(nyc311$Complaint.Type[nyc311$Agency == "NYPD"]))
a1 <- as.data.frame(table(p3))
newdata4 <- head(a1[order(-a1$Freq),],5 )
vec4 <- newdata4$p
vec4 <- substring(vec4, first = 1, last = 10)
vec4 <- toupper(vec4)
titlestr = paste("Complaints Types made to \"NYPD\"")
p3<-ggplot(data=newdata4, aes(x=reorder(vec4, -newdata4$Freq), y=newdata4$Freq)) +
  geom_bar(stat="identity",color='steelblue',fill = 'steelblue') +
  theme(axis.text.x=element_text(angle=20))+
  xlab(toString(titlestr)) +
  ylab("Service Requests")
```

```
grid.arrange(p,  p1,p2,p3,ncol=2)
```

4) More on NYPD:

We as a team care about public safety. The department of police is the first thing that comes to our mind. We wanted to dig deep in the trends of complaint registered under NYPD. Areas of Complaint: We wanted to identify whether any Agency, in New York had a disproportionate amount of incident calls within a given timeframe. We retained the latitude/longitude data to be able to generate a geographical visualization.: **To visualize most frequent areas of complaint** We plotted a countmap for the service requests which were made to NYPD department over all the years. It is difficult to display leaflet image by rendering the r markdown to pdf. As a workaround, we saved the image as png file in the current folder and displaying it with help of library "png". Manhattan is the highest area of complaint followed by Brooklyn

```
library(leaflet)

center_lon = median(nyc311$Longitude,na.rm = TRUE)
center_lat = median(nyc311$Latitude,na.rm = TRUE)

NYC311NYPDSampleAll = nyc311_NYPD %>% sample_n(20000) %>%
  filter(!is.na(Latitude) ) %>%
  filter(!is.na(Longitude))
 NYC311NYPDSampleAll %>%
   leaflet() %>%
   addProviderTiles("Esri.NatGeoWorldMap")  %>%
   addTiles() %>%
```

```
    addMarkers(lng = ~Longitude, lat = ~Latitude,clusterOptions =
                 markerClusterOptions()) %>%
    setView(lng=center_lon, lat=center_lat,zoom = 12)
```

```
library("png")
pp <- readPNG("countmap.png")
plot.new()
rasterImage(pp,0,0,1,1)
```

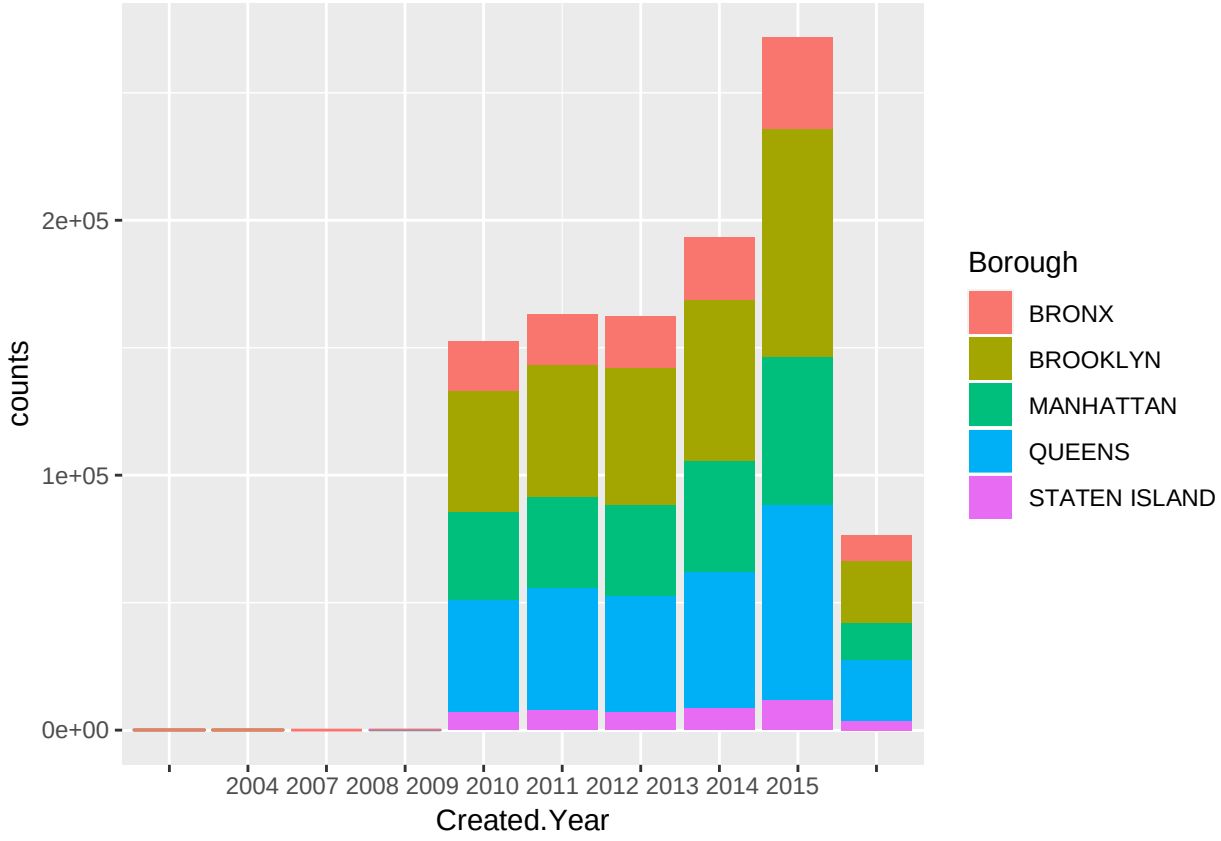
5) Yearly Trends of Requests made to NYPD:

We wanted to see how the Requests registered to NYPD Department changes with every year.
To Visualize this:
We grouped the data set on Created year and Borough and plotted the corresponding frequencies using ggplot, geom Plot.

Observation:
It can be observed that there are very few observations for NYPD before year 2010. This might be because, beginning in 2010, NYC launched an initiative to expose government data via NYC Open Data in an effort to improve the accessibility, transparency, and accountability of City government. After 2010, Most complaints were registered in the year 2014 and the highest were from the borough Brooklyn

```
nyc311_NYPD <- na.omit(nyc311_NYPD, cols = "Borough")
df <- nyc311_NYPD %>%
  group_by(Created.Year, Borough) %>%
  summarise(counts = n())
```

```
ggplot(df, aes(x = Created.Year, y = counts)) +
  geom_bar(
    aes(color = Borough, fill = Borough),
    stat = "identity", position = position_stack()
    )
```



6) Hourly Trends of requests

We wanted to see how the Requests registered to NYPD Department changes with hours of the day.

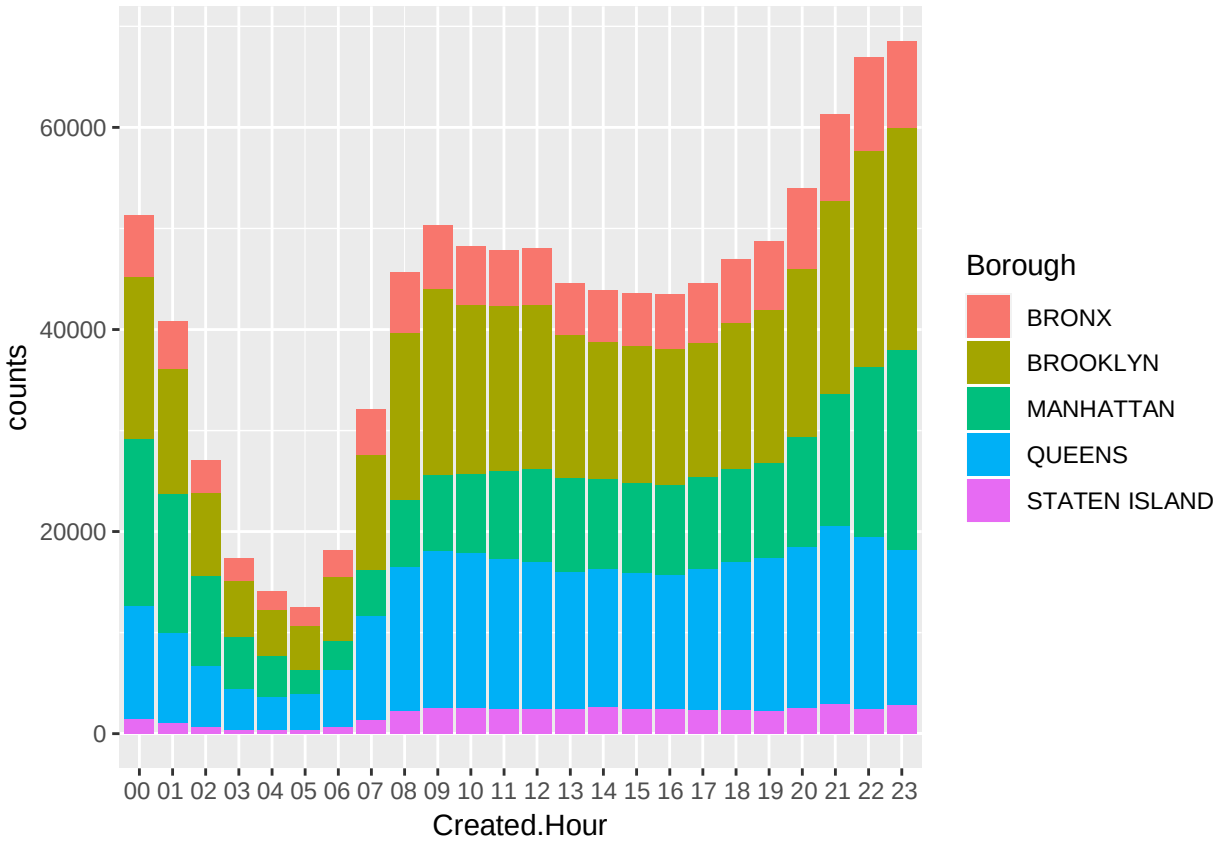
To Visulaize this:

We grouped the data set on Created Hour and Borough and plotted the corresponding frequencies using ggplot, geom Plot.

Observation:

It can be observed that most complaints are registered in late hours such as 9 - 12 PM. Least number of complaints are registered in early morning hours such as 3 - 6 AM

```
df <- nyc311_NYPD %>%
  group_by(Created.Hour, Borough) %>%
  summarise(counts = n())
ggplot(df, aes(x = Created.Hour, y = counts)) +
  geom_bar(
    aes(color = Borough, fill = Borough),
    stat = "identity", position = position_stack()
    )
```



7) Types of complaint per borough to NYPD for nyc311 dataset

Then we wanted to Visualize Most frequent Complaint Types per Borough.

To Visualize this:

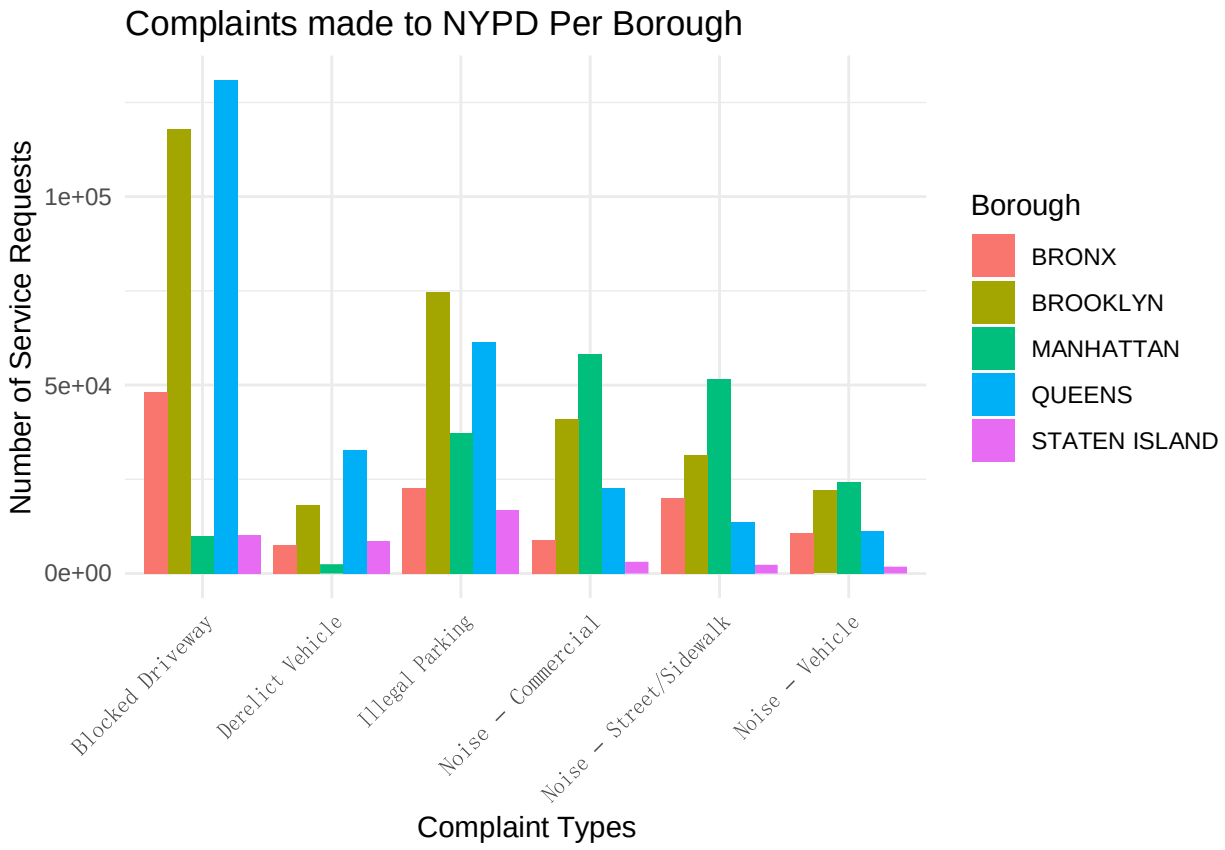
We grouped the complaint Types and stored the 6 most frequent complaint types from the dataset. Filtering by these 6 complaint types we got a new data set containing only those 6 type of complaints. Then we grouped on Borough and Complaint Type to achieve the following plot.

Observation:

It can be observed that most complaints are from Brooklyn and queens. Queens has the highest complaint regarding blocked highway and Derelict Vehicle. Brooklyn had highest number of complaints in Illegal Parking. Most complaints in Manhattan are for noise commercial followed by noise street/sidewalk and street parking. Which makes sense as Manhattan is one of the busiest city of the worlds.

```
nyc311_NYPD <-nyc311_NYPD[!(is.na(nyc311_NYPD$Complaint.Type) |
                            nyc311_NYPD$Complaint.Type==""), ]
exp_dt <- nyc311_NYPD%>%
  group_by(Complaint.Type)  %>%
  filter(!is.na(Complaint.Type))  %≥%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(Complaint.Type = reorder(Complaint.Type,Count)) %≥%
  arrange(desc(Count)) %>%
  head(6)

NYCFiltered <- filter(nyc311_NYPD, nyc311_NYPD$Complaint.Type %in%exp_dt$Complaint.Type)
```

```
NYCFiltered %>%
  group_by(Borough, Complaint.Type) %>%
  filter(!is.na(Complaint.Type))  %>%
  summarise(num_call = n()) %>%
  ungroup() %>%
  arrange(desc(num_call))   %>%

   ggplot(aes(y=num_call, x=Complaint.Type, fill = Borough))  +
  geom_histogram(stat="identity",position = 'dodge') +
  ggtitle("Complaints made to NYPD Per Borough") + xlab("Complaint Types") +
  ylab("Number of Service Requests") +
  theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



8) NYC_Crimes:

NYC_311 data set has all the non-emergency complaints. We wanted to visualize how the emergency requests made to NYPD varies with Boroughs.

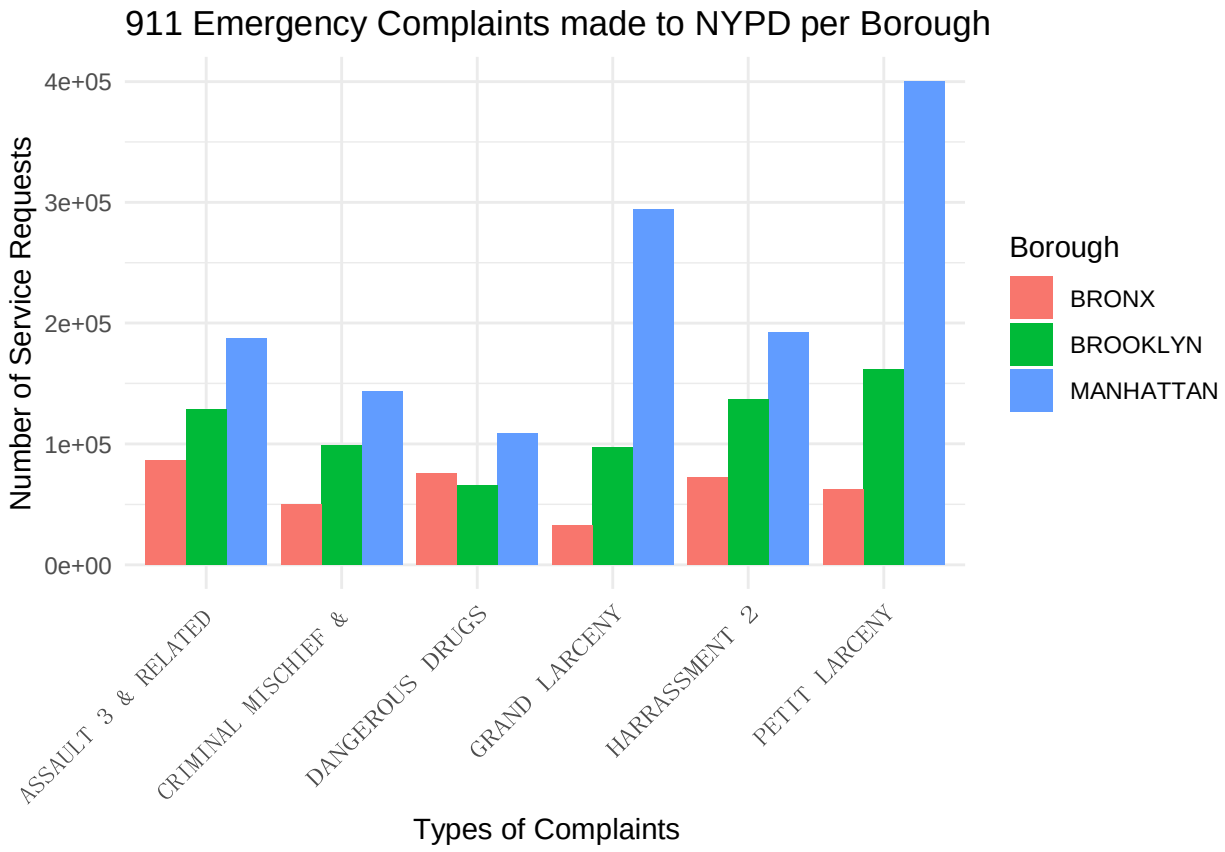
To Visualize this:

We picked a new data set called NYC_Crimes which has the records of all the emergency complaints made to NYPD. We joined the datasets using, Boroughs, Latitude and Longitude to see, what kind of emergency requests are made in the same locations as that of NYC_311 Dataset. We were interested in the most frequent offenses/Complaint Types Registered. We grouped the data by offenses per borough and filtered the data set for top 6 offenses. Then we plotted the filtered dataset for the boroughs which had highest complaint volume.

Observation:

It can be observed that most requests are for Petit Larceny from Manhattan. Petty larceny is the crime of theft of another's property or money under a statutorily defined value, in which the value is below the grand larceny limit. Similarly, Manhattan has the most complaints for grand Larceny as well, followed by harassments, Assaults, Criminal Mischief and Drugs. Complaint Trend is almost similar in Brooklyn except for Grand Larceny Complaint types. Whereas Bronx has different trends. The highest reported complaint in Bronx is assault, followed by harassment, dangerous drugs, petit larceny, criminal mischief and grand larceny.

```
join_nyc_crimes <-Merged_Data[!(is.na(Merged_Data$OFNS_DESC) |
                                Merged_Data$OFNS_DESC==""), ]
exp_dt <- join_nyc_crimes %>%
  group_by(OFNS_DESC) %>%
  filter(!is.na(OFNS_DESC)) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(OFNS_DESC = reorder(OFNS_DESC,Count)) %>%
  arrange(desc(Count)) %>%
  head(6)

NYCFiltered <- filter(join_nyc_crimes, join_nyc_crimes$OFNS_DESC %in% exp_dt$OFNS_DESC)


exp_dt <- NYCFiltered %>%
  group_by(Borough) %>%
  filter(!is.na(Borough)) %>%
  summarise(Count = n()) %>%
  ungroup() %>%
  mutate(Borough = reorder(Borough,Count)) %>%
  arrange(desc(Count)) %>%
  head(3)

NYCFiltered2 <- filter(NYCFiltered, NYCFiltered$Borough %in% exp_dt$Borough)

NYCFiltered2 %>%
  group_by(Borough, OFNS_DESC) %>%
  filter(!is.na(OFNS_DESC)) %>%
  summarise(num_call = n()) %>%
  ungroup() %>%
  arrange(desc(num_call)) %>%
  mutate(OFNS_DESC = substring(OFNS_DESC, first = 1, last = 20)) %>%
  mutate(OFNS_DESC = toupper(OFNS_DESC)) %>%


   ggplot(aes(y=num_call, x=OFNS_DESC, fill = Borough)) +
  geom_histogram(stat="identity",position = 'dodge') +
  ggtitle("911 Emergency Complaints made to NYPD per Borough") +
  xlab("Types of Complaints") + ylab("Number of Service Requests") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion:

THE ABOVE ARE JUST A FEW OF THE INSIGHTS OBTAINED WITHOUT ACTUALLY HAVING DEEP KNOWLEDGE OF THE WORKINGS OF NYC DEPARTMENTS.

We tried to dive in the complaints related to NYPD Agency. Someone who is actually working for any of the NYC agencies can better "mine" the data using this visualization to quickly determine: -what areas are experiencing what kinds of problems and -when do these occur more frequently. This in turn should help NYC authorities (the target audience) to optimize the allocation of it's resources by location, time and by expertise/skills needed.

References

1. https://mycourses.rit.edu/d2l/le/content/703420/viewContent/5270471/View?ou=703420
2. https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
3. https://www.kaggle.com/adamschroeder/crimes-new-york-city
4. Book - Wickham2017(R for data Science)

APPENDIX

Describing each column of the data (Data Dictionary)

Table 1: Data Dictionary

Column.Name	Description	Expected.Values	Notes
Unique Key	Unique identifier of a Service Request	Integer	This is NOT the Service Request (SR) # provided to the initiating customer. SR #s are not available in this data set.
Created Date	Creation date of the service request	Date in format MM/DD/YY	
Created Hour	Creation Hour of the service request	Hour in 24 hour format	
Closed Date	Closing date of the service request	Date in format MM/DD/YY HH:MM:SS AM/PM	
Agency	Acronym of responding City Government Agency	String	64
Complaint Type	This is the fist level of identifying the topic of the incident or condition.	String	245 distinct. Lowest = adopt a basket, agency issues, air quality, animal abuse. Highest - water system, window guard.All possible compaaint type in the link
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.	String	Lowest : 10 little basets
Location Type	Describes the type of location used of the address information	String	
Incident Zip	Incident location's zip code	5 digit Integer	highest is brooklyn : 11226
Street Name	Street name of incident address provided by the submitter	String	
Cross Street 1	First Cross street based on the geo validated incident location	String	
Cross Street 2	Second Cross Street based on the geo validated incident location	String	

Table 1: Data Dictionary *(continued)*

Column.Name	Description	Expected.Values	Notes
Intersection Street 1	First intersecting street based on geo validated incident location	String	
Intersection Street 2	Second intersecting street based on geo validated incident location	String	
Address Type	Type of incident location information available.	String Values: Intersection; LatLong; Placename; Address; Block face	6 distinct values , Address; Block face; Intersection; LatLong; Placename
City	City of the incident location provided by geovalidation.	String	Brooklyn highest
Facility Type	If available, this field describes the type of city facility associated to the SR	String	6 distinct values.1 Precinct 12 N/A 22 137 DSNY Garage 439 School 5527008 School District
Status	Status of SR submitted	String	22 distinct values like, closed, open, pending, Assigned, Email sent, started, Unassigned, to be routed, closed-testing , closed - In person, Closed - by phone...etc.
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal Service Level Agreements (SLAs).	Date in format MM/DD/YY HH:MM:SS AM/PM	
Resolution Action Updated Date	Date when responding agency last updated the SR.	Date in format MM/DD/YY HH:MM:SS AM/PM	
Community Board	Provided by geovalidation.	String	Community Boards are the foundation of democratic, community-based planning in New York City.

Table 1: Data Dictionary *(continued)*

Column.Name	Description	Expected.Values	Notes
Borough	Provided by the submitter and confirmed by geovalidation.	String. New York City encompasses five county-level administrative divisions called boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island.	String. New York City encompasses five county-level administrative divisions called boroughs: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island.
X Coordinate (State Plane)	X coordinate of the incident location.	Integer	
Y Coordinate (State Plane)	Y coordinate of the incident location.	Integer	
Park Facility Name	Name of the Park Facility of the incident location is Parks Dept Facility	String	if it is a park or not like . Unspecified :9067682 Central Park : 2215 Riverside Park : 991 Prospect Park : 763 Rockaway Beach Boardwalk : 651 Flushing Meadows Corona Park: 600 (Other) : 52035 . If not then unspecified.
School Name	Name of the School, if the incident took place in Dept. of Education.	String	Sometimes Park values appear.Possible values: 101 Street Soccer Field 107th Street Pier 115th Street Playground 146 St Playground
School Region	Region in which the school is located if the incident occurred in school	String	
School City	City of the facility of the incident	String	
School State	State of the facility of the incident	possible values : NY : 57255 Unspecified : 9067682	
School Zip	Zipcode of the School of the incident.	5 digit Integer	
School or Citywide Complaint	This field is for a school or citywide issue if the incident is realted to a school.	Y; N; BLANK	

Table 1: Data Dictionary *(continued)*

Column.Name	Description	Expected.Values	Notes
Vehicle Type	Specifies the type of TLC vehicle, if the incident occurred in taxi	Car Service; Commuter Van; Green Taxi	
Taxi Company Borough	Displays the borough of the taxi company.	String	
Taxi Pick Up Location	Shows the taxi pickup location as taxi, if the inciddent took place in taxi	Grand Central Station; Intersection; JKF Airport; La Guardia Airport; New York-Penn Station; Other; Port Authority Bus Terminal	
Bridge Highway Name	Name of the Bridge Highway, when the incient took place on bridge Highway	String	
Bridge Highway Direction	Direction of bridge/highway of incident's location	String	
Road Ramp	Tells if the incident took place on road or the road ramp	Roadway; Ramp	
Garage Lot Name	Shows garage, in which the meter is located	String	
Ferry Direction	Indicates the direction of the Ferry, used when the incident location is within a Ferry	Manhattan Bound; Staten Island Bound	
Ferry Terminal Name	Ferry Terminal where the incident took place, used when the incident took place at ferry location	String	
Latitude	Latitude of the incident location	Numeric	
Longitude	Longitude of the incident location	Numeric	
Location	Latitude & Longitude of the incident location	Numeric	
CMPLNT_NUM	Randomly generated ID for each complaint	Numeric	
COMPLAINT FROM DATE	Exact date of occurrence for the reported event (or starting date of occurrence, if COM-PLAINT.TO.DATE exists)	Date MM/DD/YYYY	

Table 1: Data Dictionary *(continued)*

Column.Name	Description	Expected.Values	Notes
COMPLAINT FROM TIME	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)	Time HH:MM:SS AM/PM	
COMPLAINT TO DATE	Ending date of occurrence for the reported event, if exact time of occurrence is unknown	Date MM/DD/YYYY	
COMPLAINT TO TIME	Ending time of occurrence for the reported event, if exact time of occurrence is unknown	Time HH:MM:SS AM/PM	
RPT_DT	Date event was reported to police	Date MM/DD/YYYY	
KY_CD	Three digit offense classification code	Integer	
OFNS_DESC	Description of offense corresponding with key code	String	
PD_CD	Three digit internal classification code (more granular than Key Code)	Integer	
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)	String	
CRM ATPT CPTD CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely	String	
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation	String	
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.	String	
ADDR PCT CD	The precinct in which the incident occurred	Integer	

Table 1: Data Dictionary *(continued)*

Column.Name	Description	Expected.Values	Notes
LOC OF OCCUR DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of	String	
PREM TYP DESC	Specific description of premises; grocery store, residence, street, etc.	String	
PARKS NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)	String	