homework vi

Parthivi Shrivastava, Khavya Seshadri

2019-10-11

Contents

INTRODUCTION

In this report, we are performing explorations on the following datasets: 311 NYC Service call requests and NYC Crimes data. 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 2003 to 2015. It contains around 243 complaint types been reported to 311. The relatable dataset which we chose was NYPD NYC crimes data. We took a sample of size 95,593 from the original data source which was around 5.5 million. This data contains three major categories of crime: Felony, Violation and Misdemeanor. Each record corresponds to the crime information being reported in New York city.

NYC 311 data

Initialization

Here we load the required packages 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
```

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

```
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
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

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

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year

## The following object is masked from 'package:base':
##
##     date
```

```r
nyc311<-fread("311_Service_Requests_from_2010_to_Present.csv",
              na.strings = c("","NA","N/A"))
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 nyc311 dataset.

```r
nyc311 <- nyc311[,c(-1,-10:-19,-23, -25:-49)]
nyc311nodups <- distinct(nyc311)
names(nyc311nodups)
```

```
##  [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"
```

Handling missing values

In the following snippet, we have handled the missing values and the infelicities in the columns of the data. Intially, we replaced the invalid zip codes with NA. The criteria we used to ensure the validity of the zip code in the data are as follows: 1. Zipcode length should be 5 or 10. 2. If the zipcode length is 10, then it should satisfy the format of xxxxx-xxxx. Apart from the above rules, we also found zipcodes like 00000, 10000 which were invalid, hence replaced them with NA. Considering the closed date column, we had dates that were defaulted to 01/01/1900 and also there were around 100K records with closed date lesser than the created date, which seems to be invalid and hence we replaced them with NA. For borough, there were around 800K records with unspecified values, out of which 600K had valid zip codes, so we found the boroughs for those records using the valid zipcode information and remaining we filled with NA.

```r
# Replacing invalid zipcodes with NA
nyc311nodups[Incident.Zip=="00000" | (str_length(str_trim(Incident.Zip))<5 |
        (str_length(str_trim(Incident.Zip)) > 5 &
           str_length(str_trim(Incident.Zip)) < 10)  |
          Incident.Zip=="10000","Incident.Zip"] <- NA

nyc311nodups[as.Date(nyc311nodups$Closed.Date, format="%m/%d/%Y")==
                as.Date("01/01/1900", format="%m/%d/%Y") |
                as.Date(nyc311nodups$Closed.Date, format="%m/%d/%Y")<
                  as.Date(nyc311nodups$Created.Date, format="%m/%d/%Y"),
             c("Closed.Date") ] <- NA

unspecifiedBro <- nyc311nodups %>%
  select(Incident.Zip, Borough) %>%
  filter(Borough=="Unspecified" & !is.na(Incident.Zip))

zipCodeTable <- nyc311nodups %>%
  select(Incident.Zip, Borough) %>%
  filter(Borough!="Unspecified" & (str_length(str_trim(Incident.Zip))==5 |
   (str_length(str_trim(Incident.Zip))==10 & (str_detect(Incident.Zip,'-')))))
zipCodeTable <- distinct(zipCodeTable)
zipCodeTable <-  zipCodeTable %>%
 group_by(Incident.Zip) %>%
 summarize(Borough = first(Borough))
joinedTab <- merge(x=unspecifiedBro, y=zipCodeTable, by = "Incident.Zip", all.x = TRUE)
joinedTab <- distinct(joinedTab)
colnames(joinedTab)[colnames(joinedTab)=="Borough.x"] <- "Borough"

nyc311nodups <- merge(x=nyc311nodups, y=joinedTab,
                  by=c("Incident.Zip", "Borough"), sort=FALSE, all.x = TRUE)
nyc311nodups[!is.na(Borough.y), "Borough"] <- nyc311nodups[!is.na(Borough.y), "Borough.y"]
nyc311nodups[Borough=="Unspecified", "Borough"] <-
  nyc311nodups[Borough=="Unspecified", "Borough.y"]
# drop the borough.y
nyc311nodups <- nyc311nodups[,-"Borough.y"]
head(nyc311nodups)
```

```
##    Incident.Zip   Borough           Created.Date            Closed.Date
## 1:        10465     BRONX 04/14/2015 02:14:40 AM 04/14/2015 03:03:22 AM
## 2:        11234  BROOKLYN 04/14/2015 02:10:12 AM                   <NA>
## 3:        11204  BROOKLYN 04/14/2015 02:03:01 AM                   <NA>
## 4:        11211  BROOKLYN 04/14/2015 02:02:40 AM                   <NA>
## 5:        10025 MANHATTAN 04/14/2015 02:00:04 AM 04/14/2015 02:47:33 AM
## 6:        11205  BROOKLYN 04/14/2015 01:52:15 AM 04/14/2015 02:11:10 AM
##    Agency                     Agency.Name          Complaint.Type
## 1:   NYPD New York City Police Department                 Vending
## 2:   NYPD New York City Police Department        Blocked Driveway
## 3:   NYPD New York City Police Department Noise - Street/Sidewalk
## 4:   NYPD New York City Police Department Noise - Street/Sidewalk
## 5:   NYPD New York City Police Department Noise - Street/Sidewalk
## 6:   NYPD New York City Police Department Noise - Street/Sidewalk
##            Descriptor   Location.Type   Status               Due.Date
## 1: In Prohibited Area Street/Sidewalk   Closed 04/14/2015 10:14:40 AM
## 2:          No Access Street/Sidewalk     Open 04/14/2015 10:10:12 AM
## 3:   Loud Music/Party Street/Sidewalk     Open 04/14/2015 10:03:01 AM
## 4:       Loud Talking Street/Sidewalk Assigned 04/14/2015 10:02:40 AM
## 5:       Loud Talking Street/Sidewalk   Closed 04/14/2015 10:00:04 AM
## 6:       Loud Talking Street/Sidewalk   Closed 04/14/2015 09:52:15 AM
##    Resolution.Action.Updated.Date Latitude Longitude
## 1:         04/14/2015 03:03:05 AM 40.82573 -73.82111
## 2:                           <NA> 40.61879 -73.93771
## 3:                           <NA> 40.61859 -73.99846
## 4:         04/14/2015 02:10:32 AM 40.71410 -73.95589
## 5:         04/14/2015 02:04:59 AM 40.79792 -73.96385
## 6:         04/14/2015 02:11:10 AM 40.68833 -73.96481
##                                     Location
## 1:   (40.8257259931145, -73.82111429330192)
## 2: (40.618794391821936, -73.93770589155426)
## 3:  (40.61859442131066, -73.99845832101916)
## 4:  (40.71409874640673, -73.95589458206499)
## 5:  (40.79791780509379, -73.96384631347463)
## 6:  (40.68832571866554, -73.96481079590191)
```

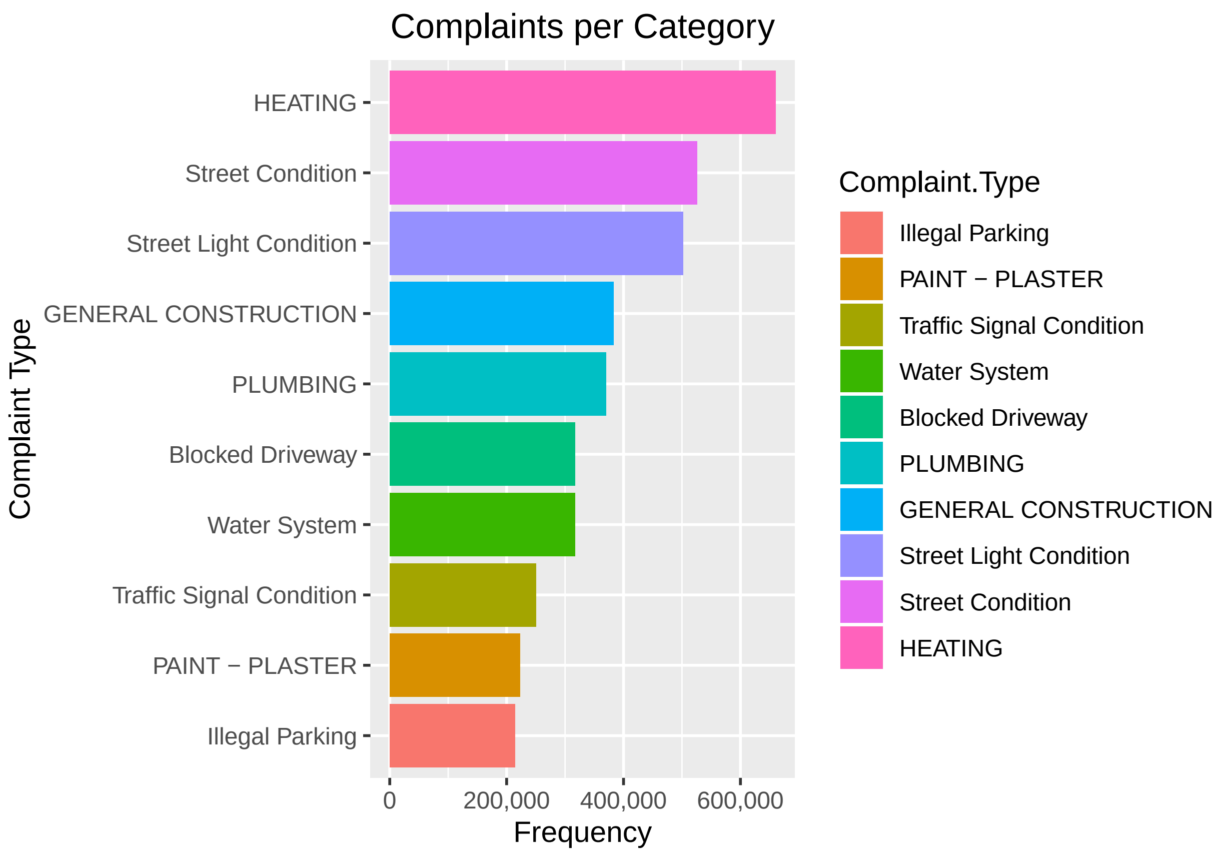
Nyc311 Exploration

The following horizontal bar chart shows the top 10 complaint types received, with the color specified for each complaint type. We see that the top complaints received in NYC are Heating, Street Condition, Street Light Condition, etc.

```
topComplaints <- nyc311nodups %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>%
  arrange(desc(count)) %>%
  top_n(10)
```

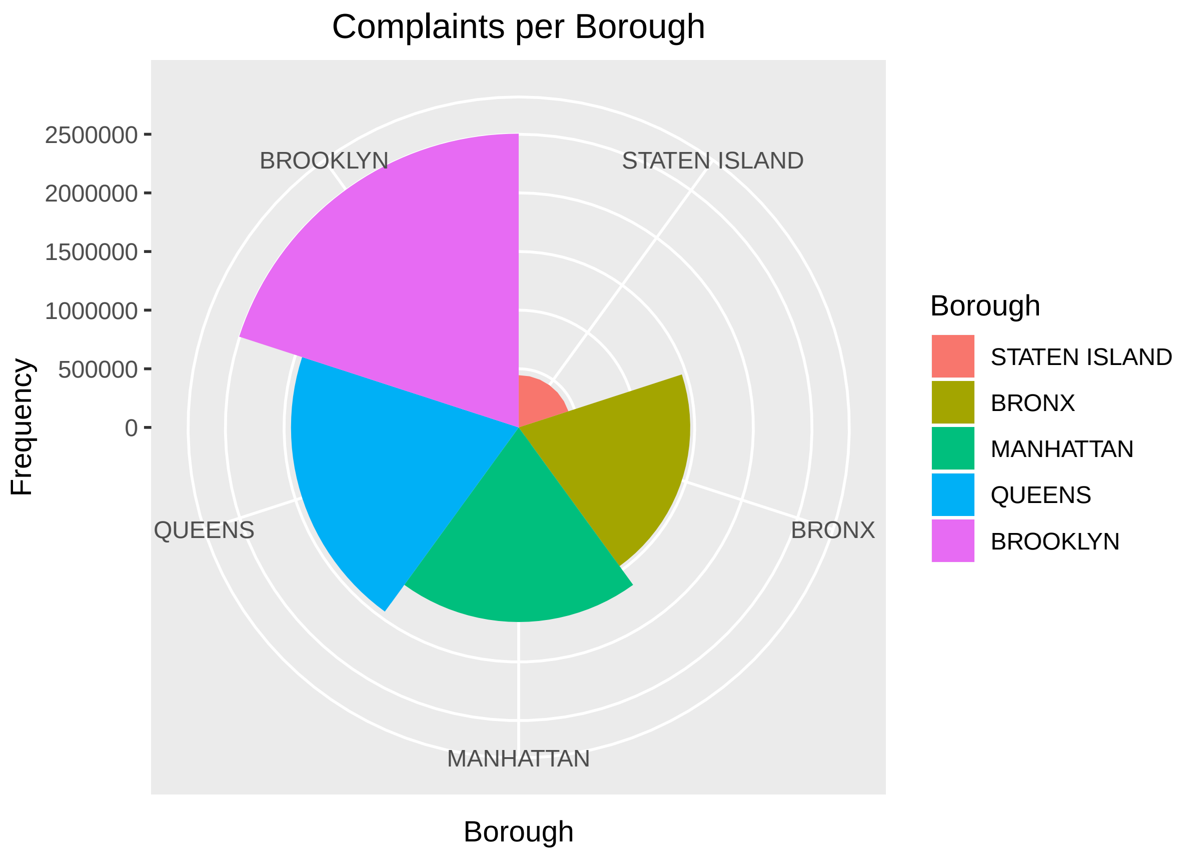
```
## Selecting by count
```

```
topComplaints$Complaint.Type<-factor(topComplaints$Complaint.Type,
  levels=topComplaints$Complaint.Type[order(topComplaints$count)])

(ggplot(topComplaints,aes(x=Complaint.Type,y=count, fill=Complaint.Type)) +
   geom_bar(stat="identity") +
   coord_flip() +
  scale_y_continuous(breaks = seq(0,700000,by = 200000), labels = comma)+
    xlab("Complaint Type") +
    ylab("Frequency") +
   ggtitle("Complaints per Category")+
     theme(plot.title = element_text(hjust = 0.5)))
```



The following coxcomb shows the boroughs that received the most service call requests.

```
boroughs <- nyc311nodups %>%
  filter(!is.na(Borough))%>%
  group_by(Borough) %>%
  summarize(count=n())
boroughs$Borough<-factor(boroughs$Borough,
  levels=boroughs$Borough[order(boroughs$count)])

(ggplot(boroughs,aes(x=Borough,y=count, fill=Borough)) +
   geom_bar(stat="identity", width=1) +
   theme(aspect.ratio = 1) +
   coord_polar() +
    ylab("Frequency") +
   ggtitle("Complaints per Borough") +
  theme(plot.title = element_text(hjust = 0.5)))
```

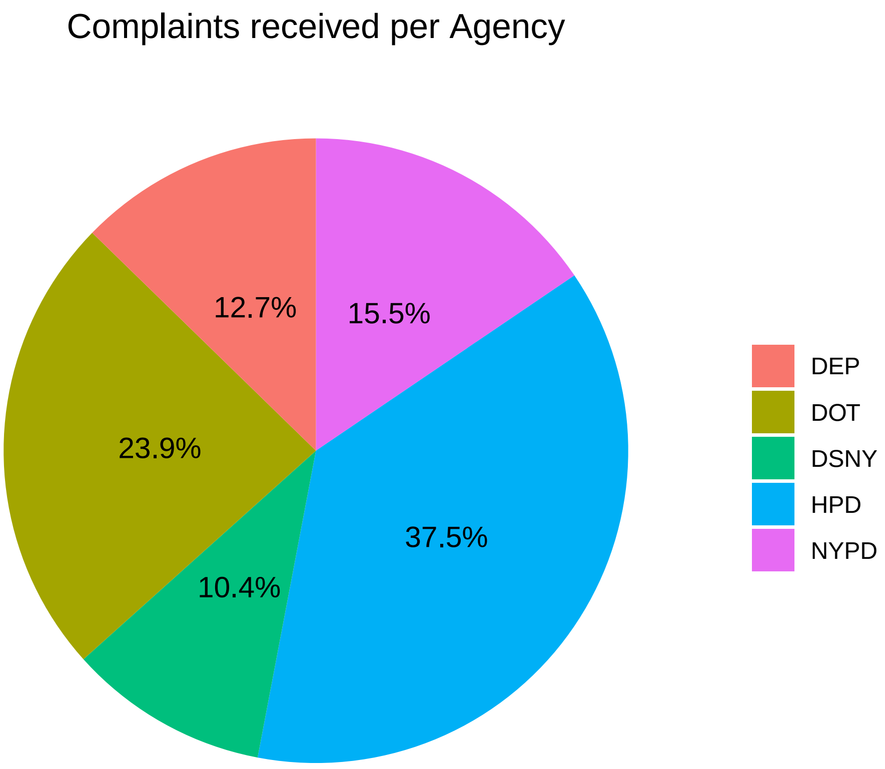


The following pie chart shows the top 5 agencies, which recieved the most complaints.

```
bigAgency <- nyc311nodups %>%
  group_by(Agency) %>%
  summarize(count=n()) %>%
  arrange(desc(count)) %>%
  top_n(5)
```

```
## Selecting by count
```

```
(ggplot(bigAgency, aes(x="", y=count, fill=Agency)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0) +
  geom_text(aes(label = paste0(round(count / sum(count) * 100, 1),"%")),
position = position_stack(vjust = 0.5)) +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Complaints received per Agency") +
  theme_classic() + theme(axis.line = element_blank(),
      axis.text = element_blank(),
      axis.ticks = element_blank(),
      plot.title = element_text(hjust = 0.5, color = "#000000")))
```

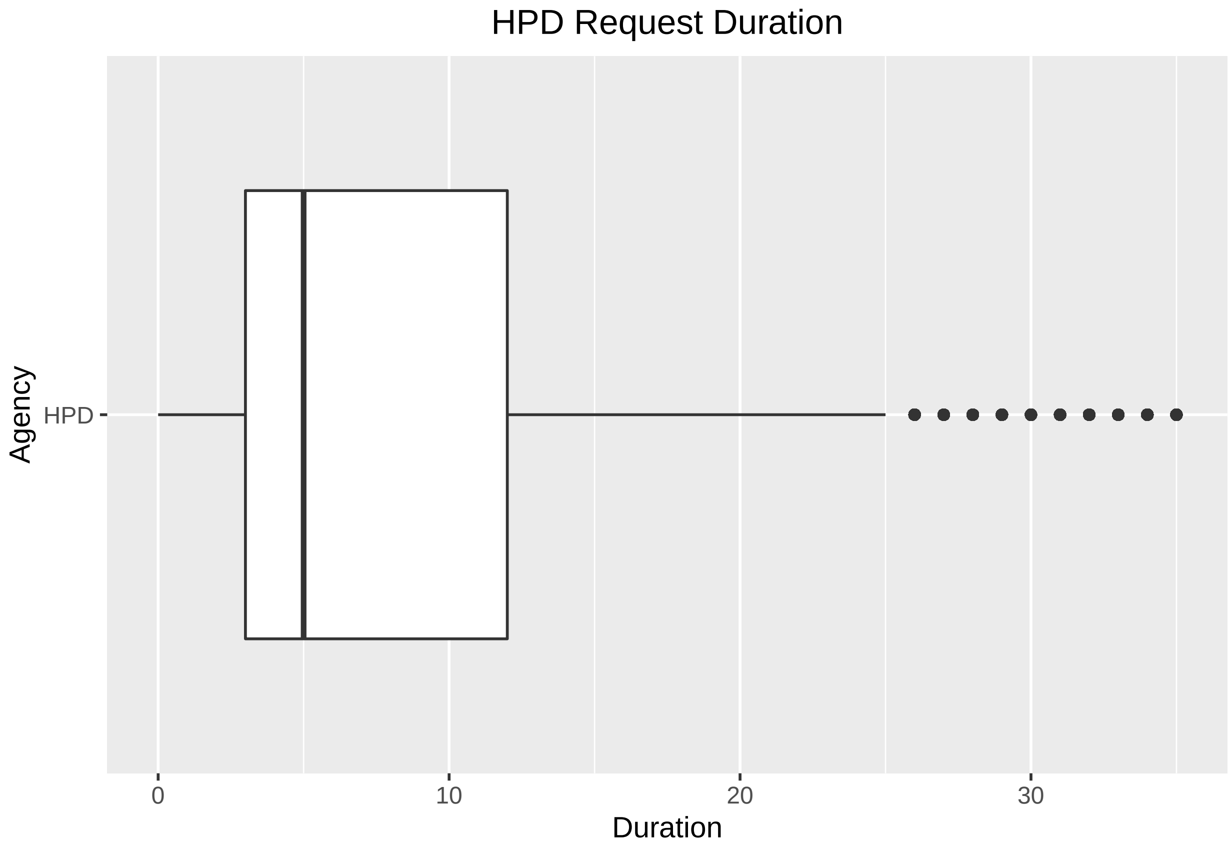


The table information shows the average time taken by the top three agencies. The number of days taken to resolve a complaint are computed using the created date and closed date. From the above, we see that HPD has received the most complaints, so dive deep into exploring the request duration of HPD in resolving the complaints.

```r
resolveComplaints <- nyc311nodups %>%
  select(Complaint.Type,
     Created.Date,
     Closed.Date,
     Due.Date,
     Agency,
     Borough)
filteredData <-dplyr::filter(resolveComplaints,
              (!is.na(Closed.Date)))
numOfDays <- (as.Date(filteredData$Closed.Date, format="%m/%d/%Y")-
              as.Date(filteredData$Created.Date, format="%m/%d/%Y"))

filteredData <- data.frame(filteredData,numOfDays)
slowAgency <- filteredData %>%
  group_by(Agency) %>%
  summarize(averageTime = as.integer(mean(numOfDays)))
slowAgency <- slowAgency[order(-slowAgency$averageTime),]


topAgencies <- dplyr::filter(slowAgency, Agency=='HPD'|Agency=='DOT'|Agency=='NYPD')
topAgencies
```

```
## # A tibble: 3 x 2
##   Agency averageTime
##   <chr>        <int>
## 1 HPD             10
## 2 DOT              8
## 3 NYPD             0
```

```
hpdComplaints <- dplyr::filter(filteredData, (Agency=="HPD"))
duration <- as.Date(hpdComplaints$Closed.Date, format="%m/%d/%Y") -
  as.Date(hpdComplaints$Created.Date, format="%m/%d/%Y")

(ggplot(hpdComplaints, aes(x=Agency, y=duration)) +
         geom_boxplot() + ylim(0,35) +
         ylab("Duration") +
         ggtitle("HPD Request Duration") +
         theme(plot.title = element_text(hjust = 0.5))+
  coord_flip())
```

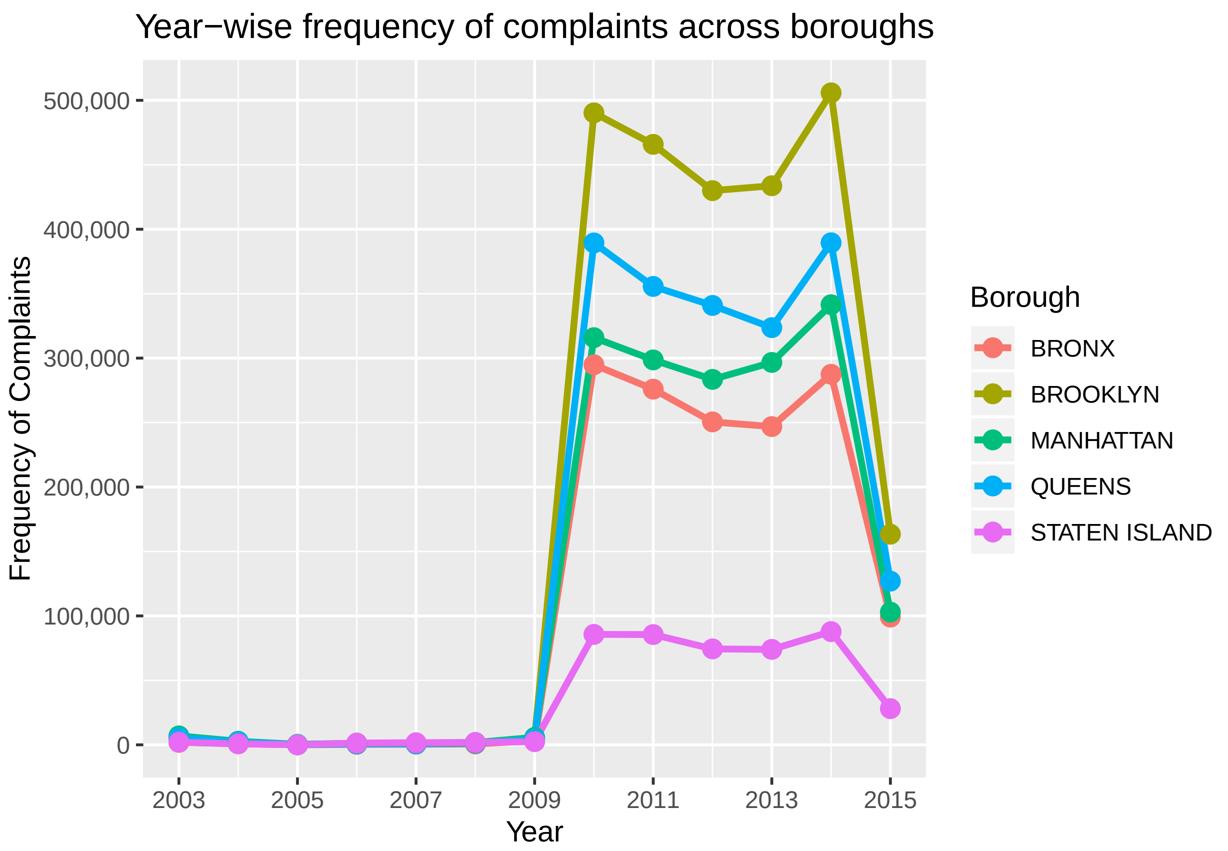


The following line graph shows the year-wise frequency of complaints accross the boroughs. We see a similar pattern across all the boroughs with respect to the increase/decrease in frequency over the years. Although, we don't have population statistics for NYC boroughs, we researched on that and we see the decreasing order with respect to population numbers are as follows: Brooklyn Queens Manhattan Bronx Staten Island We find the same decreasing order of boroughs with respect to frequency of complaints, with the highest being Brooklyn and the lowest being Staten Island.

```
boroughYear <-nyc311nodups %>%
  select( Borough , Created.Date, Complaint.Type) %>%
  filter(!is.na(Borough))
yearData <- separate(boroughYear, Created.Date, into=c("month", "day", "year"),
                     convert = T)

boroughYear <- yearData %>%
  group_by(year, Borough) %>%
  summarize(frequency=n())
(yearSpread <- boroughYear %>%
  spread(key=year, value=frequency))
```

```
## # A tibble: 5 x 14
##   Borough `2003` `2004` `2005` `2006` `2007` `2008` `2009` `2010` `2011`
##   <chr>    <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1 BRONX     1907    808      7    374    434    631   3198 294858 275932
## 2 BROOKL~   5391   2186     63    839    942   1219   5188 490283 465870
## 3 MANHAT~   6911   2744    393   1239   1251   1744   5755 315889 298611
## 4 QUEENS    5336   2314     47    696    792   1327   4331 389379 355607
## 5 STATEN~   2015    761      2   1373   1621   1855   2432  85656  85533
## # ... with 4 more variables: `2012` <int>, `2013` <int>, `2014` <int>,
## #   `2015` <int>
```

```
(ggplot(data=boroughYear, aes(x=year, y=frequency, group=Borough)) +
    scale_x_continuous(breaks = seq(2003,2015,by = 2)) +
    scale_y_continuous(breaks = seq(0,700000,by = 100000),labels = comma)+
  geom_line(linetype="solid", size=1.2, aes(color = Borough))+
  geom_point(aes(color = Borough), size=3)+
    xlab("Year")+
    ylab("Frequency of Complaints")+
    ggtitle("Year-wise frequency of complaints across boroughs")+
    theme(plot.title = element_text(hjust = 0.5)))
```



In the following, we are showing the year-wise breakdown of the top 5 complaints: general construction, heating, plumbing, street condition, and street light condition.

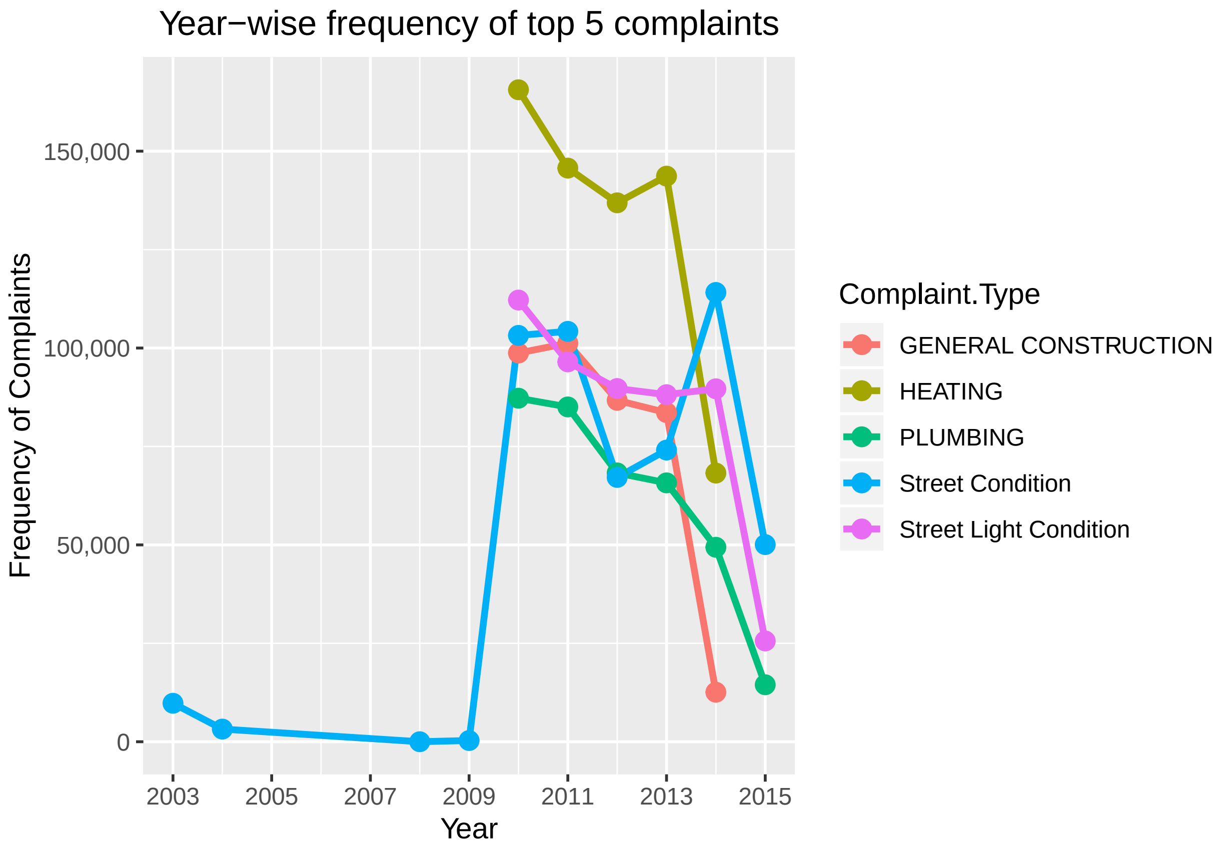
```
topComplaints <- nyc311nodups %>%
  group_by(Complaint.Type) %>%
  summarize(count=n()) %>%
  arrange(desc(count))%>%
  top_n(5)
```

```
## Selecting by count
```

```
complaintYear <-nyc311nodups %>%
  select( Created.Date, Complaint.Type)

complaintYear <- separate(complaintYear,
                Created.Date, into=c("month", "day", "year"), convert = T)
complaints <- complaintYear %>%
    filter(Complaint.Type %in% topComplaints$Complaint.Type) %>%
  group_by(Complaint.Type,year) %>%
  summarize(frequency=n())
(complSpread <- complaints %>%
  spread(key=year, value=frequency))
```

```
## # A tibble: 5 x 11
## # Groups:   Complaint.Type [5]
##   Complaint.Type `2003` `2004` `2008` `2009` `2010` `2011` `2012` `2013`
##   <chr>           <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1 GENERAL CONST~     NA     NA     NA     NA  98732 101220  86710  83599
## 2 HEATING            NA     NA     NA     NA 165604 145707 136887 143665
## 3 PLUMBING           NA     NA     NA     NA  87257  85040  68276  65755
## 4 Street Condit~   9770   3214      2    308 103212 104241  67132  74086
## 5 Street Light ~     NA     NA     NA     NA 112189  96480  89715  88161
## # ... with 2 more variables: `2014` <int>, `2015` <int>
```

```
(ggplot(data=complaints, aes(x=year, y=frequency, group=Complaint.Type)) +
    scale_x_continuous(breaks = seq(2003,2015,by = 2)) +
    scale_y_continuous(breaks = seq(0,300000,by = 50000),labels = comma)+
  geom_line(linetype="solid", size=1.2, aes(color = Complaint.Type))+
  geom_point(aes(color = Complaint.Type), size=3)+
    xlab("Year")+
    ylab("Frequency of Complaints")+
    ggtitle("Year-wise frequency of top 5 complaints")+
    theme(plot.title = element_text(hjust = 0.5)))
```



NYPD NYC Crimes data

We chose NYC Crimes data as the relatable dataset, because we found complaint types reported in 311NYC data can be categorized into the crime types such as felony, misdemeanor and violation.

Initialization

Here we load Crimes data set from the link as provided below and we fill the empty cells with NA.

```
nycCrimes <-
  fread("https://raw.githubusercontent.com/jamesjynus/Shiny/master/data/crime.csv",
                  na.strings = c("","NA"))
```

Data pre-processing

Here, we removed the irrelevant columns and duplicate records in the data, fixed the column names and displaying the head of the crimes data.

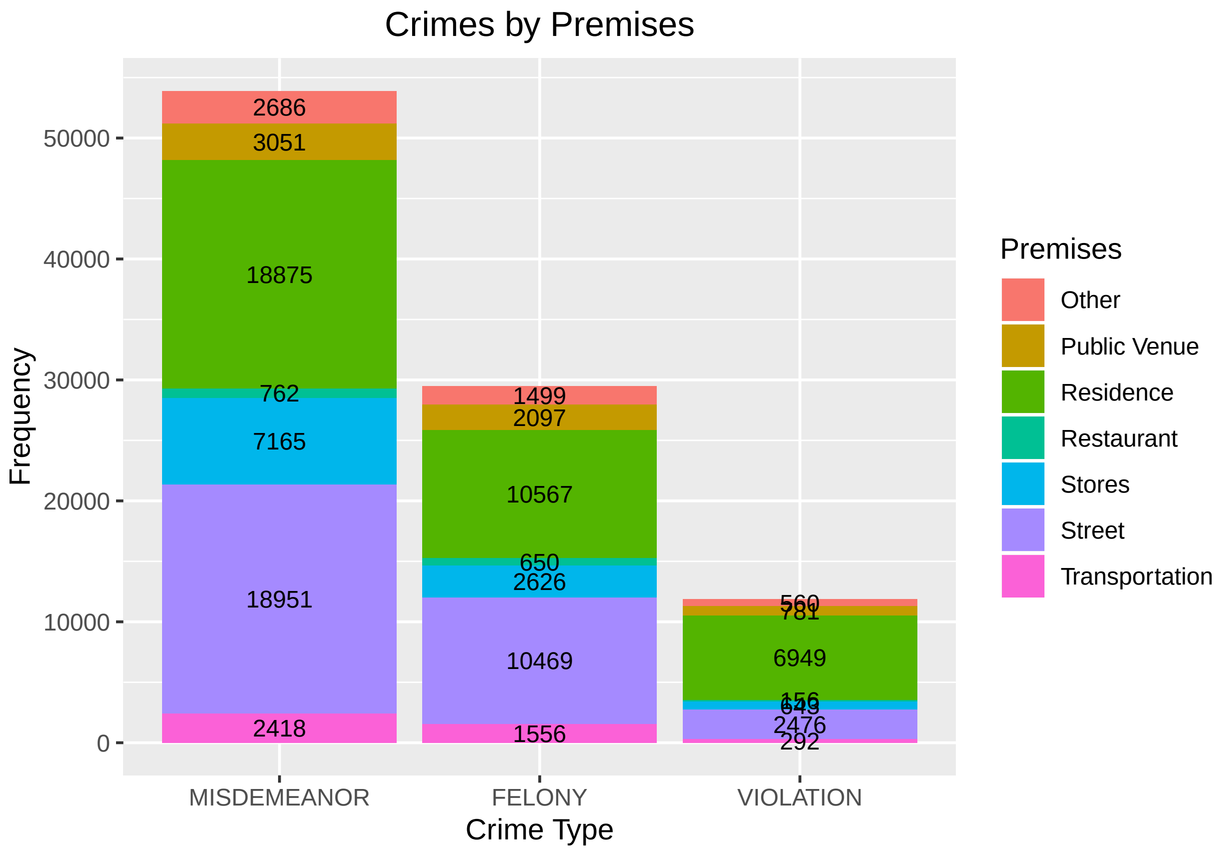
```
nycCrimes <- nycCrimes[,c(-1,-2,-13,-14,-15,-17)]
nycCrimenodups <- distinct(nycCrimes)
colnames(nycCrimenodups)[colnames(nycCrimenodups)=="Boro"] <- "Borough"
nycCrimenodups <-  nycCrimenodups[str_trim(Offense)!="",]
head(nycCrimenodups)
```

```
##          Date     Time Code                        Offense    Status
## 1: 2006-03-10 14:30:00  113                        FORGERY COMPLETED
## 2: 2012-12-19 10:00:00  344   ASSAULT 3 & RELATED OFFENSES COMPLETED
## 3: 2011-10-14 14:20:00  126        MISCELLANEOUS PENAL LAW COMPLETED
## 4: 2009-07-31 11:50:00  109                  GRAND LARCENY ATTEMPTED
## 5: 2006-01-23 17:45:00  341                  PETIT LARCENY COMPLETED
## 6: 2013-09-09 21:47:00  359 OFFENSES AGAINST PUBLIC ADMINI COMPLETED
##           Type       Borough       Premises Latitude Longitude Population
## 1:      FELONY      BROOKLYN         Street 40.66200 -73.91959    2465690
## 2: MISDEMEANOR STATEN ISLAND      Residence 40.57112 -74.09007     471000
## 3:      FELONY     MANHATTAN      Residence 40.79967 -73.94720    1595517
## 4:      FELONY        QUEENS   Public Venue 40.76480 -73.77161    2230000
## 5: MISDEMEANOR     MANHATTAN Transportation 40.77365 -73.95986    1566766
## 6: MISDEMEANOR         BRONX         Street 40.81937 -73.91828    1420414
##    Year_Month_New
## 1:        2006-03
## 2:        2012-12
## 3:        2011-10
## 4:        2009-07
## 5:        2006-01
## 6:        2013-09
```

NYPD NYC Crimes Exploration

Here, we are exploring the frequency of the following crime types: Felony, Misdemeanor, Violation. The bar chart also shows the amount of crimes happening with respect to premises like residence, restaurants, etc. depecited using the color for each Premises. We see that misdemeanor which could be petty theft, assault, intoxication, etc. has been majoring compared to other crime types and is frequently found to occur on the streets and residence(premises type).

```
crimesData <- nycCrimenodups %>%
  group_by(Type, Premises) %>%
  summarize(frequency=n()) %>%
  arrange(desc(frequency))

(ggplot(crimesData, aes(x=reorder(Type,-frequency), y=frequency, fill=Premises, label=frequency)) +
    scale_y_continuous(breaks = seq(0,60000, by=10000)) +
  geom_bar(stat ="identity") +
    xlab("Crime Type") +
    ylab("Frequency") +
  ggtitle("Crimes by Premises") +
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  theme(plot.title = element_text(hjust = 0.5)))
```



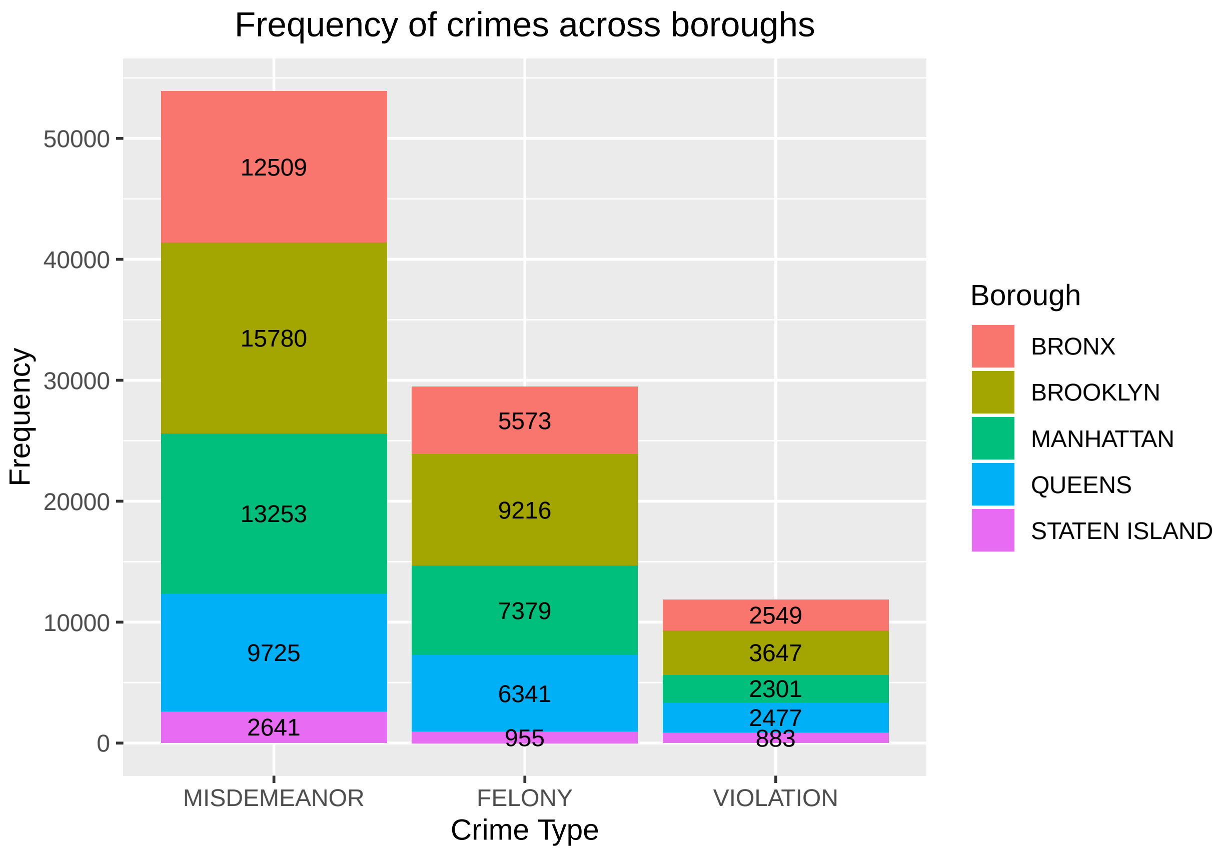
In the following snippet, we are computing the frequency of crimes across every borough with respect to every crime type, by spreading on the borough column. From the previous section as indicated with respect to the population numbers, Brooklyn being the most populated borough, we also see that it's been majoring in the number of crimes reported compared to other boroughs.

```
subsetData <- select(nycCrimenodups, Type, Borough)
subsetData <- subsetData %>%
    filter(!is.na(Borough)) %>%
  group_by(Type,Borough) %>%
  summarize(count=n()) %>%
  arrange(desc(count))

boroughSpread <- subsetData %>%
  spread(key=Borough, value=count)
```

```r
boroughSpread[is.na(boroughSpread)] <- 0
boroughSpread
```

```
## # A tibble: 3 x 6
## # Groups:   Type [3]
##   Type        BRONX BROOKLYN MANHATTAN QUEENS `STATEN ISLAND`
##   <chr>       <int>    <int>     <int>  <int>           <int>
## 1 FELONY       5573     9216      7379   6341             955
## 2 MISDEMEANOR 12509    15780     13253   9725            2641
## 3 VIOLATION    2549     3647      2301   2477             883
```

```r
(ggplot(subsetData, aes(x=reorder(Type, -count),y=count, fill=Borough, label=count)) +
    scale_y_continuous(breaks = seq(0,60000, by=10000)) +
  geom_bar(stat ="identity") +
    xlab("Crime Type") +
    ylab("Frequency") +
  ggtitle("Frequency of crimes across boroughs") +
  geom_text(size = 3, position = position_stack(vjust = 0.5)) +
  theme(plot.title = element_text(hjust = 0.5)))
```

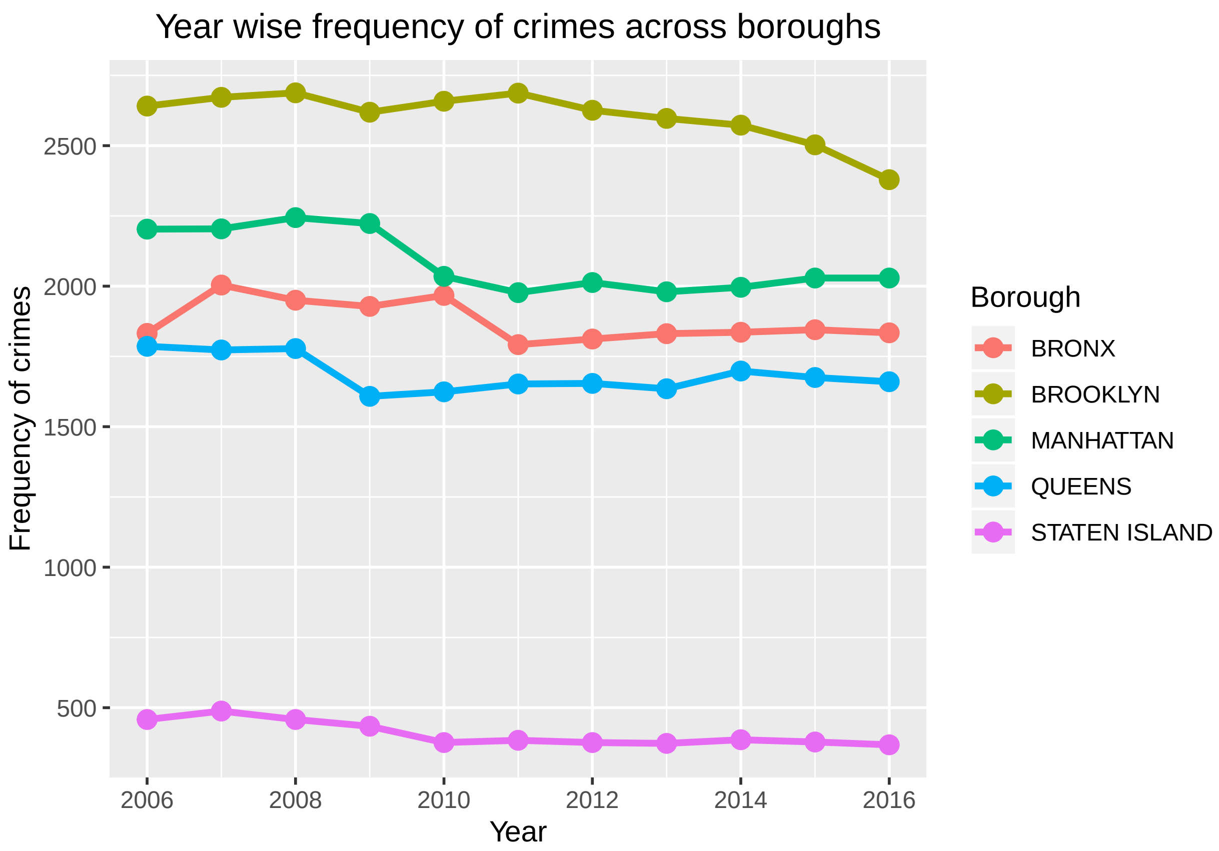


In the following snippet, we are showing a table which depicts the year wise frequency of crimes for each borough. We have achieved this by using the separate function to extract the year from the created date, and then we spread across the year, thus computing the frequency of crimes for each borough. The following line graph shows the year-wise trends of crimes across boroughs.

```r
boroYear <-nycCrimenodups %>%
  select( Borough , Year_Month_New,Type) %>%
  filter(!is.na(Borough))
yearData <- separate(boroYear, Year_Month_New, into=c("year", "month"), convert = T)

boroYear <- yearData %>%
  group_by(year,Borough) %>%
  summarize(frequency=n())

(yearSpread <- boroYear %>%
  spread(key=year, value=frequency))
```

```
## # A tibble: 5 x 12
##   Borough `2006` `2007` `2008` `2009` `2010` `2011` `2012` `2013` `2014`
##   <chr>    <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1 BRONX     1832   2004   1950   1928   1967   1792   1812   1831   1836
## 2 BROOKL~   2641   2672   2688   2619   2658   2687   2626   2597   2573
## 3 MANHAT~   2203   2204   2244   2223   2035   1977   2013   1980   1996
## 4 QUEENS    1786   1773   1778   1608   1624   1652   1654   1635   1698
## 5 STATEN~    458    488    458    434    376    384    376    373    386
## # ... with 2 more variables: `2015` <int>, `2016` <int>
```

```r
(ggplot(data=boroYear, aes(x=year, y=frequency, group=Borough)) +
    scale_x_continuous(breaks = seq(2006,2016, by=2)) +
    scale_y_continuous(breaks= seq(0,3000, by=500)) +
  geom_line(linetype="solid", size=1.2, aes(color=Borough))+
  geom_point(aes(color=Borough), size=3) +
     ggtitle("Year wise frequency of crimes across boroughs") +
    xlab("Year") +
    ylab("Frequency of crimes") +
  theme(plot.title = element_text(hjust = 0.5)))
```



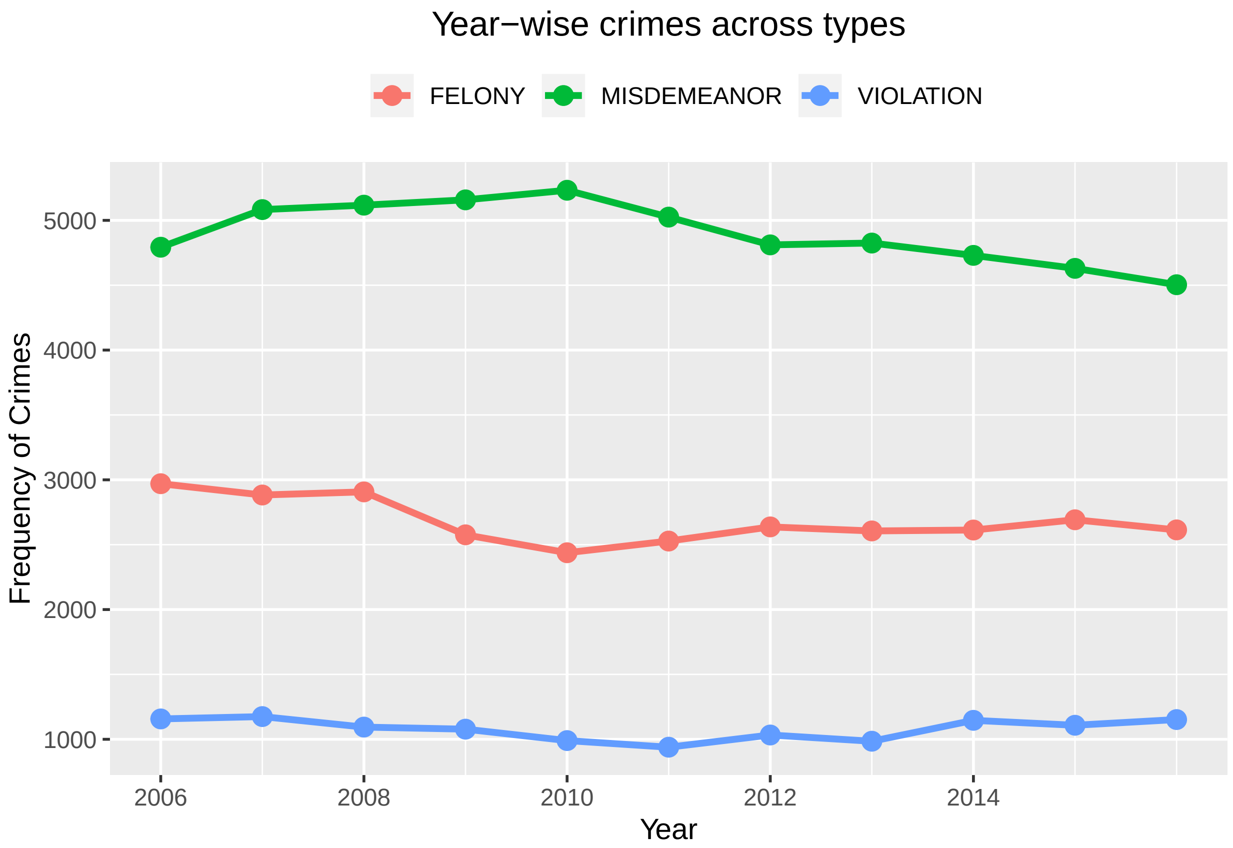
The following line graph shows the frequency of the three crime types over the years. From the year-wise trend we find that maximum crimes reported for violation was during 2007, for felony was during 2006 and misdemeanor during 2010. We then explored the month-wise breakdown of the crimes for the year which had the maximum occurrence.

```
crimeTypYear <- yearData %>%
  filter(!is.na(year) & !is.na(Type)) %>%
  group_by(Type, year) %>%
  summarize(frequency=n())

(typeSpread <- crimeTypYear %>%
  spread(key=year, value=frequency))
```

```
## # A tibble: 3 x 12
## # Groups:   Type [3]
##   Type  `2006` `2007` `2008` `2009` `2010` `2011` `2012` `2013` `2014`
##   <chr>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
## 1 FELO~   2970   2883   2907   2576   2438   2528   2637   2606   2613
## 2 MISD~   4793   5083   5117   5158   5232   5025   4811   4825   4730
## 3 VIOL~   1157   1175   1094   1078    990    939   1033    985   1146
## # ... with 2 more variables: `2015` <int>, `2016` <int>
```

```
crimeTyp <- crimeTypYear %>%
  group_by(Type) %>%
  summarize(totalCrimes= sum(frequency))
```

```
crimeTypYear <- merge(x=crimeTypYear, y=crimeTyp, by="Type")

(ggplot(data=crimeTypYear, aes(x=year, y=frequency, group=Type)) +
      scale_x_continuous(breaks = seq(2006,2015, by=2)) +
      geom_line(linetype="solid", size=1.2, aes(color=Type))+
      geom_point(aes(color=Type), size=3) +
      ggtitle("Year-wise crimes across types") +
      xlab("Year") +
      ylab("Frequency of Crimes") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "top", legend.title = element_blank()))
```

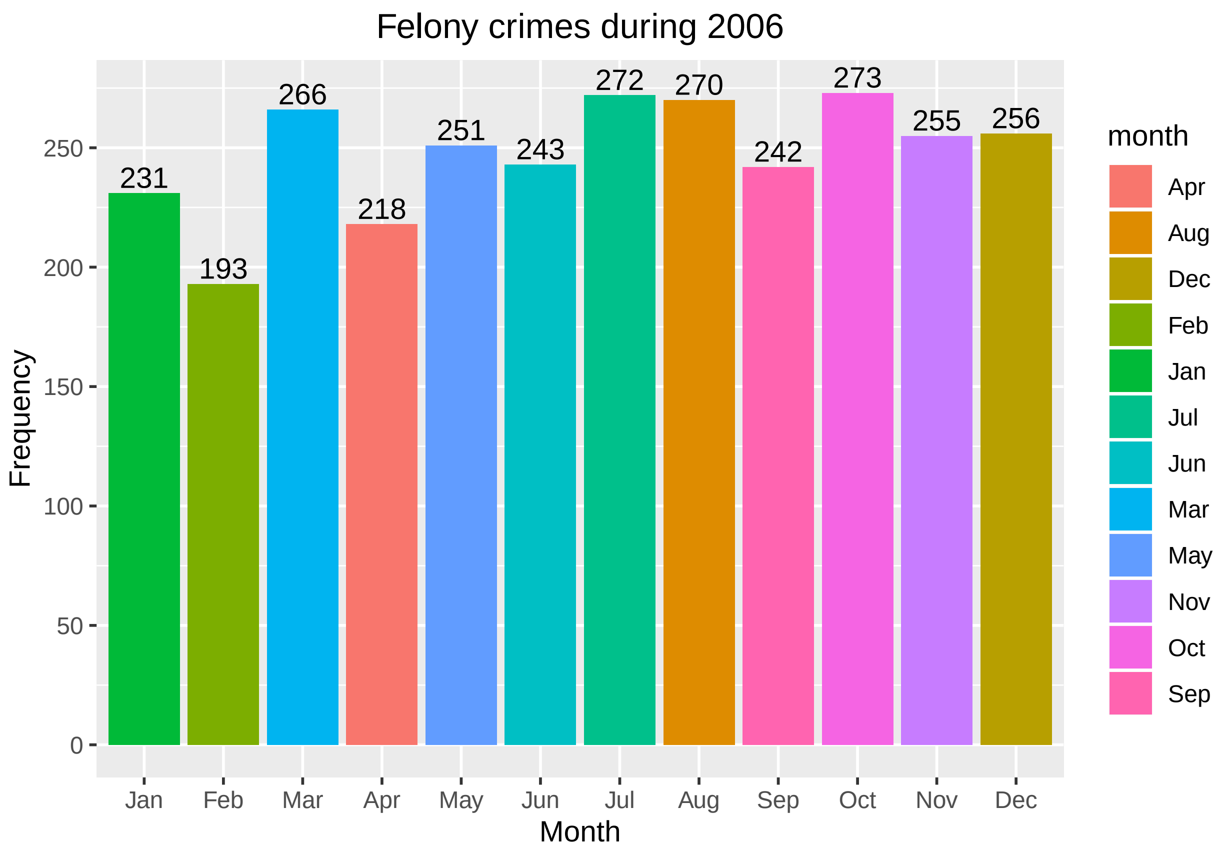


```
boroYear <- nycCrimenodups %>%
  select( Borough, Year_Month_New, Type) %>%
  filter(!is.na(Borough))
yearData <- separate(boroYear, Year_Month_New, into=c("year", "month"), convert = T)

yearStats <- yearData %>%
  group_by(Borough, Type, year) %>%
  summarize(count=n())

# Computing crime type
yearCrime <-yearStats %>%
  group_by(Type,year) %>%
  summarize(count = sum(count))
```

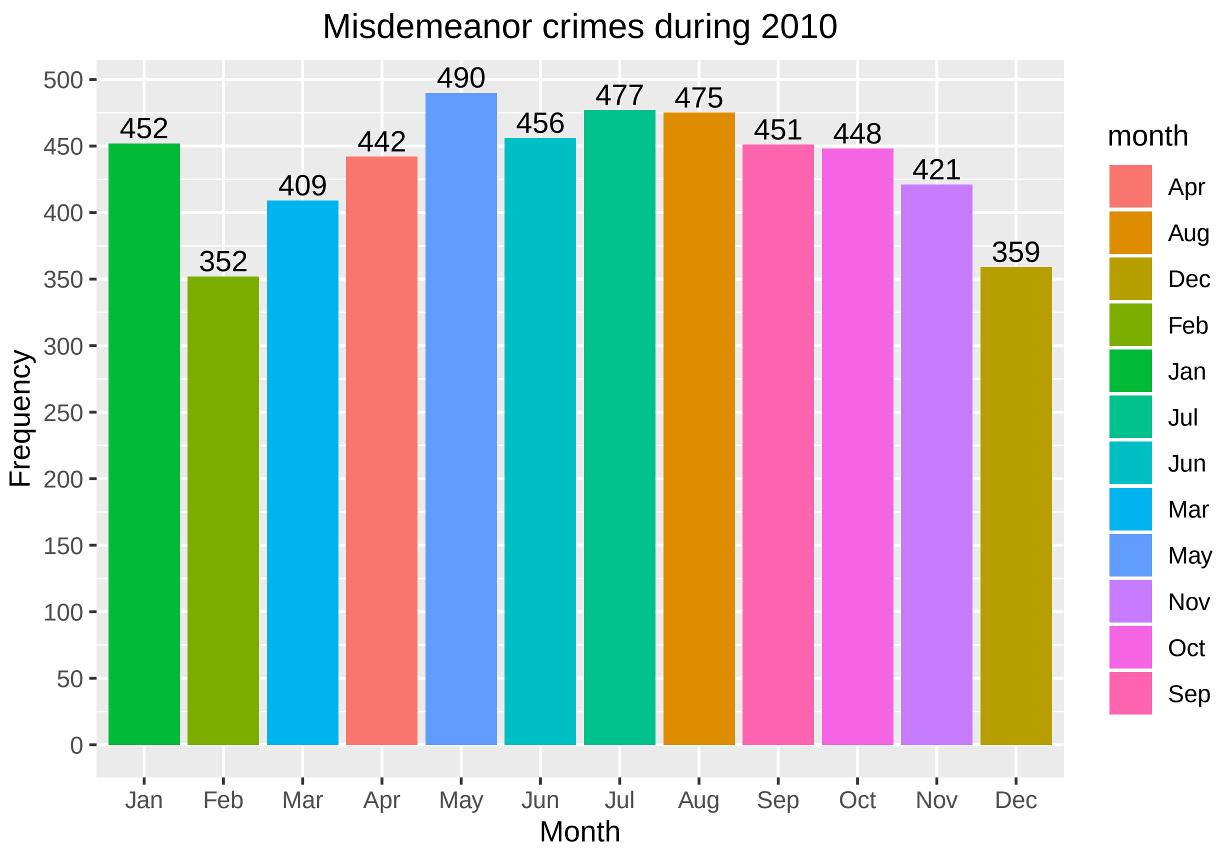
```
(maxYearCrime <- yearCrime %>%
  group_by(Type) %>%
  summarize(maxCount=max(count),
            maxYear= year[count==maxCount]))
```

```
## # A tibble: 3 x 3
##   Type        maxCount maxYear
##   <chr>          <int>   <int>
## 1 FELONY          2970    2006
## 2 MISDEMEANOR     5232    2010
## 3 VIOLATION       1175    2007
```

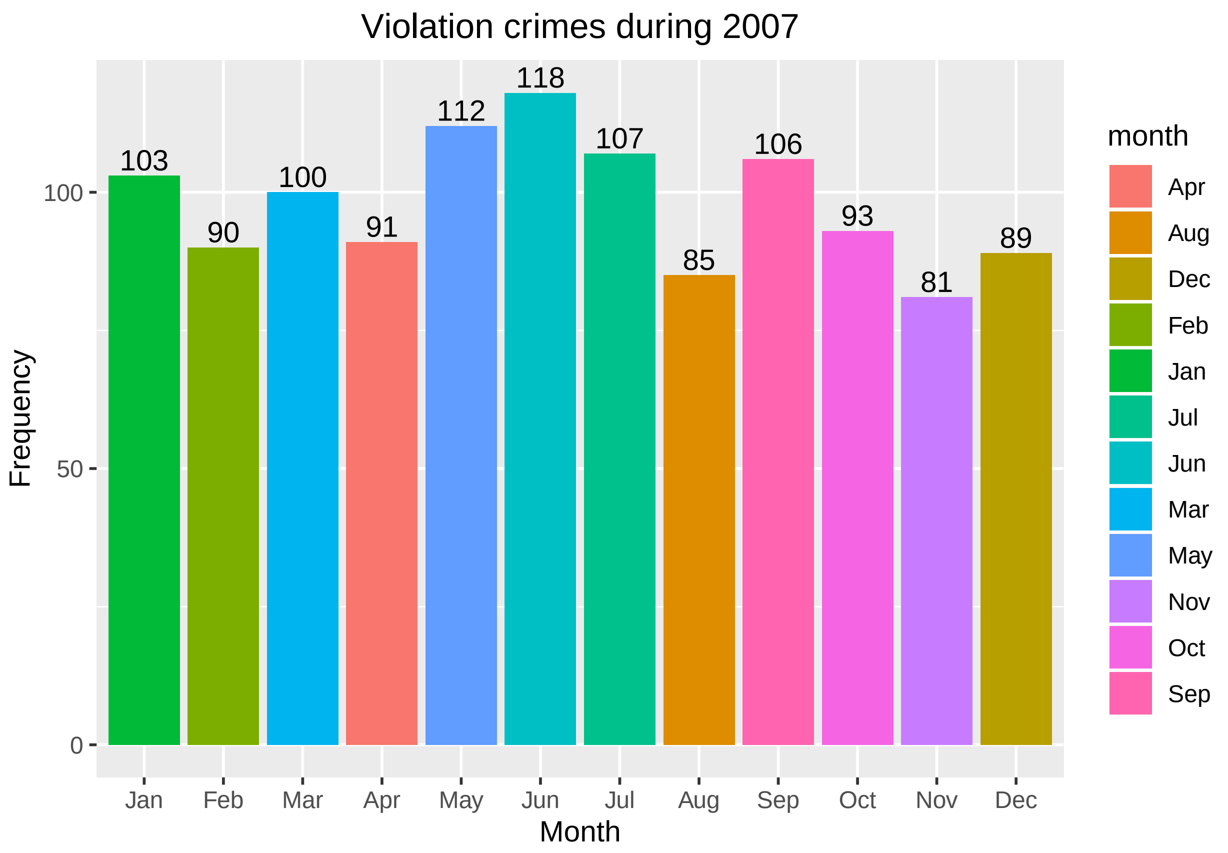
```
felonyMonthCrimes <- yearData %>%
  filter(Type=="FELONY" &
         year==maxYearCrime[maxYearCrime$Type=="FELONY","maxYear"]$maxYear) %>%
   group_by(month) %>%
  summarize(monthFrequency = n())
felonyMonthCrimes$month <- month.abb[felonyMonthCrimes$month]

misdeameanorCrimes <- yearData %>%
  filter(Type=="MISDEMEANOR" &
         year==maxYearCrime[maxYearCrime$Type=="MISDEMEANOR","maxYear"]$maxYear) %>%
   group_by(month) %>%
  summarize(monthFrequency = n())
misdeameanorCrimes$month <- month.abb[misdeameanorCrimes$month]

violationCrimes <- yearData %>%
  filter(Type=="VIOLATION" &
         year==maxYearCrime[maxYearCrime$Type=="VIOLATION","maxYear"]$maxYear) %>%
   group_by(month) %>%
  summarize(monthFrequency = n())

violationCrimes$month <- month.abb[violationCrimes$month]

(ggplot(felonyMonthCrimes,aes(x=month,y=monthFrequency, fill=month)) +
     geom_bar(stat="identity") +
      scale_y_continuous(breaks = seq(0,3000,by=50) ) +
      scale_x_discrete(limits = month.abb) +
     ggtitle(paste0("Felony crimes during ",
                    maxYearCrime[maxYearCrime$Type=="FELONY", "maxYear"]$maxYear)) +
      geom_text(aes(label=monthFrequency), position=position_dodge(width=0.9),
                vjust=-0.25) + guides(colour="none") +
      ylab("Frequency") +
      xlab("Month") +
       theme(plot.title = element_text(hjust = 0.5)))
```



```
(ggplot(misdeameanorCrimes,aes(x=month,y=monthFrequency, fill=month)) +
     geom_bar(stat="identity") +
      scale_y_continuous(breaks = seq(0,3000,by=50) ) +
      scale_x_discrete(limits = month.abb) +
      ggtitle(paste0("Misdemeanor crimes during ",
                     maxYearCrime[maxYearCrime$Type=="MISDEMEANOR", "maxYear"]$maxYear)) +
      geom_text(aes(label=monthFrequency), position=position_dodge(width=0.9),
                vjust=-0.25) + guides(colour="none") +
      ylab("Frequency") +
      xlab("Month") +
       theme(plot.title = element_text(hjust = 0.5)))
```



```
(ggplot(violationCrimes, aes(x=month,y=monthFrequency, fill=month)) +
     geom_bar(stat="identity") +
      scale_y_continuous(breaks = seq(0,3000,by=50) ) +
      scale_x_discrete(limits = month.abb) +
      ggtitle(paste0("Violation crimes during ",
                     maxYearCrime[maxYearCrime$Type=="VIOLATION", "maxYear"]$maxYear)) +
      geom_text(aes(label=monthFrequency), position=position_dodge(width=0.9),
                vjust=-0.25) + guides(colour="none") +
      ylab("Frequency") +
      xlab("Month") +
       theme(plot.title = element_text(hjust = 0.5)))
```



Crime Statistics

In the following snippet, we made use of the year statistics across boroughs. We used unite function to combine the crime type and year, forming a new variable named (Type_year) and then spreaded across that column. The following shows the head of the crime statistics information which will be used for joining with the 311NYC data.

```
(crimeStats <- yearStats %>%
  unite("Type_year", Type, year) %>%
  spread(key=Type_year, value = count))
```

```
## # A tibble: 5 x 34
## # Groups:   Borough [5]
##   Borough FELONY_2006 FELONY_2007 FELONY_2008 FELONY_2009 FELONY_2010
##   <chr>         <int>       <int>       <int>       <int>       <int>
## 1 BRONX           536         549         506         473         476
## 2 BROOKL~         892         877         934         789         766
## 3 MANHAT~         819         760         776         676         588
## 4 QUEENS          638         595         586         558         539
## 5 STATEN~          85         102         105          80          69
## # ... with 28 more variables: FELONY_2011 <int>, FELONY_2012 <int>,
## #   FELONY_2013 <int>, FELONY_2014 <int>, FELONY_2015 <int>,
## #   FELONY_2016 <int>, MISDEMEANOR_2006 <int>, MISDEMEANOR_2007 <int>,
## #   MISDEMEANOR_2008 <int>, MISDEMEANOR_2009 <int>,
## #   MISDEMEANOR_2010 <int>, MISDEMEANOR_2011 <int>,
## #   MISDEMEANOR_2012 <int>, MISDEMEANOR_2013 <int>,
## #   MISDEMEANOR_2014 <int>, MISDEMEANOR_2015 <int>,
## #   MISDEMEANOR_2016 <int>, VIOLATION_2006 <int>, VIOLATION_2007 <int>,
## #   VIOLATION_2008 <int>, VIOLATION_2009 <int>, VIOLATION_2010 <int>,
## #   VIOLATION_2011 <int>, VIOLATION_2012 <int>, VIOLATION_2013 <int>,
## #   VIOLATION_2014 <int>, VIOLATION_2015 <int>, VIOLATION_2016 <int>
```

Joining NYC311 and NYCCrimes data

We perform a join on the above crime statistics data and the cleaned 311NYC data using Borough. As our focus would be narrowed down to just complaints and crimes across boroughs over the years, we have ignored other irrelevant information. The following shows the head of the joined data.

```r
complCrimeData <- inner_join(nyc311nodups, crimeStats, by="Borough")
complCrimeData <- complCrimeData[,c(-1,-4,-8:-15)]
head(complCrimeData)
```

```
##     Borough           Created.Date Agency                     Agency.Name
## 1     BRONX 04/14/2015 02:14:40 AM   NYPD New York City Police Department
## 2  BROOKLYN 04/14/2015 02:10:12 AM   NYPD New York City Police Department
## 3  BROOKLYN 04/14/2015 02:03:01 AM   NYPD New York City Police Department
## 4  BROOKLYN 04/14/2015 02:02:40 AM   NYPD New York City Police Department
## 5 MANHATTAN 04/14/2015 02:00:04 AM   NYPD New York City Police Department
## 6  BROOKLYN 04/14/2015 01:52:15 AM   NYPD New York City Police Department
##            Complaint.Type FELONY_2006 FELONY_2007 FELONY_2008 FELONY_2009
## 1                 Vending         536         549         506         473
## 2        Blocked Driveway         892         877         934         789
## 3 Noise - Street/Sidewalk         892         877         934         789
## 4 Noise - Street/Sidewalk         892         877         934         789
## 5 Noise - Street/Sidewalk         819         760         776         676
## 6 Noise - Street/Sidewalk         892         877         934         789
##   FELONY_2010 FELONY_2011 FELONY_2012 FELONY_2013 FELONY_2014 FELONY_2015
## 1         476         486         486         507         499         521
## 2         766         845         852         841         825         814
## 3         766         845         852         841         825         814
## 4         766         845         852         841         825         814
## 5         588         562         644         598         623         667
## 6         766         845         852         841         825         814
##   FELONY_2016 MISDEMEANOR_2006 MISDEMEANOR_2007 MISDEMEANOR_2008
## 1         534             1038             1185             1203
## 2         781             1395             1453             1445
## 3         781             1395             1453             1445
## 4         781             1395             1453             1445
## 5         666             1177             1219             1252
## 6         781             1395             1453             1445
##   MISDEMEANOR_2009 MISDEMEANOR_2010 MISDEMEANOR_2011 MISDEMEANOR_2012
## 1             1224             1286             1126             1103
## 2             1508             1568             1538             1466
## 3             1508             1568             1538             1466
## 4             1508             1568             1538             1466
## 5             1314             1258             1223             1152
## 6             1508             1568             1538             1466
##   MISDEMEANOR_2013 MISDEMEANOR_2014 MISDEMEANOR_2015 MISDEMEANOR_2016
## 1             1111             1090             1091             1052
## 2             1446             1382             1328             1251
## 3             1446             1382             1328             1251
## 4             1446             1382             1328             1251
## 5             1208             1152             1153             1145
## 6             1446             1382             1328             1251
##   VIOLATION_2006 VIOLATION_2007 VIOLATION_2008 VIOLATION_2009
## 1            258            270            241            231
## 2            354            342            309            322
## 3            354            342            309            322
## 4            354            342            309            322
## 5            207            225            216            233
## 6            354            342            309            322
##   VIOLATION_2010 VIOLATION_2011 VIOLATION_2012 VIOLATION_2013
## 1            205            180            223            213
## 2            324            304            308            310
## 3            324            304            308            310
## 4            324            304            308            310
## 5            189            192            217            174
## 6            324            304            308            310
##   VIOLATION_2014 VIOLATION_2015 VIOLATION_2016
## 1            247            233            248
## 2            366            361            347
## 3            366            361            347
## 4            366            361            347
## 5            221            209            218
## 6            366            361            347
```

Exploration on joined datasets

The following gives a small overview of the following crime types: Violation - The action of breaking regulations especially law, agreement, principles. For example: breaking the traffic rules, illegal parking, smoking in prohibited areas, etc. Misdemeanor - This type of crime is a minor wrong doing. For example: theft, drug trafficking, animal abuse, etc. Felony - This type of crime involves extreme violence which is considered as more serious than misdemeanor. For example: murder, hit and run accident cases, rape cases, etc.

Now, we are classifying the complaint types into felony, violation and misdemeanor crimes.

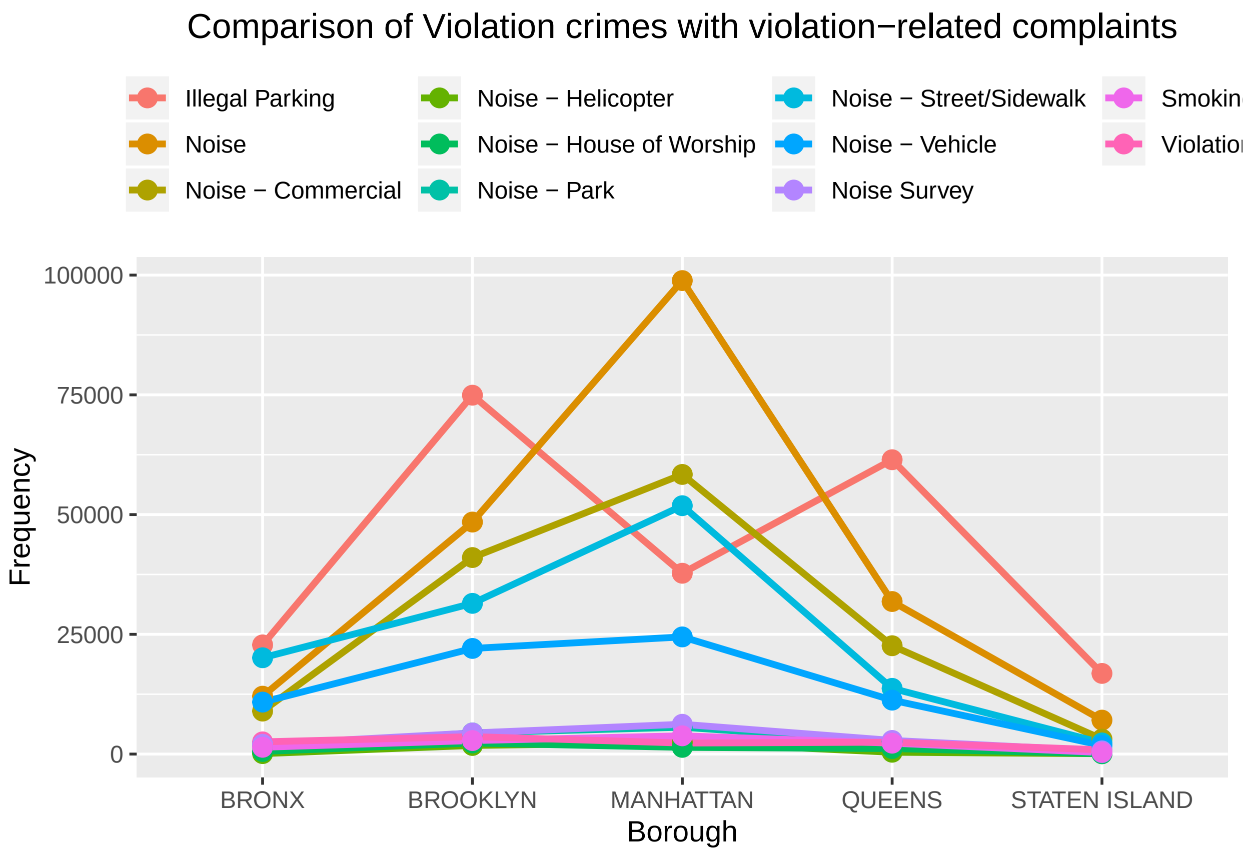
Considering violation, some of the relatable complaints could be illegal parking, smoking and noise complaints. The reason for choosing the above complaints being relevant to violation is because all these complaints are related to breaking the basic rules and regulations. The following shows trends across the boroughs for the violation related complaints and violation crimes.

```
#Illegal Parking, Smoking, Noise complaints

voilationCompl <- complCrimeData %>%
  select(Borough, Complaint.Type, Created.Date) %>%
  filter(Complaint.Type=="Illegal Parking" |
           Complaint.Type=="Smoking" | str_starts(Complaint.Type,"Noise"))%>%
  group_by(Borough, Complaint.Type) %>%
  summarize(frequency=n())
```

```
(complSpread <- voilationCompl %>%
  spread(key=Complaint.Type, value=frequency))
```

```
## # A tibble: 5 x 11
## # Groups:   Borough [5]
##   Borough `Illegal Parkin~ Noise `Noise - Commer~ `Noise - Helico~
##   <chr>              <int> <int>            <int>            <int>
## 1 BRONX              22796 12085             8971               95
## 2 BROOKL~            74929 48440            41030             1798
## 3 MANHAT~            37752 98859            58383             2403
## 4 QUEENS             61451 31848            22617              380
## 5 STATEN~            16839  7086             3126               80
## # ... with 6 more variables: `Noise - House of Worship` <int>, `Noise -
## #   Park` <int>, `Noise - Street/Sidewalk` <int>, `Noise - Vehicle` <int>,
## #   `Noise Survey` <int>, Smoking <int>
```

```
violationBoro <- complCrimeData %>%
  select(Borough, c(28:38))

violationBoro <- distinct(violationBoro)%>%
  gather(key="typeYear",value="frequency", c(2:length(names(violationBoro))))

violationBoro <- violationBoro%>%
  group_by(Borough)%>%
  summarize(Violation=sum(frequency))

violationBoro <- merge(violationBoro, complSpread, by="Borough")
violationGather <- violationBoro %>%
  gather(key="Violation.Type", value = "frequency", c(2:length(names(violationBoro))))

(ggplot(data=violationGather, aes(x=Borough, y=frequency, group=Violation.Type)) +
      geom_line(linetype="solid", size=1.2, aes(color=Violation.Type))+
      geom_point(aes(color=Violation.Type), size=3) +
      ggtitle("Comparison of Violation crimes with violation-related complaints") +
      xlab("Borough") +
      ylab("Frequency") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "top", legend.title = element_blank()))
```

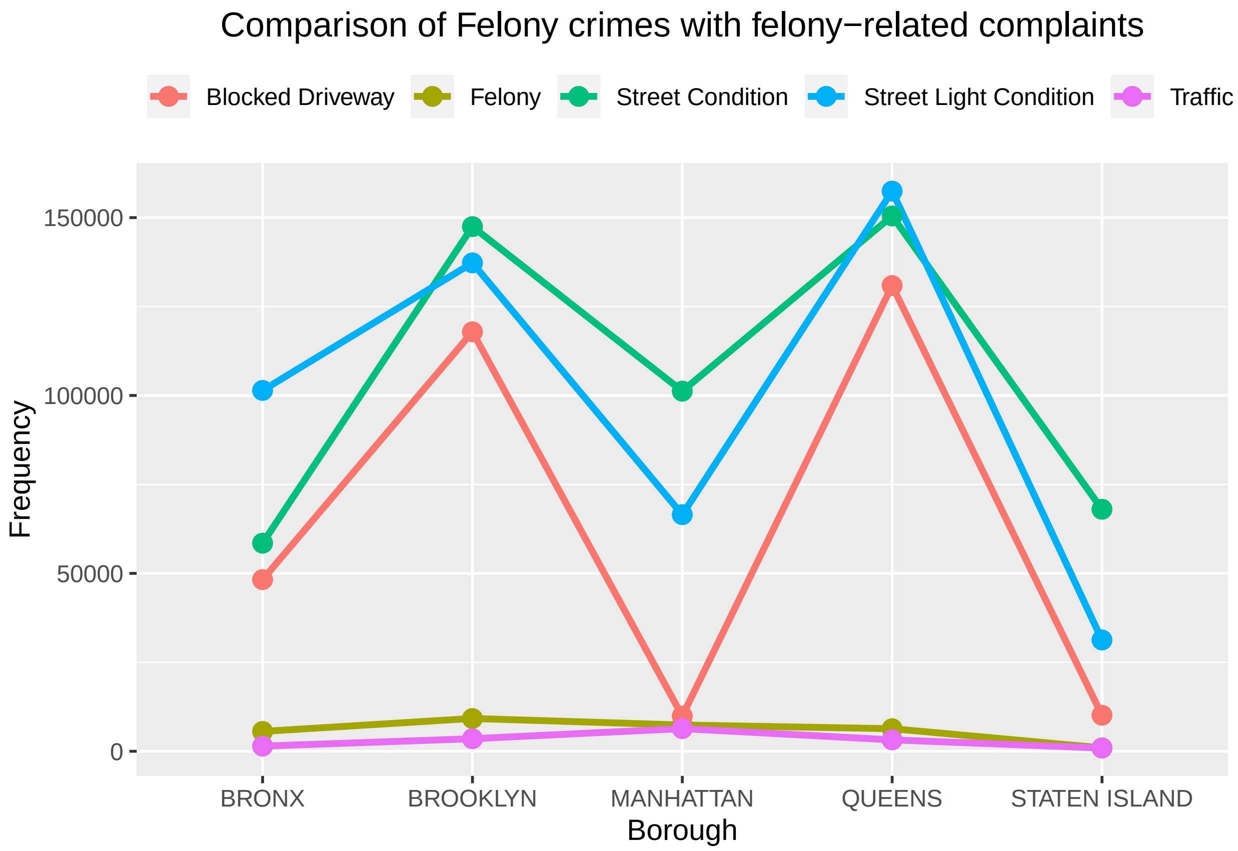


Considering felony, some of the relatable complaints could be blocked driveway, traffic, street condition and street light condition. The reason for choosing the above complaints being relevant to felony is that there are could be accidents due to improper street conditions, heavy traffic that also caused blocked driveway. Even murders can occur on the street which may lead to traffic and blocked driveway. Assuming these criteria, we find high correlation between felony and the above mentioned complaints. The following shows trends across the boroughs for the felony related complaints and felony crimes.

```
# Blocked Driveway, Traffic, Street Condition, Street Light Condition

felonyCompl <- complCrimeData %>%
  select(Borough, Complaint.Type) %>%
  filter(Complaint.Type=="Blocked Driveway" |
           Complaint.Type=="Traffic" | Complaint.Type=="Street Condition" |
           Complaint.Type=="Street Light Condition")%>%
  group_by(Borough, Complaint.Type) %>%
  summarize(frequency=n())

(complSpread <- felonyCompl %>%
  spread(key=Complaint.Type, value=frequency))
```

```
## # A tibble: 5 x 5
## # Groups:   Borough [5]
##   Borough    `Blocked Drivewa~ `Street Conditi~ `Street Light Cond~ Traffic
##   <chr>                  <int>            <int>               <int>   <int>
## 1 BRONX                  48247            58490              101425    1447
## 2 BROOKLYN              117895           147471              137270    3522
## 3 MANHATTAN              9894           101222               66506     6367
## 4 QUEENS               130899           150456              157445     3207
## 5 STATEN IS~            10139            68014               31282      901
```

```
felonyBoro <- complCrimeData %>%
  select(Borough, c(6:16))
felonyBoro <- distinct(felonyBoro)%>%
  gather(key="typeYear",value="frequency", c(2:length(names(felonyBoro))))

felonyBoro <- felonyBoro%>%
  group_by(Borough)%>%
  summarize(Felony=sum(frequency))

felonyBoro <- merge(felonyBoro, complSpread, by="Borough")

felonyGather <- felonyBoro %>%
  gather(key="Felony.Type", value = "frequency", c(2:length(names(felonyBoro))))

(ggplot(data=felonyGather, aes(x=Borough, y=frequency, group=Felony.Type)) +
      geom_line(linetype="solid", size=1.2, aes(color=Felony.Type))+
      geom_point(aes(color=Felony.Type), size=3) +
      ggtitle("Comparison of Felony crimes with felony-related complaints") +
      xlab("Borough") +
      ylab("Frequency") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "top", legend.title = element_blank()))
```

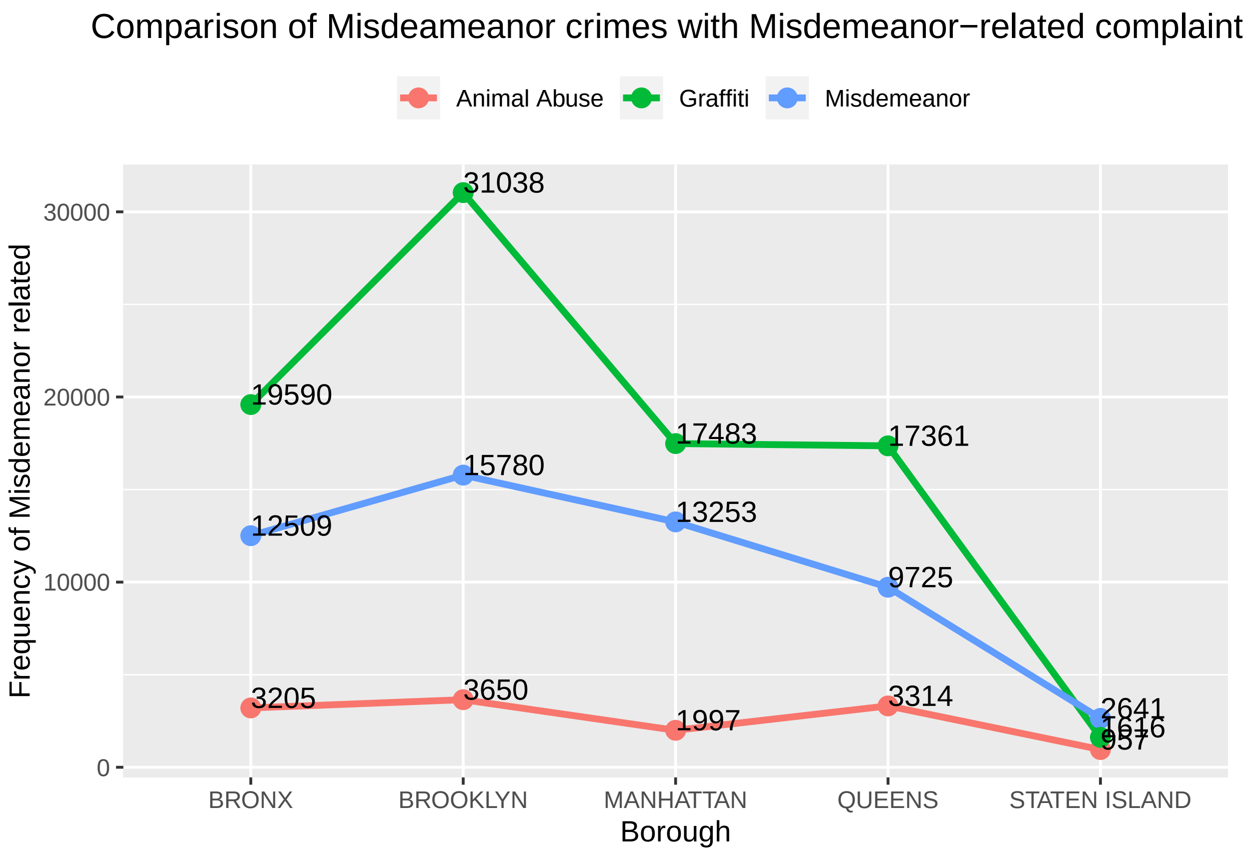


Considering misdemeanor, some of the relatable complaints could be lost property(theft), graffiti and animal abuse. The reason for choosing the above complaints being relevant to misdemeanor is because these complaints are consider as minor wrong doings and doesn't cause any fatal outcomes. The following shows trends across the boroughs for the misdemeanor related complaints and misdemeanor crimes.

```
# Graffitti, Animal abuse

misdemeanorCompl <- complCrimeData %>%
  select(Borough, Complaint.Type, Created.Date) %>%
  filter(Complaint.Type=="Graffiti" |  Complaint.Type=="Animal Abuse")%>%
  group_by(Borough, Complaint.Type) %>%
  summarize(frequency=n())

(complSpread <- misdemeanorCompl %>%
  spread(key=Complaint.Type, value=frequency))
```

```
## # A tibble: 5 x 3
## # Groups:   Borough [5]
##   Borough       `Animal Abuse` Graffiti
##   <chr>                  <int>    <int>
## 1 BRONX                   3205    19590
## 2 BROOKLYN                3650    31038
## 3 MANHATTAN               1997    17483
## 4 QUEENS                  3314    17361
## 5 STATEN ISLAND            957     1616
```

```
misdemeanorBoro <- complCrimeData %>%
  select(Borough, c(17:27))

misdemeanorBoro <- distinct(misdemeanorBoro)%>%
  gather(key="typeYear",value="frequency", c(2:length(names(misdemeanorBoro))))

misdemeanorBoro <- misdemeanorBoro%>%
  group_by(Borough)%>%
  summarize(Misdemeanor=sum(frequency))

misdemeanorBoro <- merge(misdemeanorBoro, complSpread, by="Borough")
misdemeanorGather <- misdemeanorBoro %>%
  gather(key="Misdemeanor.Type", value = "frequency",
         c(2:length(names(misdemeanorBoro))))

(ggplot(data=misdemeanorGather, aes(x=Borough, y=frequency, group=Misdemeanor.Type)) +
      geom_line(linetype="solid", size=1.2, aes(color=Misdemeanor.Type)) +
      geom_point(aes(color=Misdemeanor.Type), size=3) +
      geom_text(aes(label=frequency), hjust=0, vjust=0) +
      ggtitle("Comparison of Misdeameanor crimes with Misdemeanor-related complaints") +
      xlab("Borough") +
      ylab("Frequency of Misdemeanor related") +
    theme(plot.title = element_text(hjust = 0.5),
          legend.position = "top", legend.title = element_blank()))
```



CONCLUSION

In this document, we have explored both the NYC 311 data and the NYPD NYC Crimes data by showing various visualization graphs. We are joining them using borough as a common column and continued to explore the connections between them. We depicted the correlations between the 311 complaints and crime types with sound reasoning of why we found them relevant.

APPENDIX

Data dictionary of joined data

- Borough – town/ district of the NYC provided by submitter (Values: BRONX, BROOKLYN, MANHATTAN, QUEENS, STATEN ISLAND).
- Created.Date – The date when the service request was created (Type: timestamp (mm/dd/yyyy hh:mm:ss)).
- 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).
- FELONY_2006 - Frequency of “FELONY” crime type during 2006.

- FELONY_2007 - Frequency of “FELONY” crime type during 2007.
- FELONY_2008 - Frequency of “FELONY” crime type during 2008.
- FELONY_2009 - Frequency of “FELONY” crime type during 2009.
- FELONY_2010 - Frequency of “FELONY” crime type during 2010.
- FELONY_2011 - Frequency of “FELONY” crime type during 2011.
- FELONY_2012 - Frequency of “FELONY” crime type during 2012.
- FELONY_2013 - Frequency of “FELONY” crime type during 2013.
- FELONY_2014 - Frequency of “FELONY” crime type during 2014.
- FELONY_2015 - Frequency of “FELONY” crime type during 2015.
- FELONY_2016 - Frequency of “FELONY” crime type during 2016.
- MISDEMEANOR_2006 - Frequency of “MISDEMEANOR” crime type during 2006.
- MISDEMEANOR_2007 - Frequency of “MISDEMEANOR” crime type during 2007.
- MISDEMEANOR_2008 - Frequency of “MISDEMEANOR” crime type during 2008.
- MISDEMEANOR_2009 - Frequency of “MISDEMEANOR” crime type during 2009.
- MISDEMEANOR_2010 - Frequency of “MISDEMEANOR” crime type during 2010.
- MISDEMEANOR_2011 - Frequency of “MISDEMEANOR” crime type during 2011.
- MISDEMEANOR_2012 - Frequency of “MISDEMEANOR” crime type during 2012.
- MISDEMEANOR_2013 - Frequency of “MISDEMEANOR” crime type during 2013.
- MISDEMEANOR_2014 - Frequency of “MISDEMEANOR” crime type during 2014.
- MISDEMEANOR_2015 - Frequency of “MISDEMEANOR” crime type during 2015.
- MISDEMEANOR_2016 - Frequency of “MISDEMEANOR” crime type during 2016.
- VIOLATION_2006 - Frequency of “VIOLATION” crime type during 2006.
- VIOLATION_2007 - Frequency of “VIOLATION” crime type during 2007.
- VIOLATION_2008 - Frequency of “VIOLATION” crime type during 2008.
- VIOLATION_2009 - Frequency of “VIOLATION” crime type during 2009.
- VIOLATION_2010 - Frequency of “VIOLATION” crime type during 2010.
- VIOLATION_2011 - Frequency of “VIOLATION” crime type during 2011.
- VIOLATION_2012 - Frequency of “VIOLATION” crime type during 2012.
- VIOLATION_2013 - Frequency of “VIOLATION” crime type during 2013.
- VIOLATION_2014 - Frequency of “VIOLATION” crime type during 2014.
- VIOLATION_2015 - Frequency of “VIOLATION” crime type during 2015.
- VIOLATION_2016 - Frequency of “VIOLATION” crime type during 2016.