# homework v

*Sachin Mohan Sujir*

*2020-10-02*

## Introduction

In this document, I will be computing crime data statistics which focuses on yearwise frequency of crimes for every borough. I will then joining the cleaned 311Nyc data and the crime statistics data using join functions and ignoring the irrelevant columns from the final joined data.

## Initialization

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

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

```
library(xtable)
options(xtable.comment=FALSE)
options(xtable.booktabs=TRUE)
nyc311<-nyc311 %>%
  select(Agency,
     Agency.Name,
     Created.Date,
     Closed.Date,
     Incident.Zip,
     Due.Date,
     Latitude,
     Longitude,
     Complaint.Type,
     Descriptor,

     Status,
     Borough)
xtable(head(nyc311))
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rlllllllrrllll}
##   \toprule
##  & Agency & Agency.Name & Created.Date & Closed.Date & Incident.Zip & Due.Date & Latitude & Longitude
##   \midrule
## 1 & NYPD & New York City Police Department & 04/14/2015 02:14:40 AM & 04/14/2015 03:03:22 AM & 10465
##   2 & NYPD & New York City Police Department & 04/14/2015 02:10:12 AM &  & 11234 & 04/14/2015 10:10:
##   3 & NYPD & New York City Police Department & 04/14/2015 02:03:01 AM &  & 11204 & 04/14/2015 10:03:
##   4 & NYPD & New York City Police Department & 04/14/2015 02:02:40 AM &  & 11211 & 04/14/2015 10:02:
##   5 & NYPD & New York City Police Department & 04/14/2015 02:00:04 AM & 04/14/2015 02:47:33 AM & 100
##   6 & NYPD & New York City Police Department & 04/14/2015 01:52:15 AM & 04/14/2015 02:11:10 AM & 112
##    \bottomrule
## \end{tabular}
## \end{table}
```

```
nyc311 <- distinct(nyc311)
names(nyc311)
```

```
##  [1] "Agency"         "Agency.Name"    "Created.Date"   "Closed.Date"
##  [5] "Incident.Zip"   "Due.Date"       "Latitude"       "Longitude"
##  [9] "Complaint.Type" "Descriptor"     "Status"         "Borough"
```

## Handling missing values in 311NYC

In the following snippet, I have handled the missing values and the errornous records in the columns of the data. Intially, I have replaced the invalid zip codes with NA if the zip code length is not 5 or 10 and if the zip code length is 10 then it should satisfy the "xxxxx-xxxx" format. Besides, I could find zipcodes like 00000, 10000 which were invalid, hence replaced them with NA. Now considering the closed date column,there were 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 I replaced them with NA. For borough, there were around 800K records with unspecified values, out of which 600K had valid zip codes, so I found the boroughs for those records using the valid zipcode information and remaining was filled with NA. I could match the zip code that had missing borough and the zip code with the borough specified and filled the missing borough information.

```
# Replacing invalid zipcodes with NA
nyc311[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

nyc311[as.Date(nyc311$Closed.Date, format="%m/%d/%Y")==
                as.Date("01/01/1900", format="%m/%d/%Y") |
                as.Date(nyc311$Closed.Date, format="%m/%d/%Y")<
                  as.Date(nyc311$Created.Date, format="%m/%d/%Y"),
             c("Closed.Date") ] <- NA

unspecifiedBro <- nyc311 %>%
  select(Incident.Zip, Borough) %>%
  filter(Borough=="Unspecified" & !is.na(Incident.Zip))

zipCodeTable <- nyc311 %>%
  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"

nyc311 <- merge(x=nyc311, y=joinedTab,
                by=c("Incident.Zip", "Borough"), sort=FALSE, all.x = TRUE)
nyc311[!is.na(Borough.y), "Borough"] <- nyc311[!is.na(Borough.y), "Borough.y"]
nyc311[Borough=="Unspecified", "Borough"] <-
  nyc311[Borough=="Unspecified", "Borough.y"]
# drop the borough.y
nyc311 <- nyc311[,-"Borough.y"]
```

# Relatable data set - NYPD NYC Crimes data

## Description

I have used the NYPD NYC crimes data which is a sample of size approx 95K records taken from the original data source.This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD). I found this dataset not only relevant to nyc311 but also interesting.

## Initialization

Here I load the NYC Crimes data set from the link as provided below and I 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 of NYC Crimes data

Here, I removed the irrelevant columns and duplicate records in the data, fixed the column name for borough and I am showing the head and data dictionary.

```
names(nycCrimes)
```

```
##  [1] "V1"              "ID"              "Date"            "Time"
##  [5] "Code"            "Offense"         "Status"          "Type"
##  [9] "Boro"            "Premises"        "Latitude"        "Longitude"
## [13] "Year"            "Month"           "Hour"            "Population"
## [17] "Year_Month"      "Year_Month_New"
```

```
library(xtable)
options(xtable.comment=FALSE)
options(xtable.booktabs=TRUE)
nycCrimes<-nycCrimes %>%
  select(Date,
         Time,Code,Offense,Status,Type,
         Boro,Latitude,Longitude,Latitude,
         Population,Year_Month_New)
xtable(head(nycCrimes))
```

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrlrlllllrrrl}
##   \toprule
##  & Date & Time & Code & Offense & Status & Type & Boro & Latitude & Longitude & Population & Year\_M
##   \midrule
## 1 & 13217 & 14:30:00 & 113 & FORGERY & COMPLETED & FELONY & BROOKLYN & 40.66 & -73.92 & 2465690 & 20
##   2 & 15693 & 10:00:00 & 344 & ASSAULT 3 \& RELATED OFFENSES & COMPLETED & MISDEMEANOR & STATEN ISLA
##   3 & 15261 & 14:20:00 & 126 & MISCELLANEOUS PENAL LAW & COMPLETED & FELONY & MANHATTAN & 40.80 & -73
##   4 & 14456 & 11:50:00 & 109 & GRAND LARCENY & ATTEMPTED & FELONY & QUEENS & 40.76 & -73.77 & 223000
##   5 & 13171 & 17:45:00 & 341 & PETIT LARCENY & COMPLETED & MISDEMEANOR & MANHATTAN & 40.77 & -73.96 &
##   6 & 15957 & 21:47:00 & 359 & OFFENSES AGAINST PUBLIC ADMINI & COMPLETED & MISDEMEANOR & BRONX & 40
##    \bottomrule
## \end{tabular}
## \end{table}
```

```
nycCrimes1 <- distinct(nycCrimes)
colnames(nycCrimes1)[colnames(nycCrimes1)=="Boro"] <- "Borough"
nycCrimes1 <-  nycCrimes1[str_trim(Offense)!="",]
names(nycCrimes1)
```

```
##  [1] "Date"            "Time"            "Code"            "Offense"
##  [5] "Status"          "Type"            "Borough"         "Latitude"
##  [9] "Longitude"       "Population"      "Year_Month_New"
```

```
head(nycCrimes1)
```

```
##          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 Latitude Longitude Population Year_Month_New
## 1:      FELONY      BROOKLYN 40.66200 -73.91959    2465690        2006-03
## 2: MISDEMEANOR STATEN ISLAND 40.57112 -74.09007     471000        2012-12
## 3:      FELONY     MANHATTAN 40.79967 -73.94720    1595517        2011-10
## 4:      FELONY        QUEENS 40.76480 -73.77161    2230000        2009-07
## 5: MISDEMEANOR     MANHATTAN 40.77365 -73.95986    1566766        2006-01
## 6: MISDEMEANOR         BRONX 40.81937 -73.91828    1420414        2013-09
```

# Computing Crime statistics from NYC Crimes data

In this NYPD NYC Crimes data, there are these following three crime types: Felony, Misdemeanor and Violation. In the following snippet, I will be computing the yearwise frequency of crimes for every borough in NYC using group_by function. Then I will unite the crime type and year, forming a new variable named (Type_year) and then spread across that column. The following shows the head of the crime statistics information which will be used for joining with the 311NYC data.

```
boroYear <- nycCrimes1 %>%
  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())
```

```
(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 data and removing irrelevant columns

In the following snippet I have joined the above crime statistics data along with the 311NYC data and dropped the irrelevant columns from them. As our focus would be narrowed down to just complaints and crimes across boroughs during every year, I have ignored other irrelevant information.

```
complCrimeData <- inner_join(nyc311, crimeStats, by="Borough")
names(complCrimeData)
```

```
##  [1] "Incident.Zip"     "Borough"          "Agency"          
##  [4] "Agency.Name"      "Created.Date"     "Closed.Date"     
##  [7] "Due.Date"         "Latitude"         "Longitude"       
## [10] "Complaint.Type"   "Descriptor"       "Status"          
## [13] "FELONY_2006"      "FELONY_2007"      "FELONY_2008"     
## [16] "FELONY_2009"      "FELONY_2010"      "FELONY_2011"     
## [19] "FELONY_2012"      "FELONY_2013"      "FELONY_2014"     
## [22] "FELONY_2015"      "FELONY_2016"      "MISDEMEANOR_2006"
## [25] "MISDEMEANOR_2007" "MISDEMEANOR_2008" "MISDEMEANOR_2009"
## [28] "MISDEMEANOR_2010" "MISDEMEANOR_2011" "MISDEMEANOR_2012"
## [31] "MISDEMEANOR_2013" "MISDEMEANOR_2014" "MISDEMEANOR_2015"
## [34] "MISDEMEANOR_2016" "VIOLATION_2006"   "VIOLATION_2007"  
## [37] "VIOLATION_2008"   "VIOLATION_2009"   "VIOLATION_2010"  
## [40] "VIOLATION_2011"   "VIOLATION_2012"   "VIOLATION_2013"  
## [43] "VIOLATION_2014"   "VIOLATION_2015"   "VIOLATION_2016"
```

```
complCrimeData <- complCrimeData[,c(-1,-6,-7,-8,-9,-11,-12)]
head(complCrimeData)
```

```
##     Borough Agency                     Agency.Name           Created.Date
## 1     BRONX   NYPD New York City Police Department 04/14/2015 02:14:40 AM
## 2  BROOKLYN   NYPD New York City Police Department 04/14/2015 02:10:12 AM
## 3  BROOKLYN   NYPD New York City Police Department 04/14/2015 02:03:01 AM
## 4  BROOKLYN   NYPD New York City Police Department 04/14/2015 02:02:40 AM
## 5 MANHATTAN   NYPD New York City Police Department 04/14/2015 02:00:04 AM
## 6  BROOKLYN   NYPD New York City Police Department 04/14/2015 01:52:15 AM
##            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             1110             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
```

# Data Dictionary after joining datasets

- 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.

## Conclusion

In this document, I first created data statistics for the cleaned NYPD NYC crime data. Then computed the yearwise frequency of each crime type for every borough. I used this statistics to join with the 311NYC cleaned data and removed irrelevant columns. Finally, I provided the data dictionary of the joined data set.