**PGDDS June 2018 Cohort**

Team Scooby Data Doo

**Team members:**

Siva Subramaniam Padmanabhan

Shruti Saxena Das

Shailesh Kumar Jha

Chinmaya Sahu

**Case Study on NYC Parking Tickets: An Exploratory Analysis**

**BACKGROUND.**

New York City is a thriving metropolis. Just like most other metros that size, one of the biggest problems its citizens face is parking. The classic combination of a huge number of cars and cramped geography is the exact recipe that leads to a huge number of parking tickets.

**OBJECTIVE.**

To perform exploratory analysis, using Spark, on the given data segment of the year 2017 and to scientifically estimate and analyse the probable reasons behind the high proportion of parking tickets.

**DATA UNDERSTANDING.**

Total Number of rows: 10803028

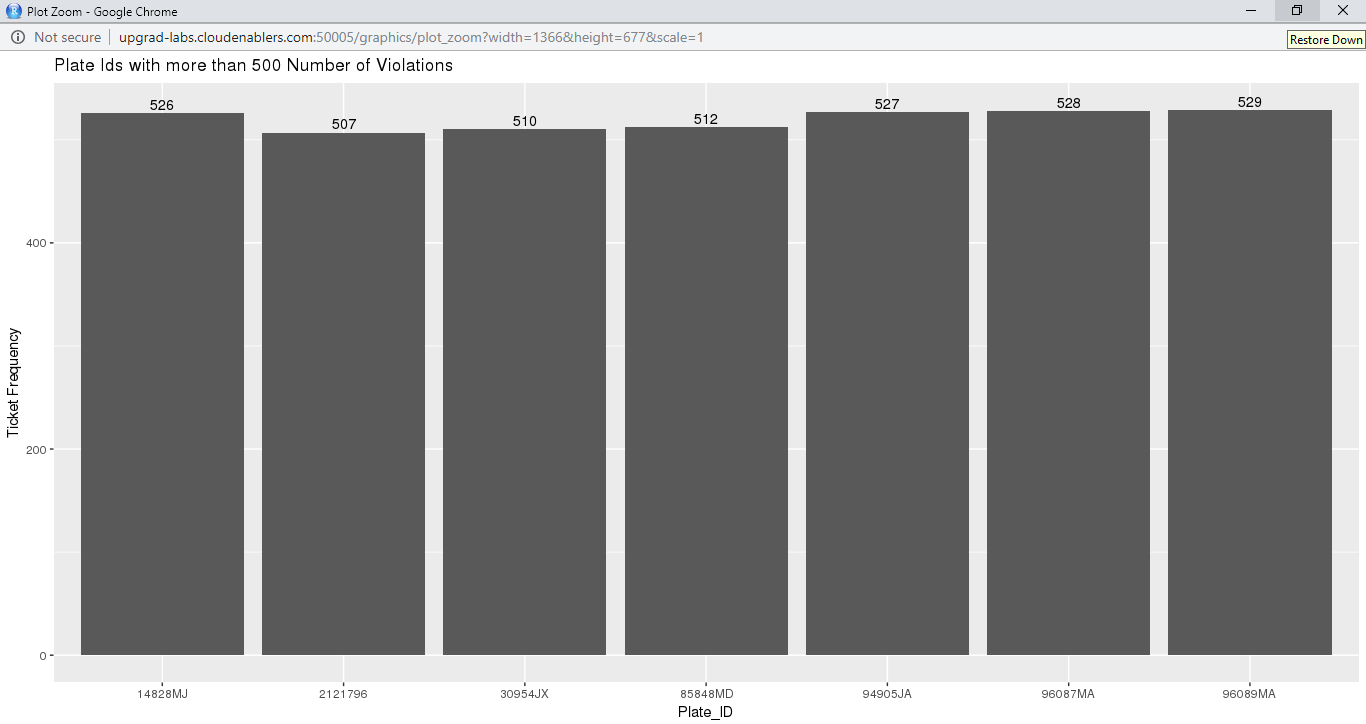
Total Number of Columns: 10

**DATA CLEANING.**

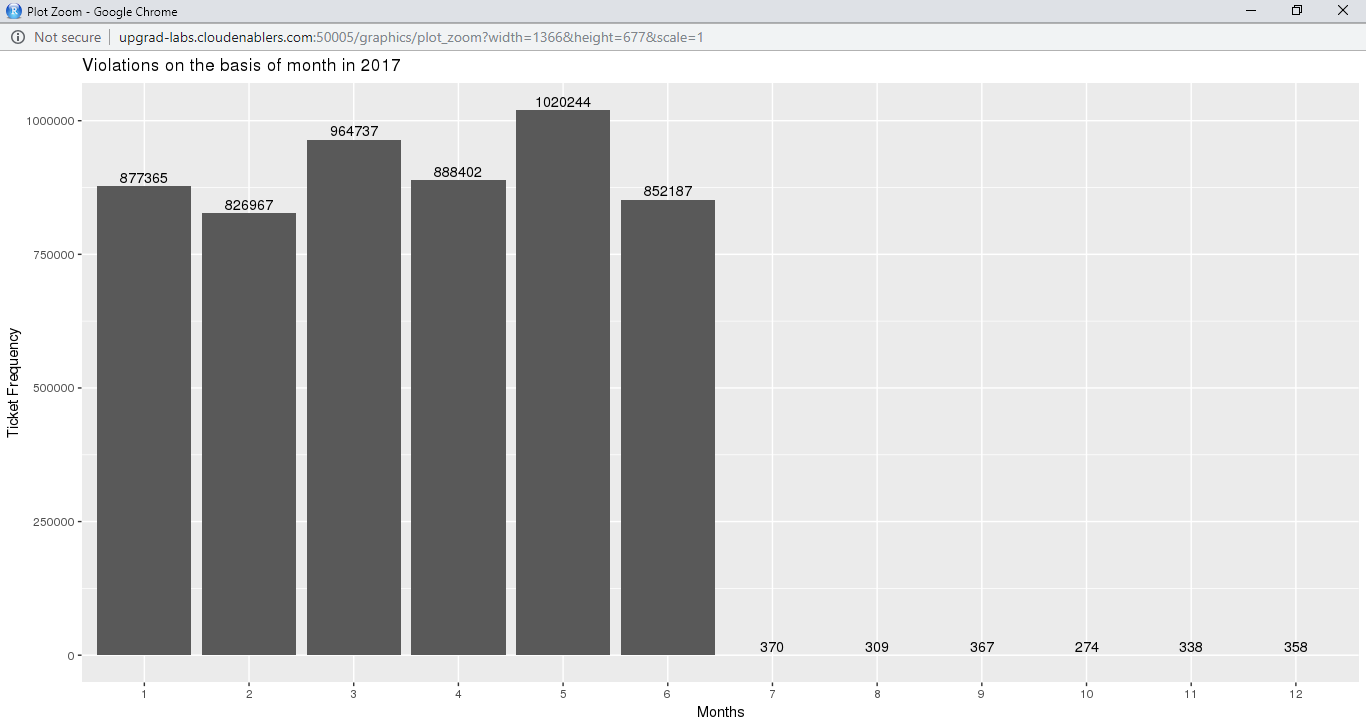
1. We have realized that the given data set contains data for 55 distinct years. Per the given objective, the scientific analysis is to be restricted for the data pertaining to the year 2017 hence we have filtered the data accordingly.
2. We have extracted two new columns Issue\_Year and Issue\_Month from the field Issue\_Date.
3. We have observed that there are no NULL value in any of the column.
4. For certain Data serials, Plate\_ID contains a value-'BLANKPLATE' which has not been considered in our analysis.

**OBSERVATIONS.**

1. There are around 200 Plate\_Ids that have been booked for parking violation for more than a 100 times in the year 2017.
2. Data Analysis reveals (through Plate\_Id) that there are exactly seven cars that have been booked for parking violation for more than 500 times in the year 2017.



3. Data Analysis reveals that there is a significant drop in number of parking violations after the month of June in the year 2017.



Finally total number of rows :5426657, columns:12

**Examine the data**

1. Find the total number of tickets for the year.

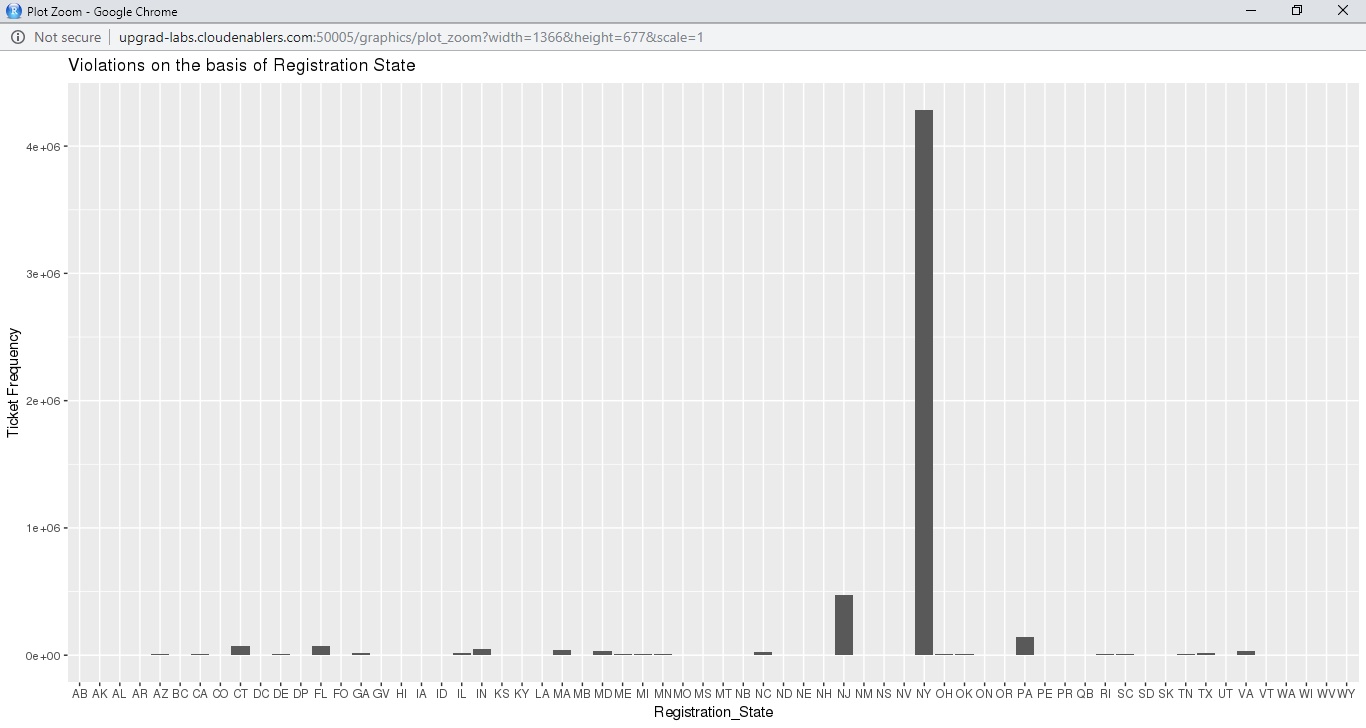
**Answer** Total number of tickets in year 2017 = **5426657**

In our data set, Summons\_Number is the unique field. As after data cleaning, number of rows are 5426657 which is same as number of tickets in the year 2017 that means there is no duplicate Summons\_Number.

1. Find out the number of unique states from where the cars that got parking tickets came from. (Hint: Use the column 'Registration State')  
   There is a numeric entry '99' in the column which should be corrected. Replace it with the state having maximum entries. Give the number of unique states again.

**Answer** There are 65 unique Registration\_State including 99 as one of the Registration\_State. It is specified that we have to replace it with the state having maximum entries i.e. NY. After replacing 99 with NY:-

Number of unique Registration States: **64**



**Aggregation tasks**

1. How often does each violation code occur? Display the frequency of the top five violation codes.

**Answer**- There are 100 distinct violation codes. The frequency of top 5 violation codes is as follows:

Violation\_Code Ticket\_Frequency

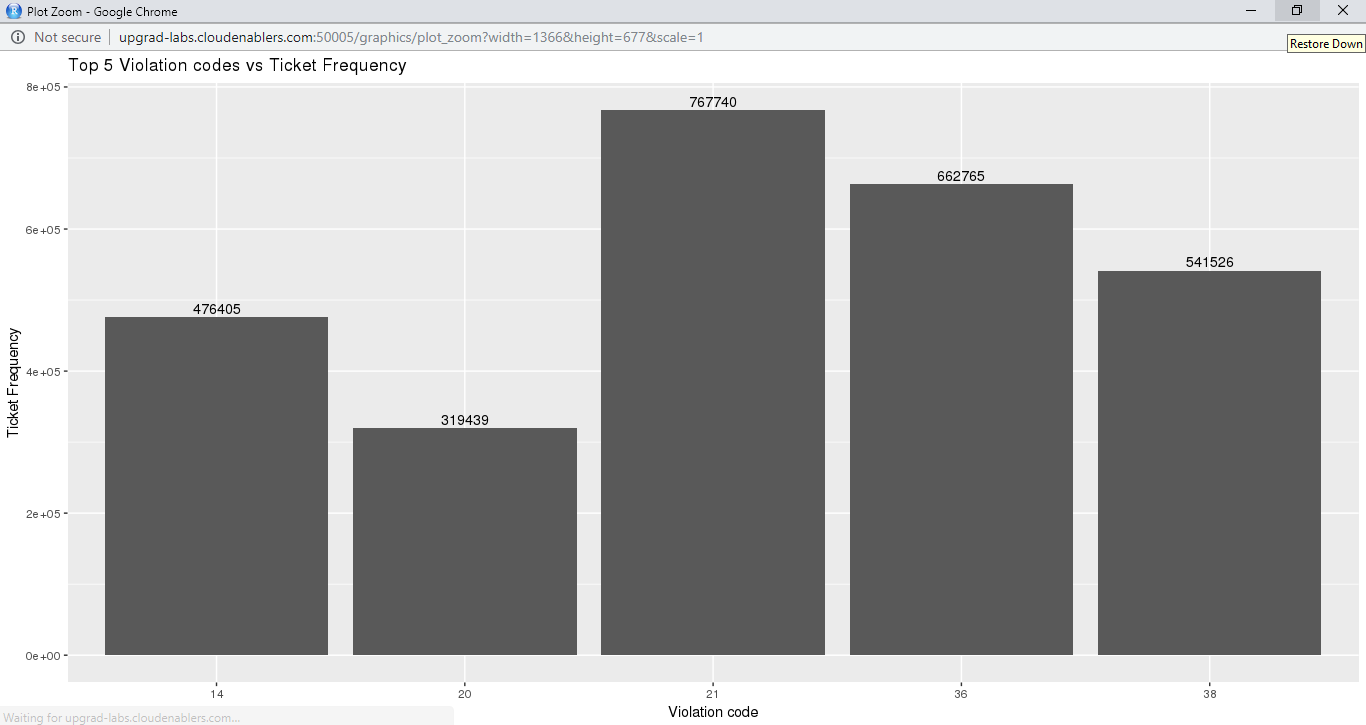
1 21 767740

2 36 662765

3 38 541526

4 14 476405

5 20 319439



1. How often does each 'vehicle body type' get a parking ticket? How about the 'vehicle make'? (Hint: find the top 5 for both)

**Answer**-

Vehicle\_Body\_Type Ticket\_Frequency

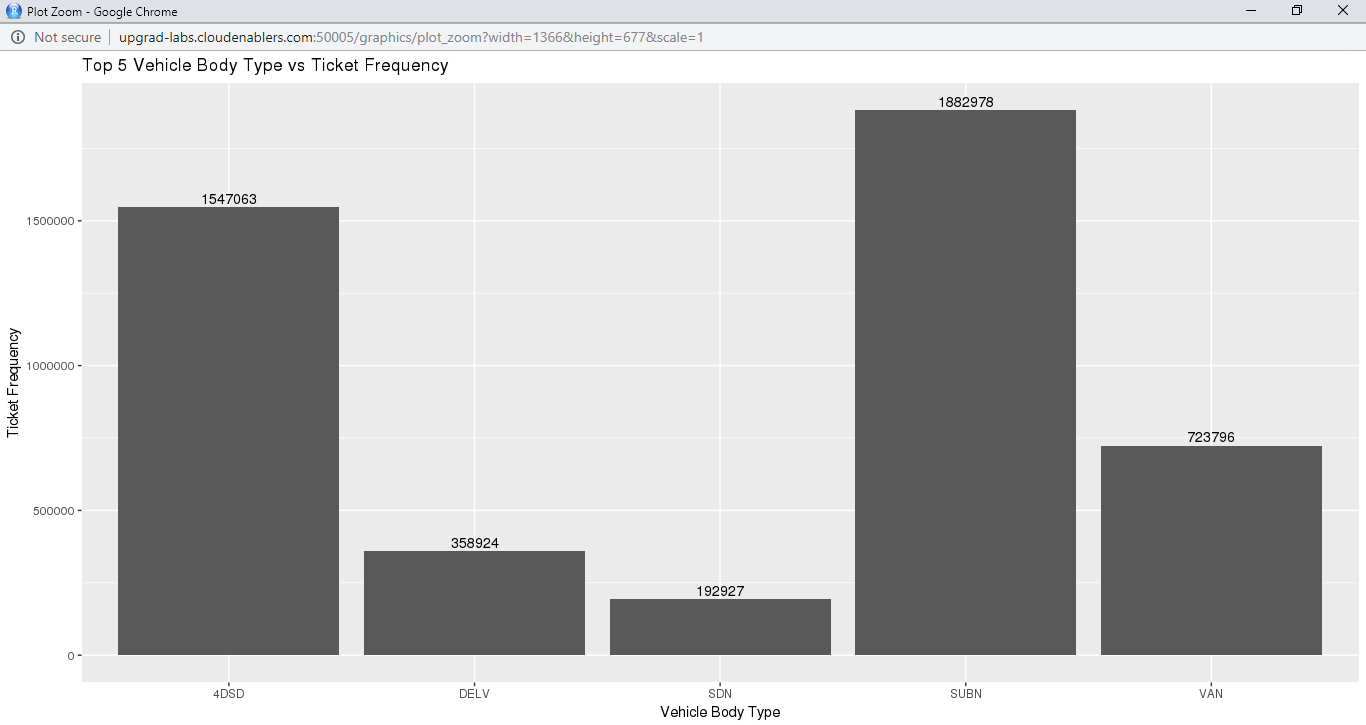
1 SUBN 1882978

2 4DSD 1547063

3 VAN 723796

4 DELV 358924

5 SDN 192927



Vehicle\_Make Ticket\_Frequency

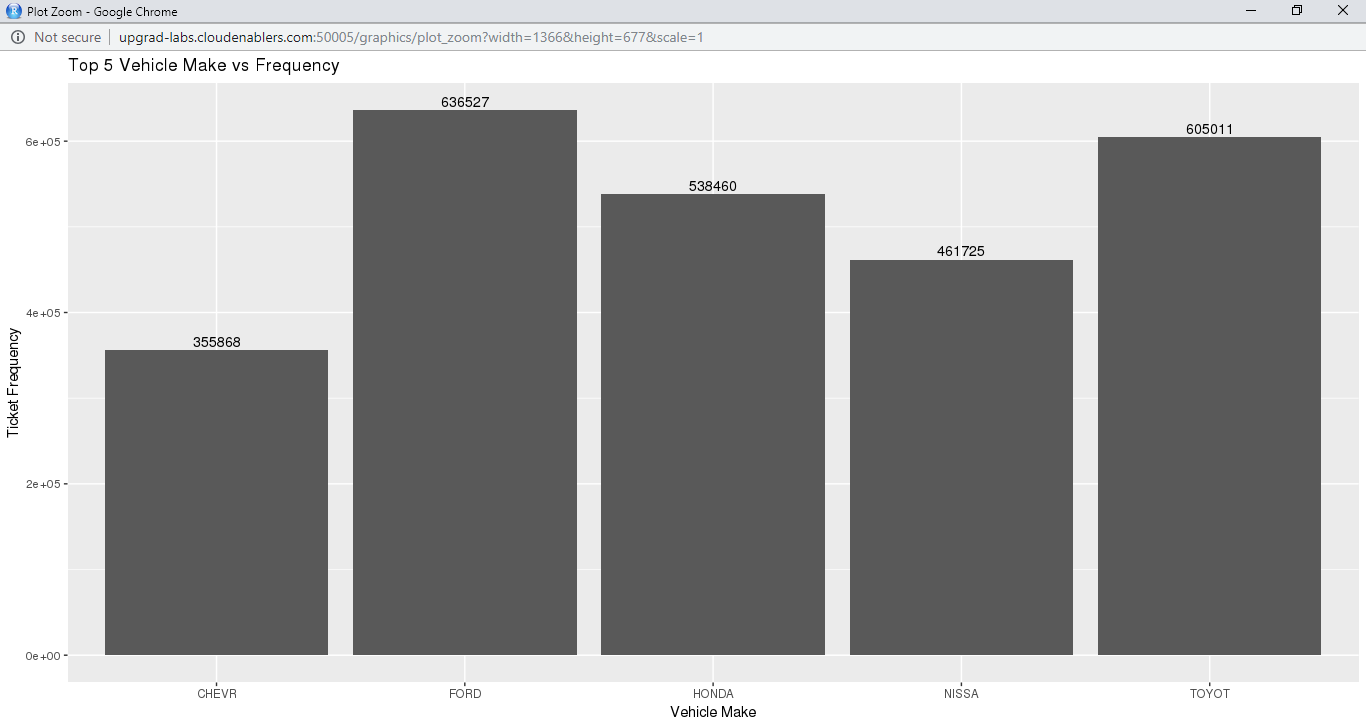
1 FORD 636527

2 TOYOT 605011

3 HONDA 538460

4 NISSA 461725

5 CHEVR 355868



1. A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequency of tickets for each of the following:'Violation Precinct' (this is the precinct of the zone where the violation occurred). Using this, can you make any insights for parking violations in any specific areas of the city?

**Answer**

Violation\_Precinct Ticket\_Frequency

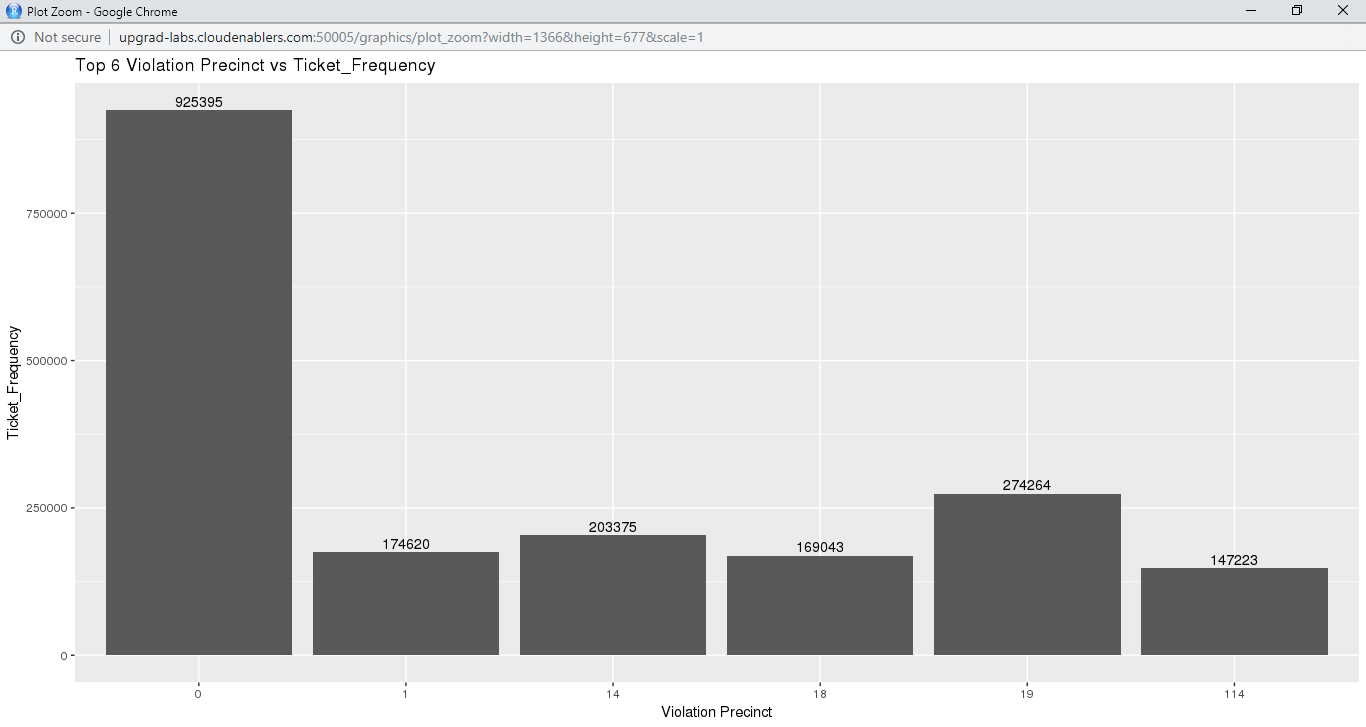
1 0 925395

2 19 274264

3 14 203375

4 1 174620

5 18 169043



'Issuer Precinct' (this is the precinct that issued the ticket)  
Here you would have noticed that the dataframe has 'Violating Precinct' or 'Issuing Precinct' as '0'. These are the erroneous entries. Hence, provide the record for five correct precincts. (Hint: Print top six entries after sorting)

Issuer\_Precinct Ticket\_Frequency

1 0 1077884

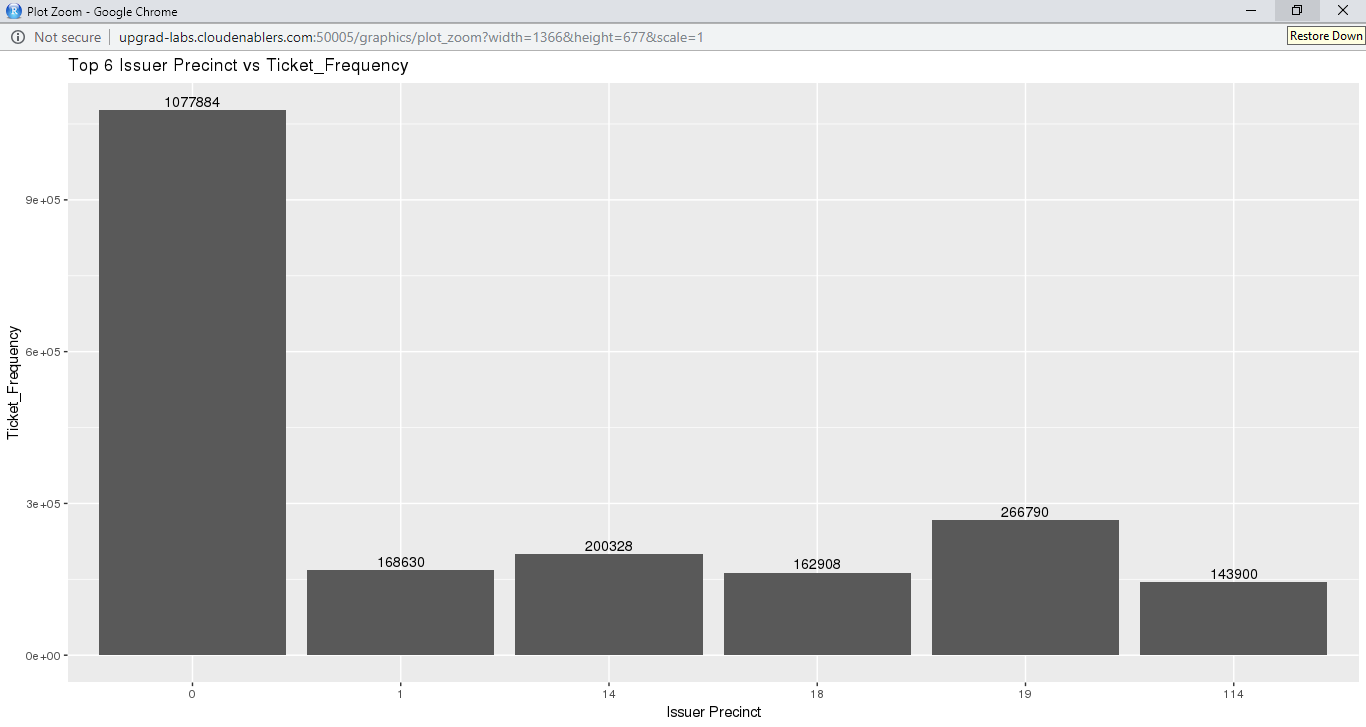
2 19 266790

3 14 200328

4 1 168630

5 18 162908

6 114 143900



1. Find the violation code frequency across three precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes? Are these codes common across precincts?   
   Hint: In the SQL view, use the 'where' attribute to filter among three precincts.

**Answer**

The three precincts which have issued the most number of tickets are: **18,19,14.**

The top 5 violation codes that are common in these precincts:

Violation\_Code Frequency

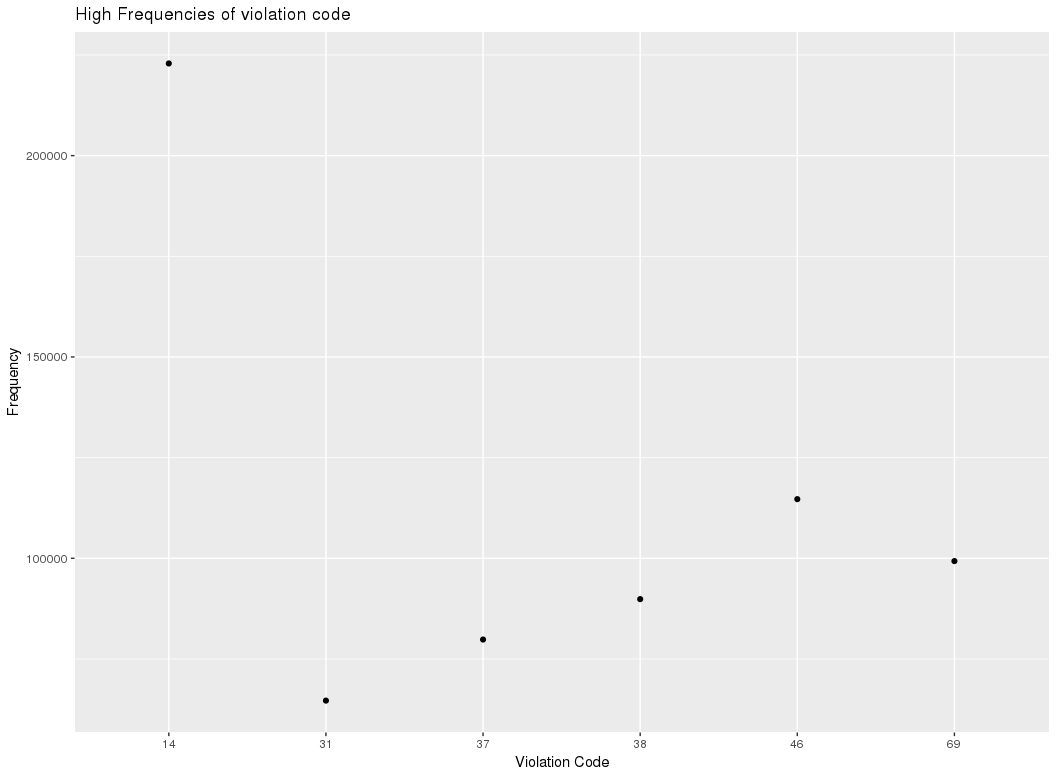
1 14 222885

2 46 114693

3 69 99315

4 38 89863

5 37 79829



1. You’d want to find out the properties of parking violations across different times of the day:Find a way to deal with missing values, if any.  
   Hint: Check for the null values using 'isNull' under the SQL. Also, to remove the null values, check the 'dropna' command in the API documentation

The Violation Time field is specified in a strange format. Find a way to make this into a time attribute that you can use to divide into groups.

Divide 24 hours into six equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the three most commonly occurring violations.  
Hint: Use the CASE-WHEN in SQL view to segregate into bins. For finding the most commonly occurring violations, a similar approach can be used as mention in the hint for question 4.

Now, try another direction. For the three most commonly occurring violation codes, find the most common time of the day (in terms of the bins from the previous part)

**Answer**

Bin 1

Violation\_Code Violation\_Time\_bin count

1 21 1 36934

2 40 1 25810

3 78 1 15513

Bin 2

Violation\_Code Violation\_Time\_bin count

1 14 2 74095

2 40 2 60609

3 21 2 57876

Bin 3

Violation\_Code Violation\_Time\_bin count

1 21 3 597827

2 36 3 348165

3 38 3 176310

Bin 4

Violation\_Code Violation\_Time\_bin count

1 36 4 286284

2 38 4 240511

3 37 4 166969

Bin 5

Violation\_Code Violation\_Time\_bin count

1 38 5 102784

2 14 5 75856

3 37 5 70317

Bin 6

Violation\_Code Violation\_Time\_bin count

1 7 6 26293

2 40 6 22301

3 14 6 21026

**6) Let’s try and find some seasonality in this data**

**First, divide the year into some number of seasons, and find frequencies of tickets for each season. (Hint: Use Issue Date to segregate into seasons)**

**Then, find the three most common violations for each of these seasons.**

**Answer:**

**Divided into two season based on number of violation.**

**S1 1st 6 months with high violations.**

**S2 last 6 months with low violations.**

Distribution\_on\_season<- SparkR::sql("SELECT Issue\_month,violation\_code,count(\*) as Ticket\_Frequency FROM nyc\_parking\_view GROUP BY Issue\_month,violation\_code order by Ticket\_Frequency desc")

>Distribution\_on\_season$season<-ifelse(Distribution\_on\_season$Issue\_month<=6,"S1","S2")

**# b)find the three most common violations for each of these seasons.**

**# Top 3 violation for season-1**

> Distribution\_on\_S1<-where(Distribution\_on\_season,Distribution\_on\_season$season=="S1")

> head(Distribution\_on\_S1,3)

Issue\_month violation\_code Ticket\_Frequency season

1 5 21 154794 S1

2 5 36 138881 S1

3 3 21 135665 S1

**# Top 3 violation for season-2**

>Distribution\_on\_S2<-where(Distribution\_on\_season,Distribution\_on\_season$season=="S2")

> head(Distribution\_on\_S2,3)

Issue\_month violation\_code Ticket\_Frequency season

1 9 46 89 S2

2 8 21 84 S2

3 12 46 77 S2

7) Find the total amount collected for the three violation codes with maximum tickets. State the code which has the highest total collection.

>common\_Violation<- SparkR::sql("select Violation\_Code, count(\*) as Frequency from nyc\_parking\_view group by Violation\_Code order by Frequency desc")

> head(common\_Violation,3)

# Violation Code Frequency

#1 21 767740

#2 36 662765

#3 38 541526

**# Code with 21 had the highest collection.**

common\_Violation$fine<-ifelse(common\_Violation$Violation\_Code==21,common\_Violation$frequency\*55,common\_Violation$frequency\*50)

> head(common\_Violation,3)

# Violation\_Code Frequency fine

#1 21 767740 42225700

#2 36 662765 33138250

#3 38 541526 27076300

**What can you intuitively infer from these findings?**

**#Jan to June had the major violation & July to Dec has a drastic drop.**

**#Highest violation &collection was by Code-21(No parking where parking is not allowed by sign, street marking or traffic control device.)**

sparkR.stop()