NYC Parking Tickets: An Exploratory Analysis

Introduction

The purpose of the case study is to perform exploratory data analysis on the parking ticket data collected by the New York Police Department using Spark.

The scope of the analysis is 2015, 2016 and 2017. We have considered the fiscal year as per the files.

The analysis has been performed on RStudio on the Corestack cluster using the SparkR library.

Data cleaning:

There are three csv files, each for 2015, 2016 and 2017. The size of each file is as follows:

2015	11809233 rows	51 columns
2016	10626899 rows	51 columns
2017	10803028 rows	43 columns

The following data issues have been addressed:

1. Column names have spaces between them. Therefore, the spaces are removed.
2. Columns with null values are removed.

TimeFirstObserved
IntersectingStreet
ViolationLegalCode
UnregisteredVehicle?

Examine the data:

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

2015	11809233

2016	10626899
2017	10803028

The number of parking tickets was maximum in 2015. It reduced by 10% in 2016 and increased by 1.6% in 2017.

2. Find out the number of unique states from where the cars that got parking tickets came from.
   The state with the maximum tickets is NY for all three years.
   The number of states for which tickets were issued in 2015 was 68, 2016 was 67 and 2017 was 66.
3. Some parking tickets don't have the address for violation location on them, which is a cause for concern. Write a query to check the number of such tickets.
   The number of tickets without violation location is as follows:

2015	1799170
2016	1868656
2017	2072400

The number increased by 3.8% in 2016 and by 10.9% in 2017. This is a steep increase.

Aggregation tasks

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

   The frequency of violation codes in 2015 is as follows:

Violation Code	Count
21	1630912
38	1418627
14	988469
36	839197
37	795918

The frequency of violation codes in 2016 is as follows:

Violation Code	Count
21	1531587
36	1253512
38	1143696
14	875614
37	686610

The frequency of violation codes in 2017 is as follows:

Violation Code	Count
21	1528588
36	1400614
38	1062304
14	893498
37	618593

Wrong Parking (21), Failing to show a parking ticket (38) and Speeding over the limit (36) are consistently high in all three years.

2. How often does each 'vehicle body type' get a parking ticket? How about the 'vehicle make'?

   The vehicle body type and number of parking tickets for 2015 is shown below:

Vehicle Body Type	Count
SUBN	3729346
4DSD	3340014
VAN	1709091
DELV	892781
SDN	524596

Vehicle make and number of parking tickets for 2015 is shown below:

Vehicle Make	Count
FORD	1521874
TOYOT	1217087
HONDA	1102614
NISSA	908783
CHEVR	897845

The vehicle body type and number of parking tickets for 2016 is shown below:

Vehicle Body Type	Count
SUBN	3466037
4DSD	2992107
VAN	1518303
DELV	755282
SDN	424043

Vehicle make and number of parking tickets for 2016 is shown below:

Vehicle Make	Count
FORD	1324774
TOYOT	1154790
HONDA	1014074
NISSA	834833
CHEVR	759663

The vehicle body type and number of parking tickets for 2017 is shown below:

Vehicle Body Type	Count
SUBN	3719802

4DSD	3082020
VAN	1411970
DELV	687330
SDN	438191

Vehicle make and number of parking tickets for 2017 is shown below:

Vehicle Make	Count
FORD	1280958
TOYOT	1211451
HONDA	1079238
NISSA	918590
CHEVR	714655

Observation: Sub-Urban and 4 Door Sedans consistently got the highest tickets in all three years. From a make point of view, Ford , Toyota and Honda makes attracted the highest tickets. This could possibly be because of the distribution of total no of cars in US

3. 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?

      The data for 2015 is as follows:

Violation Precinct	Count
19	598351
18	427510
14	409064
1	329009
114	320963

The data for 2016 is as follows:

Violation Precinct	Count
19	554465
18	331704
14	324467
1	303850
114	291336

The data for 2017 is as follows:

Violation Precinct	Count
19	535671
14	352450
1	331810
18	306920
114	296514

Based on the above observations in the 3 years, we can say that there were many parking violations in Precincts 19, 14, 1, 18 and 114.

- 'Issuer Precinct' (this is the precinct that issued the ticket)

The data for 2015 is as follows:

Issuer Precinct	Count
19	579998
18	417329
14	392922
1	318778
114	314437
13	296403

The data for 2016 is as follows:

Issuer Precinct	Count

19	540569
18	323132
14	315311
1	295013
114	286924
13	282635

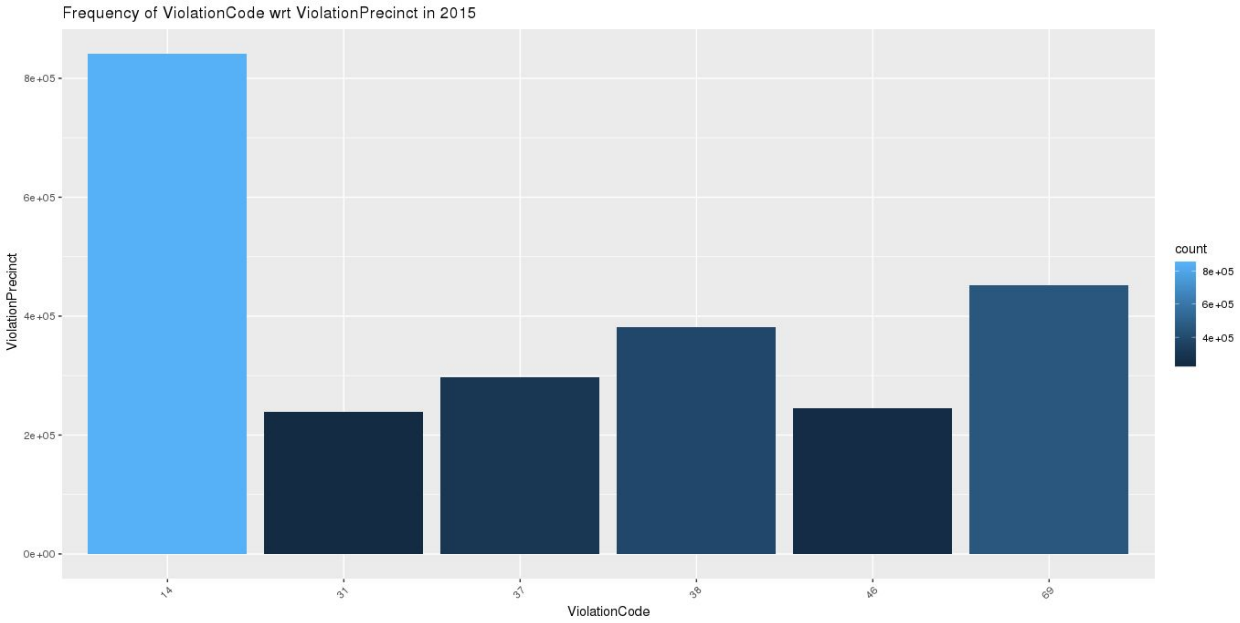
The data for 2017 is as follows:

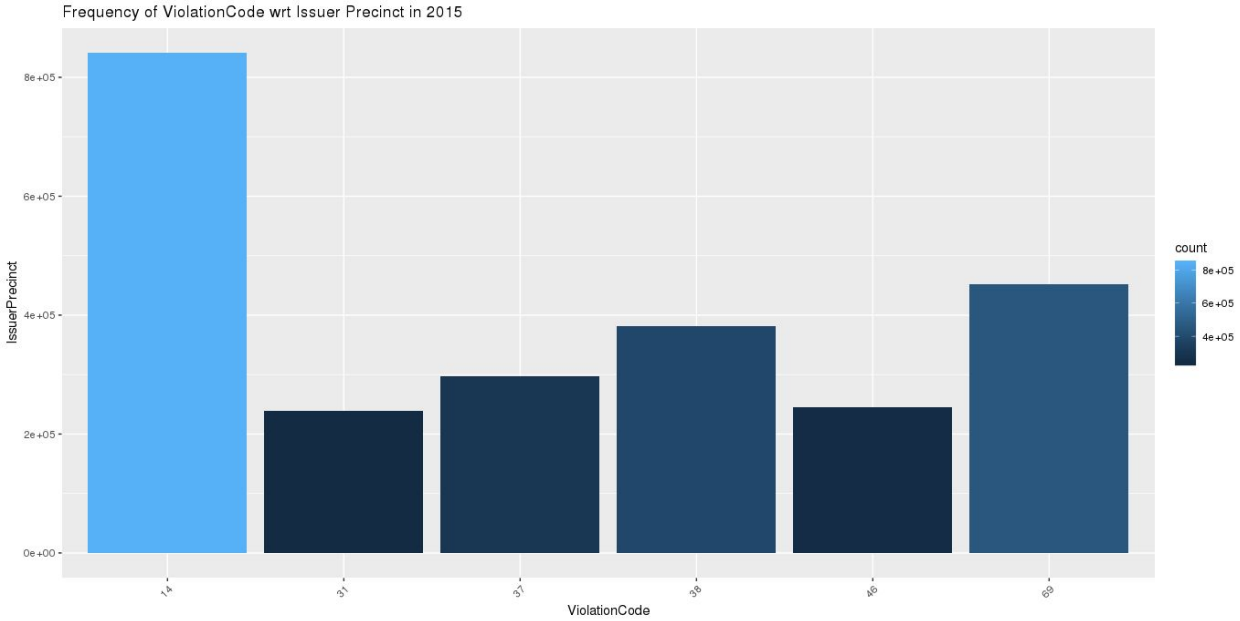
Issuer Precinct	Count
19	521513
14	344977
1	321170
18	296553
114	289950
13	240833

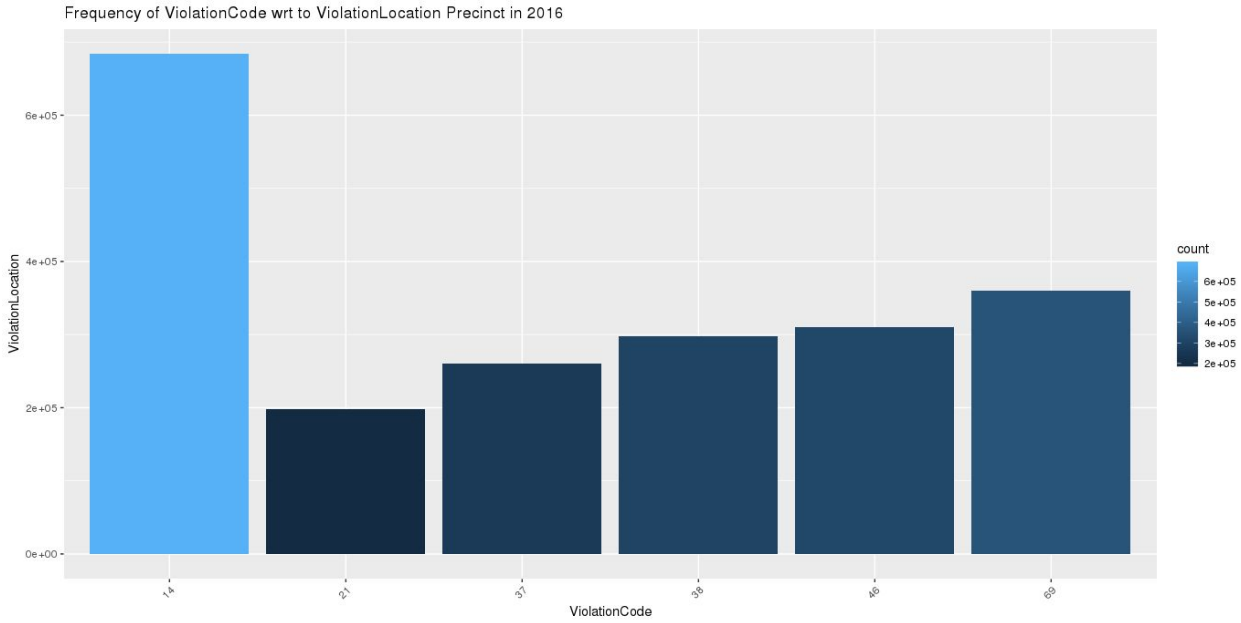
4. Find the violation code frequency across three precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes? Are these codes common across precincts?

   From the results of Question 3, it is evident that precincts 19,18 and 14 issued most number of tickets.

   Violation codes 14, 69, 46, 38 are common across precincts


Frequency of ViolationCode wrt ViolationPrecinct in 2015
ViolationPrecinct
8e+05
6e+05
4e+05
2e+05
0e+00
14
31
37
38
46
69
ViolationCode
count
8e+05
6e+05
4e+05


Frequency of ViolationCode wrt Issuer Precinct in 2015
IssuerPrecinct
8e+05
6e+05
4e+05
2e+05
0e+00
14
31
37
38
46
69
ViolationCode
count
8e+05
6e+05
4e+05



5. You'd want to find out the properties of parking violations across different times of the day:
    - 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.

      The data for 2015 is as follows:

12am to 4am	21,38,14
4am to 8am	21,14,19
8am to 12pm	21,38,14
12pm to 4pm	38,37,14
4pm to 8pm	38,37,14
8pm to 12am	40,38,14

The data for 2016 is as follows:

12am to 4am	21,38,14
4am to 8am	21,14,20
8am to 12pm	21,38,14
12pm to 4pm	38,37,14

4pm to 8pm	38,37,14
8pm to 12am	40,38,14

The data for 2017 is as follows:

12am to 4am	21,38,14
4am to 8am	21,14,19
8am to 12pm	21,38,14
12pm to 4pm	38,37,14
4pm to 8pm	38,37,14
8pm to 12am	40,38,14

Across 2015 , 2016 and 2017, the 3 most common violations are 21,38,14

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.

      For 2015, 2016 and 2017, from the graphs, we have data only for spring, winter and summer. The 3 most common violations in seasons are 21, 38 and 14.
7. The fines collected from all the parking violation constitute a revenue source for the NYC police department.
    - Violation Code 14 has the highest collection in 2015
    - Violation Code 21 has the highest collection in 2016
    - Violation Code 21 has the highest collection in 2017