New York City Parking Violations

Exploratory Data Analysis using SparkR

Submitted by

- Rohit Sipani
- Prabhakar Kalaiselvan
- Raghu Teja
- Natarajan Ganapathi

1 Data Understanding

Objective of the data analysis is to understand the trends, causes, patterns in parking tickets issues to different vehicles in NYC over the period of 3 years with combination of data analysis using SparkR and plots using ggplot2 package.

Our Understanding of the Data and the Domain is summarized below.

1.1 Business Understanding

NYC police have been issuing parking tickets for different types of parking violations done by vehicle drivers in the city of New York and outskirts.

- There are around 70+ types parking violations which are eligible for fines.
- Fines charge for a particular violation depends on whether the address where violation happened, where parking offences in the Manhattan attracting higher penalties compared to other localities within NYC.
- The tickets are issued by different precincts, which is police station responsible for a certain locality/street within NYC.
- Some of the fine amounts also depends on whether the offence is first time or repeat offence.

Files for 2015 and 2016 had 51 Columns each whereas File for 2017 had only 43 columns.

1.2 Data Understanding

- Data for the parking tickets issues have been provided in 3 different files, with one each for each year.
- Files for 2015 and 2016 had 51 Columns each whereas File for 2017 had only 43 columns.
- Files are provided in the fiscal year format
    - *2015 File: From 01 July 2014 to 30 June 2015*
    - *2016 File: From 01 July 2015 to 30 June 2016*
    - *2017 File: From 01 July 2016 to 30 June 2017*
- Duplicate values are found in 2015 files. These are removed from the files for further analysis.
- Date apart from fiscal year are observed in all three (2015, 2016, 2017) files. These are removed from the files for further analysis.

Year	Initial Rows	After Removing Duplicates and Additional Dates
2015	11809233	10598036
2016	10626899	10396984
2017	10803028	10539563

2 Data Quality Issues & Assumptions

2.1 Data Quality Issues Observed

Data Quality issues identified in the data set are mentioned below.

- Files for 2015 and 2016 had 51 Columns each whereas File for 2017 had only 43 columns.
- Duplicate values found in the 2015 and 2017 files. These are removed from the files for further analysis.
- Date apart from fiscal year period are observed in all three (2015, 2016, 2017) files. These are removed from the files for further analysis.
- Street Address (Home Number or Street name) are not found in some of records in each file. However they are not excluded from Analysis.
- Precinct value is found to be 0 for many of the records in all the files. Upon looking up valid values for Precinct on NYC.gov website, we understand that 0 is not a valid value and possibly indicate absence of precinct information, indicating that the same is not captured in the Tickets data.
- There are many variations and out of range values observed in the Issue Time field of all three files.

2.2 Assumptions for Data Analysis

Following Assumptions have been made as a part of the Data Analysis.

- All our analysis done based on the Fiscal Year. In any of the analysis results/plots, year 2015 indicate the fiscal year starting from 2014-Jul to 2015-Jun and so on.
- For Issue regarding street address, we have assumed that if either street name OR house number is not present, then we consider that as an issue. Our reported numbers are based on this assumption.
- Regarding the NYC Fines data, we have taken the average price as mentioned in the case study guidelines.
- For some offences, there is a variation between a first time offence and repeat offences. For this analysis, we have taken the first time offence fees as the value for simplicity.
- There are many patterns found in the Issue Time field, related to format of the Time stamp column. We have documented the detailed understanding and our approach as a part of Chapter 3 and Section 5.

3 Observations & Insights

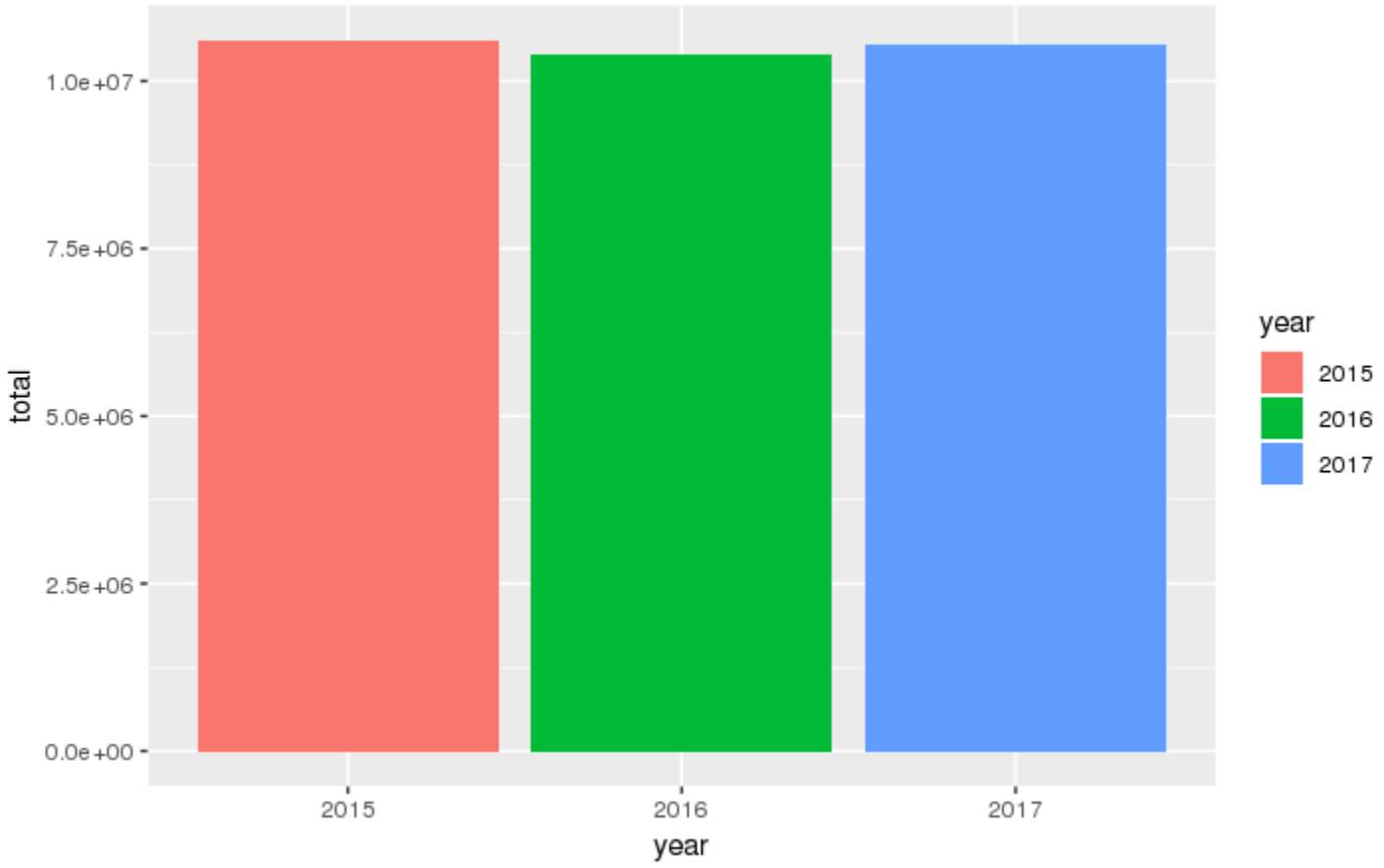
3.1 Examine the data

Note: The results are after removing the duplicates and additional dates from each of the files

1. Total Tickets each year

Please find the numbers.

Year	# Tickets
2015	10598035
2016	10396984
2017	10539563



Conclusion:

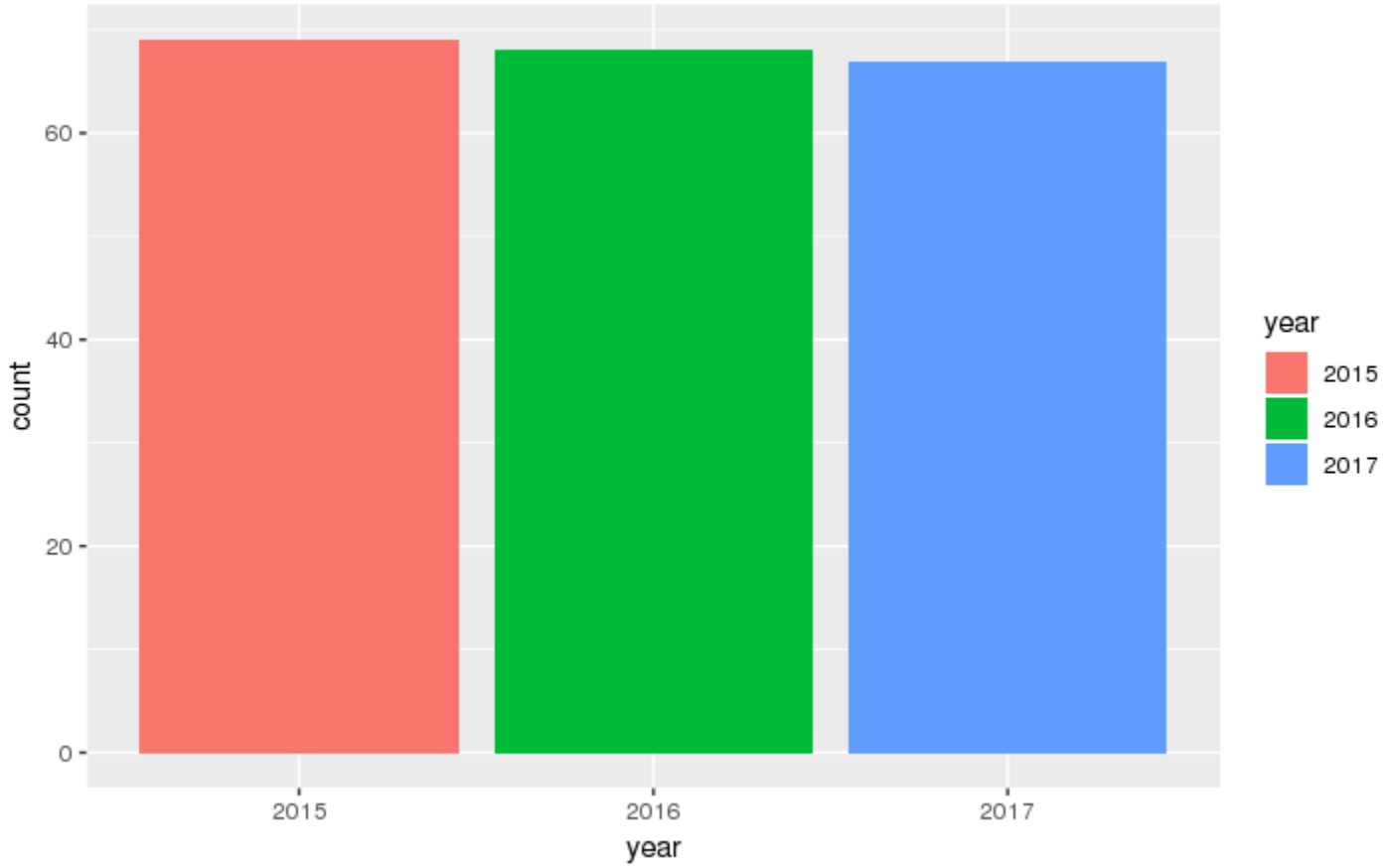
We could observe the following from above plot and table:

- Decrease in number of parking tickets during fiscal year 2016 while comparing with fiscal year 2015
- Increase in number of parking tickets during fiscal year 2017 while comparing with fiscal year 2016
- Increase of parking tickets is cause of concern and action to be taken for reducing it

2. Vehicles & States

Find out how many unique states the cars which got parking tickets came from.

Year	# Unique States
2015	69
2016	68
2017	67



Conclusion:

We could observe the following from above plot and table:

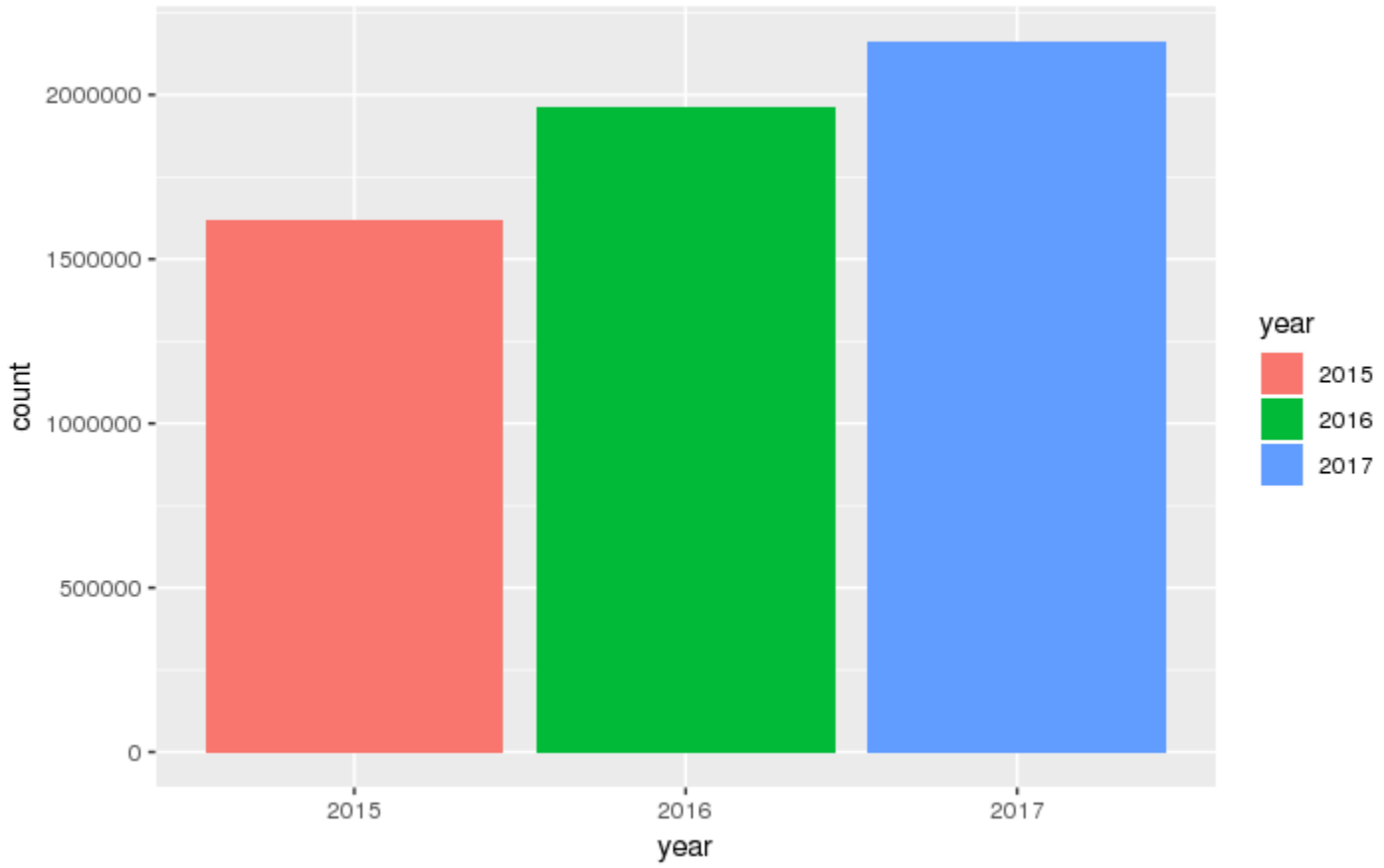
- Decreasing trend could be observed from fiscal year 2015 but the number of unique states decreased were minor

3. Tickets without address data

Some parking tickets don't have addresses on them, which is cause for concern. Find out how many such tickets there are.

Note: Either house name or street name missing was considered as missing address because address is incomplete without either of them

Year	# Tickets without Addresses
2015	1622076
2016	1963921
2017	2160639



Conclusion:
We could observe the following from above plot and table:

- Increasing trend could be observed from fiscal year 2015 in number of tickets without address
- Increase in parking tickets without address is cause of concern and action should be taken to have all the violation with address

3.2 Aggregation Tasks

1. Frequency of Violations

How often does each violation code occur? (Frequency of violation codes - find the top 5)

2015

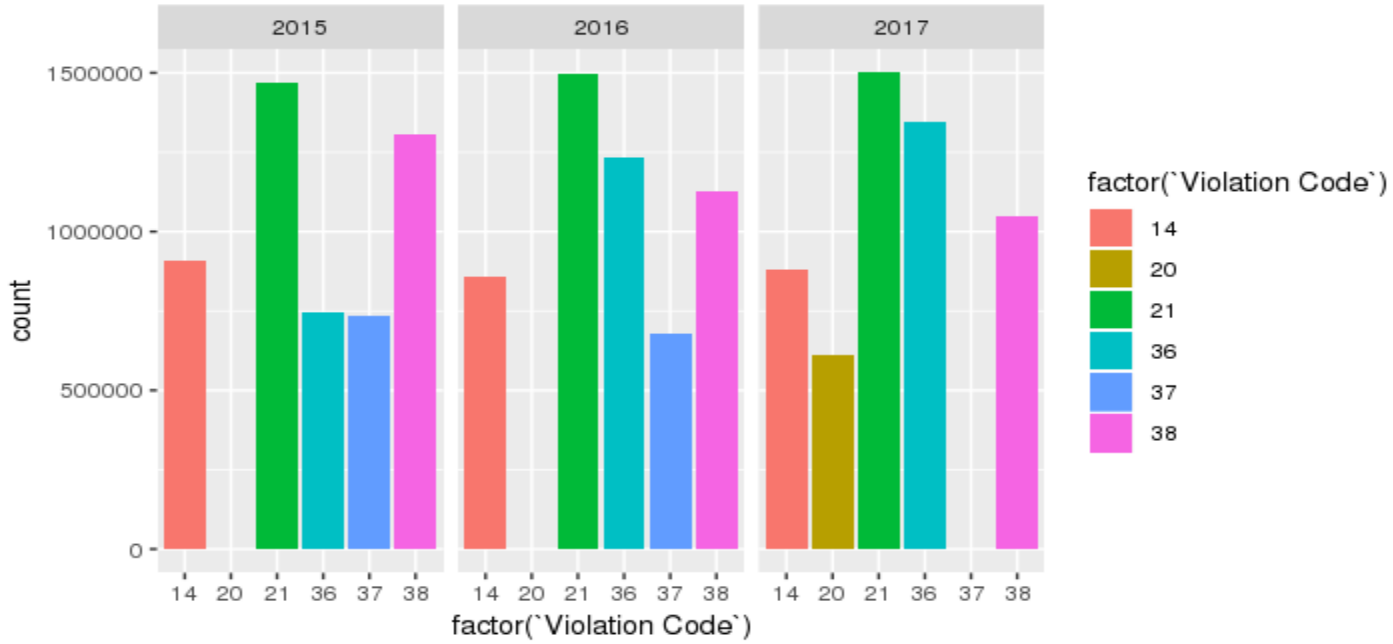
Violation Code	Frequency
21	1469228
38	1305007
14	908418
36	747098
37	735600

2016

Violation Code	Frequency
21	1497269
36	1232952
38	1126835
14	860045
37	677805

2017

Violation Code	Frequency
21	1500396
36	1345237
38	1050418
14	880152
20	609231



Conclusion:

We could observe the following from above plot and table:

- **"Street Cleaning**: No parking where parking is not allowed by sign, street marking or traffic control device" has the highest occurrence for violation for all the three fiscal years
- After street cleaning following violations are having highest occurrences in all three fiscal years
  - Exceeding the posted speed limit in or near a designated school zone.
  - Muni Meter
  - **General No Standing**: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.
- ➢ Action needs to be taken on reducing the highest occurrences violation which would reduce the number of parking tickets

2. Violations Frequency by Vehicle Body & Vehicle Make

How often does each vehicle body type get a parking ticket? How about the vehicle make? (Find the top 5 for both)

Vehicle Body Type

2015

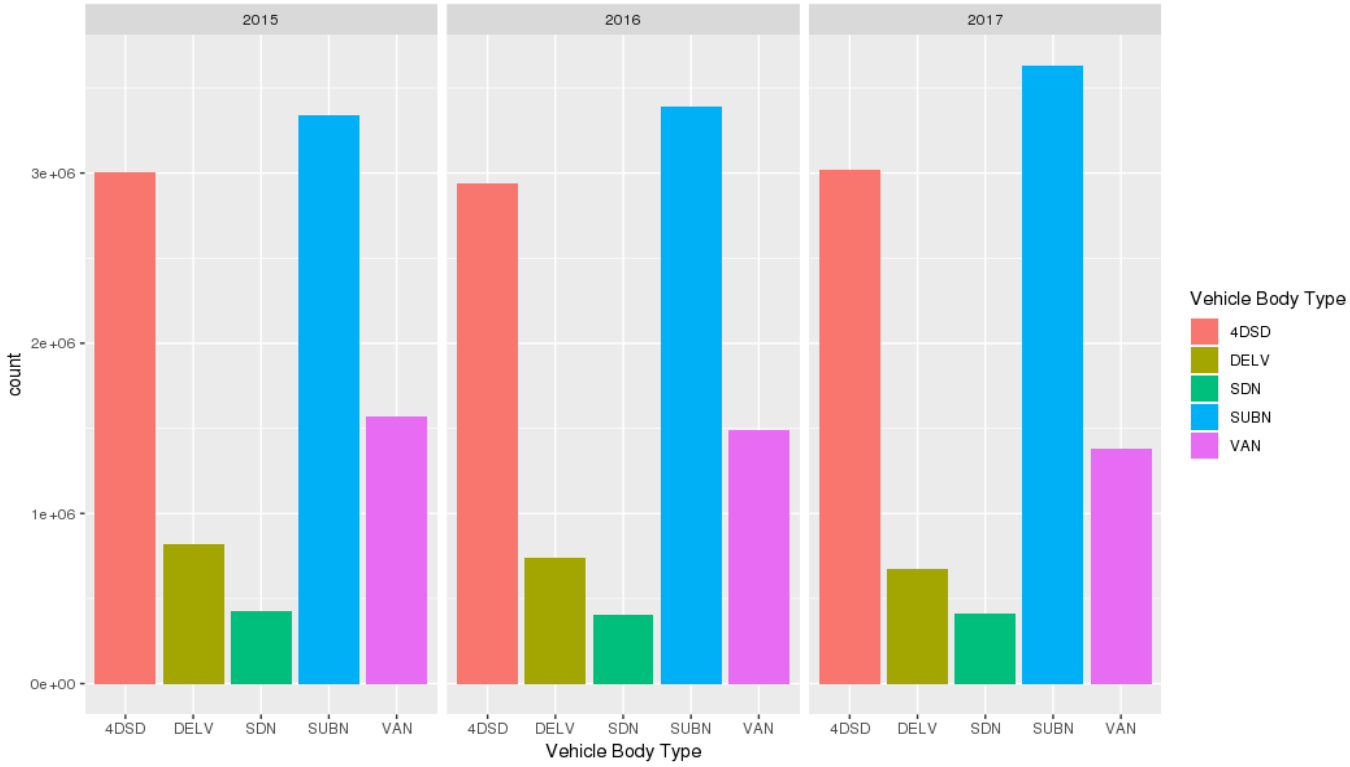
Vehicle Body	Frequency
SUBN	3341110
4DSD	3001810
VAN	1570227
DELV	822041
SDN	428571

2016

Vehicle Body	Frequency
SUBN	3393838
4DSD	2936729
VAN	1489924
DELV	738747
SDN	401750

2017

Vehicle Body	Frequency
SUBN	3632003
4DSD	3017372
VAN	1384121
DELV	672123
SDN	414984



Conclusion:

We could observe the following from above plot and table:

- SUBN, 4DSD, VAN, DELV and SDN in sequence are the top five vehicles body type for all the three fiscal years

Vehicle Make

2015

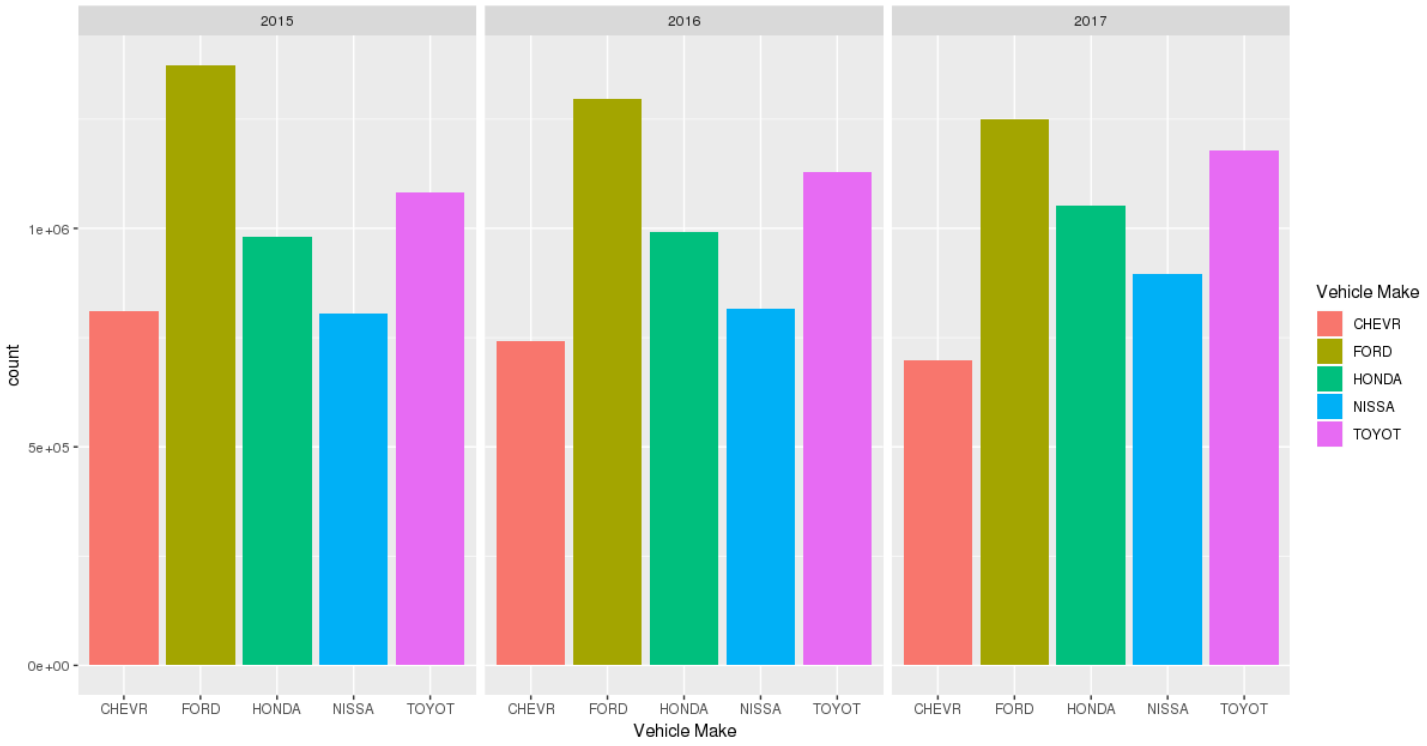
Vehicle Make	Tickets
FORD	1373157
TOYOT	1082206
HONDA	982130
CHEVR	811659
NISSA	805572

2016

Vehicle Make	Tickets
FORD	1297363
TOYOT	1128909
HONDA	991735
NISSA	815963
CHEVR	743416

2017

Vehicle Make	Tickets
FORD	1250777
TOYOT	1179265
HONDA	1052006
NISSA	895225
CHEVR	698024



Conclusion:

We could observe the following from above plot and table:

- FORD, TOYOT, HONDA in sequence are the top three vehicles make for all the three fiscal years
- CHEVR was holding fourth position and NISSA was holding fifth position in 2015 which was changed in 2016. CHEVR was moved to fifth position and NISSA came to fourth position in 2016.
- In 2017 the sequence remains the same as 2016

3. Analysis by Precinct

A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

a. Violating Precincts (this is the precinct of the zone where the violation occurred)
b. Issuing Precincts (this is the precinct that issued the ticket)

Violating Precincts

2015

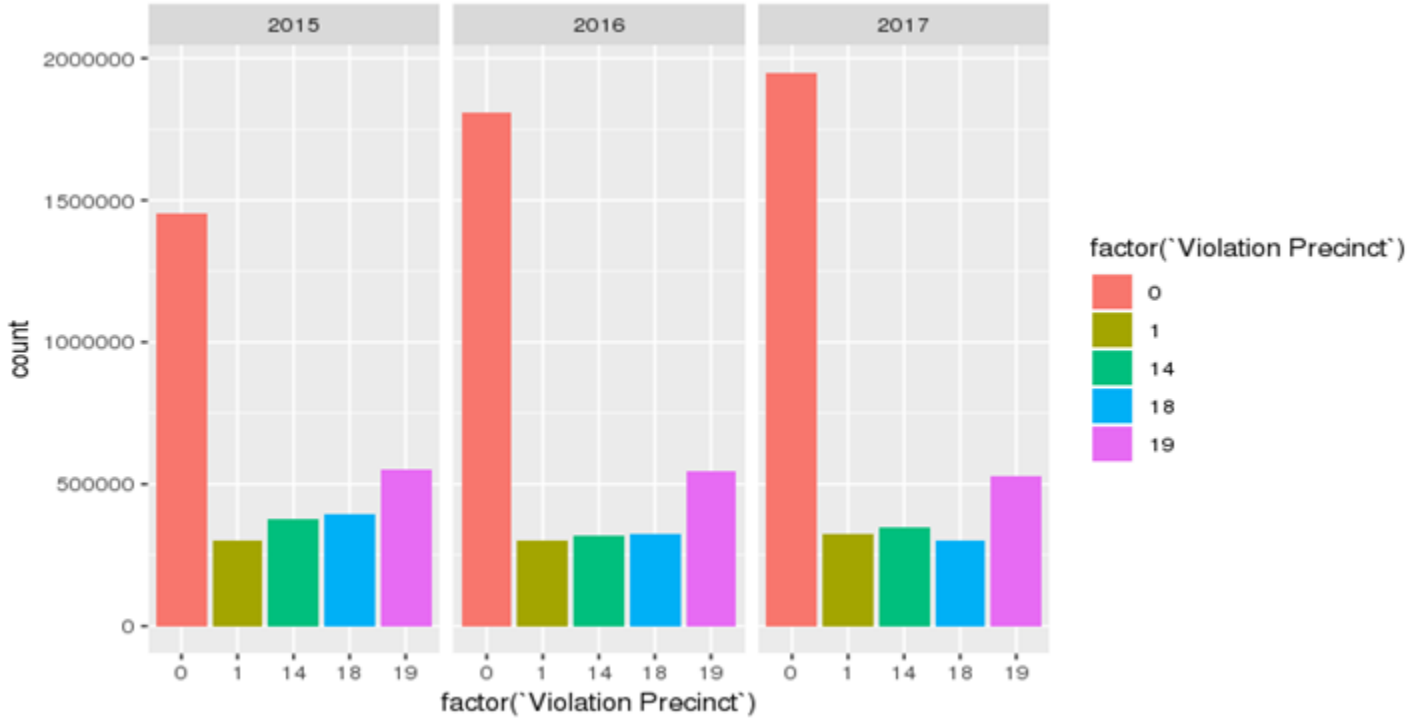
Violating Precincts	Frequency
0	1455166
19	550797
18	393802
14	377750
1	302737

2016

Violating Precincts	Frequency
0	1807139
19	545669
18	325559
14	318193
1	299074

2017

Violating Precincts	Frequency
0	1950083
19	528317
14	347736
1	326961
18	302008



Conclusion:

We could observe the following from above plot and table:

- Precinct 0, Precinct 19 in sequence are the top two violating precincts in all the three fiscal years
- Precinct 18 was holding the third position, Precinct 14 was holding the fourth position and Precinct 1 was holding fifth position in 2015 and 2016 which was changed in 2017. Precinct 18 was moved to fifth position, Precinct 14 came to third position and Precinct 1 came to fourth position in 2017

Issuing Precincts

2015

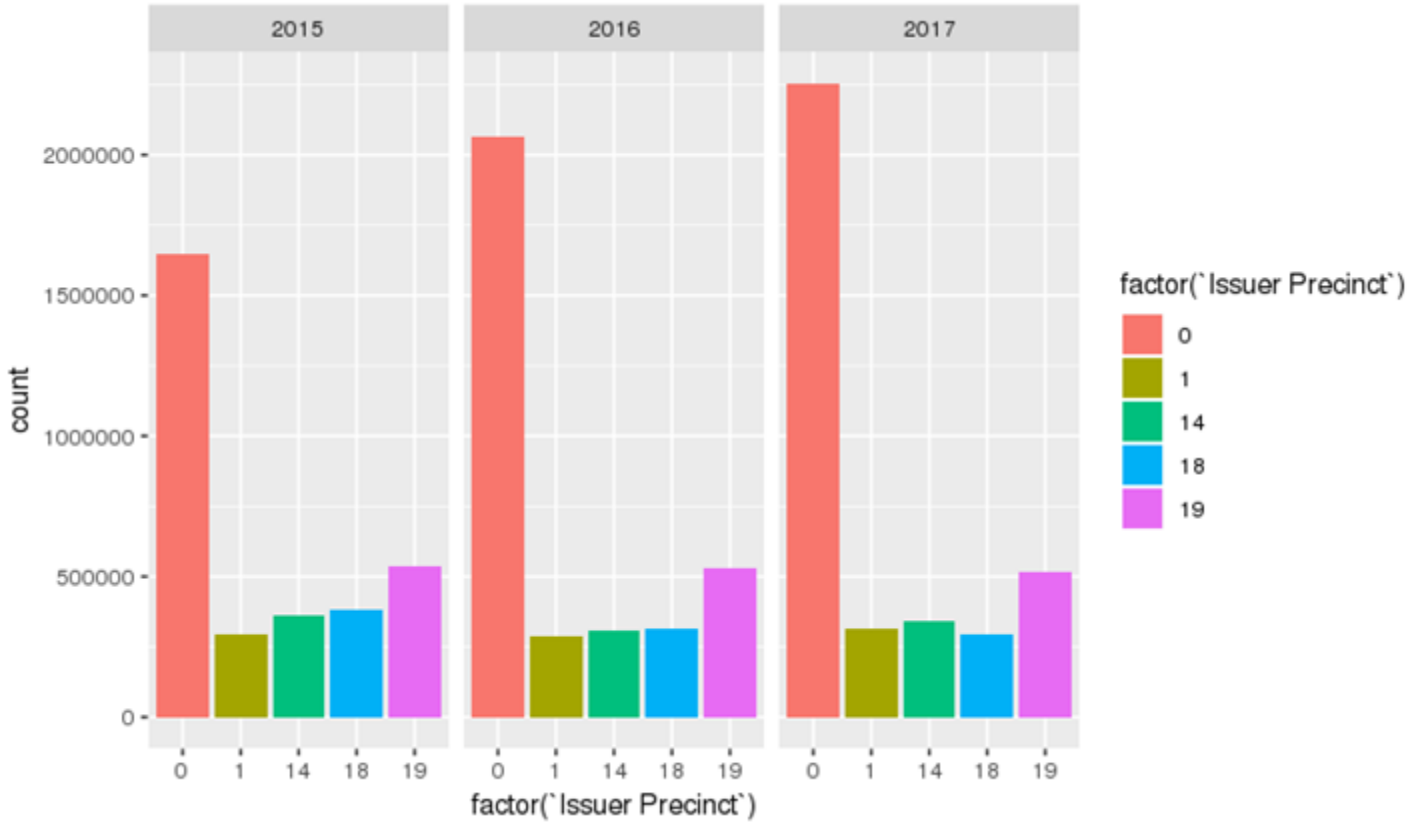
Issuing Precincts	Frequency
0	1648671
19	536627
18	384863
14	363734
1	293942

2016

Issuing Precincts	Frequency
0	2067219
19	532298
18	317451
14	309727
1	290472

2017

Issuing Precincts	Frequency
0	2255086
19	514786
14	340862
1	316776
18	292237



Conclusion:

We could observe the following from above plot and table:

- Issuing Precincts top 5 were same as the Violating Precincts for all the three fiscal years

4. Violations across top 3 precincts

Find the violation code frequency across 3 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?

Violation Code Frequency across 3 Precincts (Top 5)

2015

Violation Code	Frequency
36	747098
7	567953
21	232101
14	183000
5	127154

2016

Violation Code	Frequency
36	1232951
7	457871
21	287729
14	164816
5	106617

2017

Violation Code	Frequency
36	1345237
7	464691
21	314929
14	136728
5	130964

Conclusion:

We could observe the following from above table:

- All three fiscal years have the same violation code as top 5 occurrence category

Are These Codes Common Across Precincts?

2015

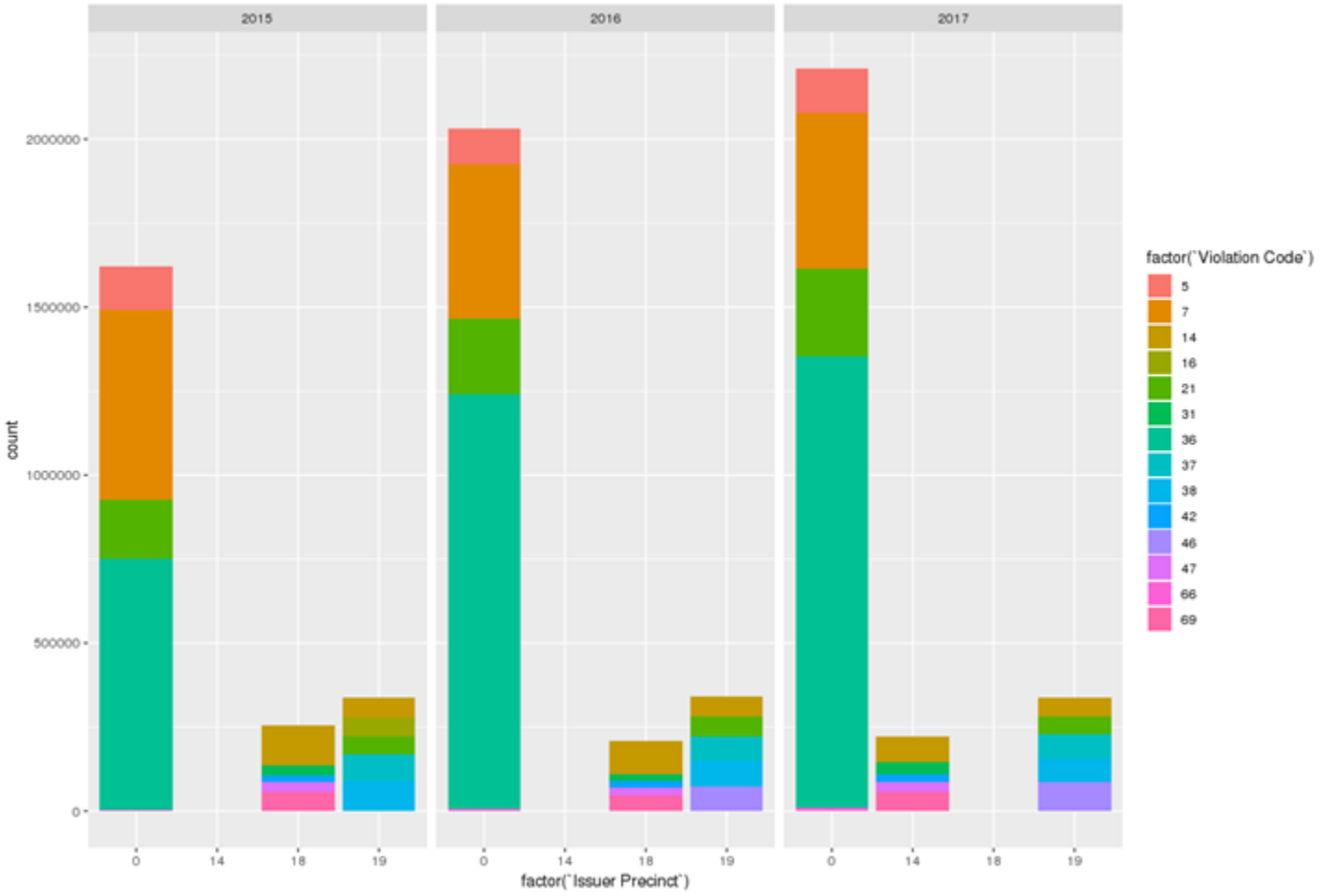
Issuer Precinct	Violation Code	Frequency
19	38	89102
19	37	78716
19	14	59915
19	16	55762
19	21	55296
0	36	747098
0	7	567951
0	21	173191
0	5	127153
0	66	4703
18	14	119078
18	69	56436
18	31	30030
18	47	28724
18	42	19522

2016

Issuer Precinct	Violation Code	Frequency
19	38	76178
19	37	74758
19	46	71509
19	14	60856
19	21	57601
0	36	1232951
0	7	457871
0	21	226687
0	5	106617
0	66	7275
18	14	98160
18	69	47129
18	47	23618
18	31	22413
18	42	17416

2017

Issuer Precinct	Violation Code	Frequency
19	46	84789
19	38	71631
19	37	71592
19	14	56873
19	21	54033
14	14	73007
14	69	57316
14	31	39430
14	47	30200
14	42	20402
0	36	1345237
0	7	464690
0	21	258771
0	5	130963
0	66	9281



Conclusion:

We could observe the following from above plot and table:

- All the codes are not common across precincts. Few codes could be common but most of them differs across precincts.

5. Violations by Hour of the day

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

a. 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.
b. Find a way to deal with missing values, if any.
c. Divide 24 hours into 6 equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the 3 most commonly occurring violations
d. Now, try another direction. For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part)

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.

a. **Understanding:**
   1. The first two characters stand for hours, third and fourth stand for minutes, and the last one determines whether it is AM or PM.
b. **Ways to correct the format**
   1. Violation time was separated into different columns (hour, day and convention) using substr function
   2. Converting the hour into 24-hour format and exclusion of A and P
   3. Concat hour and minutes separated by colon

Find a way to deal with missing values, if any.

a. **Understanding:**
   1. As the format is in AM or PM format, the hours could be in between the value 1-12, minutes would be less than a value of 60 and classification of "A" or "P" should be present

b. **Identifying and Fixing Missing Values:**
   1. Hours other than value 1 to 12 were observed and some of the values were also with special characters were observed
   2. Some of values under minutes column were observed with special characters
   3. Some of values under convention column was without "A" or "P" classification
   4. All the above issues were filtered out before performing the analysis as the number were not significant and considering the huge dataset, it would not have much impact

Divide 24 hours into 6 equal discrete bins of time

Bins	Timing
Early Morning	5:00 to 7:59
Morning	8:00 to 11:59
Afternoon	12:00 to 16:59
Evening	17:00 to 20:59
Night	21:00 to 23:59
Late Night	0:00 to 4:59

3 Most Commonly Occurring Violations for Each Groups

2015

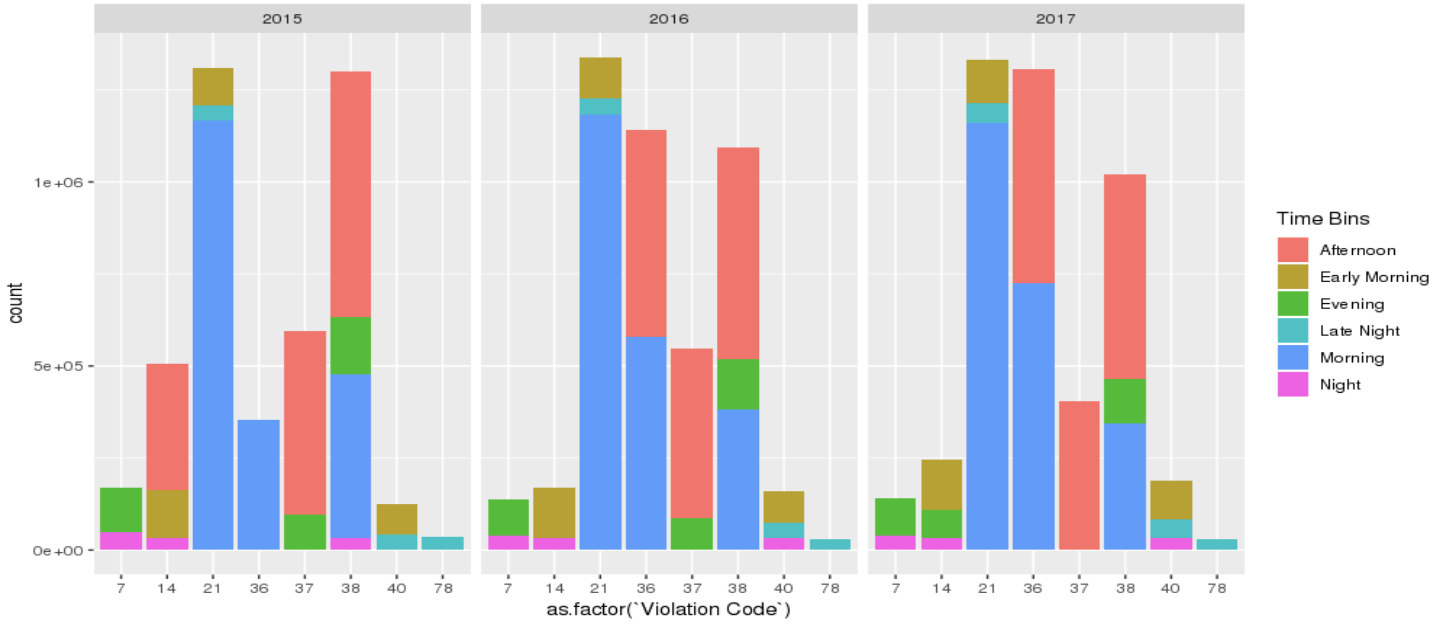
Groups	Violation Code	Frequency
Evening	38	155790
Evening	7	122958
Evening	37	98028
Morning	21	1167094
Morning	38	442655
Morning	36	353555
Early Morning	14	130419
Early Morning	21	102689
Early Morning	40	84539
Afternoon	38	668865
Afternoon	37	496972
Afternoon	14	343298
Night	7	48129
Night	38	34338
Night	14	33672
Late Night	40	41587
Late Night	21	41140
Late Night	78	36298

2016

Groups	Violation Code	Frequency
Evening	38	135471
Evening	7	98834
Evening	37	88334
Morning	21	1183374
Morning	36	578035
Morning	38	382100
Early Morning	14	136222
Early Morning	21	109753
Early Morning	40	85250
Afternoon	38	575504
Afternoon	36	564438
Afternoon	37	457571
Night	7	38990
Night	40	33679
Night	14	32289
Late Night	21	44878
Late Night	40	39806
Late Night	78	29639

2017

Groups	Violation Code	Frequency
Evening	38	122642
Evening	7	100221
Evening	14	76871
Morning	21	1161225
Morning	36	726513
Morning	38	343204
Early Morning	14	137262
Early Morning	21	115073
Early Morning	40	104859
Afternoon	36	579804
Afternoon	38	553366
Afternoon	37	405221
Night	7	40953
Night	40	33935
Night	14	32517
Late Night	21	54067
Late Night	40	49109
Late Night	78	30463



Conclusion:

We could observe the following from above plot and table:

- No parking in the street cleaning zone seems is the highest number of offences in the morning. Indicating that people are either parking in the night or leaving it there.
- Muni meter receipts seems to be pretty high in the afternoon, possible post lunch as drivers exceeding the minimum time allowed.

Most Common Times of Day for 3 Most Commonly Occurring Violation Codes

2015

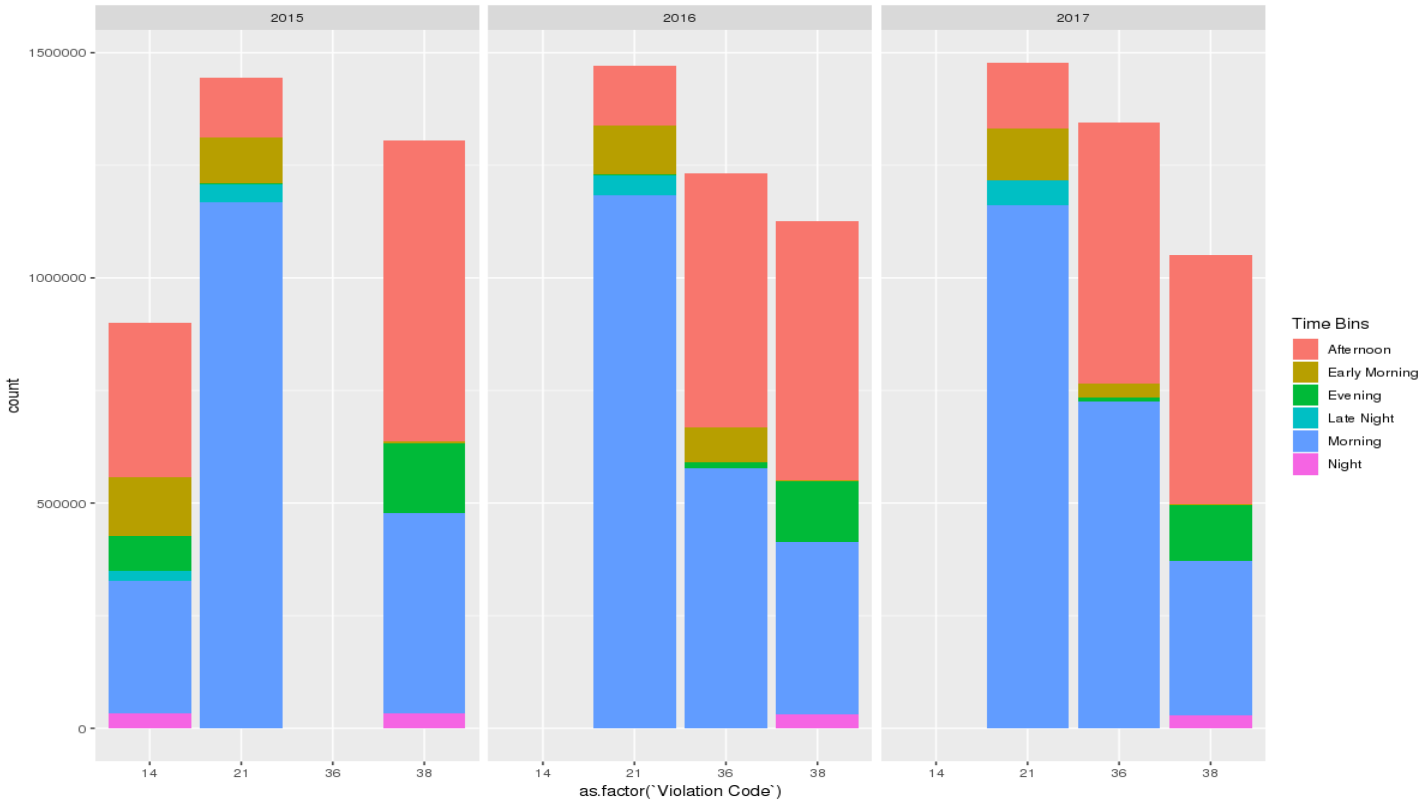
Violation Code	Times of Day	Frequency
14	Afternoon	343298
14	Morning	292903
14	Early Morning	130419
14	Evening	76747
14	Night	33672
14	Late Night	23738
21	Morning	1167094
21	Afternoon	131476
21	Early Morning	102689
21	Late Night	41140
21	Evening	573
21	Night	414
38	Afternoon	668865
38	Morning	442655
38	Evening	155790
38	Night	34338
38	Early Morning	2640
38	Late Night	703

2016

Violation Code	Times of Day	Frequency
21	Morning	1183374
21	Afternoon	131754
21	Early Morning	109753
21	Late Night	44878
21	Evening	414
21	Night	332
36	Morning	578035
36	Afternoon	564438
36	Early Morning	77656
36	Evening	12820
36	Late Night	2
38	Afternoon	575504
38	Morning	382100
38	Evening	135471
38	Night	31207
38	Early Morning	2054
38	Late Night	493

2017

Violation Code	Times of Day	Frequency
21	Morning	1161225
21	Afternoon	145791
21	Early Morning	115073
21	Late Night	54067
21	Evening	391
21	Night	275
36	Morning	726513
36	Afternoon	579804
36	Early Morning	30072
36	Evening	8848
38	Afternoon	553366
38	Morning	343204
38	Evening	122642
38	Night	28465
38	Early Morning	2157
38	Late Night	568



Conclusion:

We could observe the following from above plot and table:

- General no standing offence has reduced over the period and no longer in the top 3 category.
- Almost 100% jump in offence category 38 in 2017, which is to do with muni meter receipts, where people stayed/parked beyond allowed time limit.

6. Violations by Seasonality

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.
- Then, find the 3 most common violations for each of these season

Divide the Year into Some Number of Seasons

Seasons	Months
Summer	June to August
Fall	September to November
Winter	December to February
Spring	March to May

Frequencies of Tickets for Each Season

2015

Seasons	Frequency
Spring	2860987
Summer	2838306
Fall	2718502
Winter	2180241

2016

Seasons	Frequency
Fall	2971672
Spring	2789066
Winter	2421620
Summer	2214536

2017

Seasons	Frequency
Spring	2873383
Fall	2829224
Winter	2483036
Summer	2353920

3 Most Common Violations for Each of These Season

2015

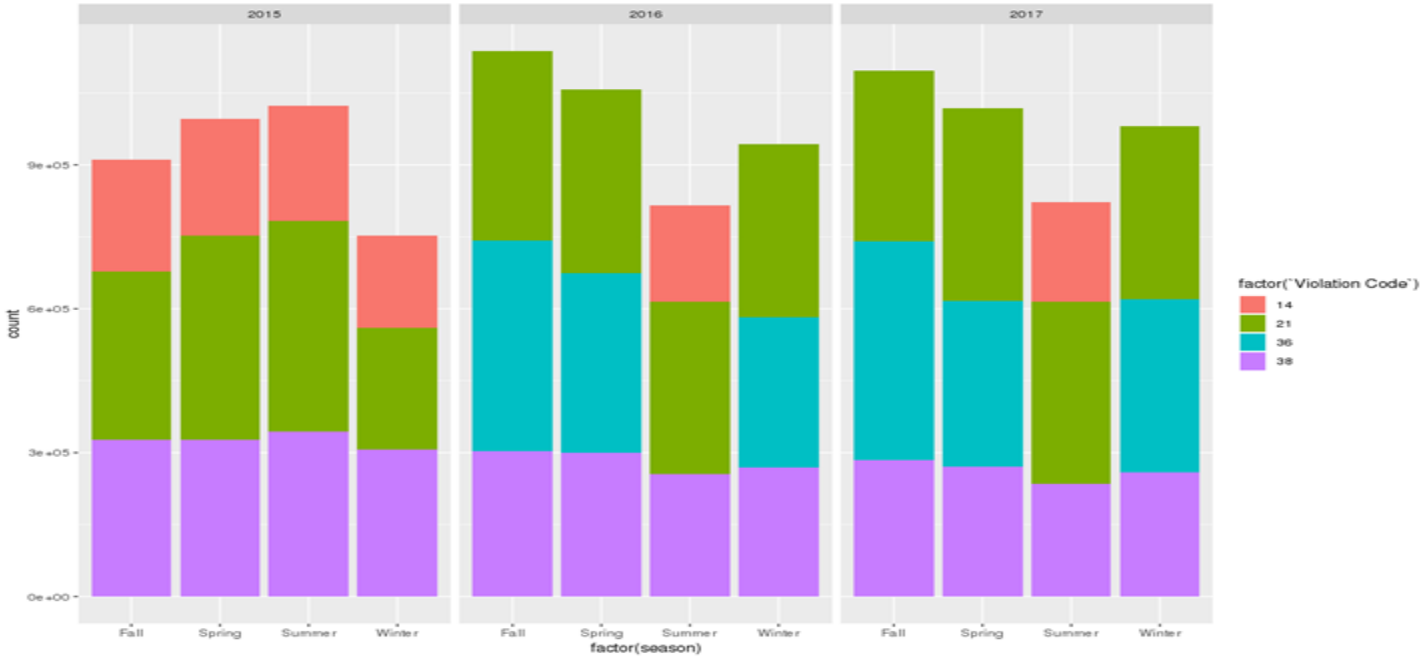
Seasons	Violation Code	Frequency
Spring	21	425163
Spring	38	327048
Spring	14	243622
Summer	21	439632
Summer	38	344262
Summer	14	239339
Fall	21	351390
Fall	38	326700
Fall	14	232300
Winter	38	306997
Winter	21	253043
Winter	14	193157

2016

Seasons	Violation Code	Frequency
Spring	21	383448
Spring	36	374362
Spring	38	299439
Summer	21	358896
Summer	38	255600
Summer	14	200608
Fall	36	438320
Fall	21	395020
Fall	38	303387
Winter	21	359905
Winter	36	314765
Winter	38	268409

2017

Seasons	Violation Code	Frequency
Spring	21	402424
Spring	36	344834
Spring	38	271167
Summer	21	378699
Summer	38	235725
Summer	14	207495
Fall	36	456046
Fall	21	357257
Fall	38	283816
Winter	21	362016
Winter	36	359338
Winter	38	259710



Conclusion:

We could observe the following from above plot and table:

- Spring season across all three years more violations are occurred followed by fall, winter and summer
- Violation code 38 and 21 has occurred as the most common violation in all the season for three fiscal years

7. Top Violations by Count and Fine Amount Analysis

The fines collected from all the parking violation constitute a revenue source for the NYC police department. Let's take an example of estimating that for the 3 most commonly occurring codes.

- Find total occurrences of the 3 most common violation codes
- Then, search the internet for NYC parking violation code fines. You will find a website (on the nyc.gov URL) that lists these fines. They're divided into two categories, one for the highest-density locations of the city, the other for the rest of the city. For simplicity, take an average of the two.
- Using this information, find the total amount collected for all of the fines. State the code which has the highest total collection.
- What can you intuitively infer from these findings?

Total Occurrences of 3 Most Common Violation Codes

2015

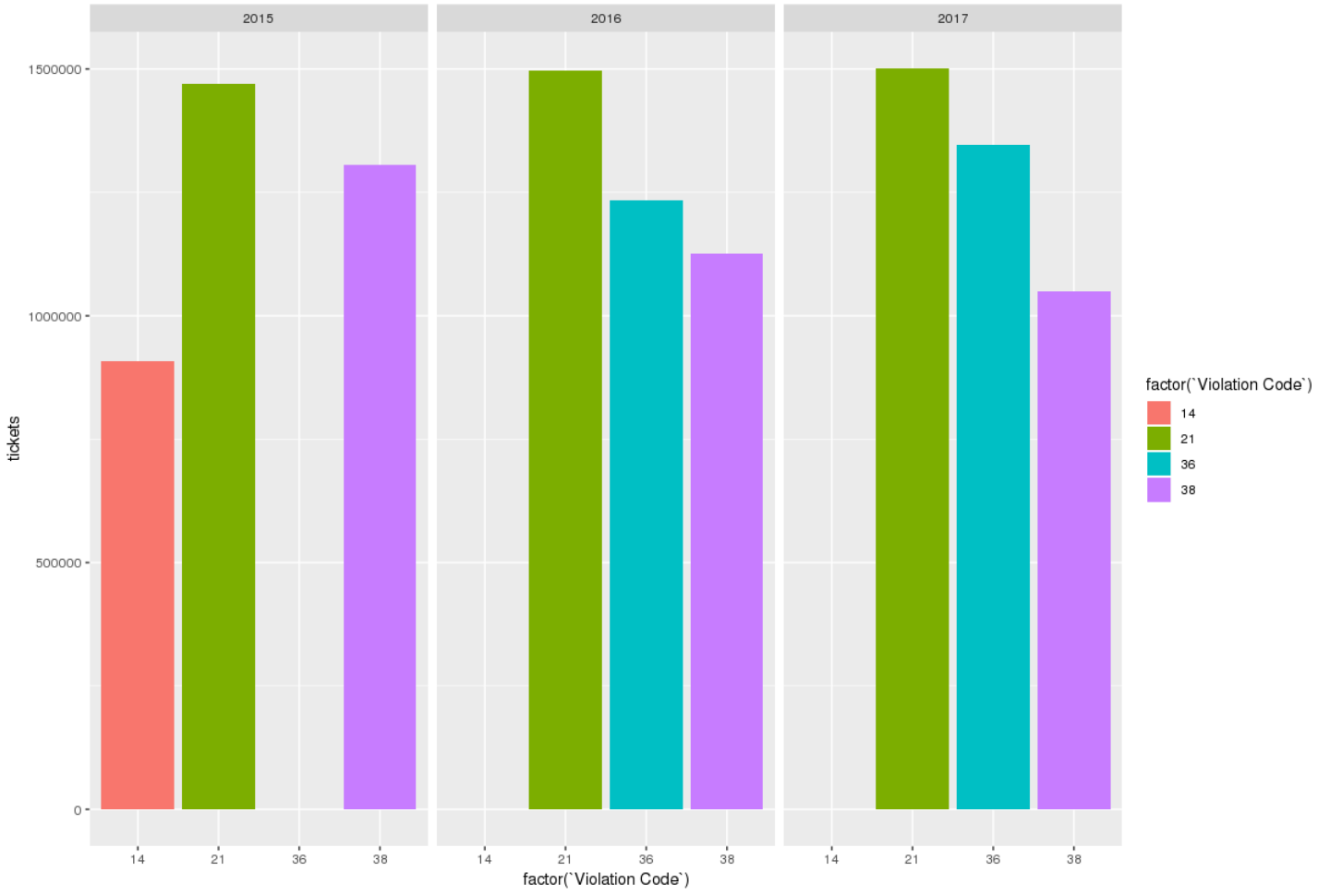
Violation Code	Frequency
21	1469228
38	1305007
14	908418

2016

Violation Code	Frequency
21	1497269
36	1232952
38	1126835

2017

Violation Code	Frequency
21	1500396
36	1345237
38	1050418



Code which has the Highest Total Collection

2015

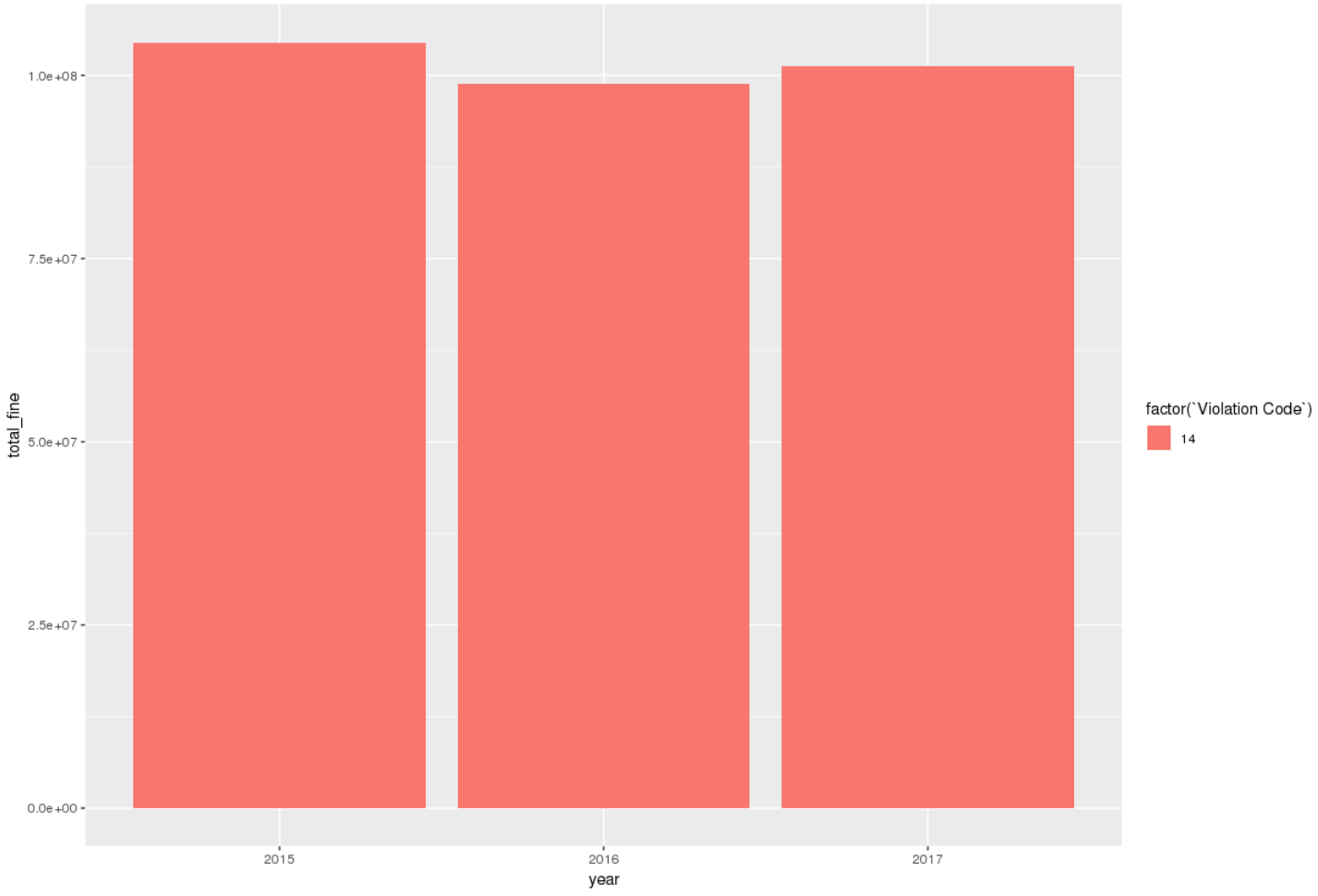
Violation Code	# Tickets	Average Fine	Total Collections
14	908418	115	104468070

2016

Violation Code	# Tickets	Average Fine	Total Collections
14	860045	115	98905175

2017

Violation Code	# Tickets	Average Fine	Total Collections
14	880152	115	101217480



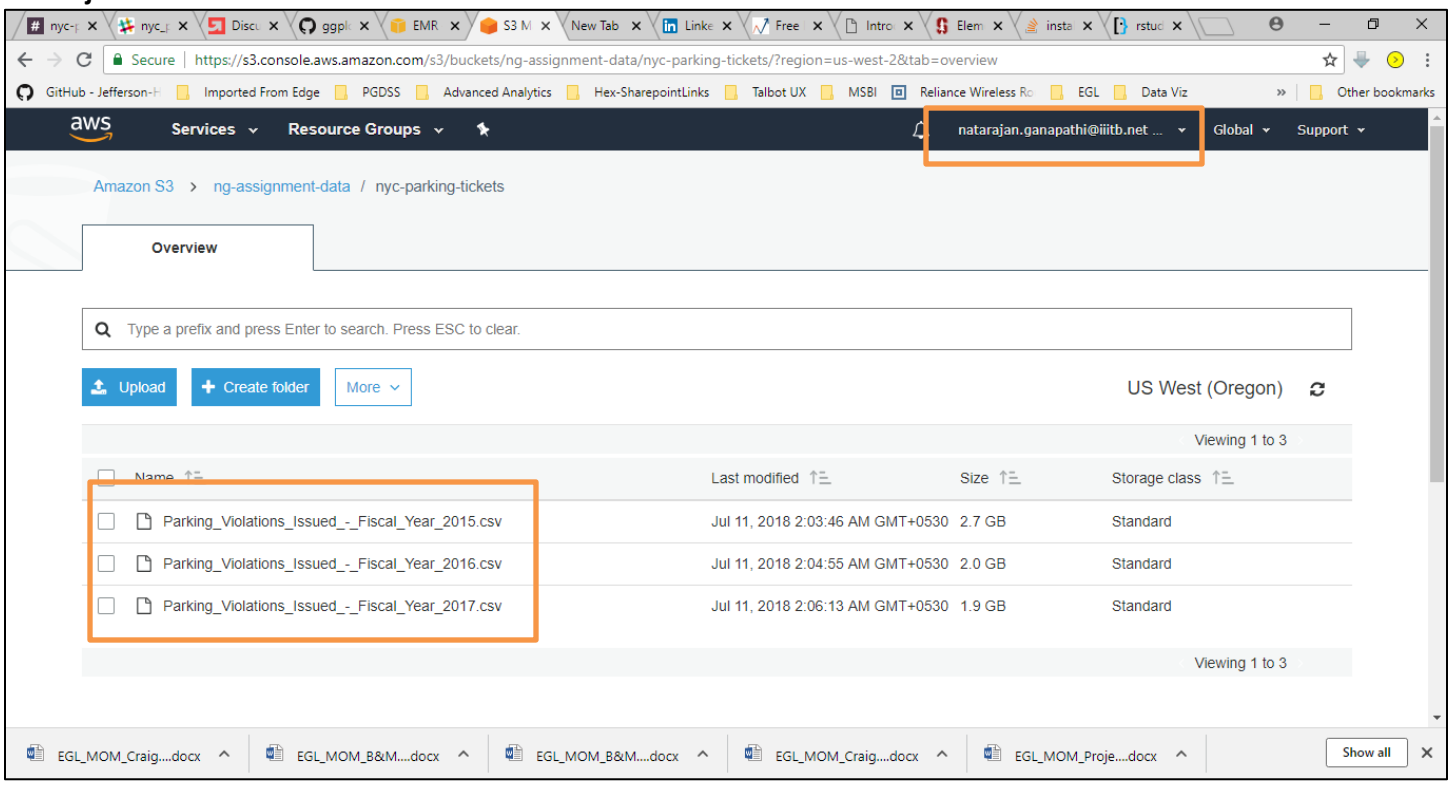
Inferences:

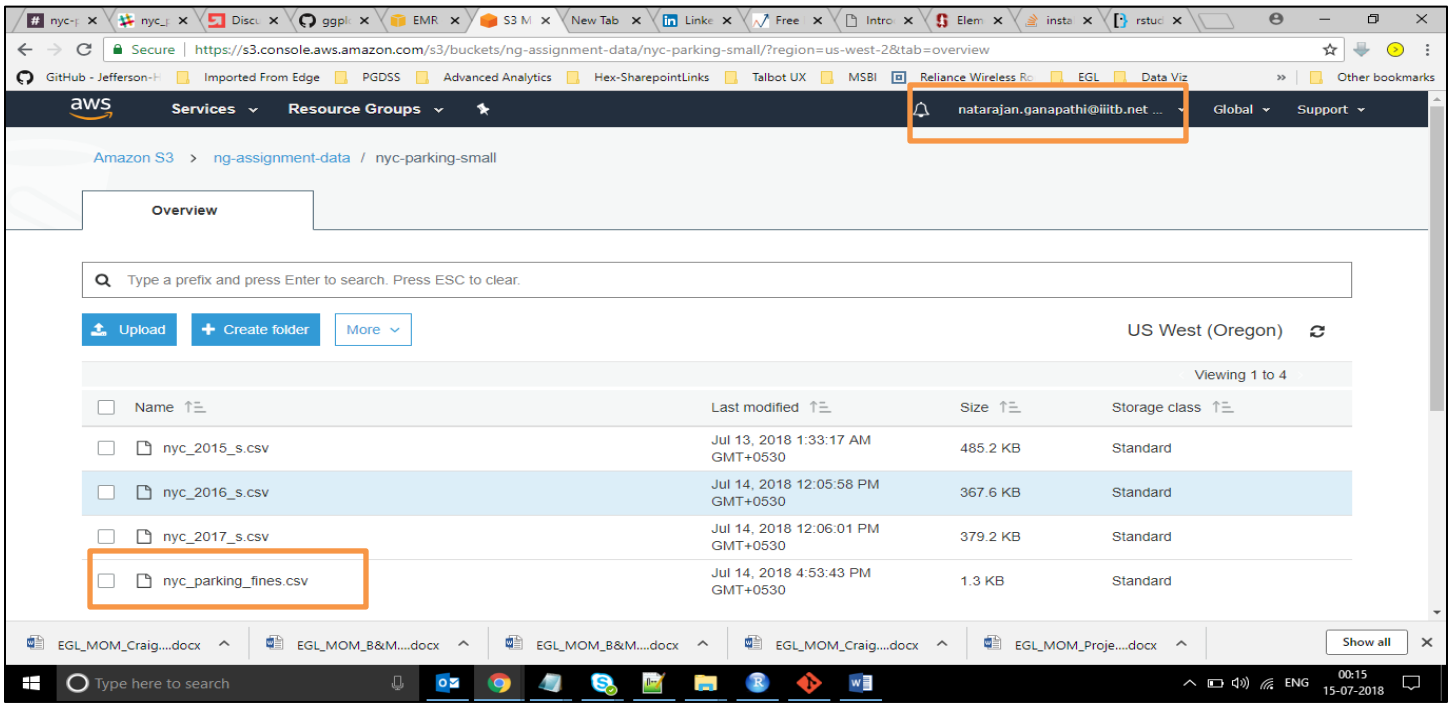
1. Violation code 21 has most number of occurrences in all three (2015, 2016, 2017) datasets
2. Violation code 14 has most number of collections in all three (2015, 2016, 2017) datasets

4 Screenshots & Evidences

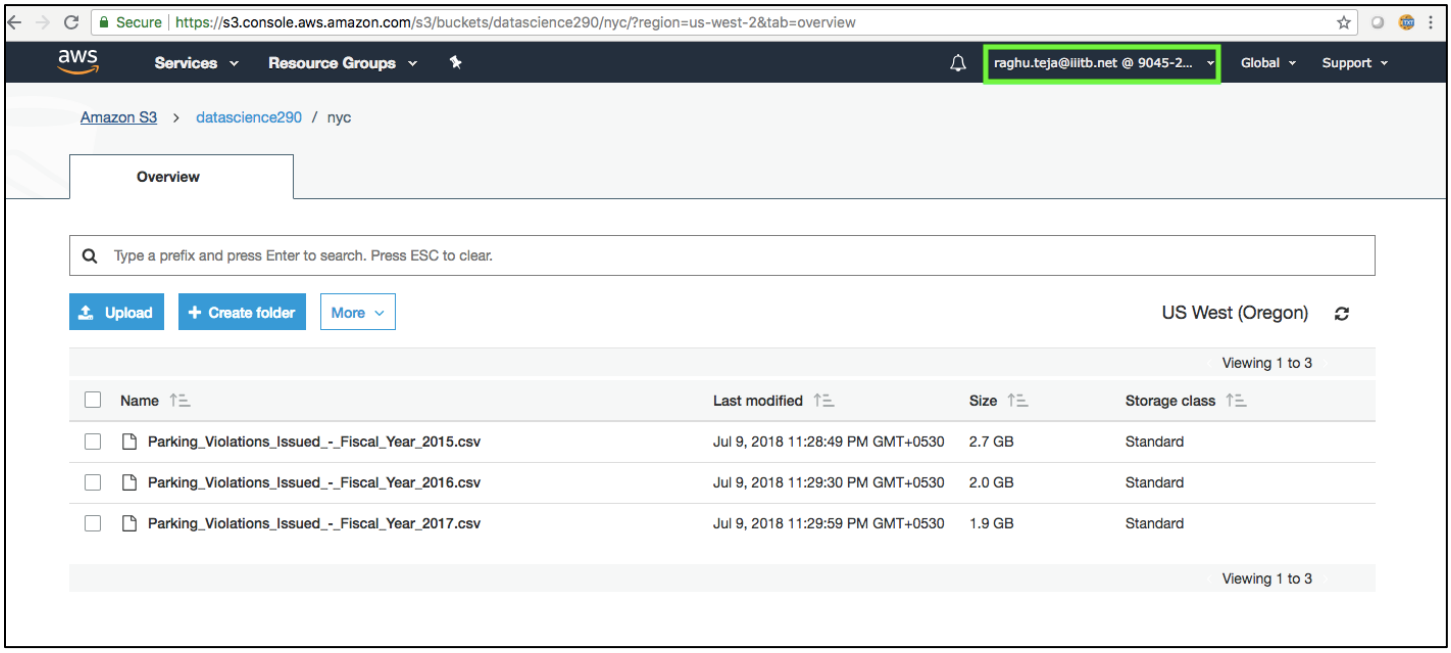
Screenshots of S3 Buckets of Team Members:

Natarajan:

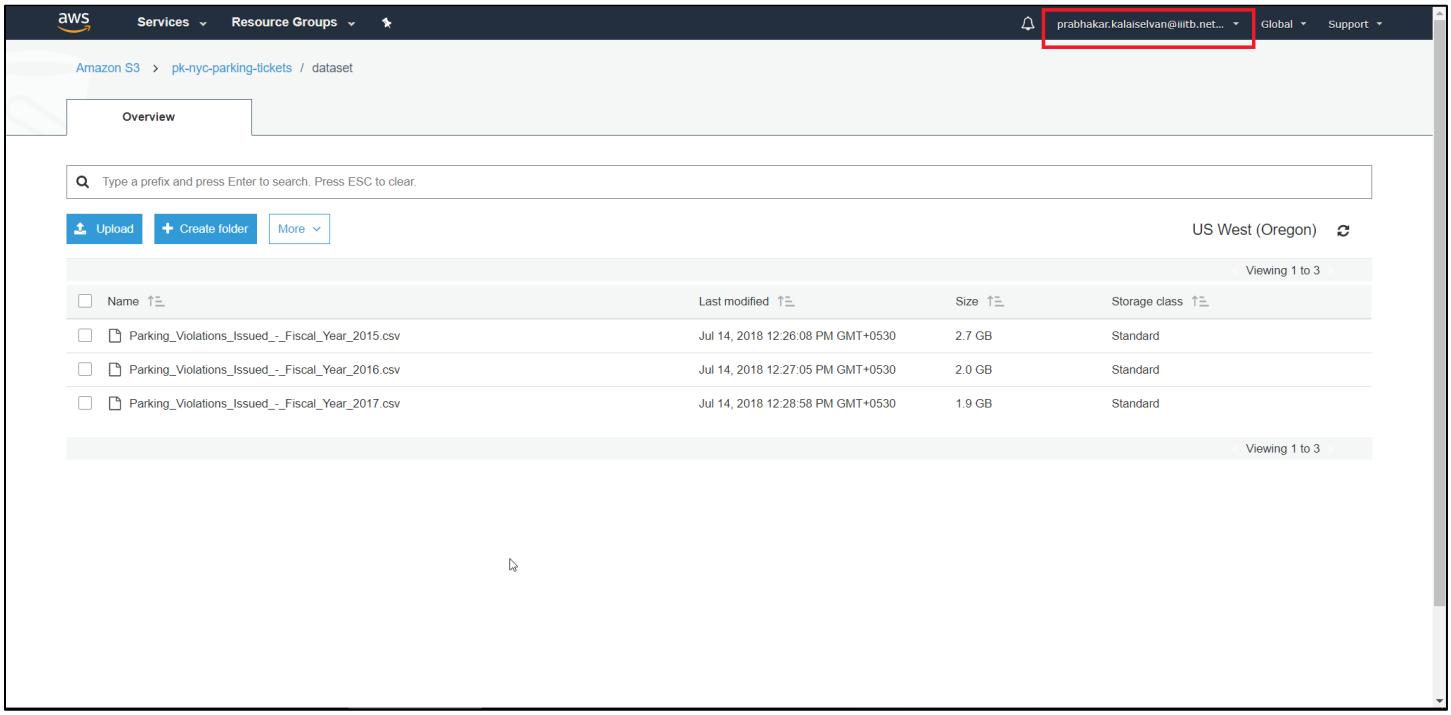




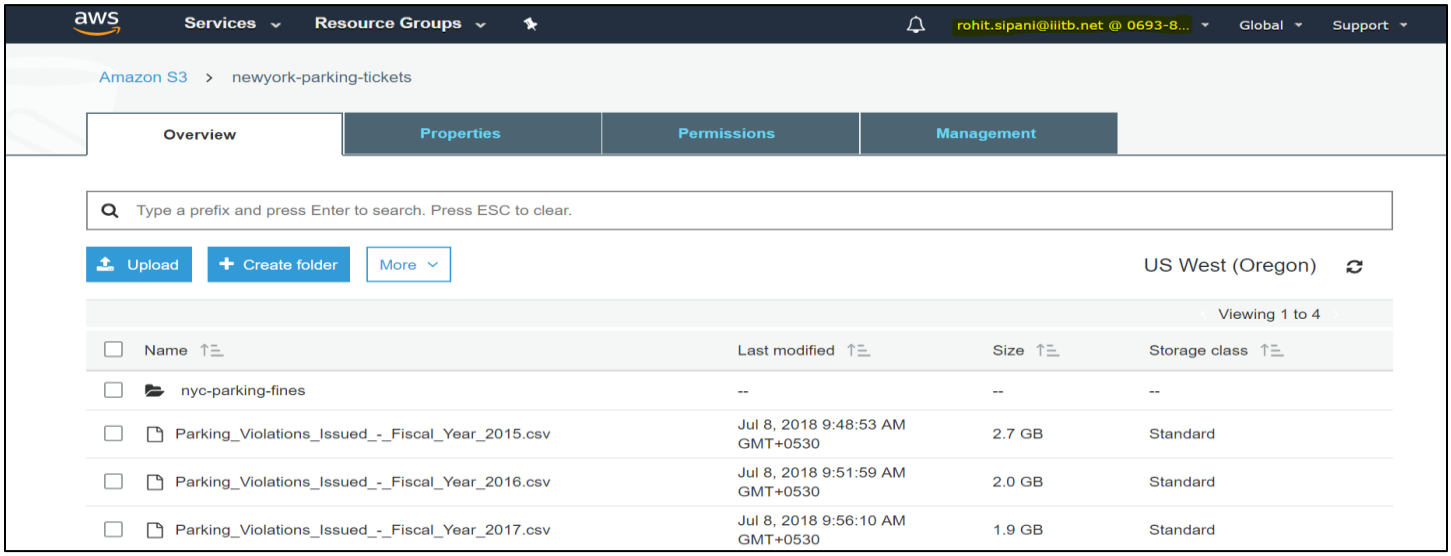
Raghu:

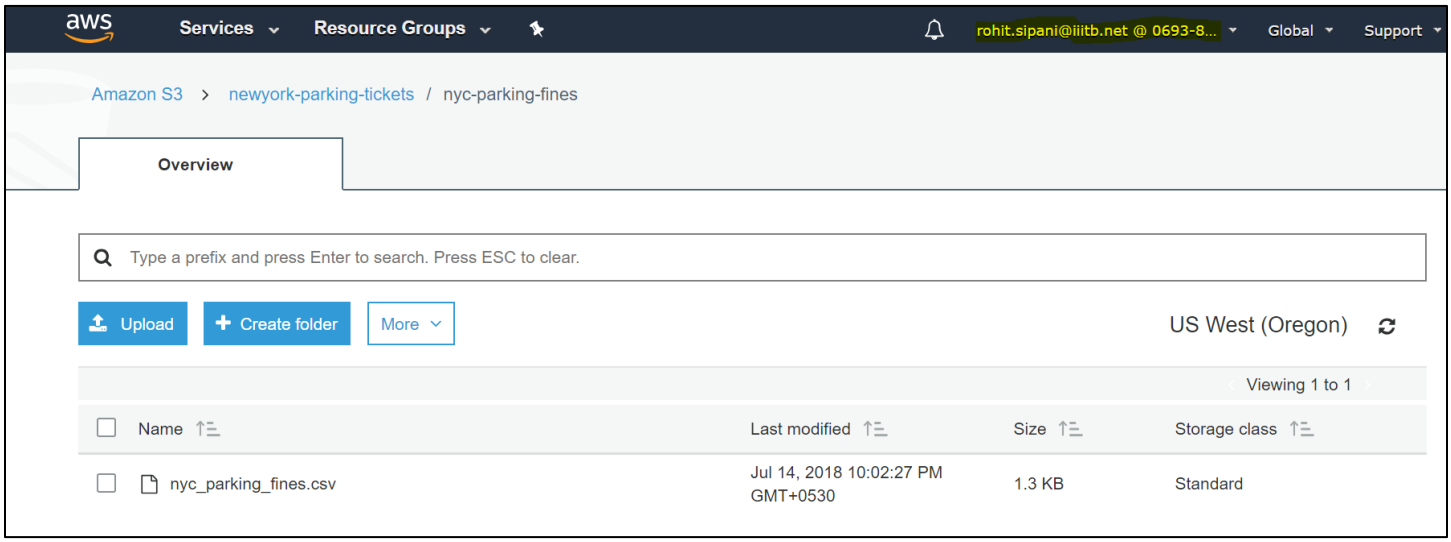


Prabhakar:



Rohit:


aws
Services
Resource Groups
rohit.sipani@iiitb.net @ 0693-8...
Global
Support
Amazon S3 > newyork-parking-tickets
Overview
Properties
Permissions
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US West (Oregon)
Viewing 1 to 4
Name
Last modified
Size
Storage class
nyc-parking-fines -- -- --
Parking_Violations_Issued_-_Fiscal_Year_2015.csv Jul 8, 2018 9:48:53 AM GMT+0530 2.7 GB Standard
Parking_Violations_Issued_-_Fiscal_Year_2016.csv Jul 8, 2018 9:51:59 AM GMT+0530 2.0 GB Standard
Parking_Violations_Issued_-_Fiscal_Year_2017.csv Jul 8, 2018 9:56:10 AM GMT+0530 1.9 GB Standard


aws
Services
Resource Groups
rohit.sipani@iiitb.net @ 0693-8...
Global
Support
Amazon S3 > newyork-parking-tickets / nyc-parking-fines
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US West (Oregon)
Viewing 1 to 1
Name
Last modified
Size
Storage class
nyc_parking_fines.csv Jul 14, 2018 10:02:27 PM GMT+0530 1.3 KB Standard