NYC Parking Tickets

AN EXPLORATORY ANALYSIS

Objective

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 a cramped geography is the exact recipe that leads to a huge number of parking tickets.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has collected data for parking tickets. We will try and perform some exploratory analysis on this data. Spark will allow us to analyse the full files at high speeds, as opposed to taking a series of random samples that will approximate the population.

The purpose of this case study is to conduct an exploratory data analysis that helps us understand the data.

Data Source

NYC Police Department has collected data for parking tickets. Out of these, the data files from 2014 to 2017 are publicly available on Kaggle.

As part of the assignment, we are provided with data at '/common_folder/nyc_parking/Parking_Violations_Issued_-_Fiscal_Year_201x.csv' where 'x' is between {5, 6, 7}.

We will try and perform some exploratory analysis on this data. For the scope of this analysis, we wish to compare the phenomenon related to parking tickets over three different years - 2015, 2016, 2017. All the analysis steps will be done for three different years.

Methodology

Analysis will be performed on RStudio mounted on Corestack cluster, using the SparkR library and will be carried out in below phases:

- **Phase1:** Loading, verification and Cleansing of source Data
- **Phase2:** Examine the data to answer the assignment questions
- **Phase3:** Performing Aggregation tasks to answer the assignment questions

Phase 1: Loading, verification and Cleansing of source Data

In this phase below steps are carried out to understand the source data and prepare the final data for analysis in phase 2 and 3:

- Data is loaded into R data frames
- Understood the structure of the source data by comparing with the data dictionary available in Kaggle.
- All the columns data types are as per data dictionary except for Issue Date and Violation Location.
- Analysis is done on column names and found out that columns names are in sentence case with spaces and unregistered vehicle column name had a special character (?).
- Removed the leading and trailing spaces, substituted the space in between with "-", removed (?) from column name and converted them to lower case in all the three data frames.
- It is observed and removed the duplicate values in summons_number column which is supposed to be a unique column.
- It is observed that latitude, longitude, community_board, community_council, census_tract, bin, bbl and nta columns are present in 2015 and 2016 data but not in 2017. It is identified that all the values in these columns are NULL, so dropped these columns from 2015 and 2016 data frames.
- It is observed that issue date is in mm/dd/yyyy format in all the three data files and however it is recorded as string, so converted the same to date format in the three source data frames.
- NULL value check analysis is done on issue_date and found zero discrepancies.
- It is observed that data is spread across years in all three files.
- 2015 data file has data prior to 2014 till June 2015.
- 2016 data file has data spread across many years including fiscal year 2015 and 2017
- 2017 data file has data spread across many years including fiscal year 2015 and 2016
- Assumption: We are assuming that some of the tickets issued in 2015, 2016 or 2017 are reported in the successive years which is practical case scenario.
- As we need to do analysis for Fiscal year or calendar year of 2015, 2016 and 2017, we have extracted all three fiscal years data from three source data file and then combined into one file each for fiscal years 2015, 2016 and 2017
- As per US law, fiscal year is from Oct 1st of previous year to 30th September of current year. From the combined data deriving fiscal year data of 2015, 2016 and 2017
- This derived fiscal year data sets will be used for answering questions in phase 2 and 3.

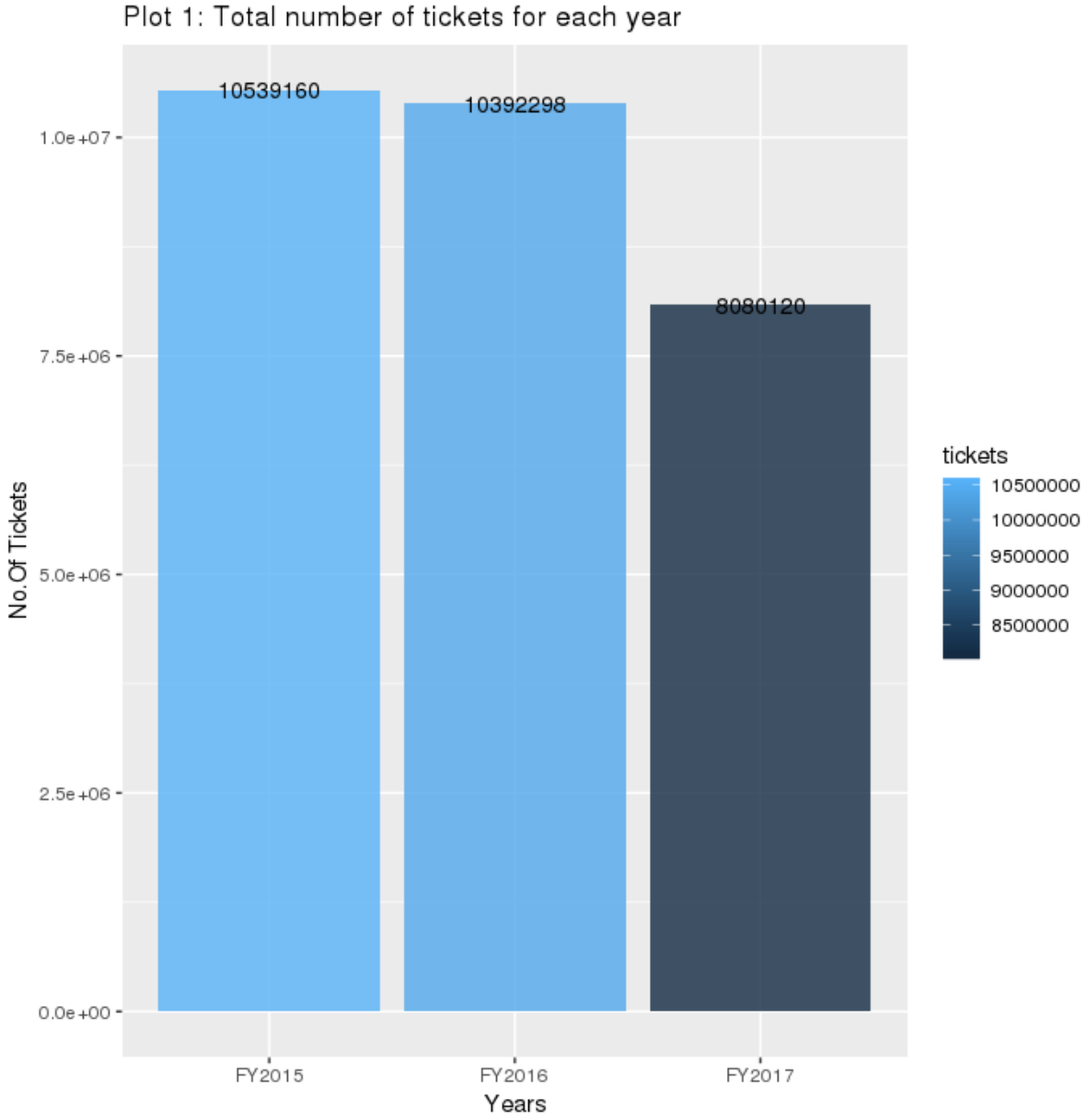
Phase 2: Examine the data to answer the assignment questions

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

Answer:
Fiscal year wise total number of tickets issued per year is below:

year	tickets
FY2015	10539160
FY2016	10392298
FY2017	8080120

From the below plot it is evident that number of tickets issued is showing a decline trend with a considerable decrease in Fiscal year 2017.



Question 2:

Find out the number of unique states from where the cars that got parking tickets came from.

Answer:

The number of unique states from where the cars that got parking tickets came from is as below:

year	Unique States
FY2015	69
FY2016	67
FY2017	67

The unique from where the cars come from is highest in 2015 while the number is equal for 2016 and 2017.

It is observed that registration state with code "NY" is the state having maximum entries. Also it is found out that one of the state code is numeric value of "99". As instructed in the assignment, replaced "99" with state having maximum entries in all the three years. After replacing, the number of unique states from where the cars that got parking tickets came from is as below:

year	Unique States
FY2015	68
FY2016	66
FY2017	66

Question 3:

Some parking tickets don't have the address for violation location on them, which is a cause for concern.

Answer:

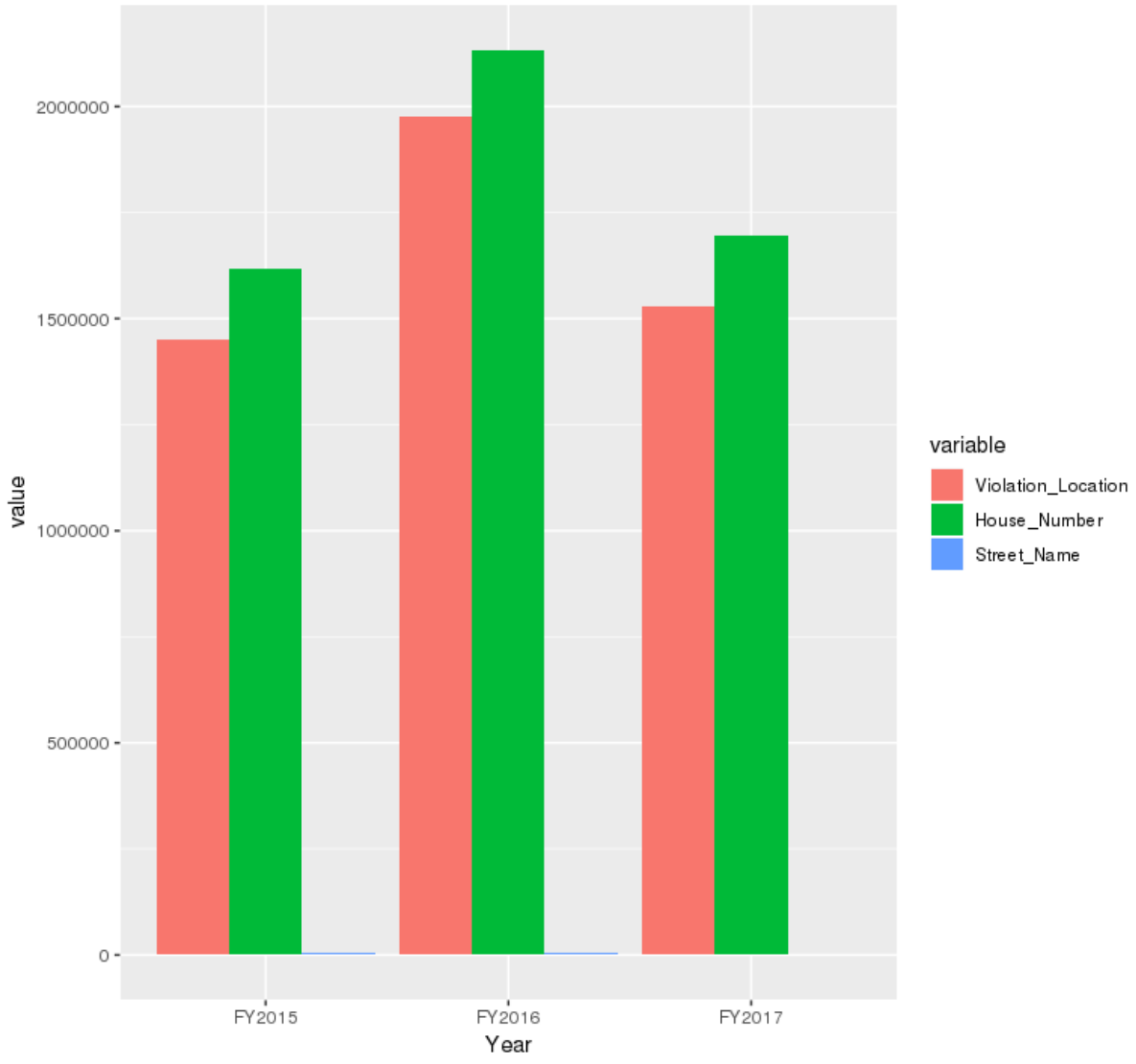
To answer this question we analysed three parameters Violation Location, Street Number and House number and found out missing values of each variable across three fiscal years. The results are:

Year	Violation_Location	House_Number	Street_Name
FY2015	1452260	1617406	6497
FY2016	1975285	2133853	4354
FY2017	1527752	1694560	2646

It is observed that :

- Violation location and house number discrepancies are highest in FY2016 followed by FY2015 and FY2017.
- Street name is highest in FY2015, followed by FY2016 and FY2017.

Plot 2: Records with missing address for each year



Phase 3: Performing aggregation tasks to answer the assignment questions

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

Answer:
Year wise top 5 violation codes and their frequency is as shown below:

Year	Violation code	count
FY2015	21	1475086
FY2015	38	1261904
FY2015	14	908916
FY2015	36	765892
FY2015	37	721594

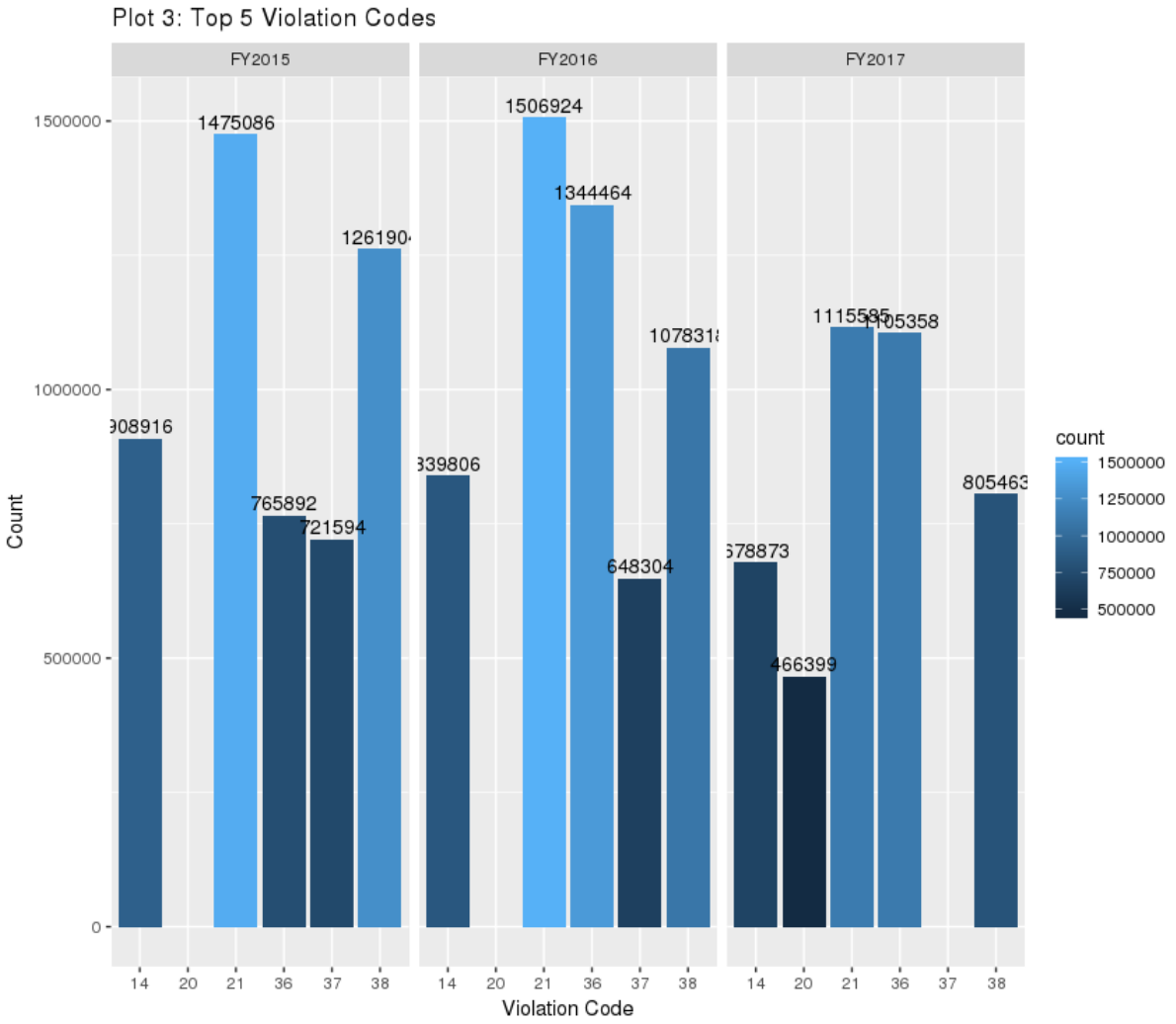
Year	Violation code	count
FY2016	21	1506924
FY2016	36	1344464
FY2016	38	1078318
FY2016	14	839806
FY2016	37	648304

Year	Violation code	count
FY2017	21	1115585
FY2017	36	1105358
FY2017	38	805463
FY2017	14	678873
FY2017	20	466399

It is clearly evident that:

- Violation Code 21 is the highest in all three years with frequency relatively less in FY2017.
- Violation Code 38 is the second highest in FY2015, whereas code 36 in FY2016 & FY2017.
- Violation Code 14 is the third highest in FY2015, whereas code 38 in FY2016 & FY2017.
- Violation Code 36 is the fourth highest in FY2015, whereas code 14 in FY2016 & FY2017.
- Violation Code 37 is the fifth highest in FY2015 & FY2016, whereas code 20 in FY2017.

Below plot helps in comparative study of top 5 violation codes across years



Note: Description of codes

21: Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.

38: Failing to show a receipt or tag in the windshield.

36: Exceeding the posted speed limit in or near a designated school zone.

37: Parking in excess of the allowed time

14: General No Standing: Standing or parking where standing is not allowed by sign

Question 2:

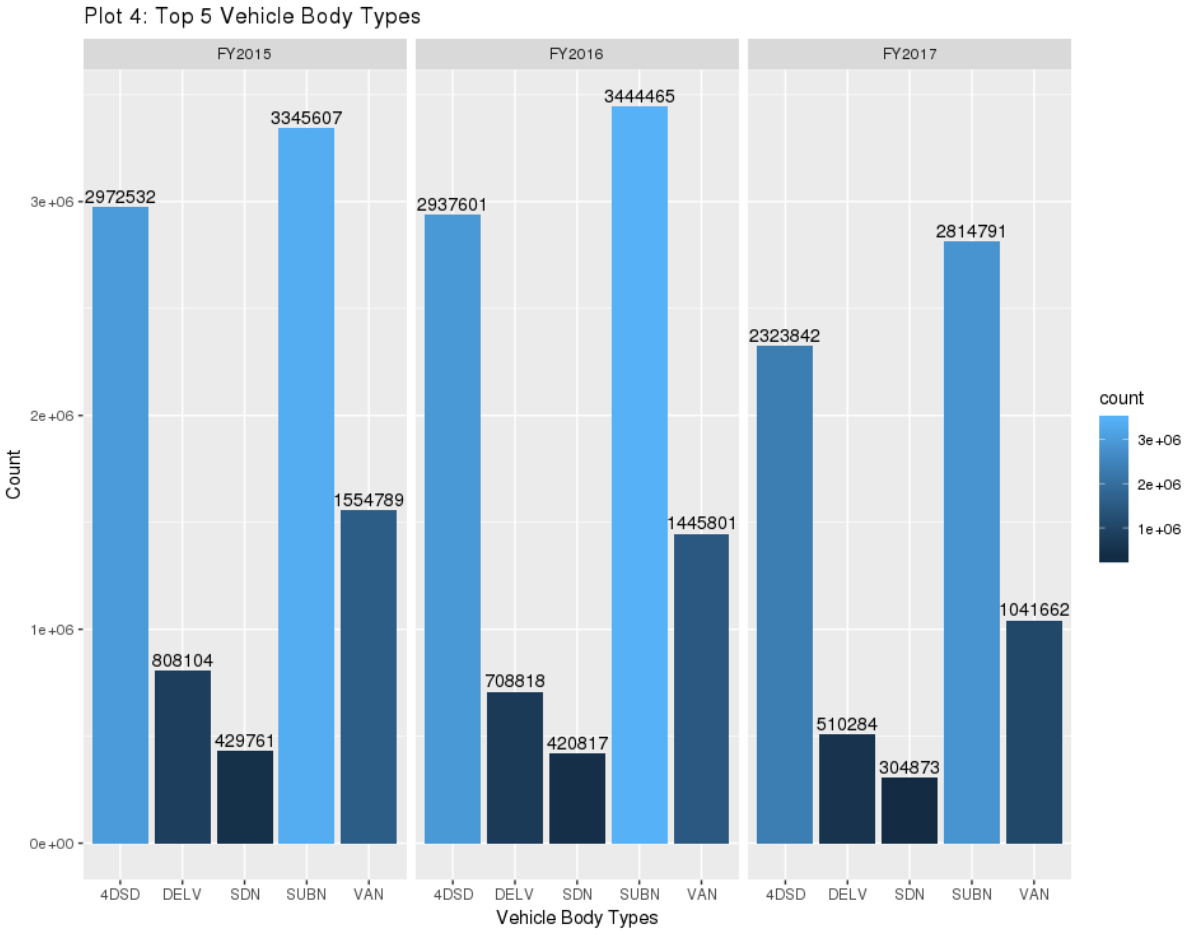
How often does each 'vehicle body type' get a parking ticket? How about the 'vehicle make'?

Answer:

2 a) Vehicle body type analysis year wise is as below:

vehicle_body_type	2015	2016	2017
SUBN	3345607	3444465	2814791
4DSD	2972532	2937601	2323842
VAN	1554789	1445801	1041662
DELV	808104	708818	510284
SDN	429761	420817	304873

Below plot helps in comparative study of top 5 violation codes across years

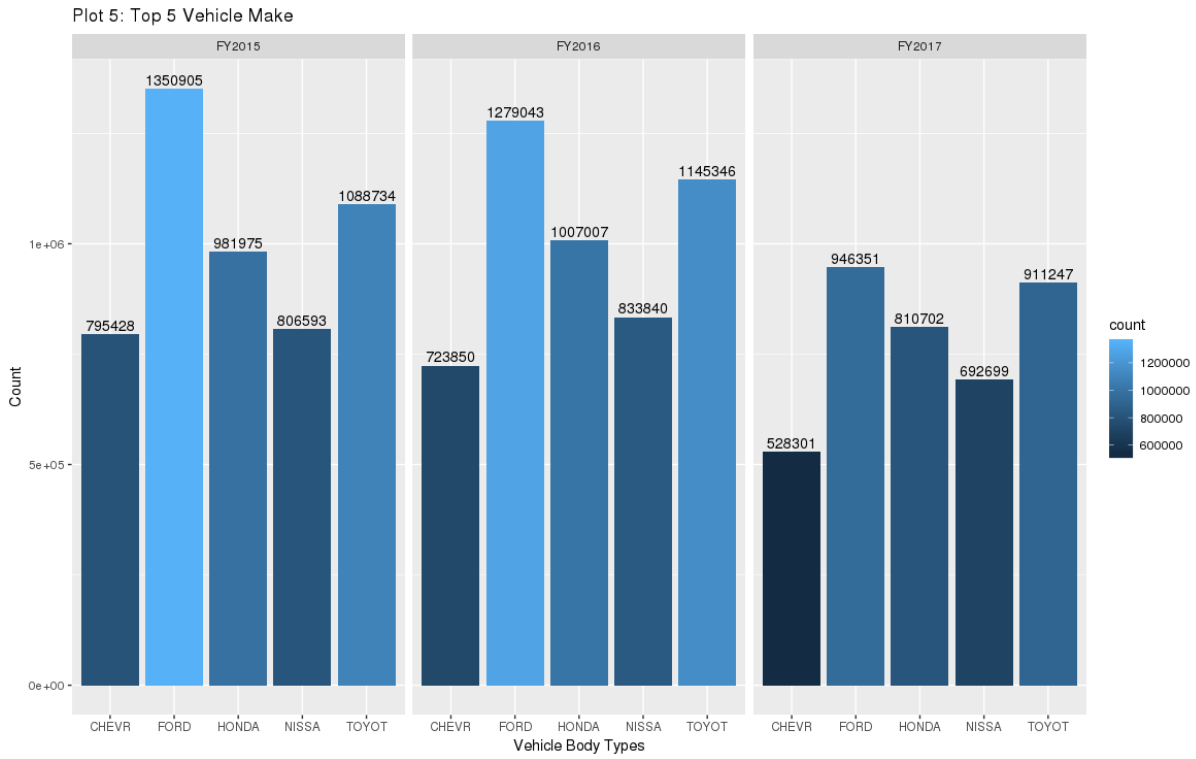


From the plot it is evident that, vehicle body types SUBN, followed by 4DSD, VAN, DELV and SDN are issued high parking tickets and remain the same across the years.

2 b) Vehicle make analysis year wise is as below:

vehicle_make	2015	2016	2017
FORD	1350905	1279043	946351
TOYOT	1088734	1145346	911247
HONDA	981975	1007007	810702
NISSA	806593	833840	692699
CHEVR	795428	723850	528301

Below plot helps in comparative study of top 5 violation codes across years



From the plot and above table it is evident that FORD is the vehicle make which had received highest number of tickets in all the three fiscal years.

FORD is followed by Toyota, Honda, Nissan and Chevrolet respectively in all the three years.

A downward trend can be observed from FY2015 to FY2017 which followed the similar with total number of tickets issued.

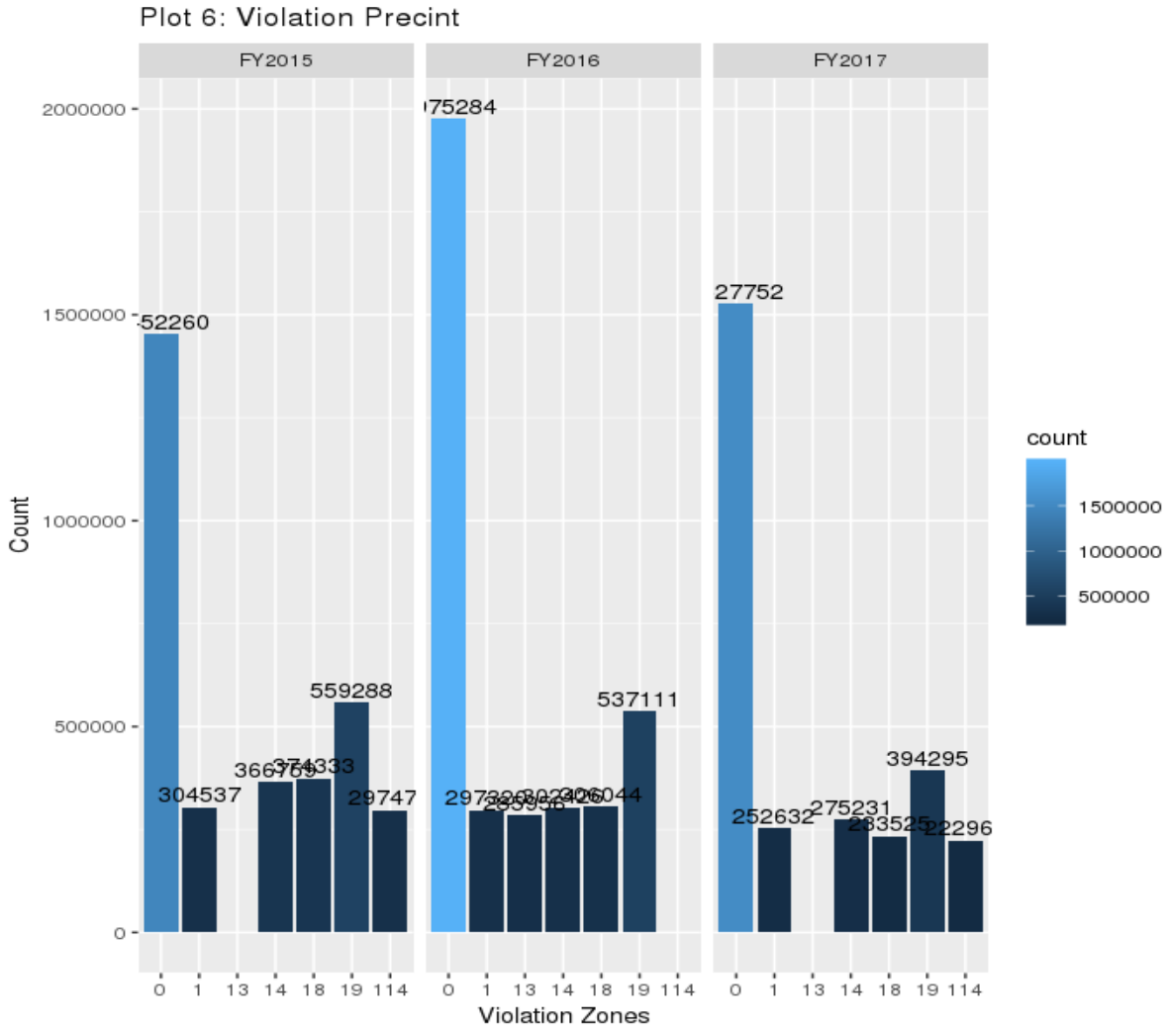
Question 3:
A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequency of tickets

Question 3a:
'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 analysis year wise is as below:

violation_precinct	2015	2016	2017
0	1452260	1975284	1527752
19	559288	537111	394295
18	374333	306044	275231
14	366759	302426	252632
1	304537	297320	233525
114	297477	285956	222967

"0" is a wrong precint code entered in the data. Hence ignoring the same from the analysis.



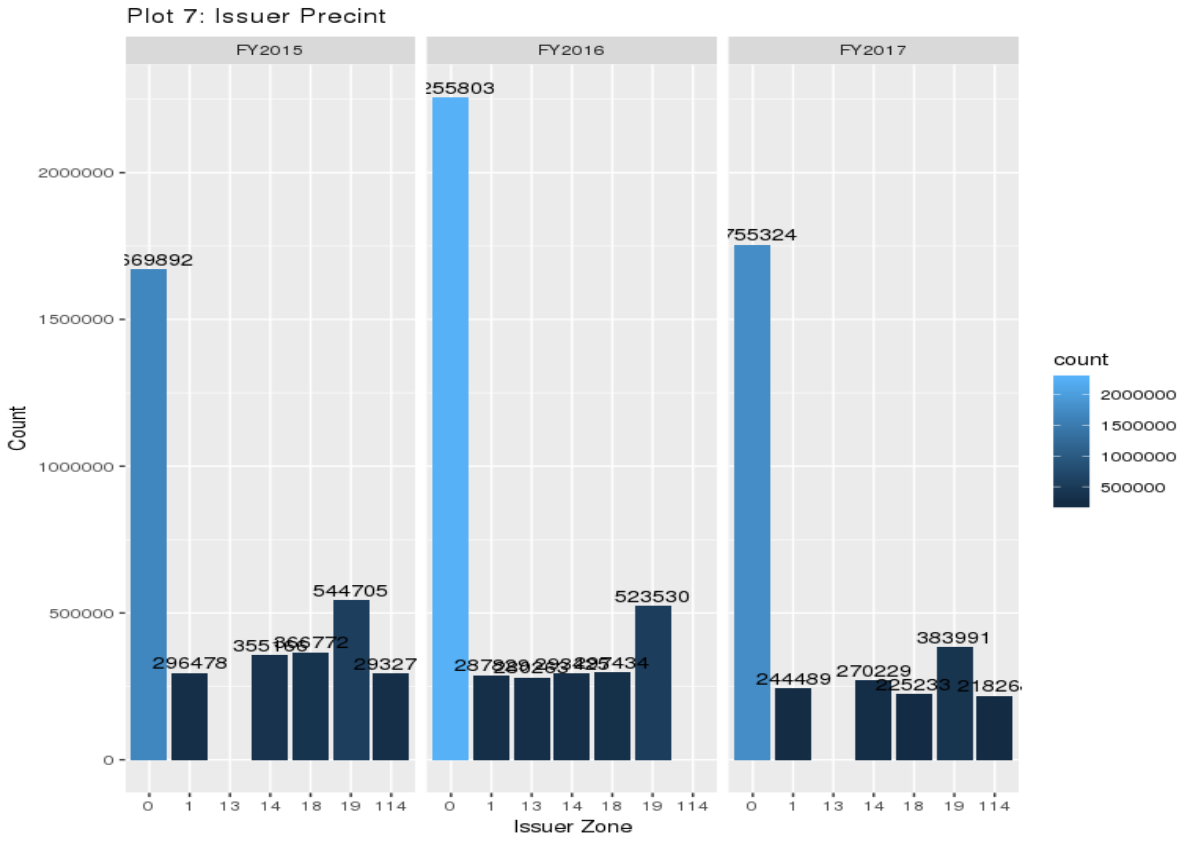
From the table and plot, it is evident that:

- Violation Precinct (Zone 19) recorded highest number of tickets across years FY2015, 16 and 17.
- Zones 18 and 14 are close to each other in ticket issuing across the three years are next to zone 19.
- Zone 1 and 114 are in 4th and 5th position across the three years.
- There is a downward trend observed from FY2015 to FY2017 in all the zones.

Question 3b:
'Issuer Precinct' (this is the precinct that issued the ticket)

Answer:
Issuer Precinct analysis year wise is as below:

FY 2015		FY2016		FY2017	
issuer_precinct	count	issuer_precinct	count	issuer_precinct	count
0	1669892	0	2255803	0	1755324
19	544705	19	523530	19	383991
18	366772	18	297434	14	270229
14	355166	14	293425	1	244489
1	296478	1	287829	18	225233
114	293277	13	280263	114	218268



From the table and plot it is evident that:

- "0" is the issuer precinct which recorded highest number of tickets. However 0 is not a valid code, so excluding the same from analysis.
- Issuer precinct "19" recorded highest number of tickets across three years.
- Issuer precinct "18" recorded second highest number of tickets for FY2015 and 16, while precinct "14" recorded second highest for FY2017.
- Issuer precinct "14" recorded third highest number of tickets for FY2015 and 16, while pre cinct "1" recorded third highest for FY2017.
- Issuer precinct "1" recorded fourth highest number of tickets for FY2015 and 16, while precinct "18" recorded fourth highest for FY2017.
- Issuer precinct "114" recorded fourth highest number of tickets for FY2015 and 17, while precinct "13" recorded fourth highest for FY2016.

Question 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?

Answer:

Violation code analysis of top issuer precincts as below:

FY2015			FY2016			FY2017		
Issuer precinct	Violation code	tickets	Issuer precinct	Violation code	tickets	Issuer precinct	Violation code	tickets
18	14	114496	18	14	91931	19	46	65062
19	38	85453	19	38	74159	14	14	59907
19	37	77840	19	37	74044	1	14	55446
14	69	77246	19	46	74027	19	38	54146
14	14	73794	14	69	61577	19	37	53131
19	14	62032	19	14	59077	14	69	42521

Across all the three years and top precincts, there are some exceptionally high frequency of violation codes is observed.

- 14 is the violation code which is observed against all the precinct in the exceptionally high cases except for 19 precinct in 2017 and 14 precinct in 2016.
- 38 is the violation code which is exceptionally high in precinct 19 across years.
- 69 is the violation code which is exceptionally high in precinct 14 across years.

Question 5:
5 a) Find a way to deal with missing values, if any.

Answer:

- More than 1000 records found in 2015 data with null or missing values
- 420 records found in 2016 data with missing values
- 43 records found in 2017 data with missing values

All the missing records in violation time are dropped.

5 b) 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.

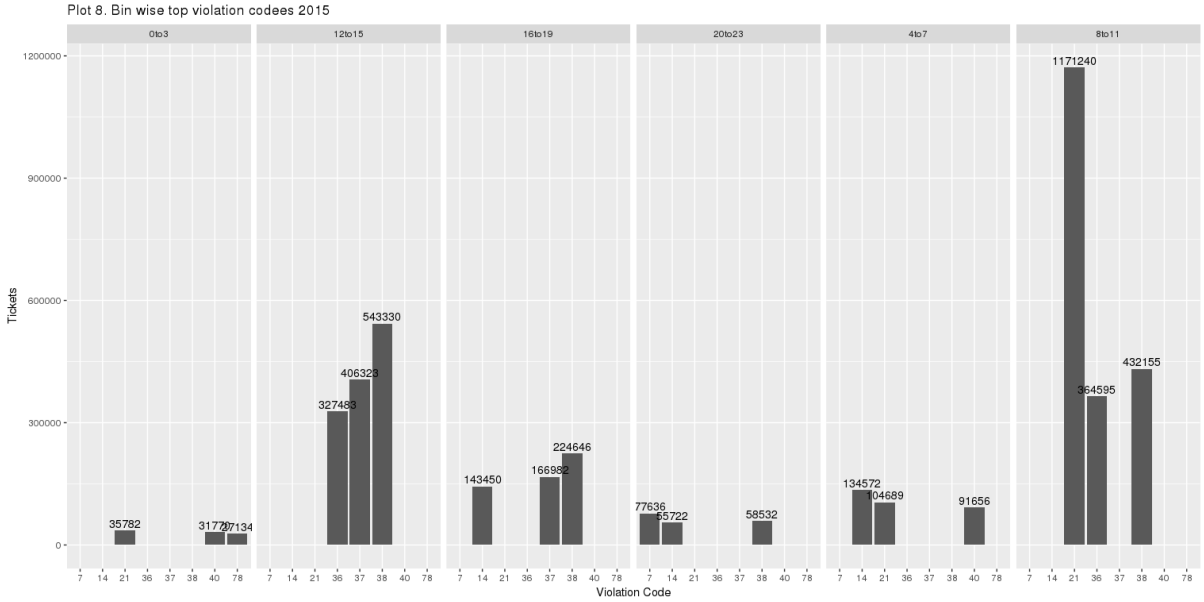
Answer:

- Tt is observed that violation time column in char with values "0953A" "0520P" "0545P" etc.
- It is assumed that first two characters are hours, next two characters are minutes
- Also, A in the end is considered as AM and P as PM
- Concatenated the column with M so that it becomes AM and PM
- Converted all the records with values as 12XXAM to 00XXAM
- Converted the entire column from string to time format.

5 c) 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.

Answer:

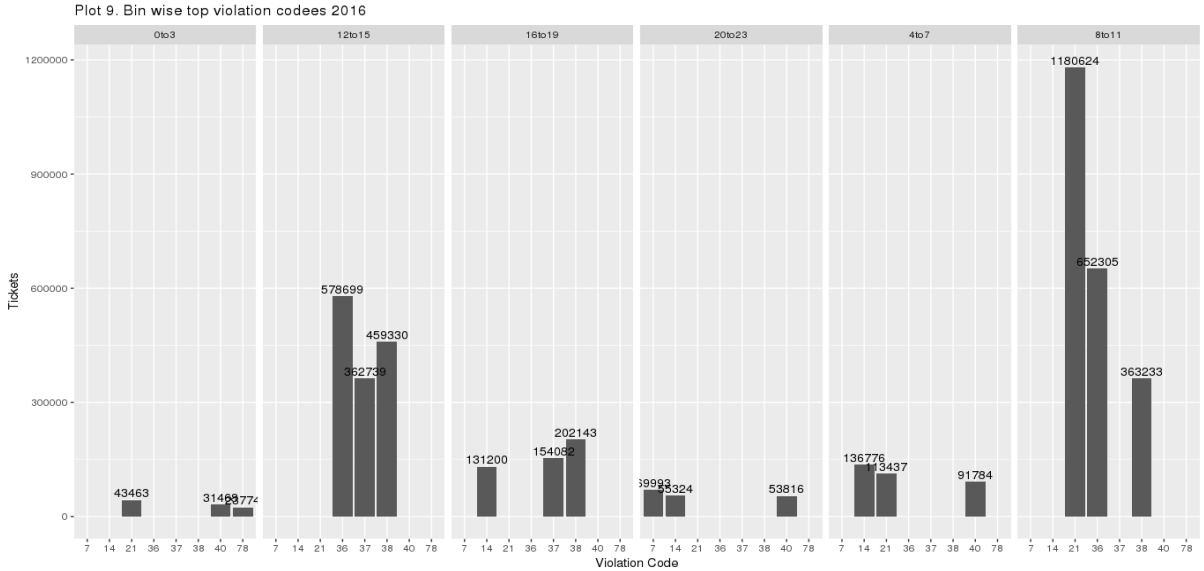
2015 Analysis:



From the plot it is evident that:

- In the bin 0 to 3 hr – 21, 40 and 78 are the major violation codes
- In the bin 4 to 7 hr – 14, 21 and 40 are the major violation codes
- In the bin 8 to 11 hr – 21, 36 and 38 are the major violation codes
- In the bin 12 to 15 hr – 36, 37 and 38 are the major violation codes
- In the bin 16 to 19 hr – 14, 37 and 38 are the major violation codes
- In the bin 20 to 23 hr – 7, 14 and 38 are the major violation codes

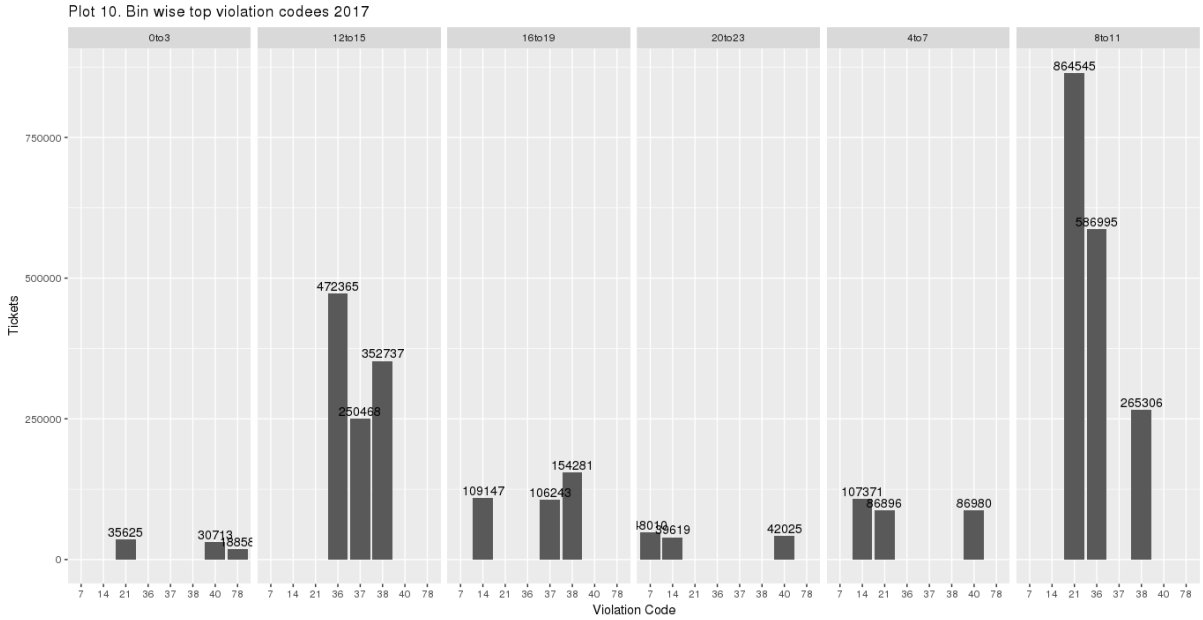
2016 Analysis:



From the plot it is evident that:

- In the bin 0 to 3 hr – 21, 40 and 78 are the major violation codes
- In the bin 4 to 7 hr – 14, 21 and 40 are the major violation codes
- In the bin 8 to 11 hr – 21, 36 and 38 are the major violation codes
- In the bin 12 to 15 hr – 36, 37 and 38 are the major violation codes
- In the bin 16 to 19 hr – 14, 37 and 38 are the major violation codes
- In the bin 20 to 23 hr – 7, 14 and 40 are the major violation codes

2017 bin wise analysis:



From the plot it is evident that:

- In the bin 0 to 3 hr – 21, 40 and 78 are the major violation codes
- In the bin 4 to 7 hr – 14, 21 and 40 are the major violation codes
- In the bin 8 to 11 hr – 21, 36 and 38 are the major violation codes
- In the bin 12 to 15 hr – 36, 37 and 38 are the major violation codes
- In the bin 16 to 19 hr – 14, 37 and 38 are the major violation codes
- In the bin 20 to 23 hr – 7, 14 and 40 are the major violation codes

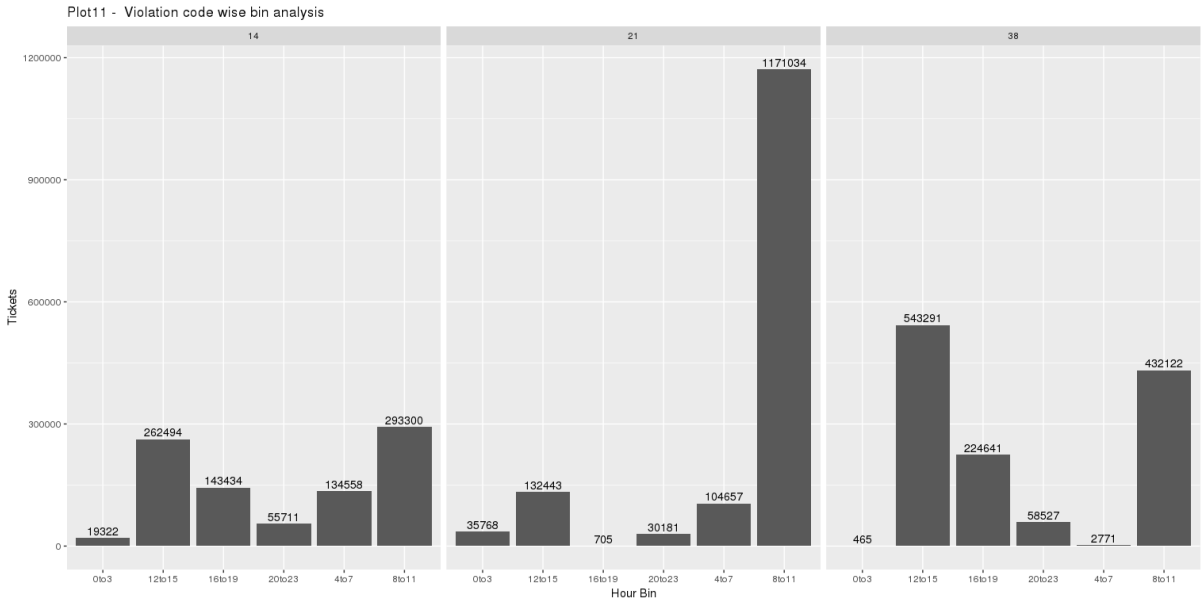
For FY 2016 and 2017 across the bins major violation codes are same.

Same is the case for FY2015 also except for bin 20 to 23.

5 d: For the 3 most commonly occurring violation codes, find the most common time of the day

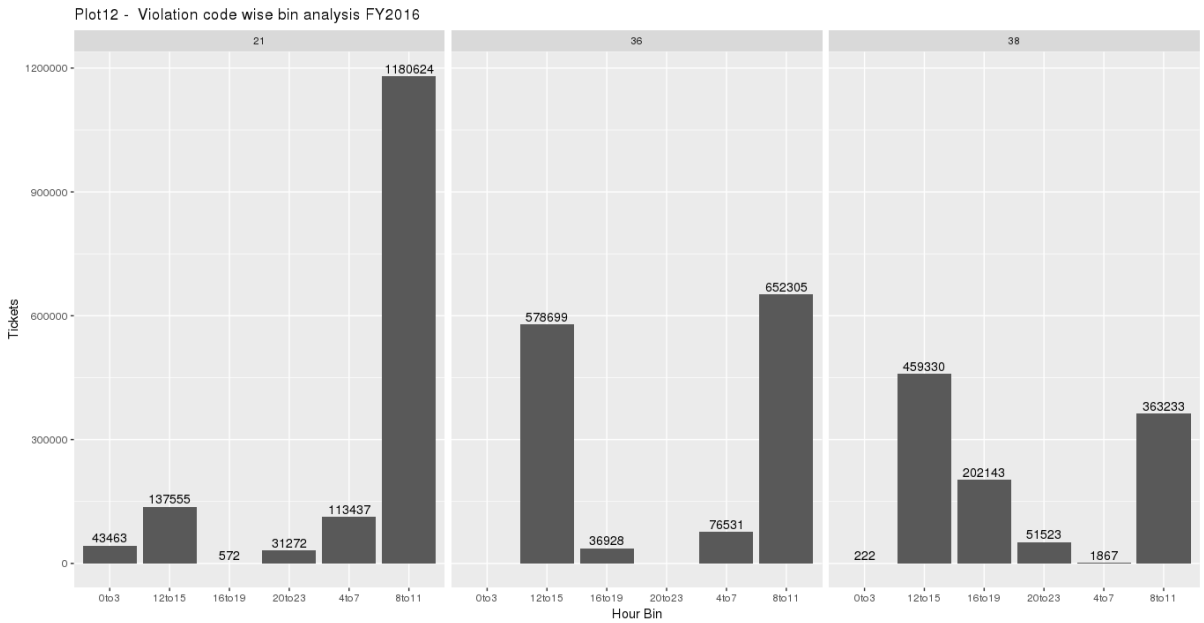
Answer:

2015 Analysis:



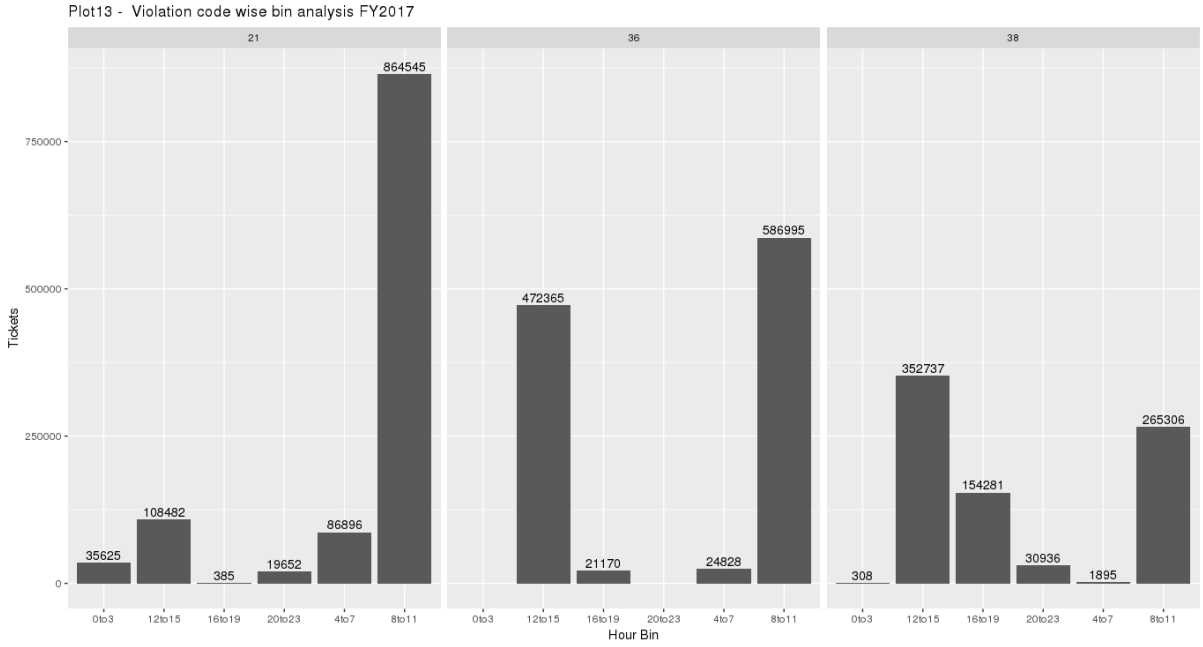
- For Violation code 14 – 8 to 11, 12 to 15 and 16 to 19 are the bins with major ticket
- For Violation code 21 - 8 to 11, 12 to 15 and 4 to 7 are the bins with major ticket
- For Violation code 24 – 12 to 15, 8 to 11 and 16 to 19 are the bins with major ticket

2016 Analysis:



- For Violation code 21 – 8 to 11, 12 to 15 and 4 to 7 are the bins with major ticket
- For Violation code 36 - 8 to 11, 12 to 15 and 4 to 7 are the bins with major ticket
- For Violation code 38 – 12 to 15, 8 to 11 and 16 to 19 are the bins with major ticket

2017 Analysis:



- For Violation code 21 – 8 to 11, 12 to 15 and 4 to 7 are the bins with major ticket
- For Violation code 36 - 8 to 11, 12 to 15 and 4 to 7 are the bins with major ticket
- For Violation code 38 – 12 to 15, 8 to 11 and 16 to 19 are the bins with major ticket

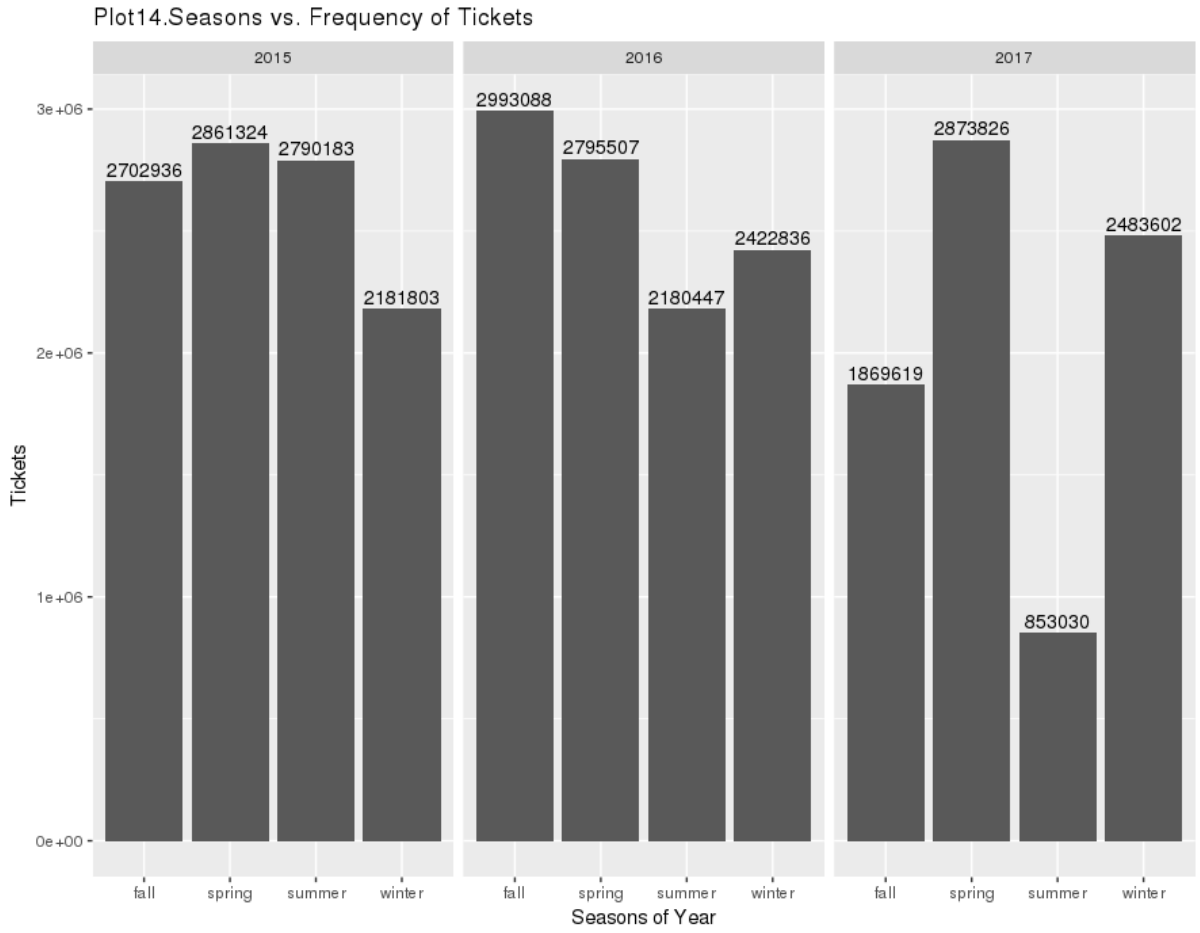
For FY2016 and 2017 violation codes (21, 36 and 38) with major tickets are same and are also issued in the same time bins. Whereas for FY2015 violation codes (14, 21 and 24) have major tickets.

Question 6:

6 a) Let's try and find some seasonality in this data. Divide the year into seasons and find frequency of tickets for each season

Answer:

- Division of year into seasons is done as stated below:
- Winter- for issue months of 12,1,2
- Spring-for issue months of 3,4,5
- Summer-for issue months of 6,7,8
- Fall-for issue months of 9,10,11



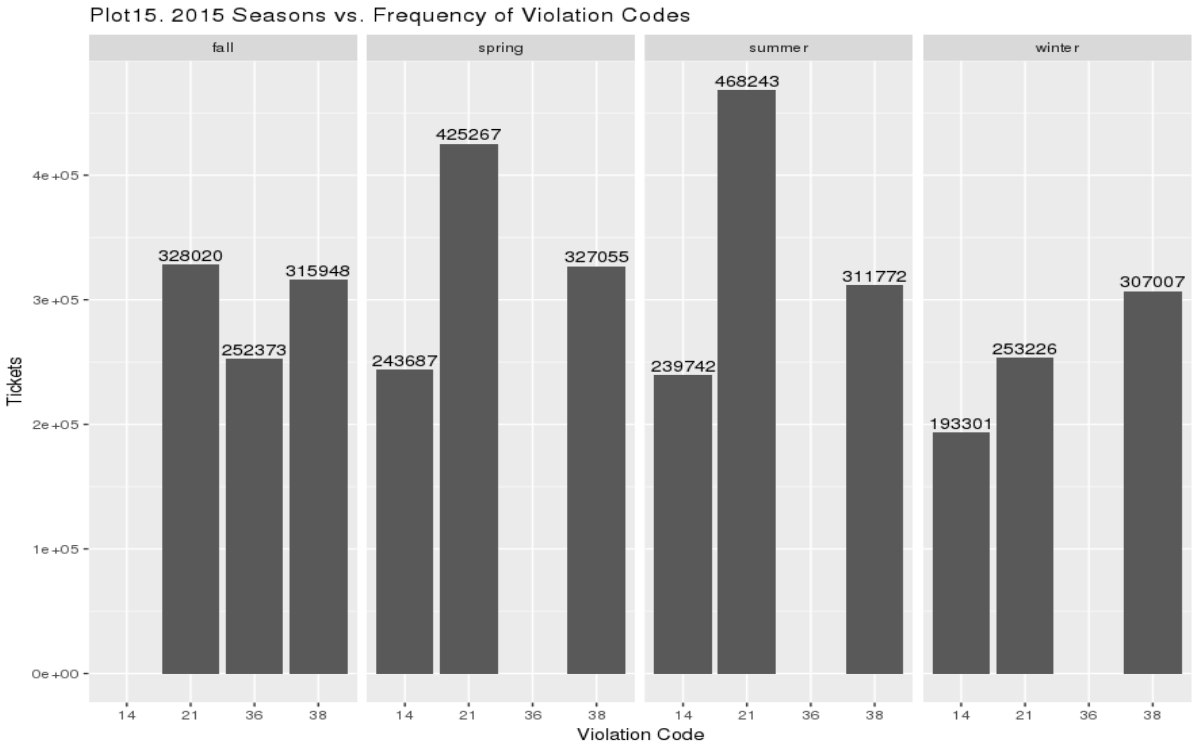
We can observe that:

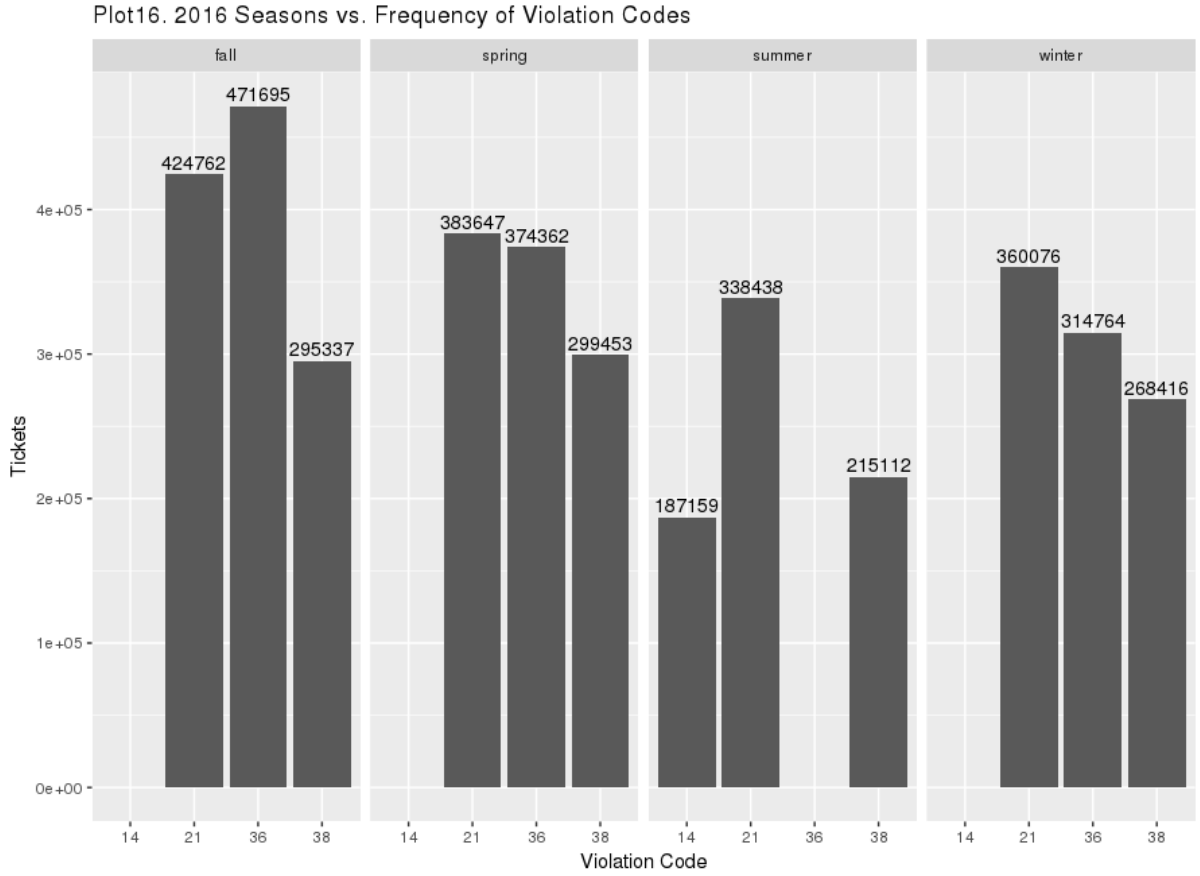
- In year 2015, Spring season has the highest frequency of tickets
- In the year 2016, Fall has the highest frequency of tickets
- In the year 2017, spring has the highest frequency of tickets

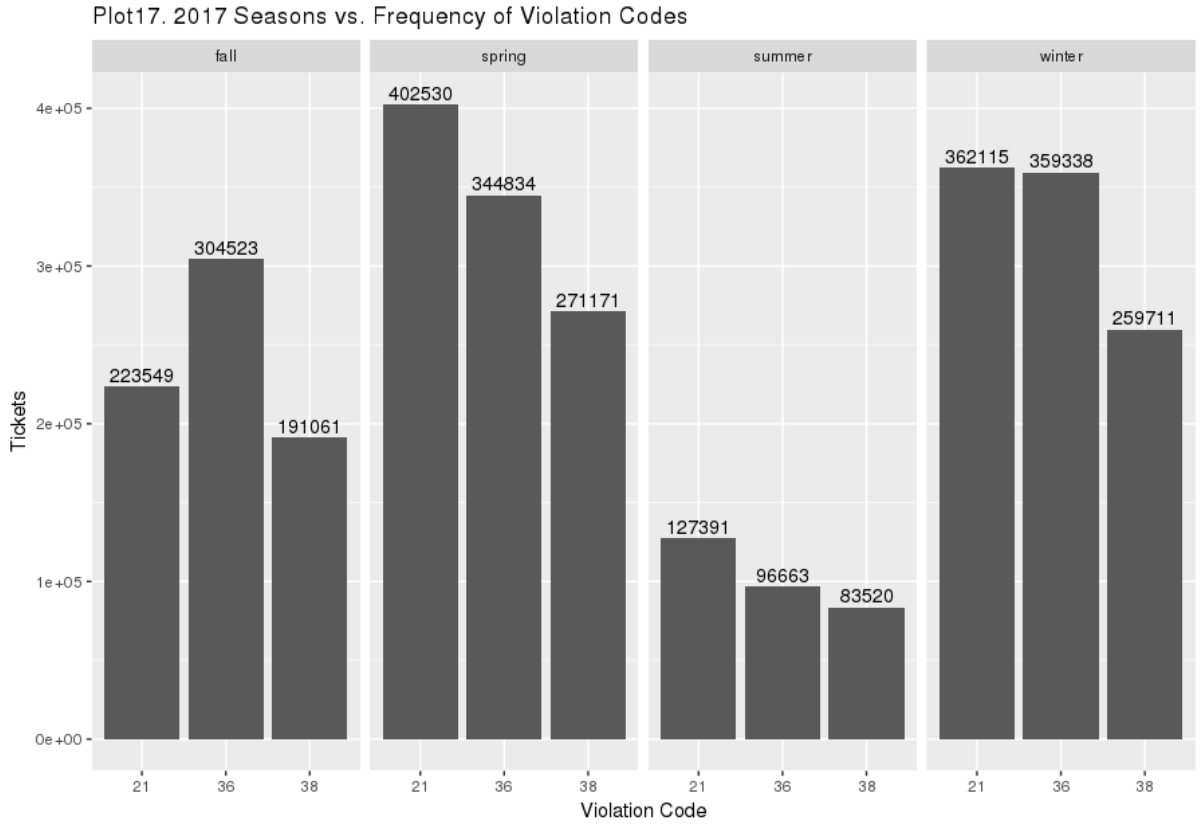
6 b) Find the 3 most common violations for each of these seasons

Answer:

Here below are the graphs which shows 3 most common violations for each season year wise.







7) Find the total occurances of the three most common violation codes. Then find the total amount collected for the three violation codes with maximum tickets.State the code which has the highest total collection. What can you intuitively infer from these findings?

Answer:

For Year 2015

Violation code	Tickets	Year	Average fine	Total fine
21	1474774	2015	55	81112570
38	1261793	2015	50	63089650
14	908797	2015	115	104511655

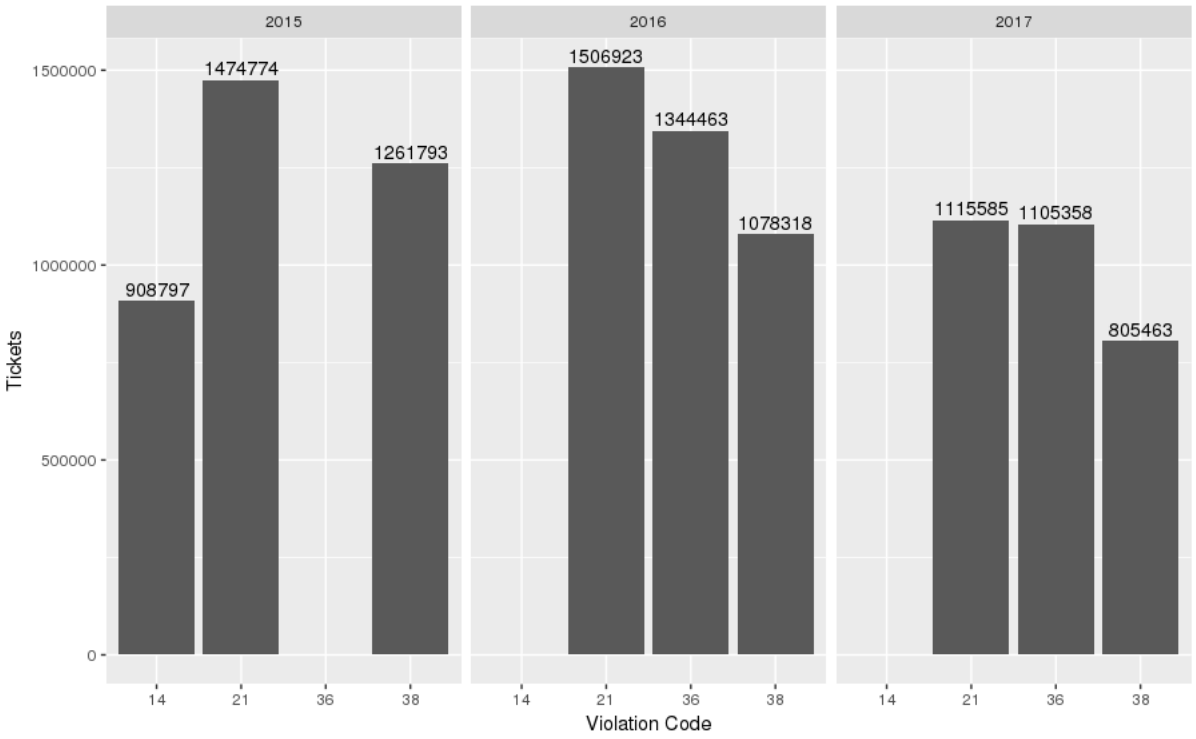
For Year 2016

Violation code	Tickets	Year	Average fine	Total fine
21	1506923	2016	55	82880765
36	1344463	2016	50	67223150
38	1078318	2016	50	53915900

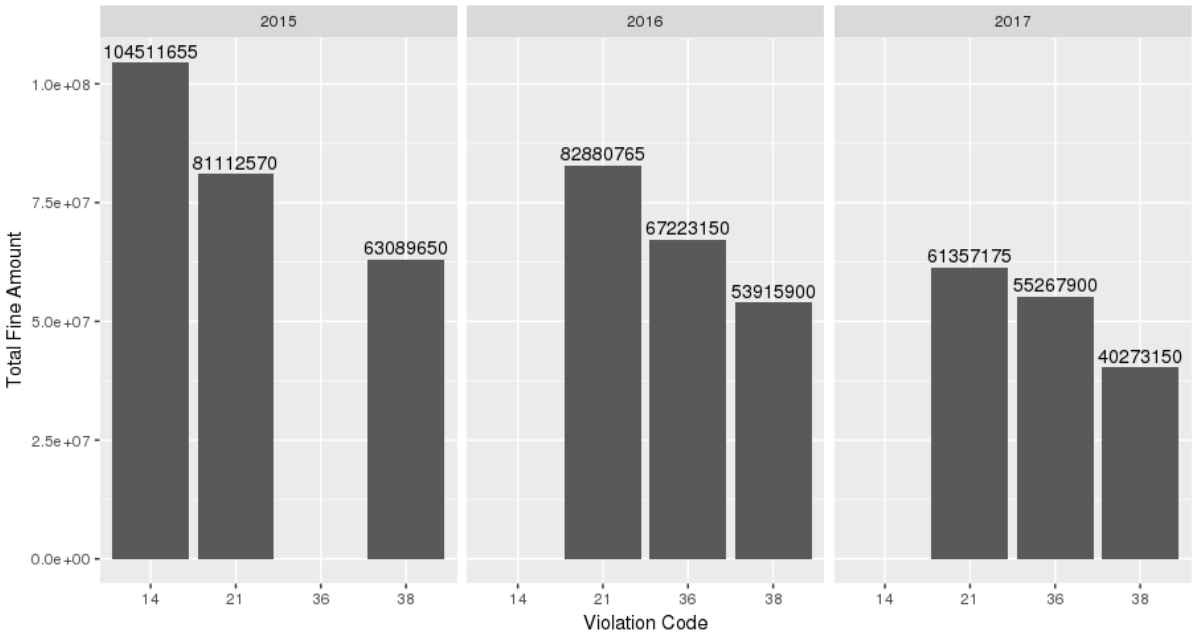
For Year 2017

Violation code	Tickets	Year	Average fine	Total fine
21	1115585	2017	55	61357175
36	1105358	2017	50	55267900
38	805463	2017	50	40273150

Plot18. Comparison of Top 3 Violation Code vs Tickets



Plot19.Top 3 Violation Code vs Total Fine Amount



We can clearly observe that frequency of tickets got increased from 2015 to 2016 whereas this frequency of tickets got down in the year 2017 which means violations has come down. However reduce in the total tickets in FY2017 can be considered as one of the main reason.