UpGrad SparkR Case Study

Swami Prem Pranav Kayashyap

NYC Parking Tickets - Case Study

September 09, 2018

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

- 2015
- 2016
- 2017

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

Violating Precincts (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?

Issuing Precincts (this is the precinct that issued the ticket)

- 2015
- 2016
- 2017

- Comparison of Issuing Precinct

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?

- 2015
- 2016
- 2017

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

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.

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

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

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)

- 2015
- 2016
- 2017

Let's try and find some seasonality in this data

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.

Overview

The NYC Department of Finance collects details on Parking tickets issued to vehicles within its jurisdiction. This data has been made public and we shall use the data for the years 2015, 2016 and 2017 to analyse trends and answer some pertinent questions with regards to Parking violations.

We shall conduct the study in SparkR and this will also allow us to understand SparkR syntax and its similarities/dissimilarities with base R.

Goals

1. **Data Quality Verification and Cleaning**

   We shall verify the accuracy and relevance of the data and remove any unnecessary fields in order to obtain clean datasets. The major steps that have been taken are

   a. Cleaning column names
   b. Removing duplicate rows.
   c. Converting date fields to Timestamp values.
   d. Removing unnecessary data and columns.
   e. Appending AM/PM to necessary fields.

2. **Overview and Examining the dataset**

   We have examined the dataset and attempted to answer the following questions

   a. The total number of tickets for each year.
   b. The number of unique states that the violating cars come from.
   c. The number of parking tickets that don’t have an address.

3. **Deriving and Comparing Metrics**

   We have performed exploratory data analysis on the datasets to derive answers for these questions

   a. The frequency of each Violation code (top 5).

b. The frequency of violation for each Vehicle body type and Vehicle make (top 5).
c. The highest number of Violating and Issuing Precincts (top 5).
d. The Violation code frequency ((top 5)) across the three most frequent Violating Precincts along with further correlation analysis.
e. Cleaning the Violating Time Field, distributing it into 6 bins, finding the 3 most commonly occurring violations and correlating them to the understand if they have a time trend.
f. Analysing seasonality trend in the data and finding the 3 most common violations for each season.
g. Find the revenue accrued by NYC Police Department through the 3 most frequent violations.

Data Specifications

1. Data Sets

The datasets that have been used are

a. https//www.kaggle.com/new-york-city/nyc-parking-tickets/data#Parking_Violations_Issued_-_Fiscal_Year_2015.csv
b. https//www.kaggle.com/new-york-city/nyc-parking-tickets/data#Parking_Violations_Issued_-_Fiscal_Year_2016.csv
c. https//www.kaggle.com/new-york-city/nyc-parking-tickets/data#Parking_Violations_Issued_-_Fiscal_Year_2017.csv

2. Tools

The tools and platforms used are

a. Hadoop platform and HDFS for storing and analysis of the datasets.

b. RStudio on Hadoop along with SparkR for EDA.

Results

- **Data Quality Verification and Cleaning**
  a. Cleaning column names
     We have removed spaces from the Column names and replaced the space between strings with "_".
  b. Removing duplicate rows

Dataset	Original Row Num	Cleaned Row Num
nyc_parking_tkts_2015	11,809,233	10,951,256
nyc_parking_tkts_2016	10,626,899	10,626,899
nyc_parking_tkts_2017	10,803,028	

  c. Converting date fields to Timestamp values
     We have converted the following fields to Timestamp
     i. Issue_Date
     ii. Vehicle_Expiration_Date
     iii. Date_First_Observed
     iv. Violation_Time
     v. Time_First_Observed
     vi. From_Hours_In_Effect
     vii. To_Hours_In_Effect

  d. Removing unnecessary data and columns
     We have removed the following rows and columns
     i. Rows with Issue_Date before the respective Fiscal year

Dataset	Original Row Num	Cleaned Row Num with dates within Fiscal period
nyc_parking_tkts_2015	10,951,256	10,598,035
nyc_parking_tkts_2016	10,626,899	10,396,894
nyc_parking_tkts_2017	10,803,028	10,539,563

ii. Columns that aren't present in all Data sets

1. No_Standing_or_Stopping_Violation
2. Hydrant_Violation
3. Double_Parking_Violation
4. Latitude
5. Longitude
6. Community_Board
7. Community_Council
8. Census_Tract
9. BIN
10. BBL
11. NTA

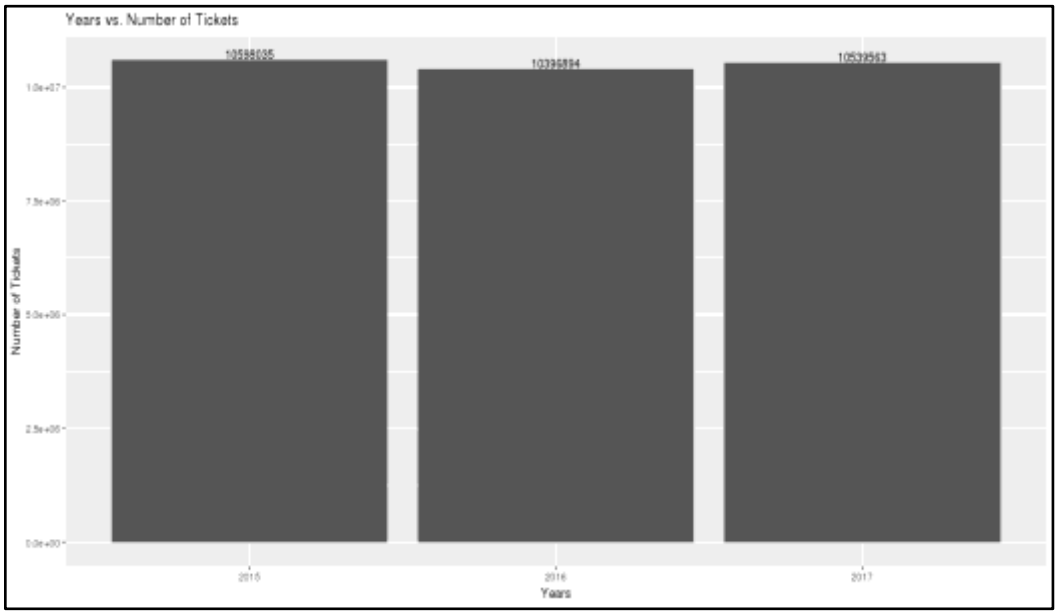
e. Appending AM/PM to necessary fields

We have appended the correct Clock cycle to these columns

i. Violation_Time
ii. Time_First_Observed
iii. From_Hours_In_Effect
iv. To_Hours_In_Effect

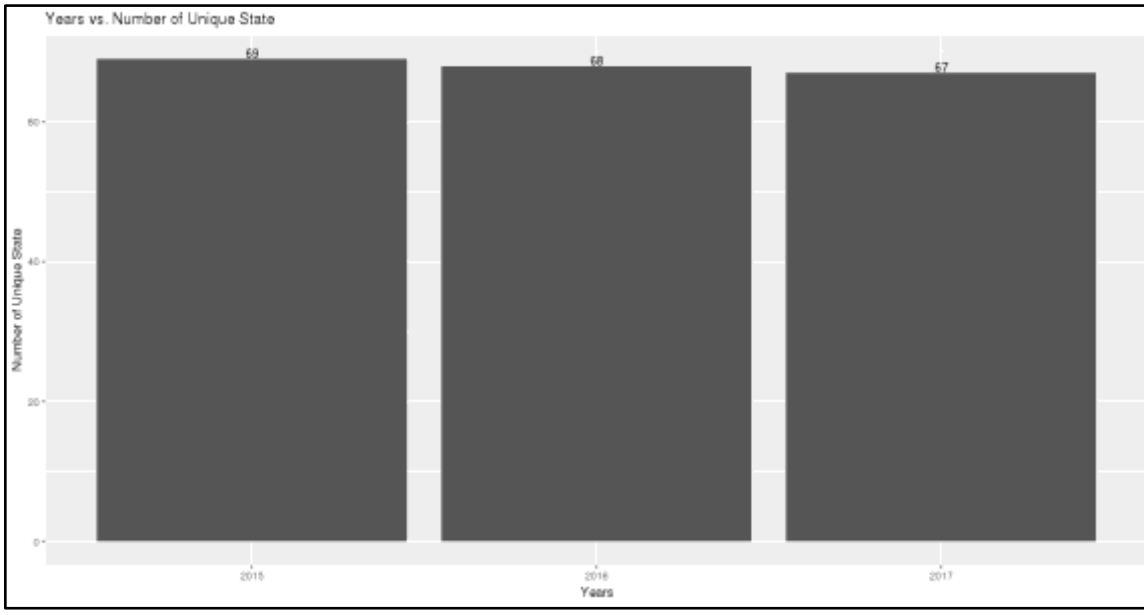
- **Data Quality Verification and Cleaning**
  a. Find total number of tickets for each year

2015	2106	2017
10,598,035	10,396,894	10,539,563



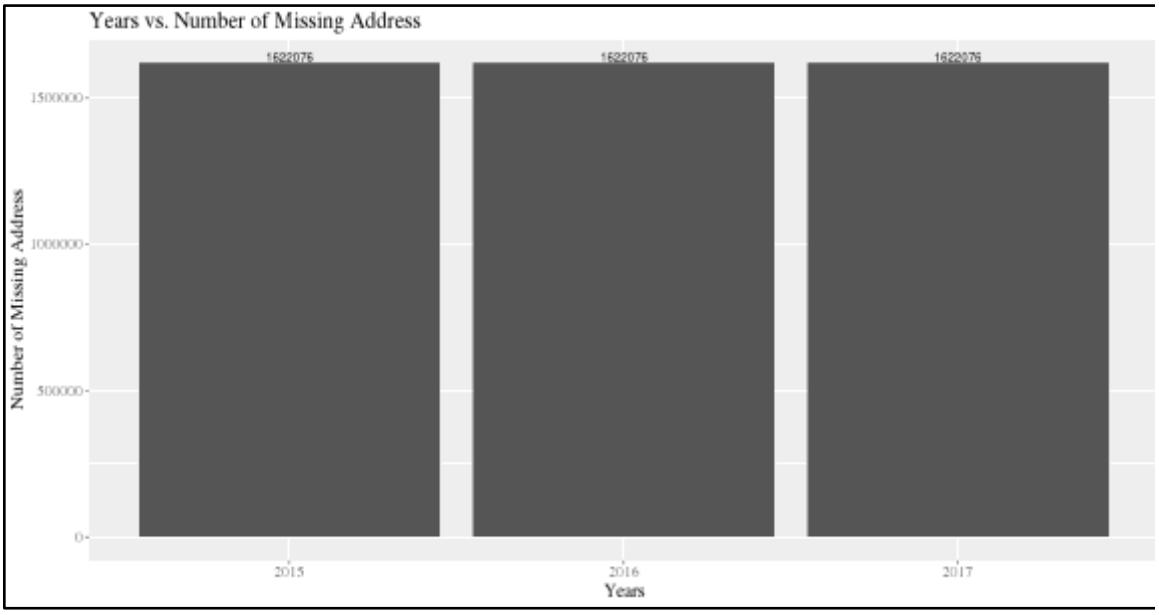
b. Find out how many unique states the cars which got parking tickets came from

2015	2106	2017
69	68	67



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

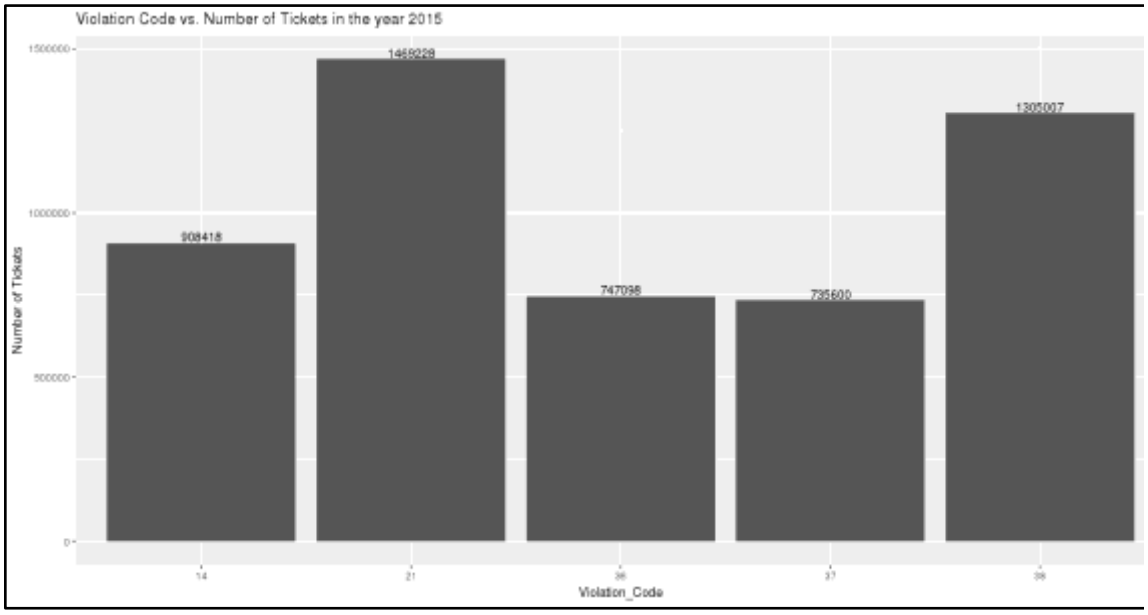
2015	2106	2017
1622076	1622076	1622076



- **Deriving and Comparing Metrics**
  a. How often does each violation code occur? (frequency of violation codes - find the top 5)

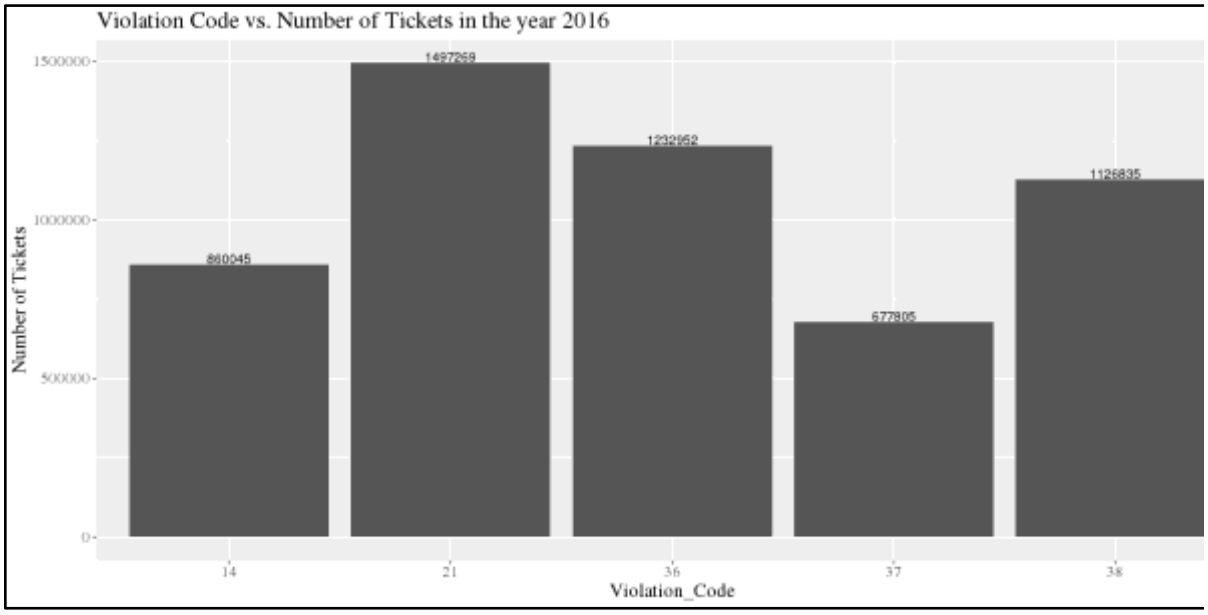
  2015

Violation Code	Frequency
21	1,469,228
38	1,305,007
14	908,418
36	747,098
37	735,600



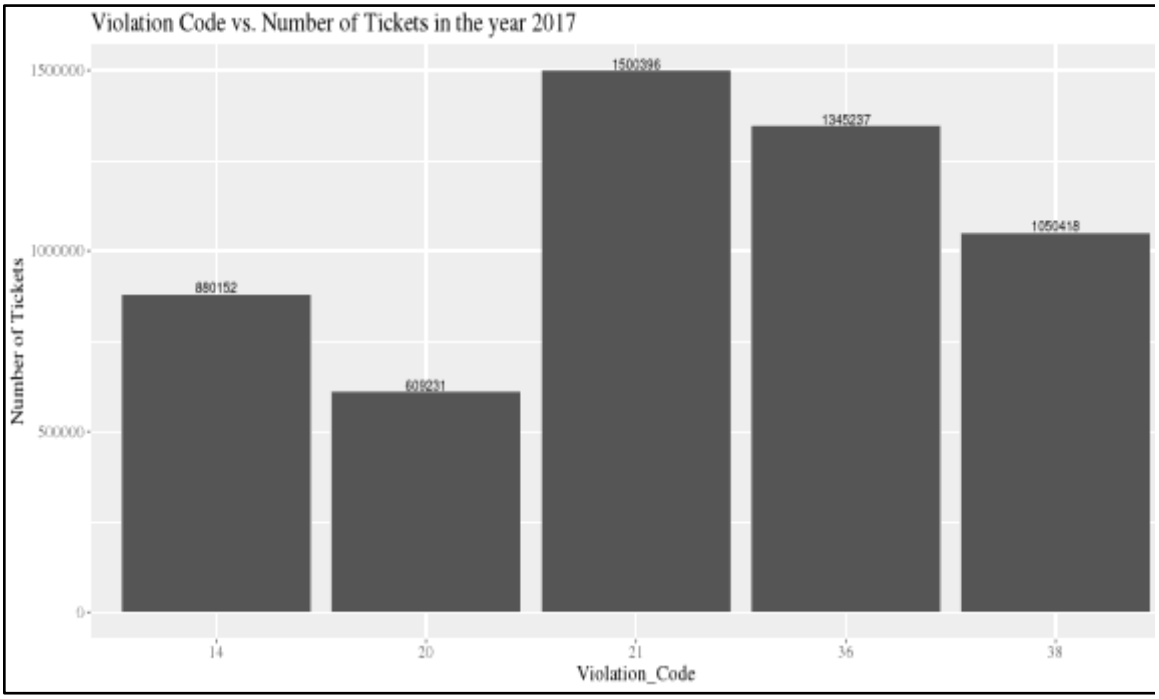
2016

Violation Code	Frequency
21	1,497,269
36	1,232,952
38	1,126,835
14	860,045
37	677,805



2017

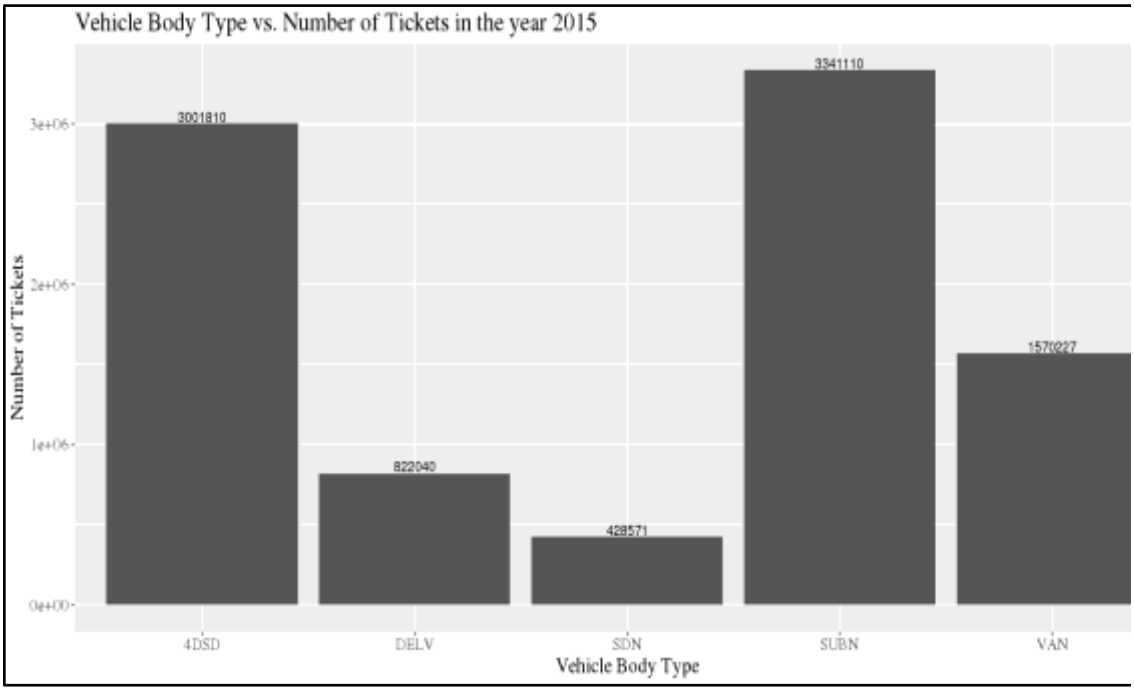
Violation Code	Frequency
21	1,500,396
36	1,345,237
38	1,050,418
14	880,152
20	609,231

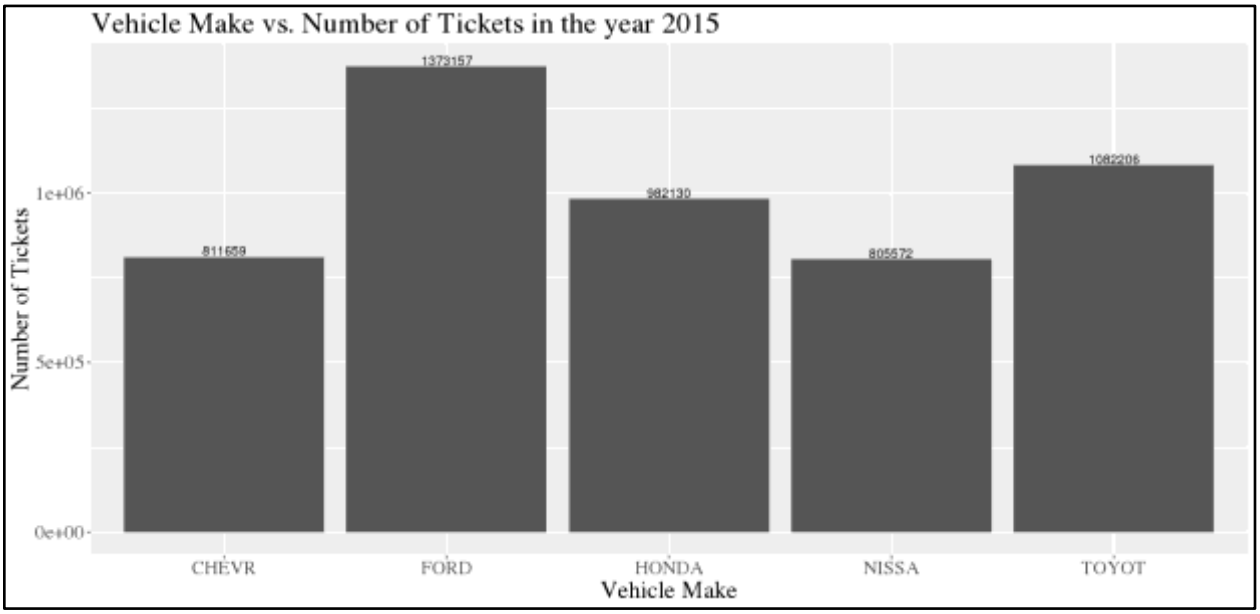


b. How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

2015

Vehicle Body Type	Frequency		Vehicle Make	Frequency
SUBN	3,341,110		FORD	1,373,157
4DSD	3,001,810		TOYOT	1,082,206
VAN	1,570,227		HONDA	982,130
DELV	822,040		CHEVR	811,659
SDN	428,571		NISSA	805,572

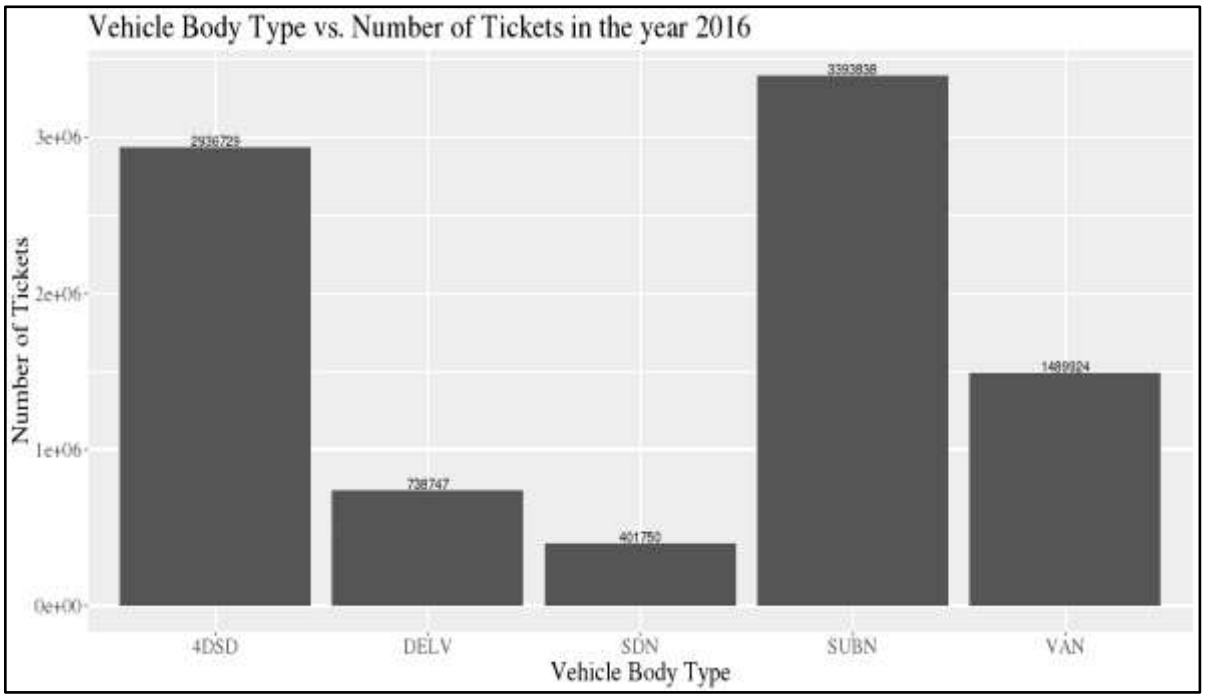


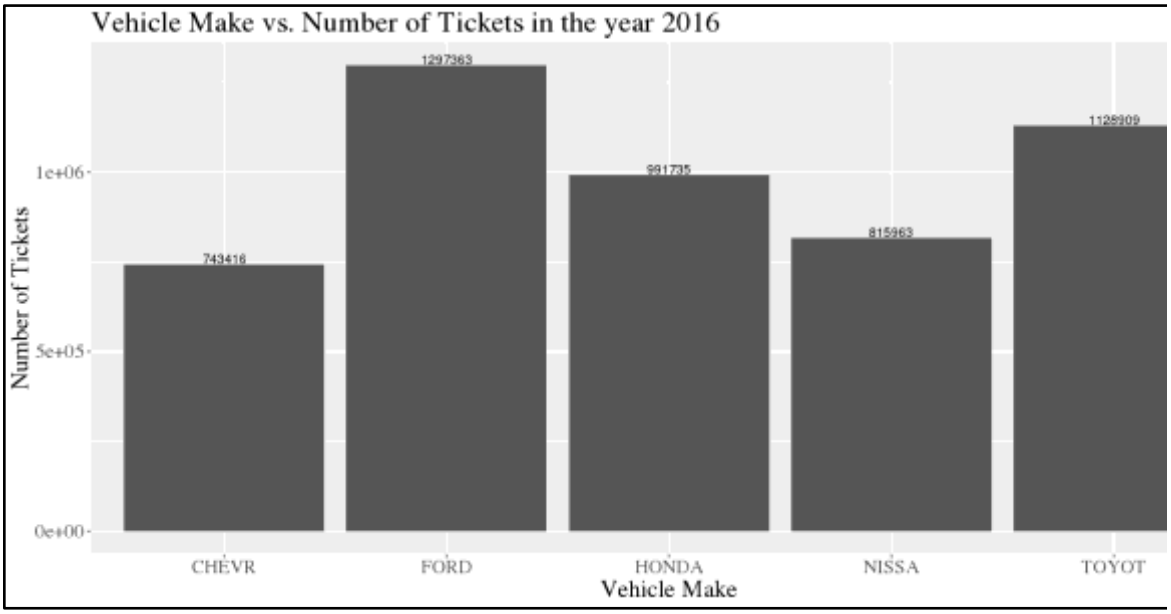


2016

Vehicle Body Type	Frequency		Vehicle Make	Frequency

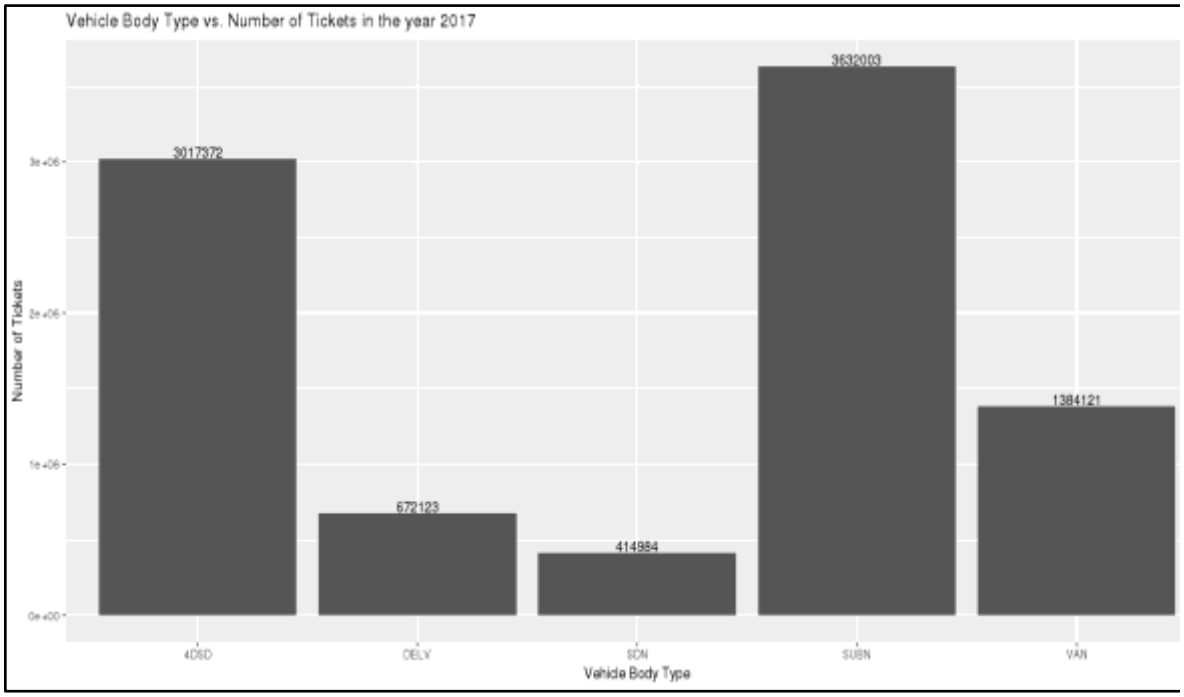
SUBN	3,393,838		FORD	1,297,363
4DSD	2,936,729		TOYOT	1,128,909
VAN	1,489,924		HONDA	991,735
DELV	738,747		NISSA	815,963
SDN	401,750		CHEVR	743,416

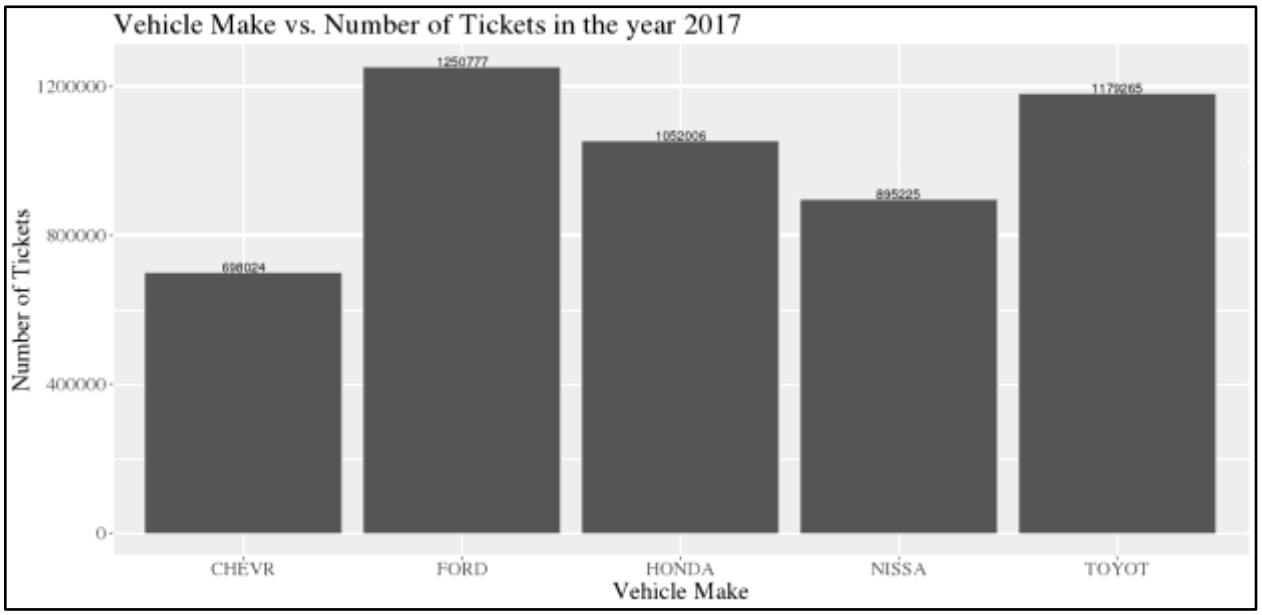




2017

Vehicle Body Type	Frequency		Vehicle Make	Frequency
SUBN	3,632,003		FORD	1,250,777
4DSD	3,017,372		TOYOT	1,179,265
VAN	1,384,121		HONDA	1,052,006
DELV	672,123		NISSA	895,225
SDN	414,984		CHEVR	698,024





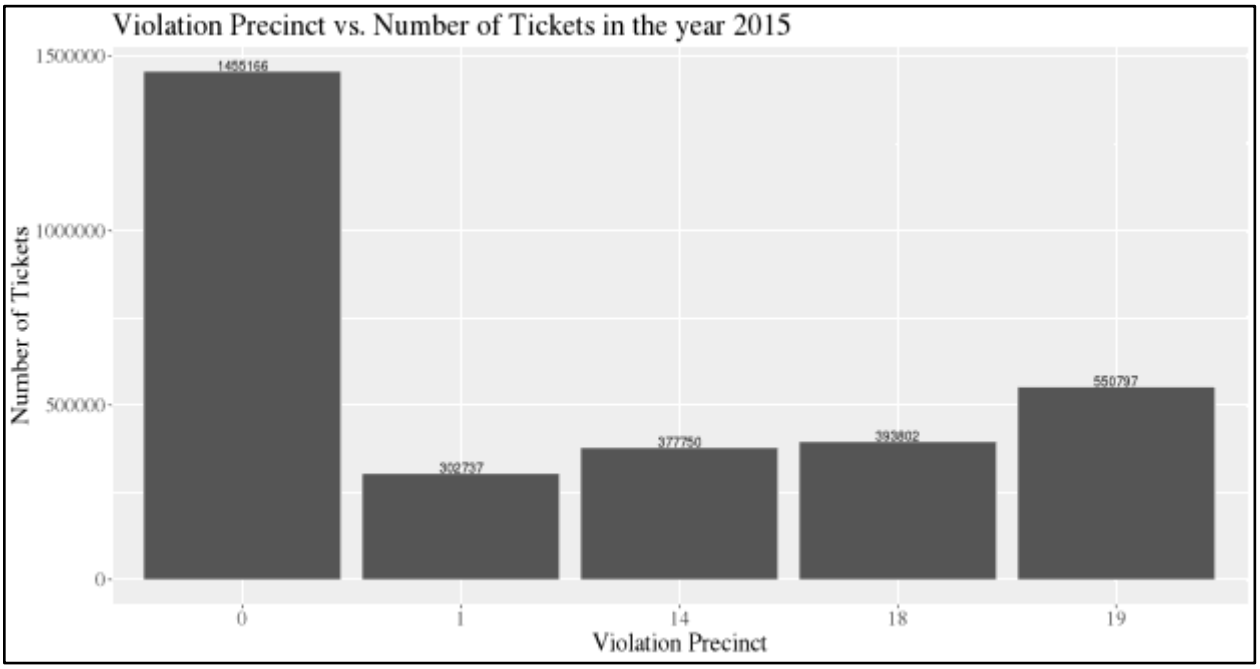
c. A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:

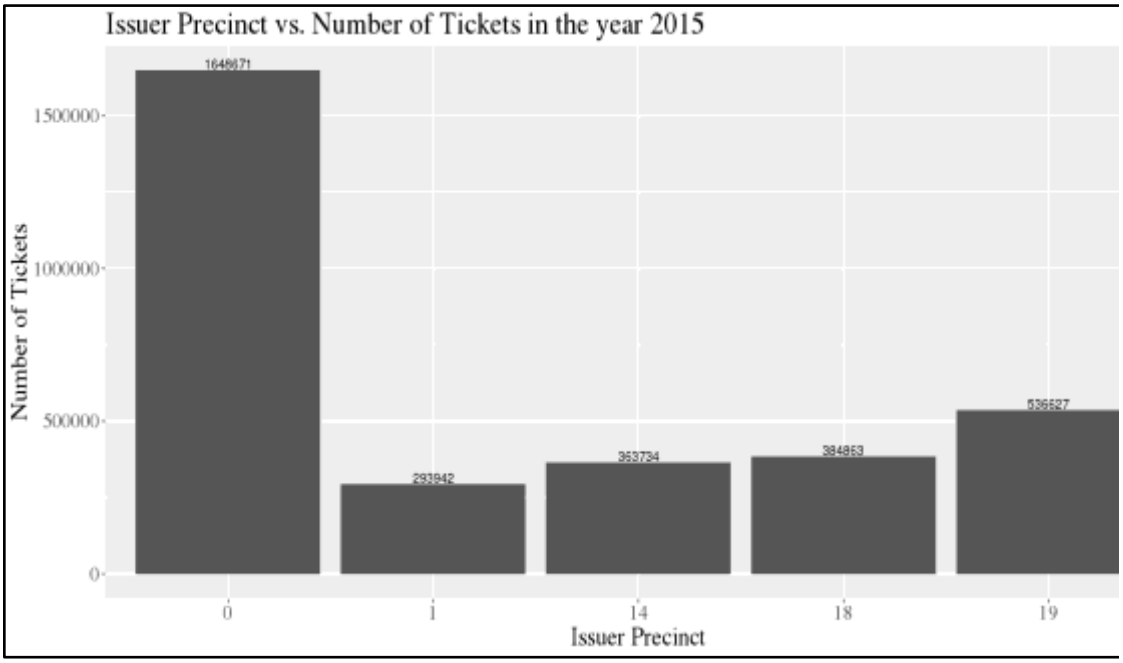
Violating Precincts (this is the precinct of the zone where the violation occurred). Using this, can you make any insights for parking violations in any specific areas of the city?

Issuing Precincts (this is the precinct that issued the ticket)

2015

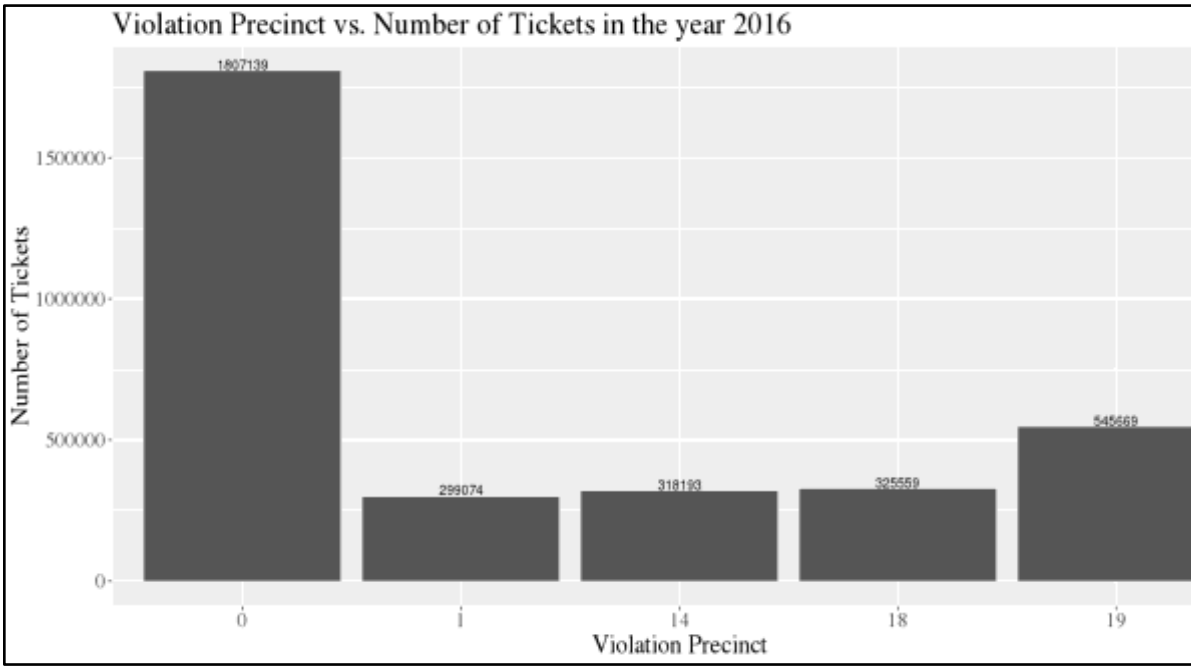
Violating Precinct	Frequency		Issuing Precinct	Frequency
0	1,455,166		0	1,648,671
19	550,797		19	536,627
18	393,802		18	384,863
14	377,750		14	363,734
1	302,737		1	293,942

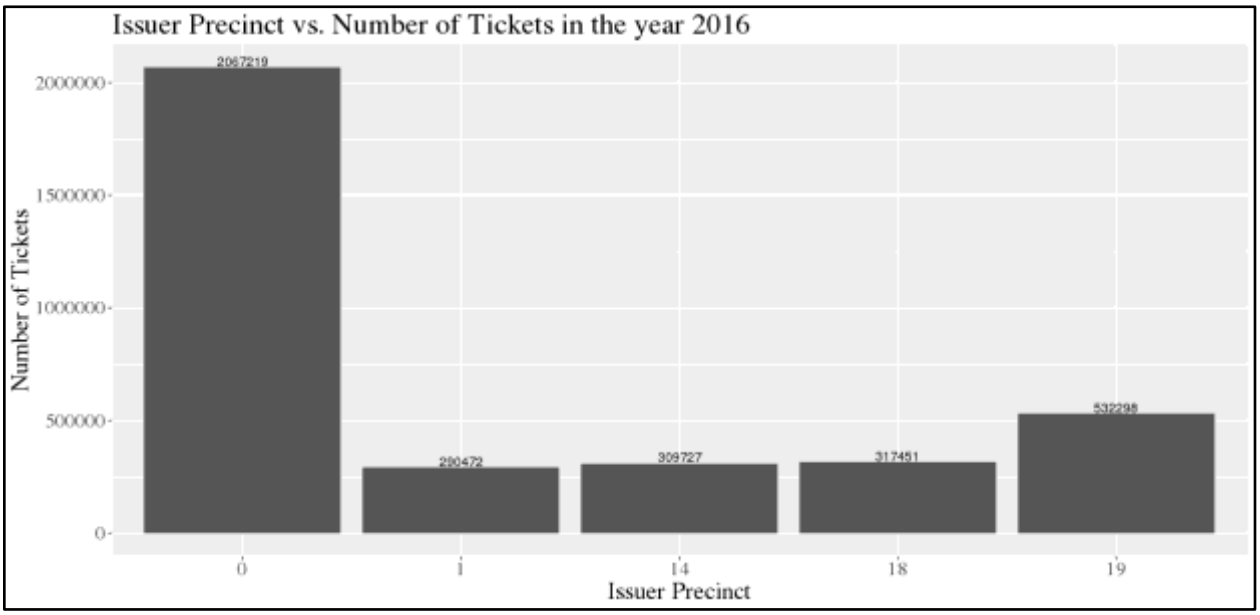




2016

Violating Precinct	Frequency		Issuing Precinct	Frequency
0	1,807,139		0	2,067,219
19	545,669		19	532,298
18	325,559		18	317,451
14	318,193		14	309,727
1	299,074		1	290,472

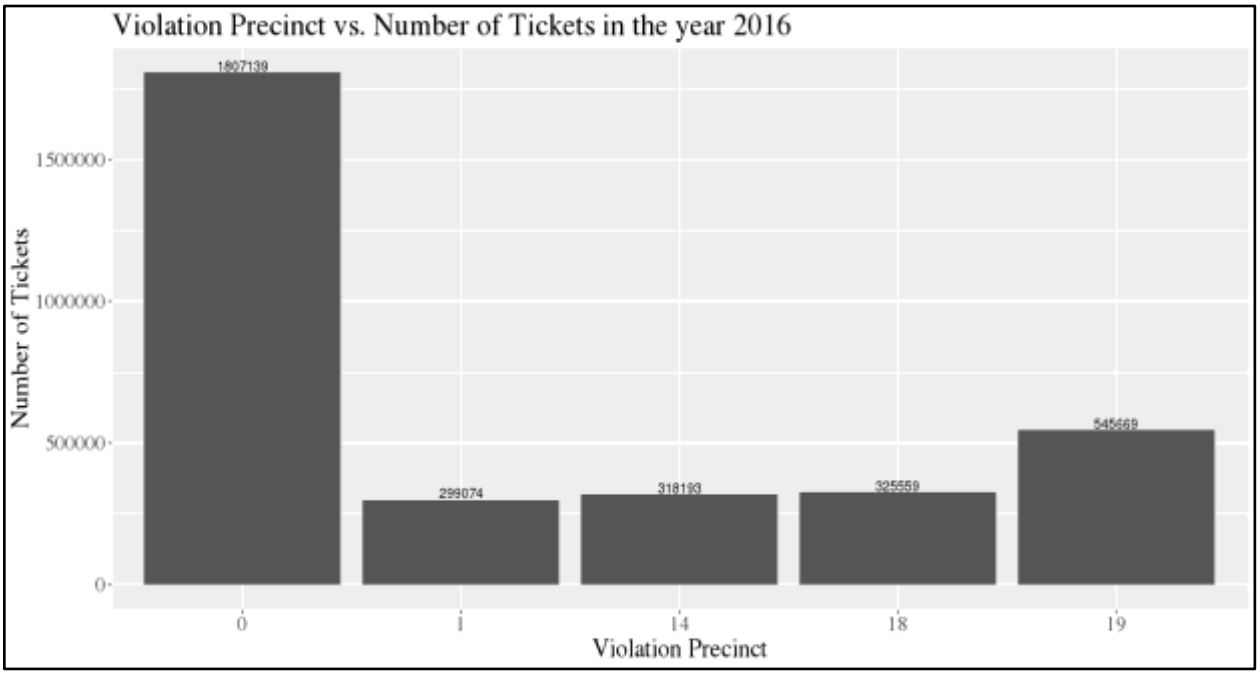


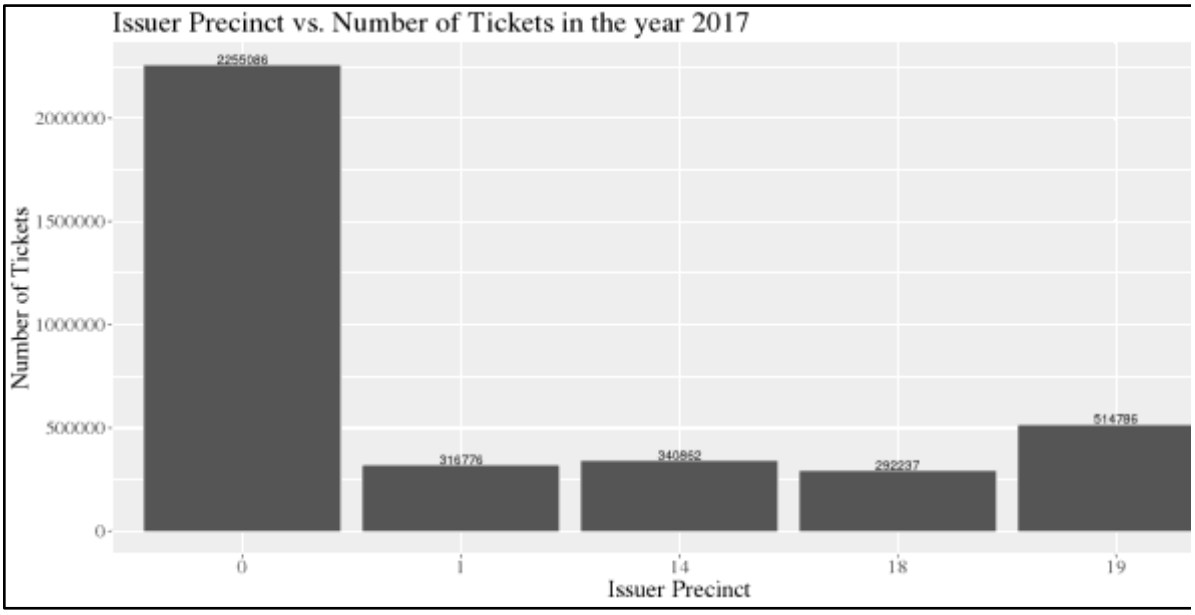


2017

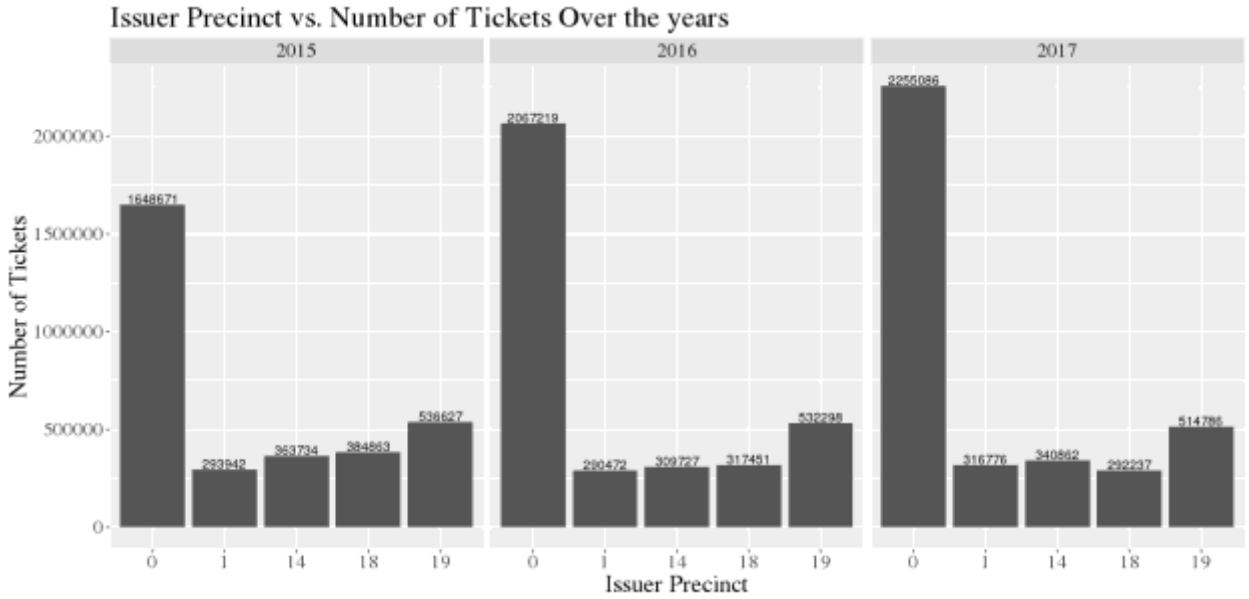
Violating Precinct	Frequency		Issuing Precinct	Frequency
0	1,950,083		0	2,255,086

19	528,317		19	514,786
14	347,736		14	340,862
1	326,961		1	316,776
18	302,008		18	292,237





Comparison of Issuing Precinct

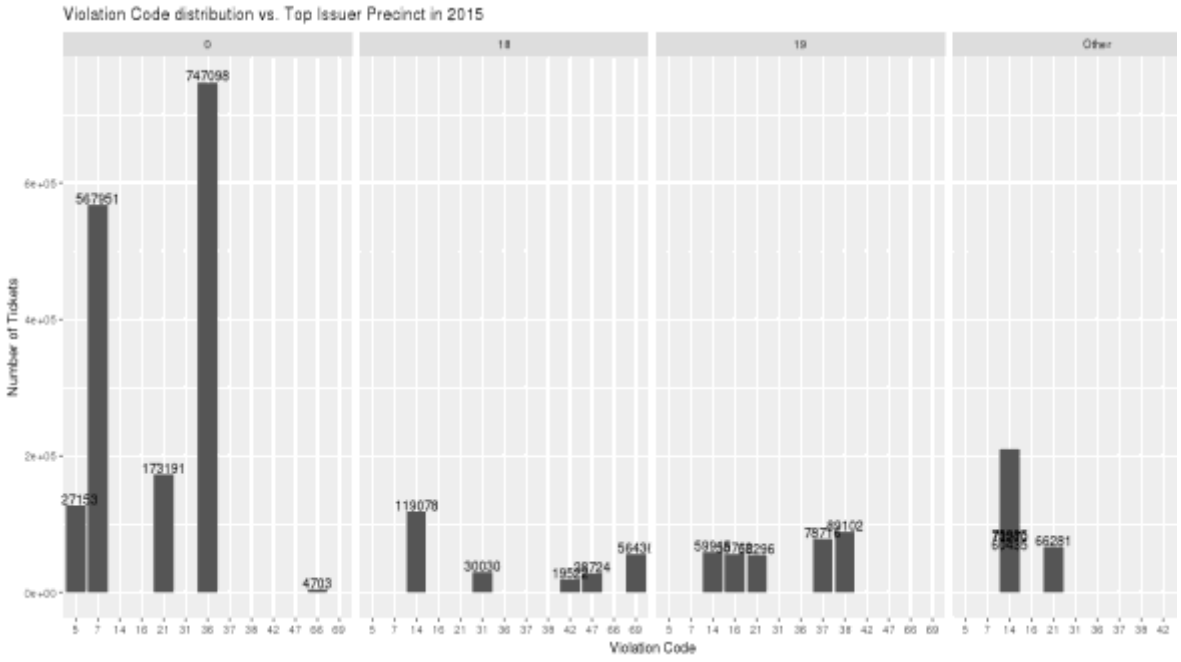


d. Find the violation code frequency across 3 precincts which have issued the most number of tickets - do these precinct zones have an exceptionally high frequency of certain violation codes? Are these codes common across precincts?

The most frequent Issuing Precincts are 0, 19 and 18.

2015

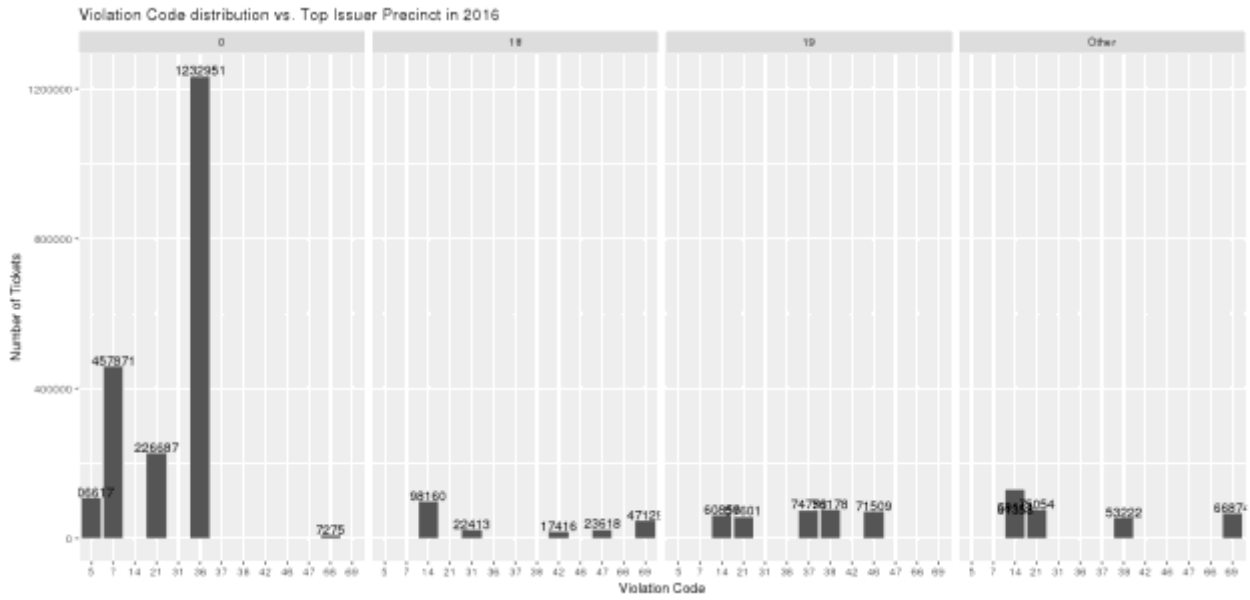
Violation_Code	no_of_tickets	Issuer_Precinct
36	747,098	0
7	567,951	0
21	1,73,191	0
5	1,27,153	0
6	4703	0
38	89,102	19
37	78,716	19
14	59,915	19
16	55,762	19
21	55,296	19
14	1,19,078	18
69	56,436	18
31	30,030	18
47	28,724	18
42	19,522	18



2016

Violation_Code	no_of_tickets	Issuer_Precinct
36	1,232,951	0
7	457,871	0
21	226,687	0
5	106,617	0
66	7,275	0
38	76,178	19
37	74,758	19
46	71,509	19
14	60,856	19
21	57,601	19

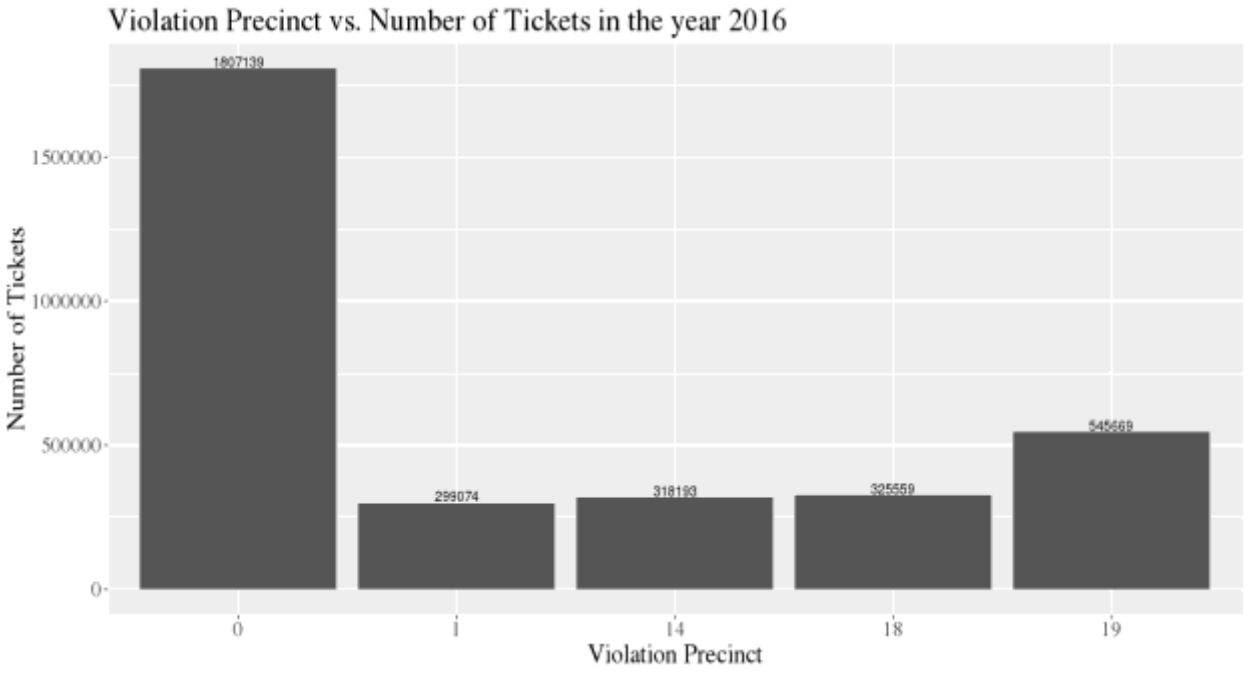
14	98,160	18
69	47,129	18
47	23,618	18
31	22,413	18
42	17,416	18



2017

Violation_Code	no_of_tickets	Issuer_Precinct
36	1,345,237	0
7	464,690	0
21	258,771	0

5	130,963	0
66	9,281	0
46	84,789	19
38	71,631	19
37	71,592	19
14	56,873	19
21	54,033	19
14	90,145	18
69	36,246	18
47	23,487	18
31	21,292	18
46	14,394	18



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

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.

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

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

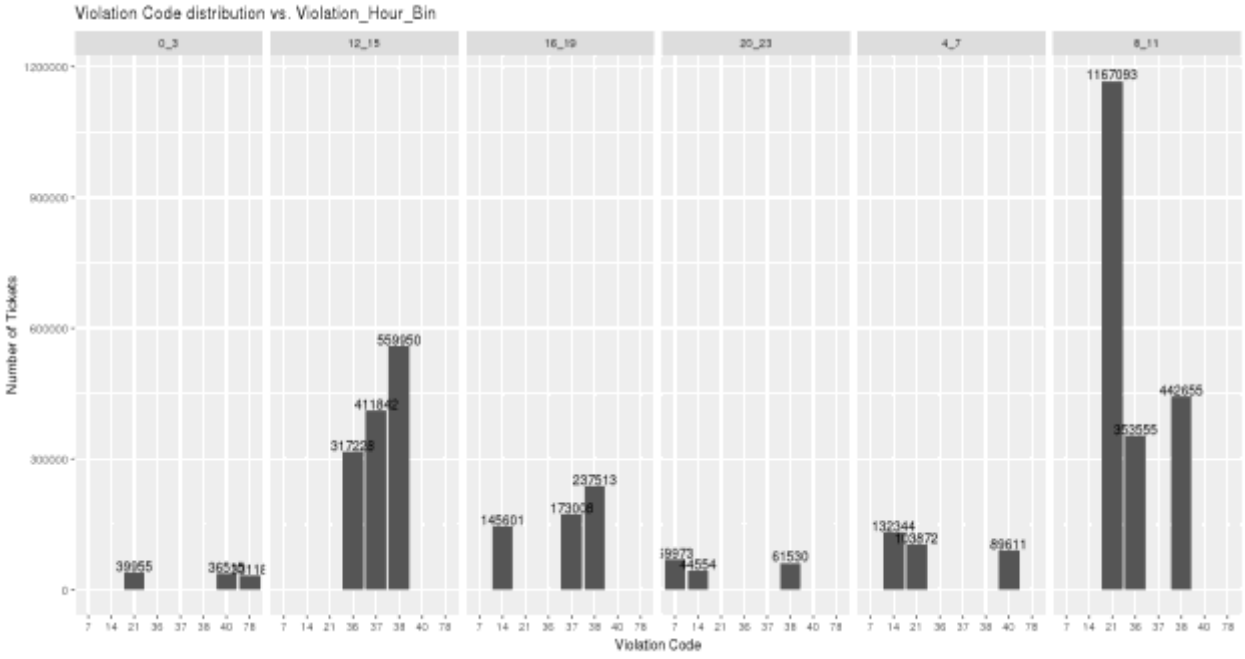
Now, try another direction. For the 3 most commonly occurring violation codes, find the most common times of day (in terms of the bins from the previous part)

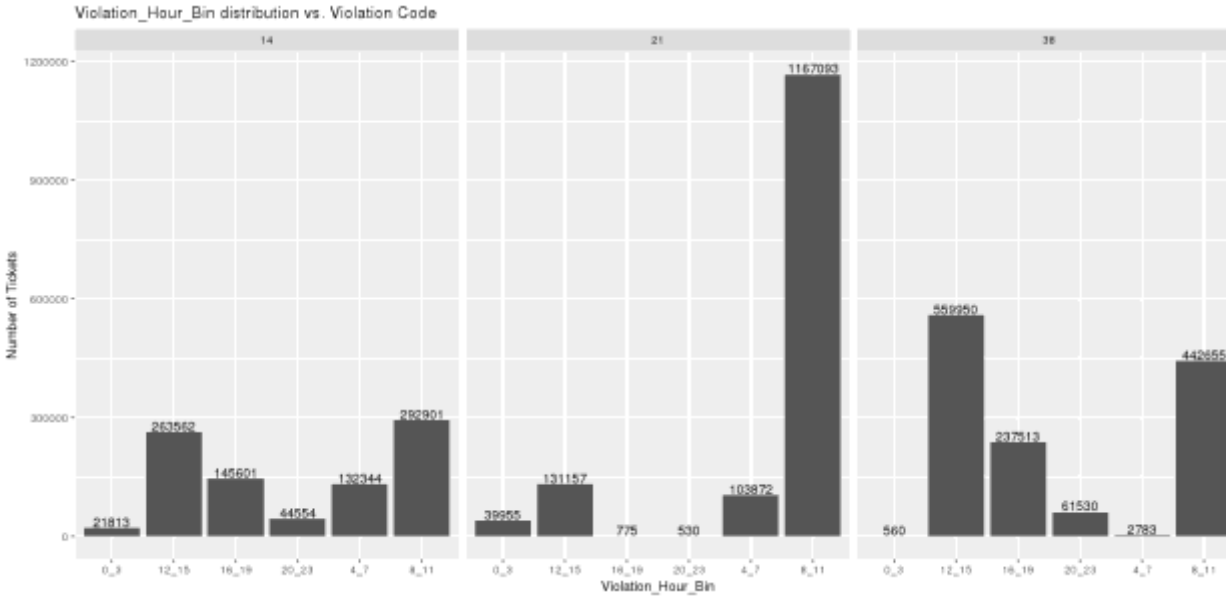
2015

Total_Count	Missing_violation_time	Percent_missing_violation_time
10,598,035	61,603	0.5812681

Violation_Hour_Bin	Violation_Code	num_of_tkts
16_19	38	237513
16_19	37	173008
16_19	14	145601
8_11	21	1167093
8_11	38	442655
8_11	36	353555
4_7	14	132344
4_7	21	103872
4_7	40	89611

12_15	38	559950
12_15	37	411842
12_15	36	317228
0_3	21	39955
0_3	40	36515
0_3	78	33118
20_23	7	69973
20_23	38	61530
20_23	14	44554

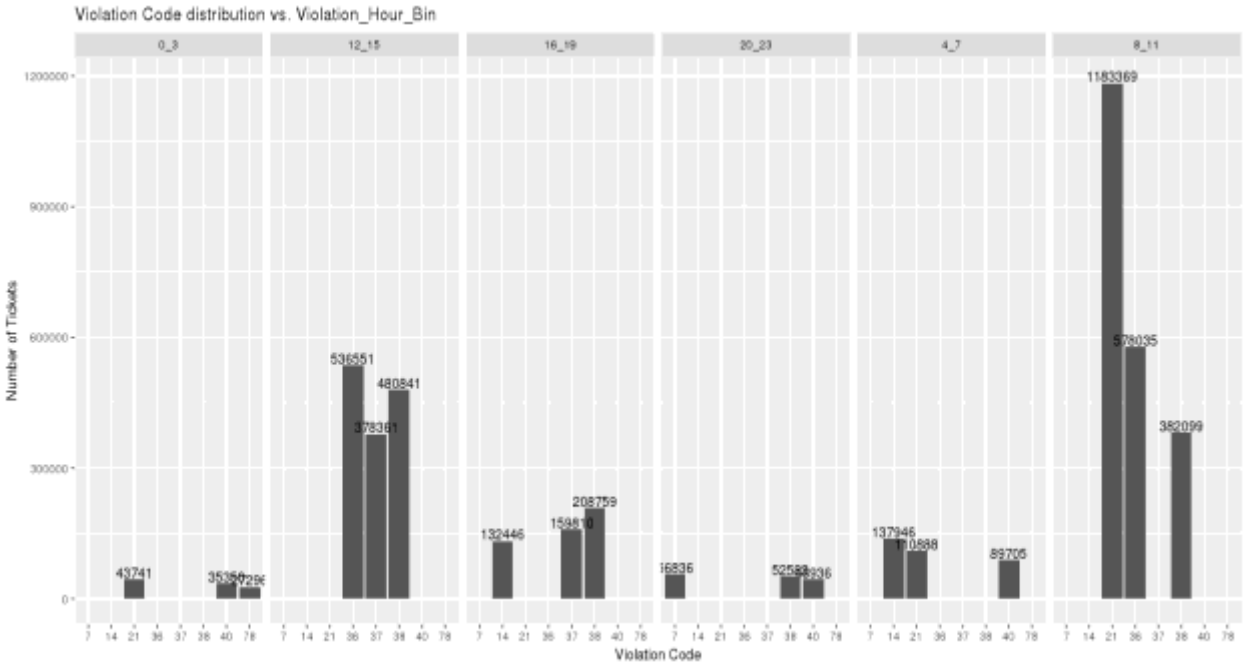


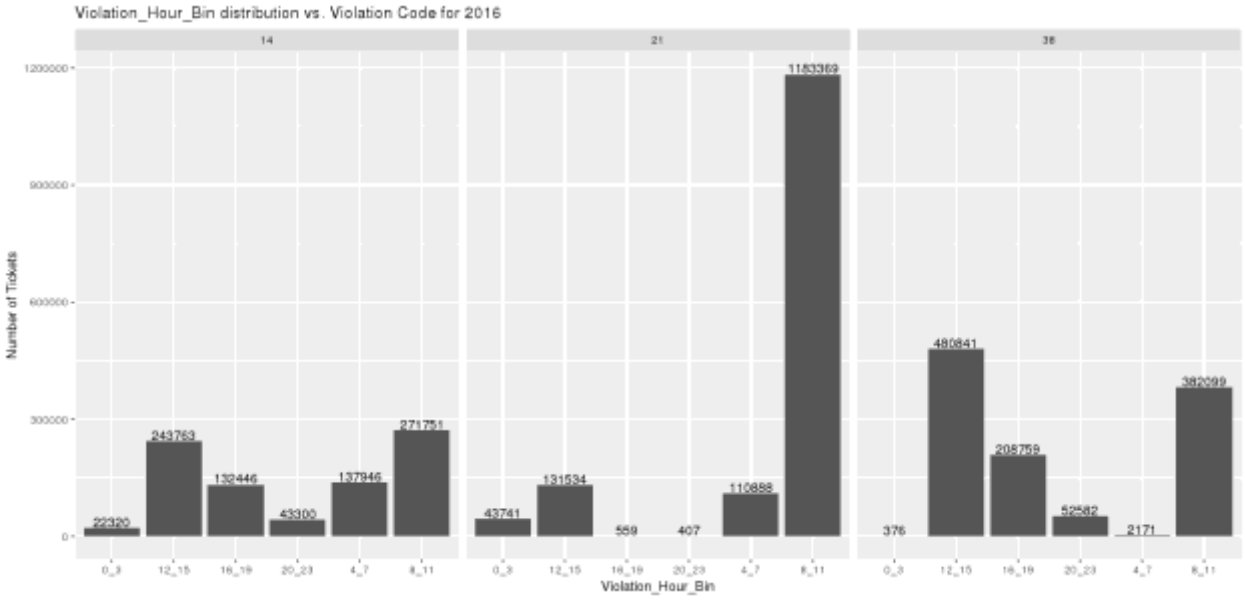


2016

Violation_Hour_Bin	Violation_Code	num_of_tkts
16_19	38	208759
16_19	37	159810
16_19	14	132446
8_11	21	1183369
8_11	36	578035
8_11	38	382099
4_7	14	137946
4_7	21	110888
4_7	40	89705
12_15	36	536551
12_15	38	480841
12_15	37	378361
0_3	21	43741
0_3	40	35350

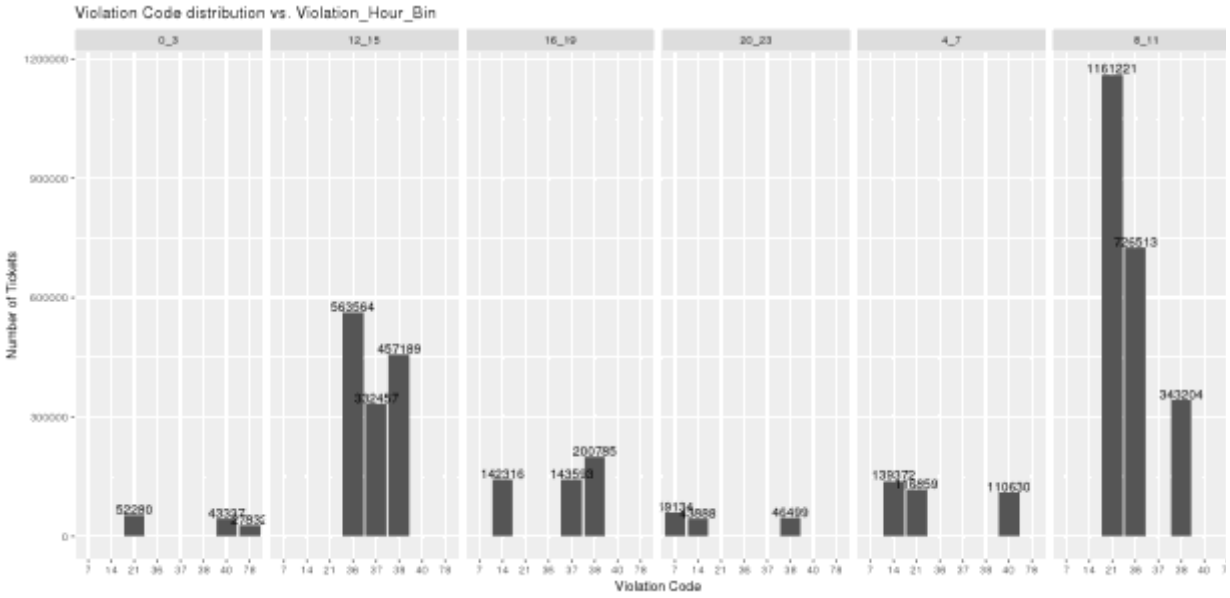
0_3	78	27296
20_23	7	56836
20_23	38	52582
20_23	40	43936

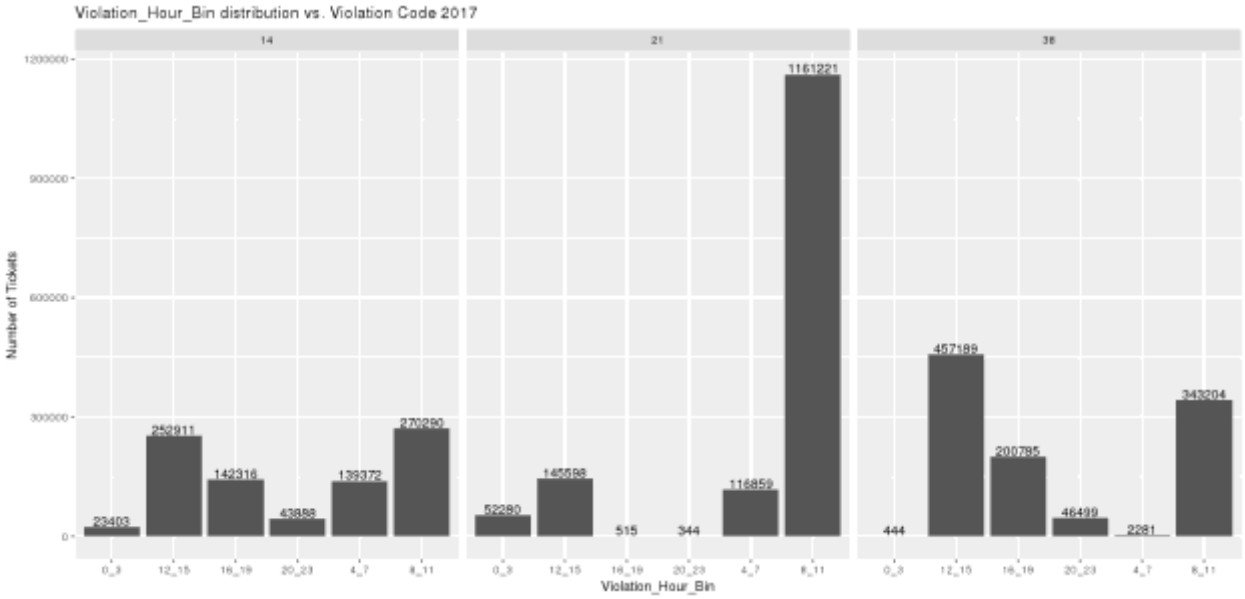




2017

Violation_Hour_Bin	Violation_Code	num_of_tkts
16_19	38	200785
16_19	37	143593
16_19	14	142316
8_11	21	1161221
8_11	36	726513
8_11	38	343204
4_7	14	139372
4_7	21	116859
4_7	40	110630
12_15	36	563564
12_15	38	457189
12_15	37	332457
0_3	21	52280
0_3	40	43337
0_3	78	27832
20_23	7	59134
20_23	38	46499
20_23	14	43888

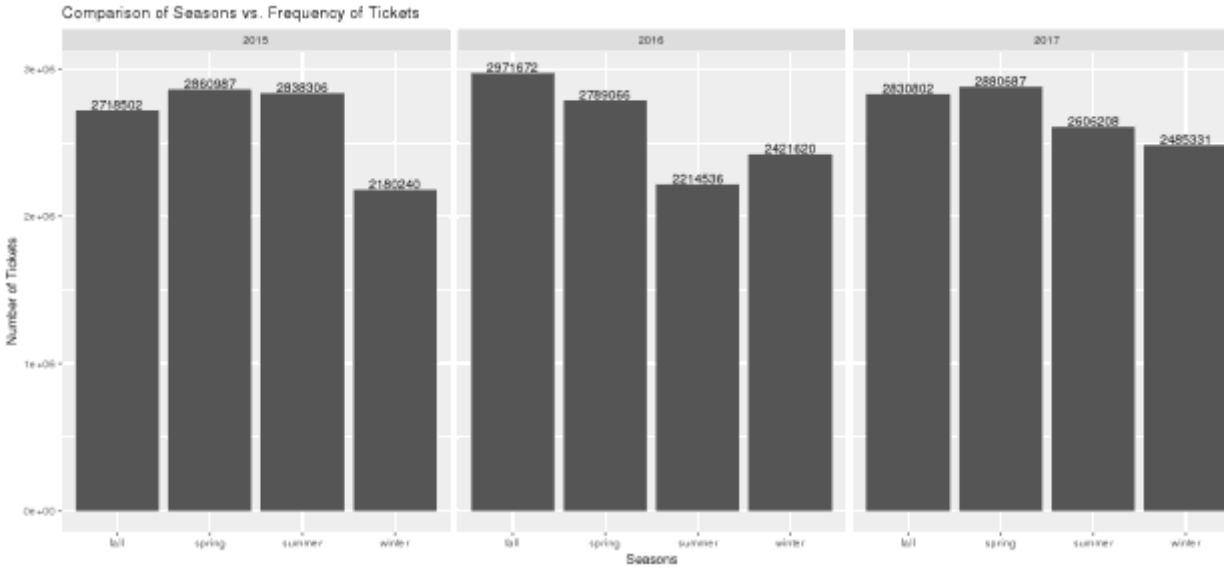


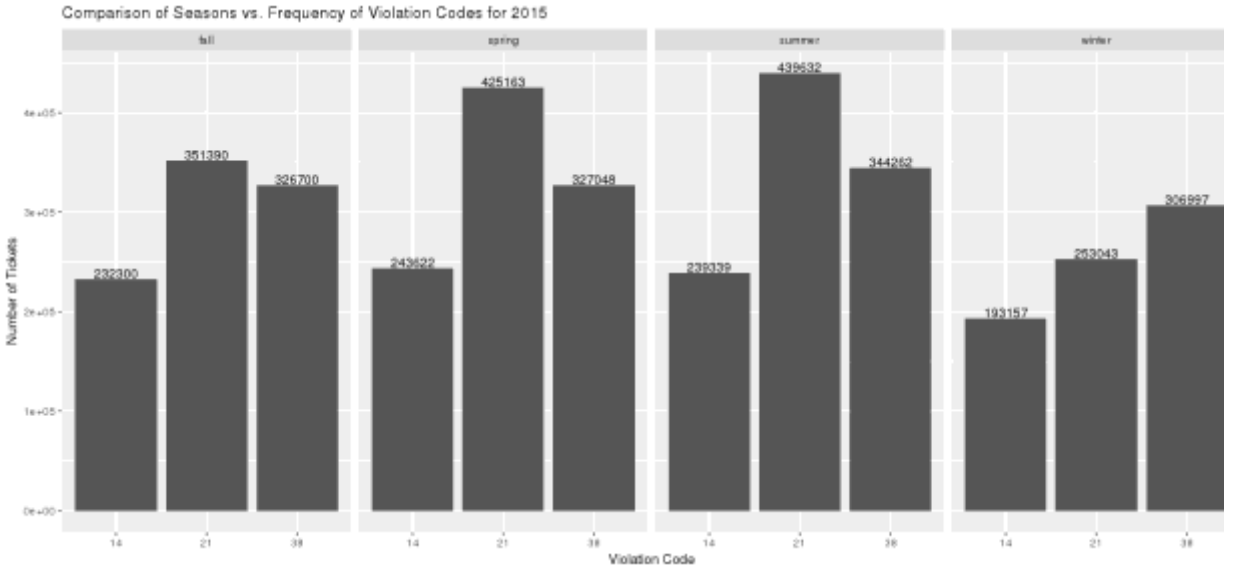


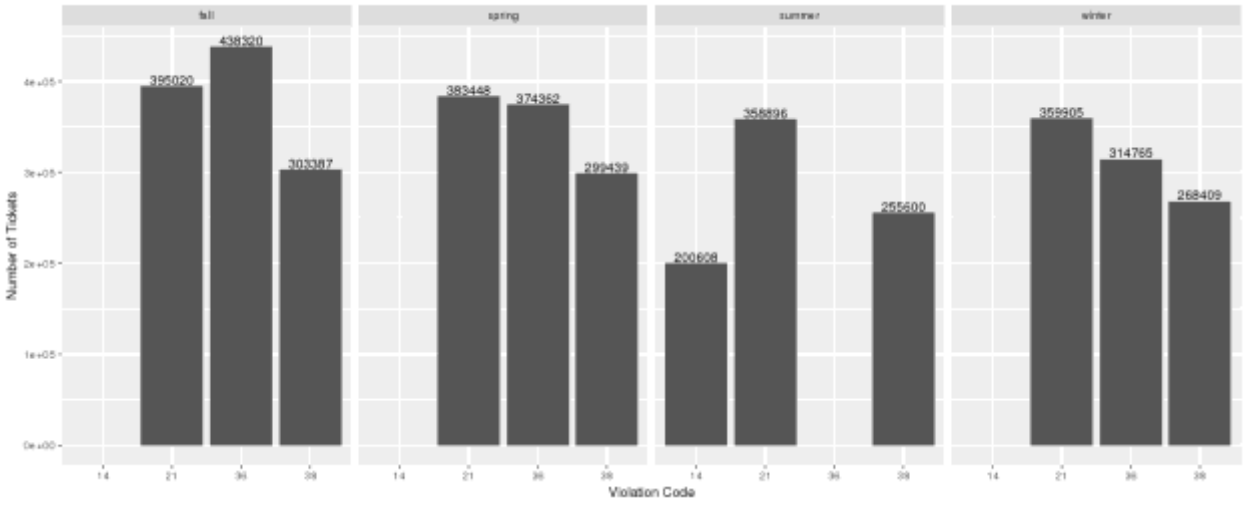
f. Let's try and find some seasonality in this data

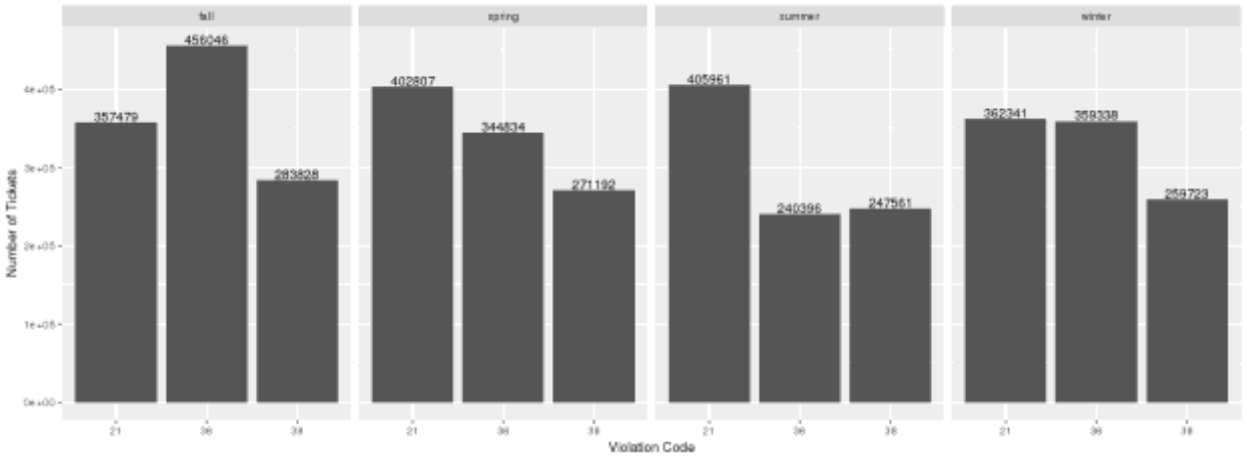
First, divide the year into some number of seasons, and find frequencies of tickets for each season.

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


Comparison of Seasons vs. Frequency of Tickets
2015
2016
2017
Number of Tickets
Seasons
fall
spring
summer
winter
2718502
2860987
2838306
2180240
2971672
2789055
2214536
2421620
2830802
2880587
2605208
2485331


Comparison of Seasons vs. Frequency of Violation Codes for 2015
fall
spring
summer
winter
Number of Tickets
Violation Code
232300
351390
326700
243622
425163
327048
239339
439532
344282
193157
253043
306997


Comparison of Seasons vs. Frequency of Violation Codes for 2016
fall
spring
summer
winter
Number of Tickets
Violation Code
395020
438320
303387
383448
374362
299439
200608
358896
255500
359905
314765
268409


Comparison of Seasons vs. Frequency of Violation Codes for 2017
fall
spring
summer
winter
Number of Tickets
Violation Code
357479
456046
283828
402807
344634
271192
405961
240396
247561
362341
359338
259723

g. 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?