Problem Statement :

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.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has **collected data for parking tickets**. For the scope of this analysis, we wish to analyse the parking tickets over the year 2017.

Location In HDFS :

The data for this case study has been placed in HDFS at the following path:**'/common_folder/nyc_parking/Parking_Violations_Issued_-_Fiscal_Year_2017.csv'**

Assumptions:

- As mentioned in the Up-grad portal, to consider the data for the 2017 Year alone, So, considered only 2017 Year (i.e. date range from 2017-01-01 to 2017-12-31).
- Removed all the rows, if any rows have the Null Values but Didn't find any such for the 2017 Year.
- Removing the Duplicate records, if the dataset contains as such for the Year 2017 but didn't find any as such for 2017 Year.
- Based on the Violation Time, Divided the Violation Time to Time Bins into 6 equal Bins by considering as 24 hrs clock as below :
    1. 0-3
    2. 4-7
    3. 8-11
    4. 12-15
    5. 16-19
    6. 20-23
- Based on the Month of the Issue date, Divided the Issue date to Time Bins into 4 equal as below :
    1. Mar-May → Spring
    2. Jun-Aug → Summer
    3. Sep-Nov → Autumn
    4. Dec-Feb → Winter

- I have taken the Average of the TWO fines (The fines are based on the two different categories) to calculate the total amount collected by NYC Police Department.

EDA :

- After filtering the entire dataset for 2017 alone, Didn’t Find any Null values in the entire 2017 Dataset.
- After filtering the entire dataset for 2017 alone, Didn’t Find any duplicates in the entire 2017 Dataset.
- Didn’t find any numeric entry with ‘99’ in the Registration State and hence not replaced with any values.
- There are total **10803028** records in the entire dataset.
- After filtering for 2017 Year alone, there are **5431909** records.
- Created the Temporary views for the further Analysis.

ANALYSIS :

1. Examine the data

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

Count
5431909

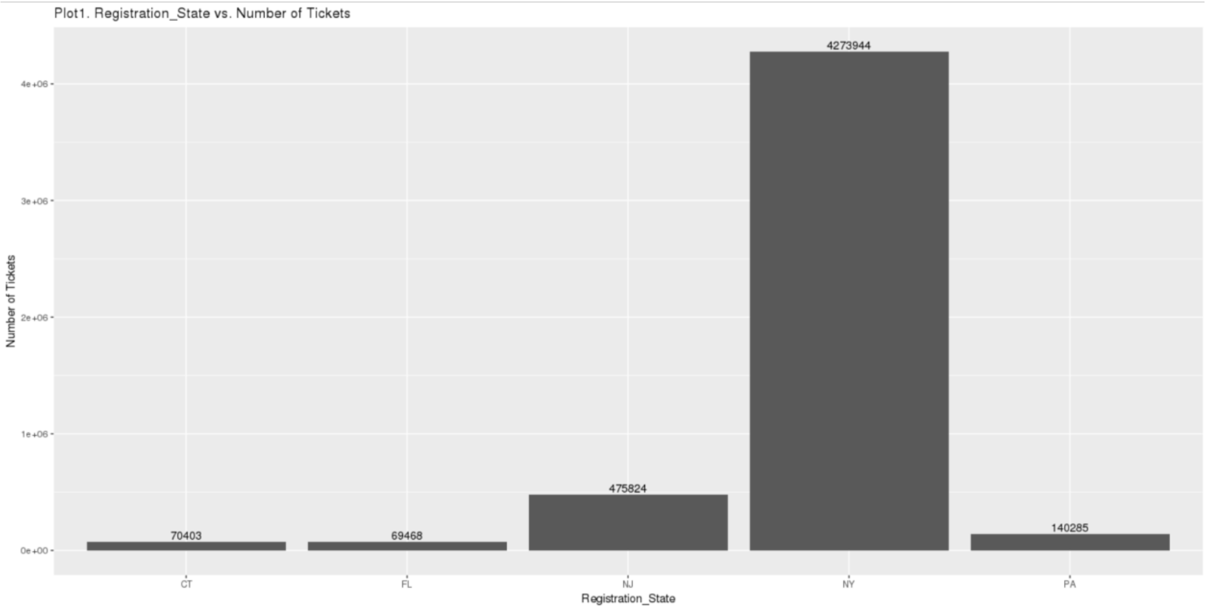
1.2. Find out the number of unique states from where the cars that got parking tickets came from.

unique_states_count
65

- Didn’t find any numeric entry with ‘99’ in the Registration State and hence not replaced with any values.

1.2.1. Find out the number of tickets from each unique states where the cars that got parking tickets came from(ONLY FOR TOP 5 STATES).

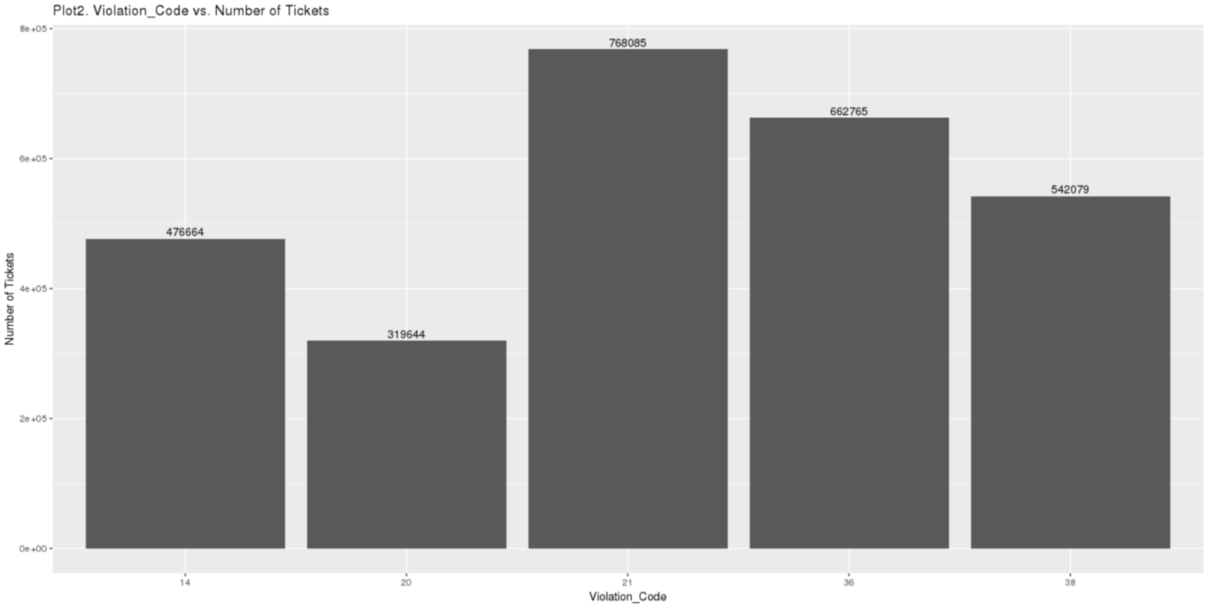
Registration_State	Ticket_count
NY	4273944
NJ	475824
PA	140285
CT	70403
FL	69468



2. Aggregation tasks

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

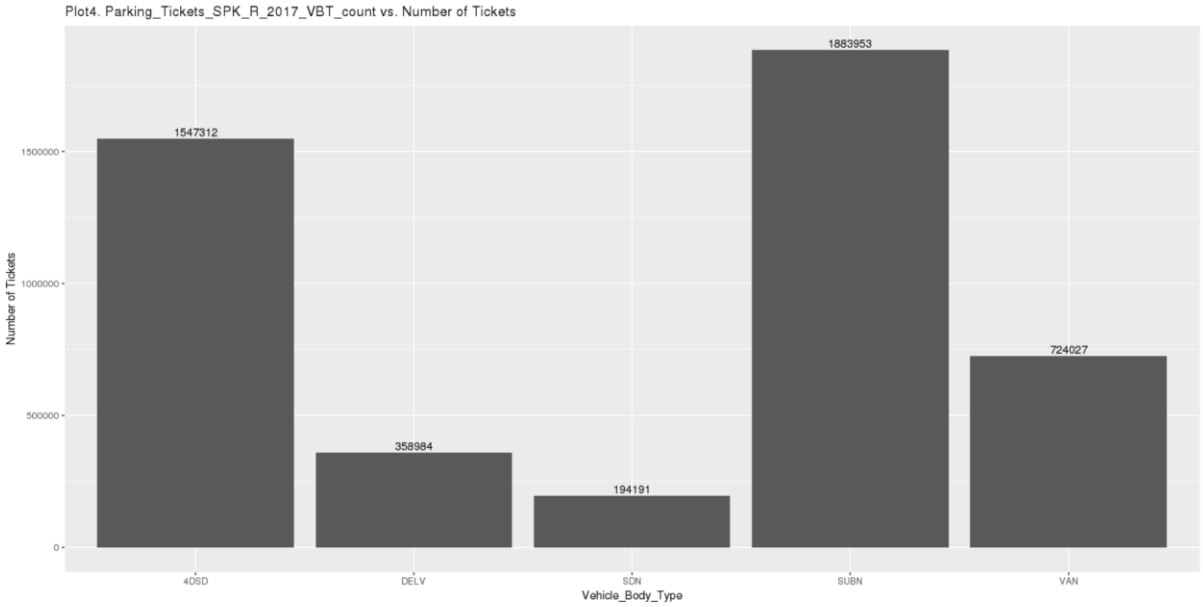
Violation_Code	count
21	768085
36	662765
38	542079
14	476664
20	319644



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

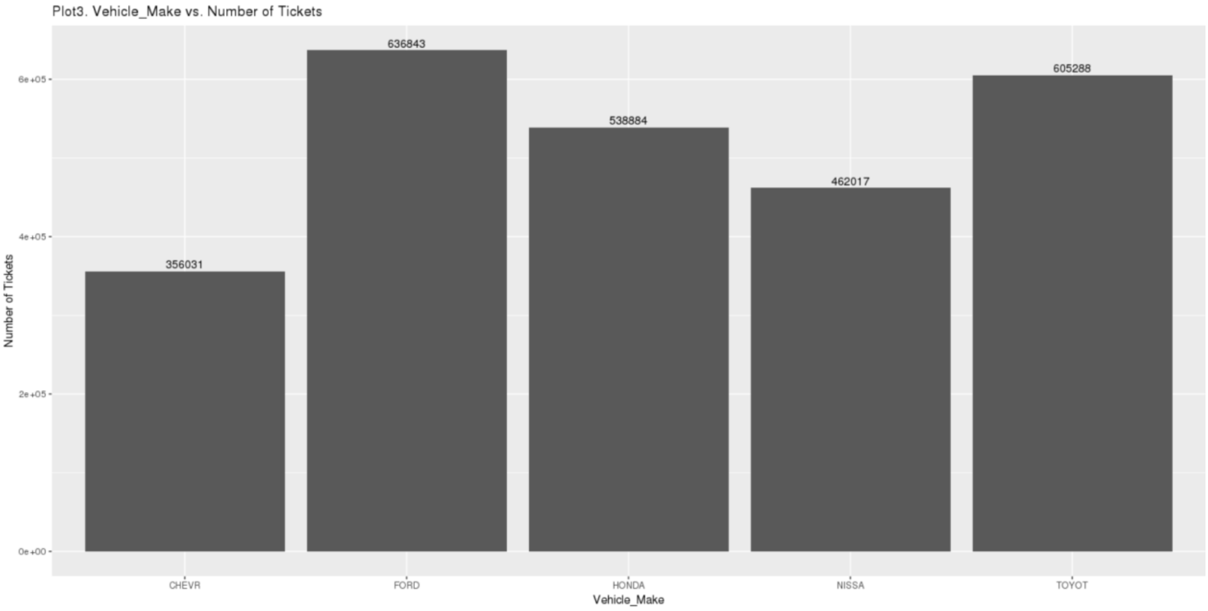
vehicle body type :

Vehicle_Body_Type	Number_of_Tickets
SUBN	1883953
4DSD	1547312
VAN	724027
DELV	358984
SDN	194191



vehicle make :

Parking_Ticket	Vehicle_Make
636843	FORD
605288	TOYOT
538884	HONDA
462017	NISSA
356031	CHEVR

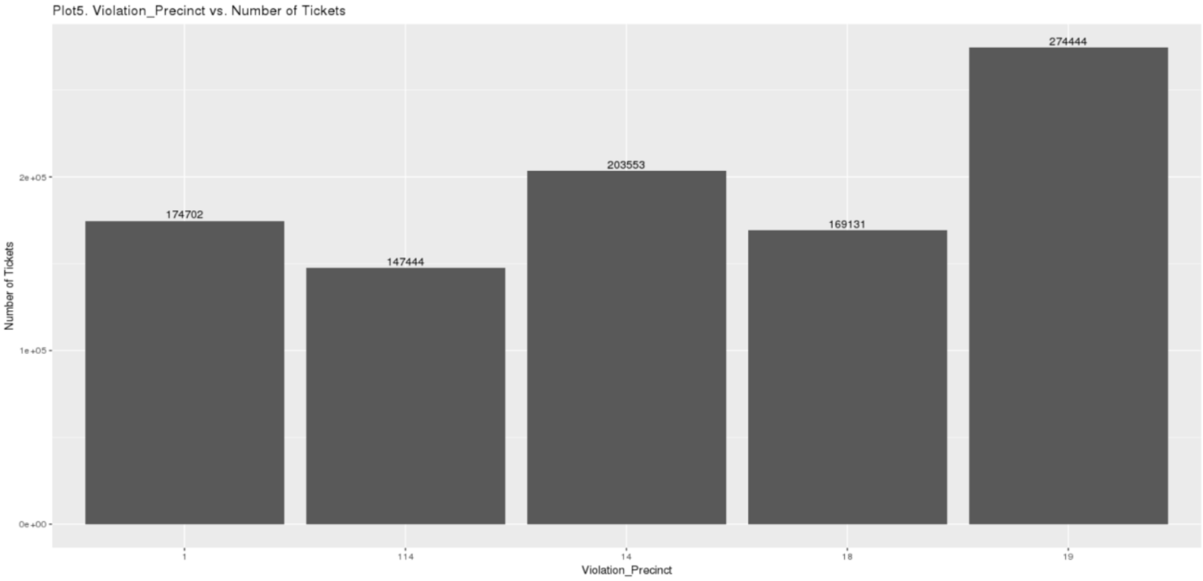


2.3. Top five Precinct Zone where violation occurred and where ticket was issued.

- While doing the analysis, we come across the Zone value as "0", which we considered as erroneous. We ignored that value and did the analysis.

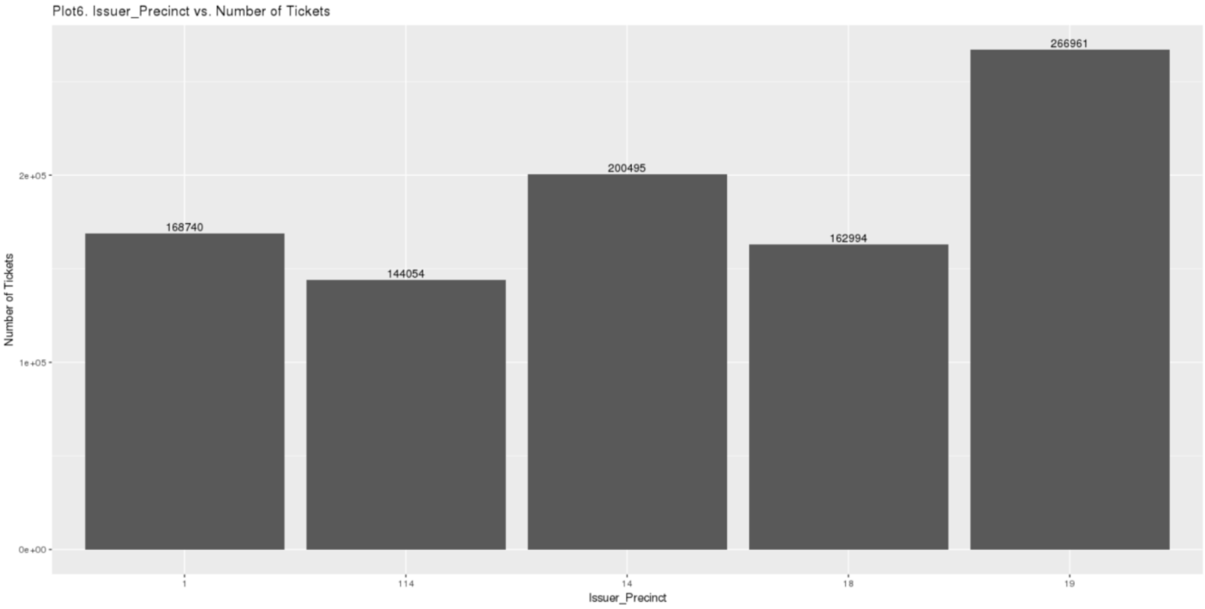
Violation Precinct :

Violation_Precinct	Number_of_Tickets
19	274444
14	203553
1	174702
18	169131
114	147444



Issuer Precinct :

Issuer_Precinct	Number_of_Tickets
19	266961
14	200495
1	168740
18	162994
114	144054



2.4. Violation Code frequency across three Precincts which have issued to the greatest number of tickets. Considering top 5 Violation Code across each Precincts

2.4.1. The violation code frequency across three precincts With Respect to top 3 Issuer_Precinct.

check_wrt_Issuer_Precinct	Violation_Code
113187	14
68869	46
48190	38
43782	37
39046	69
33499	21

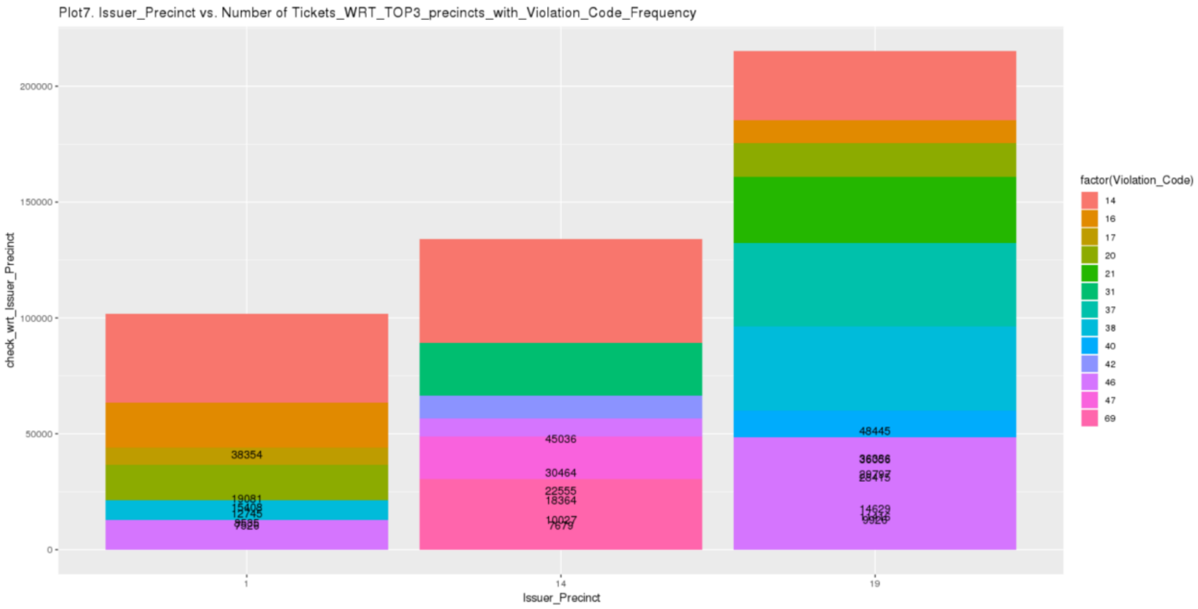
2.4.2. The violation code frequency across three precincts With Respect to top 3 Violation_Precinct.

check_wrt_Violation_Precinct	Violation_Code
116487	14
72730	46
49364	38
44219	37
39057	69
35472	21

2.4.3 Violation Code frequency across three Precincts which have issued to the greatest number of tickets. Considering top 5 Violation Code across each Precincts

check_wrt_Issuer_Precinct	Violation_Code	Issuer_Precinct
48445	46	19
36386	38	19
36056	37	19
29797	14	19
28415	21	19
45036	14	14
22555	31	14
18364	47	14
10027	42	14
7679	46	14
38354	14	1
19081	16	1
15408	20	1
12745	46	1
8535	38	1

- The Violation codes have high frequencies doesn’t have common across precincts. e.g.: The precinct 19 has large no. of tickets for violation code 46 but for precinct 14 and 1 has large no. of tickets for violation code 14.


Plot7. Issuer_Precinct vs. Number of Tickets_WRT_TOP3_precincts_with_Violation_Code_Frequency

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

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

- Didn’t Found any missing Values for the year 2017 and hence didn’t deal anything with that.

2.5.2. 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.

- Based on the Violation Time, Divided the Violation Time to Time Bins into 6 equal Bins by considering as 24 hrs clock as below :

  0-3

  4-7

  8-11

  12-15

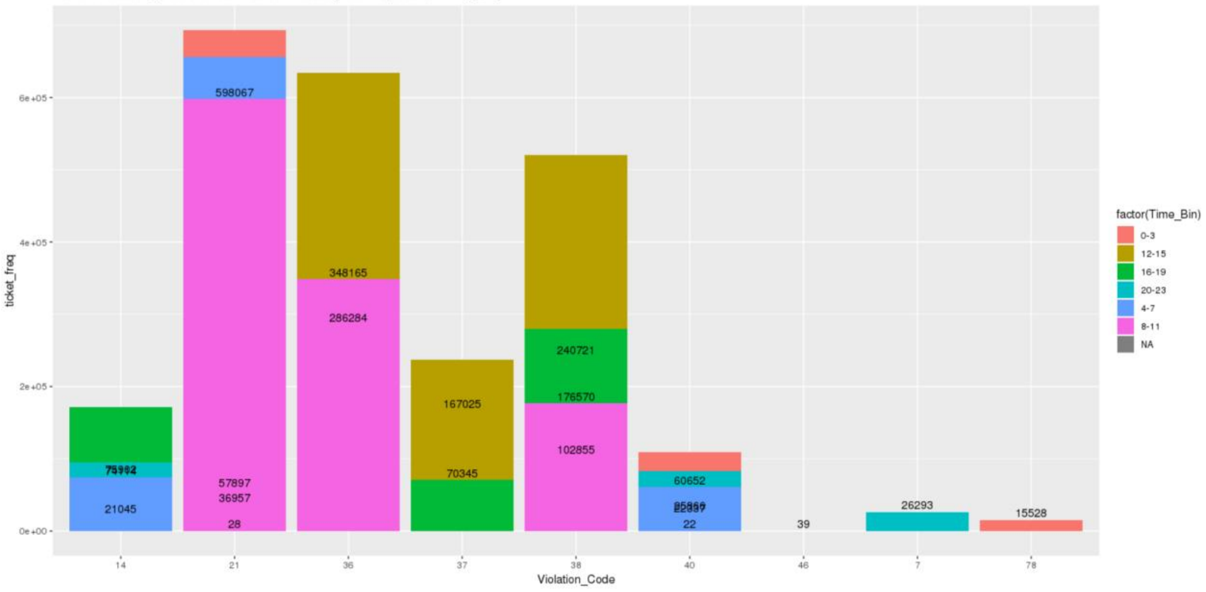
  16-19

  20-23

2.5.3. For each of these groups, find the three most commonly occurring violations.

Violation_Code	Time_Bin	ticket_freq
21	0-3	36957
40	0-3	25866
78	0-3	15528
40	4-7	60652
21	4-7	57897
14	4-7	74114
36	8-11	348165
38	8-11	176570
36	12-15	286284
38	12-15	240721
37	12-15	167025
38	16-19	102855
14	16-19	75902
37	16-19	70345
7	20-23	26293
40	20-23	22337
14	20-23	21045
21	8-11	598067

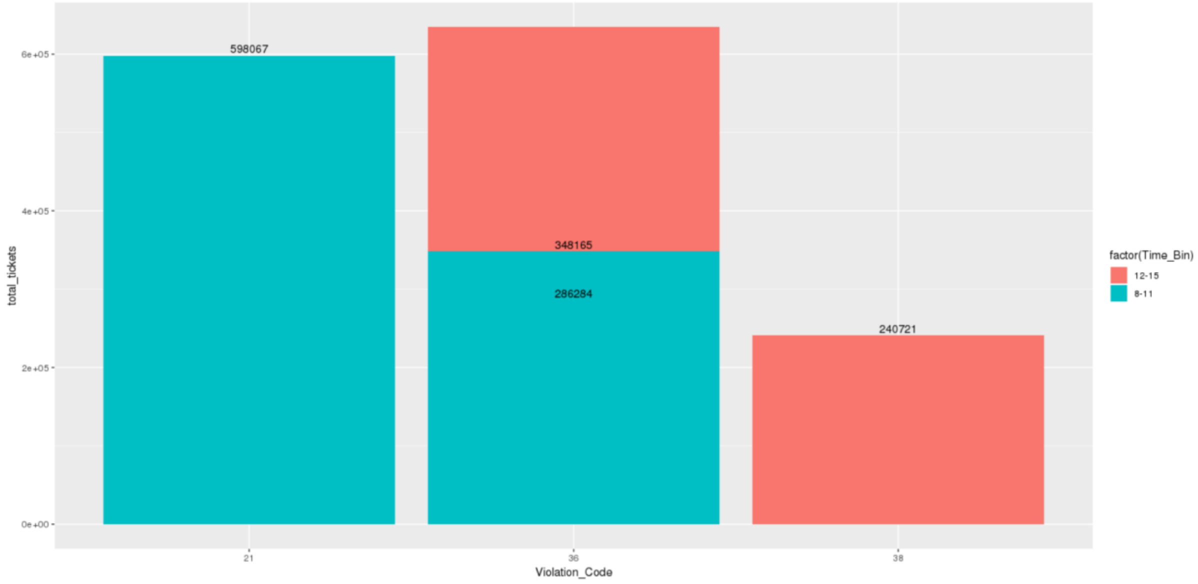
Plot8. Violation_Code vs. Number of Tickets [With respect to Time_Bin]



2.5.4. For the three most commonly occurring violation codes, find the most common time of the day (in terms of the bins)

Violation_Code	Time_Bin	total_tickets
21	8-11	598067
36	8-11	348165
36	12-15	286284
38	12-15	240721

Plot9. Time_Bin vs. Number of Tickets [With respect to top 3 Violation_Code]

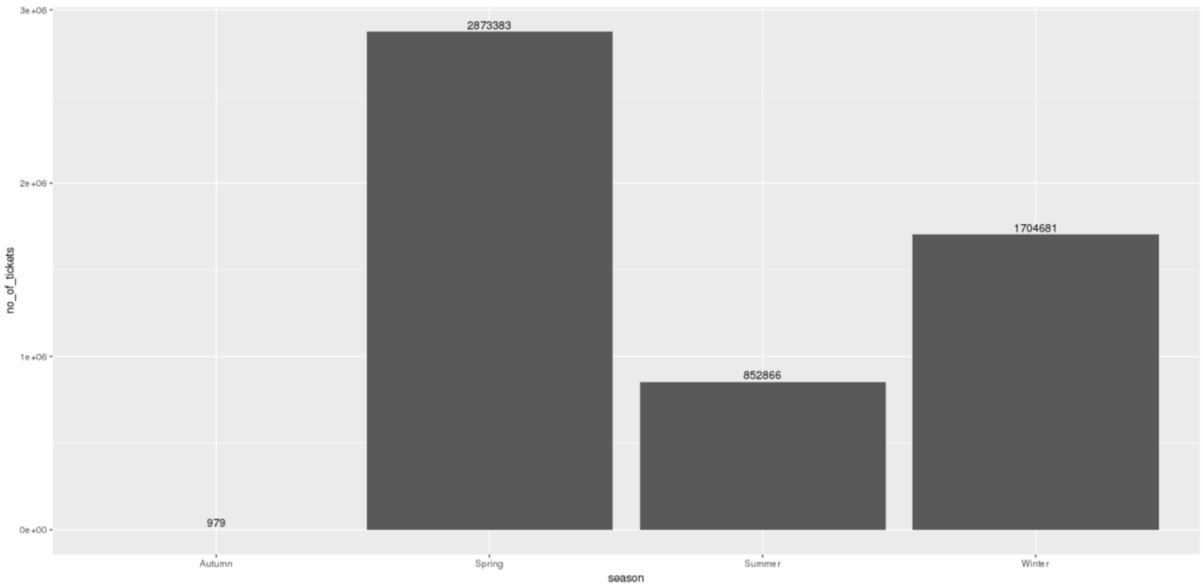


2.6. Find some seasonality in this data

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

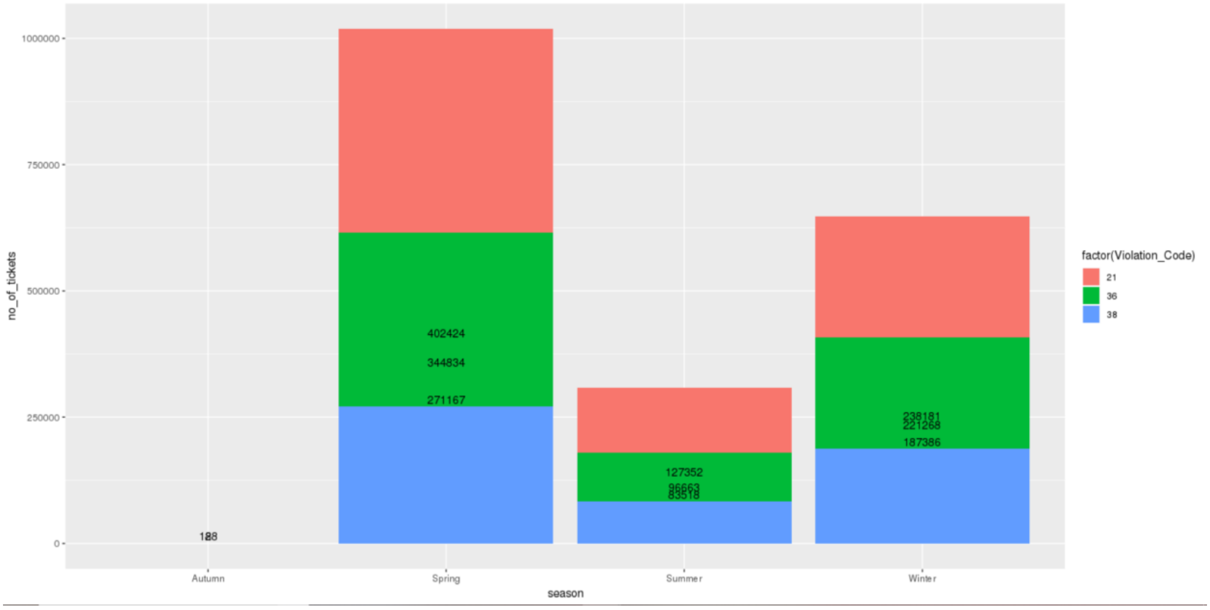
- Based on the Month of the Issue date, Divided the Issue date to Time Bins into 4 equal as below :
  1. Mar-May → Spring
  2. Jun-Aug → Summer
  3. Sep-Nov → Autumn
  4. Dec-Feb → Winter

season	no_of_tickets
Spring	2873383
Winter	1704681
Summer	852866
Autumn	979


Plot10.Seasons vs. Number of Tickets

2.6.2. Find the three most common violations for each of these seasons

Violation_Code	season	count
21	Spring	402424
36	Spring	344834
38	Spring	271167
21	Winter	238181
36	Winter	221268
38	Winter	187386
21	Summer	127352
36	Summer	96663
38	Summer	83518
21	Autumn	128
38	Autumn	8


Plot11. Seasons vs. Number of Tickets[With respect to most common Violation codes]

2.7. Estimating that for the three most commonly occurring codes.

2.7.1. Find total occurrences of the three most common violation codes.

Violation_Code	ticket_freq_wrt_VC
21	768085
36	662765
38	542079

2.7.2. Find the Average fines for the three most common violation codes.

Violation_Code	Avg_fines
21	55
38	50
36	50

2.7.3. Find the total amount collected for the three violation codes with maximum tickets.

Violation_Code	ticket_freq_wrt_VC	Avg_fines	total_amt_collected
21	768085	55	42244675
36	662765	50	33138250
38	542079	50	27103950

2.7.4. What can you intuitively infer from these findings?

- Violation Code “**21**” (Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.) has the maximum collection of amount $ 42244675 for the 2017 Year alone.

- NYC Police Department are getting the major revenue from the unparking Areas where the People are parking and tickets are getting issued.

- Violation Code “**36**”(Exceeding the posted speed limit in or near a designated school zone.) has the next major collection of amount $ 33138250 for the 2017 Year alone followed by the Violation Code “**21**”.

- Next Major revenue getting to NYC Police Department is by Exceeding the Speed limits followed by the Parking Violation.