NYC Parking Ticket - Analysis

Submitted By [Avishek Sengupta, Harkirat Dhillon, Shubhra Karmahe, Taranveer Singh Arora]

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 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**. Out of these, the data files from 2014 to 2017 are publicly available on Kaggle. 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.

For the scope of this analysis, we will compare phenomenon related to parking tickets over three different years - **2015, 2016, 2017**. All the analysis steps mentioned below is done for 3 different years with each metric derived compared across the 3 years.
The purpose of this case study is to conduct an **exploratory analysis** that helps us understand the data using RStudio and Apache SparkR.

Data Understanding & Assumptions :

Data Source url - https://www.kaggle.com/new-york-city/nyc-parking-tickets/data

1. We have taken 03 datasets for analysis.They are listed as below:-

   1.1 Parking_Violations_Issued_-_Fiscal_Year_2015.csv
   1.2 Parking_Violations_Issued_-_Fiscal_Year_2016.csv
   1.3 Parking_Violations_Issued_-_Fiscal_Year_2017.csv

2. There are 10M+ rows and 43+ columns within each dataset.

3. There are spaces and special characters in the column names in the dataset. We have removed leading/trailing and in-between spaces and special characters from column names for consistency across the datasets.

4. We have ignored and dropped columns having mostly NA values from dataset for consistency across the three datasets.

4.1 Fiscal Year 2015 - No_Standing_or_Stopping_Violation,Latitude,Longitude,
Hydrant_Violation,Double_Parking_Violation,Community_Board,
Community_Council,Census_Tract,BBL,BIN,NTA

4.2 Fiscal Year 2016 - No_Standing_or_Stopping_Violation,Latitude,Longitude,
Hydrant_Violation,Double_Parking_Violation,Community_Board,
Community_Council,Census_Tract,BBL,BIN,NTA

4.3 Fiscal Year 2017 - No_Standing_or_Stopping_Violation,
Hydrant_Violation,Double_Parking_Violation

5. We have considered **Summons_Number** and **Issue_Date** as key columns for Analysis.

6. On the basis of unique **Summons_Number**, we have dropped duplicate observations for further analysis.

7. We observe the Datasets have records (Issue_Date) with dates spanning from year 1975 to 2069. The individual Issue_Date ranges(Year-Month-Day) are as below:
Fiscal Year 2015 : 1985-07-16 to 2015-06-30
Fiscal Year 2015 :   1970-04-13 to 2069-10-02
Fiscal Year 2015 :   1972-03-30 to 2069-11-19

8. We have considered the fiscal year for the state of New York from April to March.
Reference url - https://en.wikipedia.org/wiki/Fiscal_year

9. Date range used to subset data from datasets are as:

8.1 Fiscal Year 2015 : 01/Apr/2014 to 31/Mar/2015
8.2 Fiscal Year 2015 : 01/Apr/2015 to 31/Mar/2016
8.3 Fiscal Year 2015 : 01/Apr/2016 to 31/Mar/2017

10. We have converted the date format for the (Issue_Date) to 'MM/dd/yyyy' format.

11. We have considered only House_Number and Street_Number as address while counting tickets with missing address. Intersecting_Street and County_Name alone cannot identify the correct address even though they are present on the ticket hence not considered them for calculating the tickets with missing address.

12. We are using columns Violation_Precinct and Issuer_Precinct to analyze the frequency of the violations. We are not considering records for Violation Precinct and Issuer Precinct which have value 0. As per the precincts shared on this url:
https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page

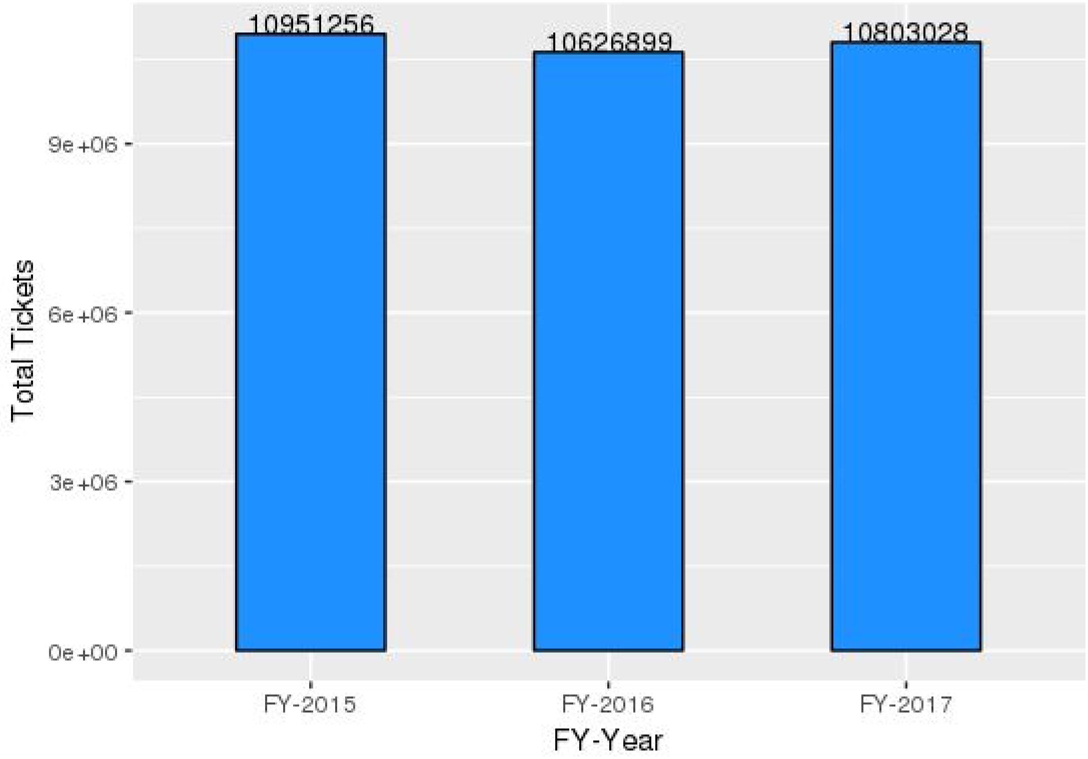
13. Violation_Time missing values are very few and hence we have ignored them to analyze frequency of violations.For our analysis under Aggregation we have Extracted Violation Hour,Minute and Part of Day from Violation_Time.

14. The total amount collected for all the parking tickets for different violations was calculated using the average of the two fines based on the precincts location as mentioned on this url :
https://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page

Examine the data:

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

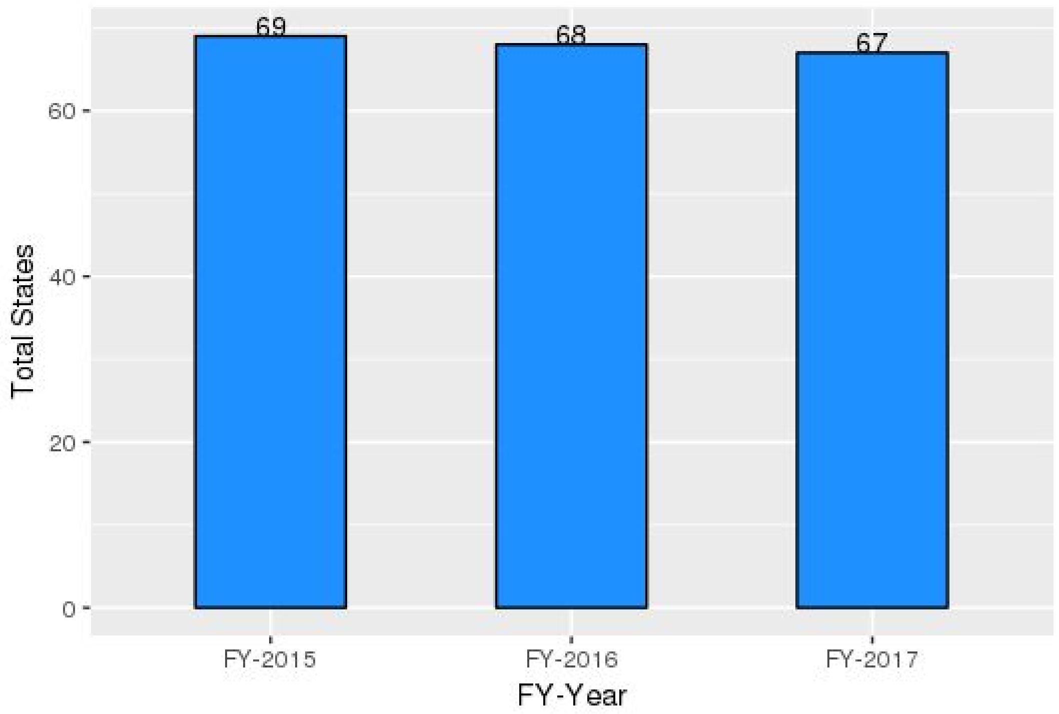
FY 2015	FY 2016	FY 2017
10951256	10626899	10803028



We observe that the total number of tickets issued for all three years 2015,2016 and 2017 range between 10.62 - 10.80 Millions($) in revenue.

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

FY 2015	FY 2016	FY 2017
69	68	67

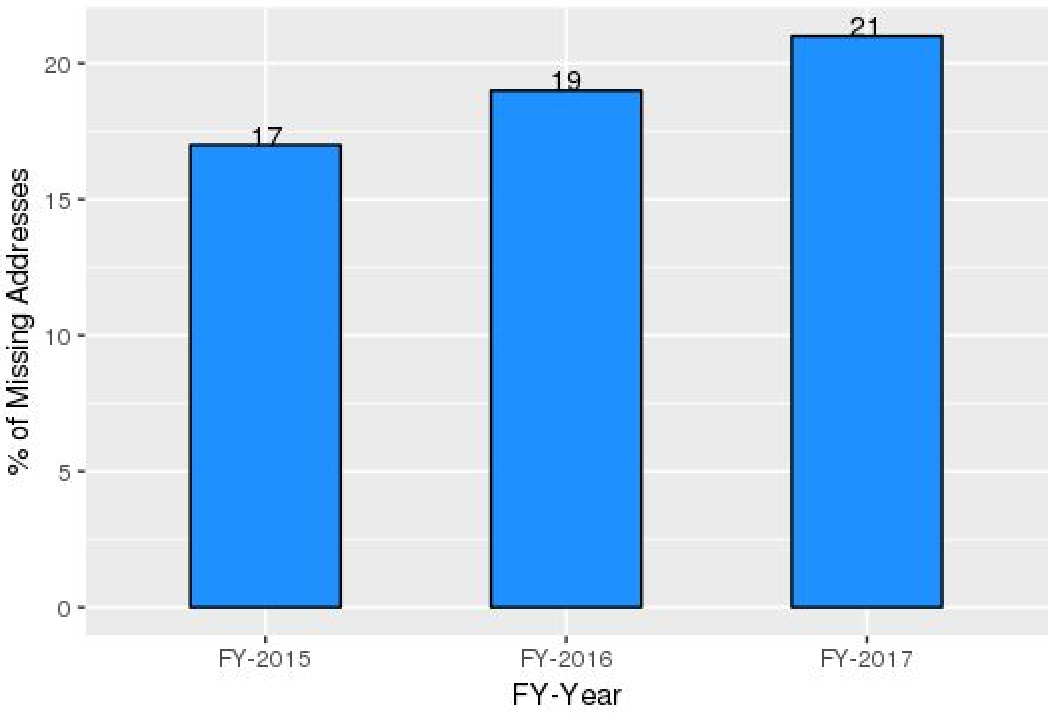


The plot indicates that the registration number of cars belonging to unique states which have been issued parking tickets have been decreasing by one from 2015 to 2017.

3. Some parking tickets don’t have addresses on them, which is cause for

concern. Find out how many such tickets there are.

FY 2015	FY 2016	FY 2017
16.5%	19%	21%

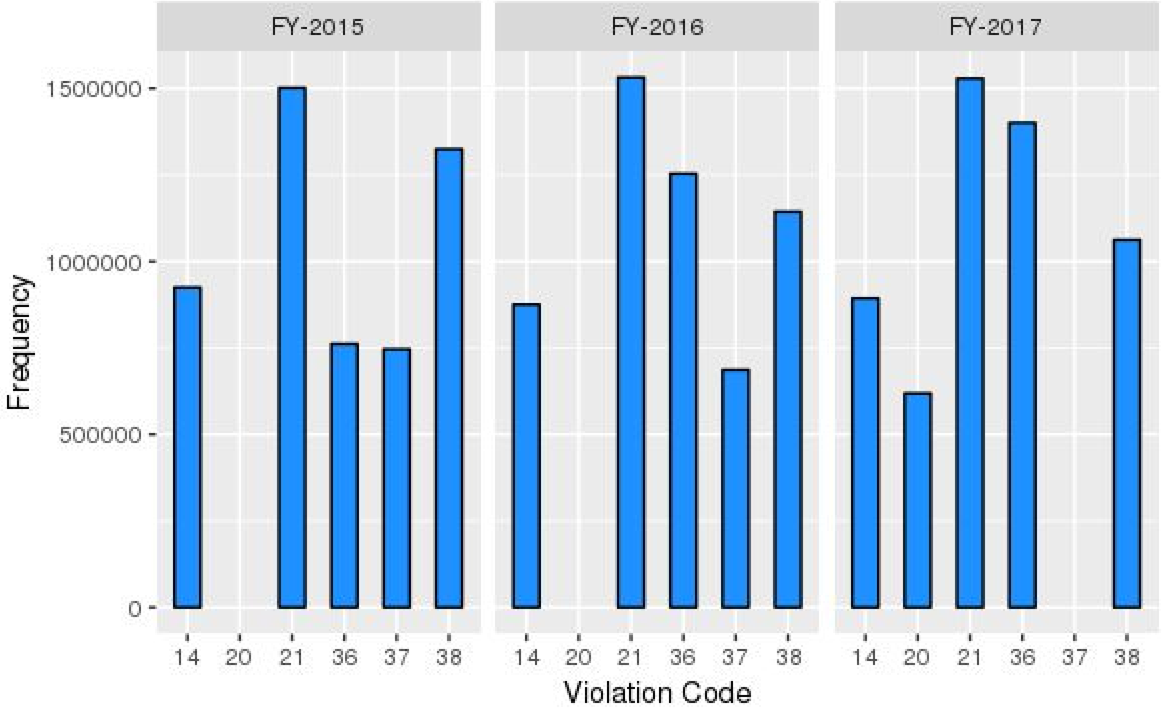


The plot clearly shows an increasing trend in the ticket numbers where the address is missing for 2015 - 2017, with the percentage rising from 16.5% in 2015 to 21% in 2017. This is cause for concern and needs to be looked into to avoid prospective revenue loss.

Aggregation Task:

1. **How often does each violation code occur? (frequency of violation codes - find the top 5)**

FY 2015	FY 2016	FY 2017
Code # 21 - 1501614	Code # 21 - 1531587	Code # 21 - 1528588
Code # 38 - 1324586	Code # 36 - 1253512	Code # 36 - 1400614
Code # 14 - 924627	Code # 38 - 1143696	Code # 38 - 1062304
Code # 36 - 761571	Code # 14 - 875614	Code # 14 - 893498
Code # 37 - 746278	Code # 37 - 686610	Code # 20 - 618593



Overall, violation code 21 is the one occurring the most among the three years. Also, we notice violation code 37 is no longer the fifth most common violation code in 2017 unlike 2015 and 2016.

As per https://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page

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

Violation Code 14 : General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.

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

Violation Code 37-38 :Muni Meter --

(37) Parking in excess of the allowed time

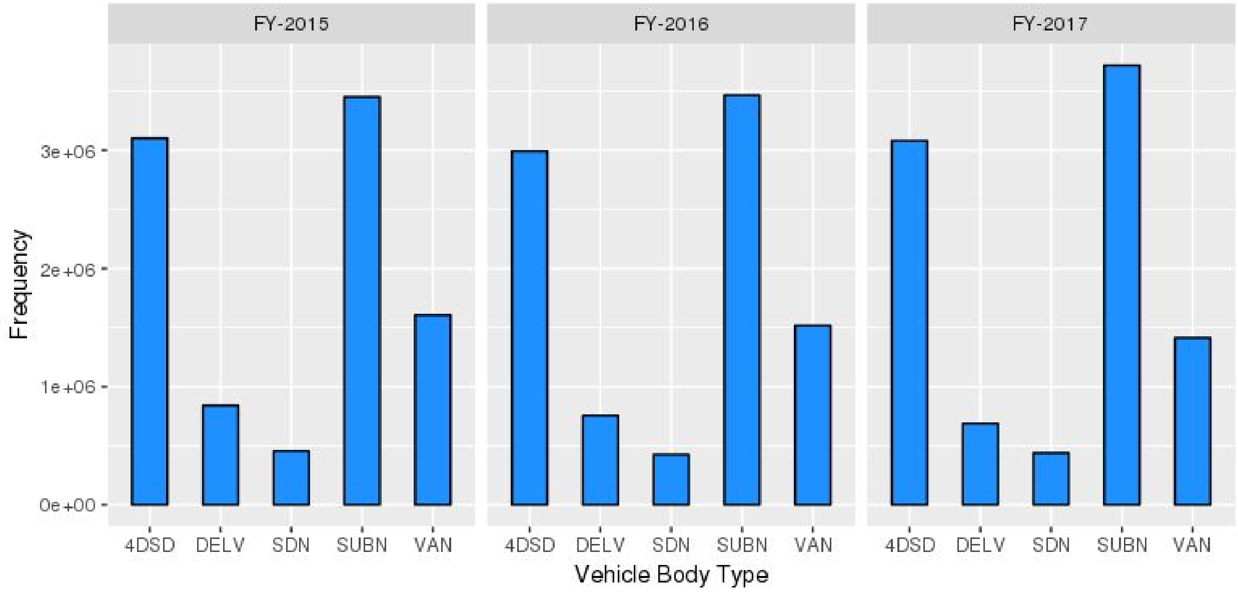
(38) Failing to show a receipt or tag in the windshield.

Drivers get a 5-minute grace period past the expired time on Muni-Meter receipts.

This suggests most of the parking tickets are issued for parking in places with No parking allowed, exceeding the posted speed limit near school zones and expired time on Muni Meter receipts.

2.1 How often does each vehicle body type get a parking ticket?(find the top 5)

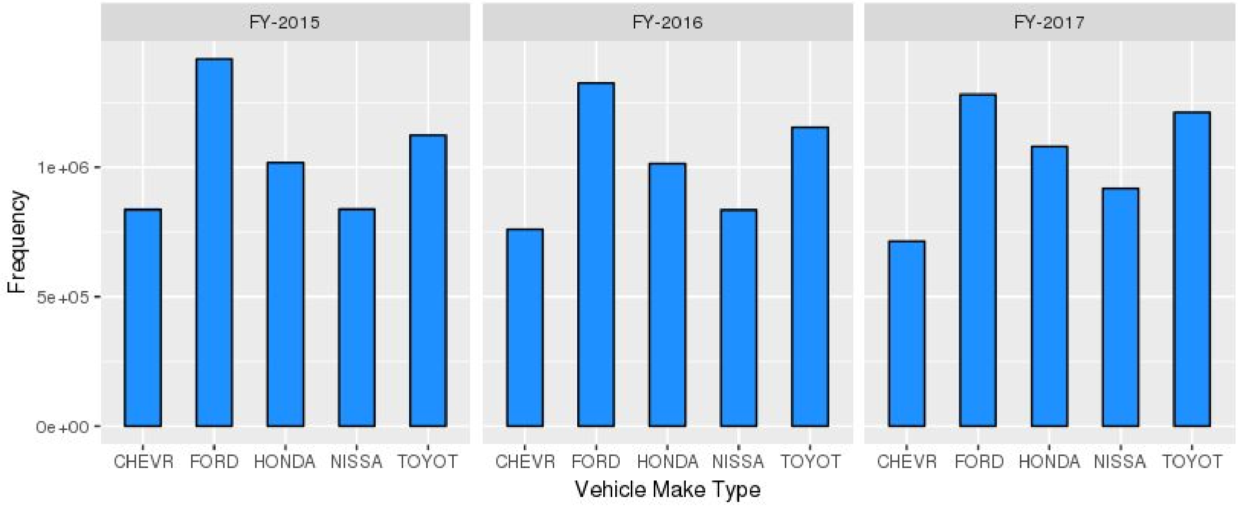
FY 2015	FY 2016	FY 2017
SUBN - 3451963	SUBN - 3466037	SUBN - 3719802
4DSD - 3102510	4DSD - 2992107	4DSD - 3082020
VAN - 1605228	VAN - 1518303	VAN - 1411970
DELV - 840441	DELV - 755282	DELV - 687330
SDN - 453992	SDN - 424043	SDN - 438191



The five vehicle types with parking tickets remain same for the three years as per the following order SUBN,4DSD,VAN,DELV,SDN.
SUBN vehicle types have the highest number of parking tickets. The law defines a suburban as a vehicle that can be used to carry passengers and cargo.

2.2 How about the vehicle make? (find the top 5)

FY 2015	FY 2016	FY 2017
FORD - 1417303	FORD - 1324774	FORD - 1280958
TOYOT - 1123523	TOYOT - 1154790	TOYOT - 1211451
HONDA - 1018049	HONDA - 1014074	HONDA - 1079238
NISSA - 837569	NISSA - 834833	NISSA - 918590
CHEVR - 836389	CHEVR - 759663	CHEVR - 714655



The five vehicle makes with parking tickets remain same for the three years as per the following order FORD,TOYOT,HONDA, NISSA, CHEVR.
FORD vehicles are issued the highest number of parking tickets for all the three years 2015 - 2017.

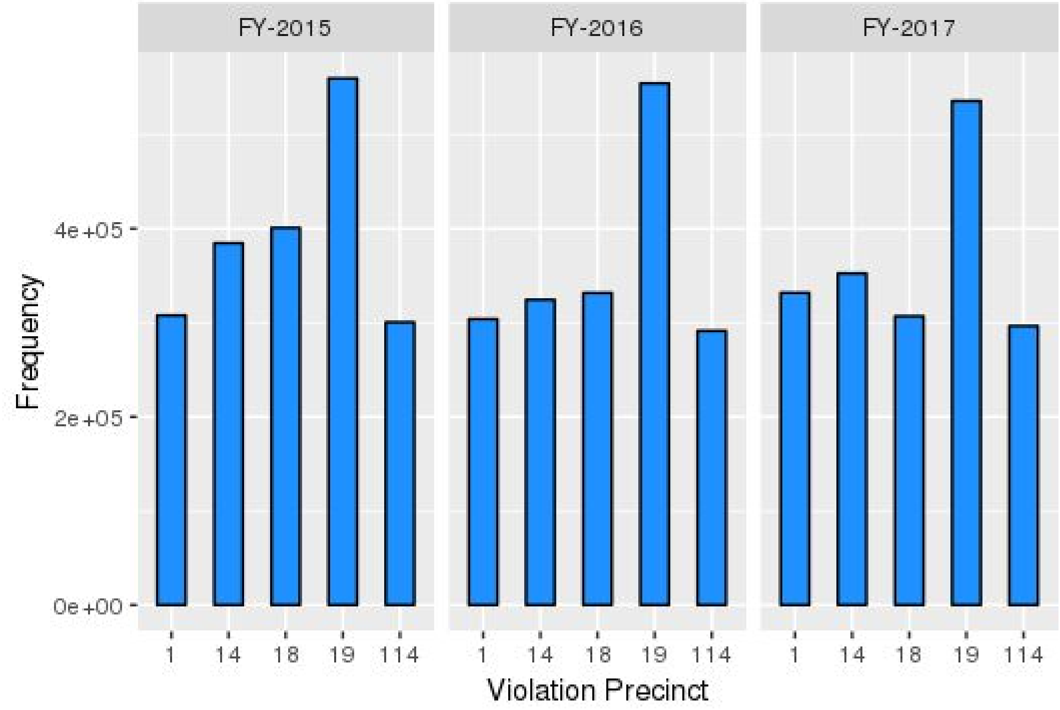
3. A precinct is a police station that has a certain zone of the city under its command.

Find the (5 highest) frequencies of:

3.1 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?

FY 2015	FY 2016	FY 2017
19 - 559716	19 - 554465	19 - 535671
14 - 400887	14 - 331704	14 - 352450

18 - 384596	18 - 324467	1 - 331810
1 - 307808	1 - 303850	18 - 306920
114 - 300557	114 - 291336	114 - 296514



We observe from the plot that the maximum violations occur in the 19th precinct for all three years 2015-2017. The other precincts 1st,14th,18th have interchangeable positions across the three years and 114th is the fifth highest violation precinct.

The 19,18,1 and 14 precincts fall in Manhattan and 114 in Queens borough of NYC. The 19th Precinct command serves the Upper East Side of Manhattan,one of the most densely populated residential areas in Manhattan.The southern part of the precinct has a large commercial area and features Madison, Lexington, and 3rd Avenues, which are well known for their shopping.

18th Precinct is now Midtown North,which serves the area of Midtown, Manhattan, just south of Central Park.The precinct encompasses the Diamond District, St. Patrick's Cathedral, the Theatre District, Restaurant Row, Radio City Music Hall, and Rockefeller Plaza.

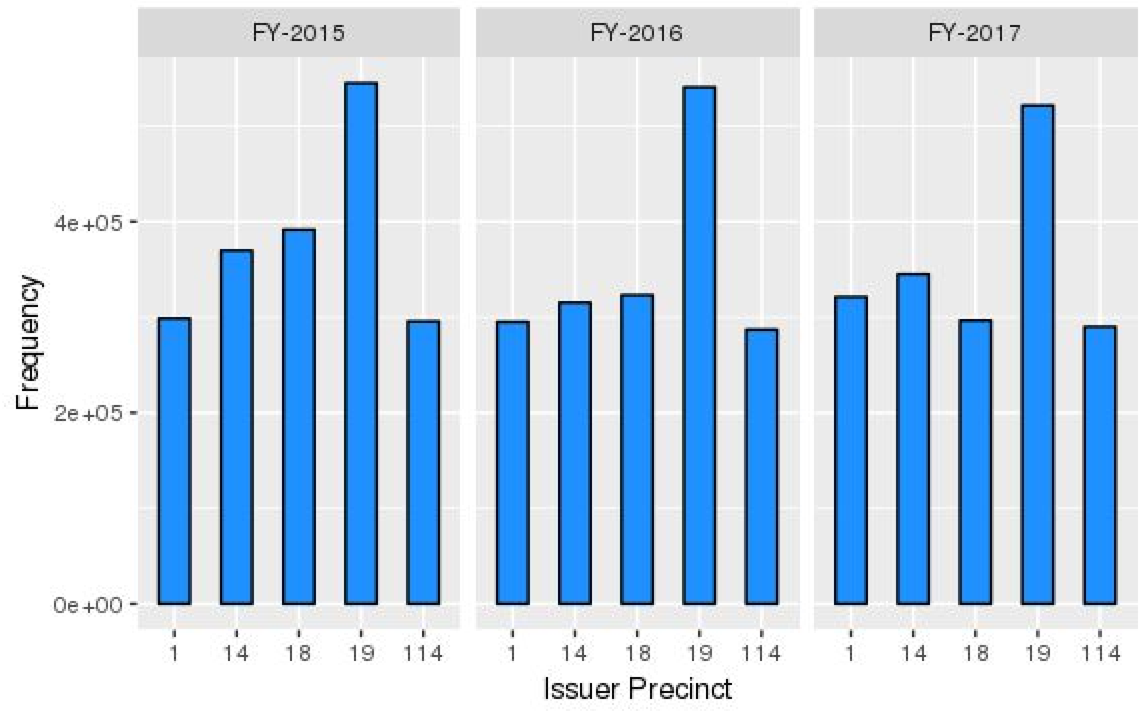
The 1st Precinct serves an area that consists of a square mile on the southernmost tip of Manhattan. The precinct is home to the World Trade Center, SOHO, Tribeca, and Wall Street which is home to office spaces and very popular with tourists.

The 14th Precinct is now Midtown South, which serves the southern portion of Midtown, Manhattan.The area contains commercial offices, hotels, Times Square, Grand Central Terminal, Penn Station,Madison Square Garden, Koreatown section, and the Manhattan Mall Plaza.

The 114th Precinct is located in the northwestern portion of Queens, and covers Astoria, Long Island City, Woodside, and Jackson Heights.

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

FY 2015	FY 2016	FY 2017
19 - 544946	19 - 540569	19 - 521513
14 - 391501	14 - 323132	14 - 344977
18 - 369725	18 - 315311	1 - 321170
1 - 298594	1 - 295013	18 - 296553
114 - 295601	114 - 286924	114 - 289950

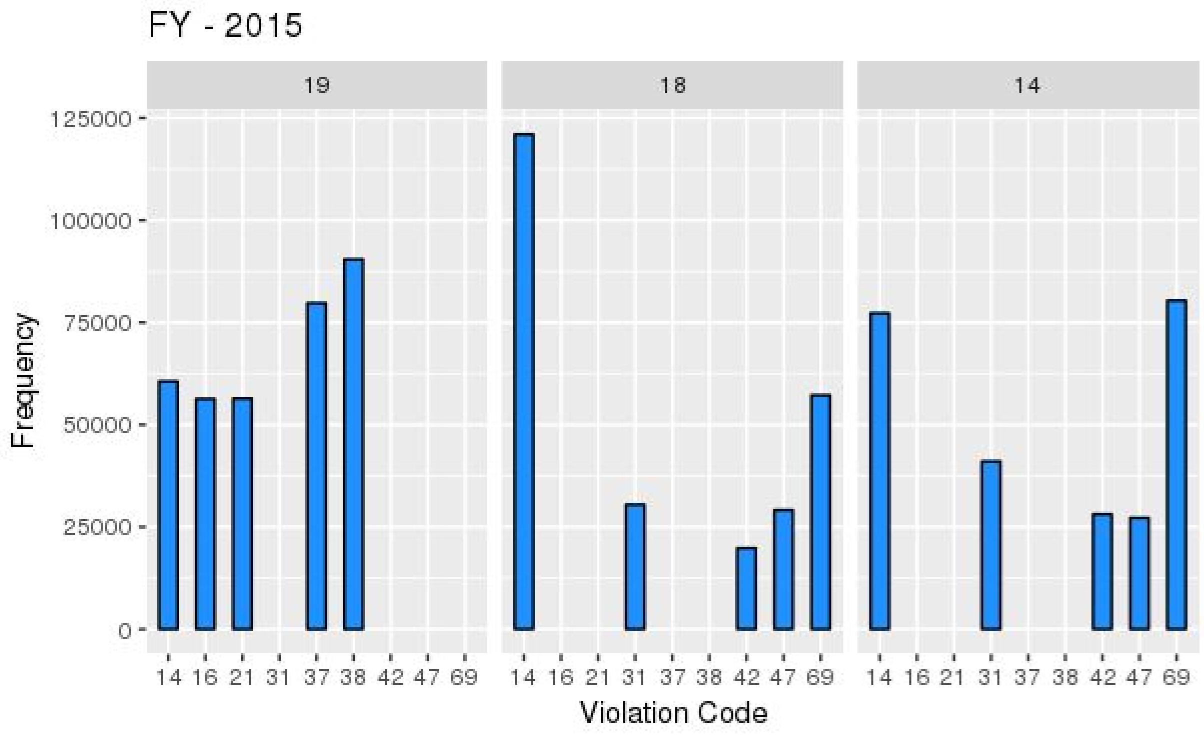


We observe from the plot that the maximum violations occur in the 19th precinct for all three years 2015-2017. The other precincts 1st,14th,18th have interchangeable positions across the three years and 114th is the fifth highest violation precinct.

4. 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?

From previous analysis (3.) we know, the top 3 issuer precincts which issued the most number of tickets are Precincts 19,18 and 14 in that order for 2015.

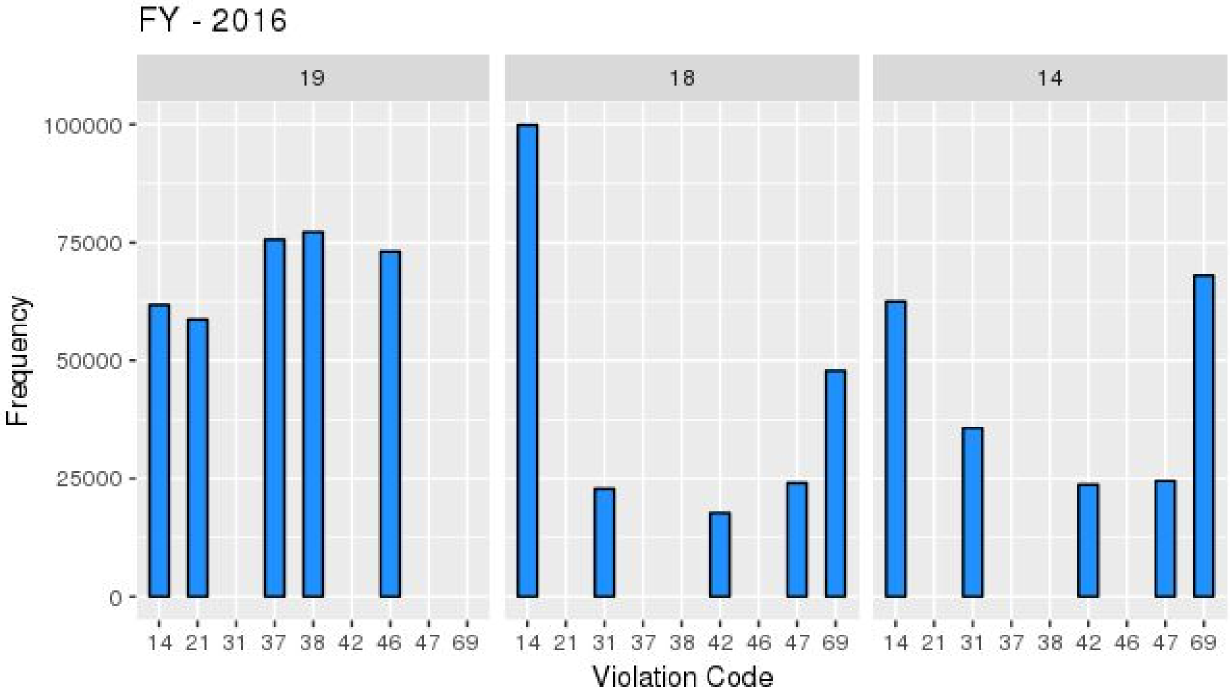
FY - 2015		
Issuer Precinct # 19	**Issuer Precinct # 18**	**Issuer Precinct # 14**
Violation Code # 38 - 90437	Violation Code # 14 - 121004	Violation Code # 69 - 80368
Violation Code # 37 - 79738	Violation Code # 69 - 57218	Violation Code # 14 - 77269
Violation Code # 14 - 60589	Violation Code # 31 - 30447	Violation Code # 31 - 41049
Violation Code # 21 - 56416	Violation Code # 47 - 29124	Violation Code # 42 - 28114
Violation Code # 16 - 56318	Violation Code # 42 - 19820	Violation Code # 47 - 27229



From previous analysis (3.) we know, the top 3 issuer precincts which issued the most number of tickets are Precincts 19,18 and 14 in that order for 2016.

FY - 2016		
Issuer Precinct # 19	**Issuer Precinct # 18**	**Issuer Precinct # 14**

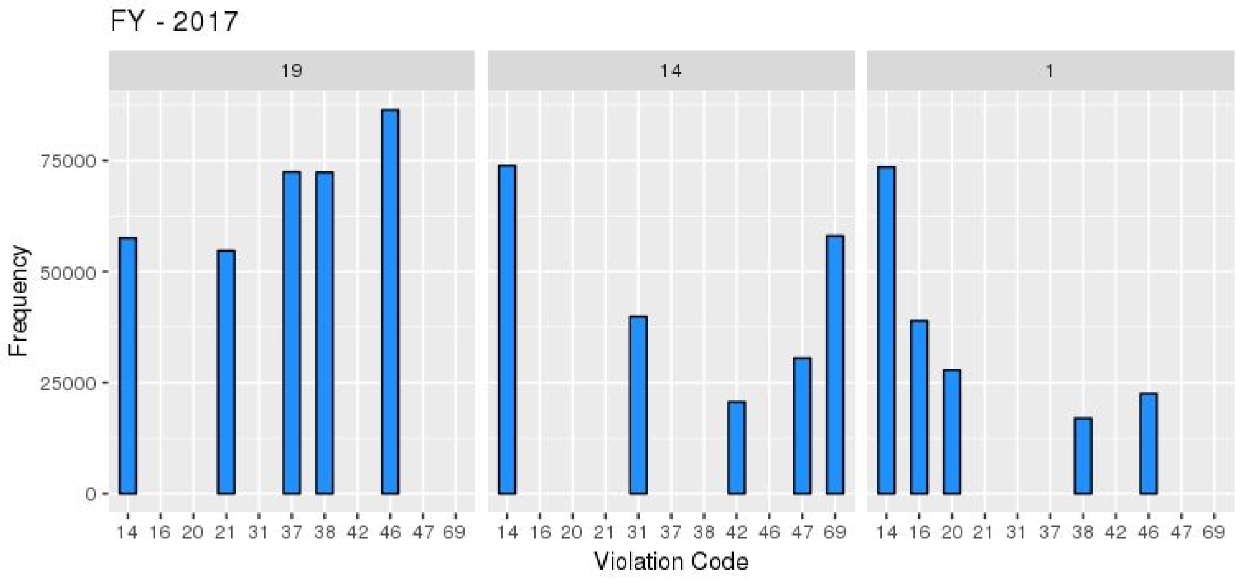
Violation Code # 38 - 77183	Violation Code # 14 - 99857	Violation Code # 69 - 67932
Violation Code # 37 - 75641	Violation Code # 69 - 47881	Violation Code # 14 - 62426
Violation Code # 46 - 73016	Violation Code # 47 - 24009	Violation Code # 31 - 35711
Violation Code # 14 - 61742	Violation Code # 31 - 22809	Violation Code # 47- 24450
Violation Code # 21 - 58719	Violation Code # 42 - 17678	Violation Code # 42 - 23662



From previous analysis (3.) we know, the top 3 issuer precincts which issued the most number of tickets are Precincts 19,14 and 1 in that order for 2015.

FY - 2017		
Issuer Precinct # 19	**Issuer Precinct # 14**	**Issuer Precinct # 1**
Violation Code # 46 - 86390	Violation Code # 14 - 73837	Violation Code #14 - 73522

Violation Code # 37 - 72437	Violation Code # 69 -58026	Violation Code # 16 - 38937
Violation Code # 38 - 72344	Violation Code # 31 - 39857	Violation Code # 20 - 27841
Violation Code # 14 - 57563	Violation Code # 47 - 30540	Violation Code # 46 - 22534
Violation Code # 21 - 54700	Violation Code # 42 - 20663	Violation Code # 38 - 16989



5. 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.

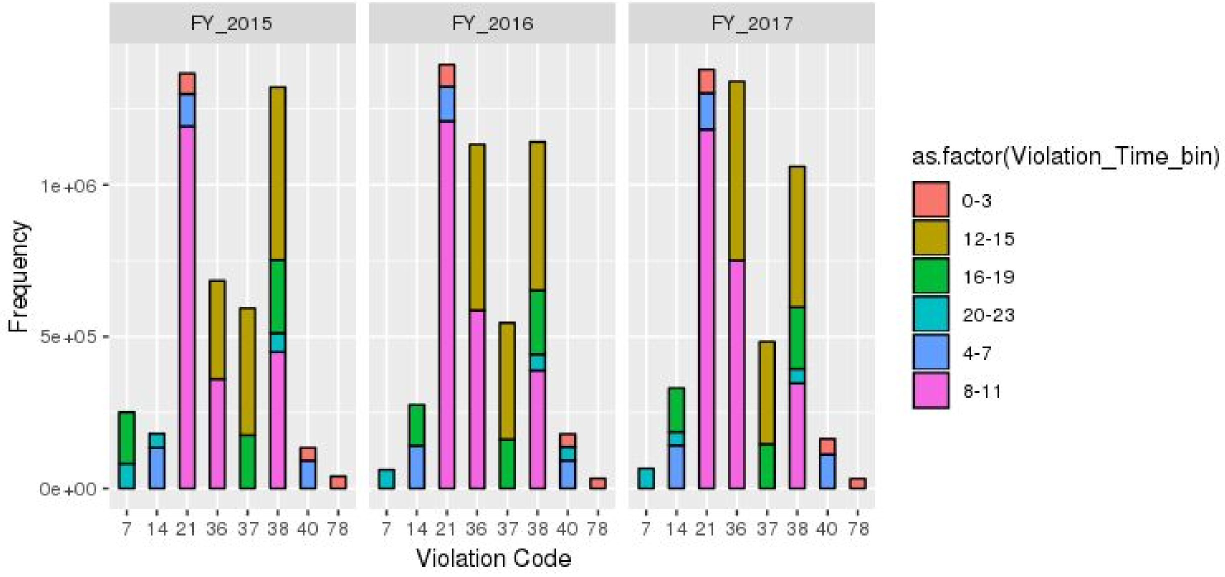
5.1 Find a way to deal with missing values, if any.

Missing values are negligible,so we are ignoring them from our analysis.

5.2 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

Time Bin	FY 2015	FY 2016	FY 2017
0-3	Violation Code # 21 - 67431	Violation Code # 21 - 72109	Violation Code # 21 - 77461
0-3	Violation Code # 40 - 42406	Violation Code # 40 - 42098	Violation Code # 40 - 50948
0-3	Violation Code # 78 - 39521	Violation Code # 78 - 32806	Violation Code # 78 - 32243
4-7	Violation Code # 14 - 134458	Violation Code # 14 - 140111	Violation Code # 14 - 141276
4-7	Violation Code # 21 - 106858	Violation Code # 21 - 114029	Violation Code # 21 - 119469
4-7	Violation Code # 40 - 91344	Violation Code # 40 - 91692	Violation Code # 40 - 112186
8-11	Violation Code # 21 - 1192163	Violation Code # 21 - 1209243	Violation Code # 21 - 1182689
8-11	Violation Code # 38 - 449070	Violation Code # 36 - 586791	Violation Code # 36 - 751422
8-11	Violation Code # 36 - 360365	Violation Code # 38-388099	Violation Code # 38-346518
12-15	Violation Code # 38 - 568272	Violation Code # 36-545717	Violation Code # 36-588395
12-15	Violation Code # 37-417613	Violation Code # 38-488302	Violation Code # 38-462758
12-15	Violation Code # 36 - 323524	Violation Code # 37 - 383361	Violation Code # 37 - 337075
16-19	Violation Code # 38 - 241327	Violation Code # 38 - 211267	Violation Code # 38 - 203232
16-19	Violation Code # 37-175802	Violation Code # 37-161655	Violation Code # 37-145784
16-19	Violation Code # 7 -	Violation Code # 14	Violation Code # 14 -

	168888	- 134976	144749
20-23	Violation Code # 7 - 81981	Violation Code # 7 - 60924	Violation Code # 7 - 65593
20-23	Violation Code # 38-62418	Violation Code # 38-53174	Violation Code # 38-47029
20-23	Violation Code # 14 - 45821	Violation Code # 40-44973	Violation Code # 14-44779



The maximum parking violations occur for Violation Code 21 in the morning between 8am - 11am, followed by Violation Codes 38 and 36 from 12pm - 3pm. The least violations occur post midnight until 3 am for FY 2015.

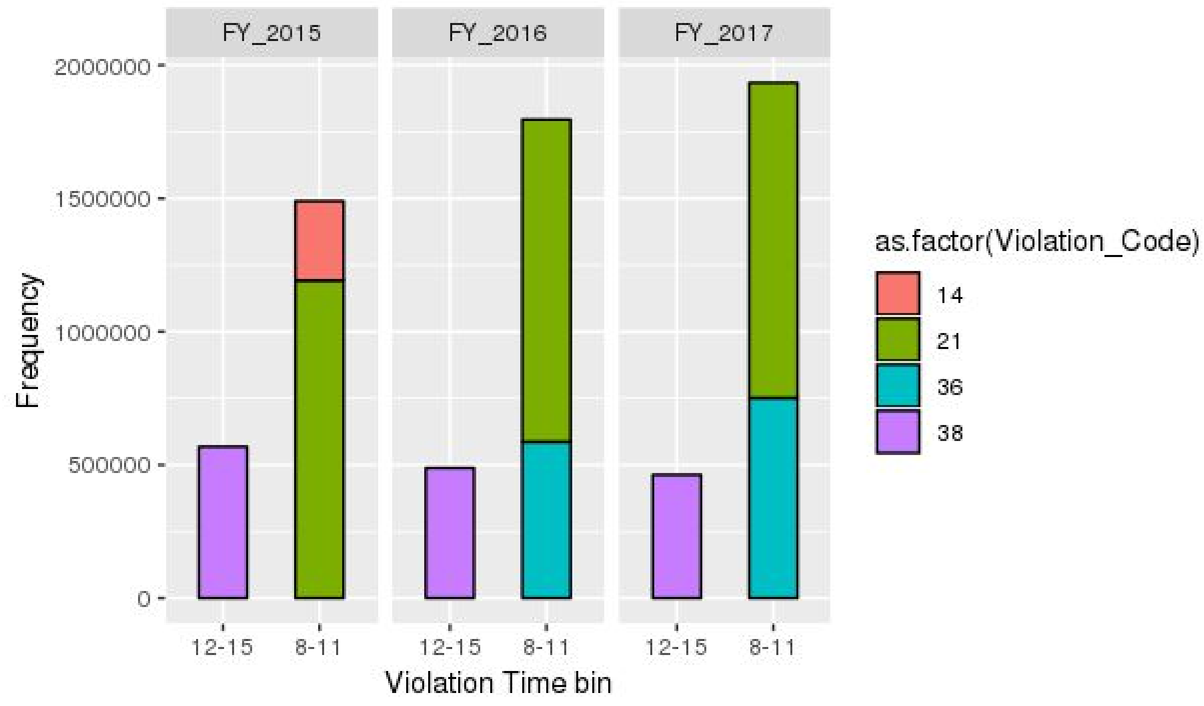
The maximum parking violations occur for Violation Code 21 in the morning between 8am - 11am, followed by Violation Codes 36 and 38 from 12pm - 3pm. The least violations occur post midnight until 3 am for FY 2016.

The maximum parking violations occur for Violation Code 21 in the morning between 8am - 11am, followed by Violation Codes 36 and 38 from 12pm - 3pm. The least violations occur post midnight until 3 am for FY 2017.

These findings suggest that common violations that occur in the morning imply the office rush hour and school pick up-drop off of children leading to tickets for no parking in parking zones,expired time receipts of Muni-Meters and overspeeding in school zones.

5.3 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)

FY 2015	FY 2016	FY 2017
Violation Code # 38 - (Time bin 12-15) - 568272	Violation Code # 38 - (Time bin 12-15) - 488302	Violation Code # 38 - (Time bin 12-15) - 462758
Violation Code # 21- (Time bin 8-11) - 1192163	Violation Code # 21- (Time bin 8-11) - 1209243	Violation Code # 21- (Time bin 8-11) - 1182689
Violation Code # 14 - (Time bin 8-11) - 297711	Violation Code # 36 - (Time bin 8-11) - 586791	Violation Code # 36 - (Time bin 8-11) - 751422



The most number of tickets are issued in the morning between 8am -11am for Violation 21[No Parking] , 14[No Parking] and 36[Exceeding the speed limit near school zones].These have increased rapidly from FY 2015 to FY 2017.

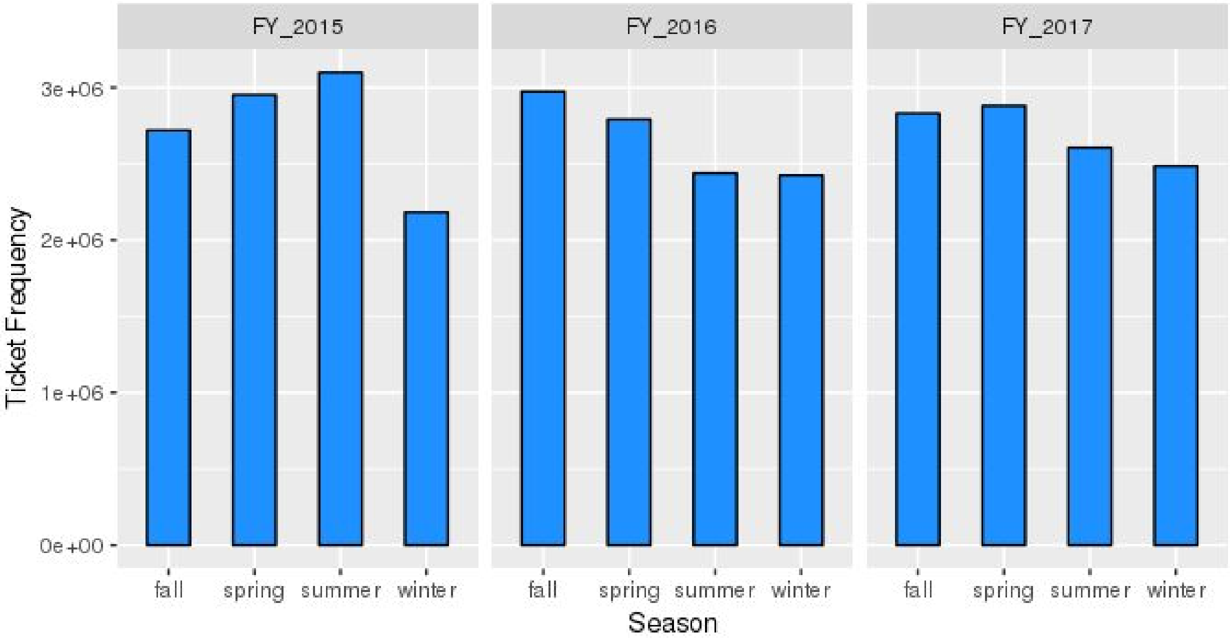
The tickets issued for Violation Code 38[Expired time on Muni-Meter receipts] is decreasing at a small pace across these three years.

This indicate the morning rush hour traffic[8-11] and parents picking up and dropping off children in morning[8-11] and afternoon [12-3 pm]

6. Let’s try and find some seasonality in this data

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

Season	FY 2015	FY 2016	FY 2017
Summer	3098729	2438069	2606208
Spring	2951328	2790946	2880687
Fall	2718868	2973396	2830802
Winter	2182331	2424488	2485331



Most number of tickets were issued in the summer followed by Spring ,Fall and least number of tickets in winter for FY 2015.

Most number of tickets were issued in Fall, followed closely by Spring, Summer and Winter for FY 2016.

Most number of tickets were issued in Spring, followed closely by Fall, Summer and Winter for FY 2017.

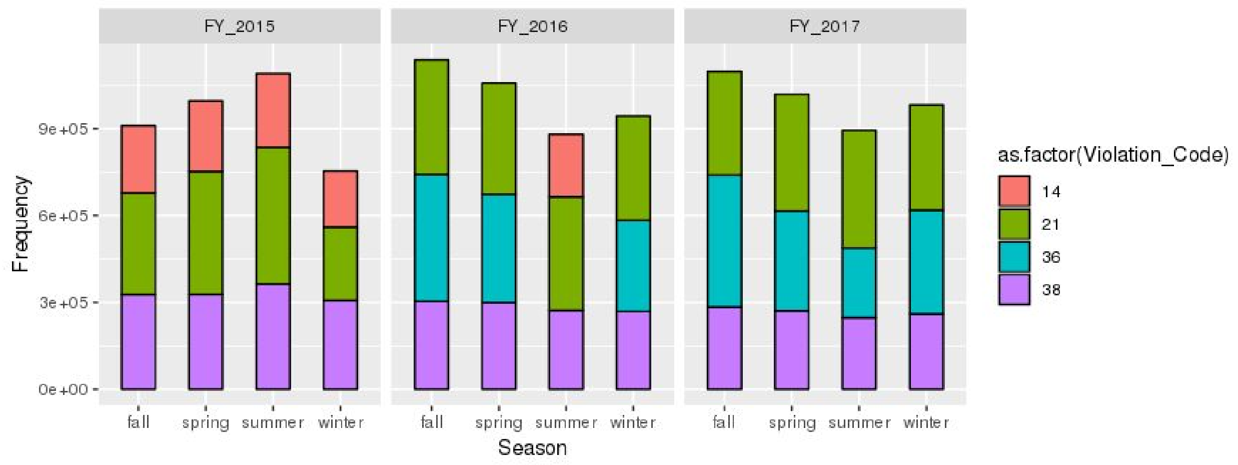
Overall Trend:

Fall and Spring season account for the highest number of parking tickets for FY 2016-2017 unlike summer season for FY 2015. We can infer that summer temperatures have been increasing over the years making Spring and Fall more common with people for travel and outdoor activities.

Winter season has the lowest number of parking tickets for all three years 2015-2017, with cold temperatures keeping people mostly indoors.

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

Season	FY 2015	FY 2016	FY 2017
Winter	Code #38 - 307012	Code #21 - 360268	Code #21 - 362341
Winter	Code #21 - 253214	Code #36 - 314765	Code #36 - 359338
Winter	Code #14 - 193337	Code #38 - 268421	Code #38 - 259723
Summer	Code #21 - 471627	Code #21 - 392205	Code #21 - 405961
Summer	Code #38 - 363815	Code #38 - 272419	Code #38 - 247561
Summer	Code #14 - 255182	Code #14 - 215683	Code #36 - 240396
Spring	Code #21 - 425350	Code #21 - 383757	Code #21 - 402807
Spring	Code #38 - 327057	Code #36 - 374362	Code #36 - 344834
Spring	Code #14 - 243769	Code #38 - 299459	Code #38 - 271192
Fall	Code #21 - 351423	Code #36 - 438320	Code #36 - 456046
Fall	Code #38 - 326702	Code #21 - 395357	Code #21 - 357479
Fall	Code #14 - 232339	Code #38 - 303397	Code #38 - 283828



The three most common violation codes for all seasons in FY 2015 are 14,21 and 38. Violation code 21 shows significant change with parking tickets dropping significantly in winter season compared to summer and spring. Violation codes 14 and 38 remain almost same across all seasons.

The three most common violation codes for fall,spring and winter seasons in FY 2016 are 21,36 and 38. We observe that for summer season there are cases of violation code 14 apart from violation codes 21 and 38.Violation Code 21 increases significantly in summer and remains almost same for spring and fall. Violation Code 14 is seen only in summer season. Violation Code 36 has the highest count in fall with dropping cases for spring and winter season.Violation Code 38 is almost same for all seasons in FY 2016.

The three most common violation codes for all seasons in FY 2017 are 21,36 and 38. Violation code 21 frequency is highest for summer and least for winter.Violation Code 36 drops significantly in summer season and is highest for fall. Violation Code 38 remains almost same across the four seasons.

Overall Trend:

Violation code 21 tickets increase in summer and is lowest for winter.

Violation Code 14 tickets frequency shows an abrupt change with no tickets issued for FY 2017(in top 3) and seen only in summer season for FY 2016.

Violation Code 36 drops in summer and peaks in fall for FY 2017, with not being in top 3 for FY 2015 and summer 2016. It increases in fall and drops in winter for FY 2016.

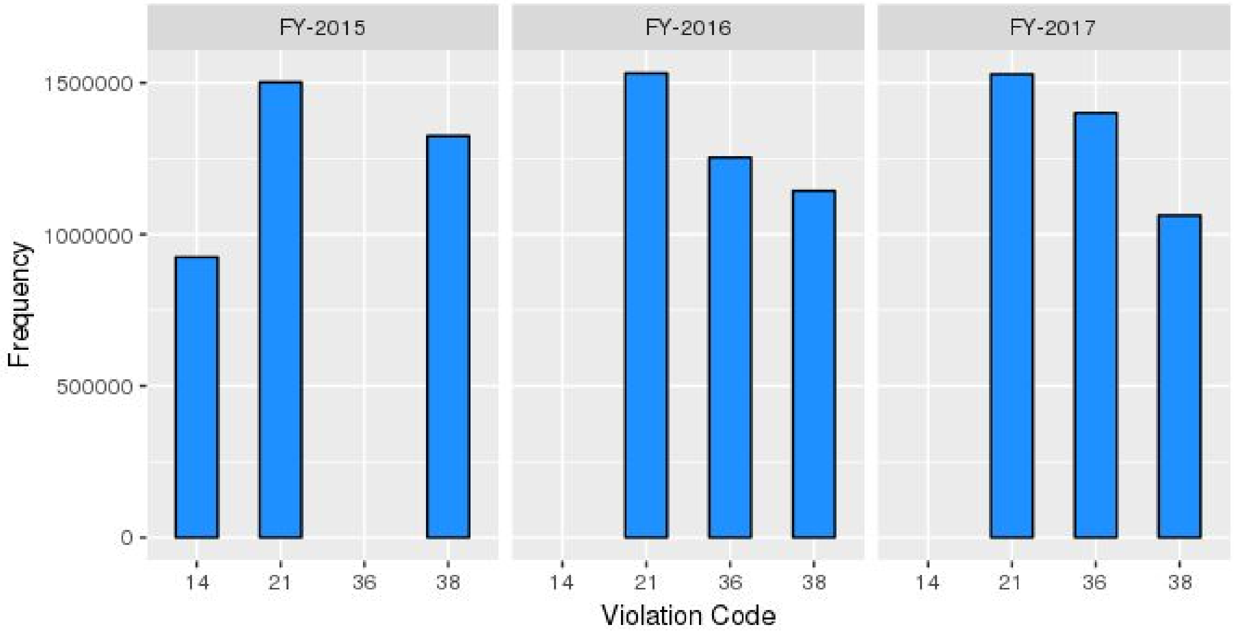
Violation code 38 tickets remain almost same across the seasons for all three years FY 2015 - FY 2017.

All violation codes show a drop in numbers for winter season.

7.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.

7.1 Find total occurrences of the 3 most common violation codes

FY 2015	FY 2016	FY 2017
Code # 21 - 1501614	Code # 21 - 1531587	Code # 21 - 1528588
Code # 38 - 1324586	Code # 36 - 1253512	Code # 36 - 1400614
Code # 14 - 924627	Code # 38 - 1143696	Code # 38 - 1062304



The three most commonly occuring violation codes for 2015 are 21,38 and 14.

The frequency of violation code 21 is highest compared to violation Codes 38 and 14 for 2015.

The three most commonly occuring violation codes for 2016 are 21,36 and 38.

The frequency of violation code 21 is highest compared to violation Codes 36 and 38 for 2016.

The three most commonly occuring violation codes for 2017 are 21,36 and 38.

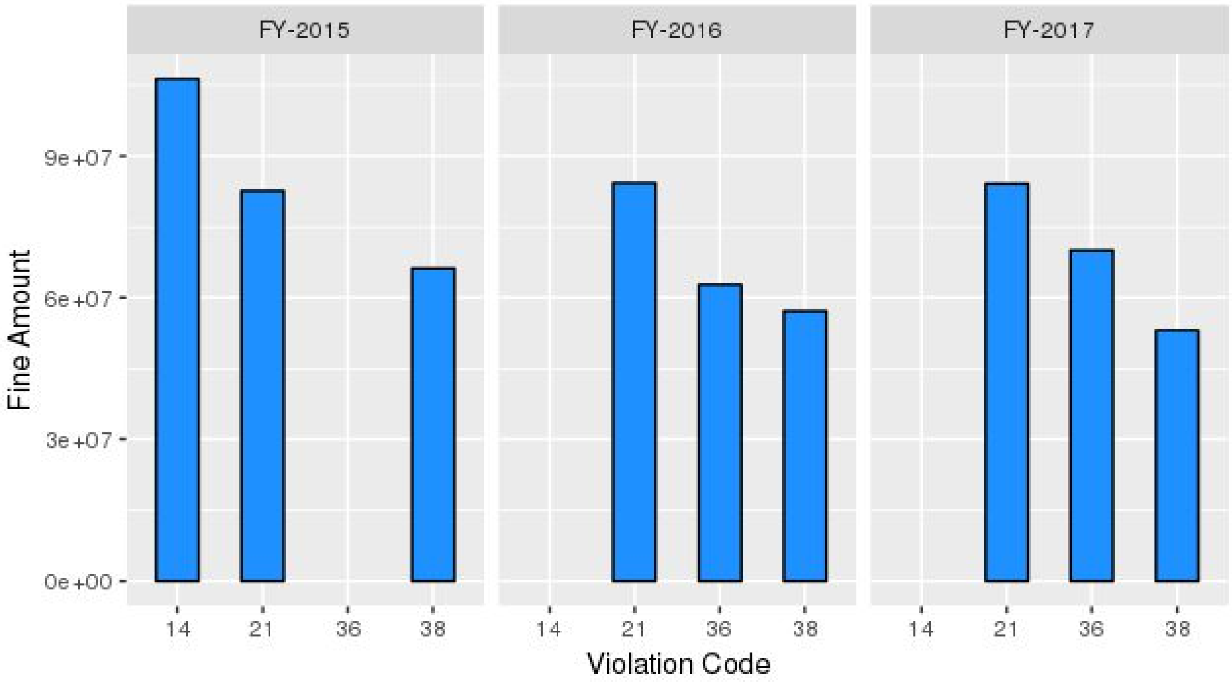
The frequency of violation code 21 is highest compared to violation Codes 36 and 38 for 2016.

Overall, we notice that the frequency of violation code 21[No Parking] is increasing over the years and frequency of violation code 38[Muni-Meter expired receipts] is decreasing over the the three years. The frequency of Violation Code 38[Exceeding the speed limit near school zones] is increasing rapidly for 2016 and 2017.

7.2 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.

FY 2015	FY 2016	FY 2017
Code # 14 - 106332105	Code # 21 - 84237285	Code # 21 - 84072340
Code # 21 - 82588770	Code # 36 - 62675600	Code # 36 - 70030700
Code # 38 - 66229300	Code # 38 - 57184800	Code # 38 - 53115200



The Violation Code 14 has generated the maximum revenue for FY 2015, followed by earnings from violation codes 21 and 38.

The Violation Code 21 has generated the maximum revenue for FY 2016, followed by earnings from violation codes 36 and 38.

The Violation Code 21 has generated the maximum revenue for FY 2017, followed by earnings from violation codes 36 and 38.

Overall Trend: We notice that the amount in fines generated by Violation code 21 is highest for 2016 and 2017 and remains almost constant for these two years. The amount in fines is decreasing for Violation Code 38 consistently from 2015 to 2017. Even though frequency of Violation Code 14 is least for 2015, it generated the highest mount in fines. The amount in fines generated from Violation Code 36 shows an increasing trend for 2016 and 2017.

7.3 What can you intuitively infer from these findings?

We notice that majority of parking tickets issued are for parking in No Parking zones(21), expired time on Muni-Meter receipts(37-38) and Exceeding the mentioned speed limit near school zones(36).

We infer that there are a huge number of cars owned by the people and the nature of the violation codes point to the lack of parking space in NYC. The existing space is not adequate to meet the demand and hence more parking space is needed.

Another, inference is that drivers with exceeding speed limit need to be made aware of the dangers of over-speeding near schools which may lead to fatalities.

The cars going more slowly will have an easier time avoiding crashes and secondly increases the chances of survival rates when crashes do happen.