NYC Parking Tickets: An Exploratory Analysis

Problem Statement

New York City is a thriving metropolis. Just like most other metros its size, one of the biggest problems its citizens face is parking. The classic combination of a huge number of cars and cramped geography leads to a huge number of parking tickets.

In an attempt to scientifically analyze this phenomenon, the NYC Police Department has collected data for parking tickets. Of these, the data files for multiple years are publicly available. it is required to perform some exploratory analysis on a part of 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 analyse the parking tickets over the year 2017.

The purpose of this case study is to conduct an exploratory data analysis that will help you understand the data. The questions given below will guide your analysis.

Data Cleanup

We should carry out following data cleanup steps on data set, before we answer the questions:

1. Check for duplicate data: There is no duplicate data in our data set
2. As our data set contains data over several years, we have filtered data for year 2017 because our analysis is over year 2017 only
3. We have derived new columns like minutes, hours etc. to carry out further in-depth analysis
4. We have dropped N/A (null) values

Examine the Data

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

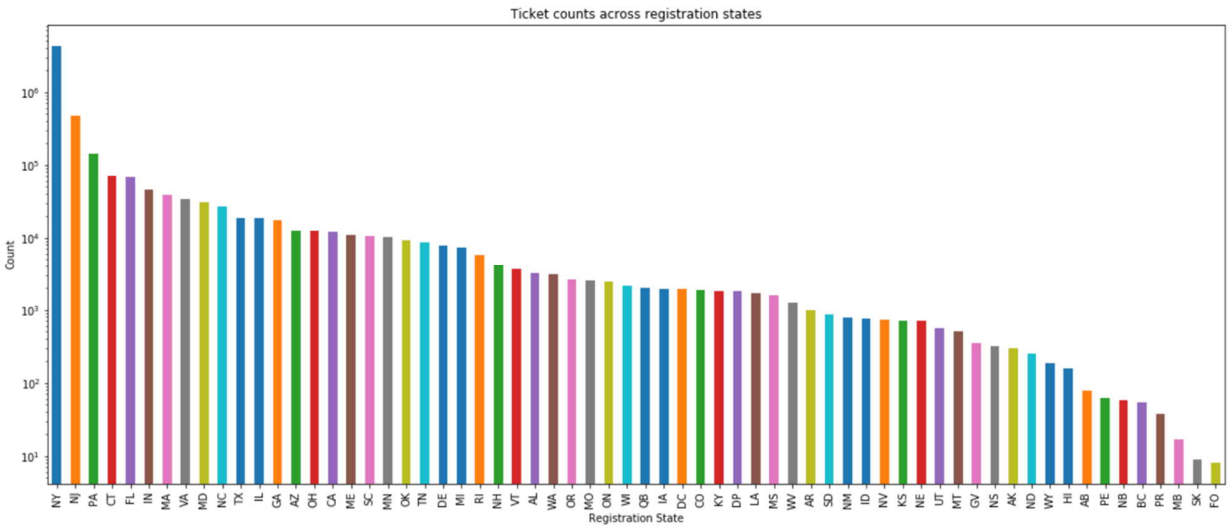
Answers 1: After performing data cleanup and filtration steps mentioned in “Data Cleanup” above, we have got total number of tickets equal to 5,431,834 over the year 2017. (Prior to data cleaning the count for the total tickets was 5,431,918)

Question 2: Find out the number of unique states from where the cars that got parking tickets came. (Hint: use the column ‘Registration State’)

There is a numeric entry ‘99’ in the column, which should be corrected. Replace it with the state having the maximum entries. Provide the number of unique states again.

Answer 2: Initial number of distinct states obtained were 65, of which one was errored (with value 99). The error values were replaced with most frequent state ‘NY’ and then the count of distinct states was reduced to 64.

Below we have added plot to show ticket distribution across different states and from below graph we can observe that the top three states for ticket counts are NY , NJ and PA respectively.



Aggregation tasks

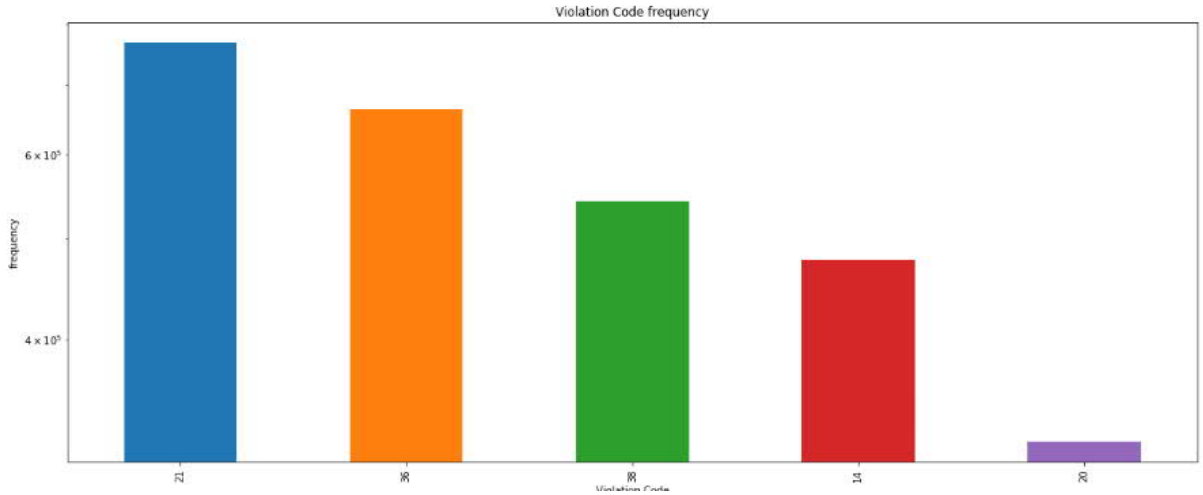
Question 1: How often does each violation code occur? Display the frequency of the top five violation codes.

Answer 1

From our analysis, we create the violation code frequency plot and from the plot we can observe that the top 5 violation code are 21, 36, 38, 14 and 20. Top 5 violation code with respective frequency are:

Violation code	Frequency
21	768,056
36	662,765
38	542,078
14	476,663
20	319,644

Plot to show frequency of top five violation code is shown below:



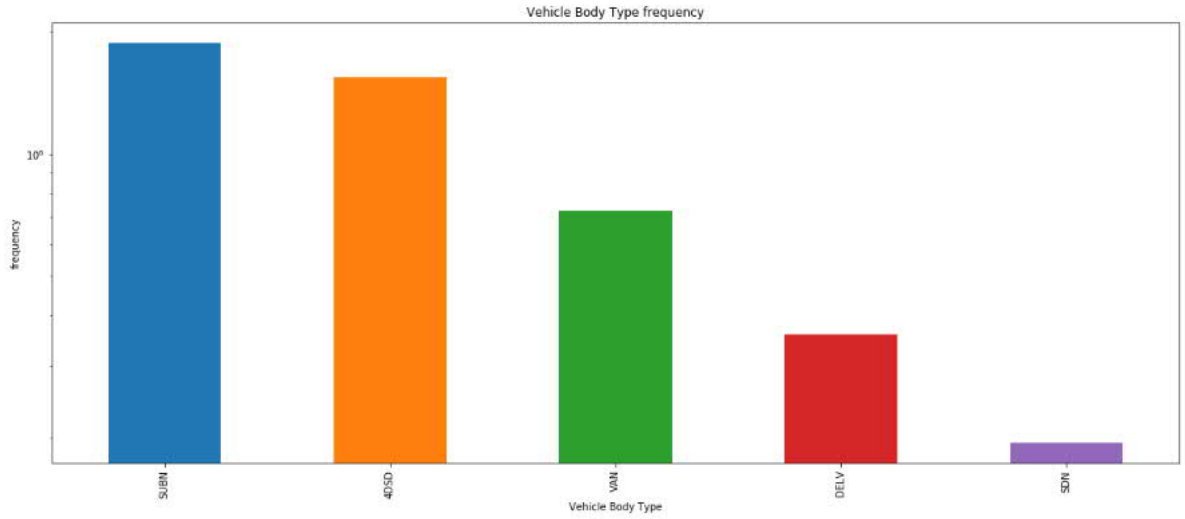
Question 2: How often does each 'vehicle body type' get a parking ticket? How about the 'vehicle make'?

Answer 2:

Top 5 vehicle body type with respective frequency is shown below:

Vehicle body type	Frequency
SUBN	1,883,925
4DSD	1,547,312
VAN	724,025
DELV	358,980
SDN	194,164

Plot to show frequency of top five vehicle body type is shown below:

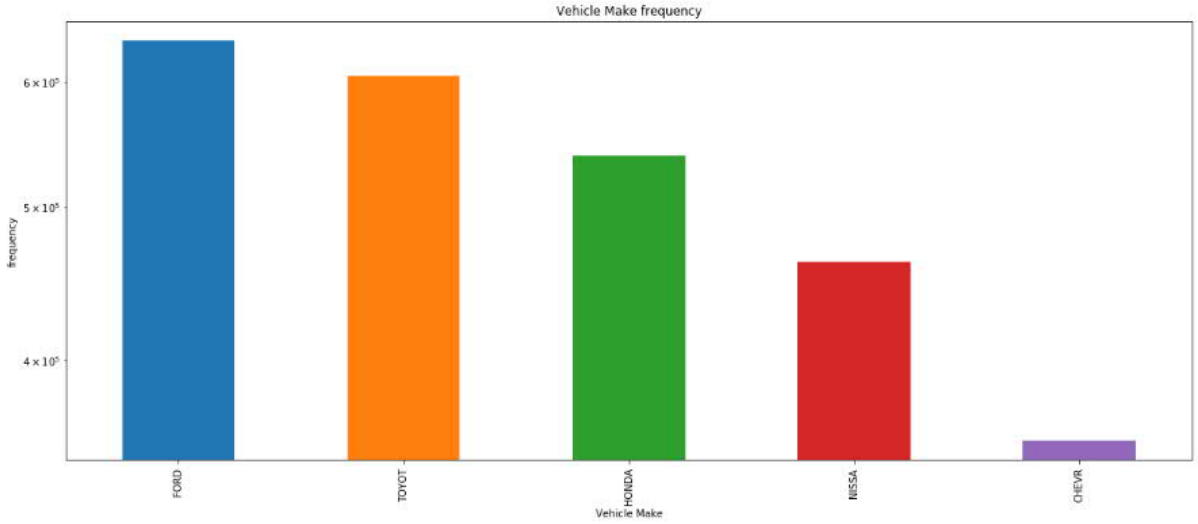


From above plot we can observe that top 5 vehicle body type based on frequency are SUBN, 4DSD, VAN, DELV, SDN.

Top 5 vehicle make with respective frequency is shown below:

Vehicle make	Frequency
FORD	636,839
TOYOT	605,283
HONDA	538,874
NISSA	462,006
CHEVR	356,026

Plot to show frequency of top five vehicle make is shown below:



From above plot we can observe that top 5 vehicle make based on frequency are FORD, TOYOT, HONDA, NISSA, CHEVR.

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

1. 'Violation Precinct' (This is the precinct of the zone where the violation occurred). Using this, can you draw any insights for parking violations in any specific areas of the city?
2. 'Issuer Precinct' (This is the precinct that issued the ticket.)

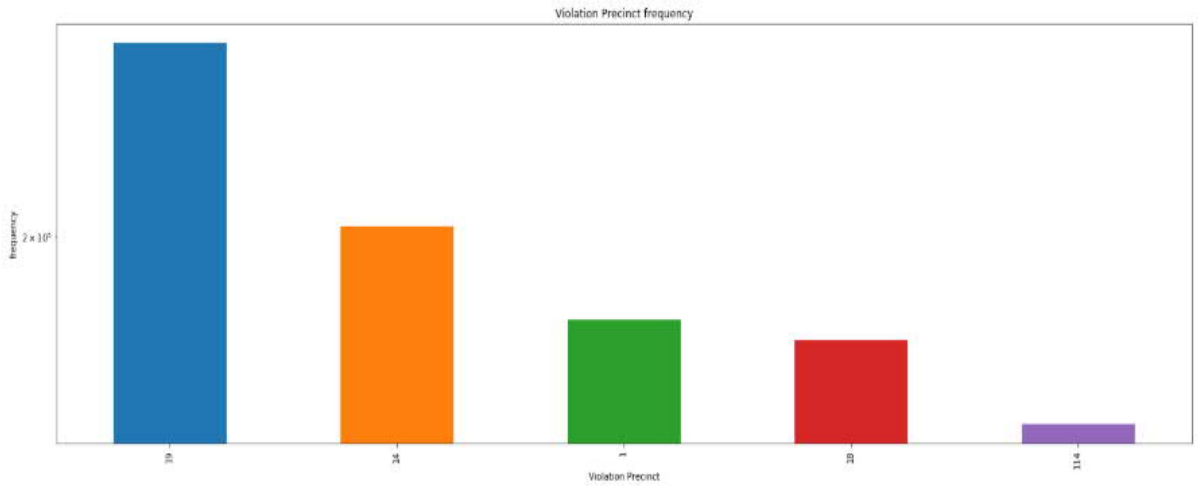
Here, you would have noticed that the data frame has the 'Violating Precinct' or 'Issuing Precinct' as '0'. These are erroneous entries. Hence, you need to provide the records for five correct precincts.

Answer 3:

Top 5 Violation Precinct with respective frequency is shown below:

Violation Precinct	Frequency
19	274,442
14	203,553
1	174,701
18	169,130
114	147,442

We plot to show most frequent Violation Precinct along with annual count, and exclude the erroneous entries

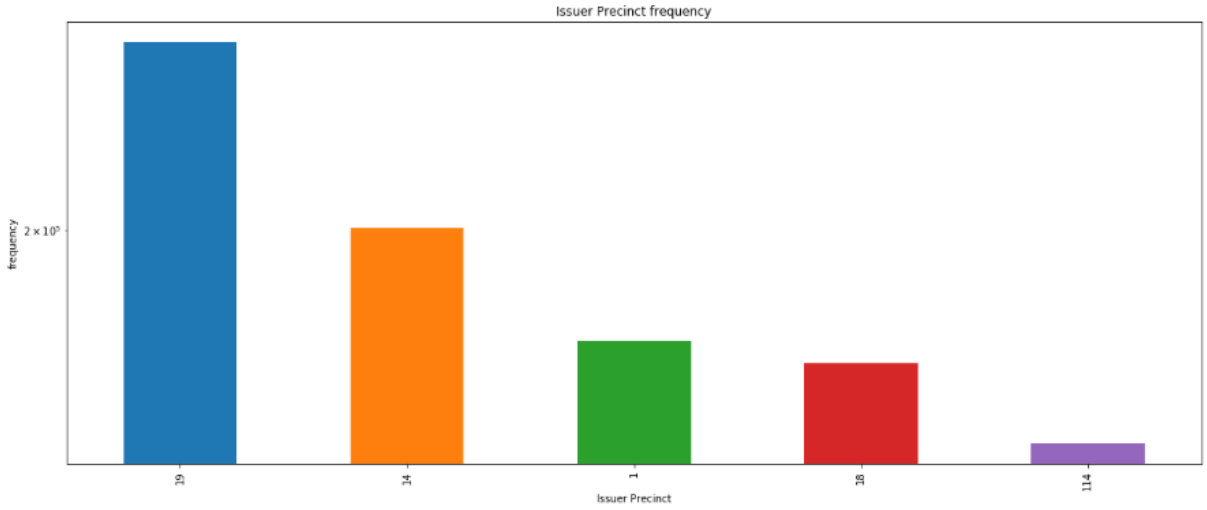


From above plot we can observe that top 5 Violation Precinct are 19, 14, 1, 18, 114.

Top 5 Issuer Precinct with respective frequency is shown below:

Issuer Precinct	Frequency
19	266,961
14	200,495
1	168,740
18	162,994
114	144,053

We now plot to show most frequent Issuer Precinct along with annual count, and exclude the erroneous entries



From above plot we can observe that top 5 Issuer Precinct are 19, 14, 1, 18, 114.

Question 4 : Find the violation code frequencies for three precincts that 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 4:

We will carry out this in two steps. In the first step, we have created a SQL view with frequencies grouped by Issuer Precinct and Violation Code. In second step we use the dense rank function with partition over Issuer Precinct to get the top 5 ranks which are the top five violation codes along with their frequencies.

From previous question top 3 Issuer Precinct are 19, 14, 1.

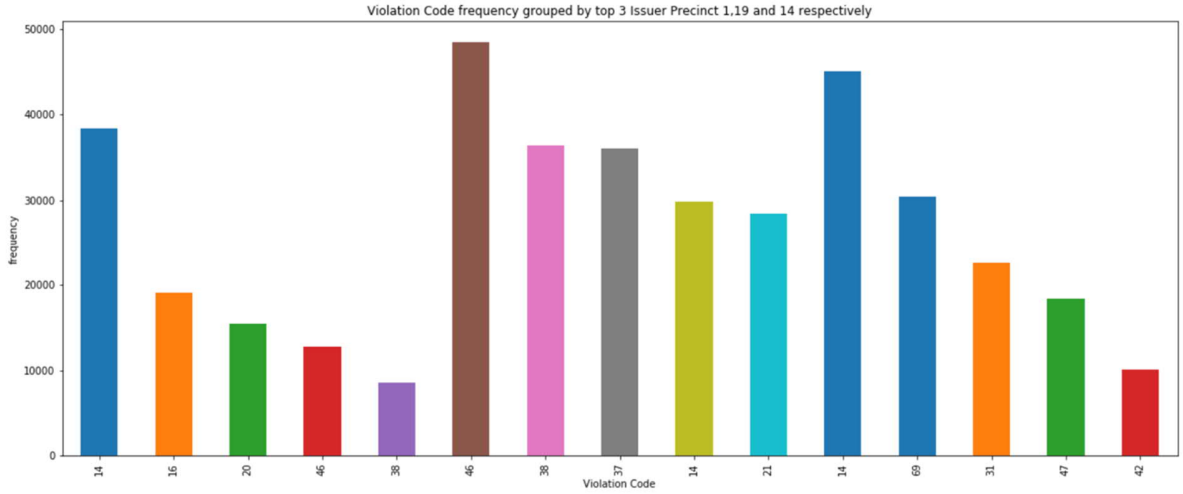
Violation code frequencies for top three precincts that have issued the most number of tickets is shown below:

Issuer Precinct	Violation Code	Frequency Per Year	Rank
1	14	38,354	1
1	16	19,081	2
1	20	15,408	3
1	46	12,745	4
1	38	8,535	5
19	46	48,445	1
19	38	36,386	2
19	37	36,056	3
19	14	29,797	4
19	21	28,415	5
14	14	45,036	1
14	69	30,464	2
14	31	22,555	3
14	47	18,364	4
14	42	10,027	5

Violation code 14 exist in top 5 violation for all three Issuer Precinct 1, 19 and 14. However, violation codes 46 and 38 exist in top 5 for both 1 and 19 Issuer Precinct.

From the frequency comparison prospective, violation code 14 has relatively higher frequencies for Issue Precinct 1 and 14. However, violation code 46 has higher frequency restricted to 19 Issue Precinct.

The summary can be realized from the below plot.



Observation based on above plot

1) First 5 violation code (14,16,20,46,38) belongs to Issuer Precinct 1
2) Next 5 violation code (46,38,37,14,21) belongs to Issuer Precinct 19
3) Next 5 violation code (14,69,31,47,42) belongs to Issuer Precinct 14

Question 5: Find out the properties of parking violations across different times of the day:

- Find a way to deal with missing values, if any.
- The Violation Time field is specified in a strange format. Find a way to make this a time attribute that you can use to divide into groups.
- Divide 24 hours into six equal discrete bins of time. Choose the intervals as you see fit. For each of these groups, find the three most commonly occurring violations.
- Now, try another direction. For the three most commonly occurring violation codes, find the most common time of the day (in terms of the bins from the previous part).

Answer 5: We have extracted the hour information into new column Violation Hour as part of data cleaning. Also we have removed the missing values as a part of data cleaning.

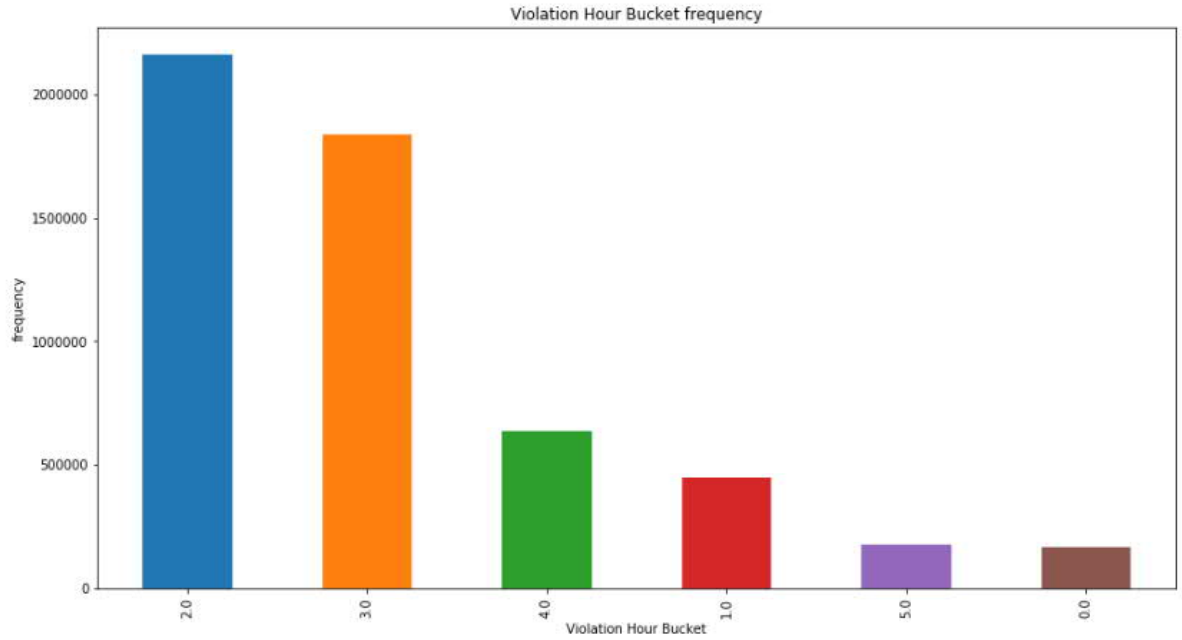
We have divided data into 6 bin as shown below:

1) Bin 0 – 0 to 4 hours
2) Bin 1 – 4 to 8 hours
3) Bin 2 – 8 to 12 hours
4) Bin 3 – 12 to 16 hours
5) Bin 4 – 16 to 20 hours
6) Bin 5 – 20 to 24 hours

Frequency for each bucket is shown below:

Violation hour bucket	Frequency
0	164,535
1	449,882
2	2,163,563
3	1,839,957
4	637,539
5	176,358

Frequency plot for each bucket is shown below and Most of the violations occur in bucket 2 duration i.e. between 8 hours to 12 hours in the day.



Three most commonly occurring violations for each bucket is shown below:

Violation Hour Bucket	Violation Code	Frequency Per Year	Rank
0.0	21	36,958	1
0.0	40	25,867	2
0.0	78	15,528	3
1.0	14	74,114	1
1.0	40	60,652	2
1.0	21	57,895	3
4.0	38	102,855	1
4.0	14	75,902	2
4.0	37	70,345	3
3.0	36	286,284	1
3.0	38	240,721	2
3.0	37	167,026	3
2.0	21	598,066	1
2.0	36	348,165	2
2.0	38	176,570	3
5.0	7	26,293	1
5.0	40	22,337	2
5.0	14	21,045	3

Comparison of top three most commonly occurring violations for each bucket is shown below

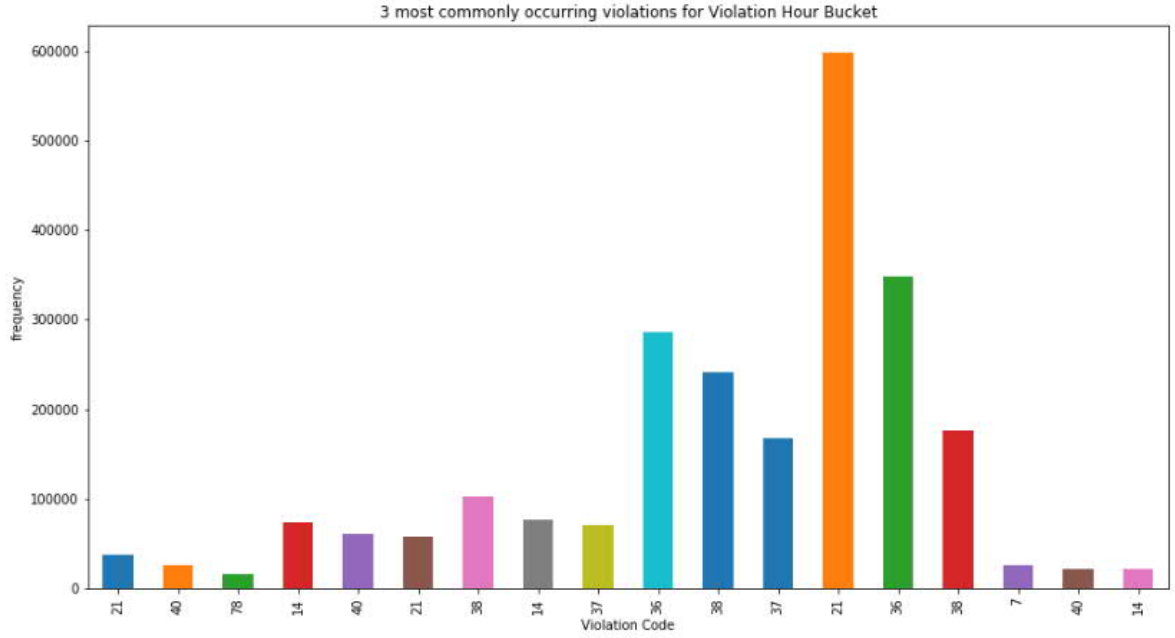


Table for most common time of the day for three most commonly occurring violation codes is shown below:

Violation Code	Frequency Per Year	Violation Hour Bucket
21	598,066	2.0
21	74,694	3.0
21	57,895	1.0
21	36,958	0.0
21	259	4.0
21	184	5.0
36	348,165	2.0
36	286,284	3.0
36	14,782	1.0
36	13,534	4.0
38	240,721	3.0
38	176,570	2.0
38	102,855	4.0
38	20,347	5.0
38	1,273	1.0
38	312	0.0

Top 3 violation codes are 21, 36, 38. Their frequencies distribution across Violation hours bucket are mentioned above.

Question 6 : Let's try and find some seasonality in this data:

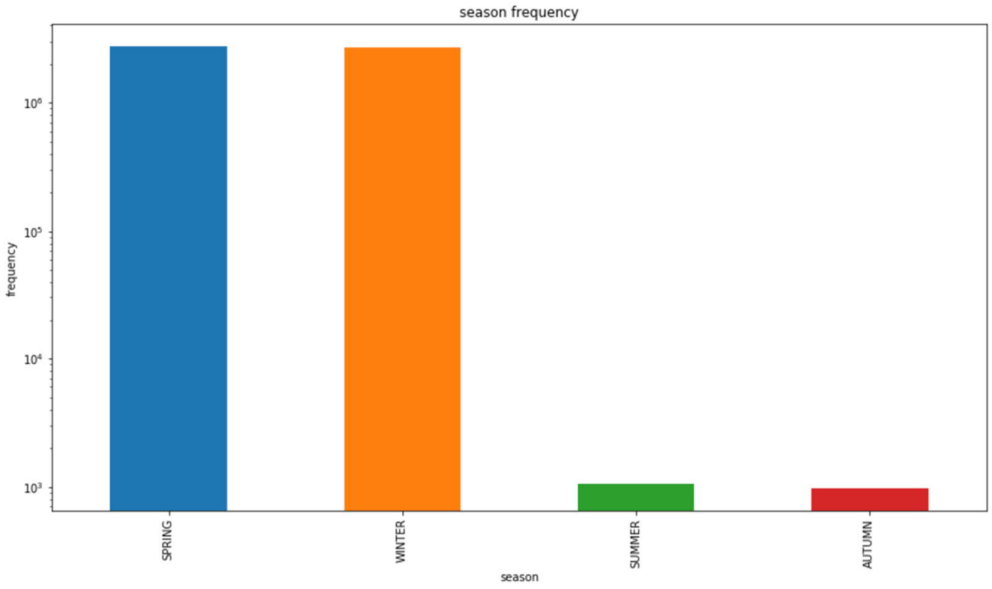
- First, divide the year into a certain number of seasons, and find the frequencies of tickets for each season.
- Then, find the three most common violations for each of these seasons.

Answer 6 : The seasons have been assigned based on the month values for Issue date. For months 1-3, season is WINTER, for months 4-6 season is SPRING, for months 7-9 season is SUMMER and for months 10-12 season is AUTUMN.

Table for frequency of each season is shown below:

Violation hour bucket	Frequency
Spring	2,760,785
Winter	2,669,033
Summer	1,046
Autumn	970

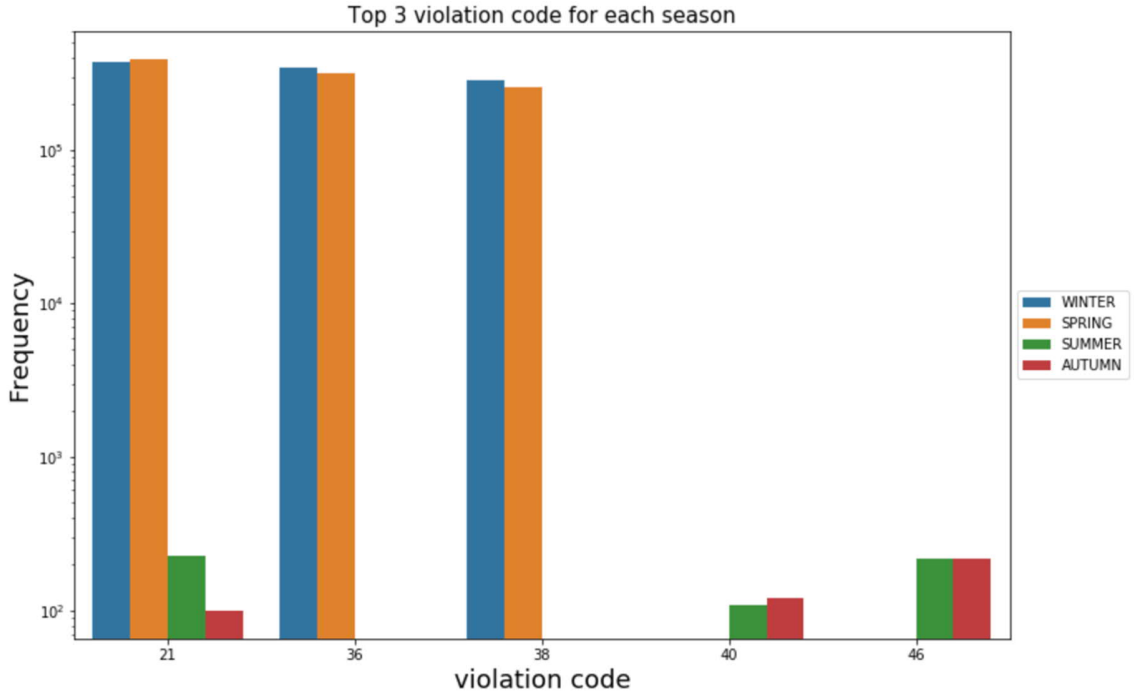
Plot for frequency distribution of each season is shown below:



Top three violation code for each season is shown in below table:

Season	Violation Code	Frequency Per Year	Rank
WINTER	21	373,862	1
WINTER	36	348,240	2
WINTER	38	286,999	3
SPRING	21	393,866	1
SPRING	36	314,525	2
SPRING	38	255,064	3
SUMMER	21	228	1
SUMMER	46	219	2
SUMMER	40	109	3
AUTUMN	46	219	1
AUTUMN	40	121	2
AUTUMN	21	100	3

Plot for comparison between top 3 violation code for each season is shown below:



Question 7 : The fines collected from all the instances of parking violation constitute a source of revenue for the NYC Police Department. Let's take an example of estimating this for the three most commonly occurring codes:

- Find the total occurrences of the three most common violation codes.
- Then, visit the website:
  http://www1.nyc.gov/site/finance/vehicles/services-violation-codes.page
  It lists the fines associated with different violation codes. They're divided into two categories: one for the highest-density locations in the city and the other for the rest of the city. For the sake of simplicity, take the average of the two.
- Using this information, find the total amount collected for the three violation codes with the maximum tickets. State the code that has the highest total collection.
- What can you intuitively infer from these findings?

Answer 7 : The total occurrences for the most three violation codes are shown in below table :

Violation Code	Frequency Per Year
38	542,078
21	768,056
36	662,765

As per the information provided in the web link, the average fine for each violation code is evaluated as :

- Average fine for Violation code 21 : (65+45)/2 = 55 $
- Average fine for Violation code 36 : (50+50)/2 = 50 $
- Average fine for Violation code 38 : (65+35)/2 = 50 $

Using the above information for frequency and average fine for each violation code, the amount collected is 102,485,230 $.

We can infer that fine collection for the violations are good source of revenue generation with big amount of $102,485,230 getting generated from top 3 violations.