# Analysis for "NYC Parking Tickets" Case Study

**Assumptions**:

- Considered only records only from the year 2017.
- Interpreted State Code 99 as NJ.

**Solutions**:

**Examine the data.**

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

   Total number of tickets for the year 2017 = 5431918.

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

   Number of unique states from where the cars that got parking tickets came = 64

**Aggregation tasks**

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

Top 5 Violation Codes:

| Violation Code | Frequency of Violation Codes |
|---|---|
| 21 | 768087 |
| 36 | 662765 |
| 38 | 542079 |
| 14 | 476664 |
| 20 | 319646 |

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

Top 5 Vehicle Body Types which got a parking ticket:

| Vehicle Body Type | Frequency of Parking Tickets |
|---|---|
| SUBN | 1883954 |
| 4DSD | 1547312 |
| VAN | 724029 |
| DELV | 358984 |
| SDN | 194197 |

Top 5 Vehicle Makes which got a parking ticket:

| Vehicle Make | Frequency of Parking Tickets |
|---|---|
| FORD | 636844 |
| TOYOT | 605291 |
| HONDA | 538884 |
| NISSA | 462017 |
| CHEVR | 356032 |

**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:

***ASSUMPTION***: 0 is not considered as a Violation Precinct as these are erroneous entries in the dataset.

- 'Violation Precinct' (This is the precinct of the zone where the violation occurred).

  Five Highest frequency of Violation Precinct:

| Violation Precinct | Frequency |
|---|---|
| 19 | 274445 |
| 14 | 203553 |
| 1 | 174702 |
| 18 | 169131 |
| 114 | 147444 |

- 'Issuer Precinct' (This is the precinct that issued the ticket.)

  Five Highest frequency of Issuer Precinct:

| Issuer Precinct | Frequency |
|---|---|
| 19 | 266961 |
| 14 | 200495 |
| 1 | 168740 |
| 18 | 162994 |
| 114 | 144054 |

**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?

***ASSUMPTION:*** 0 is not considered as a Violation Precinct as these are erroneous entries in the dataset.

| Violation Precinct | Frequency |
|---|---|
| 19 | 274445 |
| 14 | 203553 |
| 1 | 174702 |

Violation Codes for Violation Precinct = 19 along with the frequency:

| Violation Code | Frequency |
|---|---|
| 46 | 50785 |
| 38 | 37483 |
| 37 | 36468 |
| 14 | 30376 |
| 21 | 29415 |

Violation Codes for Violation Precinct = 14 along with the frequency:

| Violation Code | Frequency |
|---|---|
| 14 | 45885 |
| 69 | 30465 |
| 31 | 22649 |
| 47 | 18691 |
| 42 | 10027 |

Violation Codes for Violation Precinct = 1 along with the frequency:

| Violation Code | Frequency |
|---|---|
| 14 | 40226 |
| 16 | 19278 |
| 20 | 15743 |
| 46 | 13534 |
| 38 | 8588 |

Violation Code 14 is common across all three Precincts 19, 14 and 1. Violation Code = 46, 38 are common across Precincts 19 and 1.

**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).

***ASSUMPTION***:

Using the 'pyspark' functions isnan() and isNull(), it is not possible to determine missing values, as pyspark treats 'NA' in the dataset as string.

On manual observation, it was observed that the percentage of missing values compared to overall data set is very less. Hence, Ignored Missing Values.

Considered Hour bins as below:

o Bin 1 -- 0 to 3 hrs
o Bin 2 -- 4 to 7 hrs
o Bin 3 -- 8 to 11 hrs
o Bin 4 -- 12 to 15 hrs
o Bin 5 -- 16 to 19 hrs
o Bin 6 -- 20 to 23 hrs

The three most commonly occurring violations for each Hour Bin are given below:

| Hour Bin | Violation Code | Total Tickets |
|---|---|---|
| 1 | 21 | 34704 |
| 1 | 40 | 23629 |
| 1 | 14 | 14168 |
| 2 | 40 | 14175 |
| 2 | 14 | 7860 |
| 2 | 20 | 7046 |
| 3 | 21 | 652968 |
| 3 | 36 | 362947 |
| 3 | 14 | 215279 |
| 4 | 36 | 286284 |
| 4 | 38 | 240795 |
| 4 | 37 | 167044 |
| 5 | 38 | 102855 |
| 5 | 14 | 75902 |
| 5 | 37 | 70345 |
| 6 | 7 | 26293 |
| 6 | 40 | 22337 |
| 6 | 14 | 21045 |

The three most commonly occurring violation codes are:

| Violation Code | Frequency |
|---|---|
| 21 | 768087 |
| 36 | 662765 |
| 38 | 542079 |

The most common time of the day for Violation Code = 21 are shown below:

| Violation Code | ViolationTime24Hr Code | Frequency |
|---|---|---|
| 21 | 8.0 | 190820 |
| 21 | 9.0 | 187017 |
| 21 | 11.0 | 167647 |
| 21 | 12.0 | 74616 |
| 21 | 7.0 | 52835 |
| 21 | 10.0 | 52585 |

The most common time of the day for Violation Code = 36 are shown below:

| Violation Code | ViolationTime24 Hr Code | Frequency |
|---|---|---|
| 36 | 12.0 | 101991 |
| 36 | 13.0 | 98247 |
| 36 | 11.0 | 98077 |
| 36 | 10.0 | 97661 |
| 36 | 9.0 | 91261 |
| 36 | 8.0 | 61166 |

The most common time of the day for Violation Code = 38 are shown below:

| Violation Code | ViolationTime24Hr Code | Frequency |
|---|---|---|
| 38 | 13.0 | 73310 |
| 38 | 14.0 | 68117 |
| 38 | 12.0 | 56040 |
| 38 | 10.0 | 54230 |
| 38 | 9.0 | 53839 |
| 38 | 16.0 | 49694 |

Majorly top 3 violations occurred in between 8 AM to 9 AM, followed by in between 12 PM to 1 PM.

**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.

***ASSUMPTION***:

Considered Seasons as below:

- o Season 1 -- January to March
- o Season 2 -- April to June
- o Season 3 -- July to September
- o Season 4 -- October to December

Frequencies of tickets for each season are shown below:

| Season Bin Number | Total Tickets |
|---|---|
| 1 | 2669069 |
| 2 | 2760833 |
| 3 | 1046 |
| 4 | 970 |

Most common violations for each season

| Season Bin Number | Violation Code | Total Tickets |
|---|---|---|
| 1 | 21 | 373874 |
| 1 | 36 | 348240 |
| 1 | 38 | 287000 |
| 2 | 21 | 393885 |
| 2 | 36 | 314525 |
| 2 | 38 | 255064 |
| 3 | 21 | 228 |
| 3 | 46 | 219 |
| 3 | 40 | 109 |
| 4 | 46 | 219 |
| 4 | 40 | 121 |
| 4 | 20 | 100 |

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, 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?

   Top 3 violation codes are shown below:

| Violation Code | Frequency |
|---|---|
| 21 | 768087 |
| 36 | 662765 |
| 38 | 542079 |

As per the Website, for Violation code = 21, Manhattan 96th St. & below = $65 and All Other Areas = $45

Average fine for Violation code 21 = $55

As per the Website, for Violation code = 36, Manhattan 96th St. & below = $50 and All Other Areas = $50

Average fine for Violation code 36 = $50

As per the Website, for Violation code = 38, Manhattan 96th St. & below = $65 and All Other Areas = $35

Average fine for Violation code 38 = $50

Top 3 violation codes based on total fine collection are shown below:

| Violation Code | Total Fine Amount |
|---|---|
| 38 | 27103950 |
| 36 | 33138250 |
| 21 | 42244785 |

Violation Code that has the highest total collection = 21

**Inference:**

- For high frequency of tickets, 'Violation Precinct' and 'Issuer Precinct' are the same which include the Precinct 19, 14, 1, 18 and 114 respectively.

- Most parking tickets came from NJ.
- Violation code 14 has the highest frequency of occurrence.
- Most tickets were issues in months April, May and June followed by January, February and March.
- Violation Code 21 has the highest total fine collection.