# NYC Parking Tickets: An Exploratory Analysis

GROUP ASSIGNMENT:
BIG DATA ANALYTICS

GROUP MEMBERS:
AMAN RAI (raiaman15@gmail.com)
MUKESH TIMBREWALA (mukeshtibs@gmail.com)
UTKARSH KANT (utkarsh.kant@gmail.com)
DEVA SHARMA (deva.sharma24@gmail.com)

**THIS FILE CONTAINS THE SOLUTION AND EXPLANATION TO THE QUESTIONS ASKED IN THIS ASSIGNMENT**

**Examine the data**

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

➔ Answer: 10803028

```
+------------------------------+
|number of tickets for the year|
+------------------------------+
|                      10803028|
+------------------------------+
```

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: 66

```
+---------------------------------+
|count(DISTINCT registration_state)|
+---------------------------------+
|                               66|
+---------------------------------+
```

**Aggregation tasks**

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

➔ Answer:

```
+--------------+---------------------------------------+
|violation_code|number of tickets for the violation code|
+--------------+---------------------------------------+
|            21|                                1528588|
|            36|                                1400614|
|            38|                                1062304|
|            14|                                 893498|
|            20|                                 618593|
+--------------+---------------------------------------+
```

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

➔ Answer:

Frequency of parking tickets with respect to 'vehicle body type':

| vehicle_body_type | number of tickets for the vehicle body type |
|---|---|
| SUBN | 3719802 |
| 4DSD | 3082020 |
| VAN | 1411970 |
| DELV | 687330 |
| SDN | 438191 |

Frequency of parking tickets with respect to 'vehicle make':

| vehicle_make | number of tickets for the vehicle make |
|---|---|
| FORD | 1280958 |
| TOYOT | 1211451 |
| HONDA | 1079238 |
| NISSA | 918590 |
| CHEVR | 714655 |

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?

    ➔ Answer:

| violation_precinct | number of tickets for the violation precinct |
|---|---|
| 0 | 2072400 |
| 19 | 535671 |
| 14 | 352450 |
| 1 | 331810 |
| 18 | 306920 |
| 114 | 296514 |

- We are not considering violation_precinct=0, since that is erroneous entry, hence displaying top 6.
- We observe that precinct 19, 14, 1, 18 and 114 have been common precinct in both the queries
- We have very high number of violations in the precinct 19 (almost double of 114)

2. 'Issuer Precinct' (This is the precinct that issued the ticket.)
   *Here, you would have noticed that the dataframe has the'Violating Precinct' or 'Issuing Precinct' as '0'. These are erroneous entries. Hence, you need to provide the records for five correct precincts.*
   *(**Hint**: Print the top six entries after sorting.)*

➔ Answer:

| issuer_precinct | number of tickets for the issuer precinct |
|---|---|
| 0 | 2388479 |
| 19 | 521513 |
| 14 | 344977 |
| 1 | 321170 |
| 18 | 296553 |
| 114 | 289950 |

- We are not considering issuer_precinct =0, since that is erroneous entry, hence displaying top 6.
- We observe that precinct 19, 14, 1, 18 and 114 have been common precinct in both the queries
- We have very high number of issuers from the precinct 19 (almost double of 114)

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?
   ***(Hint**: In the SQL view, use the 'where' attribute to filter among three precincts.)*

➔ Answer:

| violation_code | number of tickets for the issuer precinct = 19 for violation code |
|---|---|
| 46 | 86390 |
| 37 | 72437 |
| 38 | 72344 |
| 14 | 57563 |
| 21 | 54700 |
| 16 | 31353 |
| 20 | 27352 |
| 40 | 21513 |
| 71 | 15107 |
| 19 | 12896 |

| violation_code | number of tickets for the issuer precinct = 14 for violation code |
|---|---|
| 14 | 73837 |
| 69 | 58026 |
| 31 | 39857 |
| 47 | 30540 |
| 42 | 20663 |
| 46 | 13435 |
| 84 | 11111 |
| 19 | 11062 |
| 82 | 8853 |
| 17 | 6160 |

| violation_code | number of tickets for the issuer precinct = 1 for violation code |
|---|---|
| 14 | 73522 |
| 16 | 38937 |
| 20 | 27841 |
| 46 | 22534 |
| 38 | 16989 |
| 17 | 13811 |
| 37 | 13513 |
| 69 | 11165 |
| 31 | 11047 |
| 19 | 10487 |

SUMMARY:

- We do have Violation codes with higher overall frequency like Violation Code 14, 46, 38 etc.

- For issuer precinct = 19,
    - Most common violation code is 46 (double parking)
    - Other frequent violation codes are 37 and 38 (Parking Meter)

- For issuer precinct = 14,
    - Most common violation code is 14 (General No Standing)
    - Next frequent violation code is 69 (Failing to show a parking meter receipt, commercial meter zone)
    - Next frequent violation code is 31 (Standing of a non-commercial vehicle in a commercial metered zone)

- For issuer precinct = 1,
    - Most common violation code is 14 (General No Standing)
    - Next frequent violation code is 16 (Truck Loading/Unloading)
    - Next frequent violation code is 20 (General No Parking)

- Violation Code 14 (General No Standing) is common across all the three precincts

- Violation Code 46 (double parking) is also common in two of three precincts

- Violation Code 38 (Failing to show a receipt or tag in the windshield) is also common in two of three precincts

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

   - Find a way to deal with missing values, if any.
     ***(Hint**: Check for the null values using 'isNull' under the SQL. Also, to remove the null values, check the 'dropna' command in the API documentation.)*

   ➔ Answer:

   - There are 0 null values in the Dataframe
   - **We have cleaned the data for violation_time column with clean data in violation_time_new_formatted column**, however, other columns may still contain error values (including 'null', etc.) Python would fail to recognize the strings having values as "NULL" or "NA". This would be considered as non-NA value which is being considered fine for this analysis.

   - 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.

   ➔ Answer:

   Using various data cleaning and string manipulation techniques to obtain usable time field.

   - Converting to 24-hour format for easy binning
   - Also handling many kinds of faulty entries like 13:00 AM or 13:00 PM; since given time must be in 12-hour format (Assumption: Considering it a entry error and treating such values directly as 24-hour time format)
   - Assumption: Imputing extremely error values to 00:00

```
+--------------+------------------+---------------------------+
|violation_time|violation_time_new|violation_time_new_formatted|
+--------------+------------------+---------------------------+
|         0758A|          07:58:00|                     7:58:0|
|         0157P|          13:57:00|                    13:57:0|
|         0649A|          06:49:00|                     6:49:0|
|         1053A|          10:53:00|                    10:53:0|
|         1118A|          11:18:00|                    11:18:0|
+--------------+------------------+---------------------------+
```

   - 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.
     ***(Hint**: Use the CASE-WHEN in SQL view to segregate into bins. To find the most commonly occurring violations, you can use an approach similar to the one mentioned in the hint for question 4.)*

   ➔ Answer:

```
violation_time_bin="00:00 - 03:59"
+--------------+-------------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-------------------------------------------------------+
|            21|                                                 216890|
|            36|                                                 211434|
|            38|                                                 106871|
+--------------+-------------------------------------------------------+
```

```
violation_time_bin="04:00 - 07:59"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-----------------------------------------------------+
|            14|                                               141275|
|            21|                                               119466|
|            40|                                               112186|
+--------------+-----------------------------------------------------+

violation_time_bin="08:00 - 11:59"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-----------------------------------------------------+
|            21|                                              1182676|
|            36|                                               751422|
|            38|                                               346518|
+--------------+-----------------------------------------------------+

violation_time_bin="12:00 - 15:59"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-----------------------------------------------------+
|            36|                                               376961|
|            38|                                               356354|
|            37|                                               265869|
+--------------+-----------------------------------------------------+

violation_time_bin="16:00 - 19:59"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-----------------------------------------------------+
|            38|                                               203232|
|            37|                                               145784|
|            14|                                               144748|
+--------------+-----------------------------------------------------+

violation_time_bin="20:00 - 23:59"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the time bin for violation code|
+--------------+-----------------------------------------------------+
|             7|                                                65593|
|            38|                                                47029|
|            14|                                                44778|
+--------------+-----------------------------------------------------+
```

- For violation_time_bin="00:00 - 03:59", common violations were:
    - 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
    - 36 - Exceeding the posted speed limit in or near a designated school zone.
    - 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.

- For violation_time_bin="04:00 - 07:59", common violations were:
    - 14 - General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.
    - 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
    - 40 - Stopping, standing or parking closer than 15 feet of a fire hydrant. Between sunrise and sunset, a passenger vehicle may stand alongside a fire hydrant as long as a driver remains behind the wheel and is ready to move the vehicle if required to do so.

- For violation_time_bin="08:00 - 11:59", common violations were:
    - 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
    - 36 - Exceeding the posted speed limit in or near a designated school zone.
    - 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.

- For violation_time_bin="12:00 - 15:59", common violations were:
    - 36 - Exceeding the posted speed limit in or near a designated school zone.
    - 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.
    - 37 - Parking in excess of the allowed time

- For violation_time_bin="16:00 - 19:59", common violations were:
    - 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.
    - 37 - Parking in excess of the allowed time
    - 14 - General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.

- For violation_time_bin="20:00 - 23:59", common violations were:
    - 7 - Vehicles photographed going through a red light at an intersection.
    - 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.
    - 14 - General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device.

- 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:

```
violation_code=38
+------------------+-----------------------------------------------------+
|violation_time_bin|number of tickets for the time bin for violation code|
+------------------+-----------------------------------------------------+
|     12:00 - 15:59|                                               356354|
+------------------+-----------------------------------------------------+

violation_code=21
+------------------+-----------------------------------------------------+
|violation_time_bin|number of tickets for the time bin for violation code|
+------------------+-----------------------------------------------------+
|     08:00 - 11:59|                                              1182676|
+------------------+-----------------------------------------------------+

violation_code=36
+------------------+-----------------------------------------------------+
|violation_time_bin|number of tickets for the time bin for violation code|
+------------------+-----------------------------------------------------+
|     08:00 - 11:59|                                               751422|
+------------------+-----------------------------------------------------+
```

- Based on above observations, our most common 3 codes were 38, 21, 36
- For violation code 38, most common time of the day is "12:00 - 15:59" i.e. Afternoon
- For violation code 21, most common time of the day is "08:00 - 11:59" i.e. Night
- For violation code 36, most common time of the day is "08:00 - 11:59" i.e. Night

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. *(**Hint**: Use Issue Date to segregate into seasons.)*

   ➔ Answer:

   Using data available at https://www.nyc.com/visitor_guide/weather_facts.75835/

   - Winter: December, January, February
   - Spring: March, April, May
   - Summer: June, July, August
   - Fall: September, October, November

   - Then, find the three most common violations for each of these seasons.
     *(**Hint**: You can use an approach similar to the one mentioned in the hint for question 4.)*

   ➔ Answer:

```
season="Winter"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the season for violation code|
+--------------+-----------------------------------------------------+
|            21|                                               362341|
|            36|                                               359338|
|            38|                                               259723|
+--------------+-----------------------------------------------------+

season="Spring"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the season for violation code|
+--------------+-----------------------------------------------------+
|            21|                                               402807|
|            36|                                               344834|
|            38|                                               271192|
+--------------+-----------------------------------------------------+

season="Summer"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the season for violation code|
+--------------+-----------------------------------------------------+
|            21|                                               405961|
|            38|                                               247561|
|            36|                                               240396|
+--------------+-----------------------------------------------------+
```

```
season="Fall"
+--------------+-----------------------------------------------------+
|violation_code|number of tickets for the season for violation code|
+--------------+-----------------------------------------------------+
|            36|                                               456046|
|            21|                                               357479|
|            38|                                               283828|
+--------------+-----------------------------------------------------+
```

For season="**Winter**", the following violation codes were most frequent:

- 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
- 36 - Exceeding the posted speed limit in or near a designated school zone.
- 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.

For season="**Spring**", the following violation codes were most frequent:

- 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
- 36 - Exceeding the posted speed limit in or near a designated school zone.
- 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.

For season="**Summer**", the following violation codes were most frequent:

- 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
- 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.
- 36 - Exceeding the posted speed limit in or near a designated school zone.

For season="**Fall**", the following violation codes were most frequent:

- 36 - Exceeding the posted speed limit in or near a designated school zone.
- 21 - Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.
- 38 - Failing to show a receipt or tag in the windshield. Drivers get a 5-minute grace period past the expired time on parking meter receipts.
- For all the seasons, the top 3 violation codes are same with some change in order over seasons.

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.

    ➔ Answer:

```
+--------------+---------------------------------------+
|violation_code|number of tickets for the violation code|
+--------------+---------------------------------------+
|            21|                                1528588|
|            36|                                1400614|
|            38|                                1062304|
+--------------+---------------------------------------+
```

- 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.

➔ Answer:

```
Revenue generation from violation code 21 (count:1528588 , avg. fine: $
55) = $84072340
Revenue generation from violation code 36 (count:1400614 , avg. fine: $
50) = $70030700
Revenue generation from violation code 38 (count:1062304 , avg. fine: $
50) = $53115200
```

- 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.

➔ Answer:

- Revenue generation from violation code 21 (count: 1528588, avg. fine: $55) = $84072340
- Revenue generation from violation code 36 (count: 1400614, avg. fine: $50) = $70030700
- Revenue generation from violation code 38 (count: 1062304, avg. fine: $50) = $53115200
- **Violation Code 21 has the highest total collection of $84072340**

- What can you intuitively infer from these findings?

➔ Answer:

We can intuitively infer the following from these findings:

- Most violations are due to **parking related issues**.
- There is certainly **parking related problem in NYC**.
- Many people are **violating the posted speed limit**.