***Solution to “NYC Parking Tickets: An Exploratory Analysis”***

**Assumptions –**

1. We are considering all records belong to Fiscal\_Year\_2017.
2. Assuming the time format of “**Violation Time**” as first 2 character as hour, next 2 char is Minutes , next 1 char is A or P

A stands for AM

P stands for PM

Hour is in 1-12

Minute is in 0-59

All other records that are not compliant with above assumption will be filtered out.

1. **Discrete bins of time** -24 hour is divided in bins using below consideration –

We divided the time in six bins (1,2,3,4,5,6).

0-4 -> 1

4-8 -> 2

8-12 -> 3

12-16 -> 4

16-20 -> 5

20-24 -> 6

1. **Number of seasons -**  There will be four seasons in a year

1. Spring - runs from March 1 to May 31;

2. Summer - runs from June 1 to August 31;

3. Autumn - runs from September 1 to November 30

4. Winter - runs from December 1 to February 28 (February 29 in a leap year).

We will neglect the time to calculate season. we will only consider day and month to calculate season

Issue data is in yyyy-MM-dd format

1. **Charges for violation code** –

It is calculated by averaging out the two different charges given as per the location.

**Questions –**

**Examine the data**

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

**Answer**- Total tickets for the year is **10803028**

1. 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-** Number of unique states from where the cars got parking tickets came is **66**.

**Aggregation tasks**

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

**Answer-**

+--------------+-------------------+

|Violation Code|freq\_violation\_code|

+--------------+-------------------+

| 21| 1528588|

| 36| 1400614|

| 38| 1062304|

| 14| 893498|

| 20| 618593|

+--------------+-------------------+

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

**Answer-**

Parking ticket count as per **Vehicle Body Type**

+-----------------+------------------+

|Vehicle Body Type|freq\_veh\_body\_type|

+-----------------+------------------+

| SUBN| 3719802|

| 4DSD| 3082020|

| VAN| 1411970|

| DELV| 687330|

| SDN| 438191|

+-----------------+------------------+

Parking ticket count as per **Vehicle Make**

+------------+-------------+

|Vehicle Make|freq\_veh\_make|

+------------+-------------+

| FORD| 1280958|

| TOYOT| 1211451|

| HONDA| 1079238|

| NISSA| 918590|

| CHEVR| 714655|

+------------+-------------+

1. 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:
2. '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**- We see that Most parking violations is happening in ‘19’ violation precinct

and top 5 Violation Precincts are 19, 14, 1, 18, 114.

+------------------+------------------+

|Violation Precinct|freq\_voilation\_pre|

+------------------+------------------+

| 19| 535671|

| 14| 352450|

| 1| 331810|

| 18| 306920|

| 114| 296514|

+------------------+------------------+

1. '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** – we see that top 5 Issuer Precinct are 19, 14, 1, 18, 114

+---------------+---------------+

|Issuer Precinct|freq\_issuer\_pre|

+---------------+---------------+

| 19| 521513|

| 14| 344977|

| 1| 321170|

| 18| 296553|

| 114| 289950|

+---------------+---------------+

**Conclusion** – **Precincts (19,14,1,18,114) are common for both type of precincts according to no of tickets issued.**

1. 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** – From the previous questions answers we got three precincts that have issued the most number of tickets are – 19,14,1

Violation precinct frequency in Violation Precinct –

+------------------+--------------+-------------------+

|Violation Precinct|Violation Code|freq\_voilation\_code|

+------------------+--------------+-------------------+

| 19| 46| 90530|

| 1| 14| 76375|

| 14| 14| 75850|

| 19| 38| 74926|

| 19| 37| 73359|

+------------------+--------------+-------------------+

Violation precinct frequency in Issuer Precinct –

+---------------+--------------+-------------------+

|Issuer Precinct|Violation Code|freq\_voilation\_code|

+---------------+--------------+-------------------+

| 19| 46| 86390|

| 14| 14| 73837|

| 1| 14| 73522|

| 19| 37| 72437|

| 19| 38| 72344|

+---------------+--------------+-------------------+

**Insights from the results** –

1. Issuer precinct zone -19 has exceptional no of violations happening.
2. Violation precinct zone - 19 has exceptional no of violations happening
3. Most of the violation codes (like 46,14,37,38) are common across precinct.
4. 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.)

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.

**Question**-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.)

**Answe**r –

We divided the time in Six slots as per below –

Hours between 0-4 -> 1

Hours between 4-8 -> 2

Hours between 8-12 -> 3

Hours between 12-16 -> 4

Hours between 16-20 -> 5

Hours between 20-24 -> 6

Violation time is converted to Human readable 24hr format.

+--------------+--------------+

|Violation Time|converted\_time|

+--------------+--------------+

| 0143A| 01:43|

| 0400P| 16:00|

| 0233P| 14:33|

| 1120A| 11:20|

| 0555P| 17:55|

| 0852P| 20:52|

| 0215A| 02:15|

+--------------+--------------+

Three most commonly occurring violations for **hour\_bin=1**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 1| 21| 53600|

| 1| 40| 44737|

| 1| 78| 28716|

+--------+--------------+--------------------+

Three most commonly occurring violations for **hour\_bin=2**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 2| 14| 141275|

| 2| 21| 119466|

| 2| 40| 112186|

+--------+--------------+--------------------+

Three most commonly occurring violations for **hour\_bin=3**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 3| 21| 1182676|

| 3| 36| 751422|

| 3| 38| 346518|

+--------+--------------+--------------------+

Three most commonly occurring violations for **hour\_bin=4**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 4| 36| 588395|

| 4| 38| 462756|

| 4| 37| 337074|

+--------+--------------+--------------------+

Three most commonly occurring violations for **hour\_bin=5**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 5| 38| 203232|

| 5| 37| 145784|

| 5| 14| 144748|

+--------+--------------+--------------------+

Three most commonly occurring violations for **hour\_bin=6**

+--------+--------------+--------------------+

|hour\_bin|Violation Code|count\_violation\_code|

+--------+--------------+--------------------+

| 6| 7| 65593|

| 6| 38| 47029|

| 6| 14| 44778|

+--------+--------------+--------------------+

**Question** - 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 –** Three most commonly occurring violation codes are- **21,36,38**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 21| 1504663|

| 36| 1400614|

| 38| 1062287|

| 14| 885469|

| 20| 614453|

+--------------+--------------------+

**Most common time of the day when violation is happening – 3** (we can see it in below table )

+--------+--------------+

|hour\_bin|count\_hour\_bin|

+--------+--------------+

| 3| 2280616|

| 4| 1199158|

| 5| 230641|

| 2| 155705|

| 1| 54052|

| 6| 47392|

+--------+--------------+

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

**Question** -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** – we assume that there will be four seasons in a year

1. Spring - runs from March 1 to May 31;

2. Summer - runs from June 1 to August 31;

3. Autumn - runs from September 1 to November 30

4. Winter - runs from December 1 to February 28 (February 29 in a leap year).

Date is divided accordingly.

Frequencies of tickets for each season is as below –

+------+-----------------------+

|season|ticket\_count\_per\_season|

+------+-----------------------+

|Spring| 2880687|

|Autumn| 2830802|

|Summer| 2606208|

|Winter| 2485331|

+------+-----------------------+

**Question** - 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**-

Three most common violations for **Spring**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 21| 402807|

| 36| 344834|

| 38| 271192|

+--------------+--------------------+

Three most common violations for **Autumn**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 36| 456046|

| 21| 357479|

| 38| 283828|

+--------------+--------------------+

Three most common violations for **Summer**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 21| 405961|

| 38| 247561|

| 36| 240396|

+--------------+--------------------+

Three most common violations for **Winter**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 21| 362341|

| 36| 359338|

| 38| 259723|

+--------------+--------------------+

1. 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:

**Question - Find the total occurrences of the three most common violation codes.**

**Answer**- Three most common violation codes are - **21, 36, 38**

+--------------+--------------------+

|Violation Code|count\_violation\_code|

+--------------+--------------------+

| 21| 1528588|

| 36| 1400614|

| 38| 1062304|

| 14| 893498|

| 20| 618593|

+--------------+--------------------+

**Question** - 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** – Average Fines are calculated using given link in the question.

**Below is the fine for codes 21,36,38**

code 21 - charges((65+45)/2=55)

code 36 - charges((50+50)/2=50)

code 38 - charges((65+35)/2=50)

Revenue generated from fine of above violation codes –

+--------------+--------+

|Violation Code| Revenue|

+--------------+--------+

| 21|84072340|

| 36|70030700|

| 38|53115200|

+--------------+--------+

**Finding**- Most fine is collected from “**No parking where parking is not allowed by sign, street marking or traffic control device” ie Violation Code 21.**

#End of Solution