Examine the data

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

   Assumption: We have filtered data for Issue Date. We have considered records which have Year of Issue Date as 2017 as stated in the question. This assumption is used throughout the assignment.

   We have **5431918** tickets issued.

2. **Find out the number of unique states from where the cars that got parking tickets came. *(Hint: Use the column 'Registration State'.)***

   Total unique states = 65

   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.

   Total unique states after numeric entry ‘99’ is handled = 64.

   (State wise count shown in notebook).

Aggregation tasks:

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

   Following is the Counts for top 5 violation codes.

Violation Code	Count
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 type with ticket count

V_body_Type	Count
SUBN	1883954
4DSD	1547312
VAN	724029
DELV	358984
SDN	194197

Top 5-vehicle make with ticket count

V_make	Count
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: '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?**

Following are the top violating precincts. We can exclude 0 as it is erroneous entry.

Violation_precinct	Count
0	925596
19	274445
14	203553
1	174702
18	169131
114	147444

'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.)

Following are the top issuing precincts. We can exclude 0 as it is erroneous entry.

Issuer_precinct	Count
0	1078406
19	266961
14	200495
1	168740
18	162994
114	144054

We can see that both for Violating Precincts and Issuer precincts top 5 are the same.

Top 5 Violating and Issuer precincts are **19, 14, 1, 18, 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.)

   Top 3 precincts are 19, 14, 1. So, count for top 5 violation codes for these precincts are:

Violation_code	Issuer_precinct	count	fraction_of_total
46	19	48445	18.15
38	19	36386	13.63
37	19	36056	13.51
14	19	29797	11.16
21	19	28415	10.64

Violation_code	Issuer_precinct	count	fraction_of_total
14	14	45036	22.46
69	14	30464	15.19
31	14	22555	11.25
47	14	18364	9.16
42	14	10027	5.0

Violation_code	Issuer_precinct	count	fraction_of_total
14	1	38354	22.73
16	1	19081	11.31
20	1	15408	9.13
46	1	12745	7.55
38	1	8535	5.06

Observations:

1. Overall top violation codes for all precincts as seen from query done earlier is (21, 36, 38, 14, 20) in this order desc.
2. For precinct 19, top 5 violation codes (46, 38, 37, 14, 21) in given order accounts for 67.09% of total tickets issues by this precinct. Here 46, 38 and 37, 14 have slightly higher fractions than others.
3. For precinct 14, top 5 violation codes (14, 69, 31, 47, 42) in given order accounts for 63.06% of total tickets issues by this precinct. Here 14, 69 and 31 have slightly higher fractions than others.
4. For precinct 1, top 5 violation codes (14, 16, 20, 46, 38) in given order accounts for 55.78% of total tickets issues by this precinct. Here 14 and 16 have slightly higher fractions than others.

Violation code 14 is in top 5 of all three precincts 19, 14 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.

(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.)

We have checked and there are no nulls in the Violation time.

We can see that there are some records (16) which are having NaN values. We have dropped these records.

We have also dropped any records which is inconsistent with the given format of time in the assignment.

Format: HHMMF

HH : Hour of day
MM: Minutes of day
F : Flag for AM or PM (A/P)

We have taken valid format as HH < 12 and F in (‘A’, ‘P’) as we only need hour of day. Based on this condition there are 63 inconsistent records. We have dropped them as well as it is very low fraction considering overall records count.

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.

We have used a case statement to divide time based on hour and AM PM flag into 6 different groups.

Bucket Definitions:

Time Slot	Time Group
12 AM – 4 AM	Late Night
4 AM – 8 AM	Early Morning
8 AM – 12 PM	Morning
12 PM – 4 PM	Afternoon
4 PM- 8 PM	Evening
8 PM - 12 AM	Night

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.)

Violation_Time_Of_Day	Violation_code	Count
Early_Morning	14	74114
Early_Morning	40	60652
Early_Morning	21	57897

Violation_Time_Of_Day	Violation_code	Count
Morning	21	598069
Morning	36	348165
Morning	38	176570

Violation_Time_Of_Day	Violation_code	Count
After_Noon	36	286284
After_Noon	38	240721
After_Noon	37	167026

Violation_Time_Of_Day	Violation_code	Count
Evening	38	102855
Evening	14	75902
Evening	37	70345

Violation_Time_Of_Day	Violation_code	Count
Night	7	26293
Night	40	22337
Night	14	21045

Violation_Time_Of_Day	Violation_code	Count
Late_Night	21	36958
Late_Night	40	25867
Late_Night	78	15528

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).

Violation_Time_Of_Day	Count
Morning	1122804
After_Noon	601700
Evening	116648
Early_Morning	73952
Late_Night	37270
Night	20531

Highest number of violations are happening during Morning Hours, which are 8-11 AM.

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.)

This data is for New York. Hence defining seasons based on following:

1. Winter: December, January, February
2. Spring: March, April, May
3. Summer: June, July, August
4. Fall: September, October, November

Violation_Season	Count
Spring	2873337
Winter	1704669
Summer	852854
Fall	979

We have very few violation records for Fall Months.

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.)

Violation_Season	Violation_code	Count
Spring	21	402407
Spring	36	344834
Spring	38	271167

Violation_Season	Violation_code	Count
Winter	21	238180
Winter	36	221268
Winter	38	187385

Violation_Season	Violation_code	Count
Summer	21	127347
Summer	36	96663
Summer	38	83518

Violation_Season	Violation_code	Count
Fall	46	231
Fall	21	128
Fall	40	116

Top Violation codes are consistent for Spring, Summer and Winter and are 21, 36, 38 in descending order. Whereas during fall season top violation codes are 46, 21, 40 in descending order.

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.

Violation_code	Count
21	768062
36	662765
38	542078

These are the top 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.

Top 3 violation codes with descriptions and average fine amounts are:

Violation_code	Violation Code Description	Avg. Fine Amount
21	Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.	$55
36	Exceeding the posted speed limit in or near a designated school zone.	$50
38	Failing to show a receipt or tag in the windshield.	$50

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.

Violation_code	Violation Code Description	Parking Fine Per Violation	Count	Total_Parking_Fine
21	Street Cleaning: No parking where parking is not allowed by sign, street marking or traffic control device.	$55	768062	$42,243,410
36	Exceeding the posted speed limit in or near a designated school zone.	$50	662765	$33,138,250
38	Failing to show a receipt or tag in the windshield.	$50	542078	$27,103,900

What can you intuitively infer from these findings?

Inferences:

1. Most commonly occurring parking violations are 21, 36 and 38.
2. Traffic violations are highest in Spring Months (March, April May) followed by Wi nter Months (December, January, February).
3. Traffic violations are highest during Morning Hours followed by After-noon hours.
4. Traffic violations are highest in precincts 19, 14, 1.