Spark Group Assignment

BNS Analytics:
1.Biswajit Dash
2.Sanyam Jain
3.Beltur Rangarajan Navneet

Examine the data

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

   5431918

2. Find out the number of unique states from where the cars that got parking tickets came from. *(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 maximum entries. Give the number of unique states again.

   64

Aggregation tasks

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

   Per Day

	Violation Code	Count_per_day
1	21	2104.3479
2	36	1815.7945
3	38	1485.1479
4	14	1305.9288
5	20	875.7425

   Per Month

	Violation Code	Count_per_mnth
1	21	64007.25
2	36	55230.42
3	38	45173.25
4	14	39722.00
5	20	26637.17

   Per Year

	Violation Code	Count_in_2017
1	21	768087
2	36	662765
3	38	542079
4	14	476664
5	20	319646

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

Body Type
Per Year

#	Vehicle Body Type	Count_in_2017
# 1	SUBN	1883954
# 2	4DSD	1547312
# 3	VAN	724029
# 4	DELV	358984
# 5	SDN	194197

Per Day

#	Vehicle_Body_Type	Count_per_day
#1	SUBN	5161.5178
#2	4DSD	4239.2110
#3	VAN	1983.6411
#4	DELV	983.5178
#5	SDN	532.0466

Per Month

#	Vehicle_Body_Type	Count_per_mnth
#1	SUBN	156996.17
#2	4DSD	128942.67
#3	VAN	60335.75
#4	DELV	29915.33
#5	SDN	16183.08

Vehicle Make
Per Year

#	Vehicle Make	Count_in_2017
# 1	FORD	636844
# 2	TOYOT	605291
# 3	HONDA	538884
# 4	NISSA	462017
# 5	CHEVR	356032

Per Day

#	Vehicle_Make	Count_per_day
#1	FORD	1744.7781
#2	TOYOT	1658.3315
#3	HONDA	1476.3945
#4	NISSA	1265.8000
#5	CHEVR	975.4301

Per Month
#Vehicle_Make Count_per_mnth
#1 FORD 53070.33
#2 TOYOT 50440.92
#3 HONDA 44907.00
#4 NISSA 38501.42
#5 CHEVR 29669.33

3. A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequency 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 make 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 dataframe has 'Violating Precinct' or 'Issuing Precinct' as '0'. These are the erroneous entries. Hence, provide the record for five correct precincts. (**Hint**: Print top six entries after sorting)*

Violation Precinct

Per Day
#Violation_Precinct Count_per_day
#1 0 2535.8795
#2 19 751.9041
#3 14 557.6795
#4 1 478.6356
#5 18 463.3726
#6 114 403.9562

 # Top 5 Highest Violation Precinct are 19 , 14, 1, 18 & 114 (Not considering 0 Violation Precinct as it is erraneous)

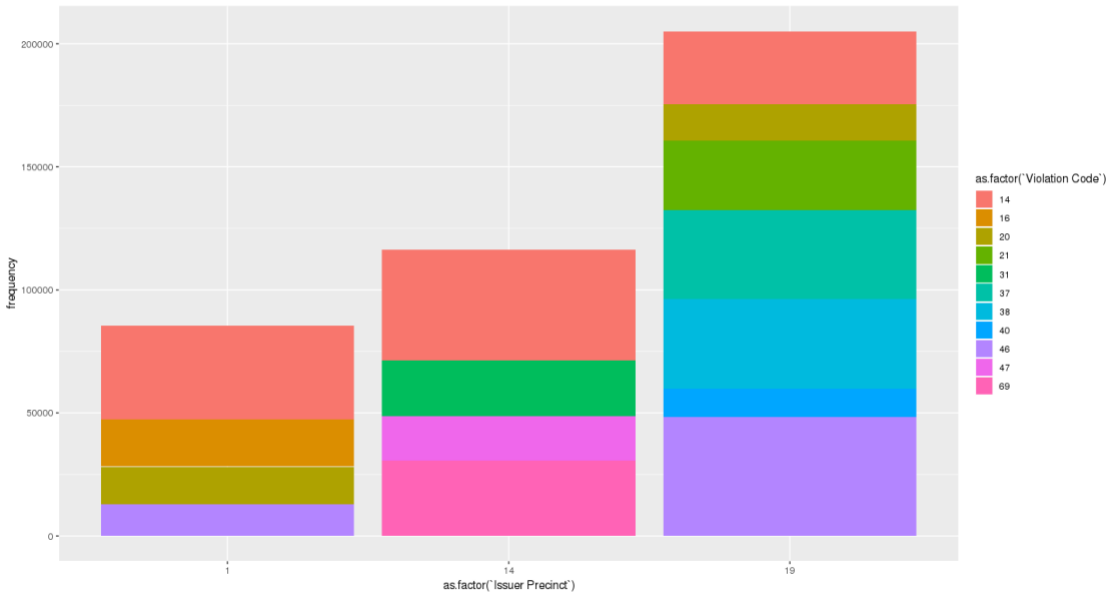
Issuer Precinct

Per Day
#Issuer_Precinct Count_per_day
#1 0 2954.5370
#2 19 731.4000
#3 14 549.3014
#4 1 462.3014
#5 18 446.5589
#6 114 394.6685

 # Top 5 Highest Issues Precinct are 19 , 14, 1, 18 & 114 (Not considering 0 Issues Precinct as it is erraneous)

4. Find the violation code frequency across three precincts which 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*

   # Issuer_Precinct Violation_Code frequency
   #1 19 46 48445
   #2 14 14 45036
   #3 1 14 38354
   #4 19 38 36386
   #5 19 37 36056
   #6 14 69 30464



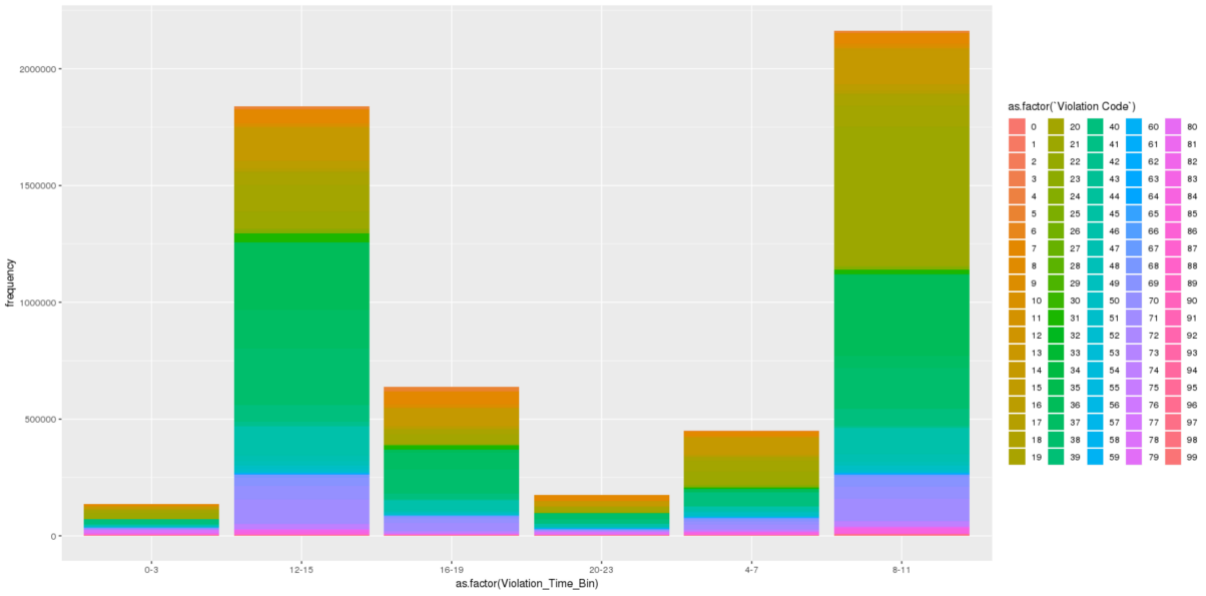
   # Zone 1 and 14 has highest Violation code 14
   # Violation code 14 is the most common across zone 19,14,1 , which are the highest Issues Precinct.

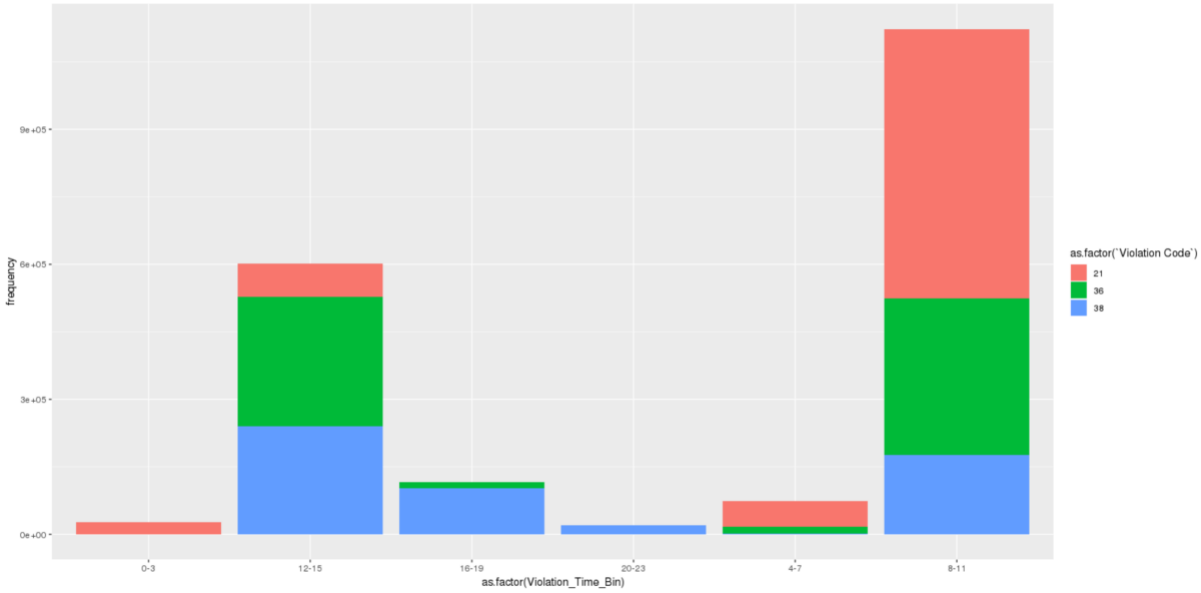
5. You'd want to find out the properties of parking violations across different times of the day:
   - Find a way to deal with missing values, if any.
     ***Hint**: Check for the null values using 'isNull' under the SQL. Also, to remove the null values, check the 'dropna' command in the API documentation.*
   - The Violation Time field is specified in a strange format. Find a way to make this into a time attribute that you can use to divide into groups.
   - Divide 24 hours into six equal discrete bins of time. The intervals you choose are at your discretion. For each of these groups, find the three most commonly occurring violations.
     ***Hint**: Use the CASE-WHEN in SQL view to segregate into bins. For finding the most commonly occurring violations, a similar approach can be used as mention in the hint for question 4.*

- 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)

Missing value analysis
#There are no NULL Values

Violation Time analysis
Steps Followed:
#Correcting Violation_Hour of 00 AM to 12 AM.
 # Correcting Violation Time : Concating Violation_Hour, Violation_Minute & Violation_AM_PM
 # Converting Violation time to time stamp
#Check for NULL Violation time.
 # Dropping Null values
 # Extracting Hours for bucketing from Violation Time
 # Bucketing Violation_Hour in 6 bins

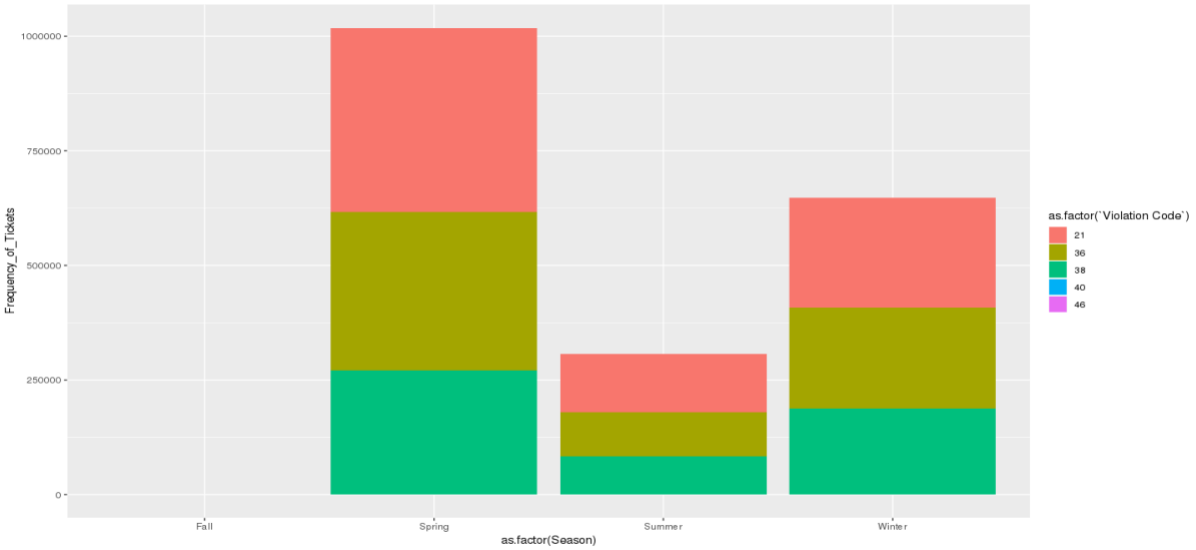




Observations:
#Most commonly Violation are between 8 to 11 and 12 to 15 bins.
#Most commonly Violation codes are 21, 36 and 38.

6. Let's try and find some seasonality in this data
   - First, divide the year into some number of seasons, and find frequencies of tickets for each season. *(**Hint**: Use Issue Date to segregate into seasons)*
   - Then, find the three most common violations for each of these seasons. *(**Hint**: A similar approach can be used as mention in the hint for question 4.)*

#	Season	Violation Code	Frequency_of_Tickets
# 1	Fall	46	231
# 2	Fall	21	128
# 3	Fall	40	116
# 4	Spring	21	402399
# 5	Spring	36	344834
# 6	Spring	38	271167
# 7	Summer	21	127344
# 8	Summer	36	96663
# 9	Summer	38	83518
# 10	Winter	21	238179
# 11	Winter	36	221268
# 12	Winter	38	187385



7. The fines collected from all the parking violation constitute a revenue source for the NYC police department. Let's take an example of estimating that for the three most commonly occurring codes.
    - Find 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 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 the three violation codes with maximum tickets. State the code which has the highest total collection.
    - What can you intuitively infer from these findings?

The three most common violation codes are 21,36,38.

Calculating Fine
violation code 21
highest-density locations of the city = $65
rest of the city = $45
#--- Average fine = ($65 + $45)/2 = $110/2 = $55 --#

violation code 36
highest-density locations of the city = $50
rest of the city = $50
#--- Average fine = ($50 + $50)/2 = $100/2 = $50 --#

violation code 38
highest-density locations of the city = $65
rest of the city = $35
#--- Average fine = ($65 + $35)/2 = $100/2 = $50 --#

	Violation.Code	violation_frq	fine	collection
1	21	768087	55	42244785
2	36	662765	50	33138250
3	38	542079	50	27103950

Violation code 21 has highest number of frequencies = 768087
Violation code 21 has highest total collection of 42244785