NYC Parking Case Study: Apache Spark

Contents

NYC Parking Case Study: Apache Spark

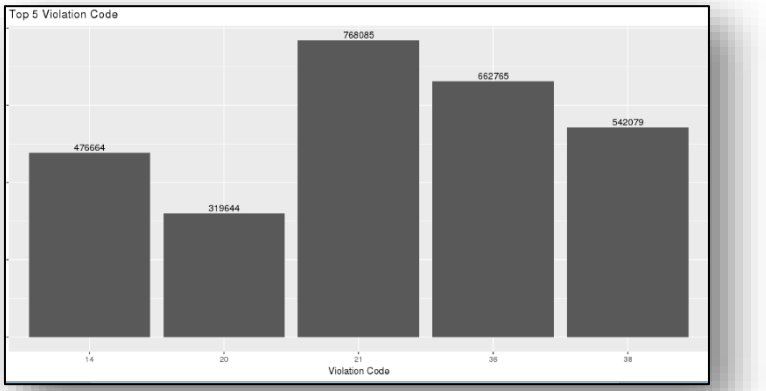
Examine the Data

Question 1 : Find the total number of tickets for the year.	Result : 5431909 Tickets were issued in the year 2017.
Question 2 : Find out the number of unique states from where the cars that got parking tickets came from	16055 rows are having with value "99". *There is a numeric entry '99' in the column which should be corrected. Replaced it with the state having maximum entries NY* NY is having maximum rows -> 4273944. The unique states are 64

Aggregation tasks

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

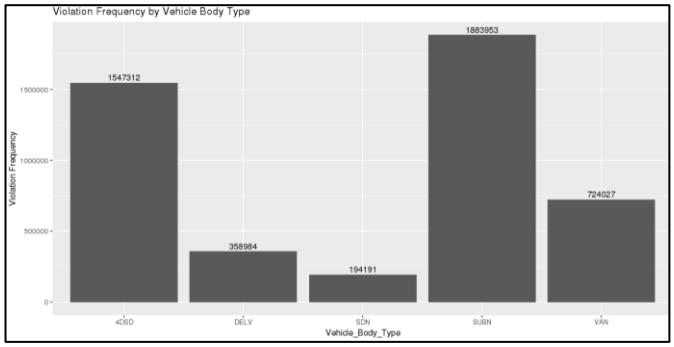
Order	Violation Code	Frequency
1	21	768085
2	36	662765
3	38	542079
4	14	476664
5	20	319644
6	46	312327



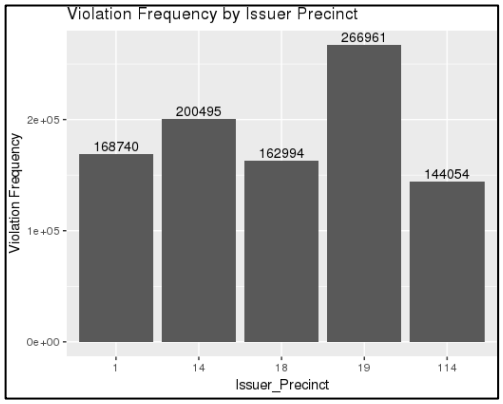
Question 2 : How often does each 'vehicle body type' get a parking ticket? How about the 'vehicle make'?

Question 2.1 : Parking Tickets by Vehicle Body Type

```
  Vehicle_Body_Type   Frequency
1              SUBN 1883953
2              4DSD 1547312
3               VAN  724027
4              DELV  358984
5               SDN  194191
```



Question 2.2 : Parking Tickets by Vehicle Make Type

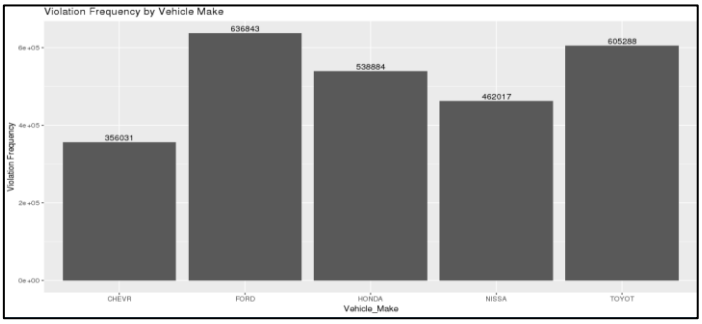


	Vehicle_Make	count
1	FORD	636843
2	TOYOT	605288
3	HONDA	538884
4	NISSA	462017
5	CHEVR	356031

Question 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: Here you would have noticed that the dataframe has 'Violating Precinct' or 'Issuing Precinct' as '0'. These are the erroneous entries

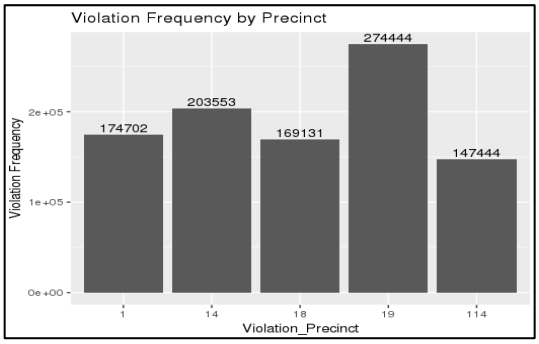
Question 3.1 : Violation Precinct' (this is the precinct of the zone where the violation occurred). Using this, can you make any insights for parking violations in any specific areas of the city?

	Violation_Precinct	Violation
1	19	274444
2	14	203553
3	1	174702
4	18	169131
5	114	147444



Question 3.2 : 'Issuer Precinct' (this is the precinct that issued the ticket)

	Issuer_Precinct	count
1	19	266961
2	14	200495
3	1	168740
4	18	162994
5	114	144054



Question 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?

Top 20 by total frequency

	Violation_Code	Top1_Precinct	Top2_Precinct	Top3_Precinct	totalFreq
1	14	29797	45036	38354	113187
2	46	48445	7679	12745	68869
3	38	36386	3269	8535	48190
4	37	36056	1256	6470	43782
5	69	2910	30464	5672	39046
6	21	28415	1029	4055	33499
7	20	14629	2761	15408	32798
8	31	2080	22555	5853	30488
9	16	9926	940	19081	29947
10	40	11416	3582	4592	19590
11	19	6856	7031	5375	19262
12	47	702	18364	32	19098
13	84	4910	6743	3310	14963
14	71	7493	2757	3581	13831
15	42	903	10027	2708	13638
16	17	1464	3534	7526	12524
17	10	5643	1319	4712	11674
18	70	4459	1461	2183	8103
19	82	888	5052	775	6715
20	48	1460	2439	1907	5806

- Top 3 IssuerPrecinct are 19 (`Top1_Precinct`), 14 (`Top2_Precinct`) and 1 (`Top3_Precinct`).
    - Precinct 14 has the most violations by issuer precinct
- Precinct 19 - Top 5 Violation codes are 46, 36, 37, 14 and 21
- Precinct 14 - Top 5 Violation codes are 14, 69, 31, 47 and 42
- Precinct 1 - Top 5 Violation codes are 14, 16,20, 46 and 38

Figure 1 -> Precinct 19 - Top 5 Violation codes are 46, 36, 37, 14 and 21

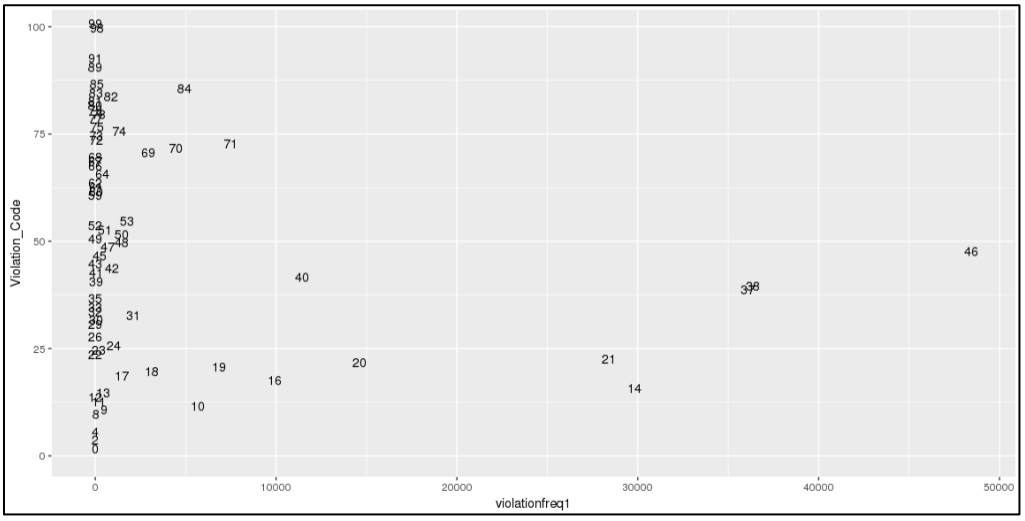


Figure 2-> Precinct 14 - Top 5 Violation codes are 14, 69, 31, 47 and 42

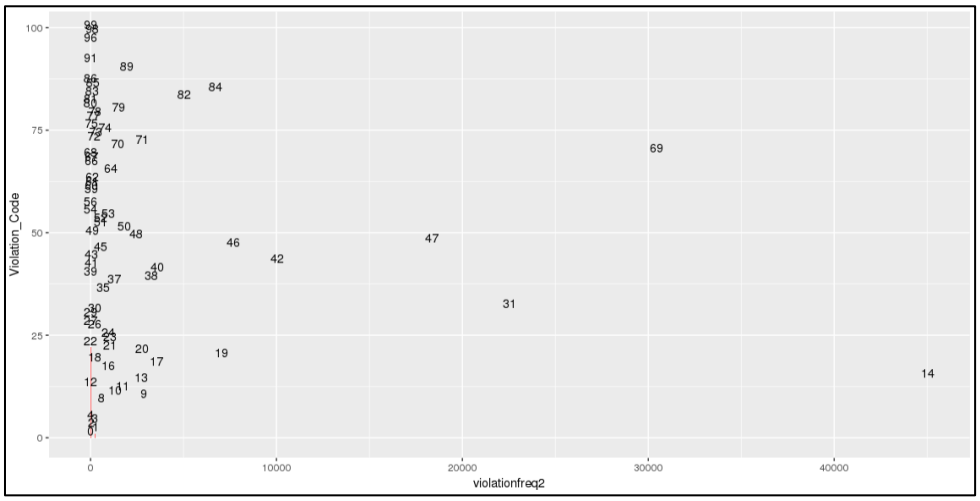
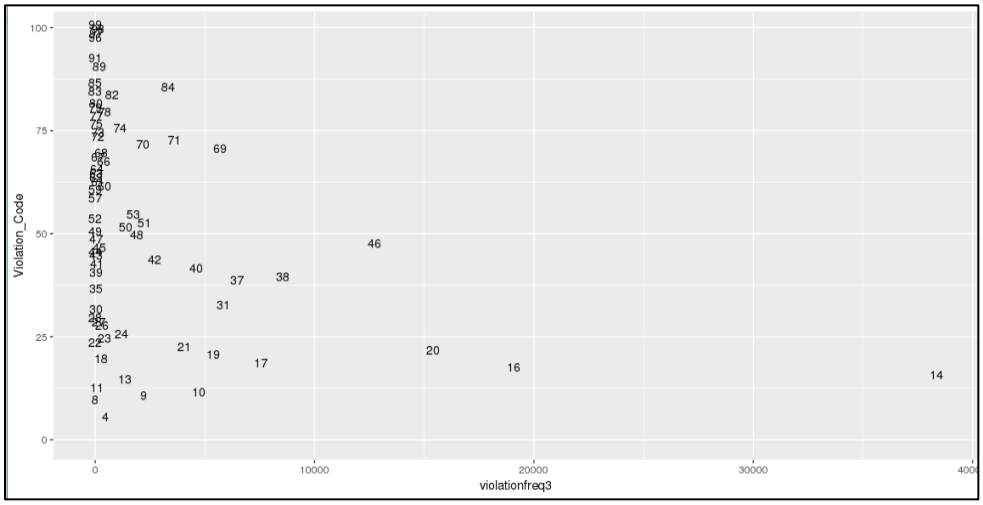


Figure 3 -> Precinct 1 - Top 5 Violation codes are 14, 16,20, 46 and 38



Question 5 : You'd want to find out the properties of parking violations across different times of the day

Question 5.1 : Find a way to deal with missing values, if any.

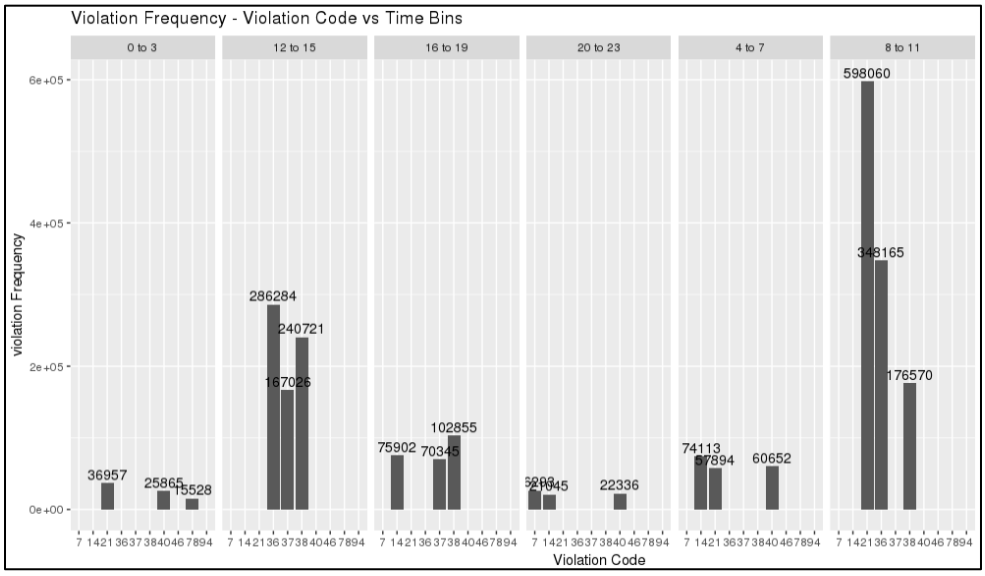
- ✓ 0 records with Null Issue Date as we already removed NA values using nycpark_2017<-na.omit(nycpark_2017)

Question 5.2 : 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.

- ✓ Violation Time is in text format HH, MM and AM/PM format but We have time as 00??A, need to convert into 12??A

Question 5.3 : 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.

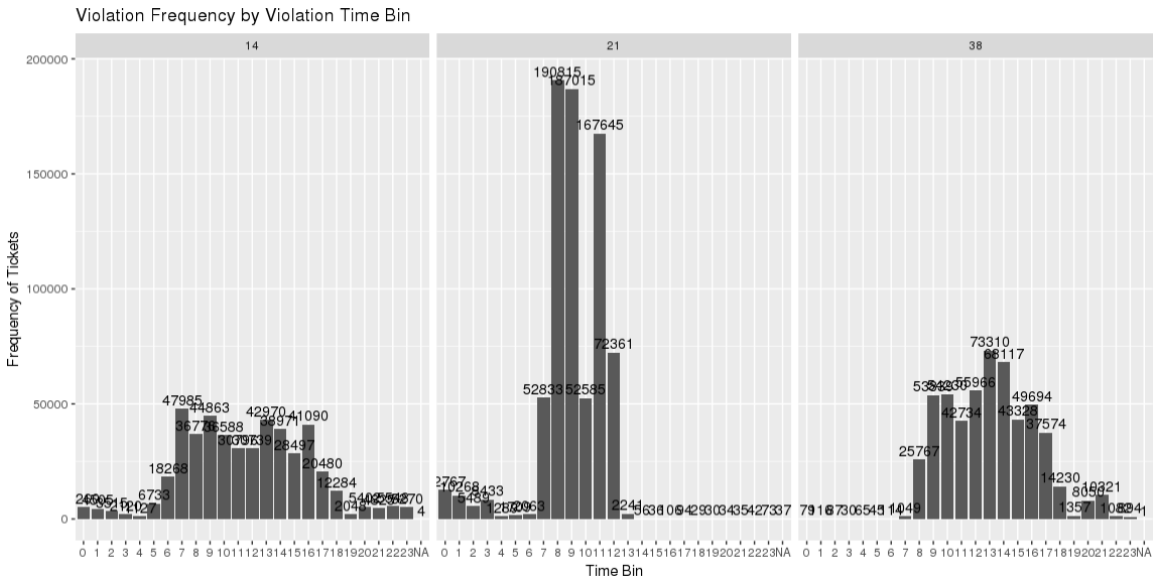
	ViolationtimeBins	Violation_Code	violationFrq
1	16 to 19	38	102855
2	16 to 19	14	75902
3	16 to 19	37	70345
4	12 to 15	36	286284
5	12 to 15	38	240721
6	12 to 15	37	167026
10	4 to 7	14	74113
11	4 to 7	40	60652
12	4 to 7	21	57894
13	0 to 3	21	36957
14	0 to 3	40	25865
15	0 to 3	78	15528
16	20 to 23	7	26293
17	20 to 23	40	22336
18	20 to 23	14	21045
19	8 to 11	21	598060
20	8 to 11	36	348165
21	8 to 11	38	176570



Question 5.4 : 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)

- ✓ Top 3 Violation Codes are 21 with 768085 violations, 36 with 662765 violations and 38 with 542079 violations

Figure 4->Most of the violations are during the day time



Question 6 : Let's try and find some seasonality in this data

Question 6.1 : First, divide the year into some number of seasons, and find frequencies of tickets for each season

- ✓ Seasons will be categorised as below and with Results::
  - Summer: June to August.
  - Autumn: September to November
  - Winter: December to February
  - Spring: March to May

Violation Code by Season and Count of Summons Number

	Summons_Number	Violation_Code	Season
1	8478629828	47	Summer
2	5096917368	7	Summer
3	1407740258	78	Winter
4	1413656420	40	Winter
5	8480309064	64	Winter
6	1416638830	20	Spring

Count of Violation by Season

	Season	Tickets
1	Spring	2873383
2	Winter	1704681
3	Summer	852866
4	Autumn	979

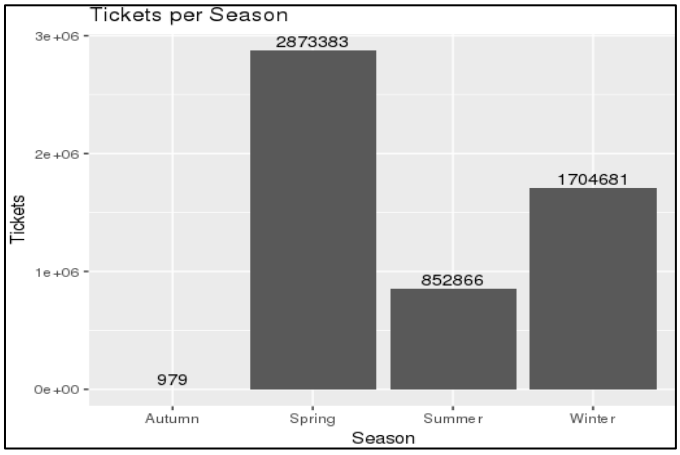


Figure 5-> frequencies of tickets for each season

Question 6.2 : Then, find the three most common violations for each of these seasons.

Season	Top 3 Violations by Season
Summer	Violation_code Tickets 1 21 127352 2 36 96663 3 38 83518
Winter	Violation_code Tickets 1 21 238181 2 36 221268 3 38 187386
Autum	Violation_code Tickets 1 46 231 2 21 128 3 40 116

Question 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.

Question 7.1 : Find total occurrences of the three most common violation codes

- ✓ Following are the most common violation codes

```
  Violation_code Tickets
1             21  768085
2             36  662765
3             38  542079
```

Question 7.2 : 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**

- ✓ Average of Violation Code 21 is $ 55 {Average of 65 and 45}
- ✓ Average of Violation Code 36 is $ 50 {Average of 50 and 50}
- ✓ Average of Violation Code 38 is $ 50 {Average of 65 and 35}

Question 7.3 : 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.

- ✓ Top Collection is for Violation Code of 21 with total collection of $42,244,675

	Violation_code	Tickets	fine_df	collection
1	21	768085	55	42244675
2	36	662765	50	33138250
3	38	542079	50	27103950

Question 7.4 : What can you intuitively infer from these findings?

- ✓ Higher Violations because of Parking in No Parking Zone (Code 21)
- ✓ Over speeding (Code 36)
- ✓ Parking Meter violations with respect to exceeding allotted parking meter time or not displaying parking meter receipt (Code 38)

_____End of Assignment_____