NYC Spark Case Study

Problem Statement:

New York City is a thriving metropolis. Just like most other metros that size, one of the biggest problems its citizens face is parking. The classic combination of a huge number of cars and a cramped geography is the exact recipe that leads to a huge number of parking tickets.

In an attempt to scientifically analyse this phenomenon, the NYC Police Department has collected data for parking tickets. Out of these, the data files from 2014 to 2017 are publicly available. We will try and perform some exploratory analysis on this data.

We are considering Fiscal Year for doing Analysis on data and created separate data frames for three-year 2015, 2016 & 2017 respectively.

Dataset location in HDFS:

The data for this case study has been placed in following path:

'/common_folder/nyc_parking/Parking_Violations_Issued_-_Fiscal_Year_201x.csv'

where 'x' is between {5,6,7}

Initialization and Loading of Spark and Data frames:

First, we have initialized and loaded the Spark, then we have created separate data frames for year 2015, 2016 and 2017 respectively.

Initial Exploration:

Total Columns in the 3 data frames: 51

Total number of rows in 2015, 2016 and 2017 data frames: 11809233, 10626899 and 10803028 respectively

Data Cleaning:

1. Remove the spaces in column names.
2. Dropped the rows having all NA values in all the columns.
3. Considering that "Summons Number" is a unique column and contains the ticket number, so dropped all the duplicate rows based on this column.
4. Checked NULL values in relevant columns i.e. Registration State, Violation Location, Violation Code, Vehicle Body Type, Vehicle Make, Violation Precinct and Issuer Precinct. Below are the number of null values we get in each column for three years data frames.

	Year		
Columns	2015	2016	2017
Registration State	0	0	0
Violation Location	1633006(13.82%)	1868656(17.58%)	2072400(19.18%)
Violation Code	0	0	0
Vehicle Body Type	411970.34%)	39271(0.36%)	42695(0.39%)
Vehicle Make	68276(0.59%)	63758(0.59%)	73047(0.67%)
Violation Precinct	0	1	0
Issuer Precinct	0	1	0

5. Checked in each data frame of 2015, 2016 & 2017 if there is any row which is not that of fiscal year. There is no such row.
6. Created new data frames of 2015, 2016 & 2017 fiscal years in which we removed the rows contains any NULL value in Vehicle Body Type, Vehicle Make, Violation Precinct, Issuer Precinct columns. Considered that these are erroneous rows and out of total data these are miniscule and it should not affect the analysis.
7. Created Temporary views of 2015, 2016 & 2017 years for further analysis.

Data Analysis:

1. Data Examination
   a. Total Number of tickets for each fiscal year i.e. 2015. 2016 & 2017.

2015	2016	2017
10853283	10533559	10698235

   b. Number of unique states for each year from where the cars that got parking tickets came from.

2015	2016	2017
69	68	67

   There is an erroneous entry in Registration State i.e. “99”, replacing it with state having maximum entries. Below is the Registration State with which we are replacing it.

2015	2016	2017
NY	NY	NY

   So, total unique registration will be left as below.

2015	2016	2017
68	67	66

   c. Number of parking tickets which don’t have the address for violation location.

2015	2016	2017
1627827(15.00%)	1864336(17.70%)	2067584(19.33%)

2. Data Aggregation
   a. Top 5 Violation codes in terms of their frequencies

	Violation Code	Frequency
2015	21	1490694
	38	1323329
	14	917658
	36	760582
	37	745545
2016	21	1518437
	36	1251826
	38	1142736
	14	868949
	37	686045
2017	21	1511576
	36	1398699
	38	1059855
	14	885245
	20	614181

b. Top five Vehicle Body Type and Vehicle Make that got parking tickets.

Vehicle Body Type

	Vehicle Body Type	Frequency
2015	SUBN	3445204
	4DSD	3102508
	VAN	1597628
	DELV	825212
	SDN	444251
2016	SUBN	3460015
	4DSD	2992104
	VAN	1512149
	DELV	740947
	SDN	415295
2017	SUBN	3713303
	4DSD	3081720
	VAN	1403553
	DELV	671930
	SDN	429420

Vehicle Make

	Vehicle Make	Frequency
2015	FORD	1413943
	TOYOT	1121173
	HONDA	1015865
	NISSA	835718
	CHEVR	834512
2016	FORD	1321814
	TOYOT	1152385
	HONDA	1011789
	NISSA	832874
	CHEVR	758078
2017	FORD	1277651
	TOYOT	1208878
	HONDA	1076547
	NISSA	916337
	CHEVR	712969

Top 5 Vehicle body type and Vehicle Make are consistent across 3 years.

c. Top five Precinct Zone where violation occurred and where ticket was issued.

While doing the analysis, we come across the Zone value as “0”, which we considered as erroneous. We ignored that value and did the analysis.

Violation Occurred Zone:

	Violation Occurred Zone	Frequency
2015	19	556374
	18	398540
	14	381175
	1	303554
	114	299599
2016	19	551536
	18	329409
	14	321059
	1	299709
	114	290442
2017	19	531578
	14	347240
	1	326576
	18	302558
	114	295294

Ticket Issuer Zone:

	Ticket Issuer Zone	Frequency
2015	19	542703
	18	389948
	14	367593
	1	295934
	114	294878
2016	19	538293
	18	321685
	14	312887
	1	292569
	114	286237
2017	19	517984
	14	340578
	1	317357
	18	293047
	114	288964

d. Violation Code frequency across three Precincts which have issued to the greatest number of tickets. Considering top 5 Violation Code across each Precincts

2015

	Issuer Precinct	Violation Code	Number of Tickets
1	14	69	80338
2	14	14	76713
3	14	31	41003
4	14	42	28110
5	14	47	27133
1	18	14	120577
2	18	69	57198
3	18	31	30413
4	18	47	29065
5	18	42	19816
1	19	38	90376
2	19	37	79666
3	19	14	60508
4	19	21	56389
5	19	16	56268

2016

	Issuer Precinct	Violation Code	Number of Tickets
1	14	69	67918
2	14	14	61827
3	14	31	35675
4	14	47	24374
5	14	42	23657
1	18	14	99471
2	18	69	47871
3	18	47	23995
4	18	31	22781
5	18	42	17671
1	19	38	77125
2	19	37	75603
3	19	46	71377
4	19	14	61674
5	19	21	58692

2017

	Issuer Precinct	Violation Code	Number of Tickets
1	1	14	72681
2	1	16	38919
3	1	20	27102
4	1	46	21949
5	1	38	16960
1	14	14	72840
2	14	69	57735
3	14	31	39777
4	14	47	30398
5	14	42	20644
1	19	46	84096
2	19	37	72376
3	19	38	72102
4	19	14	57329
5	19	21	54638

Certain Violation Codes have high frequencies and also, they are not common across precincts. For e.g. in 2017, Precinct 1 has large number of tickets for Violation Code “14”, where as Precinct 9 has for “46”.

e. Analysis of Parking Violation across different times of day
    i. Initially we removed all the rows in which Violation Time was missing, total rows in 2015, 2016 and 2017 were 1514, 4179 & 60 respectively.
    ii. After dropping total rows were left as 10851769, 10529380 and 10698175 in year 2015, 2016 & 2017 data frames respectively.
    iii. Violation Time column is in a strange format, created in timestamp format.
    iv. After making Violation Time column in Timestamp format, some values in all the three data frames become NULL values i.e. 60763, 63370 and 58120 in year 2015, 2016 & 2017 data frames respectively.
    v. Removed all such rows from data frames.
    vi. Divided the Violation Time into 6 slots i.e., 0 to 4 Hours, 4 to 8 Hours, 8 to 12 Hours, 12 to 16 Hours, 16 to 20 Hours, 20 to 24 Hours.
    vii. Identified the top three Violation Code based on their frequency across these time slots. They are listed below:

2015

	Time Slots	Violation Code	Number of Tickets
1	0 to 4	21	39805
2	0 to 4	40	36976
3	0 to 4	7	33946
1	4 to 8	14	133783
2	4 to 8	21	105238
3	4 to 8	40	90792
1	8 to 12	21	1185207
2	8 to 12	38	448692
3	8 to 12	36	359891
1	12 to 16	38	567768
2	12 to 16	37	417254
3	12 to 16	36	323095
1	16 to 20	38	241030
2	16 to 20	37	175553
3	16 to 20	7	168245
1	20 to 24	7	81692
2	20 to 24	38	62383
3	20 to 24	14	45112

2016

	Time Slots	Violation Code	Number of Tickets
1	0 to 4	21	43457
2	0 to 4	40	35684
3	0 to 4	78	27043
1	4 to 8	14	139412
2	4 to 8	21	112060
3	4 to 8	40	91154
1	8 to 12	21	1200730
2	8 to 12	36	585988
3	8 to 12	38	387821
1	12 to 16	36	544966
2	12 to 16	38	487902
3	12 to 16	37	383071
1	16 to 20	38	211032
2	16 to 20	37	161486
3	16 to 20	14	133564
1	20 to 24	7	60774
2	20 to 24	38	53148
3	20 to 24	40	44311

2017

	Time Slots	Violation Code	Number of Tickets
1	0 to 4	21	51961
2	0 to 4	40	43594
3	0 to 4	78	27398
1	4 to 8	14	140317
2	4 to 8	21	116819
3	4 to 8	40	111602
1	8 to 12	21	1171632
2	8 to 12	36	750339
3	8 to 12	38	345564
1	12 to 16	36	587626
2	12 to 16	38	461762
3	12 to 16	37	336719
1	16 to 20	38	202844
2	16 to 20	37	145611
3	16 to 20	14	143059
1	20 to 24	7	65461
2	20 to 24	38	46972
3	20 to 24	14	44189

viii. For the 3 most common violation codes, identified the time slots having maximum frequency

2015

	Violation Code	Time Slots	Number of Tickets
1	14	8 to 12	296049
2	21	8 to 12	1185207
3	38	12 to 16	567768

2016

	Violation Code	Time Slots	Number of Tickets
1	21	8 to 12	1200730
2	36	8 to 12	585988
3	38	12 to 16	487902

2017

	Violation Code	Time Slots	Number of Tickets
1	21	8 to 12	1171632
2	36	8 to 12	750339
3	38	12 to 16	461762

f. Analysis of Parking Violations across different seasons.
   i. First checked if the Null Values are present in Issue Date column, there were no Null values.
   ii. Created a new column "Season" based on month of Issue Date as per below:

   Winter Season - Dec, Jan and Feb

   Spring Season - March, April and May

   Summer Season - June, July and August

   Fall Season - September, October and November

   iii. Frequency of tickets for each season.

2015

Season	Number of Tickets
Winter Season	2151302
Summer Season	3051379
Spring Season	2910064
Fall Season	2678261

2016

Season	Number of Tickets
Winter Season	2386229
Summer Season	2396359
Spring Season	2750952
Fall Season	2932470

2017

Season	Number of Tickets
Winter Season	2446502
Summer Season	2565140
Spring Season	2835476
Fall Season	2792937

iv. Three most Violations for each of these sessions

2015

	Season	Violation Code	Frequency
1	Fall Season	21	342866
2	Fall Season	38	326372
3	Fall Season	14	228620
1	Spring Season	21	414886
2	Spring Season	38	326750
3	Spring Season	14	239680
1	Summer Season	21	459310
2	Summer Season	38	363452
3	Summer Season	14	251081
1	Winter Season	38	306740
2	Winter Season	21	247297
3	Winter Season	14	190557

2016

	Season	Violation Code	Frequency
1	Fall Season	36	437721
2	Fall Season	21	384978
3	Fall Season	38	303159
1	Spring Season	36	373904
2	Spring Season	21	373422
3	Spring Season	38	299206
1	Summer Season	21	380865
2	Summer Season	38	271894
3	Summer Season	14	211800
1	Winter Season	21	351224
2	Winter Season	36	314329
3	Winter Season	38	268212

2017

	Season	Violation Code	Frequency
1	Fall Season	36	455426
2	Fall Season	21	347810
3	Fall Season	38	283587
1	Spring Season	21	392600
2	Spring Season	36	344366
3	Spring Season	38	270080
1	Summer Season	21	394381
2	Summer Season	38	247071
3	Summer Season	36	240065
1	Winter Season	36	358842
2	Winter Season	21	352917
3	Winter Season	38	259101

g. Estimated the total revenue collected for three top violation codes with maximum tickets

2015

Violation Code	Number of Tickets	Average Fine	Total Amount
21	1464359	55	80539745
38	1323314	51.5	68150671
14	909938	115	104642870

Violation Code **"14"** which is "General No Standing: Standing or parking where standing is not allowed by sign, street marking or; traffic control device" has maximum collection for year 2015.

2016

Violation Code	Number of Tickets	Average Fine	Total Amount
21	1490489	55	81976895
36	1251826	50	62591300
38	1142471	51.5	58837256.5

Violation Code **"21"** which is "Street Cleaning: Standing or parking where standing is not allowed by sign, street marking or; traffic control device" has maximum collection for year 2016.

2017

Violation Code	Number of Tickets	Average Fine	Total Amount
21	1487708	55	81823940
36	1398699	50	69934950
38	1059839	51.5	54581708.5

Violation Code **"21"** which is "Street Cleaning: Standing or parking where standing is not allowed by sign, street marking or; traffic control device" has maximum collection for year 2016.

Statistics for all the years reflect that NYC Police Department are getting major amount of revenue from the Violation in which people are parking in unallowed area.