

# CS4830 – BIG DATA LABORATORY

## FINAL PROJECT REPORT

<b>Akhila D.</b>	<b>EE18B131</b>
<b>Chetana T.</b>	<b>CE18B125</b>
<b>Shania M.</b>	<b>CH18B067</b>

### PROBLEM STATEMENT

In this case study, we analyse the NYC Parking Tickets dataset. This dataset contains information about various tickets issued in New York (details can be found on <https://www.kaggle.com/new-york-city/nyc-parking-tickets>). The aim is to predict the part in the city where the ticket was issued (Prediction column name:- Violation County).

### ABOUT THE DATASET

The information is based on Parking Tickets issued as a result of parking violations in and around New York City, New York, United States. The majority of the dataset's columns are concerned with the circumstances and timing of the parking violation. The bulk of the columns in this dataset do not appear to contain numeric or ordinal data at first look. As a result, quantitative statistics from the dataset, such as mean and standard d Deviations are unimportant. The prediction column was 'Violation County,' which is supposed to indicate the place where the parking violation occurred. It was first confirmed that there are no null or missing entries in this column. Feature engineering was performed on the remaining columns.

### EXPLORATORY DATA ANALYSIS

#### DIMENSIONS

The dataset consists of 22436132 rows and 43 columns.

#### DUPLICATE ROWS

There were around 857976 duplicate entries. We'll have 22436132 different rows after deleting duplicate rows. We hunt for a property with as many different values as the number of distinct rows to identify the main key. The 'Summons Number' property meets the primary key criteria and is thus chosen as the dataset's main key.

#### NULL COLUMNS

Some of the columns in the dataset were populated, largely with or completely, with null (NaN) values. These columns had to be removed, so every column with more than 50% null value count

was considered for deletion. Columns dropped: 'Intersecting Street', 'Time First Observed', 'Violation Legal Code', 'Unregistered Vehicle?', 'Meter Number', 'No Standing or Stopping Violation', 'Hydrant Violation', 'Double Parking Violation'.

## UNINFORMATIVE COLUMNS

Some of the columns in the dataset were deemed unnecessary for defining the Violation County target feature since it is expected to be connected to the place of violation occurrence:

**Summons Number:** A one-of-a-kind identifier, a numeric number unrelated to the data. 'Summons Number' was dropped.

**Plate ID:** The plate id is often random and unrelated to the work; but, because various states employ different formats, the serial formatting of the plate id itself may provide insight on the location. This information is thought to be best represented in the column 'Vehicle Registration.' As a result, 'Plate ID' was eliminated.

Street Name and other name-based fields were eliminated since they included numerous unique values, exhibiting no pattern.

**Vehicle Expiration Date:** This field contained several items with incorrect or malformed data, and the Expiration Date of a vehicle is not presumed to be relevant to the job. As a result, this column and other time-based elements that were deemed unnecessary were removed.

**House Number:** A house number is a code that is made up of one or more values. As a result, it was unable to identify a precise pattern or categorize this column. As a result, it was deemed unnecessary and dropped.

Other numeric-based fields that were unrelated to the purpose (finding the location of the infraction) were removed.

'Violation Time' is binned into six-time intervals, and a new column called 'Time Bin' is established in its place.

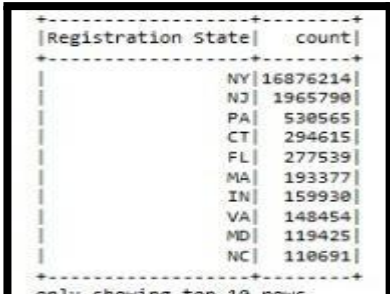
As a result of the preceding factors, the entire list of columns has been removed:

'Sub Division', 'House Number', 'Street Name', 'Date First Observed', 'From Hours In Effect', 'To Hours In Effect', 'Vehicle Color', 'Days Parking In Effect', 'Violation Post Code', 'Violation Description', 'Vehicle Year', 'Feet From Curb'

## COLUMN WISE EXPLORATION

### Exploration of Registration State:

Inference: As predicted, automobiles registered in the state of New York were responsible for the bulk of the offenses. However, other states account for a large share of registrations (primarily NJ). This might contain information about the Violation County (for example, New York City would be more likely to draw individuals from neighboring states than a less metropolitan area). As a result, the



Registration State	count
NY	16876214
NJ	1965790
PA	530565
CT	294615
FL	277539
MA	193377
IN	159930
VA	148454
MD	119425
NC	110691

only showing top 10 rows

appearance of an out-of-state registration might suggest a violation in New York City). The value 99 has been treated as a missing value and thus, has been imputed using the most frequent value of New York City.

### Exploration of Plate Type:

Conclusion: It may be useful to investigate the various Plate Types, as some plate types may be more prevalent in specific regions.

Plate Type	count
PAS	15619871
COM	4214934
OMT	682845
SRF	199389
OMS	178617
999	119095
IRP	116393
TRC	56145
MOT	54943
OMR	41540

only showing top 10 rows

### Exploration of Issue Date:

Inference: Because the Issue Dates are more evenly distributed than other categories, the top ten frequencies only account for around 3.5 percent of the total data utilized for exploration (300,000 rows). This would suggest that some feature engineering is required. The month and year of the issue date can be divided into two columns.

After Binning:

Issue Date	count
06/26/2015	77805
06/30/2015	74145
06/29/2015	66590
06/23/2015	52171
06/22/2015	51700
06/24/2015	49436
09/22/2015	49425
06/19/2015	48274
10/16/2015	48146
10/13/2015	48055

only showing top 10 rows

Issue Day	count
3	3195019
5	3168059
6	3020139
4	2940777
2	2786206
7	1965485
1	685494

Issue Month	count
3	1633397
10	1630152
8	1585730
7	1562004
6	1535580
9	1533010
4	1529735
5	1519385
11	1420574
2	1351454

only showing top 10 rows

Issue Year	count
2015	9202934
2014	4578193
2016	3976935
2017	915
2013	539
2018	428
2000	356
2012	202
2019	161
2010	136

only showing top 10 rows

### Exploration of Violation Time:

Inference: The Violation Time is a pseudo-numeric attribute in the sense that, although being represented as text entries, it may be transformed to a numeric distribution if the AM-PM time shifts are taken into consideration. Although the precise moment of violation may not be critical, having a broad notion of the time allows multiple items to be grouped together. This would be beneficial for categorizing infractions; for example, all violations happening early in the morning may be put together.

Violation Time	Bin
00:00 AM to 04:00 AM	1
04:00 AM to 08:00 AM	2
08:00 AM to 12:00 PM	3
00:00 AM to 04:00 AM	4
04:00 PM to 08:00 PM	5
08:00 PM to 00:00 AM	6

Violation Time	count
0836A	62911
1136A	60761
1140A	56277
0936A	53719
0840A	51982
1138A	51706
0906A	51666
1145A	51243
0940A	51233
1139A	51018

only showing top 10 rows

Exploration of Vehicle Body Type:

Inference: While this attribute is unlikely to be particularly useful, it will nevertheless help distinguish between different vehicle kinds in different places.

Vehicle Body Type	count
SUBN	6918000
4DSD	6094617
VAN	3123531
DELV	1595724
SDN	878035
2DSD	573384
PICK	542938
REFG	168256
UTIL	157573
TRAC	146507

only showing top 10 rows

Exploration of Vehicle Make:

Inference: This attribute, like Vehicle Body Type, isn't likely to be particularly important. However, one may argue that some manufacturers produce luxury automobiles more frequently and that these premium vehicles are more common in urban areas (such as inner NYC). As a result, this function may indirectly reveal information about the parking violation's location.

Vehicle Make	count
FORD	2742077
TOYOT	2278313
HONDA	2032123
NISSA	1672402
CHEVR	1596052
FRUEH	831844
DODGE	733021
ME/BE	719778
BMW	700668
INTER	615579

only showing top 10 rows

Issuing Agency	count
T	15753021
V	3477480
P	1711917
S	431832
X	180851
K	14745
R	1440
H	1265
C	1163
F	904

only showing top 10 rows

Exploration of Issuing Agency:

Inference: In the context of Parking Tickets, a clear understanding of the information in the Issuing Agency column was not acquired. The Issuing Agency, on the other hand, is understood to refer to the issuing authority

or jurisdiction in which the parking ticket is issued. It's thought that this will aid in determining the location of the offense.

**Exploration of Street Codes:**

Inference: In these Street Codes, the 0 value item might indicate missing values. Furthermore, there is more missing data in the Street Code 2 and Street Code 3 entries. This is to be anticipated, as some street codes are shorter than others. Despite the fact that the Street Codes are a distribution of 5-digit numbers, many of the street codes have the same value (such as the most commonly observed '13610'). This suggests that categorizing street regulations by most often might offer information about where violations occur.

Street Code1	count
0	4381771
13610	362018
10210	305127
25390	212826
24890	181676
10110	168958
10010	153047
59990	130362
10810	130324
10410	128460

only showing top 10 rows

Street Code2	count
0	6043199
40404	623286
10410	475313
13610	240518
10610	202951
10210	187331
10510	185397
10810	156634
10110	155863
24890	143122

only showing top 10 rows

Street Code3	count
0	6145642
40404	623286
10510	267760
13610	246029
10810	219889
10110	196830
10610	196515
25390	169287
10010	161008
10210	152223

only showing top 10 rows

**Exploration of Issuer Command:**

Inference: The Issuer Command and Squads are most likely based in and around New York City. As a result, the command and squad of the ticket issuer will be a good indicator of the region where the infraction happened.

Issuer Command	count
null	3477481
T103	2471202
T401	2101222
T302	2019232
T301	1335971
T402	1230709
T201	1074931
T102	1063171
T106	1016263
T105	707516

only showing top 10 rows

Issuer Squad	count
null	3478126
0000	2362284
A	1272572
M	1075494
B	1022550
C	989472
D	955552
E	910649
H	867672
J	852566

only showing top 10 rows

### Exploration of 'Violation In Front of Or Opposite:

Inference: This column is thought to provide information on the location of the vehicle that triggered the violation.

Although this column may not be very relevant to determining the region of violation, it is kept in the feature columns list because it is possible that some labels, such as 'I' instead of 'F' (both of which likely denote 'In Front Of'), are more commonly used by issuing authorities in specific regions.

Violation In Front Of Or Opposite	count
F	10925226
O	4425707
null	3648400
I	2549414
R	16043
X	13365
0	1

### Feature Transformation and handling of missing values

After using Pyspark UDF to extract the day, month, and year from the 'Issue Date' feature, additional features called 'Issue Day,' 'Issue Month,' and 'Issue Year' were generated.

Exploration Subset distribution of these new features:

The Issue Month is evenly spread throughout the year. The most commonly occurring month is irrelevant in this scenario since every month occurs with about equal frequency. The majority of the tickets appear to have been issued in the years 2014, 2015, and 2016, with 2015 appearing to be the most common.

The bulk of tickets are issued in the first seven days of the month, according to Issue Day. This column is not kept since no further conclusion can be drawn from the Issue Day. The training will use the other two columns, Issue Month and Issue Year.

Pyspark UDF was also used to divide the data set 'Violation Time' into six bins. The bins are arranged in the following order:

The majority of the infractions appear to occur between the hours of 8 a.m. and 12 p.m. in Bin 3. In comparison, early morning (bin 1) and late night (bin 6) breaches are the least common. As a result, this binning approach provides for a more accurate interpretation than the original time entries.

We utilize the most commonly occurring value [discovered via data exploration] of that related feature as a proxy for the default value of that feature in the event of missing value in any of the characteristics that are maintained. The "Issue Month" is an exception, as the data were evenly distributed and the mean was used. The following table lists the default values:

Retained Feature Name	Data Type	Imputed Value
Registration State	String	"NY"
Plate Type	String	"PAS"
Issue Month	Ordinal Integer	6
Issue Year	Ordinal Integer	2015
Vehicle Body Type	String	"SUBN"
Vehicle Make	String	"FORD"
Issuing Agency	String	"T"
Street Code1	Nominal Integer	0
Street Code2	Nominal Integer	0
Street Code3	Nominal Integer	0
Issuer Command	String	"T103"
Issuer Squad	String	"A"
Violation In Front Of Or Opposite	String	"F"
Time bin	Ordinal Integer	3