NYC Case Study

Group Name:

1. Manish Jha
2. Namrata Khatri
3. Prashant Agrawal
4. Rahul Shukla

Objectives & Goals of NYC Dataset Analysis

Business Objectives:

The purpose of this case study is to conduct an exploratory data analysis that helps you understand the data.

Overall Structure of the presentation:

- **Problem Statement** –Understand the Dataset of NYC using Spark/SparkR/Spark SQL.

Guidelines & Assumptions

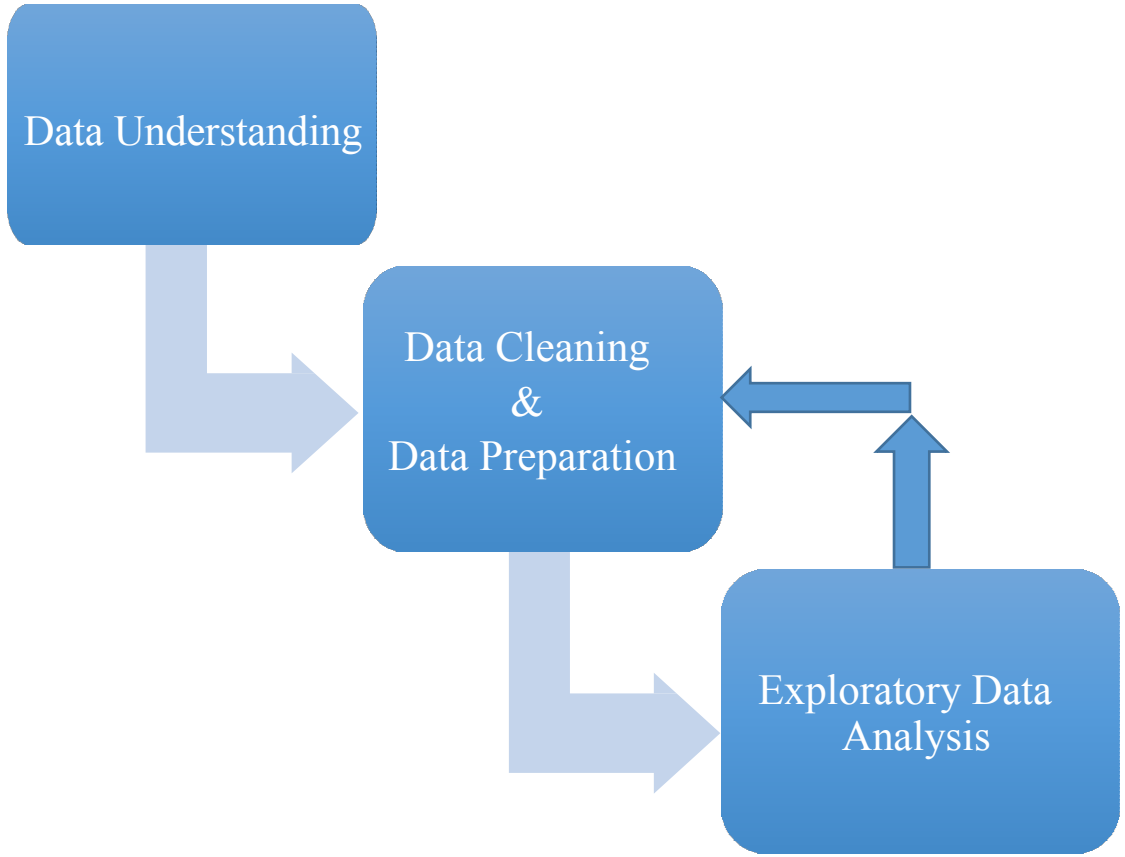
Guidelines:

- Dataset File name need not be changed.
- All code to be written in SparkR

Assumptions:

- **Baseline Code–** The code used for analysis is to be baselined and it can be used in future if follow-up questions are there from Chief Data Scientist.
- **Data** – All observations on the raw data to be explained with reasons before doing data cleaning.
- **SparkR Code-** Chief Data Scientist can ask to demonstrate additional charts.

Approach – High Level



Data Understanding & Preparation

Data Understanding:

1. Around 10million rows in the dataset
2. Dataset contains timestamp column
3. Tickets raised against violations tagged to different types of attributes.

Data Cleaning & Preparation:

1. Step wise when ever needed at different stages
2. Registration State- 99 is not a valid state. So, data cleaning exercise was done to remove this.
3. We have checked all the columns with ISNULL function and counted the rows. Null values are not there.
4. Violation Precinct, Issuer Precinct - The erroneous entries of 0 have been removed.
5. Hour value of 24hour timestamp is erroneous and any value above 24 has been cleaned up.
6. String NaN & + sign was present in the timestamp (total 4 rows). This is cleaned up.
7. Missing Values not present for any of the attributes.

Exploratory Data Analysis -1

Key Insights

- The Maximum no of tickets is from NY
- The Top-5 violations are for Violation Codes 21, 36 and 38. This gives us insights of different violations that result in tickets getting generated.

Violation	Code	No of Records
1	21	1522731
2	36	1400569
3	38	1061330
4	14	890906
5	20	616772

- The Top-5 violations from different body types

Vehicle Body	Type	No of Records
1	SUBN	3712393
2	4DSD	3081487
3	VAN	1407870
4	DELV	683415
5	SDN	427806

Exploratory Data Analysis -2

Key Insights

- The Top-5 violations are for body make. This gives us insights of different violations that result in tickets getting generated.

	Make	No of Records
1	FORD	1277575
2	TOYOT	1208877
3	HONDA	1076357
4	NISSA	916469
5	CHEVR	713135

- Top-5 Violation Precinct

	Make	No of Records
1	19	532731
2	14	350881
3	1	323077
4	18	304377
5	114	290969

Exploratory Data Analysis -3

Key Insights

- Top-5 Issuer Precinct

	Issuer Precinct	No of Records
1	19	520653
2	14	343898
3	1	319994
4	18	295759
5	114	289100

- When Violation Precinct & Issuer Precinct are 19, 14, 1 (top 3) then the maximum number of violation codes is for 14, 46 and 38.

Exploratory Data Analysis -4

- Segregation of Violation codes based on time slots and their respective counts:

	Timeslot	Violation Code	No of Records
1	0-3hr	21	52096
2	0-3hr	40	49697
3	0-3hr	14	31167
4	4-7hr	14	139842
5	4-7hr	40	111212
6	4-7hr	20	83848
7	8-11hr	21	1008666
8	8-11hr	38	345750
9	8-11hr	14	271375

Exploratory Data Analysis -5

- Segregation of Violation codes based on time slots and their respective counts (continued):

	Timeslot	Violation Code	No of Records
10	12-15hr	38	462015
11	12-15hr	37	336704
12	12-15hr	14	253866
13	16-20hr	38	202839
14	16-20hr	37	145585
15	16-20hr	14	142642
16	20-24hr	38	46989
17	20-24hr	14	44224
18	20-24hr	40	43907

Exploratory Data Analysis -6

- Segregation of timeslots for different violation codes and their respective counts (reverse of above insight).

	Violation Code	Timeslots	No of Records
1	21	8-11hr	1008666
2	21	12-15hr	123195
3	21	4-7hr	73985
4	38	12-15hr	462015
5	38	8-11hr	345750
6	38	16-20hr	202839
7	14	8-11hr	271375
8	14	12-15hr	253866
9	14	16-20hr	142642

Exploratory Data Analysis -7

- Category of seasons are derived from the dataset. the different seasons with respective months are shown below:
  11,12,1,2 -- winter
  3,4 -- spring
  5,6,7 -- summer
  8,9,10 – rainy

	Season	No of Tickets
1	Winter	2583993
2	Spring	1492922
3	Summer	2178710
4	Rainy	2122237

Exploratory Data Analysis -8

- Find the 3 most common violations in each season

Winter	Violation Code	No of Tickets
	21	399414
	38	348135
	14	263800

Spring	Violation Code	No of Tickets
	21	203609
	38	185710
	14	164835

Summer	Violation Code	No of Tickets
	21	348230
	38	250139
	14	238612

Rainy	Violation Code	No of Tickets
	21	307383
	38	276332
	14	215869

Exploratory Data Analysis -9

- Find total occurrences of 3 most common violation codes:

	Violation Code	No of Tickets
1	21	1258636
2	38	1060316
3	14	883116

- Violation code and total collection:

	Violation Code	No of Tickets	Total Fine Amount
1	21	1258636	69224980
2	38	1060316	53015800
3	14	883116	44155800

Exploratory Data Analysis -10

- Violations happens most of the office time duration 8-11,16-20 and lunch time 12-15
- Violation codes happen during these times are 21,38,14
- Maximum violations happens during Winter (4 months) then summer (3 months)
- Minimum violation happens during spring but spring has only 2 months