Driving insights from the urban jungle: Harnessing the power of Big Data to solve NYC's Parking Violation Puzzle!

ABIRAMI S AS16288 APURVA S AS17321 RAKSHANA B S RB5118

NYU TANDON SCHOOL OF ENGINEERING

NYU | TANDON SCHOOL OF ENGINEERING

OBJECTIVES

- The aim of this case study is to provide an in-depth analysis of the usage of Spark. Specifically, it will focus on gaining familiarity with PySpark's analysis process, as opposed to base Python.

- Understanding the basics of PySpark functions can be valuable in working with other libraries. Typically, the most time-consuming step in drawing insights from data is preparing the data for model building.

- Therefore, this case study will prioritize exploratory analysis to help tackle the challenge of solving NYC's Parking Violation Puzzle using Big Data.

NYU | TANDON SCHOOL OF ENGINEERING

PROBLEM STATEMENT

- The bustling city of New York is plagued with a problem that's all too common in large metropolitan areas: parking. The sheer volume of cars and limited space makes it nearly impossible to find a suitable parking spot, resulting in a staggering number of parking violations.

- Our project aims to provide insights into parking violations in 2022 by leveraging the power of big data. By conducting an exploratory data analysis, we hope to uncover patterns and trends that will help us understand the factors contributing to parking violations in the city.

- With these insights, we aim to provide recommendations and solutions that will alleviate parking-related woes for New Yorkers and enhance the city's transportation system.

NYU TANDON SCHOOL OF ENGINEERING

WHY DOES THIS PROBLEM REQUIRE BIG DATA ?

- The high volume of parking violations in NYC requires a large dataset to capture and analyze the vast amount of data.
- The complex nature of parking violations requires a comprehensive analysis of multiple variables, which can only be achieved through big data.
- Harnessing big data allows for the identification of patterns and trends that would be difficult or impossible to detect with smaller datasets, thus enabling more effective solutions to the problem.

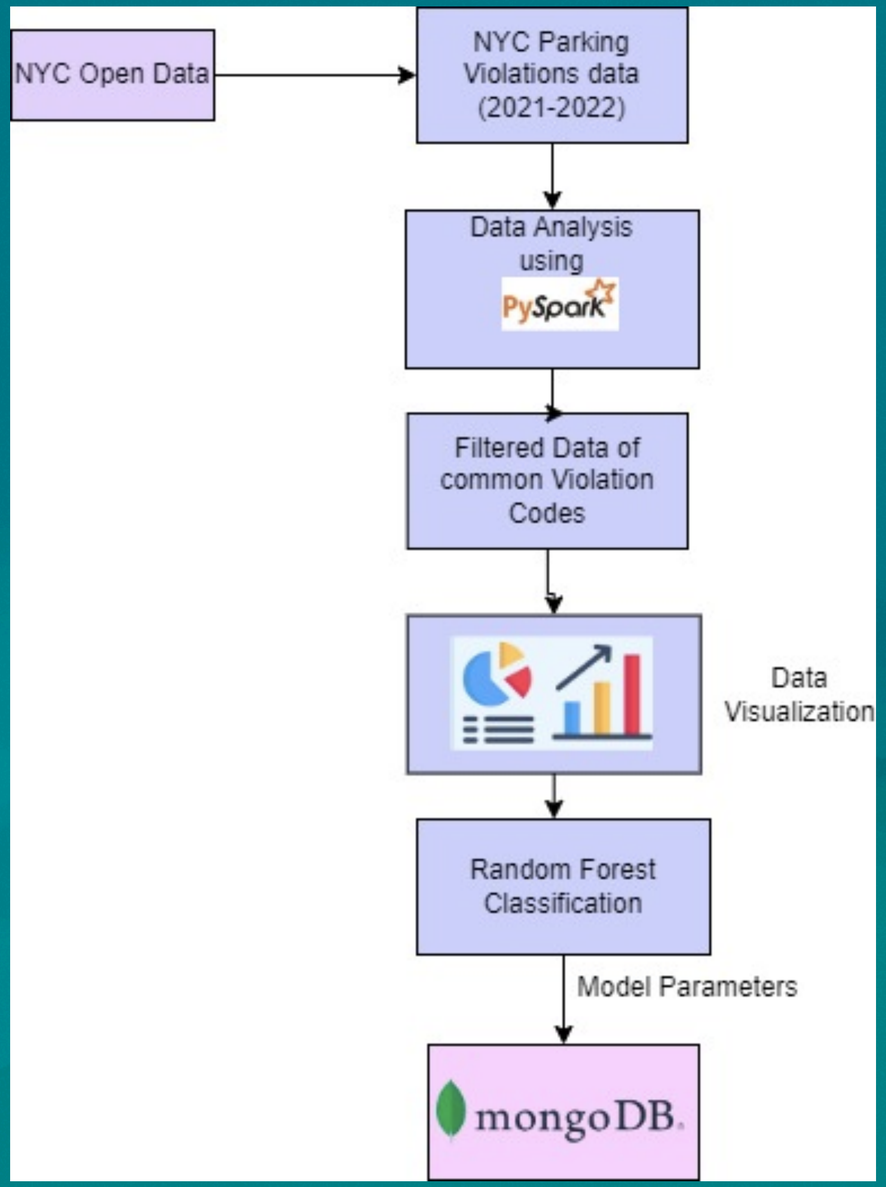
NYU TANDON SCHOOL OF ENGINEERING

TECHNOLOGIES USED

- PYSPARK
- MongoDB
- Kafka
- LIBRARIES - PANDAS, MATPLOTLIB, NUMPY, SEABORN, Folium, GEOCODER
- SPARKML

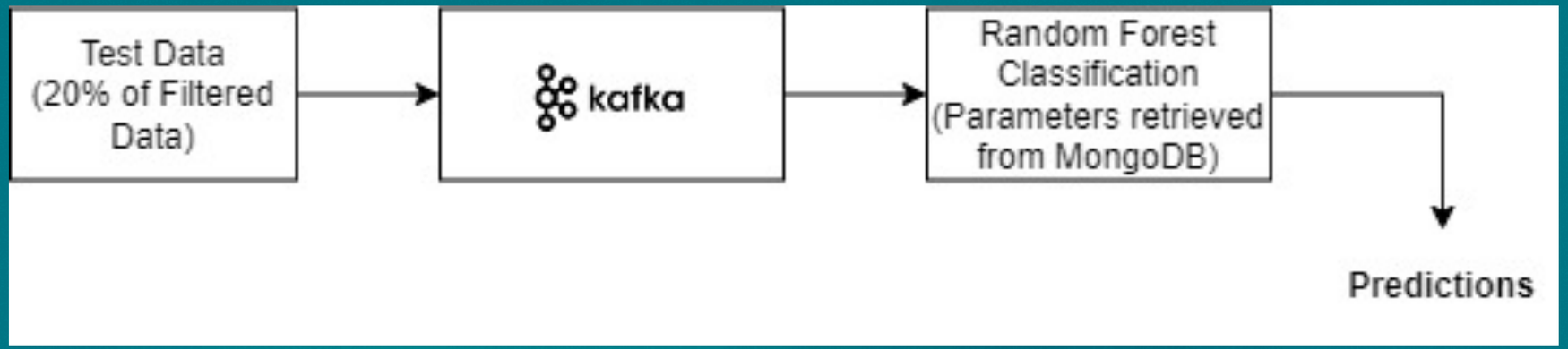
NYU TANDON SCHOOL OF ENGINEERING

ARCHITECTURE



NYU TANDON SCHOOL OF ENGINEERING

ARCHITECTURE



NYU TANDON SCHOOL OF ENGINEERING

DATASET

- Dataset sourced from https://data.cityofnewyork.us/City-Government/Parking-Violations-Issued-Fiscal-Year-2022/7mxj-7a6y
- Contains dataset consists of 15,435,607 records and 43 columns.
- The dataset contains records from June 2021 to August 2022.
- Each row represents a parking violation, and the columns provide information such as the unique identifier for the violation, license plate number, date and time of the violation, location, issuing agency, violation code, and more.
- This dataset can be used for analysis and insights into parking violations in New York City.

NYU TANDON SCHOOL OF ENGINEERING

DATA CLEANING AND PRE PROCESSING

```
Vehicle_Body_Type column has 33128 missing values
Vehicle_Make column has 10689 missing values
Violation_Location column has 6035566 missing values
Issuer_Command column has 6030198 missing values
Issuer_Squad column has 6424789 missing values
Violation_Time column has 177 missing values
Time_First_Observed column has 14643865 missing values
Violation_County column has 37194 missing values
Violation_In_Front_Of_Or_Opposite column has 6097273 missing values
House_Number column has 6164206 missing values
Street_Name column has 1517 missing values
Intersecting_Street column has 7990868 missing values
Sub_Division column has 2165 missing values
Violation_Legal_Code column has 9405407 missing values
Days_Parking_In_Effect column has 6150607 missing values
From_Hours_In_Effect column has 10441351 missing values
To_Hours_In_Effect column has 10441373 missing values
Vehicle_Color column has 1019839 missing values
Unregistered_Vehicle? column has 15118764 missing values
Meter_Number column has 13533799 missing values
Violation_Post_Code column has 6741551 missing values
Violation_Description column has 317788 missing values
No_Standing_or_Stopping_Violation column has 15435607 missing values
Hydrant_Violation column has 15435607 missing values
Double_Parking_Violation column has 15435607 missing values
```

Filtering out Irrelevant Columns

Set Threshold for missing values

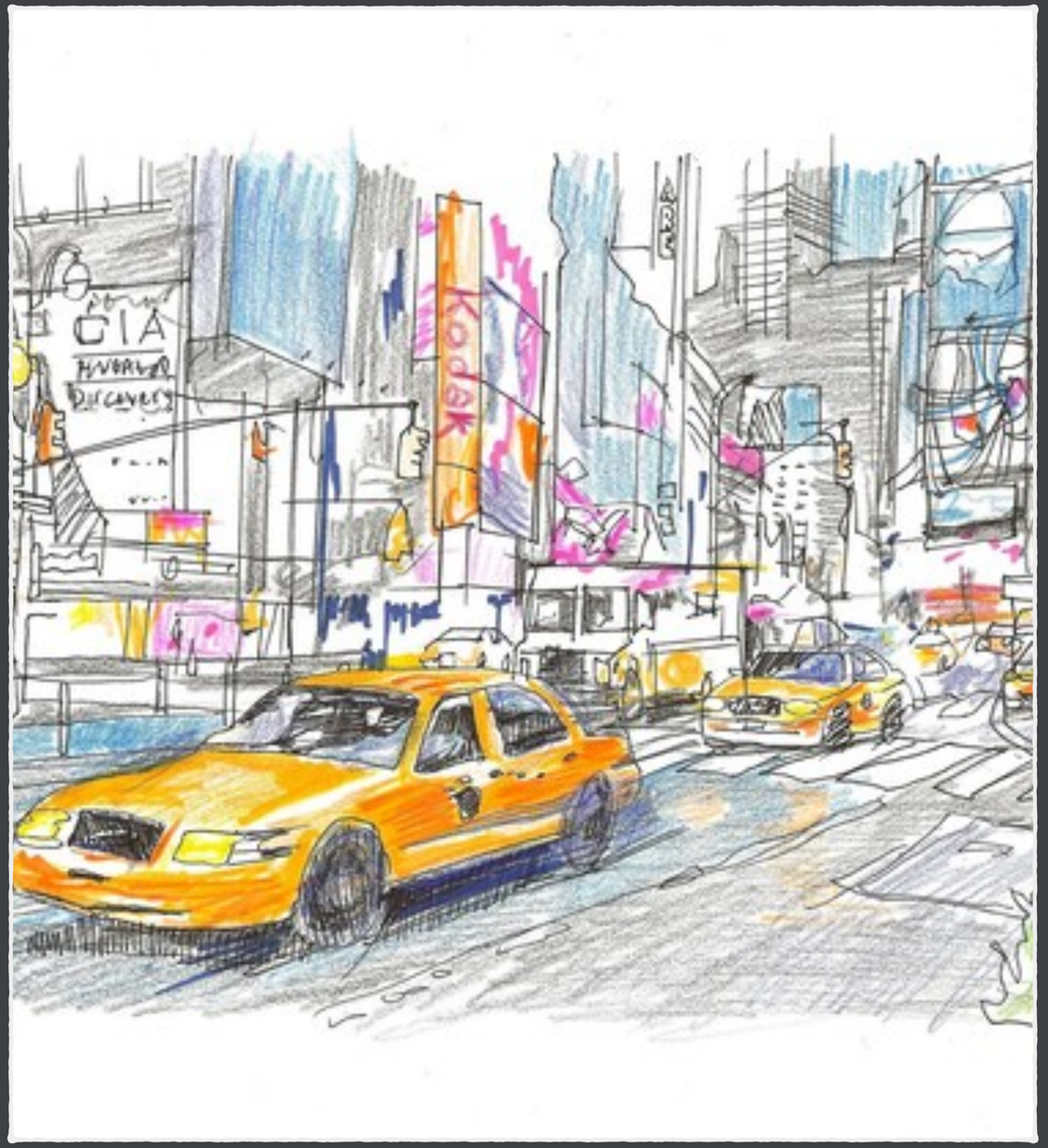
Handling Missing Values

Linear interpolation
Filling in missing values

NYU TANDON SCHOOL OF ENGINEERING

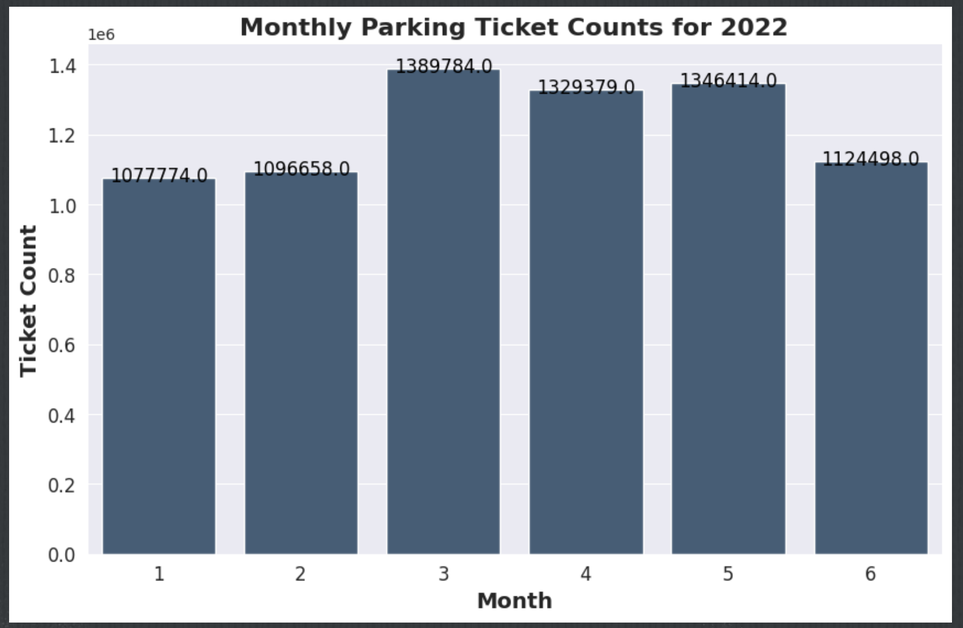
EDA

Trends and Patterns in NYC Parking



NYU TANDON SCHOOL OF ENGINEERING

plate_id	ticket_count
40404JG	657
55219MM	559
2731057	503
82536PC	453
2871114	438



month	ticket_count
1	1077774
2	1096658
3	1389784
4	1329379
5	1346414
6	1124498

NYU TANDON SCHOOL OF ENGINEERING

total_number_of_tickets
7364889

number_of_unique_states
64

registration_state	ticket_count
NY	5591980
NJ	656580
PA	229326
FL	146887
CT	115292
TX	71458
IN	62371
MA	48698
GA	47134
VA	45763
NC	41168
MD	39071
CA	28178
IL	23386
OH	21349
AZ	19180
ME	17071
SC	16864
TN	12638
DE	12331

vehicle_make	ticket_frequency
HONDA	892055
TOYOT	813451
FORD	703499
NISSA	622117
CHEVR	388320

violation_code	ticket_frequency
36	2172974
21	781055
38	549389
14	460303
40	403276

vehicle_body_type	ticket_frequency
SUBN	3124187
4DSD	2077418
VAN	640294
PICK	212023
DELV	204294

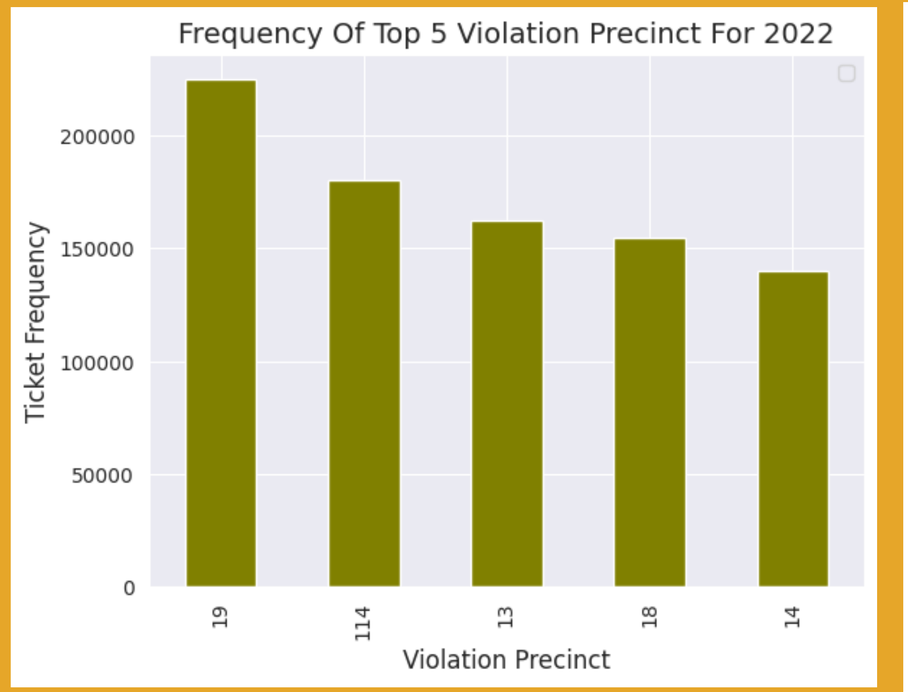
There are 50 US states mentioned in this table, and 14 provinces/territories outside the US (ON, QB, GV, DP, NS, AB, BC, NB, FO, MB, PE, SK).

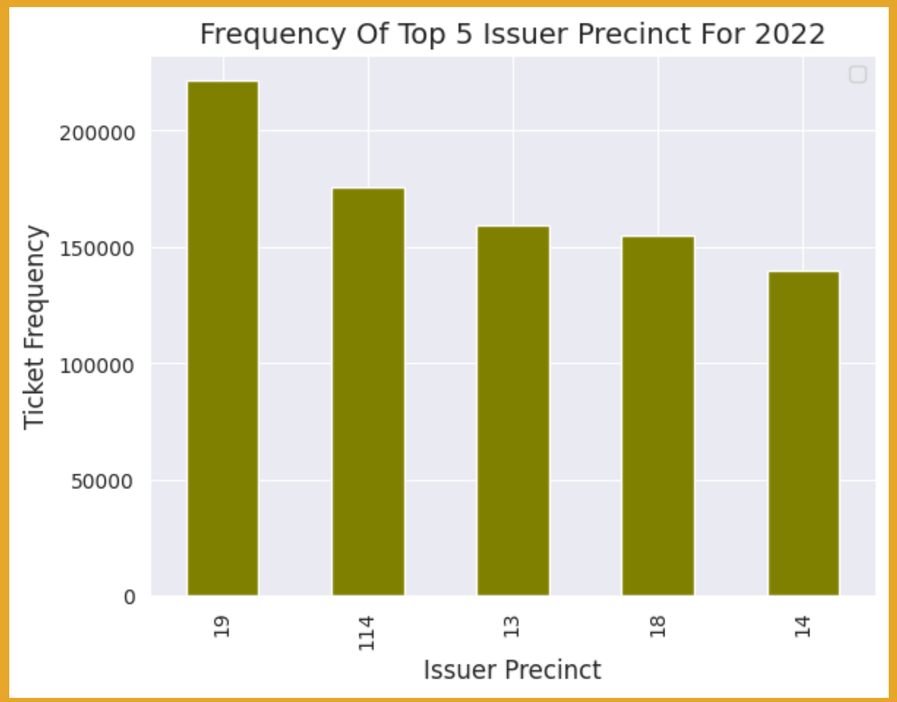
For Vehical Body Type, maximum parking violations happen for Suburban(SUBN) followed by four door sedan(4DSD) and Van - For Vehicle Make, maximum parking violations happen for HONDA follwed by Toyota and ford

NYU TANDON SCHOOL OF ENGINEERING

violation_precinct	ticket_frequency
0	2700953
19	224703
114	180404
13	162145
18	154857
14	140062

issuer_precinct	ticket_frequency
0	2914218
19	221492
114	175466
13	159279
18	154754
14	139769





Precinct 19

violation_code	ticket_frequency
38	31070
21	27571
14	23899
20	22191
40	14652

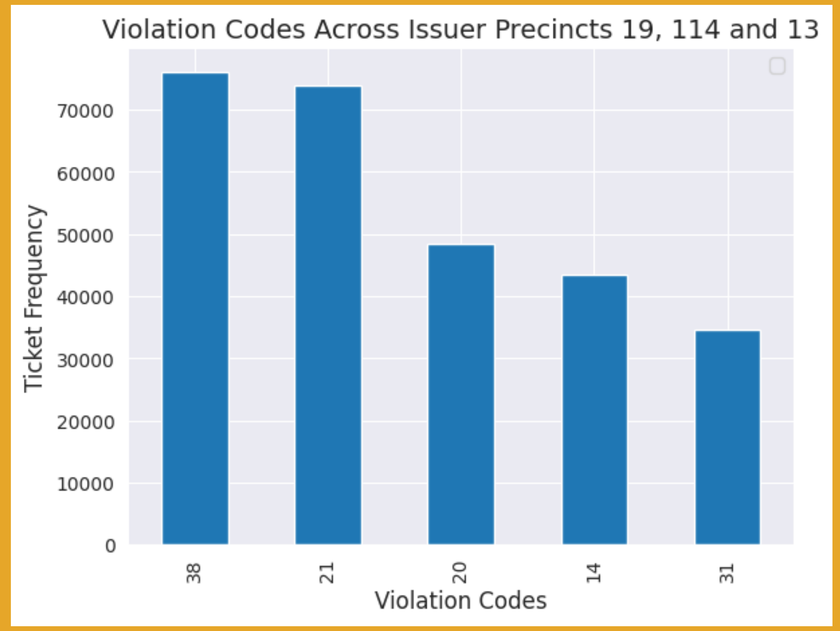
Precinct 114

violation_code	ticket_frequency
21	39114
38	33761
20	15399
40	14255
71	13981

Precinct 13

violation_code	ticket_frequency
69	24780
31	24395
47	17820
14	12718
38	11234

NYU TANDON SCHOOL OF ENGINEERING



violation_code	ticket_frequency
38	76065
21	73848
20	48363
14	43424
31	34506

A precinct is a police station that has a certain zone of the city under its command.'Violation Precinct' (This is the precinct of the zone where the violation occurred).'Issuer Precinct' (This is the precinct that issued the ticket.)

Per the results from above, it can be inferred that - The top 3 violation precincts and Issuer Precincts where maximum parking violations happen are 19, 114 and 13 .

The top 3 common violation codes across top 3 precincts 19, 114 and 13 where the parking tickets were issued are 14, 20 and 38.

Violation code 38 has exceptionally high frequency of 76065 tickets issued.

While the violation codes 14, 20, 38 are common across the issuer precinct, the violation code 14,38 occurs in all the 3 issuer precincts (19, 114 and 13) while violation codes 20 occur in issuer precincts 114 and 113

BIN 1

violation_code	violation_count
21	24903
40	23536
7	13239

BIN 2

violation_code	violation_count
36	199394
40	118172
14	91631

BIN 3

violation_code	violation_count
21	606135
36	559645
38	221069

BIN 4

violation_code	violation_count
36	636497
38	242621
20	136992

BIN 5

violation_code	violation_count
36	543529
5	90874
38	61361

BIN 6

violation_code	violation_count
36	233908
7	30969
40	28792

violation_code	violation_count
36	2172974
21	781055
38	549389

MAX VIOLATION COUNT FOR THE TOP 3 MOST VIOLATED CODE

violation_time_bin	violation_count
3	1386849
4	962923
5	605139
2	270772
6	252819
1	24913

MAX VIOLATION COUNT IN EACH TIME BIN

NYU TANDON SCHOOL OF ENGINEERING

NYU TANDON SCHOOL OF ENGINEERING

TIME BIN INFERENCE

- Based on the provided data, we can analyze the three most commonly occurring violation codes in each of the six time bins as well as across all bins.
- For bin 1 (12:00 AM to 4:00 AM), the three most commonly occurring violation codes are 21, 40, and 7.
- For bin 2 (4:00 AM to 8:00 AM), the three most commonly occurring violation codes are 36, 40, and 14.
- For bin 3 (8:00 AM to 12:00 PM), the three most commonly occurring violation codes are 21, 36, and 38.
- For bin 4 (12:00 PM to 4:00 PM), the three most commonly occurring violation codes are 36, 38, and 20.
- For bin 5 (4:00 PM to 8:00 PM), the three most commonly occurring violation codes are 36, 38, and 20.
- For bin 6 (8:00 PM to 12:00 AM), the three most commonly occurring violation codes are 36, 5, and 38.
- Across all bins, the three most commonly occurring violation codes are 36, 21, and 38. Therefore, we would consider violation codes 36, 21, and 38 as the most commonly occurring violation codes for further analysis.
- Based on the above result, the most common time of the day that violations occur for violation codes 21, 36 and 38 are between 8:00 AM to 12:00 PM followed by 12:00 PM to 4:00 PM and 4:00 PM to 8:00 PM

season	ticket_frequency
spring	4065577
winter	2174542
summer	1124593
autumn	177

TOTAL TICKETS ISSUED EACH SEASON - 2022 FISCAL YEAR

violation_code	violation_count
36	700174
21	192766
38	157576

WINTER

violation_code	violation_count
36	336455
21	151927
38	84705

SUMMER

violation_code	violation_count
98	38
40	36
46	28

AUTUMN

violation_code	violation_count
36	1136345
21	436359
38	307106

SPRING

Per the results from above, it can be inferred that - Maximum Ticket Frequency occur in Spring followed by Winter, Summer and Autumn. Autumn has the least Ticket Frequency. - Most commonly occuring violation codes during Spring, Winter and Summer are 21,36 and 38 - Most commonly occuring violation codes during Autumn are 46, 98 and 40

violation_code	ticket_frequency	fine_amount
36	2172974	108648700
21	781055	42958025
38	549389	27469450

FINE AMOUNT

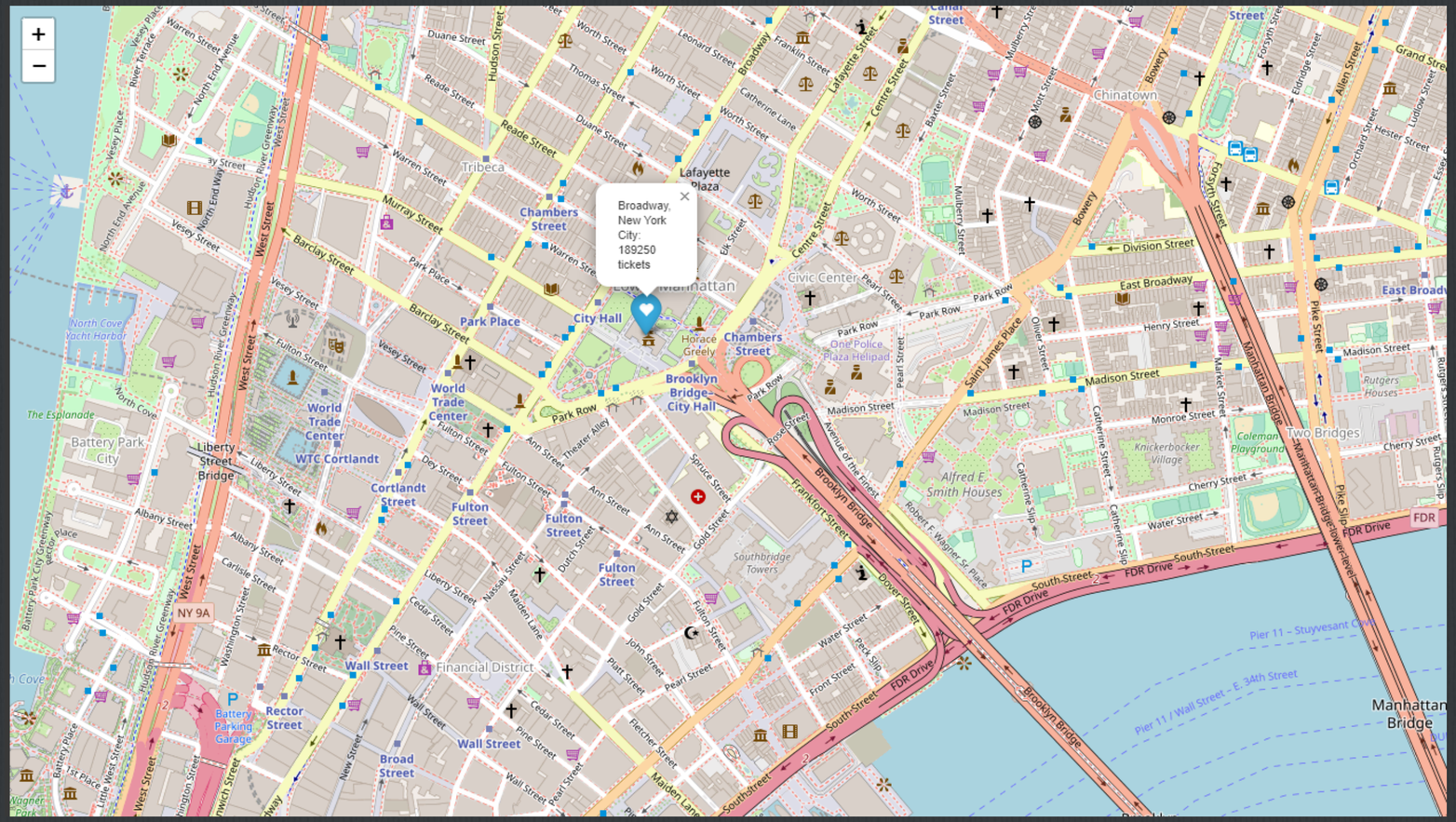
Amount collected for the three violation codes with maximum tickets
As per the website, the average prices for the three violation codes are as follows:

- For violation code 21 = (65 + 45)/2 = $55
- For violation code 36 = (50 + 50)/2 = $50
- For violation code 38 = (65 + 35)/2 = $50

Money collected - 179076175

NYU TANDON SCHOOL OF ENGINEERING

Street with the highest number of Violations

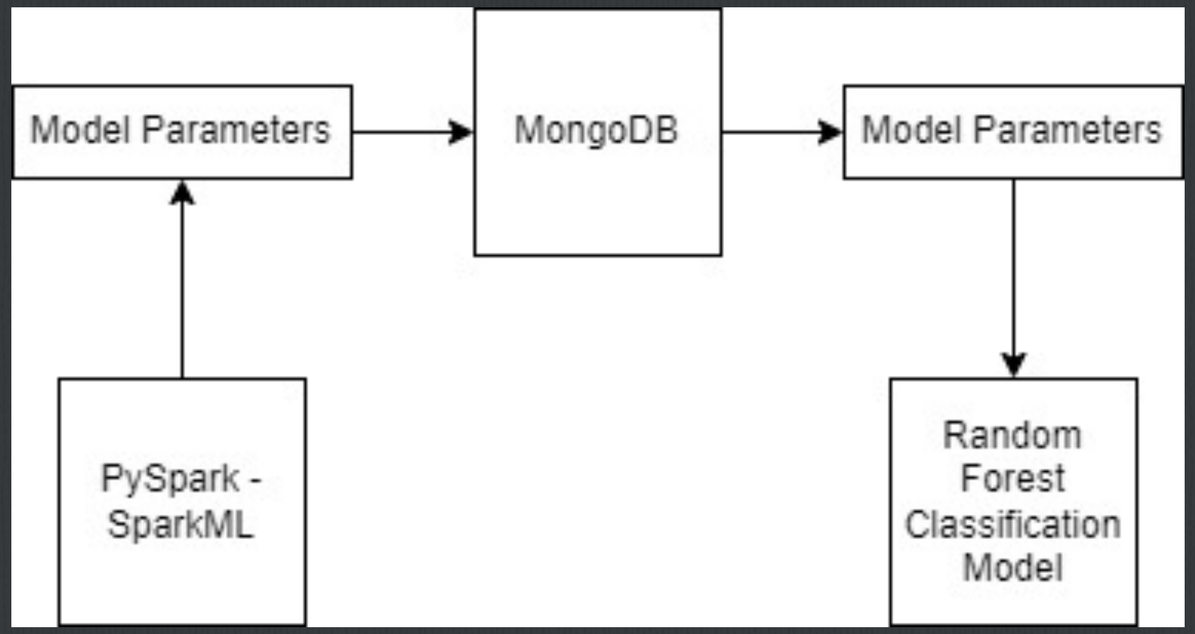


Predicting the most likely type of violation based on the available features
Multiclass classification model using random forest

```
# Convert string columns to index
string_cols = ['Plate Type', 'Vehicle Body Type', 'Vehicle Make', 'Registration State']
indexers = [StringIndexer(inputCol=col, outputCol=col+"_encoded").fit(df) for col in string_cols]
pipeline = Pipeline(stages=indexers)
parking_df = pipeline.fit(df).transform(parking_df)

# Select relevant features and target variable
features = ['Plate Type_encoded', 'Vehicle Body Type_encoded', 'Vehicle Make_encoded', 'Violation Location', 'Registration State_encoded', 'Vehicle Year']
target = 'Violation Code'
assembler = VectorAssembler(inputCols=features, outputCol="features")
df = assembler.transform(df).select("features", target)

# Split the data into training and testing sets
(training_data, testing_data) = df.randomSplit([0.8, 0.2], seed=42)

# Train a random forest classifier
clf = RandomForestClassifier(numTrees=100, seed=42, labelCol=target, featuresCol="features")
model = clf.fit(training_data)

# Make predictions on the test set
predictions = model.transform(testing_data)

# Evaluate the accuracy of the model
```

NYU TANDON SCHOOL OF ENGINEERING

Data Flow diagram



MongoDB

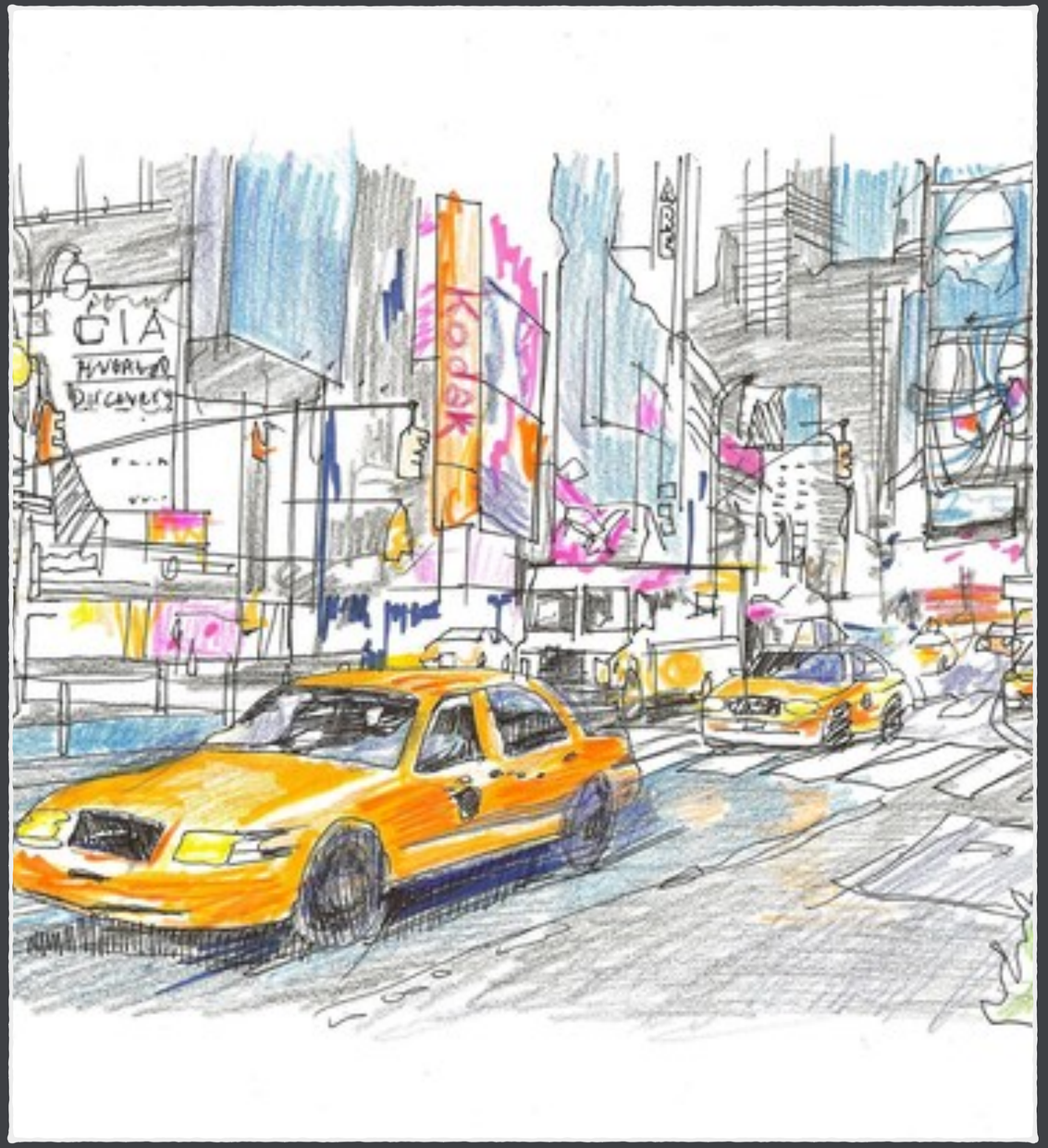
```
as17321_db> db.project.find({})
[
  {
    _id: ObjectId("645c8368a9160a666dba8ffc"),
    model_name: 'parking_violation_rf',
    model_type: 'Random Forest',
    model_params: { numTrees: 100, maxDepth: 5, seed: 42 },
    feature_cols: [
      'Plate Type_encoded',
      'Vehicle Body Type_encoded',
      'Vehicle Make_encoded',
      'Violation Location',
      'Registration State_encoded',
      'Vehicle Year'
    ],
    target_col: 'Violation Code',
    model_object: {
      type: 'RandomForestClassificationModel',
      numFeatures: 6,
      numTrees: 100,
      trees: [
        {
          algo: 'Classification',
          depth: 5,
          numNodes: 63,
          nodes: [
            {
              id: 0,
              isLeaf: false,
              leftNodeId: 1,
              rightNodeId: 8,
              split: [Object]
            },
            {
              id: 1,
              isLeaf: false,
              leftNodeId: 2,
              rightNodeId: 5,
              split: [Object]
            }
          ],
          weights: [ 0.05, 0.05 ]
        }
      ],
      treeWeights: [ 1, 1 ]
    }
  }
]
as17321_db>
```

```
doc = NYCPred.find_one({"model_name": "parking_violation_rf"})
model_byte_string = doc["model_object"]
model = pickle.loads(model_byte_string)
predictions = model.transform(test_data)
predictions_df = predictions.select("prediction").toPandas()
```

NYU TANDON SCHOOL OF ENGINEERING

Inferences

Trends and Patterns in NYC Parking



- Total Number of tickets issued for the year 2022 is 7364889
- March 2022 had the highest parking tickets issued followed by April and May
- There are 64 unique states where the cars that got parking tickets came from.
- New York has the highest parking violations followed by New Jersey and Pennsylvania.
- Plate Id 40404jg had the maximum parking violations.
- The top 3 violation codes are 21, 36 and 38 . Code 21 - No parking where parking is not allowed by sign, street marking or traffic control device. Code 36 - Exceeding the posted speed limit in or near a designated school zone. Code 38 - Parking Meter - Failing to show a receipt or tag in the windshield.Drivers get a 5-minute grace period past the expired time on parking meter receipts.

- For Vehical Body Type, maximum parking violations happen for Suburban(SUBN) followed by four door sedan(4DSD) and Van For Vehicle Make, maximum parking violations happen for Honda follwed by Toyota and Ford

- Maximum Ticket Frequency occur in Spring followed by Winter, Summer and Autumn. Autumn has the least Ticket Frequency. Most commonly occuring violation codes during Spring, Winter and Summer are 21,36 and 38.Most commonly occuring violation codes during Autumn are98,40 and 46

- The highest fine amount of $108,648,700 was for violation code 36.

- A total fine amount of $ 179,076,175 was collected for the three violation codes 21, 36 and 38

- The top 3 violation precincts and Issuer Precincts where maximum parking violations happen are 19, 114 and 1 3

- The top 3 violation codes across precincts 19, 114 and 13 where the parking tickets were issued are 38,21 and 20

- The most common time of the day that violations occur for violation codes 21, 36 and 38 are between 8:00 AM to 12:00 PM followed by 12:00 PM

FUTURE SCOPE AND CONCLUSION

- The future scope for the project could be extended by collecting the violation status along with other data and using machine learning algorithms to apply more severe charges as time passes. This can help in creating a dynamic system that adjusts the charges based on the violator's payment history.

- Moreover, the machine learning model can be trained to identify the repeat violators and track their payment history to determine the severity of the charges. This can help in reducing the number of repeat offenders, which would save time and money for the city.

- Route Optimization:Parking violations can often lead to traffic congestion. The dataset can be used to develop traffic optimization algorithms that can route traffic around congested areas and reduce traffic buildup. By analyzing the data on parking violations and occupancy rates, parking authorities can optimize routes for parking enforcement officers. This can reduce the time it takes to issue a parking ticket and improve the efficiency of parking enforcement.

- The dataset can be used to develop parking management systems that can optimize parking space allocation and reduce the number of parking violations.

- Explainability: The project could explore techniques for model explainability, such as SHAP values or LIME, to better understand the factors that contribute to predictions and improve interpretability of the model.