STEVENS
INSTITUTE OF TECHNOLOGY
1870

NYC MOTOR VEHICLE COLLISION ANALYTICS AND PREDICTIVE MODELING

BIA-678-WS
Prof. Venu Guntupalli

TEAM A

Sneha Dharne
Gunik Luthra
Neha Umathe

Table Of Contents

1. Project Overview
2. Problem Statement & Objective
3. Methodology
    a. Dataset
    b. Data Preparation
        i. Handle Null Values
        ii. Feature Selection
        iii. Data Joining
    c. Exploratory Data Analysis
    d. Data Modelling : Spatial Clustering
    e. Data Modelling : Risk Level Prediction
        i. Scaling

Project Overview

- Vehicle collisions in NYC represent a significant aspect of urban transportation dynamics, impacting both public safety and infrastructure management.

- With its densely populated streets and bustling traffic, NYC experiences a high frequency of motor vehicle collisions, ranging from minor to severe accidents resulting in injuries or fatalities.

- The consequences of these collisions extend beyond immediate injuries or damage, affecting traffic flow, emergency response services, and community well-being.

- Understanding the patterns and factors contributing to vehicle collisions in NYC is crucial for implementing effective safety measures, improving transportation infrastructure, and ultimately enhancing the quality of life for residents and visitors alike.

Objective & Problem Statements

New York City experiences a high volume of motor vehicle collisions. Today, we present a big data project leveraging NYC Open Data to analyze these collisions.

This project dives into three key datasets: collisions, vehicles, and people involved. By analyzing these interconnected datasets, we aim to gain insights into various aspects of NYC traffic accidents, including:

- Accident Patterns (Factors) : Highlighting trends in accident times, vehicle types involved, pre-accident actions, location of the victim, contributing factors, etc
- Impact Analysis (Consequences): Understanding the types of public property damaged and harm to human life
- Spatial Distribution (Location Clustering): Examining collision distribution across boroughs, identifying potential hotspots and assigning weights.
- Predictive Modeling: Developing models to predict human life loss and injuries in high-risk areas.

This project aims to provide valuable data-driven insights to improve road safety and inform traffic management strategies in New York City.

Methodology

Dataset

CRASHES.CSV

Data till : April 4 2024

COLUMNS : 29

ROWS : 2077590

VEHICLES.CSV

Data till : April 4 2024

COLUMNS : 25

ROWS : 4169890

PEOPLE.CSV

Data till : April 4 2024

COLUMNS : 21

ROWS : 5326831

Dataset

CRASHES.CSV:

```
|-- CRASH DATE: string (nullable = true)
|-- CRASH TIME: string (nullable = true)
|-- BOROUGH: string (nullable = true)
|-- ZIP CODE: string (nullable = true)
|-- LATITUDE: string (nullable = true)
|-- LONGITUDE: string (nullable = true)
|-- LOCATION: string (nullable = true)
|-- ON STREET NAME:  string (nullable = true)
|-- CROSS STREET NAME:  string (nullable = true)
|-- OFF STREET NAME:  string (nullable = true)
|-- NUMBER OF PERSONS INJURED: string (nullable = true)
|-- NUMBER OF PERSONS KILLED: string (nullable = true)
|-- NUMBER OF PEDESTRIANS INJURED: string (nullable = true)
|-- NUMBER OF PEDESTRIANS KILLED: string (nullable = true)
|-- NUMBER OF CYCLIST INJURED: string (nullable = true)
|-- NUMBER OF CYCLIST KILLED: string (nullable = true)
|-- NUMBER OF MOTORIST INJURED: string (nullable = true)
|-- NUMBER OF MOTORIST KILLED: string (nullable = true)
|-- CONTRIBUTING FACTOR VEHICLE 1: string (nullable = true)
|-- CONTRIBUTING FACTOR VEHICLE 2: string (nullable = true)
|-- CONTRIBUTING FACTOR VEHICLE 3: string (nullable = true)
|-- CONTRIBUTING FACTOR VEHICLE 4: string (nullable = true)
|-- CONTRIBUTING FACTOR VEHICLE 5: string (nullable = true)
|-- COLLISION_ID: string (nullable = true)
|-- VEHICLE TYPE CODE 1: string (nullable = true)
|-- VEHICLE TYPE CODE 2: string (nullable = true)
|-- VEHICLE TYPE CODE 3: string (nullable = true)
|-- VEHICLE TYPE CODE 4: string (nullable = true)
|-- VEHICLE TYPE CODE 5: string (nullable = true)
```

Dataset

VEHICLES.CSV:

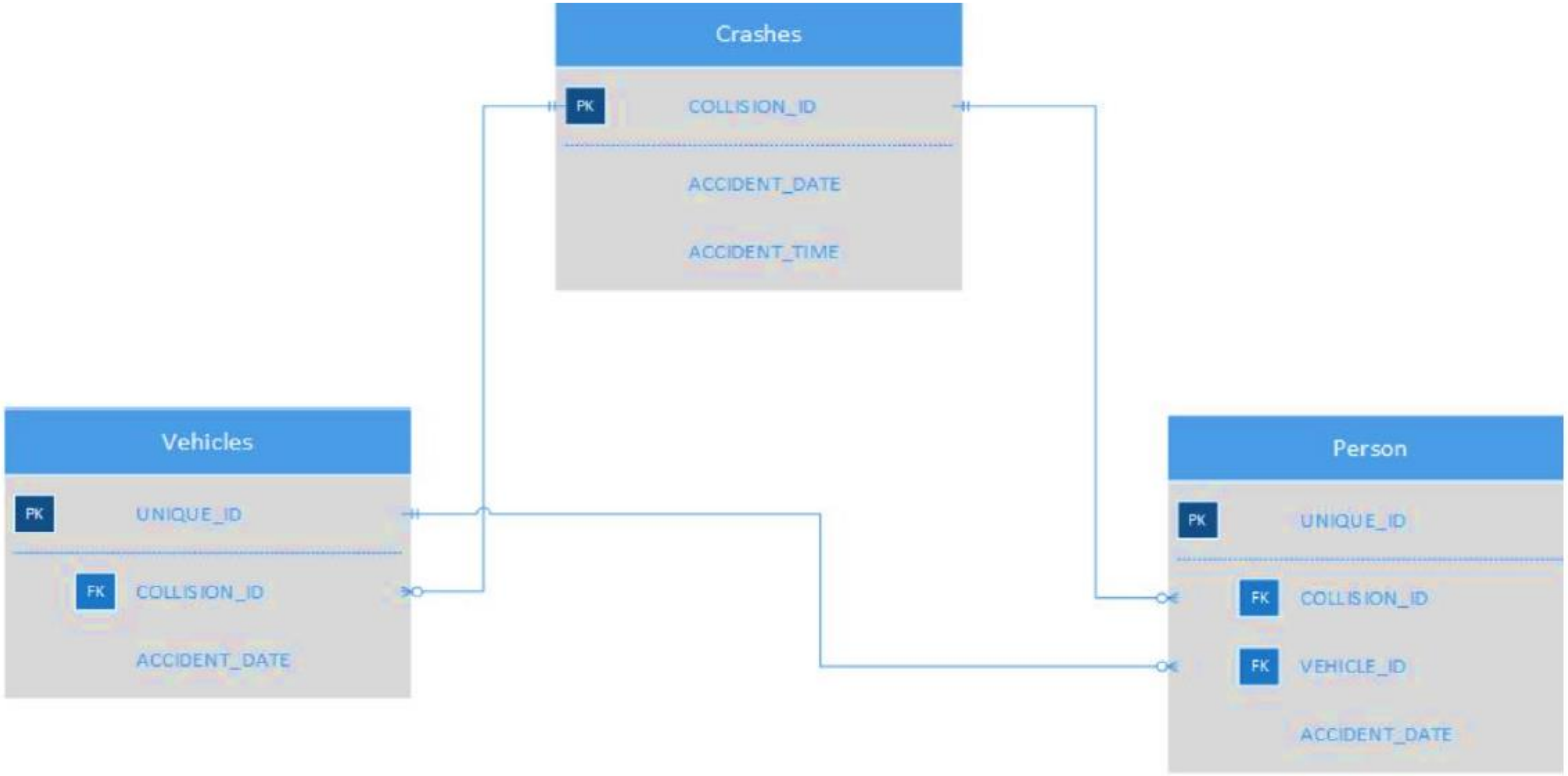
```
|-- UNIQUE_ID: string (nullable = true)
|-- COLLISION_ID: string (nullable = true)
|-- CRASH_DATE: string (nullable = true)
|-- CRASH_TIME: string (nullable = true)
|-- VEHICLE_ID: string (nullable = true)
|-- STATE_REGISTRATION: string (nullable = true)
|-- VEHICLE_TYPE: string (nullable = true)
|-- VEHICLE_MAKE: string (nullable = true)
|-- VEHICLE_MODEL: string (nullable = true)
|-- VEHICLE_YEAR: string (nullable = true)
|-- TRAVEL_DIRECTION: string (nullable = true)
|-- VEHICLE_OCCUPANTS: string (nullable = true)
|-- DRIVER_SEX: string (nullable = true)
|-- DRIVER_LICENSE_STATUS: string (nullable = true)
|-- DRIVER_LICENSE_JURISDICTION: string (nullable = true)
|-- PRE_CRASH: string (nullable = true)
|-- POINT_OF_IMPACT: string (nullable = true)
|-- VEHICLE_DAMAGE: string (nullable = true)
|-- VEHICLE_DAMAGE_1: string (nullable = true)
|-- VEHICLE_DAMAGE_2: string (nullable = true)
|-- VEHICLE_DAMAGE_3: string (nullable = true)
|-- PUBLIC_PROPERTY_DAMAGE: string (nullable = true)
|-- PUBLIC_PROPERTY_DAMAGE_TYPE: string (nullable = true)
|-- CONTRIBUTING_FACTOR_1: string (nullable = true)
|-- CONTRIBUTING_FACTOR_2: string (nullable = true
```

Dataset

PERSONS.CSV:

```
|-- UNIQUE_ID: string (nullable = true)
|-- COLLISION_ID: string (nullable = true)
|-- CRASH_DATE: string (nullable = true)
|-- CRASH_TIME: string (nullable = true)
|-- PERSON_ID: string (nullable = true)
|-- PERSON_TYPE: string (nullable = true)
|-- PERSON_INJURY: string (nullable = true)
|-- VEHICLE_ID: string (nullable = true)
|-- PERSON_AGE: string (nullable = true)
|-- EJECTION: string (nullable = true)
|-- EMOTIONAL_STATUS: string (nullable = true)
|-- BODILY_INJURY: string (nullable = true)
|-- POSITION_IN_VEHICLE: string (nullable = true)
|-- SAFETY_EQUIPMENT: string (nullable = true)
|-- PED_LOCATION: string (nullable = true)
|-- PED_ACTION: string (nullable = true)
|-- COMPLAINT: string (nullable = true)
|-- PED_ROLE: string (nullable = true)
|-- CONTRIBUTING_FACTOR_1: string (nullable = true)
|-- CONTRIBUTING_FACTOR_2: string (nullable = true)
|-- PERSON_SEX: string (nullable = true)
```

Dataset - ER Diagram



DATA PREPARATION

- Handling Null Values:
  - For the initial EDA, checked for Null values
  - As all the columns were strings, converted to the required format
  - Removed the null values where most of the data was null
  - Labelled Null as unknown or unspecified in CONTRIBUTING FACTOR,PRE_CRASH,VEHICLE_TYPE.

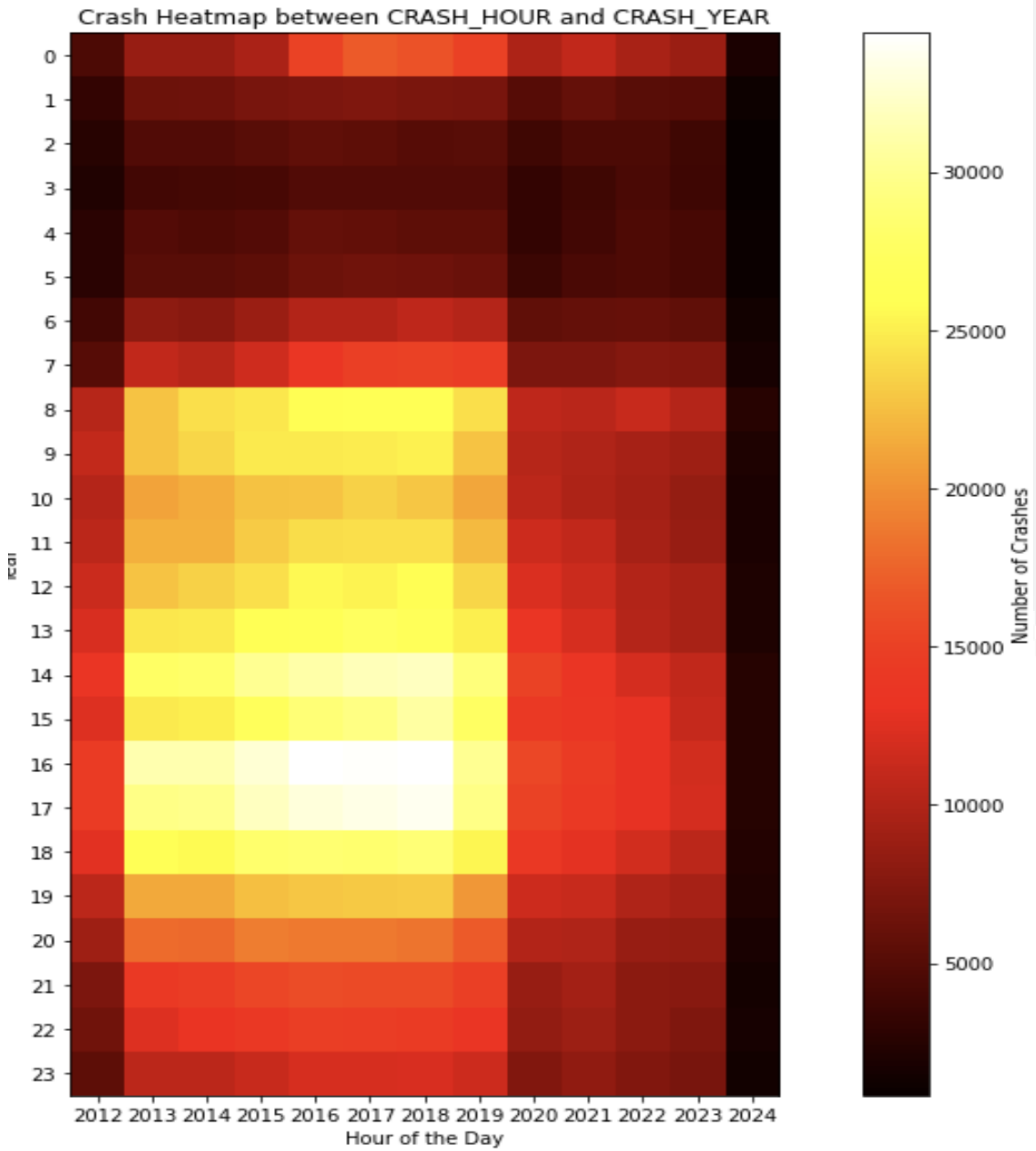
- Data Joined:
  - Left Joined the tables Vehicles and Crashes on vehicles as we took Vehicle as the center table.

- Feature Selection:
  - Selected the features that showed trend and information for the predictive modelling.
  - Based on the EDA, we transformed the columns with many values into buckets of values, with buckets of similar types.

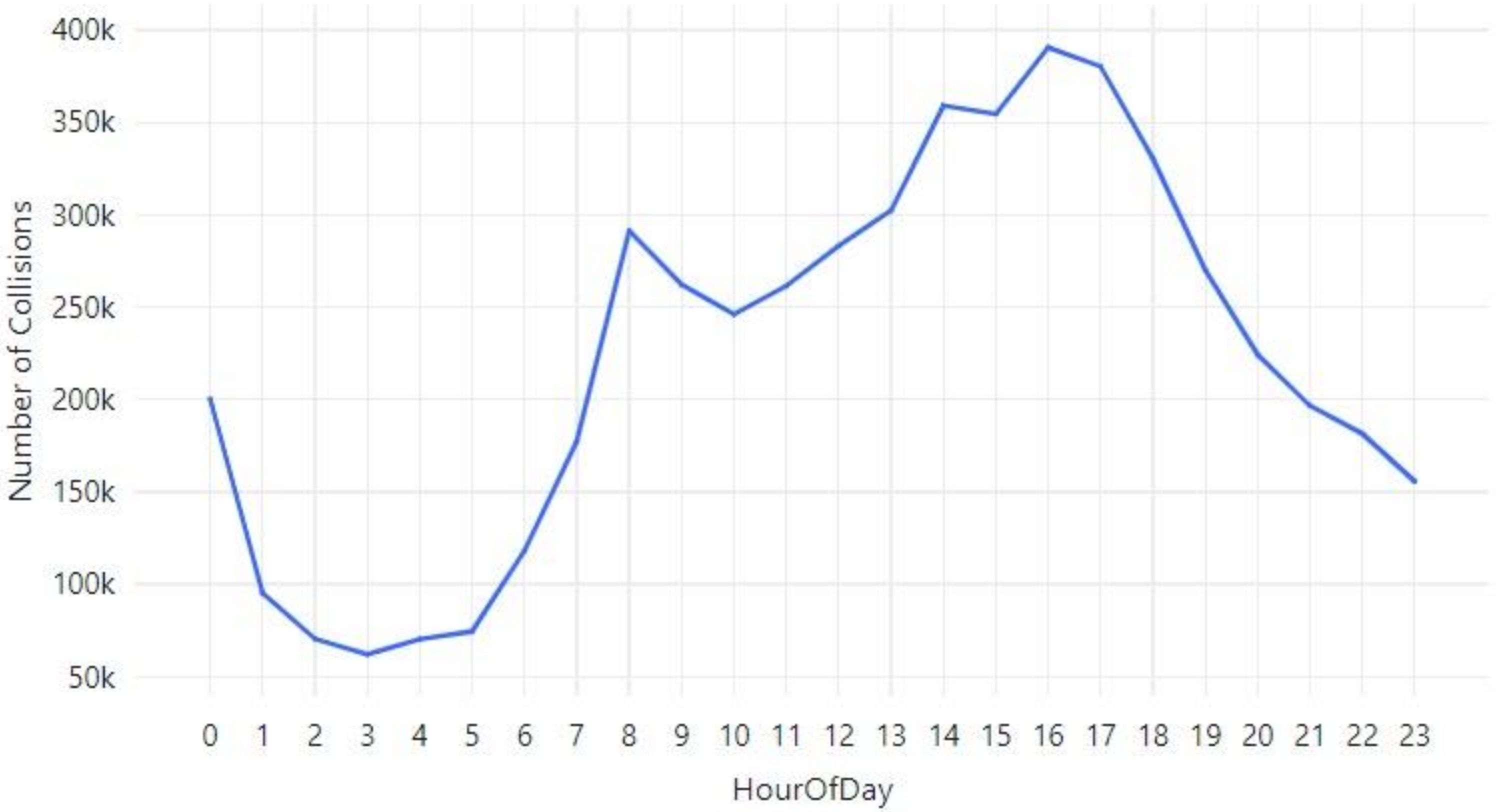
Exploratory Data Analysis

The heatmap shows the distribution of crashes on hours of day with respect to crash year

The highest number of accidents take place in daytime during 2013 to 2019



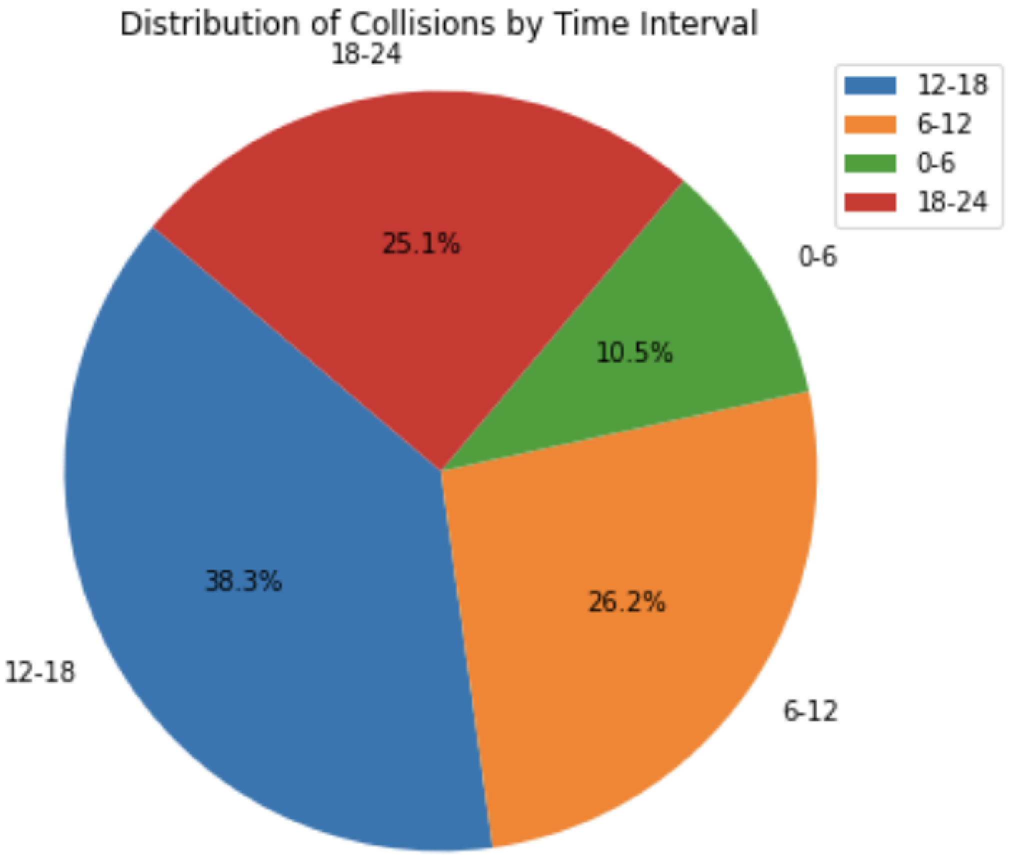
Exploratory Data Analysis



The line graph shows the number of collisions vs hour of the day.

- We notice the uneven distribution.
- Most collisions happen during 12:00 to 18:00.

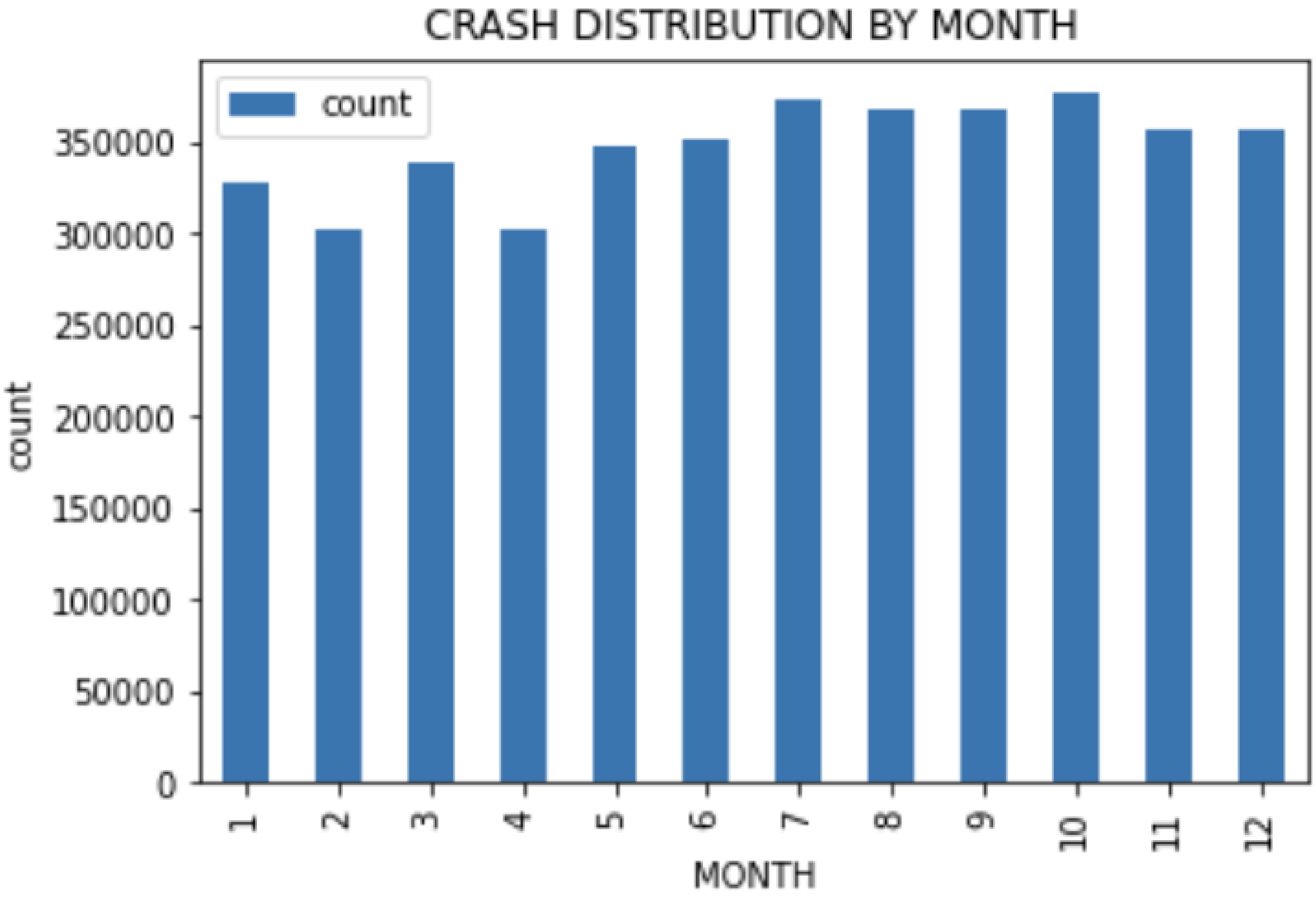
Exploratory Data Analysis



- The Piechart shows the distribution of crashes by time interval of a day

- The highest number of accidents take place during 12:00 and 18:00

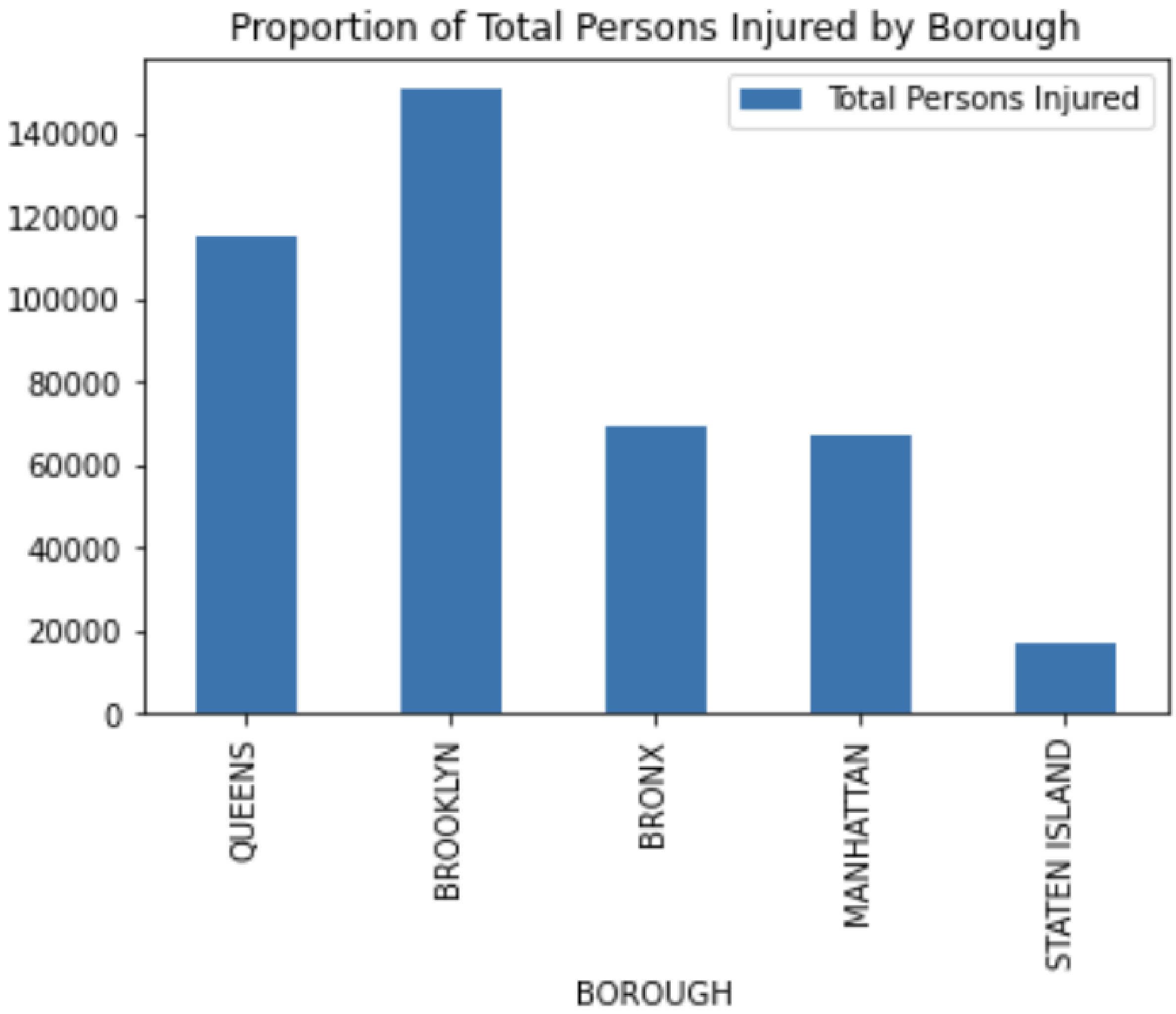
Exploratory Data Analysis



The Bar Graph shows crash distribution by month

We notice that the distribution is almost equal for every month, hence we decided not to take this as a feature.

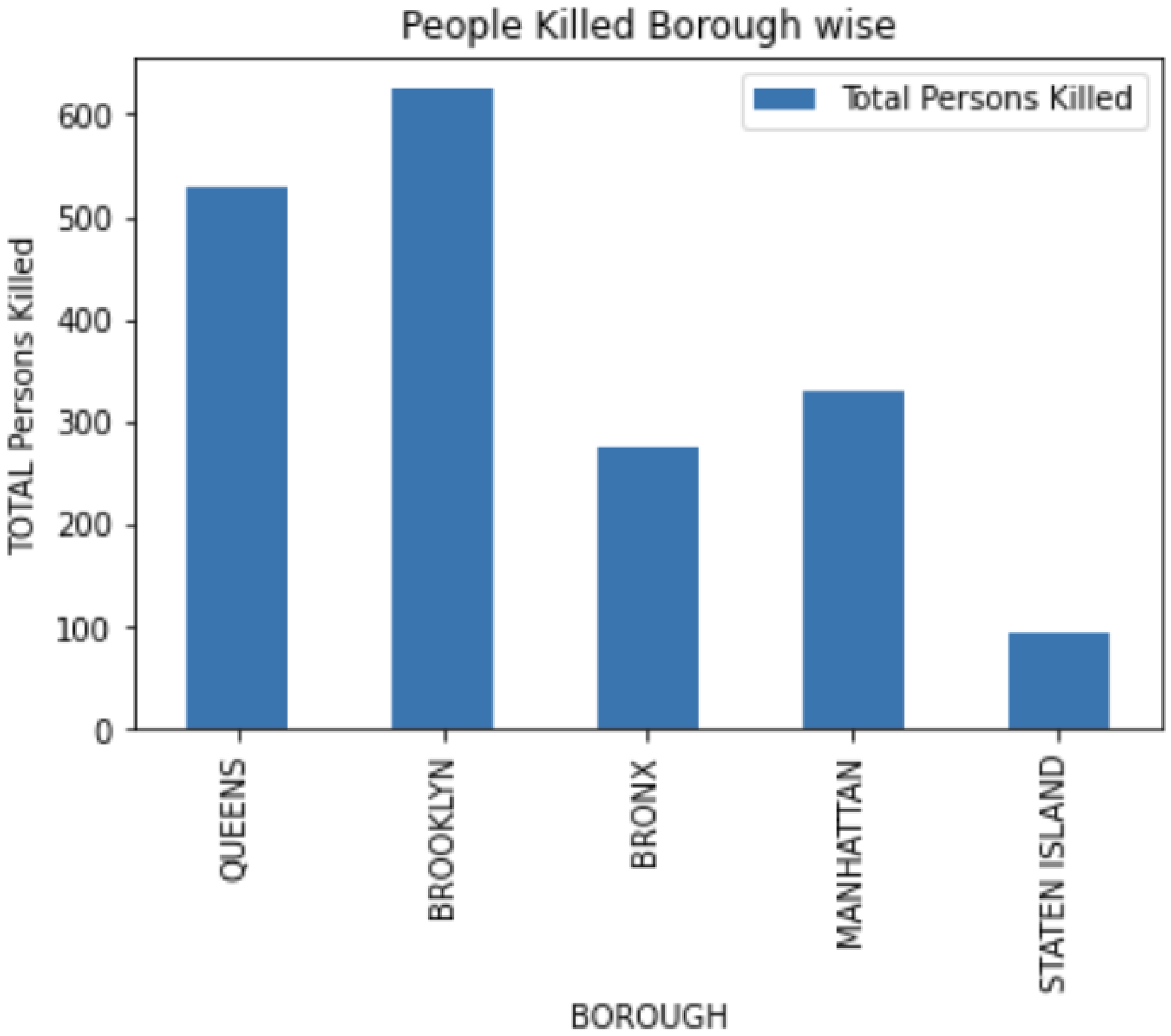
Exploratory Data Analysis



- The Bar Graph shows Proportion of Total Persons injured by Borough

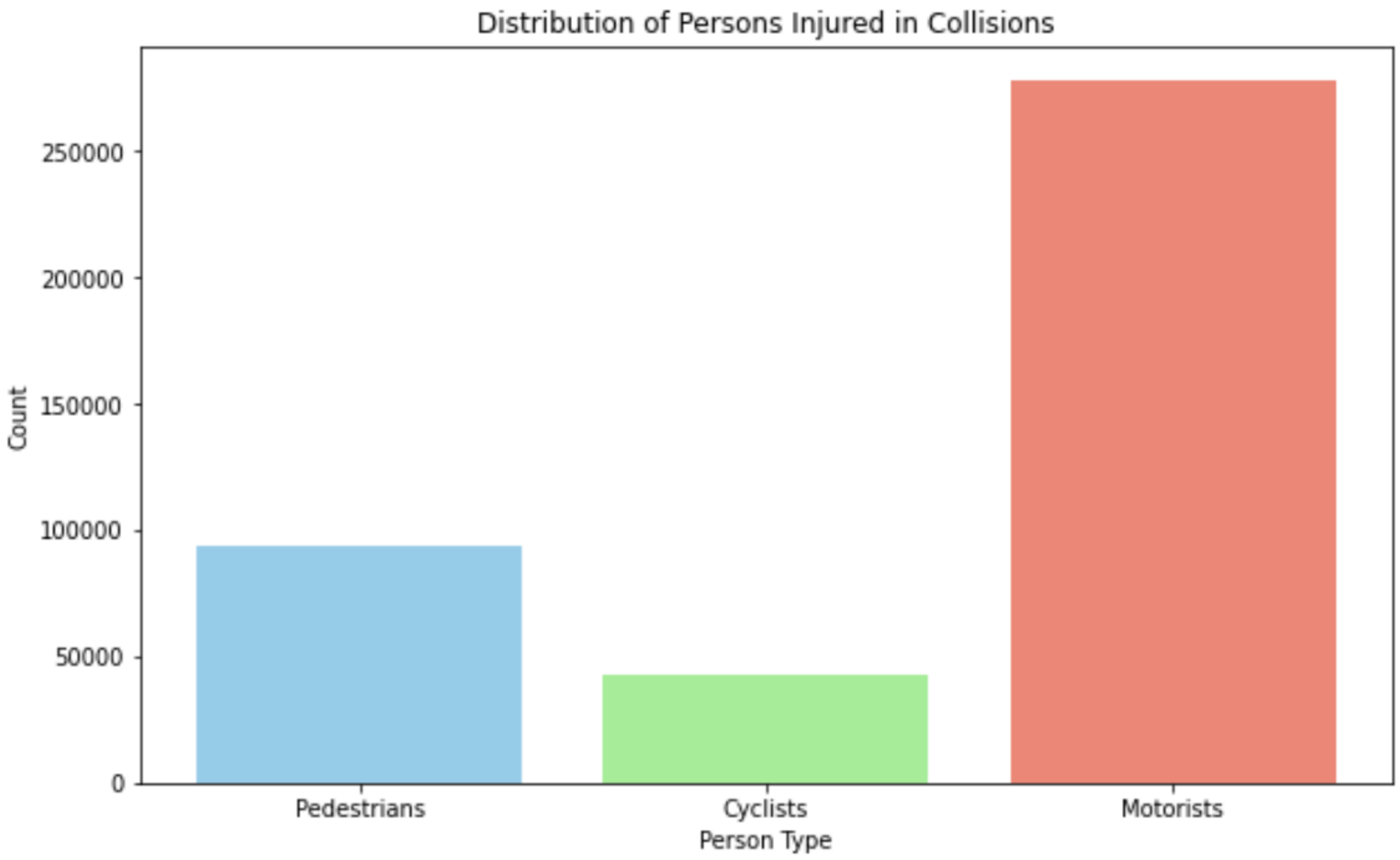
- Brooklyn has highest injuries while Staten Island has the lowest

Exploratory Data Analysis



- The Bar Graph shows Proportion of Total Persons killed by Borough
- Brooklyn has highest injuries while Staten Island has the lowest

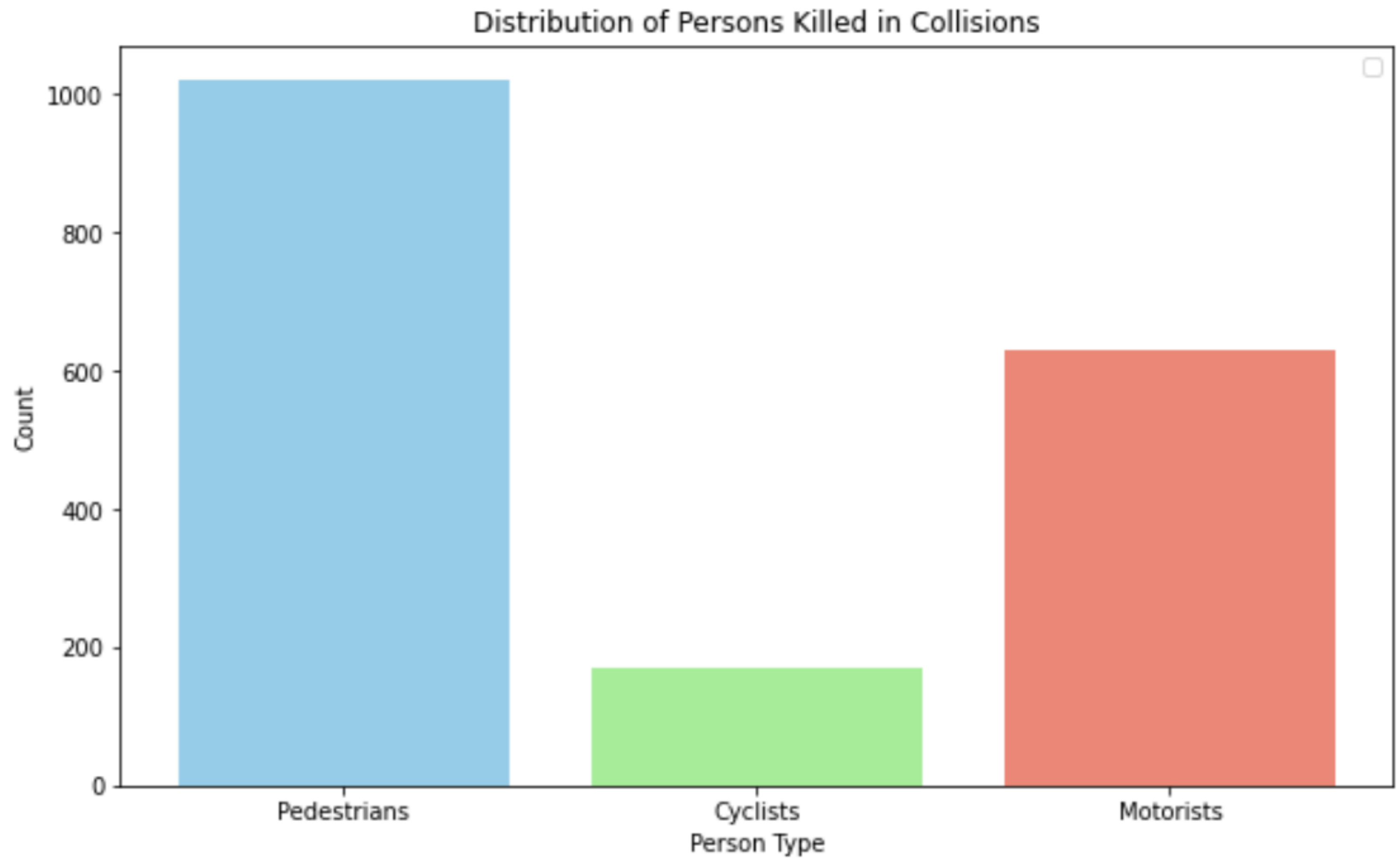
Exploratory Data Analysis



- The Bar Graph shows Proportion of Total Persons injured by Person_type

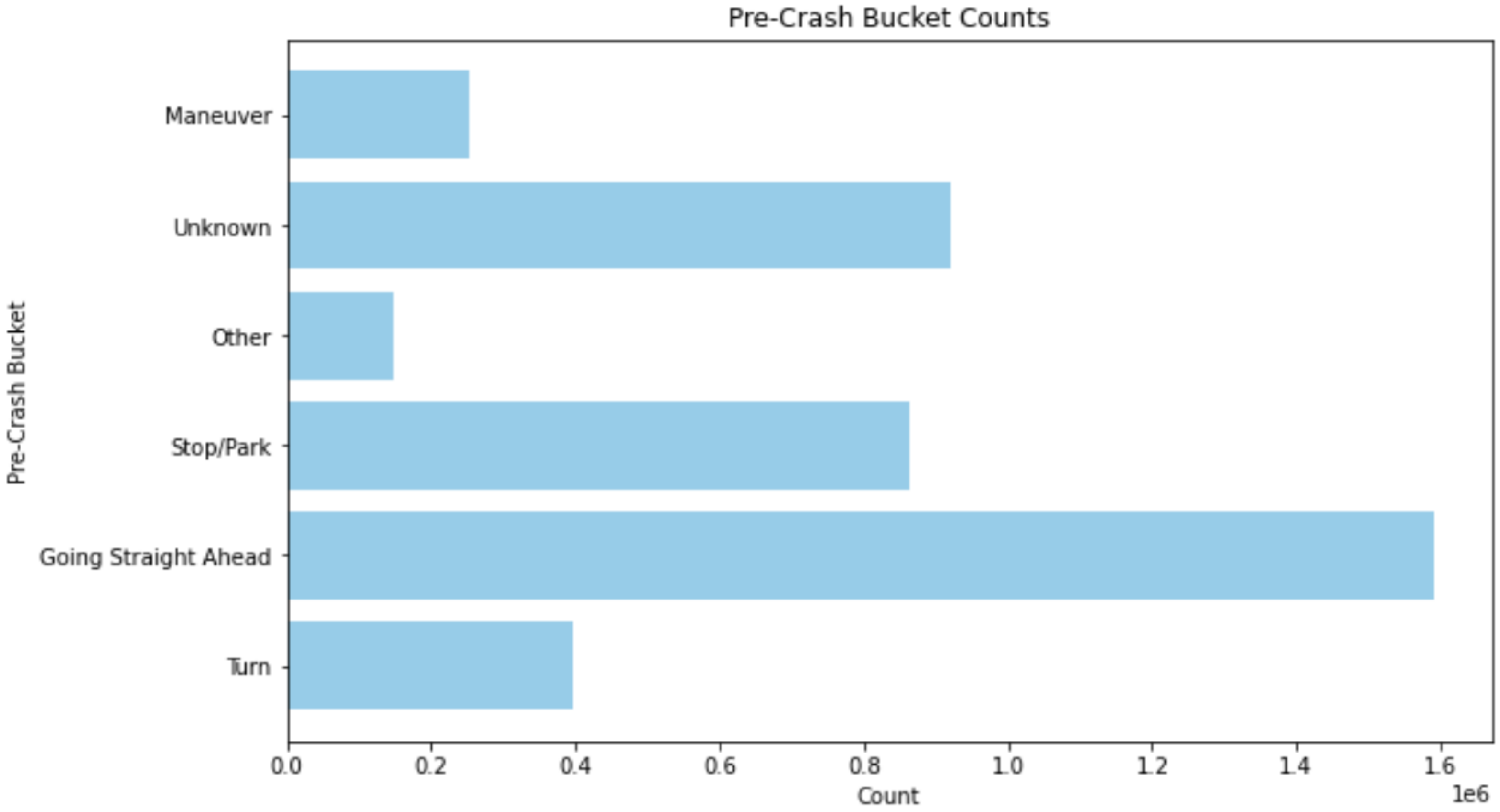
- Motorist has highest injuries while cyclist has the lowest

Exploratory Data Analysis



- The Bar Graph shows Proportion of Total Persons killed by Person_type
- Predestrians has highest fatalities while cyclist has the lowest

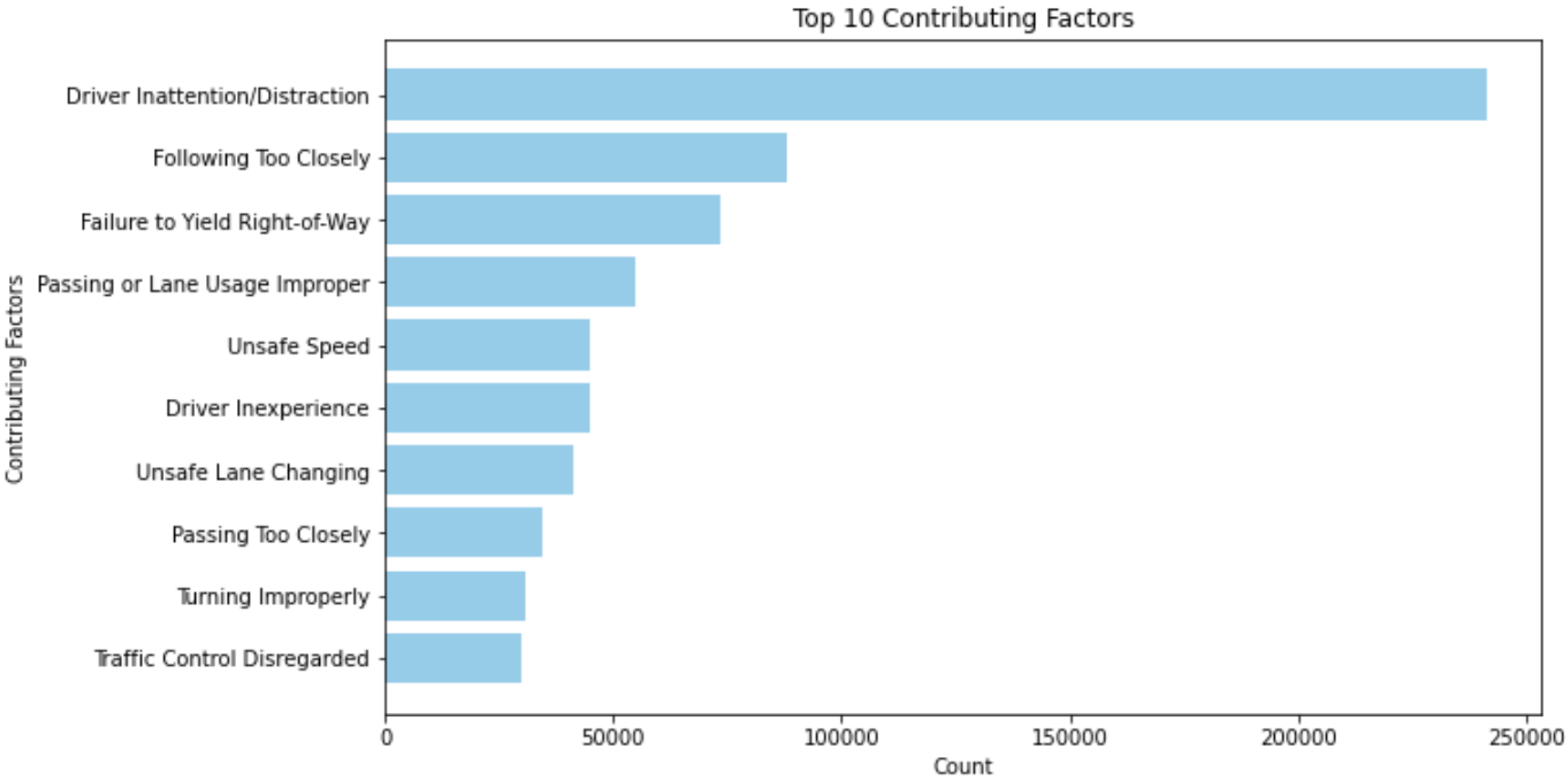
Exploratory Data Analysis



Observations:

The Bar Graph shows pre-crash action count

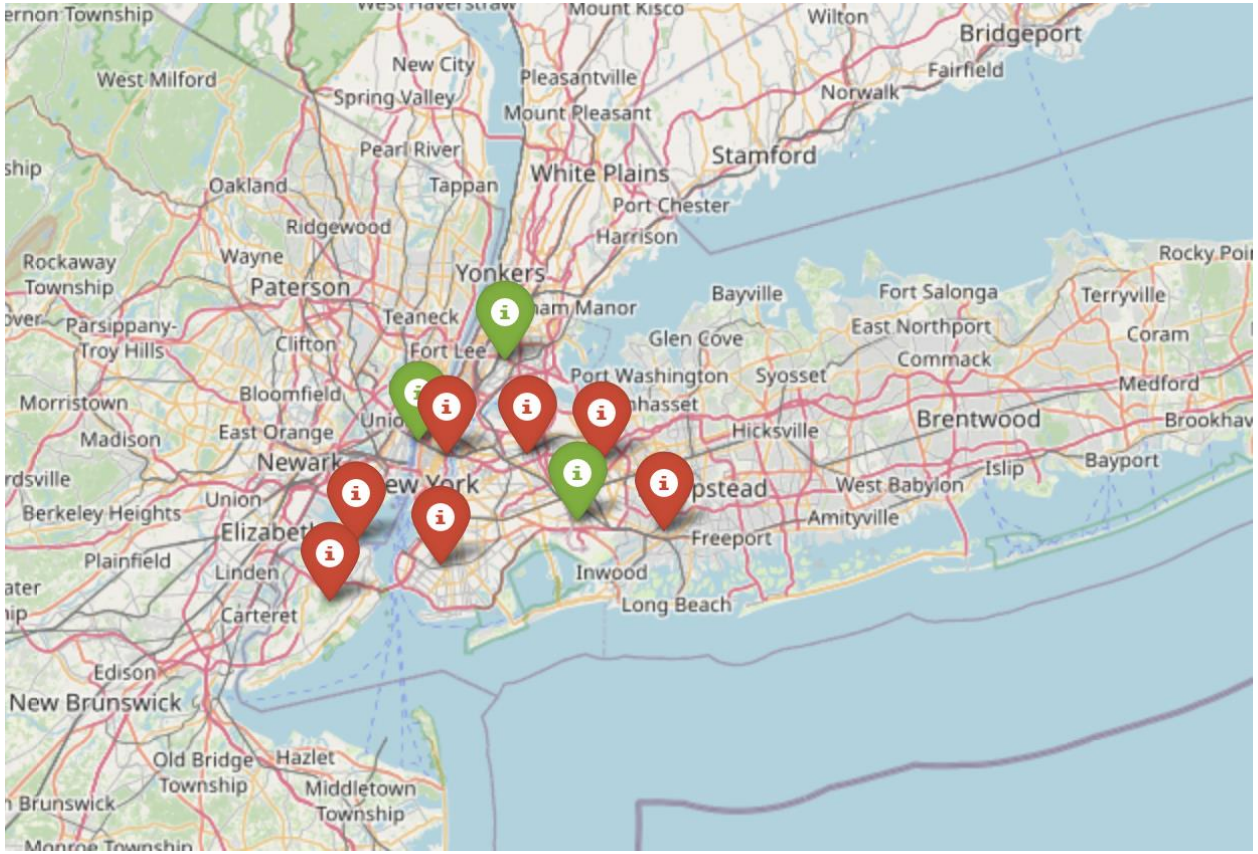
Exploratory Data Analysis



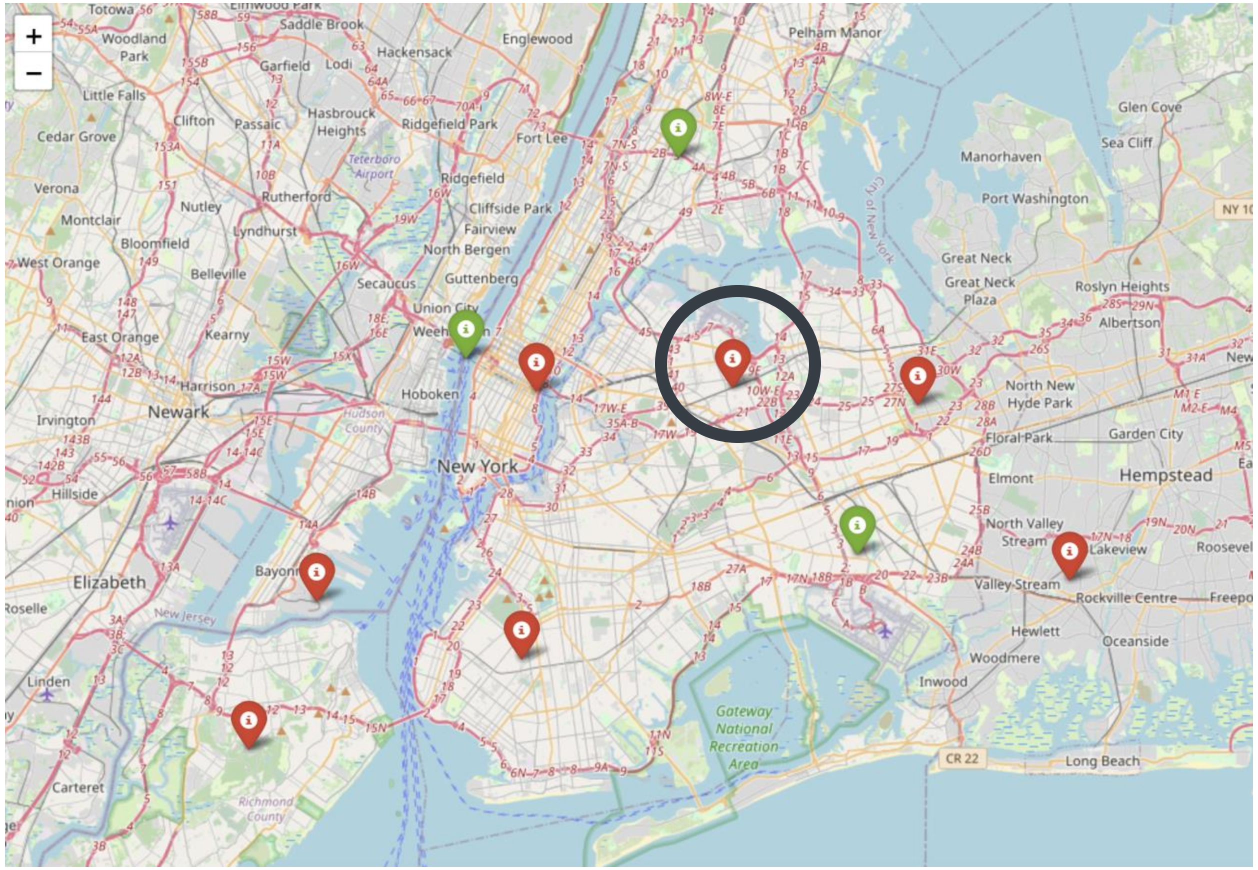
DATA MODELLING : SPATIAL CLUSTERING

- Crashes.csv ->
- LOCATION, ON STREET NAME, CROSS STREET NAME, OFF STREET NAME, COLLISION_ID
- Select rows where at least one of (LOCATION, ON STREET NAME, CROSS STREET NAME, OFF STREET NAME) is not null
- NULL Values of LOCATION -
  -> mode of rows with the same ON STREET NAME, CROSS STREET NAME, OFF STREET NAME
  ->mode of rows with the same ON STREET NAME, CROSS STREET NAME
  ->mode of rows with the same ON STREET NAME
- LOCATION -> Latitude , Longitude
- Latitude = Latitude * 100, Longitude = Longitude *100 [for better expressivity in Euclidean distance]

DATA MODELLING : SPATIAL CLUSTERING



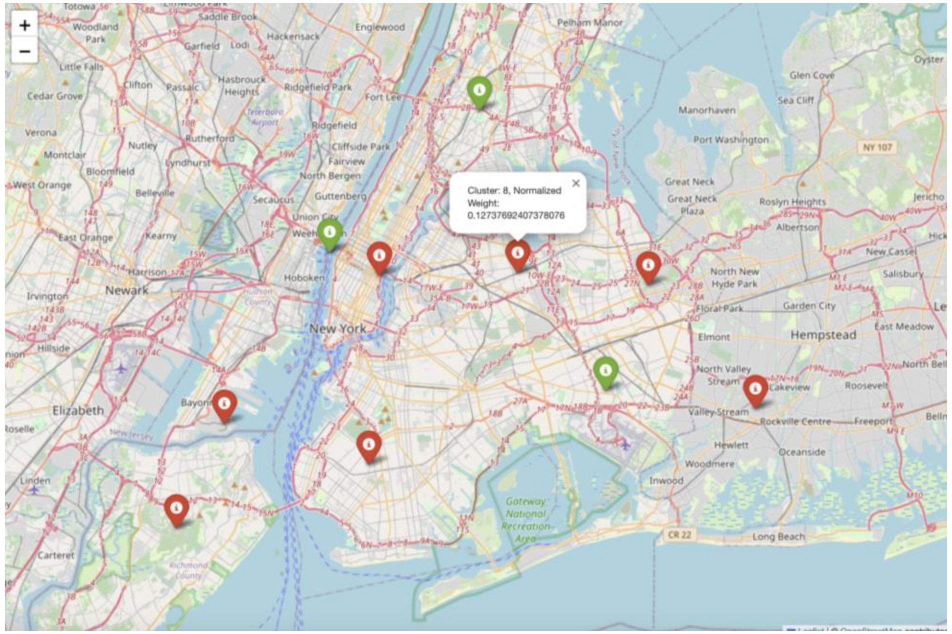
DATA MODELLING : SPATIAL CLUSTERING

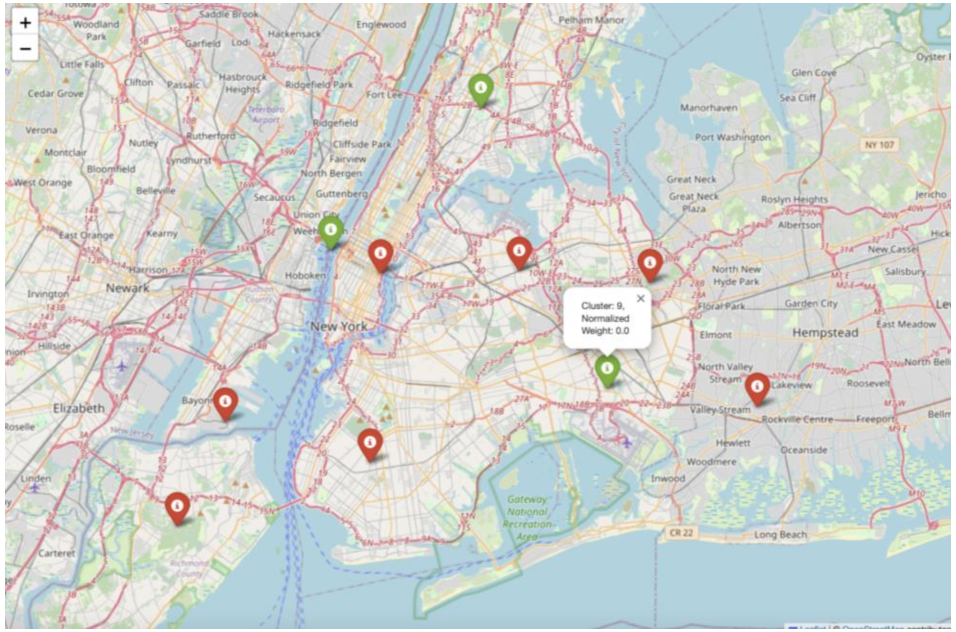


- PROXIMITY
- SEVERITY SCORE

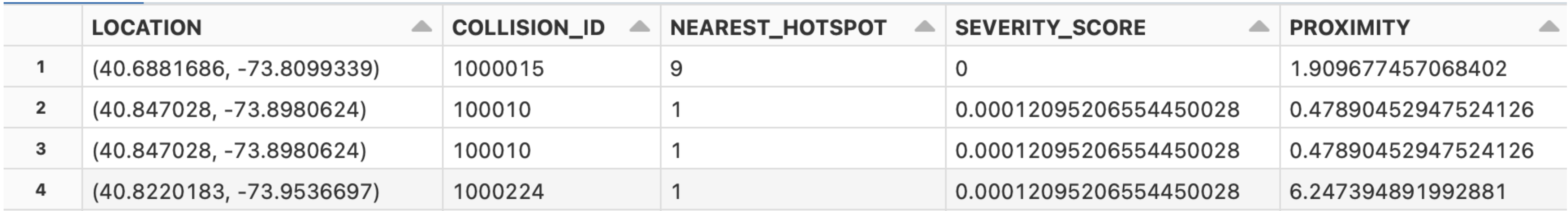
Severity Score:
Ratio of
fatal/ high risk incidents of the cluster : overall fatal / high risk incidents

DATA MODELLING : SPATIAL CLUSTERING





FEATURES FOR PREDICTIVE MODEL



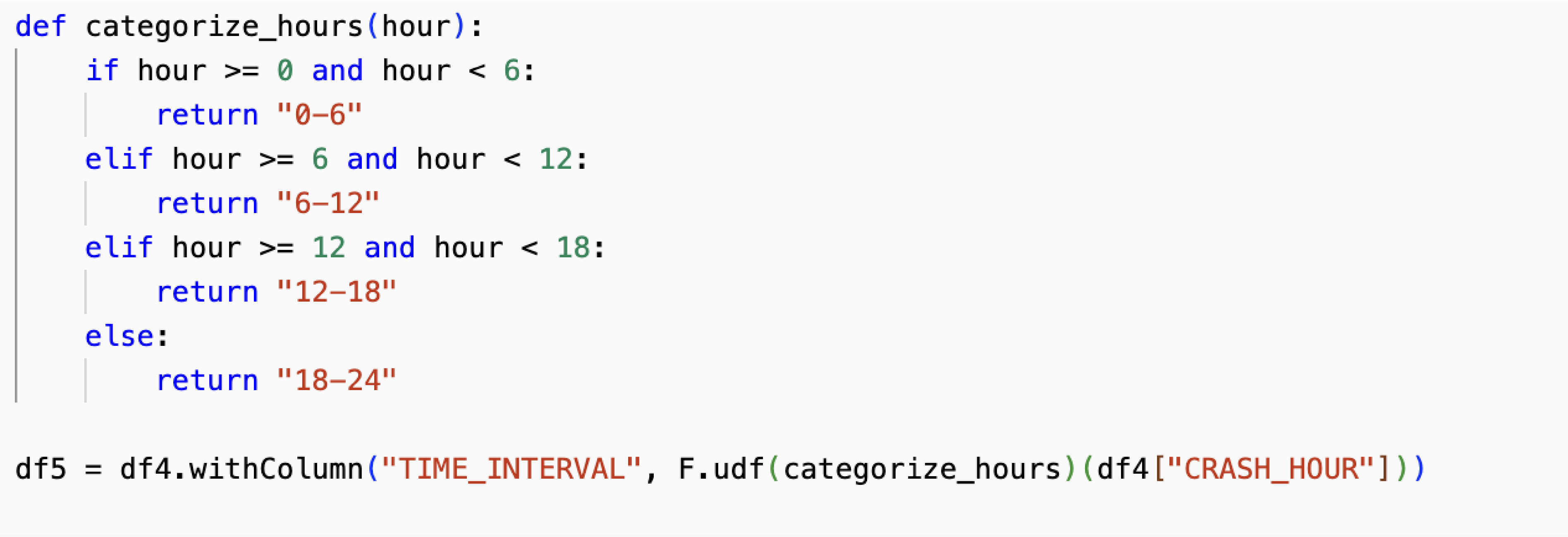
	LOCATION	COLLISION_ID	NEAREST_HOTSPOT	SEVERITY_SCORE	PROXIMITY
1	(40.6881686, -73.8099339)	1000015	9	0	1.909677457068402
2	(40.847028, -73.8980624)	100010	1	0.00012095206554450028	0.47890452947524126
3	(40.847028, -73.8980624)	100010	1	0.00012095206554450028	0.47890452947524126
4	(40.8220183, -73.9536697)	1000224	1	0.00012095206554450028	6.247394891992881

DATA MODELLING : RISK LEVEL PREDICTION

- Using the features derived from the above spatial model
- LOCATION, COLLISION_ID, NEAREST_HOTSPOT, SEVERITY_SCORE, PROXIMITY
- Adding more features to the above features such as VEHICLE_TYPE_BUCKET, CONTRIBUTING_FACTOR_1, CONTRIBUTING_FACTOR_2, PRE_CRASH_BUCKET, TIME_INTERVAL

DATA MODELLING : RISK LEVEL PREDICTION

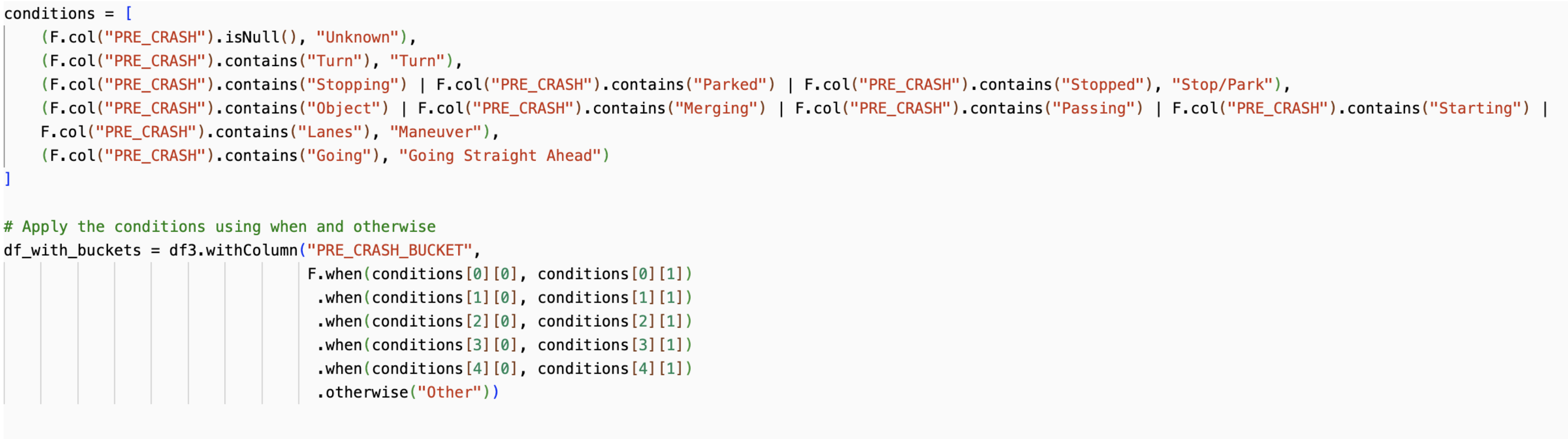
- We decided to take time interval as a feature rather than just the hours.



```python
def categorize_hours(hour):
    if hour >= 0 and hour < 6:
        return "0-6"
    elif hour >= 6 and hour < 12:
        return "6-12"
    elif hour >= 12 and hour < 18:
        return "12-18"
    else:
        return "18-24"

df5 = df4.withColumn("TIME_INTERVAL", F.udf(categorize_hours)(df4["CRASH_HOUR"]))
```

DATA MODELLING : RISK LEVEL PREDICTION

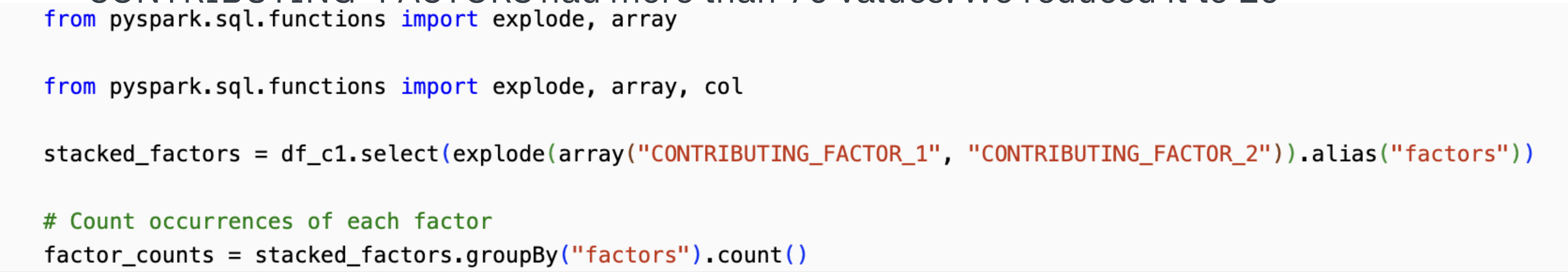
- As Pre_crash had more than 20 values, we had to get it down to 6.



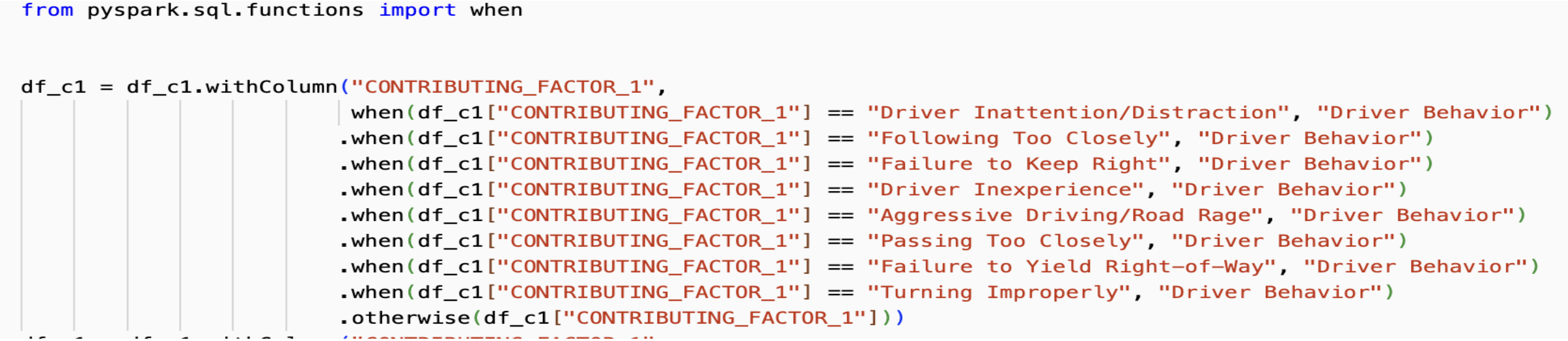
```
conditions = [
    (F.col("PRE_CRASH").isNull(), "Unknown"),
    (F.col("PRE_CRASH").contains("Turn"), "Turn"),
    (F.col("PRE_CRASH").contains("Stopping") | F.col("PRE_CRASH").contains("Parked") | F.col("PRE_CRASH").contains("Stopped"), "Stop/Park"),
    (F.col("PRE_CRASH").contains("Object") | F.col("PRE_CRASH").contains("Merging") | F.col("PRE_CRASH").contains("Passing") | F.col("PRE_CRASH").contains("Starting") |
    F.col("PRE_CRASH").contains("Lanes"), "Maneuver"),
    (F.col("PRE_CRASH").contains("Going"), "Going Straight Ahead")
]

# Apply the conditions using when and otherwise
df_with_buckets = df3.withColumn("PRE_CRASH_BUCKET",
                                 F.when(conditions[0][0], conditions[0][1])
                                  .when(conditions[1][0], conditions[1][1])
                                  .when(conditions[2][0], conditions[2][1])
                                  .when(conditions[3][0], conditions[3][1])
                                  .when(conditions[4][0], conditions[4][1])
                                  .otherwise("Other"))
```

DATA MODELLING : RISK LEVEL PREDICTION

- CONTRIBUTING FACTORS had more than 70 values. We reduced it to 10



```
from pyspark.sql.functions import explode, array

from pyspark.sql.functions import explode, array, col

stacked_factors = df_c1.select(explode(array("CONTRIBUTING_FACTOR_1", "CONTRIBUTING_FACTOR_2")).alias("factors"))

# Count occurrences of each factor
factor_counts = stacked_factors.groupBy("factors").count()
```



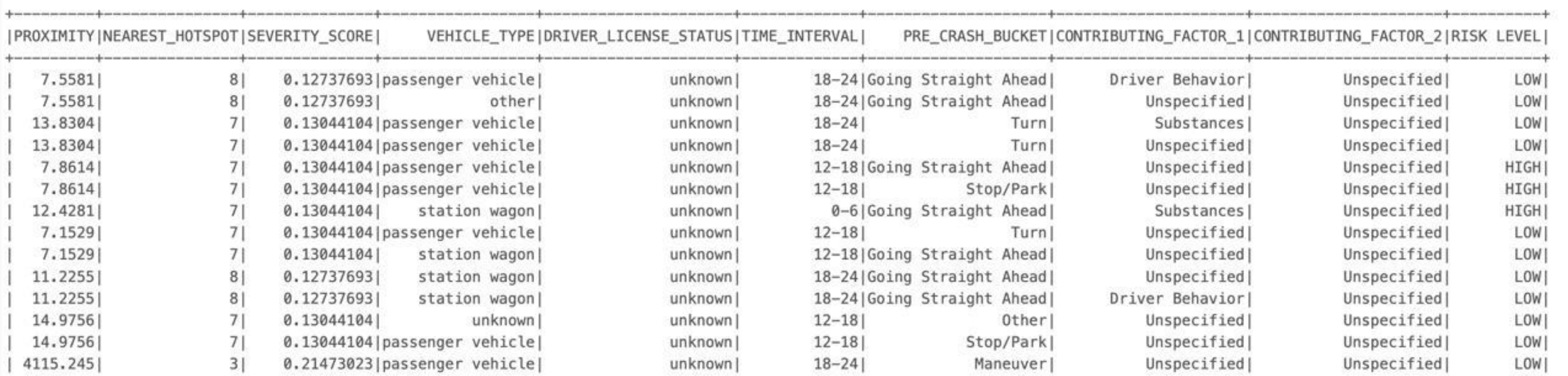
```
from pyspark.sql.functions import when


df_c1 = df_c1.withColumn("CONTRIBUTING_FACTOR_1",
                          when(df_c1["CONTRIBUTING_FACTOR_1"] == "Driver Inattention/Distraction", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Following Too Closely", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Failure to Keep Right", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Driver Inexperience", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Aggressive Driving/Road Rage", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Passing Too Closely", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Failure to Yield Right-of-Way", "Driver Behavior")
                         .when(df_c1["CONTRIBUTING_FACTOR_1"] == "Turning Improperly", "Driver Behavior")
                         .otherwise(df_c1["CONTRIBUTING_FACTOR_1"]))
```

DATA MODELLING : RISK LEVEL PREDICTION

- FOR TARGET VARIABLE:
- Assigned FATAL LEVEL if number of people killed>0
- Assigned HIGH LEVEL if number of persons injured >0
- Assigned MEDIUM LEVEL if Vehicle_damage or Property Damage is Yes
- Assigned LOW otherwise

DATA MODELLING : RISK LEVEL PREDICTION

- FINAL FEATURES



```
+---------+---------------+--------------+-----------------+---------------------+-------------+--------------------+---------------------+---------------------+----------+
|PROXIMITY|NEAREST_HOTSPOT|SEVERITY_SCORE|     VEHICLE_TYPE|DRIVER_LICENSE_STATUS|TIME_INTERVAL|    PRE_CRASH_BUCKET|CONTRIBUTING_FACTOR_1|CONTRIBUTING_FACTOR_2|RISK LEVEL|
+---------+---------------+--------------+-----------------+---------------------+-------------+--------------------+---------------------+---------------------+----------+
|   7.5581|              8|    0.12737693|passenger vehicle|              unknown|        18-24|Going Straight Ahead|      Driver Behavior|          Unspecified|       LOW|
|   7.5581|              8|    0.12737693|            other|              unknown|        18-24|Going Straight Ahead|          Unspecified|          Unspecified|       LOW|
|  13.8304|              7|    0.13044104|passenger vehicle|              unknown|        18-24|                Turn|           Substances|          Unspecified|       LOW|
|  13.8304|              7|    0.13044104|passenger vehicle|              unknown|        18-24|                Turn|          Unspecified|          Unspecified|       LOW|
|   7.8614|              7|    0.13044104|passenger vehicle|              unknown|        12-18|Going Straight Ahead|          Unspecified|          Unspecified|      HIGH|
|   7.8614|              7|    0.13044104|passenger vehicle|              unknown|        12-18|           Stop/Park|          Unspecified|          Unspecified|      HIGH|
|  12.4281|              7|    0.13044104|    station wagon|              unknown|          0-6|Going Straight Ahead|           Substances|          Unspecified|      HIGH|
|   7.1529|              7|    0.13044104|passenger vehicle|              unknown|        12-18|                Turn|          Unspecified|          Unspecified|       LOW|
|   7.1529|              7|    0.13044104|    station wagon|              unknown|        12-18|Going Straight Ahead|          Unspecified|          Unspecified|       LOW|
|  11.2255|              8|    0.12737693|    station wagon|              unknown|        18-24|Going Straight Ahead|          Unspecified|          Unspecified|       LOW|
|  11.2255|              8|    0.12737693|    station wagon|              unknown|        18-24|Going Straight Ahead|      Driver Behavior|          Unspecified|       LOW|
|  14.9756|              7|    0.13044104|          unknown|              unknown|        12-18|               Other|          Unspecified|          Unspecified|       LOW|
|  14.9756|              7|    0.13044104|passenger vehicle|              unknown|        12-18|           Stop/Park|          Unspecified|          Unspecified|       LOW|
| 4115.245|              3|    0.21473023|passenger vehicle|              unknown|        18-24|            Maneuver|          Unspecified|          Unspecified|       LOW|
```

PIPELINE FOR MODELLING

- Used STRINGINDEXER to code RISK_LEVEL as LABEL
- Divided the features into CATEGORICAL and NUMERIC
- Applied Standard Scaler to the Numeric Values using VectorAssembler
- Applied ONE HOT ENCODING to all the Categorical values
- Created a Vector Assembler for all the Inputs
- Applied the Models

RISK LEVEL PREDICTION

Main objective is to classify demand at every location in New York City.

Models

Logistic Regression

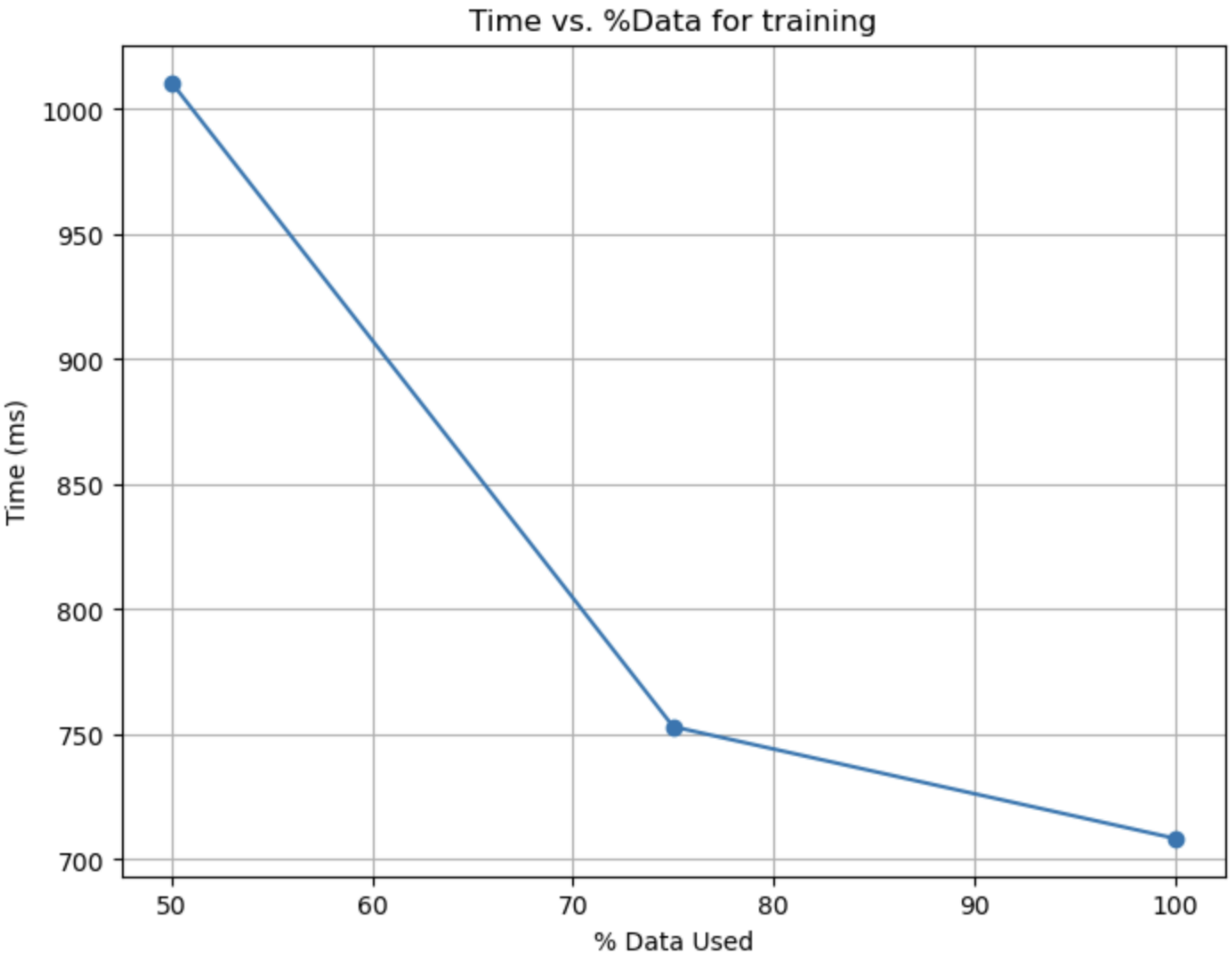
Random Forest

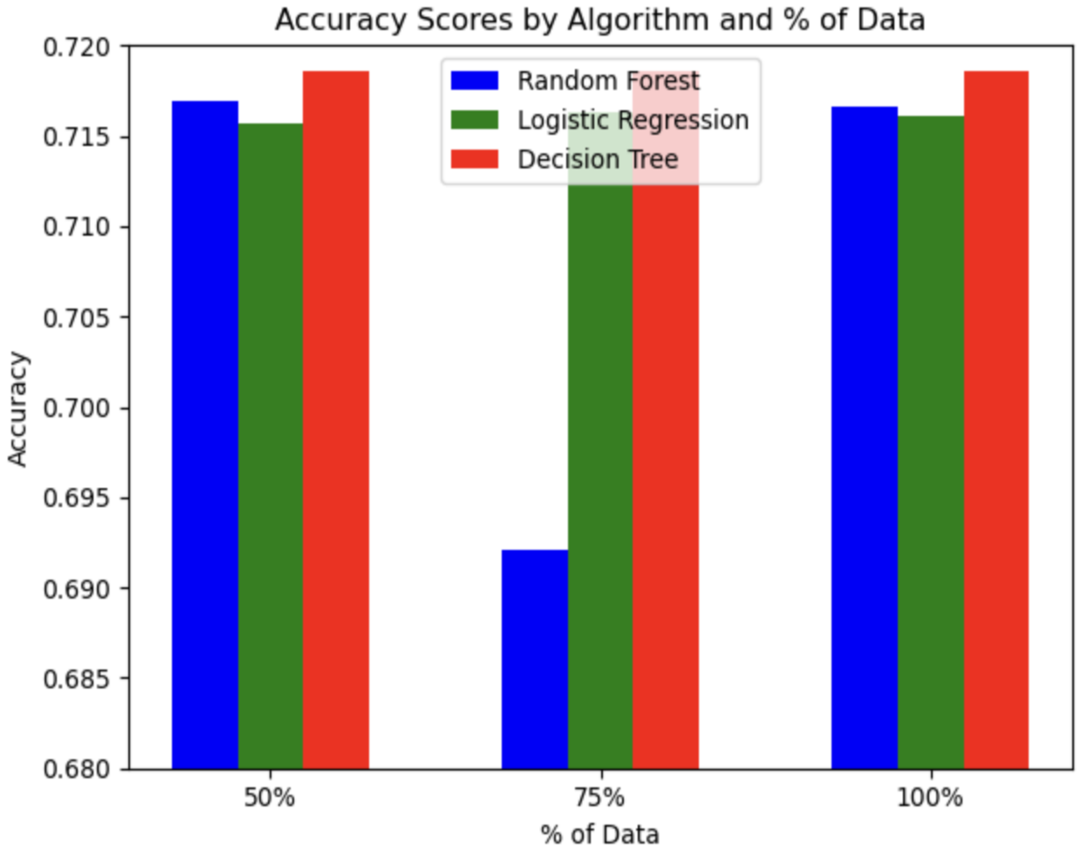
Decision Tree

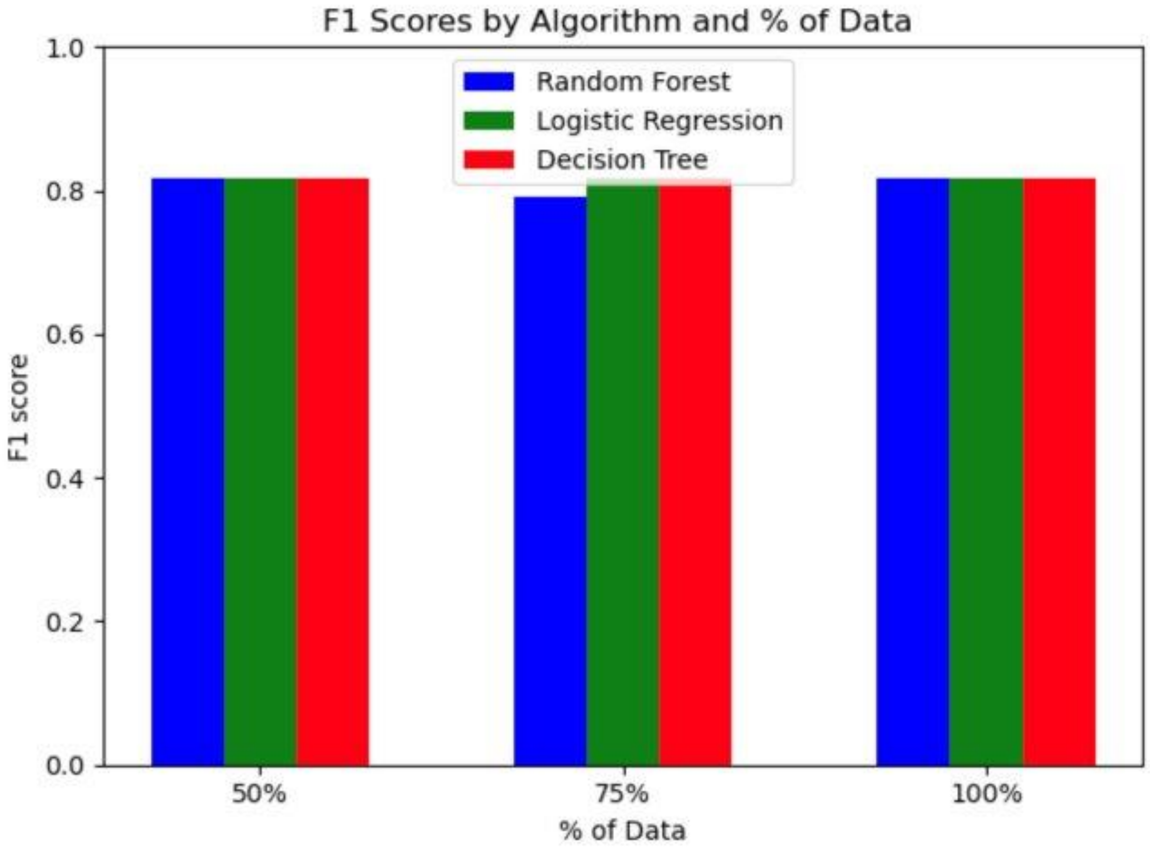
Model Evaluation

Model	Accuracy	F1 Score
Logistic Regression	0.716055	0.81752
Decision Tree Classifier	0.71856	0.81787
Random Forest Classifier	0.71663	0.81659

SCALE UP


Time vs. %Data for training
Time (ms)
1000
950
900
850
800
750
700
50
60
70
80
90
100
% Data Used


Accuracy Scores by Algorithm and % of Data
Random Forest
Logistic Regression
Decision Tree
Accuracy
0.720
0.715
0.710
0.705
0.700
0.695
0.690
0.685
0.680
50%
75%
100%
% of Data


F1 Scores by Algorithm and % of Data
Random Forest
Logistic Regression
Decision Tree
F1 score
1.0
0.8
0.6
0.4
0.2
0.0
50%
75%
100%
% of Data

RANDOM FOREST 50% data

- Time taken: 418.963305473327 64 seconds
- Accuracy: 0.7169191661486793
- Random Forest Precision: 0.7565467697308833
- Random Forest Recall: 0.8865177491392947
- Random Forest F1 Score: 0.8163917262136676

RANDOM FOREST 75% data

- Time taken: 261.78963923454285 seconds
- Accuracy: 0.692088761891803
- Random Forest Precision: 0.6929195701214325
- Random Forest Recall: 0.923536800967712
- Random Forest F1 Score: 0.7917772908237417

RANDOM FOREST 100% data

- Time taken: 237.73797464370728 seconds
- Accuracy: 0.7166315672285669
- Random Forest Precision: 0.7538214711064303
- Random Forest Recall: 0.8907834744579882
- Random Forest F1 Score: 0.8165994039653894

LOGISTIC REGRESSION 50% data

- Time taken: 297.61940908432007 seconds
- Logistic Regression Accuracy: 0.715639917905 4552
- Logistic Regression Precision: 0.7724335054004188
- Logistic Regression Recall: 0.8667506048199497
- Logistic Regression F1 Score: 0.8168785968758563

LOGISTIC REGRESSION 75% data

- Time taken: 282.52716064453125 seconds
- Logistic Regression Accuracy: 0.716267687591292
- Logistic Regression Precision: 0.7717953418241811
- Logistic Regression Recall: 0.8692280403833628
- Logistic Regression F1 Score: 0.8176192488474788

LOGISTIC REGRESSION 100% data

- Time taken: 289.8686454296112 seconds
- Logistic Regression Accuracy: 0.7160553385678399
- Logistic Regression Precision: 0.7723429623709727
- Logistic Regression Recall: 0.8683266260351726
- Logistic Regression F1 Score: 0.817527140622027

DECISION TREE 50% data

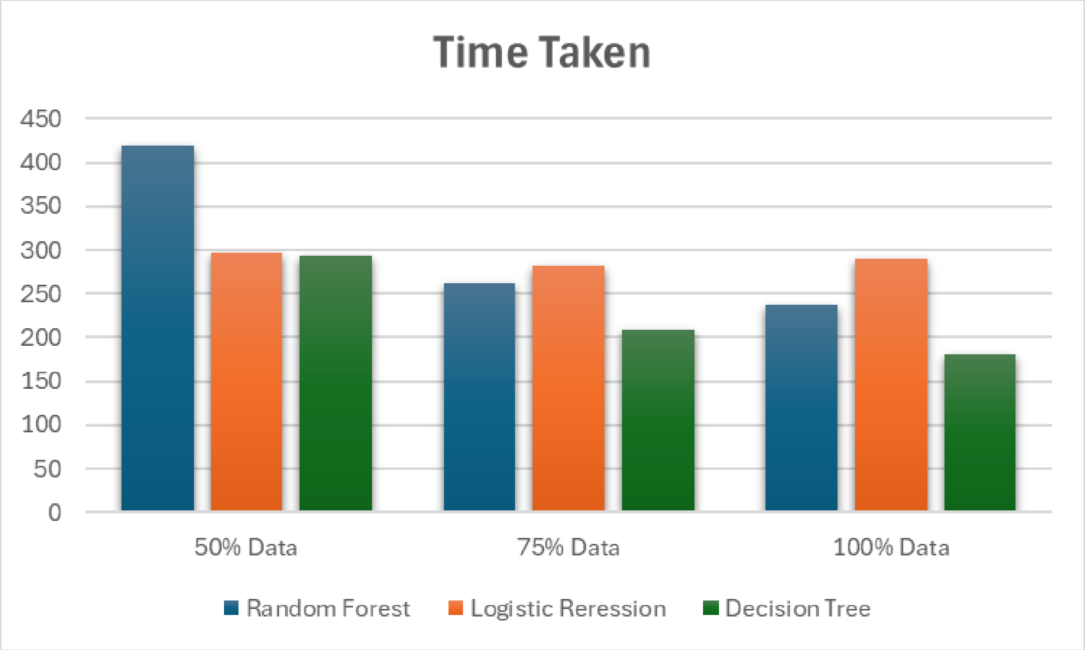
- Time taken: 294.0357983112335 seconds
- Decision Tree Accuracy: 0.7185344618756191
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274

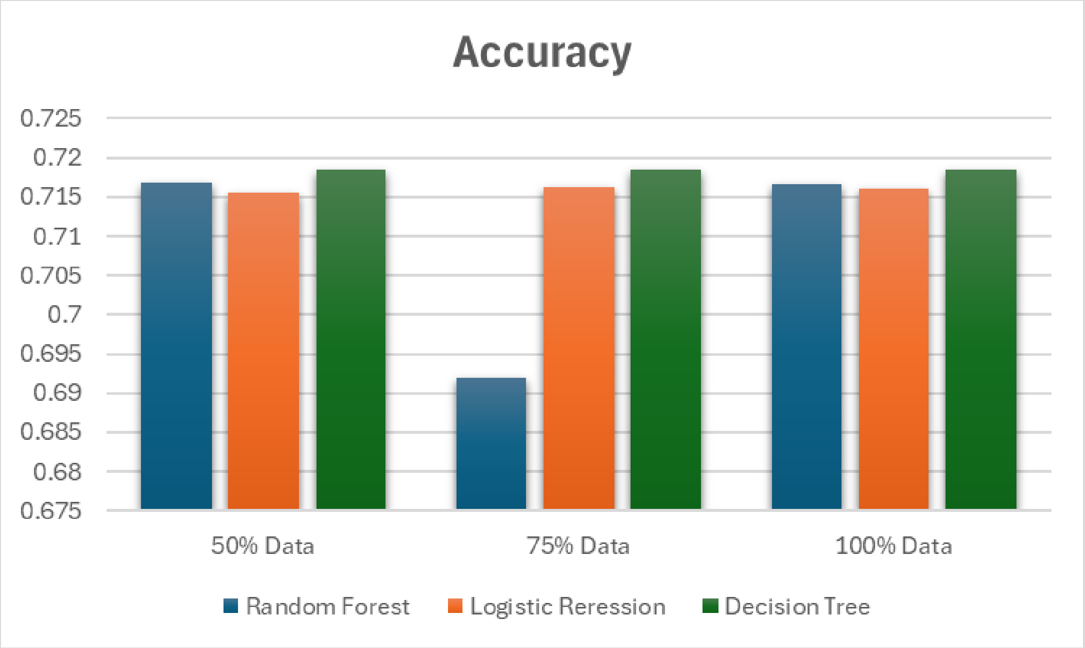
DECISION TREE 75% data

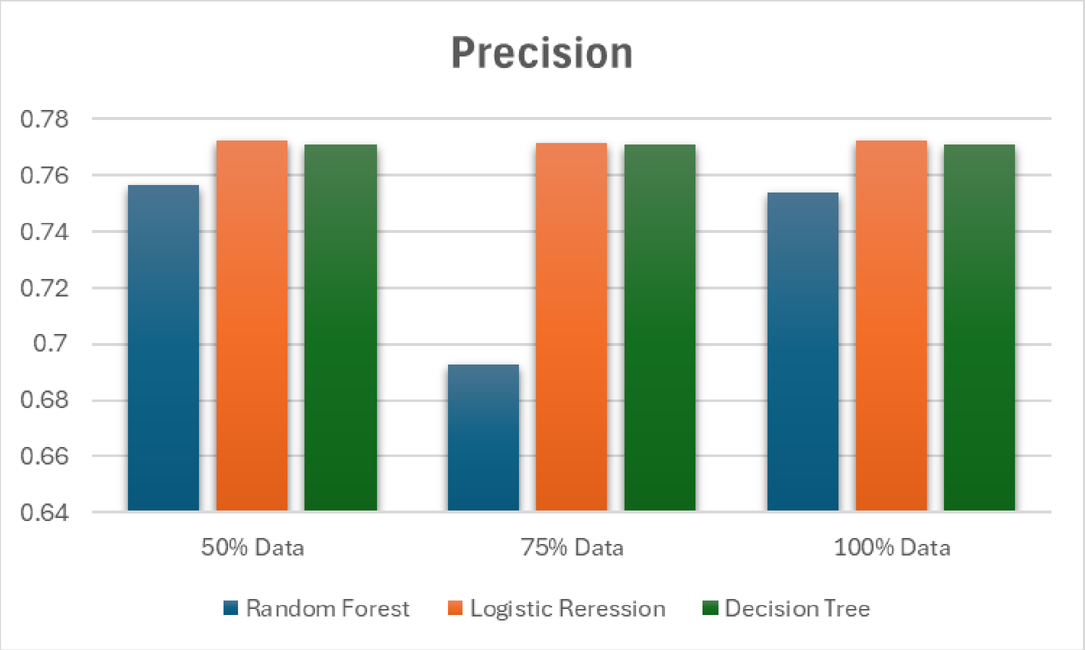
- Time taken: 208.7618627548 2178 seconds
- Decision Tree Accuracy: 0.7185427084396367
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274

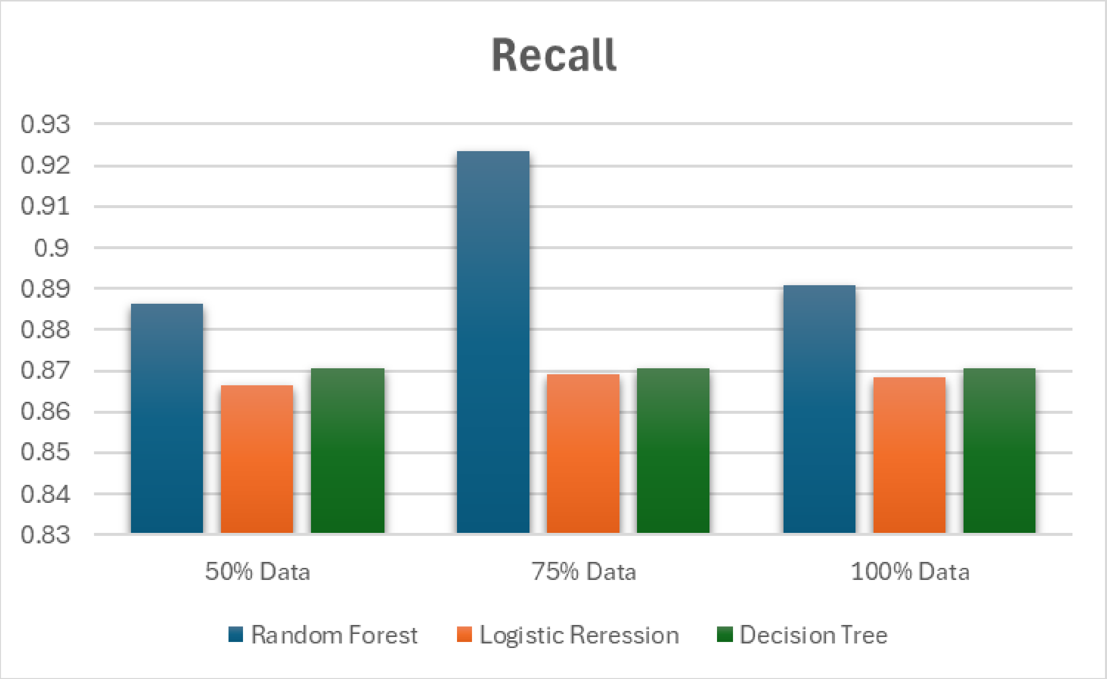
DECISION TREE 100% data

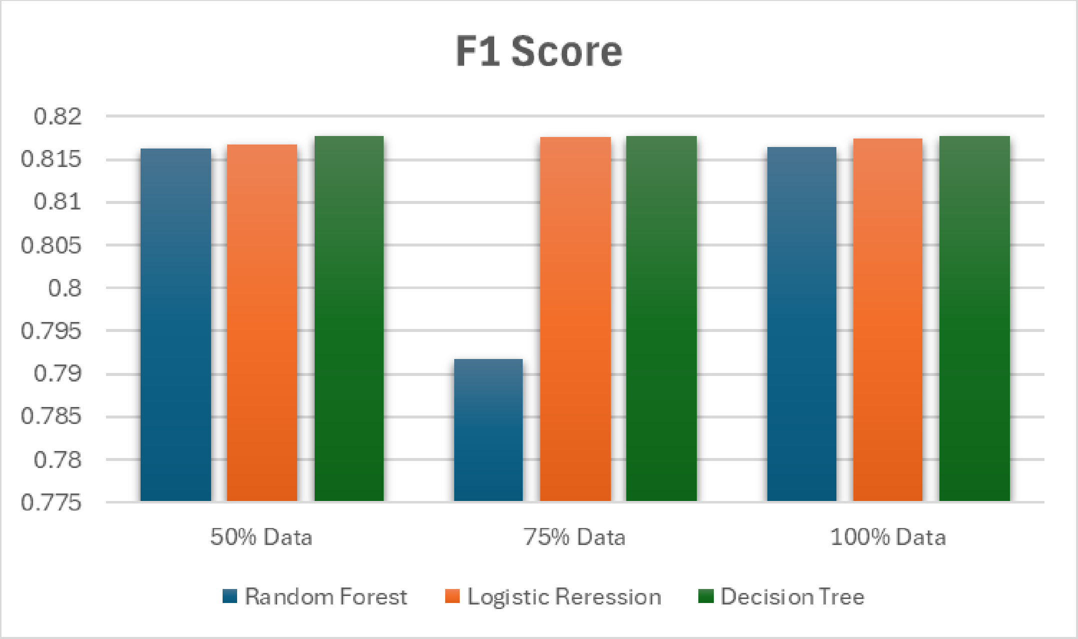
- Time taken: 180.825050354 0039 seconds
- Decision Tree Accuracy: 0.7185643556701828
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274


Time Taken
450
400
350
300
250
200
150
100
50
0
50% Data
75% Data
100% Data
Random Forest
Logistic Reression
Decision Tree

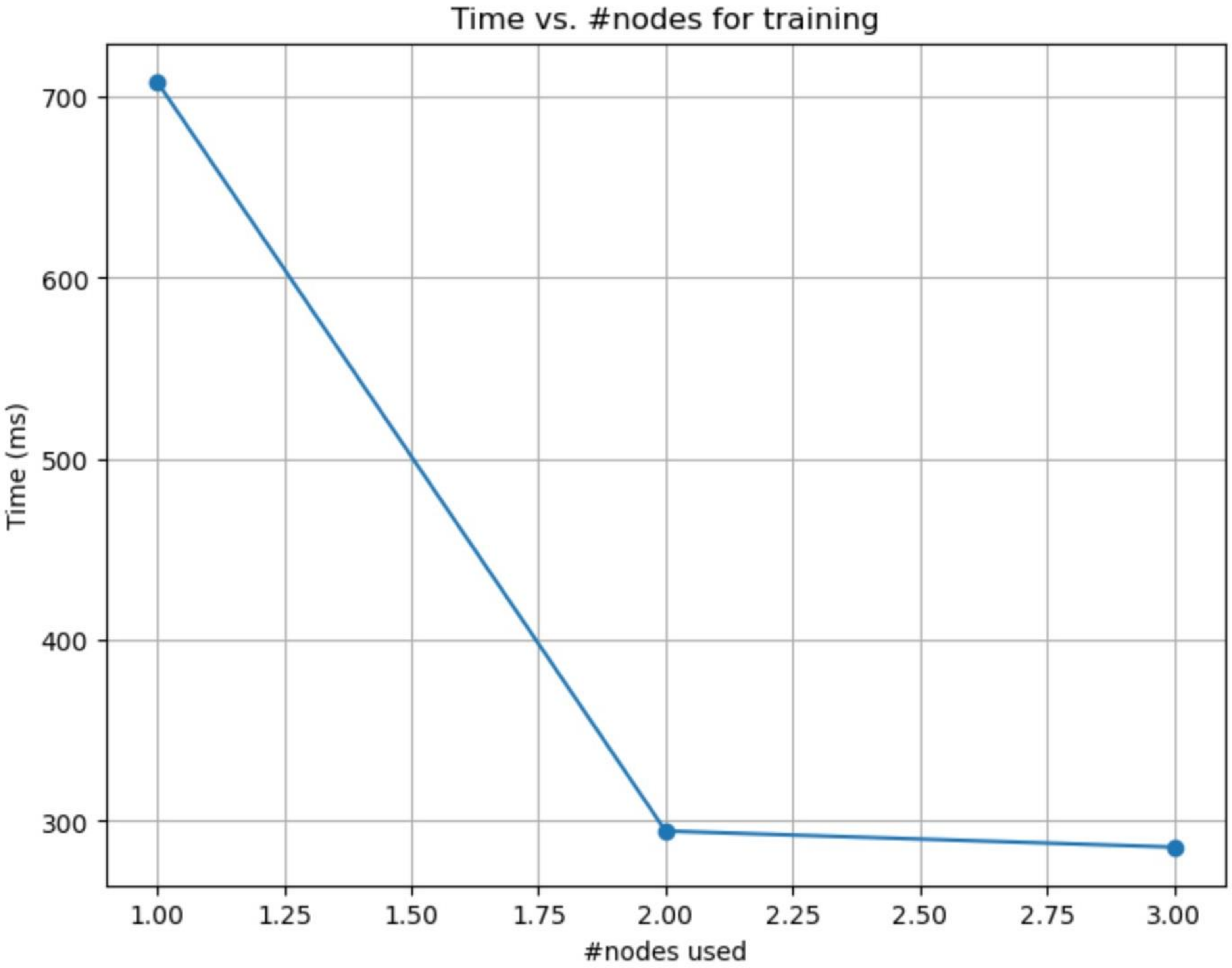

Accuracy
0.725
0.72
0.715
0.71
0.705
0.7
0.695
0.69
0.685
0.68
0.675
50% Data
75% Data
100% Data
Random Forest
Logistic Reression
Decision Tree

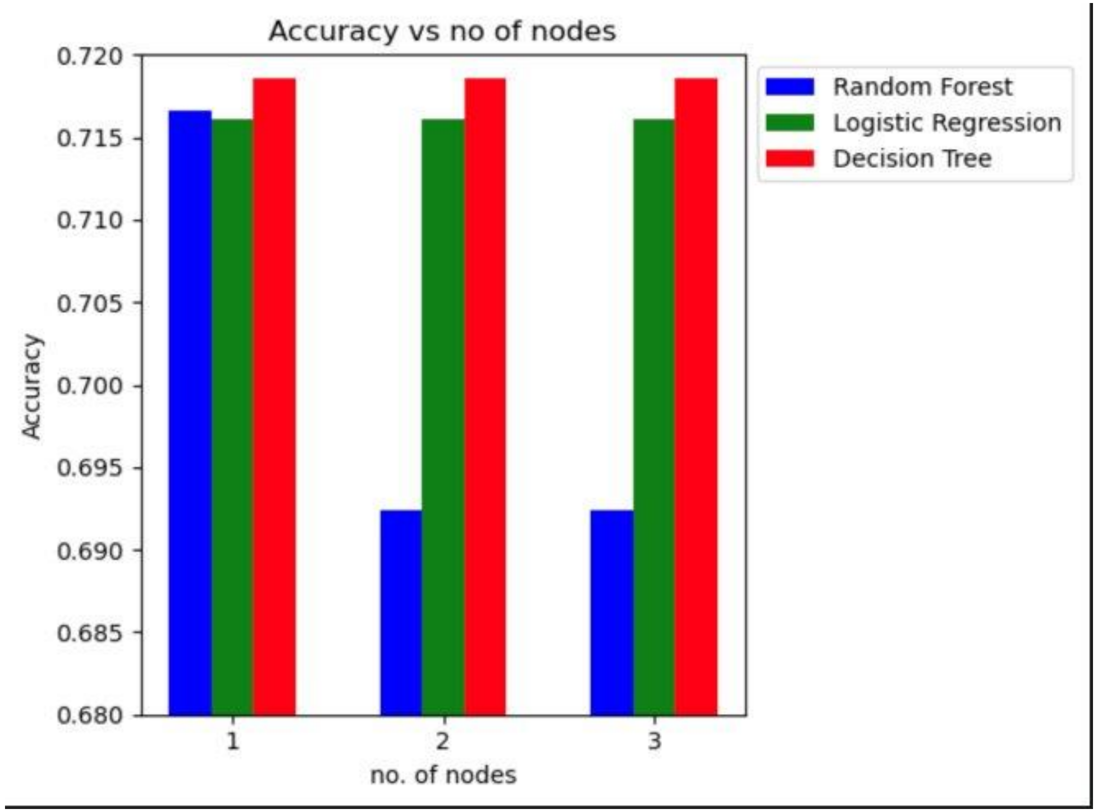

Precision
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
50% Data
75% Data
100% Data
Random Forest
Logistic Reression
Decision Tree

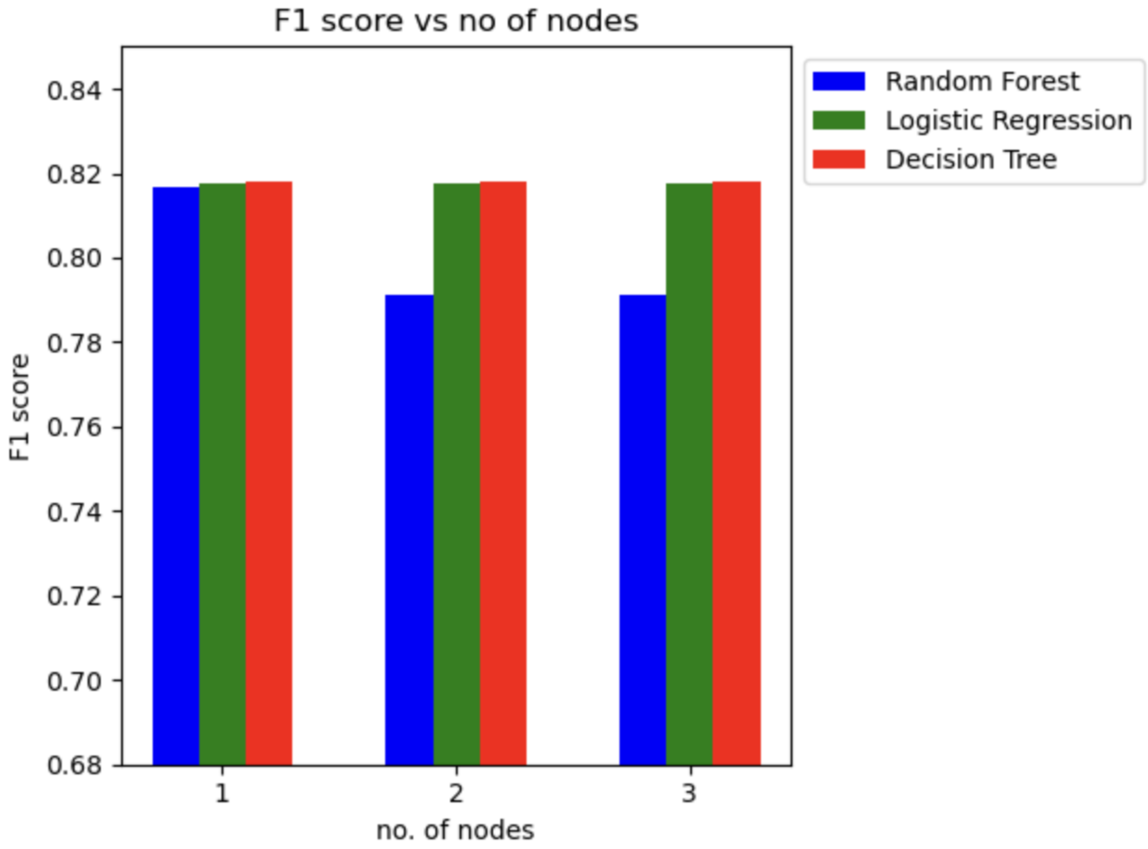

Recall
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0.85
0.84
0.83
50% Data
75% Data
100% Data
■ Random Forest
■ Logistic Reression
■ Decision Tree


F1 Score
0.82
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
50% Data
75% Data
100% Data
■ Random Forest
■ Logistic Reression
■ Decision Tree

SCALE OUT


Time vs. #nodes for training
Time (ms)
700
600
500
400
300
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
#nodes used


Accuracy vs no of nodes
0.720
0.715
0.710
0.705
0.700
0.695
0.690
0.685
0.680
Accuracy
1
2
3
no. of nodes
Random Forest
Logistic Regression
Decision Tree


F1 score vs no of nodes
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
F1 score
1
2
3
no. of nodes
Random Forest
Logistic Regression
Decision Tree

RANDOM FOREST

Nodes = 1

- Time taken: 237.73797464370728 seconds
- Accuracy: 0.7166315672285669
- Random Forest Precision: 0.7538214711064303
- Random Forest Recall: 0.8907834744579882
- Random Forest F1 Score: 0.8165994039653894

Nodes = 2

- Time taken: 125.54719972610474 seconds
- Accuracy: 0.6924402716830516
- Random Forest Precision: 0.6927531992368112
- Random Forest Recall: 0.9227429747836605
- Random Forest F1 Score: 0.7913768638198863

Nodes = 3

- Time taken: 121.74665689468384 seconds
- Accuracy: 0.6924402716830516
- Random Forest Precision: 0.6927531992368112
- Random Forest Recall: 0.9227429747836605
- Random Forest F1 Score: 0.7913768638198863

LOGISTIC REGRESSION

Nodes = 1

- Time taken: 289.8686454296112 seconds
- Logistic Regression Accuracy: 0.7160553385678399
- Logistic Regression Precision: 0.7723429623709727
- Logistic Regression Recall: 0.8683266260351726
- Logistic Regression F1 Score: 0.817527140622027

Nodes = 2

- Time taken: 96.19693279266357 seconds
- Logistic Regression Accuracy: 0.7160553385678399
- Logistic Regression Precision: 0.7723429623709727
- Logistic Regression Recall: 0.8683266260351726
- Logistic Regression F1 Score: 0.817527140622027

Nodes = 3

- Time taken: 93.86266040802002 seconds
- Logistic Regression Accuracy: 0.7160553385678
- Logistic Regression Precision: 0.7723429623709
- Logistic Regression Recall: 0.8683266260351726
- Logistic Regression F1 Score: 0.81752714062202

DECISION TREE

Nodes = 1

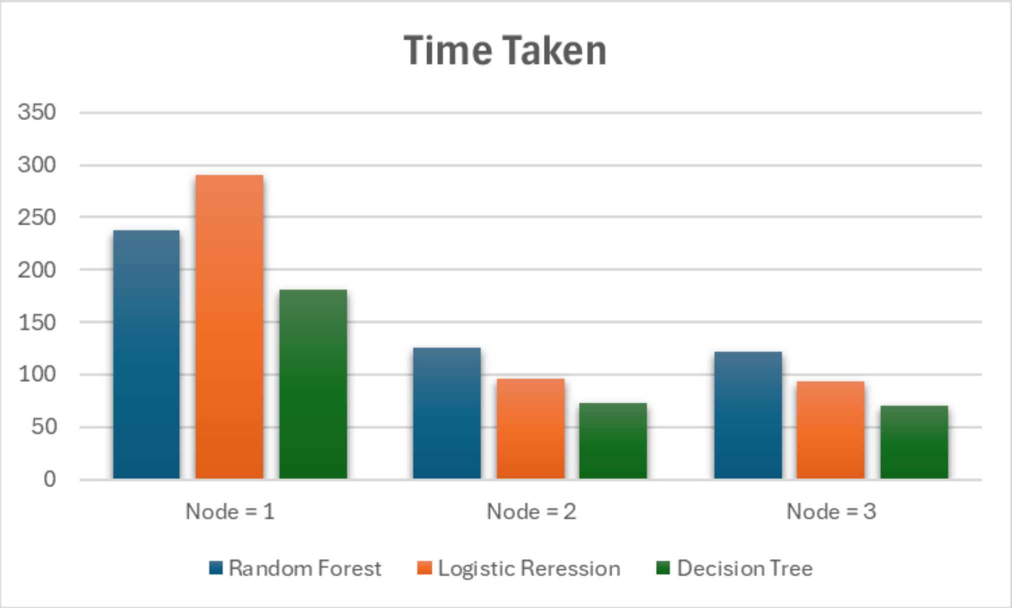
- Time taken: 180.8250503540039 seconds
- Decision Tree Accuracy: 0.7185643556701828
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274

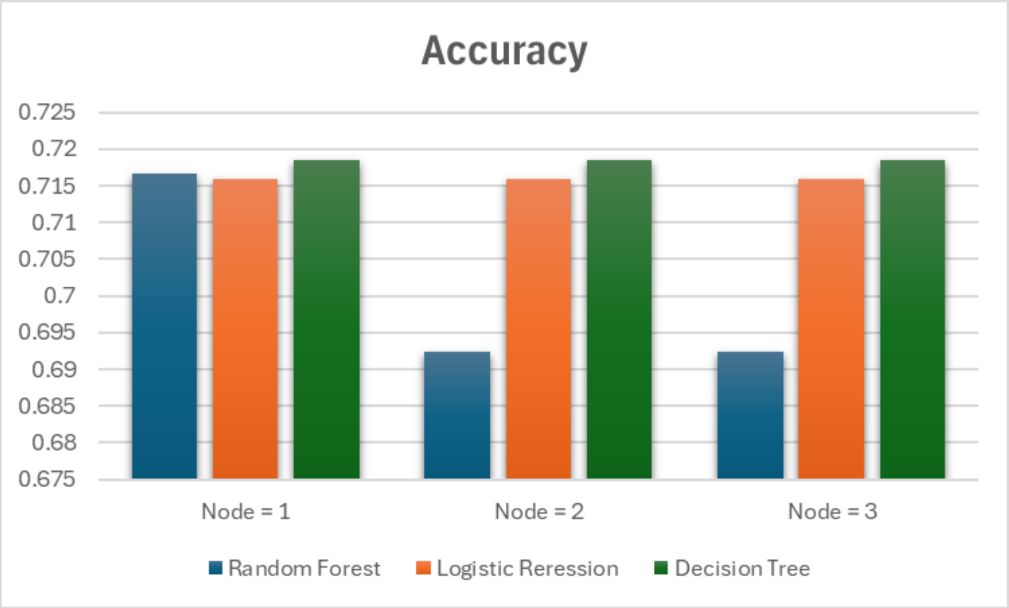
Nodes = 2

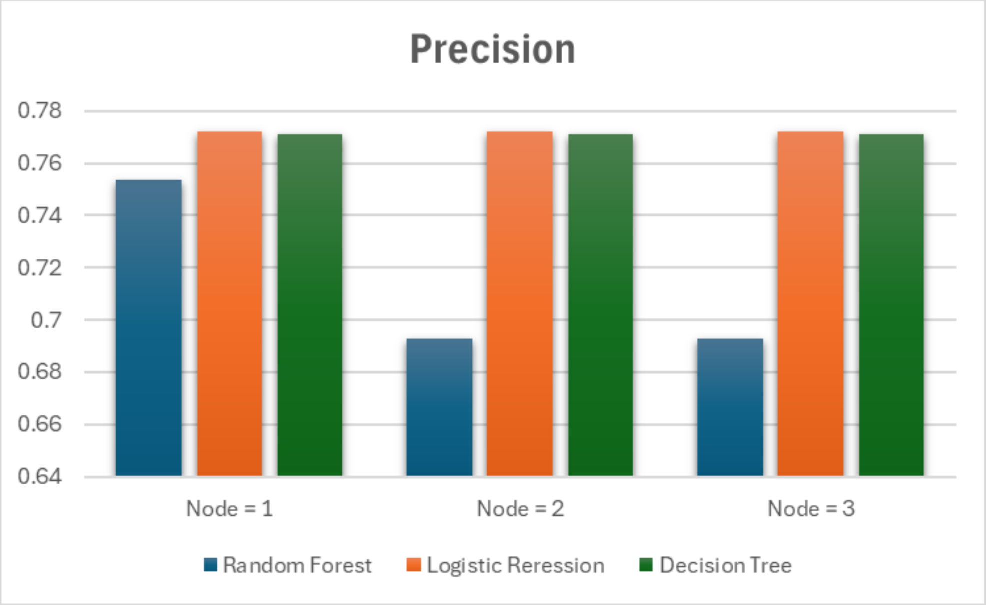
- Time taken: 72.73840665817261 seconds
- Decision Tree Accuracy: 0.7185385851576279
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274

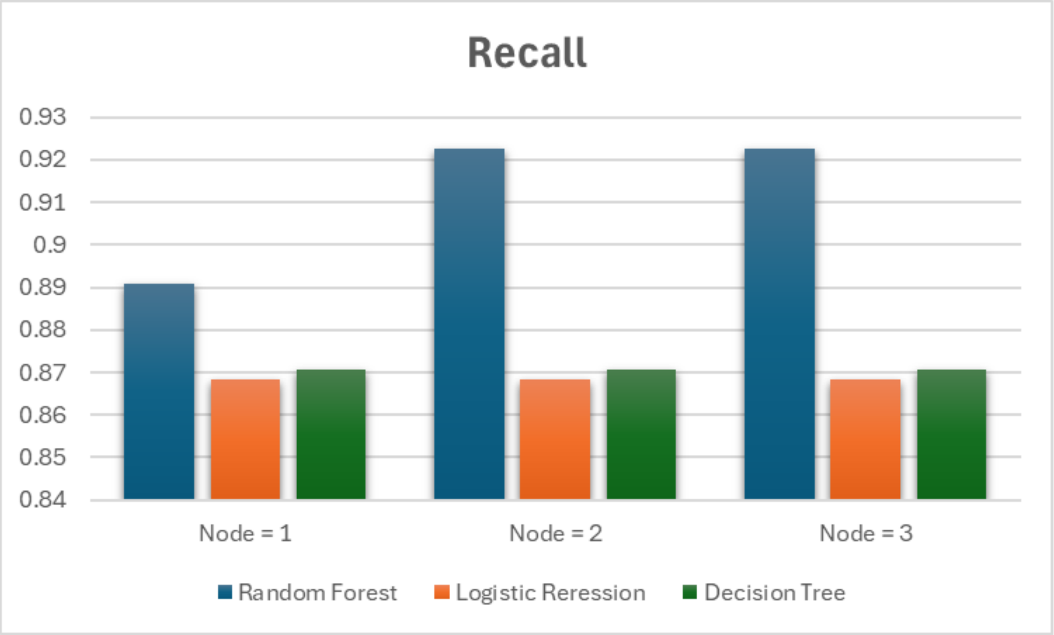
Nodes = 3

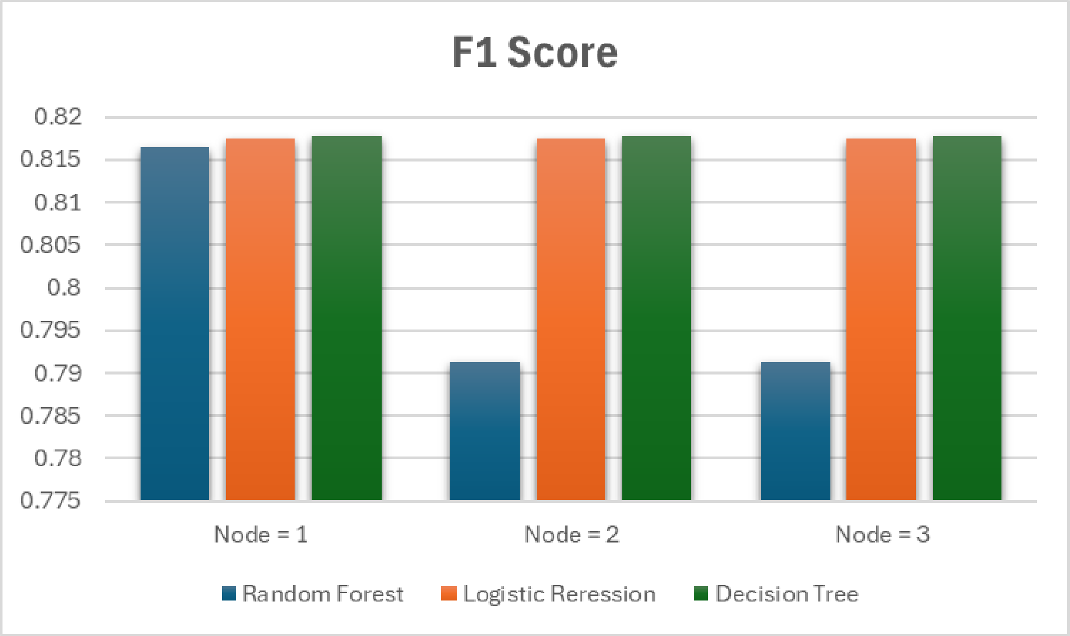
- Time taken: 70.05633354187012 seconds
- Decision Tree Accuracy: 0.7185385851576279
- Decision Tree Precision: 0.7712340342217665
- Decision Tree Recall: 0.8705132827765888
- Decision Tree F1 Score: 0.8178718661126274


Time Taken


Accuracy


Precision


Recall
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0.85
0.84
Node = 1
Node = 2
Node = 3
Random Forest
Logistic Reression
Decision Tree


F1 Score
0.82
0.815
0.81
0.805
0.8
0.795
0.79
0.785
0.78
0.775
Node = 1
Node = 2
Node = 3
Random Forest
Logistic Reression
Decision Tree

THANK YOU!

Any questions?