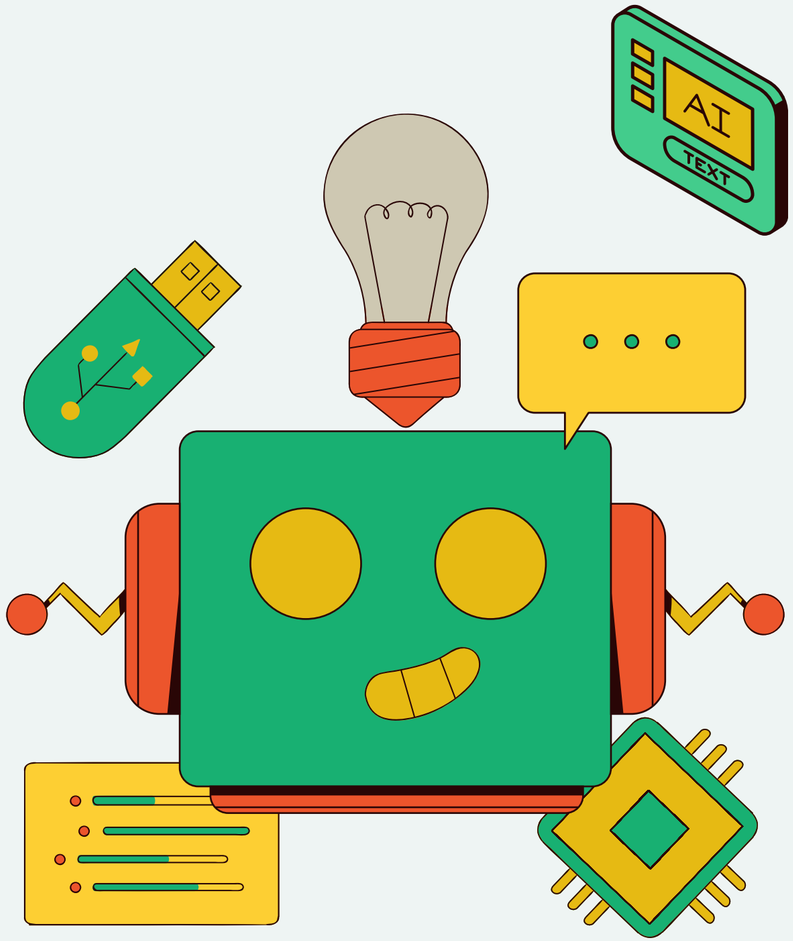


FAST TAG FRAUD DETECTION

USING MACHINE LEARNING TECHNIQUE

Presented by:

Pratiksha Saheb

PRESENTATION OUTLINE

- Introduction
- Dataset Overview
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Data Preprocessing
- Model Development
- Model Evaluation

INTRODUCTION

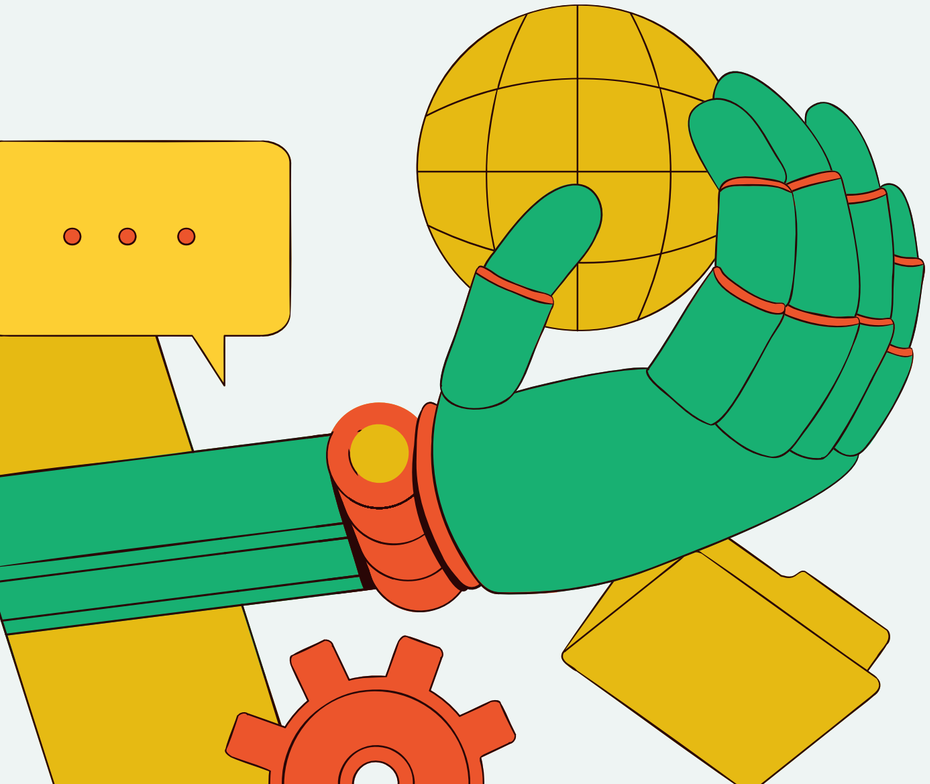
To develop an effective fraud detection system for Fastag transactions using machine learning classification techniques.

TO CREATE A ROBUST MODEL THAT CAN ACCURATELY IDENTIFY INSTANCES OF FRAUDULENT ACTIVITY, ENSURING THE INTEGRITY AND SECURITY OF FASTAG TRANSACTIONS.

DATASET OVERVIEW

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 13 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Transaction_ID         5000 non-null   int64
 1   Timestamp              5000 non-null   object
 2   Vehicle_Type           5000 non-null   object
 3   FastagID               4451 non-null   object
 4   TollBoothID            5000 non-null   object
 5   Lane_Type              5000 non-null   object
 6   Vehicle_Dimensions     5000 non-null   object
 7   Transaction_Amount     5000 non-null   int64
 8   Amount_paid            5000 non-null   int64
 9   Geographical_Location  5000 non-null   object
 10  Vehicle_Speed          5000 non-null   int64
 11  Vehicle_Plate_Number   5000 non-null   object
 12  Fraud_indicator        5000 non-null   object
dtypes: int64(4), object(9)
memory usage: 507.9+ KB
```

WHAT IS MACHINE LEARNING?



Machine learning is a subset of AI that focuses on the development of algorithms and models that enable computers to learn from and make predictions or decisions based on data.

It's a key driver of AI applications, including natural language processing, image recognition, and recommendation systems.

PROJECT KEY STEPS

DATA EXPLORATION

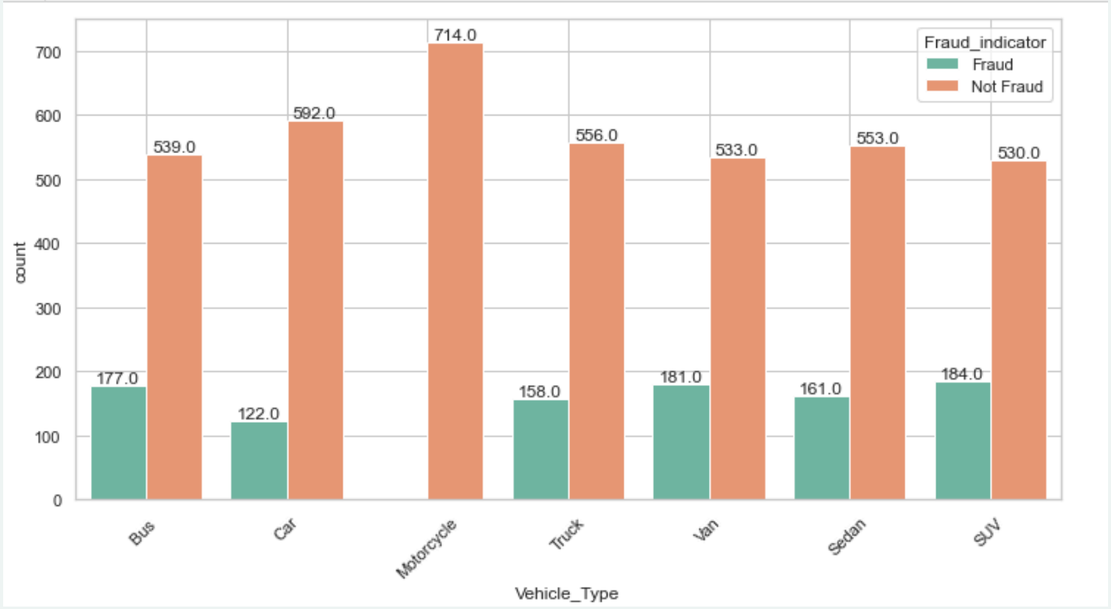
FEATURE ENGINEERING

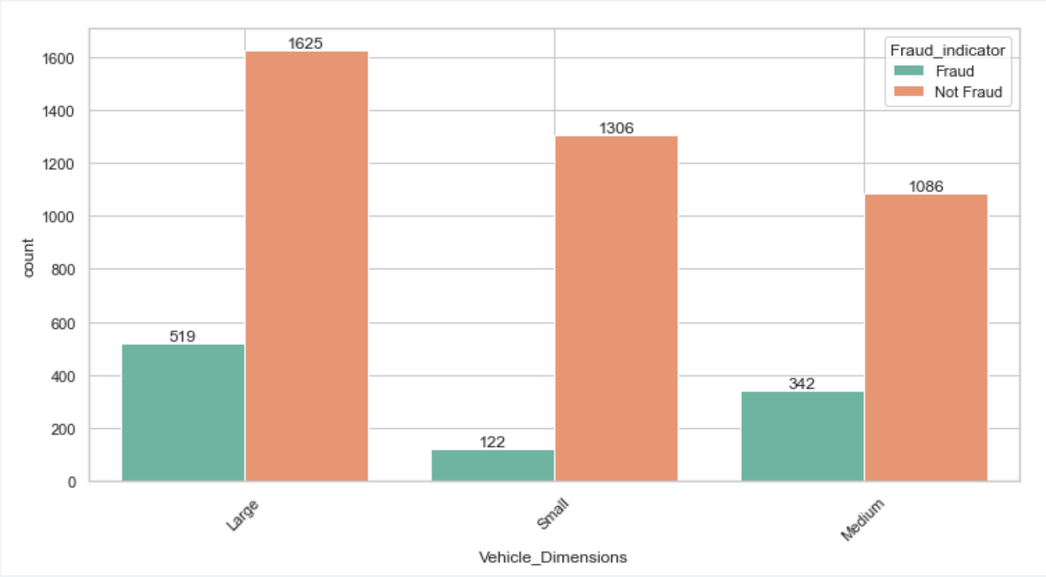
MODEL DEVELOPMENT

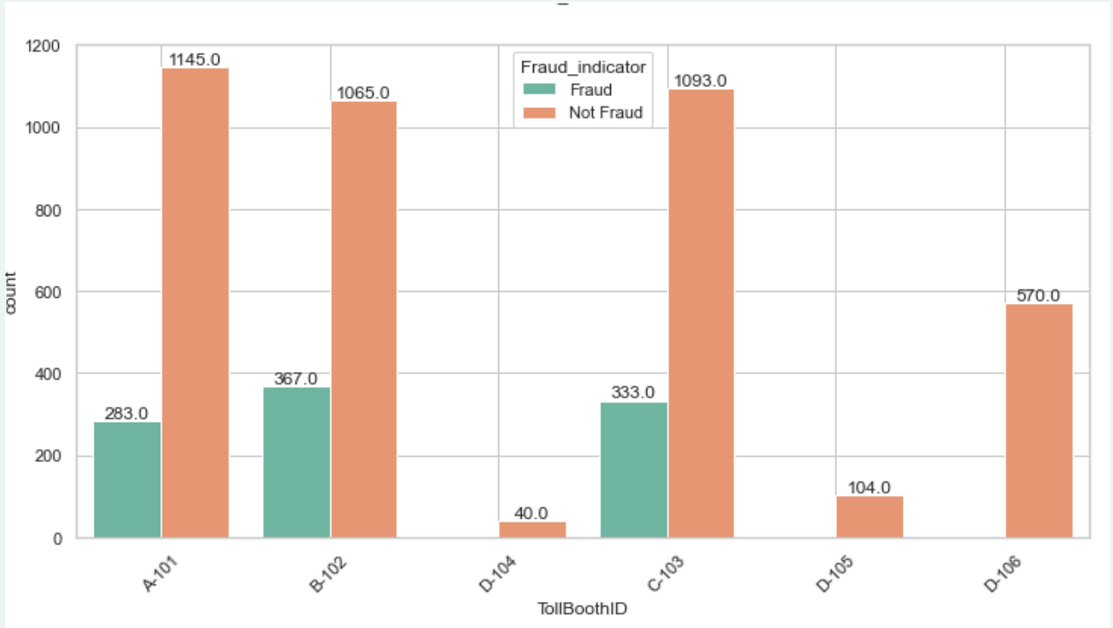
REAL TIME FRAUD DETECTION

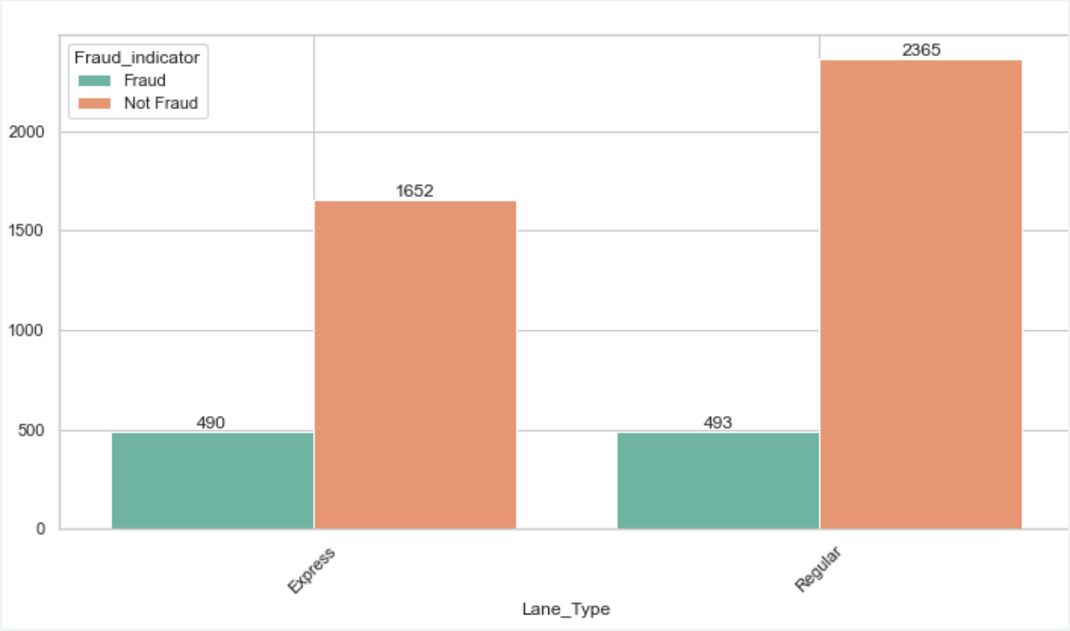
EXPLANATORY ANALYSIS

EXPLORATORY DATA ANALYSIS (EDA)

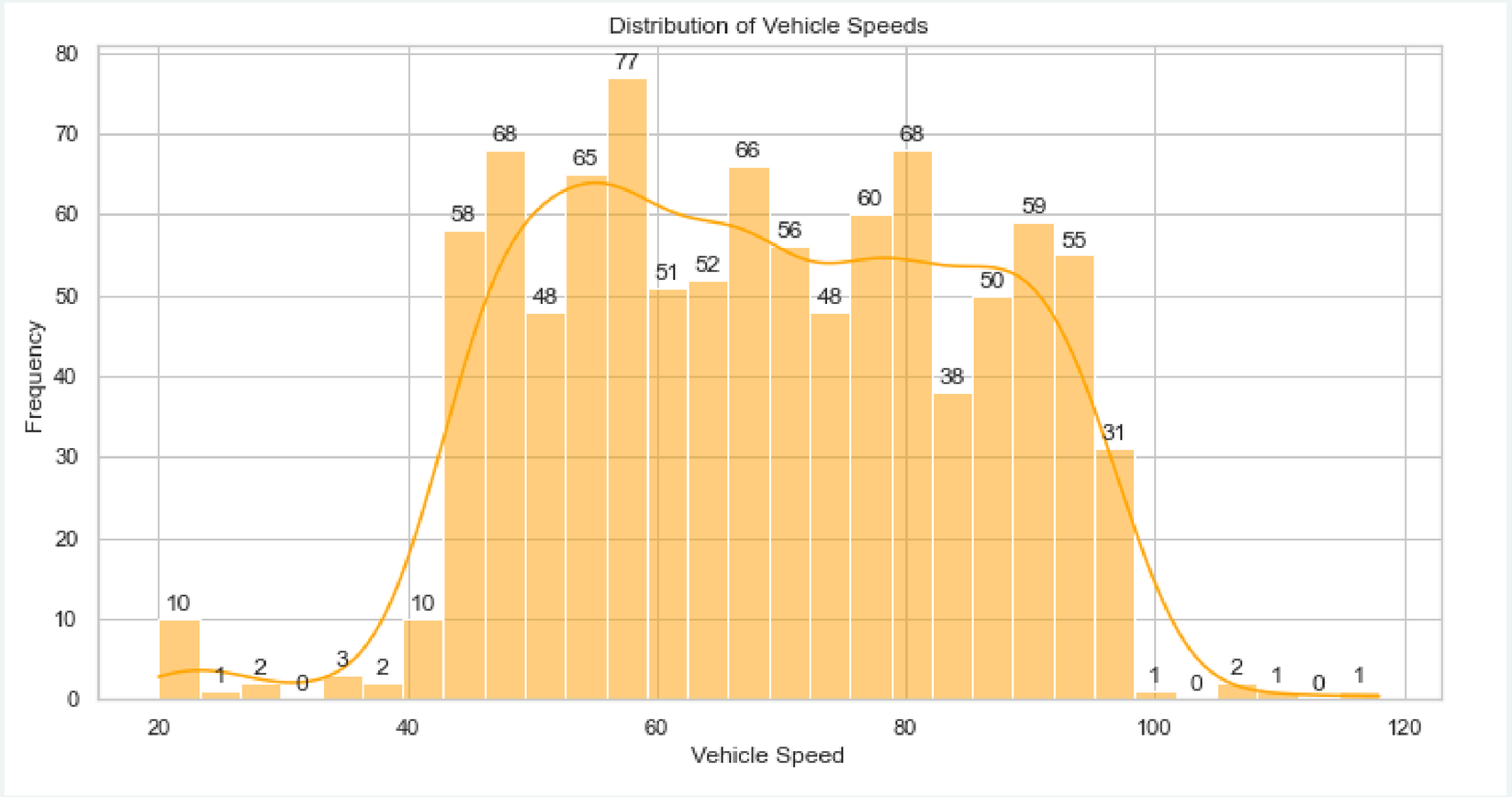








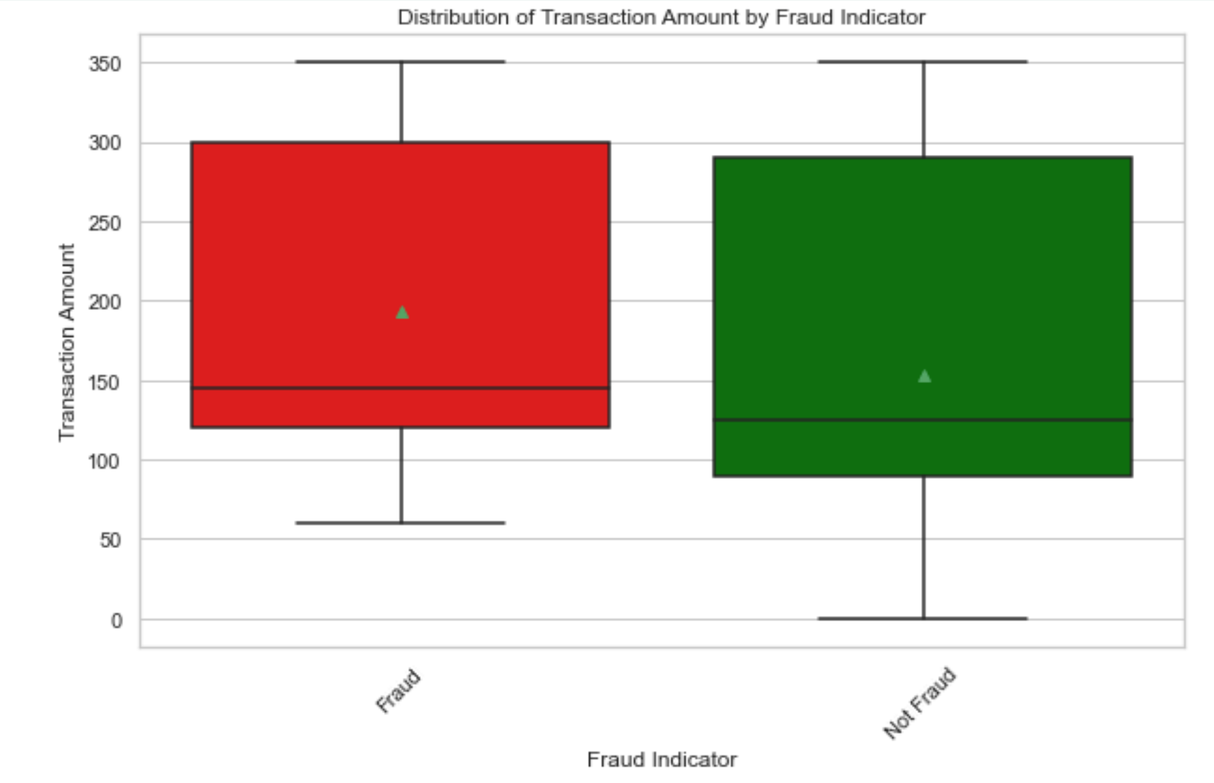
EXPLORATORY DATA ANALYSIS (EDA)



The highest frequency is observed in the 50-60 km/h range, with 77 vehicles. This indicates that the most common vehicle speed falls within this range.

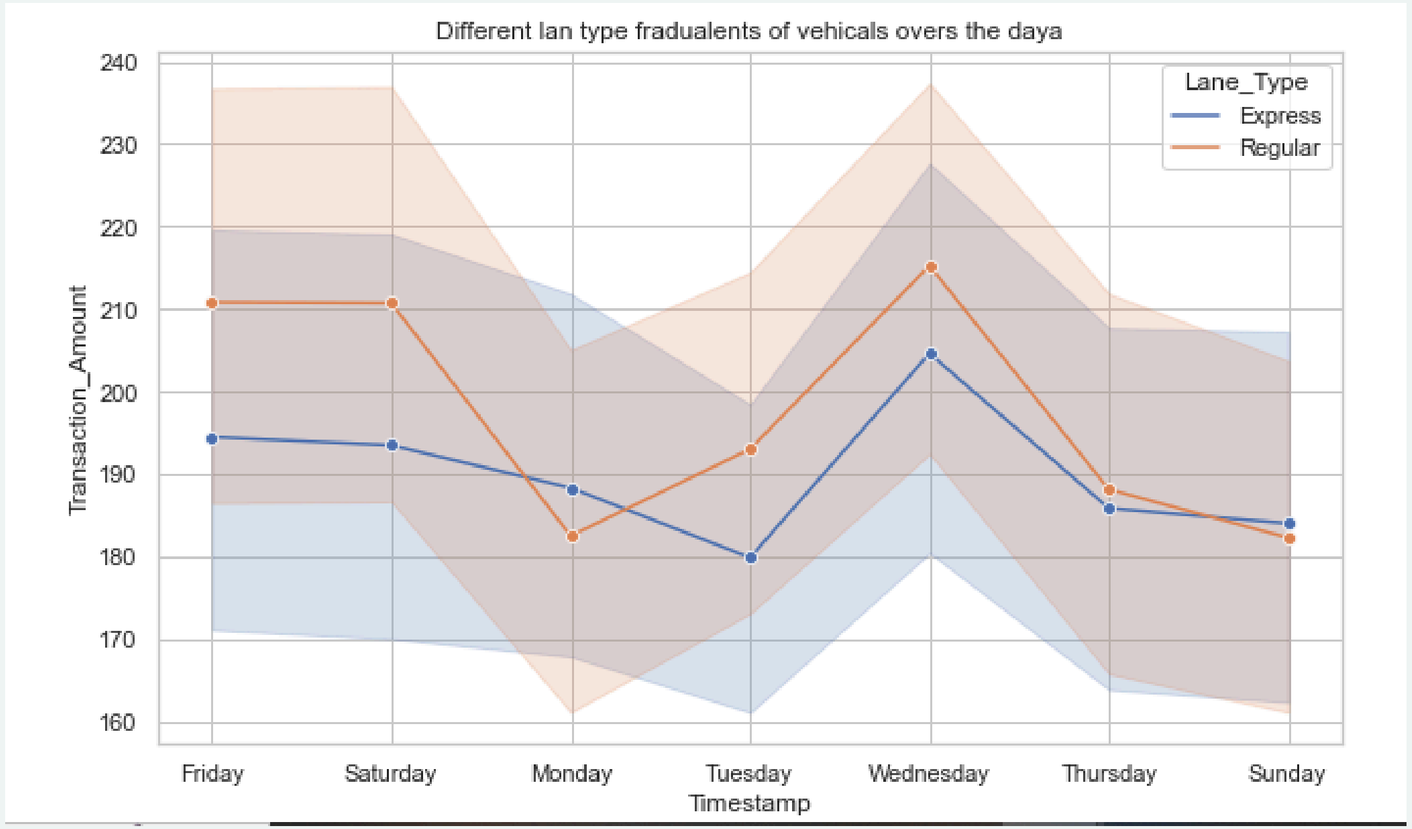
Another notable peak is in the 60-70 km/h range, with a frequency of 66 vehicles.Most vehicles are traveling within a specific range (40-80 km/h), with fewer vehicles at lower and higher speeds.

EXPLORATORY DATA ANALYSIS (EDA)



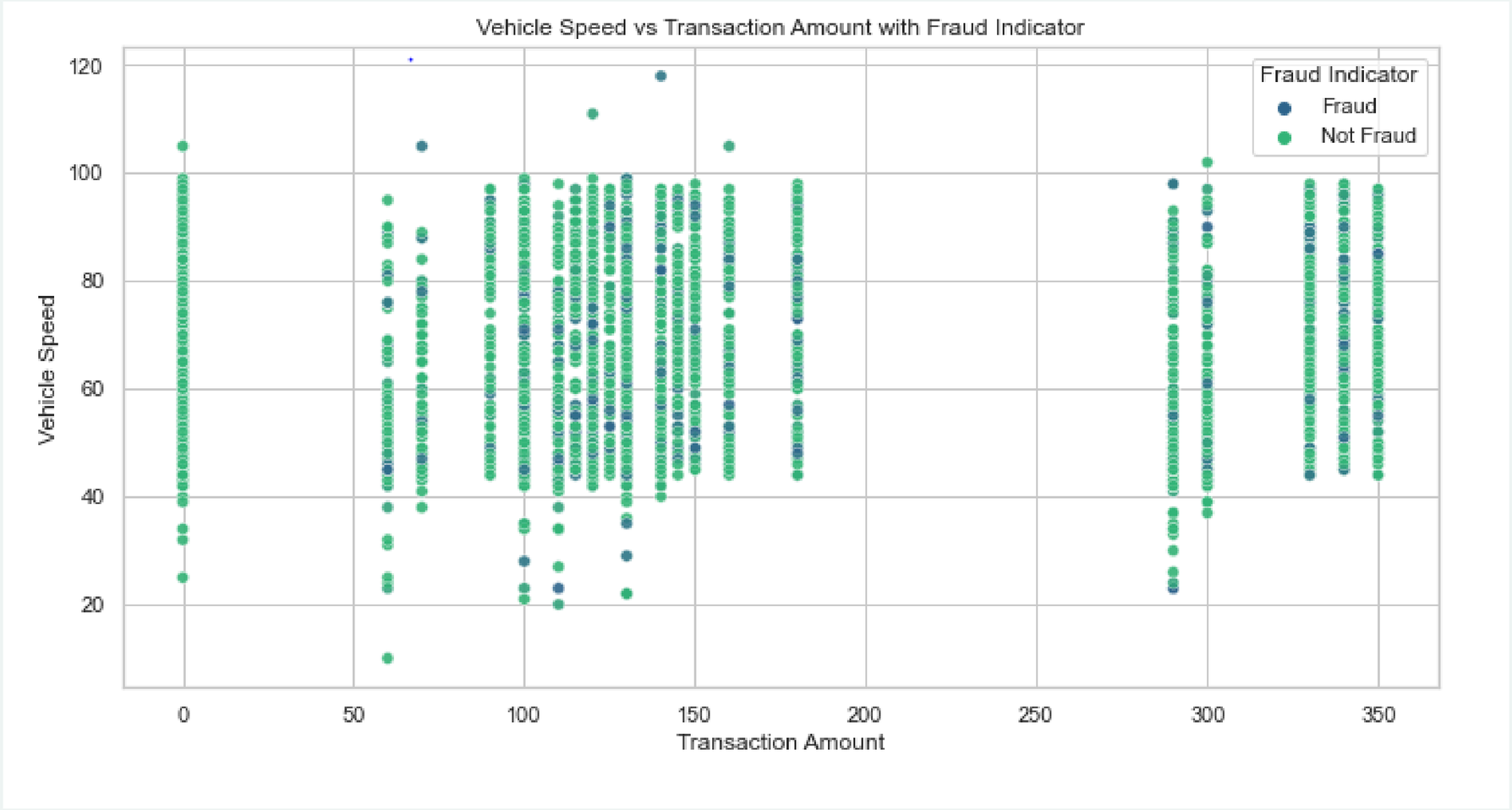
- The interquartile range (IQR), which represents the middle 50% of data, is slightly larger for fraudulent transactions compared to non-fraudulent ones. This suggests that there is more variability in the amounts of fraudulent transactions.
- The range of transaction amounts for both categories is similar, but fraudulent transactions show a slightly higher maximum.

EXPLORATORY DATA ANALYSIS (EDA)



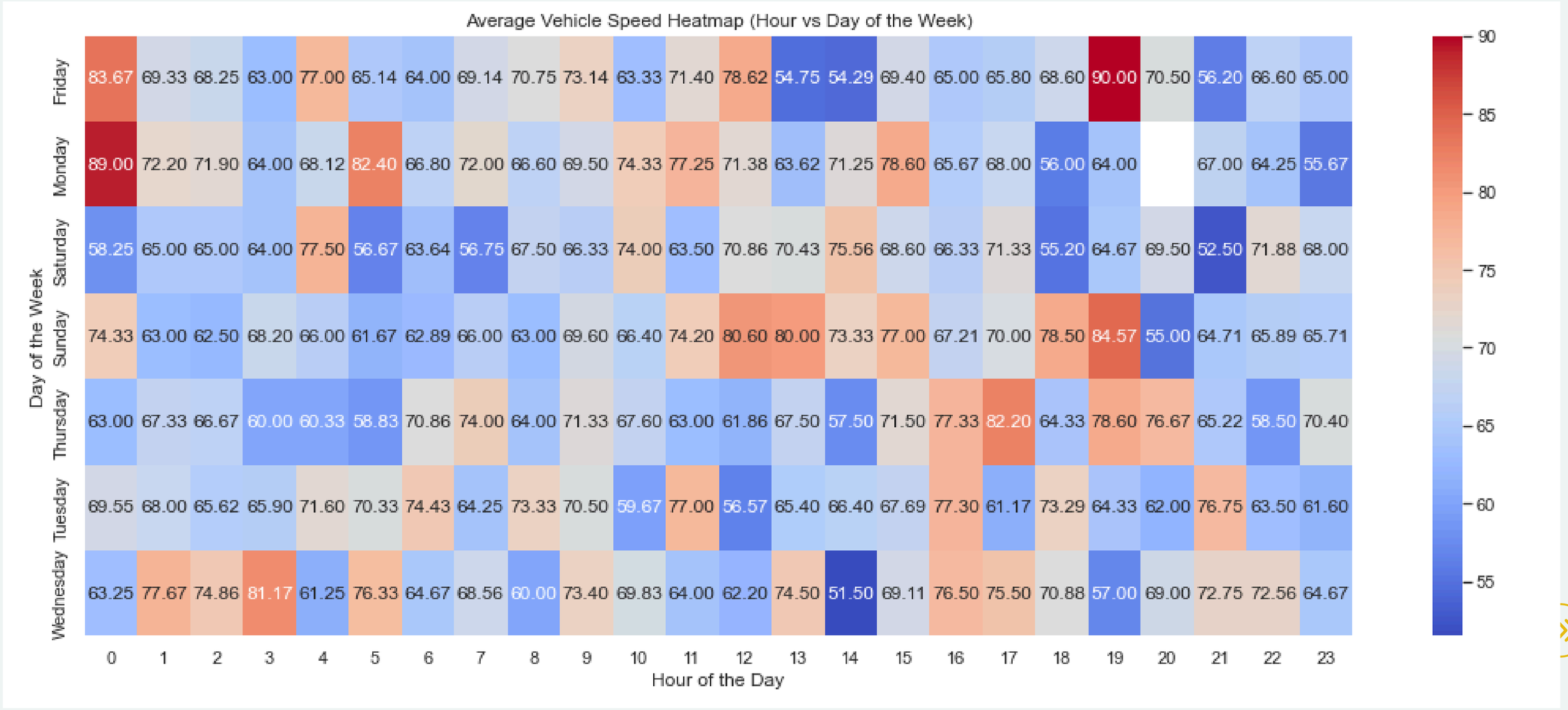
The consistently higher transaction amounts in Regular lanes might indicate that these lanes handle transactions that are either larger in amount or more frequent than Express lanes. The peak on Wednesday in Regular lanes could point to higher traffic volume or higher value transactions occurring mid-week.

EXPLORATORY DATA ANALYSIS (EDA)



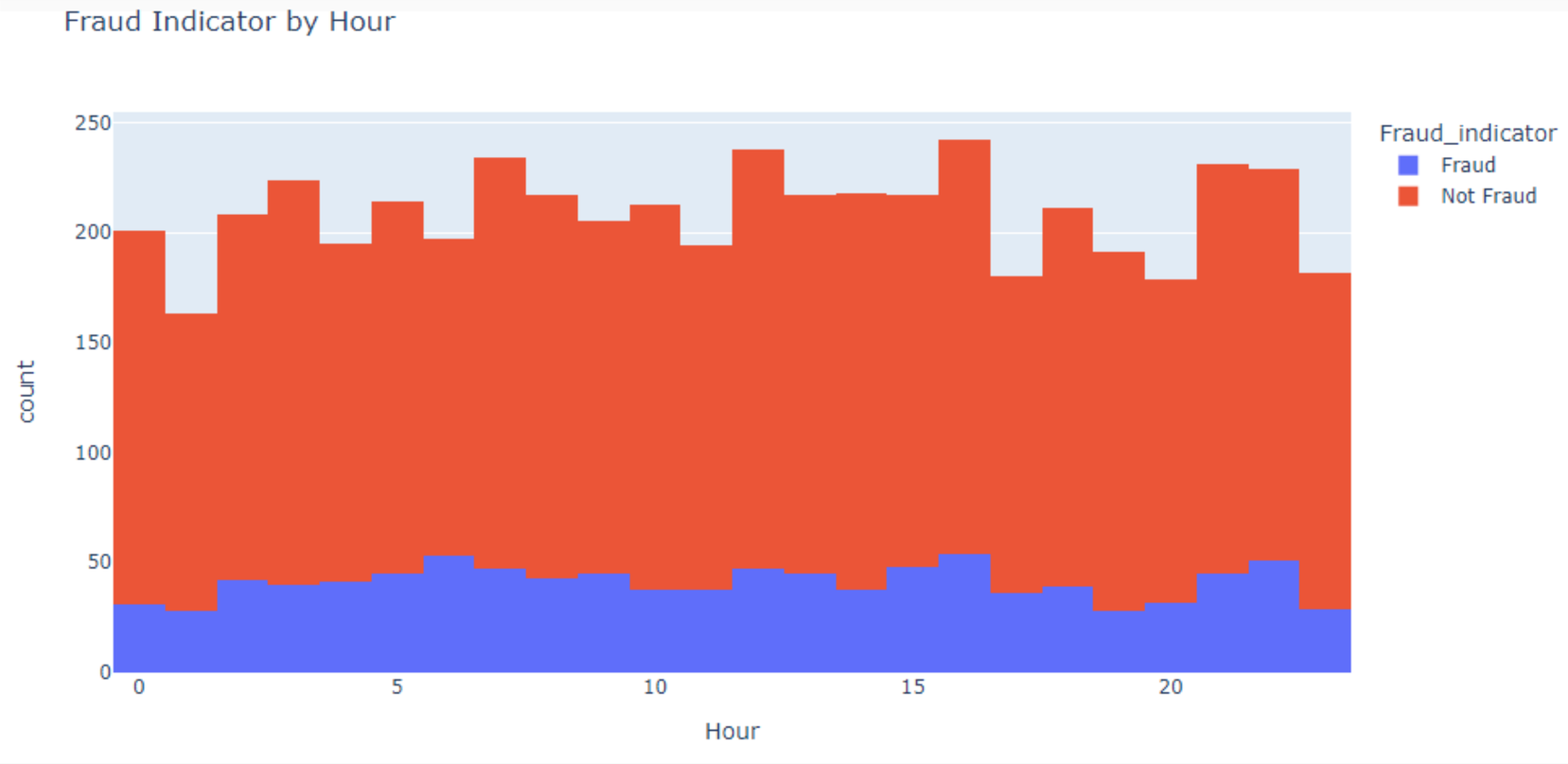
Illustrates the relationship between speed and transaction amount, highlighting transactions flagged as fraud. This can help identify unusual combinations of speed and transaction amount associated with fraudulent activities.

EXPLORATORY DATA ANALYSIS (EDA)

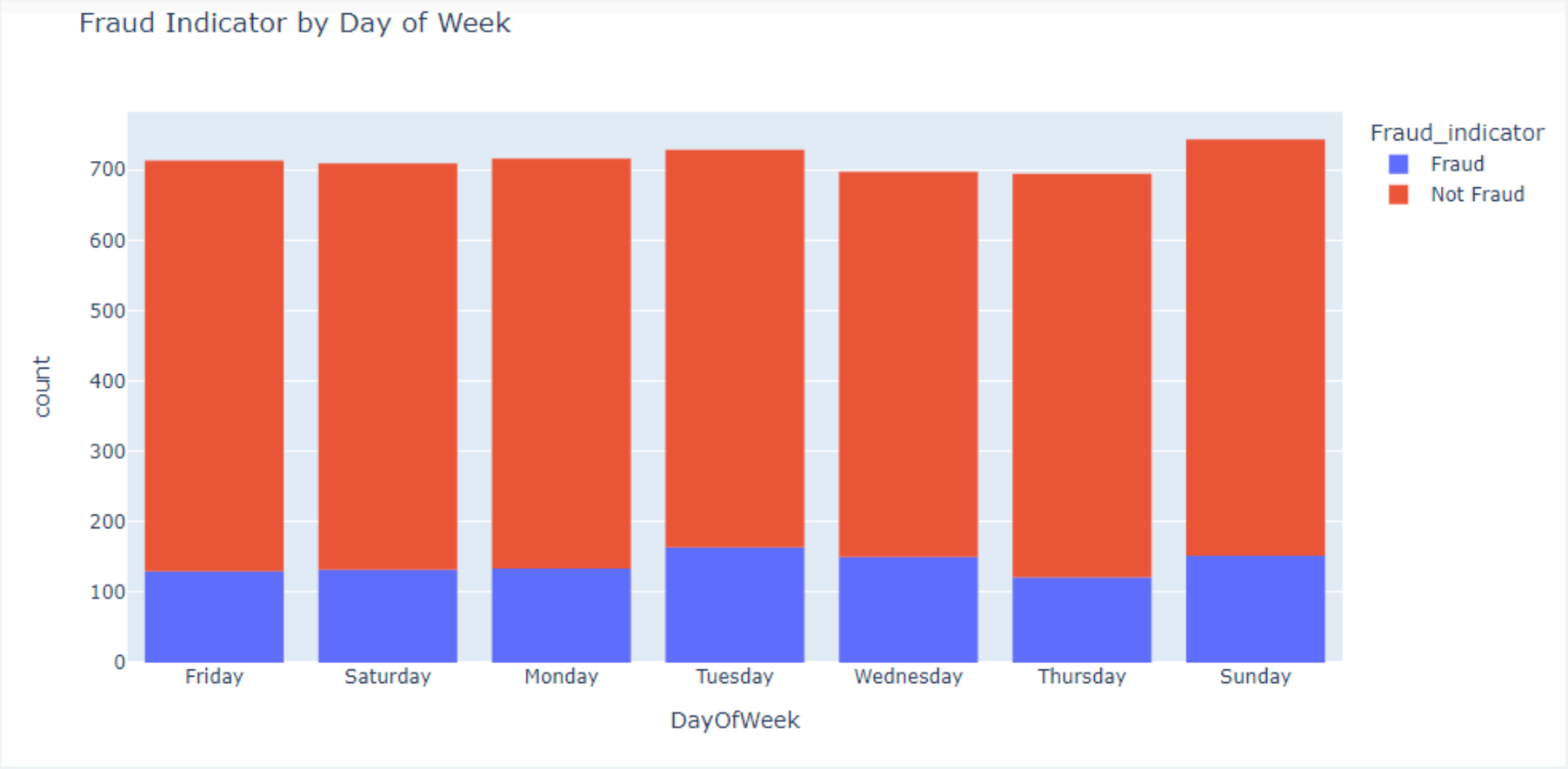


FEATURE ENGINEERING

Feature engineering involves creating new features or transforming existing ones to improve the model's performance.
By carefully engineering features, we improve the model's ability to learn from the data, leading to better predictions

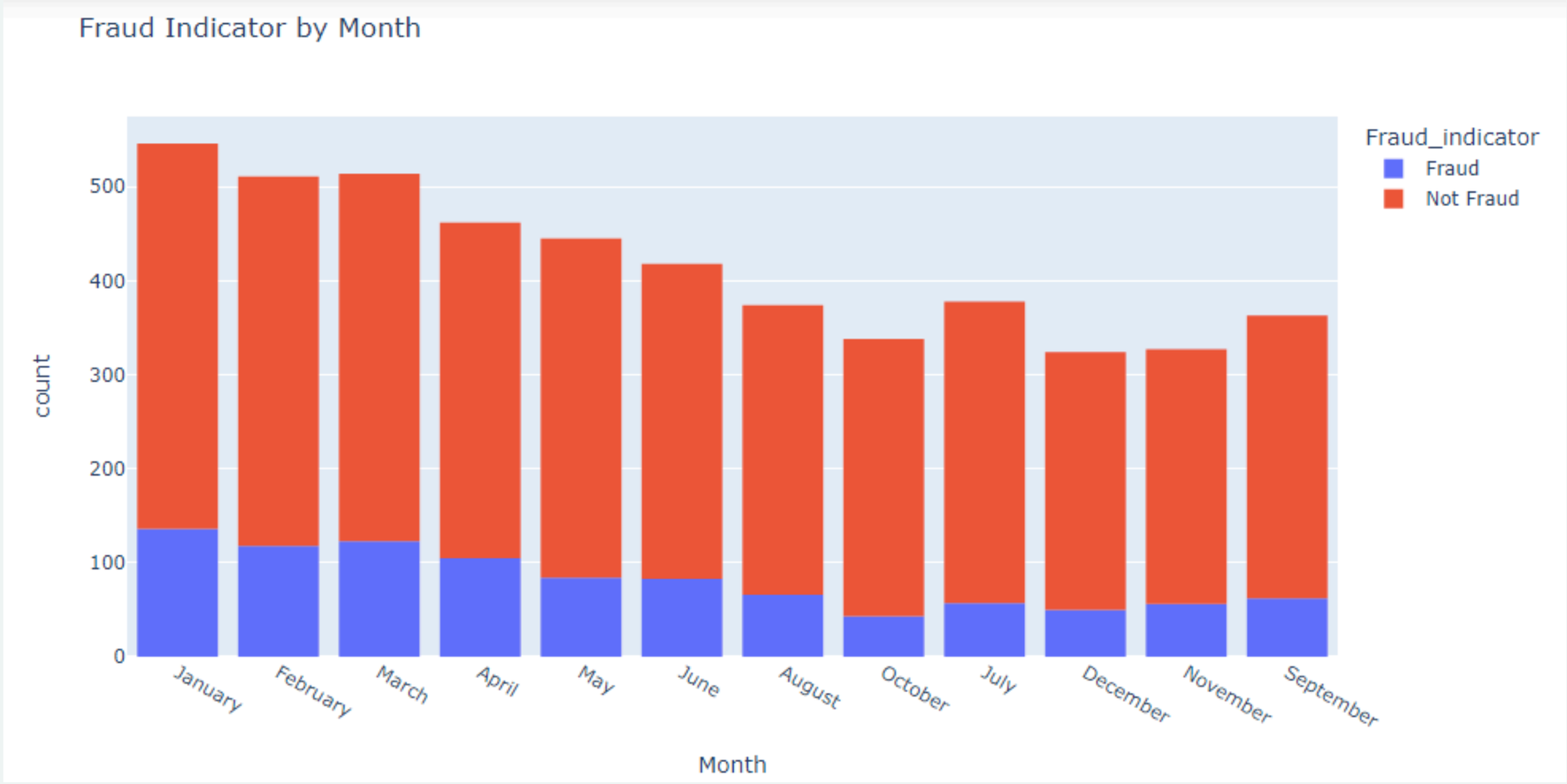


FEATURE ENGINEERING



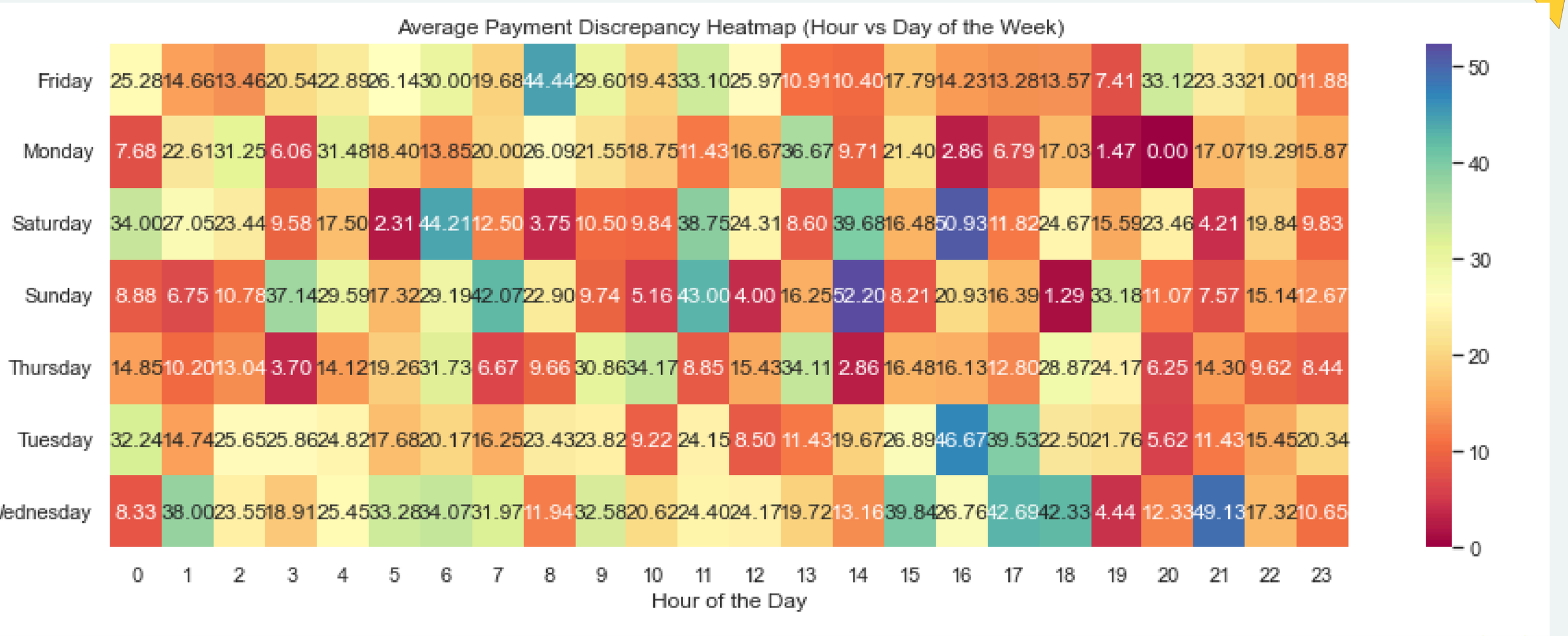
- The data suggests a possible correlation between fraudulent activity and transactions occurring on Tuesdays, Wednesdays, and Sundays.

FEATURE ENGINEERING



The data suggests a possible correlation between fraudulent activity and transactions occurring on January, February, and March.

FEATURE ENGINEERING



DATA PREPROCESSING

Handling Missing Values :

- Missing values in the dataset were handled using fillna technique to ensure the model has a complete dataset for training.
- For numerical features, we used the median value of each column to fill in the missing values.

The median is less sensitive to outliers compared to the mean, making it a robust choice for imputation

```
Transaction_ID              0
Timestamp                   0
Vehicle_Type                0
FastagID                  549
TollBoothID                 0
Lane_Type                   0
Vehicle_Dimensions          0
Transaction_Amount          0
Amount_paid                 0
Geographical_Location       0
Vehicle_Speed               0
Vehicle_Plate_Number        0
Fraud_indicator             0
dtype: int64
```

	Transaction_ID	Transaction_Amount	Amount_paid	Vehicle_Speed	Hour	Payment_Discrepancy
count	5000.000000	5000.00000	5000.000000	5000.000000	5000.000000	5000.000000
mean	2500.500000	161.06200	141.261000	67.851200	11.552800	19.801000
std	1443.520003	112.44995	106.480996	16.597547	6.817427	56.097547
min	1.000000	0.00000	0.000000	10.000000	0.000000	-80.000000
25%	1250.750000	100.00000	90.000000	54.000000	6.000000	0.000000
50%	2500.500000	130.00000	120.000000	67.000000	12.000000	0.000000
75%	3750.250000	290.00000	160.000000	82.000000	17.000000	0.000000
max	5000.000000	350.00000	350.000000	118.000000	23.000000	290.000000

MODEL DEVELOPMENT

Random Forest is a powerful and versatile ensemble learning algorithm that is used for classification and regression tasks. It builds upon the concept of decision trees, combining the output of multiple trees to improve predictive performance and control overfitting.

SMOTE (Synthetic Minority Over-sampling Technique): SMOTE is a popular method to address class imbalance. It works by generating synthetic samples for the minority class (in this case, fraudulent cases) to make the dataset more balanced. This is done by interpolating between existing minority class examples rather than simply duplicating them.

```
import joblib
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV
from imblearn.over_sampling import SMOTE
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

```
# Apply SMOTE to handle class imbalance
X_res, y_res = SMOTE(random_state=42).fit_resample(X, y)
X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.3, random_state=42)
```

MODEL EVALUATION

```
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

```
Accuracy: 0.9108910891089109
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.89      0.91       160
           1       0.88      0.94      0.91       143

    accuracy                           0.91       303
   macro avg       0.91      0.91      0.91       303
weighted avg       0.91      0.91      0.91       303

Confusion Matrix:
[[142  18]
 [  9 134]]
```

THANK YOU