New York City Taxi Trip Duration Prediction Report

Introduction

This project aims to develop a machine learning model to predict taxi trip durations based on historical trip data. Accurately estimating trip duration is essential for optimizing ride dispatch systems, improving customer service, and enhancing driver efficiency. The model incorporates multiple geographic and temporal features to make reliable predictions.

Objectives

- Develop a predictive model for taxi trip duration using key features such as distance, direction, and pickup/dropoff datetime.
- Improve model performance using Ridge regression with preprocessing techniques.
- Evaluate the model's effectiveness using relevant performance metrics.

Data Description

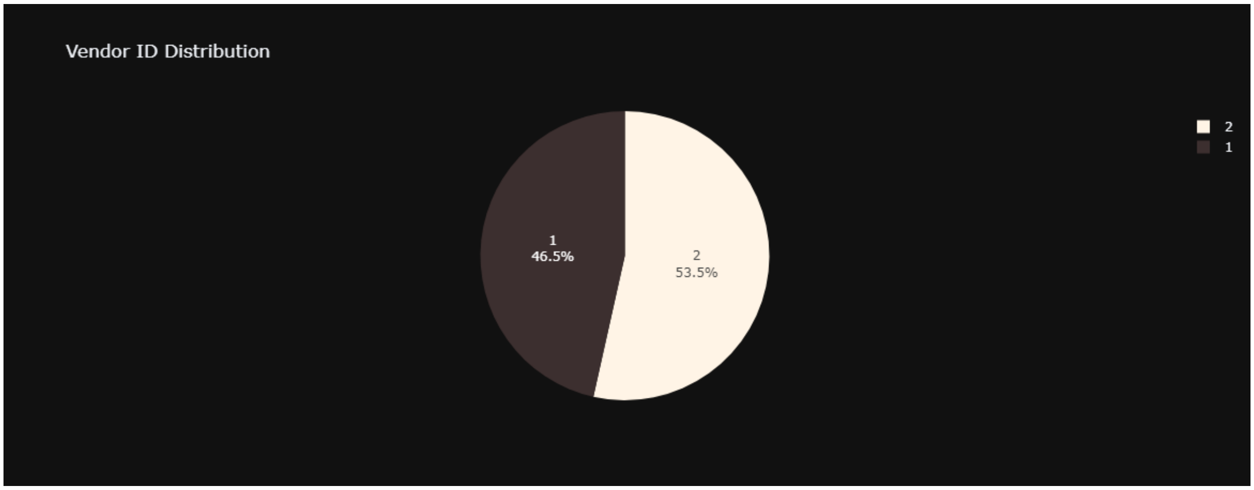
The dataset consists of various features capturing trip characteristics, including:

- **Trip ID**: A unique identifier for each trip record.
- **Vendor ID**: A code indicating the provider associated with the trip, useful for performance analysis.
- **Pickup Datetime**: The timestamp marking the start of the trip, which helps analyze peak hours and seasonal trends.
- **Dropoff Datetime**: The timestamp marking the end of the trip, allowing calculation of trip duration.
- **Passenger Count**: Number of passengers per trip, useful for studying demand patterns.
- **Pickup Longitude & Latitude**: Geographic coordinates of where the trip started, used for mapping popular pickup areas.
- **Dropoff Longitude & Latitude**: Geographic coordinates of where the trip ended, helping identify frequent dropoff locations.
- **Store and Forward Flag**: Indicates whether the trip record was stored before being sent to the vendor, which may indicate network issues.
- **Trip Duration**: The target variable representing trip time in seconds, essential for predictive modeling.
- **Additional Engineered Features**: Distance, direction, Manhattan distance, and time-based features.

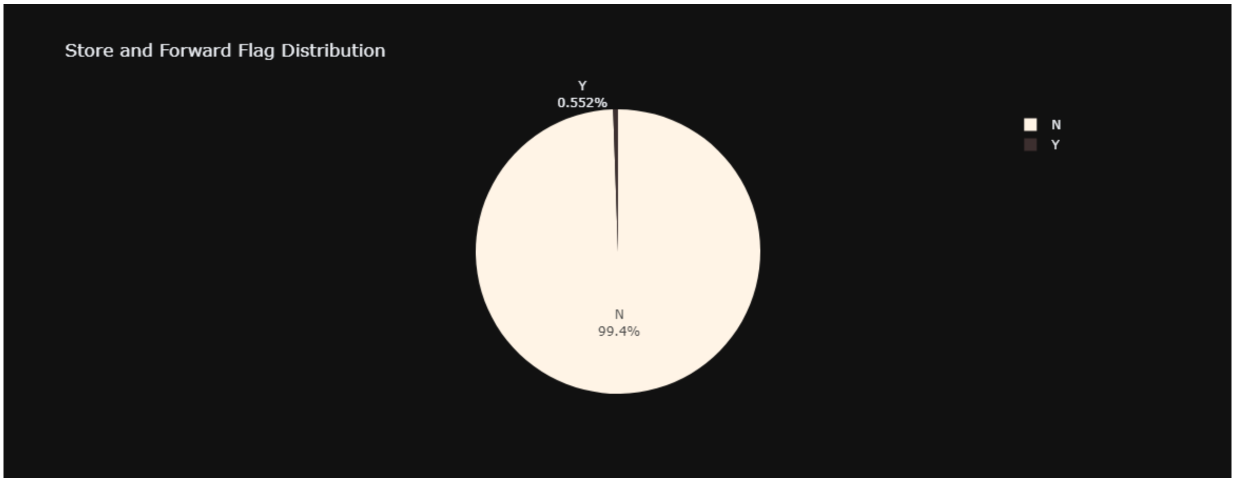
Exploratory Data Analysis (EDA)

To better understand the dataset, we conducted various exploratory analyses:

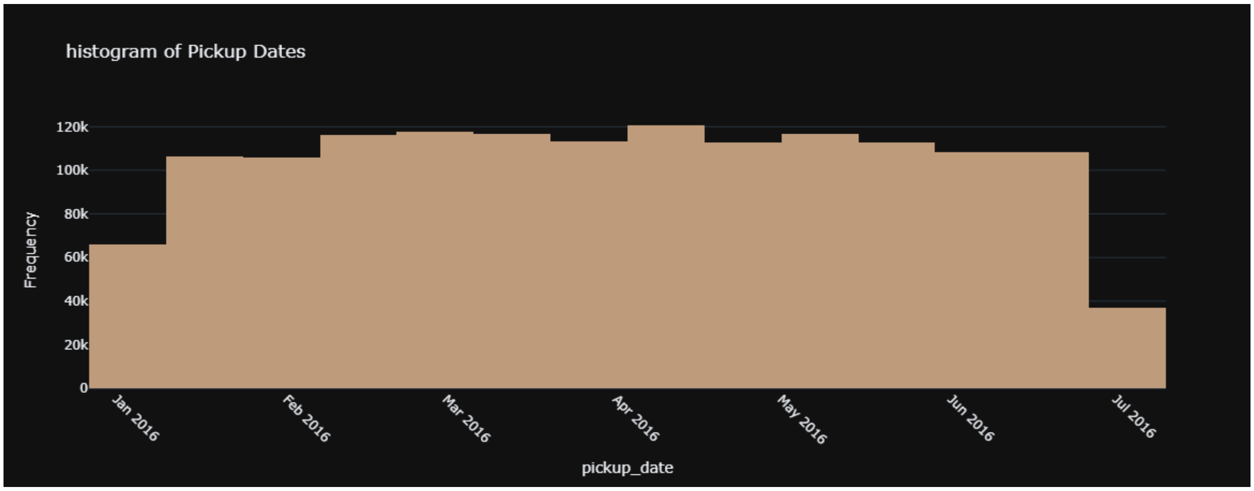
- **Vendor ID Distribution**: The trips are distributed among different vendors, helping analyze vendor-specific trends.



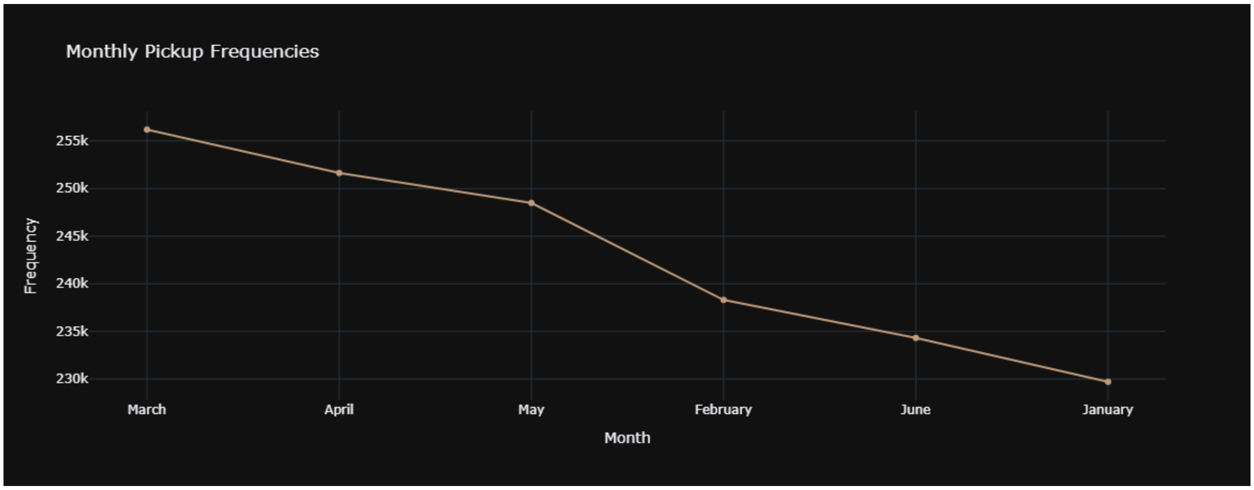
- **Store and Forward Flag Distribution**: A small fraction of trips were stored before being sent, indicating possible network issues.



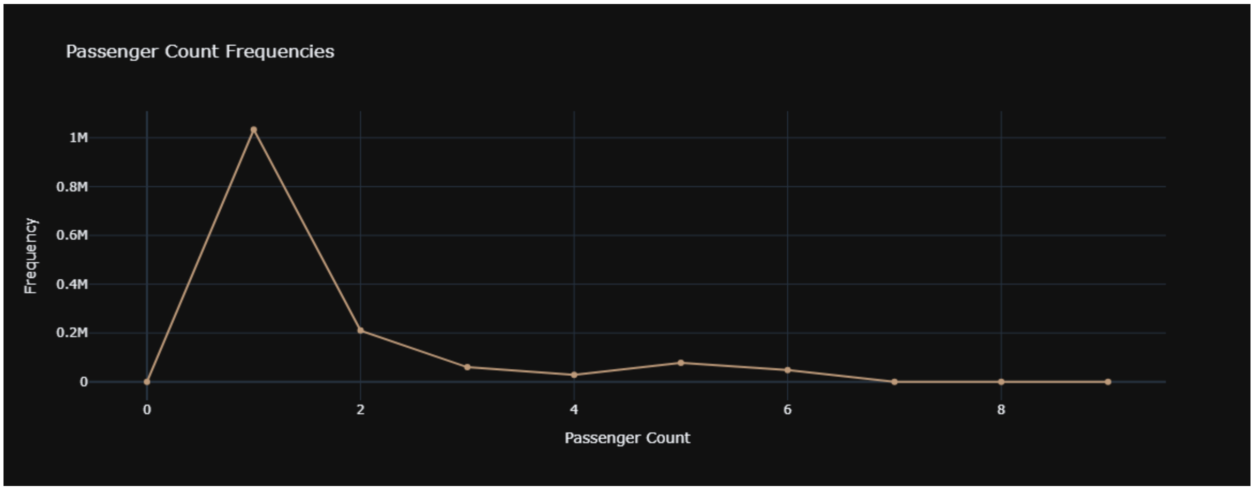
- **Histogram of Pickup Dates**: Shows the daily distribution of taxi pickups, identifying high and low-demand days.



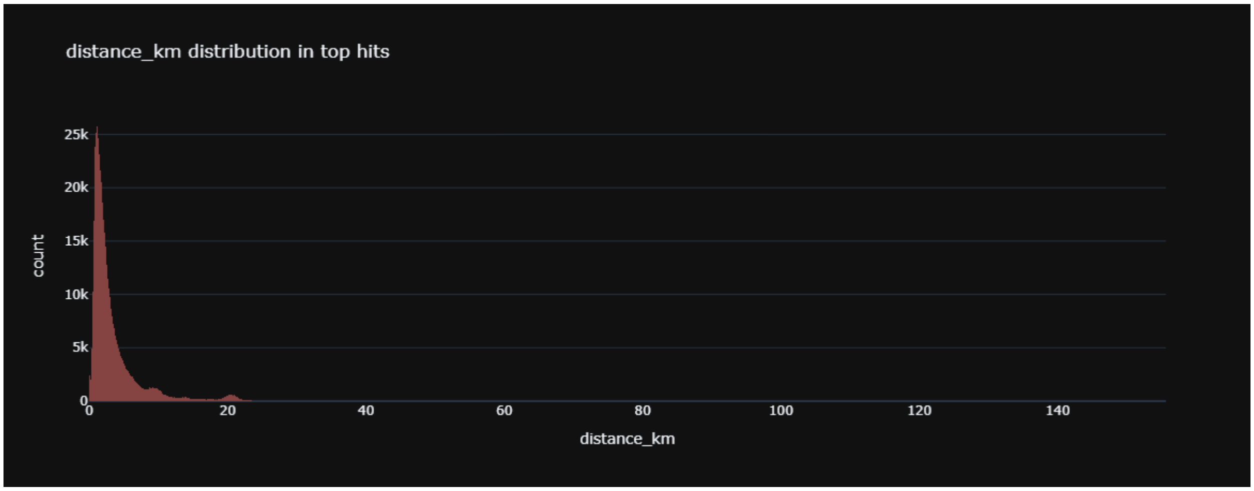
- **Monthly Pickup Frequency**: The frequency of pickups varies across different months, highlighting seasonal demand trends.



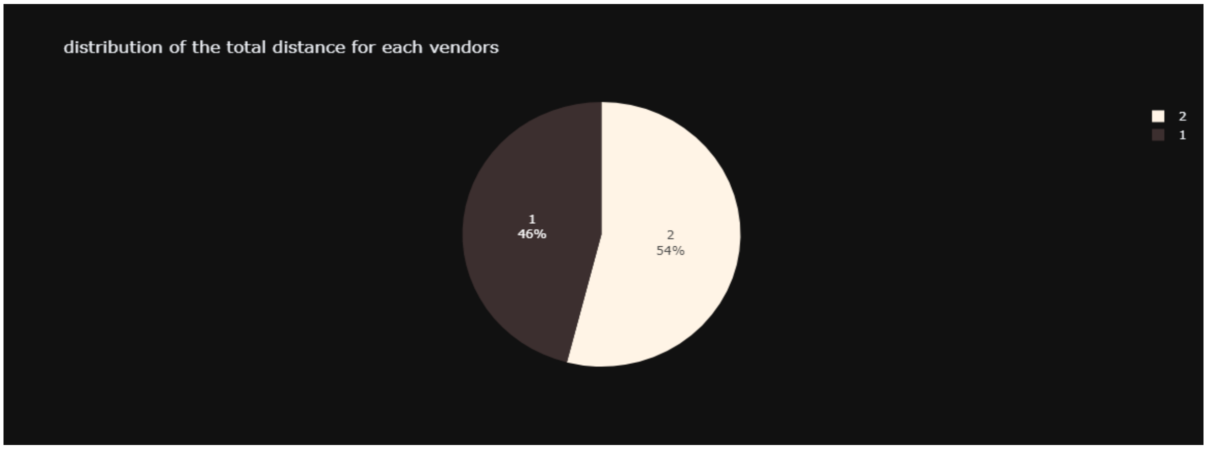
- **Passenger Count Analysis**: Most trips have 1-2 passengers, with occasional larger groups.



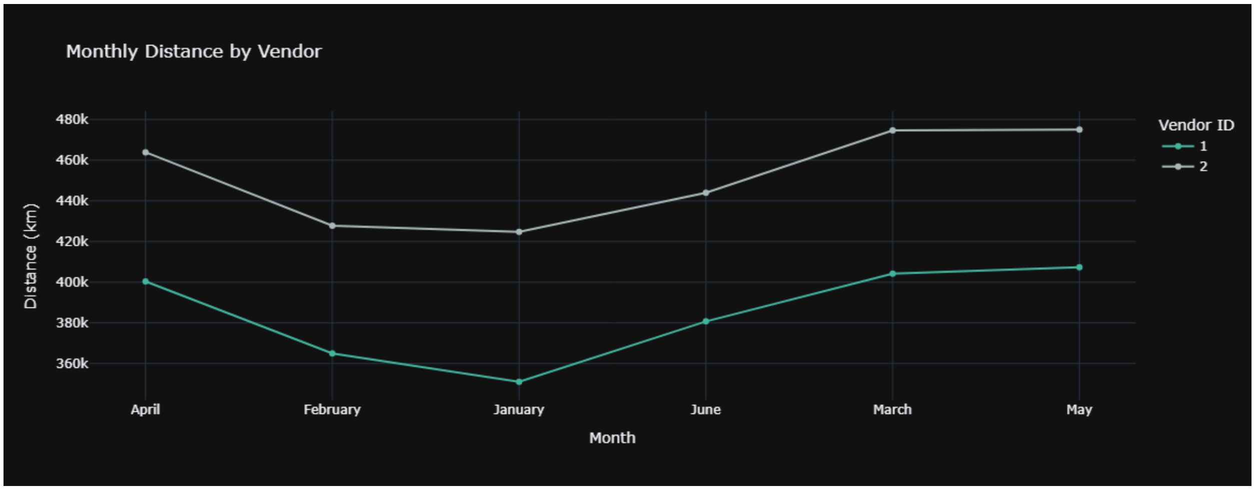
- **Distance Distribution**: Most trips are within a short distance, with a few long trips causing right skewness.



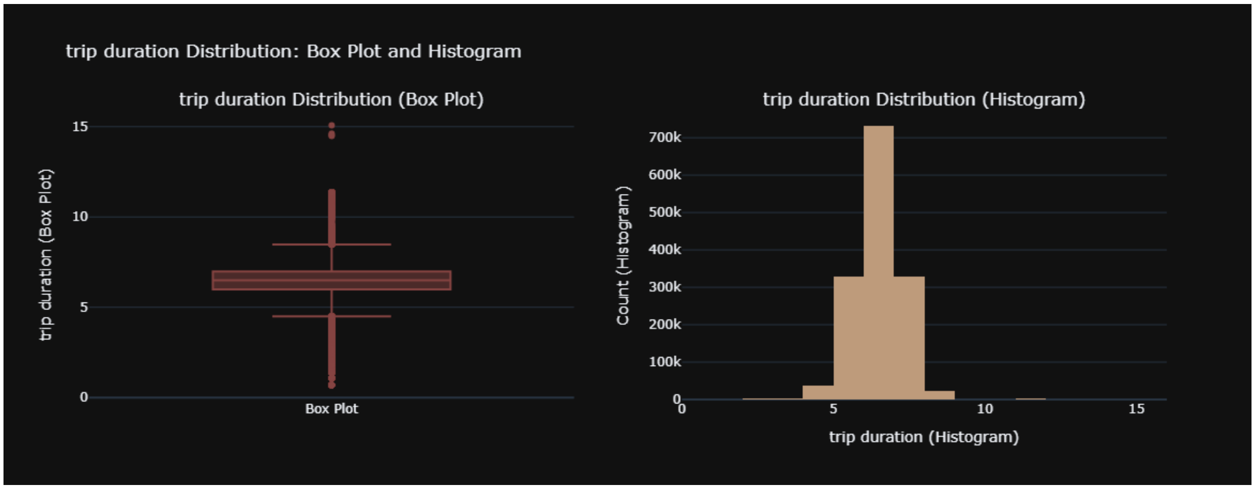
- **Total Distance for Each Vendor ID**: A comparison of the total distance traveled by different vendors shows operational variati



- **Monthly Distance by Vendor**: Examines distance of trip patterns across different vendors in each month.



- **Trip Duration Distribution**: The histogram of trip durations shows a skewed distribution, which was addressed using log transformation.



Feature Engineering

To enhance model performance, various preprocessing techniques and feature engineering steps were applied:

- **Distance Metrics**:
  - **Haversine Distance**: Measures the straight-line distance between pickup and dropoff locations.
  - **Manhattan Distance**: Accounts for urban road layouts by measuring vertical and horizontal travel.
- **Direction Calculation**: Computes the bearing between pickup and dropoff points.
- **Time-based Features**:
  - Hour of the day, day of the week, month, and season.
  - Classification of trips into morning, afternoon, evening, and night.
  - Identification of peak hours and weekends.
- **Data Transformation**:
  - Log transformation applied to distance and trip duration to reduce skewness.
  - One-hot encoding applied to categorical features like time periods and seasons.

Modeling Approach

A machine learning pipeline was developed using Ridge regression with polynomial features and scaling:

1. **Preprocessing Pipeline**:
    - Standardization of numerical features.
    - Polynomial feature expansion (degree = 3) for interaction terms.
    - One-hot encoding of categorical variables.
2. **Regression Model**:
    - Ridge Regression with regularization to prevent overfitting.
    - Hyperparameter tuning to optimize performance.

Evaluation Metrics

The model was evaluated using:

- **Root Mean Squared Error (RMSE)**: Measures the average prediction error.
- **R-squared (R^2) Score**: Assesses how well the model explains variance in trip duration.

Results

- **Training R^2 Score**: 0.677
- **Test R^2 Score**: 0.67

These scores indicate that the model explains approximately 67% of the variance in trip durations, demonstrating a strong predictive ability while leaving room for further optimization.

Conclusion

Key Findings

- **Trip distance, direction, and time-based features are critical predictors.**
- **Ridge regression with feature engineering improves accuracy while preventing overfitting.**
- **Log transformations and polynomial feature expansion enhance performance.**

Future Improvements

- **Explore advanced models** such as Gradient Boosting Machines or Random Forest for improved accuracy.
- **Incorporate real-time traffic and weather data** to enhance predictions.
- **Optimize hyperparameters further** using grid search or Bayesian optimization.

This model provides a solid foundation for trip duration prediction and can be further enhanced with additional features and fine-tuning.