***New York City Taxi Trip Duration***

1. **Introduction**
   * **The goal of this project is to predict taxi trip durations using a dataset of historical taxi trips. Accurately predicting trip duration is crucial for improving ride dispatch systems, optimizing customer service, and enhancing driver efficiency. This report presents the process of building a regression model to predict trip durations based on various features, including trip distance, passenger count, and time-based variables.**
   1. **Objectives**
      1. Develop a predictive model for taxi trip duration using features such as distance, direction, and pickup/dropoff datetime.
      2. Improve model performance using Ridge regression with optimal hyperparameters.
      3. Evaluate the model’s performance and analyze its effectiveness.
   2. **Data Description**

**The dataset consists of 100,000 rows, each representing a single taxi trip. The key columns used in this analysis include:**

* + 1. trip\_duration: The duration of the trip in seconds (target variable).
    2. pickup\_datetime and dropoff\_datetime: The timestamp of pickup and dropoff.
    3. passenger\_count: The number of passengers in the taxi.
    4. pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude: Coordinates of pickup and dropoff locations.
    5. vendor\_id: The ID of the taxi provider.
    6. store\_and\_fwd\_flag: Whether the trip record was held in the vehicle's memory before being sent to the server.
    7. Distance and direction features derived from pickup and dropoff locations.+

1. **Exploratory Data Analysis (EDA)**
   1. **Trip Duration (Target Variable)**

The Trip Duration Distribution in following the Gaussian Distribution in Figure 1,

Ane you can see outliers in the right tail.

Most of the data is around 6 and 5

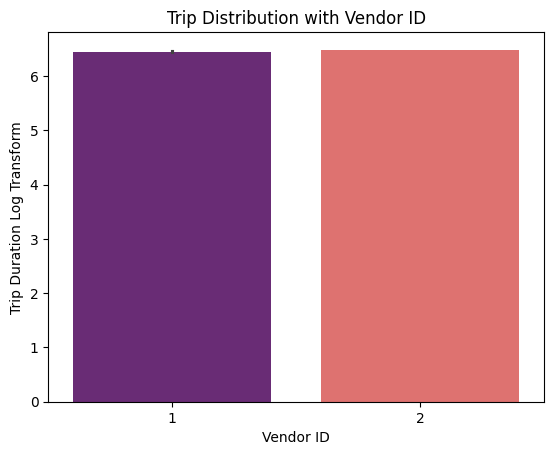
**Note: we scaled the data using log(X + 1)** for better visualization and to help us with modeling large values.

A graph of a distribution of a log

Description automatically generated

Figure 1: Distribution of Trip Duration in seconds log Transformer Applied

* 1. **Discrete Numeric features**

 A graph of a trip distribution with passenger count

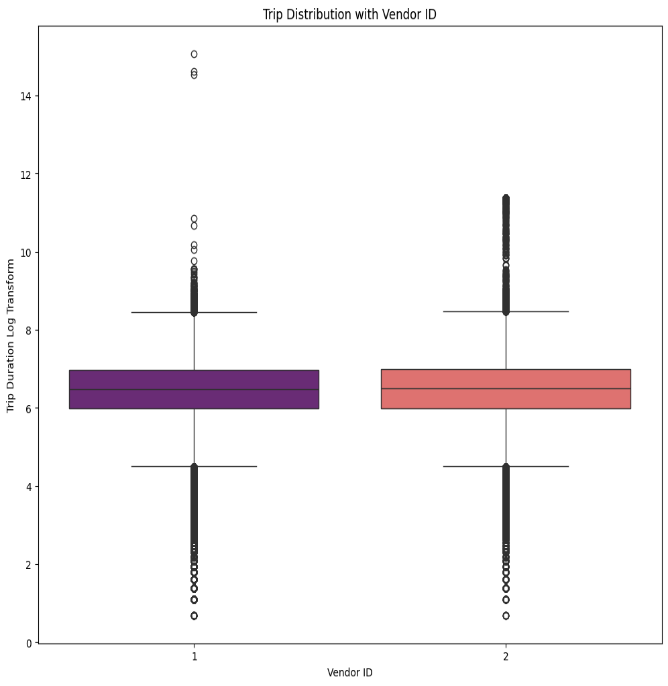
Description automatically generated

Figure 2:Dicrete Numeric Features

**Trip duration and Vendor ID:** It’s difficult to discern any clear change between trip duration and vendor ID from the left bar chart.

**Passenger Count:** When the number of passenger groups from [1 to 6] take constant trip duration and the number of passenger groups from [7 to 8] take less trip duration.

**And for the 0 Passenger maybe the driver enters the number wrong or error in the system**

**A graph of a number of colored squares

Description automatically generated with medium confidence**

Figure 3: Discrete Numeric Feature(boxplot)

**Trip Distribution:** Appears to be positive skewed in both Groups and this means that short trips is more than the longer trips

**Outliers**: There are a few outliers for both groups, represented by the circles beyond the whiskers. These are individual trips that were much longer than the majority of trips in their respective groups and this Supports conclusion passengers groups from [7 to 8] just travel less than another groups because Trip purpose.

* 1. **Geography Data**

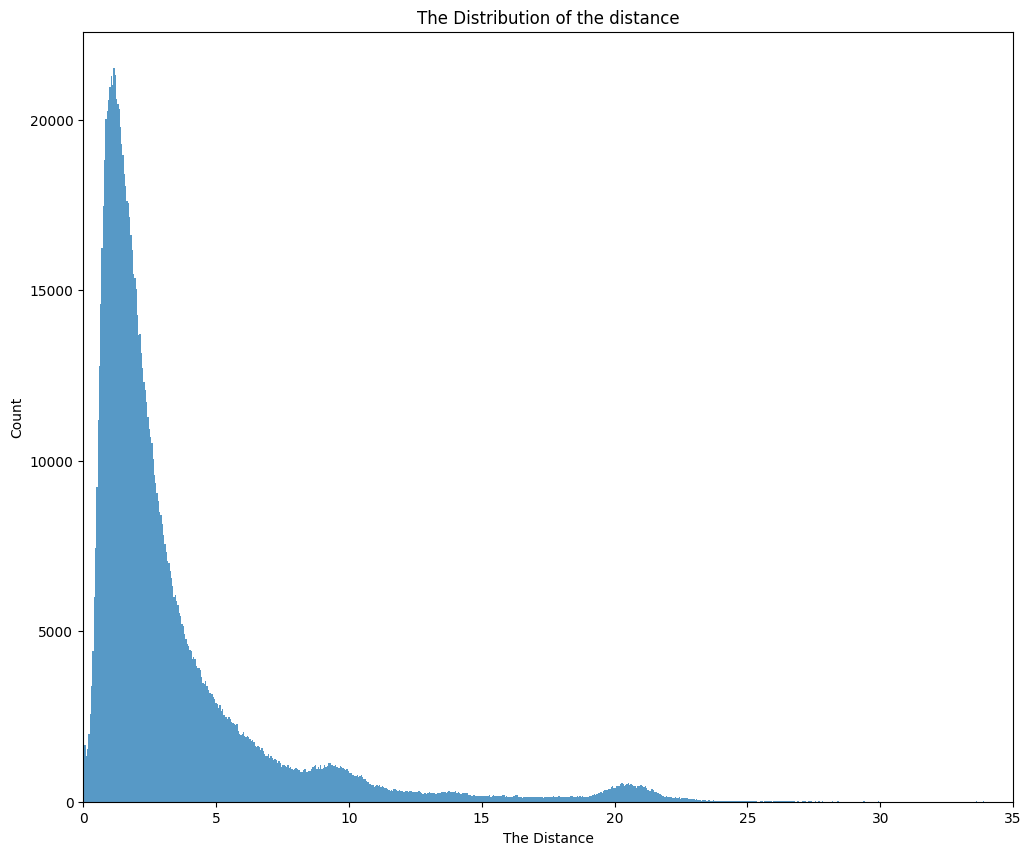
Using Coordinates of pick-up and drop-off locations to calculate the haversine distance

Figure 4: The Distribution of the Distance

**The distribution of speed:** it is right skewed distribution, so we can transform the data using **log(X+1)**

look like most of the trip's distance are from [1 to 25]km but most of them are in the range between [1 to 5]km

**Now we have the Distance and the Trip Duration we can calculate the speed**

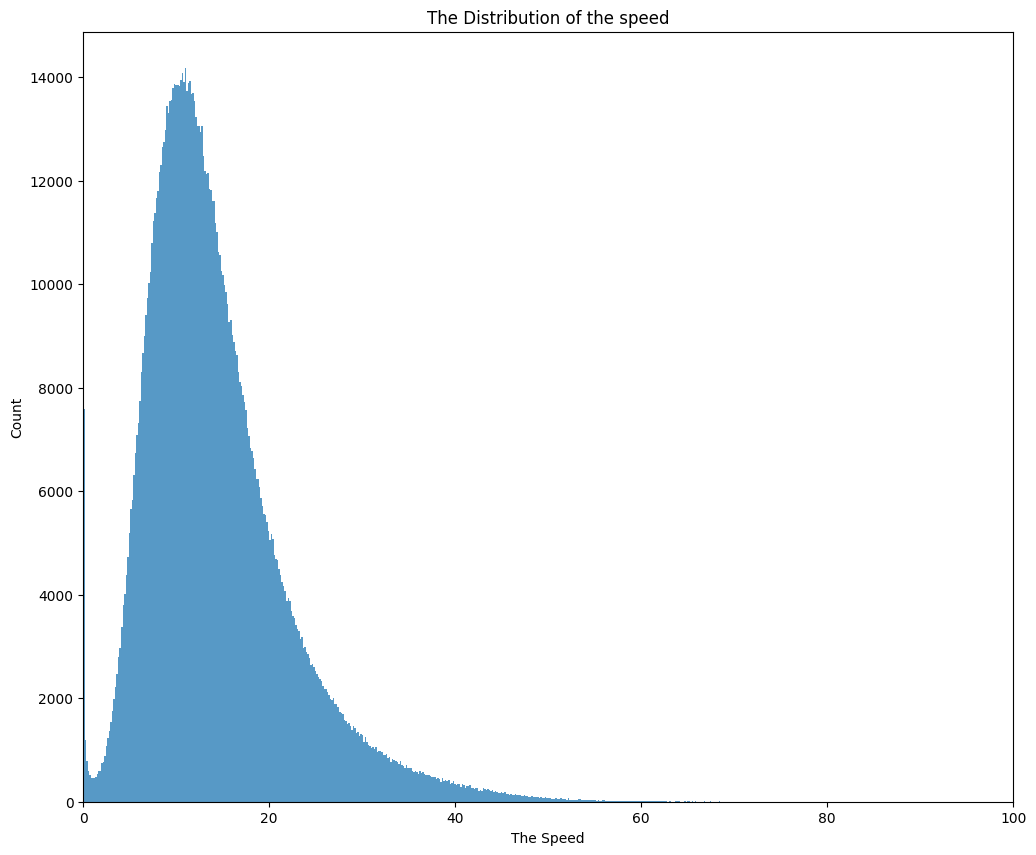
****

Figure 5: Distribution of the Speed

* 1. **Time Feature Analysis**

Using pickup\_datetime we can get new information like Months, Period of the day, season for each trip. **Correlation Analysis**

**Correlation Analysis**

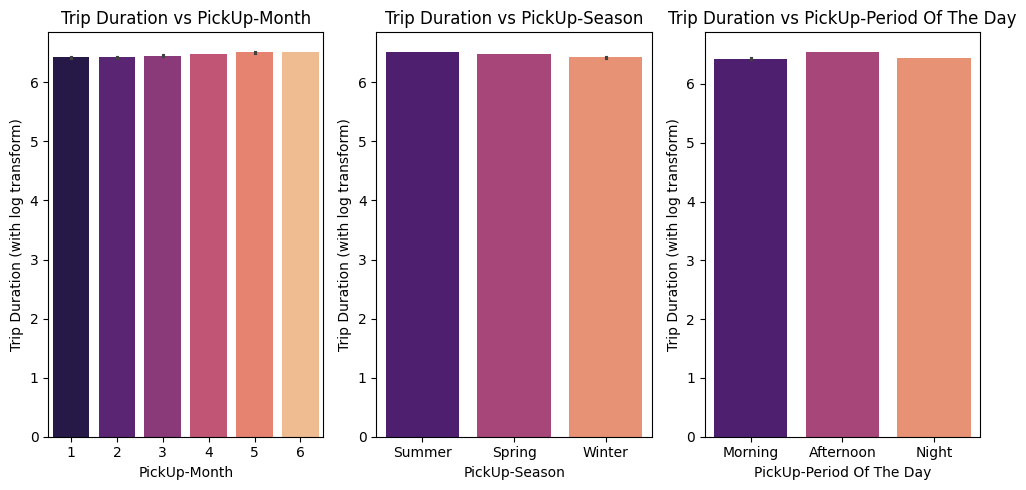
****

Figure 6: Time Feature Analysis

* We can see that in [4,5,6] months there a little increase maybe it’s because 4 and 5 are Spring season and for 6 maybe because the start of the summer vacation
* The Summer season is the highest between other seasons maybe because the start of the summer vacation

1. **Feature Engineering**
   1. **Datetime-based Features**

**We extracted several datetime-related features from the pickup\_datetime column to capture temporal patterns:**

* + 1. **Hour of Day**: Extracted the hour to capture rush hour effects.
    2. **Day of Week**: Encoded as one of seven days.
    3. **Month**: Extracted the month to observe any monthly trends.
    4. **Season**: Grouped months into four seasons: winter, spring, summer, and fall.
    5. **Time Period**: Segmented the day into morning, afternoon, evening, and night.
  1. **Distance and Direction**

**We calculated the following distance metrics between pickup and dropoff locations:**

* + 1. **Haversine Distance**: The straight-line distance between the two points on the globe.
    2. **Manhattan Distance**: The distance assuming only vertical and horizontal travel, which is more realistic for urban areas with grid-like road systems.
    3. **Direction**: The compass bearing from pickup to dropoff.

1. Correlation Analysis

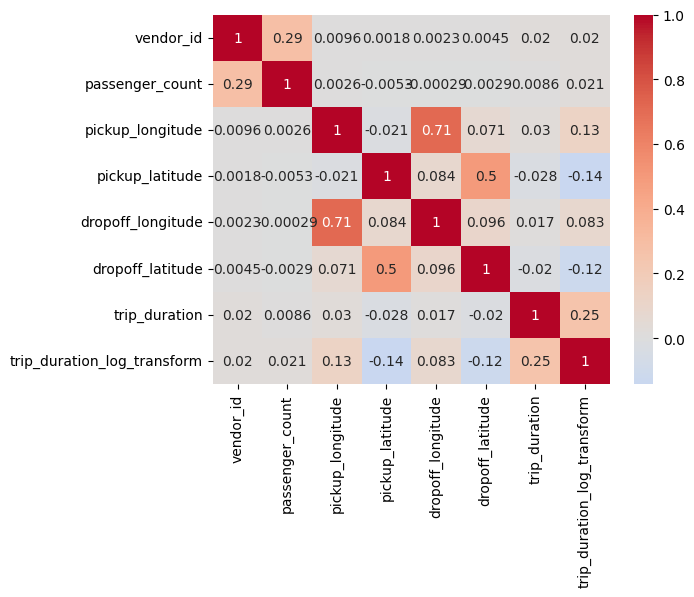


Figure 7: Categorical Data Correlation Matrix

* we have four positive relation trip duration (with log) DropOff longitude, PickUp longitude, passenger count and vendor id
* we have two negative relation trip duration (with log) DropOff latitude, PickUp latitude
  1. I found that **Haversine Distance and Manhattan Distance** are right skewed

So, I apply **log(X+1)** for better performance.

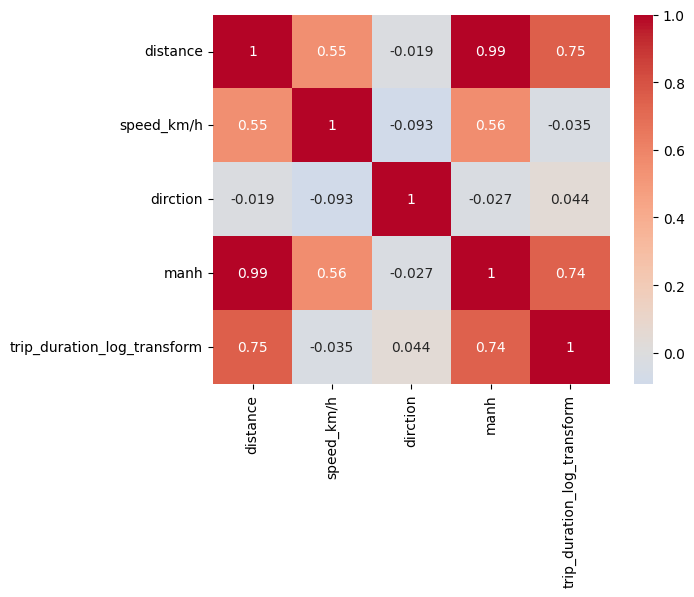


Figure 8: Numeric Data Correlation Matrix

* There is strong positive relation trip duration (with log) with distance and Manh.
* There negative relation trip duration (with log) with speed kmh.
  1. **Handling Categorical Variables**

Categorical features, such as vendor\_id, passenger\_count, and store\_and\_fwd\_flag, were encoded using one-hot encoding. For passenger\_count, infrequent categories were combined into a single “other” category to reduce noise in the data.

* 1. **Transformations and Log Scaling**

To address the right-skewed nature of trip\_duration, a log transformation was applied. This transformation reduced skewness and made the data more suitable for regression modeling.

* 1. **Outlier Handling**

Outliers in the trip\_duration variable (extremely short or long trips) were identified using the IQR method and removed. This step ensured that the model was not unduly influenced by outliers.

* **Modeling**

**Baseline Models**

**Initial models included:**

**Linear Regression**: Provided a baseline R² score of ~0.55.

* 1. **Final Model - Ridge Regression**

The final model chosen was **Ridge Regression** with an alpha value of 1. Ridge regression was selected to control overfitting by penalizing large coefficients while maintaining all features in the model. Ridge's regularization effectively balanced bias and variance.

* 1. **Model Performance**
     + **Train R² Score**: 0.6997
     + **Validation R² Score**: 0.6946

These scores indicate that the model explains about 69.9% of the variance in the trip duration, which is a significant improvement over the baseline models.

* **Conclusion**
  1. **Summary of Findings**
     1. **Key Features**: Trip distance, hour of the day, and Manhattan distance were the most important predictors of trip duration.
     2. **Model Effectiveness**: Ridge regression with alpha=1 performed well, achieving a validation R² of 0.6946. The use of regularization helped prevent overfitting, which was evident in previous models.
     3. **Feature Importance**: Distance metrics and time-based features (hour and day of the week) had the largest impact on the prediction of trip duration.
  2. **Next Steps**
* **Advanced Models**: Exploring more sophisticated models like Gradient Boosting Machines (GBM) or Random Forest could further improve prediction accuracy.
* **Feature Expansion**: Incorporating traffic data or weather conditions could add valuable information for future predictions.