New York City Taxi Trip Duration Report

By: Ahmed Reda

1. Introduction Project Goal

Develop a predictive model to forecast taxi trip durations using historical data. Accurate predictions enhance ride dispatch efficiency, customer satisfaction, and driver resource allocation.

Significance
- Optimizes operational planning for taxi companies.
- Reduces passenger wait times and improves service reliability.

1.1 Key Objectives

- Predictive Model Development:
  - Use features like trip distance, direction, and pickup/dropoff times to forecast trip duration.
- Model Optimization:
  - Apply Ridge regression with hyperparameter tuning to minimize overfitting and improve accuracy.
- Performance Evaluation:
  - Validate the model using metrics like RMSE and MAE to ensure reliability.

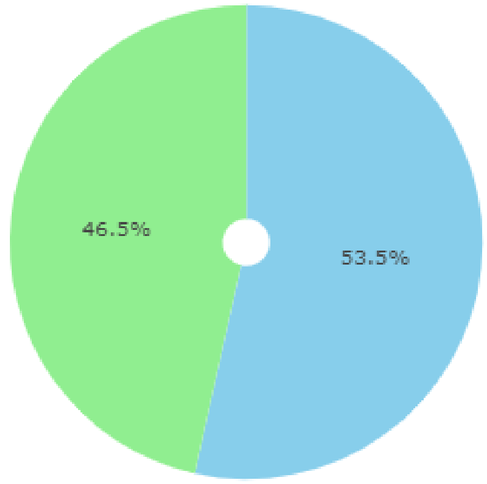
1.2 Dataset Overview Dataset Size: 100,000 taxi trip records.

Key Features:
- Target Variable:
- `trip_duration` (in seconds).
- Temporal Data:
- `pickup_datetime`, `dropoff_datetime`.
- Spatial Data:
- Pickup/dropoff coordinates (`latitude`, `longitude`).
- Operational Metrics:
- `passenger_count`, `vendor_id`, `store_and_fwd_flag`.
-Engineered Features:
- Derived distance (e.g., Haversine) and directional metrics.

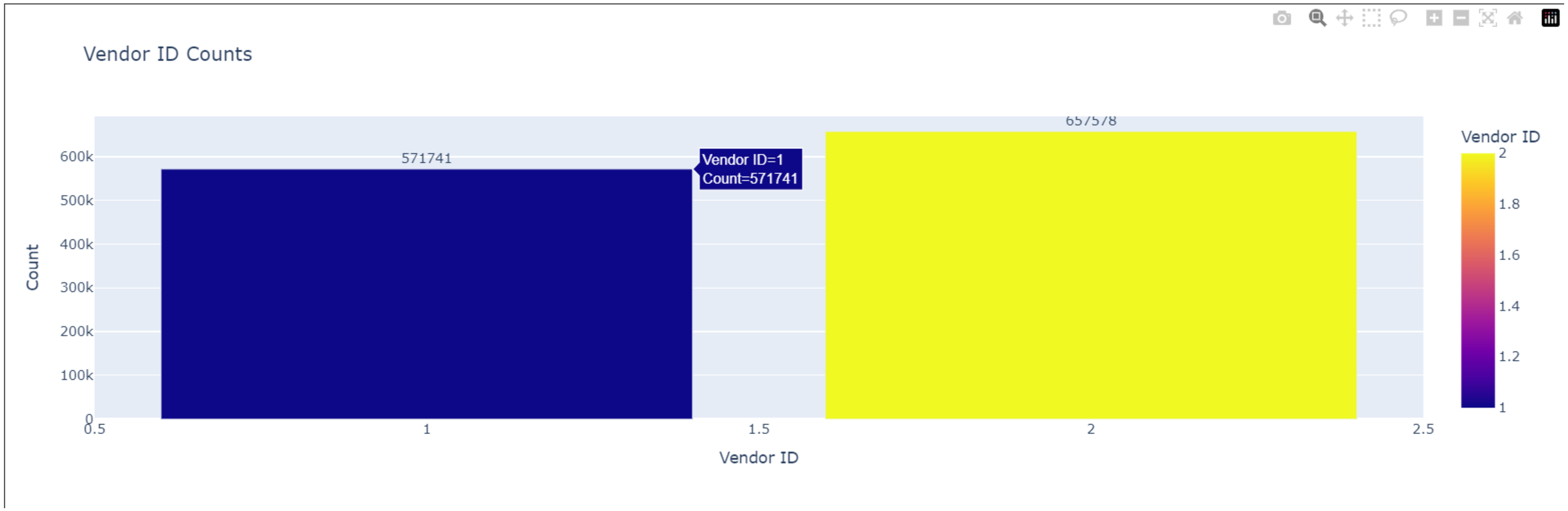
2. Exploratory Data Analysis (EDA)

2.1 Univariate Analysis

vendor_id



This graph illustrates the **proportional split between two taxi vendors** (Vendor IDs). Vendor 1 accounts for **46.5%** of trips, while Vendor 2 holds a slightly larger share at **53.5%**. The near-balanced distribution suggests no overwhelming market dominance, though Vendor 2 operates marginally more trips. Such insights help analyze service availability, competition, or operational efficiency between providers. The data likely reflects vendor-specific customer preferences or geographic coverage differences.



pickup_datetime

1. Interpretation:

- **Rows (Hour of Day)** : The rows represent the 24 hours of the day, starting from midnight (0:00) up to 23:00. Each row shows the count of trips during that specific hour for all days of the week.

- **Columns (Day of Week)** : The columns represent the days of the week. Each column corresponds to a day (e.g., "Monday", "Tuesday", etc.). The numbers inside the cells show the number of trips that occurred during that hour and on that specific day.

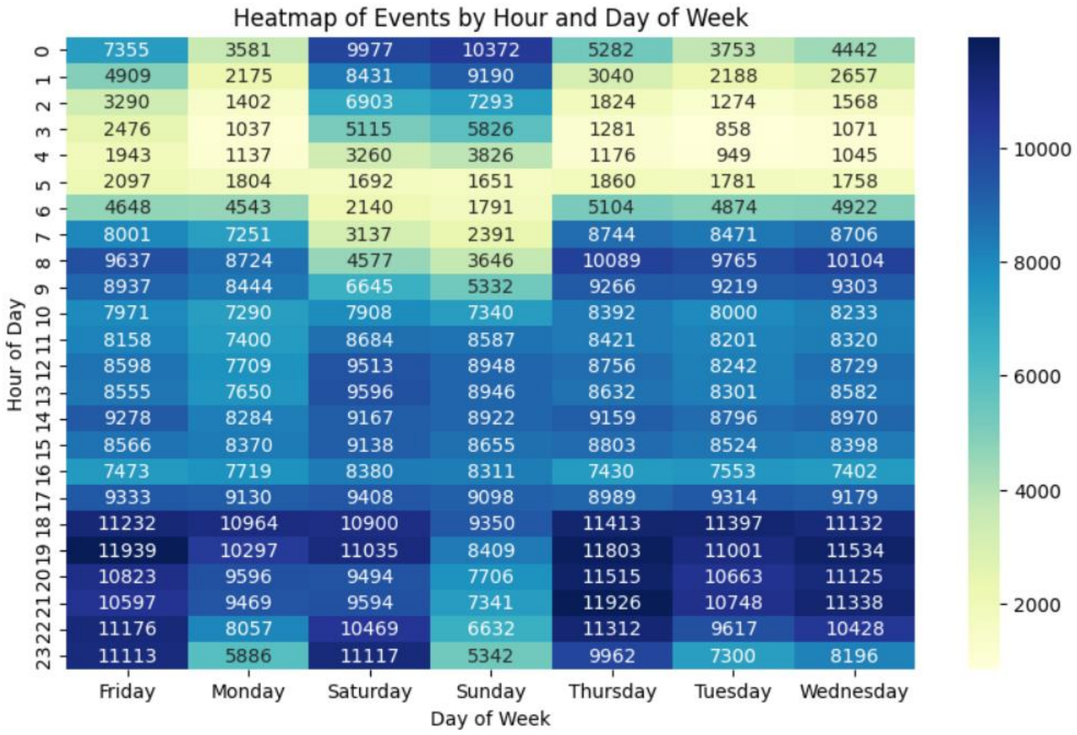
- **Cell Values**: The number inside each cell represents the count of trips that occurred during that specific hour on that specific day. The higher the number, the darker the shade of the color (as per the color map).

2. Insights:

-**Busy Times**: By looking at the heatmap, you can easily identify the hours of the day when most trips occurred (indicated by darker colors), and you can compare this across different days of the week.

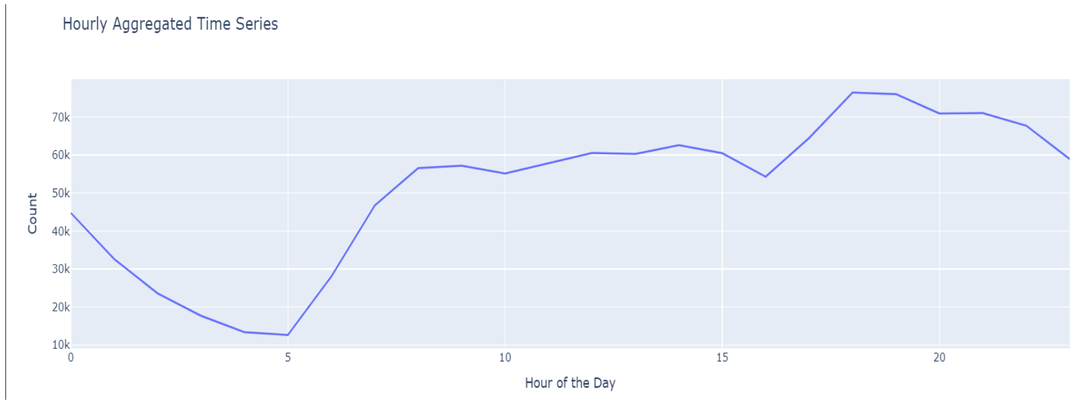
- **Trends**: You might see trends such as:
  - Weekdays like Monday to Friday having more trips during business hours.
  - Weekends (Saturday, Sunday) may have a different pattern (perhaps later or earlier in the day).

- **Outliers**: If certain time slots or days have unusually high or low numbers of trips, this might indicate outliers or patterns worth investigating.



Hourly Aggregated Time Series

This graph visualizes an **hourly aggregated time series** of a "Count" metric over a day.

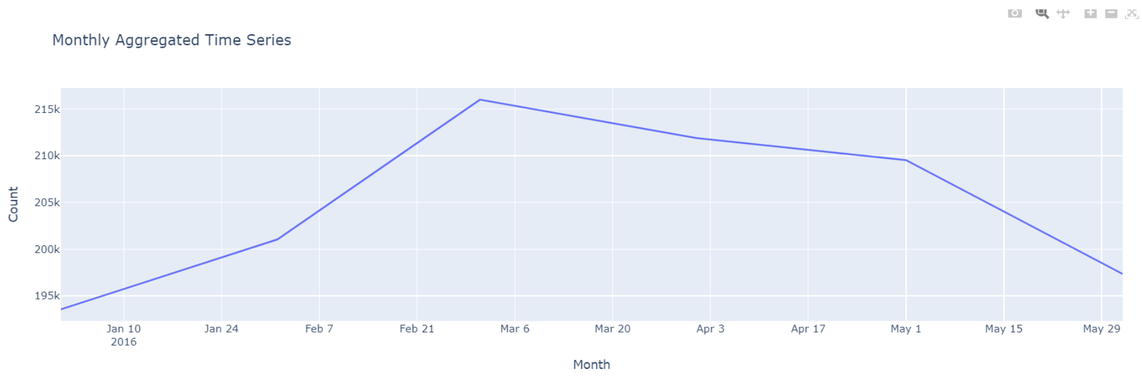


Key Observations:
- **Sharp Decline**: The count drops steeply from **70k** (at hour 5) to **0** (by hour 20).
- **X-Axis Labels**:
  - Hours of the day are marked at **5, 10, 15, 20**, likely representing specific measurement intervals (e.g., 5 AM, 10 AM, 3 PM, 8 PM).
- **Trend Significance**: The rapid decrease suggests a complete cessation or minimal activity in the tracked metric by late evening.

Summary:
A drastic downward trend in counts is observed over the course of the day, with values plummeting to zero by hour 20.

Monthly Aggregated Time Series (Jan-May 2016)

This graph visualizes a **monthly aggregated time series** of a "Count" metric over the first half of 2016.



Key Observations:
- **Downward Trend****: The count declines steadily from **215k** (Jan) to **195k** (late May).
- **X-Axis Details**:
  - Labels include specific dates (e.g., **May 29**), likely indicating intra-month data points or measurement intervals.
- **Trend Significance**: The consistent monthly decrease suggests a gradual reduction in the tracked metric.

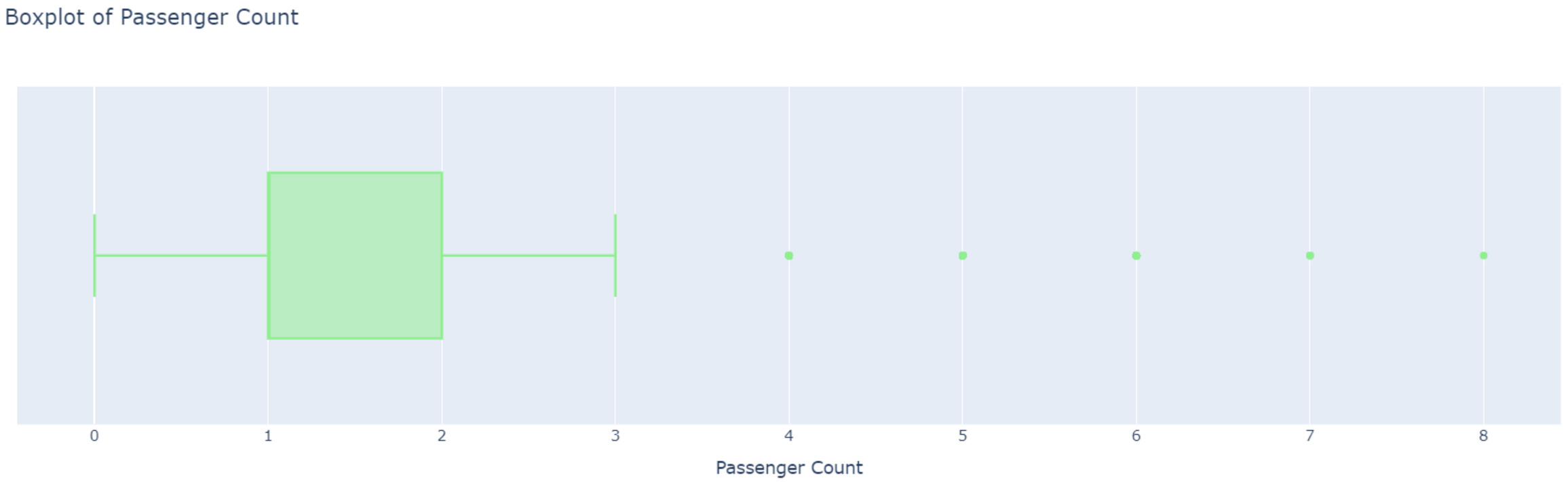
Summary:
A clear downward trend in aggregated counts is observed from January to May 2016.

Passenger Count

Boxplot of Passenger Count
This graph visualizes the distribution of passenger counts per taxi trip.

Key Observations
-Range: Passenger counts span from 0 to 8, indicating trips with no passengers up to larger groups.
- Central Tendency: The median likely falls between 1-2 passengers, reflecting common trip sizes.
- Spread: Tight clustering around lower counts suggests most trips have 1-3 passengers, with fewer outliers for larger groups.

Summary:
The majority of trips cater to small groups (1-3 passengers), with minimal demand for larger capacities.



Trip Duration (target variable)

The graph shows the **distribution of taxi trip durations (in seconds)**. The raw data is likely highly right-skewed, with most trips clustered at shorter durations and a long tail of infrequent, very long trips.

Why Log Transformation?

•Applied to reduce skewness and compress the scale, making patterns in the majority of the data (short/medium trips) more visible.

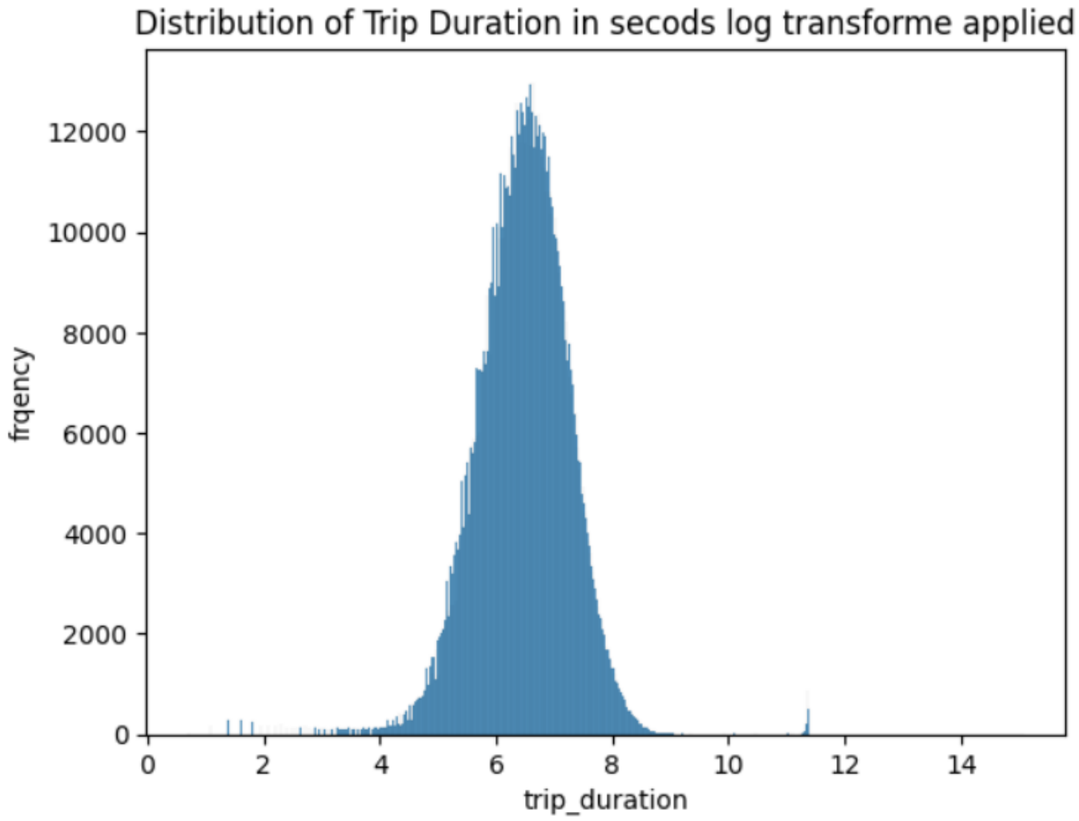
•Mitigates the impact of extreme outliers (e.g., 10-hour trips), allowing for clearer visualization of central tendencies (median, mean).

Key Insights:

•After log transformation, the distribution becomes closer to normal, simplifying statistical analysis.

•Highlights that most trips fall within a narrow log-scale range, corresponding to typical urban ride durations (e.g., 5–30 minutes).

Note: Always back-transform results (e.g., exponentiate) to interpret findings in original seconds/minutes.



The correlation matrix visualizes the relationships between different numerical features in the dataset.

Key Observations from the Correlation Matrix

-**vendor_id**: Shows a weak correlation with other variables, except for **passenger_count (0.29)**, suggesting that some vendors may serve larger groups more frequently.

- **passenger_count**: Has a weak correlation with all features, indicating that the number of passengers does not significantly influence trip duration or location-related variables.

- **pickup_longitude & dropoff_longitude (0.71 correlation)**: This strong correlation suggests that trips often follow specific longitude patterns, likely along major roads or routes.

- **pickup_latitude & dropoff_latitude (0.50 correlation)**: Exhibits a moderate correlation, meaning that pickup and drop-off locations tend to align but show more variation than longitude.

- **trip_duration`**: Has very weak correlations with other features, implying that trip duration is influenced by external factors not captured in this dataset (e.g., traffic conditions, time of day).

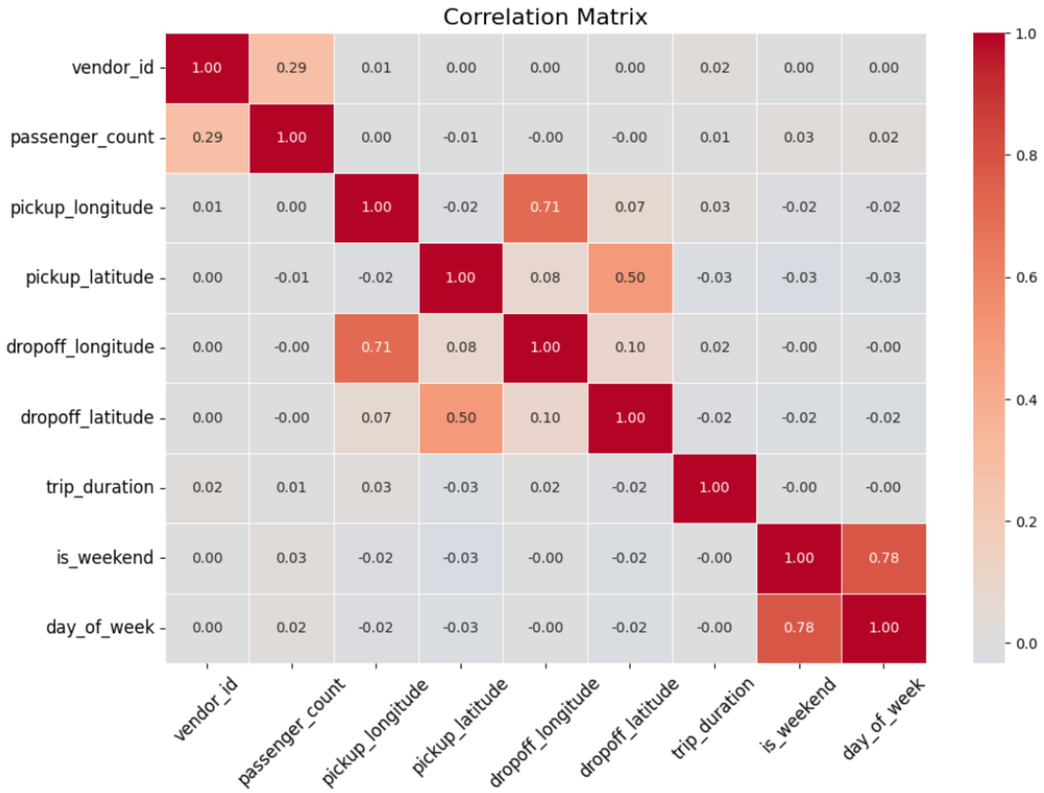
- **is_weekend & day_of_week (0.78 correlation)**: A high correlation is expected since **is_weekend** is derived from **day_of_week**, where certain values indicate weekends (Saturday & Sunday).

Insights & Implications

- The dataset does not show a **strong direct correlation** between **trip_duration** and geographical or categorical features, suggesting that additional factors (e.g., traffic, time of day, weather) may play a crucial role.

- The **high correlation between pickup and drop-off longitude/latitude** indicates that trips are often concentrated in specific areas, likely following a structured urban layout.

- Since **vendor_id** and **passenger_count** have low correlations with **trip_duration**, these features may have minimal predictive power for estimating trip time.



The correlation matrix visualizes the relationships between different numerical features in the dataset. Below is an explanation of the key columns based on their correlation values:

Key Observations from the Correlation Matrix

-**trip_duration_log_transform & distance (0.57 correlation)**: This moderate correlation suggests that longer trips generally cover more distance, though other factors may still influence trip duration.

- **trip_duration_minutes**: Has a weak correlation with most features, indicating that trip duration in minutes is not strongly dependent on individual attributes like time or day.

- **distance**: Shows a noticeable correlation (0.57) with **trip_duration_log_transform**, reinforcing the expected relationship between trip length and duration. However, it has minimal correlation with other time-based features like hour and month.

- **hour & trip_duration_log_transform (0.04 correlation)**: A near-zero correlation suggests that the hour of the trip does not significantly affect trip duration.

- **is_weekend & day_of_week (0.78 correlation)** : A high correlation is expected since **is_weekend** is derived from **day_of_week**, where certain values indicate weekends (Saturday & Sunday).

- **month**: Shows almost no correlation with trip duration, implying that seasonality does not have a major impact on trip time.

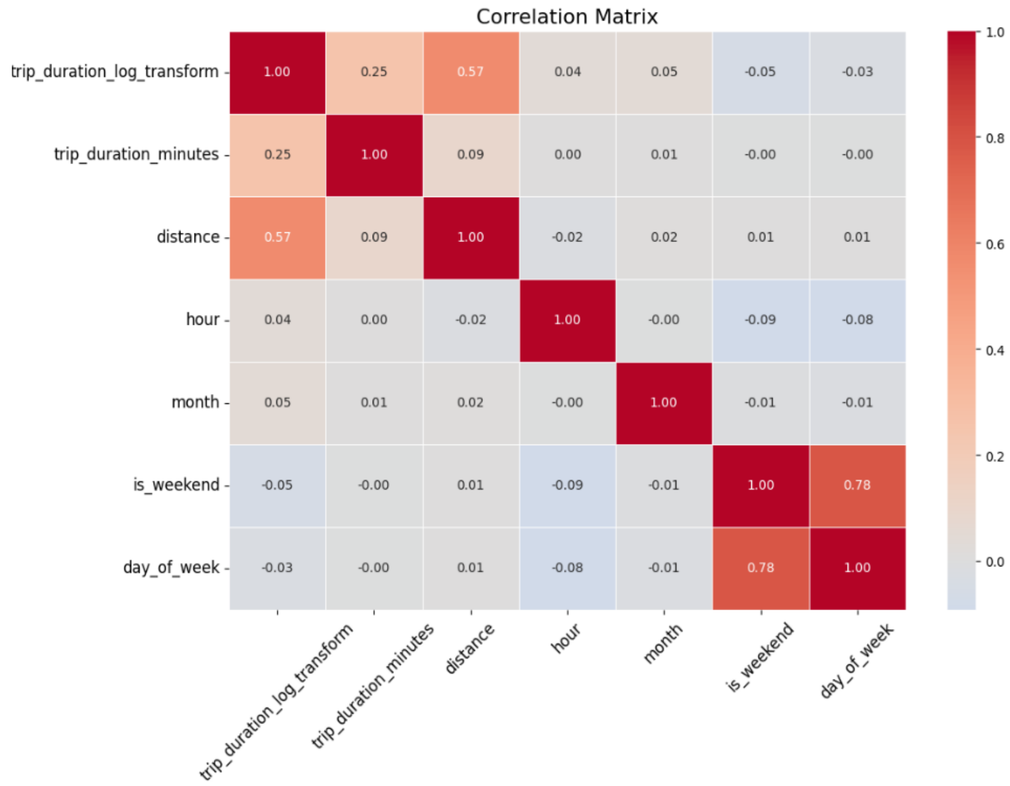
Insights & Implications

- **Trip duration is moderately correlated with distance**, but other factors, such as traffic and route selection, may still play a role in predicting trip time.

- **Time-based features (hour, month, is_weekend) show very weak correlations with trip duration****, suggesting that time alone is not a strong predictor.

- **The correlation between is_weekend and day_of_week is expected**, confirming the consistency of the derived feature.

- **Feature engineering efforts should focus on additional factors**, such as traffic patterns or weather conditions, which may improve predictive performance.



Feature Engineering & Preprocessing

Categorical Encoding

The time_of_day feature (derived from pickup_datetime hour bins) was one-hot encoded to capture temporal patterns. Rare categories in other fields were grouped into an other class to minimize sparse representations.

Nonlinear Transformations

The right-skewed trip_duration target variable was log-transformed to create trip_duration_log, stabilizing variance and improving regression assumptions. Geospatial features like distance (Haversine), mnhattan_short_path (Manhattan distance), and dirction (bearing angle) were computed to quantify spatial relationships between pickup/dropoff coordinates.

Temporal Feature Engineering

The pickup_datetime was decomposed into hour, weekday, month, and is_weekend flags. Interaction terms like hour_weekday_interaction and engineered features like is_peak_hour (based on rush-hour bins) were added to model time-dependent traffic dynamics.

Model Architecture

Pipeline Design

A ColumnTransformer integrated:

- Numeric features: distance, mnhattan_short_path, hour, and geospatial coordinates were scaled via StandardScaler, enriched with PolynomialFeatures(degree=3), and log-adjusted to handle nonlinear relationships.
- Categorical features: time_of_day was one-hot encoded to preserve temporal semantics.

Regularized Regression

A Ridge regression pipeline (α=1) with third-degree polynomial terms achieved optimal bias-variance tradeoff. Regularization suppressed overfitting from the 45+ engineered features while retaining spatial-temporal interactions.

Performance & Insights

Metrics

- Train R^2: 0.6887 | Validation R^2: 0.6871

The model explains ~69.5% of variance in trip_duration_log, demonstrating robust generalization despite feature complexity.

Key Drivers

- Geospatial metrics (distance, mnhattan_short_path) were top predictors, reflecting route geometry's impact.
- Temporal features (hour, is_peak_hour, weekday) captured recurring traffic patterns.
- Polynomial terms for hour_weekday_interaction revealed nonlinear time-of-week effects.

Future Enhancements

- Incorporate real-time traffic APIs or direction-based speed estimates to refine mnhattan_short_path assumptions.
- Test tree-based ensembles (e.g., `XGBoost`) to better model interactions between distance temporal features, and geospatial coordinates.