NYC Taxi Trip Duration Prediction Report

1. Introduction

This report presents a comprehensive exploratory data analysis (EDA) of the **New York City (NYC) Taxi Trip Duration** dataset. The primary objective is to identify **patterns, trends, and relationships** that influence taxi trip durations across the city. By analyzing a variety of features—such as **pickup and drop-off timestamps and locations**, **passenger counts**, and **vendor identifiers**—the analysis aims to extract actionable insights that contribute to a deeper understanding of **Taxi Trip Duration** in NYC.

In addition to the exploratory analysis, a **supervised machine learning model** was developed to predict taxi trip durations. This modeling phase involved extensive **data preprocessing and feature engineering**, including:

- Extraction of temporal features from datetime fields
- Calculation of geospatial distances between pickup and drop-off points
- Encoding of categorical variables
- Scaling and normalization of numerical features

These steps were essential to enhance model performance and ensure the accuracy and generalizability of the predictions.

The report outlines the following components:

- **Dataset Description**
- **Data preparation pipeline**
- **Modeling methodology**
- **Key findings** from the analysis and predictive modeling

Visualizations are incorporated throughout the report to communicate insights clearly and effectively, supporting informed and data-driven conclusions.

---

2. Dataset Description

The dataset used in this analysis contains detailed records of taxi trips conducted in **New York City**. It includes multiple fields that capture essential attributes of each trip, which

are critical for conducting meaningful exploratory analysis and building predictive models.
Below is a description of the key fields included in the dataset:

- **id**
  A unique identifier assigned to each trip record. This field serves primarily as a **primary key** and is typically excluded from direct analysis or modeling tasks.

- **vendor_id**
  A categorical code that indicates the specific taxi service provider associated with the trip. This feature is useful for analyzing differences in performance or behavior between vendors.

- **pickup_datetime**
  The exact date and time when the taxi meter was activated, marking the start of the trip. This temporal variable is vital for identifying **time-based patterns** such as peak hours or weekday vs. weekend trends.

- **passenger_count**
  An integer representing the number of passengers recorded for the trip. This can provide insights into trip dynamics related to **group vs. solo travel behavior**.

- **pickup_longitude** & **pickup_latitude**
  The geographic coordinates of the trip's origin. These features are essential for **geospatial analysis** and route-based feature engineering.

- **dropoff_longitude** & **dropoff_latitude**
  The geographic coordinates of the trip's destination. Together with pickup coordinates, they enable calculation of **distance-based features** such as Haversine or Manhattan distances.

- **store_and_fwd_flag**
  A binary flag indicating whether the trip data was **temporarily stored** in the vehicle's onboard memory before being transmitted to the server (usually due to a lack of network connectivity).
  - 'Y' indicates store-and-forward trips
  - 'N' indicates real-time transmission
- **trip_duration**
  The **target variable** for this project, representing the total duration of each trip in seconds. This is the primary outcome the machine learning model aims to predict.

The dataset comprises both training and validation records, resulting in a total of **over 1.2 million trip entries**. This extensive collection provides a robust foundation for both exploratory data analysis and model training, ensuring statistical reliability and generalizability of findings.

---

3. Data Preparation and Feature Engineering

To ensure data quality and enhance the predictive power of the machine learning model, a comprehensive data preprocessing pipeline was implemented. This process involved the transformation, enrichment, and cleaning of raw data into a format suitable for effective analysis and modeling.

3.1 Time-Based Feature Engineering

Several temporal features were extracted from the pickup_datetime field to capture seasonal, daily, and hourly patterns that influence trip durations:

- **Hour, Month, Day of Week, Day of Year**: Extracted using pandas datetime functions to represent temporal trends.
- **Period of Day**: Categorized each trip into **morning**, **afternoon**, **evening**, or **night** based on the pickup hour.
- **Weekend Indicator**: Flag indicating whether the trip occurred on a **Saturday** or **Sunday**.
- **Peak Hour Indicator**: Binary feature identifying whether the pickup occurred during common NYC rush hours (7–9 AM and 5–7 PM).
- **Season**: Each trip was assigned a season (**Winter**, **Spring**, **Summer**, or **Autumn**) based on the month of occurrence.
- **Holiday Flag**: Utilized the holidays library to mark whether the trip occurred on a **public holiday** in the United States.

3.2 Geospatial Feature Engineering

Geographic coordinates for pickup and drop-off points were used to compute various distance and direction-related features:

- **Haversine Distance**: Calculated as the shortest path between two points on a sphere, providing a baseline trip distance in kilometers.

- **Geodesic Distance**: A more precise measure of distance between pickup and drop-off points using the geopy library.
- **Manhattan Approximation**: An estimated travel distance along a grid-based route, simulating travel through a city with a block structure like NYC.
- **Bearing**: Calculated the compass direction from pickup to drop-off location, capturing directional patterns in trip behavior.
- **Trip Area**: Estimated the rectangular area (in square kilometers) covered by each trip, approximated using latitude and longitude differences.

3.3 Target Transformation

The target variable, trip_duration, originally measured in seconds, exhibited a right-skewed distribution. To stabilize variance and improve model performance, a **logarithmic transformation** (log1p) was applied.

3.4 Outlier Detection and Removal

To mitigate the influence of anomalous data points on the model, an **Interquartile Range (IQR)** method was employed to detect outliers in key numerical columns. A configurable threshold (default = 1.5) was used to identify outlier rows, which were subsequently removed from the dataset to improve model robustness.

3.5 Final Dataset Structure

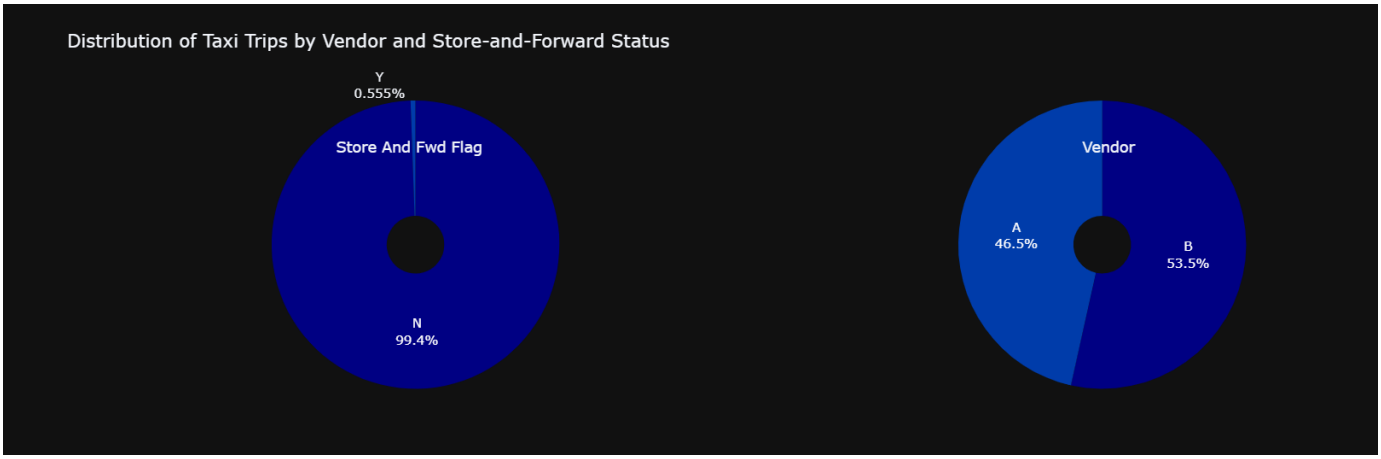
After preprocessing, the dataset was enriched with numerous engineered features capturing both temporal and spatial dimensions of taxi trips. Redundant or non-informative columns such as id and intermediate categorical labels were dropped to streamline the modeling process.

---

4. Exploratory Data Analysis and Visualizations

This section presents the **exploratory data analysis (EDA)** conducted on the preprocessed dataset. The goal of EDA is to gain a thorough understanding of the data's structure, distribution, and underlying relationships—particularly those that may influence the target variable, trip_duration.

4.1 Distribution of Taxi Trips by Vendor and Store-and-Forward Status



This plot examines the distribution of trips with respect to the taxi service provider (vendor_id) and the store-and-forward flag (store_and_fwd_flag). These attributes offer insights into system connectivity and operational characteristics across vendors.

Store-and-Forward Flag (store_and_fwd_flag)

- **99.4%** of trips were transmitted to the server in real time (N), indicating consistent and reliable network availability across the city.
- Only **0.56%** of trips were stored locally in the vehicle's memory before being forwarded to the server (Y), typically due to temporary connectivity issues.

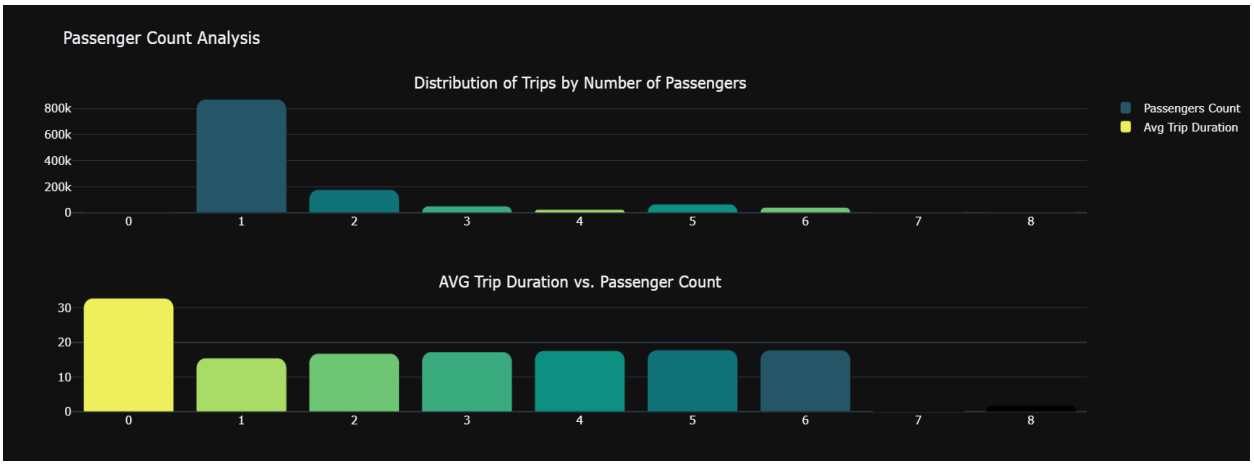
Insight:
The exceptionally high proportion of real-time data transmission suggests that NYC's taxi network operates with robust infrastructure and minimal technical disruptions.

Vendor Distribution (vendor_id)

- **Vendor 1** (Vendor A) accounted for approximately **46.5%** of all trips.
- **Vendor 2** (Vendor B) accounted for the remaining **53.5%**, indicating a slightly larger operational footprint.

Insight:
Vendor 2 appears to have a marginally higher share of trips, potentially reflecting a larger fleet, broader service coverage, or operational efficiencies. However, the distribution remains relatively balanced, allowing for fair comparative analysis between vendors.

4.2 Passenger Count Analysis



This plot explores how the number of passengers per trip influences both the frequency of trips and the average trip duration. Understanding passenger distribution is crucial for identifying demand patterns and potential capacity utilization.

Distribution of Trips by Number of Passengers

- The vast majority of trips had either **1** or **2 passengers**, with **single-passenger trips** being the most common.
- Anomalies such as trips with **0 passengers** were observed, likely due to data entry errors or system logging inconsistencies.
- Trips with **7 or more passengers** were extremely rare, reflecting the limited availability of large-capacity vehicles or the dominance of solo travel in NYC.

Insight:
Urban taxi usage in NYC is predominantly individual or small-group based. The presence of trips with zero passengers may warrant data cleaning or further investigation.

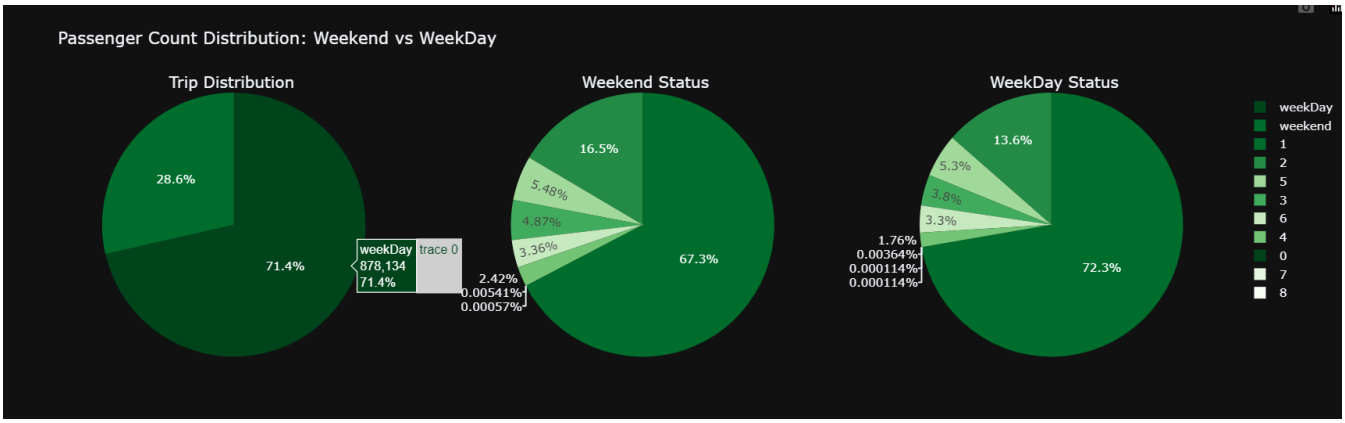
Average Trip Duration vs. Passenger Count

- **No strong correlation** was found between the number of passengers and the average trip duration..

Insight:
Passenger count does not significantly affect trip duration, indicating that time on the road is more influenced by factors like traffic, time of day, and route rather than occupancy level.

4.3 Passenger Count Distribution: Weekend vs. Weekday

This plot analyzes the variation in trip frequency and passenger count between weekends and weekdays. The goal is to uncover patterns in taxi usage based on the day of the week and trip occupancy.



1. Overall Trip Distribution

The first visualization compares the total number of taxi trips on **weekdays** versus **weekends**.

- **Weekdays** account for approximately **71.4%** of all trips.
- **Weekends** represent around **28.6%** of total trips.

Insight:
Taxi usage is significantly higher on weekdays, likely driven by work commutes, errands, and business-related travel.

---

2. Weekend Passenger Distribution

The second chart focuses on how passenger counts are distributed during **weekend** trips.

- **67.3%** of weekend trips involve **1 passenger**.
- **16.5%** of weekend trips involve **2 passengers**.
- Trips with **3 or more passengers** account for **less than 10%**.

Insight:
Despite increased opportunities for social outings, most weekend trips still involve a single passenger, indicating personal travel remains the dominant use case.

---

3. Weekday Passenger Distribution

The third visualization shows the passenger count distribution for **weekday** trips.

- **72.3%** of weekday trips involve **1 passenger**.
- **13.6%** involve **2 passengers**.
- Similar to weekends, **3+ passenger** trips are rare (under 10%).

Insight:
Weekday taxi usage is even more skewed toward solo travel, supporting the notion that taxis are mainly used for commuting or individual errands.

Summary of Findings

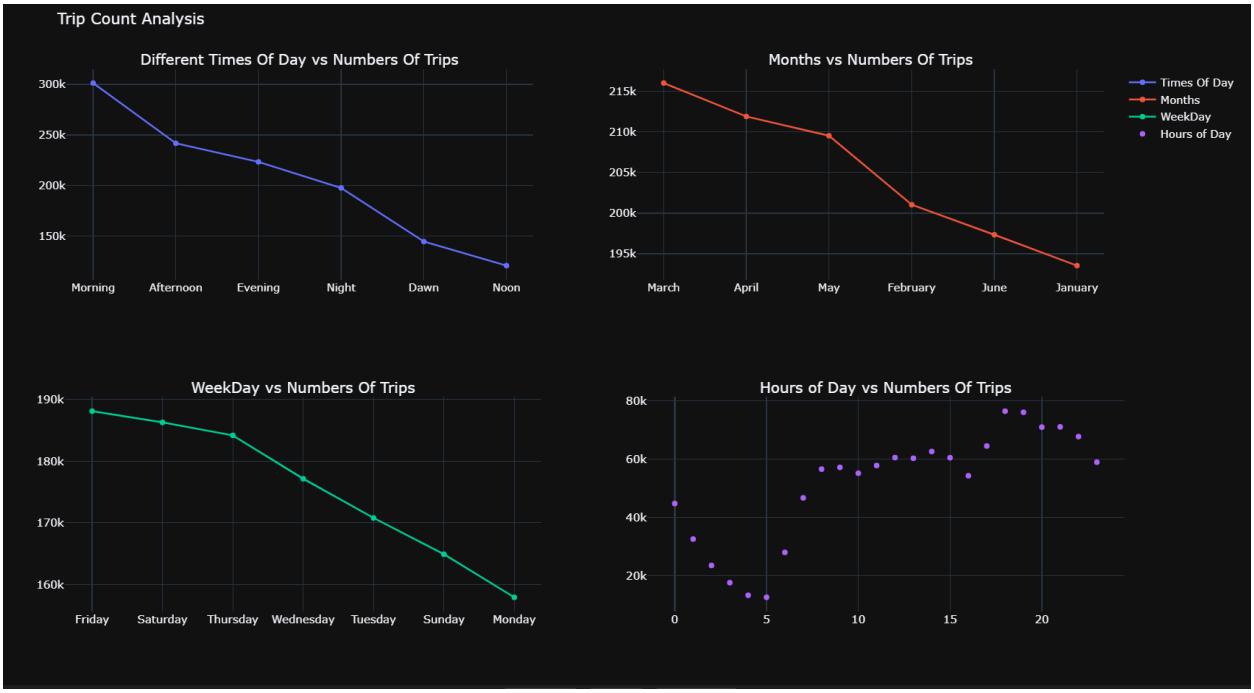
- **Trip Volume:** Taxi trips are more frequent on weekdays than weekends, highlighting workweek-related demand.
- **Occupancy Trends:** In both time frames, **single-passenger rides dominate**, with limited occurrence of high-occupancy trips.
- **Behavioral Consistency:** Despite varying purposes for travel, occupancy patterns remain consistent between weekdays and weekends.

Key Takeaways

1. **Workweek Drives Demand:** The majority of trips take place on weekdays, indicating peak operational load aligns with traditional work schedules.
2. **Single-Passenger Dominance:** The prevalence of solo trips across all days implies a user base favoring private, direct transportation.
3. **Low Multi-Passenger Incidence:** The infrequency of 3+ passenger rides suggests that taxis are not commonly used for shared or group commuting.

4.4 Trip Count Analysis

This plot explores taxi trip distribution across different temporal dimensions, including time of day, months, weekdays, and hours of the day. The objective is to identify peak demand periods and usage patterns.



1. Trip Count by Time of Day

This analysis groups trips into six time segments: **Dawn**, **Morning**, **Noon**, **Afternoon**, **Evening**, and **Night**.

- **Morning** sees the highest number of trips, approximately **300k**, likely due to commuting activity.
- Trip volume declines progressively:
    - **Afternoon** and **Evening** maintain moderate volumes.
    - **Night** and **Dawn** show significantly lower trip counts.
    - **Noon** records the **lowest** trip volume.

Insight:
Morning hours experience the greatest demand, driven by commuter traffic, while Noon and early Dawn have the lowest usage.

---

2. Trip Count by Month

Monthly analysis reveals seasonal trends in taxi usage.

- **March** has the highest number of trips (~215k), followed by **April** and **May**.

- **January** shows the **lowest** number of trips (~195k).
- Trip counts generally decline from **March to June**, with **February** and **April** falling in between.

Insight:
March may represent a seasonal high in activity, while January's lower usage could be tied to post-holiday season behavior or colder weather conditions.

---

3. Trip Count by Day of the Week

This breakdown evaluates demand fluctuations across weekdays.

- **Friday** records the highest number of trips (~190k), possibly due to weekend preparation and social events.
- **Monday** sees the lowest demand (~160k).
- Demand steadily decreases from **Friday to Monday**.

Insight:
Friday emerges as the most active day for taxi usage, while Monday reflects reduced mobility, potentially due to routine or limited social activity.

---

4. Trip Count by Hour of Day

A detailed hourly analysis illustrates how taxi usage fluctuates throughout the day.

Key Patterns:

- **Low Demand (12 AM – 5 AM):**
    - Minimal trip volume, with a trough at **3 AM** (~20k trips).
- **Morning Peak (6 AM – 9 AM):**
    - Demand surges starting at **6 AM**, peaking around **7–8 AM** (~60k trips).
- **Stable Midday (9 AM – 5 PM):**
    - Consistent trip volume (~50k–60k) during work hours.
- **Evening Peak (6 PM – 9 PM):**
    - Second demand surge, peaking at **7–8 PM** (~70k–80k trips).

- **Late-Night Decline (10 PM onward):**
    - Demand tapers off after 10 PM, returning to early morning lows.

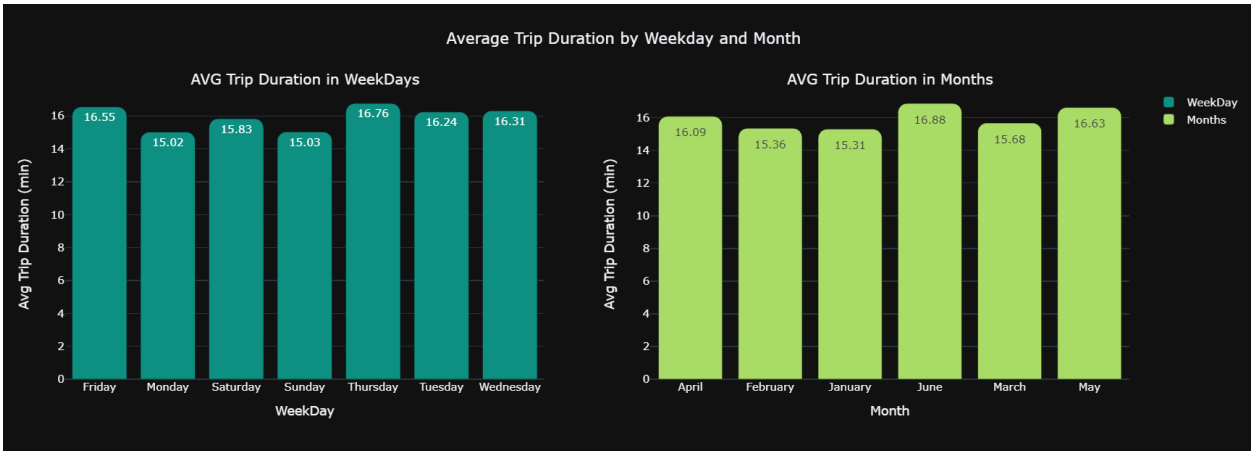
Insight:
The two main peaks align with typical **commuting hours** (7–8 AM and 7–8 PM), reinforcing the commuter-driven nature of taxi services.

Overall Insights

1. **Daily Cycles Matter:** Taxi demand is highly cyclical, with clear morning and evening peaks.
2. **Weekday vs. Weekend Behavior:** Fridays are busiest, while Mondays lag behind.
3. **Seasonal Influence:** Spring months show increased demand compared to winter months.
4. **Low Overnight Usage:** Few trips occur late at night or early morning, suggesting minimal overnight transportation needs.

4.5 Average Trip Duration by Weekday and Month

This plot explores how the **average taxi trip duration** (in minutes) varies across different **weekdays** and **months**, revealing trends influenced by commuting behavior, seasonal activities, and day-specific travel patterns.



1. Average Trip Duration by Weekday

This analysis examines the average trip duration across each day of the week.

Key Observations:

- **Friday** has the **highest average duration**, at approximately **16.76 minutes**.
- **Thursday** follows closely with **16.55 minutes**.
- **Sunday** and **Monday** record the **shortest durations**, at **15.03** and **15.02 minutes**, respectively.
- **Tuesday** and **Wednesday** lie in the middle range, averaging **15.5–15.8 minutes**.

Insight:

Longer trips on **Thursdays and Fridays** may reflect increased social outings, traffic congestion, or end-of-week errands. In contrast, **Sundays and Mondays** may involve shorter, routine trips such as errands or commutes.

---

2. Average Trip Duration by Month

This analysis investigates seasonal variations in average trip duration over the first six months of the year.

Key Observations:

- **June** has the **longest average duration** at **16.88 minutes**, followed by **May** at **16.63 minutes**.
- **January** and **February** show the **shortest durations**, with **15.31** and **15.36 minutes**, respectively.
- **March** and **April** fall between, with averages around **15.7–15.9 minutes**.

Insight:

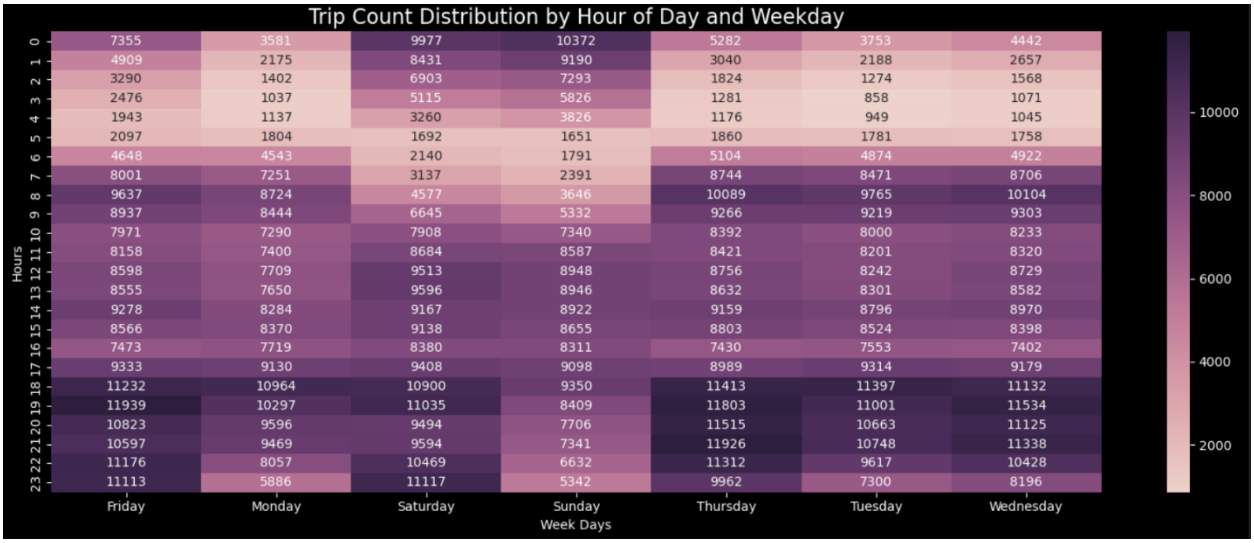
Trips in **May and June** are typically longer, likely due to warmer weather, increased tourism, or longer leisure-related travel. **Shorter durations in January and February** may be due to cold weather or a decrease in non-essential travel.

Overall Insights

1. **Weekend and End-of-Week Effect:** Trip durations increase toward the end of the week, suggesting more extensive travel activity.
2. **Seasonal Influence:** Late spring and early summer bring longer trips, potentially due to better weather and seasonal events.
3. **Efficiency in Winter:** Shorter trips in colder months may reflect efficiency or reduced non-essential travel.

4.6 Analysis of the Heatmap

This plot interprets the heatmap of taxi trip counts by **hour of the day** and **day of the week**, uncovering temporal patterns in ride demand.



1. Overall Trend – High vs. Low Activity

- **Peak activity** typically occurs during the **evening hours (5 PM – 9 PM)** across most weekdays.
- **Lowest activity** is seen in the **early morning hours (12 AM – 6 AM)**, aligning with reduced travel and resting hours.

---

2. Weekday-wise Patterns

- **Tuesday, Wednesday, and Thursday** consistently register the **highest trip volumes**, especially during **evening hours**, suggesting heightened midweek commuting or social activity.
- **Friday** shows a steady increase in trips starting from the **morning**, likely driven by flexible work hours and early weekend plans.
- **Saturday and Sunday** experience lower activity in the **morning**, followed by a rise in the **afternoon and evening**, indicative of leisure-based travel.

---

3. Hour-by-Hour Observations

- The **single highest activity peak** is observed around **6 PM on Thursday**, likely due to post-work commuting and early weekend outings.
- **Late-night hours (10 PM onward)** see a general decline in demand — **except on weekends**, where usage remains relatively strong due to nightlife and social events.
- **Morning hours (7 AM – 10 AM)** show moderate but consistent trip volumes, reflecting standard commute times.

---

4. Mondays Are the Quietest

- **Monday** records the **lowest overall activity**, both in the **morning and evening**, likely due to reduced discretionary travel following the weekend.

Key Insights & Recommendations

1. Driver Resource Optimization

- Deploy additional drivers during **evening peak hours (5 PM – 9 PM)**, particularly on **Tuesdays, Wednesdays, and Thursdays**, to meet higher demand efficiently.

2. Targeted Marketing

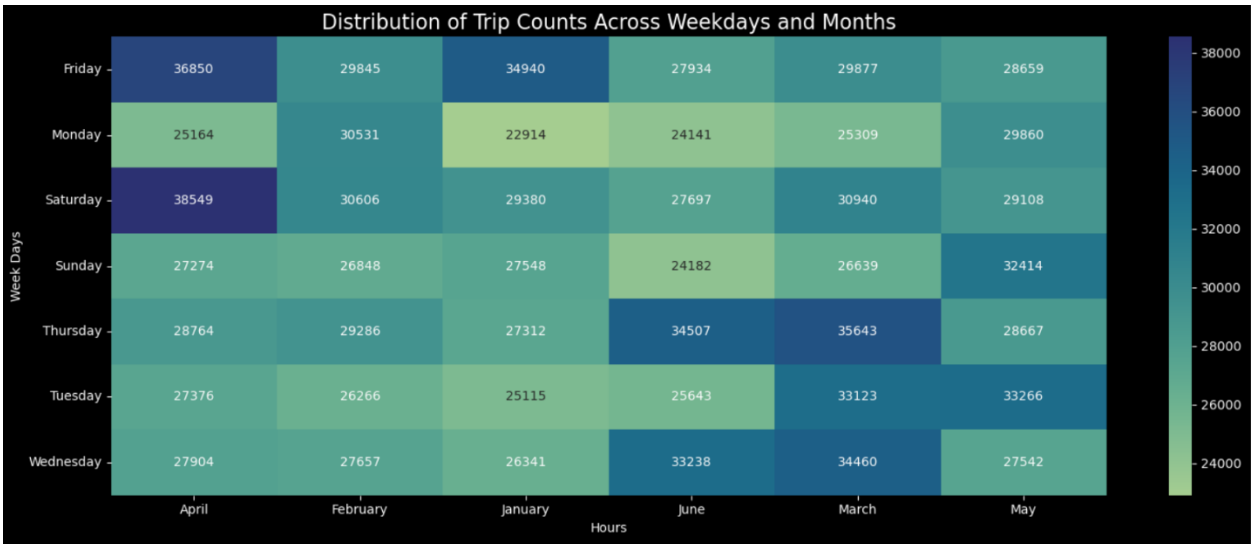
- Offer promotions or discounts during **low-demand windows**, such as **Monday mornings** or **weekday early mornings**, to encourage more rides.

3. Operational Efficiency

- Consider reducing fleet availability during **early weekday mornings (1 AM – 5 AM)** to cut idle costs.
- **Maintain or slightly increase** fleet presence during **late weekend nights**, as demand remains higher compared to weekdays.

4.7 Detailed Heatmap Analysis

This plot presents an in-depth examination of monthly and weekday trip trends based on heatmap data.



1. Monthly Trends

- **April** recorded the highest overall number of trips, particularly on:
  - **Fridays**: 36,850 trips
  - **Saturdays**: 38,549 trips
    This reflects a significant increase in travel demand during weekends in this month.
- **March** and **May** followed closely, showing consistently strong weekday activity—especially on **Tuesdays**, **Wednesdays**, and **Thursdays**.
- **January** had the lowest overall trip counts, notably on:
  - **Mondays**: 22,914 trips
  - **Tuesdays**: 25,115 trips
    The reduced volume may be attributed to colder weather or a general decline in mobility following the holiday season.

2. Weekday Patterns

- **Saturday** emerged as the busiest day of the week across all months, likely due to increased leisure and recreational travel.
- **Friday** also exhibited high trip counts, especially in April and January, suggesting elevated travel activity ahead of the weekend.
- **Monday** consistently recorded the lowest activity levels, particularly in January and April. This pattern aligns with typical behavior following the weekend.

---

3. Consistency and Spikes

- **Tuesday through Thursday** demonstrated stable demand across all months.
- Notable peaks were observed in **March** and **May**, potentially influenced by favorable weather or seasonal events:
    - **Wednesday (March)**: 34,460 trips
    - **Thursday (March)**: 35,643 trips
      These midweek increases indicate periods of heightened commuter or general travel activity.

Insights and Recommendations

1. Resource Allocation for High-Demand Periods

Increase operational capacity and driver availability during **March, April, and May**, particularly on **weekends** and **midweek days**, to accommodate higher demand.

2. Monday Demand Strategy

Implement promotional campaigns or service incentives on **Mondays** to improve engagement during historically low-demand periods.

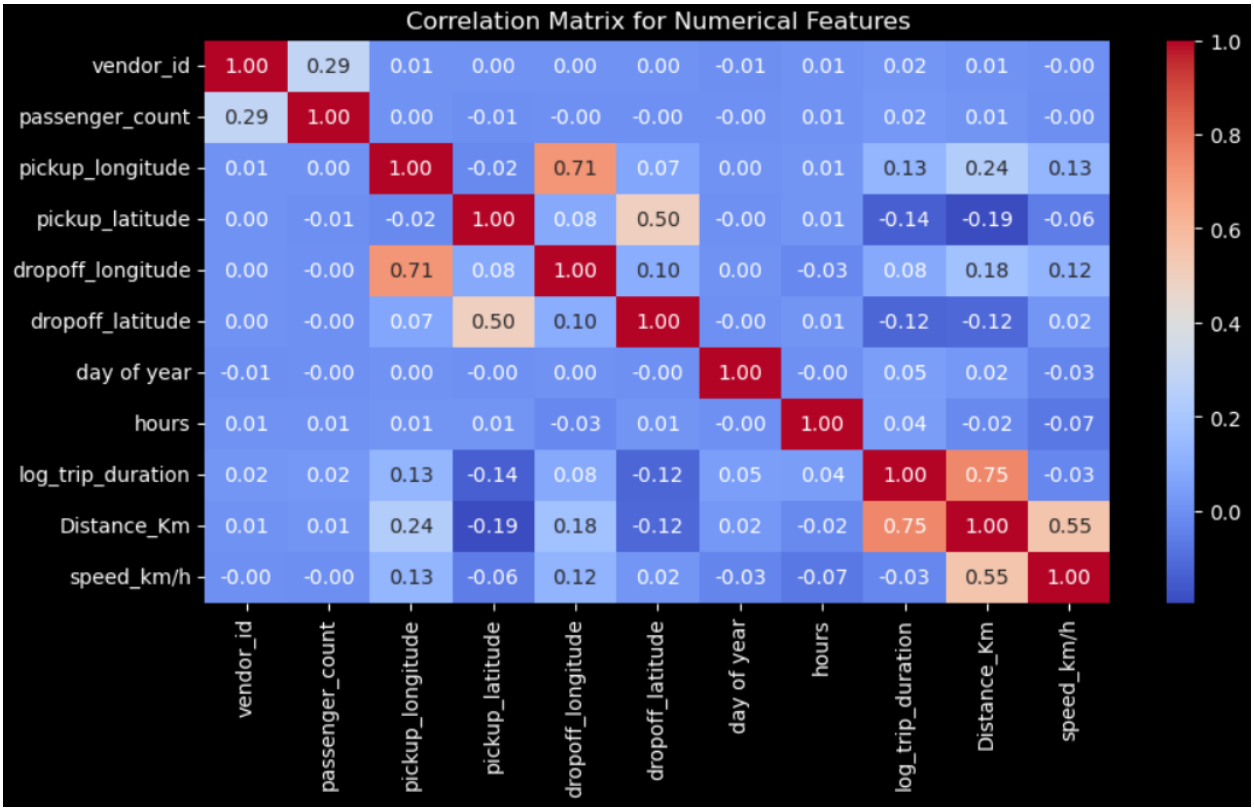
3. Weekend Optimization

Ensure optimal fleet availability and staffing on **Saturdays**, which consistently represent the highest trip volumes across all months.

4. Incorporating Seasonality into Forecasting

Integrate **monthly and weekday trends** into demand forecasting models to improve the precision of resource planning and operational efficiency, especially in preparation for **April surges**.

4.8 Correlation Analysis Summary



1. Strong Positive Correlations

These feature pairs demonstrate significant linear relationships:

- **Distance (km) vs. Log Trip Duration**: **0.75**

  A strong positive correlation, indicating that longer distances are associated with longer trip durations.

- **Pickup Longitude vs. Dropoff Longitude**: **0.71**

Suggests consistent east–west travel patterns across trips.

- **Pickup Latitude vs. Dropoff Latitude**: **0.50**

  Reflects a moderate alignment along the north–south axis, implying geographic clustering or structured city routes.

- **Distance (km) vs. Speed (km/h)**: **0.55**

  Indicates that longer trips tend to occur at higher average speeds, possibly due to more highway use or fewer interruptions.

---

2. Very Low or Negligible Correlations

These features exhibit little to no linear relationship with the target or other variables:

- **Log Trip Duration vs. Speed (km/h)**: **-0.03**

  Despite expectations, there is virtually no linear correlation, suggesting that factors like traffic, stop frequency, or trip type significantly affect duration beyond just speed.

- **Passenger Count**

  Shows negligible correlation with both log_trip_duration and Distance_Km, indicating little impact on trip duration in a linear context.

- **Vendor ID, Day of Year, Hour**

  All exhibit correlations near zero with most variables, suggesting limited utility in linear modeling, although they may still hold value in time-series, categorical, or segmented analyses.

Conclusions

- **Distance** is the most reliable linear predictor of trip duration.
- **Geographic coordinates** (longitude, latitude) show moderate alignment and may assist in spatial clustering or route prediction.

- **Speed** does not meaningfully correlate with trip duration in a linear sense, but could still play a role in **non-linear models** or when combined with other features.
- **Temporal and categorical features** like time of day, day of year, and vendor ID do not contribute strongly in a linear correlation framework but may still enhance performance in **segmented or tree-based models**.

---

5. Modeling

To predict taxi trip durations, a regularized linear regression model—**Ridge Regression**—was implemented using a custom pipeline that incorporates preprocessing, feature transformation, and model training. The approach emphasizes modularity, scalability, and robustness in handling both numerical and categorical data.

5.1 Pipeline Design

The modeling pipeline consists of the following components:

- **Numerical Features**
  - Scaled using `StandardScaler` to normalize feature distributions.
  - Transformed via `PolynomialFeatures` (degree = 5) to capture non-linear relationships.
  - Log transformation applied post-expansion using `FunctionTransformer` to reduce skewness and stabilize variance.
- **Categorical Features**
  - Encoded using `OneHotEncoder` with `handle_unknown='ignore'` to prevent issues from unseen categories during inference.

These transformations are integrated using `ColumnTransformer`, and the final pipeline is constructed using `Pipeline` from `scikit-learn`, combining preprocessing with model fitting.

5.2 Model Selection and Configuration

- **Model**: Ridge Regression
- **Regularization**: L2
- **Hyperparameter**:
  - `alpha = 1.0` (default strength of regularization)

Ridge Regression was selected for its ability to reduce model complexity and prevent overfitting, especially in the presence of polynomial features.

5.3 Important Note on Outliers

The dataset contains considerable noise and outliers, which were **not removed** prior to modeling. While high-degree polynomial features (degree = 5) often increase the risk of overfitting, in this case, the presence of noise and outliers provided a balancing effect that helped prevent overfitting. This approach allowed the model to generalize well despite the complex feature expansion and noisy data.

5.4 Evaluation Metrics

Model performance was assessed using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of errors in prediction. Lower values indicate better performance.
- **R^2 Score (Coefficient of Determination)**: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

Performance Summary

Dataset	*RMSE*	*R^2 Score*
Training	0.433	70.34%
Validation	0.443	69.32%
Test	0.433	70.45%

Conclusion

This report presented a comprehensive analysis of NYC taxi trip data, starting from detailed exploratory data analysis (EDA) to predictive modeling. Key insights were drawn from temporal, spatial, and trip duration patterns. A Ridge Regression model with polynomial features was employed, achieving strong and consistent performance across training, validation, and test sets. The results demonstrate the model's effectiveness in capturing complex relationships without overfitting.

Thank you for taking the time to review this report. Your attention and feedback are greatly appreciated.