Report: Optimizing NYC Taxi Operations

**Submission by – Sukirti Gupta**

Include your visualizations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

# Data Preparation

* 1. Loading the dataset
     1. **Sample the data and combine the files**

Primary sampling: Extraction of 500,000 records from each monthly Parquet file

Secondary refinement: Optimization of sample size to achieve a final consolidated DataFrame of approximately 1.89 million observations

This sampling methodology ensures statistical significance while maintaining computational efficiency

# Data Cleaning

### Fixing Columns

To enhance data consistency and analytical reliability, a comprehensive column standardization process was executed:

* Systematic removal of extraneous whitespace
* Implementation of uniform naming conventions
* Standardization of capitalization patterns
* Application of consistent formatting protocols across all variables

**2.1.1. Fix the index**

A significant data quality issue was identified and addressed regarding airport fee information:

Problem Identification:

* Discovery of redundant columns: 'airport\_fee' and 'Airport\_fee'
* Inconsistency attributed to varying naming conventions across monthly data files

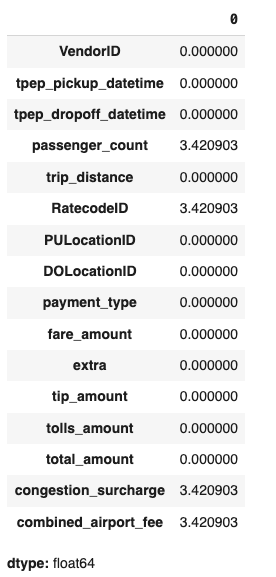
Resolution Strategy:

1. Creation of a new consolidated 'Airport\_fee' column
   * Implementation of a maximum value function between the two existing columns
   * This approach ensures preservation of all relevant fee information
2. Data Consolidation Process:
   * Systematic comparison of values between original columns
   * Selection of maximum value to prevent potential data loss
   * Validation of consolidated data accuracy
3. Column Optimization:
   * Removal of original redundant columns ('airport\_fee' and 'Airport\_fee')
   * Implementation of the new consolidated column as the singular source of airport fee data
   * Verification of data integrity post-consolidation

This methodological approach ensures data quality while maintaining the comprehensive nature of the dataset, facilitating more reliable subsequent analyses.

### Handling Missing Values

* + 1. **Find the proportion of missing values in each column**

****

* + 1. **Handling missing values in passenger\_count**

The treatment of missing values in the passenger\_count variable was executed using a mode-based imputation strategy:

Rationale:

* Passenger\_count represents a discrete numerical variable
* Mode imputation selected due to the variable's categorical nature
* Statistical validation confirms '1' as the predominant passenger count

Methodology:

* Identification and quantification of null entries
* Application of mode imputation to maintain data distribution integrity
* Validation of imputed values against existing data patterns
* Post-imputation distribution analysis to ensure consistency
  + 1. **Handle missing values in RatecodeID**

For the RatecodeID variable, a mode-based imputation strategy was implemented, considering its categorical nature:

Technical Justification:

* RatecodeID represents distinct service categories
* Mode imputation preserves the natural frequency distribution
* Maintains categorical variable integrity without introducing artificial categories

Implementation Process:

* Assessment of missing value patterns
* Mode calculation from non-null entries
* Systematic replacement of null values
* Verification of categorical distribution post-imputation
  + 1. **Impute NaN in congestion\_surcharge**

The treatment of missing values in the congestion\_surcharge variable employed a median-based imputation approach:

Statistical Considerations:

* Median selected over mean to minimize outlier influence
* Preserves central tendency while maintaining data robustness
* Ensures financial data integrity

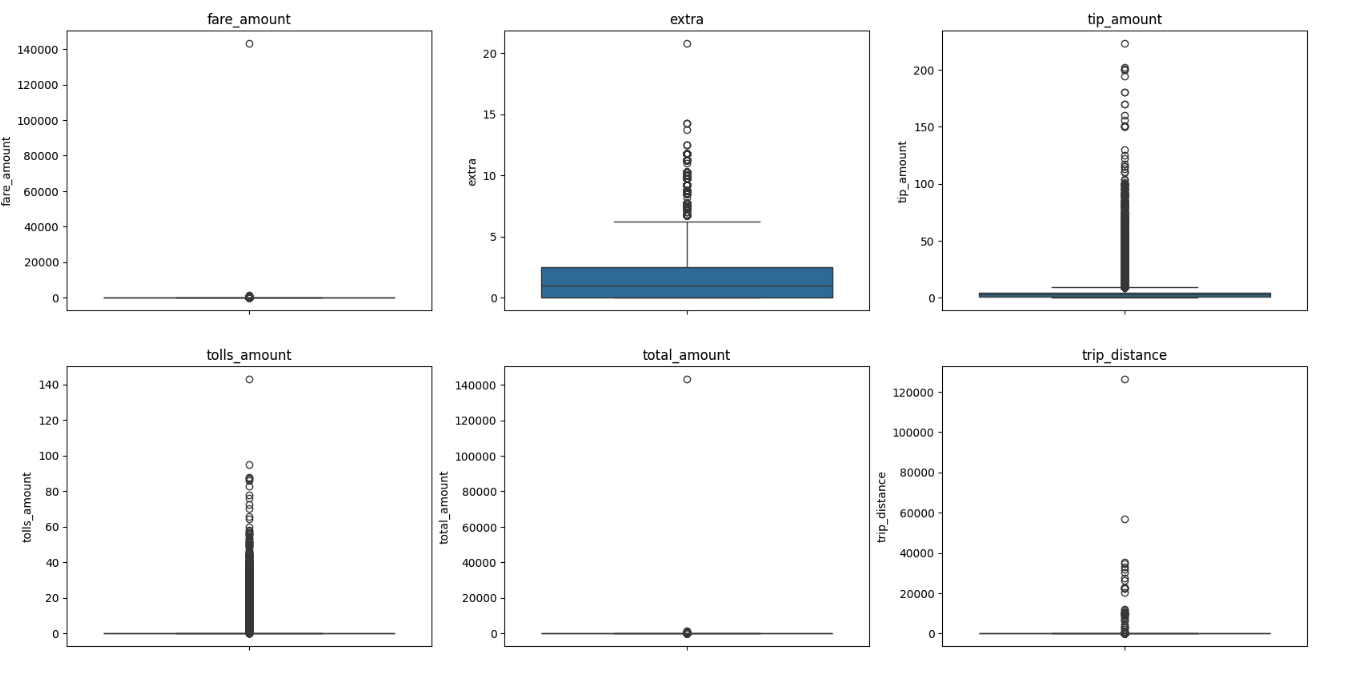
Methodological Framework:

1. Preliminary Analysis:
   * Assessment of missing value distribution
   * Evaluation of existing value patterns
   * Identification of potential outliers
2. Implementation:
   * Calculation of median from non-null values
   * Systematic application to missing entries
   * Validation of imputed values
3. Post-Implementation Verification:
   * Distribution analysis pre- and post-imputation
   * Statistical testing for distribution consistency
   * Confirmation of data integrity maintenance

This comprehensive approach to missing value treatment ensures data quality while maintaining statistical validity and analytical reliability across all variables.

### Handling Outliers and Standardizing Values

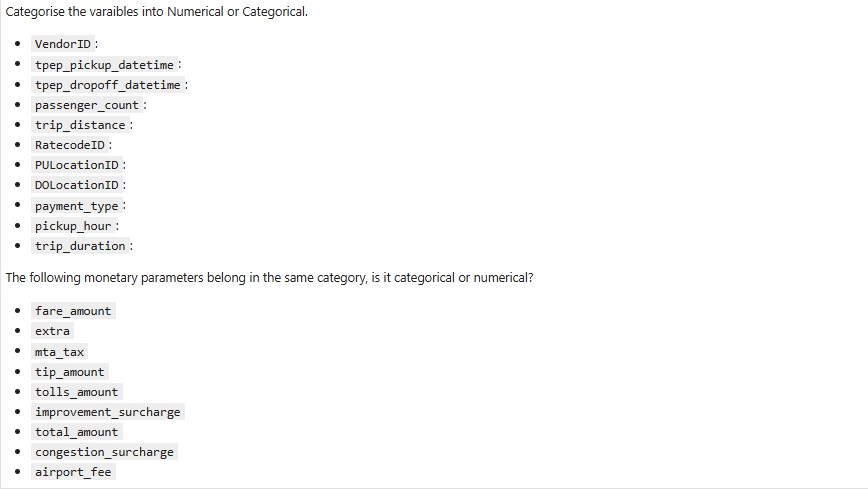
**2.3.1. Check outliers in payment type, trip distance and tip amount columns**

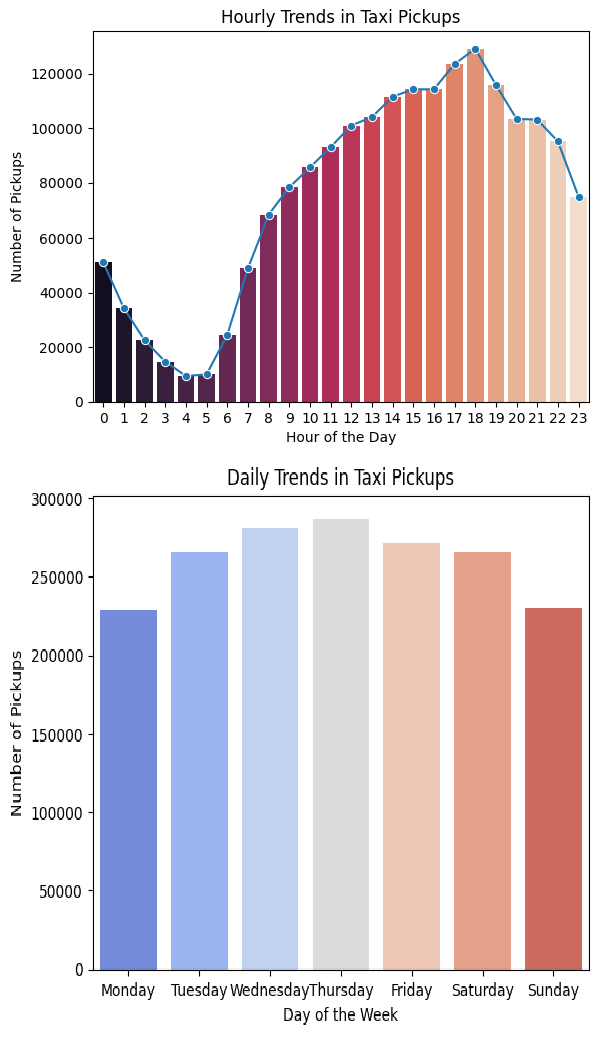
****

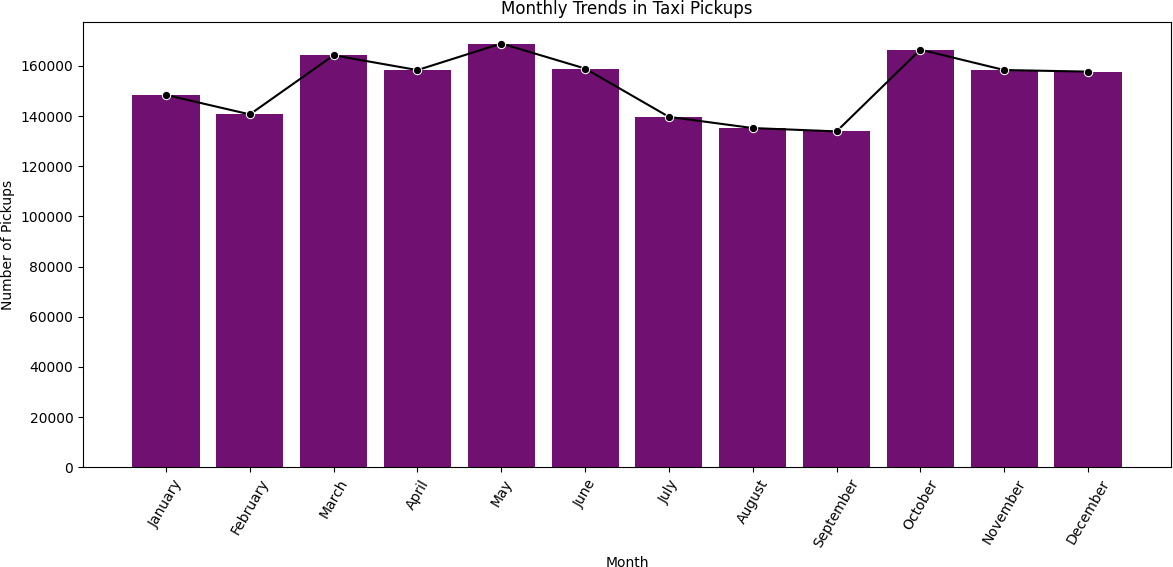
# Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**

****

* + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**



* + 1. **Filter out the zero/negative values in fares, distance and tips**

A systematic approach was implemented to enhance data quality through the careful elimination of anomalous records based on specific criteria:

1. Monetary Value Validation

1.1 Zero Fare and Total Amount Analysis:

* Records exhibiting fare\_amount = 0 or total\_amount = 0 were identified
* Rationale for Removal:
  + Such entries typically indicate transaction errors
  + May represent cancelled or invalidated trips
  + Compromise the integrity of financial analyses
* Implementation: Systematic removal of these records to maintain dataset validity

1. Distance-Location Consistency Verification

2.1 Zero Distance Anomaly Detection:

* Identification of records where:
  + trip\_distance = 0
  + AND pickup\_location ≠ dropoff\_location
* Justification for Exclusion:
  + Logical inconsistency between distance and location parameters
  + Indicates potential GPS or measurement errors
  + Compromises spatial analysis reliability

1. Tip Amount Consideration

3.1 Zero Tip Amount Treatment:

* Decision: Retention of records with tip\_amount = 0
* Analytical Justification:
  + Tipping represents a discretionary passenger behavior
  + Zero tips occur naturally in legitimate transactions
  + Verified through correlation with valid total\_amount values

3.2 Validation Methodology:

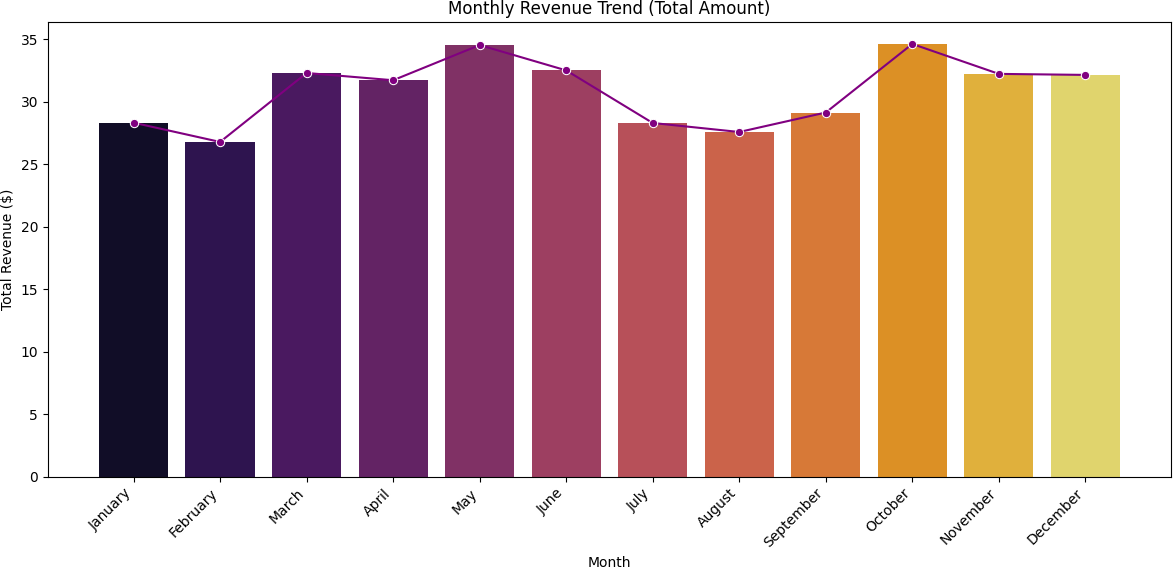
* Cross-reference analysis between tip\_amount and total\_amount
* Verification of transaction legitimacy through other parameters
* Confirmation of pattern consistency with expected customer behavior

This refined filtration protocol ensures:

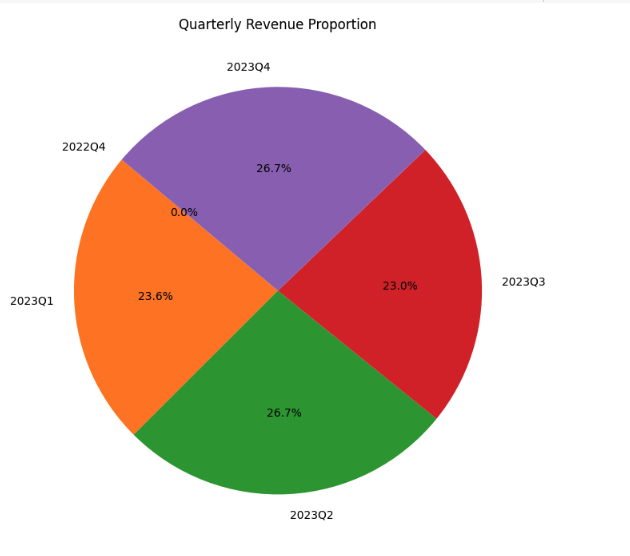
* Enhanced data integrity
* Improved analytical reliability
* Preservation of legitimate transaction patterns
* Minimization of artificial data distortion

The implementation of these criteria represents a balanced approach between maintaining data quality and preserving authentic transaction patterns, thereby establishing a robust foundation for subsequent analyses.

* + 1. **Analyse the monthly revenue trends**

****

* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**

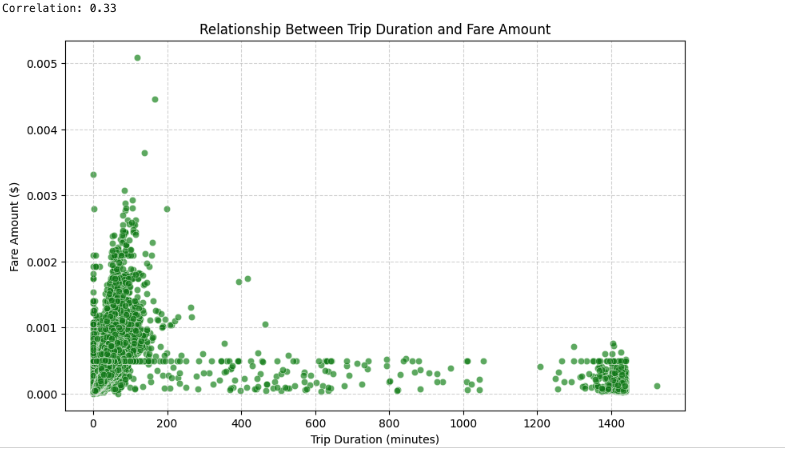
****

* + 1. **Analyze and visualize the relationship between distance and fare amount**

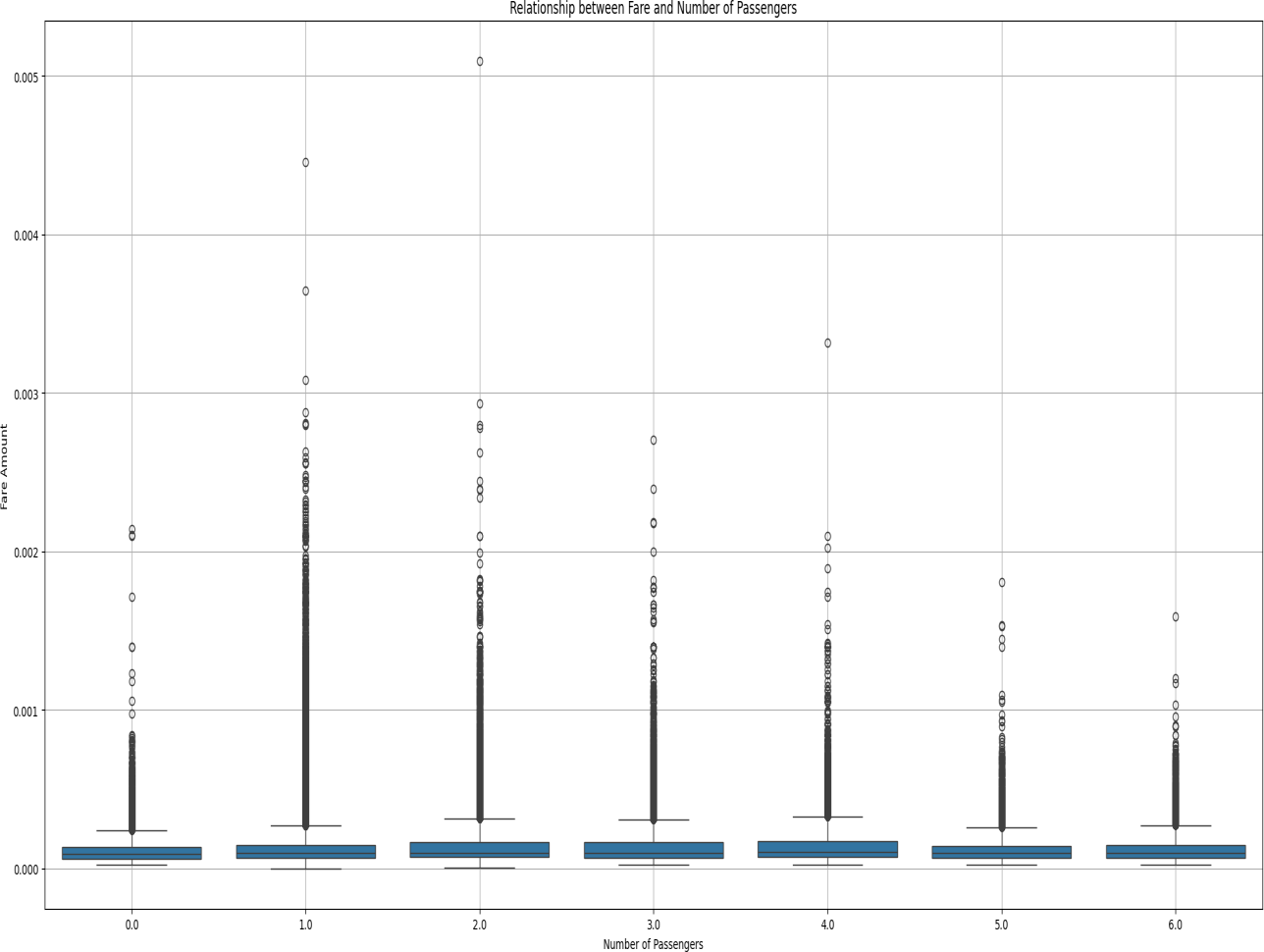
****

Correlation between Trip Distance and Fare Amount: 0.95

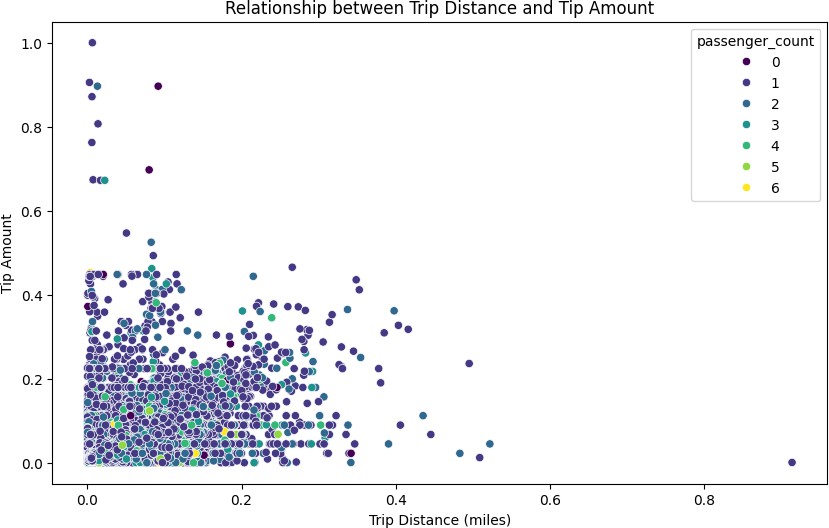
* + 1. **Analyse the relationship between fare/tips and trips/passengers**

****

Correlation between Trip Duration and Fare Amount: 0.33

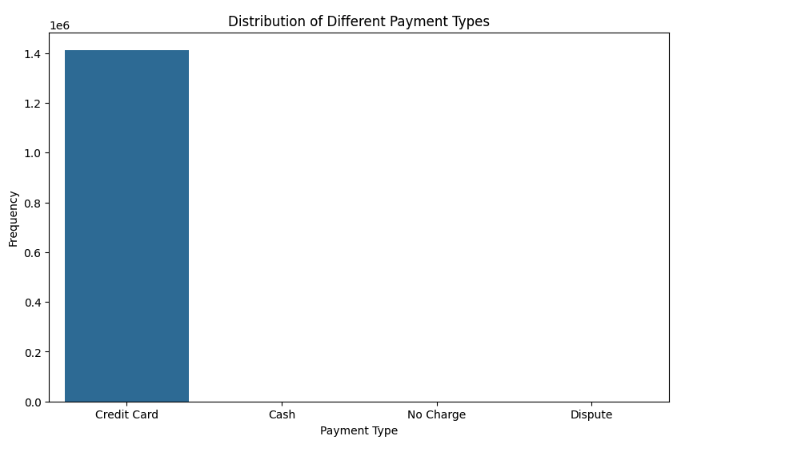


Correlation between Passenger Count and Fare Amount: 0.0403



Correlation between Trip Distance and Tip Amount: 0.80

* + 1. **Analyse the distribution of different payment types**



* + 1. **Load the taxi zones shapefile and display it**

<class 'geopandas.geodataframe.GeoDataFrame'> RangeIndex: 263 entries, 0 to 262

Data columns (total 7 columns):

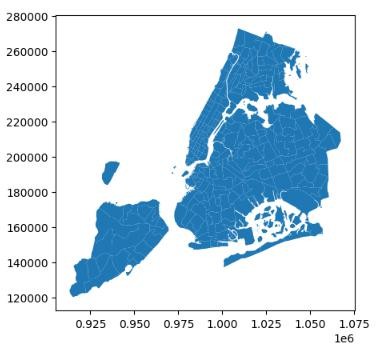
# Column Non-Null Count Dtype

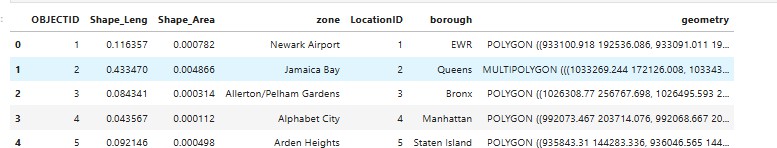
* + - 1. OBJECTID 263 non-null int32
      2. Shape\_Leng 263 non-null float64
      3. Shape\_Area 263 non-null float64
      4. zone 263 non-null object
      5. LocationID 263 non-null int32
      6. borough 263 non-null object
      7. geometry 263 non-null geometry

dtypes: float64(2), geometry(1), int32(2), object(2) memory usage: 12.5+ KB

None

<Axes: >



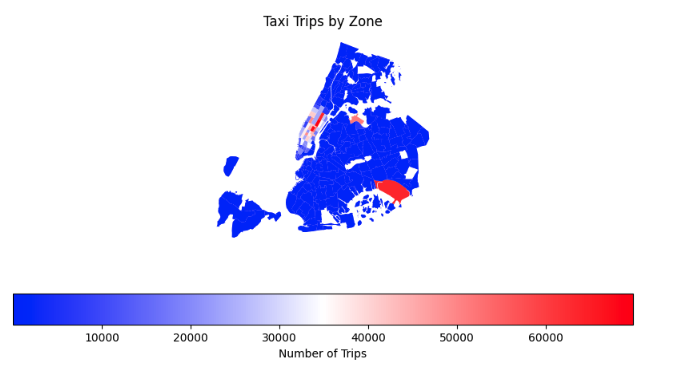


* + 1. **Merge the zone data with trips data**

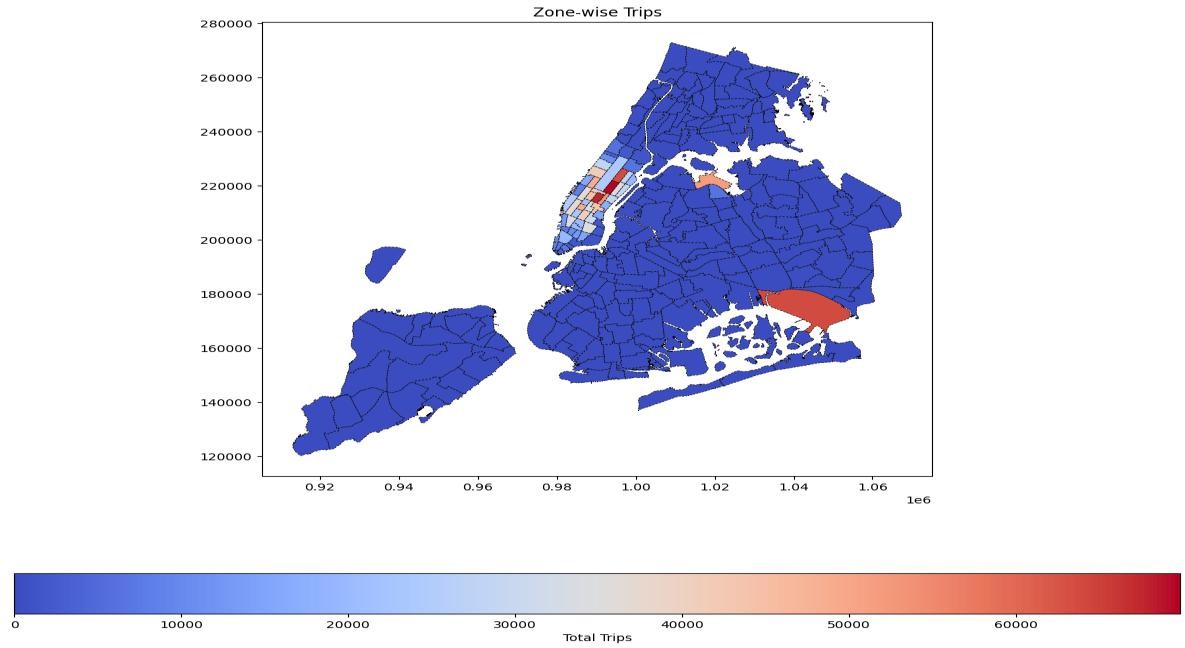
The zones dataset was merged into the trip dataset using the locationID from the zones data and the PULocationID from the trip data as the key columns.

* + 1. **Find the number of trips for each zone/location ID**

|  |  |
| --- | --- |
| **PULocationID** | **total\_Trips** |
| 1 | 35 |
| 2 | 2 |
| 4 | 1403 |
| 6 | 1 |
| 7 | 253 |

* + 1. **Add the number of trips for each zone to the zones dataframe  
         
       **

## Plot a map of the zones showing number of trips



**3.1.14 Conclude with results**

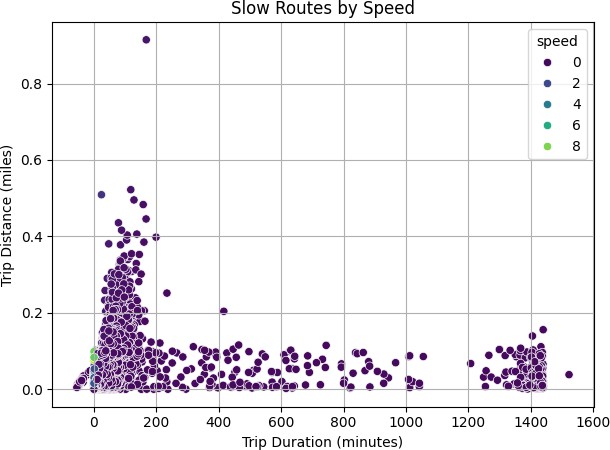
**Conclude with results** A **strong positive correlation** was observed between distance and fare, indicating that fares are predominantly **distance-driven**.

* **Weekday peak hours** align with rush hour traffic, while **weekends** exhibit a rise in **late- night activity**.
* **Airport** and **Midtown** zones show the **highest concentration of pickups and drop-offs**.
* The majority of trips involve **1–2 passengers**, with **credit cards** being the **most common payment method**.
* **Seasonal patterns** emerged, with **Q3 (July–September)** identified as the **busiest quarter**.
* Rigorous **data cleaning** was conducted to remove anomalies and **standardize numeric features**, thereby enhancing the accuracy and reliability of the analysis.

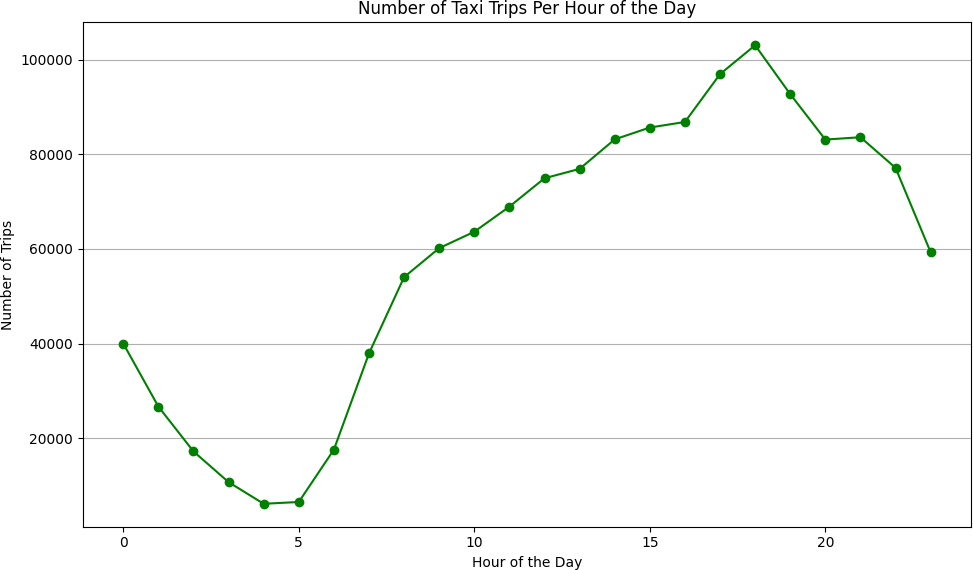
# Detailed EDA: Insights and Strategies

## Identify slow routes by comparing average speeds on different routes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PULocat ionID** | **DOLocat ionID** | **tpep\_pickup\_d**  **atetime** | **trip\_duration\_**  **derived** | **trip\_dis**  **tance** | **spee**  **d** |
| 1 | 1 | 2023-02-06  16:26:31 | 0.116667 | 0.000293 | 0.1506  26 |
| 1 | 1 | 2023-02-14  13:13:04 | 0.116667 | 0.000244 | 0.1255  21 |
| 1 | 1 | 2023-03-06  12:55:36 | 0.316667 | 0.000244 | 0.0462  45 |
| 1 | 1 | 2023-03-09  19:02:51 | 0.083333 | 0.000195 | 0.1405  84 |
| 1 | 1 | 2023-03-24  11:41:59 | 0.116667 | 0.000244 | 0.1255  21 |

****

* + 1. **Calculate the hourly number of trips and identify the busy hours**

****

The five busiest hours:

pickup\_hour

|  |  |
| --- | --- |
| 18 | 103059 |
| 17 | 96953 |
| 19 | 92730 |
| 16 | 86841 |
| 15 | 85666 |

Name: count, dtype: int64

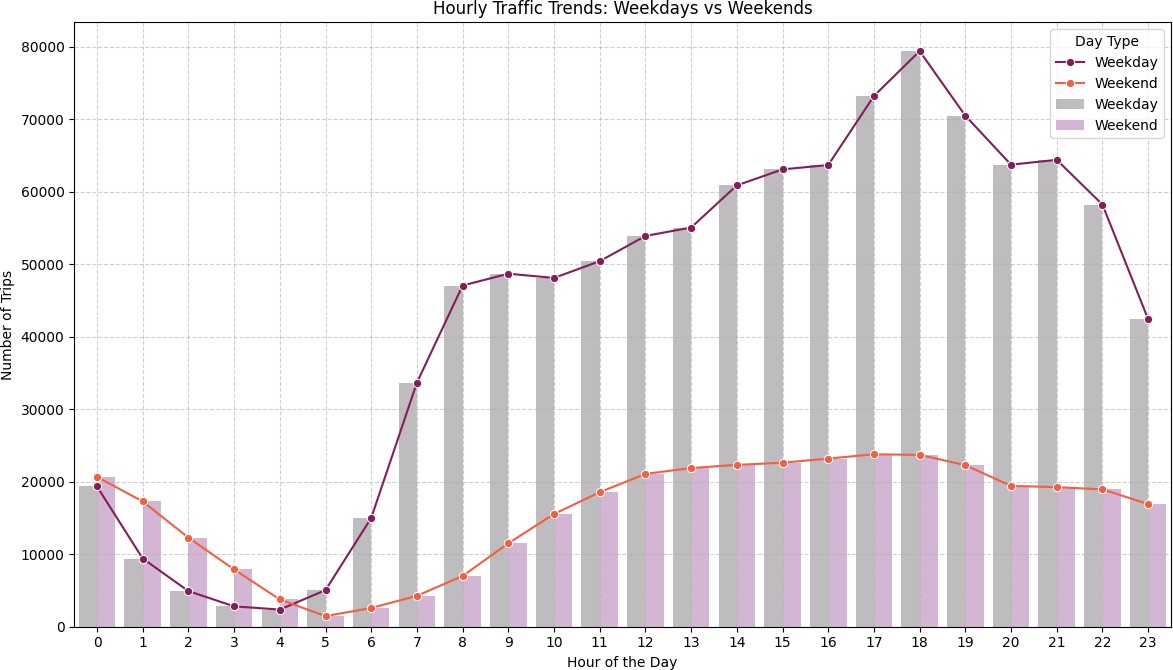
* + 1. **Scale up the number of trips from above to find the actual number of trips**

Actual number of trips in the five busiest hours (scaled from sample): pickup\_hour

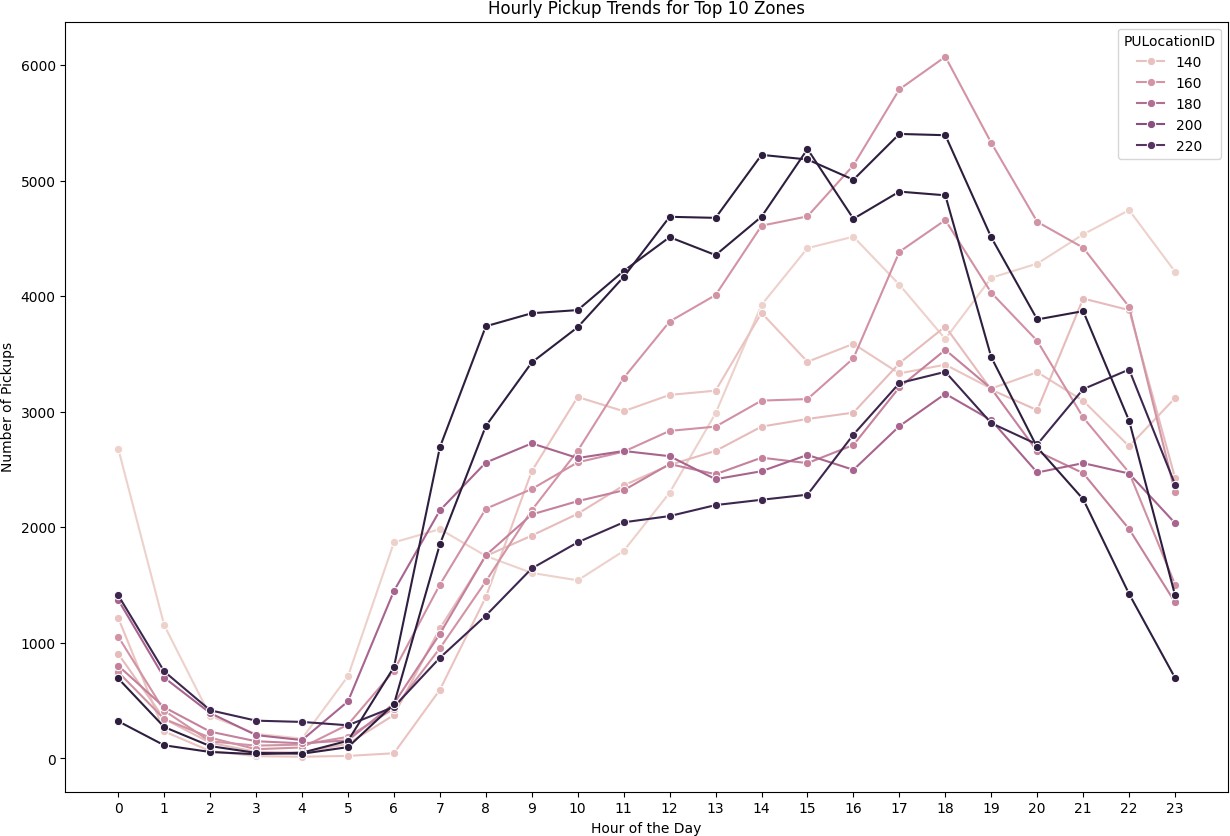
|  |  |
| --- | --- |
| 18 | 2061180 |
| 17 | 1939060 |
| 19 | 1854600 |
| 16 | 1736820 |
| 15 | 1713320 |

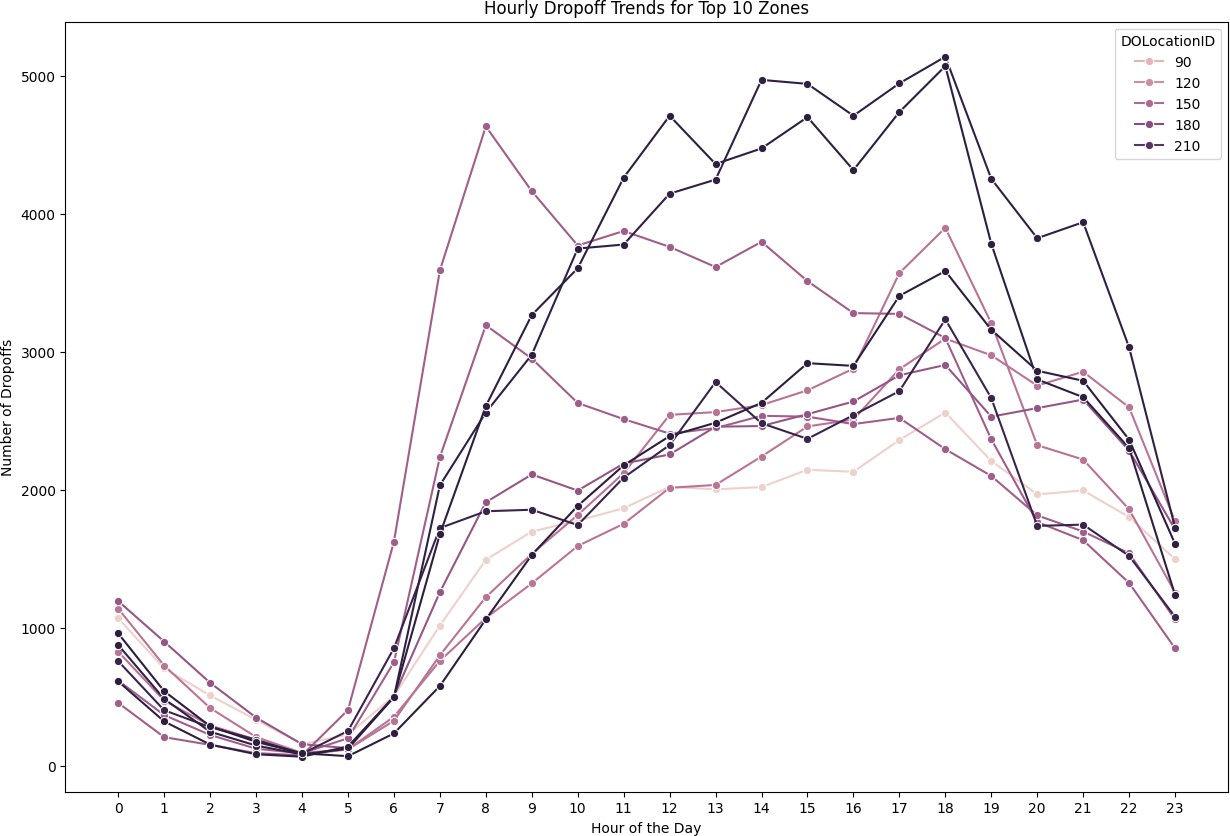
Name: count, dtype: int64

* + 1. **Compare hourly traffic on weekdays and weekends**

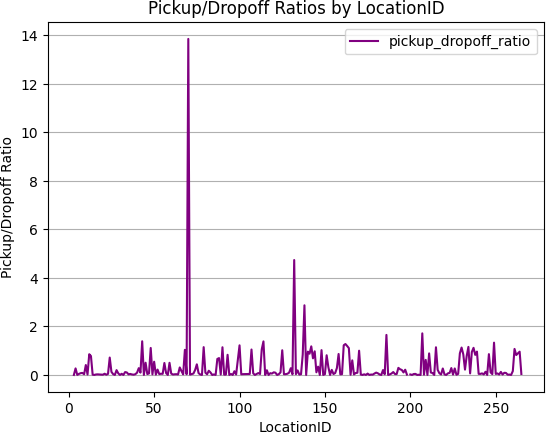
****

* + 1. **Identify the top 10 zones with high hourly pickups and drops**

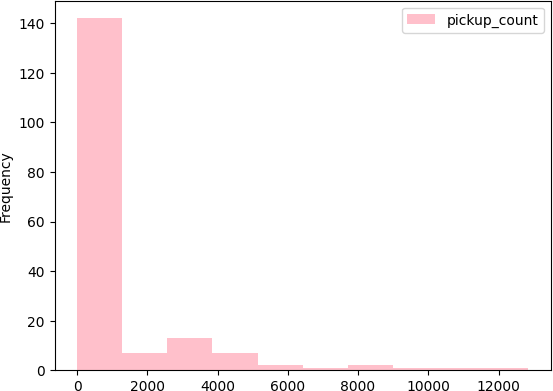
****

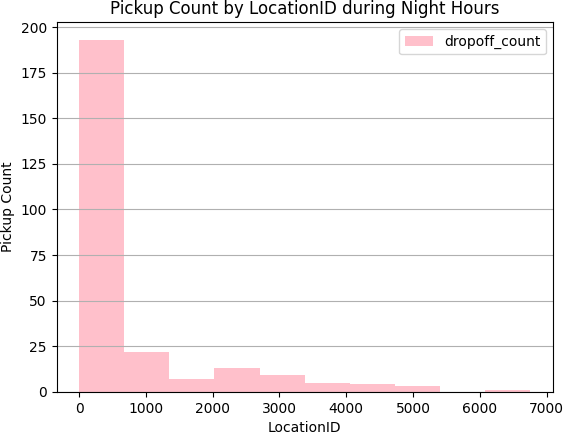


* + 1. **Find the ratio of pickups and dropoffs in each zone**

****

* + 1. **Identify the top zones with high traffic during night hours**

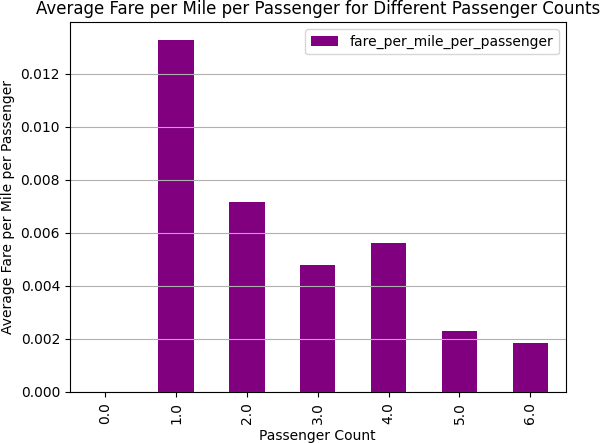
****

****

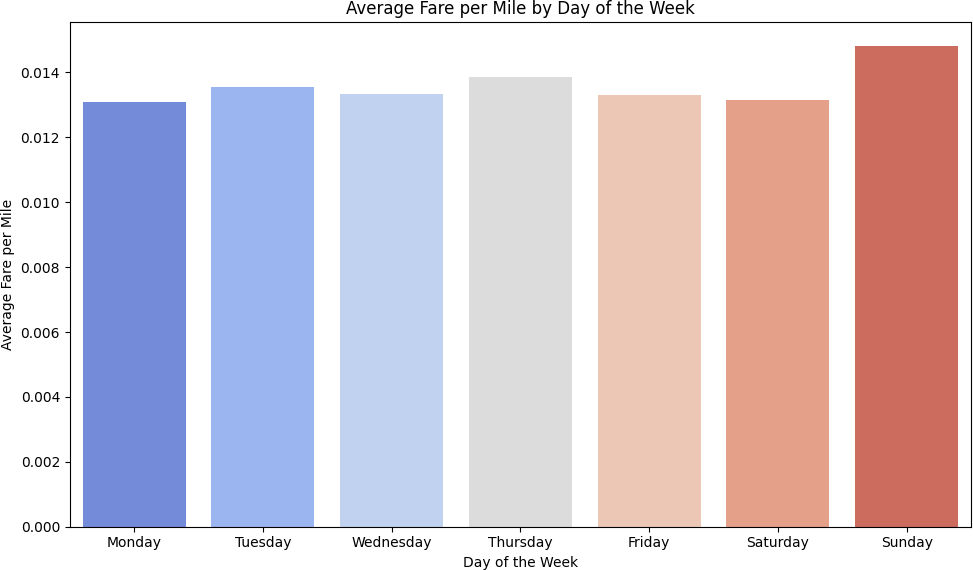
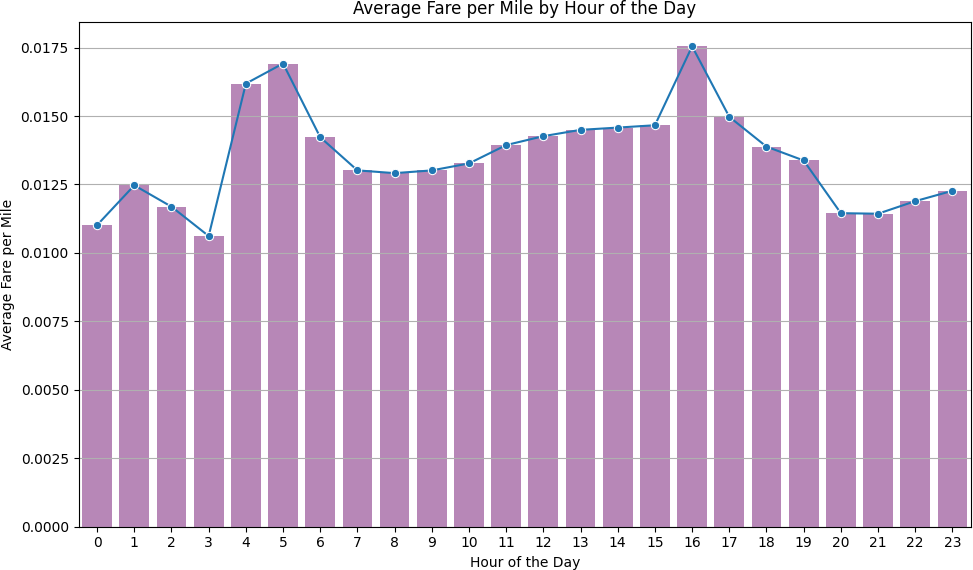
* + 1. **Find the revenue share for nighttime and daytime hours**



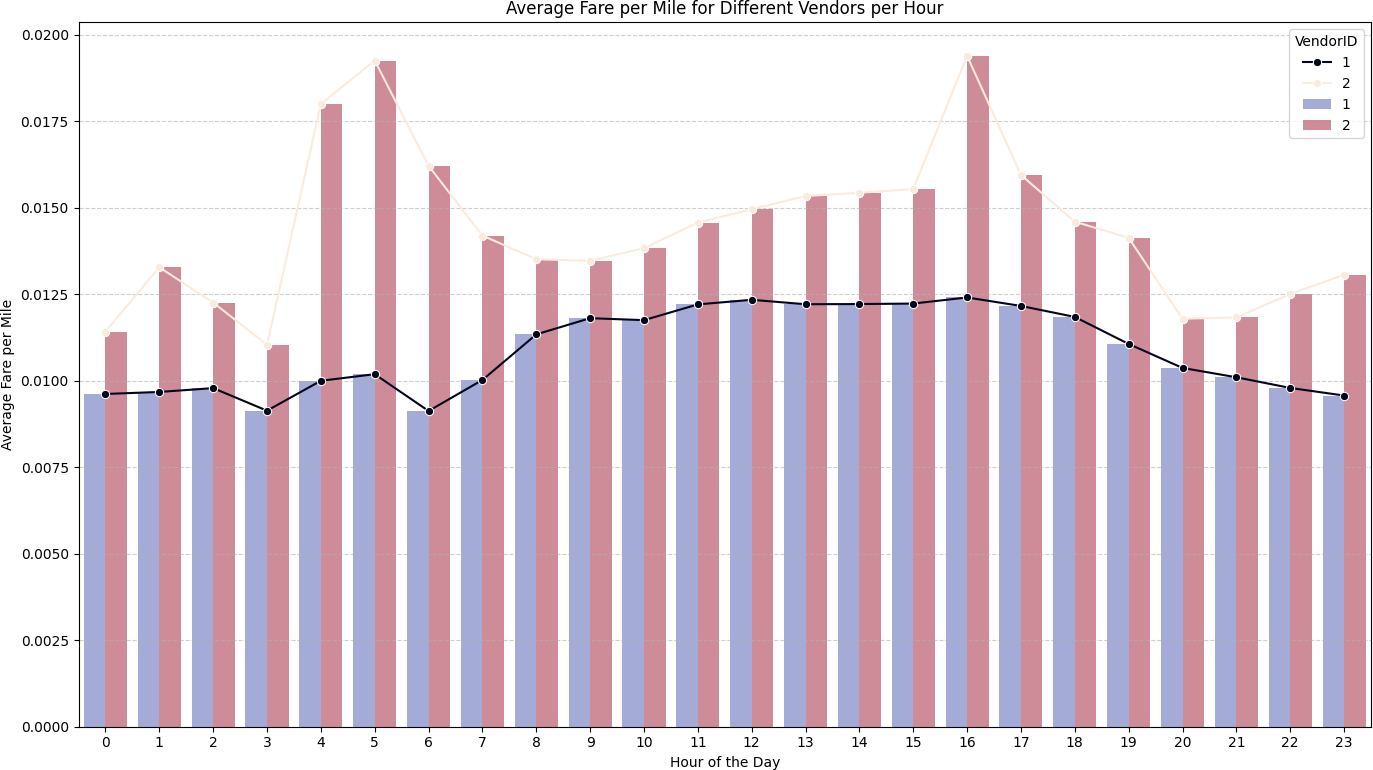
* + 1. **For the different passenger counts, find the average fare per mile per passenger**

****

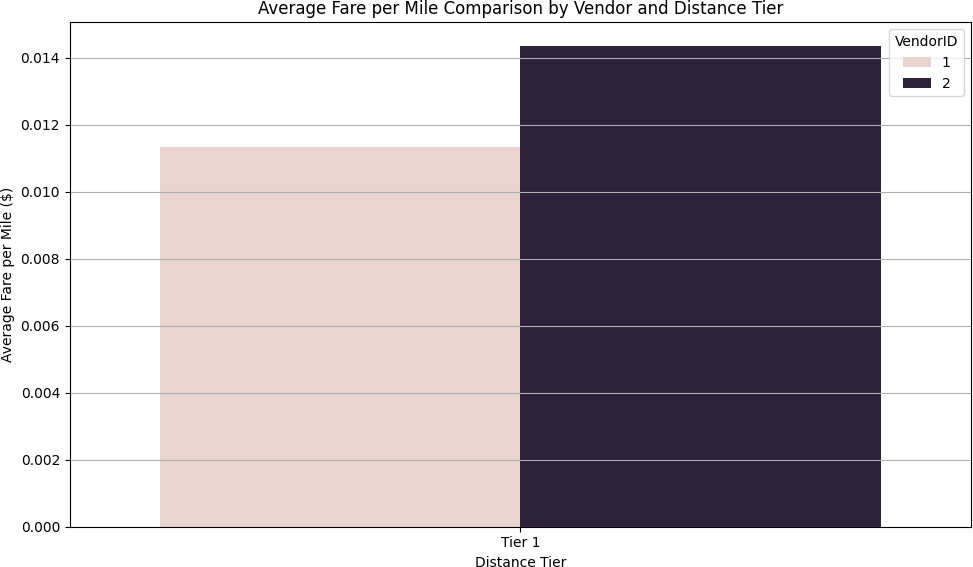
* + 1. **Find the average fare per mile by hours of the day and by days of the week**

****

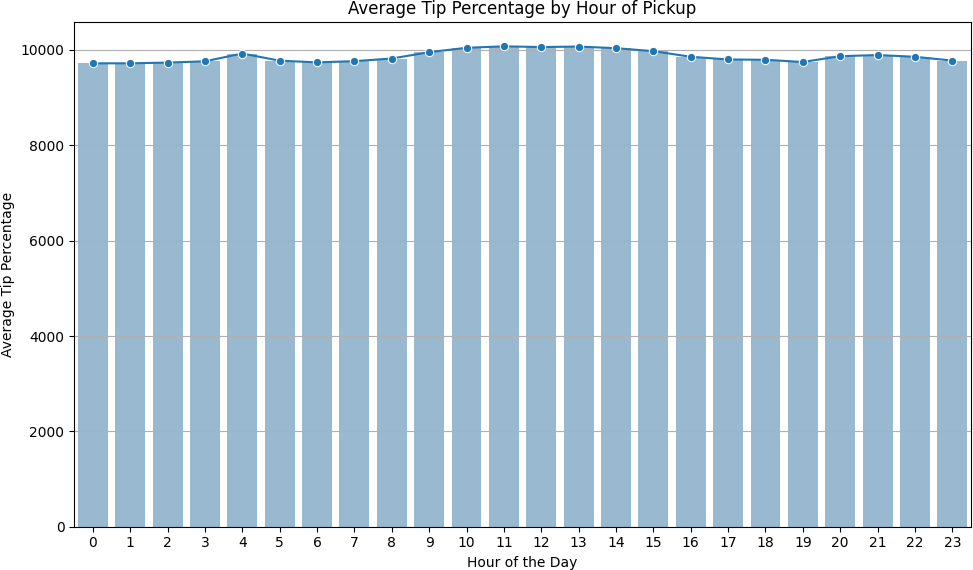
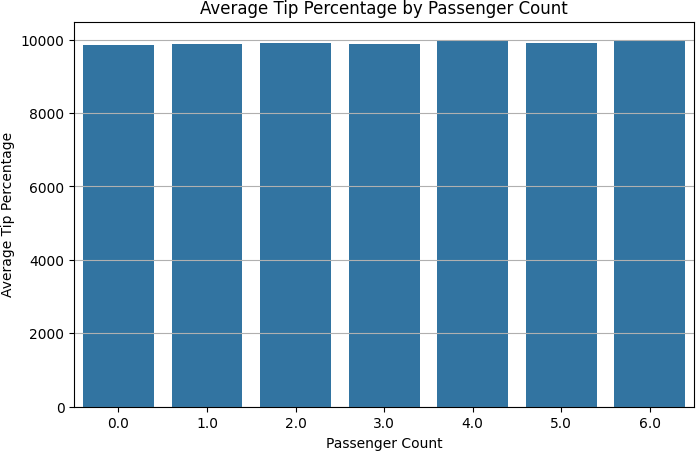
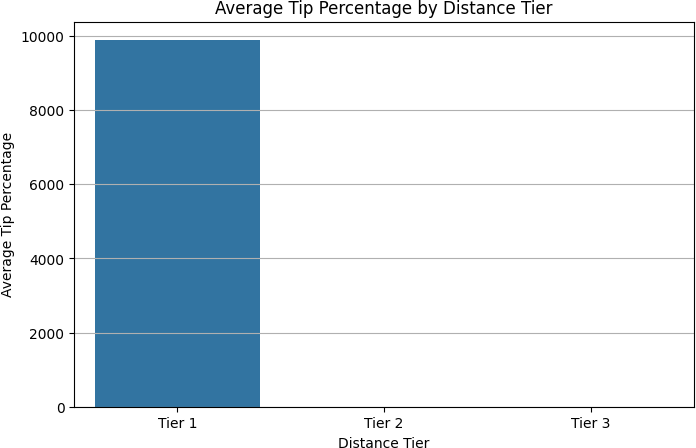
* + 1. **Analyse the average fare per mile for the different vendor**

****

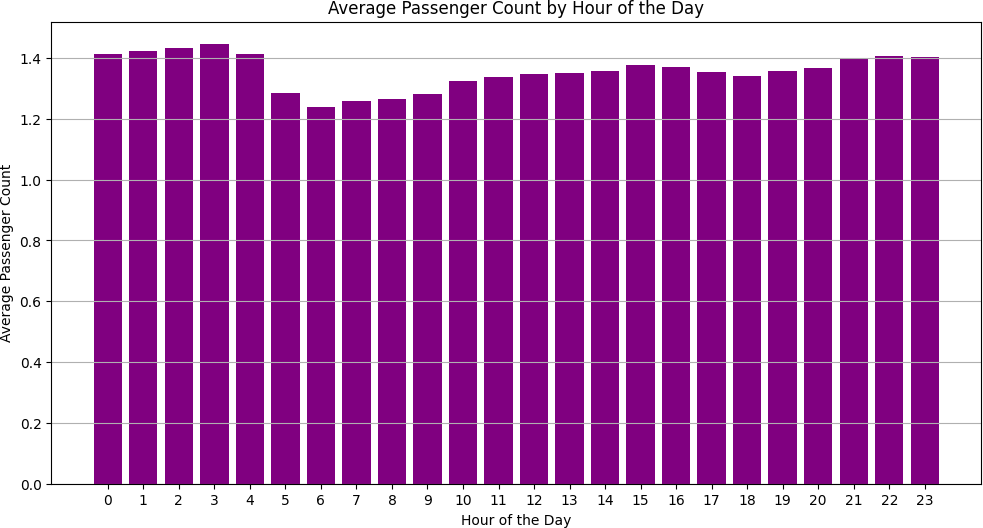
* + 1. **Compare the fare rates of different vendors in a distance-tiered fashion**

****

* + 1. **Analyse the tip percentages**

****

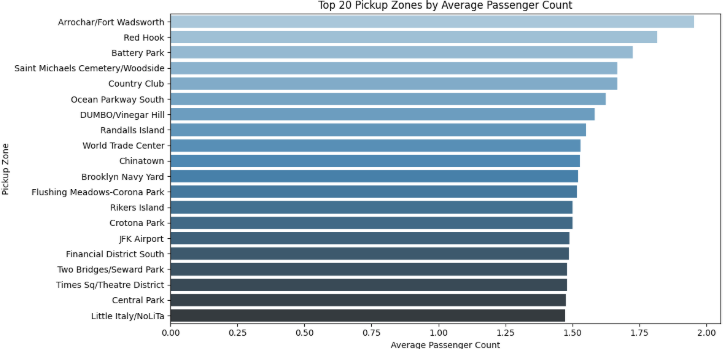
* + 1. **Analyse the trends in passenger count**

****

* + 1. **Analyse the variation of passenger counts across zones**

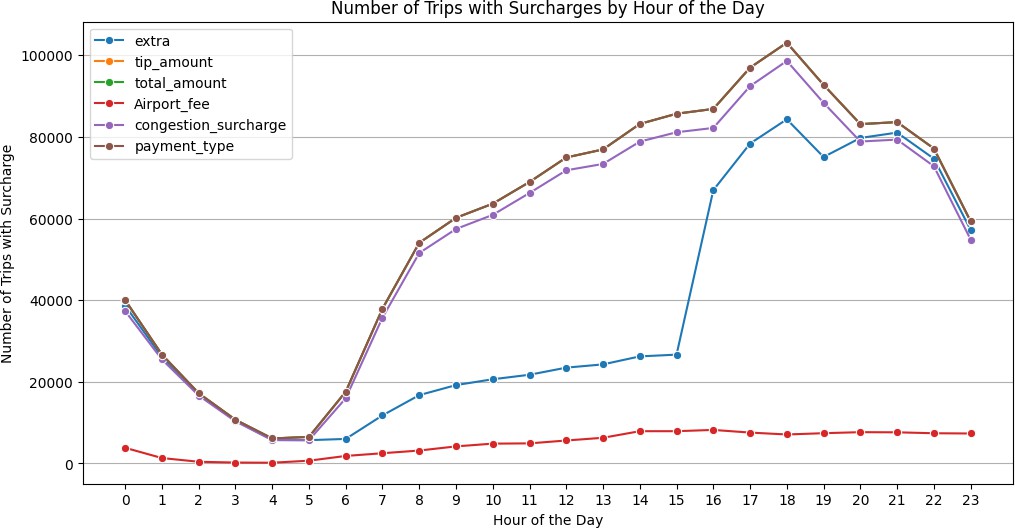
PULocationID avg\_passenger\_count\_per\_zone

|  |  |  |
| --- | --- | --- |
| 0 | 161 | 1.343836 |
| 1 | 246 | 1.388697 |
| 2 | 79 | 1.386548 |
| 3 | 79 | 1.386548 |
| 4 | 132 |  |



1.467763

* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently**

****

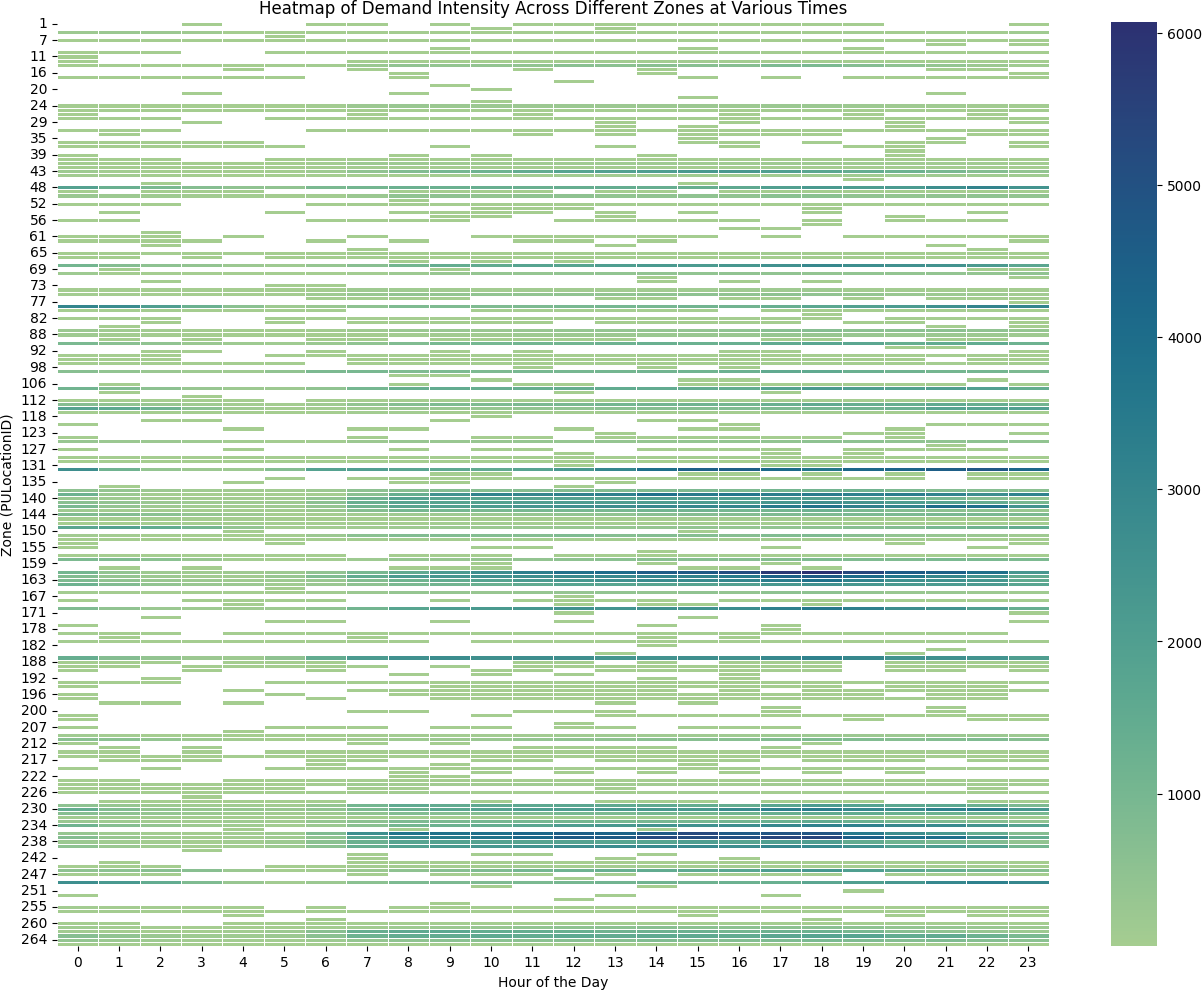
# Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

Recommendations to Optimize Routing and Dispatching:

* + - * Increase Cab Availability During Peak Daytime Hours (Based on Section 3.2.2) Deploy additional cabs between 6:00 AM and 10:00 PM to address high demand during peak daytime periods. This adjustment will help reduce passenger wait times and improve service efficiency.
      * Implement Surge Pricing in High-Demand Zones During Peak Hours Introduce dynamic pricing in areas experiencing high daytime demand to better balance supply and demand. This incentivizes drivers to move toward high-demand zones and increases profitability during busy hours.
      * Adjust Fare Rates Based on Time of Day and Day of the Week (Sections 3.2.4 & 3.2.10) Analyze average fare-per-mile trends to implement time-based and day-based pricing strategies. For example, apply higher rates during weekend evenings and weekday rush hours, while offering discounts during low-demand periods.
      * Expand Nighttime Coverage in High-Demand Zones (Section 3.2.7) Increase the number of active cabs between 11:00 PM and 5:00 AM in zones with consistent late-night demand. This improves service coverage during less active hours and caters to nightlife, airport, and shift- worker travel needs.
      * Implement Intelligent Repositioning Algorithms Introduce routing algorithms that automatically reposition idle or underutilized cabs to areas with anticipated demand surges. This data-driven dispatching approach enhances operational efficiency and maximizes cab utilization.
    1. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.**

****

#### Dentify High-Demand Zones by Time of Day

* Use the heatmap of trip counts across pickup zones and hours to pinpoint zones with consistently high demand (e.g., during morning and evening rush hours). ▪ Example: Zones showing strong activity between 7–10 AM and 4–8 PM should have increased cab presence during those intervals.
  + **Weekday vs. Weekend Demand Patterns**(Derived from Section 3.2.4) ▪ Weekday mornings and evenings often correspond to work-related commutes—focus on business districts and transportation hubs. ▪ Weekend demand may shift toward leisure zones (e.g., shopping, entertainment areas)—redeploy fleet to match this spatial pattern.

#### Match Cab Types with Trip Distances

* + - Position shorter-trip focused vehicles (e.g., sedans) in zones with high short-distance demand. ▪ Use larger or premium cabs in zones with longer average trip distances to optimize cost-efficiency and customer service.

#### Rebalancing Through Predictive Dispatch

* + - Use real-time data and historical patterns to reposition idle cabs to areas with expected demand surges. ▪ For example, after morning peaks in residential areas, shift cabs toward commercial districts for afternoon coverage.

#### Continuous Monitoring and Feedback

* + - Update zone-based positioning strategies regularly based on ongoing trip trend data. ▪ Integrate feedback loops to fine-tune allocation by time, day, and seasonal demand fluctuations.

#### Visual Aid Reference:

The heatmap titled “Heatmap of Demand Intensity Across Different Zones at Various Times” effectively reveals temporal and spatial trip density, guiding evidence-based cab deployment strategies.

* + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

The pricing strategy to maximize revenue while maintaining competitive rates with other vendors:

* + Monthly revenue is very low in July, August, September company can offer competative price as compared to other vendor during these month which can incrase pickup during that time and also revenue will increase
  + Correlation between Trip Duration and Fare Amount is 0.32 which is very low. Company can impose waiting charge for the ride which will increse the corrreation between these two variables.
  + Fare amount depended on count of pessenger can also increase the revenue for the company.
  + Consider using machine learning models to predict demand elasticity for various distances. This would allow more precise price adjustments.