STATISTICAL INSIGHTS INTO NEW YORK TAXI FARE PREDICTIONS

MA-541 B

By

Group: 7

Name	CWID
Amit Bhat Srirangapatana Guruprasad	20021872
Rohit Tiwari	20022079
Sreram Vasudev	20020902
Arun Kashyap	20022803

Date: 04/07/2024

Submitted to:

Prof. Hong Do

Table of Contents

1. INTRODUCTION

In the bustling urban landscape of New York City, taxis are a vital part of the transportation network, providing quick and convenient mobility for residents and visitors alike. With the city's ever-changing dynamics, accurately predicting taxi fares becomes crucial for both service providers and passengers, ensuring fair pricing and efficient service management. This project delves into the complex world of NYC taxi fares, aiming to harness statistical analysis and machine learning techniques to forecast fare amounts based on historical trip data. By examining a range of factors from trip distance to passenger counts, this study seeks to uncover the key determinants of taxi pricing, thereby offering valuable insights for riders, drivers, and policymakers. The overarching questions guiding this analysis include how spatial and temporal variables influence fare amounts, and the impact of passenger numbers on pricing.

The following key questions guide our analytical journey:

- **How do trip distance and day of the week impact the total taxi fare?**

  We aim to understand the combined effects of these variables on fare calculation, considering both the direct costs and the broader temporal and situational contexts.

- **Does the relationship between trip distance and total fare differ when comparing travel within the city to travel to or from locations outside the city?**

  This inquiry will explore the geographical pricing differences based on trip distance, assessing the fare variability between intra-city and extra-city travel.

- **How does a traveler's sentiment towards tipping vary, and how does this behavior change with increasing travel distances?**

  We will examine tipping trends to understand how journey length affects tipping behavior, analyzing the psychological factors at play.

2. DATASET DESCRIPTION

The dataset for this study is comprised of 6200 non-null taxi trip records, each meticulously detailed with a variety of attributes relevant to the taxi fare prediction task. The features, along with their data types, are as follows:

- *VendorID (int64)*: An integer identifier representing the taxi vendor.
- *lpep_pickup_datetime (object)*: A string timestamp indicating when the ride started.
- *lpep_dropoff_datetime (object):* A string timestamp indicating when the ride ended.
- *store_and_fwd_flag (object)*: A string flag denoting if the trip data was stored in vehicle memory before sending to the vendor.
- *RatecodeID (int64)*: An integer code signifying different types of fares that apply to the trip.
- *PULocationID (int64)*: An integer identifier for the pickup location based on taxi zone numbers.
- *DOLocationID (int64)*: An integer identifier for the drop-off location based on taxi zone numbers.
- *passenger_count (int64)*: An integer counts the number of passengers in the taxi ride.
- *trip_distance (float64)*: A float representing the recorded distance of the trip in miles.
- *fare_amount (float64)*: A float indicating the fare amount for the trip, which is the primary target variable for prediction.
- *extra (float64)*: A float for additional charges and fees.
- *mta_tax (float64)*: A float representing the metropolitan commuter transportation mobility tax.
- *tip_amount (float64)*: A float indicating the tip amount given to the driver.
- *tolls_amount (float64)*: A float representing the total amount of tolls paid during the trip.
- *improvement_surcharge (float64)*: A float for an additional charge intended for improvements to the taxi service.
- *total_amount (float64)*: A float representing the total fare amount, including all fees and taxes.
- payment_type (int64): An integer indicating the method of payment for the trip.
- *trip_type (int64):* An integer categorical field indicating the type of trip.
- *congestion_surcharge (float64)*: A float indicating the fee charged for trips passing through congested areas.

3. EDA

We plotted histograms of all the significant fields in the dataset. Below were the observations:

Passenger Count: Most rides have one or two passengers. The distribution is highly right-skewed.

Trip Distance: The majority of trips are under 10 miles, with very few longer trips. The distribution is right-skewed with a long tail.

Fare Amount: Most rides have fare amounts concentrated between 0 and 20 units, with some fares going above 50 units. The distribution is right-skewed.

Tip Amount: Most trips have little to no tips, with a few cases where the tips are significant. The distribution is highly right-skewed.

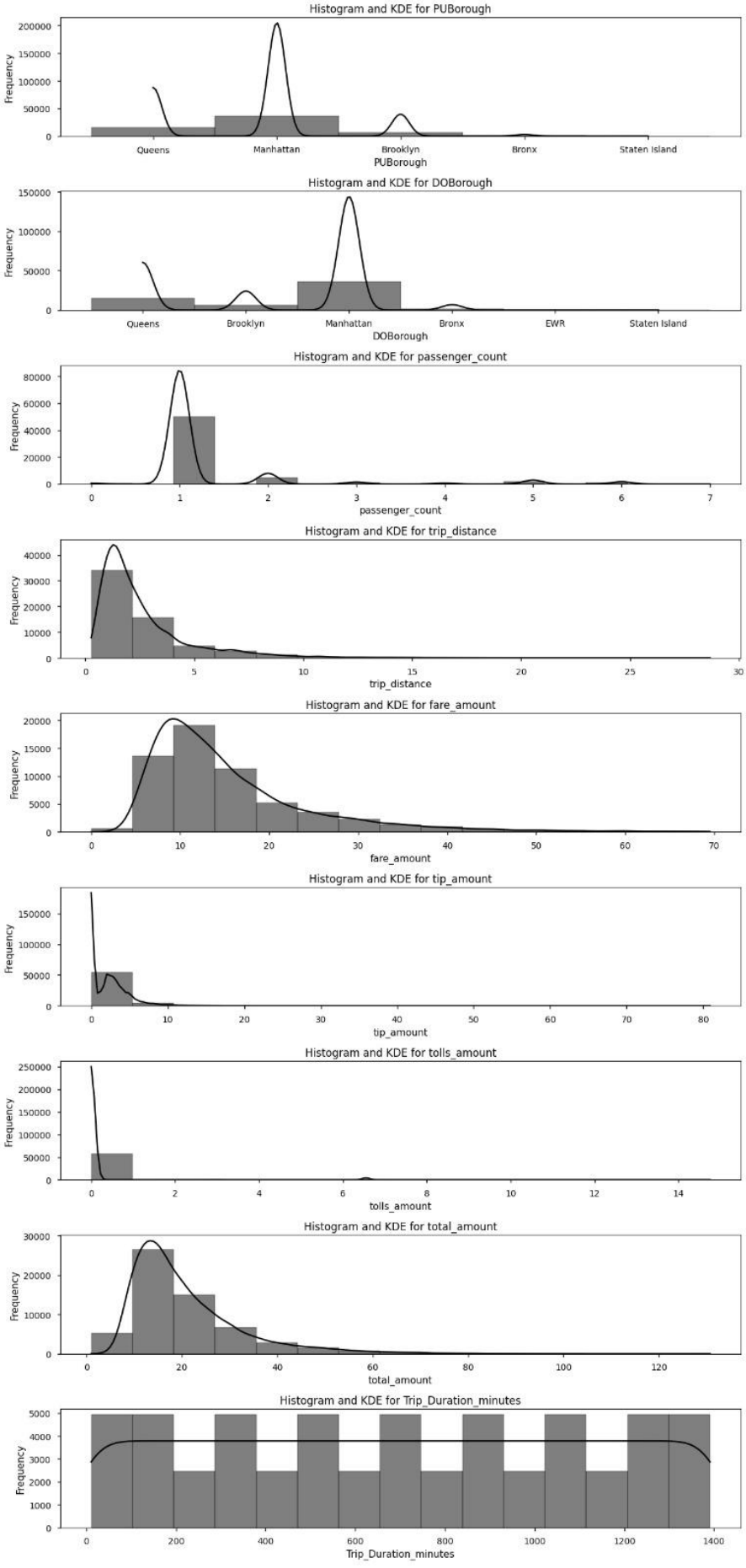
Tolls Amount: Most rides do not have tolls, and the few that do usually have low toll amounts. The distribution is very right-skewed with a very long tail.

Total Amount: Most rides have a total amount between 0 and 20 units, with some reaching up to 50 units. The distribution is right-skewed with a long tail.

Trip Duration (minutes): Most trips are shorter than 400 minutes, with some longer trips. The distribution appears somewhat uniform across the duration bins.

Pickup Borough (PUBorough): The most frequent pickup borough is Manhattan, followed by Brooklyn. Other boroughs have fewer pickups.

Dropoff Borough (DOBorough): The most frequent dropoff borough is also Manhattan, followed by Brooklyn. Other boroughs have fewer drop offs.


Histogram and KDE for PUBorough
Frequency
PUBorough
Queens
Manhattan
Brooklyn
Bronx
Staten Island
Histogram and KDE for DOBorough
Frequency
DOBorough
Queens
Brooklyn
Manhattan
Bronx
EWR
Staten Island
Histogram and KDE for passenger_count
Frequency
passenger_count
Histogram and KDE for trip_distance
Frequency
trip_distance
Histogram and KDE for fare_amount
Frequency
fare_amount
Histogram and KDE for tip_amount
Frequency
tip_amount
Histogram and KDE for tolls_amount
Frequency
tolls_amount
Histogram and KDE for total_amount
Frequency
total_amount
Histogram and KDE for Trip_Duration_minutes
Frequency
Trip_Duration_minutes

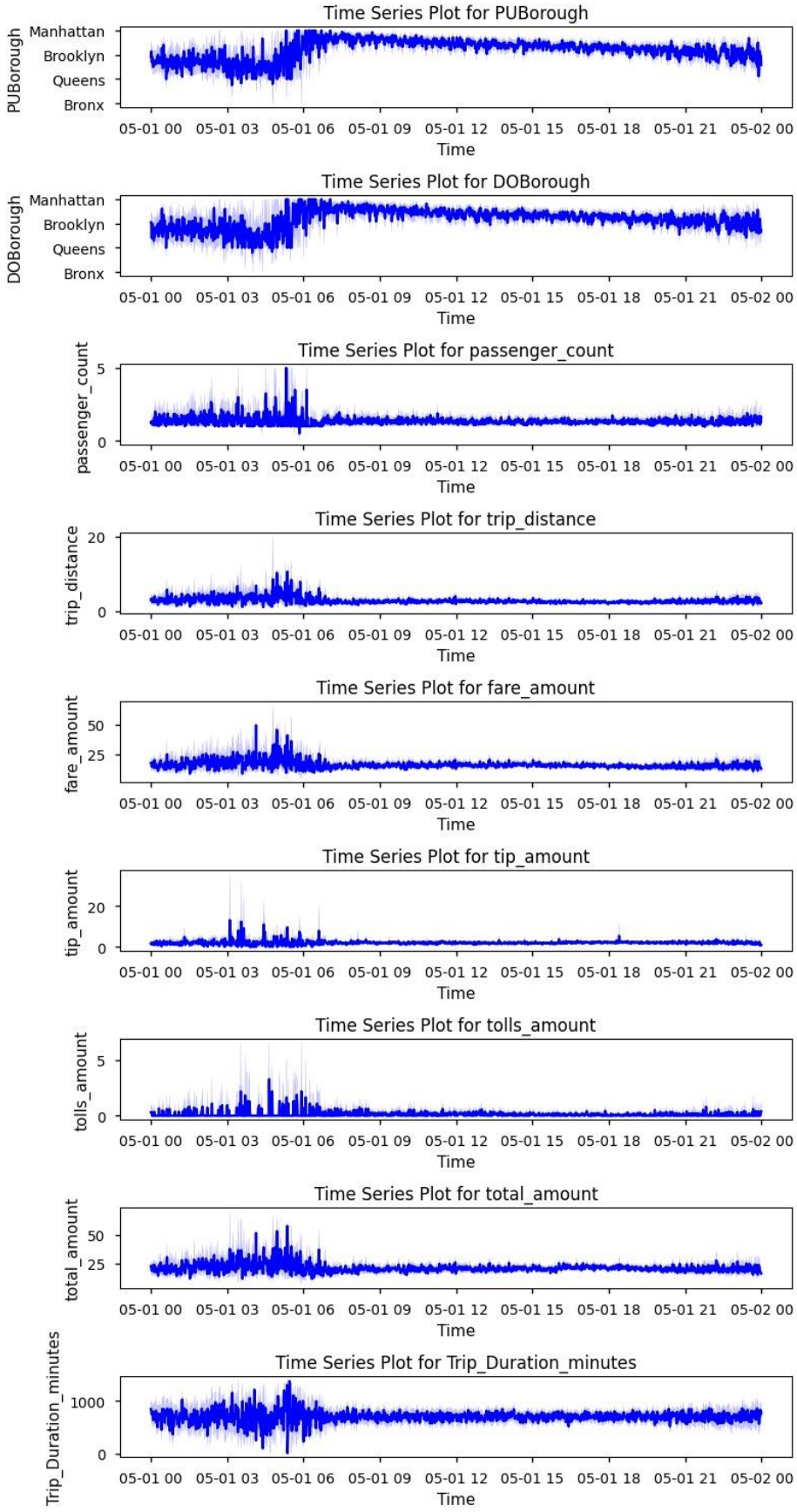
Similarly, we performed a time-series analysis of the significant fields from the dataset with the help of a time series plot as below:

Passenger Count:

The passenger count fluctuates over time but remains within the range of 1 to 5 passengers per ride.

Trip Distance: Most trips maintain a consistent distance throughout the time range, generally under 10 units. The occasional spikes indicate longer trips.
Fare Amount: The fare amounts are mostly under 50 units with some fluctuations, especially around the morning hours.
Tip Amount: The tip amounts are relatively low throughout the time range. Some fluctuations can be seen early in the time range.
Tolls Amount: The tolls amounts are generally under 5 units with consistent spikes.
Total Amount: The total amounts mostly range from 0 to 50 units, with occasional spikes.

Trip Duration (minutes): Trip durations appear mostly under 400 minutes. Some trips extend significantly beyond this range, though they are less frequent.



Through a series of visualizations and statistical examinations, we aim to distill the intricate relationships between various factors and fare amounts.

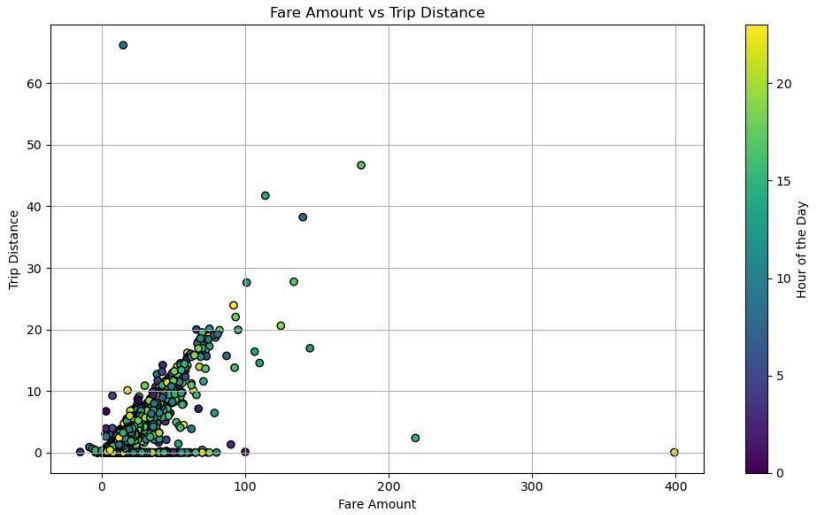


Figure 4 : Trend of Fare Amount with Trip Distance

The scatter plot analysis revealed a direct correlation between trip distance and fare amount, with an expected increase in fare corresponding to longer trips. Notably, fare variability was observed for trips of similar distances, suggesting additional factors at play. The color coding by hour suggested fare differentiation possibly due to time-of-day surge pricing.

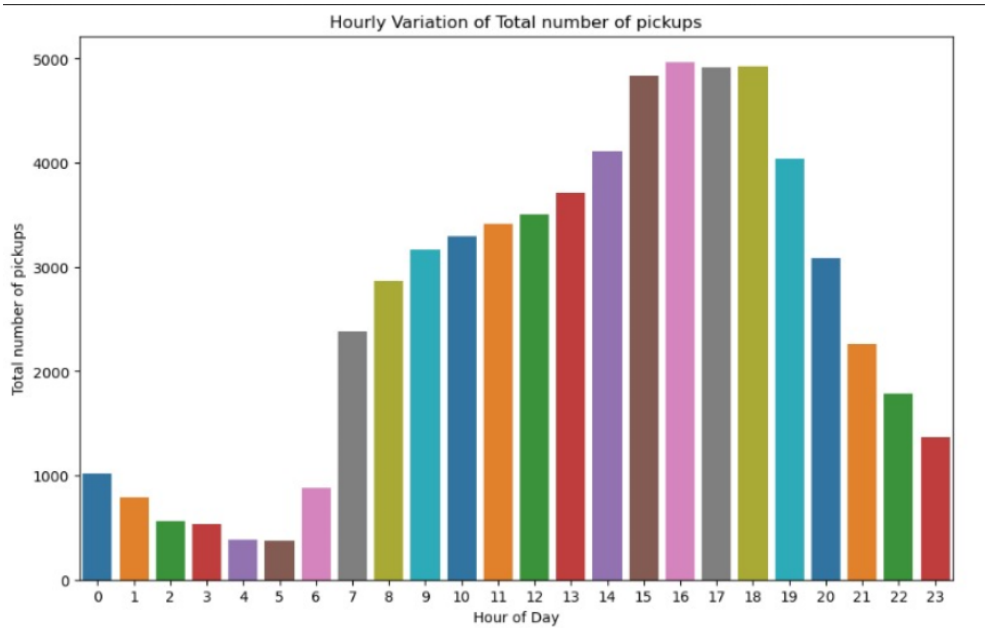
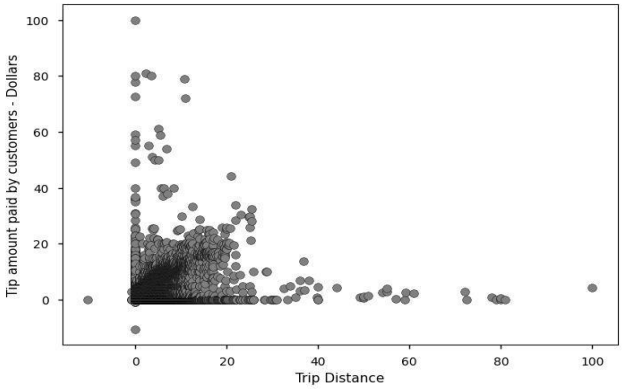


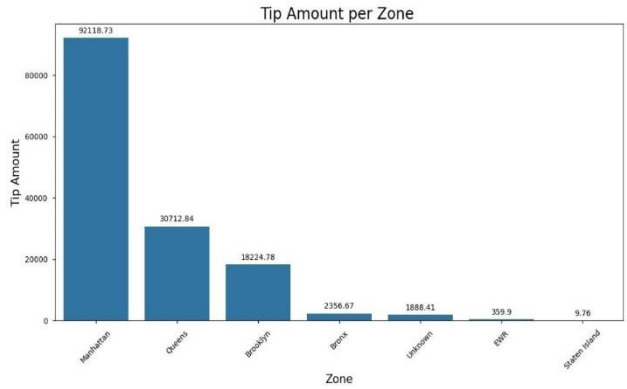
Figure 5 : Hourly trend for the total number of pickups

From the bar charts, a clear pattern of taxi demand emerged, peaking during typical rush hours and midweek days, particularly Wednesday, then declining towards the weekend. The hourly analysis showed spikes in taxi usage starting at 7 AM, which sustained into the evening, indicating peak business hours for taxi services. The data also indicates that trip location—whether within or outside the city—affects fare amounts, with longer-distance trips generally costing more. However, city trips exhibit a broader range of fares, influenced by high outliers. Tolls for outside-city trips also contribute to their higher average fares, underscoring the impact of travel distance and city boundaries on taxi fare structures.

We also utilized a bar graph displaying the total tip amount accumulated for different zones and two scatter plots that compare the tip amount with trip distance and the fare amount with the tip amount, respectively. This approach aims to discern patterns in tipping behavior relative to travel distances and zones.

The bar graph shows a significantly higher total tip amount for Manhattan, suggesting a greater volume of trips or a higher propensity to tip in this zone. The scatter plots reveal that for short to moderate distances, tips increase, but beyond a certain distance, this increase plateaus or becomes inconsistent. There is a concentration of trips with low to moderate tips regardless of distance, implying a standard tipping behavior for typical trips.





Moreover, the second scatter plot suggests that while there is a positive correlation between fare amounts and tips, the relationship isn't linear, indicating that passengers may cap their tips after a certain fare amount is reached.As travel distance increases, the tendency to tip more does increase, but only to a point, after which it stabilizes or varies independently of the distance.

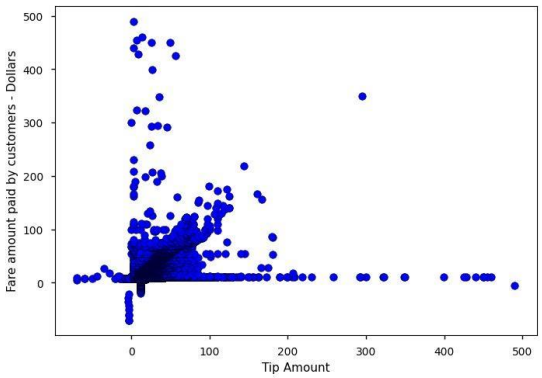


Figure 12 : Tip Amount Vs. Fare Amount

NULL VALUES

In the dataset, we examined the columns for missing values to ensure data quality and reliable model predictions. Here's a summary of the findings:

VendorID, pickup and dropoff times, Location IDs, and fare details: All these columns have complete data, with zero null values, ensuring that essential identifiers and ride details are intact for analysis.

PUZone and PUservice_zone: These columns have 66 and 201 missing values respectively, which could indicate challenges in recording pickup zones consistently. This is crucial because it might affect analysis related to geographical trends.

DOZone and DOservice_zone: With 242 and 668 missing values, respectively, similar challenges are present in identifying drop-off zones

```
df.isnull().sum()
```

```
VendorID                  0
lpep_pickup_datetime      0
lpep_dropoff_datetime     0
store_and_fwd_flag     4324
RatecodeID             4324
PULocationID              0
PUBorough                 0
PUZone                   66
PUservice_zone          201
DOLocationID              0
DOBorough                 0
DOZone                  242
DOservice_zone          668
passenger_count        4324
trip_distance             0
fare_amount               0
extra                     0
mta_tax                   0
tip_amount                0
tolls_amount              0
ehail_fee             68211
improvement_surcharge     0
total_amount              0
payment_type           4324
trip_type              4334
congestion_surcharge   4324
dtype: int64
```

Passenger count: With 4,324 non-null values, there is missing data indicating that passenger count was not recorded for some rides. This could impact analyses involving ride occupancy.

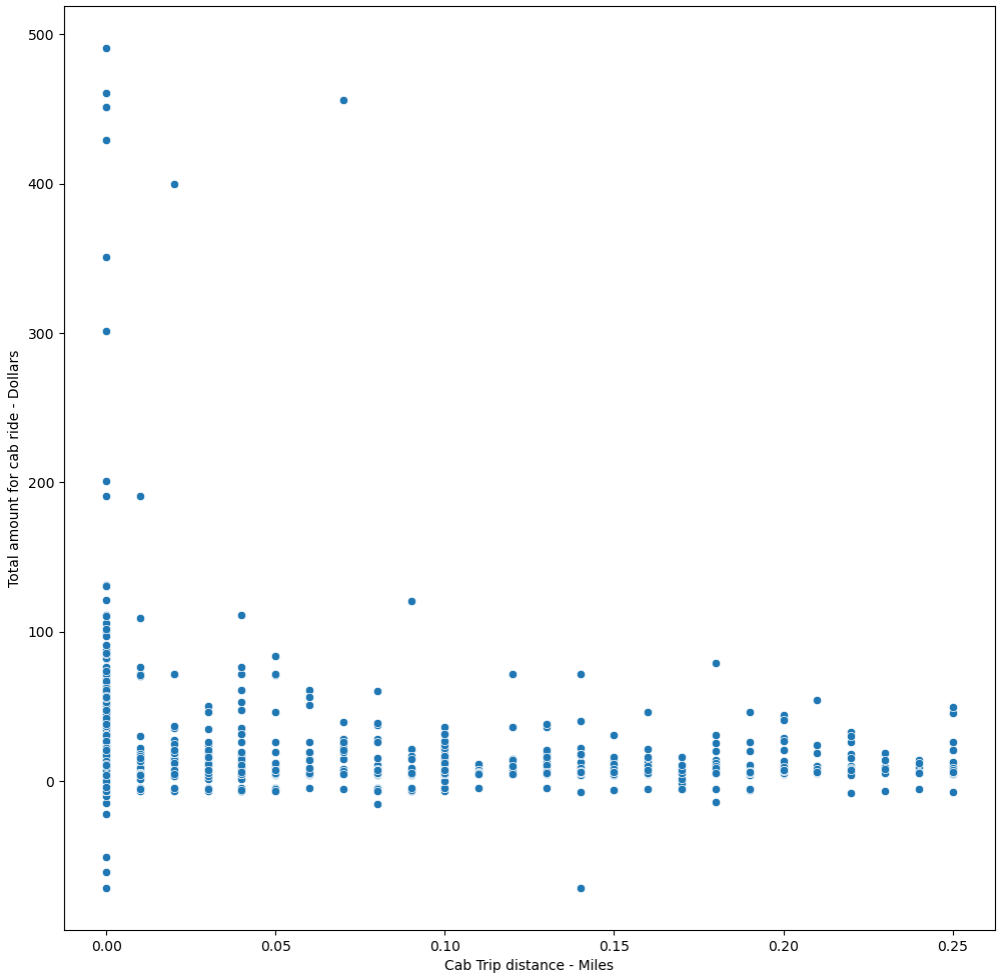
Trip Type: The data has 4,334 non-null values for this feature, indicating that a few trips might have undefined types.

Payment Type and Congestion Surcharge: Both columns have the same count of non-null values (4,324), suggesting that missing data in these columns likely stems from similar issues.

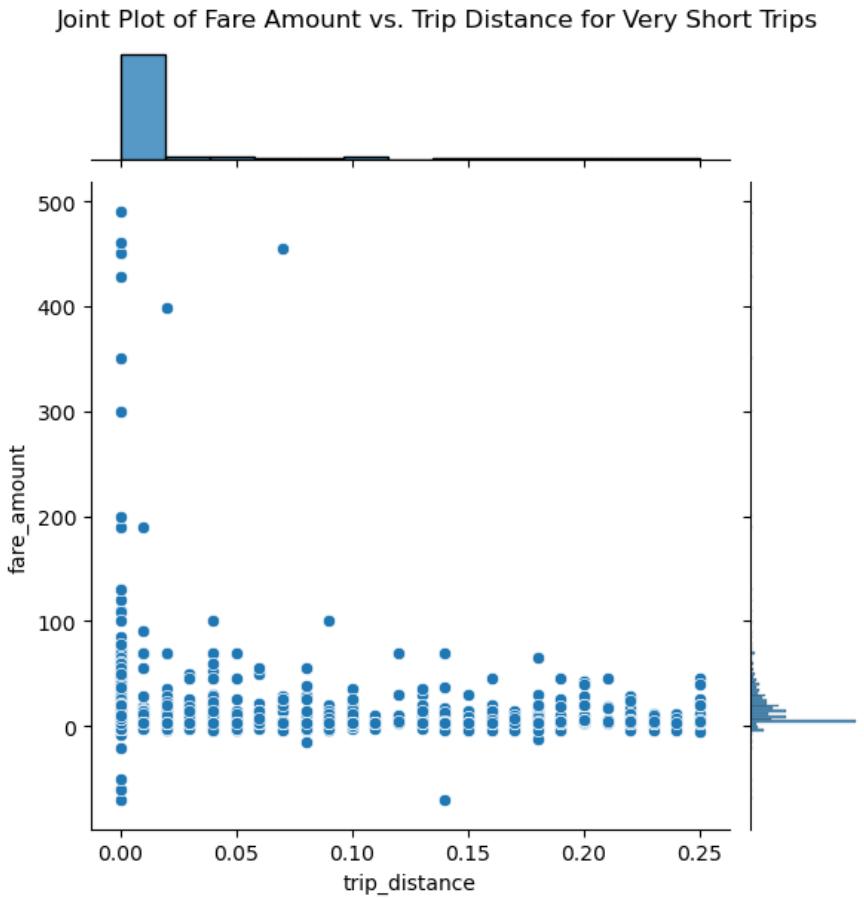
Since we have a lot of data , a lot of null value data points were dropped . The columns with numerical inputs with more than 500 null points were handled using measures of central tendency like mean . The categorical variables were handled using the mode.

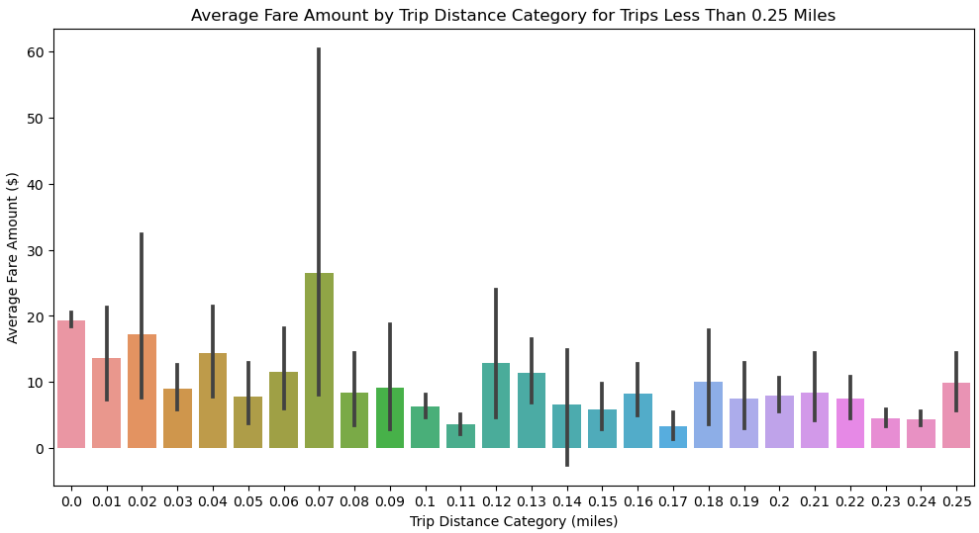
EDA for outliers

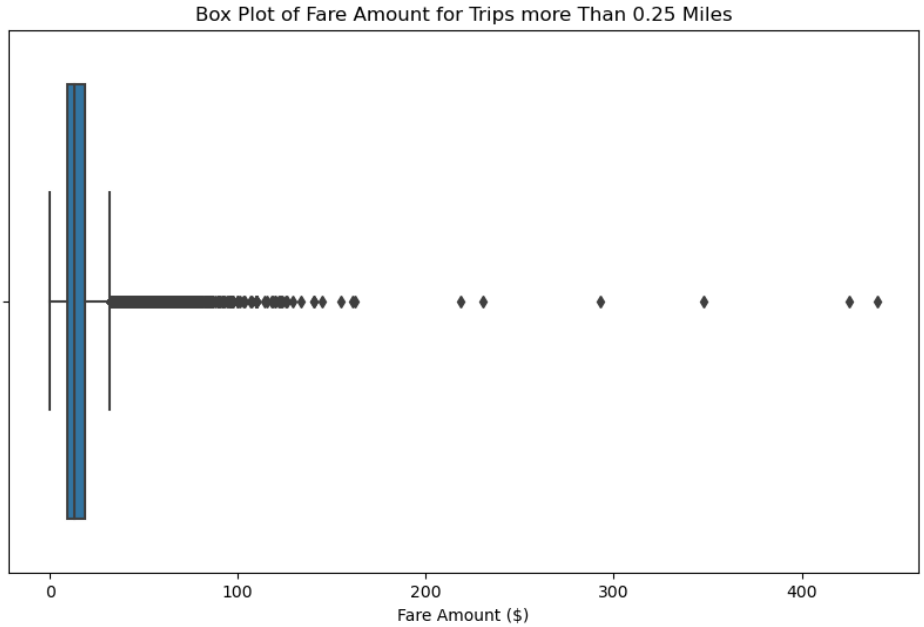
In columns such as trip distance , fare amount , month and tip we saw a lot of outliers. For fare amount and trip distance we visualized this using plots , confirmed the datapoints and handled them accordingly.



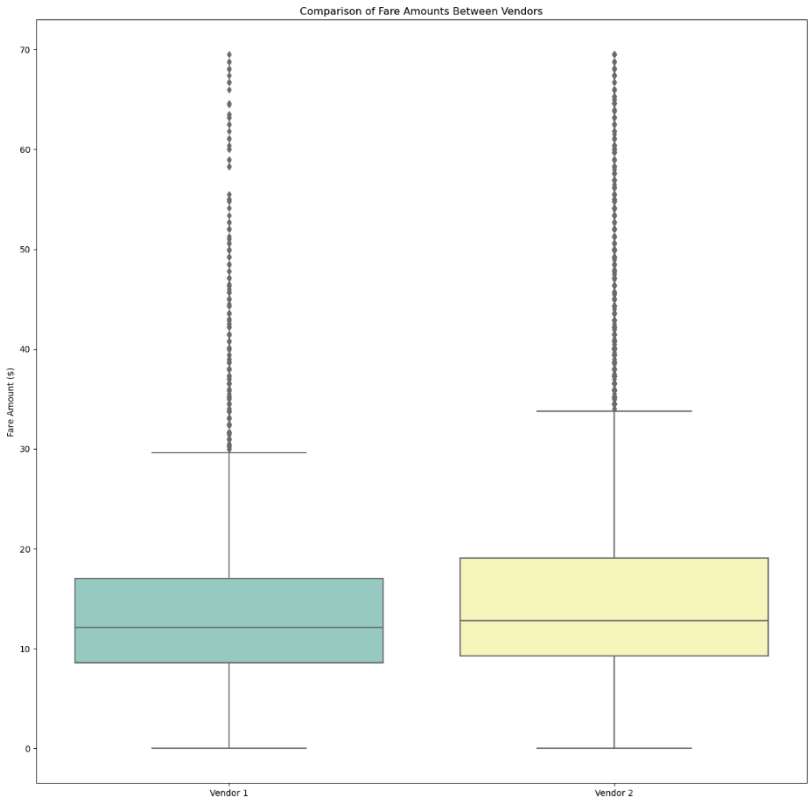
The plot highlights that there are several rides that charged abnormally high fares despite being very short in distance.On the other hand we also observed that a lot of data points had abnormally high fare for a short distance of less than 0.25 miles. Handling these points was was necessary emphasizing the importance of handling outliers carefully in data analysis to avoid skewing results.


Joint Plot of Fare Amount vs. Trip Distance for Very Short Trips
fare_amount
trip_distance
500
400
300
200
100
0
0.00
0.05
0.10
0.15
0.20
0.25


Average Fare Amount by Trip Distance Category for Trips Less Than 0.25 Miles
Average Fare Amount ($)
Trip Distance Category (miles)
60
50
40
30
20
10
0
0.0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.1 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.2 0.21 0.22 0.23 0.24 0.25



The plots above showed that a lot of datapoints were outliers mainly caused due to incorrect recording . These points which were far deviated were removed.



We can see the same from this boxplot too that the fare amount data points are too varied.

To generalize all the trip distances less than 0.25 and more than 1000 miles were removed .

In the cab duration part anything more than 750 minutes were removed

Any tip which was more than 200$ were removed as well .

4. DESCRIPTIVE STATISTICS

The descriptive statistics for the taxi fare dataset provide insights into the distribution of passenger_count, trip_distance, fare_amount, and extra:

passenger_count		*trip_distance*		*fare_amount*		*extra*	
Mean	1.324092699	Mean	2.596900218	Mean	15.39676178	Mean	0.921453202
Standard Error	0.004079135	Standard Error	0.009064793	Standard Error	0.038457063	Standard Error	0.005367063
Median	1	Median	1.89	Median	12.8	Median	0
Mode	1	Mode	1.4	Mode	8.6	Mode	0
Standard Deviation	0.992152605	Standard Deviation	2.204795357	Standard Deviation	9.353766446	Standard Deviation	1.305410533
Sample Variance	0.984366792	Sample Variance	4.861122566	Sample Variance	87.49294672	Sample Variance	1.704096659
Kurtosis	11.40519	Kurtosis	8.046234977	Kurtosis	4.895572142	Kurtosis	3.393682164
Skewness	3.442335394	Skewness	2.403470101	Skewness	1.932686829	Skewness	1.663015793
Range	7	Range	28.4	Range	69.49	Range	7.5
Minimum	0	Minimum	0.27	Minimum	0.01	Minimum	0
Maximum	7	Maximum	28.67	Maximum	69.5	Maximum	7.5
Sum	78332	Sum	153630.02	Sum	910857.03	Sum	54512.25
Count	59159	Count	59159	Count	59159	Count	59159

Passenger Count: Most trips carry one or two passengers, indicated by a mean of 1.32 and a mode of 1. High kurtosis and skewness suggest a concentration of data around fewer passengers.

Trip Distance: Averages 2.6 units with a substantial variation (standard deviation of 2.36), reflecting diverse trip lengths. The data is moderately right-skewed, suggesting longer trips are less frequent but impact the average.

Fare Amount: Mean fare is $15.40, with considerable spread (standard deviation of 9.35), influenced by varying trip distances and additional charges.

Extra: Most trips do not involve extra charges as shown by the mode of 0. However, the range indicates that when applied, extras can vary significantly.

The descriptive statistics for mta_tax, tip_amount, tolls_amount, and improvement_surcharge from the taxi fare dataset elucidate their distribution and typical values:

mta_tax		*tip_amount*		*tolls_amount*		*improvement_surcharge*	
Mean	0.62108893	Mean	2.080443719	Mean	0.105883974	Mean	0.956006694
Standard Error	0.001391541	Standard Error	0.011548983	Standard Error	0.003378174	Standard Error	0.000701887
Median	0.5	Median	1.74	Median	0	Median	1
Mode	0.5	Mode	0	Mode	0	Mode	1
Standard Deviation	0.338459366	Standard Deviation	2.80901561	Standard Deviation	0.821660545	Standard Deviation	0.170717241
Sample Variance	0.114554742	Sample Variance	7.890568695	Sample Variance	0.675126051	Sample Variance	0.029144376
Kurtosis	2.888180651	Kurtosis	711.7006835	Kurtosis	59.83071923	Kurtosis	11.34462001
Skewness	2.107618926	Skewness	12.2891355	Skewness	7.776134353	Skewness	-3.639642951
Range	1.5	Range	222.22	Range	14.75	Range	1
Minimum	0	Minimum	0	Minimum	0	Minimum	0
Maximum	1.5	Maximum	222.22	Maximum	14.75	Maximum	1
Sum	36743	Sum	123076.97	Sum	6263.99	Sum	56556.4
Count	59159	Count	59159	Count	59159	Count	59159

MTA Tax: Mostly constant at 0.5 (mode), with a narrow range (0 to 1.5), indicating standard tax charges with occasional variations. High kurtosis and positive skewness suggest a peak around the modal value with few outliers.

Tip Amount: Average tip is around $2.08, but with a high standard deviation (2.81), indicating significant variability in how much passengers tip. Extremely high kurtosis and skewness show that tips are typically low but can occasionally be very high, as seen in the maximum value of $222.22.

Tolls Amount: Most trips do not incur tolls (mode of 0), but the range up to $14.75 and a standard deviation of 0.82 indicates that when tolls apply, they can vary substantially. The skewness and kurtosis confirm that higher tolls are rare but significantly impact the mean.

Improvement Surcharge: Highly consistent at $1 (mode and median), with nearly all observations showing no variation. The negative skewness and high kurtosis indicate a distribution heavily concentrated at the standard surcharge rate.

The descriptive statistics for total_amount, congestion_surcharge, and Trip_Duration_minutes from the taxi fare dataset highlight varied aspects of taxi rides:

total_amount		*congestion_surcharge*		*Trip_Duration_minutes*	
Mean	20.62094339	Mean	0.773905915	Mean	700.9883365
Standard Error	0.047197976	Standard Error	0.005084136	Standard Error	1.707581133
Median	17.5	Median	0	Median	671
Mode	10.8	Mode	0	Mode	11
Standard Deviation	11.47978588	Standard Deviation	1.23659521	Standard Deviation	415.3285251
Sample Variance	131.7854838	Sample Variance	1.529167713	Sample Variance	172497.7838
Kurtosis	6.571180081	Kurtosis	-1.055100474	Kurtosis	-1.204165823
Skewness	1.908350559	Skewness	0.97205455	Skewness	6.73773E-06
Range	231.71	Range	2.75	Range	1380
Minimum	1.01	Minimum	0	Minimum	11
Maximum	232.72	Maximum	2.75	Maximum	1391
Sum	1219914.39	Sum	45783.5	Sum	41469769
Count	59159	Count	59159	Count	59159

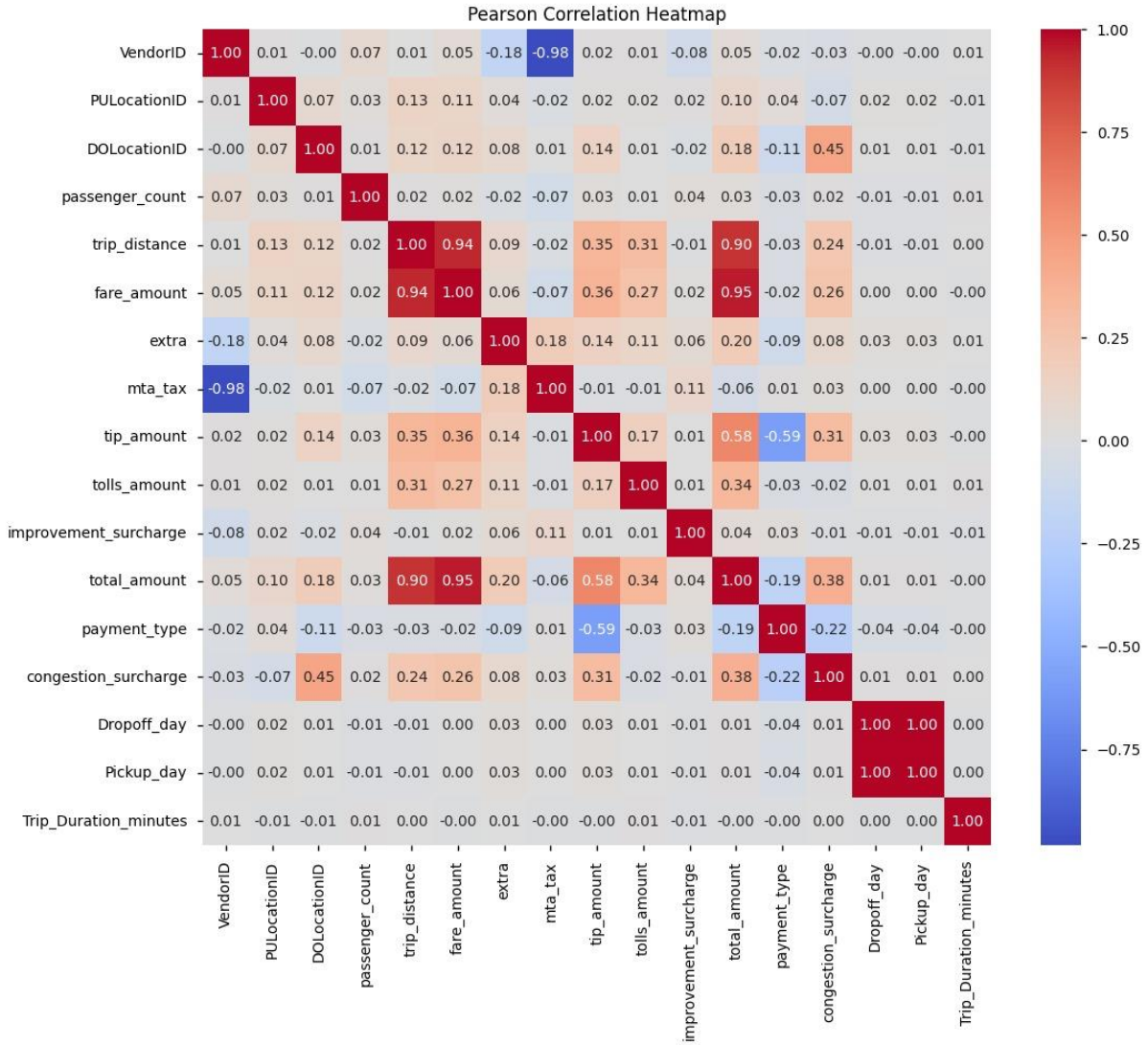
Total Amount: Shows a broad range from $1.01 to $232.72 with a mean of $20.62 and a standard deviation of $11.48, indicating diverse fare amounts influenced by different trip types and lengths.

Congestion Surcharge: Has an average of $0.77 with a standard deviation of 1.24, mostly centered around the typical $0.75 fee, reflecting its application in congestion-prone areas.

Trip Duration: The average (701 minutes) and median (671 minutes) durations are unexpectedly high, suggesting potential outliers or data recording errors, despite most trips being much shorter as indicated by a mode of 11 minutes.

5. CORRELATION

In the correlation analysis for our taxi fare prediction project, a Pearson correlation heatmap was used to visualize the relationships between several numerical features within the dataset. The heatmap reveals strong correlations between certain variables, particularly between total_amount and both fare_amount and trip_distance, which exhibit high positive correlation coefficients close to 0.95, suggesting that as trip distance increases, so does the fare amount proportionally. Notably, the mta_tax and extra show significant negative correlations with VendorID, indicating possible variability in tax and surcharge application between different vendors. Variables such as payment_type and congestion_surcharge display a mix of weak and moderate negative correlations with fare-related variables, hinting at the complex dynamics influencing fare calculations. This detailed correlation study helps in identifying key features that most significantly impact the fare, facilitating a focused model refinement for better fare estimation accuracy.



6. HYPOTHESIS TESTING

1. Hypothesis Testing for Median Tip Amounts by Trip Distance Categories

This test is appropriate as we aim to compare the median tip amounts across different trip distance categories, assuming these categories represent independent groups within our dataset. The Kruskal-Wallis test, a non-parametric method, is used due to its ability to handle data that may not follow a normal distribution and because it compares medians rather than means.

Hypothesis:

- **Null Hypothesis (H0)**: There is no difference in the median tip amounts across different trip distance categories (Less than 5 miles, 5-10 miles, 10-15 miles, 15-20 miles, 20-25 miles, and over 25 miles).
- **Alternative Hypothesis (H1)**: There is a significant difference in the median tip amounts between at least one of the specified trip distance categories.

Significance Level (α): 0.05. This represents a 5% risk of rejecting the null hypothesis when it is actually true.

The test statistic *H* is calculated using the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total sample size, k is the number of groups we are comparing, R_i is the sum of ranks for group i, and n_i is the sample size of group i.

Kruskal-Wallis Test Results:

- **Number of Categories**: 6 (representing the different trip distance categories).
- **Degrees of Freedom (DOF)**: Number of Groups - 1 = 8 - 1 = 7.
- **Kruskal-Wallis H Statistic**: 1098.429276209346. This value indicates the variance between the group medians relative to the variability within the groups.
- **P-value**: 2.926×10−2352.926×10−235. This p-value indicates the probability of observing such extreme results if the null hypothesis were true.
- **P-value vs. Significance Level**:
    - If the p-value is less than α (0.05), reject the null hypothesis.

- If the p-value is greater than or equal to α, fail to reject the null hypothesis.

Interpretation:

- Given that the p-value (2.926×10−2352.926×10−235) is significantly lower than the significance level (0.05), we reject the null hypothesis.
- **Conclusion**: Based on the dataset used in this study, there is significant evidence to suggest that the median tip amounts differ across the trip distance categories. This indicates that trip distance does play a role in influencing the median tip amount.

ANOVA Analysis :

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Manhattan	59159	36217	0.612197637	0.237415703
Bronx	59159	1700	0.028736118	0.027910825
Brooklyn	59159	6052	0.10230058	0.091836724
Queens	59159	15179	0.256579726	0.190749794
Staten Island	59159	9	0.000152132	0.000152112
EWR	59159	2	3.38072E-05	3.38066E-05
tip_amount	59159	123076.97	2.080443719	7.890568695
trip_distance	59159	153630.02	2.596900218	4.861122566

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	443397.7603	7	63342.53719	38101.3752	0	2.009610743
Within Groups	786788.9902	473264	1.662473778			
Total	1230186.751	473271				

ANOVA Summary for Testing Differences Between Boroughs and distance:

- **F-Statistic**: 38101.3752 — This is calculated by dividing the Between Groups MS by the Within Groups MS. It tests whether the group means are significantly different.
- **P-value**: 0.0 — Indicates the probability of observing the F-statistic, assuming the null hypothesis is true.
- **F critical**: 2.009610743 — This is the critical value of F at the given significance level (presumably 0.05), beyond which we would reject the null hypothesis.

Conclusion of ANOVA Analysis:

The ANOVA conducted to test for differences in the mean tip amounts across different boroughs revealed an F-statistic of 38,101.3752, indicating a significant amount of variance between the group means relative to the variance within each group.

Since p-value < significance level i.e., 0 < 0.05, we reject the null hypothesis. This conclusion is supported by the extremely low p-value. Hence, There is a significant difference in the median tip amounts between at least one of the specified trip distance categories.

2. **Hypothesis testing for relationship between passenger count and total fare with respect to pickup zones**

   The objective of this analysis is to assess whether the relationship between passenger count and total fare (total_amount) differs across various pickup zones in the NYC Taxi Fare dataset. To investigate this, we conducted a non-parametric Kruskal-Wallis test, which is suitable for comparing distributions across groups and does not require normally distributed data.

   Hypothesis:
   Null Hypothesis (H0): The Relationship between passenger count and total fare does not differ across different pickup zones

   Alternative Hypothesis (H1): The relationship between passenger count and total fare (total_amount) does differ across different pickup zones.
   The dataset used for the analysis contains information on taxi trips in NYC, including columns such as passenger_count, total_amount, and PUZone. The PUZone column, representing pickup zones like 'Manhattan' and 'Queens,' was one-hot encoded to facilitate grouping and analysis. The significance level of 0.05 was selected for the analysis

   The Kruskal-Wallis test results:
   The test was performed on total fare distributions grouped by combinations of passenger_count and PUZone levels.

- **No. of Categories:** 5 (5 different pickup zones)
- **Degrees of Freedom (DOF):** 4 (i.e. 5 – 1 = 4 DOF)
- **Kruskal – Wallis 'H' statistic:** 423.2564912554113
- **P-value:** 5.472853591558342e-71

Conclusion:

- It was observed that the p-value (5.472853591558342e-71) is significantly lower than the significance level of 0.05. Hence, we reject the null hypothesis.
- Therefore, we can conclude that the relationship between passenger count and total fare differs significantly across different pickup zones.
  Based on the p-value, we reject the null hypothesis that the relationship between passenger count and total fare is consistent across different pickup zones. This finding suggests that there may be important factors influencing total fare that vary depending on passenger count and pickup zone.

These results have implications for transportation planning and pricing strategies in the NYC taxi service. Further research may explore additional factors or alternative models to better understand the dynamics of taxi fare pricing in the city.

ANOVA Analysis:

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
passenger_count	32044	42414	1.323617526	0.984679532
total_amount	32044	661509.46	20.64378542	130.679137
PUBorough_Bronx	32044	294	0.009174885	0.00909099
PUBorough_Manhatt	32044	19627	0.61250156	0.237350806
PUBorough_Queens	32044	8387	0.261733866	0.19323528
PUBorough_brook	32044	3732	0.116464861	0.102904008
PUBorough_Staten	32044	4	0.000124828	0.000124817

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	11312101.73	6	1885350.289	99824.51529	0	2.098638066
Within Groups	4236293.599	224301	18.88664606			
Total	15548395.33	224307				

ANOVA summary:

F-statistic: 99824.515
P-Value: 0.0
F-critical: 2.0986

Conclusion of ANOVA Analysis:
ANOVA Analysis was performed to check for the relationship between passenger count and total fare differs significantly across different pickup zones. The F-statistics was found to be 99824.515 which represents significant variance between the group mean with respect to variance within the group. Here, since p-value < significance level, we reject the null hypothesis. Hence, we can conclude that there is indeed a relationship between passenger count and total fare varying significantly across different pickup zones.

3. VENDOR HYPOTHESIS TESTING

Null Hypothesis (H0) : There is no difference in the average fares between Vendor 1 and Vendor 2

Alternative Hypothesis (H1) : There is a significant difference in the average fares between Vendor 1 and Vendor 2

Descriptive Analysis

Lets first analyze the data and see how the dataset is spread across the features. We first analyze the spread of rides across each vendors

```
data_vendor1
```

	VendorID	PULocationID	PUBorough	PUZone	PUservice_zone	DOLocationID	DOBorough	DOZone	DOservice_zone	passenger_count	trip
3	1	41	Manhattan	Central Harlem	Boro Zone	238	Manhattan	Upper West Side North	Yellow Zone	1.0	
4	1	41	Manhattan	Central Harlem	Boro Zone	74	Manhattan	East Harlem North	Boro Zone	1.0	
6	1	181	Brooklyn	Park Slope	Boro Zone	45	Manhattan	Chinatown	Yellow Zone	2.0	
10	1	255	Brooklyn	Williamsburg (North Side)	Boro Zone	234	Manhattan	Union Sq	Yellow Zone	2.0	
15	1	195	Brooklyn	Red Hook	Boro Zone	210	Brooklyn	Sheepshead Bay	Boro Zone	1.0	
...	...	...	...	...	...	...	...	...	...	...	...
63823	1	129	Queens	Jackson Heights	Boro Zone	83	Queens	Elmhurst/Maspeth	Boro Zone	1.0	
63840	1	25	Brooklyn	Boerum Hill	Boro Zone	49	Brooklyn	Clinton Hill	Boro Zone	1.0	
63843	1	74	Manhattan	East Harlem North	Boro Zone	24	Manhattan	Bloomingdale	Yellow Zone	1.0	
63844	1	166	Manhattan	Morningside Heights	Boro Zone	244	Manhattan	Washington Heights South	Boro Zone	1.0	
63846	1	181	Brooklyn	Park Slope	Boro Zone	181	Brooklyn	Park Slope	Boro Zone	1.0	

7549 rows × 28 columns

There are 7549 times that people used vendor 1 as mode of transportation

Lets move to the second vendor

	VendorID	PULocationID	PUBorough	PUZone	PUservice_zone	DOLocationID	DOBorough	DOZone	DOservice_zone	passenger_count	tri
0	2	166	Manhattan	Morningside Heights	Boro Zone	143	Manhattan	Lincoln Square West	Yellow Zone	1.0	
1	2	24	Manhattan	Bloomingdale	Yellow Zone	43	Manhattan	Central Park	Yellow Zone	1.0	
5	2	41	Manhattan	Central Harlem	Boro Zone	262	Manhattan	Yorkville East	Yellow Zone	1.0	
7	2	24	Manhattan	Bloomingdale	Yellow Zone	75	Manhattan	East Harlem South	Boro Zone	1.0	
8	2	41	Manhattan	Central Harlem	Boro Zone	166	Manhattan	Morningside Heights	Boro Zone	2.0	
...	...	...	...	...	...	...	...	...	...	...	...
63882	2	130	Queens	Jamaica	Boro Zone	205	Queens	Saint Albans	Boro Zone	1.0	
63883	2	65	Brooklyn	Downtown Brooklyn/MetroTech	Boro Zone	181	Brooklyn	Park Slope	Boro Zone	1.0	
63884	2	244	Manhattan	Washington Heights South	Boro Zone	116	Manhattan	Hamilton Heights	Boro Zone	1.0	
63885	2	74	Manhattan	East Harlem North	Boro Zone	238	Manhattan	Upper West Side North	Yellow Zone	1.0	
63886	2	95	Queens	Forest Hills	Boro Zone	95	Queens	Forest Hills	Boro Zone	1.0	

51610 rows × 28 columns

There are 51610 times that people used vendor 2 as mode of transportation

The measures of central tendencies for Vendor 1 were found to be as follows :

	VendorID	PULocationID	DOLocationID	passenger_count	trip_distance	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	payment_type	trip_type	congestion_surcharge	Dropoff_year	Dropoff_day	Pickup_year	Pickup_day
count	7549.0	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.000000	7549.0	7549.000000	7549.0	7549.000000
mean	1.0	95.064114	139.308783	1.131938	2.585349	14.305999	1.517254	1.490131	1.964285	0.086111	0.992489	19.363820	1.389056	1.003047	0.847331	2023.0	16.408663	2023.0	16.406147
std	0.0	58.193190	78.759343	0.530294	2.188011	8.775822	1.653469	0.107985	2.450913	0.756506	0.086184	10.723966	0.537952	0.055117	1.269805	0.0	8.828597	0.0	8.828571
min	1.0	7.000000	1.000000	0.000000	0.300000	0.010000	0.000000	0.000000	0.000000	0.000000	0.000000	1.010000	1.000000	1.000000	0.000000	2023.0	1.000000	2023.0	1.000000
25%	1.0	74.000000	74.000000	1.000000	1.300000	8.600000	0.000000	1.500000	0.000000	0.000000	1.000000	12.100000	1.000000	1.000000	0.000000	2023.0	9.000000	2023.0	9.000000
50%	1.0	74.000000	141.000000	1.000000	1.900000	12.100000	1.000000	1.500000	1.640000	0.000000	1.000000	16.400000	1.000000	1.000000	0.000000	2023.0	17.000000	2023.0	17.000000
75%	1.0	97.000000	229.000000	1.000000	3.100000	17.000000	2.750000	1.500000	3.200000	0.000000	1.000000	23.500000	2.000000	1.000000	2.750000	2023.0	24.000000	2023.0	24.000000
max	1.0	263.000000	263.000000	6.000000	22.100000	69.500000	7.500000	1.500000	57.000000	14.750000	1.000000	93.000000	4.000000	2.000000	2.750000	2023.0	31.000000	2023.0	31.000000

Vendor 1

•**Trip Distance**: The mean trip distance is approximately 2.58 miles with a standard deviation (std) of 2.18 miles, indicating moderate variability. The maximum trip distance is significantly longer at 22.1 miles.

•**Fare Amount**: The mean fare amount is about $14.305999 with a std of $8.775822, suggesting variable fare charges. The fares range from a minimum of $0.010000 to a maximum of $69.50.(only 1 trip)

•**Tip Amount**: Average tips are $1.964285 with a relatively high std of $2.450913, indicating a wide variation in how much passengers tip. The maximum tip is $57.000000.

•**Tolls Amount**: The average tolls paid are around $0.086111, with most trips (75th percentile) not incurring tolls, but a maximum of $14.750000 suggests some routes involve significant toll charges

count	51609.0	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.000000	51609.0	51609.000000	51609.0	51609.000000
mean	2.0	96.295956	137.838768	1.352206	2.598625	15.556483	0.834273	0.493974	2.093169	0.108778	0.950669	20.800717	1.360732	1.010793	0.763181	2023.0	16.351702	2023.0	16.343738
std	0.0	58.524227	76.014936	1.039737	2.207262	9.424846	1.222279	0.054560	2.687973	0.830738	0.179161	11.537326	0.482512	0.103327	1.231316	0.0	8.758119	0.0	8.757777
min	2.0	3.000000	1.000000	0.000000	0.270000	0.050000	0.000000	0.000000	0.000000	0.000000	0.300000	1.150000	1.000000	1.000000	0.000000	2023.0	1.000000	2023.0	1.000000
25%	2.0	74.000000	74.000000	1.000000	1.220000	9.300000	0.000000	0.500000	0.000000	0.000000	1.000000	12.900000	1.000000	1.000000	0.000000	2023.0	9.000000	2023.0	9.000000
50%	2.0	75.000000	138.000000	1.000000	1.890000	12.800000	0.000000	0.500000	1.740000	0.000000	1.000000	17.600000	1.000000	1.000000	0.000000	2023.0	17.000000	2023.0	17.000000
75%	2.0	112.000000	216.000000	1.000000	3.140000	19.100000	1.000000	0.500000	3.330000	0.000000	1.000000	25.350000	2.000000	1.000000	2.750000	2023.0	24.000000	2023.0	24.000000
max	2.0	263.000000	263.000000	7.000000	28.670000	69.500000	7.500000	0.500000	80.880000	13.100000	1.000000	130.700000	4.000000	2.000000	2.750000	2023.0	31.000000	2023.0	31.000000

Vendor 2

•**Trip Distance:** The mean trip distance is slightly higher at about 2.598590 miles with a std of 2.207256 miles, showing similar variability to Vendor 1. The maximum trip distance is substantially longer at 28.670000 miles.

•**Fare Amount:** The mean fare is approximately $15.556308 with a std of $9.424839, which is higher than Vendor 1, and the range of fares (up to $69.500000) is similar to Vendor 1.But the number of trips which was for 69.5 is much more for vendor 2(12 in count)

•**Tip Amount:** Tips average about $2.097434 with a std of $2.857261, again showing considerable variability. The maximum tip amount is $222.00, significantly higher than that for Vendor 1, indicating potential high-value service instances.Tolls Amount: The tolls average $0.108776 with a std of $0.830730. Like Vendor 1, most trips do not involve tolls, but the maximum tolls paid are $13.1, similar to Vendor 1.

Comparisons and insights

•**Pricing Strategy:** Both vendors have a broad range of fare amounts, but Vendor 2 tends to have slightly higher fares on average and also shows higher maximum values in tips, which could indicate either higher service quality or a customer base that includes higher spending segments.

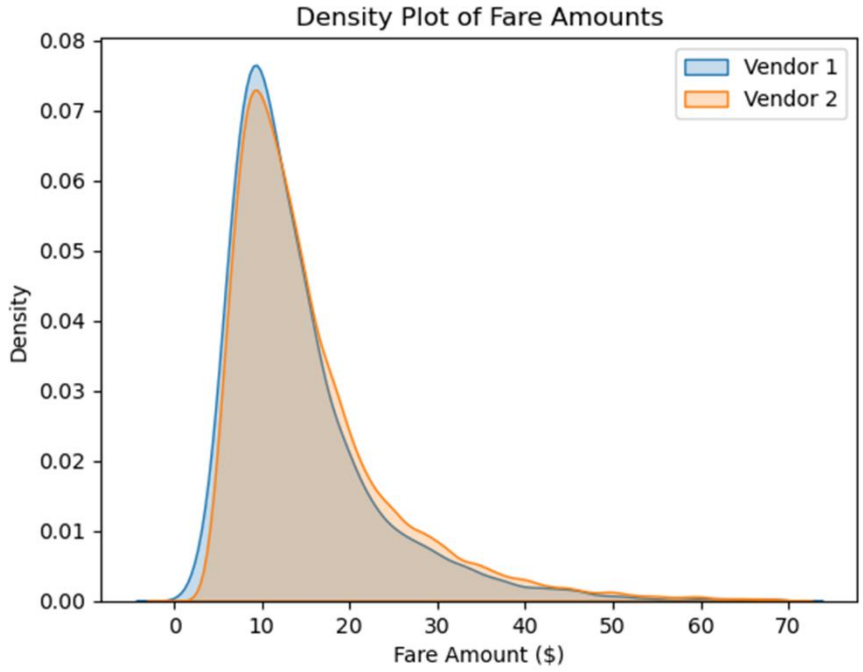
•**Service Variation:** Both vendors show significant variability in fares and tips, suggesting diverse types of trips or customer preferences. Vendor 2, however, has shown to reach significantly higher tips, possibly indicating better customer satisfaction or premium service options that encourage higher tipping.

•**Operational Differences:** The maximum values in trip distance, tips, and tolls are generally higher for Vendor 2, suggesting that Vendor 2 might be covering longer routes or possibly has a different operational focus compared to Vendor 1.

Since we have an idea of how the data is spread and how the variables are related and the distribution they might follow , now we can move on to the Visual analysis of the data and confirm the same

Visual Analysis

Density plot for fare amount



Observations from the Density Plot

- **Peak Comparison:**

Vendor 1 has a sharper and narrower peak compared to Vendor 2. This indicates that most of Vendor 1's fares are concentrated around a specific value, which is slightly above $10.

Vendor 2 has a broader peak that is slightly lower in height, centered around $10 as well, but with a wider spread. This suggests a more varied range of fare amounts.

- **2.Spread of Data:**

1.The spread of the distribution for Vendor 1 is much narrower, indicating less variability in fare amounts. This could mean that Vendor 1's pricing strategy is more consistent or standardized.

2.Vendor 2 shows a wider spread in fare amounts, indicating greater variability. This could suggest that Vendor 2 offers a wider range of services or has a more flexible pricing model that adapts to different trip conditions or customer preferences.

- **Tails of the Distribution:**

1. Vendor 1 shows a quicker decline in density as the fare amount increases beyond the peak, with very few fares approaching $70.
2. Vendor 2, while also declining, does so more gradually, indicating that higher fare amounts are more common compared to Vendor 1. This tail suggests that Vendor 2 occasionally charges significantly higher fares, potentially for longer or more premium services.

Interpretation and Implications

- **Pricing Strategy:**

Vendor 1's strategy appears to be focused on maintaining a consistent fare structure, which could appeal to customers looking for predictability in pricing.

In contrast, Vendor 2 might be targeting a more diverse customer base or offering a range of services (such as luxury or specialized transportation) that justify a higher and more varied pricing structure.

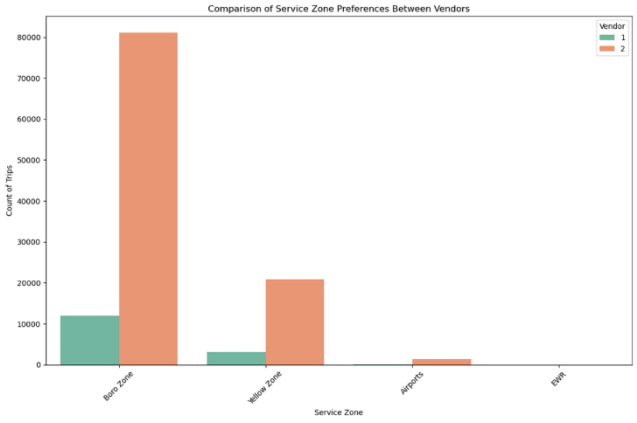
- **Market Positioning:**

The differences in fare distribution could also reflect different market positioning. Vendor 1 might be focusing on short, standard trips within a city, while Vendor 2 could be serving both urban and longer-distance routes, including airport runs, which typically command higher fares.

- **Customer Decision-making:**

Customers who prioritize budget over service diversity might prefer Vendor 1, while those who value service options or have varying trip needs might find Vendor 2 more appealing.

Since we see a wider spread with Vendor 2 , let's confirm if this is due to their service being taken up in more zones . Lets analyze this via a comparison of service zones across vendors



Operational Focus and Strategy:

- Vendor 2 shows a clear strategy of serving a broader geographical area, covering both the densely populated city center (Yellow Zone) and the surrounding boroughs (Boro Zone). This suggests that Vendor 2 targets a diverse customer base, likely offering varied services that cater to both urban and suburban clientele.
- Vendor 1 appears to focus more on the city center but with significantly lesser operations compared to Vendor 2. Their strategy might be more niche or focused on specific types of trips or customer segments within these zones.

Market Segmentation:

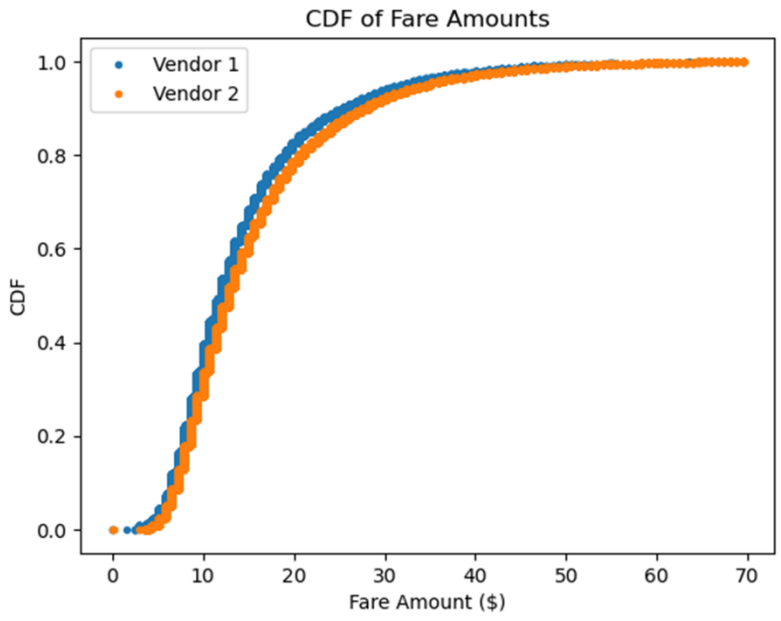
The dominance of Vendor 2 in both the Boro and Yellow zones can indicate a robust operational network that appeals to a wide range of customers, possibly offering competitive rates, more vehicle options, or better availability.

Vendor 1's strategy might involve focusing on premium services or specific times of day, given the relatively lower volume of trips but presence across zones.

Service Adaptation:

Vendor 2's presence in airport zones, although not overwhelming, shows an adaptability to varied travel needs, including airport pickups and drop-offs, which can be part of a service differentiation strategy.

Lets now move to an analysis of fare amount using CDF plot . This will further give us an idea on which fare segment each of the vendor has most of the rides



Observations from the CDF Plot

- **Initial Fare Accumulation:**

1. Both Vendor 1 and Vendor 2 have fare accumulations starting at lower fare values, with similar initial rates of increase. This suggests that both vendors offer trips at lower fare ranges, possibly covering short distances or basic service types.

2. The curves start to diverge slightly as the fare amount increases, which is indicative of differing pricing structures as fares rise.

- **Middle Fare Range:**

1. Around the fare range of $10 to $30, the CDF for Vendor 2 rises more steeply compared to Vendor 1. This indicates that a larger proportion of Vendor 2's fares fall within this middle range compared to Vendor 1.

- **High Fare Range:**

1. Beyond $30, Vendor 1's CDF continues to rise steadily, though at a slower rate, and flattens out as it approaches the higher fare amounts.
2. Vendor 2's CDF overtakes Vendor 1 slightly beyond the $30 mark and continues to accumulate more quickly until it reaches near saturation. This implies that Vendor 2 has a higher proportion of fares in the higher fare range compared to Vendor 1.

- **Distribution Saturation:**

1. Both CDFs approach 1.0 (or 100% of their fares) around the $60 fare amount, but Vendor 2 reaches this saturation point slightly faster. This suggests that Vendor 2 not only has a higher proportion of fares in the higher range but also covers a wider range of higher fare amounts more frequently than Vendor 1.

Implications and Insights

- **Pricing Strategy**

Vendor 2's broader and quicker accumulation in the middle to high fare ranges suggests a more diversified pricing strategy that perhaps caters to a wider variety of trip types or service options, including more premium services that command higher fares.

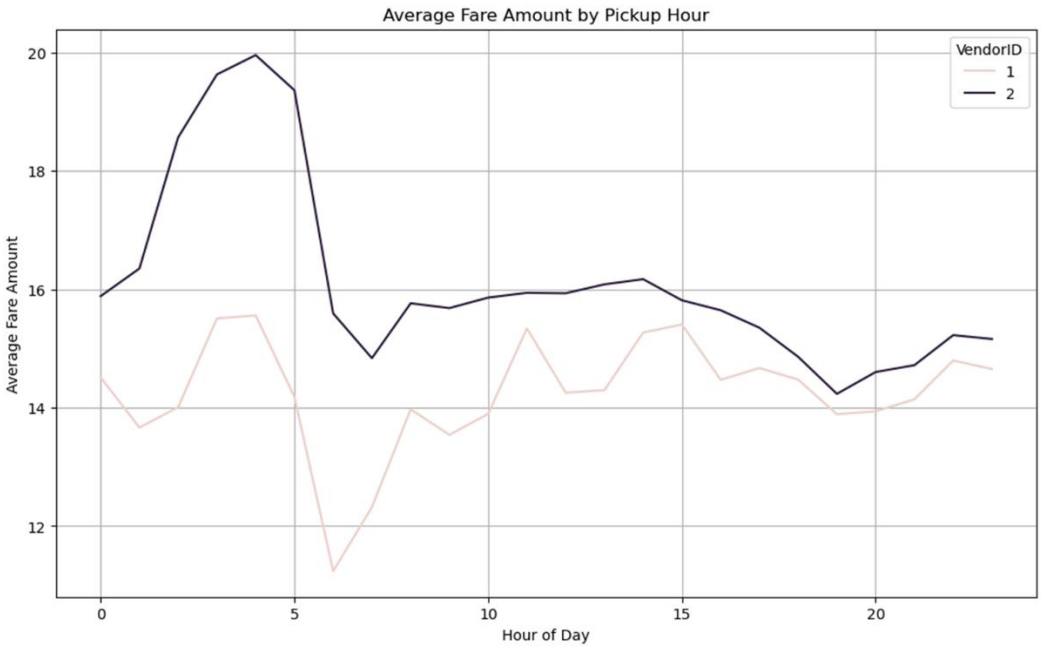
- **Market Positioning**:

Vendor 1 seems to be positioned more conservatively in the fare market, with a smoother and more gradual fare increase, indicating a focus perhaps on standard or budget-conscious market segments.

• Customer Segmentation: The CDF suggests that Vendor 2 might be targeting or appealing to segments of the market that require or are willing to pay for longer

distances or more premium service offerings, while Vendor 1 maintains a steady offering across a narrower fare range.

Now lets analyze what time does each of the vendor charge their customers . This will give us an insight into their pricing strategy



- **Early Morning (Midnight to 6 AM):**

1. Vendor 2 shows a peak around early morning hours, particularly sharp around 5 AM. This indicates high fares during these hours, which could be due to premium charges for rides when fewer drivers are available or when there is significant demand, such as early flights or business commutes.
2. Vendor 1, in contrast, maintains relatively steady and lower fares, suggesting they might offer more consistent pricing, which could be appealing for budget-conscious riders during these hours.

- **Morning to Midday (6 AM to 12 PM):**

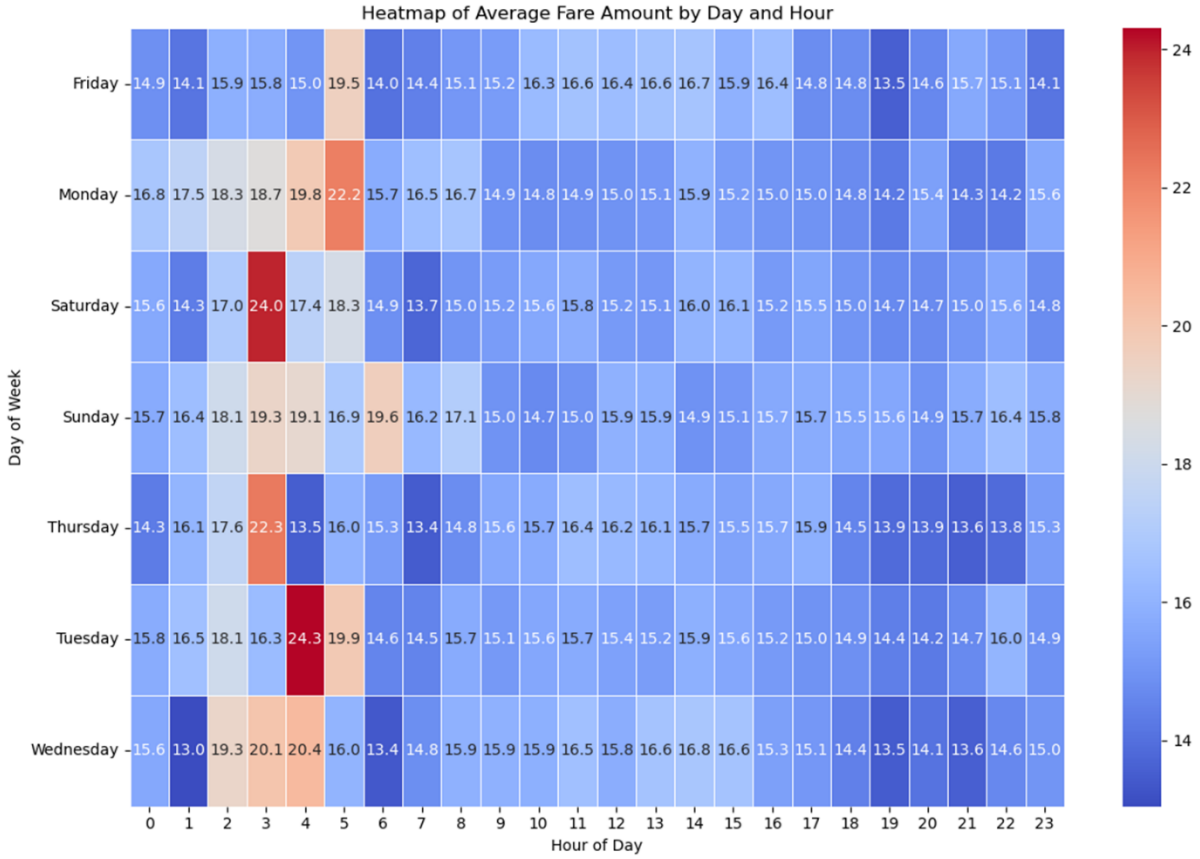
1. Vendor 2 experiences a drop after the early morning peak and then stabilizes with a slight increase around 9 AM, possibly aligning with the morning rush.
2. Vendor 1 also shows some variability but remains consistently below Vendor 2, indicating a more uniform fare structure through the morning.

- **Afternoon to Evening (12 PM to 6 PM):**

1. Both vendors show lower average fares during the early afternoon, which likely reflects a dip in demand post-lunch and pre-evening commute.
2. Vendor 2 starts to increase fares towards the late afternoon, suggesting an anticipation of higher evening demand.

- **Night (6 PM to Midnight):**

1. Vendor 2 continues with higher fares into the evening, possibly capitalizing on nightlife or other evening activities.
2. Vendor 1 maintains a smoother curve with slightly rising fares in the evening but remains lower than Vendor 2, indicating a more consistent pricing approach.



Key Observations from the Heatmap

1.Peak Fare Times:

1. Early Morning Surge: There are noticeable fare increases in the early morning hours, particularly at 5 AM on weekdays like Tuesday and Thursday. This could indicate higher demand during early commutes or possibly fewer drivers available, leading to surge pricing.

2. Late Night Premiums: On days like Saturday, fare increases are also notable around midnight and the early hours (1-3 AM), possibly due to nightlife activities when people are heading home.

2.Weekday Variations:

1. Tuesday and Thursday Mornings: The early morning hours on these days show significantly higher fares compared to other times, with peaks at 5 AM reaching up to around $24, suggesting a consistent demand for early rides or limited service availability.

2. Midday and Afternoon Hours: Weekdays tend to show more uniform fare distributions during midday and afternoon hours, with moderate averages that do not fluctuate as much as during the morning hours.

3.Weekend Patterns:

1. Saturday and Sunday: The weekend shows a different pattern, especially with a peak late on Saturday night/early Sunday morning, which might cater to a weekend crowd returning from social outings.

2. Sunday: Overall, fares on Sunday are somewhat stable, with a slight increase in the evening hours, potentially aligning with people preparing for the upcoming workweek.

Insights and Implications

•**Dynamic Pricing Indicators:** The early morning and late-night surges on specific days suggest dynamic pricing models that adjust fares based on real-time demand and supply. Such models are likely in play, especially evident during commuting hours and nightlife hours.

•**Operational Strategy:**Consistency in Weekday Demands: The consistent peaks during specific weekday hours may reflect commuter patterns, suggesting that operational strategies could be optimized around these peak times to maximize efficiency and profitability.

•**Weekend Nightlife:** The different pattern observed on weekends, especially on Saturday nights, could indicate a strategic shift to accommodate social outing returns, which could involve different routing or increased fleet availability.

```
# Check for Normality

from scipy.stats import shapiro

# Normality test for Vendor 1
stat, p = shapiro(vendor1_fares.sample(min(5000, len(vendor1_fares))))  # sample due to Shapiro test limitations
print('Vendor 1 Fares Normality Test: Statistics=%.3f, p=%.3f' % (stat, p))

# Normality test for Vendor 2
stat, p = shapiro(vendor2_fares.sample(min(5000, len(vendor2_fares))))
print('Vendor 2 Fares Normality Test: Statistics=%.3f, p=%.3f' % (stat, p))
```

```
Vendor 1 Fares Normality Test: Statistics=0.826, p=0.000
Vendor 2 Fares Normality Test: Statistics=0.821, p=0.000
```

The results from the Shapiro-Wilk normality test indicate that both Vendor 1 and Vendor 2's fare distributions are not normally distributed, as the p-values are very small ($p < 0.05$). This means that we should use a non-parametric test to compare the central tendencies of the two distributions.

Krushkal Wallis Test

```
# Krushkal Wallis Test

stat, p = kruskal(vendor1_fares, vendor1_fares, vendor2_fares)

# Print the results
print(f"Kruskal-Wallis H Test: Statistics={stat:.3f}, p={p:.3f}")

# Interpretation
alpha = 0.05
if p < alpha:
    print('There is a statistically significant difference between the groups.')
else:
    print('There is no statistically significant difference between the groups.')
```

```
Kruskal-Wallis H Test: Statistics=271.784, p=0.000
There is a statistically significant difference between the groups.
```

From this we can reject our null hypothesis and confirm that there is a significant difference between the groups

We can confirm this by running the model on the vendors separately by training them on each vendor dataset, confirm and also compare the prices.

3. **Hypothesis Testing for Total Amount by Trip Distance and Days of the Week**

This test is appropriate as we aim to compare the median trip distances across different days of the week, assuming these categories represent independent groups within our dataset. The Kruskal Wallis test, a non-parametric test, is used due to its ability to

handle data that may not follow a normal distribution and because it compares medians rather than means.

Hypothesis:

- **Null hypothesis (H0)** : There is no significant relationship between total amount , trip distance and day of the week.
- **Alternative Hypothesis (H1)** : There is a significant relationship between total amount , trip distance and day of the week.

Significance Level (α): 0.05. This represents a 5% risk of rejecting the null hypothesis when it is actually true.The test statistic *H* is calculated using the formula:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total sample size, k is the number of groups we are comparing, R_i is the sum of ranks for group i, and n_i is the sample size of group i.

Kruskal-Wallis Test Results:

- **Number of Categories**: 7 (representing the different trip distance categories).
- **Degrees of Freedom (DOF)**: Number of Groups = 8 - 1 = 7.
- **Kruskal-Wallis H Statistic and p-value** :

Day	Kruskal Wallis H Statistic	P - Value
Monday	10133.637621807226	0.0
Tuesday	12053.7868967797	0.0
Wednesday	13387.789836390742	0.0
Thursday	11954.968293776332	0.0
Friday	12251.915993879173	0.0
Saturday	12012.869687779823	0.0
Sunday	10409.878148190419	0.0

Interpretation:

- Given that the p-value (0) is significantly lower than the significance level (0.05), we reject the null hypothesis.

Conclusion: Based on the dataset used in this study, there is significant evidence to suggest that there is a significant relationship between total amount, trip distance and day of the week.

ANOVA Analysis :

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Monday	58978	7141	0.121079046	0.106420715
Tuesday	58978	8653	0.146715725	0.125192343
Wednesday	58978	9789	0.165977144	0.138431079
Thursday	58978	8588	0.145613619	0.124412402
Friday	58978	8821	0.149564244	0.127196938
Saturday	58978	8620	0.146156194	0.124796677
Sunday	58978	7366	0.124894028	0.109297363
trip_distance	58978	153044.62	2.594944216	4.851523497
total_amount	58978	1215405.47	20.60777697	130.7967686

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	21613828.56	8	2701728.57	178130.6781	0	1.938431492
Within Groups	8050598.742	530793	15.16711551			
Total	29664427.3	530801				

Conclusion of ANOVA Analysis:

The ANOVA conducted to test for differences in the total amount against trip distance and days of the week revealed an F-statistic of 178130.3752, indicating a significant amount of variance between the group means relative to the variance within each group.

Since p-value < significance level i.e., 0 < 0.05, we reject the null hypothesis. This conclusion is supported by the extremely low p-value. Hence, there is a significant relationship between the total amount with trip distance or day of the week.

7. MODEL ANALYSIS

1. Linear Regression

In our project on New York taxi fare predictions, we conducted a simple linear regression to explore the relationship between trip_distance and total_amount. This analysis, based on over 59,000 observations, revealed a strong positive correlation, with an R-squared value of 0.8132, indicating that approximately 81.32% of the variance in taxi fares can be explained by trip distance alone. The regression model demonstrated significant predictive power, as reflected by an extremely high F-statistic. Each additional unit increase in trip distance is associated with an increase of approximately $4.94 in fare, showcasing trip distance as a reliable predictor of fare amount. These findings provide a quantitative foundation for fare estimation tools and strategic pricing decisions in the taxi industry.

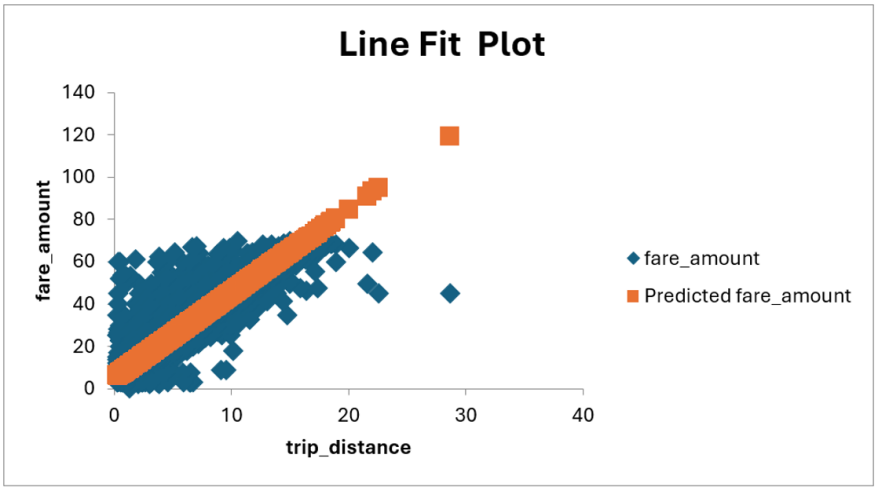
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.901796047
R Square	0.81323611
Adjusted R Square	0.813232943
Standard Error	4.942522388
Observations	58978

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6273304.178	6273304.178	256802.3867	0
Residual	58976	1440696.841	24.42852755		
Total	58977	7714001.019			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8.457231049	0.031449922	268.9110368	0	8.39558907	8.518873028	8.39558907	8.518873028
trip_distance	4.682391953	0.009239921	506.7567332	0	4.66428167	4.700502236	4.66428167	4.700502236



The line fit plot shows the actual fare amounts (fare_amount) versus the predicted fare amounts (Predicted fare_amount) based on the model. The plot illustrates a clear linear

trend, with the predicted values closely aligning with the actual values as the trip distance increases. The orange square outliers indicate instances where the model's predictions deviate significantly from the actual data, suggesting potential anomalies or exceptional cases.

$$\textbf{\textit{Predicted Fare}} = \beta 0 + \beta 1 \times trip\_distance$$

where $\beta 0$ is the intercept, and $\beta 1$ is the slope of the regression line. These parameters are estimated using the least squares method, which minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Conclusion:

The high R-squared value coupled with the F-statistic near zero provides strong evidence that trip distance is an excellent predictor of fare amount.

Effect of Distance on Fare: The positive coefficient for trip distance confirms that as the distance of the trip increases, so does the fare amount linearly.

Model Accuracy: The standard error of the estimate indicates that the typical prediction error is about $3.29, which, depending on the fare scale, might be considered acceptable for practical purposes.

2. Multiple Regression

We utilized a multiple linear regression model to predict total fares based on a variety of factors, including trip distance, passenger count, day of the week, and pickup or drop-off boroughs.

SUMMARY OUTPUT

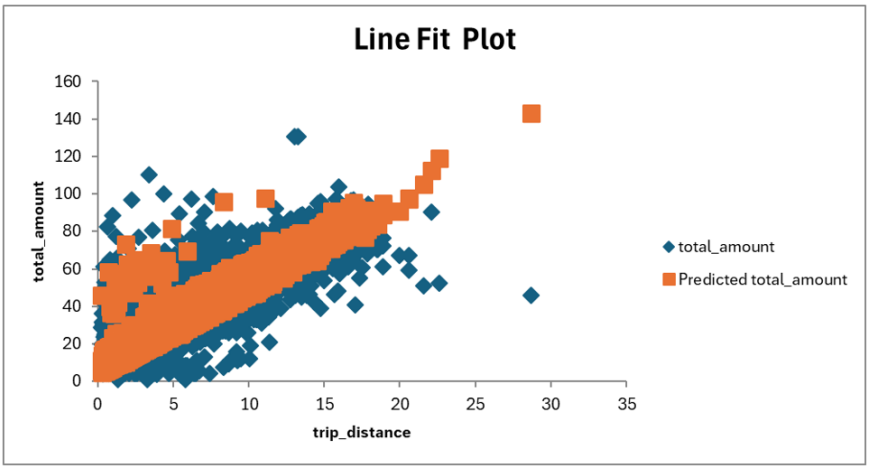
Regression Statistics	
Multiple R	0.936436718
R Square	0.876913727
Adjusted R Square	0.87686963
Standard Error	4.012833937
Observations	58978

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	6764513.385	483179.5275	32314.00835	0
Residual	58964	949487.6338	16.1028362		
Total	58978	7714001.019			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.092485274	2.84430306	0.735675921	0.461930873	-3.482360722	7.667331269	-3.482360722	7.667331269
passenger_count	0.053697484	0.016717253	3.212099777	0.001318393	0.020931598	0.08646337	0.020931598	0.08646337
trip_distance	4.034495992	0.009274436	435.012553	0	4.016318058	4.052673925	4.016318058	4.052673925
extra	0.991453154	0.013312331	74.47630104	0	0.96536093	1.017545379	0.96536093	1.017545379
mta_tax	-2.543528923	0.0508871	-49.98376623	0	-2.643267854	-2.443789992	-2.643267854	-2.443789992
tolls_amount	1.113468539	0.021625917	51.48769244	0	1.071081651	1.155855426	1.071081651	1.155855426
improvement_surcharge	3.584682214	0.097910185	36.611944	3.3916E-290	3.392777839	3.77658659	3.392777839	3.77658659
congestion_surcharge	1.718965793	0.016458684	104.4412663	0	1.686706703	1.751224883	1.686706703	1.751224883
ride_duration_minutes	0.075975064	0.001261698	60.21650939	0	0.07350213	0.078447998	0.07350213	0.078447998
DOBorough_Bronx	1.409258932	2.84331328	0.495639697	0.620150487	-4.16364709	6.982164954	-4.16364709	6.982164954
DOBorough_Brooklyn	4.666731689	2.842445512	1.641801635	0.100636453	-0.904473503	10.23793688	-0.904473503	10.23793688
DOBorough_EWR	0	0	65535	#NUM!	0	0	0	0
DOBorough_Manhattan	2.583982178	2.84217414	0.909156881	#NUM!	-2.986691125	8.15465548	-2.986691125	8.15465548
DOBorough_Queens	2.517865352	2.842011741	0.88594474	0.375650919	-3.052489647	8.088220351	-3.052489647	8.088220351
DOBorough_Staten Island	-0.560140801	3.139366493	-0.178424788	0.858389997	-6.713312369	5.593030766	-6.713312369	5.593030766

The accompanying line fit plot visually confirms the model's efficacy, showing a strong correlation between the predicted and actual total amounts as a function of trip distance.



Predicted Fare$=5.6257-0.2377\times Manhattan-1.1663\times Bronx+2.0669\times Brooklyn-0.1605\times Queens+0.0613\times Passenger\ Count+4.2275\times Trip\ Distance+0.9691\times Extra\ Charges-2.5960\times MTA\ Tax+1.0772\times Tolls\ Amount+3.4748\times Improvement\ Surcharge+1.7979\times Congestion\ Surcharge-0.000254\times Total\ Minutes$

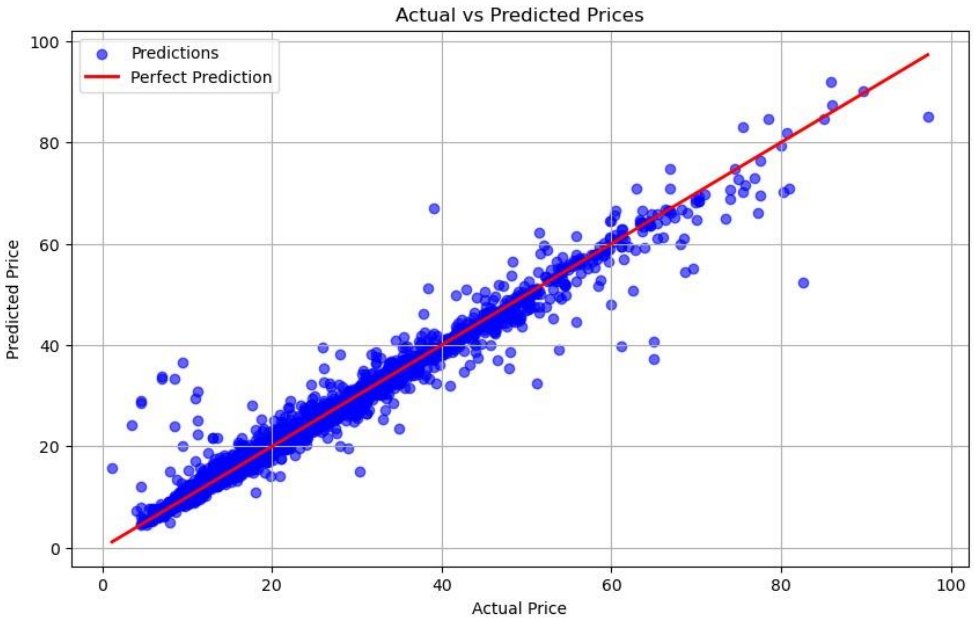
This formula can be utilized to estimate the fare for a taxi ride given specific values for each of the variables listed. It is essential to input accurate values for each variable, especially noting that borough variables are treated as binary indicators (0 or 1), depending on the location of the taxi ride. This model provides a comprehensive tool for analyzing fare structures and making informed decisions related to taxi service operations.

3. Random Forest

- The dataset was split into train (90%) and test (10%) sets.
- The target variable was the total fare, while all other features were treated as predictors.
- One-hot encoding was applied to categorical columns like PUBorough and DOBorough, which contain locations like Manhattan, Brooklyn, Queens, Bronx, and Staten Island.

```
The prediction metric using Random Forest are :
Mean Squared Error (MSE): 3.5895829185686314
Root Mean Squared Error (RMSE): 1.894619465372567
R2 Score: 0.9738761847060177
```

Metrics Interpretation:
MSE (Mean Squared Error): 3.59, which reflects the average squared difference between predicted and actual values.
RMSE (Root Mean Squared Error): 1.89, representing the standard deviation of prediction errors.
R^2 Score: 0.97, indicating that the model explains 97% of the variance in the data. This is a high R^2 score, suggesting a strong fit.



This plot compares the predicted prices (vertical axis) to the actual prices (horizontal axis). Here's what the chart suggests:

Red Line: Represents the ideal scenario where predicted prices perfectly match the actual prices. Any point lying on this line indicates a perfect prediction.
Blue Dots: Represent individual fare predictions from your model. They generally cluster around the red line, implying a high correlation between predicted and actual prices.
Observations:
The majority of points lie close to the line, suggesting the model is generally accurate in its predictions.
There are some outliers (points that deviate significantly from the red line), indicating instances where the model's predictions differ notably from the actual prices.
Overall Interpretation: The plot indicates a high degree of accuracy in the predictions, as reflected by the close alignment of the majority of points to the line. However, there are some cases where the model's predictions were not as accurate, as evidenced by the scattered outliers.

8. CONCLUSION

In the course of this analysis, we have extensively examined the determinants of taxi fares within the New York City metropolitan area. Our analysis has elucidated that trip distance exhibits a significant correlation with fare cost, conforming to the economic principle of distance-based pricing. Temporal variables also emerged as a noteworthy influence, with fare amounts displaying variation across different days of the week and hours of the day, suggesting a dynamic pricing model responsive to demand fluctuations inherent in urban transit patterns.

The current study's conclusions emphasize the intricate relationship between journey distance, duration and fare, underscored by the moderating effects of temporal factors. The data indicates that the temporal dimension, inclusive of the hour and weekday, introduces variability in pricing, possibly attributable to surge pricing mechanisms during peak traffic periods, and differing demand on weekdays versus weekends.

The hypotheses tested across various scenarios led to the rejection of the null hypotheses (H0) in all cases. Here's a summary of the conclusions drawn:

Relationship between Total Amount, Trip Distance, and Day of the Week:

Conclusion: There is a significant relationship between the total fare, the distance traveled, and the day of the week. This means fare pricing can vary significantly based on the day and distance, perhaps due to differences in demand.

Average Fares Between Vendors:

Conclusion: A significant difference exists in the average fares between Vendor 1 and Vendor 2. This suggests that the two vendors employ different pricing strategies or have variations in services offered that lead to different fare structures.

Median Tips Across Different Trip Distances:

Conclusion: The median tip amounts differ significantly across various trip distance categories. Passengers tend to tip differently based on the trip length, possibly reflecting customer satisfaction, trip complexity, or expectations.

Relationship Between Passenger Count and Total Fare:

Conclusion: The relationship between passenger count and total fare varies significantly across different pickup zones. This implies that the impact of passenger count on fares is not uniform and could depend on the zone's characteristics or specific market factors.

Future Scope:

1. Vendor-Specific Models:

Develop separate machine learning models for each vendor to capture unique pricing patterns and operational differences more effectively.

This vendor-specific modeling will provide tailored predictions that can reflect each vendor's unique pricing structures and fare distribution patterns, leading to more accurate fare predictions.

2. Dynamic Pricing Analysis:

- Train models to simulate dynamic pricing strategies, capturing the impact of peak times, demand surges, and specific events on fare structures.

- Incorporate real-time data to refine dynamic pricing models, enabling vendors to adjust fares dynamically based on current market conditions.

3. Dataset Expansion:

- Expand the dataset to include more data points and extend the analysis to multiple years to identify trends and seasonal variations.

- This expansion will enhance model robustness and ensure that predictions account for long-term patterns and changes in the market, thereby improving predictive accuracy.

4. Handling Missing Data:

- Use sophisticated imputation techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to address missing data for categorical variables and enhance model performance.

- Apply mean/mode imputation for numerical features to fill in gaps, ensuring the model has access to as much information as possible.

- Implement these techniques will ensure that missing values don't adversely affect the model, leading to improvements in performance metrics like RMSE and MSE.