Final Report: NYC Yellow Taxi Fare Analysis

By Aditya Pendyala, Sukaina Alkhalidy, Michael Cherry, Josh Elvy, Alexander Proefke, and Max Rehandorf

Background

Our dataset deals with taxi fares in New York in 2023. Our dataset has over 33 million observations and 19 features. Our main goal is to analyze the effects of the features on the fare of the taxi ride itself. However, through our analysis, we are also investigating the effect of the features on each other and how those interactions combine to influence the fare.

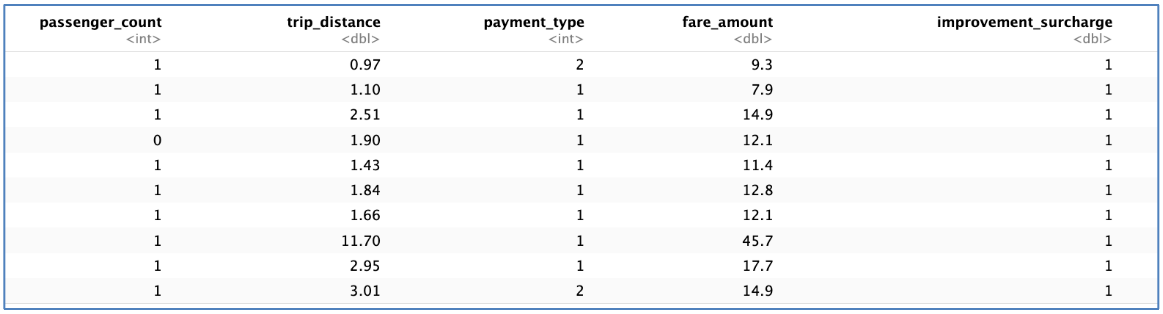
About the Dataset

Dataset Overview: The 2023 NYC Yellow Taxi Trip dataset contains trip-level details of yellow taxi rides, including pick up/drop-off locations, timestamps, fares, and trip distances.

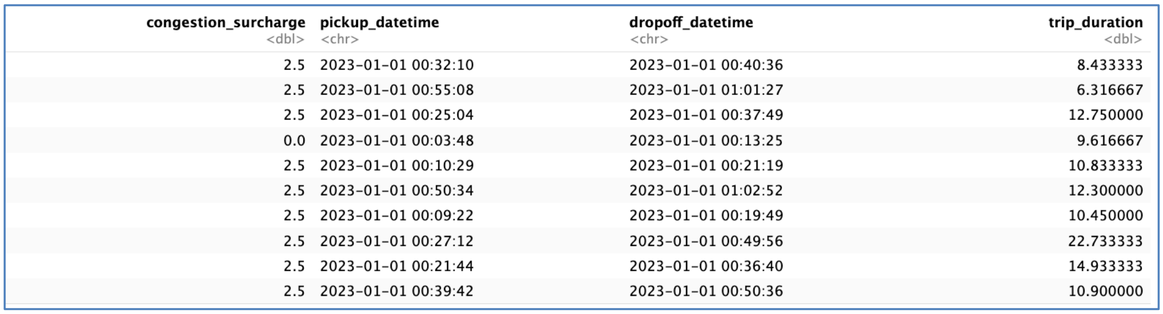
Data Source & Size: Provided by the NYC Taxi & Limousine Commission (TLC), this dataset includes millions of taxi trips recorded throughout 2023.

After preprocessing, the cleaned dataset has 6.9 million rows, with the following filters applied:

- trip distance between 0 and 100 miles.
- fare amount between 0 and $1000.
- trip duration between 0 and 1440 minutes.



passenger_count	trip_distance	payment_type	fare_amount	improvement_surcharge
<int>	<dbl>	<int>	<dbl>	<dbl>
1	0.97	2	9.3	1
1	1.10	1	7.9	1
1	2.51	1	14.9	1
0	1.90	1	12.1	1
1	1.43	1	11.4	1
1	1.84	1	12.8	1
1	1.66	1	12.1	1
1	11.70	1	45.7	1
1	2.95	1	17.7	1
1	3.01	2	14.9	1



congestion_surcharge	pickup_datetime	dropoff_datetime	trip_duration
<dbl>	<chr>	<chr>	<dbl>
2.5	2023-01-01 00:32:10	2023-01-01 00:40:36	8.433333
2.5	2023-01-01 00:55:08	2023-01-01 01:01:27	6.316667
2.5	2023-01-01 00:25:04	2023-01-01 00:37:49	12.750000
0.0	2023-01-01 00:03:48	2023-01-01 00:13:25	9.616667
2.5	2023-01-01 00:10:29	2023-01-01 00:21:19	10.833333
2.5	2023-01-01 00:50:34	2023-01-01 01:02:52	12.300000
2.5	2023-01-01 00:09:22	2023-01-01 00:19:49	10.450000
2.5	2023-01-01 00:27:12	2023-01-01 00:49:56	22.733333
2.5	2023-01-01 00:21:44	2023-01-01 00:36:40	14.933333
2.5	2023-01-01 00:39:42	2023-01-01 00:50:36	10.900000

Objective

This project explores four key questions about fare pricing in New York City's yellow taxi system:

1. What variables significantly affect the fare amount?

2. Does the average fare differ by day of the week?

3. Do night-time trips cost more than day-time trips?

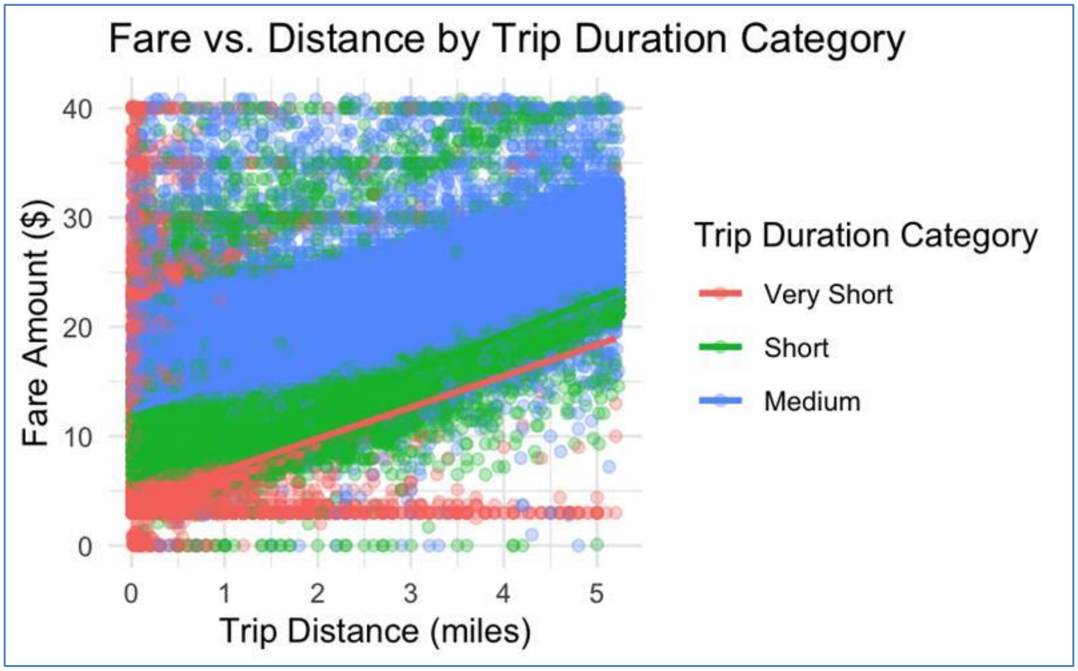
4. How do fare patterns vary by both day type (weekday vs weekend) and time of day (day vs night)?

These questions address fare transparency, behavioral trends, and system-wide patterns. The findings can help inform predictive pricing tools, transportation policy, and urban planning.

Question 1: What affects the fare amount?

We use a multiple linear regression model to explore whether variables like trip distance, duration, and passenger count significantly influence fare.

Plot: Fare vs. Trip Distance



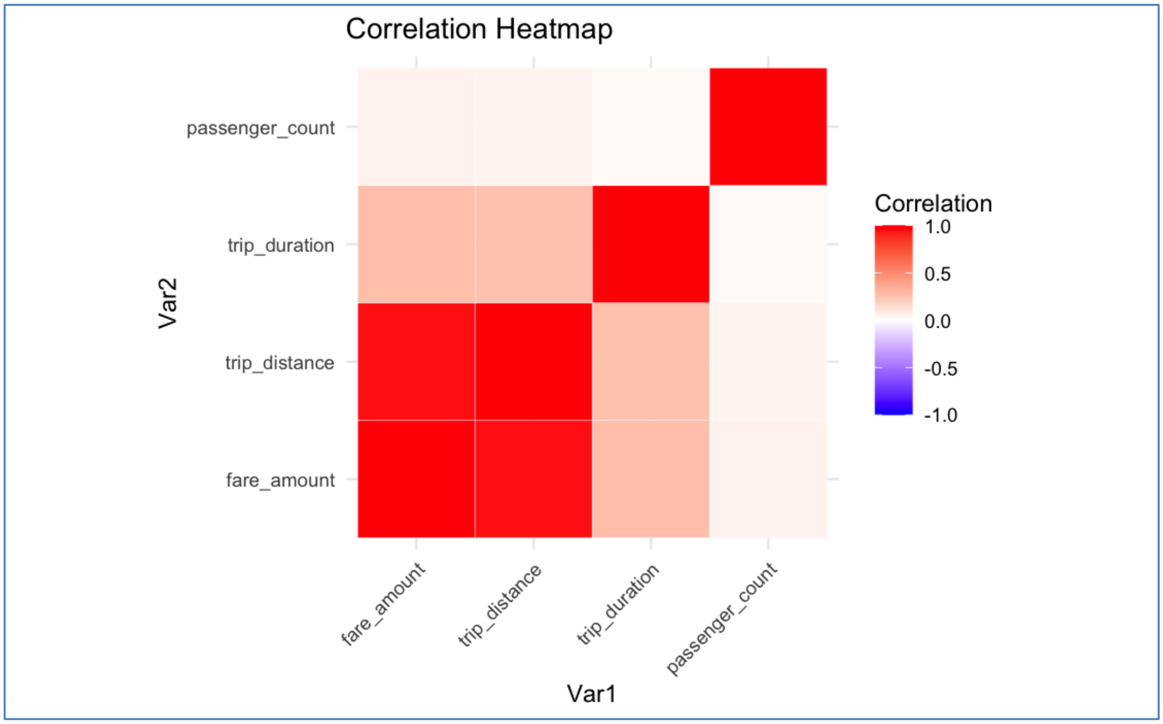
Regression Equation

The most significant variable was `trip_distance`.

Estimated Fare Amount = 6.331 + 3.685 trip distance + 1.393 trip duration + 0.159 passenger count

Where trip distance is expected to be the dominant predictor.

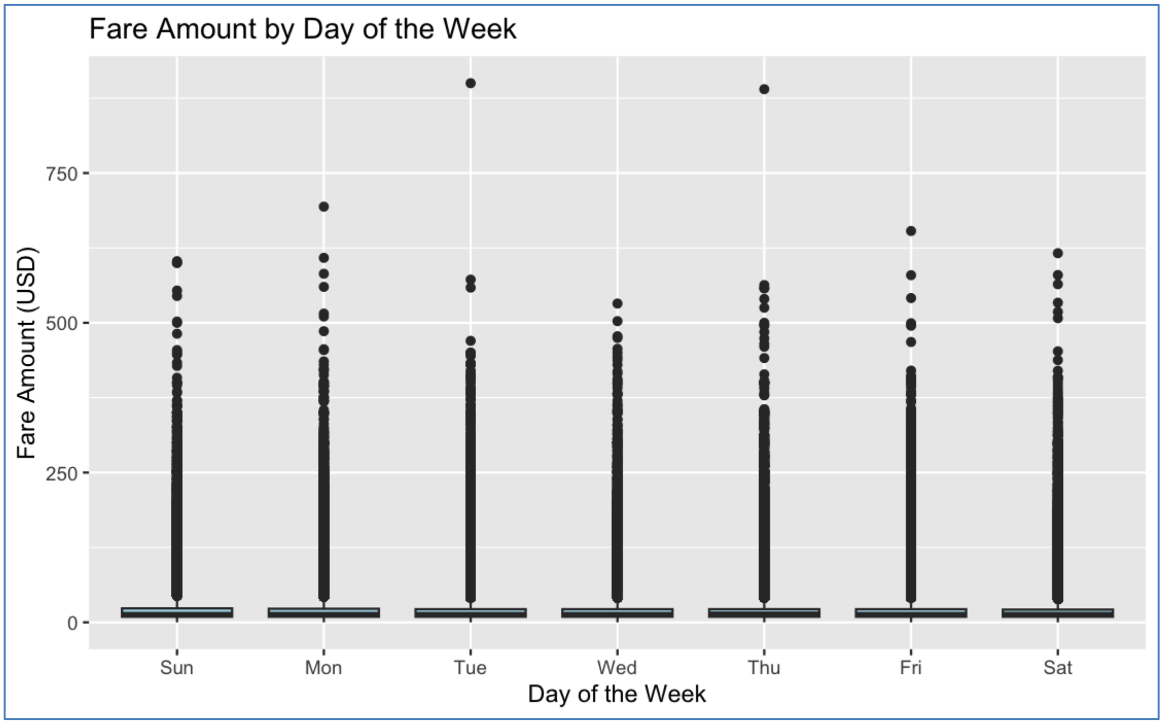
Correlation Heatmap



Question 2: Does fare amount differ by day?

We use ANOVA to test if average fare amount differs across weekdays.

Plot: Fare Amount by Day of the Week



ANOVA and Post-hoc Test

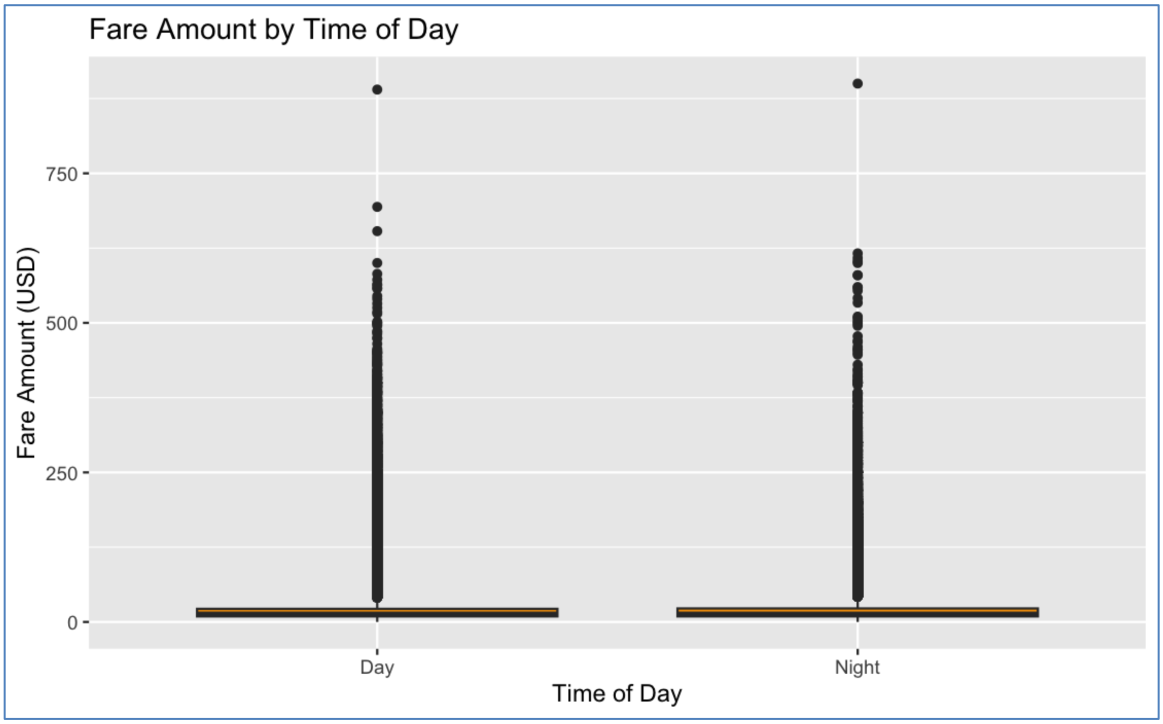
Interpretation

We conducted a one-way ANOVA to examine whether taxi fares differ by day of the week. The results showed a highly significant difference in average fares across weekdays (F = 1098, p < 0.0001). A Tukey HSD post-hoc test revealed that Saturday fares were significantly higher than all other days, with an average difference of $1.85 compared to Sunday. In contrast, Tuesday and Wednesday had the lowest average fares, with no significant difference between them. These patterns likely reflect fluctuations in demand, such as increased weekend activity and lower mid-week travel.

Question 3: Do Night-Time Trips Cost More?

We compare night-time (8 PM–6 AM) vs daytime (6 AM–8 PM) trips using a two-sample t-test.

Plot: Fare Amount by Time of Day



Two-Sample t-Test

Interpretation

We used a Welch two-sample t-test to compare taxi fare amounts between daytime (6 AM–8 PM) and night-time (8 PM–6 AM) trips. The test showed a statistically significant difference (t = 4.09, p < 0.001), with daytime fares slightly higher on average. However, the difference in means was small (approximately $0.06), suggesting that while time of day affects fare prices, the practical impact is minimal.

Question 4: Weekday vs Weekend (Day vs Night)

We examine how fare amounts vary across combinations of day type (Weekday vs Weekend) and time of day (Day: 6 AM–8 PM, Night: 8 PM–6 AM) using a two-way ANOVA**.**

Two-Way ANOVA and Tukey test

We ran a two-way ANOVA to assess:

- Main effect of day type
- Main effect of time of day
- Interaction effect between the two

All effects were statistically significant ($p < 0.001$). The Tukey post-hoc test revealed that weekend night fares were significantly lower than all other groups.

Interpretation

The two-way ANOVA showed strong evidence that fare amounts vary not just by day or time alone, but by their combination. While weekday and daytime fares were slightly higher on average, weekend night fares were the lowest across all groups — up to $0.74 lower than weekend day fares. This contradicts the assumption that night-time or weekend travel is always more expensive, suggesting complex rider behavior and possible pricing adjustments.

Conclusion and Future Work

We found that fare amounts are most influenced by trip distance, followed by moderate effects of time and weekday. The most important takeaway is the interaction effect between weekend/weekday and time of day — especially the unexpectedly low fares during weekend nights.

Future Directions

- Incorporate geospatial data (pickup/drop-off zones)
- Include weather or traffic conditions
- Build a machine learning model for real-time fare prediction
- Compare results across years or boroughs