Exploratory Data Analysis

Marisa Thomas, Keaton Magnuson, Emily Szolnoki, Klodje Toure

For this project, the original dataset contained approximately 200,000 rows and 8 columns. We removed rows with missing dropoff coordinates, filtered out trips with invalid or extreme fares, and excluded trips where pickup and dropoff locations were identical. We also filtered for trips within a typical NYC bounding box to remove geographic outliers. We created three new features: *trip_distance* (using the Haversine formula), *pickup_hour*, and *pickup_dayofweek.* The cleaned dataset now contains 193,515 records and 10 variables. These 10 variables are shown below:

Column Name	Description	Type	Notes
fare_amount	Cost of the trip in USD	float	Target variable
pickup_datetime	Timestamp of the pickup	datetime	Used to extract hour
pickup_longitude	Longitude at trip start	float	Validated between -180 and 180
pickup_latitude	Latitude at trip start	float	Validated between -90 and 90
dropoff_longitude	Longitude at trip end	float	Validated between -180 and 180
dropoff_latitude	Latitude at trip end	float	Validated between -90 and 90
passenger_count	Number of passengers	int	Mainly between 1-6
trip_distance	Distance in between pickup and drop off	float	Calculated with Haversine Formula
pickup_hour	Hour of day (0-23) when trip started	int	From the timestamp

pickup_dayofweek	Day of pickup	string	E.g., Monday, Friday

To better understand the individual characteristics of each variable, we conducted univariate analyses on key fields including *fare_amount, trip_distance, pickup_hour,* and *pickup_dayofweek.*

We began by importing the *uber_cleaned.csv* file into a pandas DataFrame and confirmed the correct data types for all relevant columns. Once verified, we visualized the distribution of *fare_amount* and *trip_distance* using histograms with overlaid Kernel Density Estimate (KDE) curves. These plots highlighted a clear right skew in both variables, revealing that the majority of Uber rides are relatively short and inexpensive. Specifically, most trips cost less than $20 and cover fewer than 5 kilometers in distance.

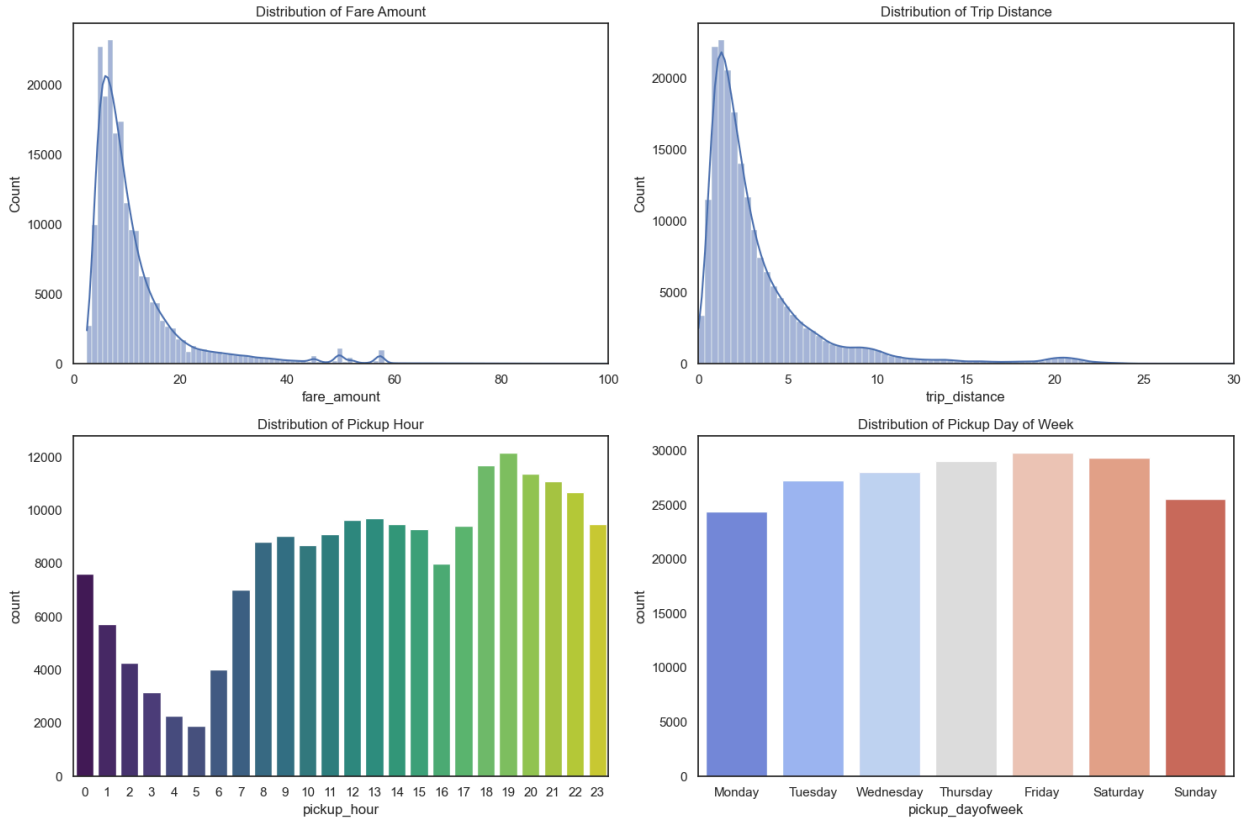

Figure 2.1 Univariate Distributions of Key Variables

Top left: Fare Amount, Top right: Trip Distance, Bottom left: Pickup Hour, Bottom right: Pickup Day of Week

These plots highlight skewed distributions in fare and distance, peak ride times in the evening, and higher activity on weekends.

To evaluate temporal ride patterns, we created bar plots for *pickup_hour* and *pickup_dayofweek* using countplot functions. As shown in Figure 2.1, the distribution of *pickup_hour* indicated pearl Uber usage between 7:00 pm and 10:00 pm, while *pickup_dayofweek* showed increased ride volume on Fridays and Saturdays. These insights suggest a strong correlation with leisure related activity during evenings and weekends.

```
IQR-Based Outlier Detection Summary:

Outlier Analysis for Fare Amount:   Outlier Analysis for Trip Distance:
25th Percentile (Q1): 6.00          25th Percentile (Q1): 1.28
75th Percentile (Q3): 12.50         75th Percentile (Q3): 3.93
Interquartile Range (IQR): 6.50     Interquartile Range (IQR): 2.65
Lower Threshold: -3.75              Lower Threshold: -2.70
Upper Threshold: 22.25              Upper Threshold: 7.91
Total Outliers Detected: 16411      Total Outliers Detected: 16029
Outlier Proportion: 8.48%           Outlier Proportion: 8.28%
```

Figure 2.2: IQR-Based Outlier Summary for Fare and Distance
This table summarizes the IQR method used to identify outliers in fare amounts and trip distances. Approximately 8.48% of fares and 8.28% of distances fall outside expected ranges.

```
pickup_dayofweek
Friday       29845
Saturday     29320
Thursday     29053
Wednesday    28072
Tuesday      27253
Sunday       25575
Monday       24397
Name: count, dtype: int64
```

Figure 2.4: Frequency of Uber Pickups by Day of Week
Friday and Saturday had the highest ride counts, with Monday being the least active. This aligns with typical leisure-related ride behavior

We also conducted an outlier analysis on *fare_amount* and *trip_distance* using the Interquartile Range (IQR) method. This identified approximately 8.48% of fare amounts and 8.28% of trip

distances as statistical outliers. These may represent legitimate long distance travel or pricing anomalies due to surge conditions or data irregularities.

Finally, summary statistics were generated using *.describe()* for numeric variables and *.value_counts()* for categorical ones like *pickup_dayofweek.* This provided additional context for interpreting the visualizations and confirmed our observations about variable distributions.

	fare_amount	passenger_count	trip_distance	pickup_hour
count	193515.000000	193515.000000	193515.000000	193515.000000
mean	11.285889	1.684753	3.342293	13.487766
std	9.323122	1.388372	3.561238	6.515784
min	2.500000	0.000000	0.000000	0.000000
25%	6.000000	1.000000	1.280000	9.000000
50%	8.500000	1.000000	2.180000	14.000000
75%	12.500000	2.000000	3.930000	19.000000
max	100.000000	208.000000	36.690000	23.000000

Figure 2.3: Summary Statistics for Key Numeric Variables
*This table provides descriptive statistics for **fare_amount, trip_distance, passenger_count**, and **pickup_hour.***

Based on the visualizations, outlier detection, and summary statistics, we observed several key trends. Most Uber trips are short and inexpensive, with the majority of fares falling under $20. This is reflected in the right-skewed distributions of both *fare_amount* and *trip_distance.* Peak activity consistently occurs during evening hours (7:00 PM to 10:00 PM) and on weekends, particularly Fridays and Saturdays. This suggests a higher demand during leisure and social hours. We also identified significant outliers, including fares over $100 and trips exceeding 30 miles. These may result from long distance travel, surge pricing, fare miscalculations, or potential data anomalies. These insights can help Uber optimize driver deployment, refine dynamic pricing strategies, and investigate outliers to uncover niche demand, improve service quality, and address data integrity issues.

With a strong understanding of each variable's individual distribution, we next turned our attention to examining how these variables interact with one another. Next, we explored both numerical and categorical relationships to determine which features most directly influence fare amount. The bivariate and multivariate analysis not only helps confirm patterns observed earlier, but also prepares us for model development by identifying key predictors and potential data challenges.

As part of our exploratory data analysis, we wanted to focus on understanding which variables had the most influence on Uber fare prices. In order to conduct this, we explored numerical and categorical variables using multiple visualizations to truly understand the relationships. Each chart was carefully constructed in order to help us identify these trends and to test our hypotheses, as well as prepare us for future modeling.

The first visualization we created was a scatter plot comparing *trip_distance* to *fare_amount*. Since *fare_amount* is our target variable, we wanted to analyze which other variables had the most impact on that target. This scatter plot helped us test the assumption that longer trips result in higher fares. We can observe that there seems to be a strong positive correlation between the two; as trip distance increases, the fare amount also increases. We also noticed some interesting patterns.

We also noticed some horizontal lines of data points, indicating that the same fare amount was charged regardless of the trip distance. These likely represent fixed-rate or capped fares, such as airport rides. This suggests that not all fares are strictly linear when it comes to trip distance, but it is a key insight that is important to understand when we move onto building our predictive models.

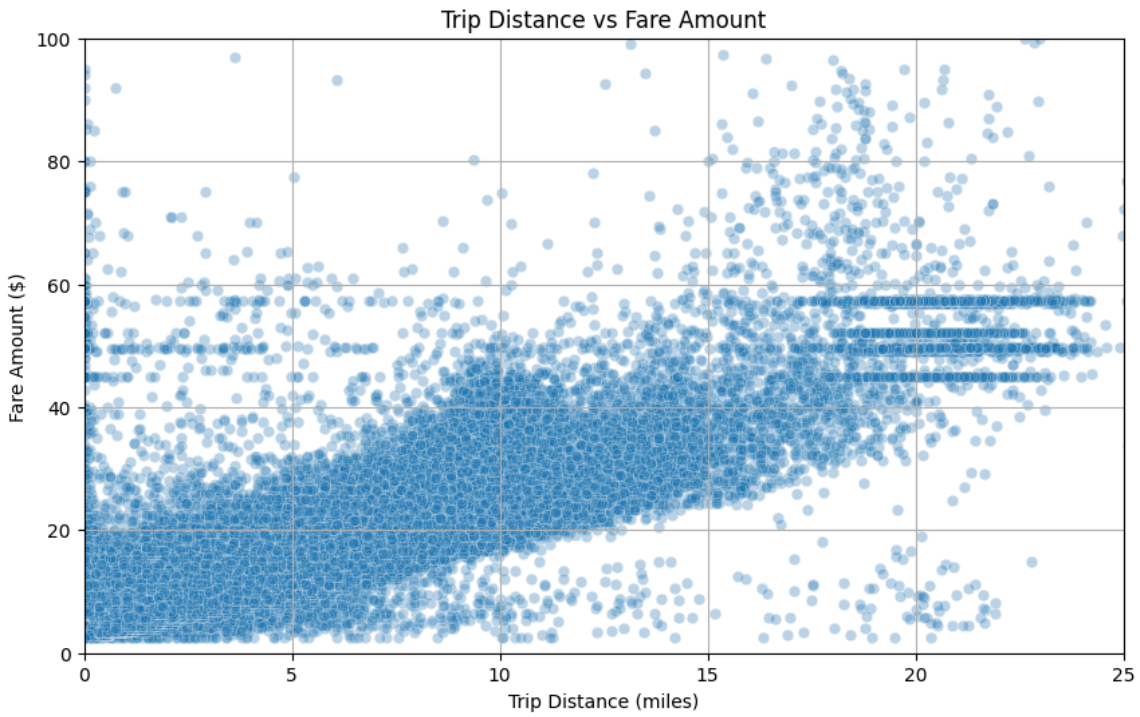

Figure 3.1: Scatter Plot of Trip Distance vs Fare Amount

This plot shows a strong positive relationship between trip distance and fare amount. Horizontal lines near the top of the chart indicate fixed-rate or capped fare scenarios, such as airport rides.

The next visualization will be comparing the average fare price to the hour of day. We created a derived variable, *pickup_hour,* from extracting the hour from the variable *pickup_datetime.* This allows us to analyze if there are any patterns by targeting only the hour. We found that there is a consistent average fare across most hours of the day, but with a spike at 5 AM. This unusual peak, shown in Figure 3.2, is possibly due to early morning airport trips or limited driver supply, which will cause the fare to be more expensive.

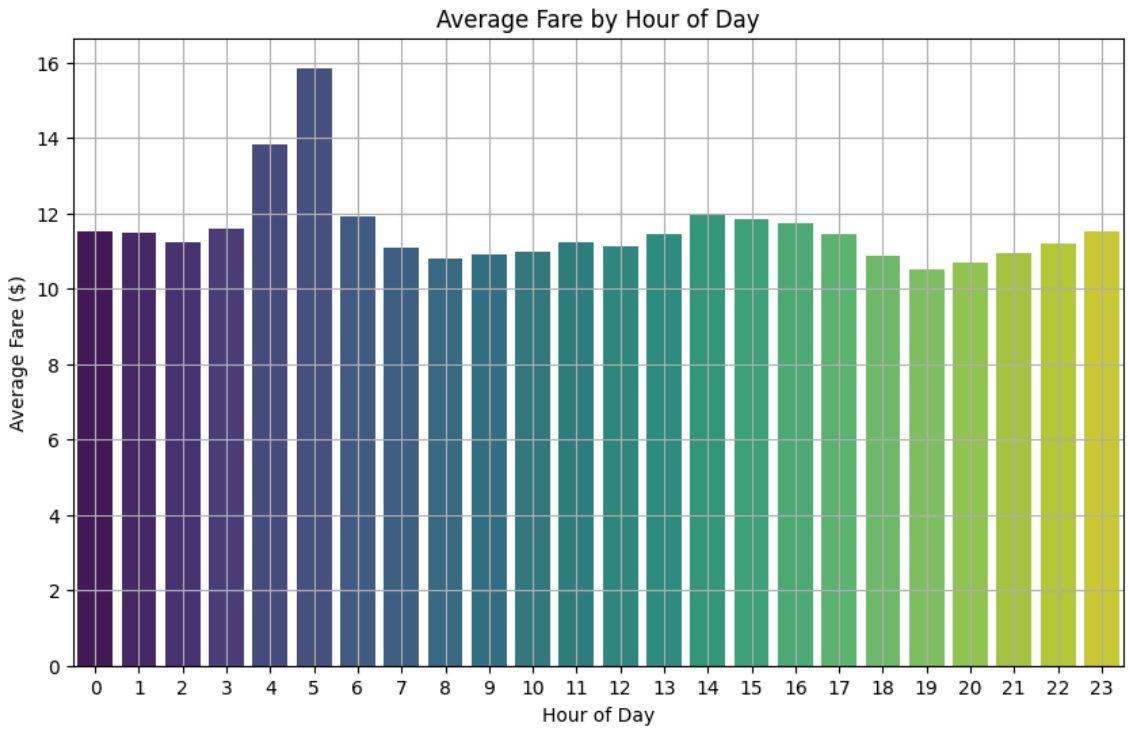

Figure 3.2: Average Fare by Hour of Day

This bar chart shows consistent fares across most hours, with a notable spike at 5:00 AM, possibly due to airport travel or low driver availability during early morning hours.

The other derived variable we created from *pickup_datetime* helps us visualize the next bar chart. In this bar chart, we are comparing the average fare price to the day of week using *pickup_dayofweek.* The results of this chart are very subtle but still give us very insightful information. As shown in Figure 3.3, Sunday has the highest daily average fare, while Tuesday has the lowest. This suggests that the day of week also has an influence on fare pricing. These trends support our decision to include our derived variables, *pickup_hour* and *pickup_dayofweek,* as features in our model.

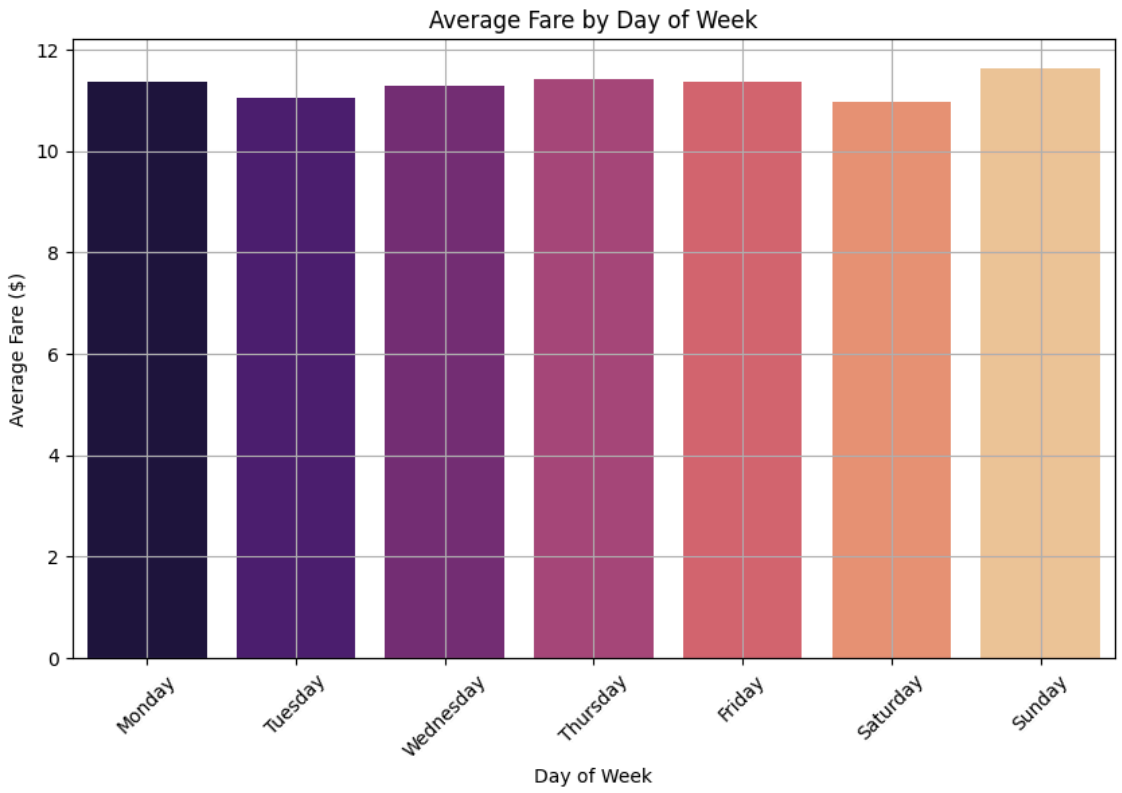

Figure 3.3: Average Fare Amount by Day of Week

This bar chart shows slight variations in average fare based on the day of the week. Sunday has the highest average fare, while Tuesday has the lowest, supporting the inclusion of pickup_dayofweek as a meaningful feature in predictive modeling.

We moved onto creating a correlation matrix that will help us analyze all variables at once. As shown in Figure 3.4, the matrix confirms that *trip_distance* is the most influential factor affecting *fare_amount,* with a very strong positive correlation coefficient of 0.89. This supports our earlier findings and hypothesis. Another thing to note are the variables, *passenger_count* and *pickup_hour,* that show little to no correlation with fare prices, besides the spike in fare amount at 5 AM. However, this suggests that time-based variables are best explored through group trends rather than individual correlations.

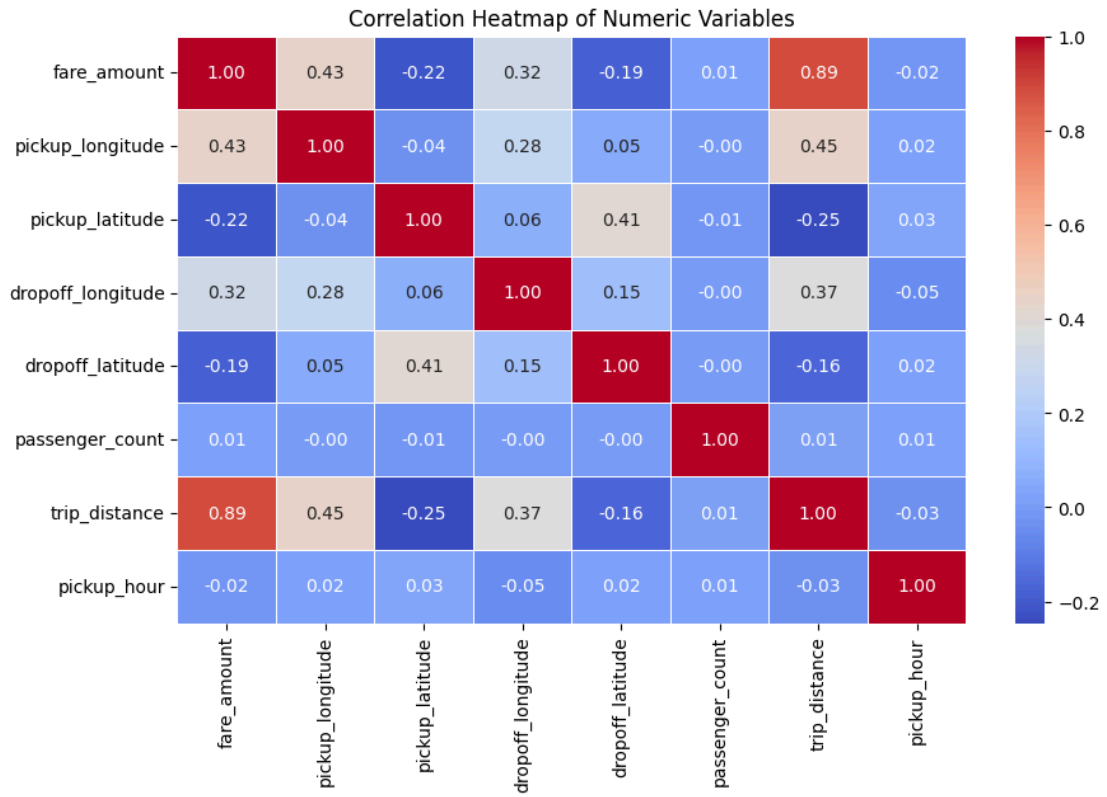

Figure 3.4: Correlation Matrix of Key Variables
This heatmap shows the strength of pairwise relationships among variables. ***Trip_distance*** *has the strongest correlation with* ***fare_amount****(r = 0.89), while* ***passenger_count*** *and* ***pickup_hour*** *show little to no correlation; suggesting the importance of analyzing time-based patterns through grouped visualizations instead.*

Lastly, we created a heatmap that shows Uber pickup density across the given longitude and latitude coordinates. The brightest red hotspot is centered in Midtown Manhattan. This reflects a consistent high demand for Uber rides from possible commuters and tourists. This hotspot is also near major transit hubs, such as Penn Station and Grand Central Station, where consumers are requesting pickups. Additionally, warmer hotspots leak into Lower Manhattan. This is likely due to activity and nightlife in the Financial District. There are also notably smaller clusters to JFK and LaGuardia Airports, and Brooklyn suggesting frequent airport travel and interborough travel. This visual (Figure 3.5) highlights key areas where Uber service is most active and supports the importance of incorporating geographic location into our predictive models.

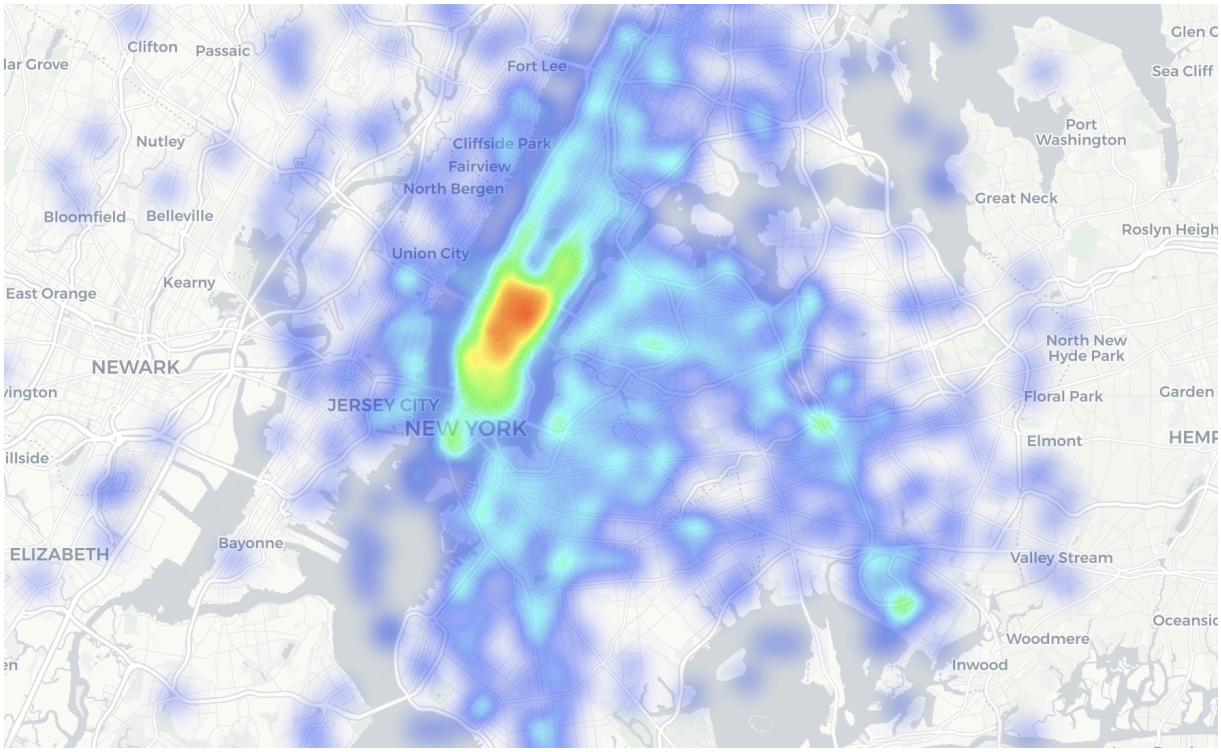

Figure 3.5: Heatmap of Uber Pickup Density by Location
This heatmap displays the concentration of Uber pickups based on geographic coordinates. Midtown Manhattan shows the highest density, followed by Lower Manhattan and airport zones such as JFK and LaGuardia. These spatial trends emphasize the value of including location data in predictive modeling.

As we move into the modeling phase, several key takeaways from our EDA stand out. Most notably, *trip_distance* emerges as the strongest predictor and will likely play a central role in our final model. Our derived variables, *pickup_hour* and *pickup_dayofweek* also revealed meaningful patterns and are best represented as categorical features. Additionally, geographic features, such as pickup and dropoff locations, suggest potential for spatial clustering that may enhance model performance. Finally, special attention should be given to outliers and fixed-rate trips, as observed in the scatter plots, since these cases could distort model accuracy if not properly addressed.

In this final section, we explored several spatial and time-based patterns in Uber pricing to better understand what drives fare differences and efficiency.

Figure 4.1 compares the average fare-per-mile across NYC boroughs. We calculated fare-per-mile as *fare_amount / trip_distance,* and assigned boroughs using latitude and longitude ranges. The plot reveals significant differences, with Bronx and Staten Island showing higher fare-per-mile on average, likely influenced by fewer trips and short-distance base fares. It is important to note that the dataset is heavily weighted toward Manhattan trips, which may affect the relative averages for outer boroughs.

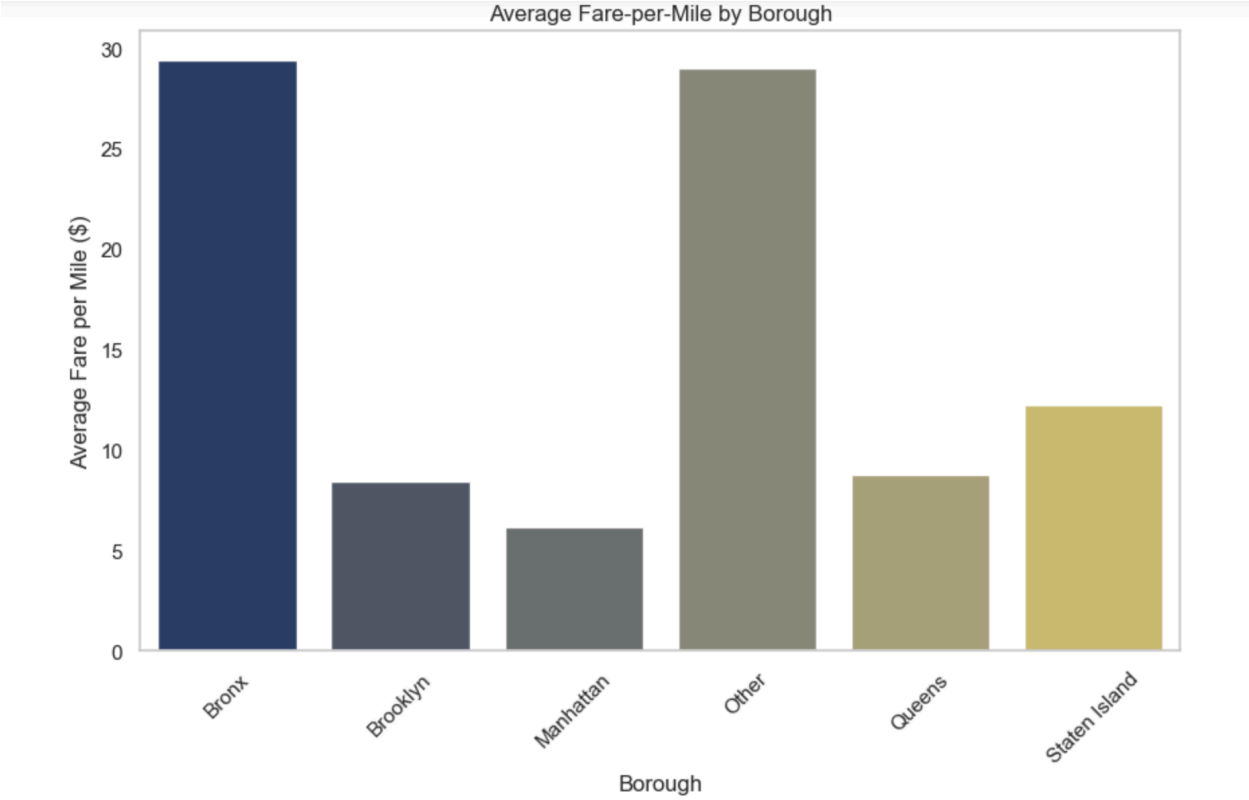

Figure 4.1: Average Fare per Mile by Borough

Bronx and Staten Island have higher fare-per-mile, likely due to fewer trips and base-fare dominance in short-distance travel

Next, we grouped trips by *pickup_hour* and calculated the average *fare_per_mile* across each hour. Figure 4.2 indicates that fare-per-mile tends to be higher during early morning hours (e.g. 6-7 AM) and late afternoon, potentially reflecting traffic conditions, demand surges, and the effect of shorter trips at certain times of day. Similar, but slightly different findings from our look

at fare per total trip. Trips with extremely short durations or distances may disproportionately influence fare-per-mile in certain hours.

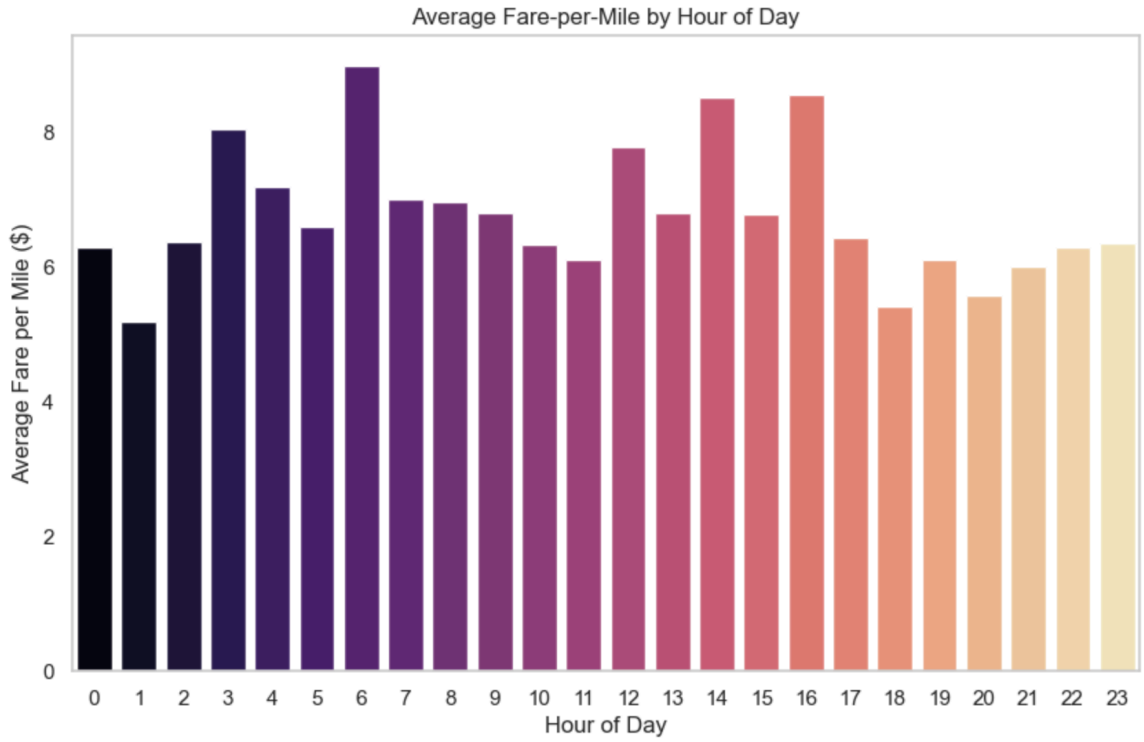

Figure 4.2: Average Fare per Mile by Hour of Day

Fare-per-mile peaks during early morning and late afternoon, potentially reflecting demand surges or shorter trip types.

Figure 4.3 shows the relationship between fare-per-minute and trip duration. We estimated trip duration based on an assumed average speed of 12 mph and calculated fare-per-minute. The trend-line highlights how fare-per-minute sharply decreases as trip duration increases, reflecting Uber's pricing structure where base fare dominates short trips. This also means per-minute and per-mile rates drive longer trip costs. The estimation of trip duration based on average speed introduces some approximation error, and outliers with very short distances may inflate fare-per-minute for short durations.

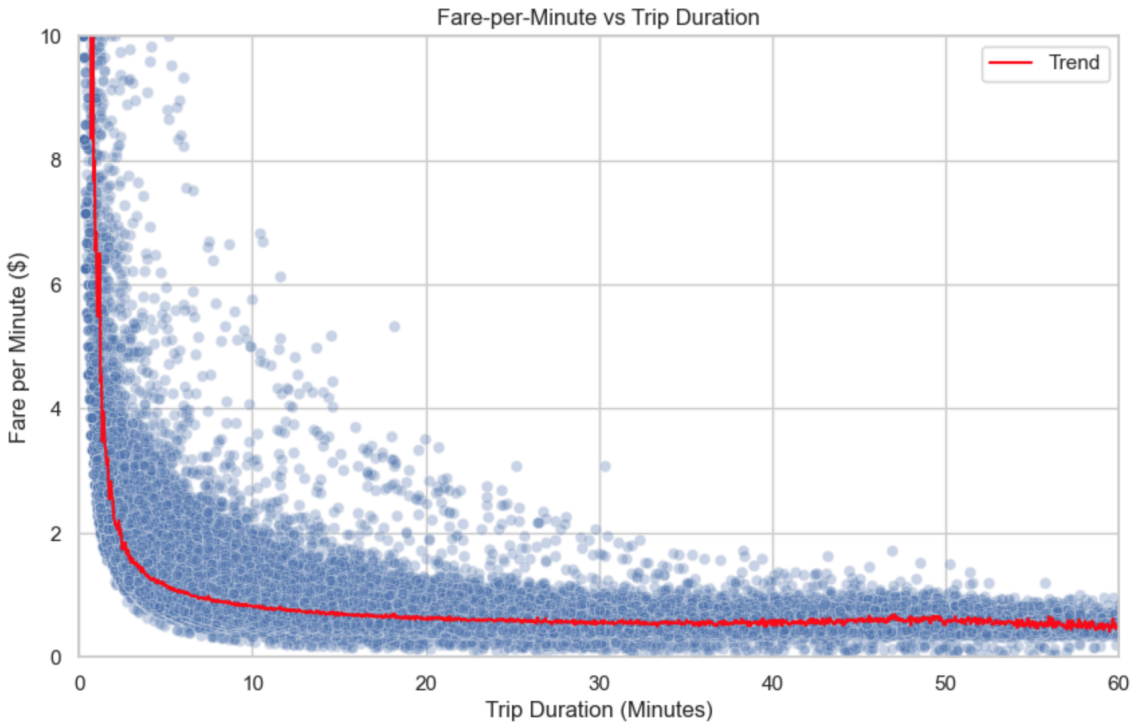

Figure 4.3: Fare per Minute vs Estimated Trip Duration

Fare-per-minute declines sharply as trip length increases, illustrating the role of base fare in short rides.

We also visualized *fare_per_minute* across boroughs using a boxplot. Figure 4.4 combines spatial and time-based pricing analysis, showing not just averages but the variance within each borough. Manhattan, Brooklyn, and Queens exhibit relatively consistent fare-per-minute patterns, while Bronx and Staten Island display much higher variability – likely due to fewer trips and a wider mix of trip types. The lower number of trips in Bronx and Staten Island should be considered when interpreting the variability in these boroughs.

This analysis confirms that trip distance and trip distance are major drivers of Uber fare pricing. Shorter trips tend to have much higher fare-per-mile and fare-per-minute rates because of the disproportionate impact of the base fare, whereas longer trips distribute costs more evenly, resulting in lower rates per unit of time and distance. We also noticed that pricing varies by borough– Manhattan, Brooklyn, and Queens show pretty consistent patterns, but Bronx and Staten Island have a lot more variation. We think this is because there are fewer trips in those

areas, and probably more unusual trip types. Certain hours of the day also show higher fare-per-mile, especially early mornings and late afternoons.

These patterns could be useful for dynamic pricing decisions. For example, Uber might want to adjust prices or promotions in areas with more inconsistent pricing, or rethink how fares are structured for short trips. The way fare-per-minute drops with trip duration is also something with modeling, since it helps explain why some rides feel more expensive than others. Moving forward, we'll focus on including these kinds of spatial and time-based patterns in any pricing models we build.

Overall, this exploratory data analysis has provided a strong foundation for modeling Uber fare prices. By identifying key drivers such as trip distance, time of day, and pickup location, we've not only uncovered valuable pricing patterns but also aligned our findings with our original research questions. With this understanding, we are now ready to move into building and evaluating predictive models.