Group members: Cindy Sun, Xiaoman Xu, Molly Han, Jenny Zhu

1. Finalized Research Question

How well can we predict ride-sharing costs in Boston using distance alone versus models that incorporate additional factors such as weather conditions, temporal patterns, and service platform?

Background: Distance is theoretically the primary driver of cost; however, pricing structures differ between Uber and Lyft. This project aims to quantify the distance-cost relationship for each platform and identify whether secondary factors, specifically weather (temperature and precipitation), time of day, and location, significantly impact pricing or drive consumer preference between the two platforms.

Modeling Strategy: We plan to use multiple regression to model the relationship between distance and cost while controlling for confounders, and logistic regression to model the probability of a user choosing Uber versus Lyft based on these trip characteristics. To achieve the best prediction accuracy, we will evaluate and compare several modeling approaches, including:

- **Regression & Multiple Regression** (to establish linear baselines)
- **Polynomial Regression** (to capture non-linear relationships)
- **Decision Trees** (to handle complex interactions between categorical and numerical features)

2. Data Description

Data Source & Collection *[Data Source]*: The analysis utilizes detailed ride-level observations collected in the Boston metropolitan area. Each row represents a single ride quote containing pricing, trip attributes, and environmental conditions at the exact time of the request. Our initial dataset contains **693,017 observations** and integrates three distinct layers of information: ride characteristics, temporal/spatial context, and weather conditions.

Variable Definitions:

Ride Metrics

- ***Target Variable:*** **price (Numerical):** Quoted fare in USD.
- **distance (Numerical):** Trip length in miles (Central Explanatory Variable).
- **cab_type (Categorical):** Service platform (Uber or Lyft).
- **name (Categorical):** Specific service class (e.g., UberX, Lyft Lux) denoting quality and base price.
- **surge_multiplier (Numerical):** Dynamic pricing factor applied to the base fare.

Spatial & Temporal Context

- **source, destination (Categorical):** Pickup and drop-off neighborhoods in Boston.
- **timestamp, month, hour, day_of_month (Num/Cat):** Time and date records of the request.
- **is_weekend (Boolean):** Engineered feature distinguishing weekend leisure from weekday travel.

Weather Metrics

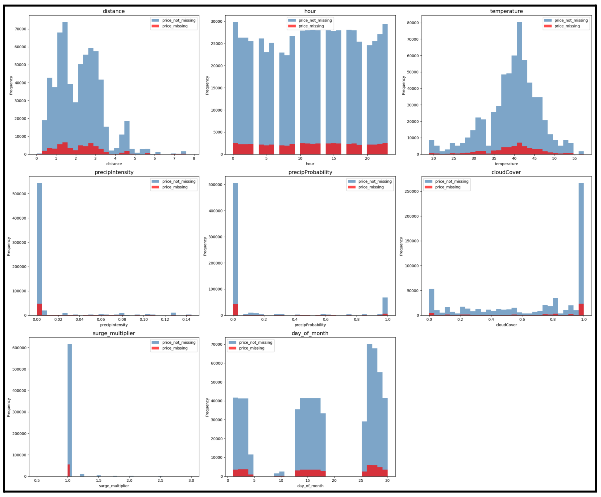
- **temperature, precipIntensity, precipProbability, cloudCover (Numerical):** Weather conditions at the time of request.
- **short_summary (Categorical):** Textual description of weather (e.g., "Light Rain").

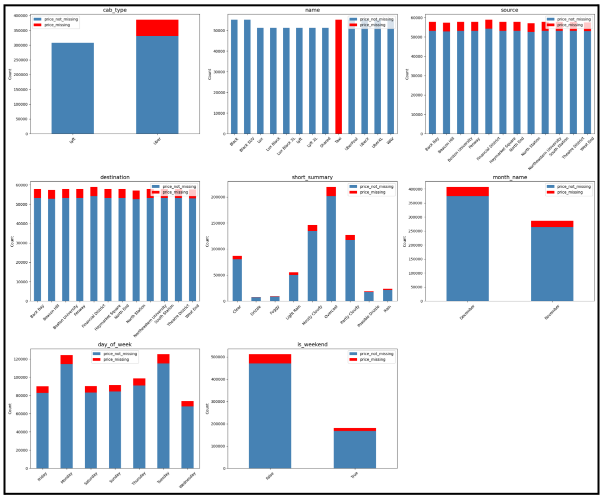
3. Deeper Understanding of the Data

To ensure robust modeling, we cleaned the data and addressed potential quality issues as follows:

a. Missing Data in Target Variable

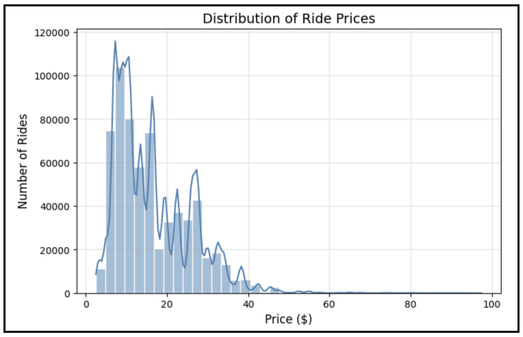
The price variable contained 55,095 missing values (7.95% of the dataset). Since mean/median imputation is unsuitable for the target variable and advanced methods (KNN/MICE) add unnecessary complexity, we removed these rows. The remaining 637,976 observations provide ample statistical power.

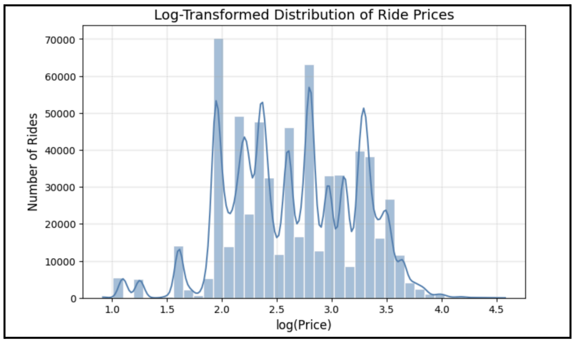




b. Class Imbalance: The cab_type distribution (55.6% Uber vs. 44.4% Lyft) shows a negligible ~5% imbalance. Given the large dataset size, the classes are sufficiently balanced for comparative analysis without requiring resampling techniques.

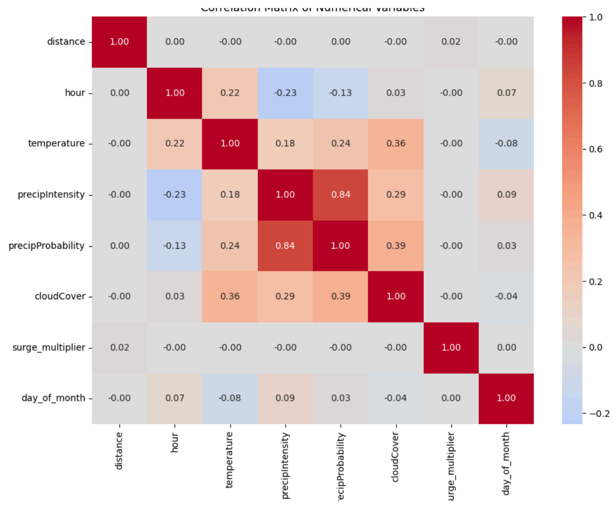
c. Addressing Target Skewness: The right-skewed price distribution violated regression normality assumptions. We normalized the data using a natural log transformation: **log(Price)**





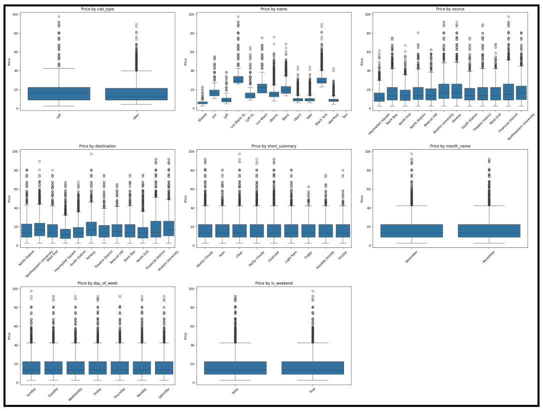
d. Temporal Feature Engineering: To better capture seasonality, we converted the raw numerical timestamp, hour, and day variables into proper datetime objects and categorical features (e.g., “Day of Week”) for modeling.

e. Multicollinearity Check: Covariance analysis revealed strong multicollinearity between **precipIntensity** and **precipProbability** (r = 0.84). To prevent variance inflation, we will drop one of these variables. All other predictors exhibited weak correlations (|r| < 0.4), posing minimal risk.



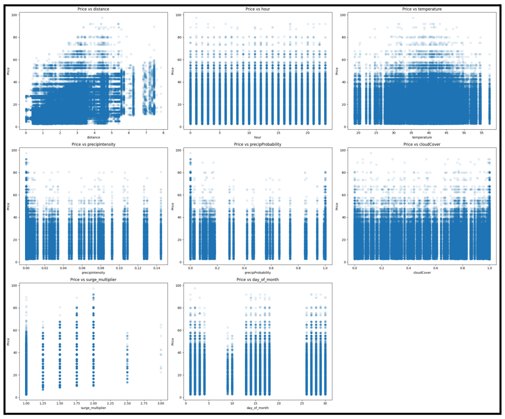
4. Meaningful Insights

a. Categorical Variable Insights



- **Service Class Variance:** Premium services (e.g., Lyft Lux) show the largest price differences, making "Service Type" a crucial control variable; otherwise, high fares could be misattributed to distance.
- **Location Impact:** Certain pickup/drop-off hubs (e.g., airports) have higher median fares, driven by longer typical distances and concentrated demand.
- **Limited Environmental Impact:** Weather categories and temporal features show only modest price differences, indicating they play a secondary role compared to distance and service level.

b. Numerical Variable Insights (PS: all predictors were standardized)



- **Price-Distance Relationship:** Distance is the only numerical predictor with a clear positive linear trend with price. The "banded" pattern suggests service classes operate on distinct base rates per mile.
- **Multimodal Distance Distribution:** Distance is multimodal, reflecting different ride types—short city trips vs. longer commutes or airport rides..
- **Weather Skewness & Outliers:** Precipitation variables are highly right-skewed with 150k+ legitimate outliers, representing rare heavy-rain conditions. These are retained because they may correspond to surge events.

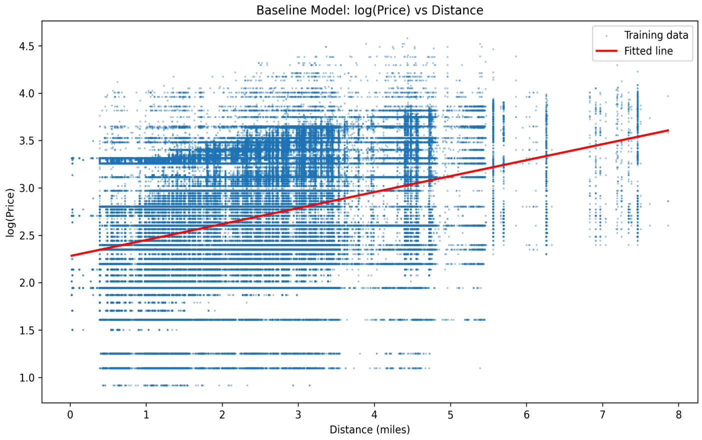
5. Baseline Model

To establish a performance benchmark, we constructed a **Simple Linear Regression** model.

log(Price) = β_0 + β_1·Distance + ε

Data Splitting: To ensure robust evaluation and prevent overfitting, we partitioned the dataset into three subsets:

1) **Training Set (60%, N=382,785):** Used to fit the model parameters; 2) **Validation Set (20%, N=127,595):** Used for tuning and unbiased evaluation during development; 3) **Test Set (20%, N=127,596):** Held out for final performance assessment.



Performance & Evaluation: The model was evaluated on both the training and validation sets:

Model fit: log(Price) = 2.2819 + 0.1685 · Distance + ε
R^2 (Coefficient of Determination): 0.114 (Train) / 0.117 (Validation)
RMSE (Root Mean Squared Error): 0.535 (Train) / 0.537 (Validation)

Interpretation: The baseline model confirms a positive linear relationship between distance and cost, with a coefficient of 0.169. However, distance alone explains only about 11–12% of the total variance in pricing (R^2 ≈ 0.117). The distinct horizontal banding observed in the residual plots, combined with the low R^2, indicates that distance is insufficient for accurate prediction. Factors such as service type (e.g., UberX vs. UberBlack), surge pricing, and contextual variables are necessary to adequately capture the ride-sharing pricing structure in future iterations.