UBER & LYFT PRICES

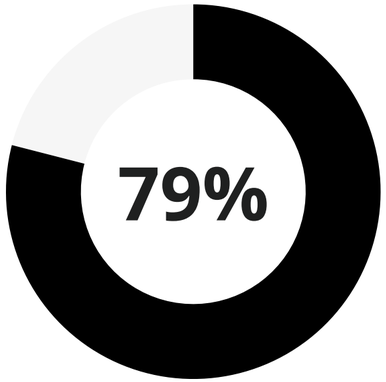
JANIE CHEN,

Our Team

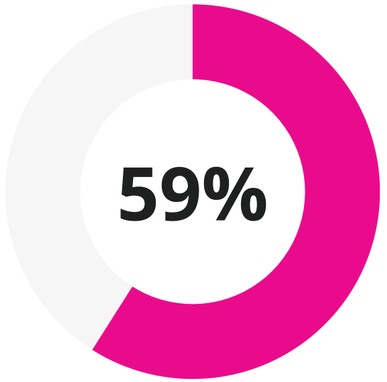
Janie Chen

Problem

Lack of Transparency and Fair Pricing Awareness in Ride-Sharing Services



OF AMERICANS USE UBER



OF AMERICANS USE LYFT

Source: United States: Ride-hailing providers used 2023 | Statista

Brown Demands Transparency from Uber and Lyft on Surge Pricing

The Official website of The United States Committee on Banking, Housing, and Urban Affairs

senate.gov / Jul 3

Project Goals

Develop A Fair pricing Comparison Model To Empower Consumers

Short-Term

Create an ethical price model based on Uber and Lyft pricing data, service, geographic, weather, and time predictors

Long-Term

Develop a platform that aggregates real-time pricing data from rideshare and transportation services to compare fares side by side

Create opportunity for small rideshare businesses to be competitive in Boston

Our Data

57 Features | 14,248 Rows
Geography: Boston, MA
Timeframe: Nov-Dec 2018

Ride-related variables

- Source
- Destination
- Cab type
- Product name
- Price
- Distance
- Surge multiplier
- Date time of ride
- etc.

Weather-related Variables

- Temperature
- Weather summary
- Precipitation
- Humidity
- Wind speed
- Visibility
- Dew point
- UV Index
- Pressure
- Cloud cover
- Ozone
- etc.

Investigate pricing patterns, assess the impact of external factors on fare fluctuations, and develop models to predict ride prices under different conditions

Data Exploration & Wrangling

Missing Value Treatment

Only null values were in price

- removed to serve as a validation set

Dropped Predictors

Non-Predictors

- Id

Redundant Variables:

- datetime, long_summary, visibility.1, product_id, month, temperatureHigh, temperatureLow, temperatureHighTime, temperatureLowTime, timezone

Non-Variance Values:

- sunsetTime

Feature Engineering

_weekday:

- new variable from days

source_destination:

- combine source and destination

distance_hour:

- product of distance and hour

temperature_precipIntensity :

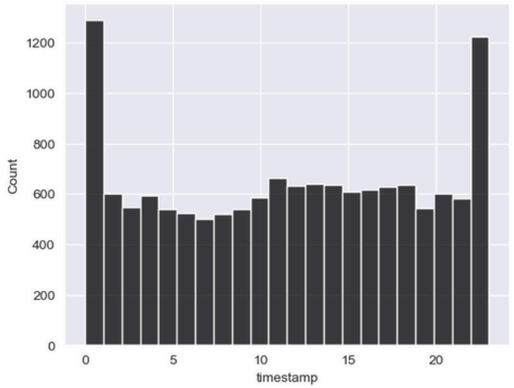
- product of temperature and precipitation intensity

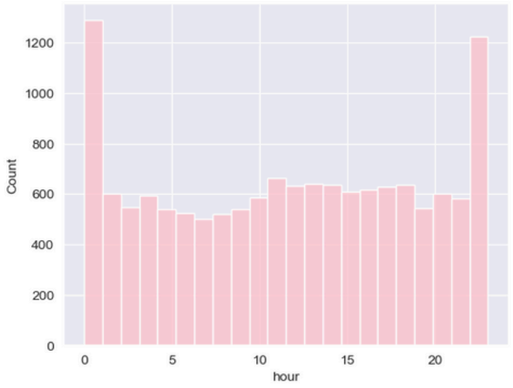
windspeed_visibility :

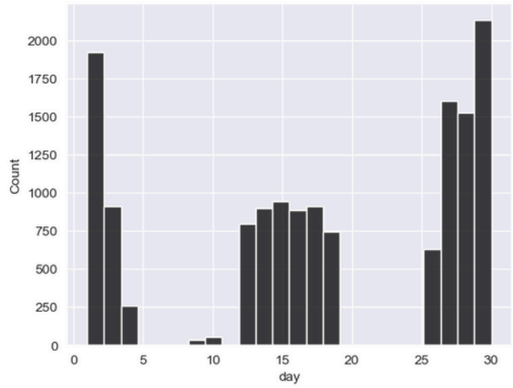
- product of windspeed and visibility

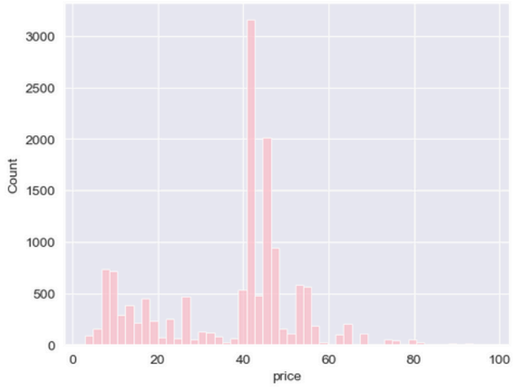
EXPLORATORY DATA ANALYSIS

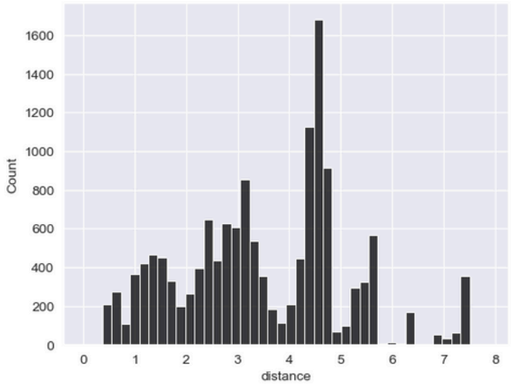
Section 2

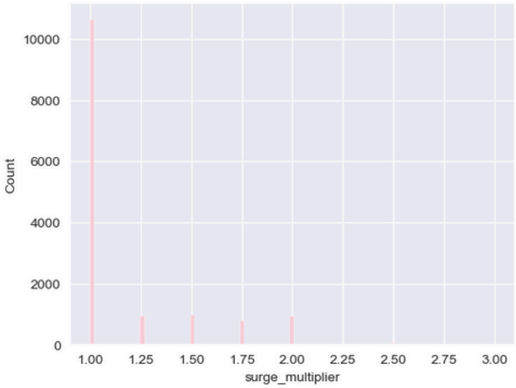






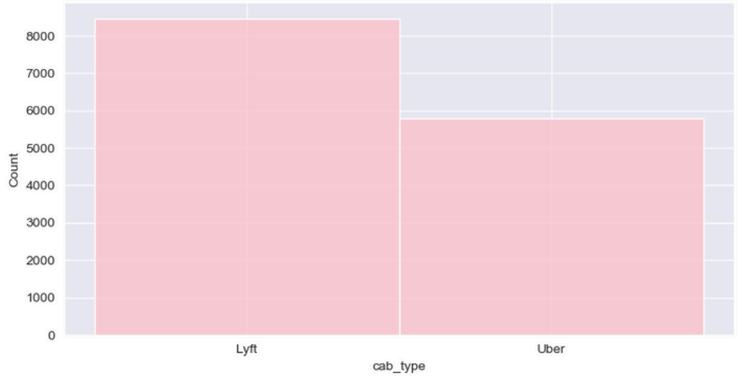


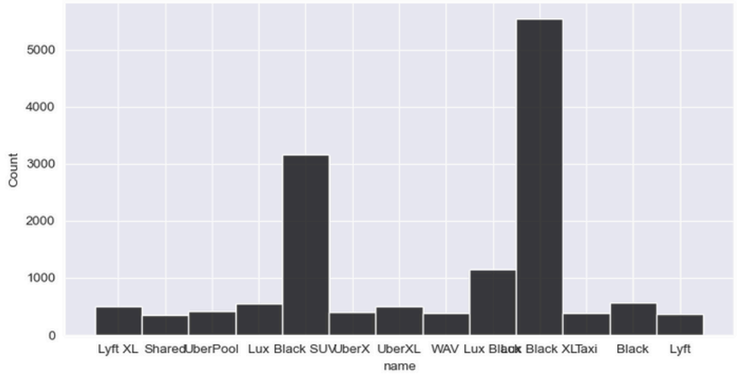


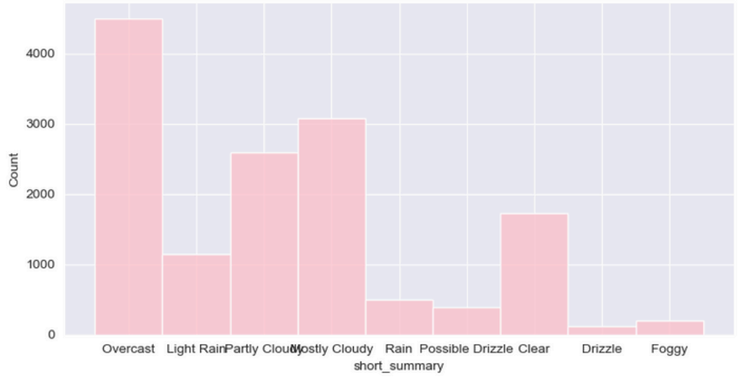


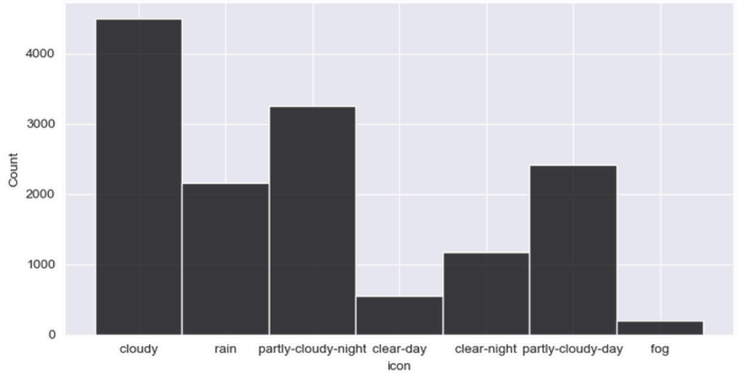
Univariate Analysis

- **timestamp & hour**: Demand is consistent throughout the day, with peaks around early morning and late evening.
- **day**: Data seems to be clustered around only beginning, middle, and end of months, missing values in between
- **price & surge_multiplier**: Ride prices spike around $40, with few expensive rides indicating possible surge pricing or long distances.
- **distance**: The values seem to be somewhat normally distributed for each ride length.





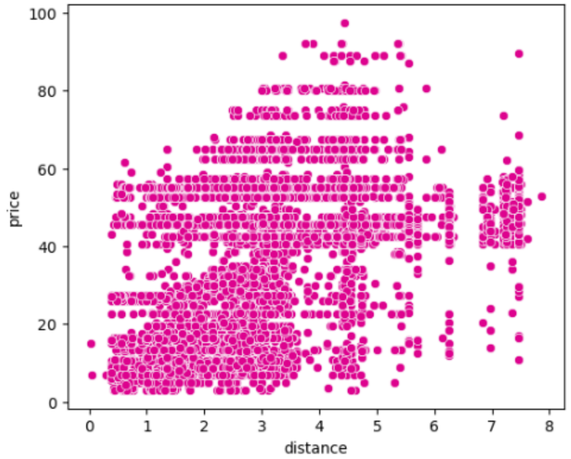


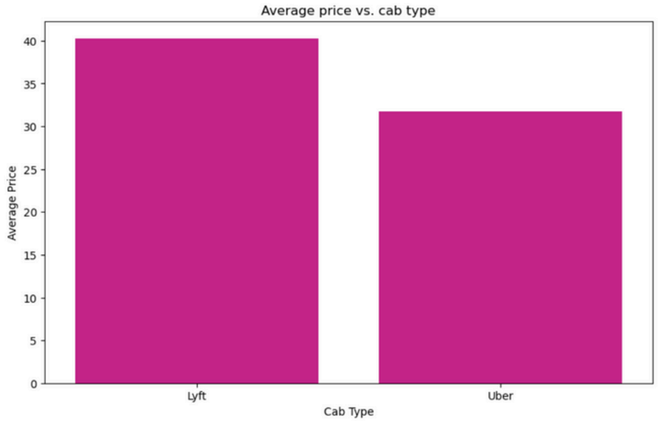


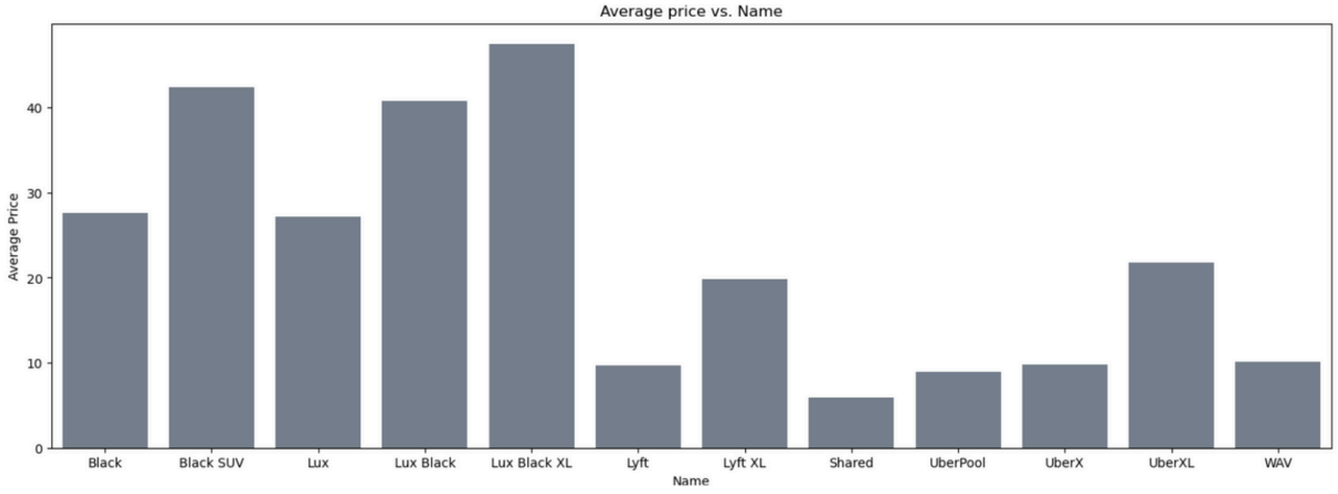
Univariate Analysis

- **cab_type**: Lyft is more popular than Uber in this dataset.
- **name**: Black XL, BlackSUV, and Lux Black seem to be the most demanded car types, indicating more ridesharing.
- **short_summary** & **icon**: Most rides occur under overcast or cloudy conditions; clear day and fog are less common.

Bivariate Analysis







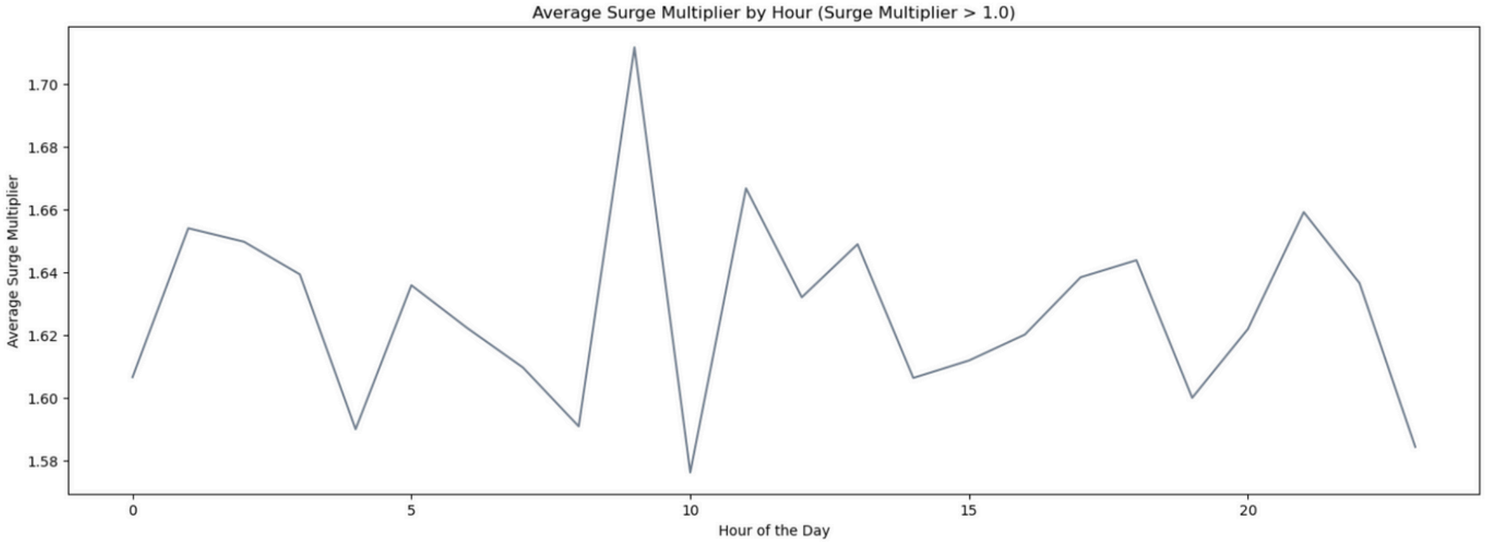
Relationships with price

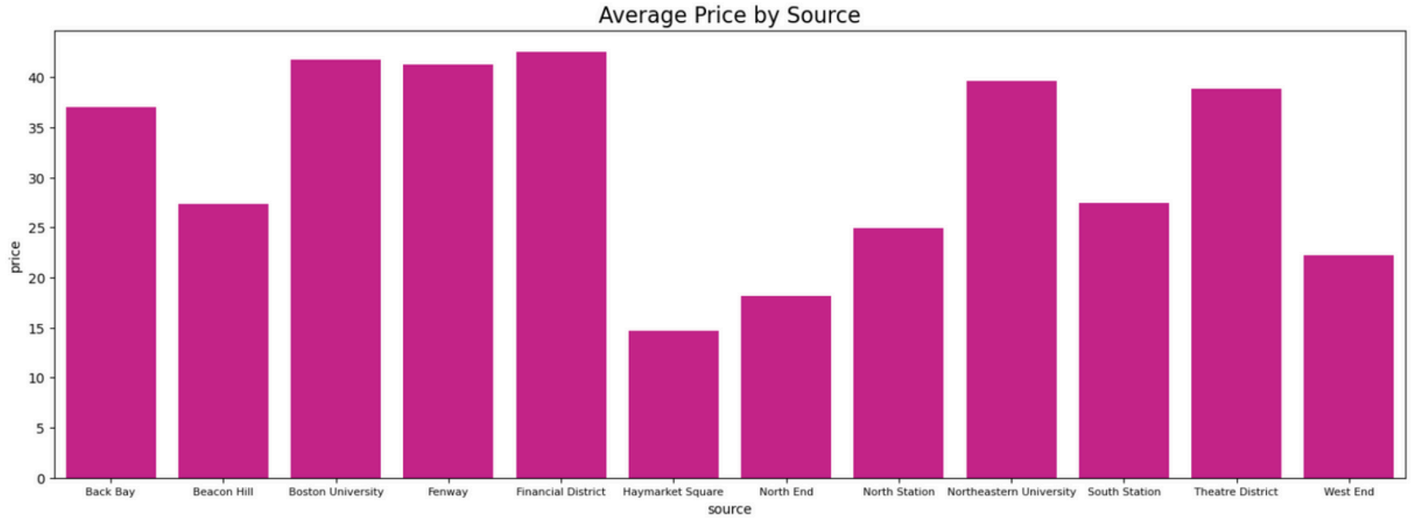
i) Continuous

- Price seems to have a strong correlation with distance of the ride
- Other variables seem to have a non-linear relationship with price

ii) Categorical

- Rides starting and ending in Fenway, Financial district and Boston University seem to be more expensive than others
- On average, price of a Lyft ride is higher than that of Uber ride
- As expected, premium cab offerings charge more than others
- None of the uber rides have surge pricing in the dataset
- Surge pricing is only assigned to Lyft rides, and it is the highest when the weather is 'Mostly cloudy'





Bivariate Analysis

Relationships with price

ii) Categorical (cont.)

- Surge pricing is only assigned to Lyft rides, and it is the highest at 9 am
- The average price for Boston University, Fenway, and Financial District are highest

Challenges

1. Data only includes two months of the year from 2018 (November and December)

2. Not all days of the month are captured (including Thanksgiving and Christmas)

3. Data lacked personal information (battery percentage, number of prior rides) we believe would have an impact on price

SOLUTION & INSIGHTS

Section 3

Linear Regression

SIGNIFICANT VARIABLES (29/86)

- Source (7 of 11 locations)
- Destination (8 of 11 locations)
- Cab Type[Uber]
- Name (all 10)
- Distance
- Surge Multiplier
- icon[cloudy]

INTERPRETABLE COEFFICIENTS

- source[Financial District]: 1.62
- destination[South Station]: -1.59
- name[UberX]: -12.84
- name[UberXL]: -5.156
- Distance: 3.43
- Surge Multiplier: 26.95

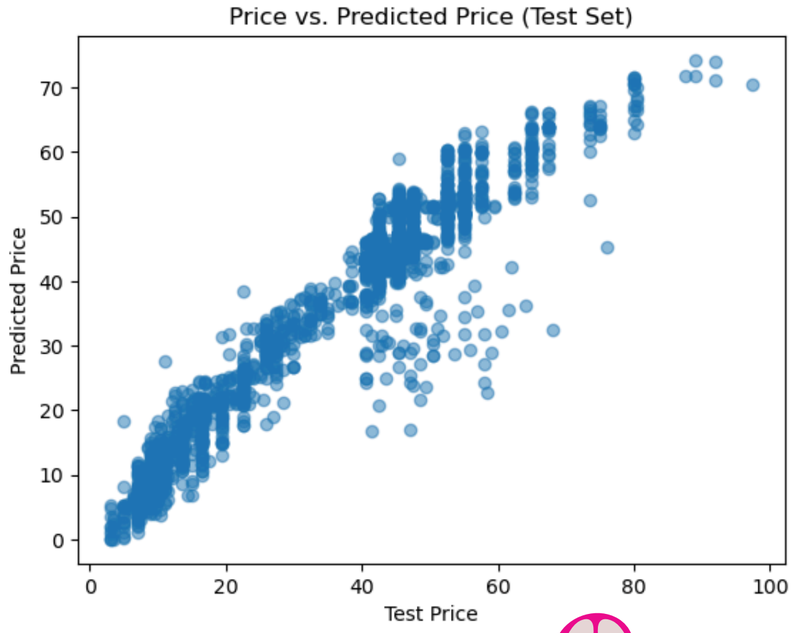
Linear Regression

Best Model Performance:

	Train	Test
Adj. R-Squared	0.93	0.92
RMSE	4.42	4.79

- Performed backward elimination to address multicollinearity concerns
- Number of significant variables was 29

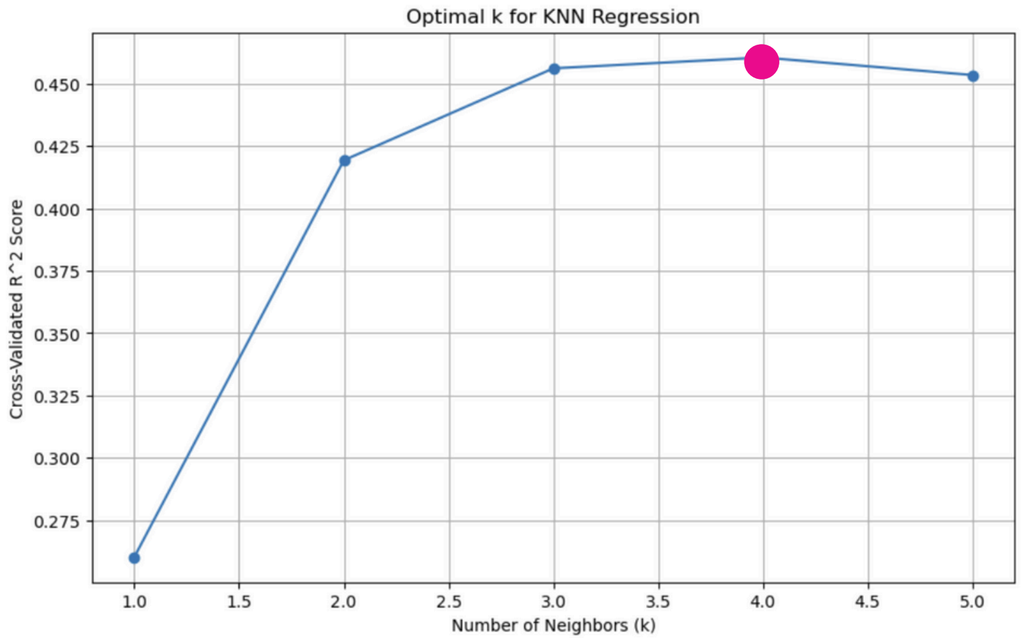
Insights: Residuals increase as fitted values increase, might not be properly capturing underlying patterns



K-Nearest Neighbors

Best Model Performance:

Metric	Train	Test
R-Squared	0.71	0.51
RMSE	8.82	11.75



Insights:

- The model performs significantly better on the training data than on the test data, which suggests that the KNN model might be overfitting
- The higher RMSE on the test set compared to the training set further confirms the presence of a generalization error

Bagging

Best Model Performance - Full Dataset:

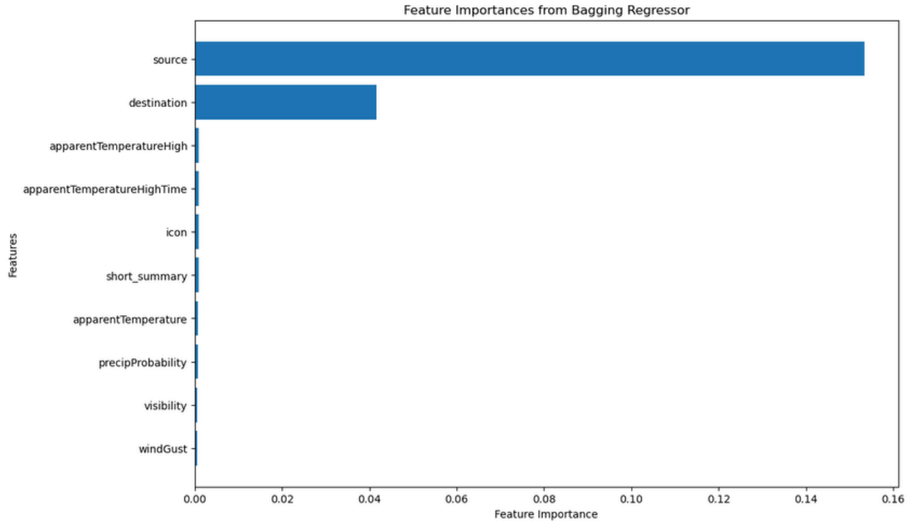
	Train	Test
R-Squared	0.98	0.97
RMSE	1.18	1.71

Best parameters after cross validation:

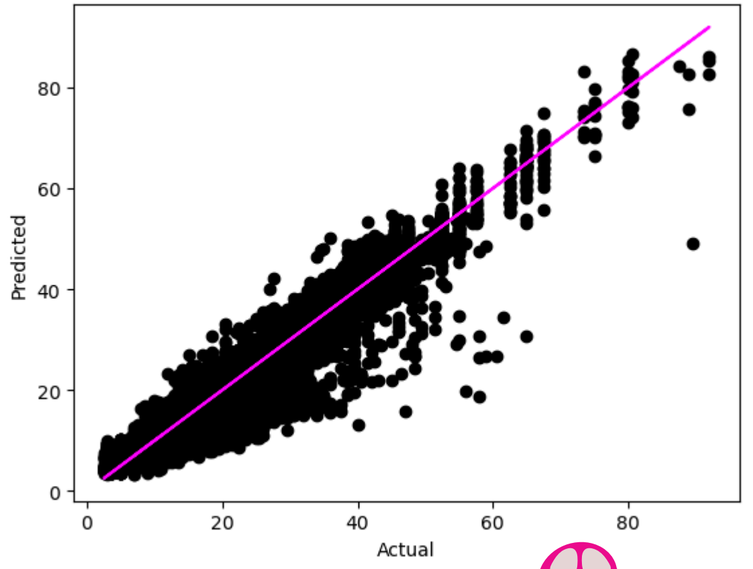
- n_estimator = 200 trees
- max_samples = 0.8 of samples
- max depth of 25

Insights: Bagging is performing well, likely due to being less prone to overfitting noise and its optimized parameters.

Feature Importances



"Predicted vs. Actual" plot



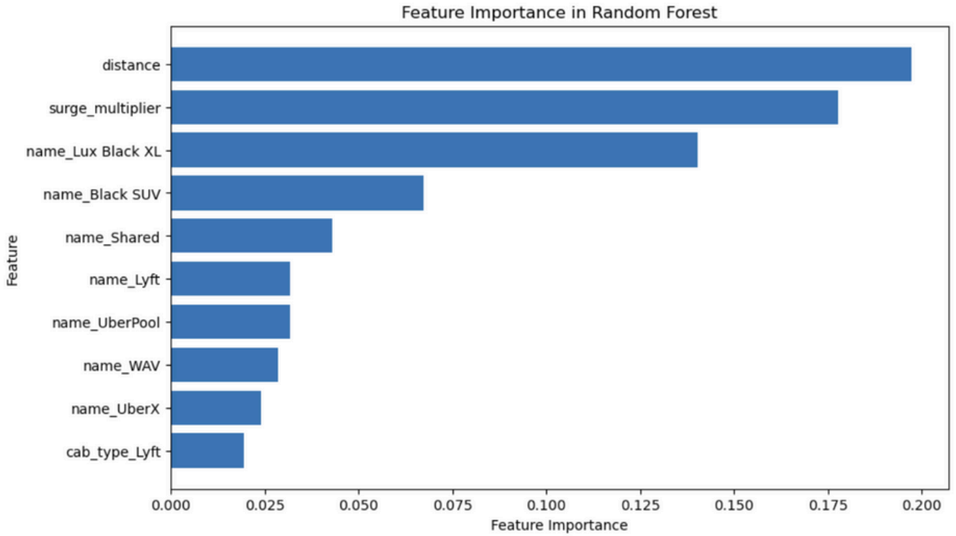
Random Forest

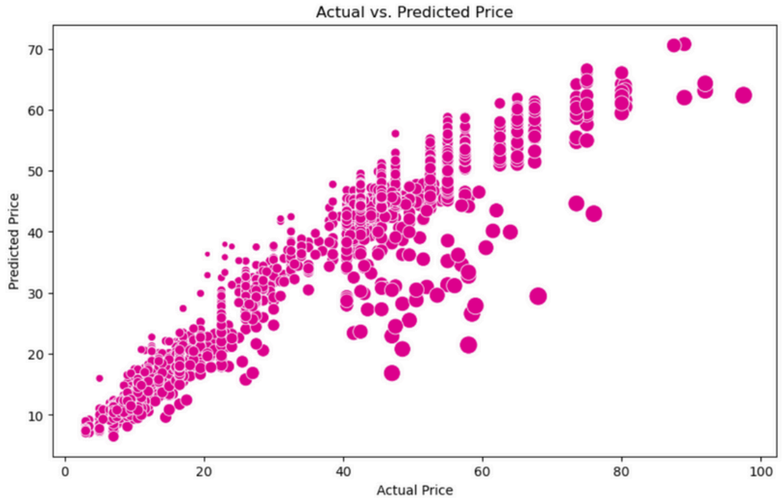
Best Model Performance:

Metric	Train	Test
R-Squared	0.97	0.93
RMSE	4.09	4.35

Best parameters after cross validation:

- Max Depth: 15
- Max Features: 10
- n_estimators: 500





Insights: The model exhibits strong performance, with an R-squared value of 0.97 on the training and 0.93 on the test set, indicating a high level of explained variance.

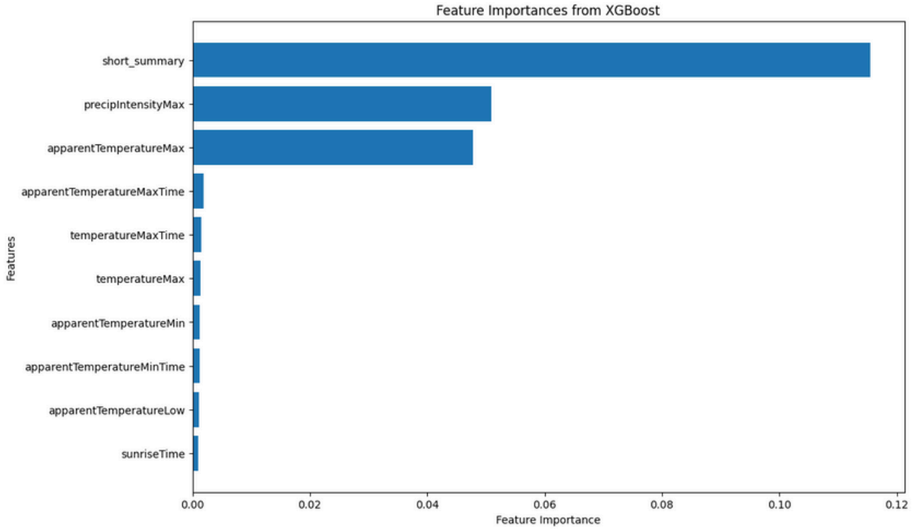
XGBoost

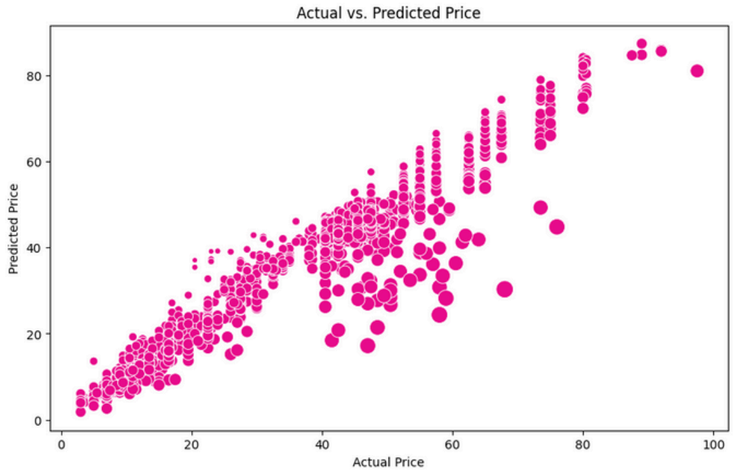
Best Model Performance:

	Train	Test
R-Squared	0.97	0.95
RMSE	2.74	3.83

Best parameters after cross validation:

- n_estimators: 200
- learning_rate: 0.1
- max_depth: 5
- colsample_bytree: 0.7
- alpha: 10



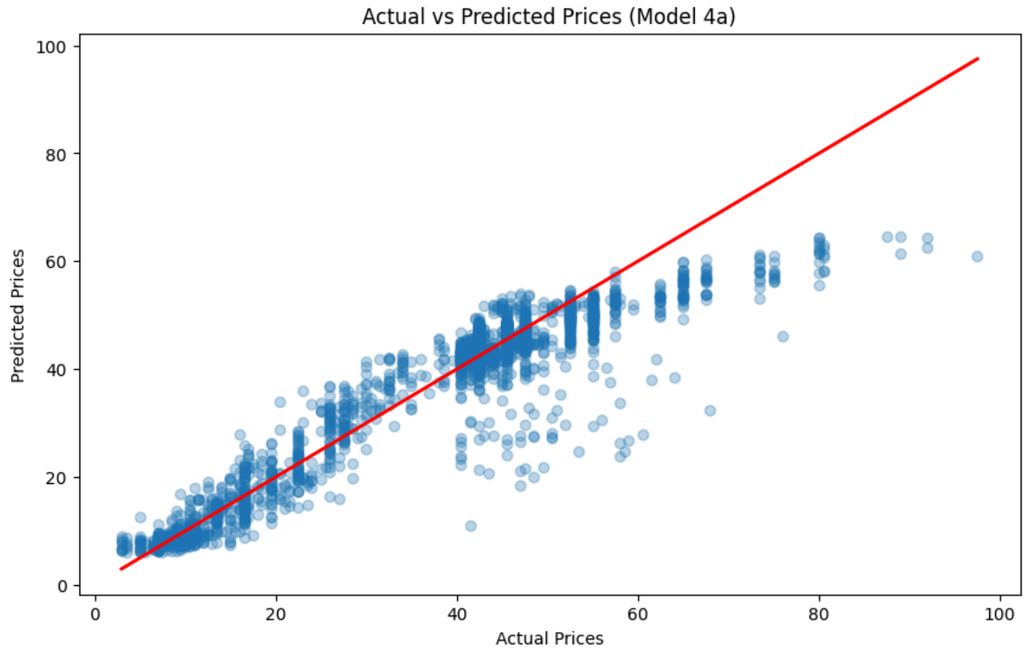


Insights: The model demonstrates strong performance, achieving an R-squared value of 0.97 on the training set and 0.95 on the test set, showing a high level of explained variance.

Neural Network

Best Model Performance:

Metric	Train	Test
R-Squared	0.90	0.91
RMSE	5.20	4.95



Architecture

- 9 Layers
- Neurons:
  - 264 -> 128 -> 64 -> 1
- Activation Functions:
  - SiLU(), ELU(), LeakyReLU()

Regularization

- Two dropout layers (10%)

Optimization:

- Adam optimizer
- 0.001 learning rate

Insight: Although neural networks are powerful for complex data, the model's accuracy decreased when trying to predict prices over $40.

Top Feature Importances in Order

	Linear Regression	Bagging	Random Forest	Boosting
1	name_WAV	source	distance	short_summary
2	name_UberPool	destination	surge_multiplier	precipIntensityMax
3	name_Lux Black XL	apparentTemperatureHigh	name_Lux_BlackXL	apparentTemperatureMax
4	name_Lux Black	apparentTemperatureHighTime	name_Black_SUV	apparentTemperatureMaxTime
5	name_Black SUV	icon	name_Shared	temperatureMaxTime

Model Comparison

	In-Sample RMSE	Out-of-Sample RMSE	Test Set R^2
Multilinear	4.42	4.79	0.92
K-Nearest Neighbors	8.82	11.75	0.51
Bagging	**1.18**	**1.71**	**0.97**
Random Forest	4.76	4.99	0.91
XG Boost	2.74	3.83	0.95
Neural Network	5.20	4.95	0.91

Solution


fair price
From
Destination
Add Stop
Time
Cab Type
Submit
Fair Pricing for your ride from Fenway to Home at 9pm
fair price
$ $ $
$14.29
UBER
$ $ $ $
$18.91
lyft
$ $ $ $ $
$20.09
Boston Cab Co
$ $ $
$14.50

Moving Forward

Our Goal

Develop A Fair Pricing Comparison Model To Empower Consumers

Next Steps

- Build a surge multiplier model to infer proprietary data
- Extending scope across the U.S.
- Create a real-time prediction model that ensembles fine-tuned models (with neural network) to create an average price prediction

To greater adventures...

THANK YOU

JANI

APPENDIX

Data Dictionary

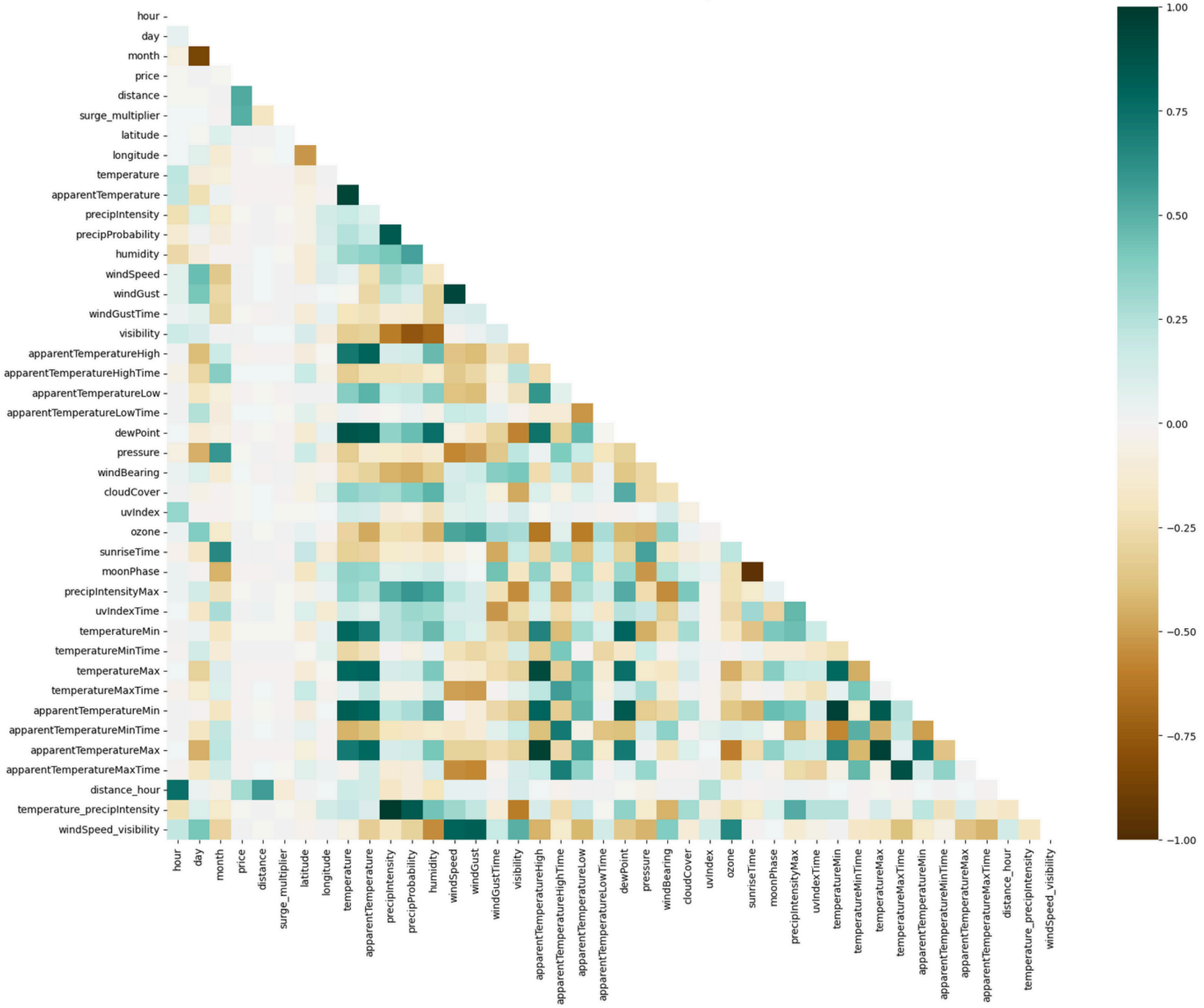
- **Time-related Variables:**
  - timestamp: The date and time when the ride was requested.
  - hour: The hour of the day when the ride was requested.
  - day: The day of the month when the ride was requested.
  - windGustTime: The time when the wind gust was recorded.
  - apparentTemperatureHighTime: Time when the highest perceived temperature was recorded.
  - apparentTemperatureLowTime: Time when the lowest perceived temperature was recorded.
  - sunriseTime: Time of sunrise on the day of the ride request.
  - sunsetTime: Time of sunset on the day of the ride request.
  - uvIndexTime: Time when the UV index was recorded.
  - temperatureMinTime: Time when the minimum temperature was recorded.
  - temperatureMaxTime: Time when the maximum temperature was recorded.
  - apparentTemperatureMinTime: Time when the minimum perceived temperature was recorded.
  - apparentTemperatureMaxTime: Time when the maximum perceived temperature was recorded.

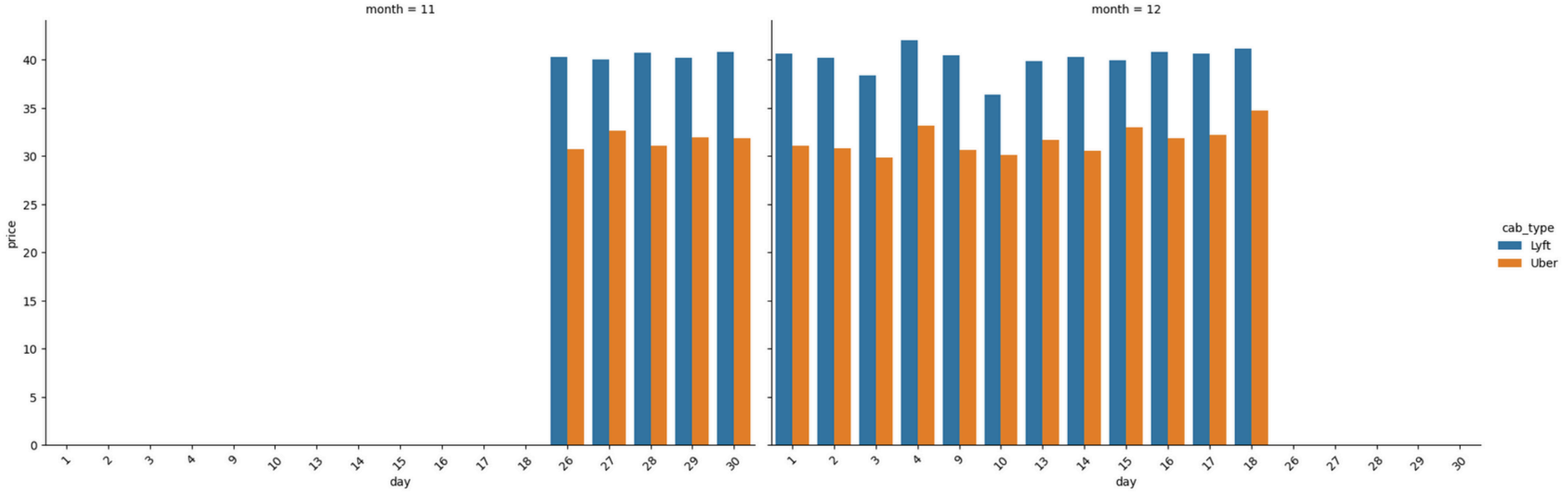
Data Dictionary

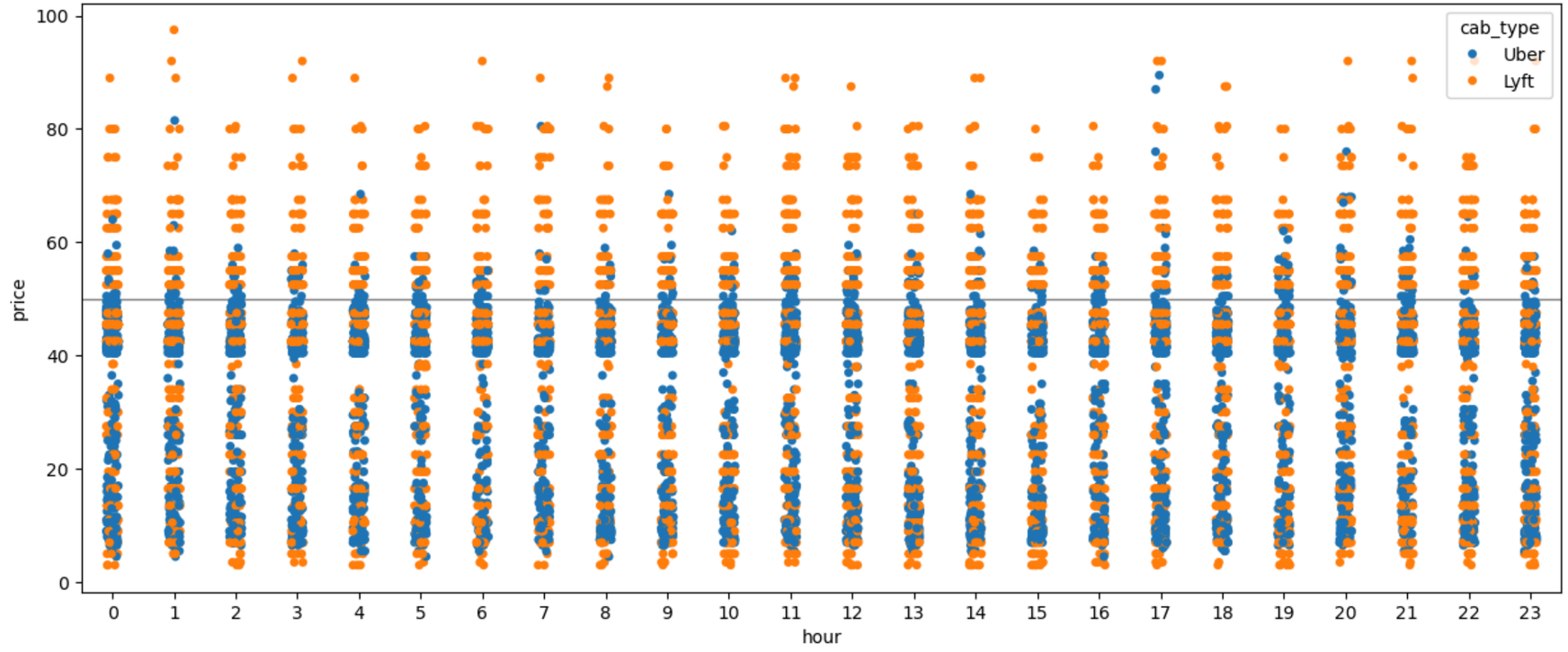
- **Ride-related Variables:**
  - cab_type: Type of cab service requested (e.g., Uber, Lyft).
  - name: Name of the ride type (e.g., UberX, Lyft Plus).
  - price: The price of the ride.
  - distance: The distance of the ride in miles.
  - surge_multiplier: Surge pricing multiplier applied to the ride.
- **Location-related Variables:**
  - source: The starting location of the ride.
  - destination: The ending location of the ride.
  - latitude: Latitude coordinate of the ride request.
  - longitude: Longitude coordinate of the ride request.

Data Dictionary

- **Weather-related Variables:**
  - temperature: The temperature at the time of the ride request.
  - apparentTemperature: The perceived temperature at the time of the ride request.
  - short_summary: Short summary of the weather conditions at the time of the ride request.
  - precipIntensity: Intensity of precipitation at the time of the ride request.
  - precipProbability: Probability of precipitation at the time of the ride request.
  - humidity: Humidity level at the time of the ride request.
  - windSpeed: Wind speed at the time of the ride request.
  - windGust: Wind gust speed at the time of the ride request.
  - visibility: Visibility at the time of the ride request.
  - apparentTemperatureHigh: Highest perceived temperature of the day.
  - apparentTemperatureLow: Lowest perceived temperature of the day.
  - icon: Weather icon representing the conditions.
  - dewPoint: Dew point at the time of the ride request.
  - pressure: Atmospheric pressure at the time of the ride request.
  - windBearing: Direction of the wind at the time of the ride request.
  - cloudCover: Cloud cover at the time of the ride request.
  - uvIndex: UV index at the time of the ride request.
  - ozone: Ozone level at the time of the ride request.
  - precipIntensityMax: Maximum intensity of precipitation for the day.
  - temperatureMin: Minimum temperature for the day.
  - temperatureMax: Maximum temperature for the day.
  - apparentTemperatureMin: Minimum perceived temperature for the day.
  - apparentTemperatureMax: Maximum perceived temperature for the day.
  - moonPhase: Phase of the moon on the day of the ride request.


Correlation Heatmap
hour
day
month
price
distance
surge_multiplier
latitude
longitude
temperature
apparentTemperature
precipIntensity
precipProbability
humidity
windSpeed
windGust
windGustTime
visibility
apparentTemperatureHigh
apparentTemperatureHighTime
apparentTemperatureLow
apparentTemperatureLowTime
dewPoint
pressure
windBearing
cloudCover
uvIndex
ozone
sunriseTime
moonPhase
precipIntensityMax
uvIndexTime
temperatureMin
temperatureMinTime
temperatureMax
temperatureMaxTime
apparentTemperatureMin
apparentTemperatureMinTime
apparentTemperatureMax
apparentTemperatureMaxTime
distance_hour
temperature_precipIntensity
windSpeed_visibility
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


month = 11
month = 12
price
day
cab_type
Lyft
Uber


cab_type
Uber
Lyft
price
hour

		price		surge_multiplier		
		mean	min	mean	median	max
cab_type	short_summary					
Lyft	Clear	41.037055	1.0	1.284889	1.0	2.5
	Drizzle	42.911765	1.0	1.257353	1.0	2.0
	Foggy	38.184000	1.0	1.220000	1.0	2.0
	Light Rain	39.804949	1.0	1.259098	1.0	2.5
	Mostly Cloudy	40.875462	1.0	1.286957	1.0	3.0
	Overcast	39.868243	1.0	1.260135	1.0	3.0
	Partly Cloudy	40.044041	1.0	1.257610	1.0	3.0
	Possible Drizzle	39.595982	1.0	1.263393	1.0	2.5
	Rain	41.025180	1.0	1.262590	1.0	3.0
Uber	Clear	30.915504	1.0	1.000000	1.0	1.0
	Drizzle	29.235294	1.0	1.000000	1.0	1.0
	Foggy	28.537313	1.0	1.000000	1.0	1.0
	Light Rain	31.403002	1.0	1.000000	1.0	1.0
	Mostly Cloudy	31.974547	1.0	1.000000	1.0	1.0
	Overcast	32.271004	1.0	1.000000	1.0	1.0
	Partly Cloudy	31.812178	1.0	1.000000	1.0	1.0
	Possible Drizzle	30.506211	1.0	1.000000	1.0	1.0
	Rain	31.416667	1.0	1.000000	1.0	1.0

3]:

		correlation
temperature_precipIntensity	precipIntensity	0.999265
precipIntensity	temperature_precipIntensity	0.999265
apparentTemperatureHigh	apparentTemperatureMax	0.974866
apparentTemperatureMax	apparentTemperatureHigh	0.974866
temperatureMax	apparentTemperatureMax	0.956148
apparentTemperatureMax	temperatureMax	0.956148
sunriseTime	moonPhase	0.955418
moonPhase	sunriseTime	0.955418
temperatureMin	apparentTemperatureMin	0.953486