# Uber/ lyft Price Prediction

Brown University Data Science Institute
Winnie Zhang
10-19-2023

https://github.com/WWWW0203/data1030Project

# Introduction

Uber & Lyft - Ride Sharing company that bridge the gap between private transportation service and people needing a ride.

**Predict Uber/Lyft ride price based on features such as distance travelled, weather, time of day, surge multiplier, etc. (Regression)**

**Why is it important?**

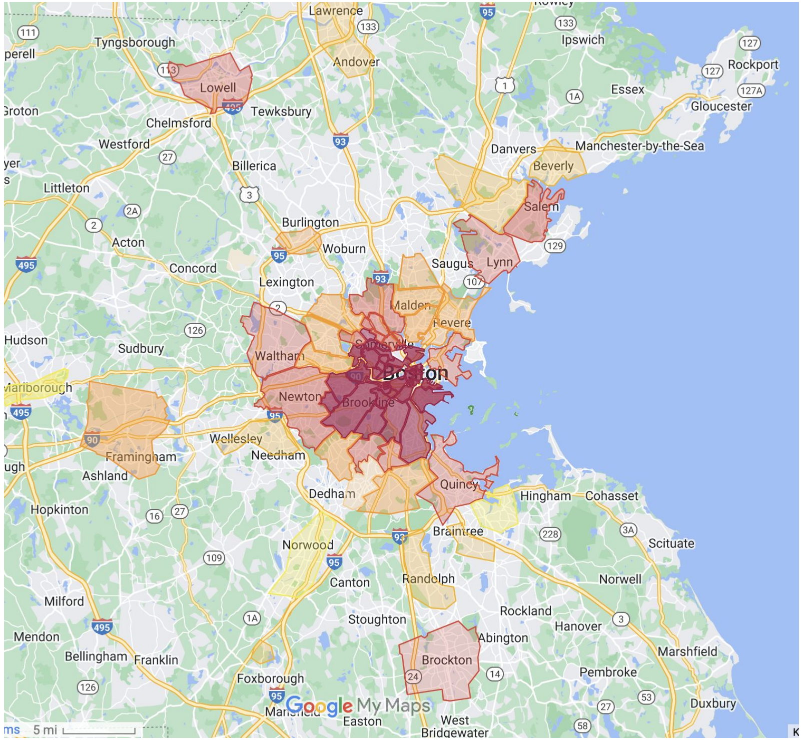
- Understanding how the prices of Uber & Lyft changes under different circumstances will give us better insights & help determine our travel plans
  - For example, it can help us decide whether to take uber/lyft, rent a car, or take other public transportations when we travel to different places.

# How Data Was Collected ?

- This dataset is not from Uber/Lyft because they do not make their data publicly available

- The author of the dataset collected real time data using Uber & Lyft API queries and corresponding weather conditions.

- Data was collected in a few hot spot in Boston (as shown in the map) for over a week from Nov.18, 2018.



Kaggle link of the data:
https://www.kaggle.com/datasets/ravi72munde/uber-lyft

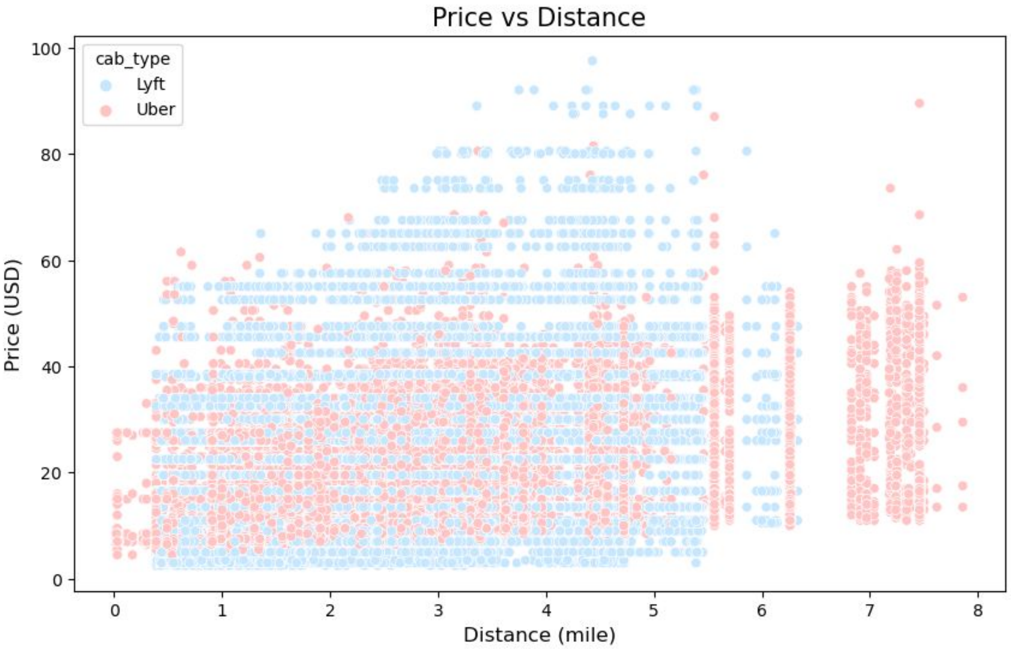
# Exploratory Data Analysis

**Our guess:**

- a clear positive linear relationship between distance and price

**Finding:**

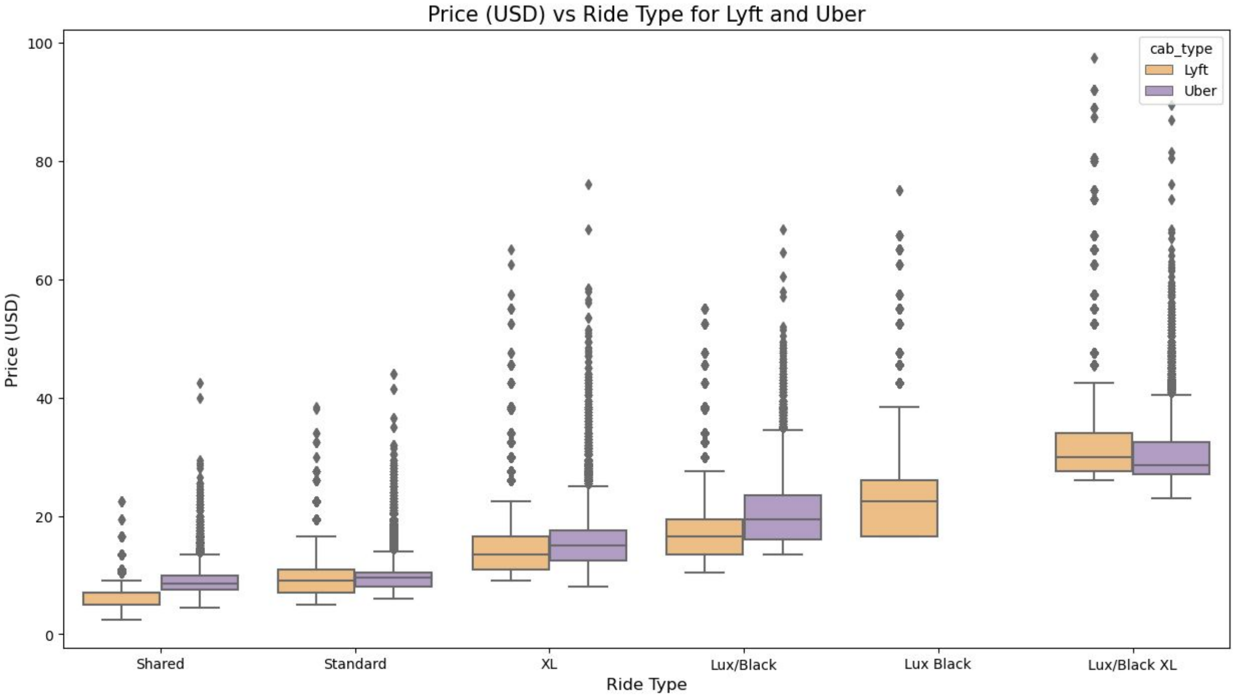
- Linear relationship seems not clear
- Uber price tends to be more affected by the distance of the trip than Lyft



# Exploratory Data Analysis

- As the ride type changes from ‘regular’ to ‘premium’, price of the trip also increases.

- Lyft tends to be cheaper overall than Uber for ‘regular’ ride types.

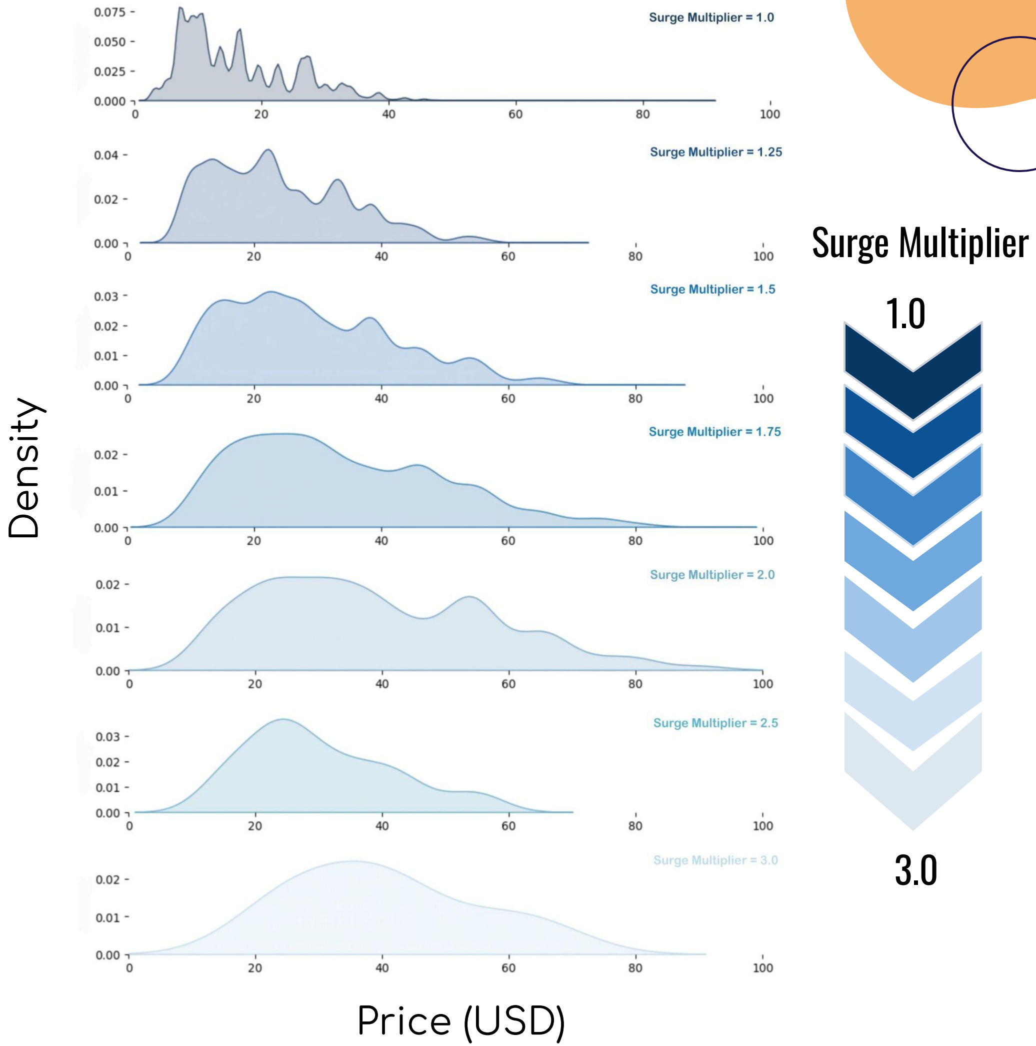


# Distribution of Price for Surge Multiplier from Lowest to Highest

- When demand for rides is higher than  the supply of cars, surge pricing comes in, increasing the price.

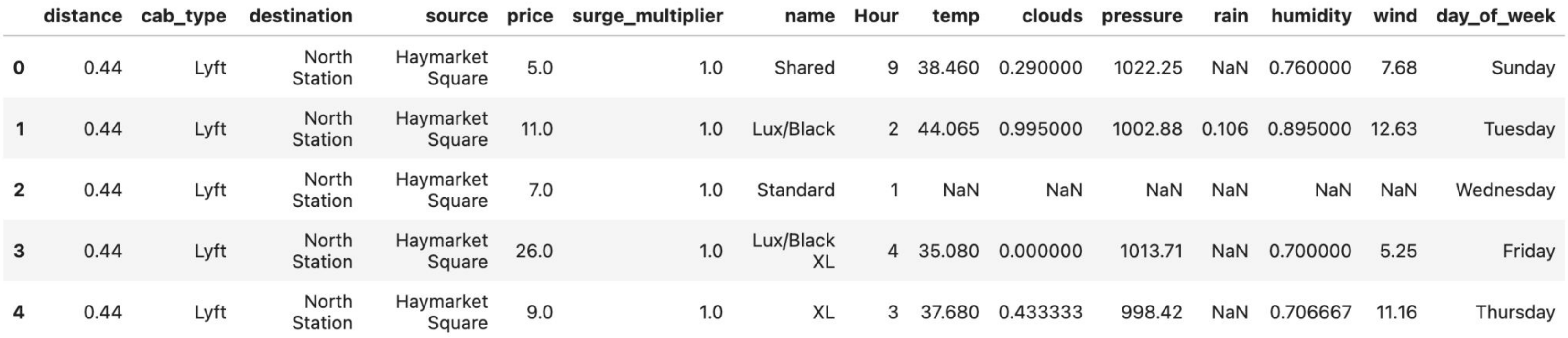
- Distribution of price is skewed to the right at the lowest surge multiplier …

- And becomes more like normal distribution as the surge multiplier increases.



# Splitting the Data

- Data Dimension: 637,976 x 15 (Large Data Set)
- Each row represents one trip (i.i.d)
- Target variable is price in USD, X variables are columns in the dataset excluding price.
- Basic split to split the data randomly into train (60%), validation (20%), and test set (20%)
  - Training: 382785 x 15, Validation: 127595 x 15,  Test: 127595 x 15



| | distance | cab_type | destination | source | price | surge_multiplier | name | Hour | temp | clouds | pressure | rain | humidity | wind | day_of_week |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.44 | Lyft | North Station | Haymarket Square | 5.0 | 1.0 | Shared | 9 | 38.460 | 0.290000 | 1022.25 | NaN | 0.760000 | 7.68 | Sunday |
| **1** | 0.44 | Lyft | North Station | Haymarket Square | 11.0 | 1.0 | Lux/Black | 2 | 44.065 | 0.995000 | 1002.88 | 0.106 | 0.895000 | 12.63 | Tuesday |
| **2** | 0.44 | Lyft | North Station | Haymarket Square | 7.0 | 1.0 | Standard | 1 | NaN | NaN | NaN | NaN | NaN | NaN | Wednesday |
| **3** | 0.44 | Lyft | North Station | Haymarket Square | 26.0 | 1.0 | Lux/Black XL | 4 | 35.080 | 0.000000 | 1013.71 | NaN | 0.700000 | 5.25 | Friday |
| **4** | 0.44 | Lyft | North Station | Haymarket Square | 9.0 | 1.0 | XL | 3 | 37.680 | 0.433333 | 998.42 | NaN | 0.706667 | 11.16 | Thursday |

# Preprocessing

# OrdinalEncoder

### Ride Type

Shared

Standard

XL

Luxury or Black

Luxury and Black

Luxury or Black XL

### Surge Multiplier

1.0

1.25

1.5

1.75

2.0

2.5

3.0

### Hour of Day

0 (12 am)

1

2

⋮

23 (11 pm)

### Day of Week

Monday

Tuesday

Wednesday

Thursday

Friday

Saturday

Sunday

# OneHotEncoder

## Cab Type

Uber
Lyft

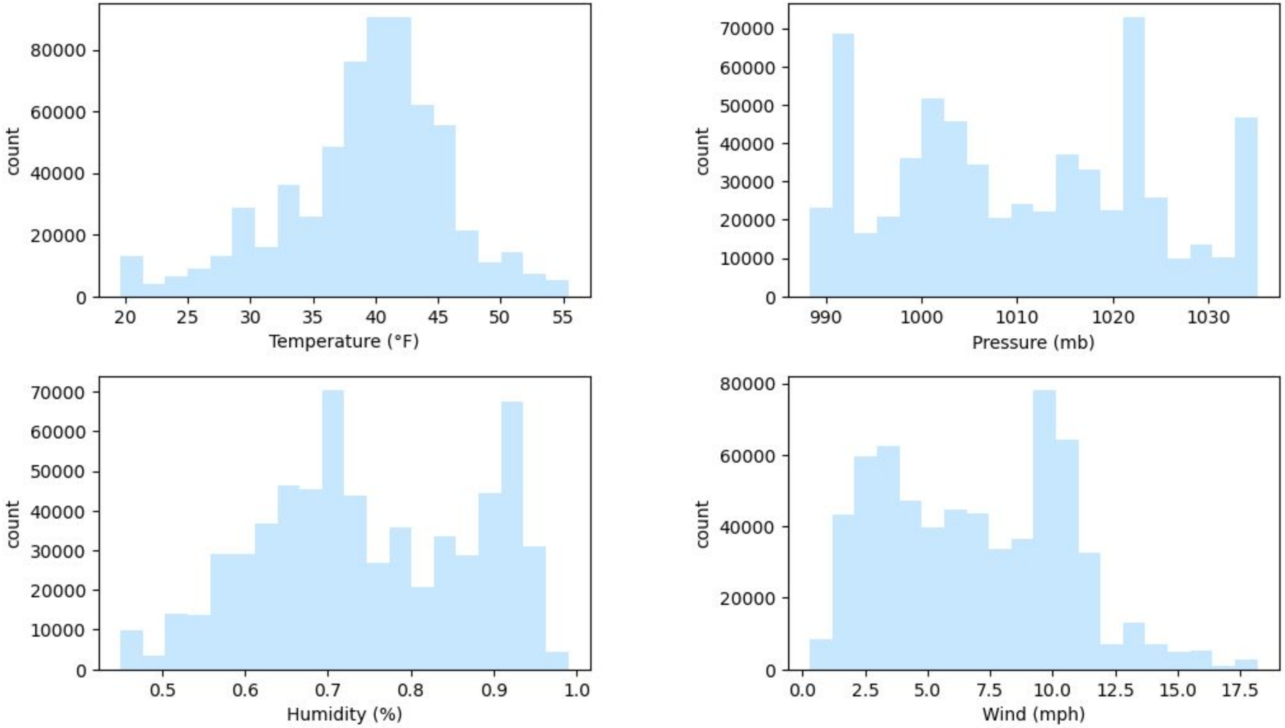
## Destination

Financial District
Back Bay
Theatre District
Haymarket Square
Boston University
Fenway
Northeastern University
North End
South Station
West End
Beacon Hill
North Station

## Source

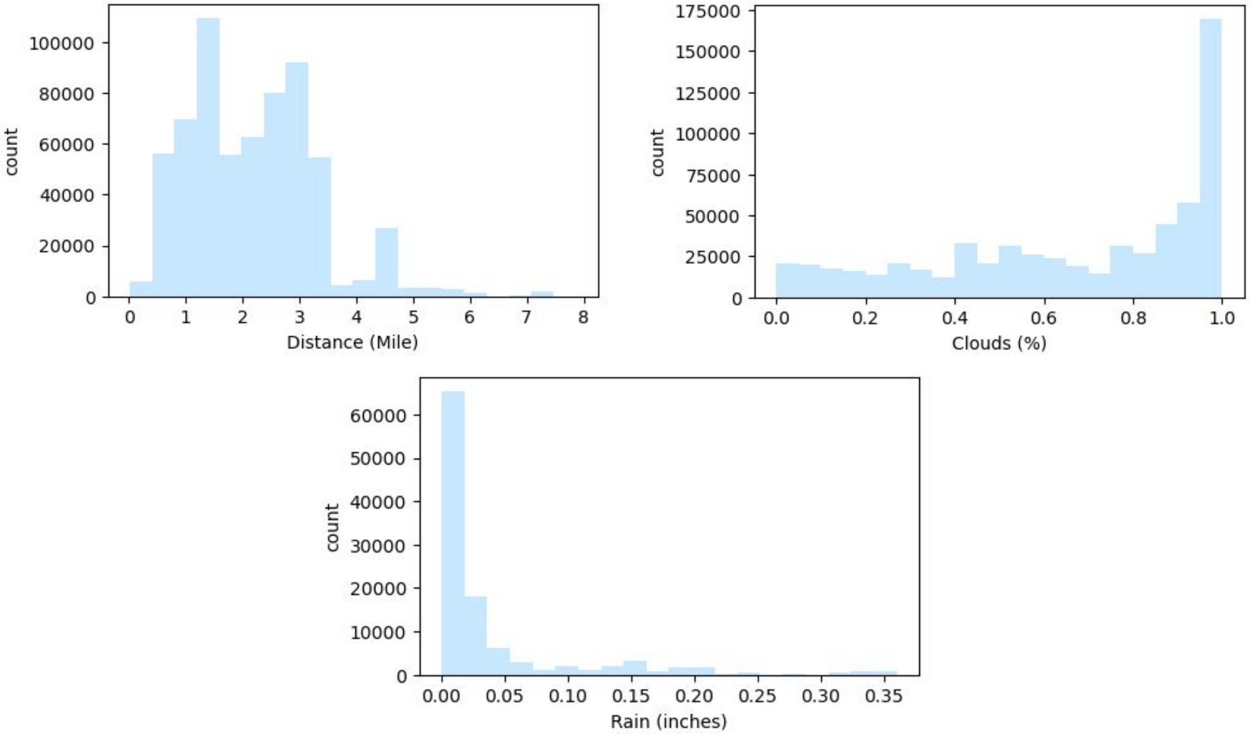
Financial District
Back Bay
Theatre District
Boston University
North End
Fenway
Northeastern University
South Station
Haymarket Square
West End
Beacon Hill
North Station

# MinMaxScaler



Distribution plot of Temperature, Pressure, Humidity, and Wind (Continuous & Bounded in a range)

# StandardScaler



Distribution plot of Distance, Clouds, and Rain (Heavy Tailed Distribution)

# Data Dimensions & Missing Data

Training Set Before Preprocessing: 14 features
Training Set After Preprocessing: 37 features

**Fraction of missing values in features (continuous):**

| | |
|---|---|
| Temperature: | 4% |
| Clouds: | 4% |
| Pressure: | 4% |
| Humidity: | 4% |
| Wind: | 4% |
| Rain: | 83% |

83% of all data points have missing values.

# THANKS

## Questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**