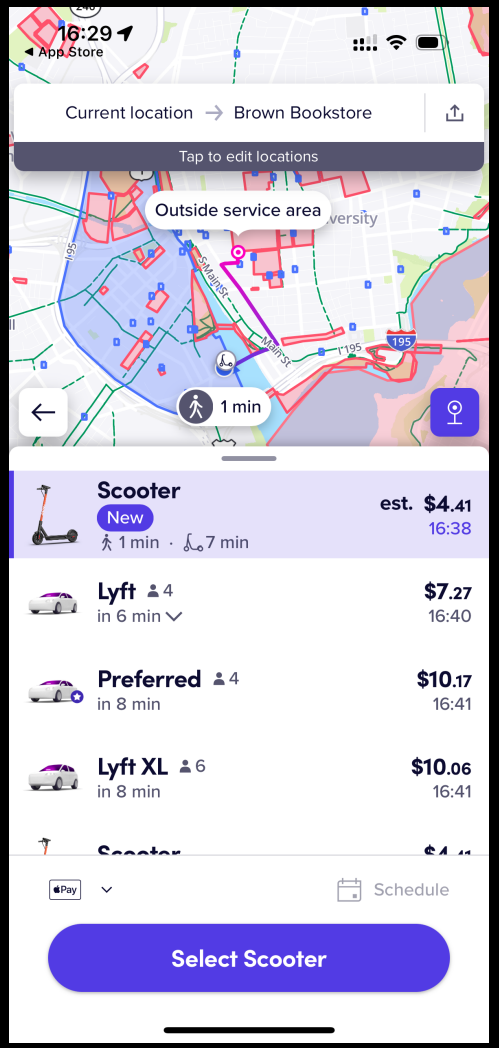
Uber & Lyft: Can we discover the secret behind dynamic pricing?

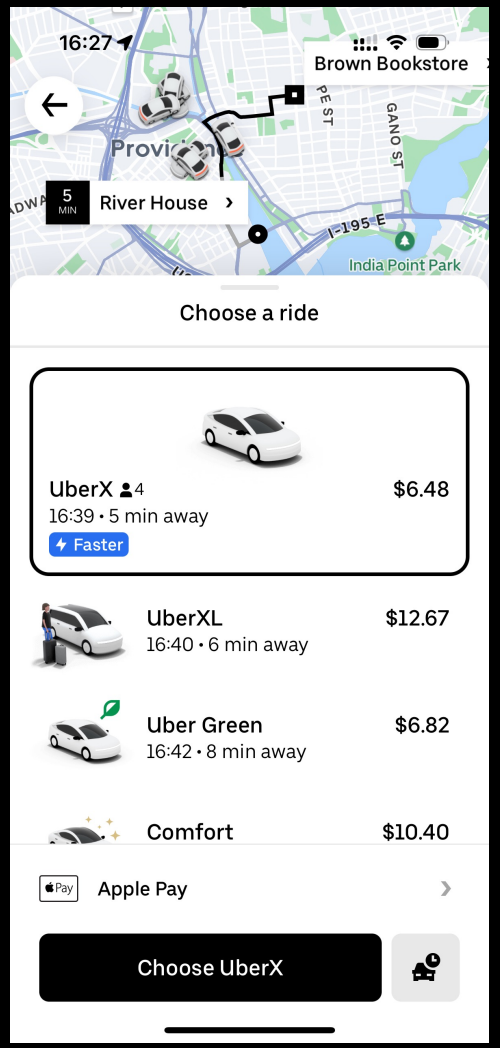
https://stablediffusionweb.com/#demo

PROJECT MIDTERM PRESENTATION
FOR DATA 1030

Yijia Xue
10/18/2023
Brown University
Repo link: https://github.com/Ikea-179/Uber_Lyft_price_prediction#uber_lyft_price_prediction

INTRODUCTION





- How exactly are these prices calculated?
- Why do the prices vary for the same journey on different dates?
- On what weather conditions is it most cost-effective for me to travel?
- Is there a pattern to their dynamic pricing rules?

Predict and compare the price of Uber and Lyft rideshares -> Regression Problem

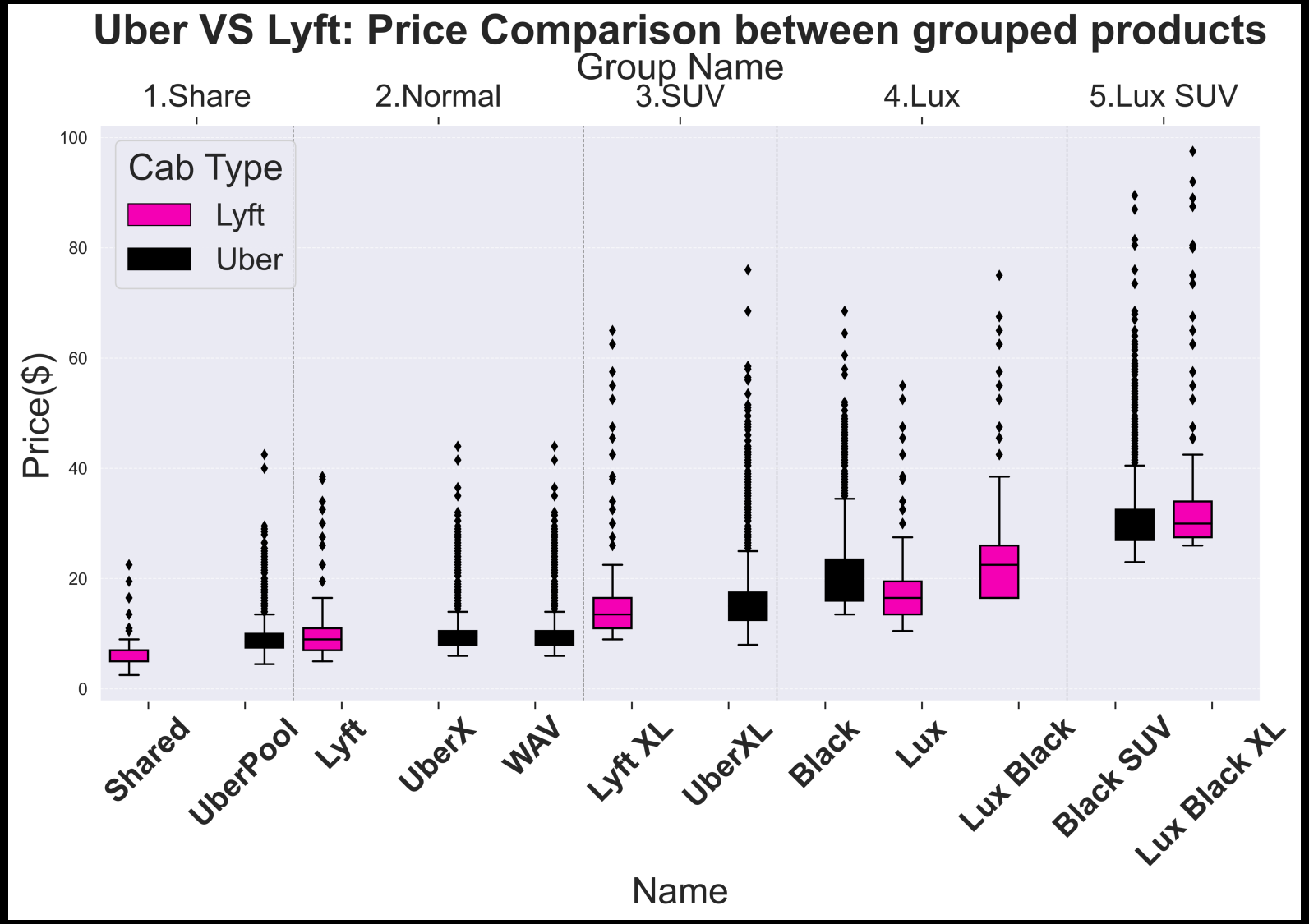
Exploratory Data Analysis I

Dataset Intro

The data is from Kaggle: Uber and Lyft Dataset Boston, MA

- 693071 data points with 57 features
- Price, Source and Destination, time, weather
- Redundant information about weather: about 30 columns weather information
- Wrong information in geo location: latitude, longitude
- Missing value in target variable: price

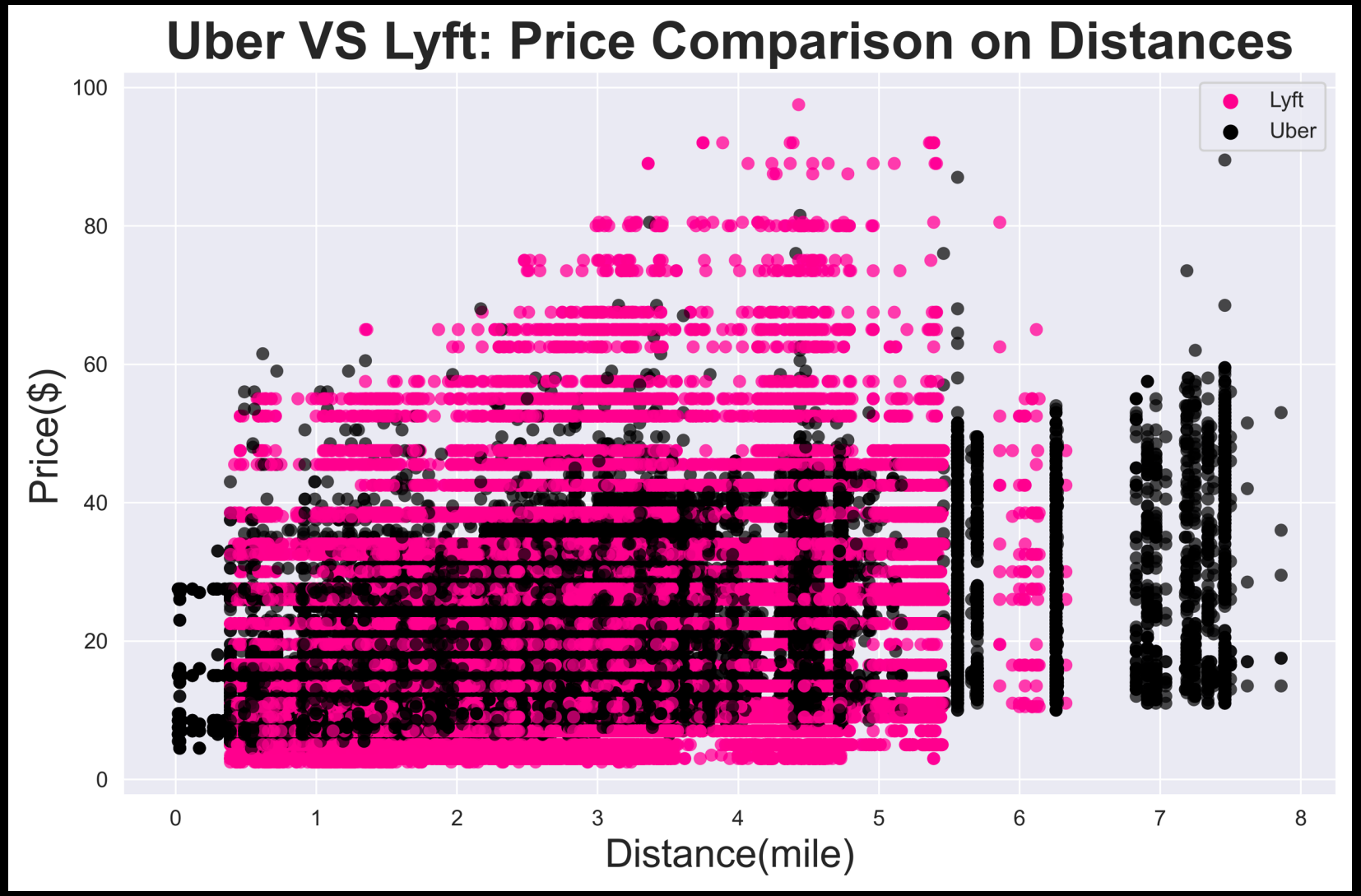
Exploratory Data Analysis II: Visualization I



Lyft is
Cheaper than Uber for shared, and SUV cars
Similar price in Normal cars
Giving more options in Luxury cars
More expensive in Luxury SUV car.

There are lots of outlier for each type

Exploratory Data Analysis II: Visualization II



Lyft has smaller range in distance larger gap between price levels

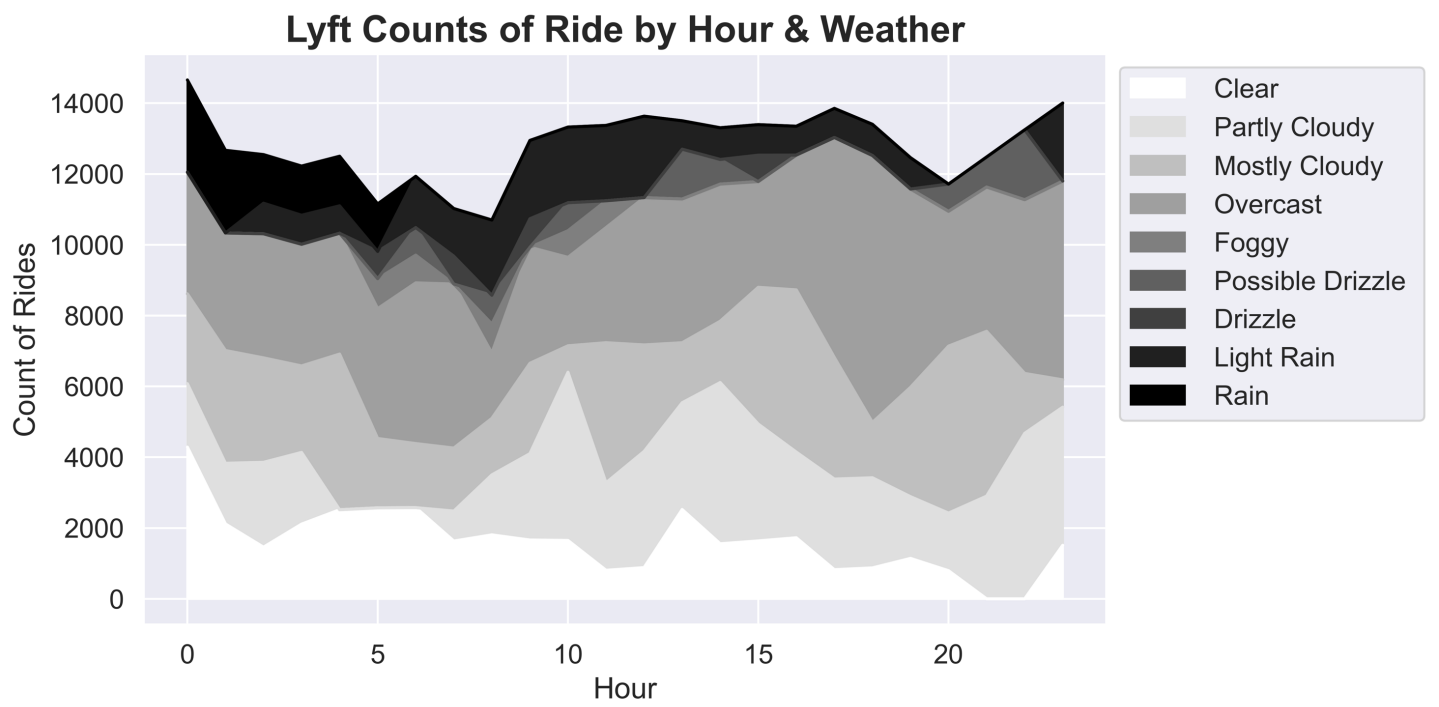
Price and Distance are moderately correlated

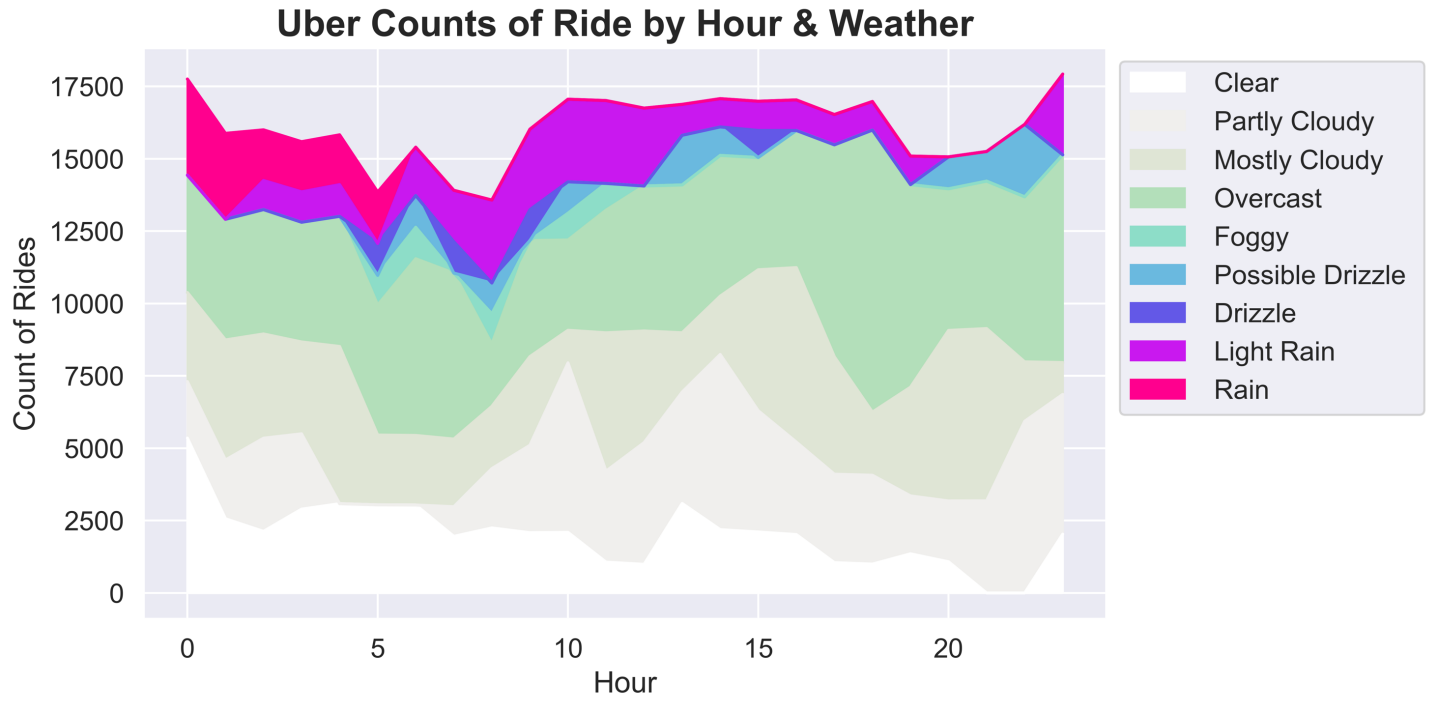
Exploratory Data Analysis II
Visualization III

Peak Hour Period:
10am-19pm, 10pm-1am

Pattern:
Similar between Uber and Lyft

Weather:
It‘s mostly gloomy in Boston in October/November when people took a ride





Splitting and Processing

Splitting

- iid dataset: each row contains info for one ride
- KFold(n_splits=3,shuffle=True)

Preprocess

- Data shape before preprocessing & splitting: (693071, 57)
- Extract the day_of_week column from the datetime variable
- Extract whether holiday/weekend column from the datetime variable
- Extract whether morning/evening/night from the hour variable
- Drop several redundant columns
- Correct the Geographical info for ride location visualization
- Dropping the missing value for target variable: price

Splitting and Processing

Preprocess

- Using onehot encoder for categorical columns, Ordinal encoder for ordinal columns and Standard Scaler for continuous columns
  9 categorical columns:'hour','day','month','day_of_week','cab_type','destination','source','short_summary','name'
  1 ordinal column: cab_type_group
  11 numerical columns:'surge_multiplier','distance','temperature','humidity','windSpeed','windGust','visibility','pressure','cloudCover','moonPhase','precipIntensityMax'

- Data shape after preprocessing & splitting : X_train:(340254, 108),X_val: (170126, 108), X_test:(127596, 108)

Thank you

https://stablediffusionweb.com/#demo