Project Summary

Maksuda Toma, Israt Zarin, Shadman Shakib

March 16, 2025

Context

Uber and Lyft's ride prices are not constant like public transport. They are greatly affected by the demand and supply of rides at a given time. So what exactly drives this demand? The first guess would be the time of the day; times around 9 am and 5 pm should see the highest surges on account of people commuting to work/home. Another guess would be the weather; rain/snow should cause more people to take rides.

Content

With no public data of rides/prices shared by any entity, we tried to collect real-time data using Uber&Lyft api queries and corresponding weather conditions. We chose a few hot locations in Boston from this map We built a custom application in Scala to query data at regular intervals and saved it to DynamoDB. The project can be found here on GitHub We queried cab ride estimates every 5 mins and weather data every 1 hr.

The data is approx. for a week of Nov '18 (I actually have included data collected while I was testing the 'querying' application so might have data spread out over more than a week. I didn't consider this as a time-series problem so did not worry about regular interval. The chosen interval was to query as much as data possible without unnecessary redundancy. So data can go from end week of Nov to few in Dec)

The Cab ride data covers various types of cabs for Uber & Lyft and their price for the given location. You can also find if there was any surge in the price during that time. Weather data contains weather attributes like temperature, rain, cloud, etc for all the locations taken into consideration.

Objective

Our aim was to try to analyze the prices of these ride-sharing apps and try to figure out what factors are driving the demand. Do Mondays have more demand than Sunday at 9 am? Do people avoid cabs on a sunny day? Was there a Red Sox match at Fenway that caused more people coming in? We have provided a small dataset as well as a mechanism to collect more data. We would love to see more conclusions drawn.

Problem Statement

The transformations in urban transportation through Uber and Lyft ridesharing have been significant, yet customers experience uncertainties in fare pricing because of changing factors. The proposed model development will create predictions for taxi fares by assessing relevant parameters, which include scheduling data alongside customer demand together with weather elements and transportation specifications. Price variations within the historical data become understandable through the model, which enables service providers and customers to make smart choices.

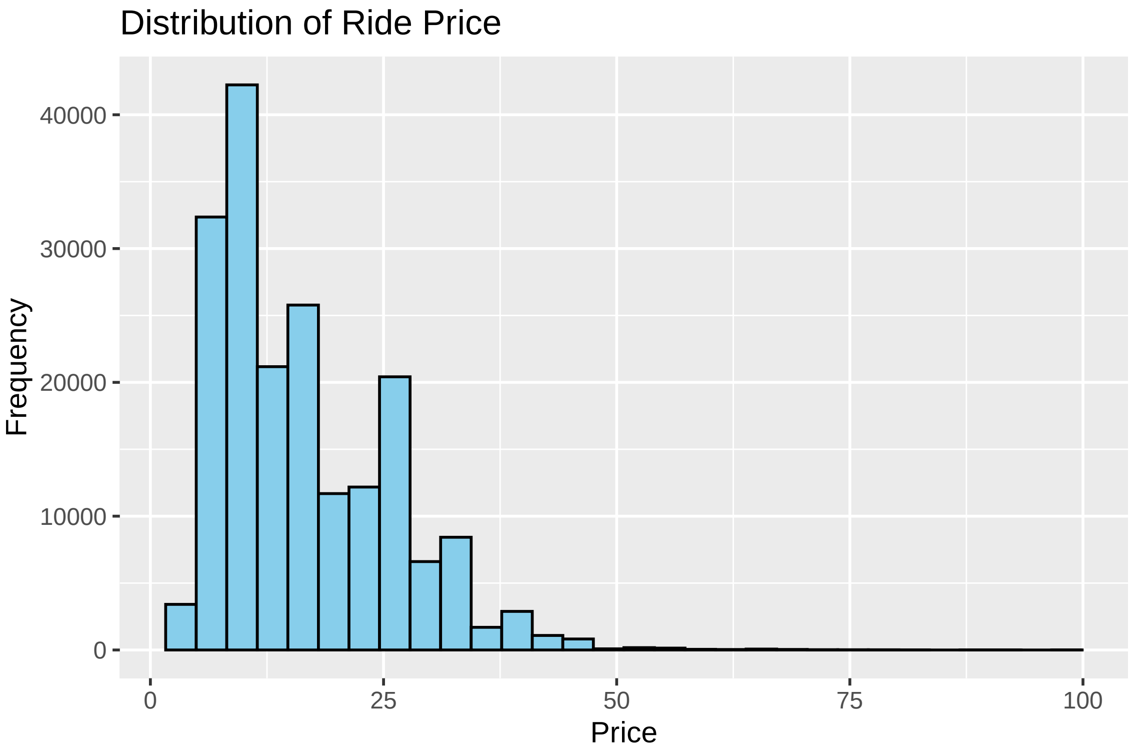
Operating with extensive data sets poses challenges because processing becomes complex and leads to suboptimal model accuracy because of data abnormalities together with database capacity limitations. The solution to these problems depends on selecting essential features alongside effective data management practices and using resistant machine learning approaches.

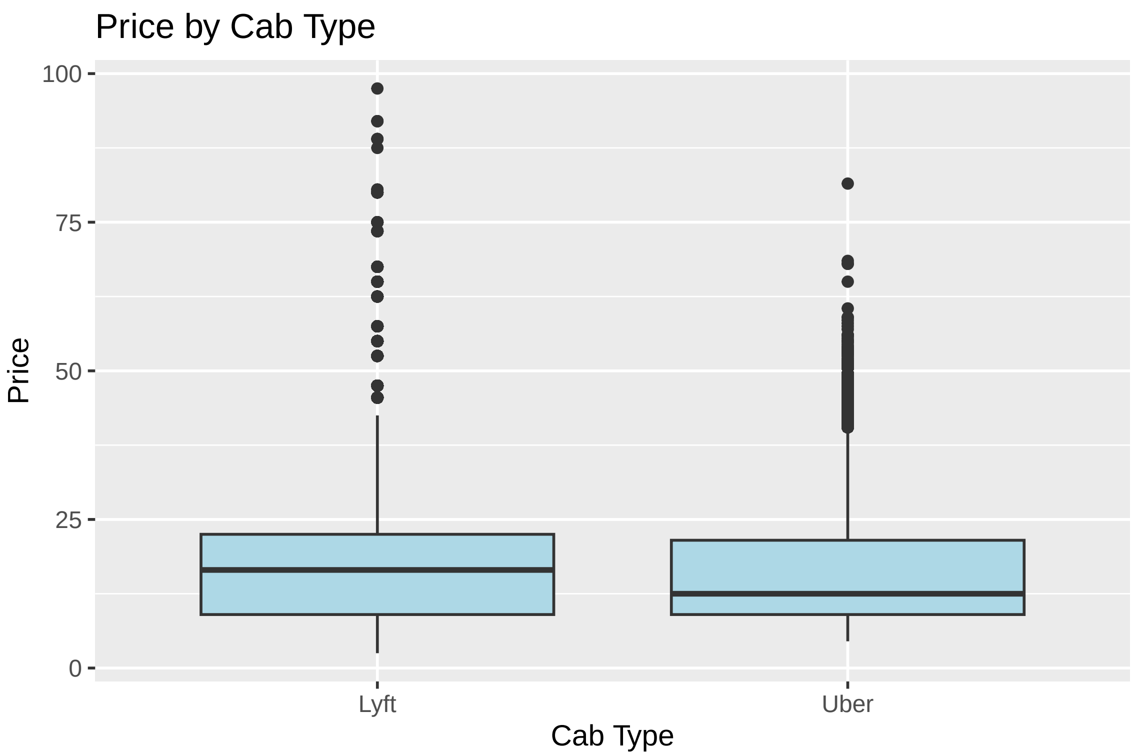
2. Dataset and Preprocessing

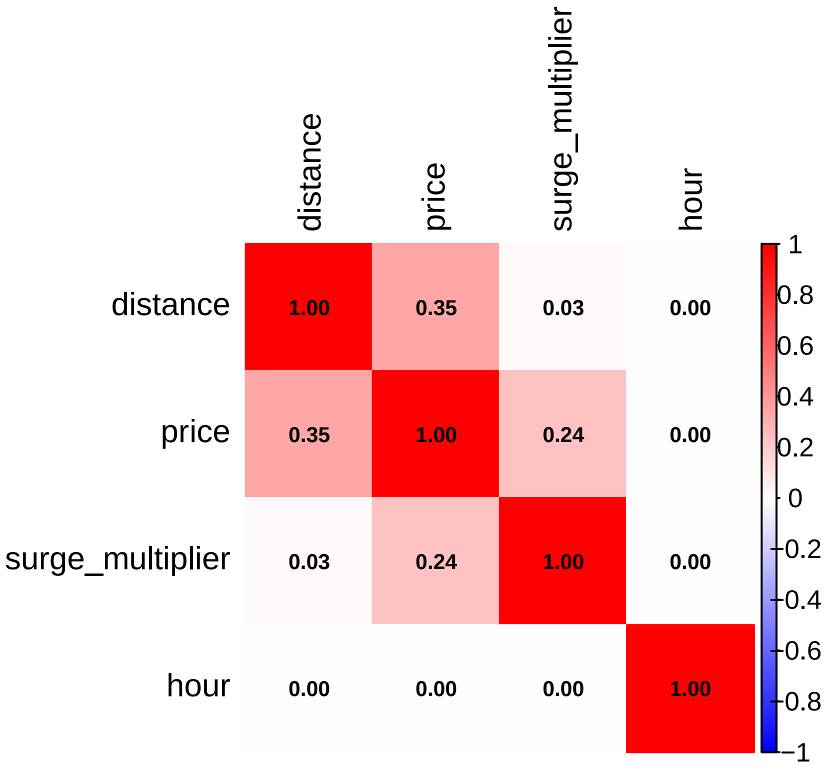
```
#| echo: false

# Load libraries
library(readr)
library(dplyr)
library(caret)

# Load data
data <- read.csv("Reduced_data.csv")
head(data)
Reduced_data <- read_csv("Reduced_data.csv")
head(Reduced_data)
plot(Reduced_data)
summary(Reduced_data)
```

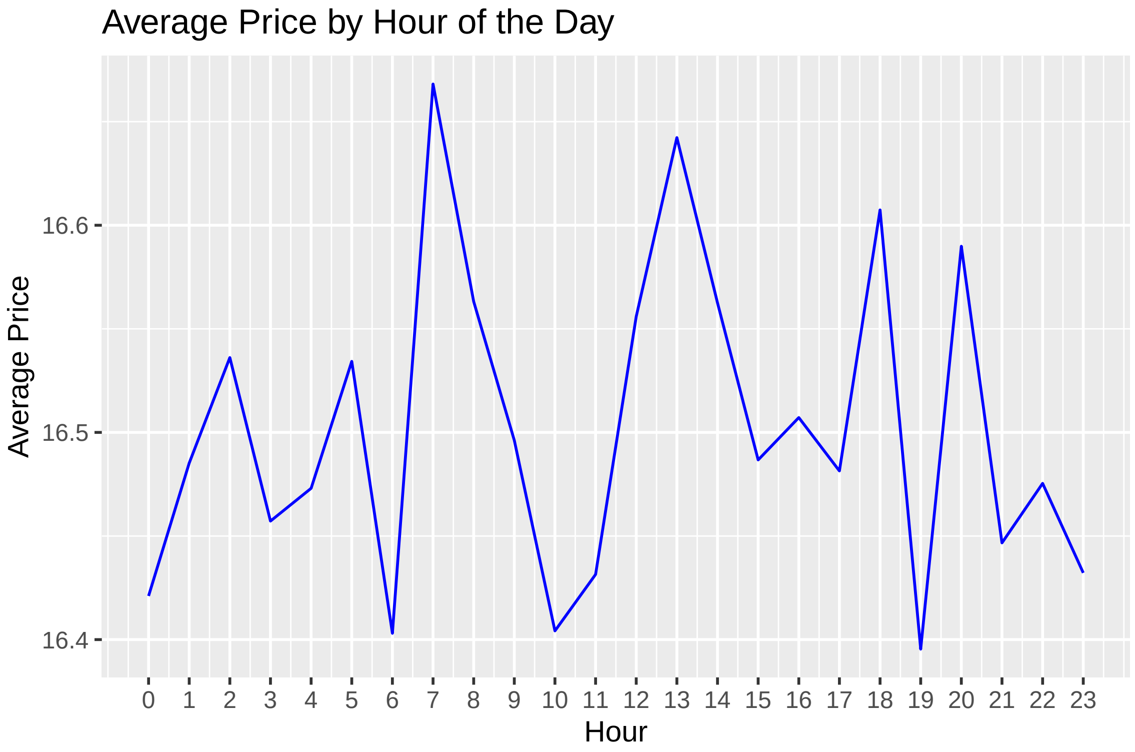
Missing Value imputation

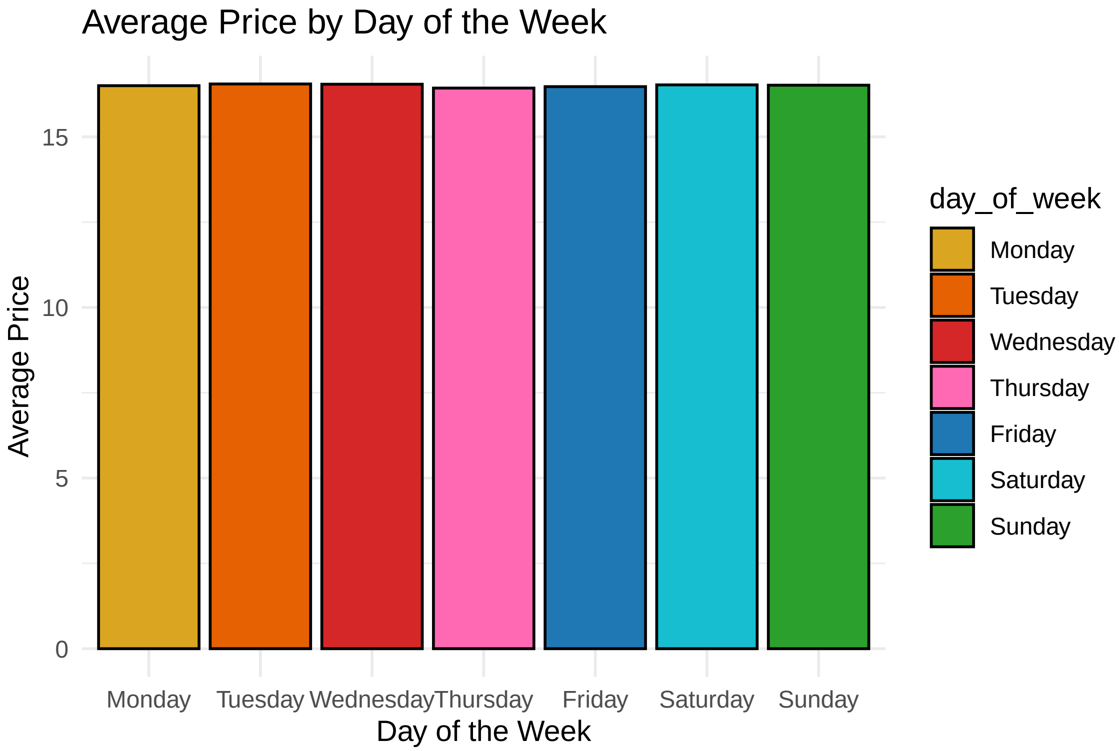
```
# Load libraries
library(ggplot2)

# Calculate missing value percentage
missing_percent <- colSums(is.na(data)) / nrow(data) * 100

# Create bar plot
ggplot(data = data.frame(Variable = names(missing_percent), Percent = missing_percent),
       aes(x = reorder(Variable, -Percent), y = Percent)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Percentage of Missing Values per Column", x = "Variable", y = "Percentage
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Price by Cab Type
Price
100
75
50
25
0
Lyft
Uber
Cab Type


distance
price
surge_multiplier
hour
distance
1.00
0.35
0.03
0.00
price
0.35
1.00
0.24
0.00
surge_multiplier
0.03
0.24
1.00
0.00
hour
0.00
0.00
0.00
1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Average Price by Hour of the Day
Average Price
16.6
16.5
16.4
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour



Models Used

```
# Train Linear Regression
lm_model <- lm(price ~ ., data = train)

# Train Random Forest
library(randomForest)
rf_model <- randomForest(price ~ ., data = train)

# Predict
lm_pred <- predict(lm_model, test)
lm_pred
rf_pred <- predict(rf_model, test)
```

Performance Metrics and Model Evaluation

```
# Load libraries
library(Metrics)

# Linear Regression Metrics
mae(lm_pred, test$price)
mse(lm_pred, test$price)
R2(lm_pred, test$price)

# Random Forest Metrics
mae(rf_pred, test$price)
mse(rf_pred, test$price)
R2(rf_pred, test$price)
```

Justification for Model Choice