Factors Influencing Ride-Sharing Pricing in Boston: A Study on Lyft and Uber

6101 Intro to Data Science Midterm Presentation

Aaron Yang Jianjun Gao Luhuan Wang

Why this topic?

1. Ride-sharing Popularity
2. Easy Data Access
3. Dynamic Pricing Model
4. Comparative Analysis

Uber

lyft

Team Contributions

Aaron Yang	Jianjun Gao	Luhuan Wang
Outline & Proposal	Data Collection	Data Summary
Data Analysis(with Jianjun)	Data Analysis(with Aaron)	Slide Creation
Presentation	Presentation	Presentation
Paper Co-author	Paper Co-author	Paper Co-author

Key Concepts & Principles

Data Science and Statistical Methods	Programming	Visualization
Data Collection & Cleaning	R	Graphical Representation
Exploratory Data Analysis (EDA)	Python	
Regression Model		

Project Workflow Overview 1

Step No.	Phase	Tasks & Details
1.	Topic Decision	- Team Meeting - Choose a topic based on available data
2.	Data Collection	- Identify Data Sources - Gather data - Initial Data Cleaning
3.	Research & Background	- Literature Review - Industry Knowledge
4.	Data Processing & EDA	- Data Cleaning & Transformation - Descriptive Statistics - Visualization
5.	Data Analysis & Modeling	- Hypothesis Testing - Regression Analysis - Advanced Modeling - Model Evaluation

Project Workflow Overview 2

Step No.	Phase	Tasks & Details
6.	Interpretation of Results	- Discuss Findings - Statistical Significance
7.	Presentation Preparation	- Slide Development - Rehearsal
8.	Presentation & Feedback	- Deliver Presentation - Collect Feedback
9.	Final Paper	- Write a paper

Data Collection & Filtering

BM · UPDATED 4 YEARS AGO

123 | New Notebook | Download (47 MB)

Uber and Lyft Dataset Boston, MA

From 11-26-2018 to 12-18-2018

VS

Data Card | Code (23) | Discussion (3)

About Dataset

Uber vs Lyft

This is a very beginner-friendly dataset. It does contain a lot of NA values. It is a good dataset if you want to use a Linear Regression Model to see the pattern between different predectors such as `hour` and `price`.

A really amazing part of this dataset is that I have included the corresponding `weather data` for that `hour` with a short summary of the `weather`. Other important factors are `temperature`, `wind`, and `sunset`.

Usability
9.12

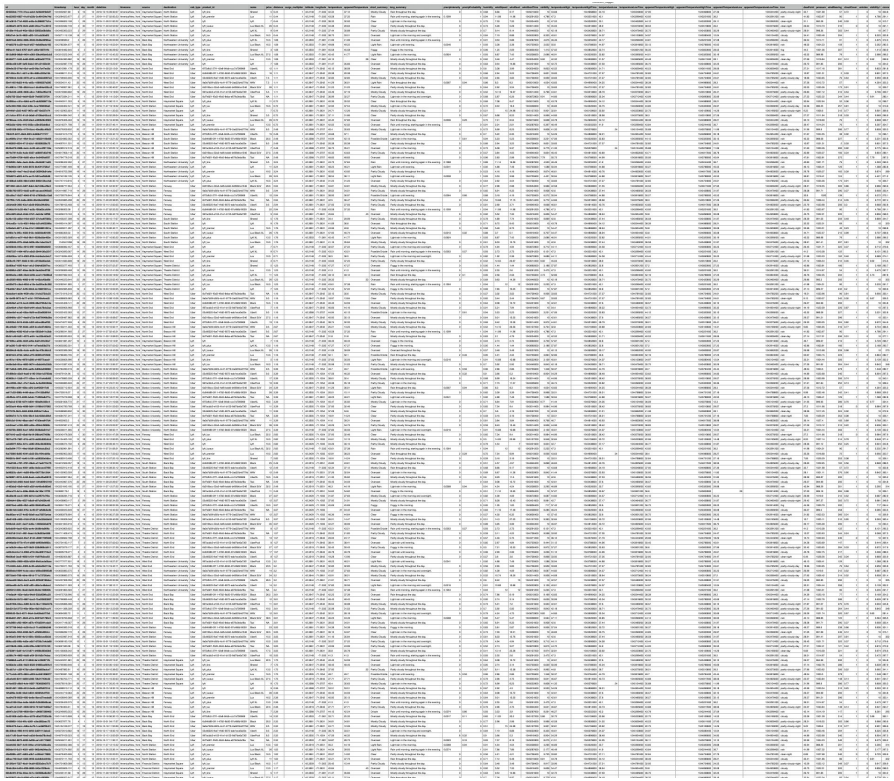
License
CC0: Public Domain

Expected update frequency
Quarterly

Tags
Travel

Kaggle DataBase

Data Collection & Filtering



More than 300,000 rows and 55 columns data

Uber

lyft

Data Collection & Filtering

Original Dataset

	id	timestamp	hour	day	month	datetime	timezone	source	destination	cab_type	...	precipIntensityMax	uvIndexTime	temperatureMin	temperatureMinTime	temperatureMax	temperatureMaxTime	apparentTemperatureMin	apparentTemperatureMinTime	apparentTempe
0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1.544953e+09	9	16	12	2018-12-16 09:30:07	America/New_York	Haymarket Square	North Station	Lyft	...	0.1276	1544979600	39.89	1545012000	43.68	1544968800	33.73	1545012000	
1	4bd23055-6827-41c6-b23b-3c491f24e74d	1.543284e+09	2	27	11	2018-11-27 02:00:23	America/New_York	Haymarket Square	North Station	Lyft	...	0.1300	1543251600	40.49	1543233600	47.30	1543251600	36.20	1543291200	
2	981a3613-77af-4620-a42a-0c0866077d1e	1.543367e+09	1	28	11	2018-11-28 01:00:22	America/New_York	Haymarket Square	North Station	Lyft	...	0.1064	1543338000	35.36	1543377600	47.55	1543320000	31.04	1543377600	
3	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	1.543554e+09	4	30	11	2018-11-30 04:53:02	America/New_York	Haymarket Square	North Station	Lyft	...	0.0000	1543507200	34.67	1543550400	45.03	1543510800	30.30	1543550400	
4	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	1.543463e+09	3	29	11	2018-11-29 03:49:20	America/New_York	Haymarket Square	North Station	Lyft	...	0.0001	1543420800	33.10	1543402800	42.18	1543420800	29.11	1543392000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
693066	616d3611-1820-450a-9845-a9ff304a4842	1.543708e+09	23	1	12	2018-12-01 23:53:05	America/New_York	West End	North End	Uber	...	0.0000	1543683600	31.42	1543658400	44.76	1543690800	27.77	1543658400	
693067	633a3fc3-1f86-4b9e-9d48-2b7132112341	1.543708e+09	23	1	12	2018-12-01 23:53:05	America/New_York	West End	North End	Uber	...	0.0000	1543683600	31.42	1543658400	44.76	1543690800	27.77	1543658400	
693068	64d451d0-639f-47a4-9b7c-6fd92fbd264f	1.543708e+09	23	1	12	2018-12-01 23:53:05	America/New_York	West End	North End	Uber	...	0.0000	1543683600	31.42	1543658400	44.76	1543690800	27.77	1543658400	
693069	727e5f07-a96b-4ad1-a2c7-9abc3ad55b4e	1.543708e+09	23	1	12	2018-12-01 23:53:05	America/New_York	West End	North End	Uber	...	0.0000	1543683600	31.42	1543658400	44.76	1543690800	27.77	1543658400	
693070	e7fdc087-fe86-40a5-a3c3-3b2a8badcbda	1.543708e+09	23	1	12	2018-12-01 23:53:05	America/New_York	West End	North End	Uber	...	0.0000	1543683600	31.42	1543658400	44.76	1543690800	27.77	1543658400	

693071 rows × 57 columns

Data Collection & Filtering

Drop null values, random sampling(random_state = 42),choose 6 columns

	hour	price	distance	surge_multiplier	cab_type	name
0	6	7.0	4.51	1.0	Lyft	Shared
1	0	10.5	2.80	1.0	Uber	WAV
2	12	7.0	1.09	1.0	Lyft	Lyft
3	3	15.5	0.92	1.0	Uber	Black
4	9	16.5	1.12	1.0	Lyft	Lux Black
...	...	...	...	...	...	...
5995	3	19.5	2.32	1.0	Uber	Black
5996	18	26.0	0.54	1.0	Lyft	Lux Black XL
5997	10	14.0	1.89	1.0	Uber	UberX
5998	13	9.0	2.08	1.0	Lyft	Lyft
5999	4	22.5	2.36	1.0	Uber	Black

6000 rows × 6 columns

Columns Adjustments

cab_type <chr>	name <chr>
Lyft	Shared
Uber	WAV
Lyft	Lyft
Uber	Black
Lyft	Lux Black
Uber	WAV

cab_company <chr>	cab_type <chr>
Lyft	Shared
Uber	WAV
Lyft	Lyft
Uber	Black
Lyft	Lux Black
Uber	WAV

Descriptive Stats

```
unique(df$cab_company) #see unique values in company
```

```
## [1] "Lyft" "Uber"
```

```
unique(df$cab_type) #see unique values in cab_type
```

```
##  [1] "Shared"       "WAV"          "Lyft"         "Black"        "Lux Black"
##  [6] "Lyft XL"      "Lux"          "UberX"        "Black SUV"    "UberPool"
## [11] "Lux Black XL" "UberXL"
```

```
summary(df) # descriptive stats
```

```
##       hour          price          distance     surge_multiplier
##  Min.   : 0.0   Min.   : 2.5   Min.   :0.02   Min.   :1.00
##  1st Qu.: 6.0   1st Qu.: 9.0   1st Qu.:1.30   1st Qu.:1.00
##  Median :12.0   Median :13.5   Median :2.19   Median :1.00
##  Mean   :11.7   Mean   :16.5   Mean   :2.20   Mean   :1.02
##  3rd Qu.:18.0   3rd Qu.:22.5   3rd Qu.:2.94   3rd Qu.:1.00
##  Max.   :23.0   Max.   :73.5   Max.   :7.46   Max.   :2.50
##    cab_type             name
##  Length:6000        Length:6000
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
```

```{r, results='markup', echo=TRUE}
# measure of variance
sd(df$price)
sd(df$distance)
sd(df$surge_multiplier)
```

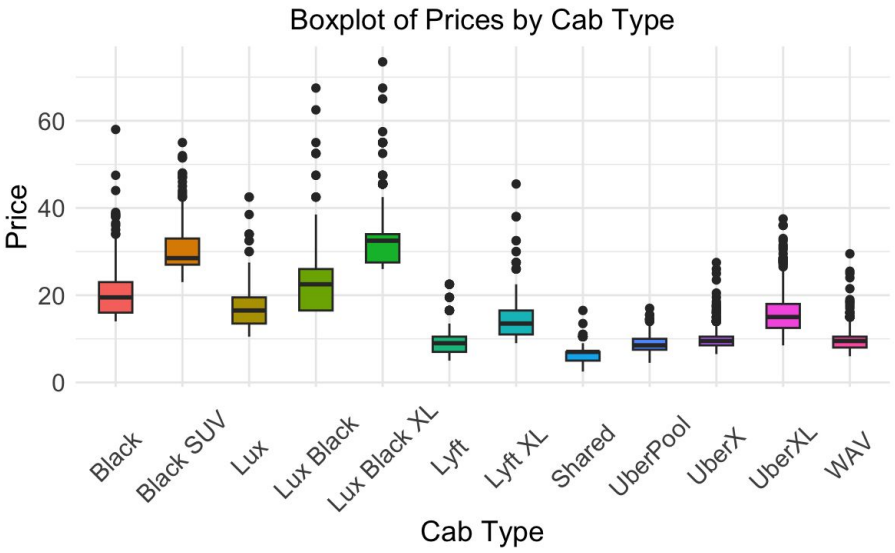
```
[1] 9.43
[1] 1.14
[1] 0.103
```

Data Visualization

0-8: Relatively High

15-23: Relatively Low

Data Visualization



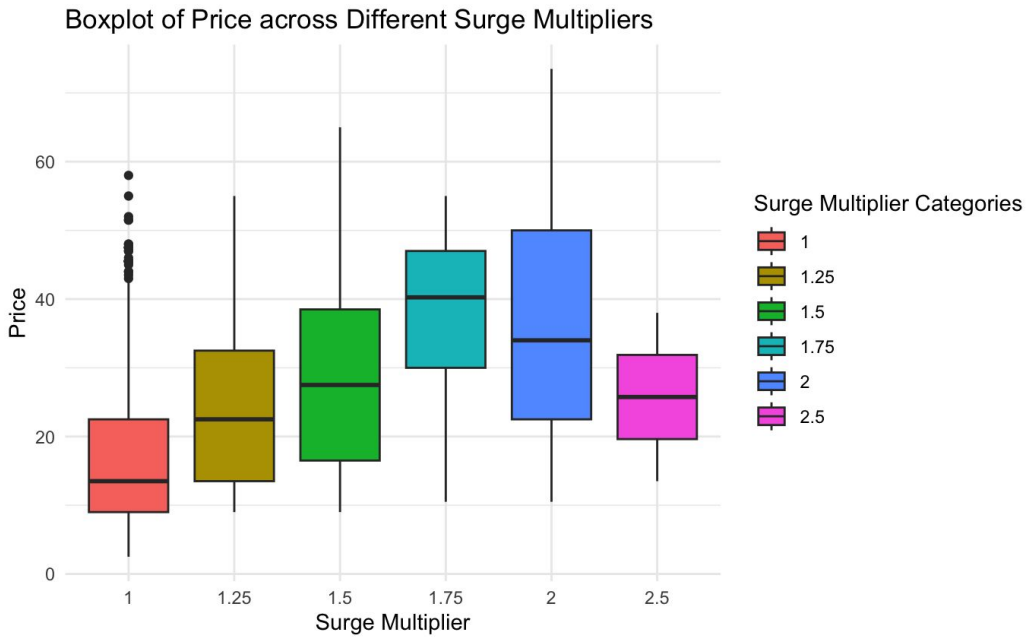
Highest Price:

Black SUV, Lux Black XL

Lowest Price:

Shared

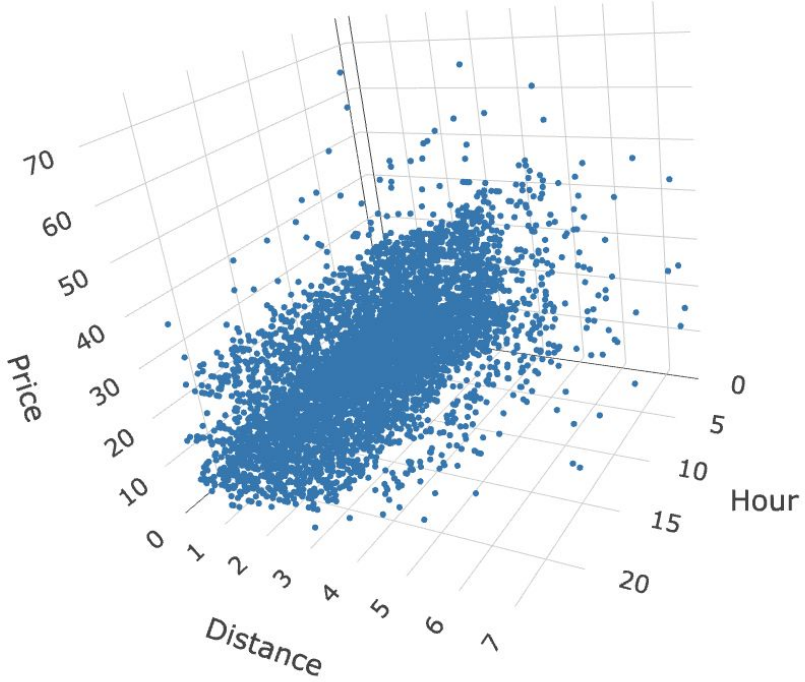
Data Visualization



Highest Median when Tipped 1.75

Lowest Median when Tipped 1

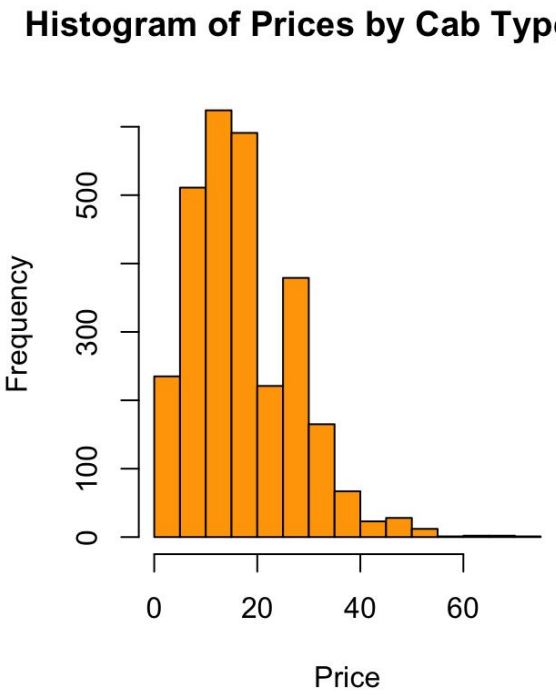
Data Visualization

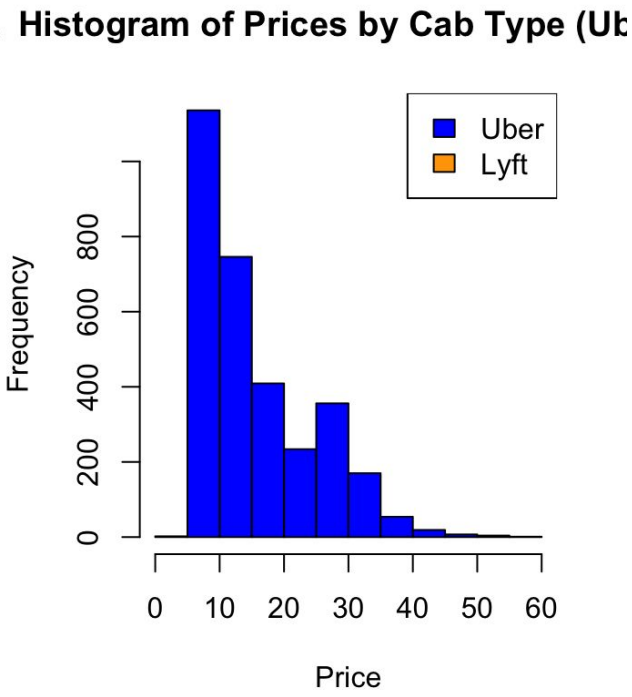


Data Visualization

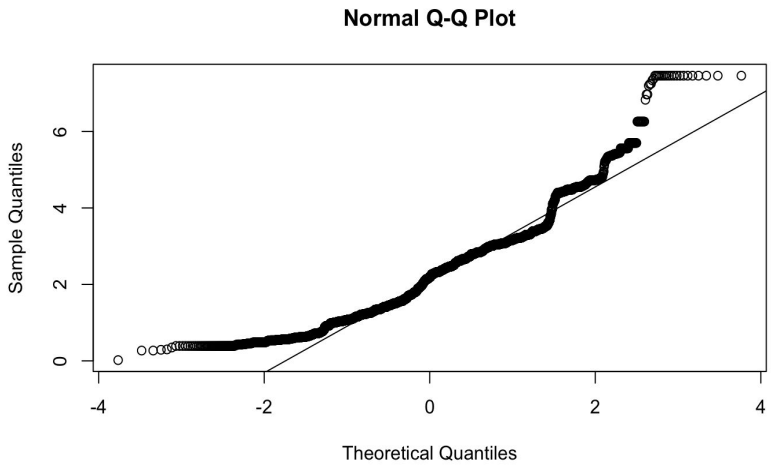
A Trend That is not Obvious

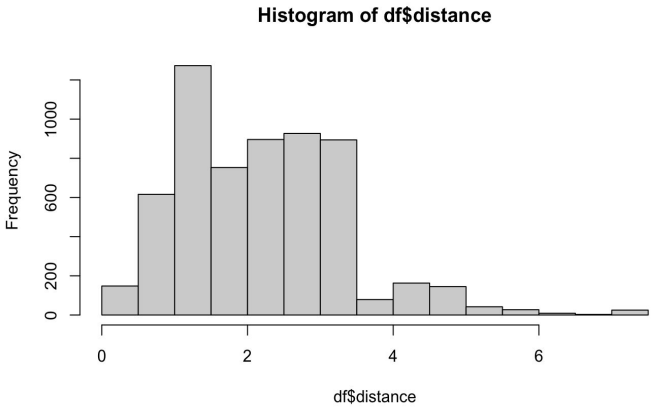
Data Visualization

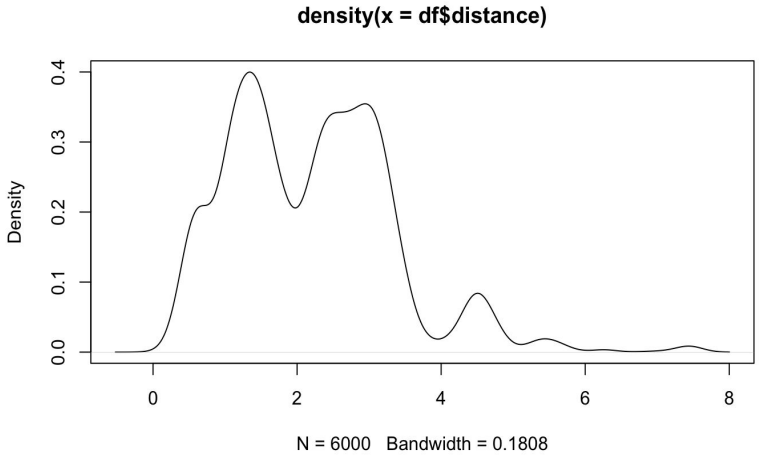




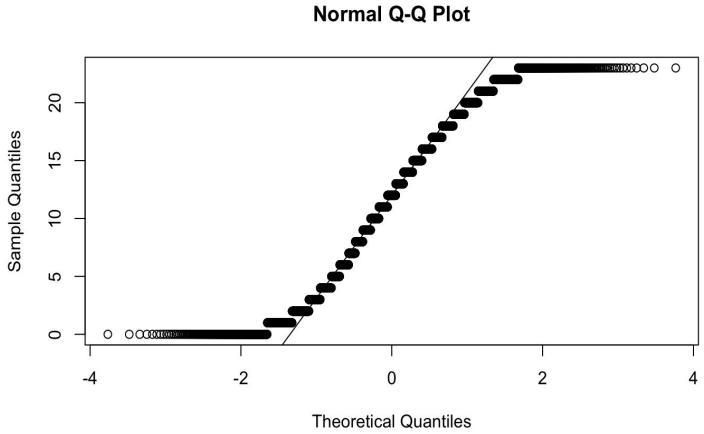
Normality Test For Distance

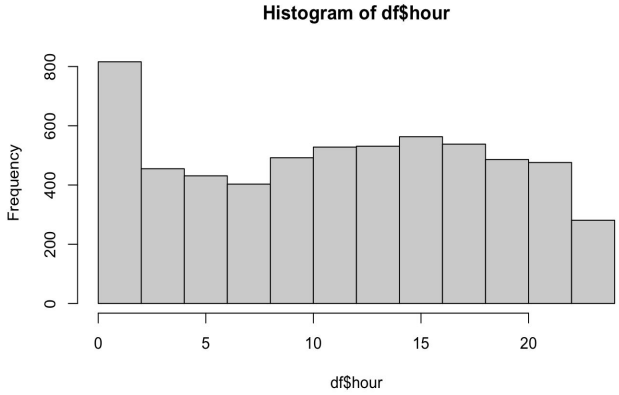


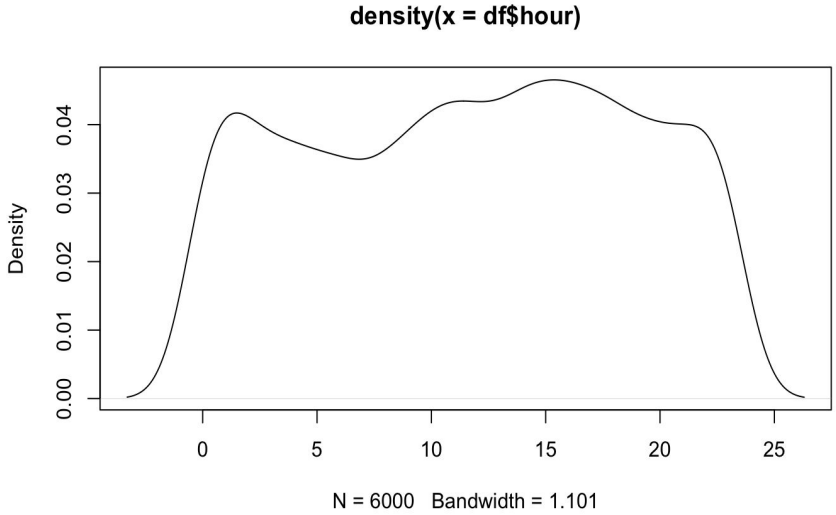




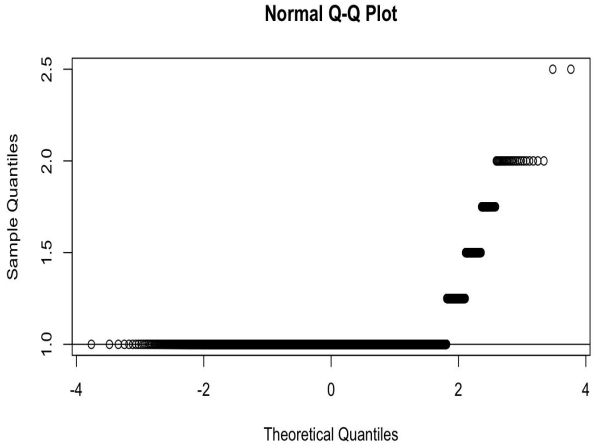
Normality Test For Hour

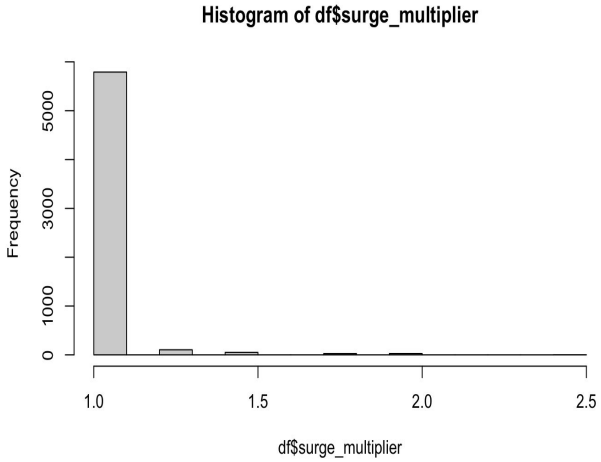


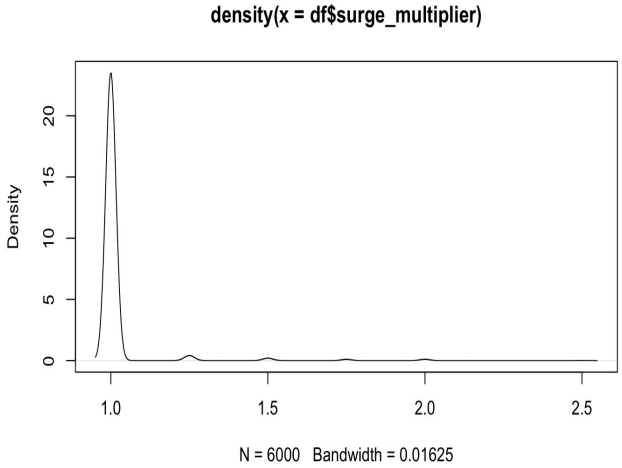




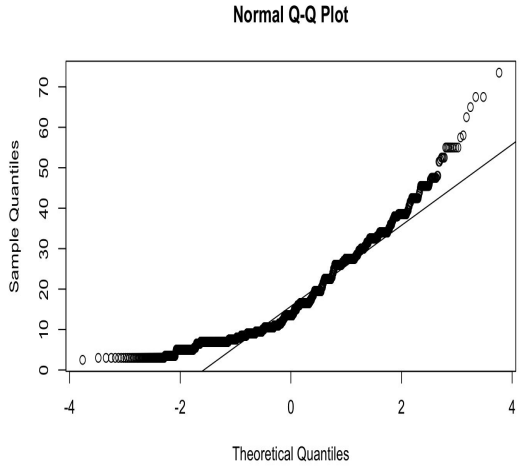
Normality Test For Surge_mutiplier

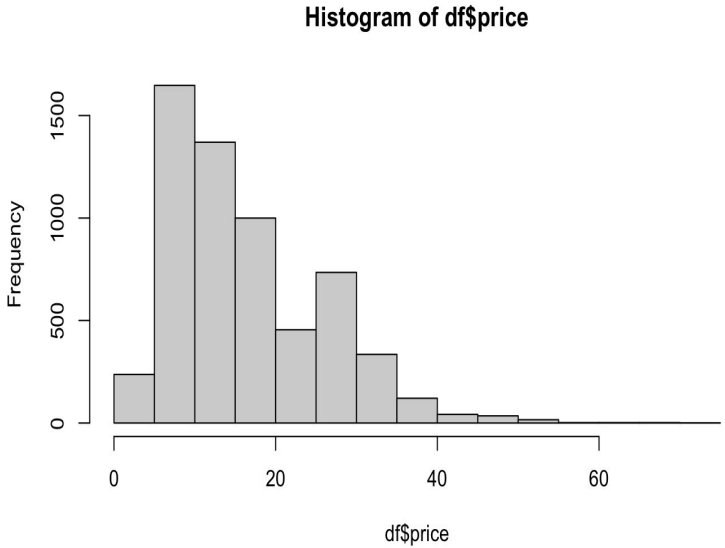


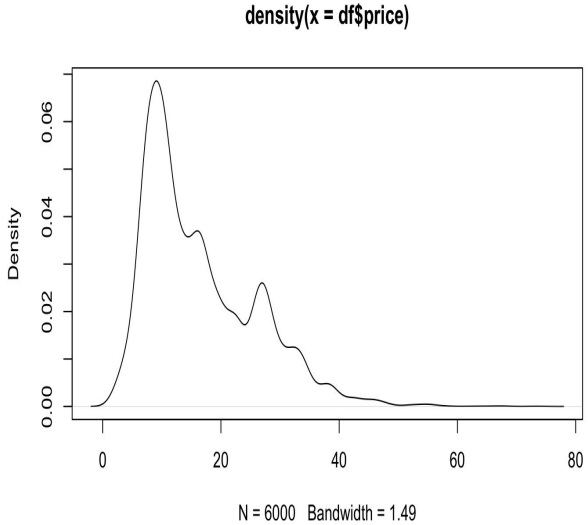




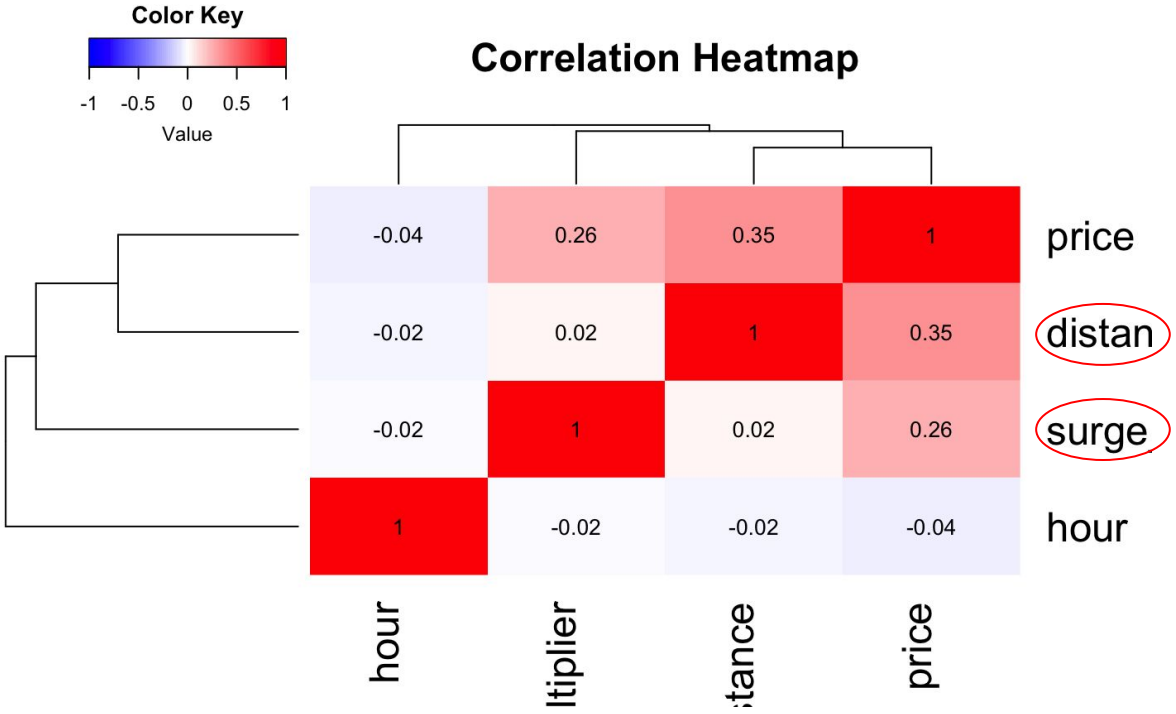
Normality Test For Price







Correlation Heatmap for Numeric Columns



Chi-Squared Test for Cab Company & Cab Type

```r
# Chi-squared test
chisq_result <- chisq.test(df$cab_company, df$cab_type)

chisq_result
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$cab_company and df$cab_type
## X-squared = 6000, df = 11, p-value <2e-16
```

Cab_company and cab_type are significantly related.

Linear Regression Analysis

```
Call:
lm(formula = price ~ distance + surge_multiplier, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-32.53  -6.75  -1.61   4.98  38.56 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -12.9730     1.1094   -11.7   <2e-16 ***
distance           2.8232     0.0962    29.4   <2e-16 ***
surge_multiplier  22.8997     1.0703    21.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.52 on 5997 degrees of freedom
Multiple R-squared:  0.184,	Adjusted R-squared:  0.183 
F-statistic:  674 on 2 and 5997 DF,  p-value: <2e-16
```

Distance, surge_multiplier are good predictors for price.

In terms of 0.184 R-squared, the model is not so good.

Summary

-the reasons we chose this topic

With the development of social technology and fast-paced life, online shared-vehicle-services have attracted more and more attention.

Selected graphics

- Used python to do the initial data cutting

- Normality Test

  Boxplot

  Heatmap

Final result (according to EDA)

- **Weak relation to: Price and Hours**

- **Price Fluctuations Throughout the Day**

- **Possible Peak Hours: the first peak close to the 5th hour might /the subsequent decrease might signify a lull during mid-morning.**

- **Midday Drop & Evening Rise**

Final result (according to EDA)

Strong relation to: Surge_multiplier and Distance with Price

Hours is not a Key-point

Final result (according to EDA)

Strong relation to: Cab_company and Cab_type

Cab_company: This may result in software design and vehicle design being different from each other:

Uber is the cheaper company.

Uber is available all over the world, while Lyft is only available in the United States and Canada.

Reference Page

1. Brllrb. (2019). Uber and Lyft Dataset: Boston, MA. Kaggle. https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma/data
2. Peng, R. D., & Matsui, E. (2016). *Art of Data Science*.

Thank you for your attention!