Lyft vs. Uber

Explaining Cab Prices Using Weather Conditions

DROMB8114: Applied Regression Analysis

Group Members:
Anunay Sanganal
Bihui Sun
Haoran Sun
Sarthak Tiwari
Skand Upmanyu
Youngkeun Yoon

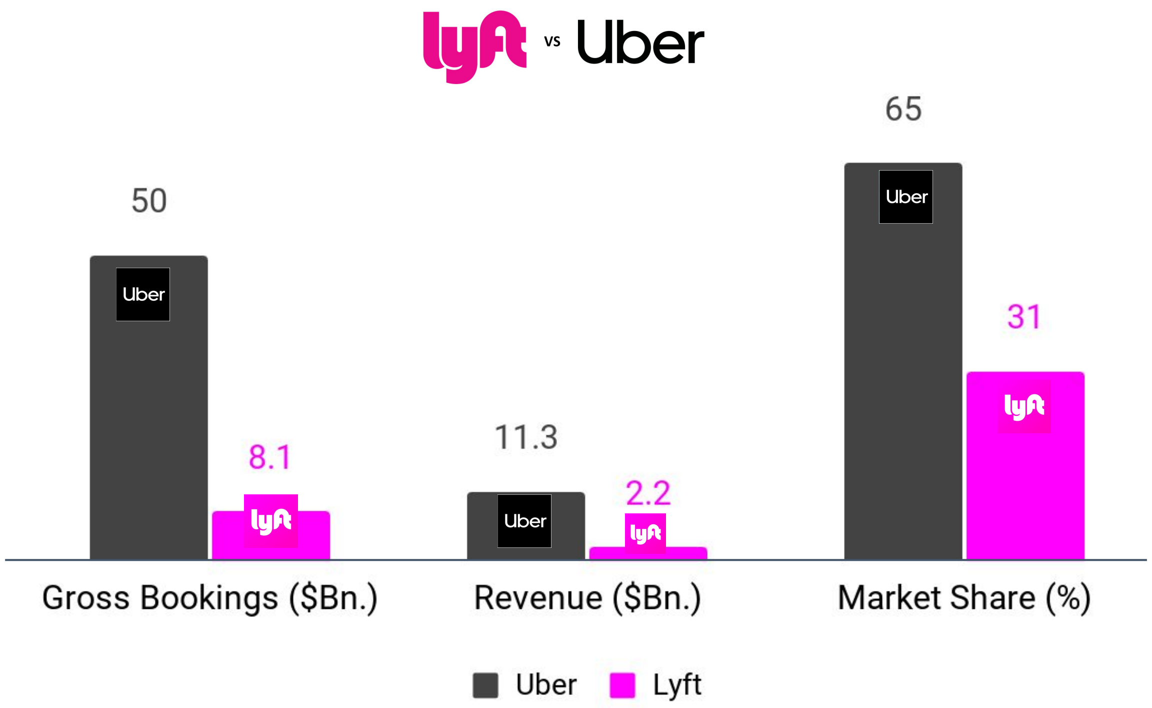
Overview

Introduction and Data Collection

Battle of Rideshare Companies: Uber’s Success

What are the Reasons for Uber’s success?

Key Performance Metrics for Uber and Lyft (2018)




BuzzFeed News This Is How Uber Will Take Over The World

TECH

This Is How Uber Will Take Over The World

The ride-hail giant is expanding globally — and it's doing so by replicating the hyperlocal approach it already uses in the U.S. Auto-rickshaws for India, Lamborghinis for Singapore.


DOERS EMPIRE SERVICES OUR WORK

WHY IS UBER WINNING THEIR COMPETITION?

Posted by Mohit Soni | Nov 12, 2019 | Entrepreneurs | 0


MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

TRADING NATION

In the battle between Lyft and Uber, analyst sees a clear winner

Source: https://www.statista.com/chart/17261/lyft-vs-uber/

Framing the Right Questions

Using Cab Rides data to answer the following questions

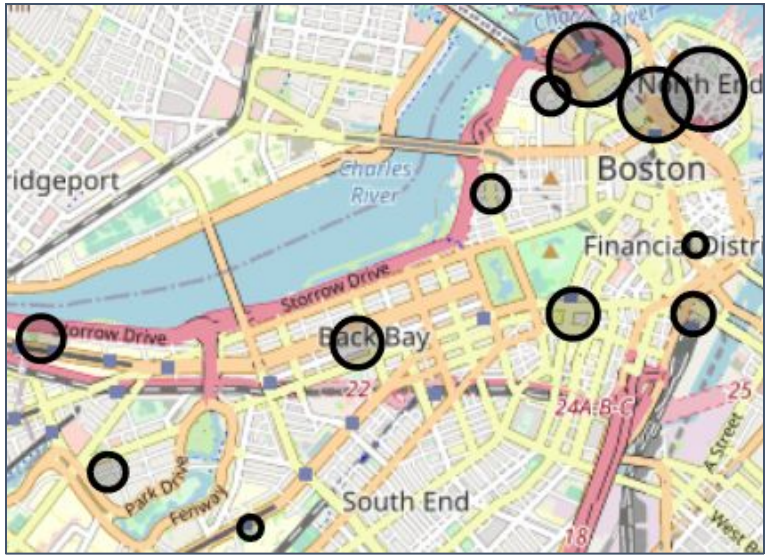
- What are the reasons for the success of Uber over Lyft?
- Is there an overall difference in the cab prices of Uber and Lyft?
  - If yes, which one is more economical?
  - Is the difference statistically significant?
- How does weather play a role in the pricing of cab rides?
  - Is the effect of weather conditions different for Uber and Lyft?
- In order to help a customer to achieve the lowest price for cab rides, what should be the recommendations?

Collecting Data for Cab Rides and Weather Conditions

Data Collection Approach

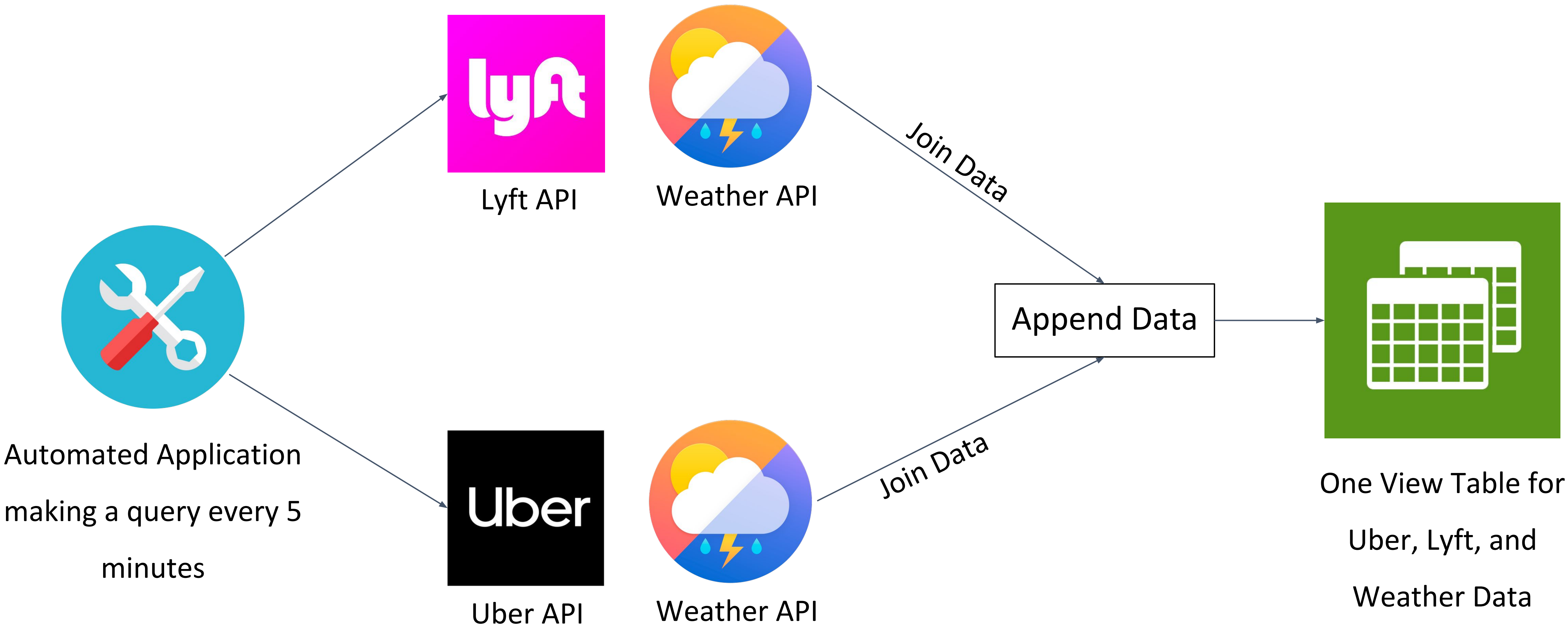
- With no publicly available data of rides/prices, the data was collected using Uber & Lyft API queries and corresponding weather conditions. Some of the hot locations in Boston (MA) were chosen:

1. Haymarket Square	7. Fenway
2. Back Bay	8. South Station
3. North End	9. Theatre District
4. North Station	10. West End
5. Beacon Hill	11. Financial District
6. Boston University	12. Northeastern University



- An automated application makes a query to the Uber and Lyft API to find out the price of all types of cab rides from location A to B for both Lyft and Uber and also queries the weather conditions at the same time. This helps us to know what the cab prices and the corresponding weather conditions are at the same time. This data is available at: https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

Data Collection Process in Action



*One row represents one ride and corresponding weather conditions

Overview of Collected Data

Data Summary

Duration	26-Nov-2018 to 18-Dec-2018 (not all days covered)
# Days	17
# Rows	693,071
# Columns	56
Data Categories	Timestamp of ride, Cab category, Distance, Temperature, Humidity, UV Index, Precipitation Probability, Wind Speed, Visibility, Pressure etc.
Tool Used for Analysis	Python

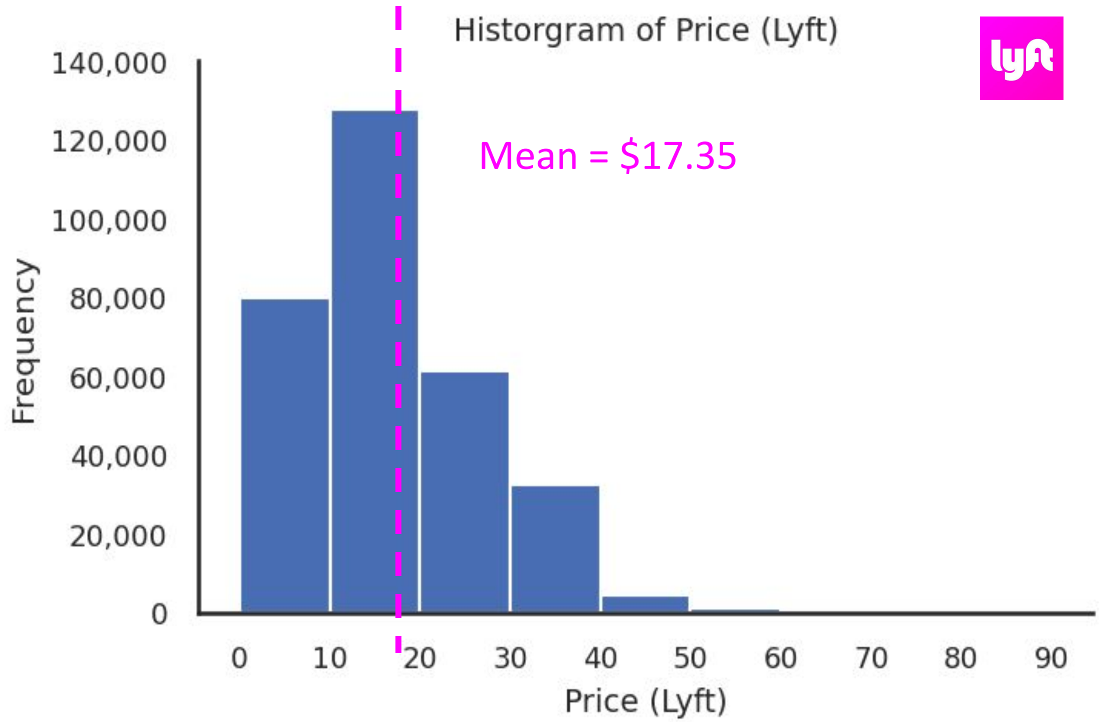
- **Note:** One row represents the cab price and weather conditions at a particular timestamp. It does not represent an actual ride taken by the customer. The demand at a particular time is indicated by the price of the cab at that time

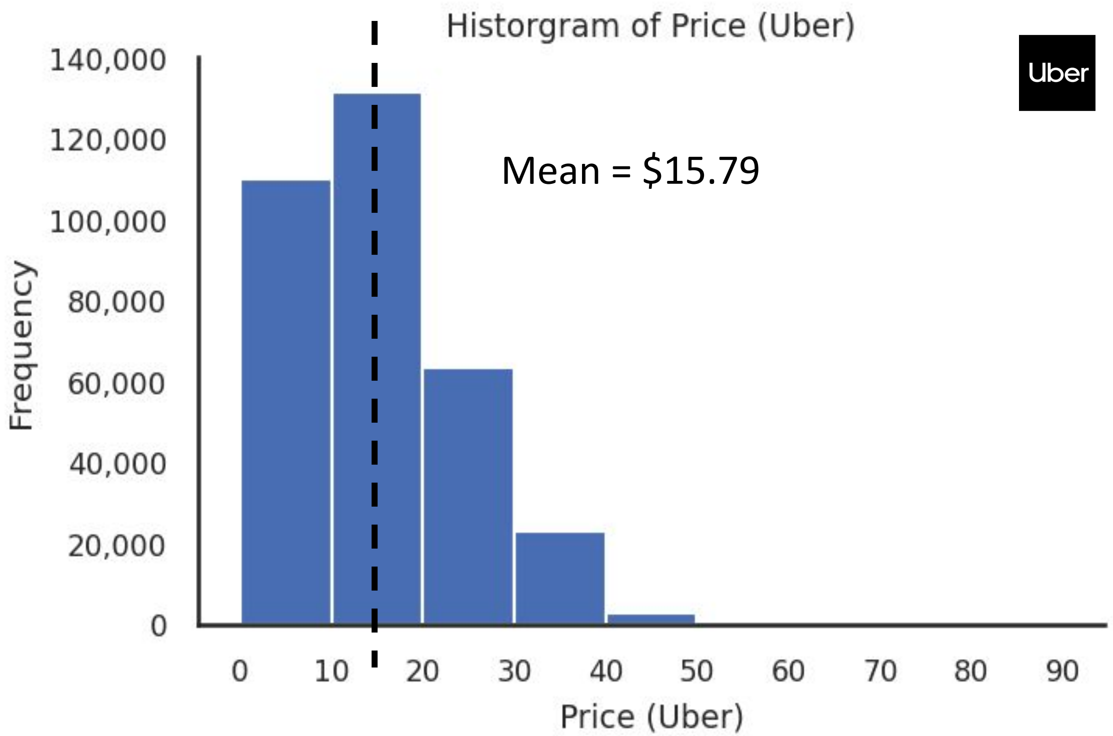
Data Exploration

Performing EDA to determine the relationship between features and price

On an Average, Lyft is Priced Lower than Uber

Univariate: Price of Cab Ride





Is this difference statistically significant?

Lyft Prices are **Significantly Lower Priced** than Uber

F-Test: Single Factor ANOVA for Cab Company

Hypotheses:

H_0: There is no difference in the average price by cab company

H_a: There is difference in the average price by cab company

Since the p-value < 0.05, we **reject the Null Hypothesis** and conclude that **Cab Company** is a **significant feature** for differentiating the price of the cab ride

One possible reason for Uber's success can be its **significantly lower price**

SUMMARY

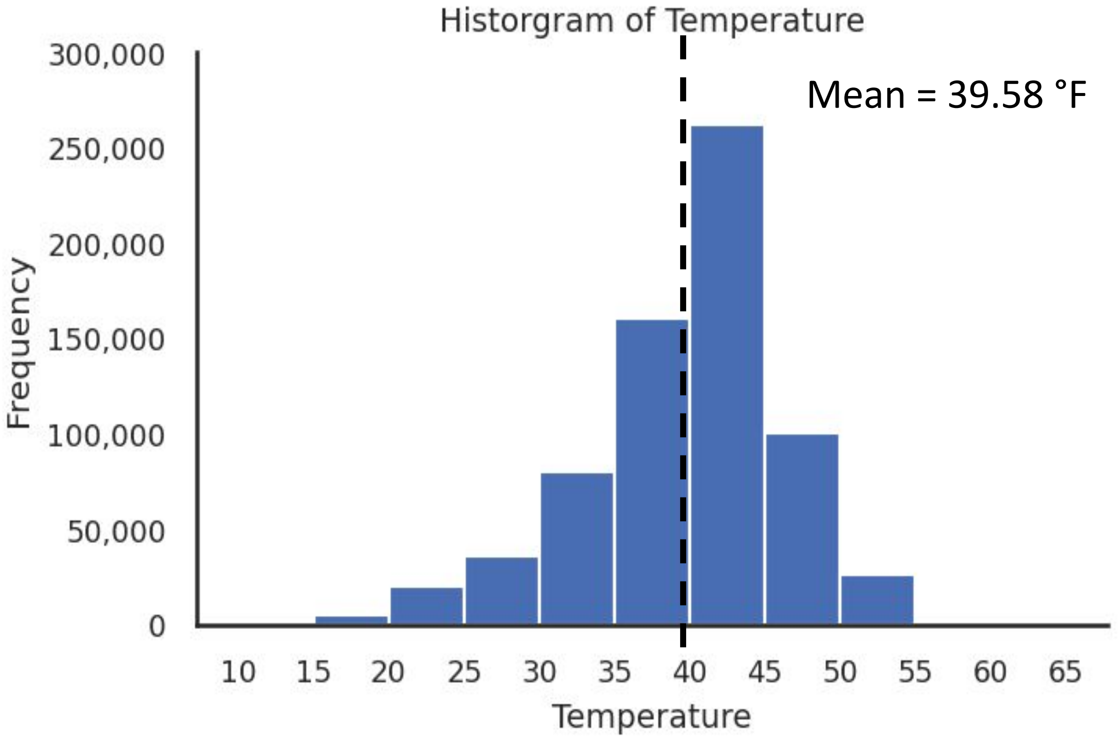
Groups	*Count*	*Sum*	*Average*	*Variance*
Lyft	307408	5333957.98	**17.35**	100.38
Uber	330568	5221435.00	**15.80**	73.28

ANOVA

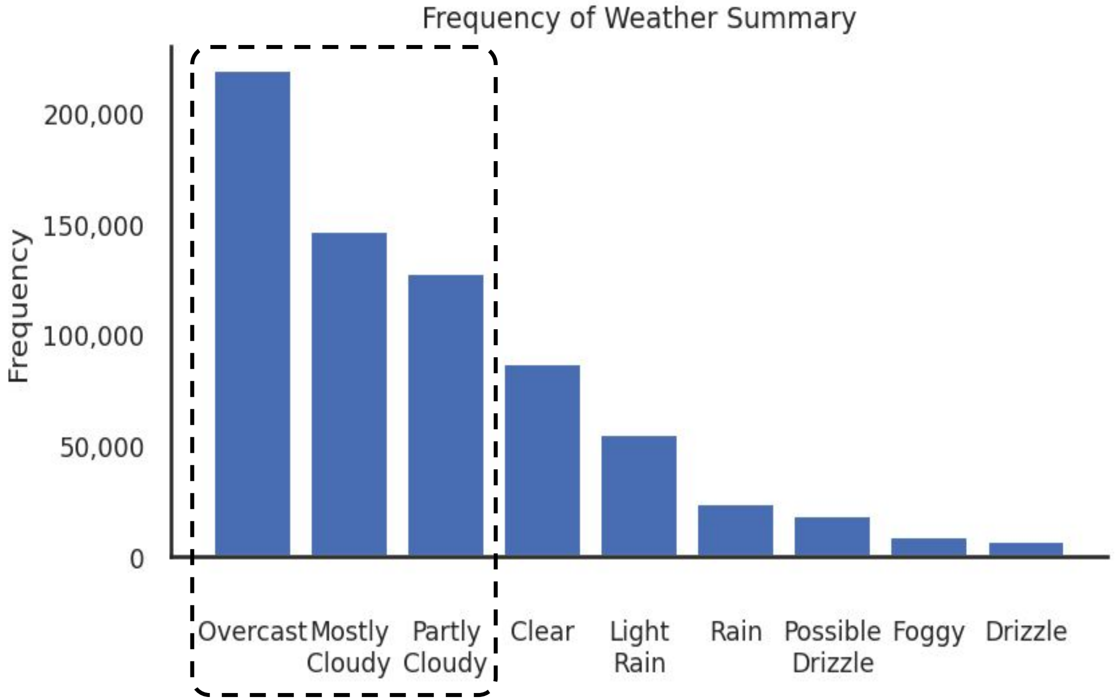
Source of Variation	*SS*	*df*	*F*	*P-value*
Between Groups	385674.02	1	**4466.96**	**0.00**
Within Groups	55082209.36	637974		
Total	55467883.37	637975		

Cold Weather Conditions with less Sunlight in Winter in Boston, MA

Univariate: Temperature and Weather Conditions



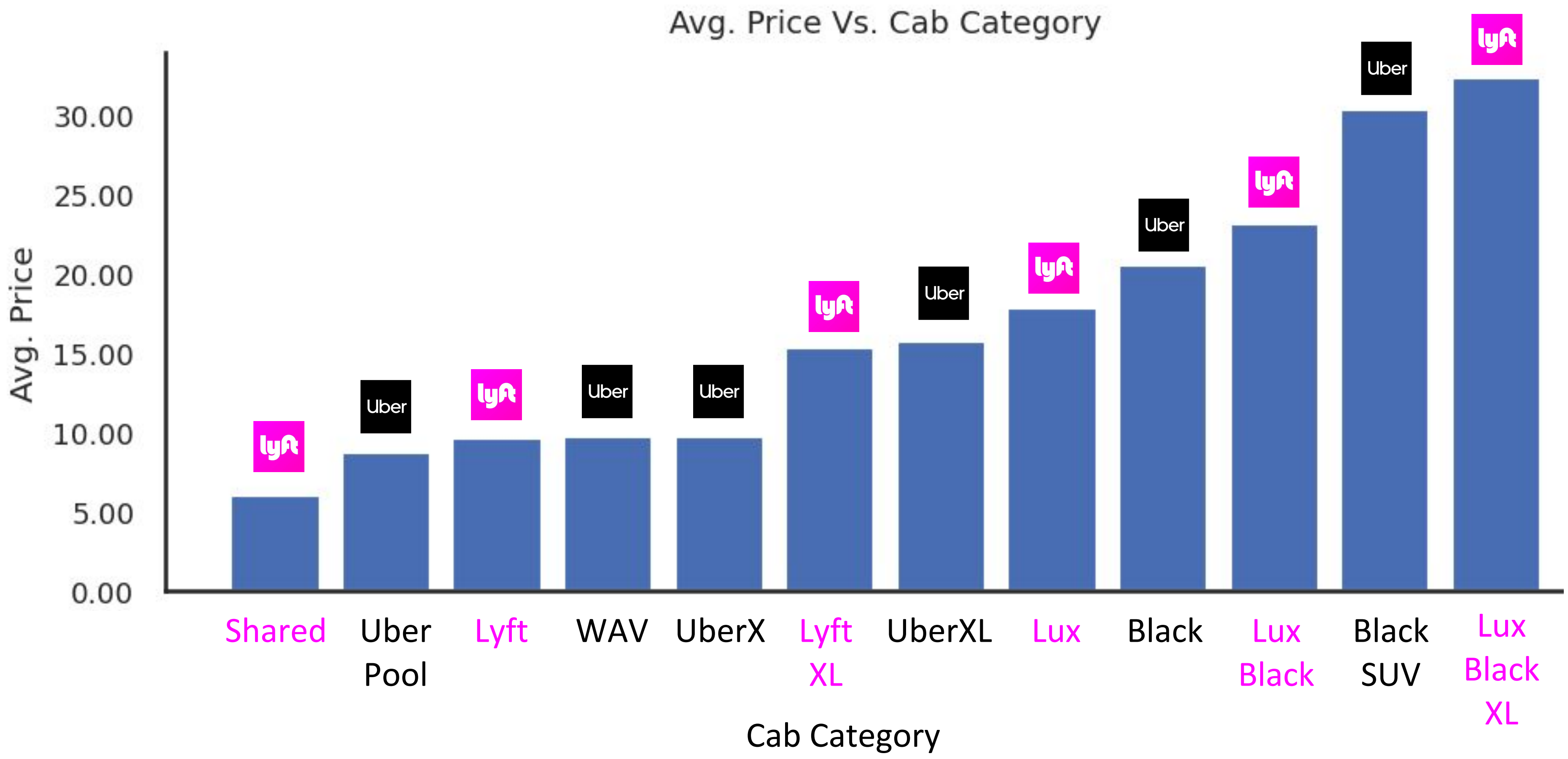
Cold Weather Conditions



Mostly Overcast/Cloudy

Lyft Covers a Broad Spectrum of Low Priced to High Priced Options

Bivariate: Avg. Price Vs. Cab Category



Cab Category is **Significant** for Determining Price

F-Test: Single Factor ANOVA for Cab Category

Hypotheses:

H_0: There is no difference in the average price by cab category

H_a: There is difference in the average price by cab category

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Black	55095	1130758.00	**20.52**	24.52
Black SUV	55095	1668679.50	**30.29**	23.39
...	...	...	...	...

ANOVA

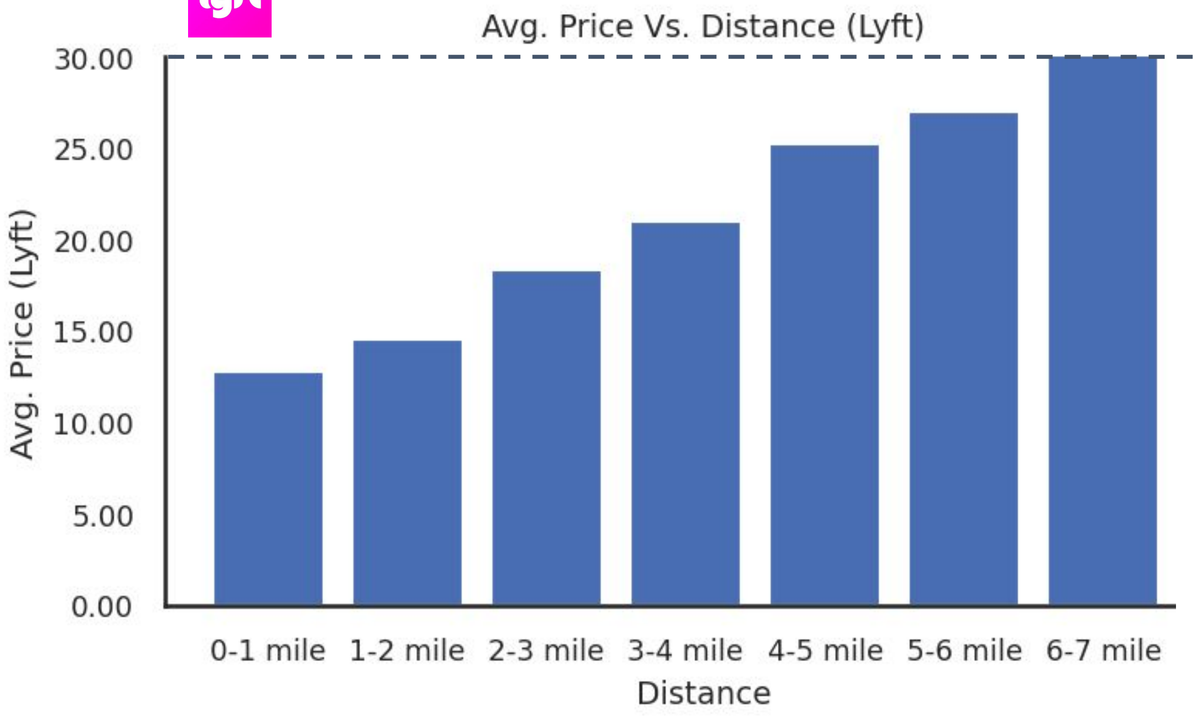
Source of Variation	*SS*	*df*	*F*	*P-value*
Between Groups	46848473.37	12	**198877.84**	**0.00**
Within Groups	12523449.45	637964		
Total	59371922.82	637976		

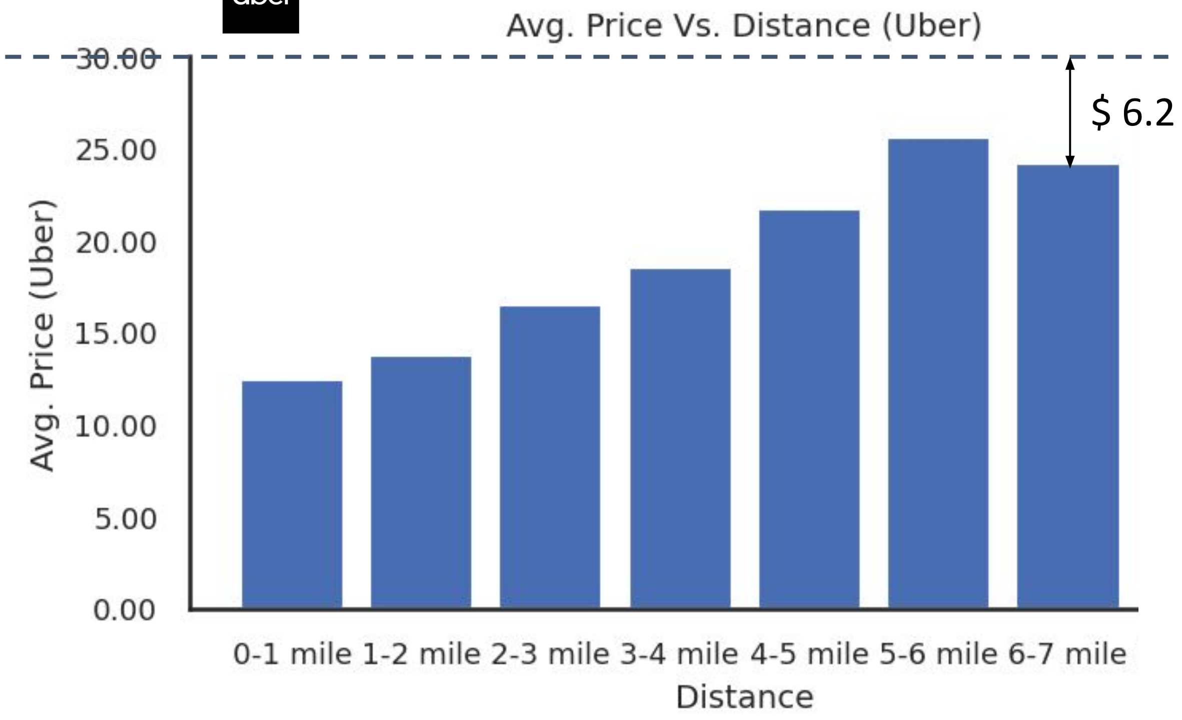
Since the p-value < 0.05, we **reject the Null Hypothesis** and conclude that **Cab Category** is a **significant feature** for differentiating the price of the cab ride

Note: Other categorical variables were also tested but turned out to be insignificant *(refer Appendix)*

Uber is More Economical for Long Distance Rides

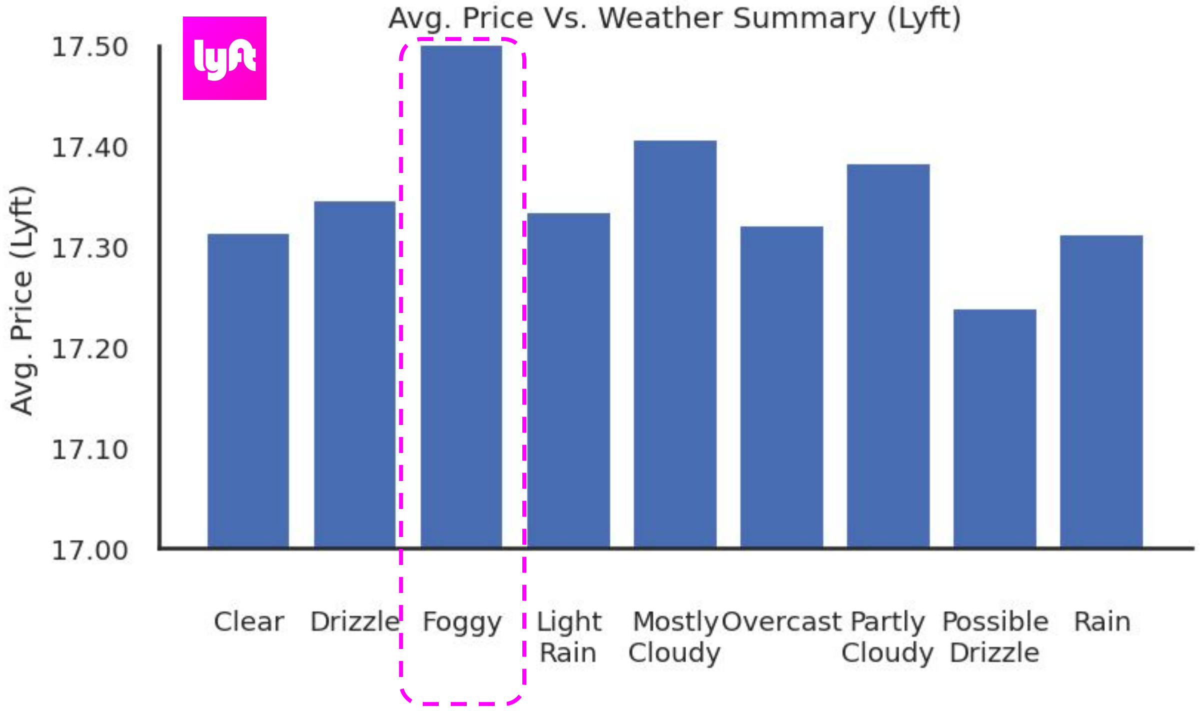
Bivariate: Avg. Price vs. Distance

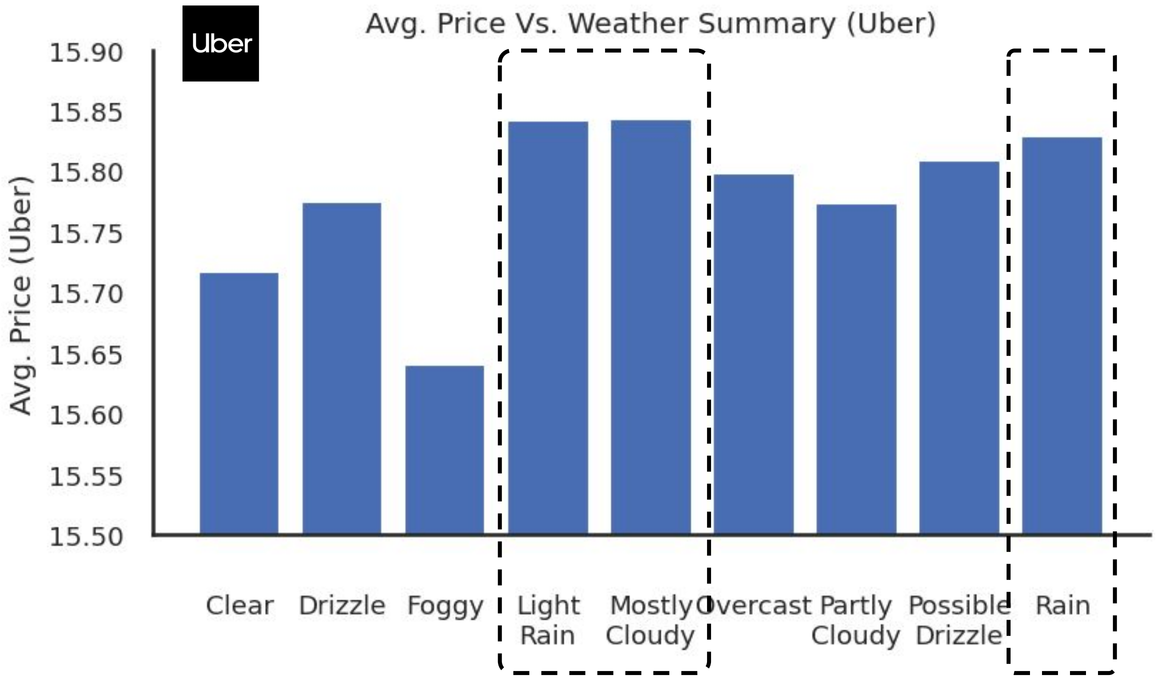




Lyft Prices Peak during Fog and Uber Prices Peak During Rains

Bivariate: Avg. Price vs. Weather Summary





Regression Modeling

Regression models to determine the price of a cab ride using ride data and weather data

Modeling Framework

Data Import	Imported CSV data into Python for data preparation
Missing Value Treatment	**Dropped rows** with missing values as we have a large dataset
Feature Engineering	Created **additional features** for modeling (Time from Max Temp., Day of Week, Hour, etc.)
Correlation and VIF Analysis	Removed **Correlated Variables** and variables with **High VIF values**
Model Iterations and Variable Selection	Explored **Linear** and **Exponential Regression** and selected variables using **Forward Selection** 1. Model for explaining price of **all cab rides** together 2. Individual model for explaining prices of **Lyft rides** only 3. Individual model for explaining prices of **Uber rides** only
Insights from Model	Selected model with the **highest Adj. R Squared** and **interpreted coefficients** to create insights
Regression Diagnostics	Analyzed the residuals for **Homoscedasticity** and **patterns**

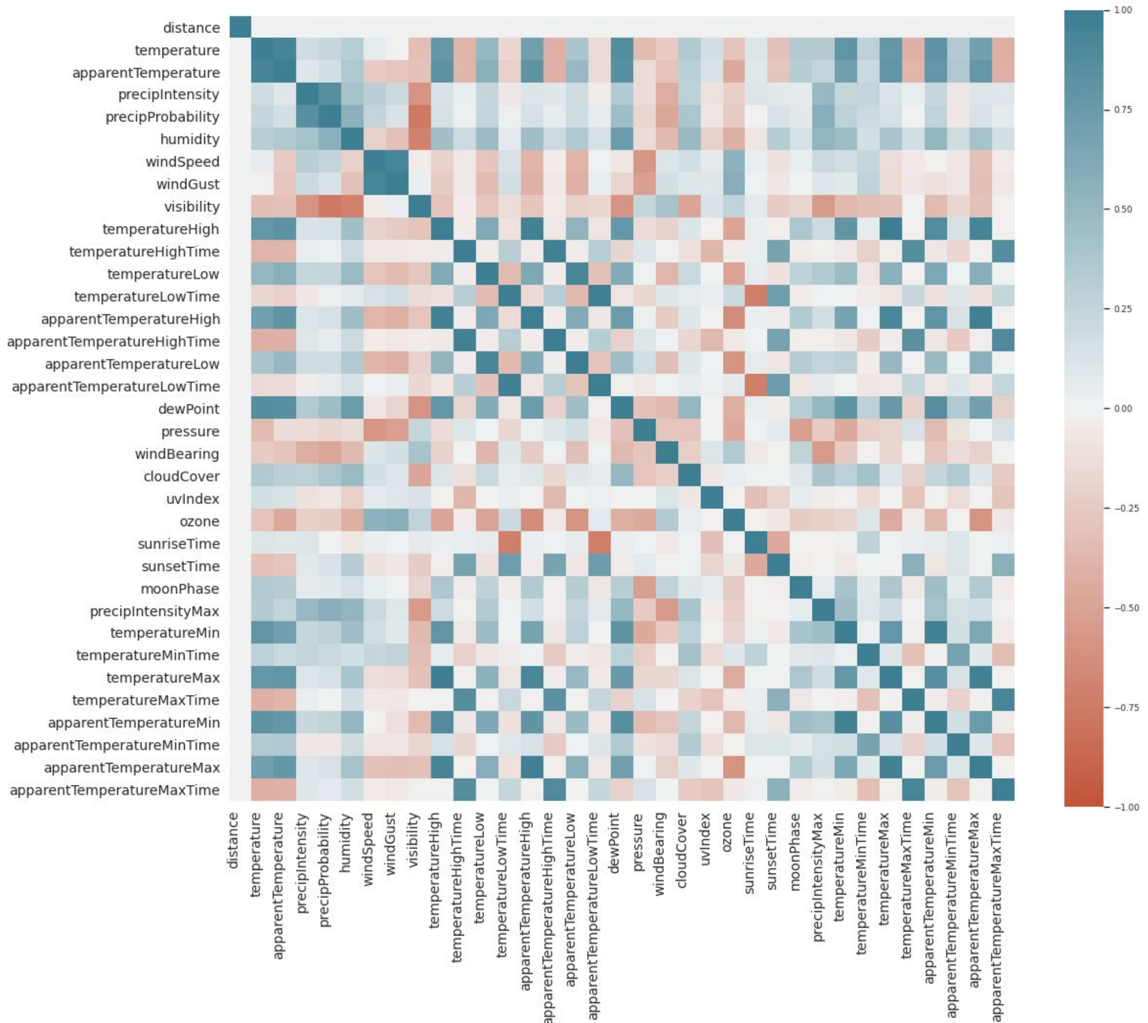
Feature Engineering

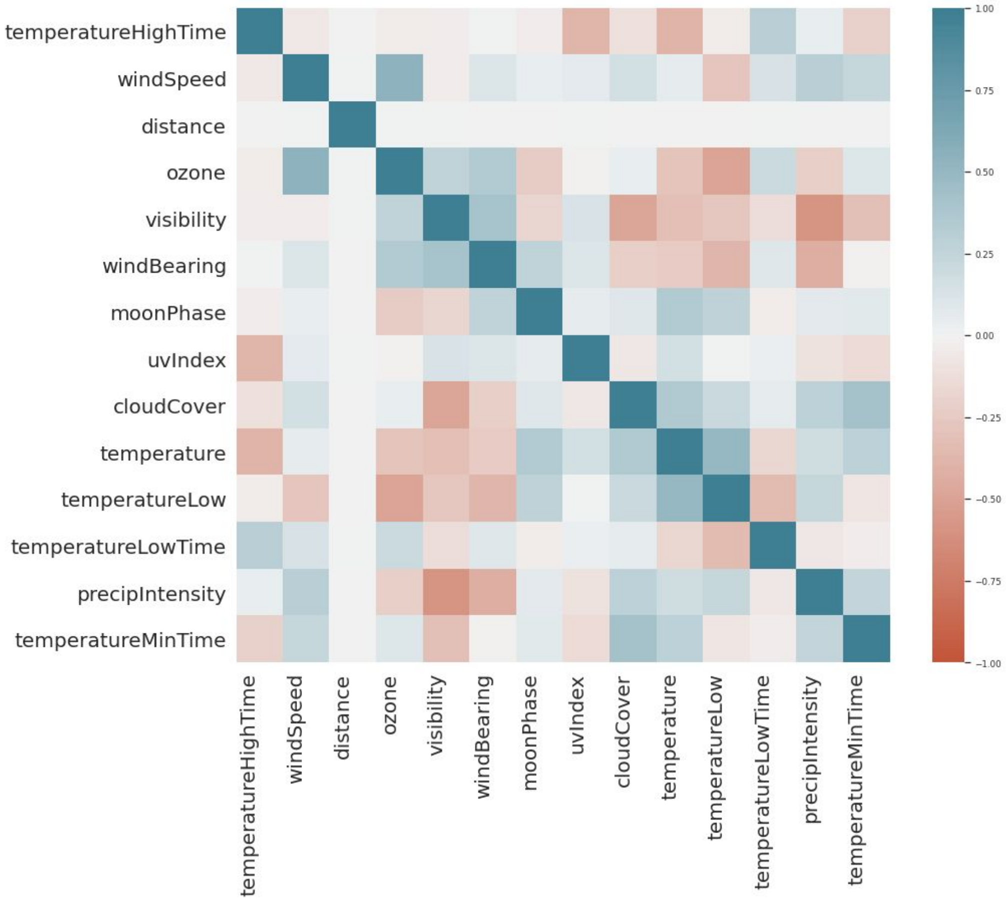
Additional Features Created for Modeling

- Features from Time of Cab Booking:
  - Hour
  - Day of Week
- Time Difference between Time of Cab Booking and:
  - Time of Maximum Temperature during that day
  - Time of Minimum Temperature during that day
  - Time of Sunrise during that day
  - Time of Sunset during that day
- One Hot Encoding (Dummy variables) created for categorical features

Correlation Analysis (21 variables removed)

Removed variables with High Correlation (> 0.6 or < -0.6)





Correlation Analysis (21 variables removed)

Correlated Variables

- Sunset/Sunrise Time correlated with Max Temp. Time and Min. Temp. Time
- Apparent Temp. High/Low Time correlated with Temp. High/Low Time
- Wind Gust correlated with Wind Speed
- Precipitation, Dew Point, Humidity, Visibility, and Pressure all correlated
- Temperature, High Temperature, and Low Temperature all correlated

VIF Analysis (8 variables removed)

Removed variables with High Variable Inflation Factor (> 5.0)

Variable	VIF
cab_type_Uber	inf
name_Lux	inf
name_LyftXL	inf
name_Lyft	inf
name_Shared	inf
name_LuxBlackXL	inf
name_LuxBlack	inf
short_summary_Overcast	46.36
cloudCover	28.72
short_summary_LightRain	24.15
short_summary_Rain	20.22
short_summary_MostlyCloudy	19.88
precipIntensity	17.93
temperatureLowTime	15.25
uvIndex	10.67
...	...

Variable	VIF
dayOfWeek_Tuesday	3.79
dayOfWeek_Monday	3.14
temperature	3.13
temperatureLow	2.73
dayOfWeek_Thursday	2.51
dayOfWeek_Sunday	2.51
windBearing	2.47
windSpeed	2.36
cloudCover	2.33
dayOfWeek_Wednesday	2.30
temperatureMinTime	2.17
dayOfWeek_Saturday	2.09
hour_10	2.04
hour_11	2.03
hour_17	2.02
...	...

Model Iterations (All Cab Rides)

Model Iterations

lyft + Uber

# Rows = 637,976	v1.0: Full Model	v1.1: Linear Regression with Forward Selection	v1.2: Log(Price), Log(Distance) with Forward Selection
Summary	Full model with all features	Forward selection using all features	Changed price to log(price) and distance to log(distance)
# Features	56	20	20
Significant features	20	20	20
R^2	0.893	0.893	**0.917**
Adj. R^2	0.893	0.893	**0.917**

Insights from the Proposed Model for All Cab Rides

Regression Statistics	
R Square	0.917
Adjusted R Square	0.917

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.7973	0.001	3260.592	0.000
name_UberPool	-0.8521	0.001	-861.067	0.000
name_Shared	-1.2671	0.001	-1256.908	0.000
name_BlackSUV	0.4042	0.001	408.411	0.000
name_LuxBlackXL	0.4550	0.001	451.39	0.000
name_UberXL	-0.2797	0.001	-282.602	0.000
name_WAV	-0.7438	0.001	-751.627	0.000
name_Lux	-0.1636	0.001	-162.337	0.000
name_UberX	-0.7438	0.001	-751.615	0.000
name_LuxBlack	0.1033	0.001	102.432	0.000
name_Lyft	-0.7686	0.001	-762.419	0.000
name_LyftXL	-0.3121	0.001	-309.553	0.000
log_distance	0.3169	0.000	905.318	0.000
short_summary_MostlyCloudy	0.0013	0.001	2.600	0.009
cloudCover	-0.0019	0.001	-3.061	0.002
temperatureMinTime	8.543e-05	3.65e-05	2.339	0.019
hour_13	0.0031	0.001	3.057	0.002
hour_17	0.0031	0.001	3.073	0.002
hour_20	0.0025	0.001	2.281	0.023
hour_21	0.0024	0.001	2.267	0.023
hour_2	0.0021	0.001	2.016	0.044

Shared Cab is the most economical (Least coefficient)

Lux Black XL is the most expensive (Highest coefficient)

Rate of increase of log(price) with log(distance)

Cloudy Weather leads to increase in Price: Customers prefer cabs in Cloudy weather (due to possibility of rain)

Busy hours: 1:00pm, 5:00pm, 8:00pm, and 9:00pm
Less availability of cabs: 2:00am

Model Iterations (Lyft Rides)

Model Iterations

# Rows = 307,408	**v2.0: Full Model**	**v2.1: Linear Regression with Forward Selection**	**v2.2: Log(Price), Log(Distance) with Forward Selection**
Summary	Full model with all features	Forward selection using all features	Changed price to log(price) and distance to log(distance)
# Features	56	11	10
Significant features	18	11	10
R^2	0.877	0.877	**0.919**
Adj. R^2	0.877	0.877	**0.919**

Insights from the Proposed Model for Lyft Rides

Regression Statistics	
R Square	0.919
Adjusted R Square	0.919

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.4969	0.001	1677.051	0.000
name_LuxBlackXL	0.4985	0.001	447.744	0.000
name_Lux	1.3703	0.001	1230.807	0.000
name_LuxBlack	1.1034	0.001	991.076	0.000
name_Lyft	1.7221	0.001	1546.771	0.000
name_LyftXL	0.9550	0.001	857.773	0.000
log_distance	0.3684	0.001	645.495	0.000
short_summary_MostlyCloudy	0.0022	0.001	2.733	0.006
hour_13	0.0043	0.002	2.754	0.006
hour_18	-0.0036	0.002	-2.253	0.024
hour_20	0.0035	0.002	2.094	0.036

Shared Cab is the most economical

Lux Black XL is the most expensive (Highest coefficient)

Higher Rate of increase with distance: (0.3169 for overall model)

Cloudy weather increasing price due to chances of rain

1:00pm and 8:00pm are *busier*. However, *6:00pm is the most economical* time for Lyft

Model Iterations (Uber Rides)

Uber

Model Iterations

# Rows = 330,568	v3.0: Full Model	v3.1: Linear Regression with Forward Selection	v3.2: Log(Price), Log(Distance) with Forward Selection
Summary	Full model with all features	Forward selection using all features	Changed price to log(price) and distance to log(distance)
# Features	56	10	11
Significant features	10	10	11
R^2	0.920	0.920	**0.918**
Adj. R^2	0.920	0.920	**0.918**

Insights from the Proposed Model for Uber Rides

Uber

Regression Statistics	
R Square	0.918
Adjusted R Square	0.918

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.8228	0.001	3301.083	0.000
name_UberPool	-0.8521	0.001	-972.925	0.000
name_BlackSUV	0.4042	0.001	461.458	0.000
name_UberXL	-0.2797	0.001	-319.312	0.000
name_WAV	-0.7438	0.001	-849.272	0.000
name_UberX	-0.7438	0.001	-849.263	0.000
log_distance	0.2760	0.000	664.979	0.000
temperatureMinTime	0.0001	4.46E-05	3.076	0.002
cloudCover	-0.0018	0.001	-2.256	0.024
hour_17	0.0038	0.001	3.004	0.003
hour_2	0.0026	0.001	2.013	0.044
hour_1	-0.0027	0.001	-2.134	0.033

UberPool is the most economical option for Uber

BlackSUV is the most expensive option for Uber

Lower Rate of increase with distance: (0.3684 for Lyft)

More demand when away from Temp Min. Time / Night time

Reduced price of Uber during Cloudy weather: Reliability

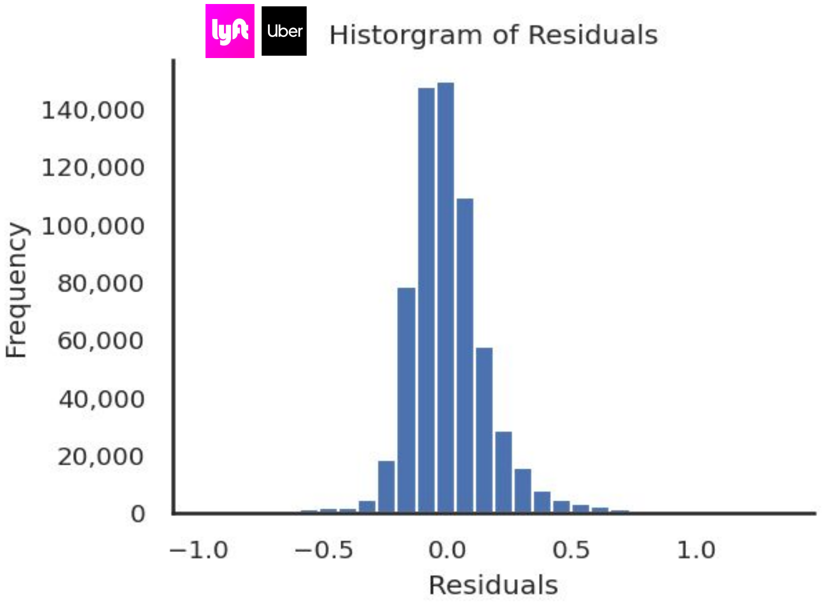
5:00pm is a busy time for Uber

Fluctuation at Night: 1:00am is the most economical while 2:00am is busy. One possibility being less demand at 1:00am but many drivers heading home after 2:00am reducing cab availability

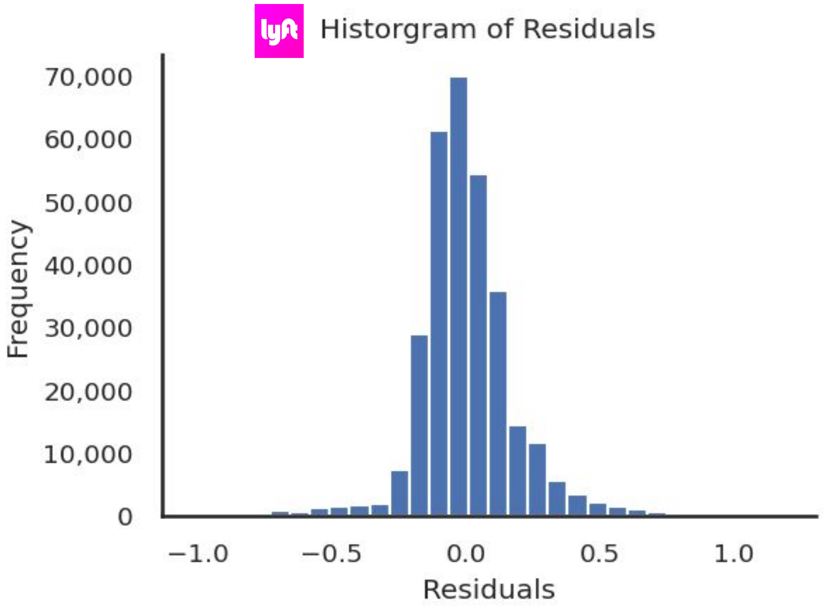
Regression Diagnostics

Distribution of Residuals

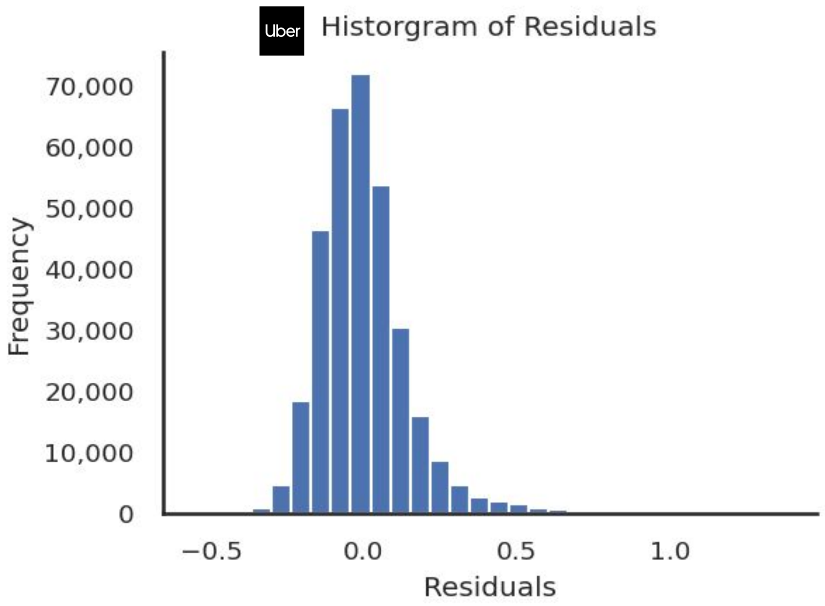
Model for All Rides (v1.2)



Model for Lyft Rides (v3.2)



Model for Uber Rides (v3.2)



The distribution of residuals suggests that the residuals are normally distributed

Recommendations and Takeaways

Final Recommendations, learnings and scope for improvement

Recommendations for a Consumer to Achieve Lowest Cab Prices

Achieving the Most Economical Options

- Uber has a **significantly lower price** than Lyft
- Uber is also more economical for **long distance rides**
- More **reliable** during **bad weather conditions**
- Consider taking advantage of **fluctuation of price** around 1:00am - 2:00am at night

However, consider Lyft during the following conditions:

- **Shared cab** of Lyft is even more economical than UberPool
- **6:00pm** is a sweet spot for booking Lyft rides to get home from office. This time is expensive for Uber

Note: Results only applicable to hot areas of Boston, MA. There might be different trends in different geographic locations

Learnings from Project and Scope for Improvement

Learnings and Way Forward

Python

- Reproducible code and easy to replicate results
- Easier for collaboration in a team (Google Colab)

- No UI with drag drop features and requires coding background
- Google Sheets as a substitute for collaboration but has limited features

Target Variable

- Different Target Variables can be considered for analysis:
  - Price
  - Price / Distance
  - Surge Multiplier

Data Background and Research

- Important to research about the data collection approach in order to understand the data that we are working with. The number of rides per cab category did not make sense initially
- Is this lower price of Uber sustainable?

Thank you

Appendix

Day of Week is **Almost Significant** for Determining Price

Single Factor ANOVA for Day of Week

Hypotheses:

H_0: There is no difference in the price by day of week

H_a: There is difference in the price by day of week

Since the p-value = 0.07, **we can almost reject the Null Hypothesis** and conclude that **Day of Week** is **almost a significant feature** for determining the price of the cab ride

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Monday	114239	1884137.85	**16.49**	86.20
Tuesday	115091	1909410.80	**16.59**	88.06
...	...	...	...	...

ANOVA

Source of Variation	*SS*	*df*	*F*	*P-value*
Between Groups	1008.98	6	**1.93**	**0.07**
Within Groups	55466874.39	637969		
Total	55467883.37	637975		

Weather Summary alone is **Not Significant** for Determining Price

Single Factor ANOVA for Weather Summary

<u>Hypotheses:</u>

H_0: There is no difference in the price by weather summary

H_a: There is difference in the price by weather summary

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Overcast	201429	3330651.85	**16.54**	86.81
Mostly Cloudy	134603	2233658.63	**16.59**	87.89
…	…	…	…	…

ANOVA

Source of Variation	*SS*	*df*	*F*	*P-value*
Between Groups	725.33	8	**1.04**	**0.40**
Within Groups	55467158.05	637969		
Total	55467883.37	637967		

Since the p-value > 0.05, **we cannot reject the Null Hypothesis** and conclude that **Weather Summary** alone is **not a significant feature** for determining the price of the cab ride

Hour of Cab Ride alone is **Not Significant** for Determining Price

Single Factor ANOVA for Hour

<u>Hypotheses:</u>

H_0: There is no difference in the price by hour of cab ride

H_a: There is difference in the price by hour of cab ride

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
0	29872	495121.50	**16.57**	87.69
1	26310	434477.50	**16.51**	87.12
...	...	...	...	...

ANOVA

Source of Variation	*SS*	*df*	*F*	*P-value*
Between Groups	898.61	23	**0.45**	**0.99**
Within Groups	55466984.76	637952		
Total	55467883.37	637975		

Since the p-value > 0.05, **we cannot reject the Null Hypothesis** and conclude that **Hour of Cab Ride** alone is **not a significant feature** for determining the price of the cab ride

Final Selected Variables

- hour_6
- hour_13
- dayOfWeek_Tuesday
- moonPhase
- hour_22
- cloudCover
- temperature
- hour_9
- short_summary_Rain
- name_UberPool
- dayOfWeek_Thursday
- short_summary_Drizzle
- hour_3
- name_Shared
- windBearing
- dayOfWeek_Saturday
- temperatureMinTime
- dayOfWeek_Monday
- short_summary_PartlyCloudy
- name_BlackSUV
- hour_17
- name_LuxBlackXL
- short_summary_PossibleDrizzle
- distance
- dayOfWeek_Wednesday
- hour_14
- hour_12
- short_summary_Foggy
- temperatureLow
- hour_15
- hour_7
- name_UberXL
- windSpeed
- hour_5
- name_WAV
- price
- short_summary_LightRain
- short_summary_MostlyCloudy
- name_Lux
- name_UberX
- hour_21
- name_LuxBlack
- dayOfWeek_Sunday
- hour_4
- hour_10
- hour_8
- name_Lyft
- hour_20
- hour_18
- hour_11
- name_LyftXL
- hour_23
- hour_2
- hour_16
- hour_1
- hour_19

Variables Removed due to high correlation

- sunsetTime
- sunriseTime
- apparentTemperatureHighTime
- apparentTemperatureMaxTime
- apparentTemperatureLowTime
- apparentTemperature
- apparentTemperatureLow
- apparentTemperatureMin
- apparentTemperatureHigh
- apparentTemperatureMax
- windGust
- temperatureMaxTime
- temperatureMax
- precipProbability
- humidity
- pressure
- precipIntensityMax
- dewPoint
- temperatureHigh
- temperatureMin
- apparentTemperatureMinTime

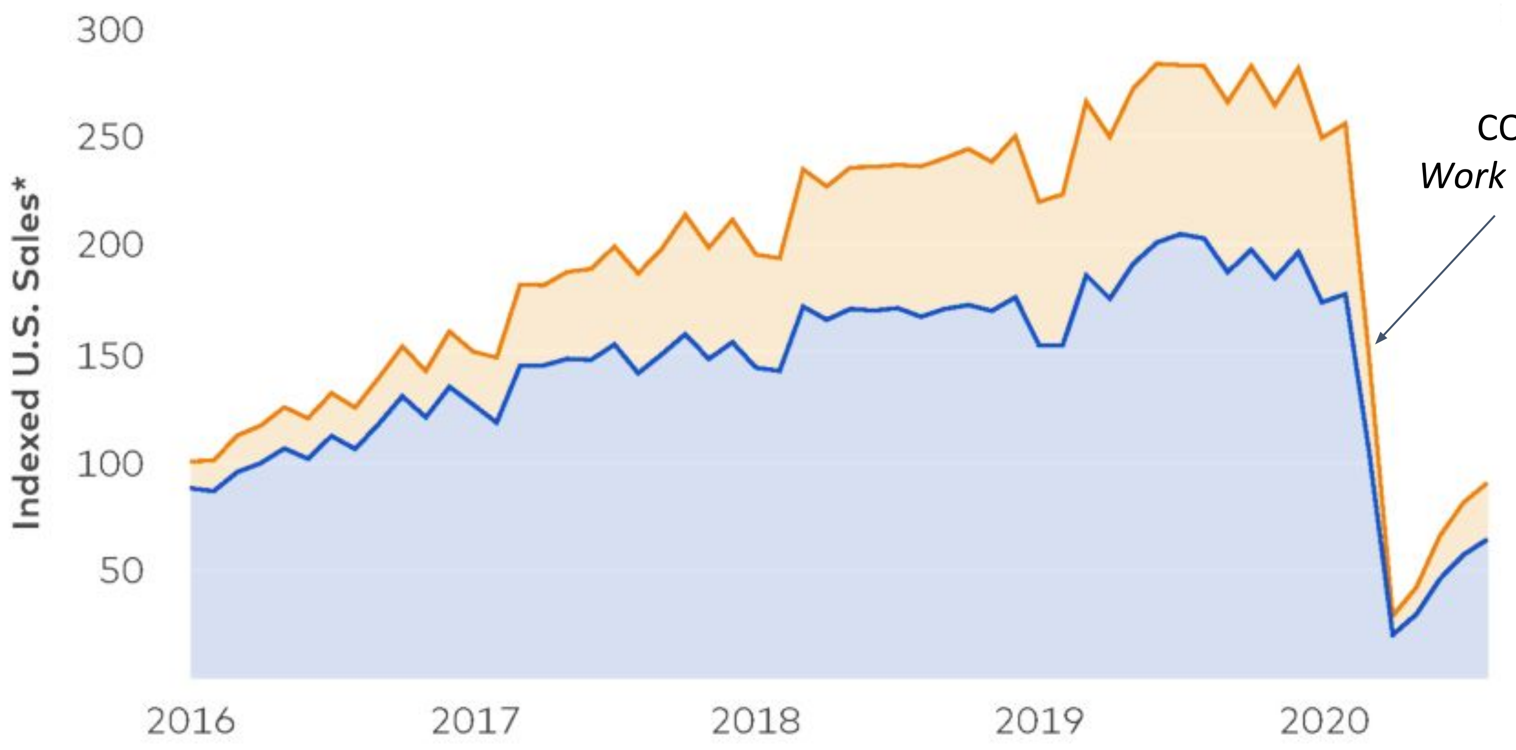
Variables Removed due to high VIF

- cab_type_Uber
- short_summary_Overcast
- precipIntensity
- temperatureLowTime
- uvIndex
- ozone
- visibility
- temperatureHighTime

Market Share

Market Share

Monthly Sales for Uber and Lyft in the U.S.



August 2020

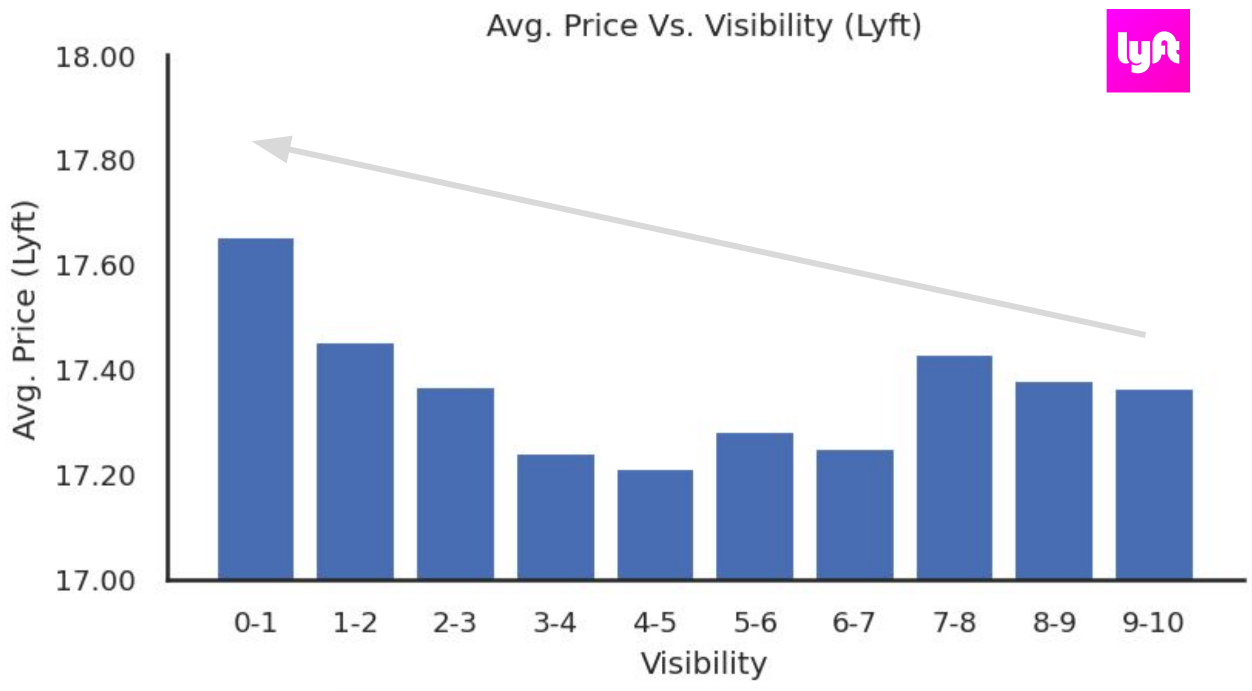
Lyft 29%

Uber 71%

Source: https://secondmeasure.com/datapoints/rideshare-industry-overview/

Visibility Plays an Important Role for Lyft but not so much for Uber

Bivariate: Avg. Price vs. Visibility

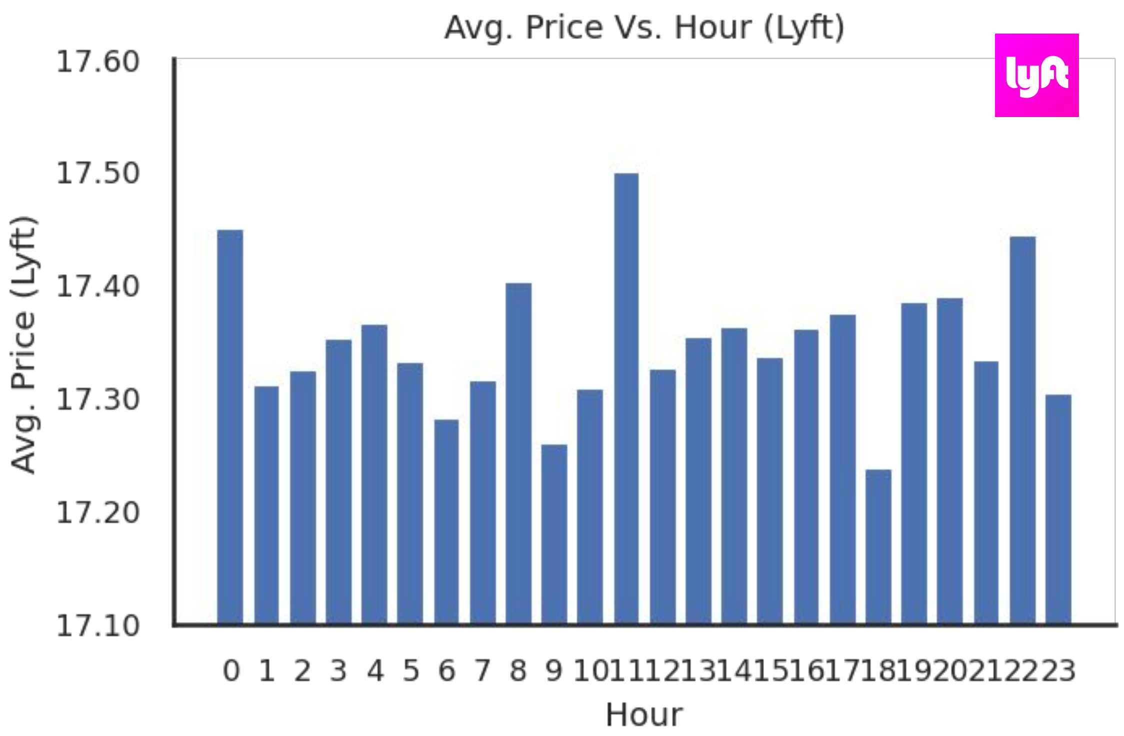


Price decreasing with Increase in Visibility

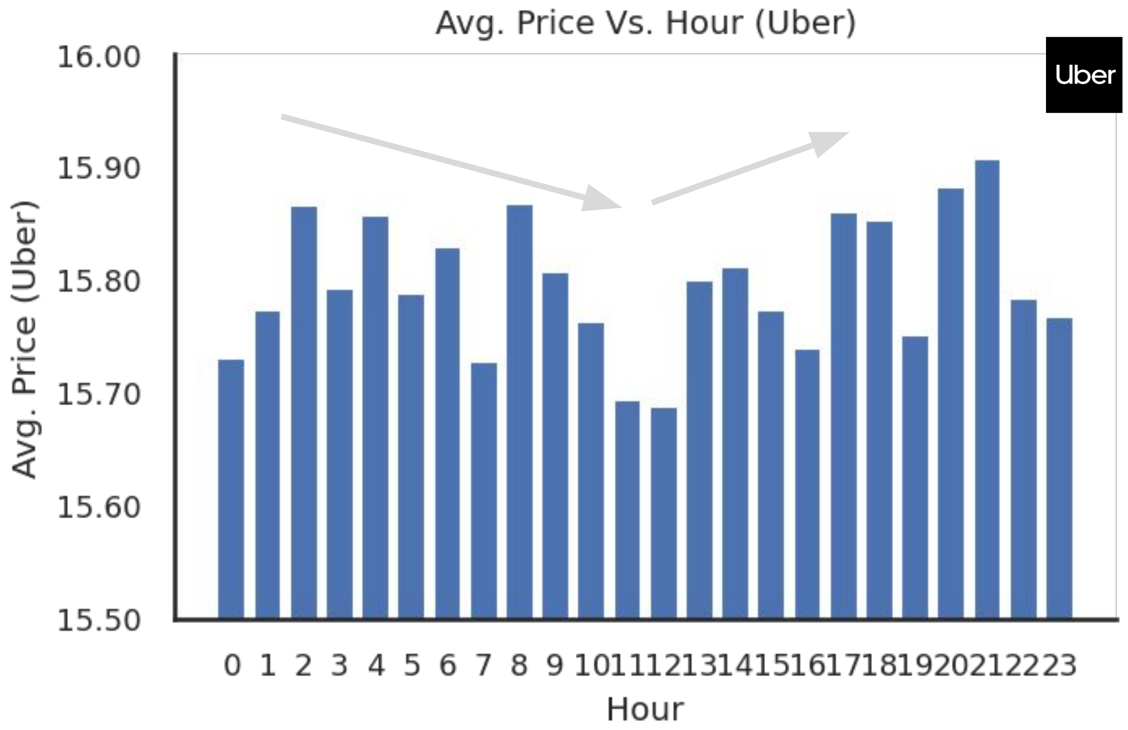
Not a Clear Pattern of Price with Visibility

Price plateaued at noon for Uber and no clear pattern for Lyft

Bivariate: Avg. Price vs. Hour



No clear pattern, but price spiked at 11am, 10pm, 0am



Price drops approaching noon before picking up again

High Humidity and Wind Speed in Winter in Boston, MA

Univariate: Humidity and Wind Speed