

Uber vs Lyft Ultimate battle!

PROJECT SUMMARY

Project Title:

To compare the fare between Lyft and Uber

Abstract:

This project involves a detailed analysis and visualization of ride-hailing data from Uber and Lyft to compare their performance across various metrics. The dataset includes ride prices, durations, distances, and additional contextual factors like weather and location. The objective is to uncover patterns, differences, and similarities between the two services, providing insights into their market dynamics and customer preferences. This analysis serves as a comprehensive comparison of Uber and Lyft, contributing valuable information for stakeholders and researchers in the transportation industry.

Initial Data Collection Report

Data source

Transaction data: CSV export

- Format: CSV
- System: [ERP system name]
- Location: <https://www.kaggle.com/code/hkhoi91/data-visualization-uber-vs-lyft-ultimate-battle/input?select=weather.csv>
- Authentication: anyone with Admin role
- Cost and future availability: [details]

DATA DESCRIPTION REPORT

Dictionary

Dataset 1: **cab_rides** (Dataset of cab rides collected for a week in Nov - Dec '18. Collected at a regular interval of 5 mins)

- **Distance:** distance between source and destination
- **cab_type** : Specifies the type of ride service, either Uber or Lyft.
- **time_stamp:** epoch time when data was queried
- **Destination:** destination of the ride
- **Source:** the starting point of the ride
- **Price:** price estimate for the ride in USD
- **surge_multiplier:** the multiplier by which price was increased, default 1
- **Id:** unique identifier
- **product_id:** uber/Lyft identifier for cab-type
- **Name:** Visible type of the cab e.g.: Uber Pool, UberXL

Dataset 2: **weather** (Weather data for all the locations in consideration. Collected every hour)

- **Temp:** Temperature in F
- **Location:** Location name
- **Clouds:** Clouds
- **Pressure:** pressure in mb
- **Rain:** rain in inches for the last hour
- **time_stamp:** epoch time when row data was collected

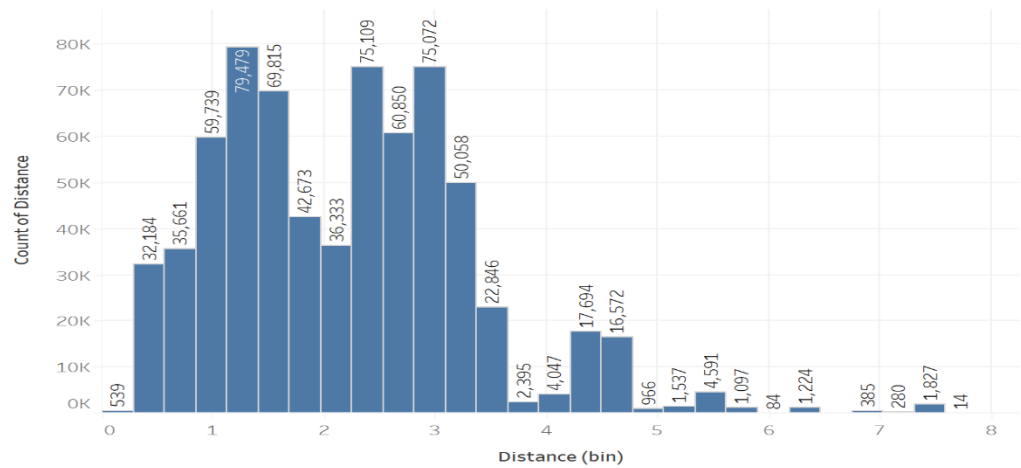
- **Humidity:** humidity in %
- **Wind:** wind speed in mph

Univariate Properties

Feature	Type	Count	Missing	Unique	Min	Q1	Med	Q3	Max	Mean	SD	Skew	Kurt
distance	float	693071	0	693071	0.020	1.280	2.160	2.920	7.860	2.189	1.139	0.83	1.23
Cab type	String		0										
Time stamp	integer	693071	0									0.43	-1.56
destination	String		0										
source	String		0										
price	Integer	637976	773	637976	2.50	9.00	13.50	22.50	97.50	16.55	9.32	1.05	1.22
surge multiplier	Integer	693071	0	693071	1	1	1	1	3	1.014	0.092	8.32	80.53
id	String		0										
Product id	String		0										
name	String		0										
temp	Integer	6276	0	6276	19.62	36.08	40.13	42.83	55.41	39.09	6.02	-0.64	0.95
location			0										
clouds	Integer	6276	0	6276	0	0.440	0.780	0.970	1	0.678	0.314	-0.63	-0.95
pressure	Float	6276	0	6276	988.25	997.75	1007.66	1018.48	1035.12	1008.45	12.87	0.26	-0.97
rain	float	894	5382	894	0.0002	0.0049	0.0149	0.0609	0.7807	0.0577	0.1008	3.80	20.04
Time stamp	Integer	6276	0	6276								0.87	-0.94
humidity	float	6276	0	6276	0.4500	0.6700	0.7600	0.8900	0.9900	0.7640	0.1273	-0.23	-0.95
wind	float	6276	0	6276	0.29	3.52	6.57	9.92	18.18	6.80	3.63	0.20	-0.77

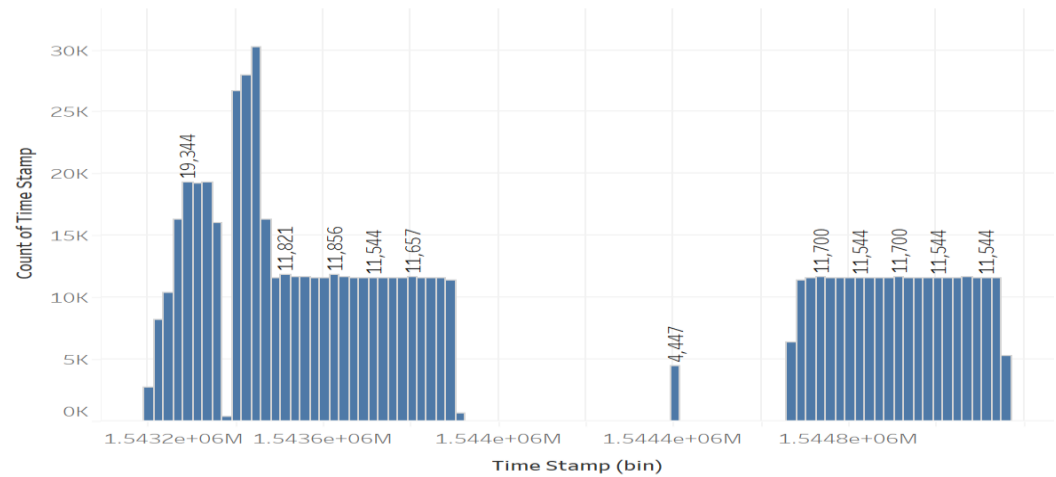
Univariate Visualizations

Histogram - Distance

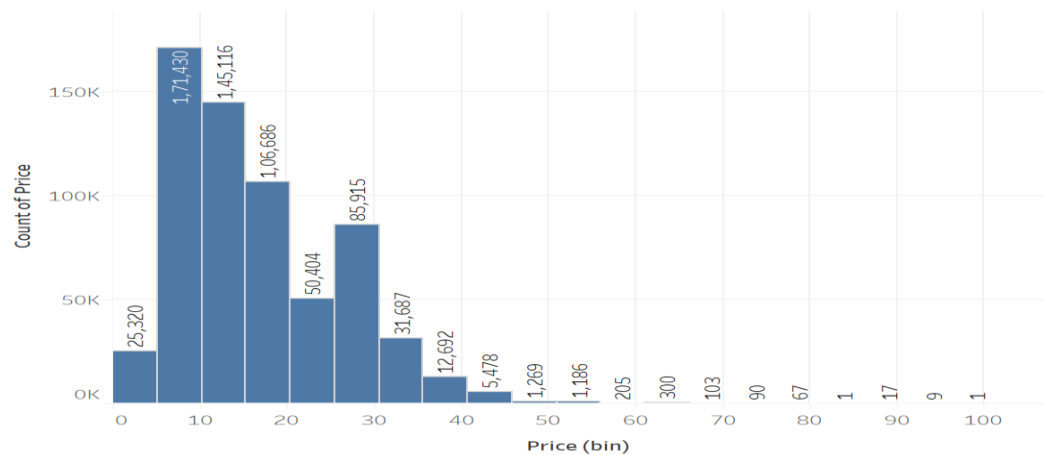


The trend of count of Distance for Distance (bin).

Histogram - Timestamp for Cabs

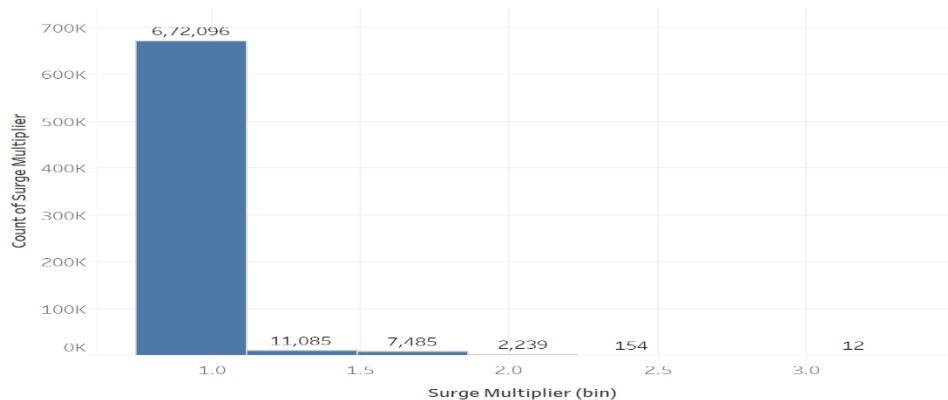


Histogram - Price



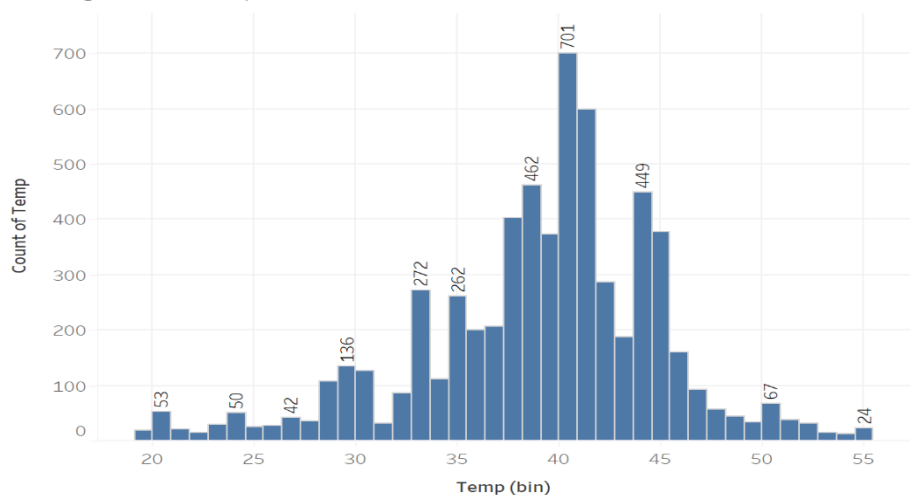
The trend of count of Price for Price (bin).

Histogram - Surge Multiplier



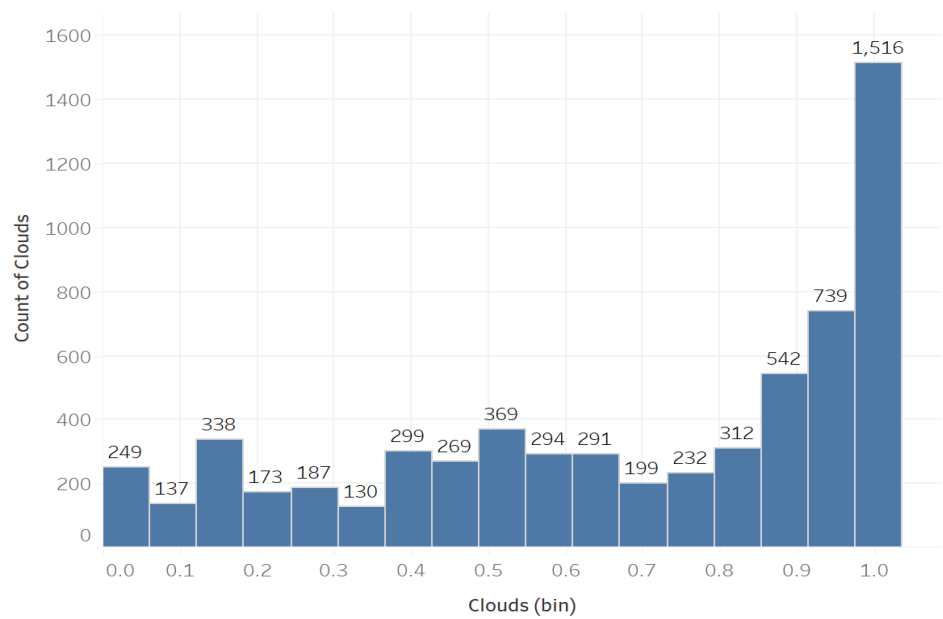
The trend of count of Surge Multiplier for Surge Multiplier (bin).

Histogram - Temp



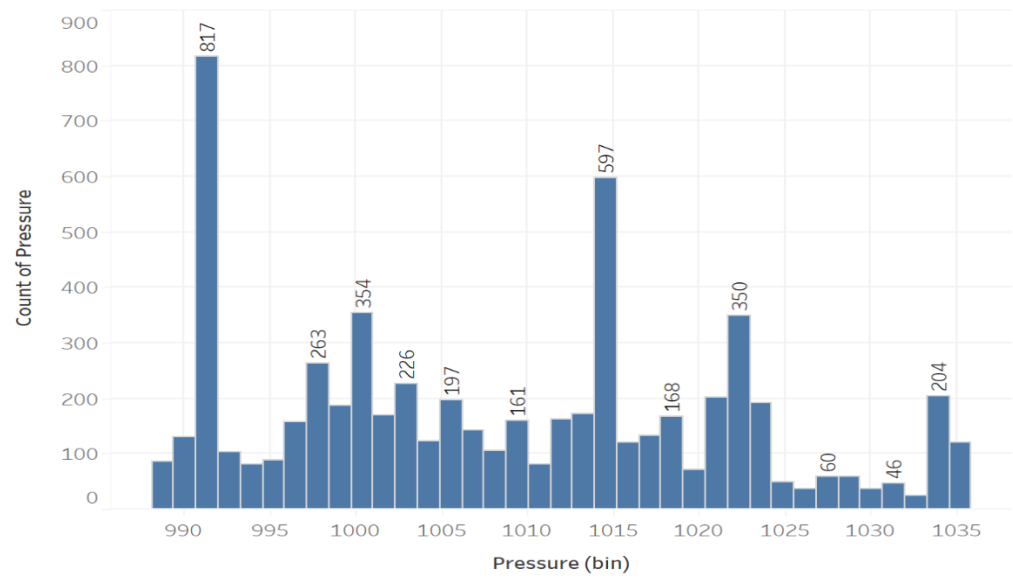
The trend of count of Temp for Temp (bin).

Histogram - Clouds



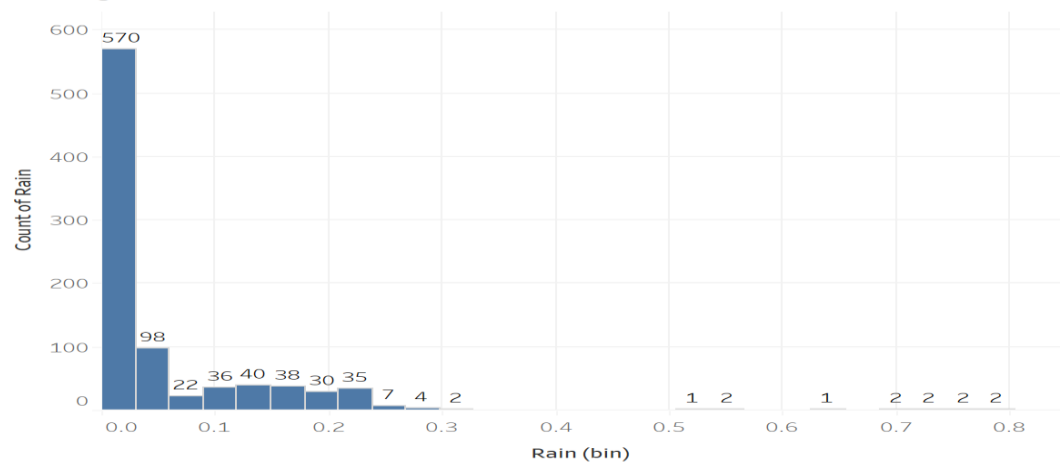
The trend of count of Clouds for Clouds (bin).

Histogram - Pressure



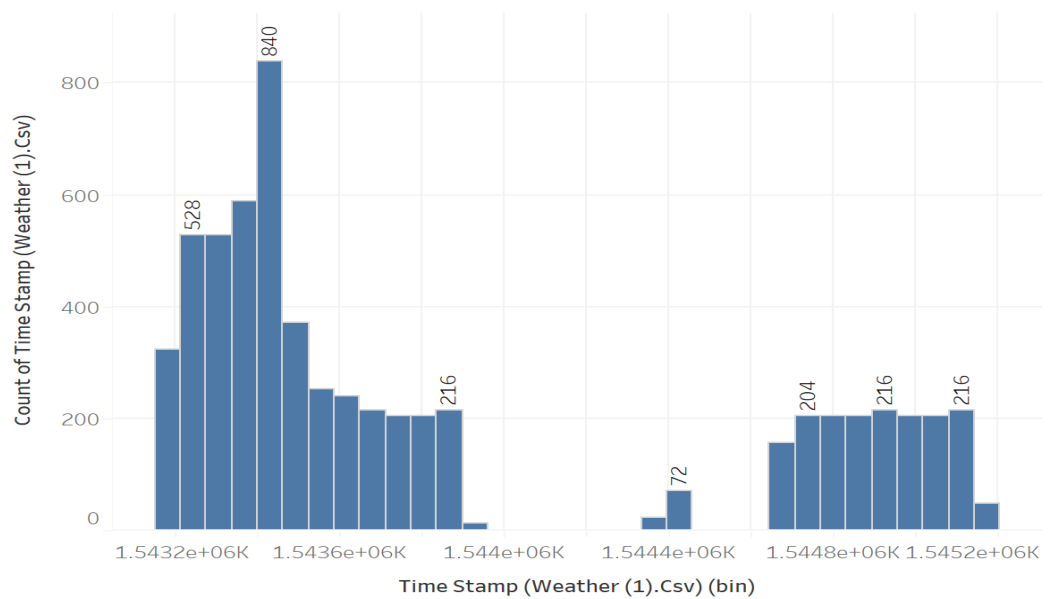
The trend of count of Pressure for Pressure (bin).

Histogram - Rain

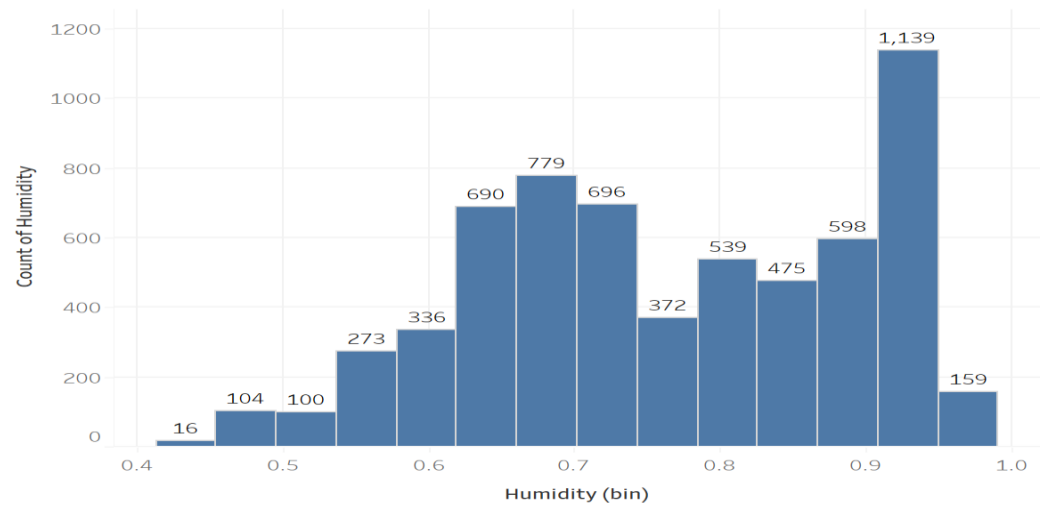


The trend of count of Rain for Rain (bin).

Histogram - Timestamp Weather

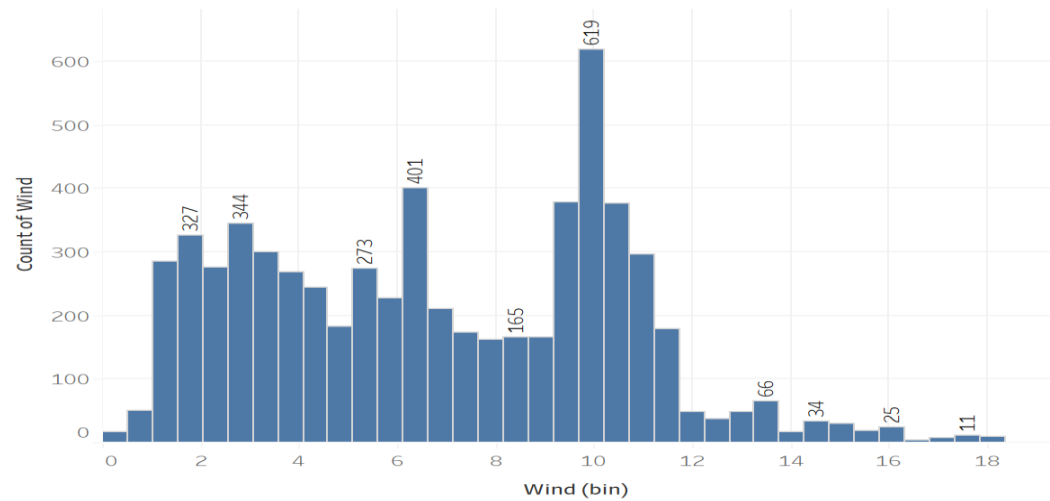


Histogram - Humidity



The trend of count of Humidity for Humidity (bin).

Histogram - Wind



The trend of count of Wind for Wind (bin).

Data Exploration Report

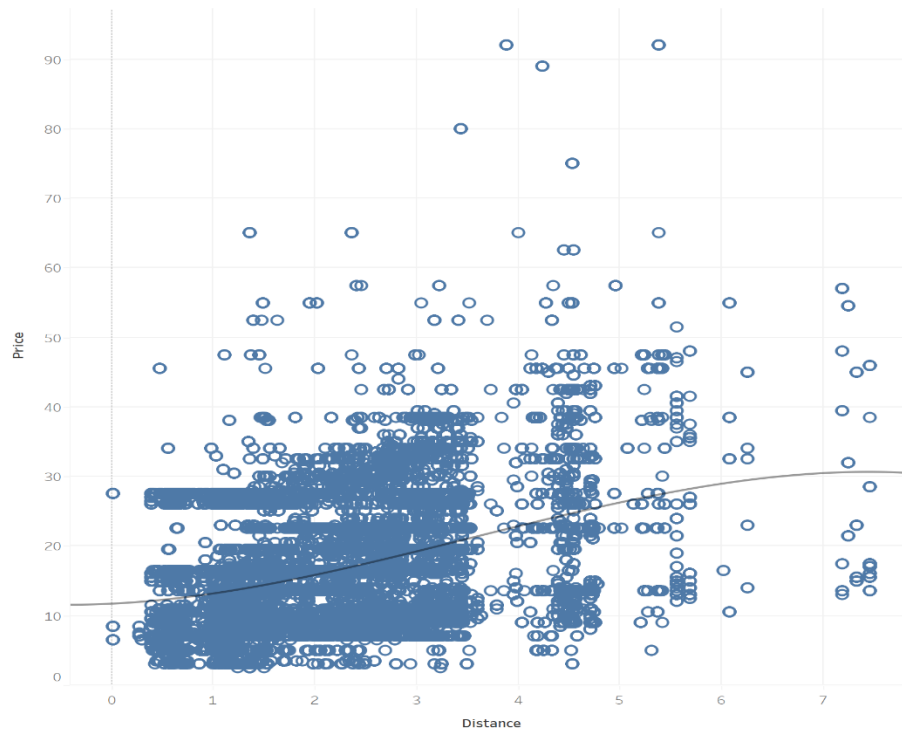
This report details the relationship between each potential feature with the label “Price”

Feature	Analysis	Effect size	P-value
distance	R squared	0.135859	< 0.0001
Cab type	One way ANOVA	23128.24	0.00
Time stamp (Cab)	R squared	0.0003086	0.0043483
destination	One way ANOVA	1064.34	0.00
source	One way ANOVA	1338.03	0.00
surge multiplier	R squared	0.0892004	< 0.0001
id	(primary key)	-	-
Product id	One way ANOVA	198877.84	0.00
name	One way ANOVA	198877.84	0.00
temp	R squared	0.0004131	< 0.0001
clouds	R squared	0.0008131	< 0.0001
pressure	R squared	0.0002461	0.0012014
rain	R squared	0.0058639	< 0.0001
Timestamp(Weather)	R squared	0.0003084	0.0043483
humidity	R squared	0.00078	< 0.0001
wind	R squared	0.0002954	0.0056161

The remainder of the report includes greater details on each relationship. We find that all the features are worth including during the modeling phase.

Distance

Scatter Plot - Distance



Distance vs. Price.

R-Squared: 0.135859

P-value: < 0.0001

Equation:

$$\text{Price} = -0.0787875 * \text{Distance}^3 + 0.83285 * \text{Distance}^2 + 0.71203 * \text{Distance} + 11.6935$$

Coefficients

Term	Value	StdErr	t-value	p-value
Distance^3	-0.0787875	0.010697	-7.36537	< 0.0001
Distance^2	0.83285	0.10464	7.95916	< 0.0001
Distance	0.71203	0.293115	2.42918	0.0151371
intercept	11.6935	0.232339	50.3293	< 0.000

Trend Lines Model

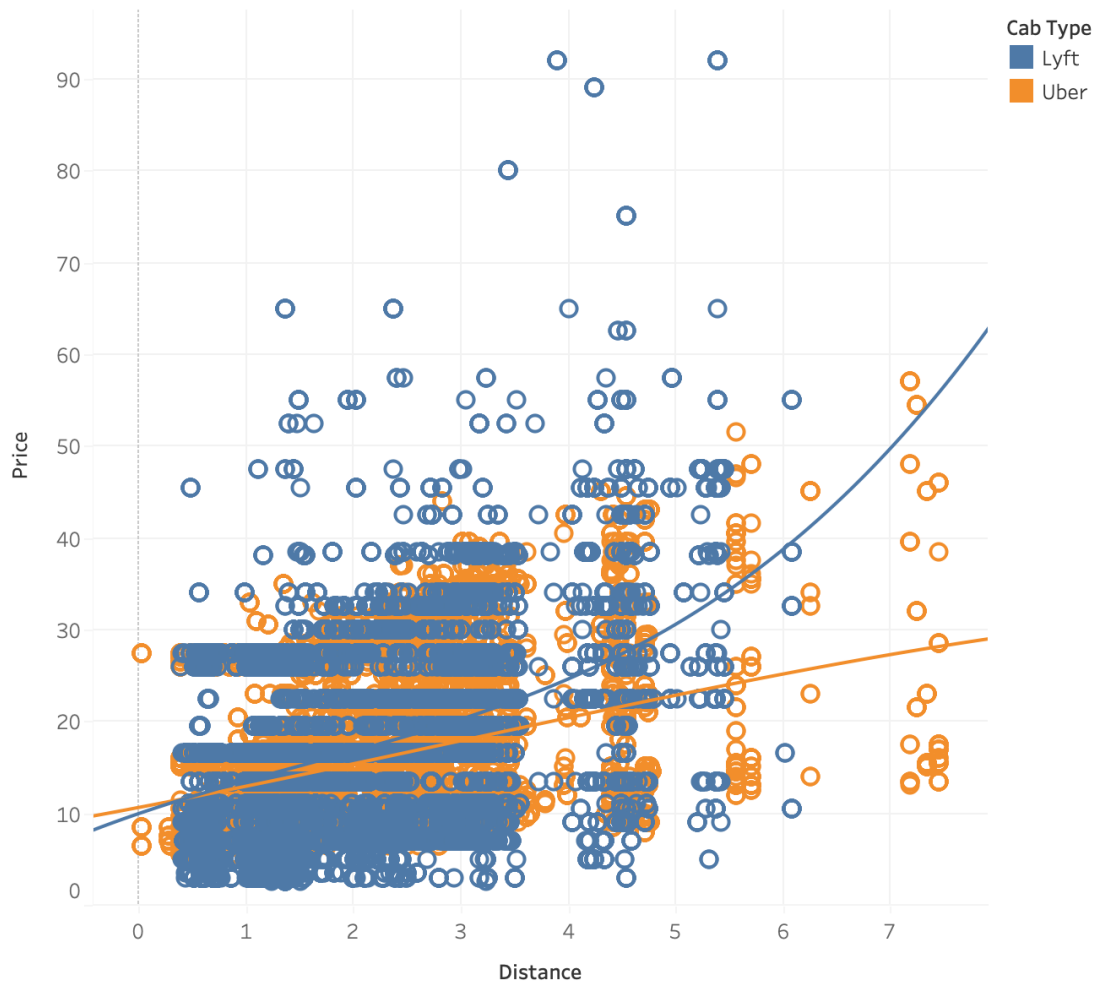
A polynomial trend model of degree 3 is computed for Price given Distance. The model may be significant at $p \leq 0.05$.

Inference from Graph:

Now, from the above visualization we learn that distance and price are not strongly correlated. As the distance increases, price does not increase linearly.

After comparing Prince vs Distance individually for Uber and Lyft we get the following visualization as below:

Distance - Scatter



Distance vs. Price. Color shows details about Cab Type.

From the above table we see that distance has the highest R squared value (0.135859) and has the some effect on the price even though they are not strongly correlated. And therefore, we drafted a new revised version of scatterplot for the same as you can see above. Looking at the scatterplot, there appears to be a consistent variance in prices for cab types. However, scatterplot and with trend lines clearly reveals that Lyft costs more than Uber.

Cab type

Bar Chart - Cab Type




Average of Price for each Cab Type.

Mean price of Lyft : 17.606

Mean price of uber: 15.915

In Summary we can conclude that Lyft cab costs more than Uber.

Price of Cab types

Cab Type	Product Id		Price
Lyft	lyft	33,425	 19,200 114,969
	lyft_line	19,200	
	lyft_lux	81,402	
	lyft_luxsuv	114,969	
	lyft_plus	53,515	
	lyft_premier	59,351	
Uber	6c84fd89-3f11-4782-9b5..	77,278	
	6d318bcc-22a3-4af6-bdd..	112,845	
	6f72dfc5-27f1-42e8-84db..	58,750	
	8cf7e821-f0d3-49c6-8eba..		
	9a0e7b09-b92b-4c41-977..	35,375	
	55c66225-fbe7-4fd5-907..	33,543	
	997acbb5-e102-41e1-b15..	33,272	

Sum of Price broken down by Cab Type and Product Id. Color shows sum of Price. The marks are labeled by sum of Price.

Inference from Heat map and BarCharts:

Lyft cab types generally have higher prices compared to Uber cab types. This is evident from the sum of prices displayed for each service. The Lyft lux SUV cab type has a significantly higher price compared to any Uber cab type. This is highlighted by the price difference where Lyft lux SUV costs \$114,969, which is markedly higher than the Uber options. On average, Lyft services tend to be more expensive than Uber services. This is supported by the mean price of Uber being 15.9, which is lower than Lyft's prices.

The heatmap visually confirms that Lyft's premium services (like Lyft lux and Lyft lux SUV) are priced higher than similar Uber services. Overall, Lyft's cab options, particularly the luxury ones, have higher costs compared to Uber's offerings.

Anova: Single Factor

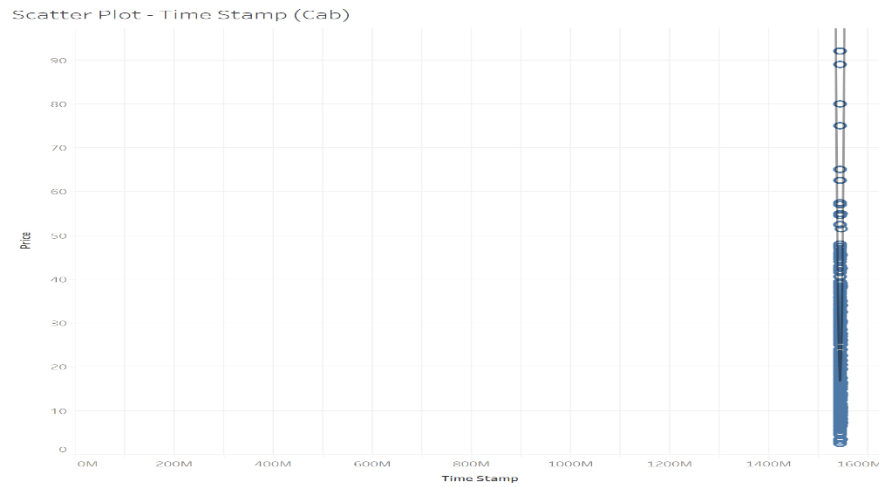
SUMMARY

<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Lyft	307408.00	5333957.98	17.35	100.38
Uber	307408.00	4160258.00	13.53	93.37

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	2240624.26	1.00	2240624.26	23128.24	0.00	3.84
Within Groups	59562117.24	614814.00	96.88			
Total	61802741.49	614815.00				

Time stamp (Cab Ride)



R-Squared: 0.0003086

P-value: 0.0043483

Equation:

$$\text{Price} = -1.57246e-21 \cdot \text{Time Stamp}^3 + 8.34268e-12 \cdot \text{Time Stamp}^2 + -0.0145148 \cdot \text{Time Stamp} + 8.31041e+06$$

Trend Lines Model:

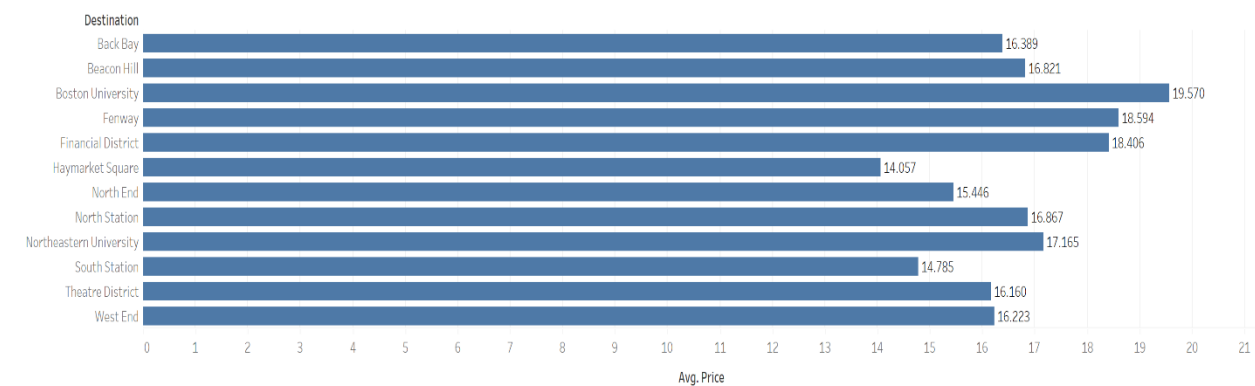
A polynomial trend model of degree 3 is computed for Price given Time Stamp. The model may be significant at $p \leq 0.05$.

Inference from Graph:

Since the data of cab rides has been collected for one hour, we could see that all scatter plots to be concentrated in between that time as in the report. Given the low R-squared value, factors other than timestamp must be influencing the price significantly. But since we are having significant p value there is some relation between time when people are using cabs and price of the cabs

Destination

Bar Chart - Destination



Average of Price for each Destination.

Inference from BarCharts

The bar chart compares different destinations, displaying the sum of prices for each destination. - The destinations with the highest sums are "Boston University" and "Fenway," indicating these areas have higher average prices. - The destination "Haymarket Square" has the lowest sum, suggesting it is the least expensive among the listed destinations

Anova: Single Factor

SUMMARY

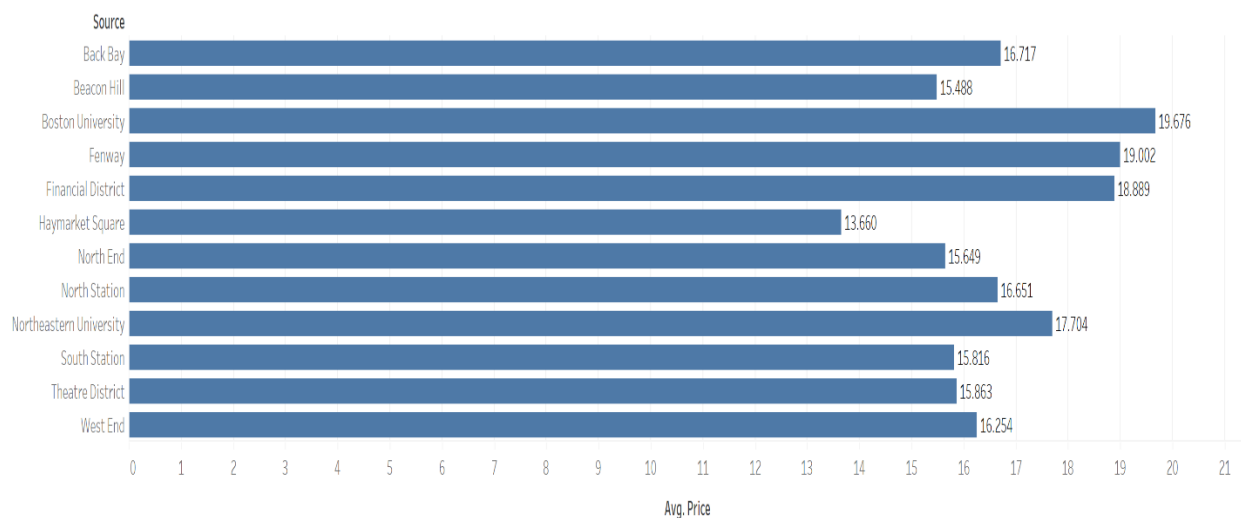
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Back Bay	57780	862218.00	14.92	89.43
Beacon Hill	57403	858562.00	14.96	83.41
Boston University	57764	1007172.35	17.44	126.59
Fenway	57757	964772.50	16.70	112.35
Financial District	58851	977964.00	16.62	140.62
Haymarket Square	57764	757982.00	13.12	73.93
North End	57756	797577.50	13.81	80.98
North Station	57119	883569.00	15.47	107.77
Northeastern University	57755	947799.93	16.41	107.91
South Station	57749	788270.85	13.65	78.80
Theatre District	57798	849665.00	14.70	84.89
West End	57575	859839.85	14.93	93.36

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1152065.01	11.00	104733.18	1064.34	0.00	1.79
Within Groups	68198678.00	693059.00	98.40			
Total	69350743.01	693070.00				

Source

Bar Chart - Source



Average of Price for each Source.

Inference from BarCharts

The bar chart compares different sources, displaying the sum of prices for each destination. - The destinations with the highest sums are "Boston University" and "Fenway," indicating these areas have higher average prices. - The sources "Haymarket Square" has the lowest sum, suggesting it is the least expensive among the listed destinations

Anova: Single Factor

SUMMARY

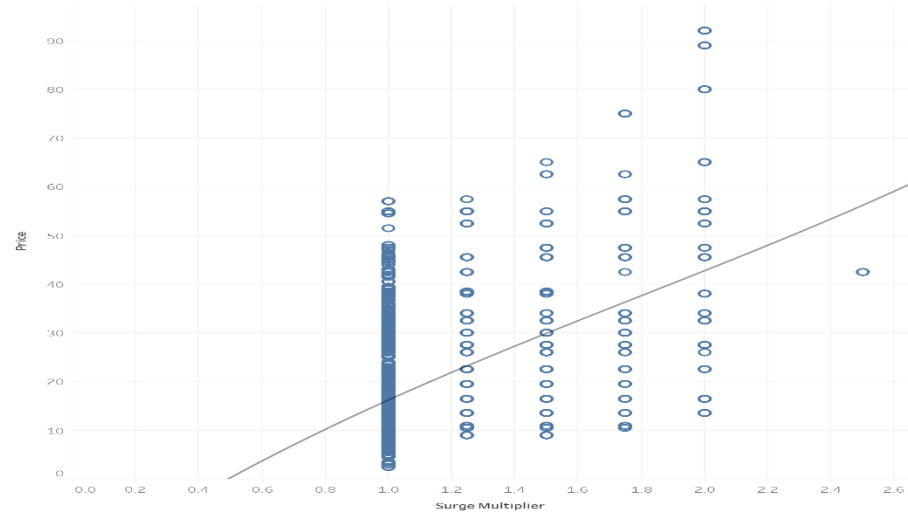
<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>
Boston University	57764	1002453.50	17.35	125.89
Fenway	57757	977164.00	16.92	118.70
Financial District	58857	985375.80	16.74	142.39
Haymarket Square	57736	721636.00	12.50	66.61
North End	57763	805720.00	13.95	77.30
North Station	57118	860354.00	15.06	93.54
Northeastern University	57756	951695.35	16.48	111.41
South Station	57750	833149.00	14.43	84.23
Theatre District	57813	882976.70	15.27	100.66
West End	57562	853428.20	14.83	86.81

ANOVA

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Groups	1214432.28	9.00	134936.92	1338.03	0.00	1.88
Within Groups	58276300.68	577866.00	100.85			
Total	59490732.96	577875.00				

Surge Multiplier

Scatter Plot - Surge Multiplier



Surge Multiplier vs. Price.

R-Squared: 0.0892004

P-value: < 0.0001

Equation:

Price = 1.90004*Surge Multiplier^3 + -10.2339*Surge Multiplier^2 + 43.9387*Surge Multiplier + -19.3968

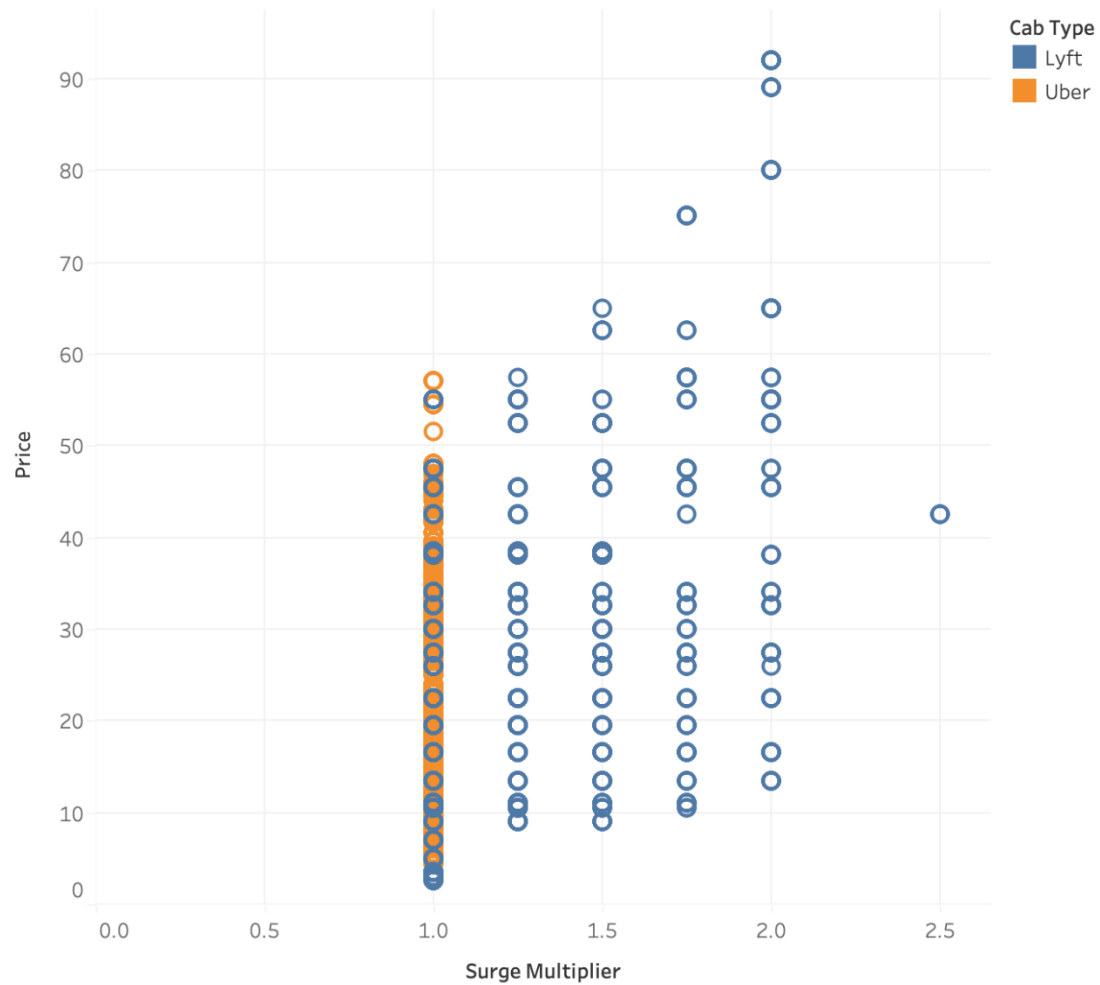
Coefficients

Term	Value	StdErr	t-value	p-value
Surge Multiplier^3	1.90004	4.45891	0.426123	0.67002
Surge Multiplier^2	-10.2339	19.9399	-0.513236	0.607789
Surge Multiplier	43.9387	28.7028	1.53082	0.125822
intercept	-19.3968	13.249	-1.46402	0.143196

Trend Lines Model:

A polynomial trend model of degree 3 is computed for Price given Surge Multiplier. The model may be significant at $p \leq 0.05$.

Surge multiplierVs Price

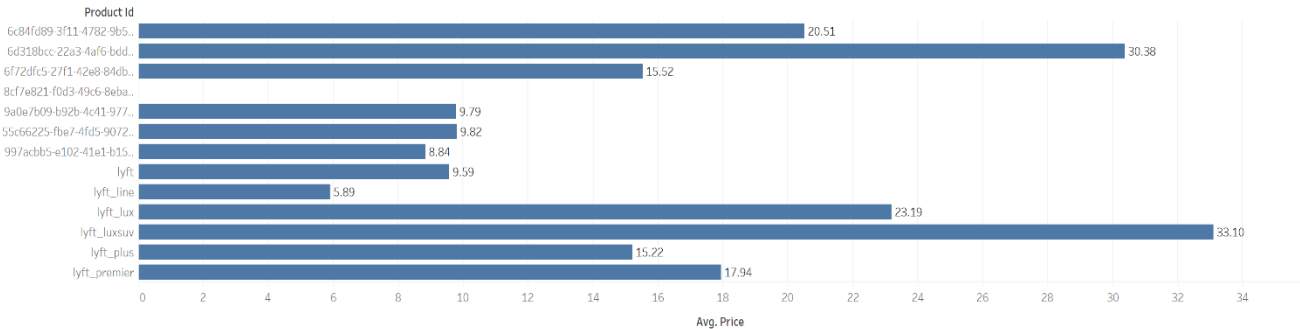


Surge Multiplier vs. Price. Color shows details about Cab Type.

We could also see that Uber follows a pattern of surge multiplier (1) and it does not vary much. From scatterplots we could understand that Lyft Surge multiplier Varies up to two folds. Also, the effect size (0.0892004) is next highest to distance which plays a major role in label price

Product id

Bar Chart - Product ID



Average of Price for each Product Id.

Summary Statistics:				
Groups	Count	Sum	Average	Variance
55c66225-fbe7-4fd5-9072-eab1ece5e23e	55094	537997	9.765074	6.076722
6c84fd89-3f11-4782-9b50-97c468b19529	55095	1130758	20.523786	24.522599
6d318bcc-22a3-4af6-bddd-b409bfce1546	55096	1668679.5	30.286763	23.387831
6f72dfc5-27f1-42e8-84db-ccc7a75f6969	55096	863803	15.678144	20.465451
997acbb5-e102-41e1-b155-9df7de0a73f2	55091	482184	8.7525	4.436292
9a0e7b09-b92b-4c41-9779-2ad22b4d779d	55096	538013.5	9.765019	6.076642
lyft	51235	492413.68	9.610885	6.402442
lyft_line	51233	308929.5	6.029893	4.442526
lyft_lux	51235	1181605.5	23.062468	41.89821
lyft_luxsuv	51235	1656124.5	32.324086	51.568108
lyft_plus	51235	784375.2	15.309363	20.752702
lyft_premier	51235	910509.5	17.77124	28.005086

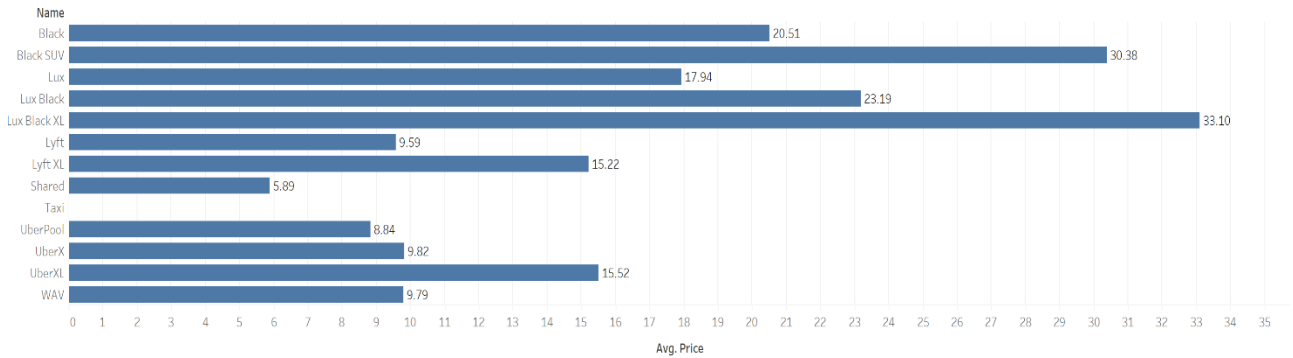
ANOVA Table:						
Source of Variation	SS	df	MS	F	P-value	F-crit
Between Groups	46848470	12	3904039.17	198877.8436	0	1.7539
Within Groups	12523450	637964	19.63			

Inference from BarCharts

The Lyft lux SUV (product ID `lyft_luxsuv`) also shows a high average price of around 33.10

Name

Bar Chart- Name



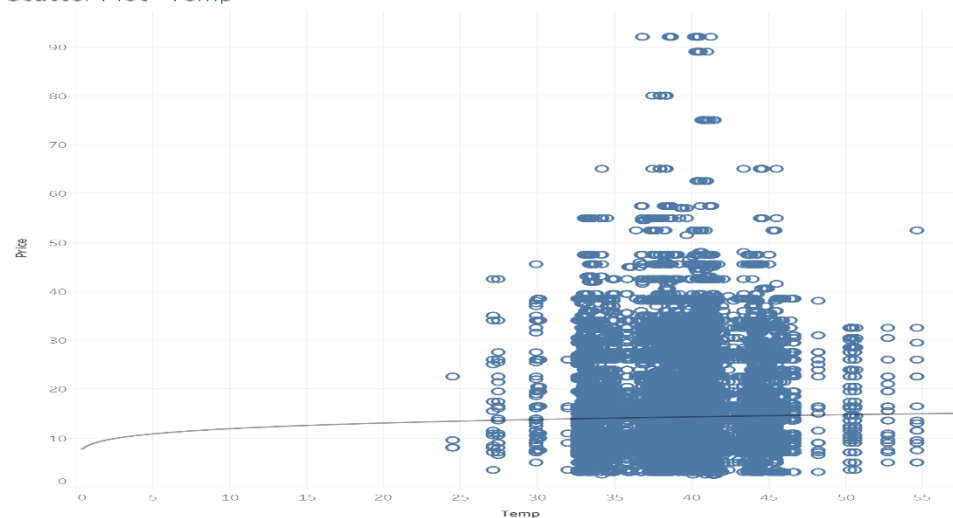
Average of Price for each Name.

Summary Statistics:				
Groups	Count	Sum	Average	Variance
Black	55095	1130758	20.52	24.52
Black SUV	55096	1668679.5	30.29	23.39
Lux	51235	910509.5	17.77	28.01
Lux Black	51235	1181605.55	23.06	41.9
Lux Black XL	51235	1656124.55	32.32	51.57
Lyft	51235	492413.68	9.61	6.4
Lyft XL	51235	784375.2	15.31	20.75
Shared	51233	308929.5	6.03	4.44
UberPool	55091	482184	8.75	4.44
UberX	55094	537997	9.77	6.08
UberXL	55096	863803	15.68	20.47
WAV	55096	538013.5	9.77	6.08

ANOVA Table:						
Source of Variation	SS	df	MS	F	P-value	F-crit
Between Groups	46848470	12	3904039.17	198877.84	0	1.7539
Within Groups	12523450	637964	19.63			
Total	59371920	637976				

Temp

Scatter Plot - Temp



Temp vs. Price.

R-Squared: 0.0004131

P-value: < 0.0001

Equation:

$$\ln(\text{Price}) = 0.134354 \cdot \ln(\text{Temp}) + 2.16627$$

Coefficients

Term	Value	StdErr	t-value	p-value
ln(Temp)	0.134354	0.0320159	4.19648	< 0.0001
intercept	2.16627	0.117315	18.4655	< 0.0001

Trend Lines Model:

A linear trend model is computed for natural log of Price given natural log of Temp. The model may be significant at $p \leq 0.05$.

Inference from Graph

The graph shows that people take the cabs around the temperature greater than 25 to 55 degrees and we see that there are some dependencies of temperature on the price as we have a significant P value which suggests that Temperatures has some effect on the prices of the cab fares

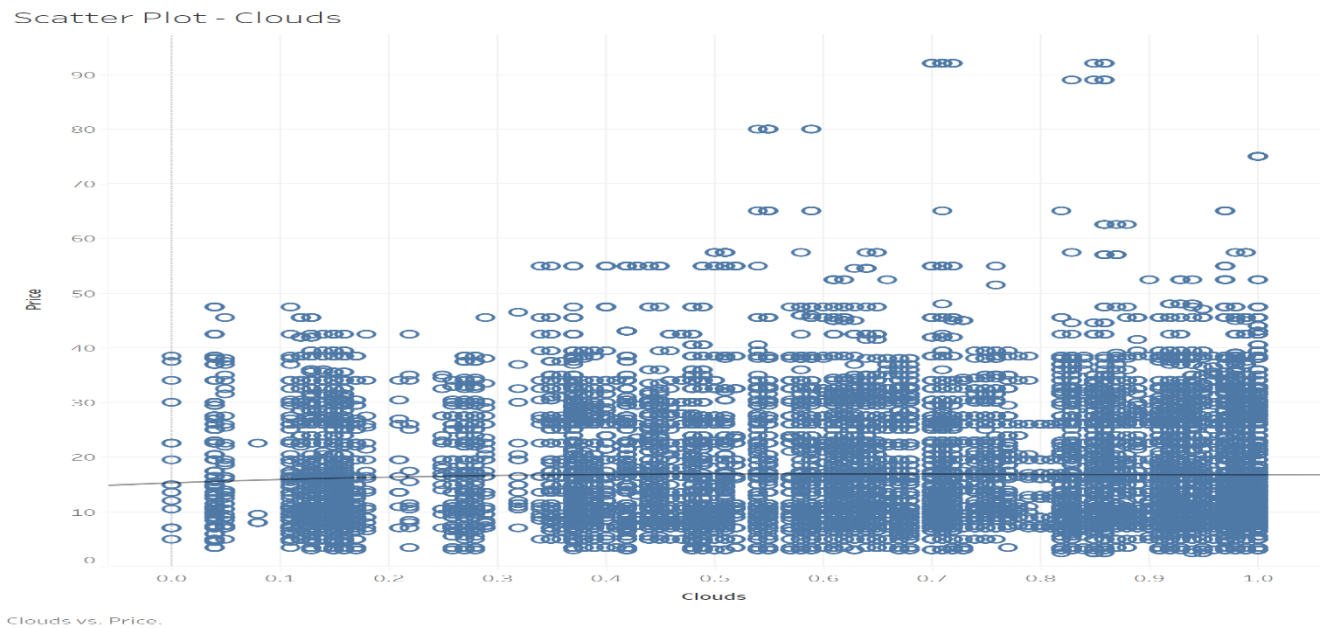
Location

Bar Chart - Location



Average of Price for each Location.

Clouds



R-Squared: 0.0008131

P-value: < 0.0001

Equation:

$$\text{Price} = 4.19631 \cdot \text{Clouds}^3 + -9.78413 \cdot \text{Clouds}^2 + 7.086 \cdot \text{Clouds} + 15.2828$$

Coefficients

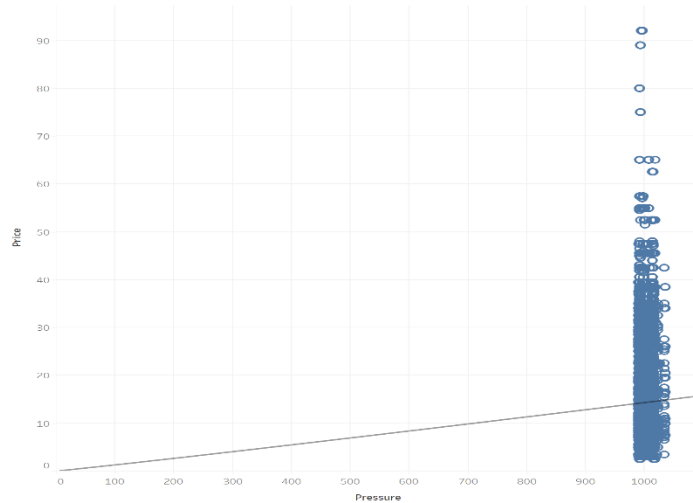
Term	Value	StdErr	t-value	p-value
Clouds^3	4.19631	2.76051	1.52012	0.128488
Clouds^2	-9.78413	4.6714	-2.09448	0.0362235
Clouds	7.086	2.26833	3.12389	0.001786
intercept	15.2828	0.301995	50.6062	< 0.0001

Trend Lines Model:

A polynomial trend model of degree 3 is computed for Price given Clouds. The model may be significant at $p \leq 0.05$.

Pressure

Scatter Plot - Pressure



Pressure vs. Price.

R squared value: 0.0002461

P-value :0.0012014

Equation:

$$\ln(\text{Price}) = 1.04954 * \ln(\text{Pressure}) + -4.59$$

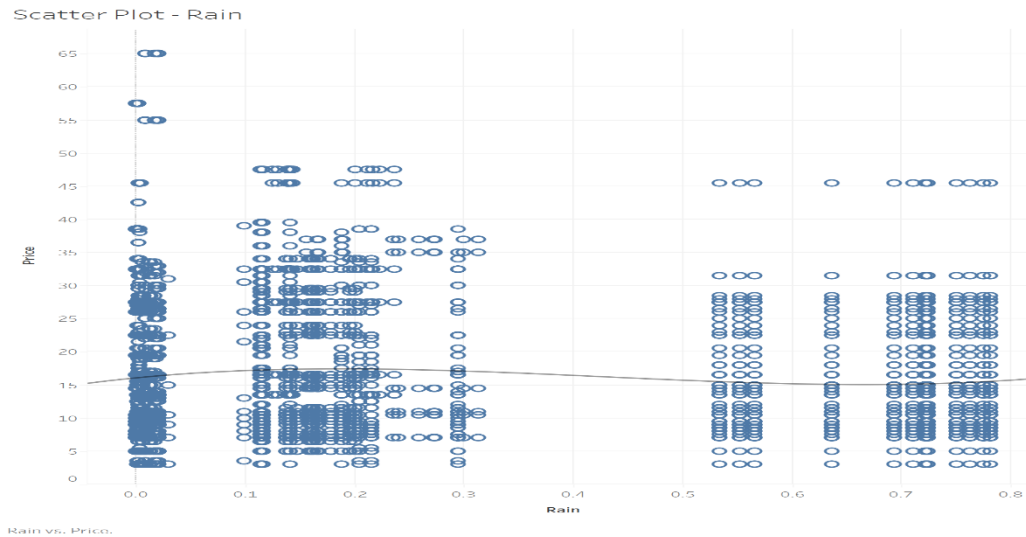
Coefficients

Term	Value	StdErr	t-value	p-value
ln(Pressure)	1.04954	0.324056	3.23877	0.0012014
intercept	-4.59	2.23802	-2.05092	0.0402813

Trend Lines Model:

A linear trend model is computed for natural log of Price given natural log of Pressure. The model may be significant at $p \leq 0.05$.

Rain



R-Squared: 0.0058639

P-value: < 0.0001

Equation:

$$\text{Price} = 43.1689 \cdot \text{Rain}^3 + -55.2639 \cdot \text{Rain}^2 + 16.1509 \cdot \text{Rain} + 16.0449$$

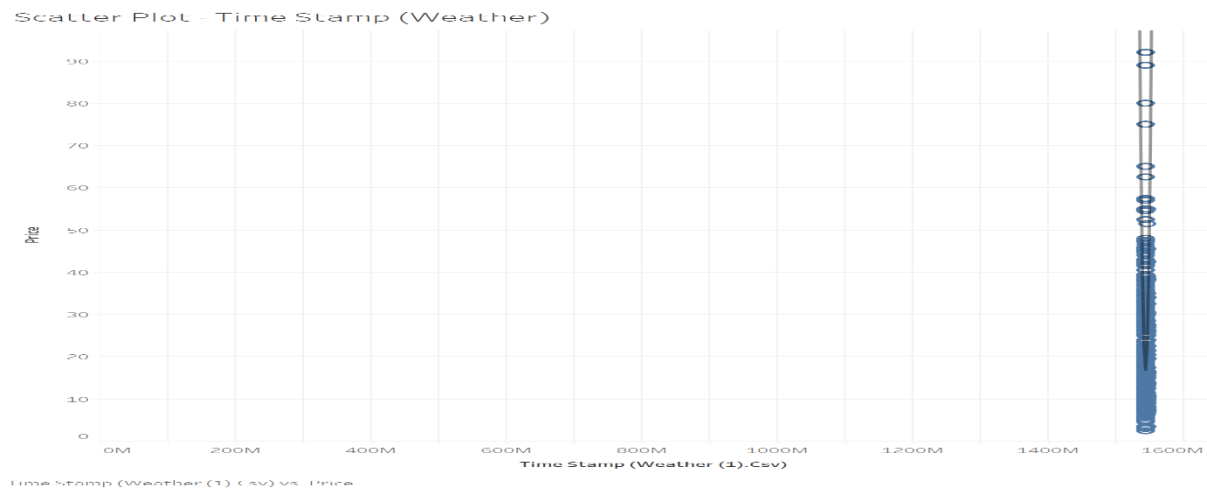
Coefficients

Term	Value	StdErr	t-value	p-value
Rain^3	43.1689	16.7501	2.57724	0.0099933
Rain^2	-55.2639	17.2053	-3.21203	0.001328
Rain	16.1509	4.24282	3.80663	0.0001429
intercept	16.0449	0.214515	74.7961	< 0.0001

Trend Lines Model:

A polynomial trend model of degree 3 is computed for Price given Rain. The model may be significant at $p \leq 0.05$.

Time stamp (Weather)



R-Squared: 0.0003084

P-value: 0.0043483

Equation:

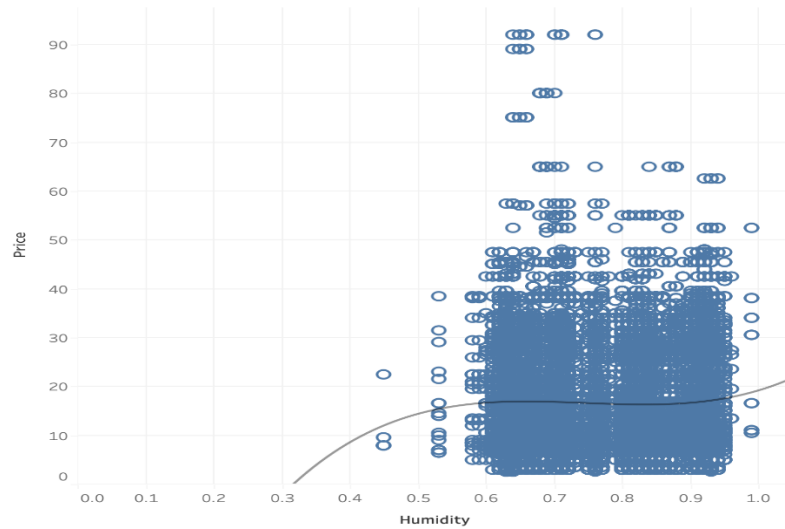
Price = $-1.57246e-21 \cdot \text{Time Stamp (Weather (1).Csv)}^3 + 8.34268e-12 \cdot \text{Time Stamp (Weather (1).Csv)}^2 - 0.0145148 \cdot \text{Time Stamp (Weather (1).Csv)} + 8.31041e+06$

Trend Lines Model:

A polynomial trend model of degree 3 is computed for Price given Time Stamp (Weather (1).Csv). The model may be significant at $p \leq 0.05$.

Humidity

Scatter plot-Humidity



Humidity vs. Price.

R-squared 0.00078

Pvalue: < 0.0001

Equation:

$$\text{Price} = 240.551 \cdot \text{Humidity}^3 + -538.12 \cdot \text{Humidity}^2 + 395.798 \cdot \text{Humidity} + -79.059$$

Coefficients

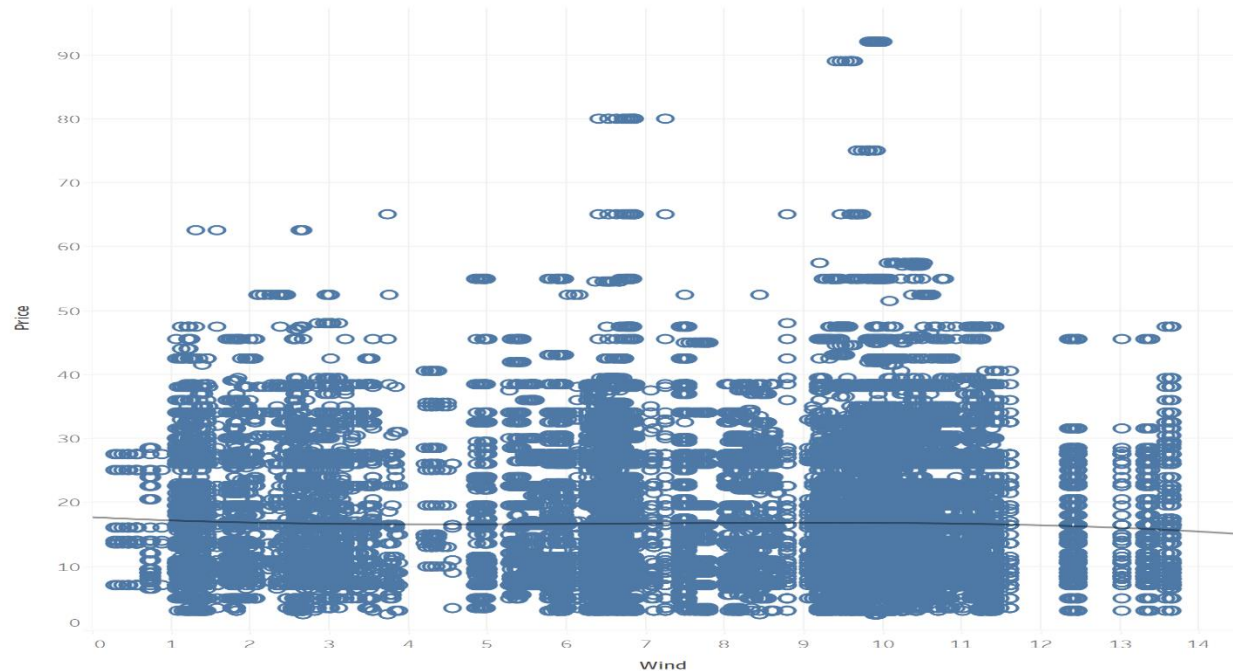
Term	Value	StdErr
Humidity^3	240.551	55.8401
Humidity^2	-538.12	129.656
Humidity	395.798	99.5199
intercept	-79.059	25.2545

Trend Lines Model

A polynomial trend model of degree 3 is computed for Price given Humidity. The model may be significant at $p \leq 0.05$.

Wind

Scatter Plot - Wind



Wind vs. Price.

R-Squared: 0.0002954

P-value: 0.0056161

Equation:

$$\text{Price} = -0.00491493 \cdot \text{Wind}^3 + 0.10022 \cdot \text{Wind}^2 - 0.600275 \cdot \text{Wind} + 17.6747$$

Coefficients

Term	Value	StdErr	t-value	p-value
Wind^3	-0.0049149	0.0014926	-3.29278	0.0009928
Wind^2	0.10022	0.0311632	3.21596	0.0013011
Wind	-0.600275	0.192844	-3.11275	0.0018547
intercept	17.6747	0.32577	54.2552	< 0.0001

Trend Lines Model:

A polynomial trend model of degree 3 is computed for Price given Wind. The model may be significant at $p \leq 0.05$.

DATA CLEANING REPORT

No cleaning was done before importing the data on AMLS.

3 pills were used to clean the data within the AMLS Designer.

- **Clean missing data pill:** This pill was used to clean the label column(price) which had 723 missing values, and one feature (rain), which had 5382 missing values. The price misses were replaced by the median since the data is skewed, whilst the rain misses were replaced by mode.
- **Apply Math Transformation pill:** This label(price) column was extremely positively skewed and therefore this pill was used to correct skewness using the Logarithmic (LN) math function.
- **Clip values pill:** Addressing of outliers was done on the label column (price) using the Empirical rule.

MODEL TESTING

Features	Algorithm	MAE	RMSE	R-squared
All	Linear Regression	0.10258	0.154488	0.921464
All	Boosted Decision Tree	0.065151	0.100145	0.966998
All	Decision Forest Regression	0.017452	0.057717	0.989038
All	Neural Network Regression	0.097109	0.146668	0.929213
All	Poisson Regression	0.108882	0.162427	0.913185
All except Location and rain	Decision Forest Regression	0.01591	0.05477	0.98995

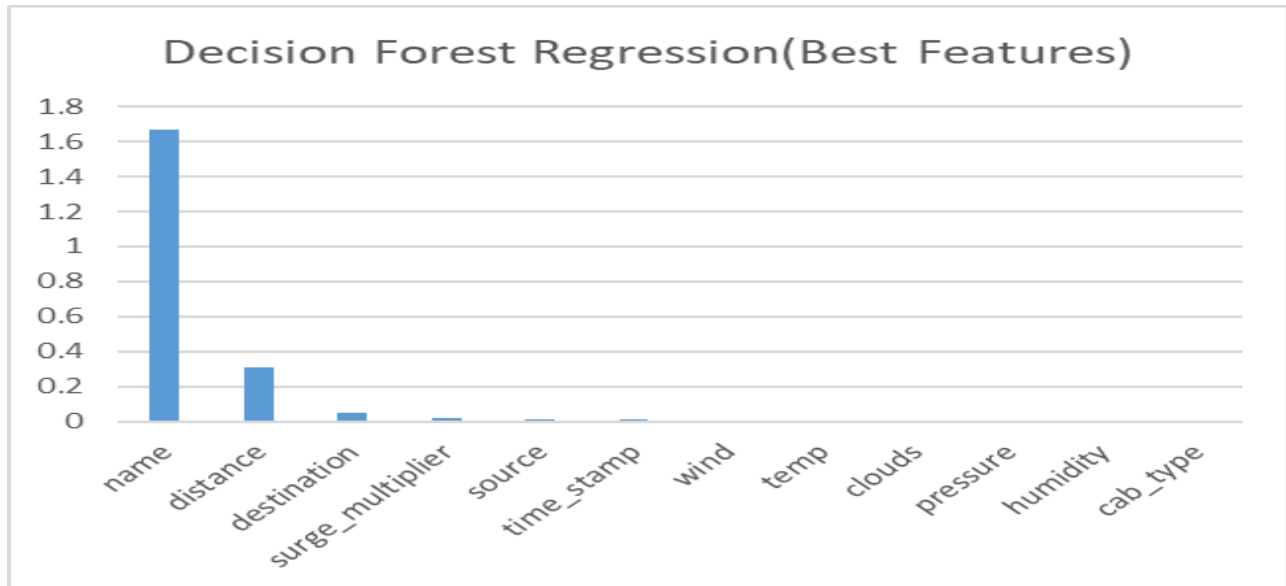
Inference from Table:

Looking at the above table we can conclude that Decision Forest Regression has the Highest values for R square.

To further enhance the value, we removed 2 features and found the value to increase.

Here is a List of best features based on Decision Forest Regression

Feature	Score
name	1.666156
distance	0.31271
destination	0.05113
surge_multiplier	0.022999
source	0.014481
time_stamp	0.01025
wind	0.008012
temp	0.007368
clouds	0.007108
pressure	0.006867
humidity	0.003345
cab_type	0.001761



HYPERPARAMETERS

Runs	Algorithm	Features	Random Sweeps	MAE	RMSE	Rsquared
1	Decision Forest Regression	All except Location and rain	5	0.012978	0.041036	0.994359
2	Decision Forest Regression	All except Location and rain	10	0.004463	0.019883	0.998676
3	Decision Forest Regression	All except Location and rain	20	0.004463	0.019883	0.998676
4	Decision Forest Regression	All except Location and rain	50	0.004463	0.019883	0.998676

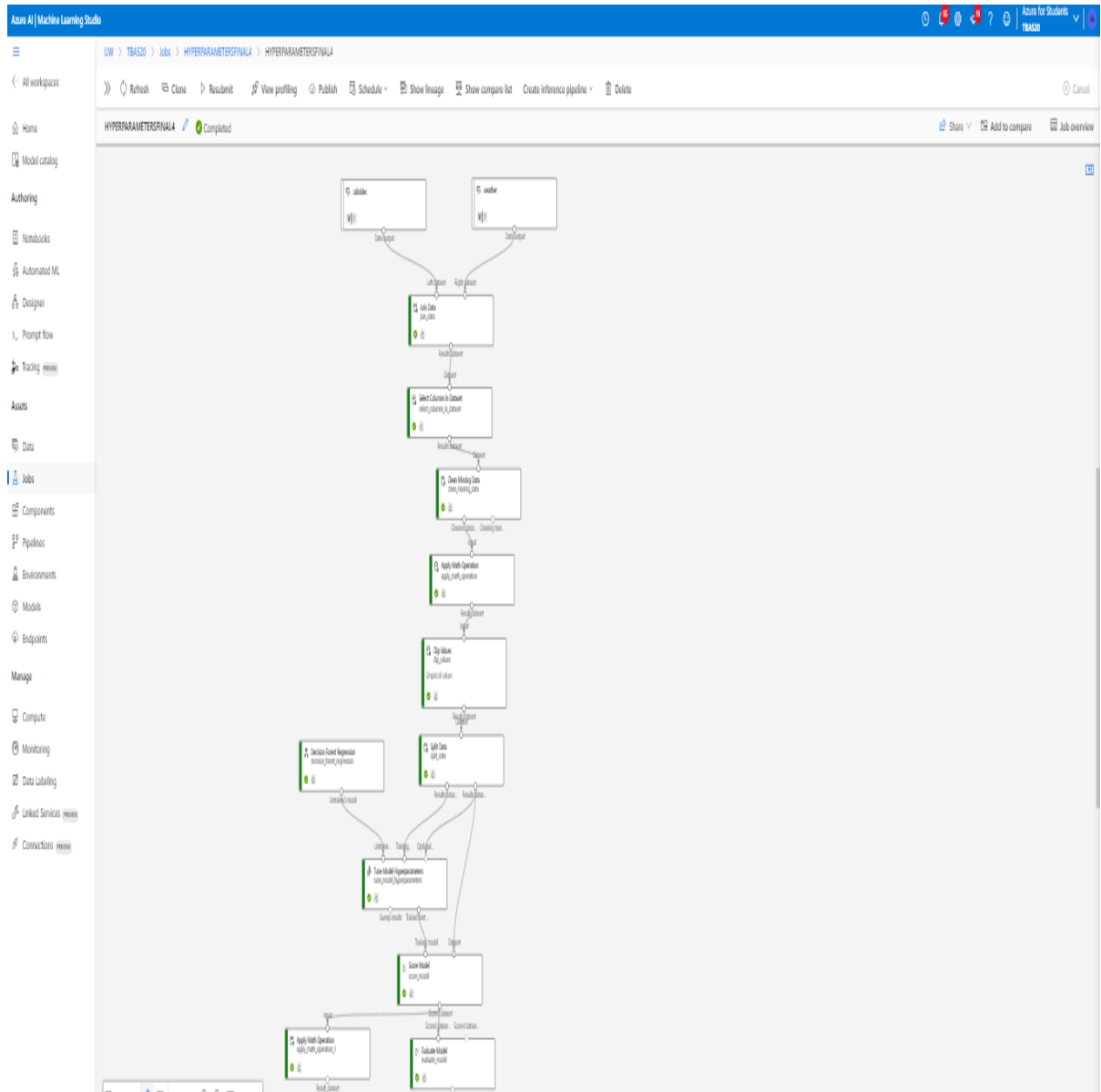
Inference from Table:

Looking at the table we can conclude that with Random sweeps = 10,20 or 50 we are achieving the same level for R square.

Best and highest R-squared achieved = **0.998676**

AMLS Designer pipeline

Here is the screenshot for our final version of the pipeline with the best features selected.



FINDINGS

Findings and Implications Report from Visualizations

1. Price Comparison Between Lyft and Uber:

- **Mean price of Lyft:** \$17.60
- **Mean price of Uber:** \$15.9
- **Implication:** Lyft rides are generally more expensive than Uber rides. This could influence customer choice based on price sensitivity, and it suggests that Uber may have a competitive edge in attracting cost-conscious riders.

2. Price Distribution Analysis:

- The price distribution for both Lyft and Uber shows that most rides fall within the \$0-\$40 range.
- **Implication:** Both services offer a range of pricing options, but Lyft has a slightly broader spread of higher prices, indicating potentially more premium service offerings.

3. Variance in Ride Prices:

- Lyft has a higher variance in ride prices compared to Uber.
- **Implication:** This suggests that Lyft may have more variability in its pricing, possibly due to a wider range of service types and dynamic pricing models.

4. Influence of Cab Types:

- Certain Lyft cab types (e.g., Lux, Lux Black XL) have significantly higher average prices than regular Lyft rides.
- **Implication:** Lyft's diverse range of premium cab options could cater to a market segment looking for higher-end ride experiences, while Uber's offerings might be more uniformly priced.

Findings and Implications Report from AMLS

- We find that **name** is the feature that plays an important role in determining cab prices.
To add to the findings Lyft cab types (e.g., Lux, Lux Black XL) have significantly higher average prices than regular Lyft rides.

Implication: Lyft's diverse range of premium cab options could cater to a market segment looking for higher-end ride experiences, while Uber's offerings might be more uniformly priced.

- **Distance** is the next important feature that affects cab prices.
Lyft has a lower minimum fare compared to Uber, at the same time Lyft has higher maximum fare. Also, Uber travel distances are generally higher than Lyft.
- The **destinations** are the next important factor that decides the price of the cab from the AIML model. To add it up from the visualization, the destinations with the highest sums are "Boston University" and "Fenway," indicating these areas have higher average prices. - The destination "Haymarket Square" has the lowest sum, suggesting it is the least expensive among the listed destinations
- **Surge multiplier** is the fourth important feature in deciding the price. Uber follows a pattern of 1 from visualization whereas Lyft values vary many folds. Uber's prices are more clustered, indicating a more consistent pricing strategy.

The Best Decision

Lyft may need to consider the balance between maximizing revenue and maintaining customer satisfaction by possibly limiting the extent of surge multipliers. - Uber might explore slight variations in its surge multipliers to capitalize on peak demand periods without alienating customers who prefer predictable pricing. Overall, the differences in surge multiplier strategies between Lyft and Uber highlight their distinct approaches to pricing and customer management, with significant implications for market positioning and customer satisfaction.

Business Decisions

Pricing Strategy:

- Lyft generally has higher fares than Uber. Lyft can consider reviewing its pricing strategy to remain competitive, particularly for premium services like Lyft Lux SUV.
- Both companies should evaluate the impact of surge pricing on customer satisfaction and demand, considering Lyft's higher variability in surge multipliers.

Market Positioning:

- Lyft can emphasize its luxury and premium services in its marketing campaigns, highlighting the unique value and experiences offered.
- Uber can leverage its lower average fares to attract price-sensitive customers and promote its cost-effective options.

Service Optimization:

- Analyze and optimize the pricing and availability of services in high-demand areas like Boston University and Fenway to maximize revenue.
- Evaluate ride demand patterns and adjust resource allocation accordingly to improve efficiency and customer satisfaction.

Customer Experience Enhancement:

- Both companies should consider customer feedback to refine service offerings and improve ride experiences, focusing on areas with high ride prices and demand.

Steps to Improve the AMLS Pipeline

Data Quality and Consistency:

- Ensure high data quality by implementing robust data cleaning and preprocessing steps to handle missing values and inconsistencies.
- Regularly update and validate data sources to maintain accuracy and reliability.

Feature Engineering:

- Incorporate additional contextual features like traffic conditions, time of day, and events that could impact ride prices and demand.
- Use advanced feature engineering techniques to capture complex relationships between variables and improve model performance.

Model Optimization:

- Experiment with different machine learning algorithms and hyperparameter tuning to enhance prediction accuracy.
- Implement ensemble methods to combine the strengths of multiple models and achieve better results.

Performance Monitoring and Evaluation:

- Establish a continuous monitoring system to track model performance and detect issues in real time.
- Use A/B testing to evaluate the impact of model updates and new features on business metrics.

Scalability and Automation:

- Ensure the AMLS pipeline is scalable to handle increasing data volumes and complexity.
- Automate data processing and model training workflows to reduce manual intervention and improve efficiency.

Implementing these steps can help in making informed business decisions and continuously improving the AMLS pipeline to better meet business objectives.