Spark Group Work B

Uber & Lyft rides analysis

General Mainframe

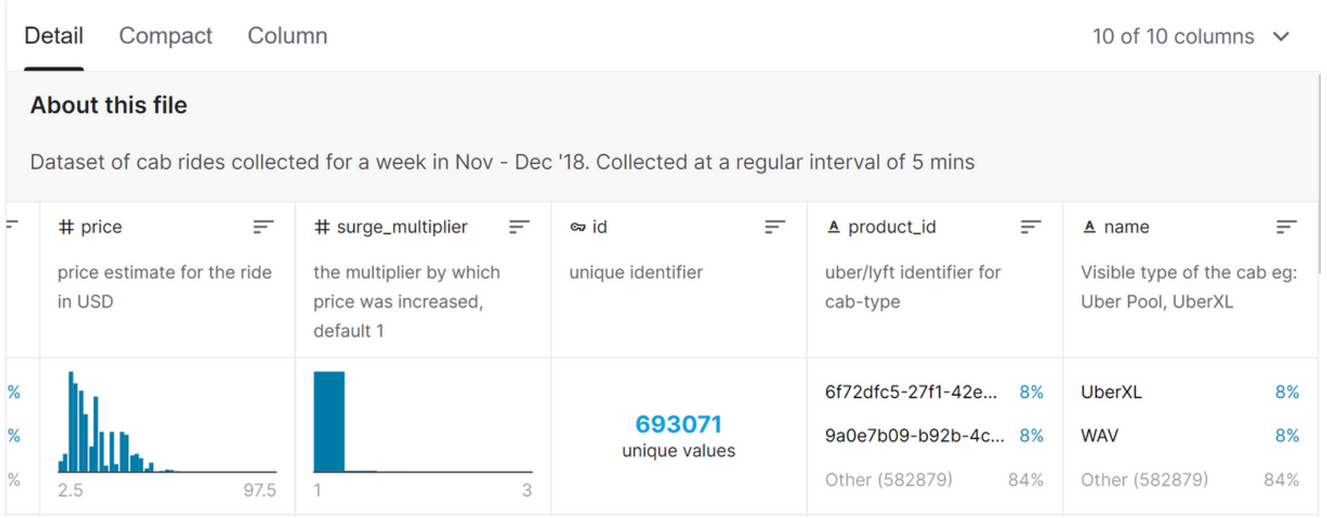
Both, Uber and Lyft are two of the biggest personal drivers companies, which tend to provide similar services and are easy to replace by their consumers. That is, in the sense of the sector that is being targeted. It is from here that we acquired a large dataset from kaggle referring to the main trips and details from a location in the US.

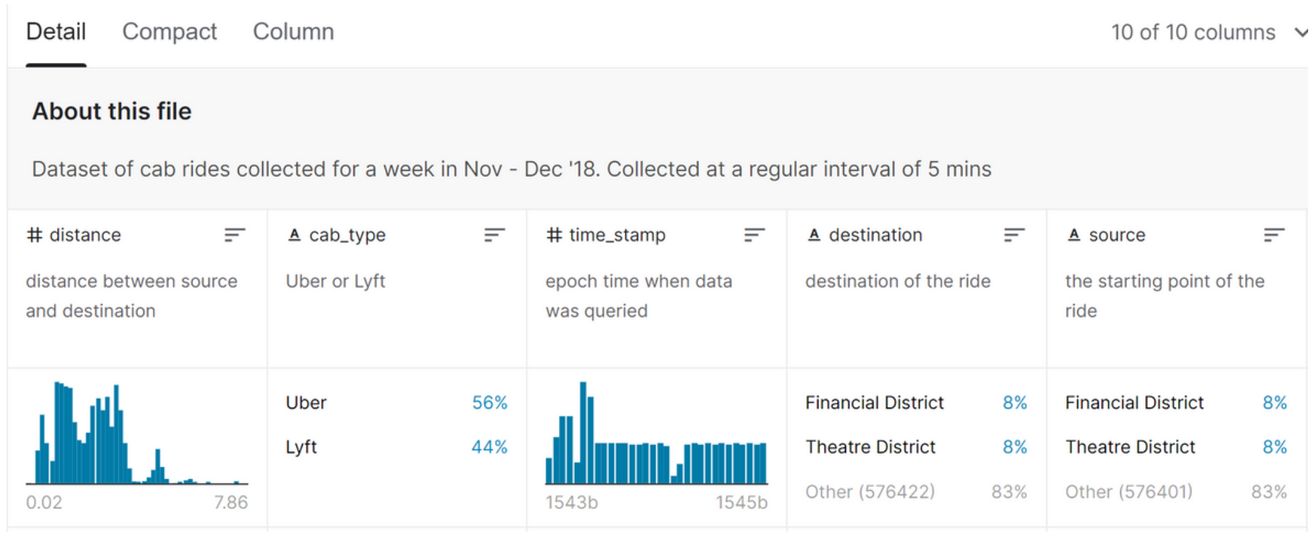
03

Description of the scenario

The dataset was obtained from https://www.kaggle.com/ravi72munde/uber-lyft-cab-prices, where the information is characterized by the following:

- A set with 693,071 rows, regarding the number of rides from one point to another considering the type of service, price, multiplier and other aspects.
- The information is from November to December 2018.
- It has 10 colums: Distance, Cab_type, time_stamp, destination, source, price, surge_multiplier, id, product_id and name.





Goal of the analysis

We want to analyze the rides done through these companies and look at the various services that they have to offer and how the variables of distance, surge, and time affects the pricing of the rides.

Business Questions to answer

Considering the dataset,we could solve the following questions:

- Which is the day of the month that has the most rides?
- Which is the most frequent destination and the least one?
- Which is the most frequent origin and the least one?
- Do people care about the multiplier effect? Hence, that means they have enough supply or they just prefer to wait until there is no multiplier?
- Which is the market share per each company?
- Which service do people prefer from each of the companies?
- Do distance has somthing to do with the type of cab they select?
- Which is the service that people use the most and tend to vary its prices? Which one maintains its range?

Analysis of the dataset

	distance	cab_type	time_stamp	destination	source	price	surge_multiplier	id	product_id	name
0	0.44	Lyft	1544952607890	North Station	Haymarket Square	5.0	1.0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	lyft_line	Shared
1	0.44	Lyft	1543284023677	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux
2	0.44	Lyft	1543366822198	North Station	Haymarket Square	7.0	1.0	981a3613-77af-4620-a42a-0c0866077d1e	lyft	Lyft
3	0.44	Lyft	1543553582749	North Station	Haymarket Square	26.0	1.0	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	lyft_luxsuv	Lux Black XL
4	0.44	Lyft	1543463360223	North Station	Haymarket Square	9.0	1.0	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	lyft_plus	Lyft XL
5	0.44	Lyft	1545071112138	North Station	Haymarket Square	16.5	1.0	f6f6d7e4-3e18-4922-a5f5-181cdd3fa6f2	lyft_lux	Lux Black
6	1.08	Lyft	1543208580200	Northeastern University	Back Bay	10.5	1.0	462816a3-820d-408b-8549-0b39e82f65ac	lyft_plus	Lyft XL
7	1.08	Lyft	1543780384677	Northeastern University	Back Bay	16.5	1.0	474d6376-bc59-4ec9-bf57-4e6d6faeb165	lyft_lux	Lux Black
8	1.08	Lyft	1543818482645	Northeastern University	Back Bay	3.0	1.0	4f9fee41-fde3-4767-bbf1-a00e108701fb	lyft_line	Shared
9	1.08	Lyft	1543315522249	Northeastern University	Back Bay	27.5	1.0	8612d909-98b8-4454-a093-30bd48de0cb3	lyft_luxsuv	Lux Black XL

Can be observed the following:

- There is a time_stamp column, which needs to be transformed in order to have better understanding of time issues.

- id column does not seem relevan as it is the identification code for the ride.

- product_id and name look similar, so we will see if we need to maintain both or if it is enough with one.

	distance	cab_type	destination	source	price
0	0.44	Lyft	North Station	Haymarket Square	5.0
1	0.44	Lyft	North Station	Haymarket Square	11.0
2	0.44	Lyft	North Station	Haymarket Square	7.0
3	0.44	Lyft	North Station	Haymarket Square	26.0
4	0.44	Lyft	North Station	Haymarket Square	9.0
5	0.44	Lyft	North Station	Haymarket Square	16.5
6	1.08	Lyft	Northeastern University	Back Bay	10.5
7	1.08	Lyft	Northeastern University	Back Bay	16.5
8	1.08	Lyft	Northeastern University	Back Bay	3.0
9	1.08	Lyft	Northeastern University	Back Bay	27.5

	surge_multiplier	id	product_id
0	1.0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	lyft_line
1	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier
2	1.0	981a3613-77af-4620-a42a-0c0866077d1e	lyft
3	1.0	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	lyft_luxsuv
4	1.0	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	lyft_plus
5	1.0	f6f6d7e4-3e18-4922-a5f5-181cdd3fa6f2	lyft_lux
6	1.0	462816a3-820d-408b-8549-0b39e82f65ac	lyft_plus
7	1.0	474d6376-bc59-4ec9-bf57-4e6d6faeb165	lyft_lux
8	1.0	4f9fee41-fde3-4767-bbf1-a00e108701fb	lyft_line
9	1.0	8612d909-98b8-4454-a093-30bd48de0cb3	lyft_luxsuv

	name	day	month	year	hour	time
0	Shared	16	12	2018	9	09:30:07.890000
1	Lux	27	11	2018	2	02:00:23.677000
2	Lyft	28	11	2018	1	01:00:22.198000
3	Lux Black XL	30	11	2018	4	04:53:02.749000
4	Lyft XL	29	11	2018	3	03:49:20.223000
5	Lux Black	17	12	2018	18	18:25:12.138000

Transformed dataset

The main process that has been done was changing the time_stamp into date_time and from there separate the day, month, year hour and time from being in one cell.

This transformation is important as we want to see the effect a particular day has on the demand or supply.

Recognition of Data Elements

So as it can be seen we identified the elements and also the potential columns for categorization. With that we can start building the respective tables for answering our business questions.

C. Data entities, metrics and dimensions

I've identified the following elements:

- **Entities:** Cab_type, destination, source, id, product_id, name
- **Metrics:** Distance, price
- **Dimensions:** surge_multiplier, destination, source, product_id, name, day, month, year, time

D. Column categorization

The following could be a potential column categorization:

- **Time related columns:** *Year*, *Month*, *day*, *time*
- **Related columns profiling:** *Destination*, *Source*, *name* and *distance*
- **Other columns for basic profiling:** *Surge_multiplier*, *product_id* and *Cab_Type*

Summary of columns Year, Month, Day and Hour:

```
+-------+------------------+------------------+--------------------+-----------------+
|summary|               day|             month|                year|             hour|
+-------+------------------+------------------+--------------------+-----------------+
|  count|            693071|            693071|              693071|           693071|
|   mean|17.794364502338144| 11.58668448109934|              2018.0|11.61913714467926|
| stddev| 9.982286013944805|0.4924288279625382|1.135905899981944...|6.948114156101857|
|    min|                 1|                11|                2018|                0|
|    25%|                13|                11|                2018|                6|
|    50%|                17|                12|                2018|               12|
|    75%|                28|                12|                2018|               18|
|    max|                30|                12|                2018|               23|
+-------+------------------+------------------+--------------------+-----------------+
```

Checking for nulls on columns Year, Month, Day and Time:

```
+---+-----+----+----+
|day|month|year|hour|
+---+-----+----+----+
|  0|    0|   0|   0|
+---+-----+----+----+
```

Checking amount of distinct values in columns Year, Month, Day and Hour:

```
+---+-----+----+----+
|day|month|year|hour|
+---+-----+----+----+
| 17|    2|   1|  24|
+---+-----+----+----+
```

Most and least frequent occurrences for Day and Hour:

leastFreqDay	mostFreqDay	leastFreqHour	mostFreqHour
9 (1674 occurrences)	27 (76121 occurrences)	8 (24275 occurrences)	0 (32413 occurrences)

Time related questions

Answering to the most and least frequent days it can be stated the following:

-The least frequent day is the 9, or the beginning of the month.
- While the most frequent day is the 27. In this case we need to consider that the dataset if from Nov and Dec; hence, it makes sense that holidays are the ones with more rides.

Also the most frequent hour is at the middle of the night, which makes sense on holidays; and the least frequent hour is at 8am.

Related Columns Profiling

Summary of columns destination, source, name and distance:

```
+-------+-----------+--------+------+------------------+
|summary|destination|  source|  name|          distance|
+-------+-----------+--------+------+------------------+
|  count|     693071|  693071|693071|            693071|
|   mean|       null|    null|  null|2.1894297553925237|
| stddev|       null|    null|  null|1.1389369868597294|
|    min|   Back Bay|Back Bay| Black|              0.02|
|    25%|       null|    null|  null|              1.28|
|    50%|       null|    null|  null|              2.16|
|    75%|       null|    null|  null|              2.92|
|    max|   West End|West End|   WAV|              7.86|
+-------+-----------+--------+------+------------------+
```

Checking for nulls on columns destination, source, name and distance:

```
+-----------+------+----+--------+
|destination|source|name|distance|
+-----------+------+----+--------+
|          0|     0|   0|       0|
+-----------+------+----+--------+
```

Checking amount of distinct values in columns destination, source, name and distance:

```
+-----------+------+----+--------+
|destination|source|name|distance|
+-----------+------+----+--------+
|         12|    12|  13|     549|
+-----------+------+----+--------+
```

Most and least frequent occurrences for destination, source, name and distance:

leastFreqDestination	mostFreqDestination	leastFreqSource	mostFreqSource
North Station (57119 occurrences)	Financial District (58851 occurrences)	North Station (57118 occurrences)	Financial District (58857 occurrences)

leastFreqName	mostFreqName	leastFreqDistance	mostFreqDistance
Shared (51233 occurrences)	Black SUV (55096 occurrences)	5.95 (6 occurrences)	2.66 (9174 occurrences)

- The least frequent destination and origin is North Station, while the Financial District is the most frequent origin and destination. This gives us the idea that several people use the rides for working issues, but not early in the morning.

- The least frequent service is Lyft Shared; meanwhile the one that is mostly use is Uber Black SUV, which reaffirms the idea of working trips and commodity.

- Now, usually the trips are done to short distances, less than 3 miles, which states that the public are interested in arriving fast into their destination. The least frequent distance is 5.95 miles.

-'

Summary of columns Surge_multiplier, Price, Product_id and Cab_type:

```
+-------+------------------+-----------------+--------------------+--------+
|summary|   surge_multiplier|            price|          product_id|cab_type|
+-------+------------------+-----------------+--------------------+--------+
|  count|            693071|           637976|              693071|  693071|
|   mean|1.0138697911180816|16.54512549061407|                null|    null|
| stddev|0.09164126209924166|9.324358581411659|                null|    null|
|    min|               1.0|              2.5|55c66225-fbe7-4fd...|    Lyft|
|    25%|               1.0|              9.0|                null|    null|
|    50%|               1.0|             13.5|                null|    null|
|    75%|               1.0|             22.5|                null|    null|
|    max|               3.0|             97.5|        lyft_premier|    Uber|
+-------+------------------+-----------------+--------------------+--------+
```

Checking for nulls on columns Surge_multiplier, Price, Product_id and Cab_type:

```
+---------------+-----+----------+--------+
|surge_multiplier|price|product_id|cab_type|
+---------------+-----+----------+--------+
|              0|55095|         0|       0|
+---------------+-----+----------+--------+
```

Checking amount of distinct values in columns Surge_multiplier, Price, Product_id and Cab_type:

```
+---------------+-----+----------+--------+
|surge_multiplier|price|product_id|cab_type|
+---------------+-----+----------+--------+
|              7|  147|        13|       2|
+---------------+-----+----------+--------+
```

Other Columns for basic profiling

In this case we can observe that most of the rides were with 1 multiplier but also something interesting appeared. There are 55095 values in the column price that are empty. This might be a mistake or not, but all of them correspond to Uber's product called Taxi. Hence, the payments were made in cash or Uber just charged a commision to the taxi.

Market Ratio of each service

We did a dataframe for obtaining the market share regarding each service. As it is observable, the tend to be pretty similar, so consumer is not only looking for a particular service, just to be able to movilize.

Cab Category	NumRides	RoundedRatio
Black SUV	55096	7.95
UberXL	55096	7.95
WAV	55096	7.95
Taxi	55095	7.95
Black	55095	7.95
UberX	55094	7.95
UberPool	55091	7.95
Lux	51235	7.39
Lyft XL	51235	7.39
Lux Black XL	51235	7.39
Lyft	51235	7.39
Lux Black	51235	7.39
Shared	51233	7.39

Companies market share and the influence of the multiplier

Cab Company	NumRides	RoundedRatio
Uber	385663	55.65
Lyft	307408	44.35

In the created dataframes it can be observed that Uber is the company with biggest market share.

Multiplier	NumRides	RoundedRatio
No Multiplier	672096	96.97
Greater than 1 or...	20809	3.0
Greater than 2 or...	166	0.02

Regarding the impact of the multiplier, almost 97% of the rides has no multiplier, which means to things:
- People prefer to wait until the multiplier is 1
- Or there is enough supply to cover the demand at key hours.

Preference for distance and prices

Most of the rides are short distances, having also a good amount in medium distances.

At the same time, most of the trips had a low price, which reiterates the part of not having a multiplier. For the purpose of this analysis, Low price is up to USD 30.

Distance Range	NumRides	RoundedRatio
Short Distance	325634	46.98
Medium Distance	318071	45.89
Long Distance	49366	7.12

Price Range	NumRides	RoundedRatio
Low Price	581170	83.85
Medium Price	56216	8.11
High Price	590	0.09

Cab Preference according to Origin:

source	Cab Category	NumberofRides	TotalRides	Cab Preference
Theatre District	Black	4612	57813	7.98
Theatre District	UberX	4612	57813	7.98
Theatre District	Black SUV	4612	57813	7.98
Theatre District	UberXL	4612	57813	7.98
Theatre District	Taxi	4612	57813	7.98
Theatre District	WAV	4612	57813	7.98
Theatre District	UberPool	4611	57813	7.98
West End	UberPool	4582	57562	7.96
West End	Black SUV	4582	57562	7.96
West End	WAV	4582	57562	7.96
West End	UberXL	4582	57562	7.96
West End	Black	4582	57562	7.96
West End	Taxi	4582	57562	7.96
West End	UberX	4582	57562	7.96
Northeastern Univ...	UberPool	4591	57756	7.95
Northeastern Univ...	UberXL	4592	57756	7.95
Northeastern Univ...	Black	4592	57756	7.95
Northeastern Univ...	Black SUV	4592	57756	7.95
Northeastern Univ...	UberX	4591	57756	7.95
Northeastern Univ...	Taxi	4592	57756	7.95

Is the place of origin significant for the type of cab you choose?

Accordin to the analysis made through Spark, the origin does not present certain significance for choosing a type of cab. The Cab preferences are relatively similar.

Distance Preference:

name	Distance Range	NumberofRides	TotalRides	Distance Preference
Black SUV	Short Distance	26205	55096	47.56
Taxi	Short Distance	26204	55095	47.56
UberX	Short Distance	26204	55094	47.56
UberXL	Short Distance	26205	55096	47.56
UberPool	Short Distance	26203	55091	47.56
Black	Short Distance	26204	55095	47.56
WAV	Short Distance	26205	55096	47.56
Lux	Medium Distance	23910	51235	46.67
Lyft XL	Medium Distance	23910	51235	46.67
Lux Black XL	Medium Distance	23910	51235	46.67
Lyft	Medium Distance	23910	51235	46.67
Shared	Medium Distance	23910	51233	46.67
Lux Black	Medium Distance	23910	51235	46.67
Lux	Short Distance	23701	51235	46.26
Lyft XL	Short Distance	23701	51235	46.26
Lux Black XL	Short Distance	23701	51235	46.26
Lyft	Short Distance	23701	51235	46.26
Shared	Short Distance	23699	51233	46.26
Lux Black	Short Distance	23701	51235	46.26
Black SUV	Medium Distance	24945	55096	45.28

Which is the effect of the distance when selecting a type of cab?

As it is visible, most of the rides are for short distances. And curiously, the Black SUV is present in the different types of distances.

Price variation according to cab:

name	Price Range	NumberofRides	TotalRides	Price Preference
Shared	Low Price	51233	51233	100.0
UberPool	Low Price	55089	55091	100.0
Lyft	Low Price	51223	51235	99.98
UberX	Low Price	55084	55094	99.98
WAV	Low Price	55086	55096	99.98
Lyft XL	Low Price	50827	51235	99.2
UberXL	Low Price	54423	55096	98.78
Lux	Low Price	50366	51235	98.3
Black	Low Price	52640	55095	95.54
Lux Black	Low Price	45869	51235	89.53
Black SUV	Low Price	32382	55096	58.77
Lux Black XL	Low Price	26948	51235	52.6
Lux Black XL	Medium Price	23799	51235	46.45
Black SUV	Medium Price	22695	55096	41.19
Lux Black	Medium Price	5290	51235	10.32
Black	Medium Price	2452	55095	4.45
Lux	Medium Price	869	51235	1.7
UberXL	Medium Price	671	55096	1.22
Lux Black XL	High Price	488	51235	0.95
Lyft XL	Medium Price	406	51235	0.79
Lux Black	High Price	76	51235	0.15
Black SUV	High Price	19	55096	0.03
Lyft	Medium Price	12	51235	0.02
UberX	Medium Price	10	55094	0.02
WAV	Medium Price	10	55096	0.02
Black	High Price	3	55095	0.01
Lyft XL	High Price	2	51235	0.0
UberXL	High Price	2	55096	0.0
UberPool	Medium Price	2	55091	0.0

Does the price changes too much considering the type of cab?

So in this case the only cab type that is taken with the same price is Shared from Lyft. Nevertheless, the Black SUV from Uber is present with low, medium and high prices, making it the major type of cab used.

Spark ML Prediction

Finally we tried to develop a Spark ML Prediction model, using random forests and Multiclassification Evaluator and got the following result:

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

evaluator = MulticlassClassificationEvaluator(labelCol='Cab Type',\
                                              predictionCol='prediction',\
                                              metricName='accuracy')
accuracy = evaluator.evaluate(predictions)
print('The accuracy of our model is ', accuracy)
```

```
[Stage 199:============================>                            (1 + 1) / 2]
```

So it is evident that the model lacks of accuracy, demonstrating what we saw regarding the similar preferences for the types of cab. This also means that we need other variables to determine which type of cab the consumer would select. The whole process can be seen in the code.

Conclusions

So some general conclusions can be the following:
- Uber Black SUV service is the predominant one between the consumers; hence they should focus more on providing that type of service.
- In the case of Lyft, its Shared cab type is the one that maintains the most the low price; however, it is not the most used. So price is not the main issue for the consumers.
- We have seen that they prefer comfort and short distances, which means moving fast from one place to another.
- Uber lacks of information regarding his service Taxi, where price is not shown but do also present a several quantity of rides. Could be interesting to make the comparison between it with Shared from Lyft.
- In general, cab types market share are pretty similar, which means that costumer just want to be transported quickly. However, there are certain cab types that can be promoted within different price ranges, such as the Black SUV.

UBER

lyft