SPARK Individual Assignment

Handed in by
Eugen Wettstein

25.09.2020

Table of Contents

1. INTRODUCTION....1

2. GOAL OF THE ANALYSIS....2

3. ANALYSIS DEEP DIVE....3

3.1 INFO AND PRE-PROCESS OF THE DATA IN DATAIKU....3
3.2 READING OF THE DATA INTO SPARK....3
3.3 PROFILING OF DIFFERENT CATEGORIES....4
3.4 ANSWERING OF SOME BUSINESS-RELATED QUESTIONS....6
3.5 CONCLUSION....9

1. Introduction

The biggest car transportation providers in the US are Uber and Lyft. Uber has in the US **67% market share** and Lyft being the second largest owns **30% market share.** Both are offering the same services with different labels for their categories.

Uber categorizes his cars in **5 categories** with one having a sub-category:

UberX: UberX is the standard entry level for all Uber rideshare services

UberPool: UberPool is normally covered by UberX cars and is the possibility to share the ride with unknown individuals that are going in the same direction

UberWAV: UberWAV are wheelchair-accessible vehicles with qualified drivers

UberXL: UberXL is Uber's large passenger-carrying service, for taking five to six passengers per ride. The income is higher since the ride rate is higher, but only large groups or families will order this option

BlackSUV: These are Uber's full luxury SUV fleet for taking five to six passengers

UberBlack: These are Uber's full luxury vehicles and cost most to the request

Lyft categorizes his cars as well in **5 categories** with a sub-category**:**

Lyft: Lyft or also called Original Lyft provides rides in regular cars with up to four passengers

Shared: Same concept as the one from Uber and normally covered by Original Lyft

Lyft XL: Lyft's large passenger-carrying service for up to 6 passengers

Lux: Lux provides premium black car service in high-end vehicles

Lux Black: Is Lyft's premium black car service that is only limited to the most luxurious makes and models

Black XL: Black XL or also Lux Black XL provides rides in premium black SUV for up to six passengers

Given the fact that Uber was founded in **2007** and Lyft **2012.** We are considering and analyzing data of the year **2018** and gathered in **Boston, USA.**

2. Goal of the Analysis

A new competitor is trying to enter the market in **Boston,USA** and wants to know if through data analytics they can be supported in their decision or not.

Data Analytics is not only the fuel which allows many other technologies to function, it also enables the industry to have greater transparency over business and portfolio performance, thus making more informed, less risky, and ultimately more profitable decisions.

Given the market share of both providers one could expect that **more or less every third person** looking for a ride should be ordering it via Lyft and the rest should be done through Uber. The ratio of provided rides we can expect should be therefore **2:1 for Uber**.

Data from **November and December of 2018** from both providers will be analyzed and compared. We will be considering parameters related to **time, rides and type of cars** that were used on certain trajectories.

Based on the results we get from the analysis, the decision can be simplified if there is any chance in entering the market in Boston, USA or not.

Since Uber is longer in the market and therefore better known by customers, we expect them to **provide the double amount** of rides than Lyft. Therefore trying to enter as new player in the market would not be beneficial.

3. Analysis Deep Dive

3.1 Info and pre-process of the data in Dataiku

For this analysis we were looking at a CSV file called ***cab_rides.csv*** provided by Kaggle.

The dataset contains different rides in Boston,USA and consists of **637976 rows and 10 columns.** Each row is an individual ride and the columns have information about each ride.

These are the columns provided in the dataset with their meaning:

Distance:	Distance between source and destination
Cab_type:	Categorical variable with values Uber and Lyft
Timestamp:	The moment where the data was queried in timestamp format
Source:	Starting point
Destination:	Endpoint
Price:	Estimated price for thr ride in USD
Surge_multiplier:	The multiplier by which the price was multiplied, default is 1
ID:	Unique identifier
Product_id:	Uber/Lyft identifier for cab_type
Name:	Visible tag of the cab

As a first step the data was analized with Dataiku. The source contained also data about regular taxi rides and those were removed (only 75 values out of 600k+ values).

The timestamp column was transformed into a date with format **yyyy-mm-dd HH:MM:SS.**

Once all necessary transformations were made, the data was read into SPARK.

3.2 Reading of the data into SPARK

Once the data is processed by SPARK the following elements are identified

Entities: Rides(main one which is measured – facts)

Metrics: Timestamp

Dimension: Distance, cab_type, destination, price, surge_multiplier, name and id

and as followed categorized

Timing related column: Timestamp

Drive related columns: distance, cab_type, source, destination, price and surge_multiplier

Company car related columns: id, product_id and name

3.3 Profiling of different categories

3.3.1 Timing related columns

Summary of column time_stamp:

summary	year	month	day	hour
count	637976	637976	637976	637976
mean	2018.0	11.589251006307448	17.762665053230844	11.51422310557137
stddev	1.135600535413938...	0.4919701589039917	10.002298741735492	6.9593165878176775
min	2018	11	1	0
25%	2018	11	13	5
50%	2018	12	17	12
75%	2018	12	28	17
max	2018	12	30	23

Having a first summary overview of this column we notice it has **one year, two month values, all the days in a month and all the hours in a day**.

Checking amount of distinct values in column time_stamp:

year	month	day	hour
1 occurrences	2 occurrences	16 occurrences	24 occurrences

By checking the distinct values in those categories, we notice there are only **like half of the days** of a normal month and the rest are what we expected and the column also doesn't contain any null values.

3.3.2 Drive related columns

Summary of columns distance, cab_type, source, destination, price and surge_multiplier:

summary	distance	cab_type	source	destination	price	surge_multiplier
count	637976	637976	637976	637976	637976	637976
mean	2.189261100730507	null	null	null	16.54512549061407	1.0150675730748493
stddev	1.1354130181861846	null	null	null	9.324358581411598	0.09542184282423667
min	0.02	Lyft	Back Bay	Back Bay	2.5	1.0
25%	1.27	null	null	null	9.0	1.0
50%	2.16	null	null	null	13.5	1.0
75%	2.93	null	null	null	22.5	1.0
max	7.86	Uber	West End	West End	97.5	3.0

We can see that **cab_type, source and destination are categorical** variables and **distance, price and surge_multiplier numerical.**

Checking amount of distinct values in columns distance, cab_type, source, destination, price and surge_multiplier:

```
+--------+--------+------+-----------+-----+----------------+
|distance|cab_type|source|destination|price|surge_multiplier|
+--------+--------+------+-----------+-----+----------------+
|     549|       2|    12|         12|  147|               7|
+--------+--------+------+-----------+-----+----------------+
```

Looking at the distinct values in each column we can see we have **two cab_types (Uber/Lyft) and 12 destinations and routes**. Same case as before there aren't any null values in these columns.

Checking frequency of dibstinct values in column source, destination and surge_multiplier:

leastFreqSource	mostFreqSource	leastFreqDestination	mostFreqDestination	leastFreqSurge_multiplier	mostFreqSurge_multiplier
North Station (52576 occurrences)	Financial District (54197 occurrences)	North Station (52577 occurrences)	Financial District (54192 occurrences)	3.0 (12 occurrences)	1.0 (617001 occurrences)

Also something we can observe is that the Financial District is the most frequent source and destination and the North Station the least frequent source and destination.

3.3.3 Company car related columns

Summary of columns id, product_id and name:

```
+-------+--------------------+--------------------+------+
|summary|                  id|          product_id|  name|
+-------+--------------------+--------------------+------+
|  count|              637976|              637976|637976|
|   mean|                null|                null|  null|
| stddev|                null|                null|  null|
|    min|00005b8c-5647-410...|55c66225-fbe7-4fd...| Black|
|    25%|                null|                null|  null|
|    50%|                null|                null|  null|
|    75%|                null|                null|  null|
|    max|ffffecd1-49b1-498...|        lyft_premier|   WAV|
+-------+--------------------+--------------------+------+
```

Looking at the summary we can see that all of the variables are categorical and that product_id might have some wrong inputs. Therefore in the further analysis we won't use it but also because it contains the same information as the name column.

Checking amount of distinct values in columns id, product_id and name:

```
+------+----------+----+
|    id|product_id|name|
+------+----------+----+
|637976|        12|  12|
+------+----------+----+
```

Checking thee distinct values in the three columns we get the expected 12 values since each provider has 6 distinct categories and again as in the other columns before we don't have any null values.

3.4 Answering of some business-related questions

3.4.1 Ratio of rides throughout the day

To get statistics of the rides throughout the day we will modify slightly the data by first dividing the data for each provider and then binning the timestamp column. The new created column **Time_of_day** is divides as follows:

Night rides: From 12am – 6am

Morning rides: From 6am – 12pm

Afternoon rides: From 12pm – 6pm

Evening rides: From 6pm – 12am

After the analysis of the data we see the following statistics for both providers and also splited by months

Uber's rides for the months november and december:

month	Time_of_day	rides	avg(price)	avg(surge_multiplier)	Ratio
11	Afternoon_ride	38802	15.739832998299057	1.0	28.552296575372704
11	Evening_ride	36951	15.877702903845634	1.0	27.190245625395516
11	Night_ride	33140	15.795368135184068	1.0	24.385936511206936
11	Morning_ride	27005	15.82234771338641	1.0	19.871521288024844

month	Time_of_day	rides	avg(price)	avg(surge_multiplier)	Ratio
12	Night_ride	51771	15.797193409437716	1.0	26.594236400061643
12	Morning_ride	49908	15.782810371082792	1.0	25.637232239173986
12	Afternoon_ride	48397	15.76082195177387	1.0	24.86104689988185
12	Evening_ride	44594	15.808371081311387	1.0	22.907484460882518

Lyft's rides for the months november and december:

month	Time_of_day	rides	avg(price)	avg(surge_multiplier)	Ratio
11	Afternoon_ride	35914	17.344922592860723	1.0321740825304895	28.46928260007927
11	Evening_ride	35020	17.346330668189605	1.0310322672758423	27.760602457391993
11	Night_ride	30260	17.350132187706542	1.0314441506939855	23.987316686484345
11	Morning_ride	24956	17.22771277448309	1.0312049206603622	19.782798256044394

month	Time_of_day	rides	avg(price)	avg(surge_multiplier)	Ratio
12	Night_ride	48344	17.355377089194107	1.031084519278504	26.671374504849442
12	Morning_ride	46131	17.361774078168693	1.0300015174177886	25.450462876121332
12	Afternoon_ride	44643	17.396557131017182	1.0320766973545685	24.629533593000033
12	Evening_ride	42140	17.371506407214046	1.031359753203607	23.248629026029196

There are several things we can observe from the above outcome. First we can see that both providers have the exactly same order for the **Time_of_day** column. In the month November they had slightly more afternoon rides and in December they had slightly more night rides.

This makes sense if we consider that in November there might be more tourists that are visiting the city in the afternoons and in December are a lot of Christmas dinners where people need to get home and might not be in a condition to drive by themselves. If we look at the ratio we can see that they are pretty similar for both providers and for both months.

The surprising fact is that although the rides of Lyft are slightly more expensive than the ones from Uber, **they have almost the same number of rides** they provided.

3.4.2 Number of rides per hour

Next we will have a look at the different hours where rides were provided. The tables will be seperatly showed and for each month

Amount of rides provided by Uber for each hour and for the months november and december:

month	hour	cab_type	rides
11	1	Uber	6936
11	23	Uber	6764
11	15	Uber	6602
11	12	Uber	6582
11	19	Uber	6536
11	17	Uber	6491
11	11	Uber	6476
11	0	Uber	6421
11	16	Uber	6412
11	14	Uber	6372
11	13	Uber	6343
11	18	Uber	6290
11	22	Uber	6013
11	10	Uber	5790
11	21	Uber	5685
11	20	Uber	5663
11	2	Uber	5659
11	5	Uber	4778
11	3	Uber	4674
11	4	Uber	4672
11	7	Uber	4203
11	8	Uber	3944
11	9	Uber	3607
11	6	Uber	2985

month	hour	cab_type	rides
12	3	Uber	9101
12	7	Uber	9009
12	0	Uber	8930
12	6	Uber	8855
12	5	Uber	8845
12	4	Uber	8634
12	1	Uber	8279
12	17	Uber	8132
12	11	Uber	8117
12	14	Uber	8102
12	15	Uber	8068
12	16	Uber	8062
12	13	Uber	8041
12	9	Uber	8030
12	19	Uber	8018
12	12	Uber	7992
12	2	Uber	7982
12	8	Uber	7962
12	10	Uber	7935
12	18	Uber	7932
12	20	Uber	7286
12	21	Uber	7220
12	23	Uber	7082
12	22	Uber	7056

Amount of rides provided by Lyft for each hour and for the months november and december:

month	hour	cab_type	rides
11	1	Lyft	6609
11	23	Lyft	6550
11	18	Lyft	6234
11	13	Lyft	6127
11	14	Lyft	6110
11	16	Lyft	6041
11	11	Lyft	5966
11	17	Lyft	5950
11	0	Lyft	5948
11	19	Lyft	5927
11	12	Lyft	5844
11	15	Lyft	5842
11	22	Lyft	5736
11	20	Lyft	5393
11	10	Lyft	5344
11	21	Lyft	5180
11	2	Lyft	4983
11	5	Lyft	4248
11	4	Lyft	4242
11	3	Lyft	4230
11	7	Lyft	4020
11	8	Lyft	3647
11	9	Lyft	3203
11	6	Lyft	2776

month	hour	cab_type	rides
12	6	Lyft	8379
12	3	Lyft	8318
12	5	Lyft	8254
12	0	Lyft	8056
12	1	Lyft	8048
12	4	Lyft	7982
12	7	Lyft	7915
12	2	Lyft	7686
12	18	Lyft	7619
12	10	Lyft	7604
12	12	Lyft	7528
12	13	Lyft	7506
12	9	Lyft	7497
12	19	Lyft	7477
12	15	Lyft	7464
12	17	Lyft	7399
12	14	Lyft	7393
12	8	Lyft	7377
12	11	Lyft	7359
12	16	Lyft	7353
12	20	Lyft	7068
12	22	Lyft	6744
12	23	Lyft	6697
12	21	Lyft	6535

Again by comparing them on a **more micrometric level we see again that the times are very similar when people order rides** and with again almost same amount of rides.

3.4.3 Most frequent routes per server

We now look at the different routes for which the people order a cab. For this analysis we also binned the distances the following way:

Short_distance: Everything below 2.5 miles

Medium_distance: Between 2.5 and 5 miles

Long_distance: Everything above 5 miles

For the 12 different sources and destinations these where the most frequent routes with their ranges.

Lyft's top destinations in november and december:

month	cab_type	Range	source	destination	rides
11	Lyft	Short_distance	North End	Beacon Hill	1918
11	Lyft	Medium_distance	Northeastern Univ...	West End	1916
11	Lyft	Short_distance	Financial District	South Station	1893
11	Lyft	Short_distance	Beacon Hill	North End	1869
11	Lyft	Short_distance	South Station	Financial District	1867

only showing top 5 rows

month	cab_type	Range	source	destination	rides
12	Lyft	Short_distance	South Station	Financial District	2759
12	Lyft	Short_distance	Financial District	South Station	2733
12	Lyft	Medium_distance	Fenway	West End	2702
12	Lyft	Short_distance	Haymarket Square	Financial District	2687
12	Lyft	Short_distance	Financial District	Haymarket Square	2676

only showing top 5 rows

Ubers's top destinations in november and december:

```
+-----+--------+--------------+------------------+------------------+-----+
|month|cab_type|         Range|            source|       destination|rides|
+-----+--------+--------------+------------------+------------------+-----+
|   11|    Uber|Short_distance|     South Station|  Theatre District| 2121|
|   11|    Uber|Short_distance|     South Station|Financial District| 2004|
|   11|    Uber|Short_distance|   Haymarket Square|Financial District| 2003|
|   11|    Uber|Short_distance|        Beacon Hill|         North End| 2001|
|   11|    Uber|Short_distance|Financial District|     South Station| 2001|
+-----+--------+--------------+------------------+------------------+-----+
only showing top 5 rows

+-----+--------+---------------+------------------+------------------+-----+
|month|cab_type|          Range|            source|       destination|rides|
+-----+--------+---------------+------------------+------------------+-----+
|   12|    Uber| Short_distance|Financial District|     South Station| 2907|
|   12|    Uber| Short_distance|     South Station|Financial District| 2904|
|   12|    Uber| Short_distance|Financial District|  Haymarket Square| 2904|
|   12|    Uber| Short_distance|          West End|     South Station| 2873|
|   12|    Uber|Medium_distance|     North Station|            Fenway| 2860|
+-----+--------+---------------+------------------+------------------+-----+
only showing top 5 rows
```

Looking at it we see that Uber and Lyft mostly where used for small distances so more in the center of Boston. We can also see that for Lyft in November the most frequent routes were **South Station-Financial District and Beacon Hill-North End** and for December **South Station-Financial District and Haymarket Square-Financial District.** For Uber the most frequent routes in November are as well **South Station-Financial District** and for December again **South Station-Financial District.**

Seeing this we could say that the **Financial District and the South Station is mostly dominated by Uber** and that **Beacon Hill and the North End is mostly dominated by Lyft**. Both providers will benefit from each other when there are no rides available in the moment where people look for a ride.

3.5 Conclusion

Overall, we can say, although Uber has more than double of market share in the USA, in Boston they have only a few more rides than Lyft and **the numbers are very similar**. A market entry could be possible since the people seem to be very open for new competitors.
Each provider has his customer bases in certain districts in the city but there are also districts where passengers use both providers equally. Putting a bigger focus with advertisements on areas like the **Theatre District, Haymarket Square or Fenway** could provide advantages in the market.