G2M insight for Cab Investment.

19 JUINE

Data Glacier

Créé par : Nada Belaidi

Data Glacier

Summary:

I. General introduction:

XYZ is a private firm in US. Due to remarkable growth in the Cab Industry in last few years and multiple key players in the market, it is planning for an investment in Cab industry. This work's objective is providing meaningful insights through analyzing the data provided in order to help XYZ choose the right company for their investment.

II. Data presentation and exploring:

1-Data presentation:

Four datasets have been provided for the analysis containing information about:

Cab_Data.csv – this file includes details of transaction for 2 cab companies

Customer_ID.csv – this is a mapping table that contains a unique identifier which links the customer's demographic details

Transaction_ID.csv – this is a mapping table that contains transaction to customer mapping and payment mode

City.csv – this file contains list of US cities, their population and number of cab users

Time period of data is from 31/01/2016 to 31/12/2018.

2-Data exploring:

	Cab_Data.csv	**Customer_ID.csv**	**Transaction_ID.csv**	**City.csv**
Shape	(359392, 7)	(49171, 4)	(440098, 3)	(20, 3)
Zeroes	0	0	0	0
Missing values	0	0	0	0

Columns	Transaction ID Date of Trave Company City KM Travelled Price Charged Cost of Trip	Customer ID Gender Age Income (USD/Month)	Transaction ID Customer ID Payment Mode	City Population Users

III. Data merging and creating a global dataset:

In order to analyze each company's customers, we will be merging the customers dataset, transactions dataset and Cab dataset.Transactions.csv has Customer_ID and Transaction_ID in common with the other two.

We will be leaving City.csv behind for now, since it doesn't seem to have any relevant information.

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mode	Gender	Age	Income (USD/Month)
0	10000011	42377	Pink Cab	ATLANTA GA	30.45	370.95	313.6350	29290	Card	Male	28	10813
1	10351127	43302	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Cash	Male	28	10813
2	10412921	43427	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Card	Male	28	10813
3	10000012	42375	Pink Cab	ATLANTA GA	28.62	358.52	334.8540	27703	Card	Male	27	9237
4	10320494	43211	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Card	Male	27	9237

IV. Data preparation:

1-Decoding the "Date of Travel" column:

Maximum value of this column is: 42371
Minimum value is: 43465
The difference is: 1094
1091=3 years
42371 is equivalent to 31/01/2016
43465 is equivalent to 31/12/2018

We will be applying this lambda expression on the whole column in order to get the number of days it's pointing on:

```
.apply(lambda x: 43465-x["Date of Travel"],axis=1)
```

The result looks like this:

```
Entrée [43]: finald["Date of Travel"]

Out[43]: 0         1088
         1          163
         2           38
         3         1090
         4          254
                   ...
         359387     358
         359388     362
         359389     361
         359390     359
         359391     361
         Name: Date of Travel, Length: 359392, dtype: int64
```

Later on, we will be counting the exact date, the result looks like this:

```
Entrée [45]: finald["Date of Travel"]

Out[45]: 0        2019-01-23
         1        2016-07-12
         2        2016-03-09
         3        2019-01-25
         4        2016-10-11
                     ...
         359387   2017-01-23
         359388   2017-01-27
         359389   2017-01-26
         359390   2017-01-24
         359391   2017-01-26
         Name: Date of Travel, Length: 359392, dtype: datetime64[ns]
```

2-Splitting data by company:

In order the study each company customer base and compare them we will be splitting our merged data set:

Pink Cab:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mo
0	10000011	2019-01-23	Pink Cab	ATLANTA GA	30.45	370.95	313.635	29290	Ca
3	10000012	2019-01-25	Pink Cab	ATLANTA GA	28.62	358.52	334.854	27703	Ca
6	10395626	2016-04-05	Pink Cab	ATLANTA GA	13.39	167.03	141.934	27703	Ca
7	10000013	2019-01-29	Pink Cab	ATLANTA GA	9.04	125.20	97.632	28712	Ca
11	10000014	2019-01-24	Pink Cab	ATLANTA GA	33.17	377.40	351.602	28020	Ca

Yellow Cab:

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip	Customer ID	Payment_Mo
1	10351127	2016-07-12	Yellow Cab	ATLANTA GA	26.19	598.70	317.4228	29290	Ca
2	10412921	2016-03-09	Yellow Cab	ATLANTA GA	42.55	792.05	597.4020	29290	Ca
4	10320494	2016-10-11	Yellow Cab	ATLANTA GA	36.38	721.10	467.1192	27703	Ca
5	10324737	2016-09-28	Yellow Cab	ATLANTA GA	6.18	138.40	87.5088	27703	Ca
8	10079404	2018-05-11	Yellow Cab	ATLANTA GA	39.60	704.30	494.2080	28712	Ca

V. Data analysis:

Using the matplotlib library, we will be visualizing and observing the different information and how are they related to the others.

We chose the following features to work on:

-Profit analysis.

-Travel frequency analysis.

-Reach rate analysis.

-Customers base analysis.

The analysis resulted in the following insights:

Profit analysis: Yellow cab makes 9 times more money than Pink cab each month and each year. It could be because its fees are higher than pink cab's fees and that it has a big customer base making it's travel frequency also high.

Travels frequency analysis: Yellow cab is on demand all the time around the year. It has higher ride frequency compared to Pink Cab.

Reach rate analysis: Yellow cab has higher customer reach in 25 cities while Pink cab has higher customer reach in only 2 cities with very low travel frequency.

Customers base analysis: Yellow cab has more customers in every age range than Pink Cab. Even the 65+ years old customers.

VI. Conclusion:

Based on the analysis performed on the data sets provided, Yellow Cab seems the most suitable for investment.