G2M Case Study

Sachin Subramanian

February 21, 2024

Background

- XYZ is a private equity 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.
- Objective: Provide actionable insights to help XYZ firm in identifying the right company for making investment.
- The analysis has been divided into four parts:
  - Data Understanding
  - Analysis of companies based on transactions, cities, and users
  - Determining the most profitable company
  - Recommendations for investments

Data Understanding

	City	Population	Users
0	NEW YORK NY	8405837	302149
1	CHICAGO IL	1955130	164468
2	LOS ANGELES CA	1595037	144132
3	MIAMI FL	1339155	17675
4	SILICON VALLEY	1177609	27247

	Transaction ID	Date of Travel	Company	City	KM Travelled	Price Charged	Cost of Trip
0	10000011	1/8/2016	Pink Cab	ATLANTA GA	30.45	370.95	313.635
1	10000012	1/8/2016	Pink Cab	ATLANTA GA	28.62	358.52	334.854
2	10000013	1/8/2016	Pink Cab	ATLANTA GA	9.04	125.20	97.632
3	10000014	1/8/2016	Pink Cab	ATLANTA GA	33.17	377.40	351.602
4	10000015	1/8/2016	Pink Cab	ATLANTA GA	8.73	114.62	97.776

	Customer ID	Gender	Age	Income (USD/Month)	Transaction ID	Payment_Mode
0	29290	Male	28	10813	10000011	Card
1	29290	Male	28	10813	10351127	Cash
2	29290	Male	28	10813	10412921	Card
3	27703	Male	27	9237	10000012	Card
4	27703	Male	27	9237	10320494	Card

	Transaction ID	Customer ID	Payment_Mode
0	10000011	29290	Card
1	10000012	27703	Card
2	10000013	28712	Cash
3	10000014	28020	Cash
4	10000015	27182	Card

- The data is broken up into 4 different files: City.csv, TransactionID.csv, CustomerID.csv, CabData.csv, as displayed above
- City looks at proportion of users from each city, TransactionID looks like the types of transactions, CustomerID analyzes the ages and incomes of the customer, and CabData provides insights on the trips and cost by each company.
- Each 4 different datasets provide unique insights on the different companies.

Data Analysis Part 1: Proportion of users in each city

	City	Population	Users	User_Proportion
0	NEW YORK NY	8405837	302149	0.035945
1	CHICAGO IL	1955130	164468	0.084121
2	LOS ANGELES CA	1595037	144132	0.090363
3	MIAMI FL	1339155	17675	0.013199
4	SILICON VALLEY	1177609	27247	0.023138
5	ORANGE COUNTY	1030185	12994	0.012613
6	SAN DIEGO CA	959307	69995	0.072964
7	PHOENIX AZ	943999	6133	0.006497
8	DALLAS TX	942908	22157	0.023499
9	ATLANTA GA	814885	24701	0.030312
10	DENVER CO	754233	12421	0.016468
11	AUSTIN TX	698371	14978	0.021447
12	SEATTLE WA	671238	25063	0.037338
13	TUCSON AZ	631442	5712	0.009046
14	SAN FRANCISCO CA	629591	213609	0.339282
15	SACRAMENTO CA	545776	7044	0.012906
16	PITTSBURGH PA	542085	3643	0.006720
17	WASHINGTON DC	418859	127001	0.303207
18	NASHVILLE TN	327225	9270	0.028329
19	BOSTON MA	248968	80021	0.321411

	City	Z_Score	Significant
0	NEW YORK NY	−0.474499	False
1	CHICAGO IL	0.411880	False
2	LOS ANGELES CA	0.461995	False
3	MIAMI FL	−3.008916	True
4	SILICON VALLEY	−1.294141	False
5	ORANGE COUNTY	−3.193687	True
6	SAN DIEGO CA	0.304199	False
7	PHOENIX AZ	−7.109819	True
8	DALLAS TX	−1.258996	False
9	ATLANTA GA	−0.751044	False
10	DENVER CO	−2.217647	True
11	AUSTIN TX	−1.474286	False
12	SEATTLE WA	−0.418690	False
13	TUCSON AZ	−4.835052	True
14	SAN FRANCISCO CA	1.279807	False
15	SACRAMENTO CA	−3.099072	True
16	PITTSBURGH PA	−6.841407	True
17	WASHINGTON DC	1.187182	False
18	NASHVILLE TN	−0.874023	False
19	BOSTON MA	1.233450	False

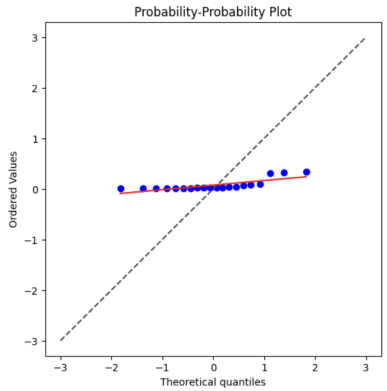

Figure: Utilizing the City.csv file, the user proportion within each city is calculated (number of users/total population). In the plot to the right, the data follows a normal distribution, so a z-score distribution for each of the samples is used. Based on the data, we find that certain cities, such as San Francisco, Washington DC, and Seattle, tend to have a higher proportion of users compared to the other cities.

Data Analysis Part 2: Types of transactions

- When referring to the Transaction.csv, we decided to test whether higher valuable customers (greater income) affects the transaction type (payment by card vs. by cash).
- Due to the large sample sizes, we are able to perform a t-test on the number of transactions by card vs. by cash.
- With the formula of the two-sample t-test given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

$\bar{X}_1, \bar{X}_2$ are the sample mean incomes of the card and cash transaction groups

s_p is the pool standard deviation,

n_1, n_2 are the sample sizes of the card and cash groups

We find our t-statistic to be 0.6872, which is less than the observed t-statistic for alpha level 0.95. Therefore, we conclude that there is no significant difference between the income levels of the customers paying by card vs. by cash.

Data Analysis Part 3: Age vs. Income

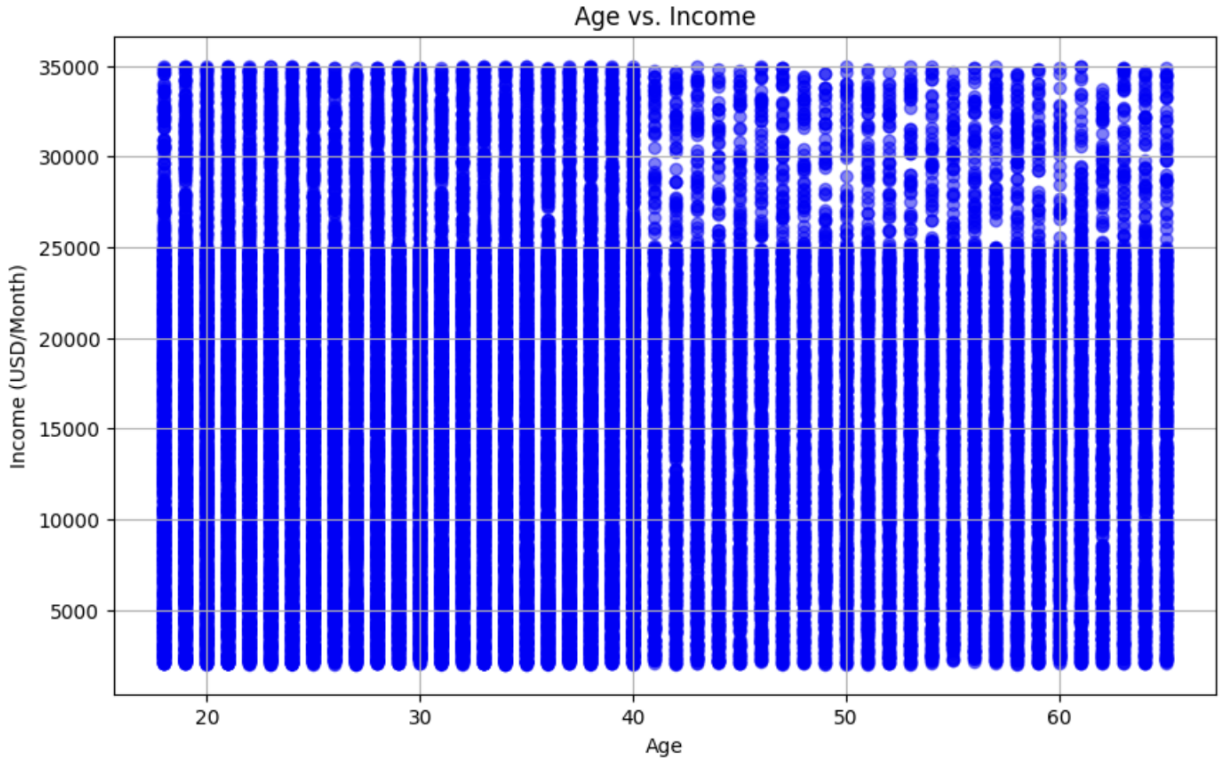

Figure: For the CustomerID.csv file, we analyzed if there's any association between the age and the transactional value (income) of customers. Above represents the plot of the ages of customers (X) and the income of the customers (Y). Incorporating the pearson correlation r (and the visual representation of the plot), we find that there is no significant correlation between the age and income of the customer.

Data Analysis Part 4: Seasonality

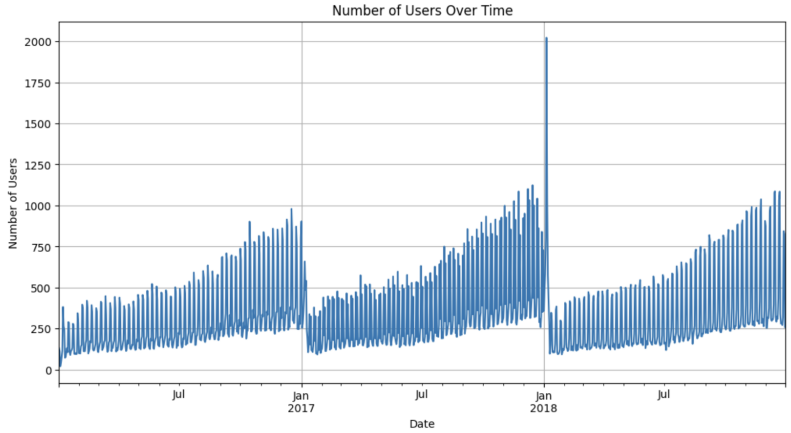

Figure: With the Cabdata.csv file, we first viewed a time series plot of the data, which helps measure seasonality of the data. Based on the plot above, we can conclude that during the early winter time (November, December, early January), there is a notable increase of users. This could be due to the weather and daylight savings time changes during this time; as the weather gets colder and daylight hours decrease, more and more people opt to obtain cab rides during this time.

Pink vs Yellow Cab companies: profit analysis

The profit made on each trip is determined by the sum of price charged minus the cost of the trip. Taking into account the CabData.csv file, we can graph the profits of both company hand in hand:

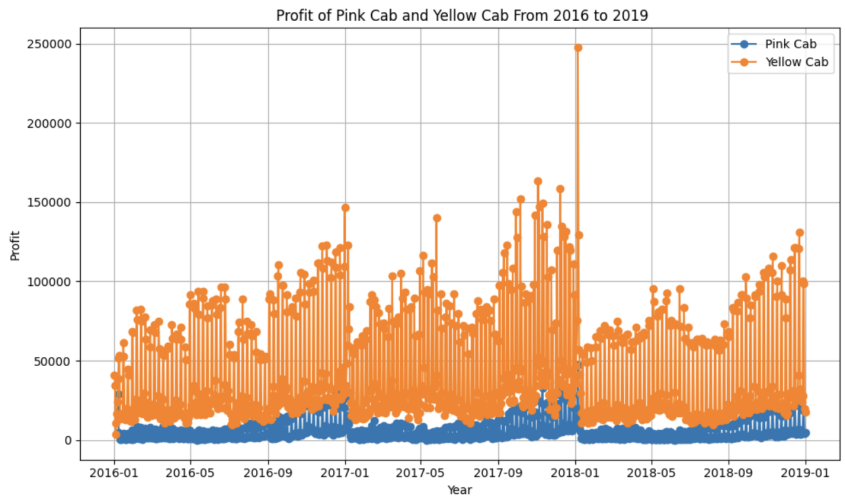

Figure: Based on the graph above, the yellow cab company seems to produce a higher profit on average compared to the pink cab company.

Which company is more profitable?

- Utilizing the graph from the previous slide, we conducted another t-test, with our null hypothesis being the pink and yellow cab companies having a similar profit and our alternative hypothesis being the yellow cab company having a greater profit than the pink cab company.
- Based on the t-test, we found our observed t-statistic to be greater than the t-statistic found at significance level 0.05,concluding the yellow cab company obtains a higher net profit

Recommendations for investments

- We believe XYZ should invest the yellow cab company due to the higher profits.
- It's important to attract more users by building hubs in cities with higher user proportions. Cities such as San Francisco, Washington DC, and Boston can serve as important centers where we can grow our user numbers, as well as the large cities like New York, Chicago, and Los Angeles.
- Having more rides available during the late fall and early winter time is key since there are more users during that time of year.
- Being able to accept multiple forms of payment can help increase profits and bring in more users. Having different sets of people, whether they are rich, poor, old or young, will be a win-win situation for the company and its customers: the company will obtain more profits while the users will be satisfied by the accessible ride services.