Abida S Bhatti

G2M Insight for Cab Investment Firm

EDA Notebook

XYZ wants to invest in a Cab company business and wants an insight into datasets provided for the yellow cab and pink cab companies.

For greater understanding of Datasets, Statistical data analysis can be performed on the dataset and the prediction target. Rigorous Data analysis include Descriptive Analysis, Correlational Analysis, and Contextual (time and agent based) Analysis. The goal of these analysis is to find the quality of features and their predictive power when contrasting with target value or label.

Descriptive analysis provides an understanding of characteristics of each attribute in a dataset. Correlation analysis allows us to analyze the relationship between two attributes and determine if they are correlated to each other. This can be done qualitatively and quantitatively. Calculation of descriptive statistics of dependent numerical or categorical attributes against each unique value of independent categorical attribute (to understand the relationship between X and Y) done in qualitative analysis. **Hypothesis-testing** framework is applied in **quantitative analysis** to determine relationship between X and Y. **Qualitative analysis** is the primary analysis to gain insight into problems and help to form hypothesis for quantitative research.

Contextual time-based analysis can be done on a dataset to determine the transaction recall and peak time of usage by customers. Contextual agent-based analysis can be done to analyze the unique Customer ID in the transaction. With ID attribute, histograms can be made to analyze average transaction per specific user in certain months.

Feature selection based on correlation analysis can be low correlation or high correlation. Feature selection based on contextual analysis can show visualizations of data across time and/or agent based contextual dimensions to help understand the context around the dataset.

Data Intake Report

Name: G2M Insight for Cab Investment Firm
Report date: 06/16/2021
Internship Batch: LISUM01
Version:1.0
Data intake by: Abida S Bhatti
Data intake reviewer: None
Data storage location: https://github.com/DataGlacier/DataSets

Tabular data details:

Cab_Data

Total number of observations	359393
Total number of files	1
Total number of features	7
Base format of the file	.csv
Size of the data	20.1 MB

City_Data

Total number of observations	21
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Customer_ ID

Total number of observations	49172
Total number of files	1
Total number of features	4
Base format of the file	.csv
Size of the data	1.00 MB

Transaction_ ID

Total number of observations	440099
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

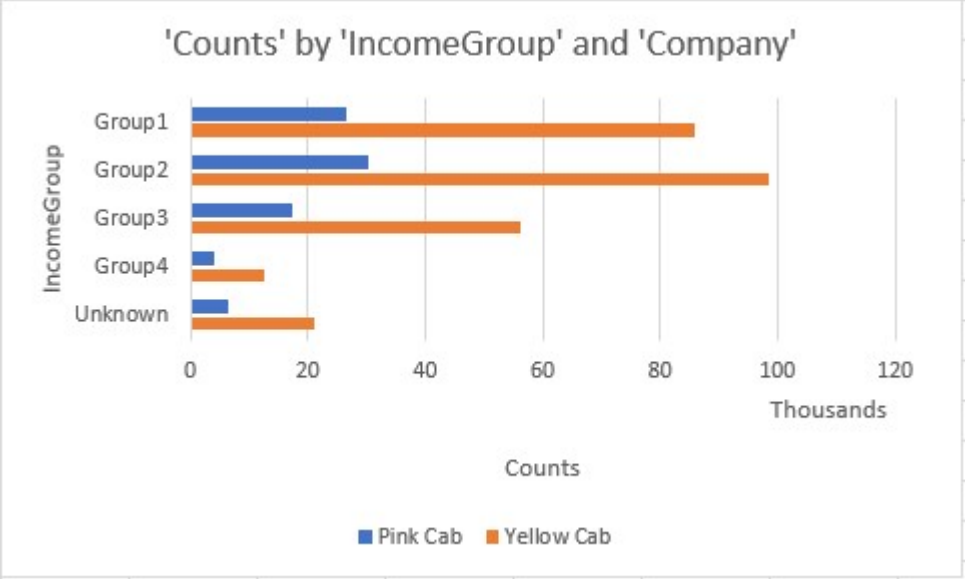
Exploratory Data Analysis (EDA):

EDA is done using the data in the intake report.

Income wise Analysis:

Range of Income analysis		
Group1		Between 1000 - 10000
Group2		Between 11000 - 20000
Group3		Between 21000 - 30000
Group4		Between 31000 - 40000
Group5		Unknown

Sum of Counts	Column Labels		
Row Labels	Pink Cab	Yellow Cab	Grand Total
Group1	26702	86008	112710
Group2	30162	98505	128667
Group3	17434	56287	73721
Group4	3938	12625	16563
Unknown	6475	21256	27731
Grand Total	**84711**	**274681**	**359392**



Hypotheses:

1. Is range of customers' income analyzed for the Yellow/Pink cab companies?
2. Do unknown income (Null values) group have any priority in riding Yellow/Pink cabs?

Outcome of Hypotheses:

1. I group the customer income into ranges of 5 groups: group 1 has 1000-10000, group 2 has 11000-20000, group 3 has 21000-30000, group 4 has 31000-40000, and group 5 has unknown income (null values). The data analysis shows all 5 groups have majority customers riding the yellow cabs.
2. The unknown group (null values) has priority of riding yellow over pink cabs.

Profit in City wise Analysis:

Hypotheses:

1. Which city shows highest profit in certain cab company business among other cities?
2. Which cab company is losing profit in big difference with the rival cab company?
3. Is there a possibility of any other big city taking the place of New York in cab company profit?

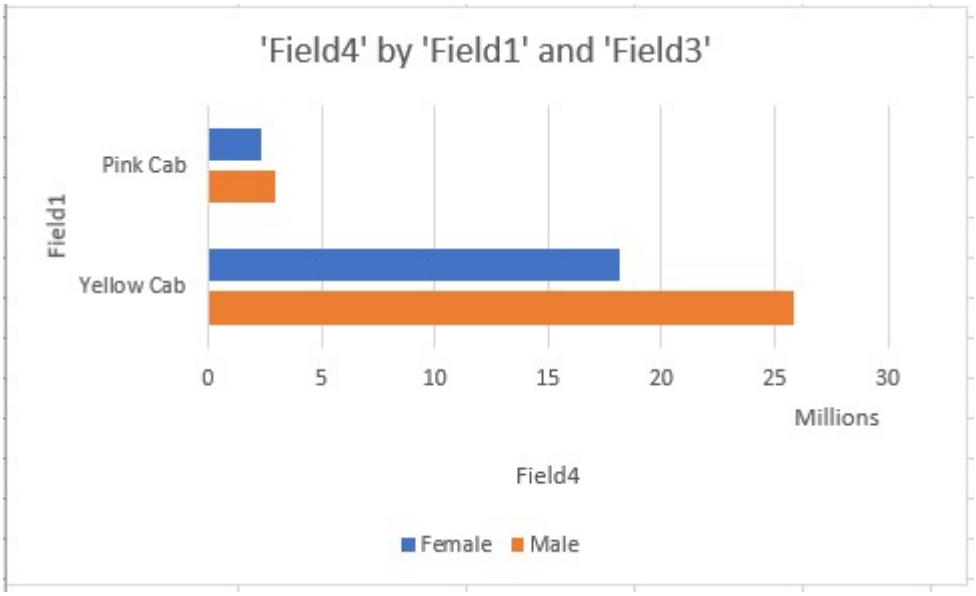
Outcome of Hypotheses:

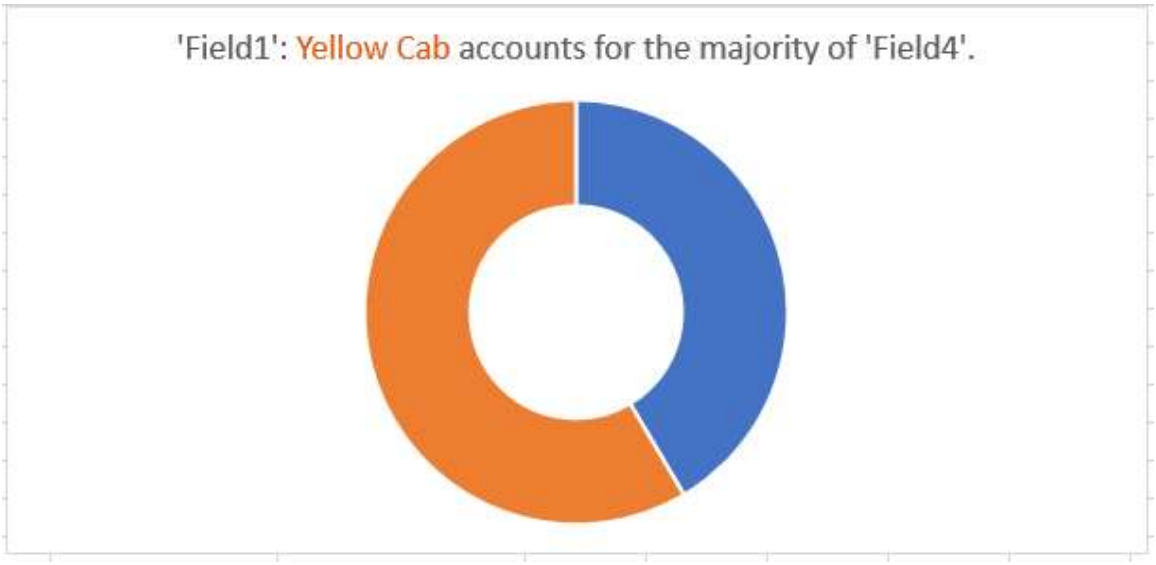
1. The data analysis clearly show New York has highest profit for the yellow cab company.
2. The pink cab is losing profit in a big difference with the yellow cab company.
3. The analyses show all other cities have less profit in cab company when compared with New York and hence cannot take the place of New York's yellow cab company profits.

Gender wise Analysis:

Sum of Field4	Column Labels		
Row Labels	Female	Male	Grand Total
Pink Cab	2330532.691	2976795.63	5307328.321
Yellow Cab	18131417.65	25888955.52	44020373.17
Grand Total	**20461950.34**	**28865751.15**	**49327701.49**

Row Labels	Sum of Field4
Female	20461950.34
Male	28865751.15
Grand Total	**49327701.49**





Hypothesis:

Is there a percentage difference for gender in yellow and pink cab companies?

Outcome Of Hypothesis:

The data analysis shows male to female ratio riding the cab is more for the yellow cab as compared to the pink cab. Therefore, yellow cab accounts for majority of the customers riding the cab.

Recommendations:

Data analysis done on customers' income shows majority of cab riders prefer yellow cabs.

The group with unknown income (null values) rides yellow cabs more than pink cabs.

The highest profit in cab company is for yellow cab in big city like New York. Other cities show more customers riding yellow cabs as compared to pink cabs.

When compared with New York, all other cities have less profit in cab company and cannot take the place of New York's yellow cab company profits.

The pink cab is losing profit in a big difference with the yellow cab company.

The analysis show gender difference in riding cabs as male to female ratio is more for riding yellow cabs as compared to pink cabs.

Based on all these recommendations, it is wise to **invest** in **yellow cab** company.