Nebula Space Organisation

CAB CASE STUDY

- By Divyaranjan Sahoo (21053386)

Highlighted Topics of the Presentation

It’s not a bug; it’s an undocumented feature. — Anonymous

1

Introduction

Introduction to the problem statement and business requirements for this case study

2

Approach

A short note on the approach to study and analyse the data set using python and it’s modules like pandas, matplotlib,etc.

3

Trends and studies

Understanding the data trends and identifying decision points to conclude the result.

4

Exploratory Data Analysis

Investigating considering hypothesis, engaging with the data, thinking critically, and using various analytical approaches to produce unique insights.

5

Conclusionary Note

Concluding the results , analytics of the dataset, insights of the case study and presenting the final decision

INTRODUCTION

Problem Description

XYZ, a private equity firm based in the United States, is strategically considering an investment in the rapidly expanding Cab Industry.

Given the significant growth witnessed in recent years and the presence of numerous key players in the market, XYZ is poised to explore opportunities within the cab sector for potential investment.

The Engagement

We have received several datasets containing information on two cab companies. Each dataset represents distinct facets of customer profiles.

Our objective is to derive actionable insights that will aid us in identifying the most suitable company for investment.

Assessing the datasets

Below are the list of datasets which are provided for the analysis with their properties and attributes

Transaction ID
Date of Trave
Company
City
KM Travelled
Price Charged
Cost of Trip

Cab_Data.csv

Transaction details for two cab companies.

Properties -

359393 Rows 7 Columns

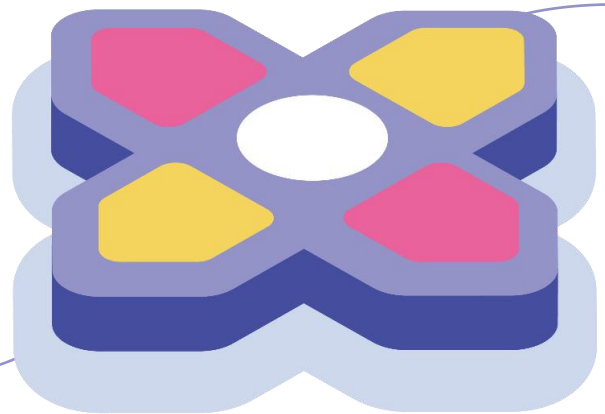
Transaction_ID.csv

Maps transactions to customers .

Properties -

440099 rows 3 columns

Transaction ID
Customer ID
Payment_Mode



Customer ID
Gender
Age
Income (USD/m)

Customer_ID.csv

Mapping table with a unique identifier linking customer demographic details.

Properties -

49172 Rows 4 Columns

City.csv

Lists US cities, their population and number of cab users

Properties -

21 Rows 3 Columns

City
Population
Users

Analysis Approach

Descriptive Statistics

Data is summarized to understand key characteristics like average fare, distance traveled, payment modes, and company-wise distributions.

Exploratory Analysis:

Explores and identifies relationships between variables correlation coefficients and statistical test at different time periods to know seasonal info.

Hypothesis Testing:

Forming and testing hypotheses using appropriate statistical tests to determine the strength of evidence and Quantify the significance of the results

The Developement Perspecive

To effectively analyze the relationship between cab company and customer preferences,analysis methods are deployed.

This combination of techniques, coupled with the use of data visualization libraries like Matplotlib and data manipulation tools like Pandas and statistical libraries like SciPy...

Provides a comprehensive understanding of the relationship, allowing you to identify key trends, test hypotheses, and extract actionable insights.

TRENDS

Studying the trends

Delving into the trends of cab company preferences reveals a fascinating interplay of factors that influence customer choices.

Data Description

Data description is the process of summarizing and organizing data to provide a clear and concise understanding of its key characteristics. It involves generating descriptive statistics, creating data visualizations, and identifying trends and patterns.

```
#Data description
merged_data.describe()
```

	Transaction ID	Customer ID	KM Travelled	Price Charged	Cost of Trip	Unnamed: 7	Age	Income (USD/Month)	Margin
count	3.593920e+05	359392.000000	359392.000000	359392.000000	359392.000000	0.0	359392.000000	359392.000000	359392.000000
mean	1.022076e+07	19191.652115	22.567254	423.443311	286.190113	NaN	35.336705	15048.822937	137.253198
std	1.268058e+05	21012.412463	12.233526	274.378911	157.993661	NaN	12.594234	7969.409482	160.311840
min	1.000001e+07	1.000000	1.900000	15.600000	19.000000	NaN	18.000000	2000.000000	-220.060000
25%	1.011081e+07	2705.000000	12.000000	206.437500	151.200000	NaN	25.000000	8424.000000	28.012000
50%	1.022104e+07	7459.000000	22.440000	386.360000	282.480000	NaN	33.000000	14685.000000	81.962000
75%	1.033094e+07	36078.000000	32.960000	583.660000	413.683200	NaN	42.000000	21035.000000	190.030000
max	1.044011e+07			2048.030000	691.200000	NaN	65.000000	35000.000000	1463.966000

Data Info

The info() method in Pandas is a descriptive tool that provides a quick overview of a DataFrame's structure and contents. It summarizes key information about the DataFrame, including missing values, descriptive statistics:, data types, etc.

```
# Dataframe information
merged_data.info();
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 359391
Data columns (total 15 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Transaction ID      359392 non-null  int64
 1   Customer ID         359392 non-null  int64
 2   Payment_Mode        359392 non-null  object
 3   Date of Travel      359392 non-null  datetime64[ns]
 4   Company             359392 non-null  object
 5   City                359392 non-null  object
 6   KM Travelled        359392 non-null  float64
 7   Price Charged       359392 non-null  float64
 8   Cost of Trip        359392 non-null  float64
 9   Unnamed: 7          0 non-null       float64
 10  Gender              359392 non-null  object
 11  Age                 359392 non-null  int64
 12  Income (USD/Month)  359392 non-null  int64
 13  Population          359392 non-null  object
 14  Users               359392 non-null  object
dtypes: datetime64[ns](1), float64(4), int64(4), object(6)
memory usage: 43.9+ MB
```

Trend Analysis Approach

Data Collection

Data is loaded into Google Colaboratory, panda is used with read_csv command.

Data Exploration

Helps understand the underlying patterns and relationships within data.

Data Imputation

The missing values in a dataset is filled using various techniques such as mean, median, mode.

Data Analysis

Helps gain valuable insights into customer preferences and assist in decision making.

HYPOTHESIS TESTING

Investigation

Hypothesis testing is a statistical method used to make inferences about a population based on sample data. Here I have done hypothesis testing by:

Formulating
a hypothesis

Connecting
it to data analysis

Emphasizing
the value of insights

Presumed Assumptions

Here, we'll procceed with our investigation on certain assumptions and get analysis results to encourage/foster our decision to invest.

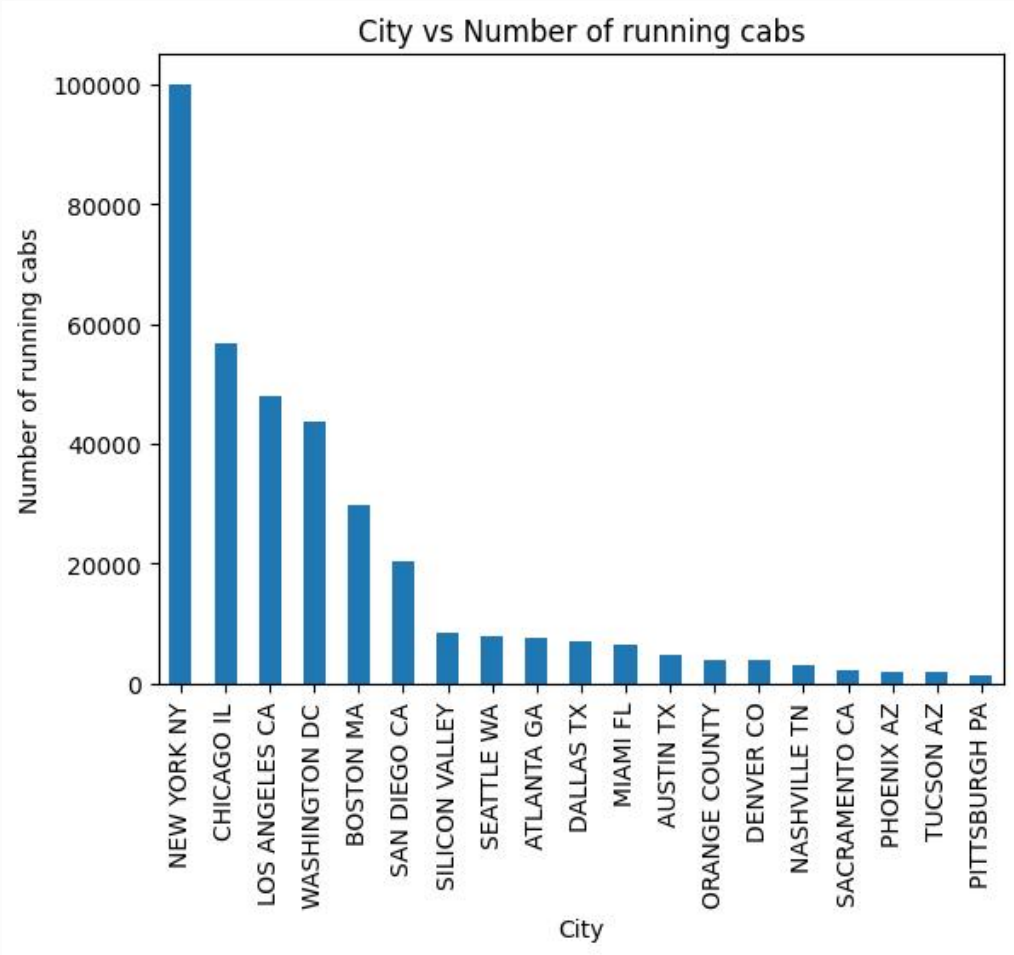
Hypotesis 1

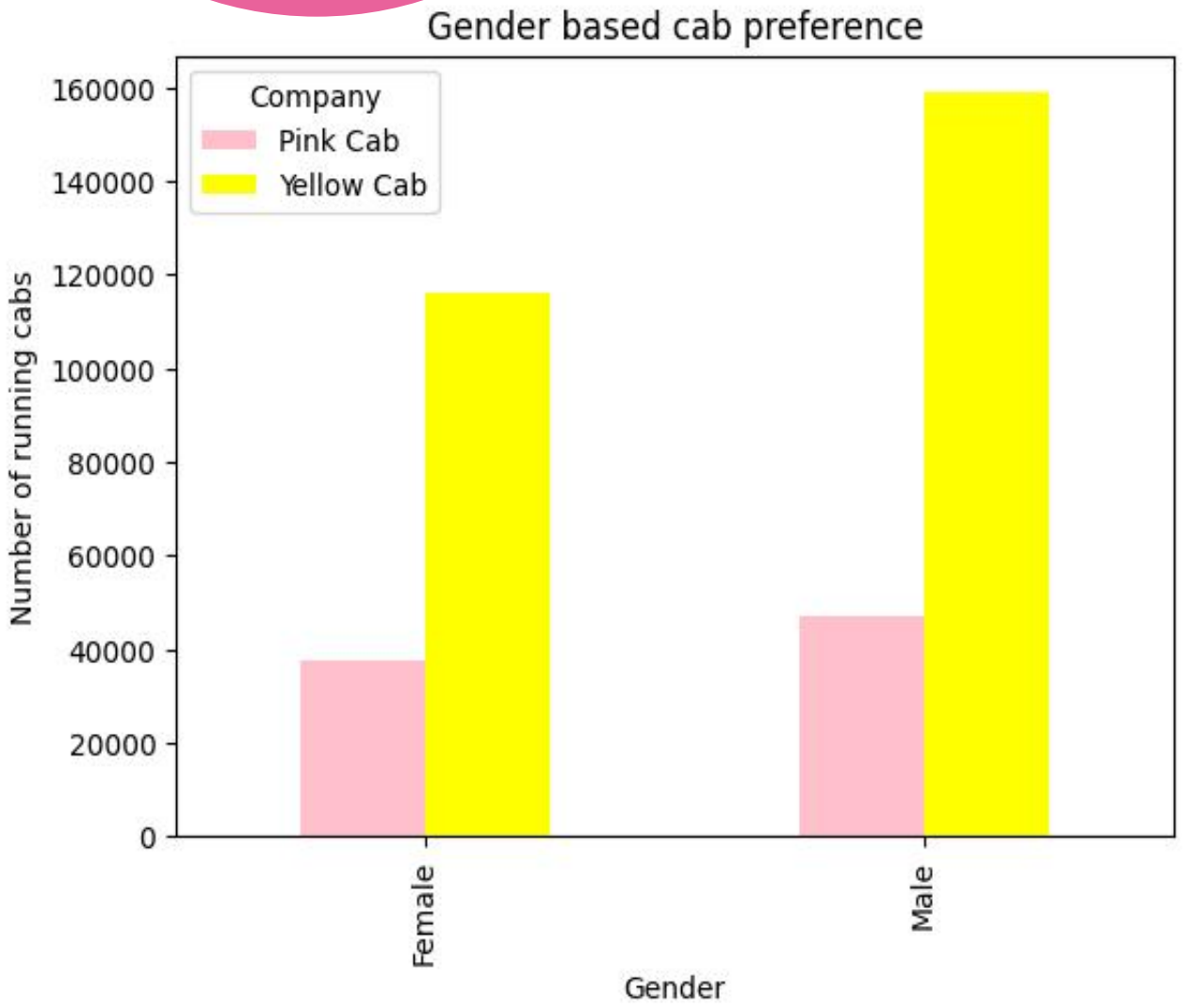
H0 ~ Los angeles is the city with maximum number of running cabs

Inference

The analysis results and plot shows New york is the city is the one with maximum number of running cabs.

Hypothesis failed to be true.





Hypotesis 1

H0 ~There's no gender based preferences for chosing cab

Inference

The analysis output shows that male are more consumers of cab in US than female.

Hypothesis failed to be true.

Nevertheless **yellow cab are more prefered** by consumer of each segment.

Hypotesis 3

H0 ~ People prefer yellow cabs more in winter (November (11) to January (1))

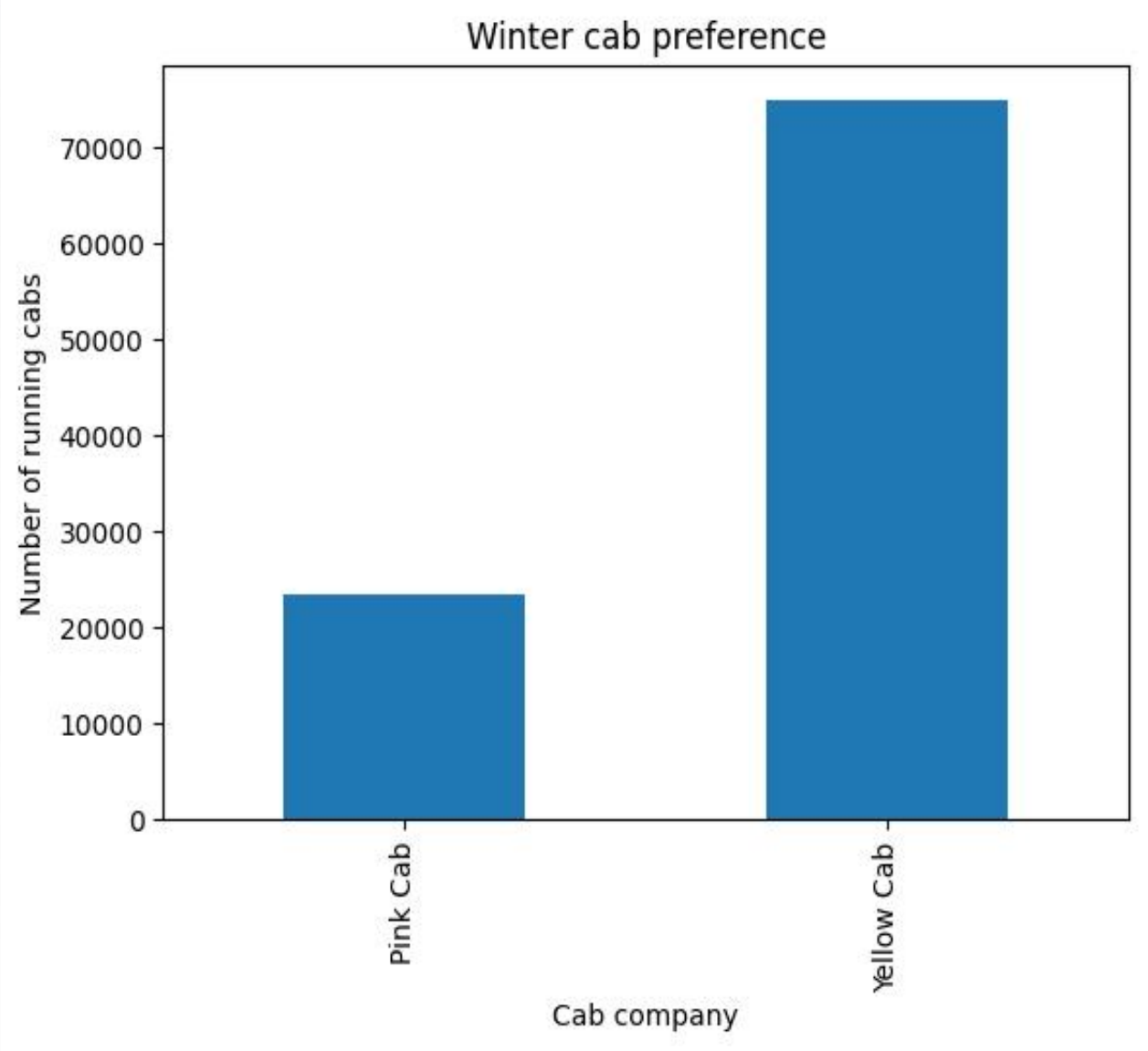
i.e showing seasonal dependency.

Inference

The analysis results and plot shows that **yellow cab is more prefered** in winter than pink.

Same also verifies in non winter seasons.

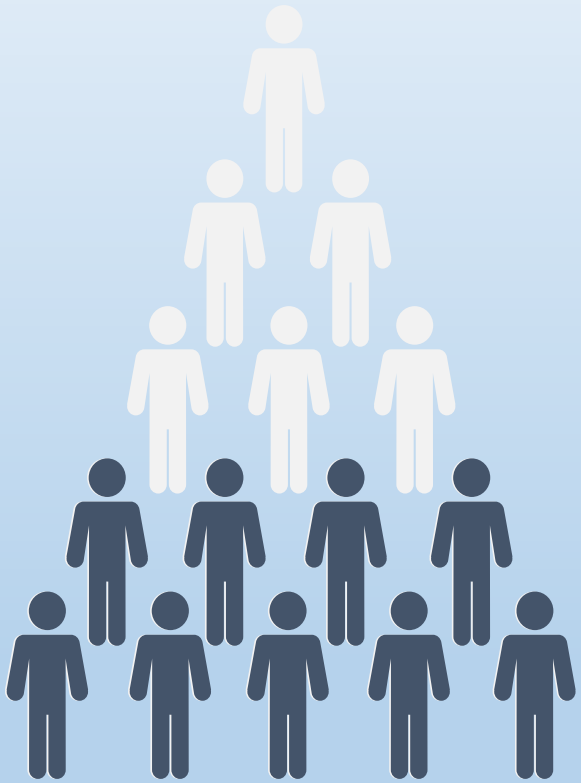
Hypothesis true.



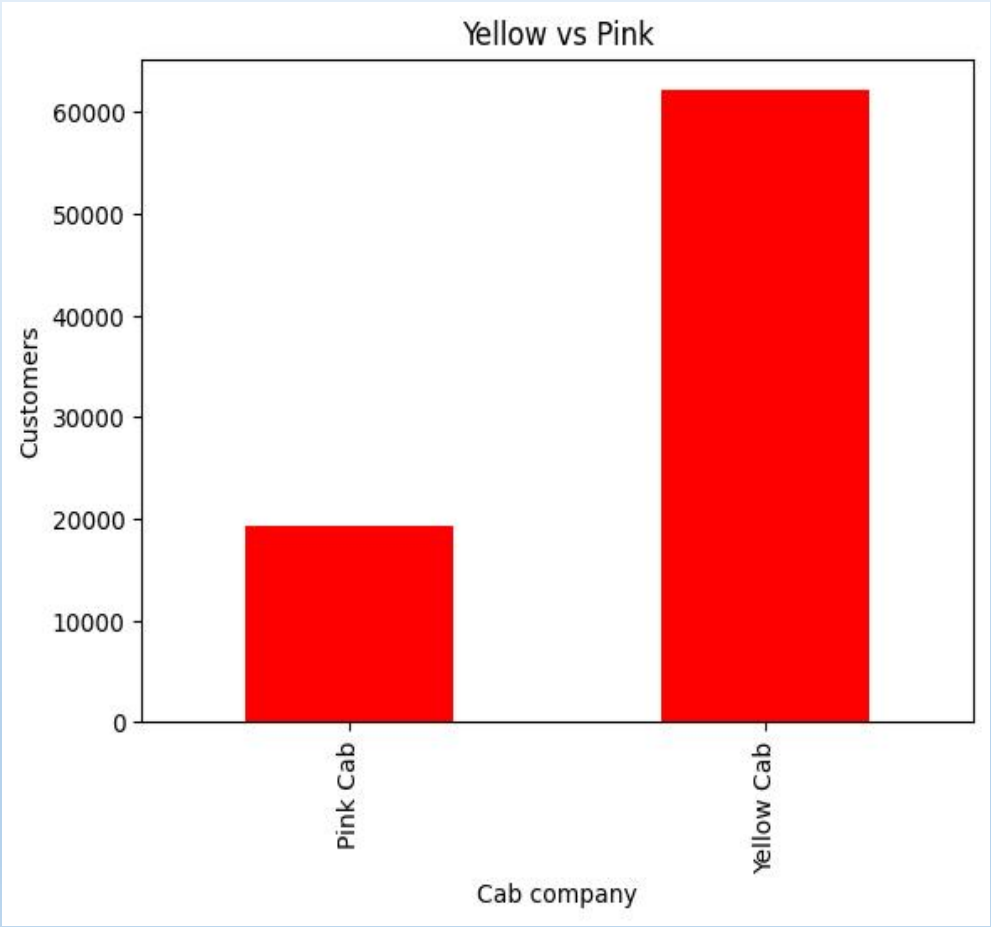
Yellow vs Pink Cab Company

The data from 2018 to 2020 gave the statistics about the yellow cab vs pink cab company as follows

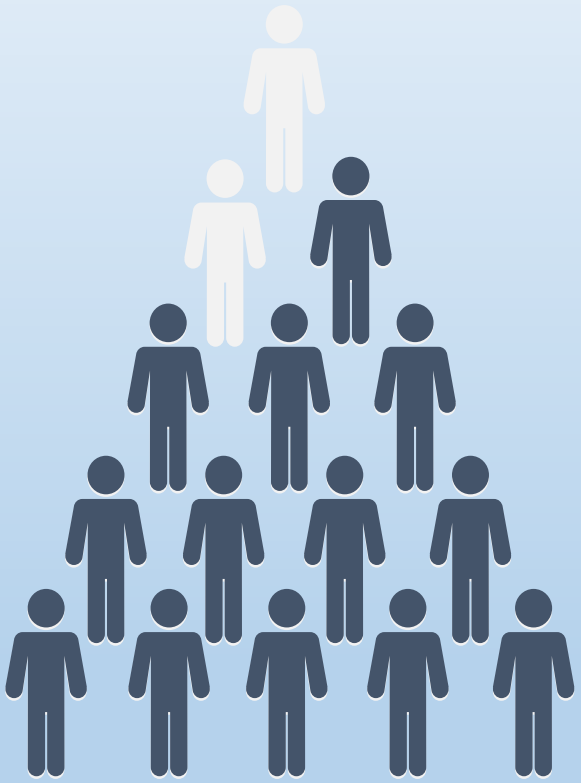
84711



Pink Cab



274681



Yellow Cab

Yellow cab company seems to have more customers than pink cab company, and hence this analytics inclines towards yellow cab.

Insights of the investigation

Lets throw lights on the insights extracted through the case study

In a nutshell

A thorough analysis of the data reveals Yellow Cab's dominance in terms of both ride count and user preference across a significant portion of the cities considered.

1. Yellow cab has higher customer preference in most cities while Pink cab has higher customer preference in few.
2. Age wise Reach: Most of the users are 20 to 30 years in age.Amongst them, most prefer yellow rides.
3. Yellow cab has maximum bookings in new york city (85918) and minimum in Pittsburg (631/)

The above facts and statistics are evident that yellow cabs are better investments than pink cabs.

XYZ should invest into yellow cab company

CONCLUSIONARY EPILOGUE

Epilogue

The conclusionary statement can be presented that the report included the insight analytics and cognitive investigation in the field of **Cab case study** ; modelling a transparent presentation for the lucid concepts and parallely demistyfing the convoluted anectdote associated with the classical approach to the cyberlife concerns.

Furthermore, the final decision can be presented that **YELLOW CAB** would be the best to invest upon in cab industry.

THANK YOU