**The Client**

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.

**The Engagement**

You have been provided with multiple data sets that contains information on 2 cab companies. Each file (data set) provided represents different aspects of the customer profile. XYZ is interested in using your actionable insights to help them identify the right company to make their investment.

The outcome of your engagement will be a **presentation to XYZ’s Executive team**. This presentation will be judged based on the visuals provided, the quality of your analysis and the value of your recommendations and insights.

**The Data**

You have been provided 4 individual data sets. **Time period of data is from 31/01/2018** to **31/12/2020.**

Below are the list of datasets which are provided for the analysis:

**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

You should fully investigate and understand each data set.

* Review the Source Documentation
* Understand the field names and data types
* Identify relationships across the files
* Field/feature transformations
* Determine which files should be joined versus which ones should be appended
* Aggregate the data into smaller files and define relationships
* Identify and remove duplicates
* Identify any biasness existing in the data

**Be Creative**

The idea is to create a hypothesis, engage with the data, think critically, and use various analytical approaches to produce unique insights.

You are not limited to only utilizing the data you have been provided.

We encourage you to find third party data sets which correspond to the overall theme and geographical properties of the data provided. **For Example:** you can leverage US weather data

Also, do research on overall cab industry in US and try to relate that with the trend in data

**Analysis**

**Create multiple hypothesis and investigate:**

**You will need to generate 5-7 hypothesis initially to investigate as some will not prove what you are expecting.**

**For Example:** *“Is there any seasonality in number of customers using the cab service?”*

Areas to investigate:

* Which company has maximum cab users at a particular time period?
* Does margin proportionally increase with increase in number of customers?
* What are the attributes of these customer segments?

**Although not required, we encourage you to document the process and findings**

* What is the business problem?
* What are the properties of the data provided?
* What steps did you take in order to create an applicable data set?
* What type of analysis is applicable to your hypothesis?
* How did you prepare and perform your analysis?
* What type of analysis did you perform?
* Why did you choose to use certain analytical techniques over others?
* What were the results?
* How do the results translate to support your hypothesis?

**Prepare a presentation that summarizes your analysis and recommendations and identify which company is performing better and is a better investment opportunity for XYZ**

**Remember, there are no wrong answers as long as the data supports them.**