**Project 1**

In this first project you will create a framework to scope out data science projects. This framework will provide you with a guide to develop a well-articulated problem statement and analysis plan that will be robust and reproducible.

**Read and evaluate the following problem statement:**

Determine which free-tier customers will convert to paying customers, using demographic data collected at signup (age, gender, location, and profession) and customer usage data (days since last log in, and activity score 1 = active user, 0= inactive user) based on Hooli data from Jan-Apr 2015.

DATASET: admissions.csv

**1. What is the outcome?**

Answer: The outcome of this assessment is to use the provided demographic data such as gender, location, age and profession to predict the outcome of paying customers

**2. What are the predictors/covariates?**

Answer: Predictors/covariates are Age, Gender, Location and Profession

**3. What timeframe is this data relevent for?**

Answer: Timeframe for this data from Jan’15 to April’15

**4. What is the hypothesis?**

Answer: Hypothesis for this data as below:

H0: There is no relationship between the paying customers and the demographic information such as age, gender, profession and location.

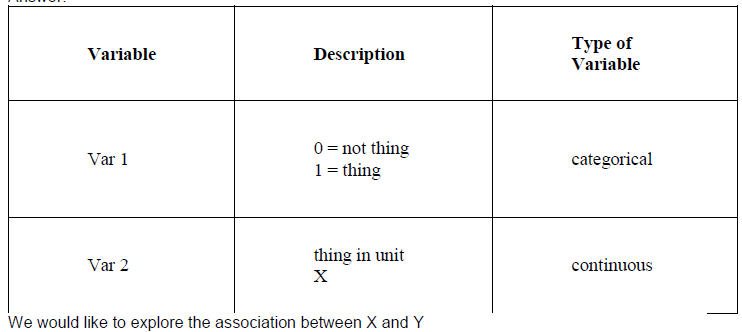
H1: There are relationship between the paying customers and the demographic information such as age, gender, profession and location.

**Let's get started with our dataset**

1. **Create a data dictionary**

Answer:

* Contain 2 variables in this dataset



**2. What is the outcome?**

Answer: Outcome is to predict the var 2 in the dataset

**3. What are the predictors/covariates?**

Answer: Predictor/covariates is var1

**4. What timeframe is this data relevent for?**

Answer: This data doesn’t contain and time frame information

**5. What is the hypothesis?**

Answer:

* H0: If Var1 is 0, meaning there is no association between variable 0 to variable 2
* H1: If Var1 is 1, meaning there is association between variable 1 to variable 2

**Problem Statement**

**Exploratory Analysis Plan**

Using the lab from a class as a guide, create an exploratory analysis plan.

**1. What are the goals of the exploratory analysis?**

Answer: The goals of the exploratory analysis also known as EDA (Exploratory Data Analysis) is to discovering patterns to foster hypothesis development and refinement. The hypothesis test is used in confirmatory data analysis. In all, the approach to analysing data sets to summarise their main characteristics with the support of visualization tools.

**2a. What are the assumptions of the distribution of data?**

Answer: The assumption of the distribution data should examine before departures from normality before the test are applied. For example, the outcome is well as long as the data is bell shape and the tail is not heavily skewed.

Another assumption of data distribution is via normal probability plot. This is a graphical method for assessing whether or not the data set is approximately normally distributed. For example, the data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Essentially, the straight line indicate the departures from normality.

**2b. How will determine the distribution of your data?**

Answer: As above, we can use either Histogram or normal probability plot (scatter plot) techniques where we can immediately see if the shape of the histogram resembles any of the widely known and used statistical distribution (Example: Gaussian), or use scatter plot method to see if the plots formed an approximate straight line which indicate the close to normality.

**3a. How might outliers impact your analysis?**

Answer: Depending on how far the outlier way from mean of the data. If the outliers stays around the data then it not so much impact the analysis. However, the outlier will have biggest effect on the data when there are significant different between the mean and outliers.

**3b. How will you test for outliers?**

Answer: Calculate the median and mode when have outliers. We can either use Histogram or normal probability plot (scatter plot) techniques to visualize if there’s any outliers.

**4a. What is colinearity?**

Answer: Colinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with substantial degree of accuracy.

**4b. How will you test for colinearity?**

Answer:

* For quantitative variable: Correlation is tested by Pearson correlation coefficient
* Association between nominal (Binominal or multinomial) can be tested via chi square
* Association between categorical and continuous variable and be tested by T-test (has 2 categories) or Annova( > 2 categories.

**5. What is your exploratory analysis plan?**

Using the above information, write an exploratory analysis plan that would allow you or a colleague to reproduce your analysis 1 year from now.

Answer: Exploratory Data Analysis Workflow

* Identify the problem
* Acquire the data
* Parse the data
* Mine the data
* Refine the data
* Build a data model
* Present the results

**Bonus Questions:**

1. Outline your analysis method for predicting your outcome

Answer:

* Chi-Square Test: There are an association between “admit” and “prestige” given the P values is smaller than the 0.05

2. Write an alternative problem statement for your dataset

Answer:

* 75% percentile of the data contain GPA > 3.5 which is highly skewed to high grade student.
* Limited input variables (only 4) for analysis.

3. Articulate the assumptions and risks of the alternative model

Answer: The data might not be sufficient for analysis due to data is highly skewed to high grade students and contain limited input variables for prediction.