## Capstone Project 1: Statistical Data Analysis

Think of the following questions and apply them to your dataset:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?

- Are there significant differences between subgroups in your data that may be relevant to your project aim?

- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

- What are the most appropriate tests to use to analyze these relationships?

Submission: Write a 1-2 page report on the steps and findings of your inferential statistical analysis. Upload this report to your GitHub and submit a link. Eventually, this report will get incorporated into your milestone report.

### Project goal:

The ecommerce company would like to predict life time value or revenue for customers acquired through the marketing organization. In the project, we are asked to predict 2-year customer life time value (24m LTV) based on the first purchase information. The prediction will help marketing orgnization to allocate marketing spend and create proper marketing strategies.

### Statistical DA: customer profiling and variable exploration

The initial analysis will involve visual analytics to determine the spread of data, variance, and measures of central tendency. This will also help determine if there are any data preparation tasks to be done (such as removing null values). Once this is completed, statistical tests can be done across the variables included in the data set. For variables with two samples, such as gender, whether or not a patient smokes, or whether or not they suffer from alcoholism, we can conduct an independent t-test. For variables that have more than one sample, such as age or wait time, we can conduct an ANOVA. This will allow us to determine if there are statistically significant differences between and among groups. Subsequently, this statistical analysis will also help determine which variables to focus on for predictive modelling.

1. Are marketing channels the same or different :

- chi-sq test on #of customer

- ANOVA test on mean of 24m LTV and first purchase

- test normality

2. Are device different?

- chi-sq test on #of customer

- ANOVA test on mean of 24m LTV and first purchase

- test normality

3. Is the shipping address type is a critial variable?

- Shipping address residential vs business - t test

- Logistic regression

4. First product purchased - ANOVA

- chi-sq test on #of customer

- ANOVA test on mean of 24m LTV and first purchase

- test normality

5. First purchase value and 24m LTV

6. Cluster analysis

7. RF variable importance analysis

8. Correlation

1. Are marketing channels the same or different:

- chi-sq test on #of customer

- ANOVA test on mean of 24m LTV and first purchase

- test normality

A screenshot of a cell phone

Description automatically generated

- chi-sq test on #of customer

Continency table

- ANOVA test on mean of 24m LTV and first purchase

A screenshot of a cell phone

Description automatically generated

- test normality