

## **DATS 6103 Project Proposal**

### **Team Member Names:**

(1)Anuradha Tidke

(3)Rakshith Reddy Eleti

(2)Priya Johny

(4)Venkata Gangadhar Naveen Palaka

### **Project Description:**

A UK-based online retail store has captured the sales data for different products for the period of one year (Nov 2016 to Dec 2017) and updated the dataset (in 2020). The organization sells gifts primarily on the online platform. The customers who make a purchase, consume directly for themselves. There are small businesses that buy in bulk and sell to other customers through the retail outlet channel. This is a trans-national dataset that contains all the transactions occurring between Nov-2016 to Dec-2017 for a UK-based online retail store.

The project objective is to find significant customers for the business who make high purchases of their favorite products. The organization wants to roll out a loyalty program to high-value customers after the identification of segments.

The dataset we used is from the following source -

**Kaggle:** <https://www.kaggle.com/vik2012kvs/high-value-customers-identification>

We need to clean the data and the Data Mining algorithm used is K-means clustering and it is in standard form. The packages we used to implement the network are NumPy, pandas, seaborn, matplotlib, sklearn and scipy.

For the k-means clustering method, the most common approach for metric evaluation is the so-called elbow method. It involves running the algorithm multiple times over a loop, with an

increasing number of cluster choices, and then plotting a clustering score as a function of the number of clusters.

### **References :**

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

<https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/>

<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

<https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html>

### **Rough Schedule:**

<u>Task</u>	Date
Github repository	April 13th
Data Preprocessing	April 14th-15th
Applying Algorithm and Analysis	April 16th - 20th
Data Visualization and Model evaluation	April 20th - 25th
Project Report	April 25th -April 29th
Project Completion (Video Recording)	April 30th