# MACHINE LEARNING ENGINEER CAPSTONE PROPOSAL PROJECT

# STARBUCKS CAPSTONE PROJECT (PROPOSED BY UDACITY)

## Domain Background

The Starbucks Capstone Project focuses on the Business Analytics domain of the American coffee company and fast-food chain, **Starbucks**. It involves analysing how promotional offers affect purchasing decisions, taking into consideration certain traits of the buyers such as age, gender, income and even whether the buyer likes to receive offers.

Analysis can be done using customer segmentation. This is therefore a classification problem where given the input data containing demographic, transactional and offer information of various customers, subgroups containing very similar data points are created and examined to deduce how demographic attributes relate to offer types and behaviour towards offers. For example, a cluster might give an output indicating that 90% of people above age 40, earning over a $1000 do not view offers but still complete them.

The results obtained from this project can be used to predict the certain types of offers a customer would most likely respond to and the method(s) to advertise the offers (which might include not advertising to the customer at all) in a supervised learning model.

## Problem Statement

Starbucks would like to know how the following types of offers affect the purchasing decisions of different customers:

1. BOGO (Buy One Get One)
2. Discount
3. Informational

This project aims at identifying which groups of people/individuals are most responsive to each type of offer, and how best to present such offer.

## Datasets and Inputs

The data used is gotten from a program that simulates how people make purchases influenced by promotional offers on the Starbucks mobile app. There are no Starbucks products stated explicitly, and it is assumed that just one product is used. The datasets used to model, train and validate are:

1. profile.json: This contains demographic data about the rewards program users. There are 17000 users and 5 fields which include:
    - gender: (categorical) M, F, O, or null

- age: (numeric) missing value encoded as 118
- id: (string/hash)
- became_member_on: (date) format YYYYMMDD
- income: (numeric)

2. portfolio.json: This contains the offers sent during the 30-days simulation period. There are 10 offers and each has 6 fields which are:
    - reward: (numeric) money awarded for the amount spent
    - channels: (list) web, email, mobile, social
    - difficulty: (numeric) money required to be spent to receive reward
    - duration: (numeric) time for offer to be open, in days
    - offer_type: (string) bogo, discount, informational
    - id: (string/hash)

3. transcript.json: This is the transactional data that shows the events such as when a user views, receives or completes an offer. It contains 306648 events and 4 fields which are:
    - person: (string/hash)
    - event: (string) offer received, offer viewed, transaction, offer completed
    - value: (dictionary) different values depending on event type
        - offer id: (string/hash) not associated with any "transaction"
        - amount: (numeric) money spent in "transaction"
        - reward: (numeric) money gained from "offer completed"
    - time: (numeric) hours after the start of test

The demographic, offer and transactional data are combined to be used to discover the group patterns in the data. These datasets and inputs are provided by Udacity.

**Solution Statement**

Using customer segmentation, identify based on demographic traits:

- How certain groups of individuals are responsive to each offer.
- The group of individuals who make purchases without considering offers (i.e those who don't receive offers and still make purchases and those who receive offers, don't view them but make purchases)

**Benchmark Model**

The proposed benchmark model for this project is the K-means algorithm employed for customer segmentation in this Medium article. It will serve as a basis for comparing the performance of the model employed in this project.

**Evaluation Metrics**

The Silhouette score would be used as an evaluation metric to determine how well the points were assigned to clusters. It returns a value between -1 and +1; the best value is +1 and the worst is -1. Values close to 0 indicate overlapping clusters. It can also be used to select the optimum number of clusters K to be used.

**Project Design**

The project workflow is as follows:

1. Data Loading and Exploration: General exploration of data. The input datasets are explored using visualization tools. For instance, histogram or bar chart showing the variation between age groups and offer types.
2. Data Cleaning and Preprocessing: Combining the datasets to form a new dataset with correlated information. Checking for missing values. Converting categorical data to numeric data and normalizing the numeric data.
3. Dimensionality reduction: Discover features that help to separate and group data, that is, features that cause the most variance in the dataset. Principal Component Analysis is chosen approach for this.
4. Feature engineering and data transformation: Selecting the features or components to use in the model based on variance. Then transforming the training data using the PCA model.
5. Clustering transformed data with either GMM or DBScan: Both models might be used and the more accurate one is selected.
6. Extracting trained model attributes and visualizing k clusters: The demographic information for each cluster is examined as it relates to the certain offer types and behaviour of customers towards offers.
7. The observations from the clusters and proposed strategies are clearly stated out as results.
8. **Optional**. Employing the clusters as features to a new supervised learning model to predict the type of offer a customer would respond to, given the demographic information.
9. **Optional**. Deploying the model created in No. 7 to a web application: A web application would serve as a platform to determine the offers a customer is most likely to respond to, and if the customer's purchasing decisions are determined by offers.