Capstone Project: Create a Customer Segmentation Report for Arvato Financial Services

- **Domain Background and Problem statement**

The project is related to data analysis in business. It provides demographic data for customers of a mail-order sales company and demographic information for the general population in Germany. The goal is to analyze data and create a customer segmentation report, that identifies the key features of the core customer base of the company. It also helps to target the company's marketing campaign for potential customers.

There are two main steps: 1. using unsupervised learning techniques for customer segmentation. 2. build a model to predict which individuals are most likely to convert into future customers for the company.

- **Datasets and inputs**
    - *Udacity_AZDIAS_052018.csv:* Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
    - *Udacity_CUSTOMERS_052018.csv*: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
    - *Udacity_MAILOUT_052018_TRAIN.csv*: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
    - *Udacity_MAILOUT_052018_TEST.csv*: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

The data has been provided by Bertelsmann Arvato Analytics. They are needed to clean up and do data analysis and preprocessing.

- **Solution statments**

I will use Python in Jupyter notebook locally and AWS sagemaker instance. The package Pandas will be used for data analysis and preprocessing. The unsupervised and predictive models will be built by sklearn.

- **Benchmark Model**

For unsupervised part, K-Means method will be used.

For the prediction model, the performance of XGBClassifier and Random Forest will be compared

- **Evaluation metrics**
    1. using PCA to identify some key features
    2. Since the problem is a binary classification problem, a metric of True Negative Rate will be used.

$$TNR = \frac{tn}{tn + fp}$$

- **Project Design**

1. Data Analysis and Preprocessing
2. Customer segmentation report by unsupervised learning
3. Predition of potential customers by supervised learning.
4. Conclution