Team Members:

- Ray Ng
  - rayng2018@gmail.com
  - Canada
  - University of British Columbia
  - Specialization: Data Science
- Rita Uzoka
  - rita.uzoka@yahoo.com
  - United Kingdom
  - Sheffield Hallam University
  - Specialization: Data Science
- Fatemeh Bagheri
  - f.bagheri13@gmail.com
  - France
  - Université Jean Monnet St Etienne - Université de Lyon
  - Your Specialization: Data Science

Problem: Customer Segmentation - XYZ bank wants to roll out Christmas offers to their customers. But the bank does not want to roll out the same offer to all customers, instead they want to roll out personalized offers to particular sets of customers. If they manually start understanding the category of customer then this will not be efficient and also they will not be able to uncover the hidden pattern in the data (pattern which groups certain kinds of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 groups as this will be inefficient for their campaign.

GitHub Repo Link:

https://github.com/faba13/VC3.git

Data Cleansing

data cleansing is an important process that needs to be done when transferring files from a database to a data warehouse. When there is lack of accuracy or being inconsistent, incomplete, duplicated, and redundant this procedure is done. On the other hand data cleansing is the process of removing unwanted data from the database to improve the consistency and accuracy of the files before they are transferred to the data warehouse.

The process of data cleaning includes:

- Correcting format
- Checking the accuracy of information
- Compiling all data information in a single area
- Standardizing data
- Identifying and fixing errors
- Removing incorrect data

The steps involved in the process of data cleansing are:

- Removing irrelevant observations
- Fixing errors in structure
- Filtering irrelevant or unwanted outliers
- Handling missing information
- Identifying the purpose of the data

Data Cleansing and Transformation Techniques:

- How to handle NA values?
    - What to assume for each column, for example?
- How to handle strange data like negative numbers which should not be in place, or strings where it’s usually numbers?
- How to handle outliers?

Share your code here (for collaborative purposes, not to be submitted, but as a ipynb notebook)

```
import pandas as pd
```