# Week 8 - Deliverables

## Team member's details

**Member 1**

Name: **Rayan Yassminh**

Email: ryassminh@yahoo.com

Country: USA

**Member 2**

Name: **Kavinilavan Muthukumar**

Email: mukavin@gmail.com

Country: United Kingdom

College: Teesside University

**Group Name : STEM Group**

**Specialization :** Data Science

**Project Name :** Customer Segmentation

## Problem description

Customer Segmentation approach-
Separate the customers into 5 different groups based on their behaviors and Lifetime values.

## Data understanding

This dataset contains the information regarding the bank customers. The format is in CSV. It contains the data of geographical, behaviors and so on. It has 47 columns and 1000000 records (customers) with the data types of object, floats and integers. It has missing values for some of the features.

## problems in the data

The problems in the dataset are follows

Missing Values

More than 80% of records are missing particular features.

Features names are inappropriate. All names in Spanish language.

Some features data types are inappropriate. For instance age and seniority months columns have float values

Some Data quality issues. Same subsets of the group have different columns (one hot encoding format). For instance term deposit and account types.problems are difficult to identify the customers which have no term deposit and number of accounts types they have while doing exploratory data analysis.

## Solution for problems

Missing values : since our columns are related to each other, we are not using the statistical methods to handle the null values, instead we are using interactive imputers methods, which consider the other columns as well when filling the null values. A more sophisticated approach is to use the interactive imputers class, which models each feature with missing values as a function of other features, and uses that estimate for imputation. It does so in an iterated round-robin fashion: at each step, a feature column is designated as output `y` and the other feature columns are treated as inputs `X`. A regressor is fit on `(X, y)` for known `y`. Then, the regressor is used to predict the missing values of `y`. This is done for each feature in an iterative fashion, and then is repeated for `max_iter` imputation rounds. The results of the final imputation round are returned.

More than 80% of records are missing particular features: simply drop such columns because those columns are not very crucial for model development.

Features names are inappropriate: all the names of the columns are in spanish, we convert into english language  with rename methods in pandas.

Some features data types are inappropriate: For instance age and seniority months columns have float values, those columns are converted into appropriate data types using pandas.

Some Data quality issues. Same subsets of the group have different columns (one hot encoding format). For instance term deposit and account

types.problems are difficult to identify the customers which have no term deposit and number of accounts types they have while doing exploratory data analysis:-

Solution : merging the different columns into single columns which belong to the same categorical group.

## Github Repo link

Link 1
https://github.com/yassmin1/Data-Science-Intern-at-Data-Glacier/tree/group_project/DS_Customer%20_Segmentation

Link 2 :
https://github.com/kavinilavanM/Data-Glacier-internship