Case Study: Retention Analysis

A Spanish Bank has been collecting data on multiple facets of customer behavior. Currently the data science team at the bank is trying to help in developing strategies to retain customers.

A team of data warehousing experts have already collected data that the management thinks could be helpful in doing the relevant analysis.

The dataset extracted has following variables:

Varaiables	**Description**
gender	Gender of Customer
age	Age of Customer
gross_income	Annual income
segment	Segment as specified by the bank
num_credit_cards	Number of credit cards issued (through cross sell)
tiprel_1mes	Customer Inactivity Flag at the beginning of the period (A-Active, I-Inactive)
ind_actividad_client e	Customer Inactivity Flag at the end of the period (6 Months) (1-Active, 0-Inactive)
num_Products	Total number of financial products that the customer has bought (through cross sell)
num_loans	Total number loans disbursed to the customer (through cross sell)
duration	Number of days since customer

This data has been extracted as a json file named 'retention.json'. In order to study the behaviour of customers who become inactive, the data from datawarehouse was pulled for a duration of 6 months and within this duration it was observed which customers were active or inactive in the beginning of the six month window and which customers are active/inactive after the 6 month time window.

You need to carry out the following tasks:

Task 1. Data Quality Check

1. Do you have all the relevant fields in the raw data file given to you?
2. Convert the given json data to a csv file
3. Create a data quality report (the format of this report must be decided by the learner) to check:
    a. The data type of each variable
    b. If a variable is numeric in nature, then, the numeric summary (Min, Max, Mean,25th percentile, Median, 75th percentile, 90th percentile, 95th percentile,

number of zeros and number of unique values, number of missing values, percentage of missing values) must be computed
    c. If a variable is string variable, then find out the number of unique values, number of missing values, percentage of missing values.
4. After the data quality report is created you need to :
    a. Check if there is any variable whose data-type needs to be changed
    b. Identify the type of data cleaning needed for different columns in the data
    c. Handle missing data appropriately
    d. In case there are extreme values present in a variable do the appropriate treatment.

Task 2. Data Exploration and business hypothesis testing

1. For people who were inactive at the start of the study and were active by the time the study ended, is there a pattern in terms of age and gender?
2. Do people with more than average annual income tend to have relatively high activity rates compared to people with less than average annual income?
3. What is the relationship between the number of products owned by customers who were active at the start and at the end vs those who were active at the start but were inactive at the end of the study period?
4. How people who display consistent behaviour (active at start and active at end, inactive at start and inactive at end) differ from people who display a change in their behaviour (active at start but inactive at the end or inactive at start but active at end)?
5. Generate elaborate profiles for the following four groups:
    a. Active at the start but inactive at the end
    b. Active at the start and active by the end
    c. Inactive at the start but active at the end
    d. Inactive at the start and inactive at the end

Deliverables:

- A well commented jupyter notebook either pushed on a github repository or on a kaggle kernel.
- All the insights must be either summarized in a PowerPoint along with relevant charts and tables or a blog entry on medium/linkedin etc. In case you are created a powerpoint, push the deck in the repository or create a slideshare link.