Data Glacier Virtual Internship
Week 8 Deliverable
Group: [group name]

Name	Wasiq Ahmed	Samuel Bailey
Email	wasiqahmed7@gmail.com	sjbrsa@gmail.com
Country	Turkey	South Africa
College	Bilkent University	University of Cape Town
Specialization	Data Analyst	Data Analyst

Problem Description

ABC Analytics' client, XYZ Credit Union, has been tasked with providing an analysis of XYZ's customer base which would indicate how they could increase their cross-selling. Cross-selling is defined as the extent to which the average XYZ customer uses multiple different products on offer from XYZ. ABC will use machine learning (ML) models to this end, but first we will conduct EDA to get a sense of what the correlates of cross-selling are likely to be.

Data Description

Having been partially prepared for a ML analysis, the data are divided into two files, Train.csv and Test.csv. Together, these files amount to 2403050342 bytes and contain data for 14576924 distinct observations. There are 24 variables in the data, of which three are date variables (`fecha_dato`, `fecha_alta` and `ult_fec_cli_1t`), two are ordinal (`age`, `antiguedad`), one is continuous (`renta`) and the rest are categorical. The variable `ult_fec_cli_1t` has blank values by design for all observations where the value of `indrel` is 1. These are not considered missing values, which are denoted by `NA` in the data. There are 14576658 observations with missing values, most of these being in the `cod_prov` or `renta` variable. There are some nonsense values as in the ordinal variable representing customer seniority, `antiguedad`, of which three values are -999999.

Other problems in the data include the number of persons older than 110 years, 60, as well as the number of persons older than 100 years, 1001. As of any date in the variable `fecha_alta`, there were only two Spaniards older than 110 years. The oldest customer according to these data is 164, 43 years older than the longest-lived person in history, Jeanne Calment.

Data Preparation Approach

The two files will be combined into a single file containing all observations. Observations with implausible ages, such as supercentenarians, will be deleted. Regarding the missing values, one approach we will take will be to impute the missing values based on the existing data in the rest of the data set. The ‘nonsense’ values will also be replaced with the appropriate statistics. In the interest of reproducibility, we aim to avoid stochastic imputation methods and opt for median imputation for the continuous and ordinal variables and mode imputation for the categorical variables. This would lead to a lower variability than is representative of the population, which can be adjusted for when constructing ML models.

The reason for the choice of median imputation for the `renta` variable is that this variable is positively skewed with a very high number of outliers (so much so that the variable is nearly log-symmetrically distributed), and that the median is more robust to skewness and outliers than the mean.