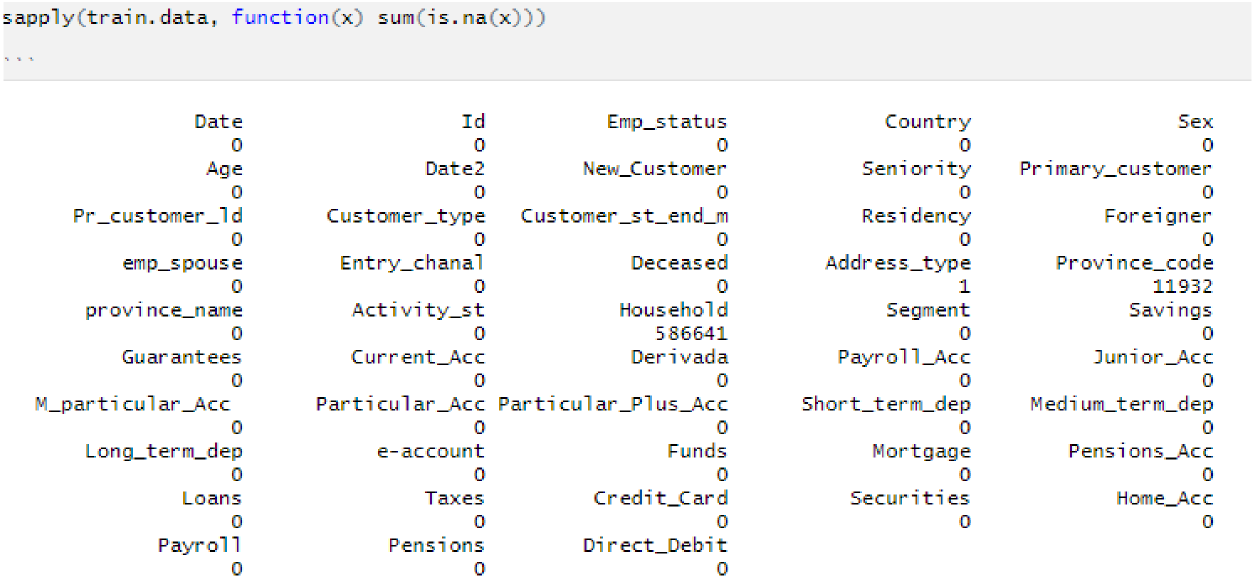
Data cleaning and data preparation

The initial step in data cleaning is to check for missing value, blank spaces or any other inconsistent data. There are 586.641 missing values for the Household and 11.932 missing values for Province_code.



~~~
sapply(train.data, function(x) sum(is.na(x)))
```
~~~

```
                Date                   Id           Emp_status              Country                  Sex
                   0                    0                    0                    0                    0
                 Age                Date2         New_Customer            Seniority     Primary_customer
                   0                    0                    0                    0                    0
      Pr_customer_Id        Customer_type    Customer_st_end_m            Residency            Foreigner
                   0                    0                    0                    0                    0
          emp_spouse         Entry_chanal             Deceased         Address_type        Province_code
                   0                    0                    0                    1                11932
       province_name          Activity_st            Household              Segment              Savings
                   0                    0               586641                    0                    0
          Guarantees          Current_Acc             Derivada          Payroll_Acc           Junior_Acc
                   0                    0                    0                    0                    0
   M_particular_Acc        Particular_Acc  Particular_Plus_Acc       Short_term_dep      Medium_term_dep
                   0                    0                    0                    0                    0
       Long_term_dep            e-account                Funds             Mortgage         Pensions_Acc
                   0                    0                    0                    0                    0
               Loans                Taxes          Credit_Card           Securities             Home_Acc
                   0                    0                    0                    0                    0
             Payroll             Pensions         Direct_Debit
                   0                    0                    0
```

Figure 2. Missing values in the Santander dataset

- Household income: has 586.641 missing values . As it is as significant number of observations, the NA’s will be replaced with the average income per province.
- Province_code has 11.932 missing value, but the variable is not necessary as the same information can pe found in province_name. As a result, the variable Province_code will be removed.

Taking a closer look at the dataset, it is observed that many of the variable have blank spaces instead of a value, which will be replace in the following way:

- Emp_spouse has the value of "S" if the customer is the spouse of an employee and "N" otherwise. Out of 2.710.381 observations, there are only 3 entry for S and 341 for N, the rest are blank spaces. Therefore , the variable emp_spouse doesn't offer a lot of information and the entire column will be removed.
- Sex column has 15 empty spaces and they will be replaced with the most comun value
- Pr_customer_ld has 2.703.492empty space, as a result the column will be removed(as most of the cells were empty)
- The Customer_type column should have the following values:1, 2, 3, 4 and P. The following replacement will tak place: 1.0 with 1, 2.0 with 2, 3.0 with 3 and 4.0 with 4, P with a value of 5 and the empty spaces will be replaced with the most comun value.
- The 47.325 empty spaces in Customer_st_end_m will be replaced as well as the most comun value.
- The Entry_chanal variable has 58026 blank spaces. They will be replaced with the most frequent value that occurs in the case of females and respectively of males.

- The blank spaces in Segment variable will be considered as different segment that it will be named “*Other*”
- The province_name variable has 11932 blank spaces. After further investigation it is clear that the customers for whom the province_name is blank, 19 of them are from Spain and the most comun value will be imputed. The rest of the customers come from other countries than Spain. We will impute the value "*International*" for the blank spaces in this case.
- Date2 :the date at which the individual became a customer of the bank is not needed as the same information is reflected in the Seniority(months)= the difference between Date and Date 2
- The purpose of recommender system is to recommend new products to the active customers. As a result, inactive and deceased customers will be removed.