**Data Presentation: Capstone Project I – Banking Product Recommendation**

In this section, analysis of the dataset is being conducted to highlight important insights necessary for feature engineering. Thus, the description of the dataset is divided in two parts:

1. Features describing the client and relation to a constructed target variable – purchaser group and non-purchaser group, labelled as Target (0,1).
2. The product basket and its dependencies to the Target are highlighted.
3. Statistical tests are applied for further steps of feature and dimensionality reduction (follows another time)
4. **Customer features**

The graphs (appendix) depict different aspects of the features describing the clients. Each frame contains four plots:

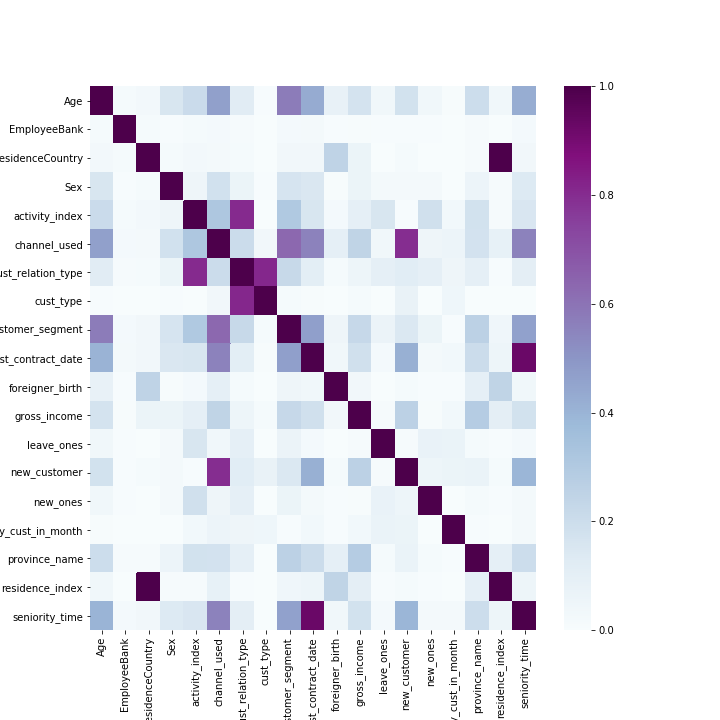
1. For categorical features, the percentage of labels gets depicted, whereas for continuous ones, a histogram shows the dispersion of values.
2. This plot adds a correction to the first one. Based on the fact, that most clients appear more than once in the dataset features get a weighting bias, i.e. if a client with age 30 shows up ten times whereas another one with age 23 shows up only three times. Such bias deteriorates the descriptive statistic and thus in the second plot all feature occurrences have been reduced to unique record entries.
3. A target variable has been introduced purchase yes or no. According to this grouping the proportion of the feature labels, for categorical and boxplot for continuous ones have been created. Thus, this plot is meant to show whether a feature might be useful for prediction.
4. The dataset contains time chronological ordering; hence, this plot is meant to spot time dependent variations of features. Therefore, for continuous features the median value over time, whereas for categorical, the percentage bar plots over time are shown.

Due to the fact, that the features, channel\_used, province\_name and ResidenceCountry have more than ten labels, only the first 6 most current ones are used for plotting. Otherwise any visual value would get lost.

The plots are in the appendix and a summary of remarkable features is provided below:

* Activity\_Index:
  + Meaning: whether a client is classified as active or not by the bank
  + Almost evenly distributed
  + Majority of customers purchasing a product are classified as active
* Age:
  + bimodal distribution around mid 20 and mid 40.
  + The group of purchasers shows a slightly higher age
  + Over time there is a slight shift in the median
* Cust\_relation\_type:
* How bank classifies the relation to customer.
* Dominated by labels Active and Inactive
* “Active” ones tend dominate the purchaser group
* Cust\_segment:
* More label variance in distinguishing purchasers from none purchasers
* First\_contract\_date:
* Bimodal distribution, slightly higher time differences for purchasing group
* New\_customers:
* Majority are old customers; slightly higher proportion of new customers in the purchasing group
* leave\_ones:
  + Show considerable variation in in purchasing group; however, most records did not sell any products
* new\_ones:
  + slight variation in purchaser group; i.e. people who bought a product are in the group of buyers ext month
* seniority-time:
  + bimodal distribution; slightly higher median in purchaser group
* the features with manipulated labels show no real dominance; i.e. manipulation was done by leaving the 6 most common labels and relabelling the remaining ones by “0”
* many features with no variation and thus no real informational content due to the prevalence of one label: “Residence\_Index”,“primary\_cust\_in\_month”,”foreigner\_birth”,”Employee Bank”,”cust\_type”. These variables are likely to get dropped in the next steps.

The graph below shows a heatmap of features:



Continuous labels have been binarized in quintiles and respective contingency tables have been constructed. Then Cramers V was calculated to measure dependency between features, thus the colormap intends to show the extent of “correlation”.

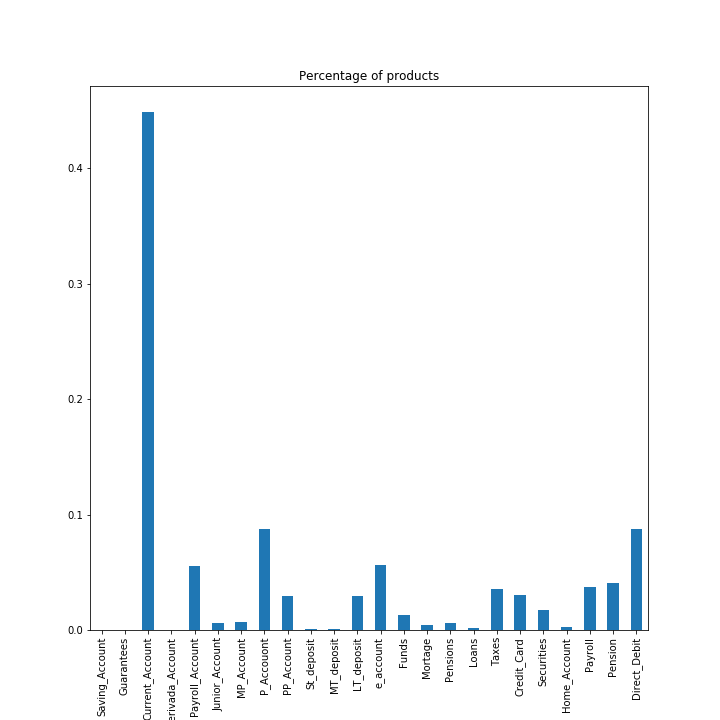
Some relations are arbitrary, like high correlation between residence index and ResidenceCountry or seniority\_time and first contract date. However, there are considerable dependencies, i.e. channel\_used to new\_customer, customer\_segment to Age or activity index to relation\_type. The table below shows Cramers V and the respective p-values of every feature against the groups “purchaser” or “no-purchaser”, i.e. “Target (0,1).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **feature** | **p-value** | **Cramers** |
| **0** | **EmployeeBank** | **0.0** | **0.0079** |
| **1** | **ResidenceCountry** | **0.0** | **0.0086** |
| **2** | **Sex** | **0.0** | **0.0235** |
| **3** | **Age** | **0.0** | **0.0665** |
| **4** | **first\_contract\_date** | **0.0** | **0.0391** |
| **5** | **new\_customer** | **0.0** | **0.0598** |
| **6** | **seniority\_time** | **0.0** | **0.0372** |
| **7** | **primary\_cust\_in\_month** | **0.0** | **0.0021** |
| **8** | **cust\_type** | **0.0** | **0.0351** |
| **9** | **cust\_relation\_type** | **0.0** | **0.1783** |
| **10** | **residence\_index** | **0.0** | **0.0054** |
| **11** | **foreigner\_birth** | **0.0** | **0.0014** |
| **12** | **channel\_used** | **0.0** | **0.1045** |
| **13** | **province\_name** | **0.0** | **0.0465** |
| **14** | **activity\_index** | **0.0** | **0.1623** |
| **15** | **gross\_income** | **0.0** | **0.0151** |
| **16** | **customer\_segment** | **0.0** | **0.082** |
| **17** | **new\_ones** | **0.0** | **0.0814** |
| **18** | **leave\_ones** | **0.0** | **0.3539** |
|  |  |  |  |

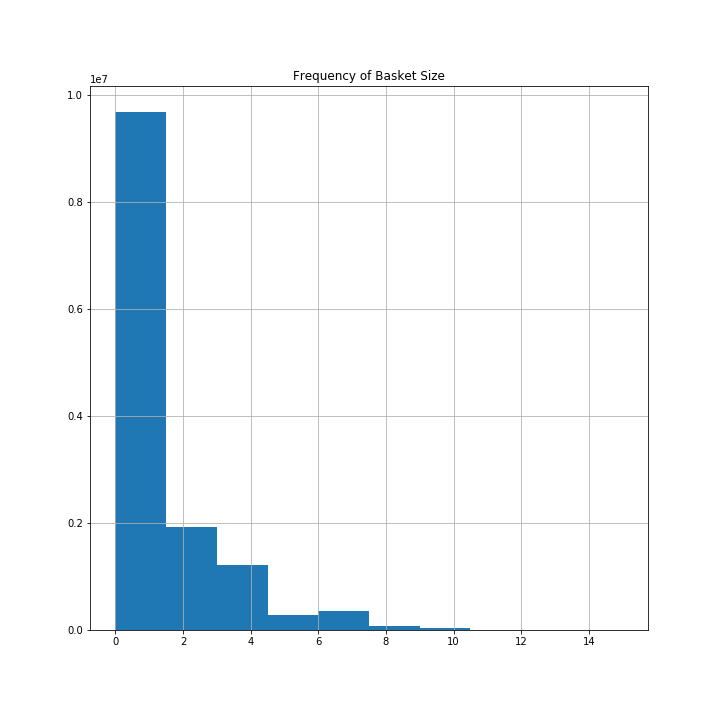
Obviously, all p-values hint at statistical significance while Cramers V does not reveal strong impact, i.e. dependency. Expect for the feature leave\_ones, i.e. measures how many products a customer has sold with respect to the last month. Obviously, this feature variation shows a stronger dependency to whether a customer will buy a product next month or not.

2) Product basket and target analysis

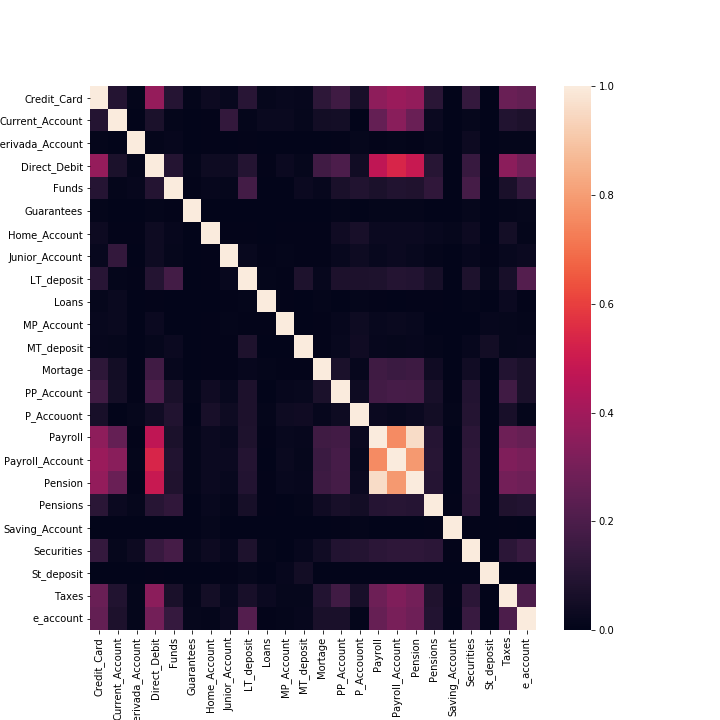
The product basket may consist of 24 diverse services and is represented as sparse matrix. The graph below shows the percentage distribution of the different services.



Current Accounts is the most prominent product followed by Personal Account and Direct Debit. The next chart shows the distribution of number of products per record.



Most customers tend to have one product in their basket and the count drops significantly after one and four products.

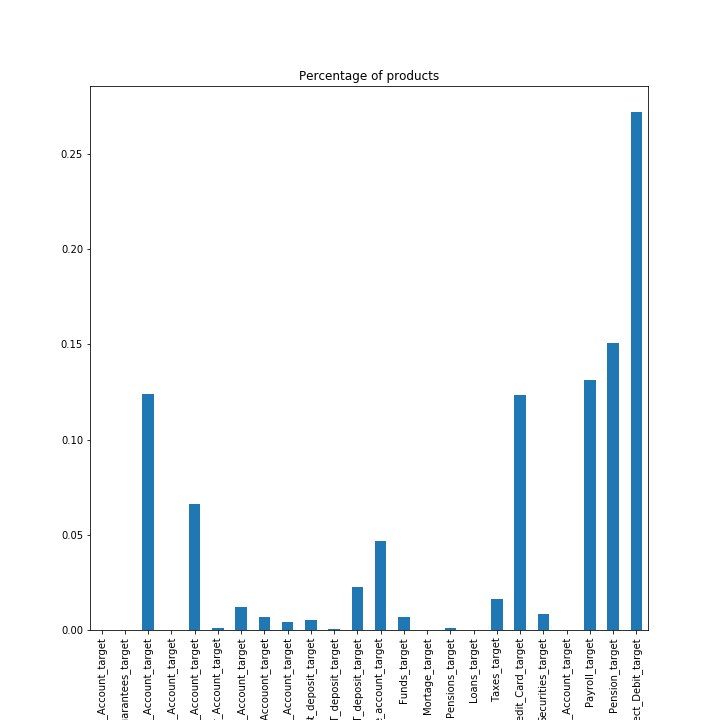


The heatmap above shows the correlation among features which is based on Cramers V derived from contingency tables. Direct Debit has some dependence to Payroll and Pension and Credit Card has some correlation to Direct Debit. However, the heatmap does not reveal very strong relational patterns overall.

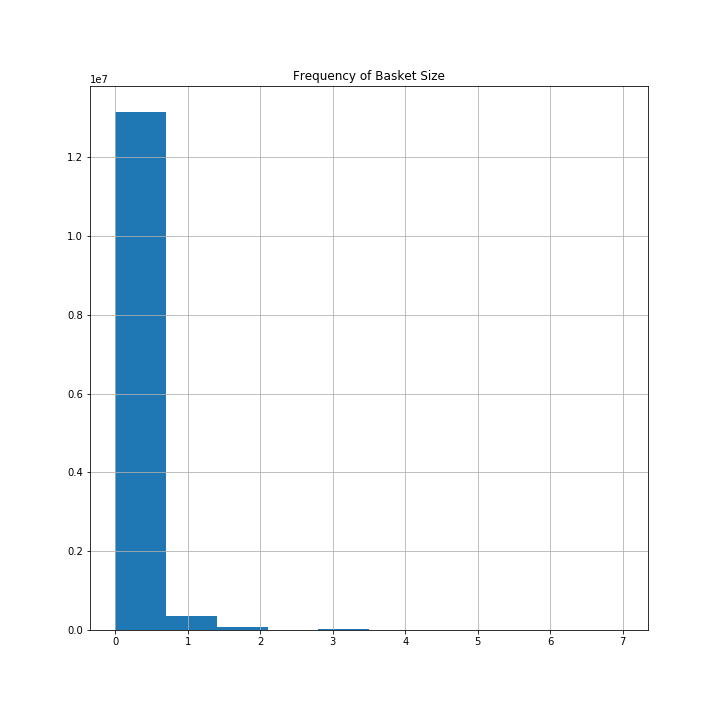
|  |  |  |  |
| --- | --- | --- | --- |
|  | **feature** | **p-value** | **Cramers** |
| **0** | **Saving\_Account** | **0.0158** | **0.0007** |
| **1** | **Guarantees** | **0.0** | **0.0019** |
| **2** | **Current\_Account** | **0.0** | **0.0555** |
| **3** | **Derivada\_Account** | **0.0** | **0.0046** |
| **4** | **Payroll\_Account** | **0.0** | **0.1624** |
| **5** | **Junior\_Account** | **0.0** | **0.014** |
| **6** | **MP\_Account** | **0.0** | **0.0439** |
| **7** | **P\_Accouont** | **0.0** | **0.0073** |
| **8** | **PP\_Account** | **0.0** | **0.0479** |
| **9** | **St\_deposit** | **0.0** | **0.034** |
| **10** | **MT\_deposit** | **0.0** | **0.0074** |
| **11** | **LT\_deposit** | **0.0** | **0.0475** |
| **12** | **e\_account** | **0.0** | **0.0902** |
| **13** | **Funds** | **0.0** | **0.0324** |
| **14** | **Mortage** | **0.0** | **0.0227** |
| **15** | **Pensions** | **0.0** | **0.0268** |
| **16** | **Loans** | **0.0** | **0.0024** |
| **17** | **Taxes** | **0.0** | **0.0855** |
| **18** | **Credit\_Card** | **0.0** | **0.0659** |
| **19** | **Securities** | **0.0** | **0.0433** |
| **20** | **Home\_Account** | **0.0** | **0.0092** |
| **21** | **Payroll** | **0.0** | **0.0804** |
| **22** | **Pension** | **0.0** | **0.0847** |
| **23** | **Direct\_Debit** | **0.0** | **0.108** |

This table shows again Cramers V and p-values of the features against the groups “purchasers” and “non-purchasers”, i.e. Target (0,1). Accordingly, all p-values signal statistical inference while the effect of dependence is overly weak. Only, Payroll Account and Direct Debit show numbers in the single digit area.

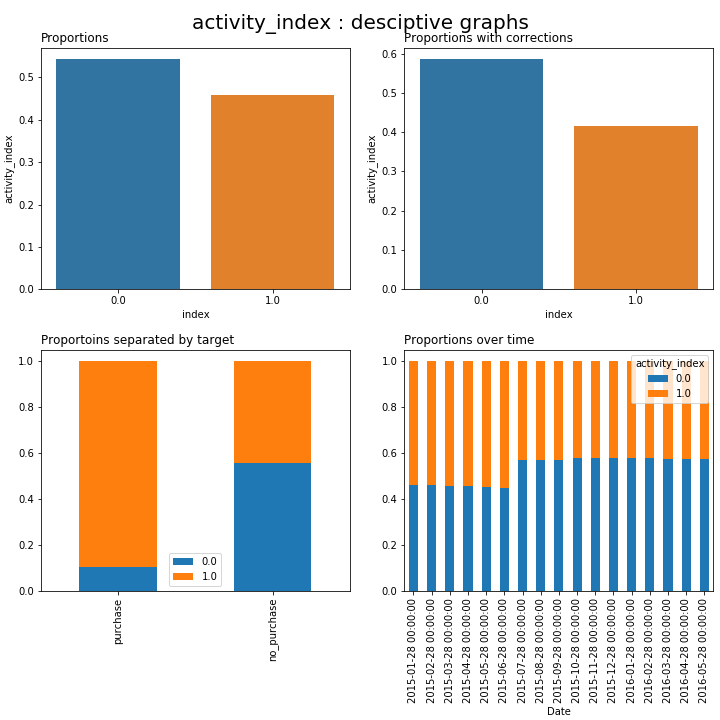
The target basket, i.e. the products which are getting purchased one month ahead are calculated over the time point change of the product basket. The graph below shows the percentage of purchased products.

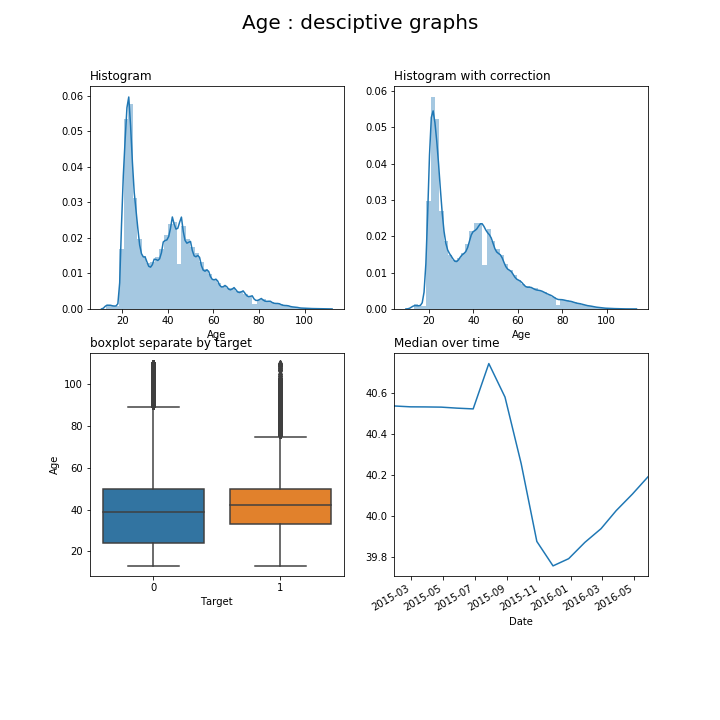


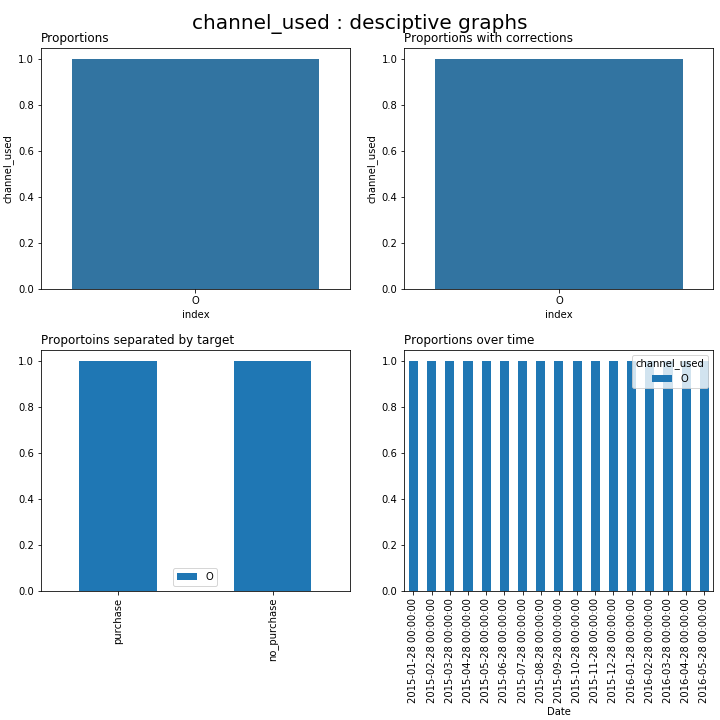
Most commonly, Direct Debit, Pension, Payroll, Current accounts and credit cards are purchased, while the number of purchases per record reaches a maximum of 7 and most often only one product was added as shown on the histogram below.

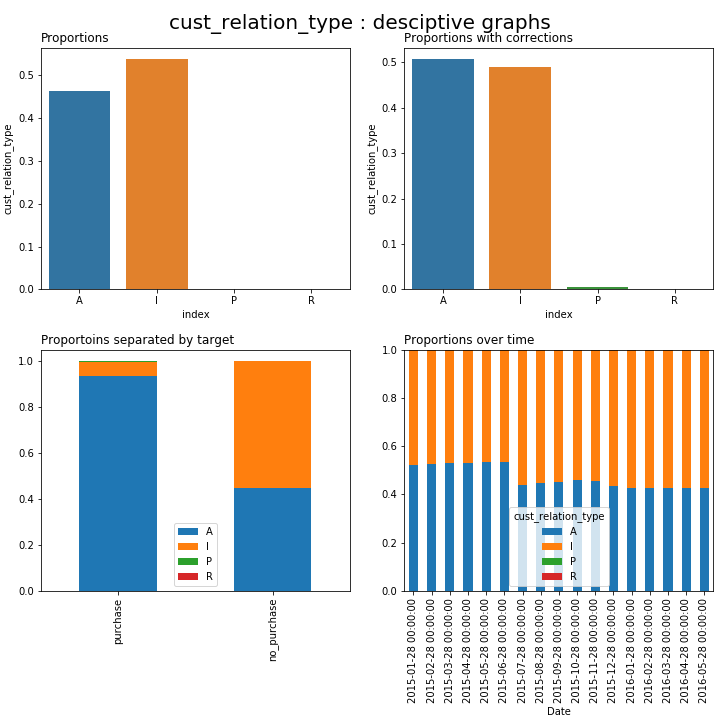


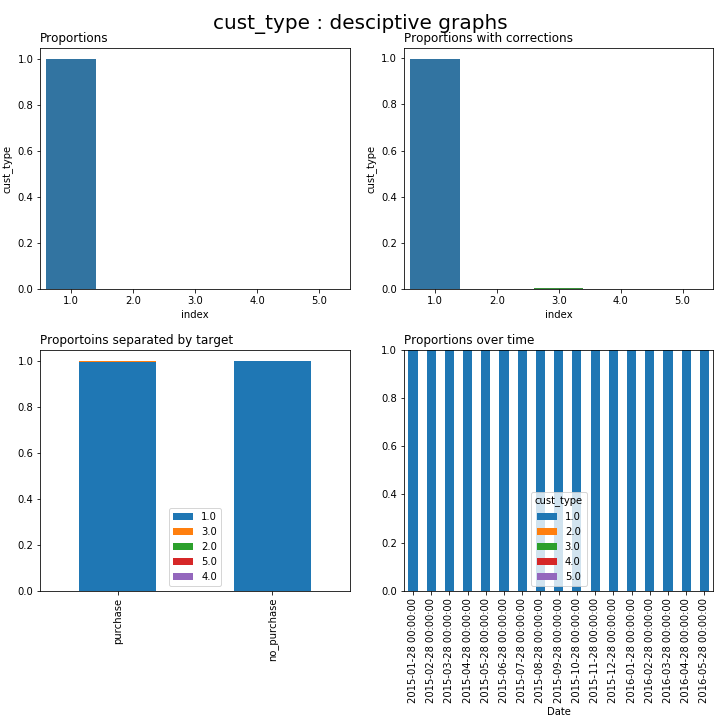
**Appendix:**

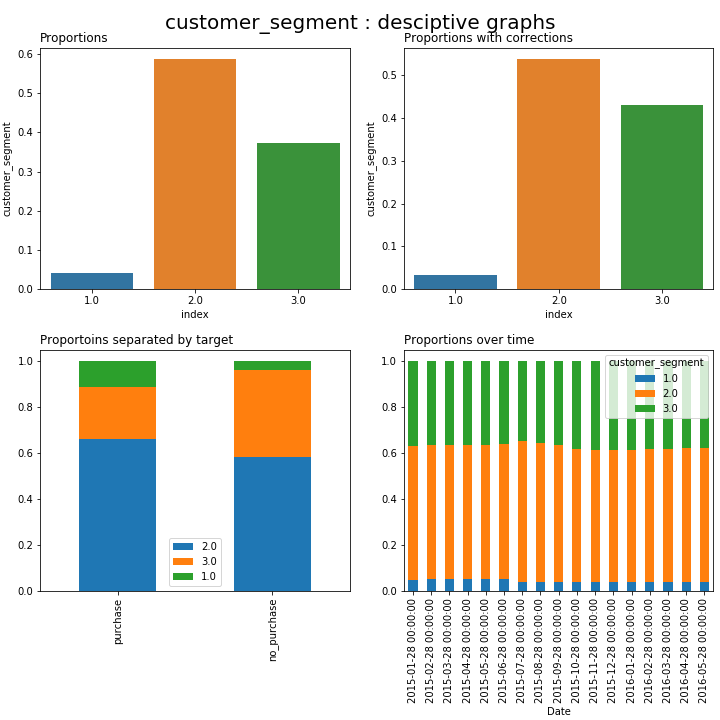


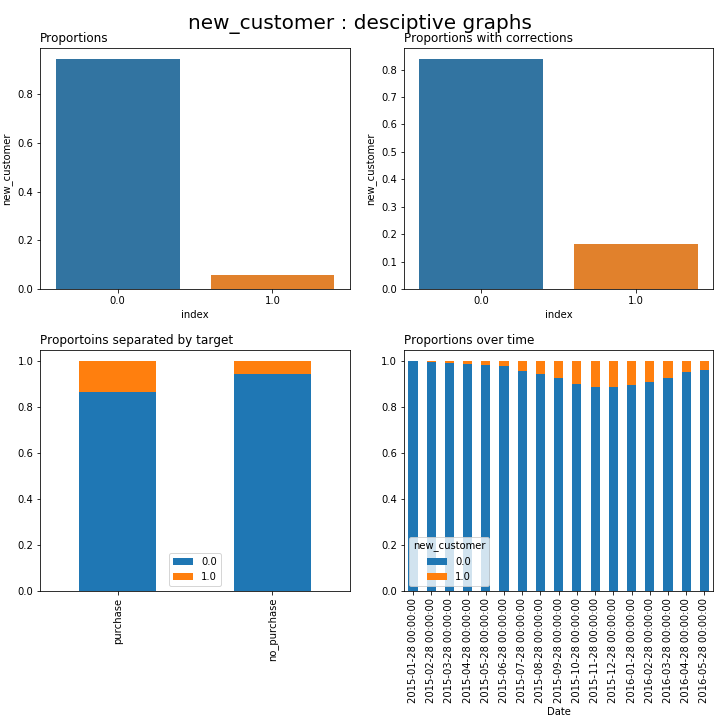
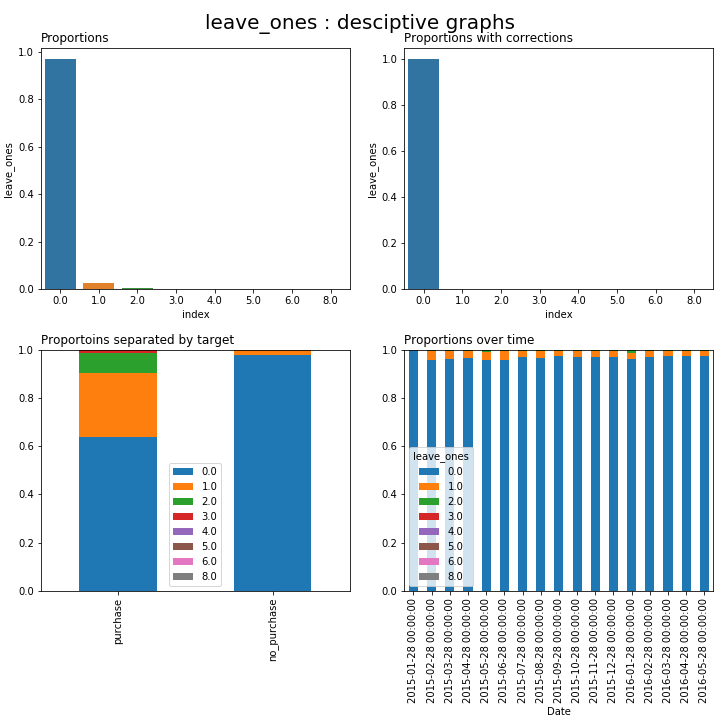
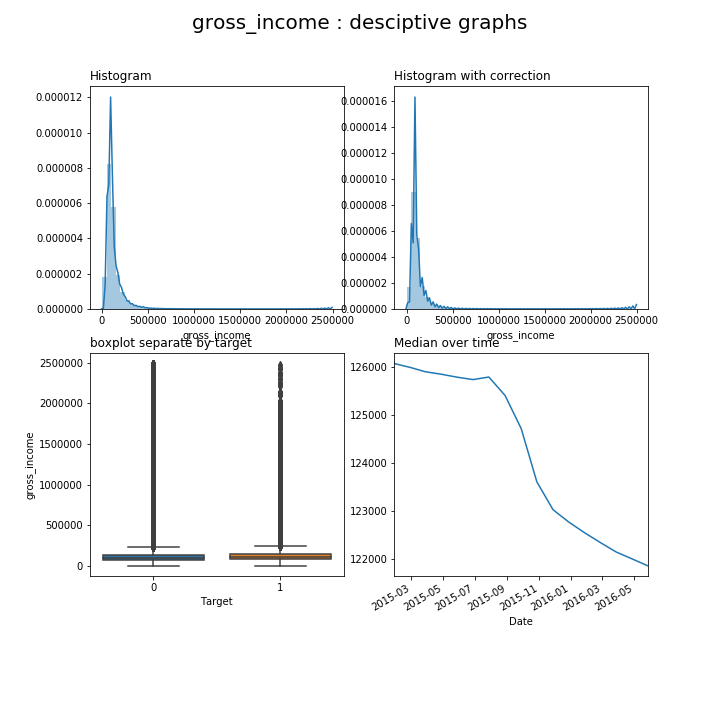
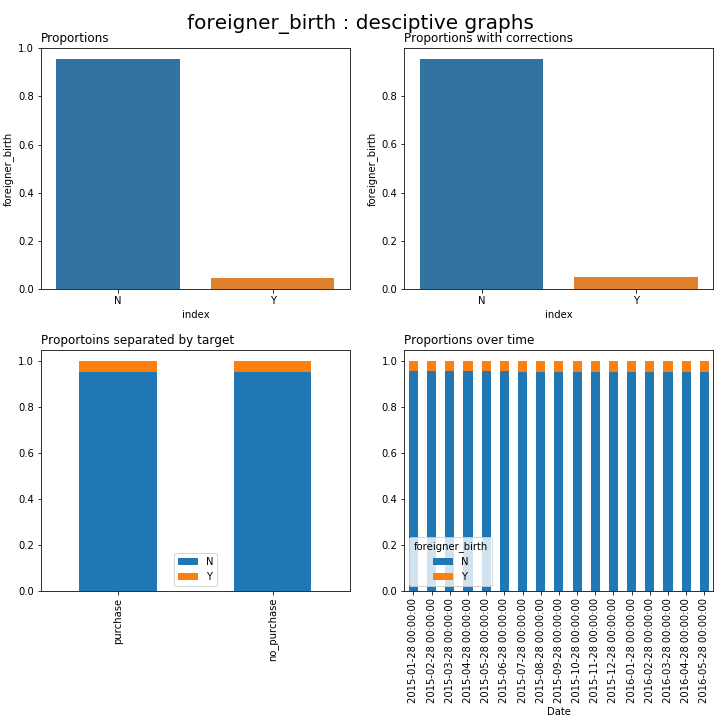
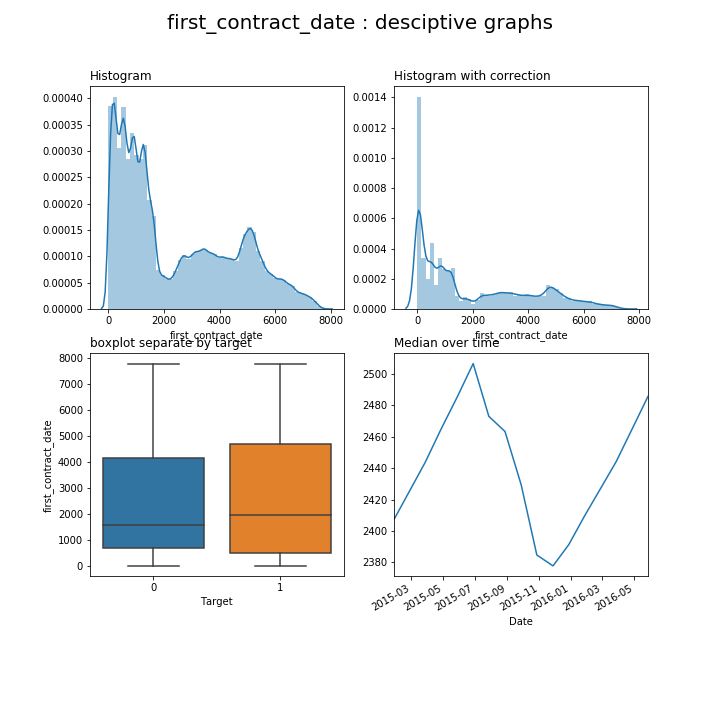
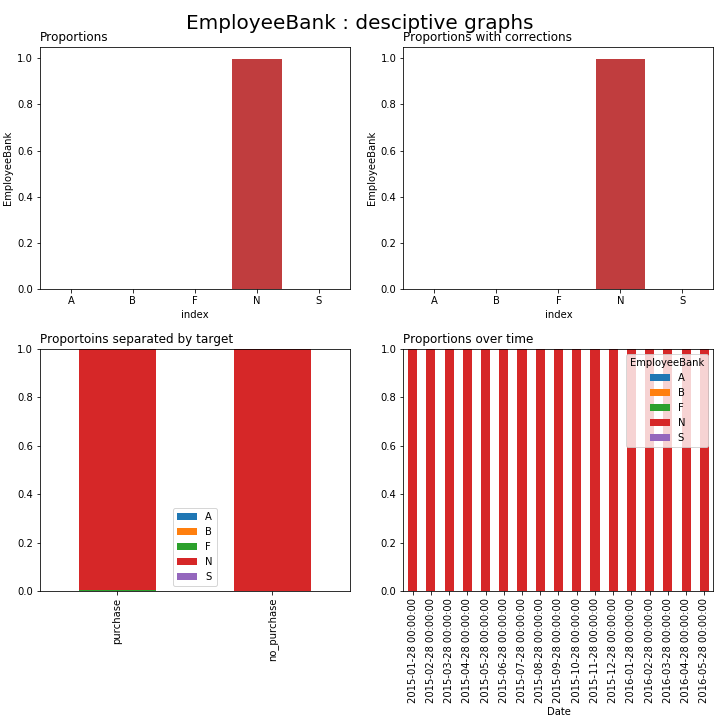


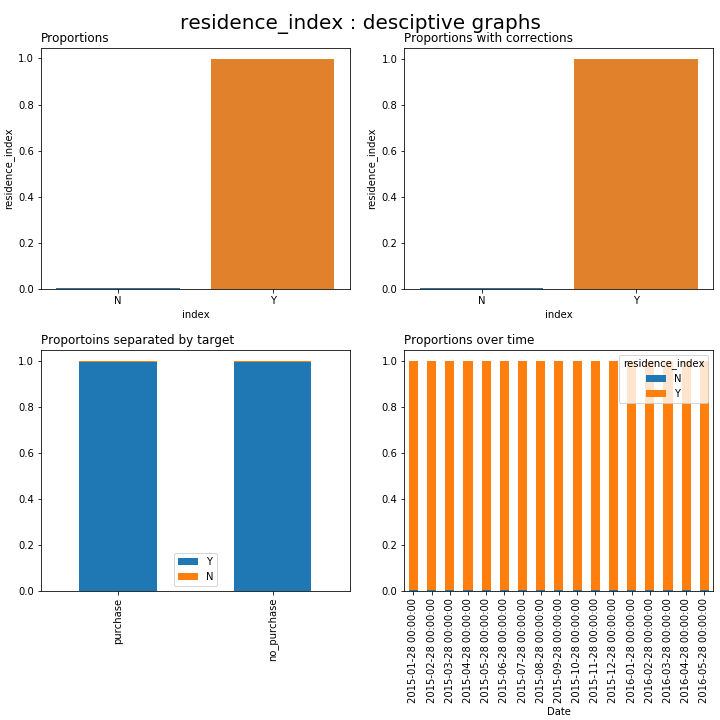
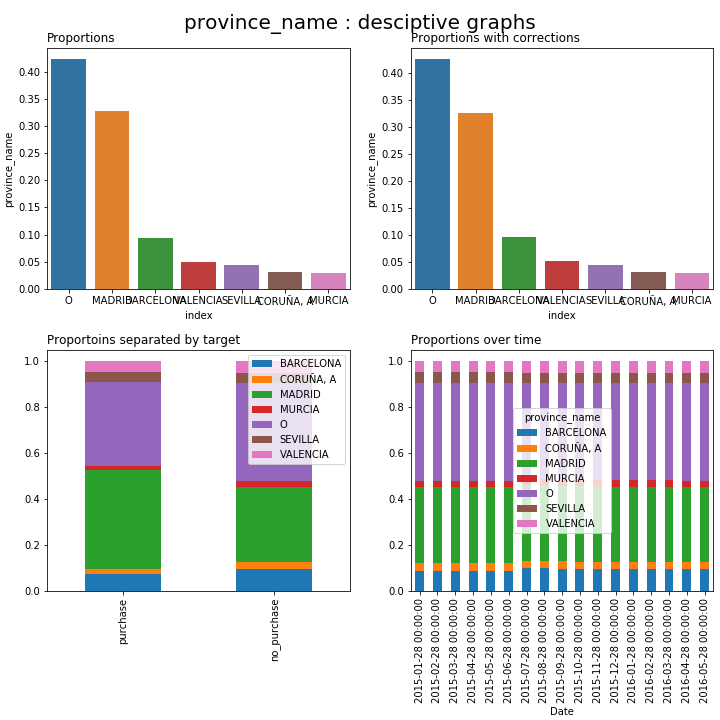
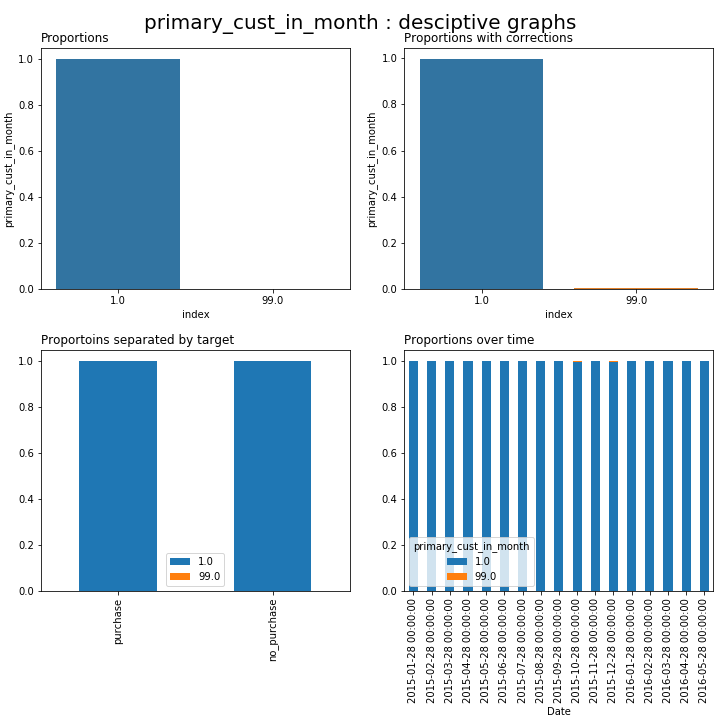
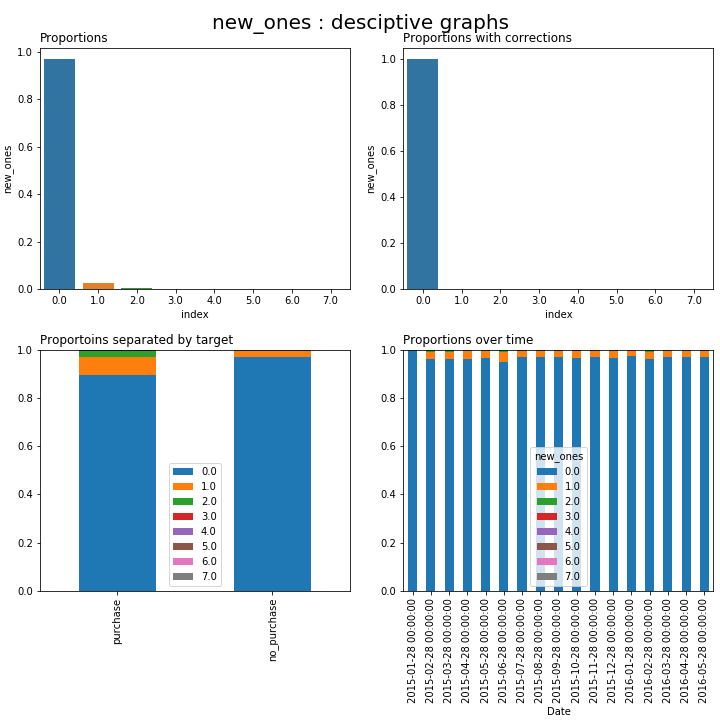


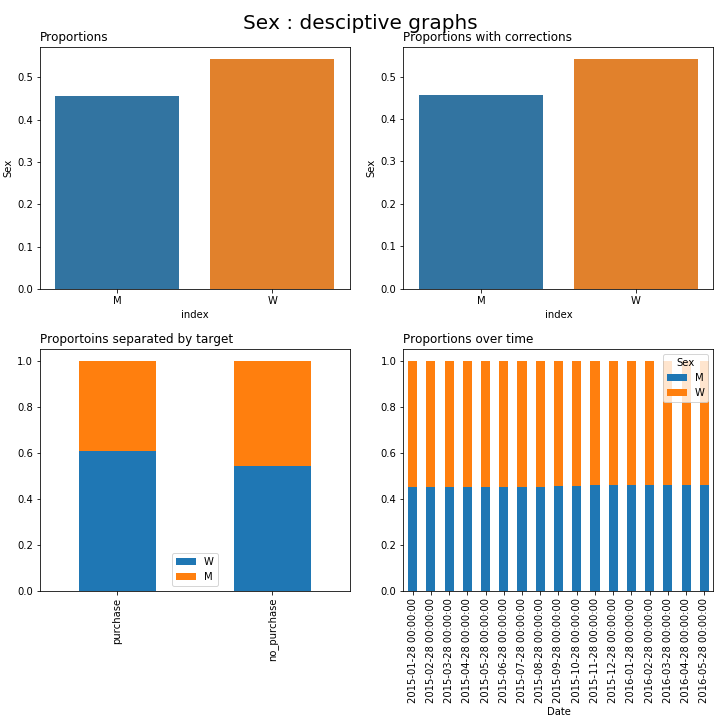
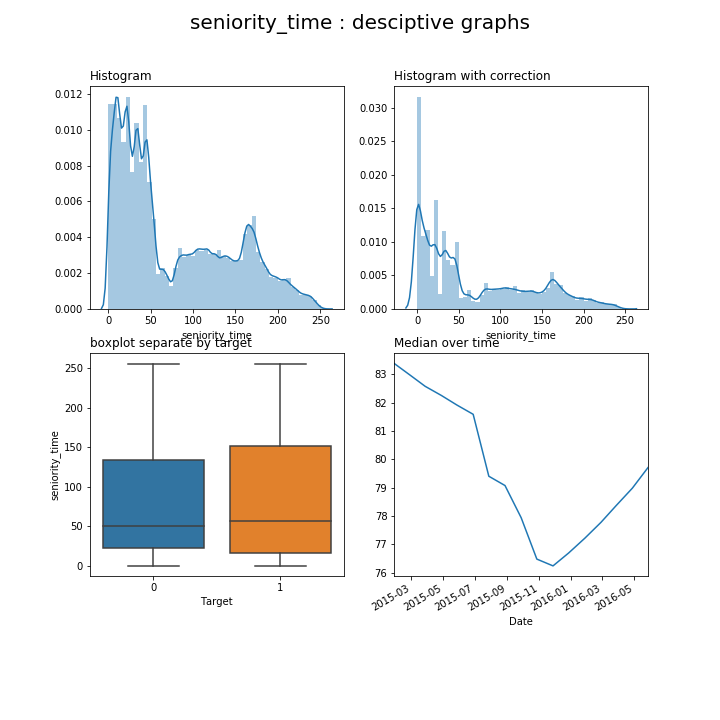
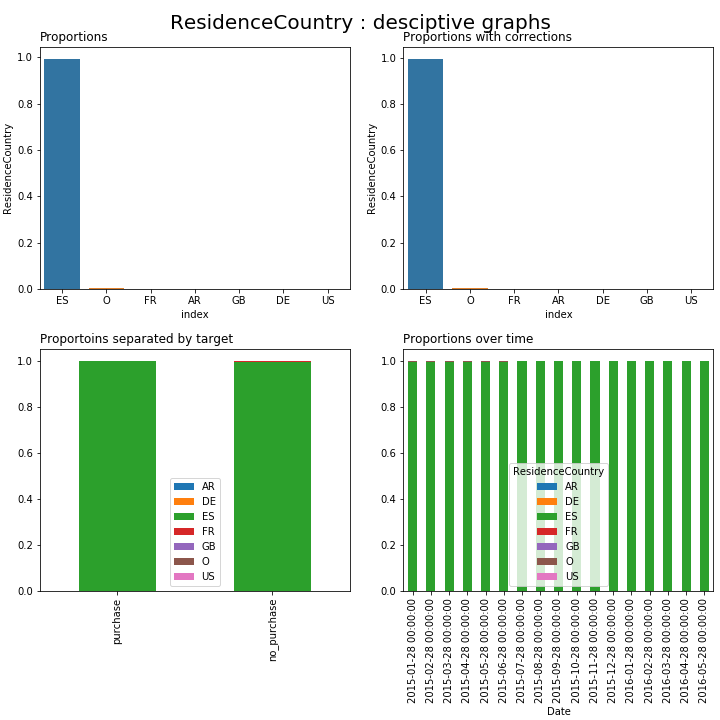
-

-









* Activity\_Index:
  + Meaning: whether a client is classifies as active or not by the bank
  + Almost evenly distributed
  + Majority of customers purchasing then next a product are classified as active
* Age:
  + bimodal distribution around mid 20 and mid 40.
  + The group of purchasers shows a slightly higher age
  + Over time there is a slight shift in the median

Cust\_relation\_type:

* How bank classifies the relation to customer.
* Dominated by labels Active and Inactive
* “Active” ones tend dominate the purchaser group

Cust\_segmeant:

* More label variance in distinguishing purchasers from none purchasers

First\_contract\_date:

* Bimodal distribution, slightly higher time differences for purchasing group

New\_customers:

* Majority is are old customers; slightly higher proportion of new\_customers in the purchasing group