Data exploration of the Tugas dataset

Data Mining II - Handout

Maria Inês Pastor Pereira da Silva, D20170088

4th April 2018

Contents

1 Introduction

TugasRWe is an on-line retailer on 5 main departments - Clothes, Housekeeping, Kitchen, Small appliances and Toys. For 2 years, they have been collecting data through their on-line platform with the aim of getting new insights on their customers and improving their marketing strategies.

The Tugas dataset, which will be explored bellow, is the result of this effort. The data set contains general information on some of TugasRWe customers, their buying habits and the results of a pilot marketing campaign performed in the past.

The present analysis has 3 main objectives:

- Obtaining a general description of the data set
- Assessing the quality of the data collected
- Investigating the success of the pilot marketing campaign

2 Tugas' general description

The dataset has 2500 observations and 21 variables (excluding the variable with the costumer ID). Table 1 displays the meta-data of the 6 categorical variables in the dataset, while table 2 describes the remaining 15 interval variables.

Tab. 1: Categorical variables' description

Variable	Description
Gender	Customer's gender
Education	Customer's education
Marital_Status	Customer's marital status
Dependents	Customer's state of dependents (Yes=1)
Recomendation	Customer's recommendation, from 1 to 5
DepVar	Customer's response to campaign (positive=1)

Tab. 2: Interval variables' description

Variable	Description
Year_Birth	Customer's year of birth
Income	Customer's household income
Dt_Customer	Date when customer joined
Rcn	Customer's recency in days, in the last 18 months
Frq	Customer's number of purchases in the last 18 months
Mnt	Amount spent by customer in the last 18 months
Clothes	% spent by the customer on Clothes
Kitchen	% spent by the customer on Kitchen
SmallAppliances	% spent by the customer on Small appliances
Housekeeping	% spent by the customer on House keeping
Toys	% spent by the customer on Toys
NetPurchase	% purchases of customer through the internet
CatPurchase	% purchases of customer through the catalogue
CostPerContact	Campaign's cost of contact
RevenuePerPositiveAnswer	Campaign's net revenue for a positive answer

To aid the analysis, a simple transformation to two interval variables was applied:

- Dt_Customer was exchanged by CustMonths, which represents the number of months since the customer joined the platform (assuming as the current month March 2018).
- Year_Birth was replaced by the corresponding customer's age in years.

2.1 Categorical variables

In order to have a first view on the categorical variables, we retrieved the number of classes and the mode of each variable. Table 3 summarises this information.

Tab. 3: Summary statistics of categorical variables

Variable	Number of classes	Mode
Gender	3	F
Education	7	Graduation
Marital_Status	7	Married
Dependents	2	1
Recomendation	6	4
DepVar	2	0

Dependents, Gender and DepVar are binary variables. From all customers in the dataset, 70.1% have dependents and 29.9% don't have any dependents. Gender shows 3 different classes because there are 0.2% of the customers with the class "?", which clearly corresponds to missing values. The remaining 76.4% of the customers are females while 23.4% are males. DepVar is highly unbalanced since only 7% of the customers responding positively to the marketing campaign.

Education, Marital_Status and Recomendation have more than 2 classes. The three frequency plots in figure 1 illustrate their distributions and we can conclude from it that all three variables present some issues. Education has 0.3% of missing values and 0.2% of a class labelled "OldSchool", which does not seem to be a type of education and most probably is an error. Marital_Status shows similar issues, with

0.7% of missing values and 0.8% of a class names "BigConfusion". Lastly, the variable Recomendation, which should be a rating between 1 and 5, has 7.7% of customers in a class named "6". This is either an error or a problem with the variables' meta-data.

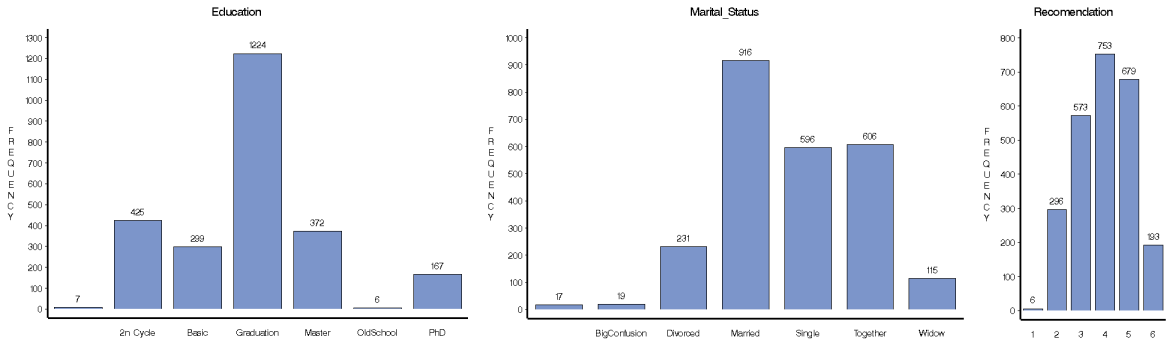

Fig. 1: Frequency plots for the variables Education, Marital_Status and Recomendation

In conclusion, all categorical variables are good enough to use in further analysis. in spite of most variables having missing values and/or errors, the weight of those issues in the dataset is not significant. The only exception is the variable Recomendation - there is a considerable percentage of customers giving a rating of 6, which is inconsistent with the variable's meta-data.

2.2 Interval variables

For the interval variables, 5 summary statistics were computed - mean, standard deviation, minimum, maximum and range. Table 4 displays these statistics for all interval variables.

Tab. 4: Summary statistics of interval variables

Variable	Mean	Standard deviation	Minimum	Maximum	Range
CustMonths	144	167	114	172	58
Income	74062.84	28807.23	10500	144204.9	133704.9
Age	51	17	21	81	60
Rcn	63	70	0	549	549
Frq	20	115	3	58	55
Mnt	654.83	676.45	8.32	3055.52	3047.2
Clothes	51.0%	23.4%	2%	98%	96 p.p.
Kitchen	7.0%	7.9%	0%	68%	68 p.p.
SmallAppliances	28.2%	12.6%	2%	69%	67 p.p.
Housekeeping	6.9%	7.6%	0%	54%	54 p.p.
Toys	6.9%	7.9%	0%	77%	77 p.p.
NetPurchase	42.1%	18.3%	5%	88%	83 p.p.
CatPurchase	57.9%	18.3%	12%	95%	83 p.p.
CostPerContact	2	0	2	2	0
RevenuePerPositiveAnswer	15	0	15	15	0

Firstly, we note that all variables seem to have values consistent with their meta-data. For instance, the variables defined as percentages do not go bellow 0 nor beyond 100 and the variable Age varies between 21 and 81. Secondly, there are two constant variables - CostPerContact is always 2 and RevenuePerPositiveAnswer is always 15. Thus, they don't need to be further explored. Thirdly, Income is

the only interval variable with missing values, representing 2.8% of the dataset.

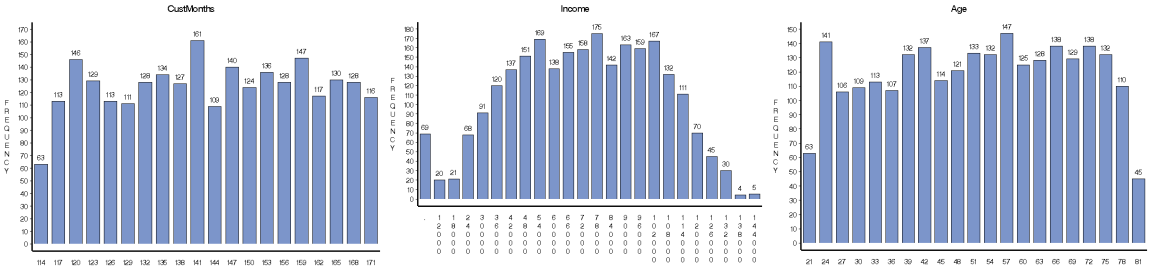

Fig. 2: Histograms for the variables CustMonths, Income and Age

The first three interval variables relate with customers' personal information. From figure 2, which depicts the histograms of these variables, we observe that CustMonths and Age have an almost uniform distribution and Income has a distribution closed to a Gaussian shape.

It is also interesting to note that the customers in this dataset are neither recent nor too old, since the variable CustMonths varies between 9.5 years and 14.25 years. Another interesting point is the Age distribution as one would not expect an uniform distribution. In fact, the frequency of customers over 60 is astonishing when compared with the frequency of young adults and middle-aged people.

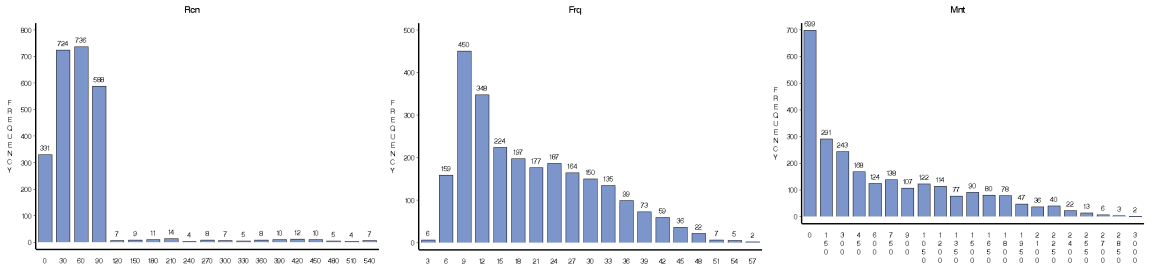

Fig. 3: Histograms for the variables Rcn, Frq and Mnt

The next three variables (Rcn, Frq, Mnt) are widely used in a priori grouping (RFM analysis) since they are usually good predictors of future purchases. Figure 3 includes a histogram for each of these variables, from which we can conclude that all three variables are right-skewed. This means that most customers have low values in all 3 variables. In other words, most customers have bought recently but tend to make few purchases and spend lower amounts.

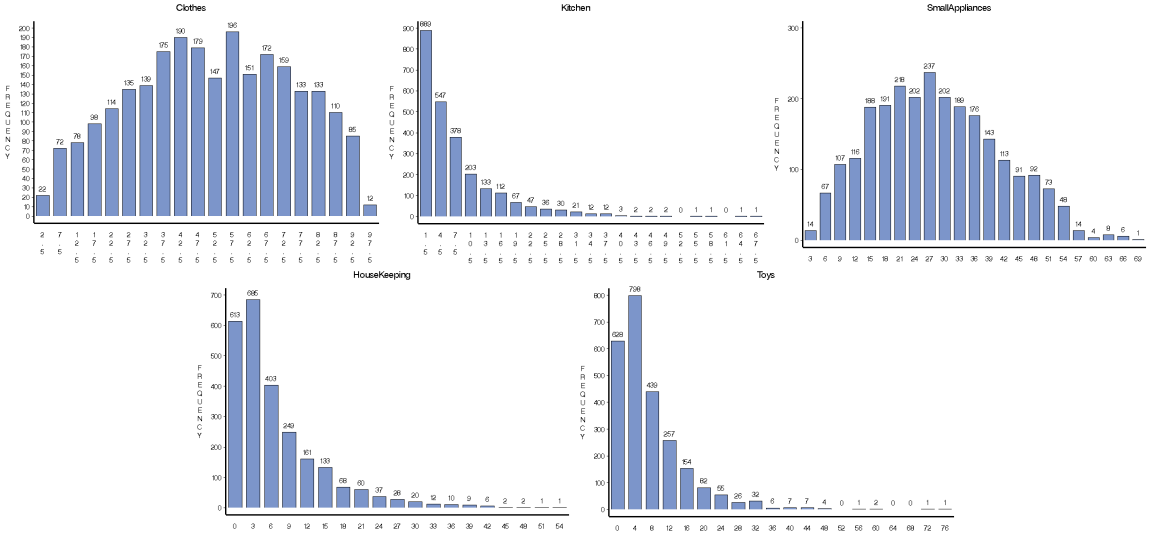

Fig. 4: Histograms for the variables Clothes, Kitchen, SmallAppliances, Housekeeping and Toys

Figure 4 represents the histograms of the first set of interval variables related with the buying patterns of TugasRWe customers.These variables express the percentage of money spent in each department of the store and, thus, for each customer, they add up to 100%.

Kitchen, HouseKeeping and Toys have right-skewed distributions. On the other hand, Clothes and SmallAppliances have more symmetric distributions and are the departments with higher percentages of spending.

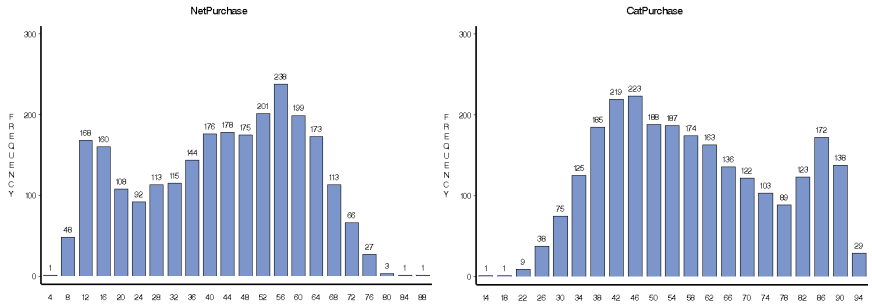

Fig. 5: Histograms for the variables NetPurchase and catPurchase

Finally, the variables NetPurchase and CatPurchase also add-up to 100% and, because of that, these variables have a strong correlation. From a quick review of figure 5, we conclude that both distributions are bi-modal, with the first mode representing the customers that buy similar amounts through the internet as through the catalogue and the second mode representing the customers that buy almost exclusively through the catalogue.

In conclusion, the set of interval variables also has a good enough quality for further analysis. Only one variable has missing values, all variables are consistent with their meta-data and, in most cases, there are not clear outliers.

3 Campaign results

From the variable DepVar, we observed that 175 customers responded positively to the pilot marketing campaign, which represents a rate of success of 7%. Given that the revenue per positive answer was 15 euros, the campaign amounted to a total revenue of $15 \times 175 = 2625$ euros.

However, we know that all 2500 customers were contacted, which resulted in a total cost of $2 \times 2500 = 5000$ euros. Therefore, the campaign did not result in a profit for TugasRWe. Instead, it led to a loss of 2375 euros.

In order to have a profitable marketing campaign, and assuming that the cost per contact and the revenue per positive answer remained the same, the campaign would need to have a rate of success of at least $\frac{2}{15} = 13.3(3)\%$.

4 Annex - SAS diagram