Data mining project report

Group 1:

Maddalena Amendola

Daniele Gadler

Riccardo Manetti

Gemma Martini

October 17, 2018

Contents

1 Data Understanding

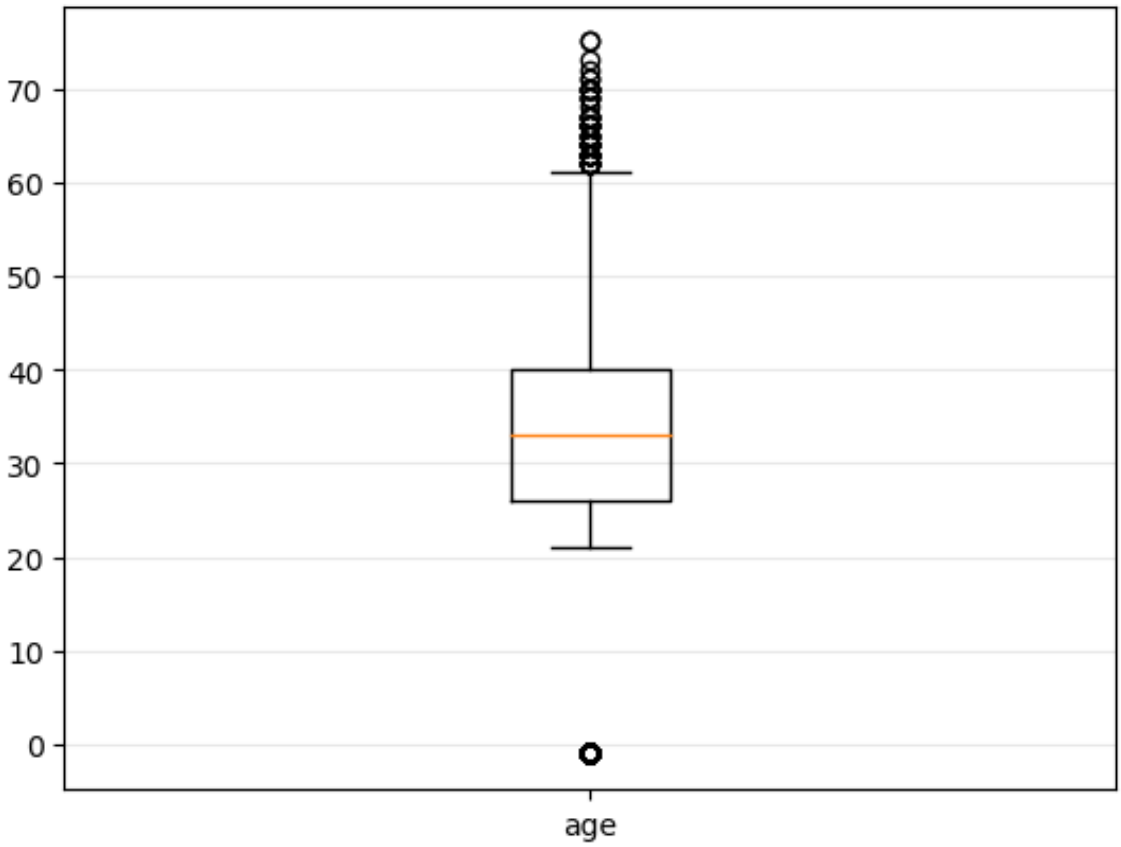
This part contains a first analysis of the data contained in the Taiwan credit card dataset. The collected data takes into account users movements on their bank accounts from April 2005 to September 2005.

1.1 Outliers' Identification

These boxplots are created with Python and give us the chance to understand the shape of the given dataset, in order to find out possible outlayers, which may give ah hint on the many characteristics of the data.

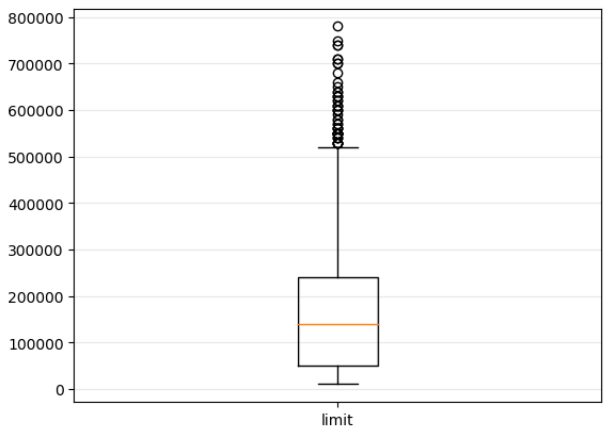
1.1.1 Age

From this plot we can observe that age has some values below 0. We can assume that the values allowed for this field are strictly positive.



1.1.2 Limit

Since the dataset takes into account Taiwanese dollars ($\sim 0,028$ €) plotting the boxplot in terms of multiples of the average Taiwanese income (NTD 49989, about USD 1700) is more informative.



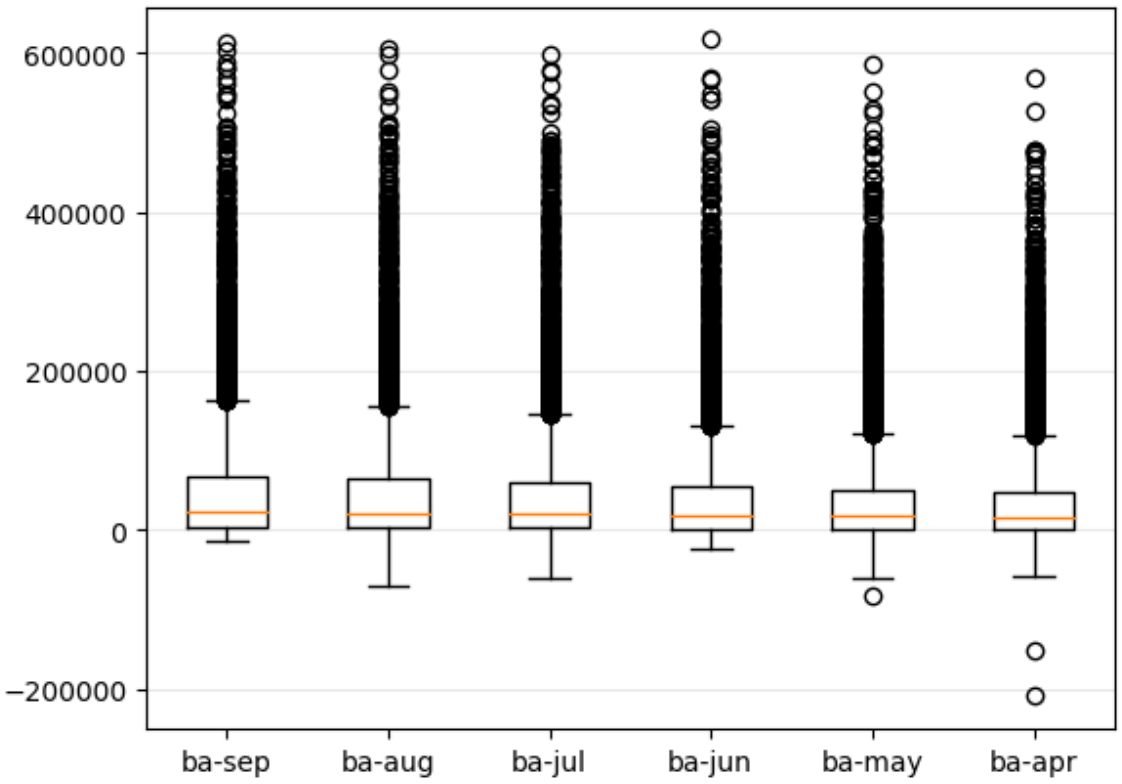
Daniele: plot limitS.png is missing

From this plot we may observe that on average the limit of the credit card plafond is 2.5 times the average income.

1.1.3 Billing amount

This values represents the balance of the bank account during the previous billing cycle. I can't understand what all these outlayers mean!!!

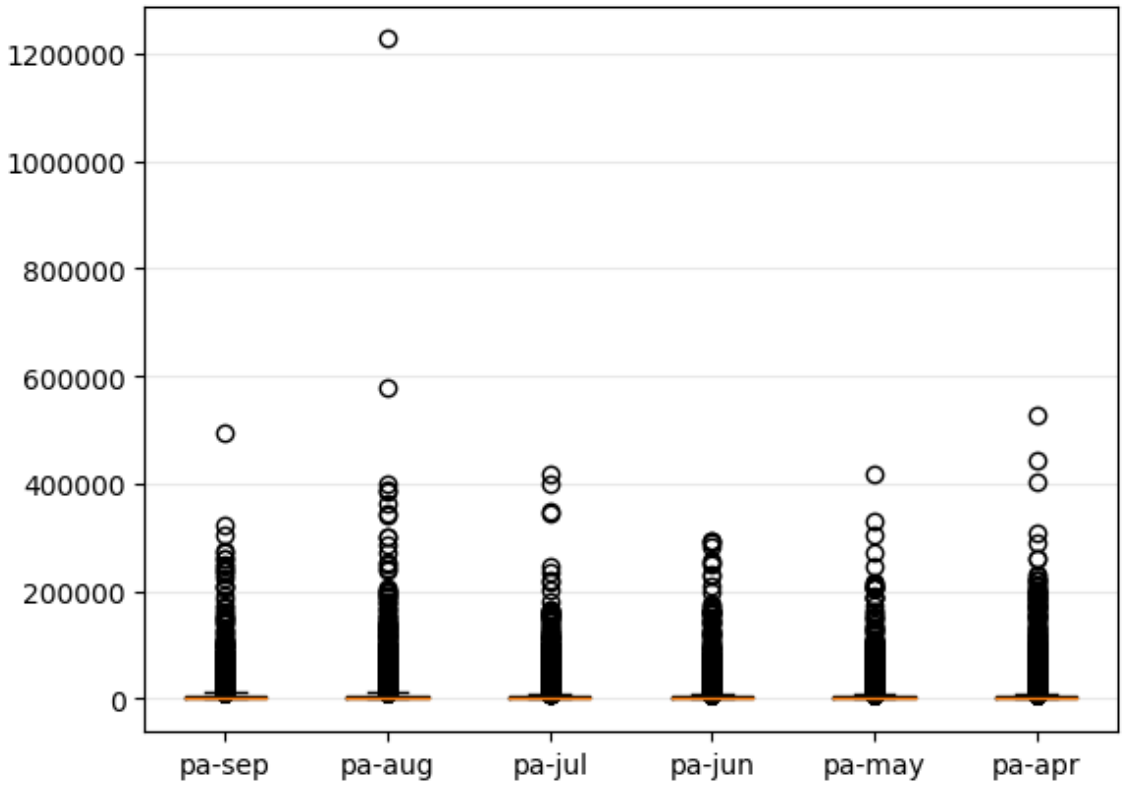
I can't understand



1.1.4 Amount of previous payments

This values tell us how much the user spent each month in the past 6 months, on a monthly basis.

Non capisco :-C



1.1.5 Payment status

This columns of the database contain the history of payments, more precisely:

Payed in time → value −1

Paid one month later → value 1

Paid two months later → value 2

Paid three months later → value 3

Paid four months later → value 4

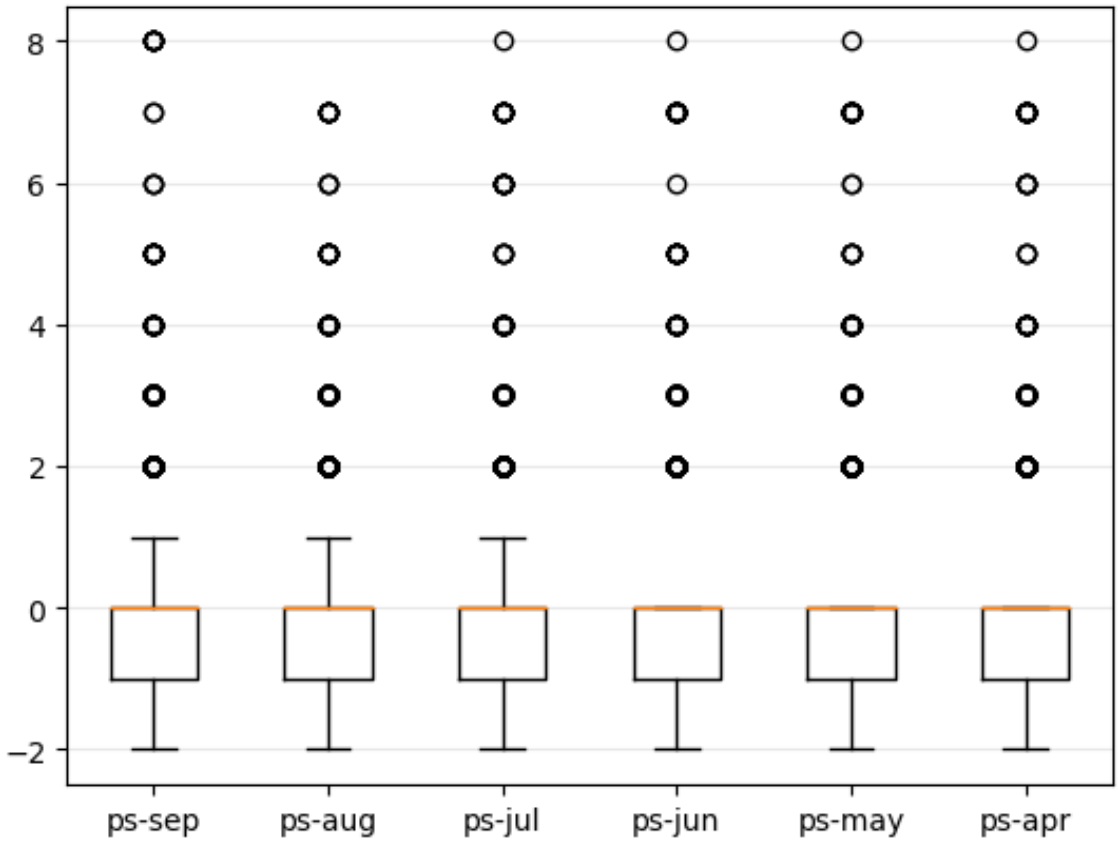
Paid five months later → value 5

Paid six months later → value 6

Paid seven months later → value 7

Paid eight months later → value 8

Paid nine months later or more → value 9



Direi che questo lo togliamo . . . ???

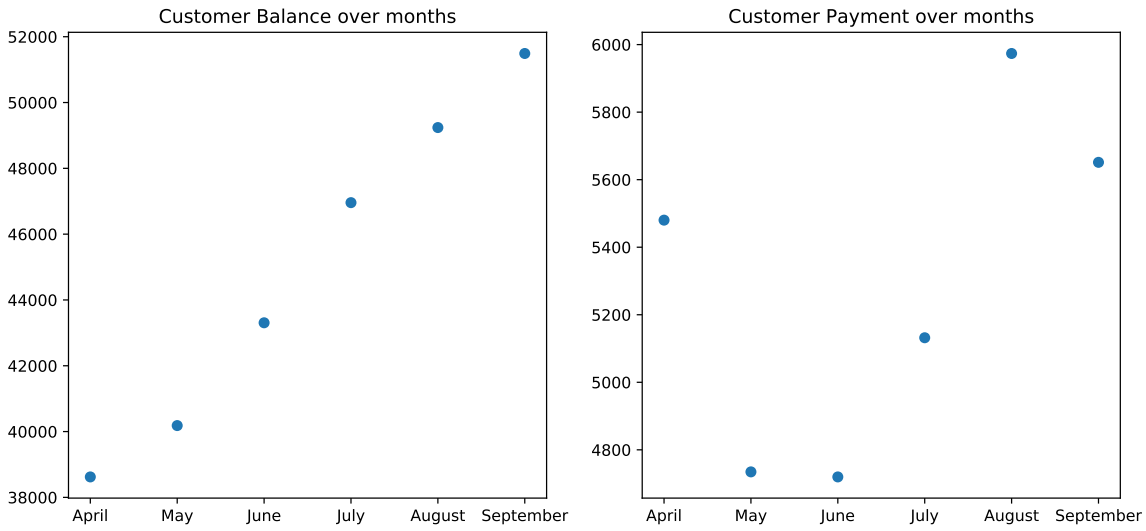


Figure 1: Plot of the average customer balance and payments over the April, May, June, July, August and September months

1.2 Context-specific Dependencies

To achieve a deeper insight into the dataset under study, let's first analyze some basic statistics and correlations about the dataset by analyzing some interesting properties of the dataset.

1.2.1 Customer Balance and Payments over months

1. **Overall Trend**

   As it is expected from a bank account dataset, the customer balance tends to increase over the months. On the other hand, the amount spent by customers is not fixed over time: it is maximum during August and September - because of an increased spending, probably on holidays - and is minimum during May and June, simple working months in Taiwan.

   Correlation between customer balance and payments: As the plotted attributes in Figure seem to be characterized by different growth patterns, the correlation between the average customer balance and the average payment is just 0.355, and thus does not suggest a strong correlation.

2. **Age**

   As it is expected, older people generally have a higher balance than young people: also, people in their 60s tend to spend more than young or middle-aged people.

3. **Education**

   Generally, people with a higher education tend to earn more than people with a lower education level and consequently enjoy a higher living standard in Taiwan. As Figure 3 shows, this fact reflects on an increased spending and increased balance availability for people that went to graduate school or university wrt. people who just attained a high school certificate.

   Add references supporting "generally"

4. **Credit Default**

   As by Figure 4, there seems to be a rather strong correlation between customer payments and a customer default. In fact, the mean value of the average customer payment over all considered months of customers who defaulted is approx. half of the average customer payment over all considered months of customers that did not default. We can hence conclude that a low customer revenue is a key factor for understanding whether a customer is going to default.

   On the other hand, the balance doesn't seem to be very strongly correlated with defaults.

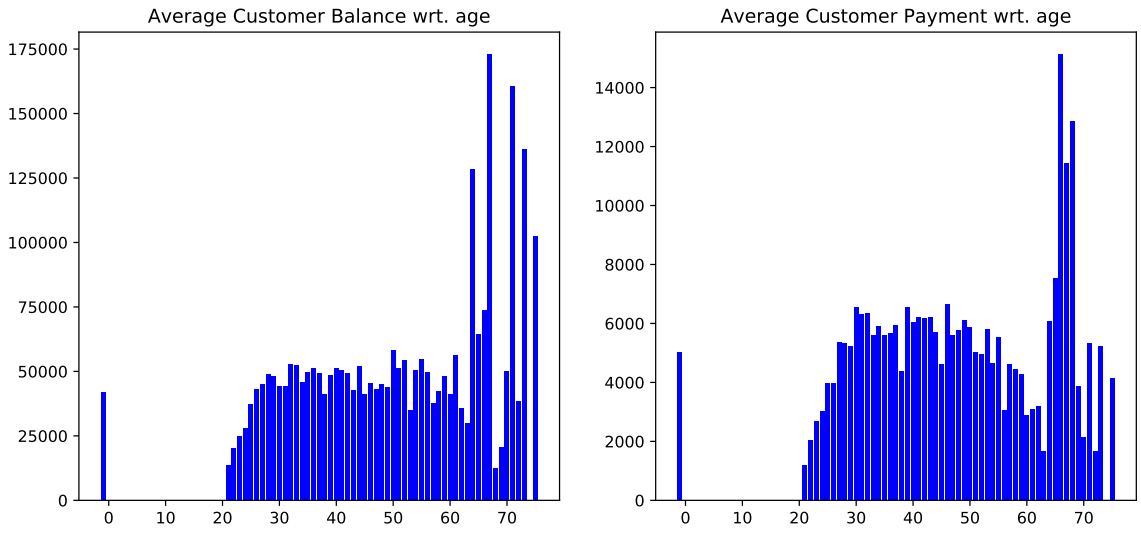


Figure 2: Plot of the average customer balance and payments over all considered months, in comparison with the age

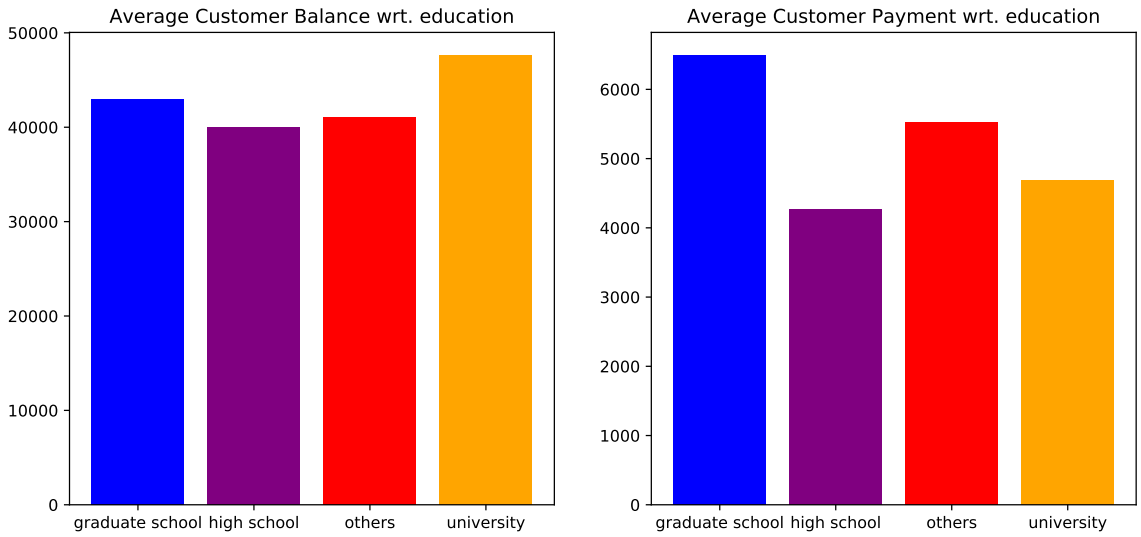


Figure 3: Plot of the average customer balance and payments over all considered months, in comparison with the education

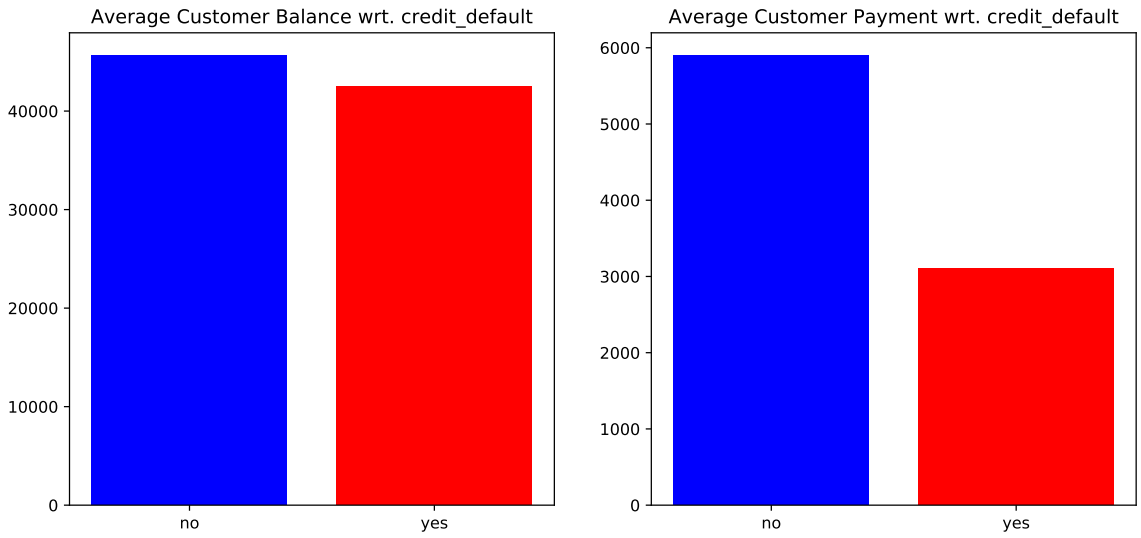
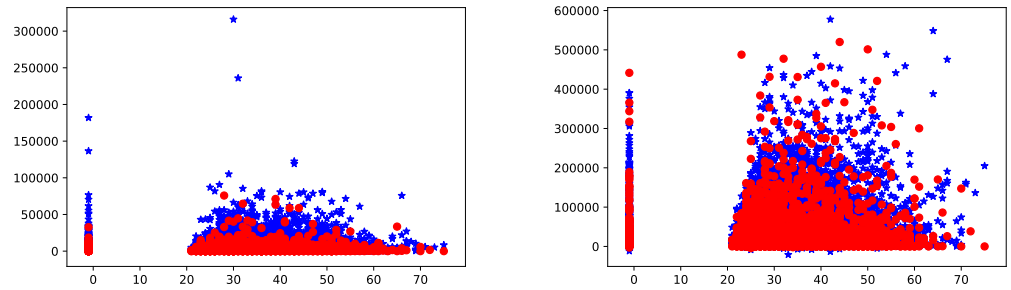


Figure 4: Plot of the average customer balance and payments over all considered months, in comparison with the education



(a) Scatter plot of the average customer payments over all considered months wrt. the age

(b) Scatter plot of the average customer balance over all considered months wrt. the age

Figure 5: Scatter plots of the customer balance and payments wrt. the age

This fact is further highlighted by the scatter plot in Figure 5a: we can see that defaults occur over all ages, however these typically occur to people with little payments' amounts. Instead, from Figure **??**, we can see that defaults are not particularly correlated to age or balance.

re-generate plots when people with age = 0 have been removed

1.3 Bank Account holders' count

We now proceed to analyze the count of bank account holders wrt. categorical attributes via bar plots.

1. **Credit Defaults, Status, Gender, Education** As it can be seen in Figure 6, most of the customers did not default. They are similarly distributed among single and married. Most of them are female, and are generally highly educated (university, graduate school).

   The prototypical customer is hence: not defaulted, female, single, university educated.

And now we analyze the count of bank account holders wrt. continuous attributes via histograms in Figure 7. We can clearly see that most of the customers are young, and in the range 20-30. The more the age increases, the fewer customers there are. The balance is also mostly up to 100000, and the limit lies in this range.

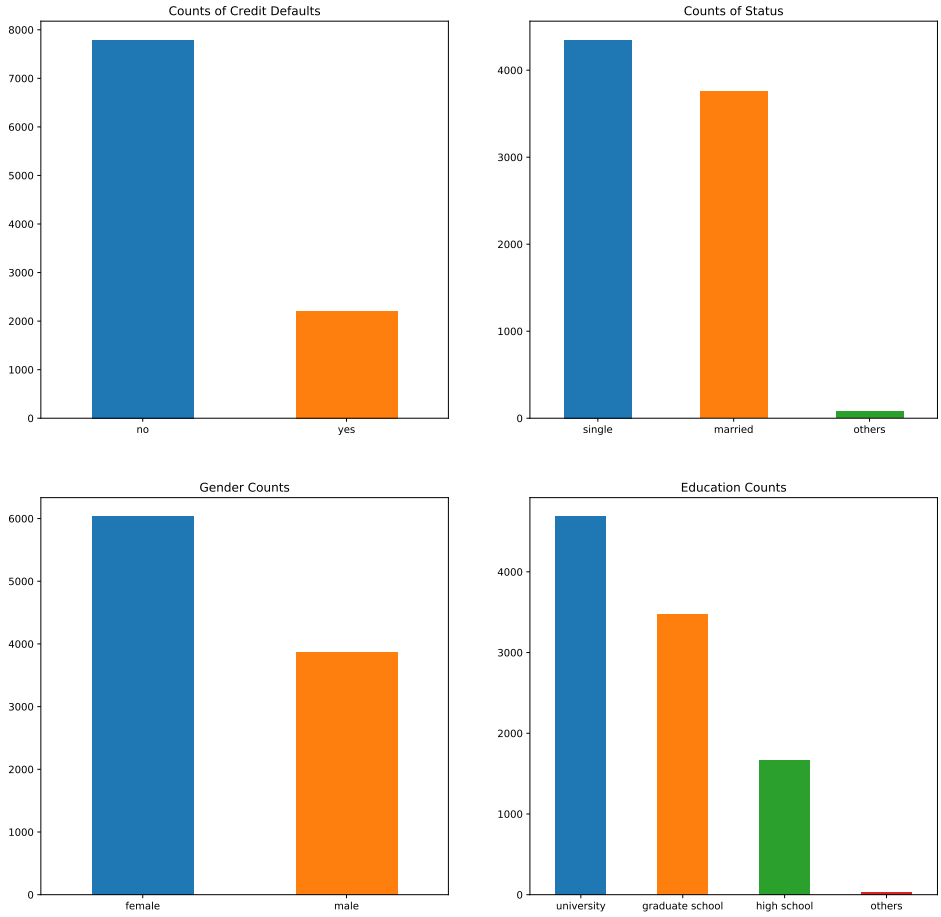


Figure 6: Plots of the amount of occurrences for the categories of discrete attributes: credit defaults, status, gender and education

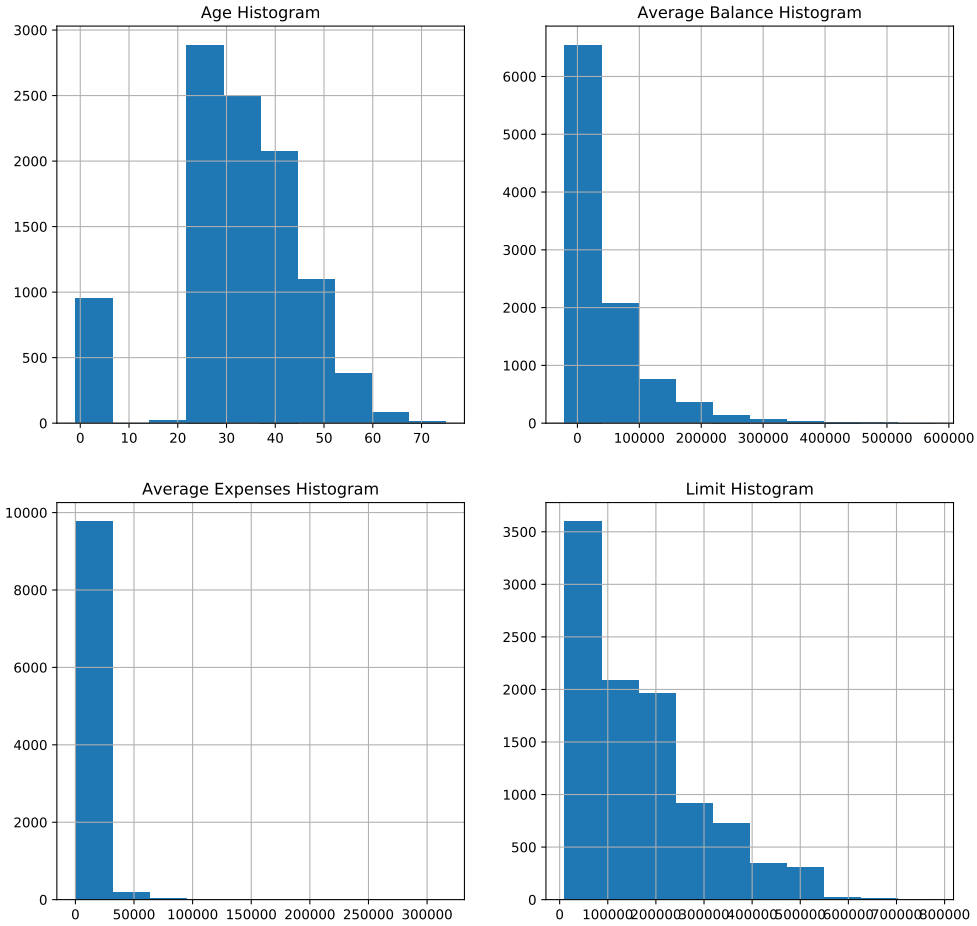
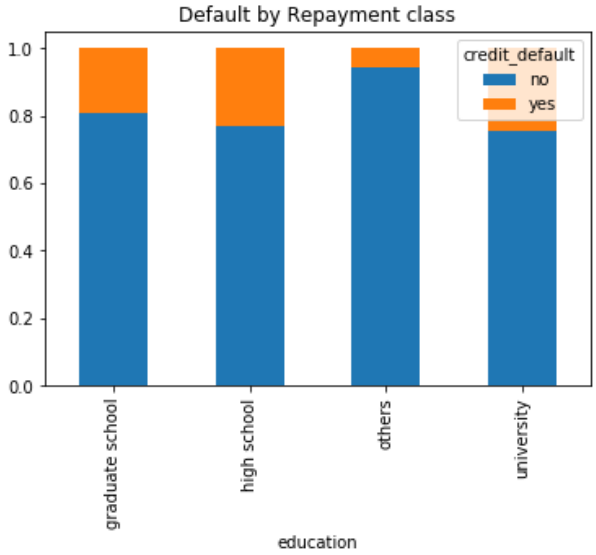
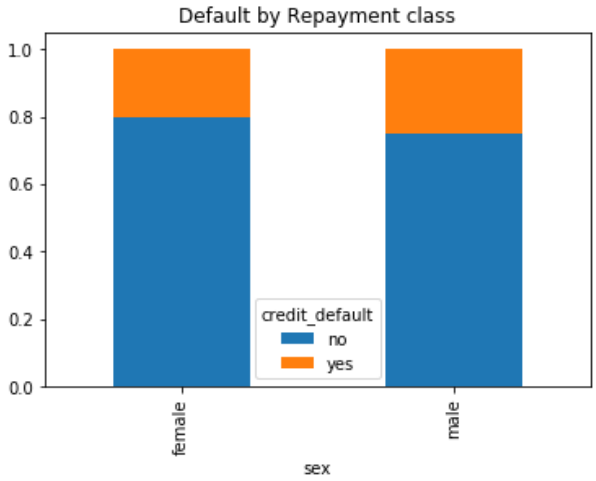


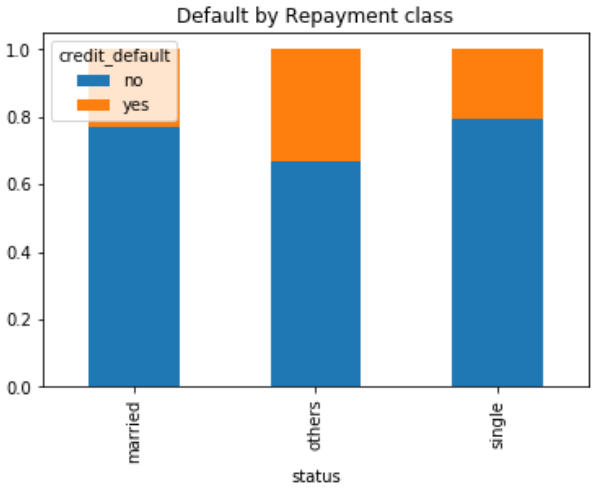
Figure 7: Plots of the amount of occurrences for continuous attributes: age, balance, expenses and limit



(a) Network 1



(b) Network 2



(c) Network 3

Figure 8: Default's occurrence wrt. education, sex, status

2. **Age, Balance, Expenses, Limit**

1.4 Credit Default Analysis

We now proceed to focus our analysis on the relation of the "credit_default" attribute wrt. different other attributes, in order to understand which correlations leading to a credit default subsist in the data.

In Figure 8a, we can see that bank account holders with an "others" education appear to be least likely to default than the other categories. From Figure 8b, we see that the gender is not a very significant factor that signalises a possible default, whereas from Figure 8c, we recognize that people in the "others" status were more likely to default than married or single people.

1.4.1 Ps-analysis

From Figure **??**, we can see that the higher the ps, the higher is the probability of a default to occur. This behaviour is typical of all the "ps" attributes in the considered dataset. This hence suggests that the "ps" is strongly correlated with the occurrence of a default.

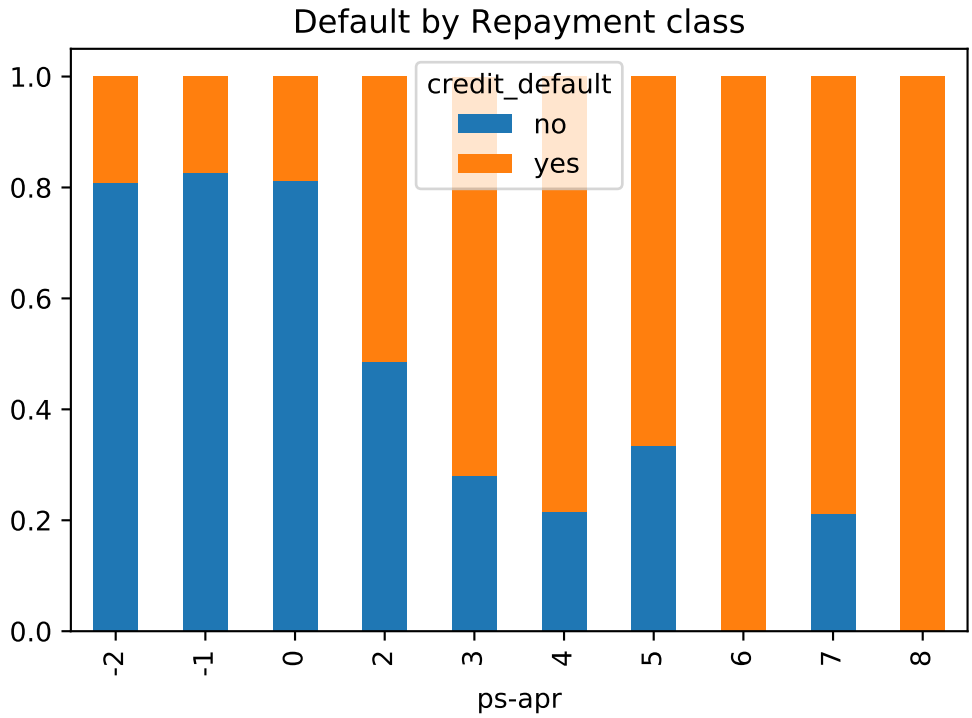


Figure 9: Plot of the credit_default occurrence by repayment class (ps)

2 Syntactic and Semantic Accurance

The goal of this analysis is to check if all the values of a specific attribute belong to the column domain and if there are typing error. The following table shows for each attribute its type and all (or in part) its values:

Name	Type	Values
limit	int	10000, 20000, 30000, 40000, . . . , 710000, 740000, 750000, 780000
sex	object	male, female, NaN
education	object	graduate school, university, high school, others, NaN
status	object	married, single, others, NaN
age	int	-1, 21, 22, 23, ..., 70, 71, 72, 73, 75
ps-sep	int	-2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8
ps-aug	int	-2, -1, 0, 1, 2, 3, 4, 5, 6, 7
ps-jul	int	-2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8
ps-jun	int	-2, -1, 0, 2, 3, 4, 5, 6, 7, 8
ps-may	int	-2, -1, 0, 2, 3, 4, 5, 6, 7, 8
ps-apr	int	-2, -1, 0, 2, 3, 4, 5, 6, 7, 8
ba-sep	int	-14386, -10682, -9802, ... , 588000, 604019, 613860
ba-aug	int	-69777, -67526, -30000, ... , 577681, 597793, 605943
ba-jul	int	-61506, -24702, -15910, ... , 577015, 578971, 597415
ba-jun	int	-24303, -15910, -15588, ... , 565669, 569034, 616836
ba-may	int	-81334, -61372, -37594, ... , 530672, 551702, 587067
ba-apr	int	-209051, -150953, -57060, ... , 478034, 527566, 568638
pa-sep	int	0, 1, 2, ... , 304815, 323014, 493358
pa-aug	int	0, 1, 2, ... , 401003, 580464, 1227082
pa-jul	int	0, 1, 2, ... , 349395, 400972, 417588
pa-jun	int	0, 1, 2, ... , 291227, 292462, 292962
pa-may	int	0, 1, 2, ... , 303512, 331788, 417990
pa-apr	int	0, 1, 2 , ... , 403500, 443001, 528666
credit_defaut	object	yes, no

From this table it's possible to notice the following problems:

1. The attributes 'Sex', 'Education' and 'Status' have a wrong value: Not a Number;
2. The attribute 'Age' has a wrong value that is '-1'.

All others attribute are semantically and syntactically correct!

References