Multivariate Data Analysis: Application to Financial Credit

Homework I

Iker Caballero & Daniel González

Multivariate Data Analysis - Prof. Dante Conti

1 Introduction

The main goal of this first homework is to analyze multivariate data for the "Credsco" dataset in order to perform a descriptive analysis of the data. This dataset shows data about 4455 credit applications to a bank and characteristics about both the clients that apply and the credit.

In the following sections, we proceed to present the implementation and results of the analysis for the dataset. We structure the first homework the following way: we first look at the dataset so that we can proceed to propose relevant topics and questions and report the data quality. We then begin the descriptive analysis, dividing the analysis by each question, and we finally present some concluding remarks.

2 The "Credsco" Dataset

In this first part, we are asked to formulate some relevant topics and questions based on the metadata and the description of the problem posed. Hence, it is logical to start by presenting the problem and then describing the data based on the metadata, so that we can logically derive some questions or topics that could be interesting.

In the "Credsco" description file, the problem proposed is to try to understand the difference between the clients whose application has not been accepted with the ones which have been. Hence, it is recommended to start preprocessing the data and then do a good description of the variables. The dataset includes 14 variables, from which 9 are numerical variables and 5 are categorical. A more precise classification of these variables is shown in Table 1.

Variables	**Type**	**Nature**
Dictamen	Categorical	Binomial
Antiguedad.trabajo	Numerical	Discrete
Vivienda	Categorical	Multinomial
Plazo	Numerical	Discrete
Edad	Numerical	Discrete
Estado.civil	Categorical	Multinomial
Registros	Categorical	Multinomial
Tipo.trabajos	Categorical	Multinomial
Gastos	Numerical	Discrete
Ingresos	Numerical	Discrete
Patrimonio	Numerical	Discrete
Cargas.patrimoniales	Numerical	Discrete
Importe.solicitado	Numerical	Discrete
Precio.del.bien.financiado	Numerical	Discrete

Table 1: Classification of variables

After tackling the relevant questions and topics, it is important to understand the data we are dealing with, as this will affect the kind of results we can get from our analysis. Hence, we will first expose the questions and then continue with a data quality report, which will help to understand and proceed better with the descriptive analysis.

2.1 Relevant Topics & Questions

We now state which are the most relevant concerns and topics that data can solve or at least inform us about. These ones are the following:

1. How are job types related to the acceptance or rejection of credit applications? Through analyzing this, we can get insights about whether more professionally stable people (such as workers with fixed

contracts) are prone to get more credit than other kinds of workers (such as autonomous or temporal), and what this mean for potential clients.

2. How are housing and civil status related to the acceptance or rejection of credit applications? By answering this question, we can see whether people with one specific type have a greater probability of being accepted than others, and hypothesize why this happens. Even though civil status might not be something relevant for a bank a priori, we can analyze with data whether some statuses could receive more accepted applications than others and give some insights about the facts.

3. How is net income related to the acceptance or rejection of credit applications? This is a very interesting question because we can obtain valuable information on the relevance banks give to income and expenditures through the net income, which is income minus expenditure, when conceding credit, surely related to the likeliness of payback.

4. Do banks concede more credit for higher-priced goods? This is the last and very relevant question, as we can get a sense of the risk aversion of banks.

All of these questions can provide a good understanding of the mechanisms inside a bank's decision to accept or reject an application of credit, so we proceed to the descriptive analysis of the data in order to get some insights.

2.2 Data Quality

From the different categories and values that these variables can take and the dataset, we can report the quality of the data we are working with.

At first, we can see that the metadata offers no further description of the variables, even though it is not necessary for some of the variables. However, it would be somehow necessary for variables such as "Vivienda" or "Tipo de Trabajo", where there is a category of others. Which categories or circumstances are included in the "others" category and why are they mixed together could shed some light on the interpretation of this category. There are also other categories that have very few individuals and are not common. Moreover, we can highlight something more obvious: the use of numbers instead of strings. Even though the codification is explained for each categorical variable, it might be clearer to use informative names for each category, so it is clear when doing exploratory analysis with the dataset.

When it comes to the dataset, we can see something interesting: the missing values are coded as "99999999" for numerical variables and as "0" for categorical variables. We note here that this coding for missing values can be misleading, as it could indicate a numerical value or a category (which is, in fact, not considered). Hence, it is important to change the coding of these values or eliminate these observations from the analysis. No other issues have been detected.

Due to our observations being consistent with the preprocessing done in the "CredscoClean" dataset, we use this dataset and do a further preprocessing of the data, focused on the categorical variables. As we stated above, categories of the relevant variables should be modified in order to obtain relevant insights. All of the modifications explained below can be found in Appendix I.

Starting with "Vivienda", we can see how there are different categories that could be grouped together and how others could be eliminated. Both "escriptura" and "contr_priv" refer to the kind of contract which is signed for buying the estate, but given that neither includes renting (like "lloguer"), we can group them in a bigger category called "comprada", referring if an individual has bought the house or not. And because there are only a few observations with no category specified or ignoring contract ("ignora_cont"), we prefer to eliminate these categories and the observations related.

We do something similar for "Estado.civil", as it divides the social status in different categories that could be grouped, such as "divorciat", "separat" or "vidu", as all three are individuals that had a partner in the past. We

merge them into a category called "tenien.par", referring to this fact, and we also change the name of the categories "casat" and "solter" to "amb.par" and "sense.par", respectively. We do also eliminate the "VivUnkown" category, as it does not provide any information for our analysis.

Finally, we eliminate the "WorkingTypeUnknown" of the "Tipo.trabajo" variable because of the same reasons stated previously. We now have our dataset preprocessed for our analysis and we now move on to the descriptive analysis.

3 Descriptive Analysis

To structure this analysis properly, we divide it into subsections that try to answer each question posed before. We summarize the result with figures and tables with the results, but we add the code needed for the graphs and computations in Appendix 1.

3.1 First Question

Our first question was about the relationship between the contract type of worker and the acceptance or rejection of an application for a credit loan. For analyzing this relation, we use both graphical and numerical methods.

We first take a look at the distribution of types of contracts in the sample. As we can see the principal type is fixed contracts, which account for 63.1% of the total sample, and temporal and autonomous workers account for the remaining 33.1% of the sample. Workers in other situations and unknown situation workers account for 3.9%, the latter category being virtually insignificant in the analysis. All of this information is summarized in Figure 1.

```
credsco$Tipo.trabajo
           Frequency Percent
altres sit       171     3.9
autonom         1016    23.0
fixe            2794    63.1
temporal         446    10.1
Total           4427   100.0
```

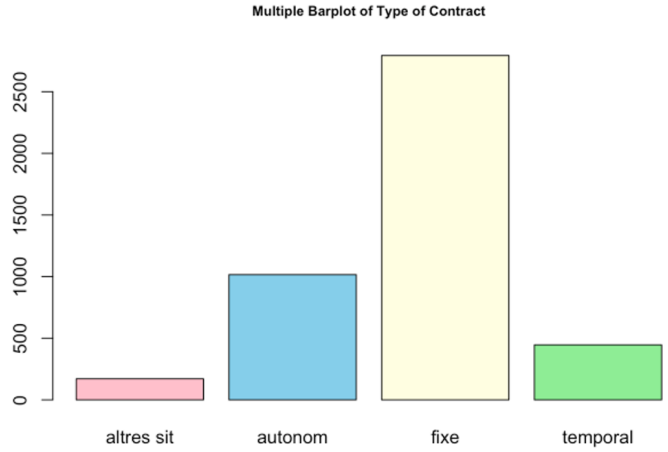


Figure 1: Frequency Table & Barplot for "Tipo.Trabajos"

Figure 2 shows all the information relevant to the joint distribution of the "Dictamen" and "Tipo.trabajo" variables. Because some categories are overrepresented, we use stacked bar plots which are grouped by the variable "Vivienda", so that one visualizes the percentage of acceptances in each category and not the percentage of the category in acceptances (which would be biased).

From the results obtained, one can see that more or less temporal workers seem to be the ones which are rejected the most, while fixed contract workers are the ones with more accepted applications. This might be due to the security or riskiness of the income flows that different workers have, or can also be because of the warrants or assets to back the credit (which would explain fixed and autonomous being more accepted). The majority of autonomous workers are also accepted, while the majority of "other situation" workers are not. A numerical analysis using frequency tables is also available for further information in Appendix I.

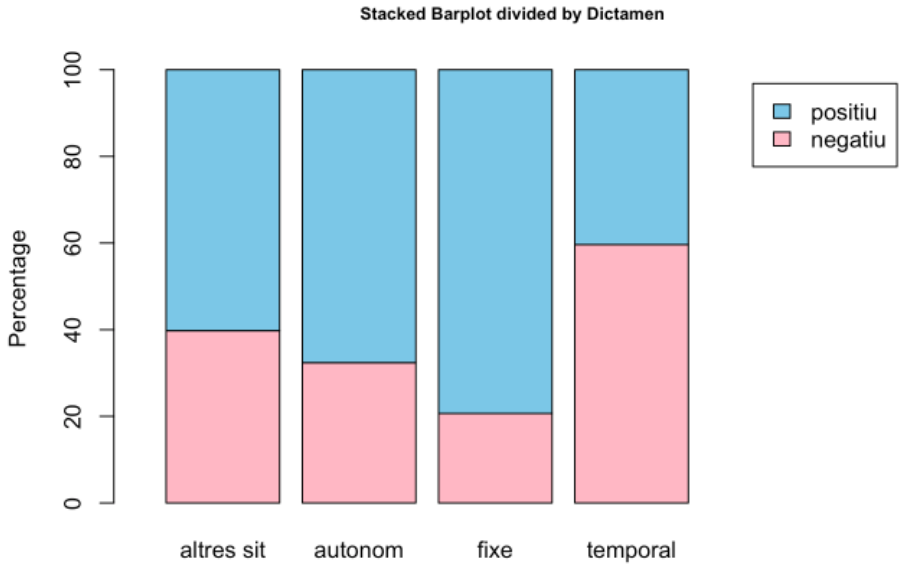


Figure 2: Stacked Barplot for "Tipo.trabajo" and "Dictamen"

In order to analyze the possible dependence between the decision and the type of contract, we have carried on a chi-square test (shown in Table 2 and available in Appendix I), which results in rejecting the independence hypothesis, meaning that there is dependence between these variables. This is a very logical result since credit screening is based on the confidence the bank puts on a client paying back, so it concedes credits to those who have a more stable professional situation like fixed workers. Autonomous workers also tend to get credit but it might be due to the fact that they have assets to back them as warrants.

Chi-Square Test	**Statistic**	**Degrees of Freedom**	**p-value**	**Conclusion**
Results:	317.46	3	0.000	Reject H_0 at 5%

Table 2: Chi-Square Test Results for "Dictamen" and "Tipo.trabajo"

All in all, we can conclude from this descriptive analysis that there is a relation between the type of worker a client is and the decision of the bank, being more stable (less risky) types (such as fixed workers) being the most preferred for lending. Banks could also consider clients in riskier positions with sufficient warrants (like autonomous) for securing payback.

3.2 Second Question

The second question is similar to the first one, but we focus on housing and civil status. Consequently, we will carry on a very similar analysis, using the same graphical and numerical tools, but diving a little deeper into the relationship between both variables.

We first take a look at the distribution of housing, civil status, and their joint distribution in the sample. Figures 3 and 4 show information about both variables in our sample. We can see how there is asymmetry regarding the observations per category in both, as coupled individuals account for 72.7% of the sample, and "comprada" is the principal category with 53.2% of the sample. However, the housing variable is more balanced than civil status.

We can now look at the joint distribution and try to detect patterns in the data than can help us. The majority of Single individuals, as expected, tend to live in their parent's house or rent, even though some own house. People that had a couple seem to be a more heterogeneous group, divided more or less equally between "comprada", "lloguer" and "pares". And finally, most of the people with a couple (married people) have bought their house. A

```
credsco$Estado.civil
           Frequency Percent
amb.par         3220    72.7
tenien.par       234     5.3
sens.par         973    22.0
Total           4427   100.0
```

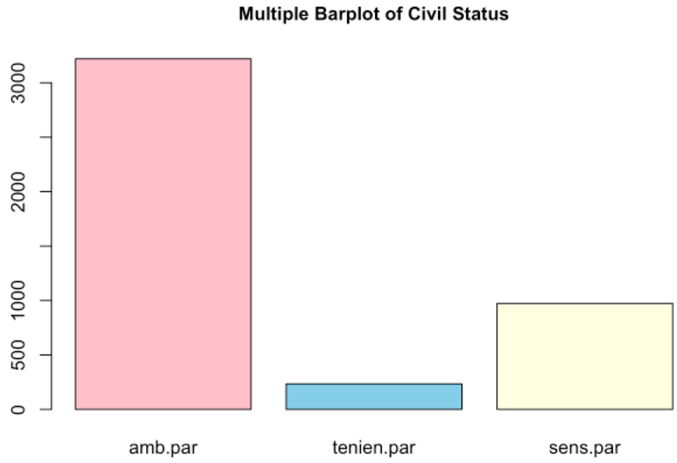


Figure 3: Frequency Table & Barplot for "Estado.civil"

```
credsco$Vivienda
           Frequency Percent
altres viv       319     7.2
comprada        2353    53.2
lloguer          973    22.0
pares            782    17.7
Total           4427   100.0
```

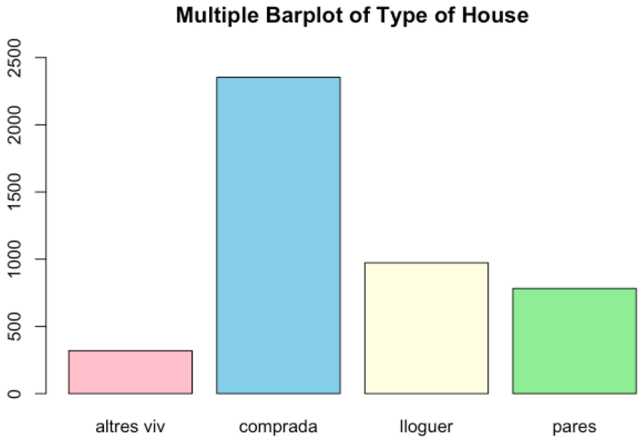


Figure 4: Frequency Table & Barplot for "Vivienda"

more detailed analysis of the joint distribution could be done using a crossed-frequency table, available in Appendix 1.

Given that there should be an obvious relationship between both variables, we carry on a chi-squared test, yielding results that support the dependence hypothesis (shown in Table 3).

Chi-Square Test	**Statistic**	**Degrees of Freedom**	**p-value**	**Conclusion**
Results:	1151.7	6	0.000	Reject H_0 at 5%

Table 3: Chi-Square Test Results for "Estado.civil" and "Vivienda"

Finally, we can analyze the joint distribution of "Dictamen" with both variables, so we can make some hypotheses with the given data. In Figure 6 we visualize the joint distributions graphically. Because of the overrepresentation of some categories, we apply the same ideas as before and use stacked bar plots divided by the categories of the variable and not by "Dictamen".

We can see how people who are coupled and have bought a house are the ones with more accepted applications. This might be due to the security that this profile gives to the bank, as we mentioned in the case of the job type so

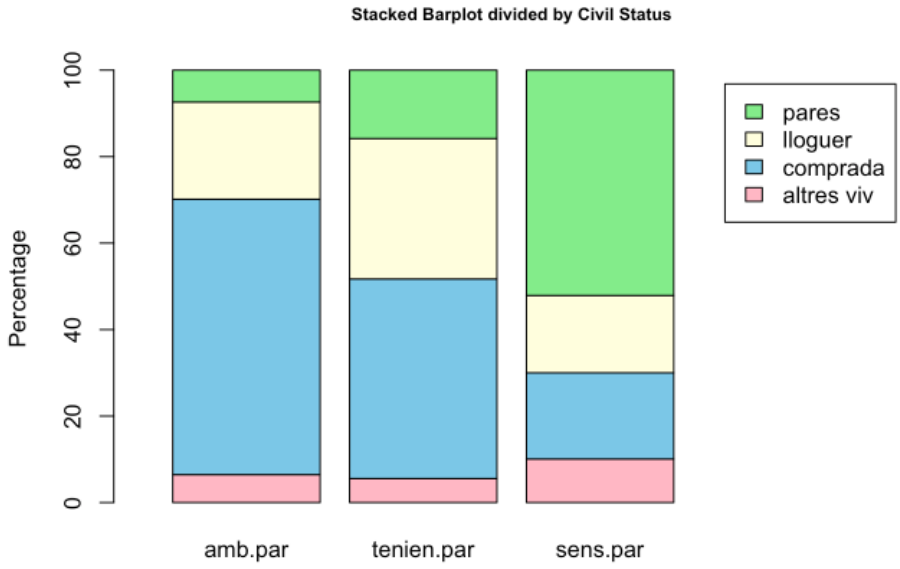


Figure 5: Stacked Barplot for "Estado.civil" and "Vivienda"

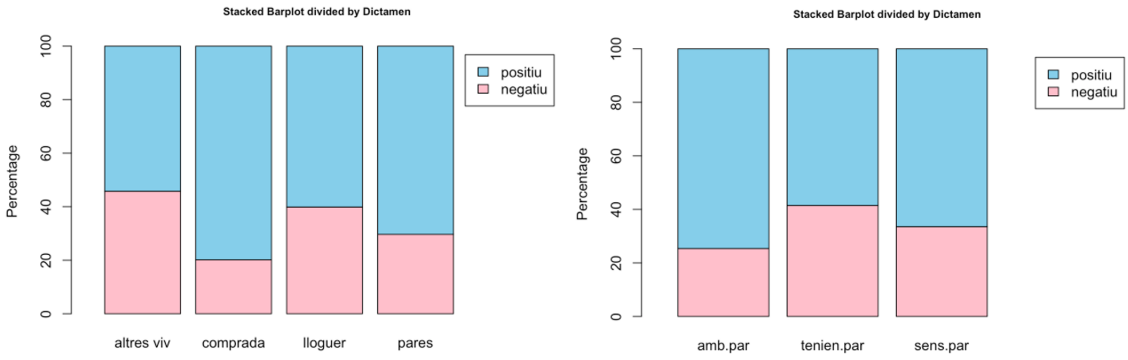


Figure 6: Stacked Barplots for the variables with "Dictamen"

that payback is more likely. When it comes to other important categories such as renting and without a couple, we can see that the percentage of accepted applications is greater than the percentage of rejected applications, but are lower than for individuals with a couple. For other categories, the percentages are more or less the same, so they are inconclusive.

If we check the dependence between these pairs of variables we obtain the same result as before: we reject independence. Hence, we can see that both variables are important when conceding credit.

Chi-Square Test	**Statistic**	**Degrees of Freedom**	**p-value**	**Conclusion**
Vivienda:	46.808	2	0.000	Reject H_0 at 5%
Estado.civil:	190.88	3	0.000	Reject H_0 at 5%

Table 4: Chi-Square Test Results for "Vivienda", "Estado.civil" and "Dictamen"

3.3 Third Question

In this question, we want to determine the relationship between net income with the decision made on applications for credit. For this question, we will first look at income and expenditure separately and then look at net income

and its relation with "Dictamen".

To study the variables separately, we use both numerical and graphical tools. At first, we can detect a great number of outliers that bias the distributions and other characteristics. In this case, the income variable has more outliers than the expenses one. These facts are represented in the histograms, which are highly skewed to the right.

In order to deal with the problems that these outliers would cause in our analysis, we eliminate some of the outliers and plot again the graphs in Figure 7. Now, it is easily observable that the distribution of income is more even, but one of the expenses remains skewed to the right. Nevertheless, the presence of outliers is now solved and we can proceed with the analysis.

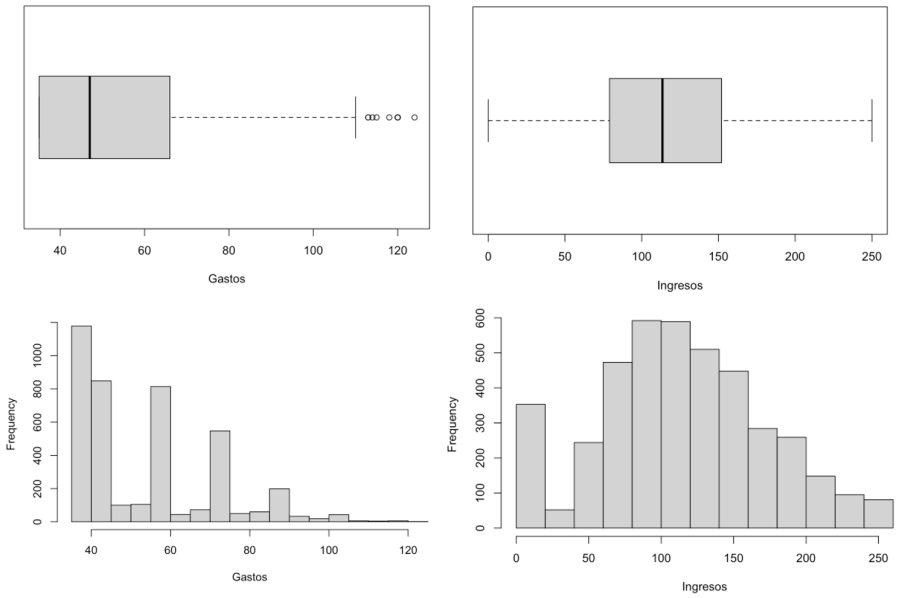


Figure 7: Boxplots & Histograms for "Ingresos" & "Gastos" without outliers

For analyzing the relationship between both variables we use the correlation matrix in Figure 8, where we can see that the positive relationship is not large. This makes sense because variables that determine income might not be too correlated with the ones that determine expenditure. For example, civil status and housing can determine expenditure, but it has little relation with the income an individual receives from his job (depends on the industry, the competitors, etc.).

	Gastos	Ingresos
Gastos	1.0000000	0.1911186
Ingresos	0.1911186	1.0000000

Figure 8: Correlation matrix for "Ingresos" & "Gastos"

Now, we can create the net income variable, which is the income minus expenditure for each individual. The hypothesis is that individuals with greater net income should be more likely to have their application accepted than others with less net income. We can check if that was the case in our sample. We present a similar analysis in Figures 9 & 10.

Here, we can still see some outliers, where some people have a very high net income and some others have negative net income (they spend more than they earn), but we can also observe how most of the observations have net income between 0 and 100 thousands of monetary units. The distribution, as one can see, is approximately normal for net income.

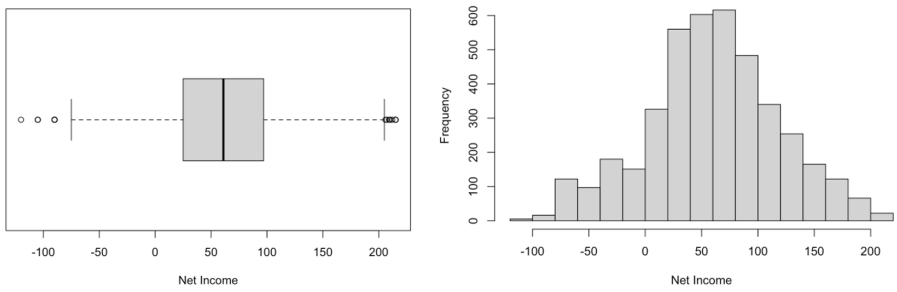


Figure 9: Boxplots & Histograms for "Net Income"

	n	mean	sd	median	min	max	range	skew	kurtosis
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	4128	60.73571	58.48346	61	-120	215	335	-0.07704517	-0.06451958

Figure 10: Summary statistics for "Net Income"

When studying the net income for the individuals with accepted applications and for those whose applications have been rejected (through Figure 11), we can observe that the variability of the income is similar, but the median values differ, indicating that the median individual whose credit has been conceded has a higher net income than those who have been rejected.

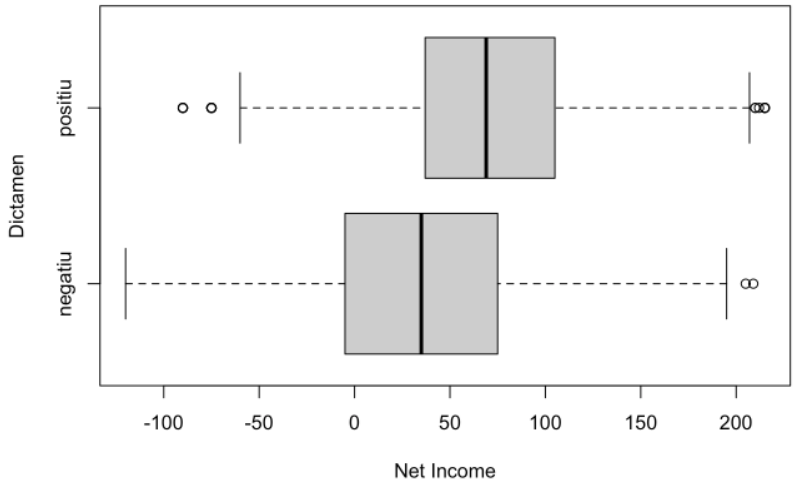


Figure 11: Boxplot for "Net Income" divided by "Dictamen"

We can test to know whether this happens too with the average individual of both groups (then, we now care about the average). By computing a hypothesis test for the comparison of means of independent groups, we can see how the means differ significantly, so we conclude that individuals who have been positively evaluated have a larger mean value of net income than the other group (in the sample).

Unpaired T-test	**Statistic**	**Degrees of Freedom**	**p-value**	**Conclusion**
Results:	17.423	2018.6	0.000	Reject H_0 at 5%

Table 5: Unpaired T-test Results for "Net Income" divided by "Dictamen"

To sum up, we can see how individuals with higher net income have had more positive evaluations than those with lower net income. This confirms our hypothesis, which is based on the economic reasoning of net income being

related to savings and to the likeliness of paying back debt.

3.4 Fourth Question

Our last question is whether banks concede more credit if the price of the good to be financed is higher. The analysis we carry on in this final section is more or less the same as before, so we start with the univariate analysis and then we see its relation to "Dictamen".

When looking at the summary statistics and the graphs of the original data, we can see that most of the prices (in thousands of monetary units) are concentrated approximately between 1000 and 1800 (with a mean equal to 1462), but the variability of these prices is high. This is caused by the presence of outliers on the left (goods with a relatively lower price) and on the right (goods with a relatively higher price).

Therefore, we eliminate these outliers (observations with more than 2500 units) and we obtain a more symmetric distribution and a clearer insight into the variability and the median value of the price of financed goods (Figure 12 & 13).

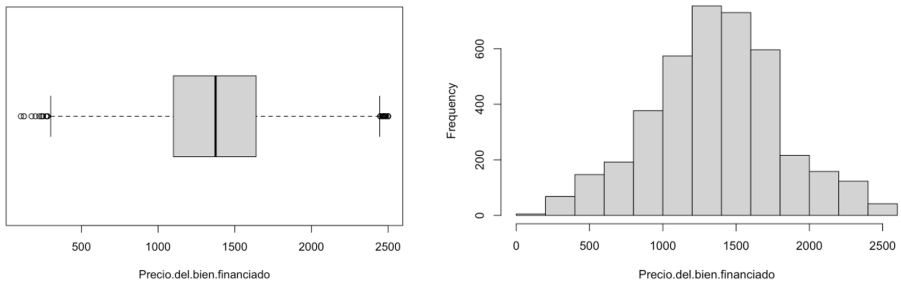


Figure 12: Boxplots for "Precio.del.bien.financiado" without outliers

	n	mean	sd	median	min	max	range	skew	kurtosis
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	3982	1368.734	437.6218	1375	105	2500	2395	-0.005729741	0.01536374

Figure 13: Summary statistics for "Precio.del.bien.financiado" without outliers

When dividing the prices by "Dictamen" we can see in Figure 14 that the distributions of both groups are not very different, having more or less the same median and dispersion. This could mean that, in our sample, the median prices of both groups do not seem different, so the rejection or acceptance of credit applications might put higher weight on relevant factors other than the price of the good to finance by the credit.

To check if this is also true for the average individual of both groups, we apply the same hypothesis testing methodology as before, and we can check that the result is that difference in means is not significant, so we cannot reject the hypothesis of both means being the same.

Unpaired T-test	Statistic	Degrees of Freedom	p-value	Conclusion
Results:	1.6852	1865.5	0.09212	Not reject H_0 at 5%

Table 5: Unpaired T-test Results for "Precio.del.bien.financiado" divided by "Dictamen"

The descriptive analysis, then, shows that the weight of this variable to accept or reject an application is not relevant, and banks look for other socio-economic factors that are more relevant. Hence, the risk aversion of the banks cannot be measured with this variable because there are no significant differences between accepted and rejected applications and we need to look for other factors (such as the previous ones).

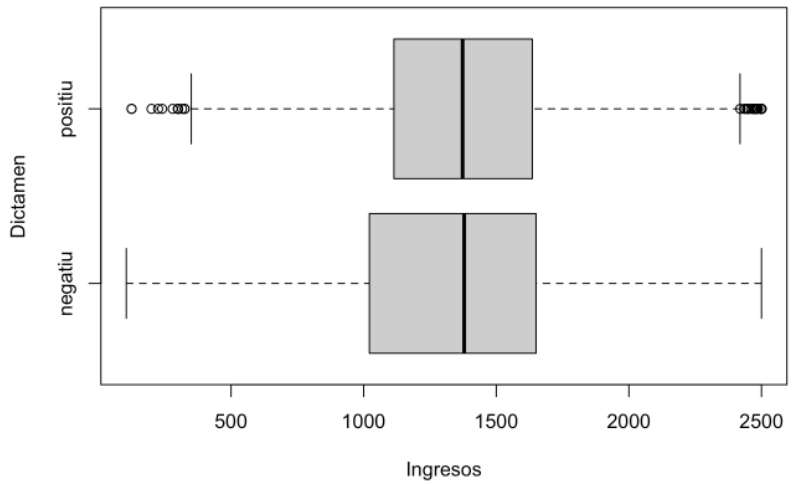


Figure 14: Boxplot for "Net Income" divided by "Dictamen"

4 Concluding Remarks

Now we summarize the results obtained throughout the descriptive analysis of the data, so the owner of the data can use this information in future analyses. We present these conclusions briefly (omitting previously mentioned details) in the following points:

- The most common profile of people that applies for credit are individuals who are married, have bought a house, and have a fixed job contract. Taking this into account will allow the user to avoid biases when analyzing data through other statistical methods.

- The proportion of accepted individuals that have a fixed job contract or are autonomous workers is relatively large in comparison to other categories. This is due to banks seeking profiles that guarantee payback of the credit through compromising the client assets (autonomous) or securing they have enough income to finance their obligations (fixed).

- Individuals that are married and have bought a house are conceded more credit relative to other categories. This might be due to the stability and security that these characteristics give to the bank related to payback.

- The type of housing and the civil status are dependent, so one could expect that using other methodologies of multivariate analysis, results show that one of the variables is enough to explain some factors of credit applications.

- An individual with a higher net income has a more positive assessment than an individual with a lower net income. This is due to the relation with the credit payback because they are more likely not to default if their net income is high (they can save more money). This has been deduced by looking at the average and the median individual, but the user should be aware that this is not always the case.

- The price of the good to be financed has not too much relevance when deciding whether to accept or reject a credit application. Again, this has been deduced by looking at the average and the median individual, but this may not always be the case.

- All in all, individuals with positive values or categories of variables that are positively related to the concept of security (in the sense of the individual paying back his debt) seem to have more accepted applications. This would mean that using other methodologies of multivariate data analysis, these variables allow us to explain factors related to credit applications.

Appendix 1. Descriptive Analysis

Iker Caballero & Daniel González

2023-03-20

Data Preprocessing

```
# For Vivienda
credsco$Vivienda <- as.factor(credsco$Vivienda)

credsco <- credsco[credsco$Vivienda!="ignora_cont"&credsco$Vivienda!="VivUnkown",]
credsco <- droplevels.data.frame(credsco,exclude=c("ignora_cont","VivUnkown"))

levels(credsco$Vivienda) <- c("altres viv","comprada","comprada","lloguer","pares")

levels(credsco$Vivienda)
```

```
## [1] "altres viv" "comprada"   "lloguer"    "pares"
```

```
# For Estado Civil

credsco$Estado.civil <- as.factor(credsco$Estado.civil)

credsco <- credsco[credsco$Estado.civil!="ECUnknown",]
credsco <- droplevels.data.frame(credsco,exclude="ECUnknown")

levels(credsco$Estado.civil) <- c("amb.par","tenien.par","tenien.par","sens.par","tenien.par")

levels(credsco$Estado.civil)
```

```
## [1] "amb.par"    "tenien.par" "sens.par"
```

```
# For Tipo Trabajo

credsco$Tipo.trabajo <- as.factor(credsco$Tipo.trabajo)

credsco <- credsco[credsco$Tipo.trabajo!="WorkingTypeUnknown",]
credsco <- droplevels.data.frame(credsco,exclude="WorkingTypeUnknown")

levels(credsco$Tipo.trabajo)
```

```
## [1] "altres sit" "autonom"    "fixe"       "temporal"
```

Question 1

```
# Graphical Analysis for Tipo.trabajo
tab_job <- table(credsco$Tipo.trabajo)
barplot(tab_job,col=c("pink","skyblue","lightyellow","lightgreen"))
title(main = "Multiple Barplot of Type of Contract", cex.main=0.75)
```

```
# Bivariate Graphical Analysis with "Dictamen"
tab_job_cred <- table(credsco$Dictamen,credsco$Tipo.trabajo)
data_percentage <- apply(tab_job_cred, 2, function(x){x*100/sum(x,na.rm=T)})
barplot(data_percentage,col=c("pink","skyblue"),legend=TRUE,xlim = c(0,6.5),ylab="Percentage")
title(main = "Stacked Barplot divided by Dictamen", cex.main=0.75)
```

```
# Numerical Analysis

(dist_job <- round(freq(credsco$Tipo.trabajo,plot=FALSE),1))
```

```
## credsco$Tipo.trabajo
##            Frequency Percent
## altres sit       171     3.9
## autonom         1016    23.0
## fixe            2794    63.1
## temporal         446    10.1
## Total           4427   100.0
```

```
# Bivariate Numerical Analysis with "Dictamen"

(dist_job_diag <- suppressWarnings(CrossTable(credsco$Tipo.trabajo,credsco$Dictamen,
                                              prop.t=TRUE,
                                              prop.r=TRUE, prop.c=TRUE,
                                              prop.chisq = F,digits=2)))
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
## ================================================
##                          credsco$Dictamen
## credsco$Tipo.trabajo    negatiu   positiu   Total
## ------------------------------------------------
## altres sit                   68       103     171
##                            0.40      0.60    0.04
##                            0.05      0.03
##                            0.02      0.02
## ------------------------------------------------
## autonom                     329       687    1016
##                            0.32      0.68    0.23
```

```
##                             0.27      0.22
##                             0.07      0.16
## ------------------------------------------------
## fixe                         577      2216    2793
##                             0.21      0.79    0.63
##                             0.47      0.70
##                             0.13      0.50
## ------------------------------------------------
## temporal                     266       180     446
##                             0.60      0.40    0.10
##                             0.21      0.06
##                             0.06      0.04
## ------------------------------------------------
## Total                       1240      3186    4426
##                             0.28      0.72
## ================================================
```

```
(chisq.test(tab_job_cred))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_job_cred
## X-squared = 317.46, df = 3, p-value < 2.2e-16
```

Question 2

```
# Graphical Analysis for both variables
tab_viv <- table(credsco$Vivienda)
barplot(tab_viv,col=c("pink","skyblue","lightyellow","lightgreen"),ylim=c(0,2500))
title(main = "Multiple Barplot of Type of House", cex.main=1.25)
```

```
tab_ec <- table(credsco$Estado.civil)
barplot(tab_ec,col=c("pink","skyblue","lightyellow"))
title(main = "Multiple Barplot of Civil Status", cex.main=1)
```

```
# Bivariate Graphical Analysis for both variables

tab_viv_cs <- table(credsco$Vivienda,credsco$Estado.civil)
data_percentage1 <- apply(tab_viv_cs, 2, function(x){x*100/sum(x,na.rm=T)})
barplot(data_percentage1,col=c("pink","skyblue","lightyellow",
                                    "lightgreen"),legend=TRUE,
                                    ylab="Percentage",xlim=c(0,5))
title(main = "Stacked Barplot divided by Civil Status", cex.main=0.75)
```

```
# Bivariate Graphical Analysis with "Dictamen"

tab_viv_dict <- table(credsco$Dictamen,credsco$Vivienda)
tab_cs_dict <- table(credsco$Dictamen,credsco$Estado.civil)
data_percentage_viv <- apply(tab_viv_dict, 2, function(x){x*100/sum(x,na.rm=T)})
```

```r
data_percentage_cs <- apply(tab_cs_dict, 2, function(x){x*100/sum(x,na.rm=T)})
barplot(data_percentage_viv,col=c("pink","skyblue"),legend=TRUE,xlim = c(0,6),
                                        ylab="Percentage")
title(main = "Stacked Barplot divided by Dictamen", cex.main=0.75)
```

```r
barplot(data_percentage_cs,col=c("pink","skyblue"),
                                       legend=TRUE,
                                       xlim = c(0,5),
        ylab="Percentage")
title(main = "Stacked Barplot divided by Dictamen", cex.main=0.75)
```

```r
# Numerical Analysis for both variables
(dist_viv <- round(freq(credsco$Vivienda,plot=FALSE),1))
```

```
## credsco$Vivienda
##            Frequency Percent
## altres viv       319     7.2
## comprada        2353    53.2
## lloguer          973    22.0
## pares            782    17.7
## Total           4427   100.0
```

```r
(dist_cs <-  round(freq(credsco$Estado.civil,plot=FALSE),1))
```

```
## credsco$Estado.civil
##            Frequency Percent
## amb.par         3220    72.7
## tenien.par       234     5.3
## sens.par         973    22.0
## Total           4427   100.0
```

```r
# Numerical Bivariate Analysis for both variables

(dist_viv_cs <- suppressWarnings(CrossTable(credsco$Vivienda,
                                            credsco$Estado.civil,
                                            prop.t=TRUE,
                                            prop.r=TRUE, prop.c=TRUE,
                                            prop.chisq = F,digits=2)))
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
## ============================================================
##                     credsco$Estado.civil
## credsco$Vivienda    amb.par   tenien.par   sens.par   Total
## ------------------------------------------------------------
```

```
## altres viv              208           13         98     319
##                        0.65         0.04       0.31    0.07
##                        0.06         0.06       0.10
##                        0.05         0.00       0.02
## ------------------------------------------------------------
## comprada               2051          108        194    2353
##                        0.87         0.05       0.08    0.53
##                        0.64         0.46       0.20
##                        0.46         0.02       0.04
## ------------------------------------------------------------
## lloguer                 723           76        174     973
##                        0.74         0.08       0.18    0.22
##                        0.22         0.32       0.18
##                        0.16         0.02       0.04
## ------------------------------------------------------------
## pares                   238           37        507     782
##                        0.30         0.05       0.65    0.18
##                        0.07         0.16       0.52
##                        0.05         0.01       0.11
## ------------------------------------------------------------
## Total                  3220          234        973    4427
##                        0.73         0.05       0.22
## ============================================================
```

```r
# Numerical Bivariate Analysis with "Dictamen"

(dist_viv_dict <- suppressWarnings(CrossTable(credsco$Vivienda,
                                              credsco$Dictamen,
                                              prop.t=TRUE, prop.r=TRUE,
                                              prop.c=TRUE,
                                              prop.chisq = F,digits=2)))
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
## ============================================
##                     credsco$Dictamen
## credsco$Vivienda    negatiu   positiu   Total
## --------------------------------------------
## altres viv              146       173     319
##                        0.46      0.54    0.07
##                        0.12      0.05
##                        0.03      0.04
## --------------------------------------------
## comprada                474      1878    2352
##                        0.20      0.80    0.53
##                        0.38      0.59
##                        0.11      0.42
## --------------------------------------------
```

```
## lloguer                  388       585     973
##                         0.40      0.60    0.22
##                         0.31      0.18
##                         0.09      0.13
## ------------------------------------------------
## pares                    232       550     782
##                         0.30      0.70    0.18
##                         0.19      0.17
##                         0.05      0.12
## ------------------------------------------------
## Total                   1240      3186    4426
##                         0.28      0.72
## ================================================
```

```
(dist_cs_dict <- suppressWarnings(CrossTable(credsco$Estado.civil,
                                             credsco$Dictamen,
                                             prop.t=TRUE, prop.r=TRUE, prop.c=TRUE,
                                             prop.chisq = F,digits=2)))
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
## ================================================
##                         credsco$Dictamen
## credsco$Estado.civil    negatiu   positiu   Total
## ------------------------------------------------
## amb.par                     817      2403    3220
##                            0.25      0.75    0.73
##                            0.66      0.75
##                            0.18      0.54
## ------------------------------------------------
## tenien.par                   97       137     234
##                            0.41      0.59    0.05
##                            0.08      0.04
##                            0.02      0.03
## ------------------------------------------------
## sens.par                    326       646     972
##                            0.34      0.66    0.22
##                            0.26      0.20
##                            0.07      0.15
## ------------------------------------------------
## Total                      1240      3186    4426
##                            0.28      0.72
## ================================================
```

```
(chisq.test(tab_viv_cs))
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  tab_viv_cs
## X-squared = 1151.7, df = 6, p-value < 2.2e-16
```

```
(chisq.test(tab_cs_dict))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_cs_dict
## X-squared = 46.808, df = 2, p-value = 6.851e-11
```

```
(chisq.test(tab_viv_dict))
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_viv_dict
## X-squared = 190.88, df = 3, p-value < 2.2e-16
```

Question 3

```
# Univariate analysis
d=describe(credsco[c(9,10)])
as.data.frame(d)[,-c(1,6,7,13)]
```

```
##             n      mean       sd median min max range     skew  kurtosis
## Gastos   4427  55.59386 19.54029     51  35 180   145 1.013965  1.406176
## Ingresos 4427 130.50666 86.25403    120   0 959   959 1.990062 10.037256
```

```
boxplot(credsco[,9], horizontal=TRUE, xlab=colnames(credsco[,9]))
```

```
boxplot(credsco[,10], horizontal=TRUE, xlab=colnames(credsco[,10]))
```

```
hist(credsco$Gastos, xlab="Gastos")
```

```
hist(credsco$Ingresos, xlab="Ingresos")
```

```
credsco<-credsco[credsco[,9]<=125,]
credsco<-credsco[credsco[,10]<=250,]

boxplot(credsco[,9], horizontal=TRUE, xlab=colnames(credsco[,9]))
```

```
boxplot(credsco[,10], horizontal=TRUE, xlab=colnames(credsco[,10]))
```

```r
hist(credsco$Gastos, xlab="Gastos")
```

```r
hist(credsco$Ingresos, xlab="Ingresos")
```

```r
hist(credsco$Gastos, xlab="Gastos")
```

```r
hist(credsco$Ingresos, xlab="Ingresos")
```

```r
cor(credsco[,c(9,10)])
```

```
##             Gastos  Ingresos
## Gastos   1.0000000 0.1911186
## Ingresos 0.1911186 1.0000000
```

```r
# Net income

net_inc <- credsco[,10]-credsco[,9]
credsco3 <- cbind(credsco,net_inc)
colnames(credsco3) <- c(colnames(credsco),"net_inc")
d_inc=describe(credsco3$net_inc)
as.data.frame(d_inc)[,-c(1,6,7,13)]
```

```
##       n     mean       sd median  min max range        skew    kurtosis
## X1 4128 60.73571 58.48346     61 -120 215   335 -0.07704517 -0.06451958
```

```r
boxplot(credsco3[,17], horizontal=TRUE, xlab="Net Income")
```

```r
hist(credsco3[,17], xlab="Net Income")
```

```r
d_pos1=describe(credsco3[(credsco$Dictamen=="positiu"),
                         17])
d_neg1=describe(credsco3[(credsco$Dictamen!="positiu"),
                         17])
as.data.frame(d_pos1)[,-c(1,6,7,13)]
```

```
##       n     mean       sd median min max range        skew  kurtosis
## X1 2945 70.76537 54.81784     69 -90 215   305 -0.08530069 0.1716811
```

```r
as.data.frame(d_neg1)[,-c(1,6,7,13)]
```

```
##       n    mean       sd median  min max range      skew   kurtosis
## X1 1182 35.7022 59.84188     35 -120 209   329 0.1761559 -0.2424536
```

```r
boxplot(credsco3$net_inc ~ credsco3$Dictamen,
        horizontal=TRUE,
        xlab="Net Income",ylab="Dictamen")
```

```r
# Tests hypothesis

t.test(credsco3[(credsco$Dictamen=="positiu"),17], credsco3[(credsco$Dictamen!="positiu"),17],
       alternative = "two.sided",
       var.equal = FALSE)
```

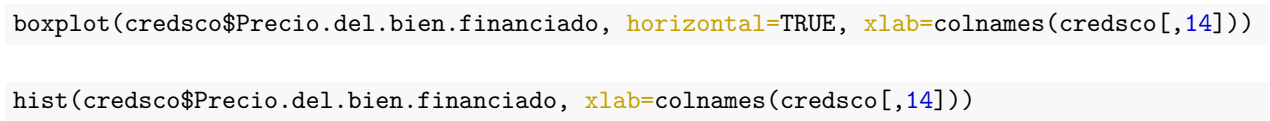
```
##
##  Welch Two Sample t-test
##
## data:  credsco3[(credsco$Dictamen == "positiu"), 17] and credsco3[(credsco$Dictamen != "positiu"), 1
## t = 17.423, df = 2018.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  31.11643 39.00990
## sample estimates:
## mean of x mean of y
##  70.76537  35.70220
```

Question 4

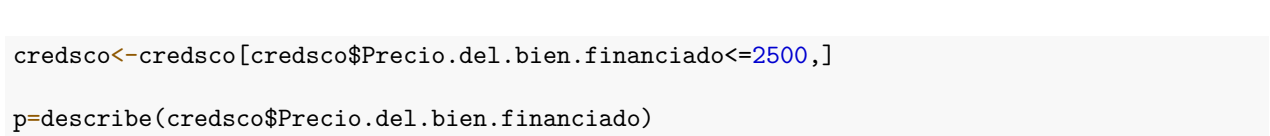
```r
# Univariate analysis

p=describe(credsco$Precio.del.bien.financiado)
as.data.frame(p)[,-c(1,6,7,13)]
```

```
##       n    mean      sd median min   max range     skew kurtosis
## X1 4128 1436.42 592.694   1395 105 11140 11035 2.760207 26.30738
```

```r
boxplot(credsco$Precio.del.bien.financiado, horizontal=TRUE, xlab=colnames(credsco[,14]))
```



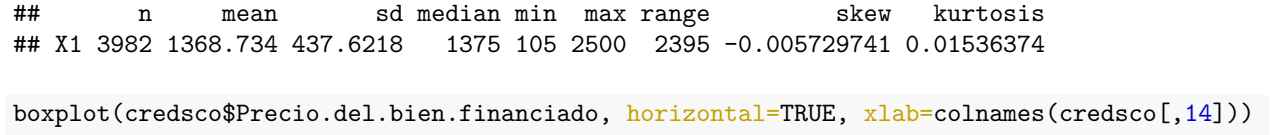
```r
hist(credsco$Precio.del.bien.financiado, xlab=colnames(credsco[,14]))
```



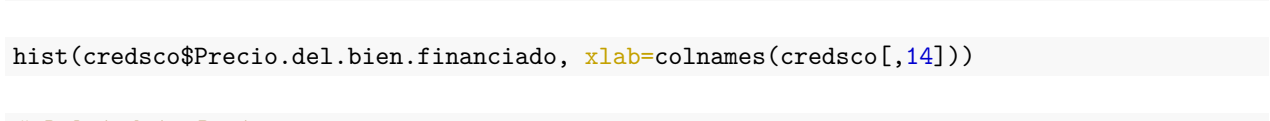
```r
credsco<-credsco[credsco$Precio.del.bien.financiado<=2500,]

p=describe(credsco$Precio.del.bien.financiado)
as.data.frame(p)[,-c(1,6,7,13)]
```

```
##       n     mean       sd median min  max range         skew   kurtosis
## X1 3982 1368.734 437.6218   1375 105 2500  2395 -0.005729741 0.01536374
```

```r
boxplot(credsco$Precio.del.bien.financiado, horizontal=TRUE, xlab=colnames(credsco[,14]))
```



```r
hist(credsco$Precio.del.bien.financiado, xlab=colnames(credsco[,14]))
```



```r
# Related to Dictamen

p_pos=describe(credsco$Precio.del.bien.financiado
```

```
               [credsco$Dictamen=="positiu"])
p_neg=describe(credsco$Precio.del.bien.financiado
               [credsco$Dictamen!="positiu"])
as.data.frame(p_pos)[,-c(1,6,7,13)]
```

```
##       n     mean       sd median min  max range       skew  kurtosis
## X1 2861 1376.263 423.5331   1373 125 2500  2375 0.03171477 0.1257875
```

```
as.data.frame(p_neg)[,-c(1,6,7,13)]
```

```
##       n     mean       sd median min  max range        skew   kurtosis
## X1 1120 1349.038 471.2643   1379 105 2500  2395 -0.04983318 -0.2641143
```

```
boxplot(credsco$Precio.del.bien.financiado
        ~ credsco$Dictamen,
        horizontal=TRUE,xlab="Ingresos",
        ylab="Dictamen")
```

```
# Tests hypothesis

t.test(
  credsco$Precio.del.bien.financiado
  [credsco$Dictamen=="positiu"],
  credsco$Precio.del.bien.financiado
  [credsco$Dictamen!="positiu"],
  alternative = "two.sided", var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  credsco$Precio.del.bien.financiado[credsco$Dictamen == "positiu"] and credsco$Precio.del.bien
## t = 1.6852, df = 1865.5, p-value = 0.09212
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.459528 58.909131
## sample estimates:
## mean of x mean of y
##  1376.263  1349.038
```