|  |  |
| --- | --- |
| **P170M109 Computational Intelligence and Decision** | **Date: 15 February 2018** |
| TOPIC 1 | **Students: Erika Gardini – Mattia Fucili** |

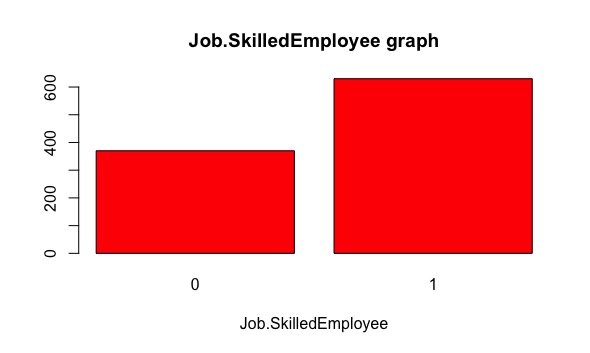
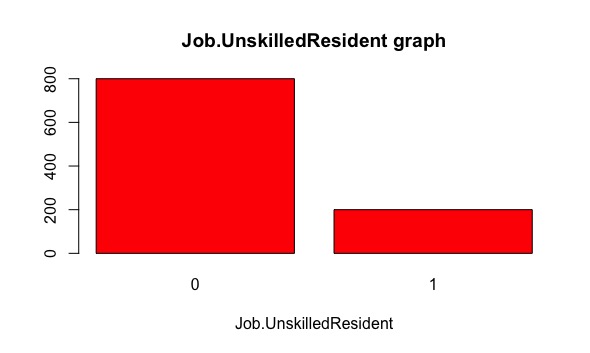
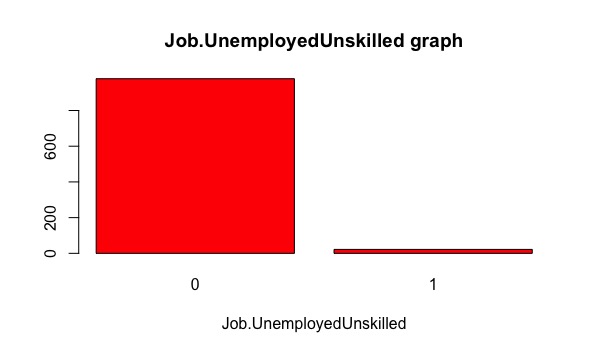
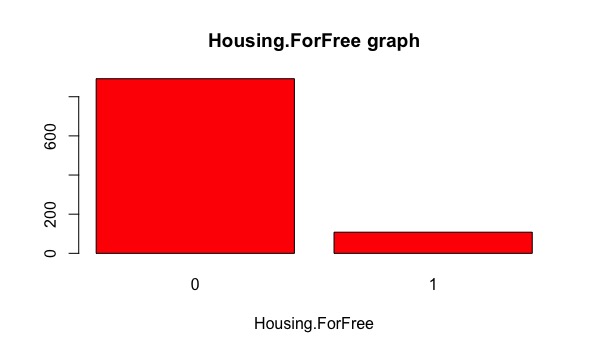
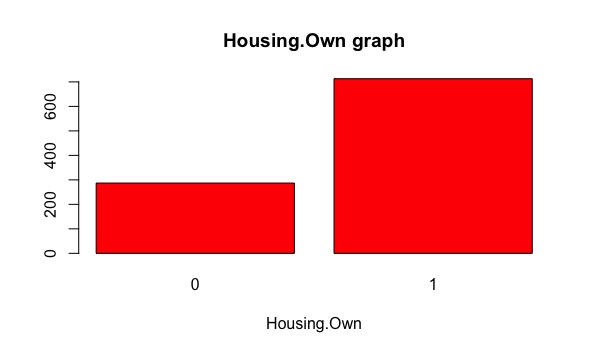
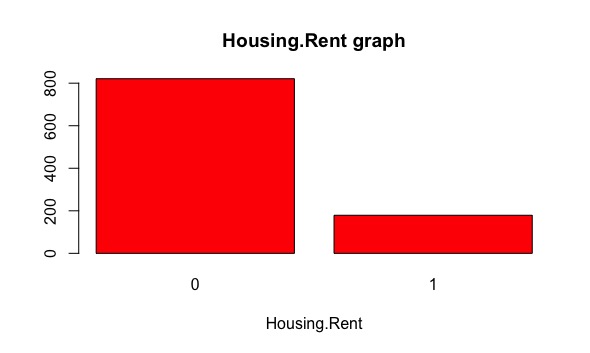
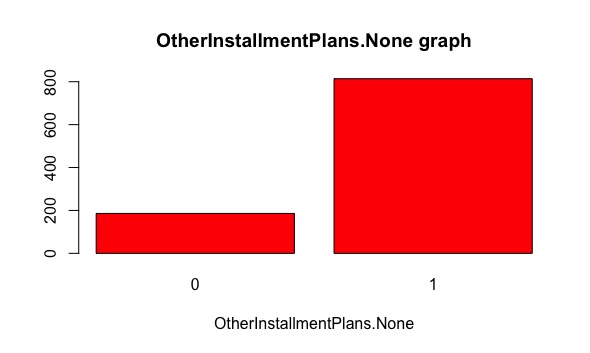
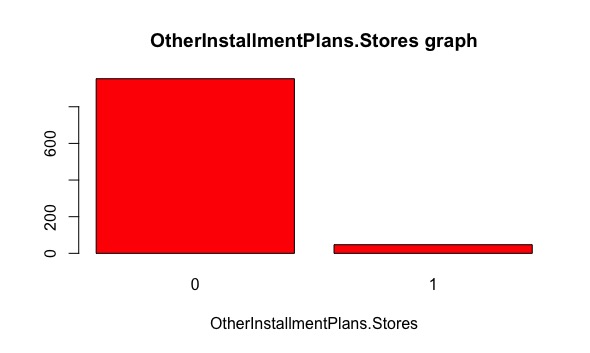
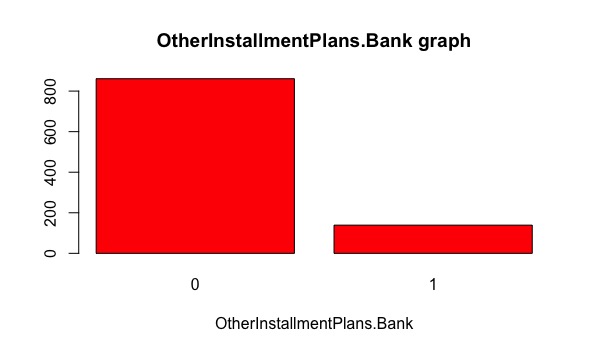
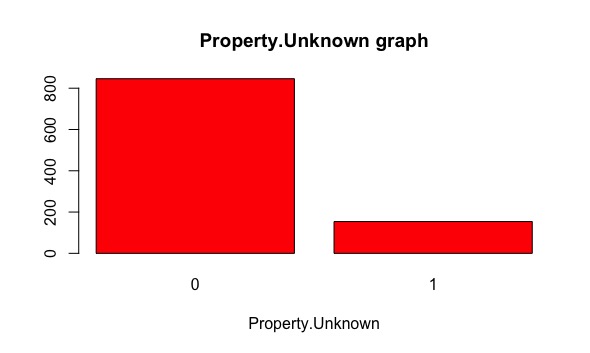
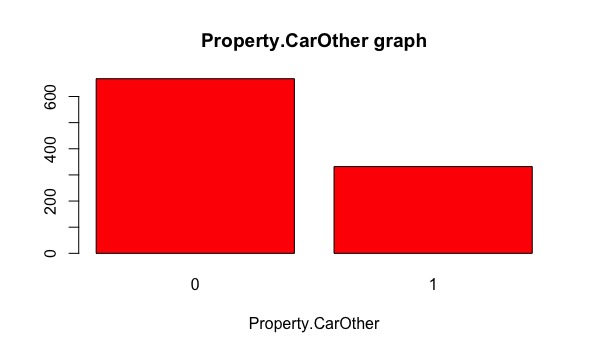
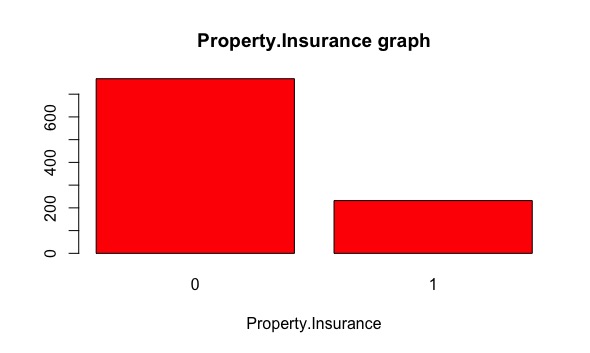
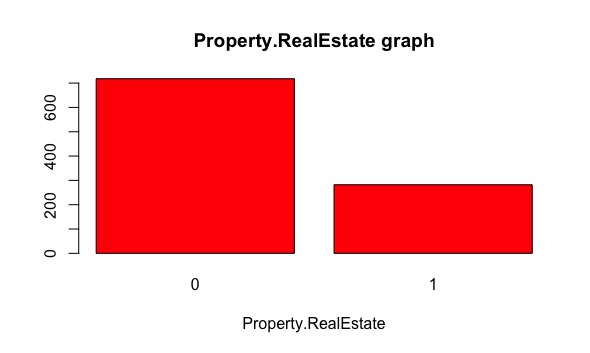
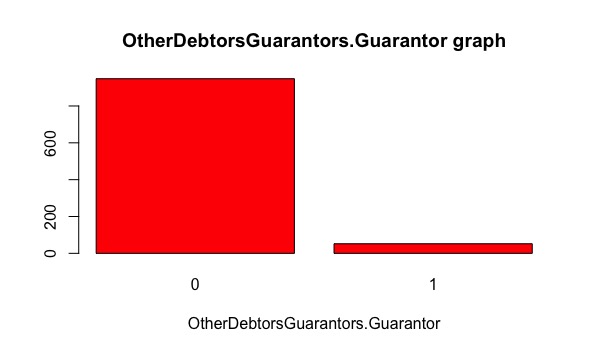
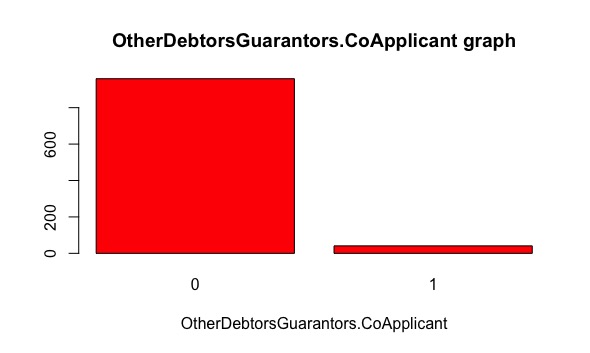
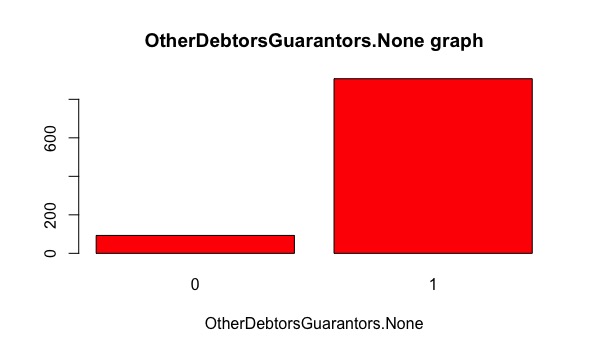
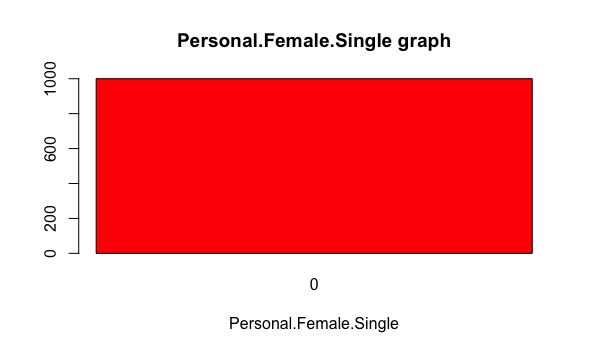
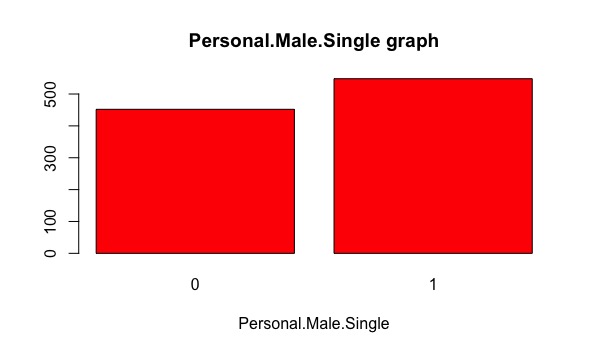
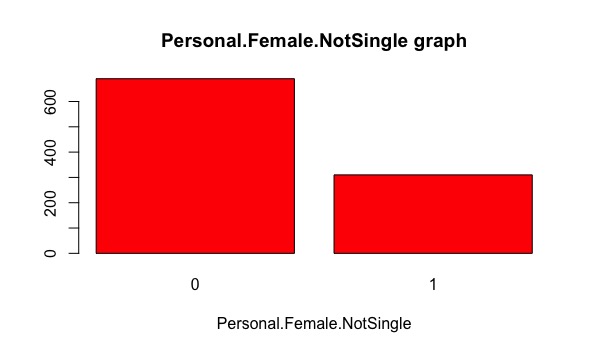
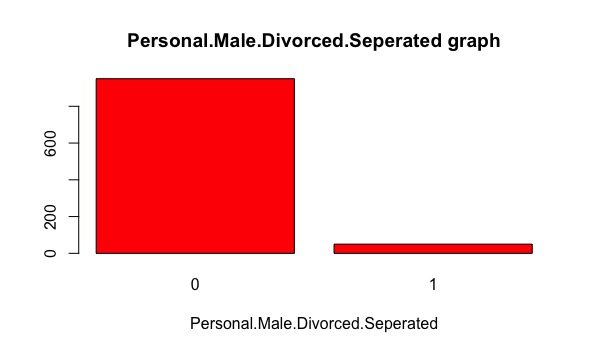
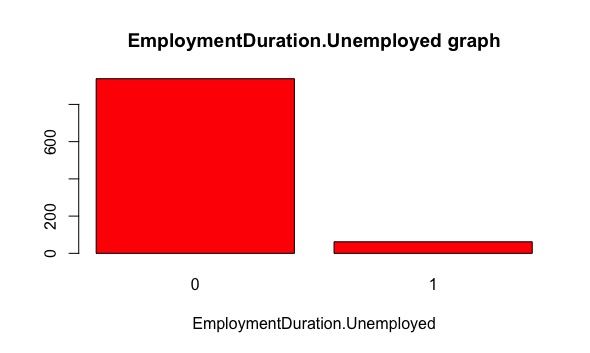
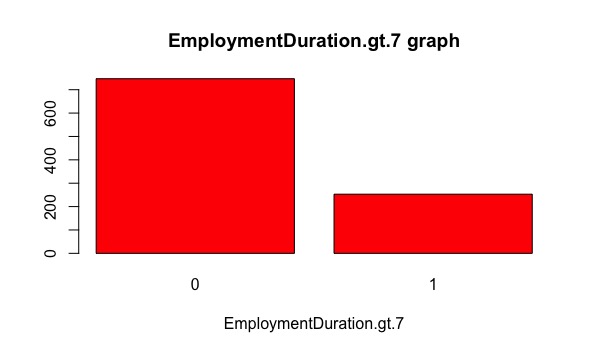
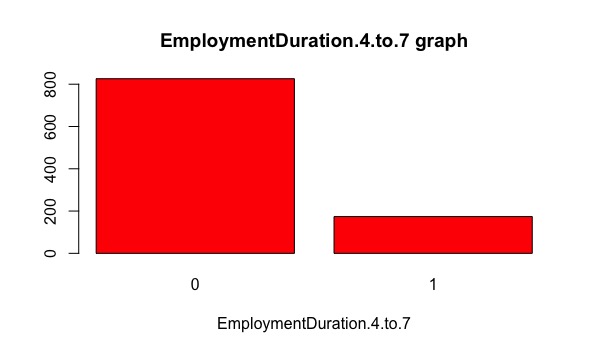
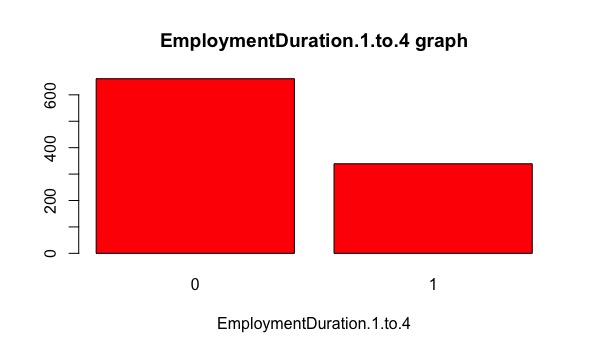
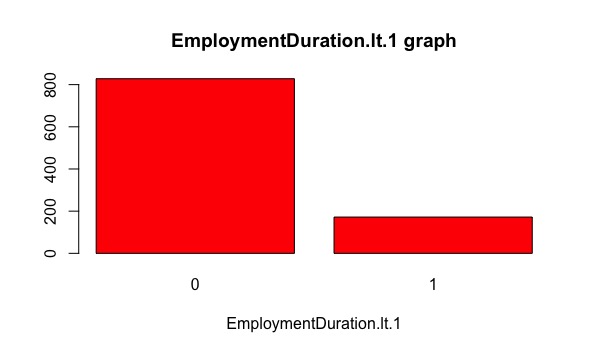
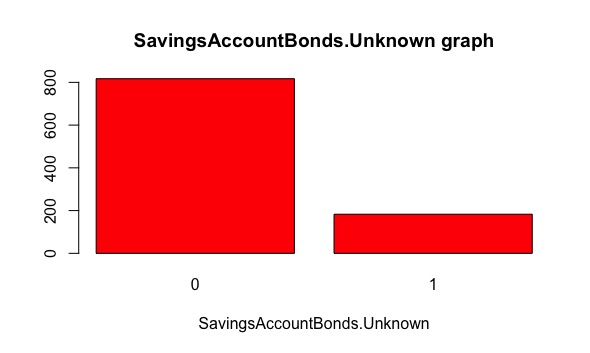
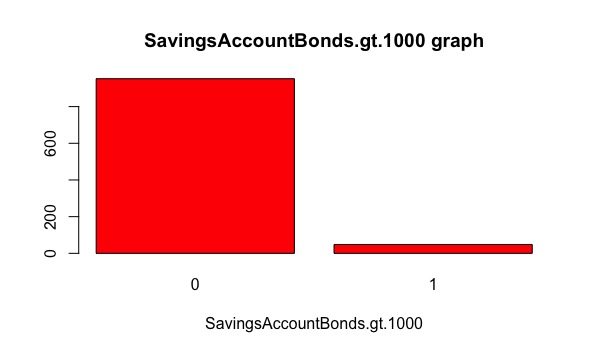
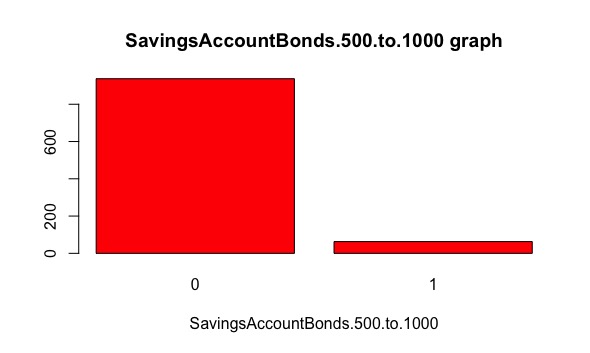
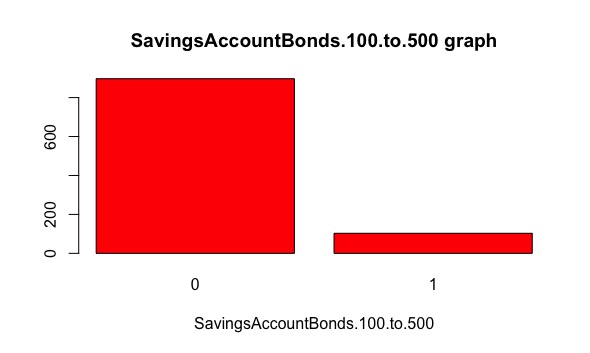
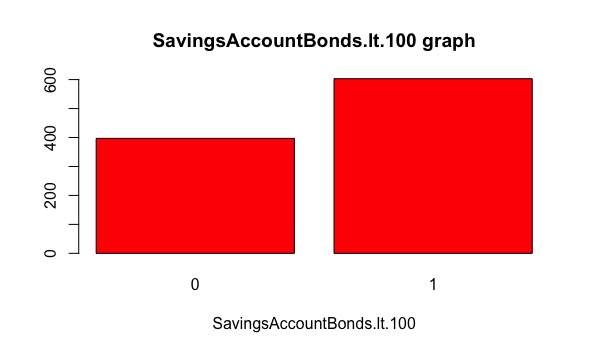
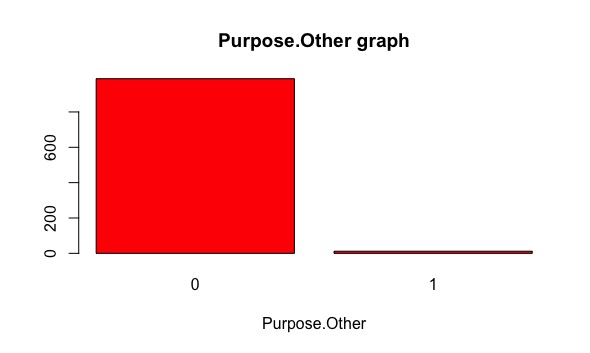
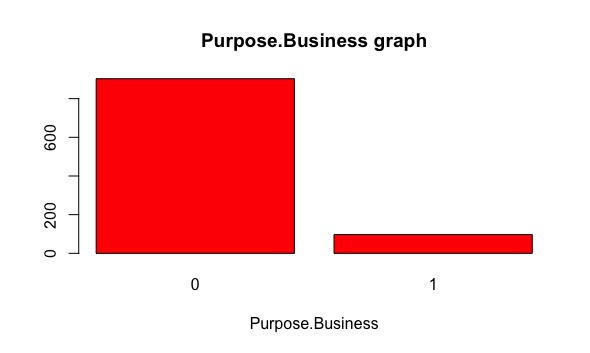
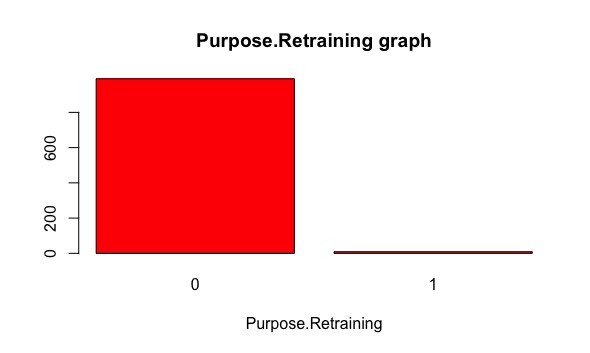
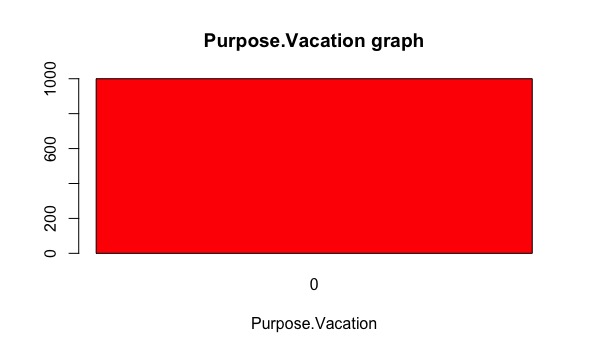
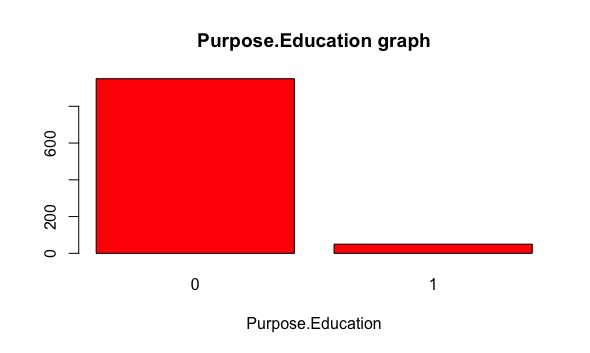
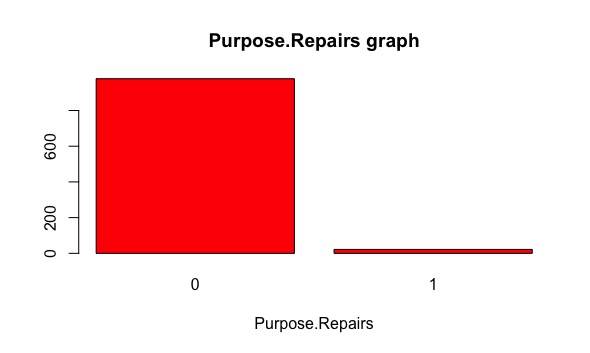
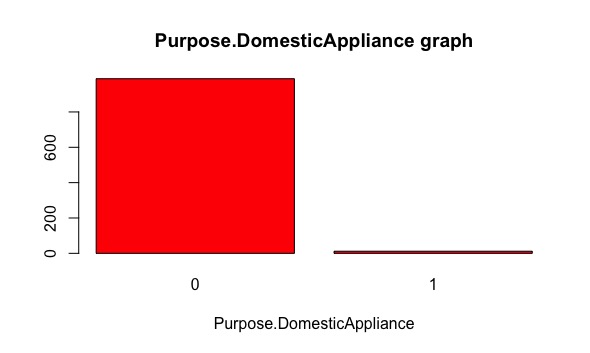
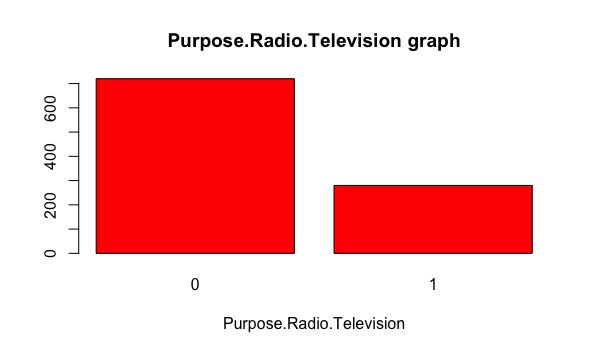
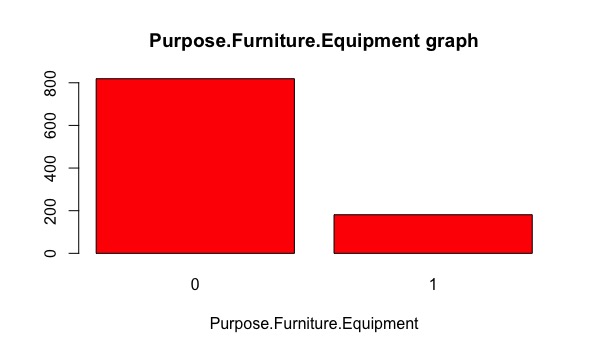
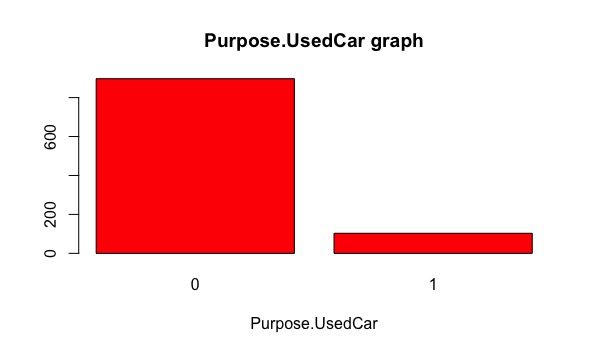
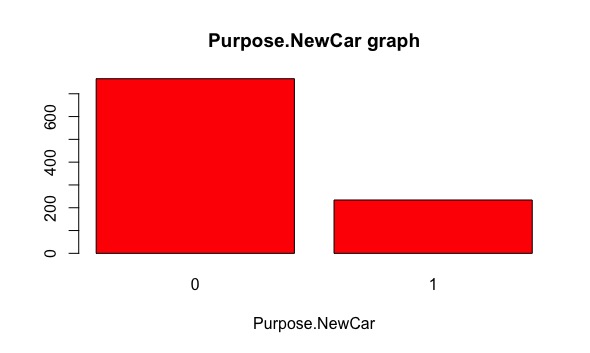
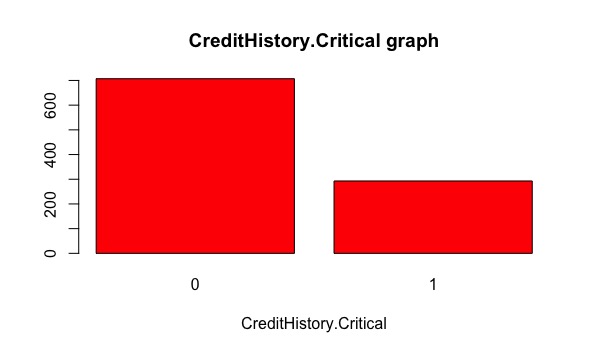
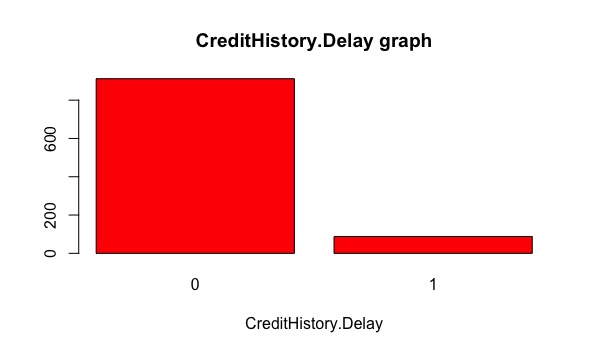
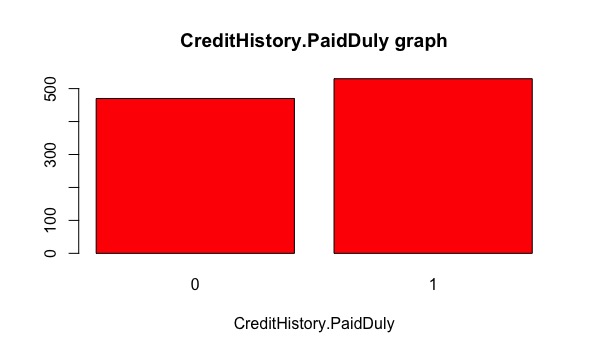
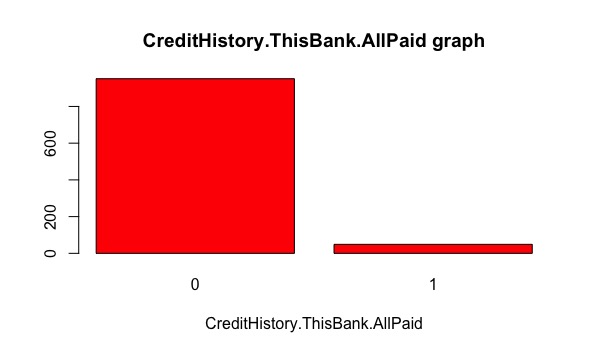
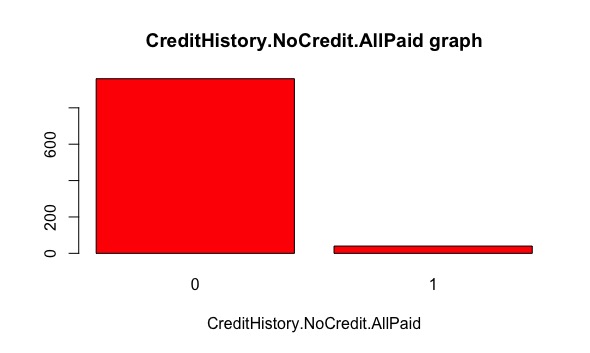
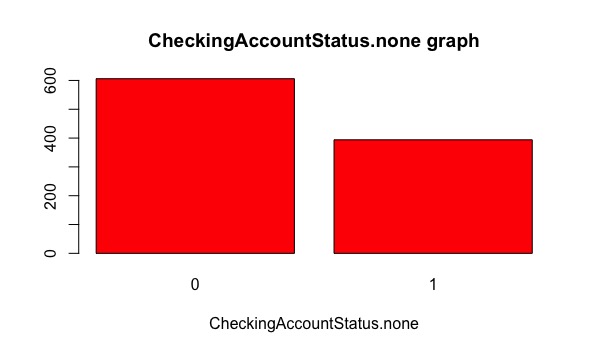
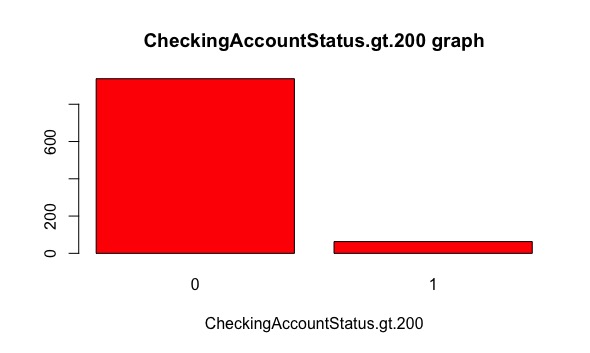
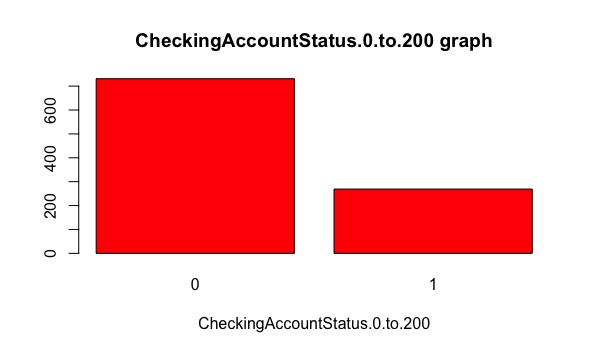
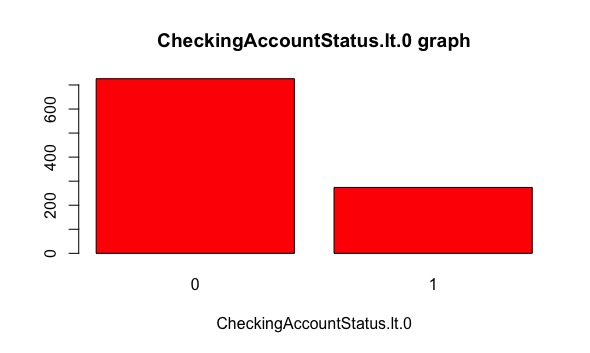
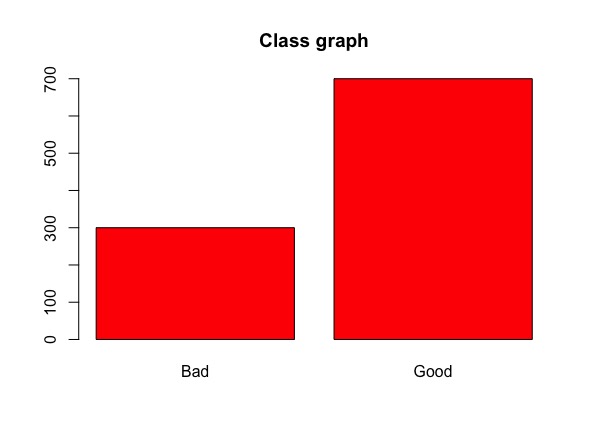
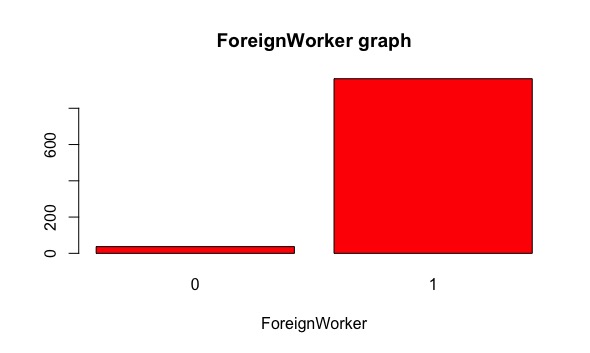
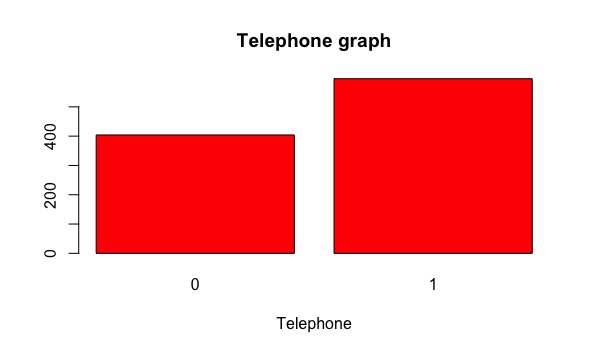
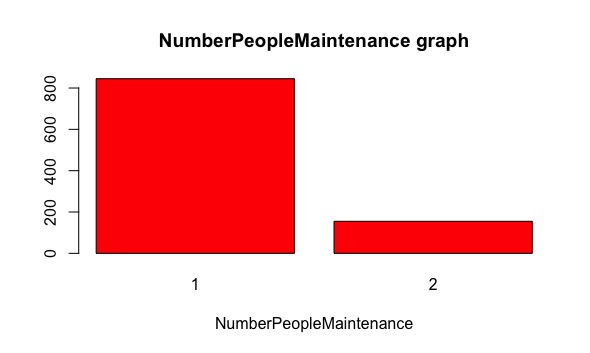
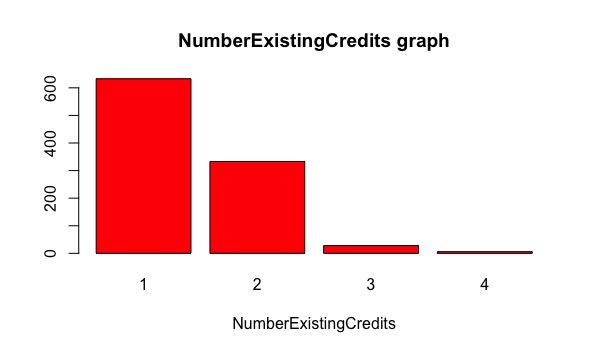
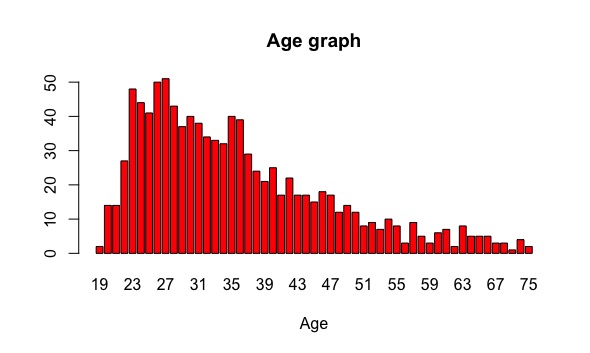
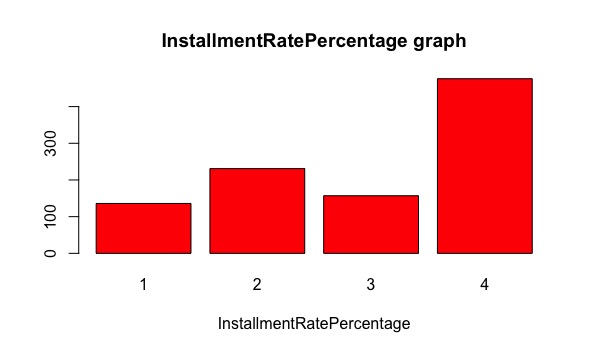
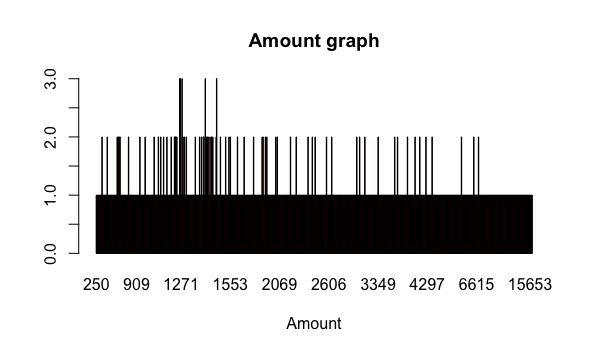
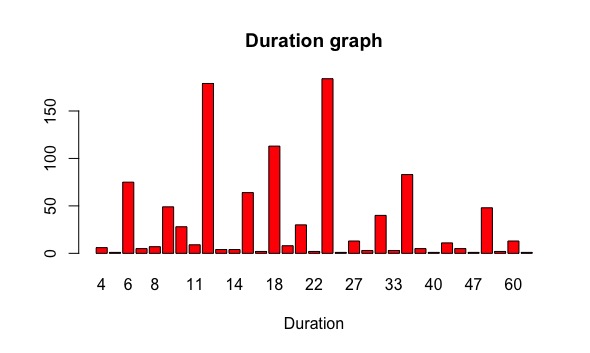
Provide the results obtained following the steps given in the lab work description (txt file). Please, plan your time to meet deadlines. This document (report) should be uploaded to Moodle system.

1. *Determine data type (categorical, binary, numerical, ...), characterize them by displaying graphically (histogram, pie, time series, ...). Chose at least 2 numerical and at least 2 categorical features to explore*

The following table contains the data type of all the feature:

|  |  |
| --- | --- |
| **Feature** | **Type** |
| Duration | Numerical |
| Amount | Numerical |
| InstallmentRatePercentage | Categorical |
| ResidenceDuration | Categorical |
| Age | Numerical |
| NumberExistingCredits | Categorical |
| NumberPeopleMaintenance | Binary |
| Telephone | Binary |
| ForeignWorker | Binary |
| Class | Categorical |
| CheckingAccountStatus.lt.0 | Binary |
| CheckingAccountStatus.0.to.200 | Binary |
| CheckingAccountStatus.gt.200 | Binary |
| CheckingAccountStatus.none | Binary |
| CreditHistory.NoCredit.AllPaid | Binary |
| CreditHistory.ThisBank.AllPaid | Binary |
| CreditHistory.PaidDuly | Binary |
| CreditHistory.Delay | Binary |
| CreditHistory.Critical | Binary |
| Purpose.NewCar | Binary |
| Purpose.UsedCar | Binary |
| Purpose.Furniture.Equipment | Binary |
| Purpose.Radio.Television | Binary |
| Purpose.DomesticAppliance | Binary |
| Purpose.Repairs | Binary |
| Purpose.Education | Binary |
| Purpose.Vacation | Binary |
| Purpose.Retraining | Binary |
| Purpose.Business | Binary |
| Purpose.Other | Binary |
| SavingsAccountBonds.lt.100 | Binary |
| SavingsAccountBonds.100.to.500 | Binary |
| SavingsAccountBonds.500.to.1000 | Binary |
| SavingsAccountBonds.Unknown | Binary |
| EmploymentDuration.lt.1 | Binary |
| EmploymentDuration.1.to.4 | Binary |
| EmploymentDuration.4.to.7 | Binary |
| EmploymentDuration.gt.7 | Binary |
| EmploymentDuration.Unemployed | Binary |
| Personal.Male.Divorced.Seperated | Binary |
| Personal.Female.NotSingle | Binary |
| Personal.Male.Single | Binary |
| Personal.Male.Married.Widowed | Binary |
| Personal.Female.Single | Binary |
| OtherDebtorsGuarantors.None | Binary |
| OtherDebtorsGuarantors.CoApplicant | Binary |
| OtherDebtorsGuarantors.Guarantor | Binary |
| Property.RealEstate | Binary |
| Property.Insurance | Binary |
| Property.CarOther | Binary |
| Property.Unknown | Binary |
| OtherInstallmentPlans.Bank | Binary |
| OtherInstallmentPlans.Stores | Binary |
| OtherInstallmentPlans.None | Binary |
| Housing.Rent | Binary |
| Housing.Own | Binary |
| Housing.ForFree | Binary |
| Job.UnemployedUnskilled | Binary |
| Job.UnskilledResident | Binary |
| Job.SkilledEmployee | Binary |
| Job.Management.SelfEmp.HighlyQualified | Binary |

What follows is the graphical representation of features:



The following table contains the features exploration of all the categorical features:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Feature*** | ***Count*** | ***Mode*** | ***Mode Freq.*** | ***Mode %*** | ***2nd Mode*** | ***2nd Mode Freq.*** | ***2nd Mode %*** |
| InstallmentRatePercentage | 1000 | 4 | 476 | 47,60% | 2 | 231 | 23,10% |
| ResidenceDuration | 1000 | 4 | 413 | 41,30% | 2 | 308 | 30,80% |
| NumberExistingCredits | 1000 | 1 | 633 | 63,30% | 2 | 333 | 33,30% |
| Class | 1000 | Good | 700 | 70,00% | Bad | 300 | 30,00% |

If the *Mode Frequency* is higher than *2nd Mode Frequency* it means that there is a dominant value. For “InstallmentRatePercentage”, “NumberExistingCredits” and “Class” we can say that there is a dominant value, respectively “4”, “1” and “Good”. The feature “ResidenceDuration” is more balanced since there is a low distance between the *Mode Frequency* and the *2nd Mode Frequency*.

The following table contains the features exploration of all the numerical features:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Feature*** | ***Count*** | ***Min.*** | ***1st Qrt.*** | ***Mean*** | ***Median*** | ***3rd Qrt.*** | ***Max.*** | ***Std. Dev.*** |
| Duration | 1000 | 4 | 12 | 20,90 | 18 | 24 | 72 | 12,05881 |
| Amount | 1000 | 250 | 1366 | 3271,00 | 2320 | 3972 | 18424 | 2.822,737 |
| Age | 1000 | 19 | 27 | 35,55 | 33 | 42 | 75 | 11,37547 |

Looking to the standard deviation we can understand how much the values spread around the mean value. In this case, the feature “Amount” has a high value of *standard deviation*, even the features “Duration” and “Age” have a low value of *standard deviation*.

1. *Explore the quality of data: find outliers and make decision about them, find missing values if any, determine the cardinality*

First of all, we are looking for missing values. Executing the following simple script, we can see that there aren’t any missing values:

missing\_value <- function(ogg) {

x = 0

for(i in 1:length(ogg)) {

t <- is.na(ogg[i])

if(t == TRUE)

x <- x + 1

}

x

}

all\_column <- function(ogg) {

x = 0

for(i in names(data)) {

res <- missing\_value(data[[i]])

if(res > 0)

x <- x + 1

}

x

}

The following table contains information about the cardinality:

|  |  |
| --- | --- |
| ***Feature*** | ***Cardinality*** |
| Duration | 33 |
| Amount | 921 |
| InstallmentRatePercentage | 4 |
| ResidenceDuration | 4 |
| Age | 53 |
| NumberExistingCredits | 4 |
| NumberPeopleMaintenance | 2 |
| Telephone | 2 |
| ForeignWorker | 2 |
| Class | 2 |
| CheckingAccountStatus.lt.0 | 2 |
| CheckingAccountStatus.0.to.200 | 2 |
| CheckingAccountStatus.gt.200 | 2 |
| CheckingAccountStatus.none | 2 |
| CreditHistory.NoCredit.AllPaid | 2 |
| CreditHistory.ThisBank.AllPaid | 2 |
| CreditHistory.PaidDuly | 2 |
| CreditHistory.Delay | 2 |
| CreditHistory.Critical | 2 |
| Purpose.NewCar | 2 |
| Purpose.UsedCar | 2 |
| Purpose.Furniture.Equipment | 2 |
| Purpose.Radio.Television | 2 |
| Purpose.DomesticAppliance | 2 |
| Purpose.Repairs | 2 |
| Purpose.Education | 2 |
| Purpose.Vacation | 1 |
| Purpose.Retraining | 2 |
| Purpose.Business | 2 |
| Purpose.Other | 2 |
| SavingsAccountBonds.lt.100 | 2 |
| SavingsAccountBonds.100.to.500 | 2 |
| SavingsAccountBonds.500.to.1000 | 2 |
| SavingsAccountBonds.Unknown | 2 |
| EmploymentDuration.lt.1 | 2 |
| EmploymentDuration.1.to.4 | 2 |
| EmploymentDuration.4.to.7 | 2 |
| EmploymentDuration.gt.7 | 2 |
| EmploymentDuration.Unemployed | 2 |
| Personal.Male.Divorced.Seperated | 2 |
| Personal.Female.NotSingle | 2 |
| Personal.Male.Single | 1 |
| Personal.Male.Married.Widowed | 2 |
| Personal.Female.Single | 2 |
| OtherDebtorsGuarantors.None | 2 |
| OtherDebtorsGuarantors.CoApplicant | 2 |
| OtherDebtorsGuarantors.Guarantor | 2 |
| Property.RealEstate | 2 |
| Property.Insurance | 2 |
| Property.CarOther | 2 |
| Property.Unknown | 2 |
| OtherInstallmentPlans.Bank | 2 |
| OtherInstallmentPlans.Stores | 2 |
| OtherInstallmentPlans.None | 2 |
| Housing.Rent | 2 |
| Housing.Own | 2 |
| Housing.ForFree | 2 |
| Job.UnemployedUnskilled | 2 |
| Job.UnskilledResident | 2 |
| Job.SkilledEmployee | 2 |
| Job.Management.SelfEmp.HighlyQualified | 2 |

We can see that the features “Personal.Male.Single” and “Purpose.Vacation” are useless because the cardinality is 1 (we can understand it also from the graphical representation of the features).

Finally, we have analysed the numerical features to find outliers (for categorical and binary features we can’t find outliers). One possible approach is to use the boxplot rule, based on the upper and lower quartiles of the data distribution.

It is based on the following nominal data range:

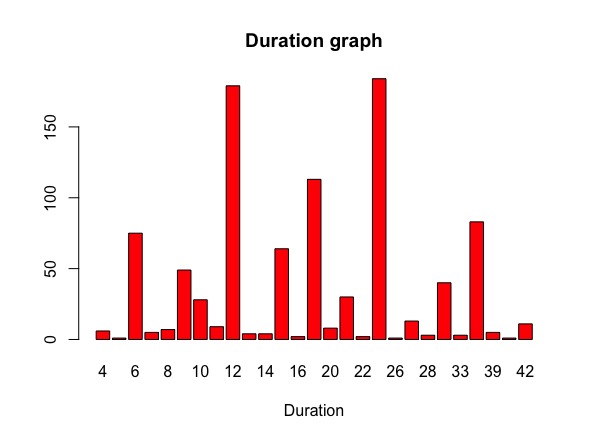
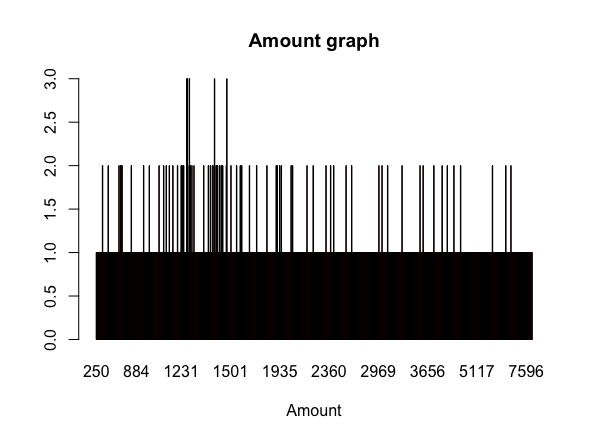
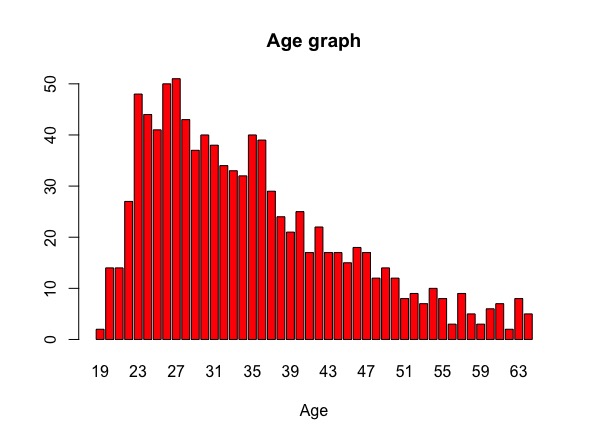
where Q1 and Q3 represent the lower and upper quartiles, respectively, of the data distribution, and

is the interquartile distance, a measure of the spread of the data similar to the standard deviation. The threshold parameter c is commonly equal to 1.5.

If we apply this procedure to our dataset we can forecast, according to the following table, that all the values under the lower threshold and above the higher threshold will be deleted.

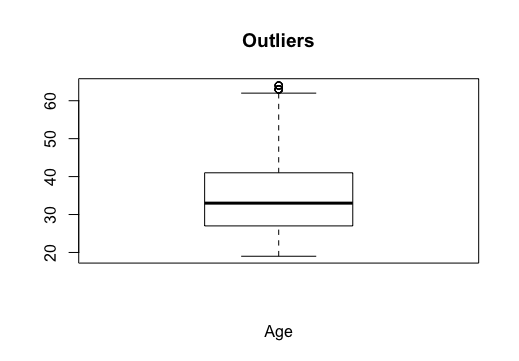
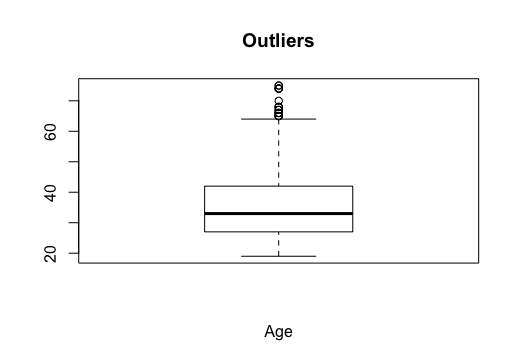
|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Duration*** | ***Amount*** | ***Age*** |
| ***Min value*** | 4 | 250 | 19 |
| ***Max value*** | 72 | 18424 | 75 |
| ***Low threshold outliers*** | -6 | -2544,625 | 4,5 |
| ***High threshold outliers*** | 42 | 7882,375 | 64,5 |

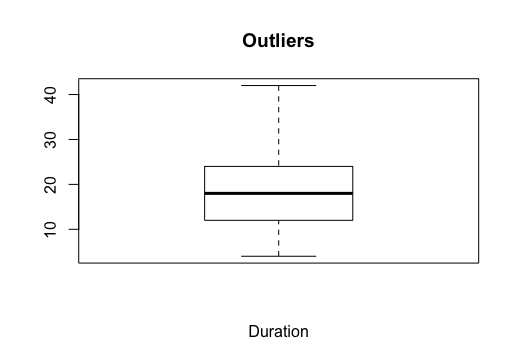
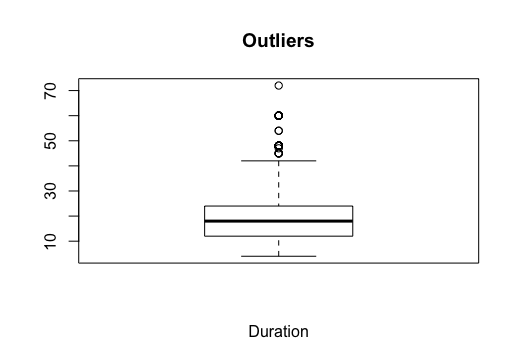
In fact, those are the histograms of the features “Age”, “Duration” and “Amount” without outliers:

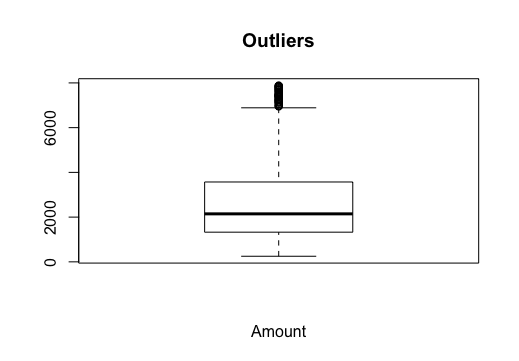
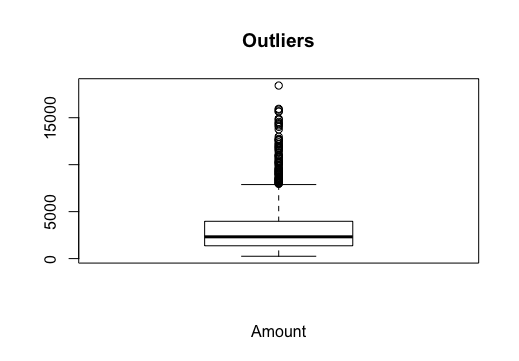


Features "Age", "Amount" and "Duration" without outliers

What follow are the boxplot of the features “Age”, “Amount” and “Duration” with and without outliers:







Features "Age", "Duration" and "Amount" with and without outliers

As we can imagine, the number of outliers is boiled down.

This result can be obtained, for example for the feature “Age”, executing the following script:

q1 <- quantile(credit$Age, 0.25)

q3 <- quantile(credit$Age, 0.75)

mean <- mean(credit$Age)

for(j in 1:length(credit$Age)) {

if(credit$Age[j] > (q3 + 1.5 \* IQR(credit$Age)) || credit$Age[j] < (q1 - 1.5 \* IQR(credit$Age))) {

credit$Age[j] <- NA

}

}

1. *Explore the shape of distribution. Perform standardization and normalization, consider the normality of data*

The following table contains information about the shape of distribution for the numerical feature:

|  |  |
| --- | --- |
| ***Feature*** | ***Shape of distribution*** |
| Duration | Multimodal |
| Amount | Uniform |
| Age | Unimodal (skewed left) |

We can understand it looking the histograms. The values of the feature “Amount” are uniformly distributed in the range, the feature “Age” tends to assume low values and the feature “Duration” is characterized by two commonly values, that are clearly separated.

For normalization we have used, for example for the feature “Age”, the following simple script:

m <- min(credit$Age, na.rm = TRUE)

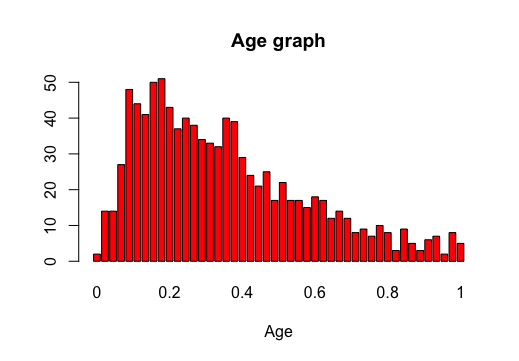
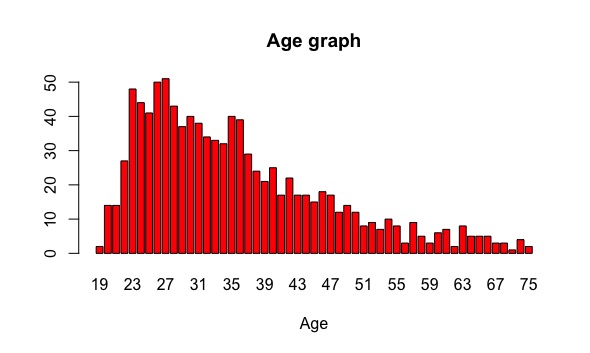
M <- max(credit$Age, na.rm = TRUE)

for(i in 1:length(credit$Age)){

credit$Age[i] <- (credit$Age[i] - m) / (M - m)

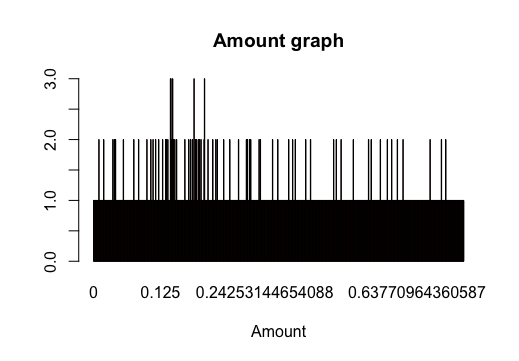
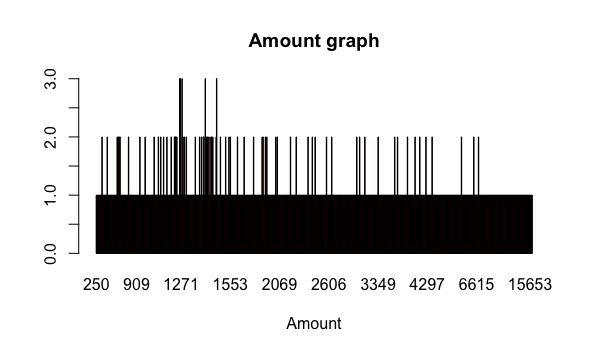
}

Using the range [0, 1], the obtained result is:

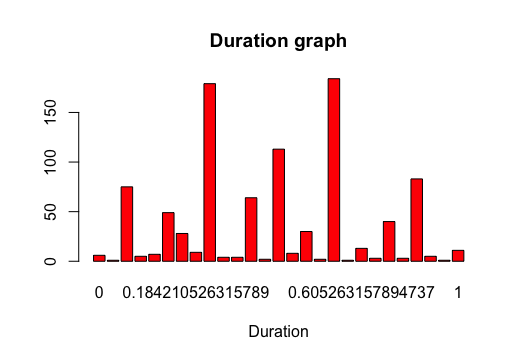
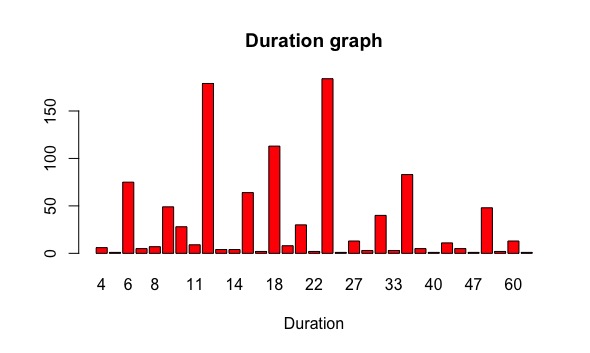


Feature "Age" before and after normalization

The same results can be obtained with the other numerical features:



Feature "Amount" before and after normalization



Feature "Duration" before and after normalization

A similar script can be used for standardization:

m <- mean(credit$Age, na.rm = TRUE)

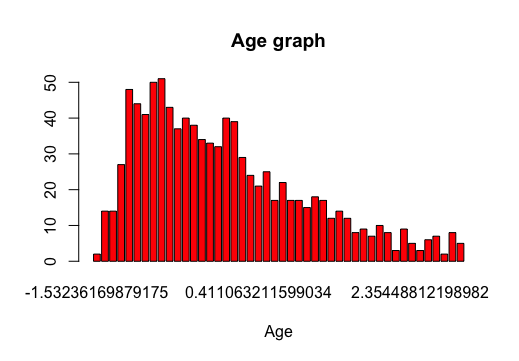
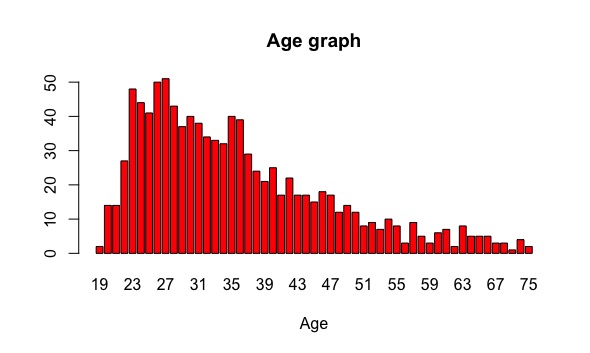
s <- sd(credit$Age, na.rm = TRUE)

for(i in 1:length(credit$Age)){

credit$Age[i] <- (credit$Age[i] - m) / s

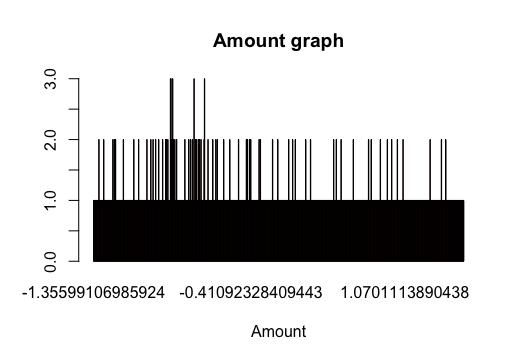
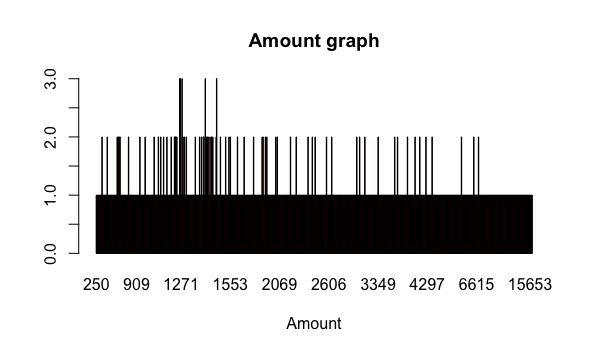
}

What follows is the result for the feature “Age”:

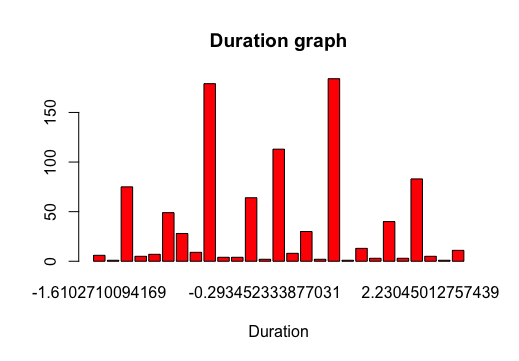
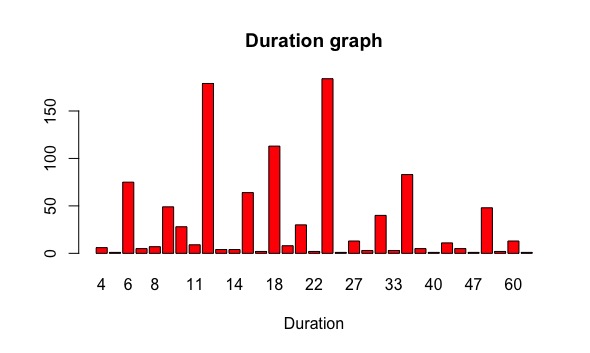


Feature "Age" before and after standardization

The same results can be obtained with the other numerical feature:



Feature "Amount" before and after standardization



Feature "Duration" before and after standardization

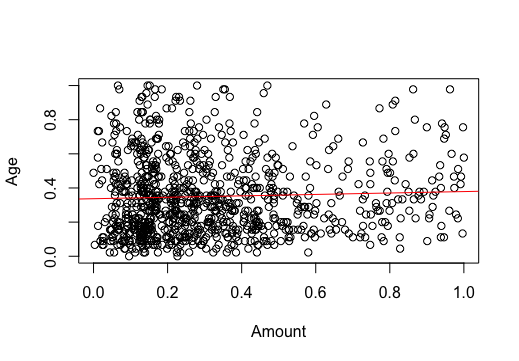
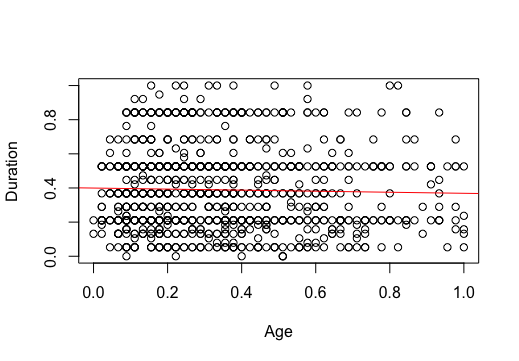
In every example we have used features without outliers.

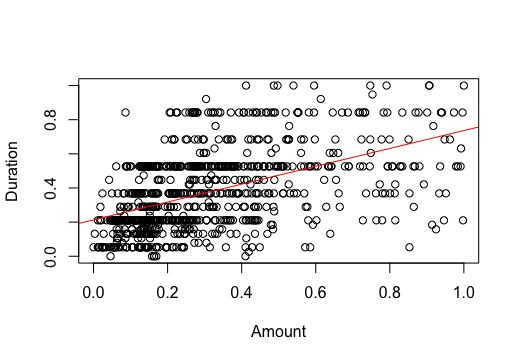
1. *Explore the characteristics of features: central tendency, spread measures and dependency (the technique depends on which data type you have)*

What follow is a table with the information about central tendency and spread measures for each feature.

|  |  |  |
| --- | --- | --- |
| ***Feature*** | ***Central tendency*** | ***Spread measures*** |
| Duration | 20,9 (mean) | 12,05881 (st. deviation) |
| Amount | 3271 (mean) | 2822,737 (st. deviation) |
| InstallmentRatePercentage | 4 (mode) | 2 (IQR) |
| ResidenceDuration | 4 (mode) | 2 (IQR) |
| Age | 42 (median) | 48 (IQR) |
| NumberExistingCredits | 1 (mode) | 1 (IQR) |
| NumberPeopleMaintenance | 1 (mode) | 0 (IQR) |
| Telephone | 1 (mode) | 1 (IQR) |
| ForeignWorker | 1 (mode) | 0 (IQR) |

What follow are a series of scatter plot to show the relationship between pairs of normalized continuous features without outliers:





Relationship between numerical feature in pairs

To obtain this result we have used the following commands:

plot(credit$Amount, credit$Duration, xlab="Amount", ylab="Duration")

abline(lm(credit$Duration~credit$Amount), col = "red")

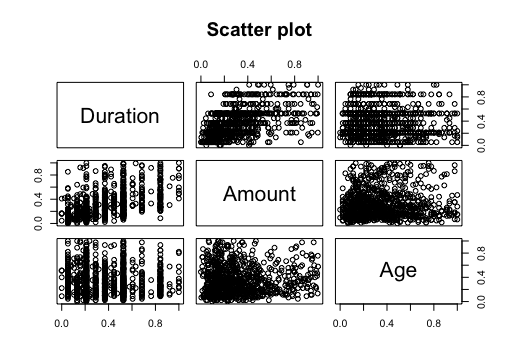
plot(credit$Amount, credit$Age, xlab="Amount", ylab="Age")

abline(lm(credit$Age~ credit$Amount), col = "red")

plot(credit$Age, credit$Duration, xlab="Age", ylab="Duration")

abline(lm(credit$Duration~ credit$Age), col = "red")

What follows is the scatter plot matrix:



Scatter plot of the numerical features

obtained by the following command:

pairs(~Duration+Amount+Age, data = credit, main = "Scatter plot")

We can also calculate the covariance matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Duration*** | ***Amount*** | ***Age*** |
| ***Duration*** | 83,04444 | 7.317,831 | -2,734148 |
| ***Amount*** | 7.317,831 | 3.201.557 | 786,8407 |
| ***Age*** | -2,734148 | 786,8407 | 105,907 |

In this case, we have used the values not normalized because the possible values of the covariance matrix are between [-∞, +∞] and we can obtain more significant results.

Looking to the covariance matrix we can say that between “Age” and “Duration” there isn’t a relationship, between “Amount” and “Age” there is a positive linear relationship and also, we have a positive relationship between “Duration” and “Amount”.

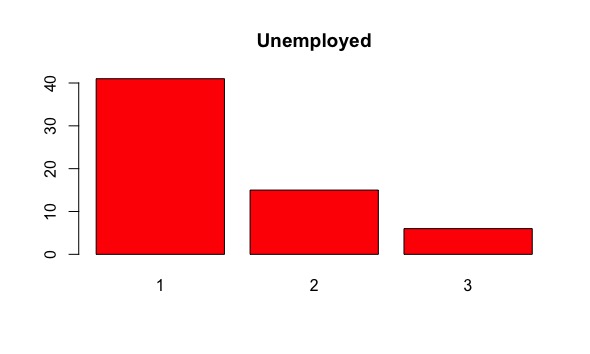
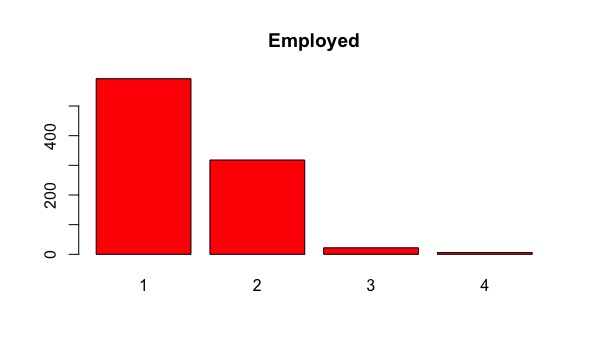
We also have added the correlation matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Duration*** | ***Amount*** | ***Age*** |
| ***Duration*** | 1 | 0,5017144 | -0,02942342 |
| ***Amount*** | 0,5017144 | 1 | 0,04286274 |
| ***Age*** | -0,02942342 | 0,04286274 | 1 |

In this case we have used the normalized values, since the covariance matrix is the same with normalized or not normalized values.

Looking to the correlation matrix we can say that between “Amount” and “Duration” there is a positive correlation (since the values is in the middle, between 0 and 1), between “Amount” and “Age” there isn’t a correlation (in particular there isn’t a linear dependency, but we don’t know nothing about other type of dependency) and also between “Age and “Duration”.

What follows is the relationship between two binary features, one describes if a person is employed or not and the other the number of credits with the bank.



Number of existing credits depending on the employed status

The commands used are:

unempl <- subset(credit, EmploymentDuration.Unemployed > 0)

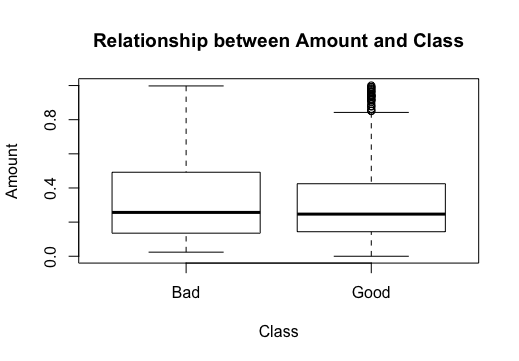
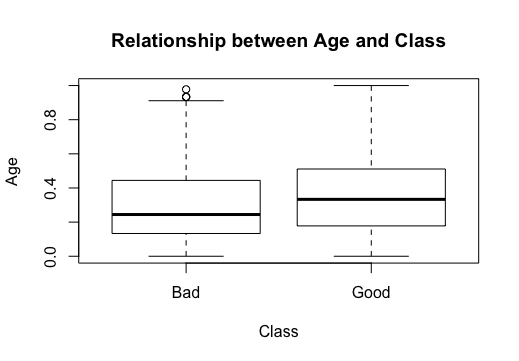
empl <- subset(credit, EmploymentDuration.Unemployed < 1)

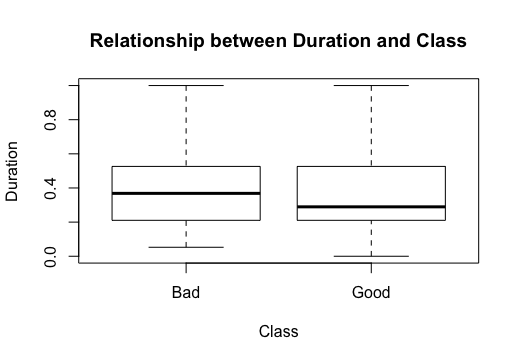
barplot(table(unempl$NumberExistingCredits), col = "red", main = "Unemployed")

barplot(table(empl$NumberExistingCredits), col = "red", main = "Employed")

Looking these histograms, we can understand that there isn’t a relationship between people employed/unemployed and the presence of credits. In fact, the majority of people have one credit irrespective of whether they’re employed or not.

Finally, we have realized some box plots to show the relationship between one categorical feature (Class) and continuous features “Age”, “Amount” and “Duration:





The command used is:

boxplot(credit$Age~credit$Class, data = credit,

main = "Relationship between Age and Class",

xlab = "Class", ylab = "Age")

Since all the boxplots aren’t separated we can say that there isn’t relationship between them, so we can’t understand nothing about the class with the information about “Age”, “Duration” and “Amount”.

Another possibility to evaluate the relationship from the boxplot is to calculate the following value:

If the sample size is 1000, if the percentage is over 10% there isn’t a relationship between the two boxplots.

For our specific case, we have obtained the following results for the relationship between Age and Class: 23,53%. So, there isn’t a relationship even in this case.

This procedure can’t be applied for the relationship between Amount/Class and Duration/Class since one boxplot is completely contained into the other.

1. *Append the data set with two derived features of different types (at least 2 out of 4)*

One possible derived feature is to aggregate the four features “CheckingAccountStatus.lt.0”, “CheckingAccountStatus.0.to.200”, “CheckingAccountStatus.gt.200” and “CheckingAccountStatus.none” into one categorical feature with the following possible values:

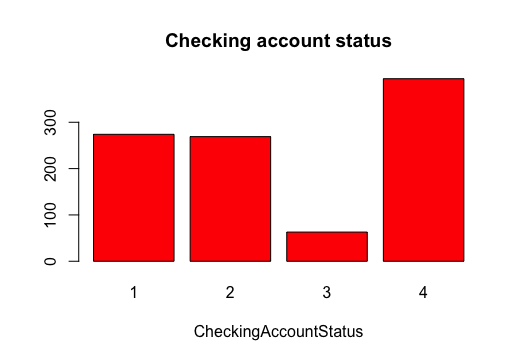
1 🡪 less then 0;

2 🡪 from 0 to 200;

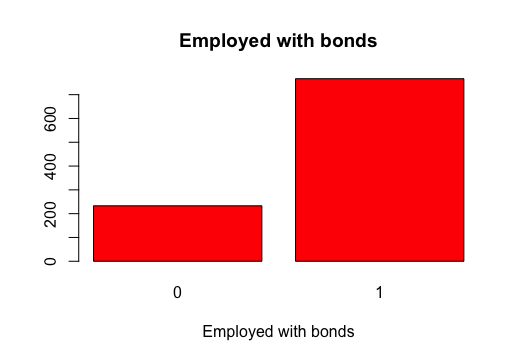
3 🡪 grater then 200;

4 🡪 none.

The following image is the histogram corresponding to the new columns:



The second possible derived feature shows all the employed people with bonds:



From the feature “EmploymentDuration.Unemployed” we know that there are 938 people employed.

Applying the logic operator NAND between this feature and the feature “SavingAccountBonds.Unkown” we can calculate the number of the employed with bonds. This could be interesting for statistic reason.

Looking the histogram, we can understand that 767 employed people have bonds; the other 233 are unemployed with bonds, employed without bonds or unemployed without bonds.