CREDIT RISK ANALYSIS

Tushar Khatri

06/09/2022

# Introduction

The kernel is taken from Lao Tse under the CC0 public licence (usability 7/10). Data is related to the credit market. Feel free to modify, discuss or improve the code.

# Aims of the analysis

The aims of the analysis is to search for statistical relationships that could give us some insigths about   
the risk of credit for both lenders and borrowers. Given that the variables and the context of the dataset are  
not fully explained, we are not going to take into account external macroeconomic events, which could completely  
change the results of the analysis. For instance, different countries could have different loan's requirement,   
or different phases of the economic cycle could end up in a very different scenario. Therefore, the aim of the   
analysis is purely descriptive and should be adapted according to circumstances.

# Loading libraries

library(haven)

## Warning: package 'haven' was built under R version 4.1.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

library(forcats)

## Warning: package 'forcats' was built under R version 4.1.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble 3.1.7 v purrr 0.3.4  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'purrr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.1.3

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

library(tidyverse)  
library(ggExtra)

## Warning: package 'ggExtra' was built under R version 4.1.3

### The dataset

credit\_risk\_dataset <- read.csv("C:/Users/TUSHAR/Downloads/Credit Risk Analysis EDA.csv")

### Transforming variables to factors

credit\_risk\_dataset$loan\_status <- as.factor(credit\_risk\_dataset$loan\_status)  
credit\_risk\_dataset$person\_home\_ownership <- as.factor(credit\_risk\_dataset$person\_home\_ownership)  
credit\_risk\_dataset$loan\_intent <- as.factor(credit\_risk\_dataset$loan\_intent)  
credit\_risk\_dataset$loan\_grade <- as.factor(credit\_risk\_dataset$loan\_grade)  
credit\_risk\_dataset$cb\_person\_default\_on\_file <- as.factor(credit\_risk\_dataset$cb\_person\_default\_on\_file)  
glimpse(credit\_risk\_dataset) # more ordered

## Rows: 32,581  
## Columns: 12  
## $ person\_age <int> 22, 21, 25, 23, 24, 21, 26, 24, 24, 21, 22,~  
## $ person\_income <int> 59000, 9600, 9600, 65500, 54400, 9900, 7710~  
## $ person\_home\_ownership <fct> RENT, OWN, MORTGAGE, RENT, RENT, OWN, RENT,~  
## $ person\_emp\_length <int> 123, 5, 1, 4, 8, 2, 8, 5, 8, 6, 6, 2, 2, 4,~  
## $ loan\_intent <fct> PERSONAL, EDUCATION, MEDICAL, MEDICAL, MEDI~  
## $ loan\_grade <fct> D, B, C, C, C, A, B, B, A, D, B, A, A, E, A~  
## $ loan\_amnt <int> 35000, 1000, 5500, 35000, 35000, 2500, 3500~  
## $ loan\_int\_rate <dbl> 16.02, 11.14, 12.87, 15.23, 14.27, 7.14, 12~  
## $ loan\_status <fct> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0~  
## $ loan\_percent\_income <dbl> 0.59, 0.10, 0.57, 0.53, 0.55, 0.25, 0.45, 0~  
## $ cb\_person\_default\_on\_file <fct> Y, N, N, N, Y, N, N, N, N, N, N, N, N, N, N~  
## $ cb\_person\_cred\_hist\_length <int> 3, 2, 3, 2, 4, 2, 3, 4, 2, 3, 4, 2, 2, 4, 4~

#str(credit\_risk\_dataset)

# Data Description

## The dataset is composed by 32581 rows (observations) and 12 columns (variables).

person\_age: is the age of the person at the time of the loan.  
person\_income: is the yearly income of the person at the time of the loan.  
person\_home\_ownership: is the type of ownership of the home.  
person\_emp\_length: is the amount of time in years that person is employed.  
loan\_intent: is the aim of the loan.  
loan\_grade: is a classification system that involves assigning a quality score to a loan based  
 on a borrower's credit history, quality of the collateral, and the likelihood of   
 repayment of the principal and interest.  
loan\_amnt: is the dimension of the loan taken.  
loan\_int\_rate: is the interest paid for the loan.  
loan\_status: is a dummy variable where 1 is default, 0 is not default.  
loan\_percent\_income: is the ratio between the loan taken and the annual income.  
cb\_person\_default\_on\_file: answers whether the person has defaulted before.  
cb\_person\_cred\_hist\_length: represents the number of years of personal history since the   
 first loan taken from that person.

### The first thing to do is to clean the dataset, looking for missing values (NAs) and for inconsistent data.

# To spot missing values  
summary(credit\_risk\_dataset)

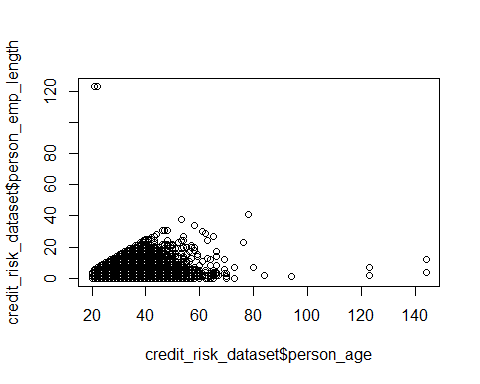
## person\_age person\_income person\_home\_ownership person\_emp\_length  
## Min. : 20.00 Min. : 4000 MORTGAGE:13444 Min. : 0.00   
## 1st Qu.: 23.00 1st Qu.: 38500 OTHER : 107 1st Qu.: 2.00   
## Median : 26.00 Median : 55000 OWN : 2584 Median : 4.00   
## Mean : 27.73 Mean : 66075 RENT :16446 Mean : 4.79   
## 3rd Qu.: 30.00 3rd Qu.: 79200 3rd Qu.: 7.00   
## Max. :144.00 Max. :6000000 Max. :123.00   
## NA's :895   
## loan\_intent loan\_grade loan\_amnt loan\_int\_rate   
## DEBTCONSOLIDATION:5212 A:10777 Min. : 500 Min. : 5.42   
## EDUCATION :6453 B:10451 1st Qu.: 5000 1st Qu.: 7.90   
## HOMEIMPROVEMENT :3605 C: 6458 Median : 8000 Median :10.99   
## MEDICAL :6071 D: 3626 Mean : 9589 Mean :11.01   
## PERSONAL :5521 E: 964 3rd Qu.:12200 3rd Qu.:13.47   
## VENTURE :5719 F: 241 Max. :35000 Max. :23.22   
## G: 64 NA's :3116   
## loan\_status loan\_percent\_income cb\_person\_default\_on\_file  
## 0:25473 Min. :0.0000 N:26836   
## 1: 7108 1st Qu.:0.0900 Y: 5745   
## Median :0.1500   
## Mean :0.1702   
## 3rd Qu.:0.2300   
## Max. :0.8300   
##   
## cb\_person\_cred\_hist\_length  
## Min. : 2.000   
## 1st Qu.: 3.000   
## Median : 4.000   
## Mean : 5.804   
## 3rd Qu.: 8.000   
## Max. :30.000   
##

# Exploratory Data Analysis

### Obervation from Dataset

Here we can notice a few important things as the summary of the variables gives us a first insight of the data:  
person\_age has a strange result: the max value is 144 which is unlikely to be true.  
person\_emp\_length has 895 NAs  
person\_emp\_length has a max value of 123 which is not possible.  
loan\_interest\_rate has 3116 NAs  
according to the data (loan\_status), 21.82% (7108/32581) of people have defaulted, value that is quite high.  
plot(credit\_risk\_dataset$person\_age, credit\_risk\_dataset$person\_emp\_length)

plot(credit\_risk\_dataset$person\_age, credit\_risk\_dataset$person\_emp\_length)



### Interpretation

Here we can see clearly that there are some issues on data by looking at outliers. It is impossible that two 20   
years old have over 120 years of experience. Therefore we proceed by cleaning our dataset and removing illogical  
outliers. We have to adjust data properly before starting the exploratory phase.

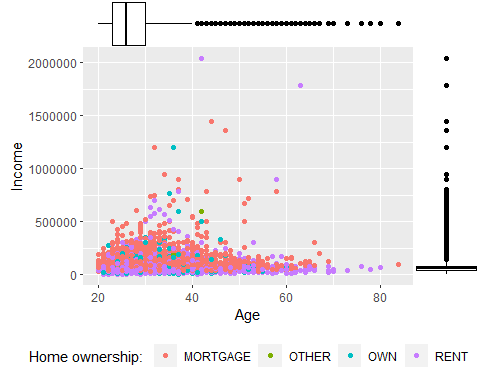
## Data Cleaning

credit\_risk <- credit\_risk\_dataset %>%  
 filter(!is.na(person\_emp\_length)) %>%  
 filter(!is.na(loan\_int\_rate)) %>%  
 filter(person\_age < 90) %>%  
 filter(person\_emp\_length < 100)  
  
summary(credit\_risk)

## person\_age person\_income person\_home\_ownership person\_emp\_length  
## Min. :20.00 Min. : 4000 MORTGAGE:11798 Min. : 0.00   
## 1st Qu.:23.00 1st Qu.: 39456 OTHER : 94 1st Qu.: 2.00   
## Median :26.00 Median : 55900 OWN : 2192 Median : 4.00   
## Mean :27.71 Mean : 66427 RENT :14548 Mean : 4.78   
## 3rd Qu.:30.00 3rd Qu.: 80000 3rd Qu.: 7.00   
## Max. :84.00 Max. :2039784 Max. :41.00   
##   
## loan\_intent loan\_grade loan\_amnt loan\_int\_rate   
## DEBTCONSOLIDATION:4565 A:9401 Min. : 500 Min. : 5.42   
## EDUCATION :5702 B:9149 1st Qu.: 5000 1st Qu.: 7.90   
## HOMEIMPROVEMENT :3198 C:5697 Median : 8000 Median :10.99   
## MEDICAL :5293 D:3247 Mean : 9655 Mean :11.04   
## PERSONAL :4875 E: 870 3rd Qu.:12500 3rd Qu.:13.48   
## VENTURE :4999 F: 209 Max. :35000 Max. :23.22   
## G: 59   
## loan\_status loan\_percent\_income cb\_person\_default\_on\_file  
## 0:22430 Min. :0.0000 N:23530   
## 1: 6202 1st Qu.:0.0900 Y: 5102   
## Median :0.1500   
## Mean :0.1695   
## 3rd Qu.:0.2300   
## Max. :0.8300   
##   
## cb\_person\_cred\_hist\_length  
## Min. : 2.000   
## 1st Qu.: 3.000   
## Median : 4.000   
## Mean : 5.794   
## 3rd Qu.: 8.000   
## Max. :30.000   
##

Now the dataset seems better.  
  
With the summary command we extrapolate some information about tha characteristics and the distribution of the   
variables that compose our data. For instance, looking at the nature of variables we can divide them in two classes:  
numerical and categorical.  
  
We could also divide the variables conceptually. For instance, person\_age, person\_income, person\_home\_ownership and  
person\_emp\_length rely on personal characteristics of people, while loan\_intent, loan\_grade, loan\_amnt, loan\_int\_rate  
, loan\_status and loan\_percent\_income rely on their credit activity. The last two variables are of interest only for  
secondary purposes.  
  
With this in mind, we can proceed with the exploration of statistical relationships by looking at the most influential  
variables.

plot1 <- ggplot(credit\_risk, aes(person\_age, person\_income, col=person\_home\_ownership)) +  
 geom\_point() + ylab("Income") + xlab("Age") + labs(color = "Home ownership: ") +  
 theme(legend.position="bottom")   
ggExtra::ggMarginal(plot1, type = "boxplot")



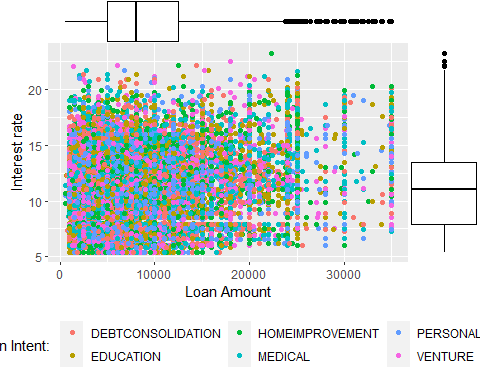
### Interpretation of Scatter Plot 1

The first graph that we have drawn is merely descriptive of people's characteristics. By looking at it we can notice a few  
important things:  
  
The distribution of the Age is right-skewed. This means that most of the values lie on the right side of the scale. In   
other words, considering the age, the dataset is populated mostly by young people.  
The distribution of the income is right-skewed. Similarly to age, most of values are in the right side of the graph, which  
in this case indicates that the distribution of income is not balanced and shows economic inequality.  
People that has a mortage seems to have an higher income on average, with respect to people that pays a rent.

### Familiarize with the dataset

First of all, we start observing statistical relationships between numerical variables. We can do this using traditional  
scatterplots and boxplots. To understand the personal characteristics of people and then their credit activity, it seems  
reasonable to compare the person age and income, the loan amount and the interest rate, the person income with the loan  
amount and finally, the person age and the interest rate. We could also add an additional dimension (with colours) to have  
a clearer overview.

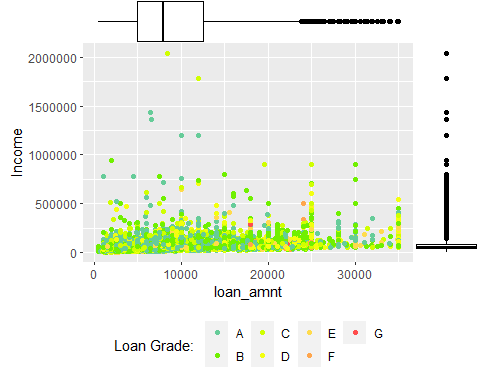
plot2 <- ggplot(credit\_risk, aes(loan\_amnt, loan\_int\_rate, col=loan\_intent)) +  
 geom\_point() + ylab("Interest rate") + xlab("Loan Amount") + labs(color = "Loan Intent: ") +  
 theme(legend.position="bottom")   
ggExtra::ggMarginal(plot2, type = "boxplot")



### Interpretation

This second graph represents the relationship between the loan amount taken by people and the interest rate applied.  
Here, we can see that there is not a clear correlation between the two variables, thus we can say that they do not   
depend from each other. Moreover, it seems that the loan intent is distributed equally in the sample, which means that  
the intent does not influence at all interest rates or the loan amount. In this case distributions are normal Gaussian  
with a slight skewness.

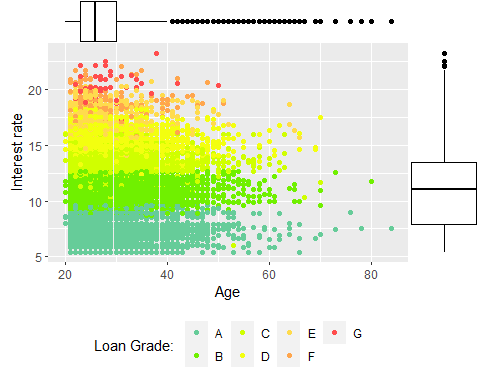
plot3 <- ggplot(credit\_risk, aes(loan\_amnt, person\_income, col=loan\_grade)) +  
 geom\_point() + ylab("Income") + labs(color = "Loan Grade: ") + theme(legend.position="bottom") +  
 scale\_color\_manual(values=c("#66cc99", "#70F000", "#D0FF00", "#F3FF0F", "#FFDB4D", "#FFA64D", "#FF4D4D"))  
ggExtra::ggMarginal(plot3, type = "boxplot")



### Interpretation

Also in this case there is not a clear relationship between income and the loan amount. By adding different colors for  
the loan grades we notice a small tendence for people with low income and with a high loan, to have a lower grade.   
However, the statistical relationship is not strong.

plot4 <- ggplot(credit\_risk, aes(person\_age, loan\_int\_rate, col=loan\_grade)) +  
 geom\_point() + ylab("Interest rate") + xlab("Age") + labs(color = "Loan Grade: ") +  
theme(legend.position="bottom") + scale\_color\_manual(values=c("#66cc99", "#70F000", "#D0FF00", "#F3FF0F", "#FFDB4D", "#FFA64D", "#FF4D4D"))  
ggExtra::ggMarginal(plot4, type = "boxplot")



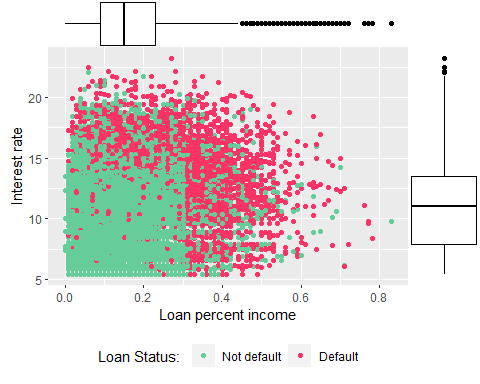
### Interpretation

This graph is more interesting: it shows the relationship between the age, the interest rate and the loan grade. Here we can clearly see that the loan grade depends on the interest rate which divide the graph in multiple sections. Basically it says that the higher the interest rate, the higher is the risk perceived. This could be due to the method of evaluation of the risk grade of people. Furthermore, we notice that people in the eldery age tend to have lower interest rates on average with respect to younger people.

### Variable of interest

Our variable of interest is the likekihood of default given certain characteristics, therefore we look at the loan\_status variable.

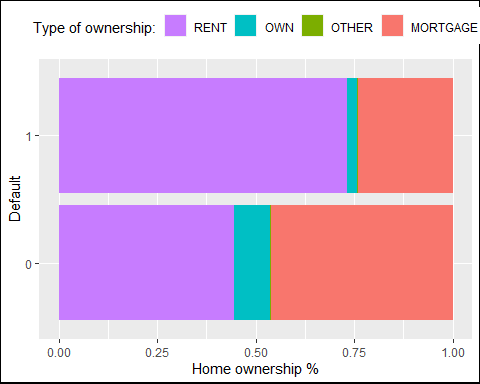
plot5 <- ggplot(credit\_risk, aes(loan\_percent\_income, loan\_int\_rate, col= factor(loan\_status, labels = c("Not default", "Default")))) +  
 geom\_point() + ylab("Interest rate") + xlab("Loan percent income") + labs(color = "Loan Status: ") +   
 theme(legend.position="bottom") + scale\_color\_manual(values=c("#66cc99", "#ff3366"))  
ggExtra::ggMarginal(plot5, type = "boxplot")

 ### Interpretation

```Here we consider again the level of interest rates, in relation to the ratio between the loan amount and income. The shape of the dots suggests that there isn’t a clear relationship between these two varibles, which means that interest rates does not depend at all on the loan percent income. However, the more we approach to an higher ratio, the lower the interest rates become.

Then, if we take into account people that have defaulted in red and people that have not defaulted in green, we can clearly see a rectangular shape with threshold levels towards 12% for interest rates, and 0.3 for the loan percent income variable. This area could be considered less risky for both lenders and borrowers. This seems to be reasonable as higher interest rates and an higher income/loan ratio means that people are less likely to repay their debt.```

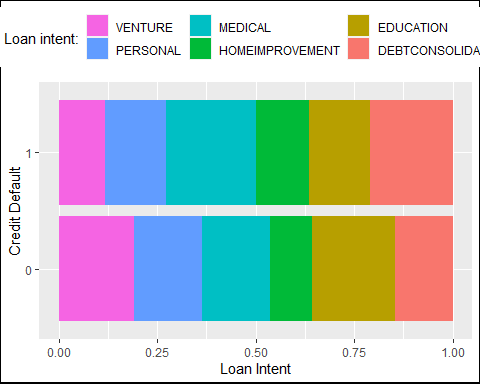
ggplot(credit\_risk, aes(x = factor(loan\_status), fill = factor(person\_home\_ownership))) +  
 geom\_bar(position = "fill") +  
 ylab("Home ownership %") + xlab("Default") + labs(fill = "Type of ownership:") +  
 theme(legend.position="top", plot.background = element\_rect(colour = "black", size = 1)) +   
 guides(fill = guide\_legend(reverse=TRUE)) +  
 coord\_flip()



### Interpretation

This graph shows the absolute frequencies of the type of home ownership with respect to people that has defaulted. It is possible to notice that tipically who pays a rent is more likely to default considering all the orher classes.

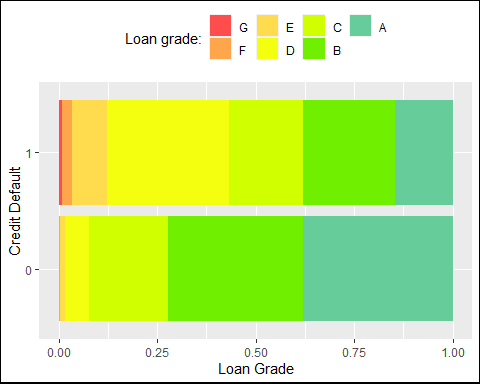
ggplot(credit\_risk, aes(x = factor(loan\_status), fill = factor(loan\_intent))) +  
 geom\_bar(position = "fill") +  
 ylab("Loan Intent") + xlab("Credit Default") + labs(fill = "Loan intent:") +  
 theme(legend.position="top", plot.background = element\_rect(colour = "black", size = 1)) +   
 guides(fill = guide\_legend(reverse=TRUE)) +  
 coord\_flip()



### Interpretation

Here we compare people that has and has not defaulted, to the purpose of the requested loan. As we have noted before, the loan intent variable is well distributed, however when the loan is dedicated to debts consolidation and for medical purposes, it seems more likely to lead to a default of the debt.

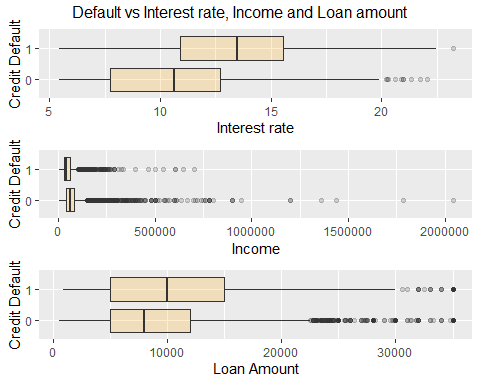
ggplot(credit\_risk, aes(x = factor(loan\_status), fill = factor(loan\_grade))) +  
 geom\_bar(position = "fill") + scale\_fill\_manual(values=c("#66cc99", "#70F000", "#D0FF00", "#F3FF0F", "#FFDB4D", "#FFA64D", "#FF4D4D")) +  
 ylab("Loan Grade") + xlab("Credit Default") + labs(fill = "Loan grade:") +  
 theme(legend.position="top", plot.background = element\_rect(colour = "black", size = 1)) +   
 guides(fill = guide\_legend(reverse=TRUE)) +  
 coord\_flip()



### Interpretation

Using again the color palette for the risk grade, it is possible to notice that this variable has some predictive power for defaults, thus, it seems to be selected appropriately from the data keeper. According to the data, when the risk grade is higher, it is less likely to have a default. The contrary applies for lower grades.

g1 <- ggplot(credit\_risk, aes(x = factor(loan\_status), y = loan\_int\_rate)) +   
 geom\_boxplot(fill = "orange", alpha = 0.2) +  
 ylab("Interest rate") + xlab("Credit Default") +  
 coord\_flip()   
  
g2 <- ggplot(credit\_risk, aes(x = factor(loan\_status), y = person\_income)) +   
 geom\_boxplot(fill = "orange", alpha = 0.2) +  
 ylab("Income") + xlab("Credit Default") +  
 coord\_flip()   
  
g3 <- ggplot(credit\_risk, aes(x = factor(loan\_status), y = loan\_amnt)) +   
 geom\_boxplot(fill = "orange", alpha = 0.2) +  
 ylab("Loan Amount") + xlab("Credit Default") +  
 coord\_flip()   
  
  
grid.arrange(g1, g2, g3, ncol = 1, nrow = 3, top = "Default vs Interest rate, Income and Loan amount")



### Interpretation

Finally we compare the risk of default with three continous variables: the interest rates, the yearly income, and the loan amount. This graph is interesting because it shows clearly that higher interest rates, a lower income and higher credit increase the possibility to have a default.

# Inferential statistics

We use information derived from the descriptive statistics to raise the right questions about our dataset, searching for the causal relationship between the regressands and regressors of our interest. Then we use inferential statistics to search confirmation about our hypothesis and to find the best model that fits with the data through statistical relevance

``` Because of the fact that we are looking for a binary dependent variable (loan\_status), we will use the logit model for statistical predictions. As independent variables, we initially use all the other columns.

# Fitting Logistic Regression to Predict the probability of Default  
  
  
```r  
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.1.3

logit <- glm(loan\_status ~ ., data = credit\_risk, family = binomial(link = 'logit'))  
summary(logit)

##   
## Call:  
## glm(formula = loan\_status ~ ., family = binomial(link = "logit"),   
## data = credit\_risk)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2671 -0.5188 -0.3003 -0.1238 3.3320   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.220e+00 2.050e-01 -20.590 < 2e-16 \*\*\*  
## person\_age -1.114e-02 6.324e-03 -1.761 0.07815 .   
## person\_income 1.969e-06 5.391e-07 3.652 0.00026 \*\*\*  
## person\_home\_ownershipOTHER 4.184e-01 3.006e-01 1.392 0.16399   
## person\_home\_ownershipOWN -1.795e+00 1.130e-01 -15.878 < 2e-16 \*\*\*  
## person\_home\_ownershipRENT 8.304e-01 4.295e-02 19.335 < 2e-16 \*\*\*  
## person\_emp\_length -1.476e-02 5.091e-03 -2.899 0.00374 \*\*   
## loan\_intentEDUCATION -8.727e-01 6.109e-02 -14.285 < 2e-16 \*\*\*  
## loan\_intentHOMEIMPROVEMENT 4.745e-02 6.779e-02 0.700 0.48400   
## loan\_intentMEDICAL -1.530e-01 5.771e-02 -2.651 0.00802 \*\*   
## loan\_intentPERSONAL -6.393e-01 6.242e-02 -10.243 < 2e-16 \*\*\*  
## loan\_intentVENTURE -1.142e+00 6.656e-02 -17.154 < 2e-16 \*\*\*  
## loan\_gradeB 1.099e-01 8.282e-02 1.327 0.18453   
## loan\_gradeC 2.407e-01 1.246e-01 1.932 0.05335 .   
## loan\_gradeD 2.332e+00 1.565e-01 14.902 < 2e-16 \*\*\*  
## loan\_gradeE 2.497e+00 1.974e-01 12.652 < 2e-16 \*\*\*  
## loan\_gradeF 2.794e+00 2.725e-01 10.252 < 2e-16 \*\*\*  
## loan\_gradeG 6.327e+00 1.053e+00 6.009 1.87e-09 \*\*\*  
## loan\_amnt -1.062e-04 4.822e-06 -22.030 < 2e-16 \*\*\*  
## loan\_int\_rate 8.863e-02 1.813e-02 4.890 1.01e-06 \*\*\*  
## loan\_percent\_income 1.344e+01 2.810e-01 47.853 < 2e-16 \*\*\*  
## cb\_person\_default\_on\_fileY 2.006e-02 5.309e-02 0.378 0.70552   
## cb\_person\_cred\_hist\_length 8.908e-03 9.566e-03 0.931 0.35173   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 29925 on 28631 degrees of freedom  
## Residual deviance: 19206 on 28609 degrees of freedom  
## AIC: 19252  
##   
## Number of Fisher Scoring iterations: 6

### Interpretation

As expected, cb\_person\_default\_on\_file and cb\_person\_cred\_hist\_length are not statistically meaningful, thus, they should be removed from the model. We are going to exclude also the variable loan\_percent\_income to avoid any kind of multicollinearity. This is because this measure is only the ratio between income and loan amount, therefore it could create redundance on calculation.

logit <- glm(loan\_status ~ person\_age + person\_income + person\_emp\_length +  
 loan\_amnt + loan\_int\_rate + person\_home\_ownership + loan\_intent +  
 loan\_grade, data = credit\_risk, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(logit)

##   
## Call:  
## glm(formula = loan\_status ~ person\_age + person\_income + person\_emp\_length +   
## loan\_amnt + loan\_int\_rate + person\_home\_ownership + loan\_intent +   
## loan\_grade, family = binomial, data = credit\_risk)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.4168 -0.5966 -0.3346 -0.0874 6.4908   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.594e+00 1.660e-01 -9.604 < 2e-16 \*\*\*  
## person\_age -5.981e-03 2.952e-03 -2.026 0.042764 \*   
## person\_income -3.450e-05 8.830e-07 -39.068 < 2e-16 \*\*\*  
## person\_emp\_length -1.641e-02 4.910e-03 -3.342 0.000833 \*\*\*  
## loan\_amnt 1.141e-04 3.311e-06 34.461 < 2e-16 \*\*\*  
## loan\_int\_rate 6.240e-02 1.735e-02 3.596 0.000324 \*\*\*  
## person\_home\_ownershipOTHER 6.366e-01 2.885e-01 2.206 0.027375 \*   
## person\_home\_ownershipOWN -1.356e+00 1.049e-01 -12.930 < 2e-16 \*\*\*  
## person\_home\_ownershipRENT 8.759e-01 4.167e-02 21.021 < 2e-16 \*\*\*  
## loan\_intentEDUCATION -8.182e-01 5.773e-02 -14.173 < 2e-16 \*\*\*  
## loan\_intentHOMEIMPROVEMENT 6.223e-02 6.510e-02 0.956 0.339079   
## loan\_intentMEDICAL -2.010e-01 5.471e-02 -3.674 0.000239 \*\*\*  
## loan\_intentPERSONAL -5.880e-01 5.909e-02 -9.951 < 2e-16 \*\*\*  
## loan\_intentVENTURE -1.022e+00 6.279e-02 -16.268 < 2e-16 \*\*\*  
## loan\_gradeB 1.301e-01 7.849e-02 1.658 0.097332 .   
## loan\_gradeC 2.977e-01 1.161e-01 2.566 0.010299 \*   
## loan\_gradeD 2.230e+00 1.478e-01 15.083 < 2e-16 \*\*\*  
## loan\_gradeE 2.434e+00 1.890e-01 12.878 < 2e-16 \*\*\*  
## loan\_gradeF 2.755e+00 2.742e-01 10.046 < 2e-16 \*\*\*  
## loan\_gradeG 7.884e+00 1.810e+00 4.357 1.32e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 29925 on 28631 degrees of freedom  
## Residual deviance: 20842 on 28612 degrees of freedom  
## AIC: 20882  
##   
## Number of Fisher Scoring iterations: 7

Now we receive a warning message: “glm.fit: fitted probabilities numerically 0 or 1 occurred”. After some research, I discovered that the problem lies on extreme values on one variable, that causes to the predicted probabilities to be indistinguishable from 0 or 1, leading to calculation's problems. We could overcome to this issue by deleting outliers from person\_income and running the algorithm again.

credit\_risk <- credit\_risk %>%  
 filter(person\_income < 800000)  
  
logit <- glm(loan\_status ~ person\_age + person\_income + person\_emp\_length +  
 loan\_amnt + loan\_int\_rate + person\_home\_ownership + loan\_intent +  
 loan\_grade, data = credit\_risk, family = binomial)  
summary(logit)

##   
## Call:  
## glm(formula = loan\_status ~ person\_age + person\_income + person\_emp\_length +   
## loan\_amnt + loan\_int\_rate + person\_home\_ownership + loan\_intent +   
## loan\_grade, family = binomial, data = credit\_risk)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.4168 -0.5968 -0.3347 -0.0883 6.4908   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.594e+00 1.660e-01 -9.604 < 2e-16 \*\*\*  
## person\_age -5.981e-03 2.952e-03 -2.026 0.042764 \*   
## person\_income -3.450e-05 8.830e-07 -39.068 < 2e-16 \*\*\*  
## person\_emp\_length -1.641e-02 4.910e-03 -3.342 0.000833 \*\*\*  
## loan\_amnt 1.141e-04 3.311e-06 34.461 < 2e-16 \*\*\*  
## loan\_int\_rate 6.240e-02 1.735e-02 3.596 0.000324 \*\*\*  
## person\_home\_ownershipOTHER 6.366e-01 2.885e-01 2.206 0.027375 \*   
## person\_home\_ownershipOWN -1.356e+00 1.049e-01 -12.930 < 2e-16 \*\*\*  
## person\_home\_ownershipRENT 8.759e-01 4.167e-02 21.021 < 2e-16 \*\*\*  
## loan\_intentEDUCATION -8.182e-01 5.773e-02 -14.173 < 2e-16 \*\*\*  
## loan\_intentHOMEIMPROVEMENT 6.223e-02 6.510e-02 0.956 0.339079   
## loan\_intentMEDICAL -2.010e-01 5.471e-02 -3.674 0.000239 \*\*\*  
## loan\_intentPERSONAL -5.880e-01 5.909e-02 -9.951 < 2e-16 \*\*\*  
## loan\_intentVENTURE -1.022e+00 6.279e-02 -16.268 < 2e-16 \*\*\*  
## loan\_gradeB 1.301e-01 7.849e-02 1.658 0.097332 .   
## loan\_gradeC 2.977e-01 1.161e-01 2.566 0.010299 \*   
## loan\_gradeD 2.230e+00 1.478e-01 15.083 < 2e-16 \*\*\*  
## loan\_gradeE 2.434e+00 1.890e-01 12.878 < 2e-16 \*\*\*  
## loan\_gradeF 2.755e+00 2.742e-01 10.046 < 2e-16 \*\*\*  
## loan\_gradeG 7.884e+00 1.810e+00 4.357 1.32e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 29920 on 28620 degrees of freedom  
## Residual deviance: 20842 on 28601 degrees of freedom  
## AIC: 20882  
##   
## Number of Fisher Scoring iterations: 7

### Interpretation

```Deviance residuals look good, even though values are not completely centered on zero and are not symmetrical. The estimates are the parameters of our interest while the Pr(>|z|) column shows the two-tailed p-values testing the null hypothesis that the coefficient is equal to zero. In other words, it shows the significance of the effects for each independent variable. In our model it seems that most of the variable are statistically significant, however, to confirm this hypotesis we proceed by using the anova test, looking at the table of deviance

The difference between the null deviance and the residual deviance shows how our model is doing against the null model, that is a model with only the intercept. The anova test says that adding these variables improves the model significantly, indeed, low p-values means that variables explain a large portion of variability.```

## Interpreting resutls

Now we can analyze the fitting and interpret what the model is telling us. As we have seen, the model suggests that most of p-values show statistical relevance on results, apart for the category HOMEIMPROVEMENT of the loan intent and the loan grade B, which are not significative.  
  
The Estimate column represents the coefficients in log-odds form. This means that when we increase the age by one unit, we could expect a change in the log-odds about -0.005981. We could also use these information looking at the sign of each estimate, to understend whether the effect of the predictor is positive or negative.

cbind(Estimate=round(coef(logit),5),  
 OR=round(exp(coef(logit)),5))

## Estimate OR  
## (Intercept) -1.59423 0.20307  
## person\_age -0.00598 0.99404  
## person\_income -0.00003 0.99997  
## person\_emp\_length -0.01641 0.98373  
## loan\_amnt 0.00011 1.00011  
## loan\_int\_rate 0.06240 1.06439  
## person\_home\_ownershipOTHER 0.63656 1.88996  
## person\_home\_ownershipOWN -1.35596 0.25770  
## person\_home\_ownershipRENT 0.87586 2.40095  
## loan\_intentEDUCATION -0.81818 0.44123  
## loan\_intentHOMEIMPROVEMENT 0.06223 1.06421  
## loan\_intentMEDICAL -0.20100 0.81791  
## loan\_intentPERSONAL -0.58798 0.55545  
## loan\_intentVENTURE -1.02152 0.36005  
## loan\_gradeB 0.13012 1.13897  
## loan\_gradeC 0.29774 1.34681  
## loan\_gradeD 2.22950 9.29523  
## loan\_gradeE 2.43363 11.40020  
## loan\_gradeF 2.75479 15.71777  
## loan\_gradeG 7.88417 2654.91157

```Based on the output that we get, when for example the interest rate parameter increases by one unit (keeping all other predictors constant), the odds of y = 1 (default) are 1.0644 higher, or, put differently, they increase about 6.44%. The same logic applies for all other variables, so for example increasing the employment length by one unit lead to a decrease in the odds of defaulting of 0.9837.

An element to be taken into consideration in our analysis is the loan\_grade variable. Indeed, here we can notice that as long as we approach to lower grades, the odds of default increase enormously. As we have said before, categorical variables should be interpreted with respect to the base category, so in this case by comparing other levels with the grade A. This variable seems to be very influential and that, of course, depends on the method of evaluation. It could be interesting to look at how they are classified, but that’s for another research. For now, we could just say that the graduation is appropriate and is helpful to identify possible default risks for borrowers and lenders.```

# Confusion Matrix

## Predicted probabilities

When developing models for prediction, the most critical metric is regarding how well the model does in predicting the target variable on out-of-sample observations. For that purpose we can compare the predicted target variable versus the observed values through the confusion matrix.```

## Precision of Model Fitted using Confusion Matrix

predicted <- predict(logit, type = "response")  
table(credit\_risk$loan\_status, predicted > 0.5)

##   
## FALSE TRUE  
## 0 21573 846  
## 1 3359 2843

table(credit\_risk$loan\_status, predicted > 0.5) %>% prop.table()

##   
## FALSE TRUE  
## 0 0.75374725 0.02955872  
## 1 0.11736138 0.09933266

Confusion\_matrix = as.matrix(table(credit\_risk$loan\_status, predicted > 0.5))  
Sensitivity\_ =Confusion\_matrix[1,1]/(Confusion\_matrix[1,1]+Confusion\_matrix[2,1])  
  
Specificity\_ = Confusion\_matrix[2,2]/(Confusion\_matrix[1,2]+Confusion\_matrix[2,2])

# Calculating Speficity and Selectivuty using Confusion matrix

Sensitivity\_

## [1] 0.8652735

Specificity\_

## [1] 0.7706696

## Conclusions

In our analysis we tried to use statistical tools such as the logistic regression to see if it is possible to predict the possibility of default for customers. First of all we have cleaned and organized our dataset by removing inconsistent values. Then, we proceeded by exploring relationships between most important variables. In this way, we discovered that most of the people in the sample is young and that the distribution of income tend to be inequal. There are not clear relationships between most of the variables, however, we found that the loan grade is strictly connected to the interest rate, so this lead us thinking that this variable depends on the method of evaluation. Nevertheless, it has a good explanatory power, therefore we leaved it in our model. By looking at descriptive statistics we also found that there is a specific area that is less risky for both borrowers and lenders and that area ranges below the 12% thresold for interest rates and below 0.3 for the income to loan ratio. Then, we have build a model which could be able to predict with an high rate, the probabilty of default of people. We have seen that the most influential variables are the interest rates, and the categorical variable of the loan grade. This reflects what we found in the exploratory phase. Also the other variables are influential and most of them are statistically significant. In conclusion, by applying the model, we are able to predict correctly about 85.3% of the cases. This is not bad, considering that the inital rate of default was 21.82%. Therefore, the model could be useful to help save money for both borrowers and lenders, making the credit market more efficient.

# Insights of above analysis done

1) Analyzed statistical relationships on data with 32581 rows and 12 variables to gain insights into credit risk  
2) Carried out Exploratory Data Analysis of Credit Risk data using ggplot2, tidyverse, and dplyr packages  
3) Fitted a Binary Classification model using Logistic Regression to predict the probability of client default  
4) Diagnosed the precision and accuracy of the Logistic Regression model using sensitivity and specificity