Contents

[**1. Introduction** 3](#_Toc180573733)

[**1.1 Data Column Identification** 3](#_Toc180573734)

[**1.2 Summary of Data Columns:** 3](#_Toc180573735)

[**1.3 Hypothesis and Objectives** 4](#_Toc180573736)

[**2. Documentation** 6](#_Toc180573737)

[**2.1 Exploratory Data Analysis** 6](#_Toc180573738)

[**2.2 Data Summary** 12](#_Toc180573739)

[**2.3 Libraries we will be using** 14](#_Toc180573740)

[**2.4 Fundamental programming techniques we will use** 14](#_Toc180573741)

[**2.5 Types of graphs we will be working on** 14](#_Toc180573742)

[**3.** **Objectives** 15](#_Toc180573743)

[**3.1 Objective 1: Distribution of various features and customer class analysis** 16](#_Toc180573744)

[**3.1.1 Analyze the distribution of entries across each class.** 16](#_Toc180573745)

[**3.1.2 Analyze the effect of checking account status on class assignment.** 17](#_Toc180573746)

[**3.1.3 Examine the interaction between loan duration and age in determining class.** 19](#_Toc180573747)

[**3.2 Objective 2: Job and loan commitment on customer class and other features** 22](#_Toc180573748)

[**3.2.1 Examine the distribution of customers based on job categories.** 22](#_Toc180573749)

[**3.2.2 Assess how job type influences installment commitments.** 23](#_Toc180573750)

[**3.2.3 Explore the relation between personal status, working type, and credit class.** 25](#_Toc180573751)

[**3.2.4 Explore the relationship between the number of dependents and existing credits in influencing class.** 28](#_Toc180573752)

[**3.3 Objective 3: Loan Characteristics and Credit Amount** 30](#_Toc180573753)

[**3.3.1 Investigate the impact of loan purpose on the credited amount.** 30](#_Toc180573754)

[**3.3.2 Explore the correlation between loan duration, credit history, and credit amount in relation to class.** 34](#_Toc180573755)

[**3.3.3 Investigate the effect of housing type on the credited amount.** 38](#_Toc180573756)

[**3.4 Objective 4: Customer Demographics and Financial Behavior** 40](#_Toc180573757)

[**3.4.1 Analyze the effect of checking account status on class assignment.** 40](#_Toc180573758)

[**3.4.2 Explore the influence of age on class classification.** 42](#_Toc180573759)

[**3.4.3 Explore relation between foreign\_worker and credit class** 44](#_Toc180573760)

[**4. Conclusion:** 45](#_Toc180573761)

[**4.1 Summary of Analysis** 45](#_Toc180573762)

[**4.2 Recommendations** 46](#_Toc180573763)

[**4.3 Limitations and Future Directions** 46](#_Toc180573764)

[**Limitations:** 46](#_Toc180573765)

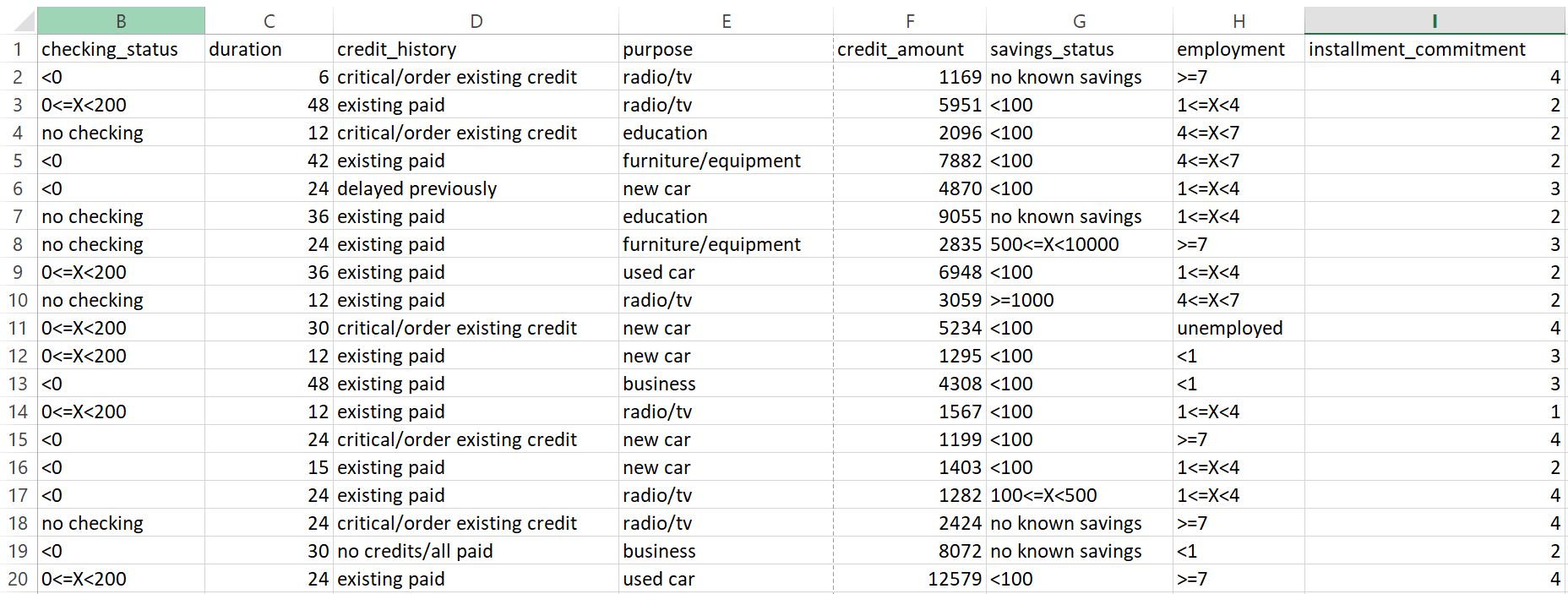
[**Future Directions:** 46](#_Toc180573766)

[**5. Workload matrix** 47](#_Toc180573767)

[**6. References** 47](#_Toc180573768)

# **1. Introduction**

## **1.1 Data Column Identification**

****bb

6000 entries, 21 columns

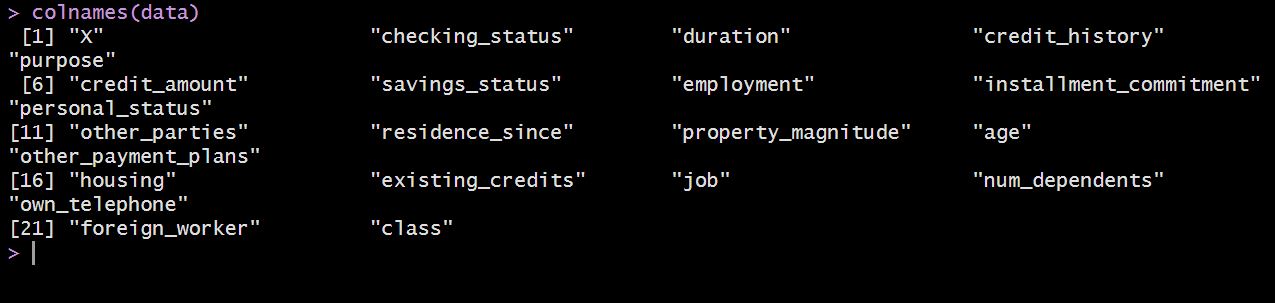
We load this csv file into our R environment in Rstudio:

C:\Users\U$ER\Desktop\checking\implementation\images\data load.JPG

By running the command:

colnames(data)

We will get the column names as follows:



## **1.2 Summary of Data Columns:**

Values in our dataset can be organized into three categories: Numerical, enums and logical.  
The numerical decimal values inside the dataset include:

1. **duration**: Loan duration (in months).
2. **credit\_amount**: Credited amount of loan
3. **installment\_commitment**: Number of installment payments.
4. **residence\_since**: Years the customer has been living in their current residence.
5. **age**: Age of the customer.
6. **existing\_credits**: Number of existing credits the customer holds.
7. **num\_dependents**: Dependent persons that the customer has to support.

The columns that can be categorized as **enums** include:

1. **checking\_status**: Status of amount inchecking account (e.g., <0, 0<=X<200, etc.).
2. **credit\_history**: Credit history categories (e.g., critical/order, existing credit).
3. **purpose**: Purpose for loan (e.g., radio/tv, car, education, etc.).
4. **savings\_status**: Savings account status (e.g., no known savings, X<100, etc.).
5. **employment**: Employment duration category (e.g., >=7, 1<=X<4, etc.).
6. **personal\_status**: Personal status (e.g., male single, female divorced/separated).
7. **other\_parties**: Other responsible parties (e.g., none, guarantor).
8. **property\_magnitude**: Type of property owned (e.g., real estate, life insurance).
9. **other\_payment\_plans**: Types of other payment plans (e.g., none, bank, stores).
10. **housing**: Type of housing (e.g., own, rent, for free).
11. **job**: Job type (e.g., skilled, unskilled resident, high qualif/self).
12. **class**: Credit classification (e.g., good, bad).

The columns that can be categorized logically are:

1. **own\_telephone**: Whether the customer owns phone(yes or no).
2. **foreign\_worker**: Whether the customer is a foreign worker (yes or no).

Numerical Columns: 7

Enum Columns: 12

Logical Columns: 2

## **1.3 Hypothesis and Objectives**

**Hypothesis**

The hypothesis of this study is that various demographic, financial, and employment related factors significantly influence customer credit behavior and classification. Factors such as checking account status, loan duration, credit history, and personal attributes like age, job type, and foreign worker status play a crucial role in determining whether a customer is classified as having a **"good" or "bad"** credit score. We hypothesize that certain patterns will emerge which will indicate that customers with stable employment, lower loan durations, and secure financial behavior, owning valuable property, good credit history are more likely to be categorized as having ‘good’ class.

By analyzing the distribution and relationships among these variables, we aim to provide a clear understanding of the factors that contribute to credit classifications. Understanding these patterns can help banks and financial institutions in making reliable decisions regarding loan approvals, risk assessment, and credit management, which will ultimately reducing financial losses.

**Objectives**

The primary objective of this report is to conduct an in-depth analysis on detailed customer credit dataset, using a set of demographic, financial, and employment-related variables. This analysis will be guided by the following specific objectives:

1. **To examine the distribution of customer credit classes**: The report will explore the distribution of credit classes, analyzing how various customer characteristics, such as checking account status, credit history, and loan duration, affect the likelihood of being classified as having good or bad credit.
2. **To investigate the impact of job type and loan commitments on credit classification**: This objective seeks to explore the relationship between customer employment status, job type, installment commitments. We will assess how these variables influence credit scores and loan repayments.
3. **To analyze the relationship between customer demographics and loan characteristics**: The report will investigate how factors such as age, foreign worker status, dependents, and personal status influence the credit amount and credit classification, aiming to uncover patterns in financial behavior across different demographics.
4. **To evaluate the effect of loan purpose and housing status on credit amount**: This objective will focus on the relationship between the intended purpose of loans, housing types, and how these factors correlate with the overall credit class.

By achieving these objectives, goal of this report to provide actionable insights into the credit behavior of customers and assist in improving credit risk assessment strategies for banks, financial departments and investors etc.

# 

# **2. Documentation**

## **2.1 Exploratory Data Analysis**

Before we prepare and clean our data for analysis, it is important to understand how data should be prepared. It includes two essential steps:

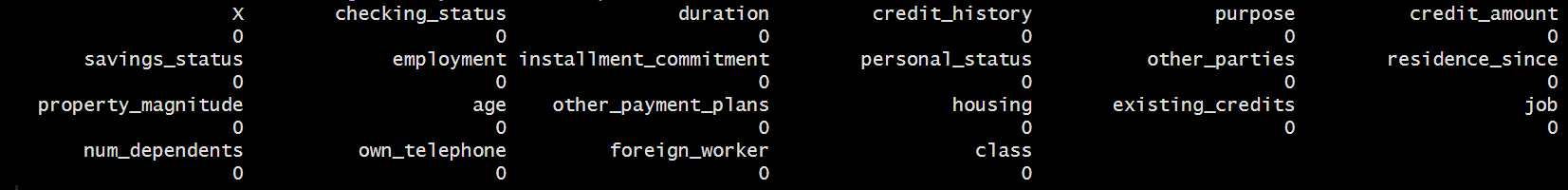
1. Data Validation
2. Missing values handling

By running a command in my Rstudio, we check for any null or empty values in our data set:

C:\Users\U$ER\Desktop\checking\implementation\images\no null values.JPG

We can also check for any null values per each column by running the command:

colSums(is.na(dataset))



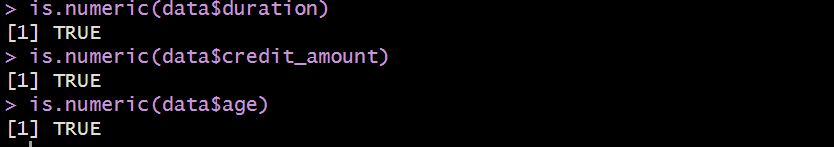
We have no null values in our data.

For data validation, we further check for duplicate and malformed values.

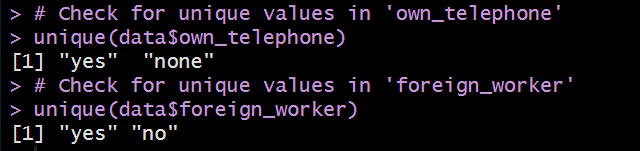
Checking duplicate values among rows:

C:\Users\U$ER\Desktop\checking\implementation\images\duplicates.JPG

We have to verify that the numeric columns such that **duration**, **credit\_amount, installment\_commitment,** etc contain only numeric data and no alphabets or alphanumeric terms.



The columns that are supposed to have only logival values also should not have any malformed or different entries.

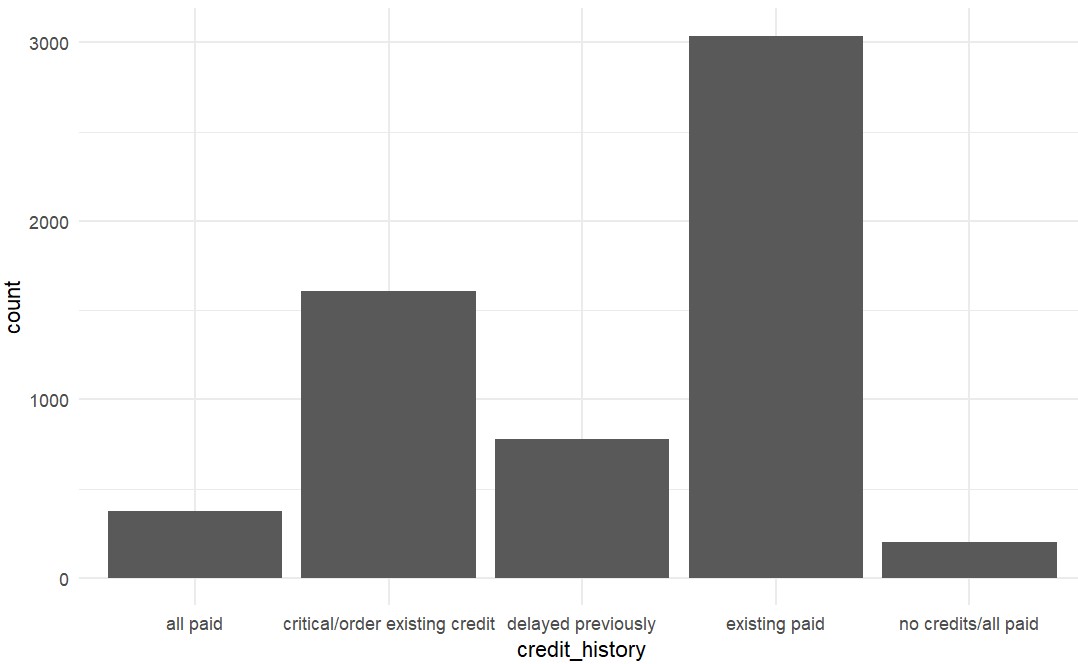


In this part we will explore how the data is distributed among different categories in different columns.

**Bar plot for credit history**

ggplot(data, aes(x = credit\_history)) + geom\_bar() + theme\_minimal()

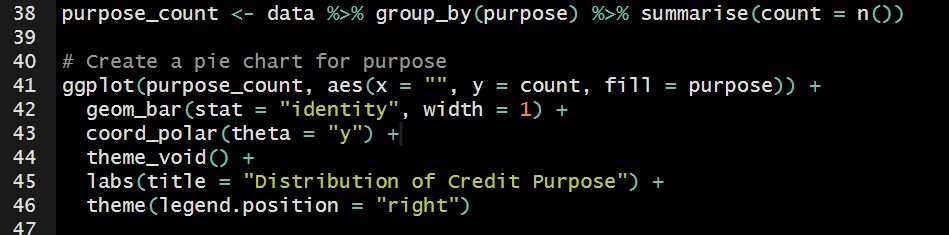
This line of code will return this graph:



It looks like most of the data consist of customers who have paid their existing load. Note that existing paid doesn’t mean that all loans are paid but it refers to a specific loan.

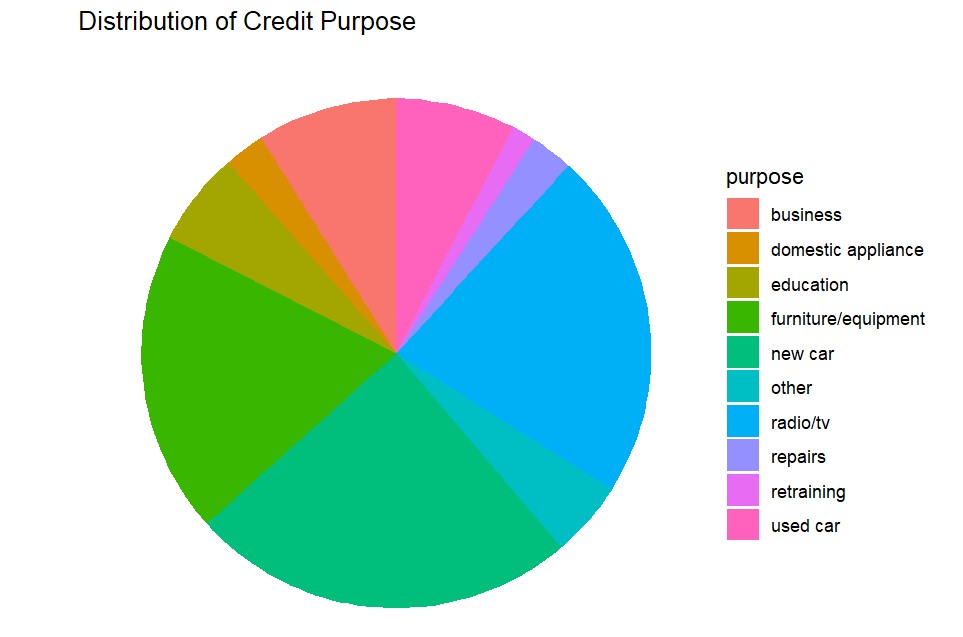
There is also considerable number of order existing credit (those who are currently paying their loans). For the category of all paid/no credits (those who have paid there loan 100% and have to further credit in their account) we have least customers in the dataset.

**Pie chart for purposes column**



* We are grouping data by ‘purpose’. Then we count how many times each purpose appear in data using n().
* Note that firstly a bar plot is mentioned in line 42. But it will be converted into pie plot using coord\_polar() in line 43.

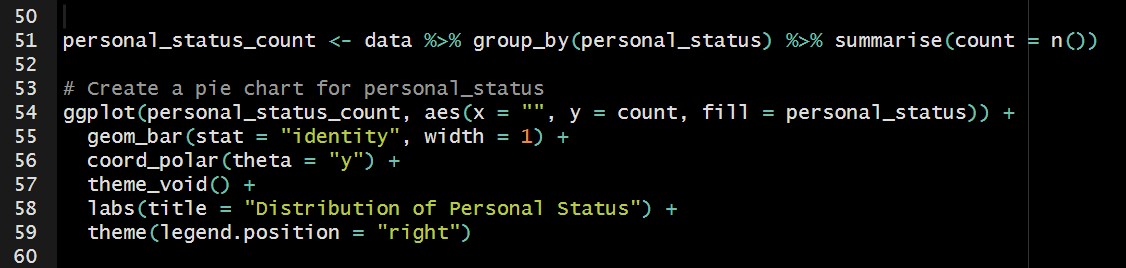
Graph looks like:



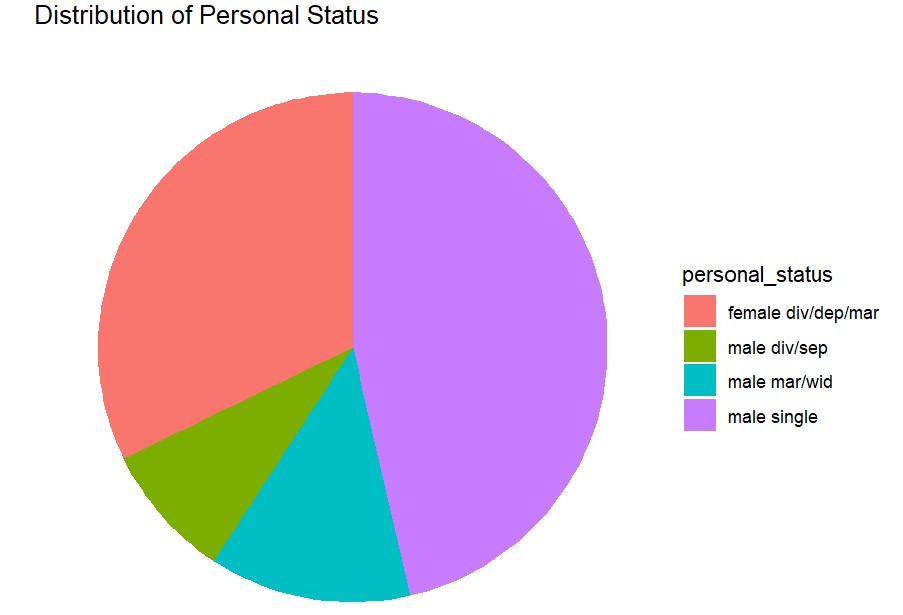
It is clear from graph that buying a new car, furniture are most dominant reasons of taking loan. The least counted reason of taking loan are repairing, retraining, domestic applications and education.

**Pie Chart for personal status column**

Using the same logic as in previous task, we can create a pie chart of personal status.



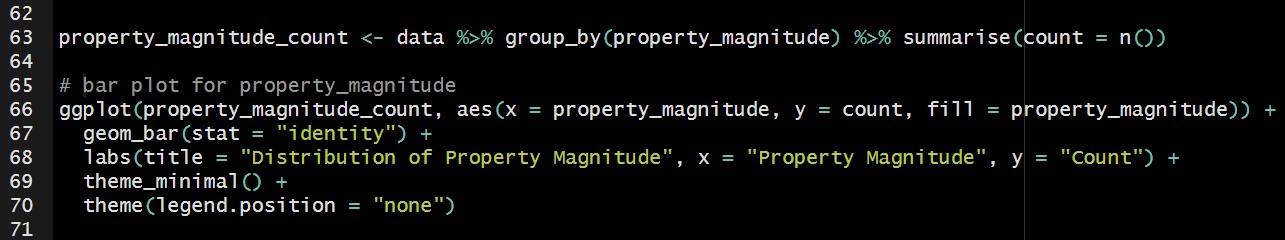
Output will looks like:



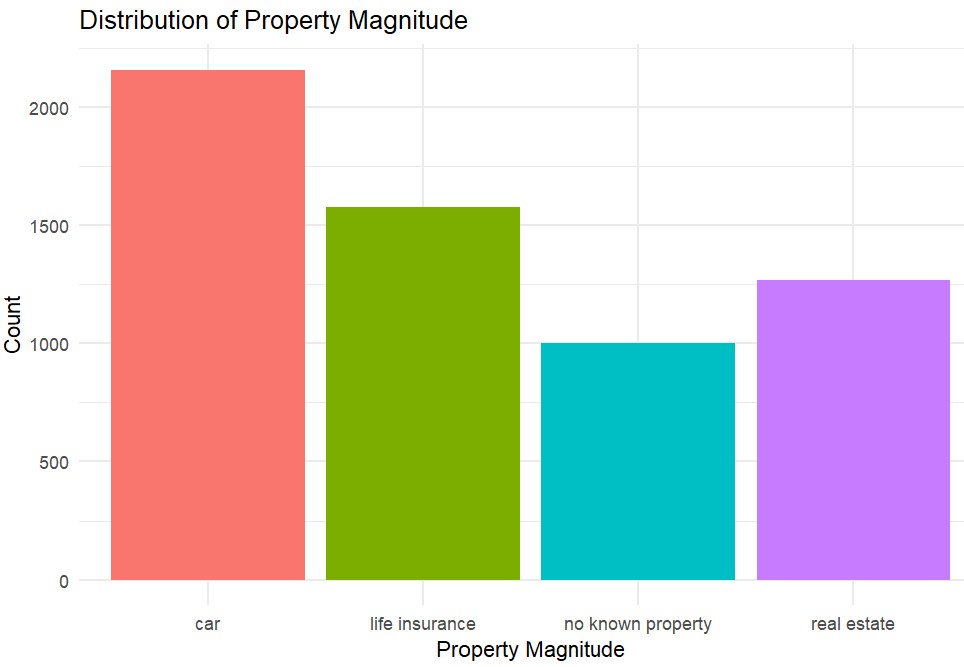
Note that male singles are of majority in the data. It means that singles males of ones who take loans mostly. Females (divorced or dependent or married) appears almost 30% in the overall data.

**Bar plot for property magnitude**

In this part we will plot each category in property\_magnitude column in bars where each height of bar will present occurrence of particular category.



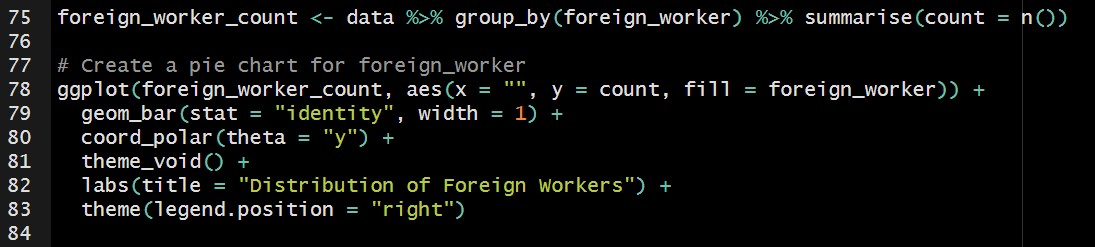
The output will looks like:



We can see that most of the individuals who take loan own a car. After that we have customers will life insurance and then real estate. Least of the customers are those who don’t own any property.

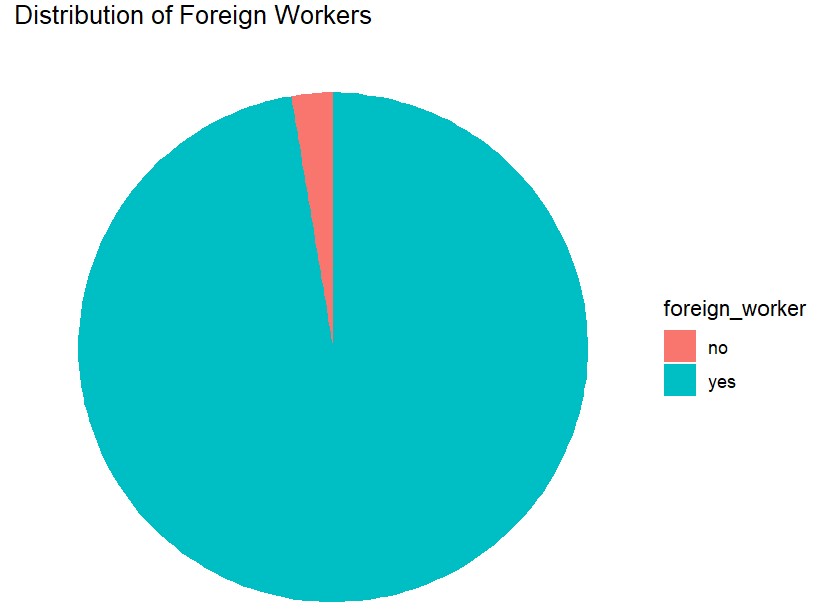
**Pie Chart of Distribution of Foreign Workers**

In this part we will create a pie chart to visualize the distribution of workers in sense of foreign workers on not.



Concepts are same as in previous tasks of pie chart.

Output will looks like:



There is a surprising trend seen in this distributions. More than 90% of the data points (individuals who take loan) are foreign workers while only a small fraction of peoples are not. Considering this investors can hunt there customers in companies where individuals work for someone out of country.

We will discuss the results and suggestions in details at the end.

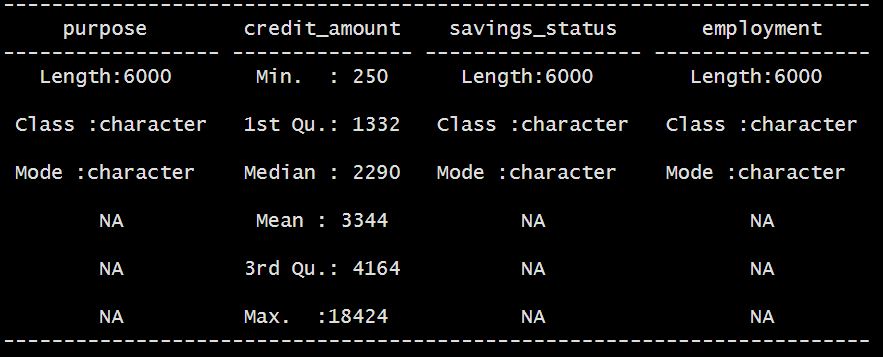
## **2.2 Data Summary**

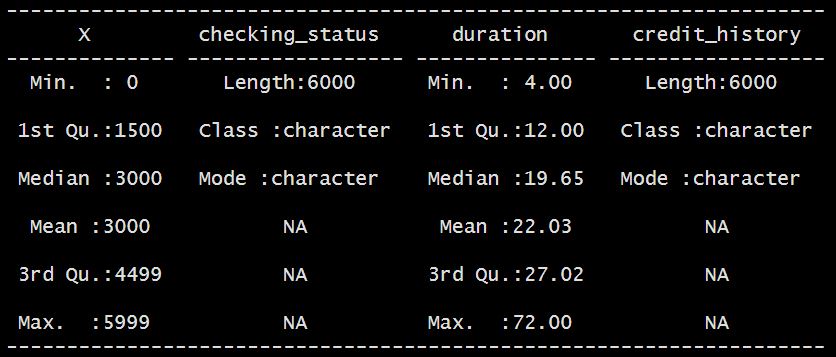
We can check the summary of our whole data set. It is displayed for each column values separately.

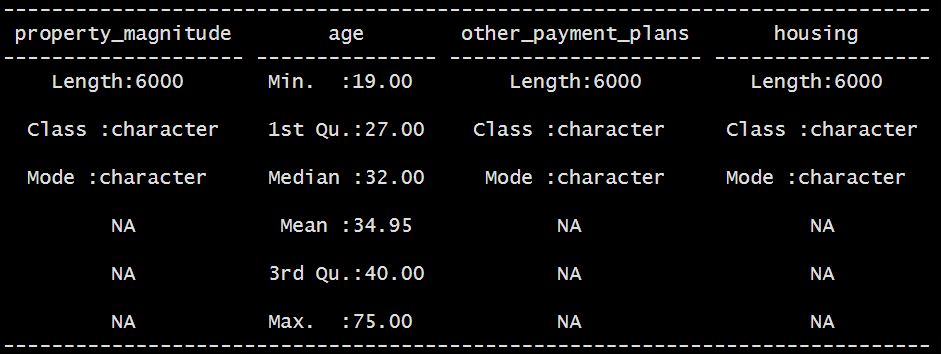
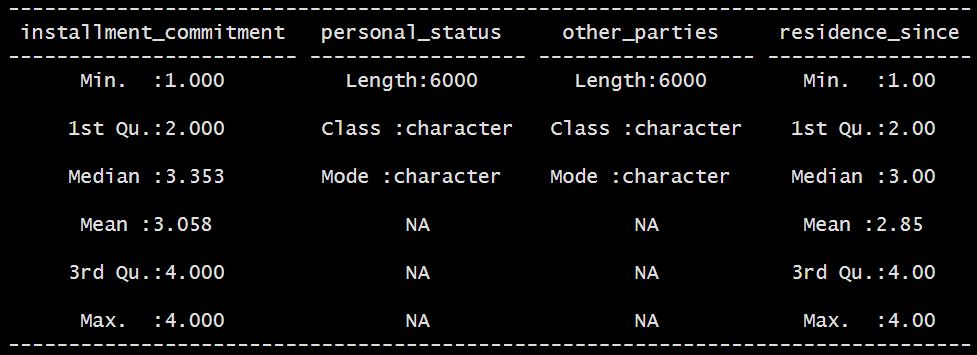
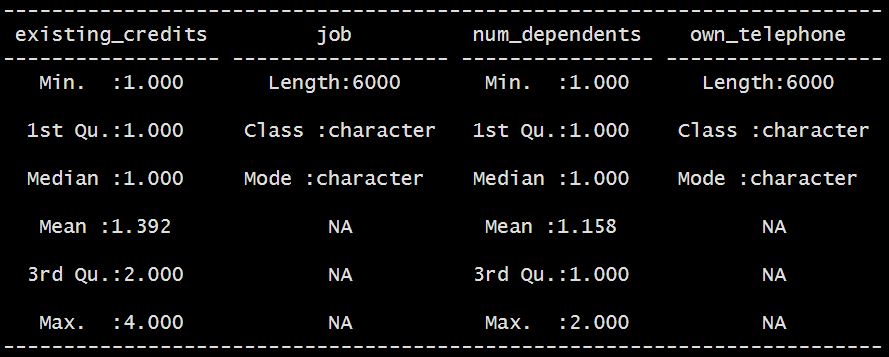
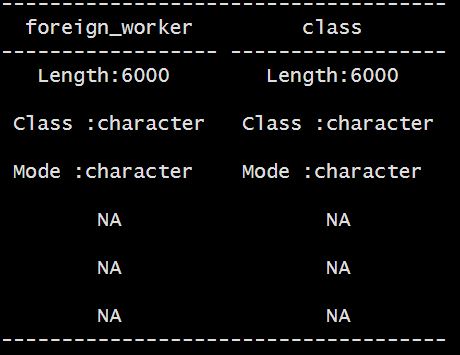
By using the command,

library(pander)

pander(summary(data))





****

The summary tables above shows following measures of all the columns in R environment.

* Mean
* Median
* 3rd Quartile
* 1st Quartile
* Minimum Value
* Maximum Value

It is to be noted that some numeric measures are available only for Numeric columns and not for enums or logical categories.

## **2.3 Libraries we will be using**

* **ggplot2:** ggplot2 is one of the most popular and powerful packages in R for creating customizable plots. We can build plots layer by layer, specifying **aesthetics** (like color, shape, size) and **geometries** (like points, lines, bars). (ggplot2, n.d.)
* **dplyr:** dplyr is a library used for **data manipulation**. It provides functions to easily filter, select, summarize, and arrange data. In most of the cases we will use it before visualization to prepare the data. (Wickham, n.d.)
* **Pander:** Pander is package that is used to convert data frames, tables, summaries into markdown format. It is majorly used to generate reports or documents. (Daróczi, 2022)

## **2.4 Fundamental programming techniques we will use**

* **$** sign in R language is used to access a particular column from dataset.
* **=** sign is used to assign something to a variable. **<-** sign can also be used for this purpose.
* **table()** keyword is used to create a table of given arguments.
* **%>%** is a way to chain multiple operations together in a readable manner. This is majorly used to manipulate and edit columns.
* **desc()** is used to sort the column given to it in descending order.
* **group\_by()** function is used to organize data into groups based on 1 or two columns. group\_by(column) will treat all rows with same values in column as a group.

## **2.5 Types of graphs we will be working on**

* **Bar charts:** presents categorical data with rectangular bars, where the height of length of bar is proportional to the values it represents. These are majorly used to compare different categories or groups. We will be working with vertical bar graphs, in which categories are on x-axis and values are on y-axis. (Yi, n.d.)
* **Stacked bar charts** are a type of bar graph where each bar is divided into multiple segments (segments represent different categories). The height or length of the bar represents the total value for that category, while the segments within each bar show the breakdown of the total value into different categories. This makes stacked bar charts useful for comparing both the overall size of categories and the contributions of individual categories. (Yi, A Complete Guide to Stacked Bar Charts, n.d.)

**For example if we are trying to display count of class of customer (good or bad) based on their property type, we will have two bars for each property type. One bar will be for good and other will be for bad class in each of the property type.**

* **Pie charts:** These are circular graphs used to represent entire column of a dataset (specifically categorical columns). Each slice of pie chart represent one particular category. These will be majorly used when will want to display percentage contribution of each category of column.

For example if we have to visualize contribution of individuals with different status (married male, married female, divorced male/female, dependent male/female etc.), we will probably use pie chart for this purpose which will present percentage of each of the category. More big is the slice of pie chart for a particular category, there is more percentage of that category in our data and so on. (pie-charts, n.d.)

* **Scatter plot:** In this visualization, we have points plotted on 2 dimensional plane. Each point present two different variables. One variable along x-axis and other variable along y-axis. Scatter plots are majorly used to analyze relation between two variables.

1. If one variable increase with other variable, it is called positive correlation.
2. If one variable decrease as other increase, it is called negative correlation.
3. If there is no relation in both variables, it is called no correlation.

For example if we want to plot our test marks and number of hours we studied, scatter plot will be a good choice in this case. Probably we will see positive correlation in this data. This is because as we increase number of hours we study, marks in exam will also increase but there can be exception. (Yi, A complete guide to scatter plots, n.d.)

* **Boxplots:** boxplots are used to present middle 50% of the data only. They are difficult to understand as compared to other type of charts.

1. It consist of a box which present 50% of the data (referred by interquartile range).
2. There is a line inside a box which represent median of data.
3. Line extending from box present data outside interquartile range (these lines are known as whiskers).
4. Data points outside the whiskers are known as outliers and displayed by dot in the boxplots.

Whenever we want to analyze data based on median and quartiles, boxplots are of better choice. These are also good to visualize spread of data and outliers in it. (Yi, A complete guide to box plots, n.d.)

* **Histograms:** Histograms distribute the data by dividing it into bins. They provide a way to understand the frequency of data points in different ranges. Histograms looks like bar charts but these are different. X-axis contain data values divided into different bins and y-axis represent frequency of data points. For better understanding the histograms, let’s work with an example.

Let say we have data of age of people in a village as:

22, 25, 25, 28, 30, 32, 32, 32, 35, 36, 40, 45, 50, 55

We can create 5 bins from this data ranging from: 20 – 30, 30 – 40, 40 – 50 and 50 – 60.

20-30: 5 (22, 25, 25, 28, 30)

30-40: 5 (32, 32, 32, 35, 36, 40)

40-50: 2 (45, 50)

50-60: 1 (55)

So first bar will have height of 5 units, second bar will have height of 6 units, third bar will have 3 units and last bar will only have 1 unit height. (what is a Histogram?, n.d.)

# **3.** **Objectives**

1. **Distribution of various features and customer class analysis**

* Analyze the distribution of entries across each class.
* Analyze the effect of checking account status on class assignment.
* Examine the interaction between loan duration and age in determining class.

2. **Job and loan commitment on customer class and other features**

* Examine the distribution of customers based on job categories.
* Assess how job type influences installment commitments.
* Explore the relation between personal status, working type, and credit class.
* Explore the relationship between the number of dependents and existing credits in influencing class.

3. **Loan Characteristics and Credit Amount**

* Investigate the impact of loan purpose on the credited amount.
* Explore the correlation between loan duration, credit history, and credit amount in relation to class.
* Investigate the effect of housing type on the credited amount.

4. **Customer Demographics and Financial Behavior**

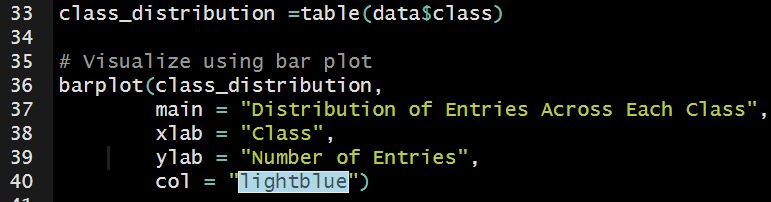
* Analyze the effect of checking account status on class assignment.
* Explore the influence of age on class classification.
* Explore relation between foreign\_worker and credit class

## **3.1 Objective 1: Distribution of various features and customer class analysis**

### **3.1.1 Analyze the distribution of entries across each class.**

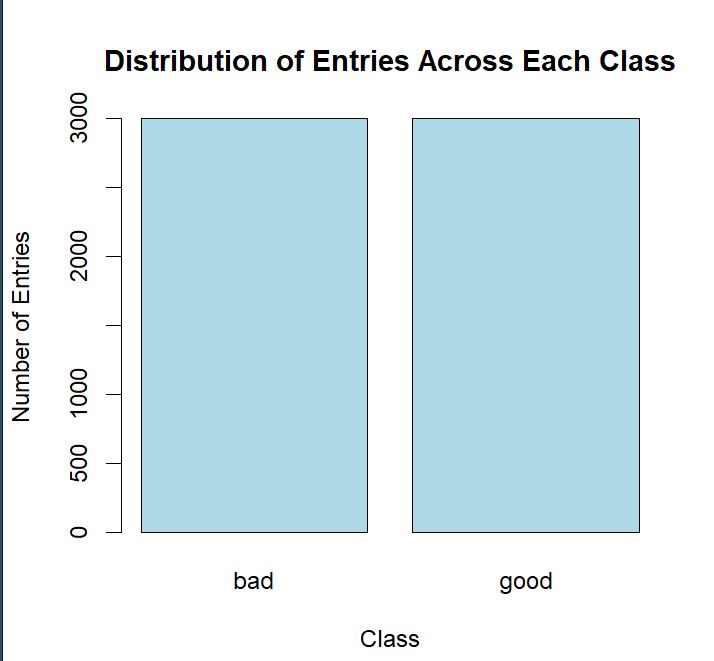
First of all, we will check the distribution of our data on the basis of column class, analyzing how many entries are “good” or “bad” respectively. This column conveys type of customer.

Running the following command in RStudio:



* We have fetched class column of our data using (data$class) in line 33.
* We make data frame table of class col and save it in variable class\_distribution. It contains frequency of each class.
* For visualizing our result, we are using bar plot from line 36 to 40.
* We are giving the title, label to x-axis and y-axis. Also, specifying color of bars.

We have this graph:



**Result:**

This graph shows that column class is equally distributed and we have equal entries for both good and bad class i.e, 3000.

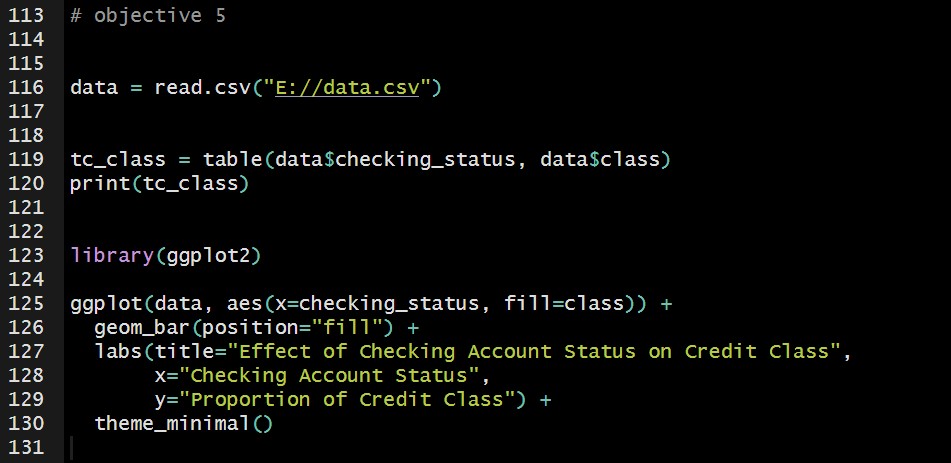
### **3.1.2 Analyze the effect of checking account status on class assignment.**

Now it’s time to analyze type of customer based on the credits they have in their checking account. We have 4 categories of checking account in our dataset. These are:

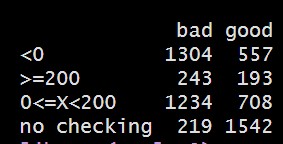
* Customers with no checking account
* Customers having credit less than 0$ (negative).
* Customers having credit between 0 and 200$.
* Customers having credit equal to or more than 200$.

We will analyze type of customers (good or bad) for each of the checking account category.

Firstly we will make table of checking status and customer type using table() keyword. Then using ggplot library, we will plot for better understanding. Note that we also make graph of relation between checking status and customer type without using the table.

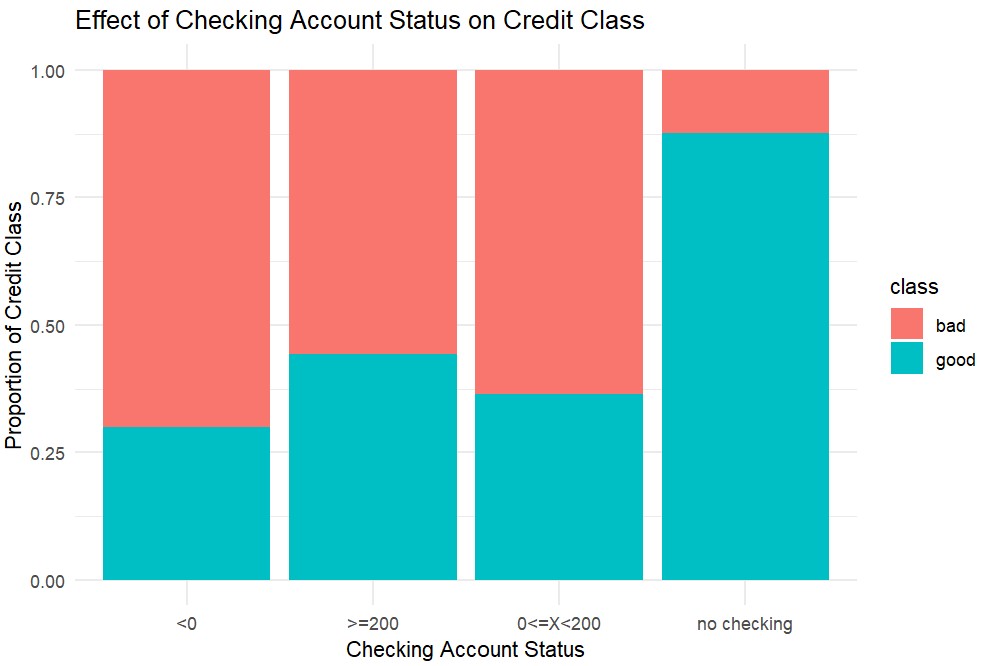


In line 119 we have constructed the required table. Output of line 120 looks like:



This output gives count of good customer and bad customer for each of the checking status category. From line 125 to 130 in code, we have plotted checking status data (on x-axis) and class data (on y-axis). Note that we have two kind of classes, there are will be two kind of bars for each of the checking status.

The plotted graph looks like:



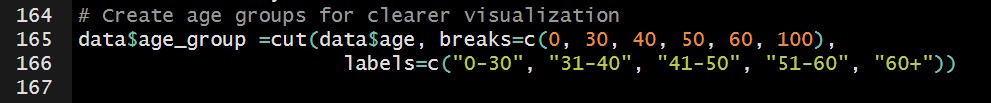
**Results derived after visualizing:**

1. More than 50% of the customers who have checking accounts are bad customers not depending on the amount they have in their checking account.
2. More than 90% customers who don’t have checking account are mentioned as good customers in our dataset.
3. Among the customers having checking account, those who have more than 200$ in there account are comparatively greater in number (for good class) than other two categories (negative balance or balance between 0 and 200$).

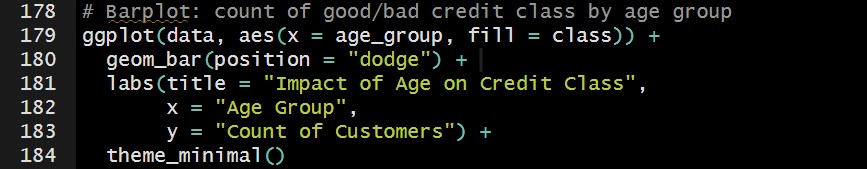
### **3.1.3 Examine the interaction between loan duration and age in determining class.**

**Impact of Age on credit class**

Firstly, we would see the impact of Age on credit class and then the comined effect of age and duration.

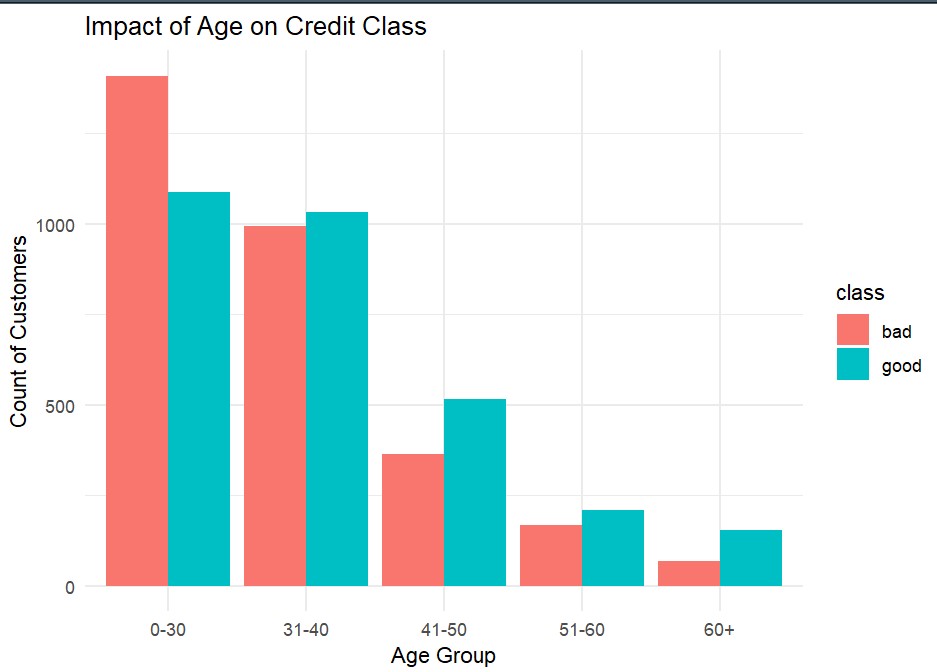


Age is a continuous variable and can take any data. If we use all exact values, it can lead to highly scattered and unorganized plot. That’s why we are grouping age into 5 groups in the above given code snippet. Group 1 that is from 0-30 involve all ages from 0 to 30, etc.



* We are using same ggplot library for plotting bars.
* For x argument, we are passing age\_groups that we have formed in above snippet in line 165-166.
* The bars get filled according to what class they are assigned in dataset.
* geom\_bar function of ggplot library is used for plotting bar plot. It visualizes the count of categorical values.
* position= “dodge”: this argument separates the bars of different categories rather than stacking them together.
* We are also giving our plot a proper title and labels to x and y axis.

The plot we get as result is:



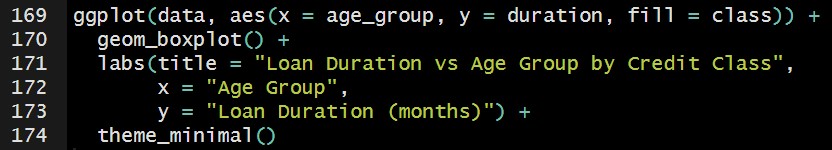
**Result from this plot:**

1. A significant number of young individuals tend to have bad credit than good credit, in age group 0-30. It means young people are more prone to having a bad credit class.
2. The age group of **31-40** seems to be a transition period where people manage credit more responsibly, with more individuals having **good** credit.
3. Older individual (more than 40) shows greater trend towards having a good credit.

**Combined impact of age and duration on credit class**

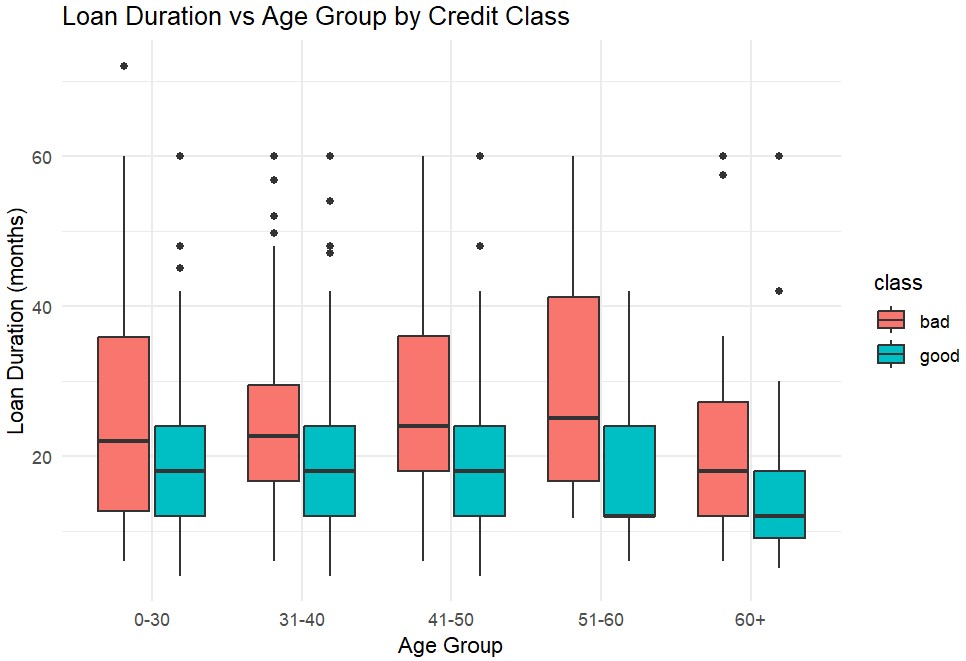
It refers to the combined effect of both features (load\_duration and Age) in classifying customers into good or bad categories. For this purpose, we would explore different combinations of duration and Age and see their combined effect on class category.

This type of analysis helps identify patterns, such as whether younger individuals with longer loan durations are more likely to fall into the bad credit category, or if older individuals with shorter loan durations tend to be more creditworthy.



* Using the same ggplot library, we pass age group as x-axis argument, duration for y-axis argument.
* Fill=class:The fill color of the boxes will be based on the credit class (good or bad
* theme\_minimal() is a function in **ggplot2** that is used for clean, simple, and minimalist appearance to the plot.
* labs( ) function allows us to add title and labels to our plot on the basis of our requirement.

The graph we get as a result is:



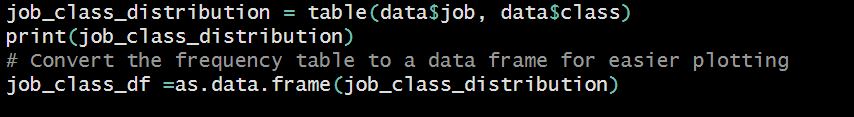
**Results from this graph:**

1. Age group 51 – 60 have longest range of duration for ‘bad’ class (18 – 28).
2. People of age 60+ have shortest loan duration range for the ‘good’ category (9 – 18).
3. In all the age groups, ‘bad’ class have relatively longer range (and higher values) of loan duration (in months).

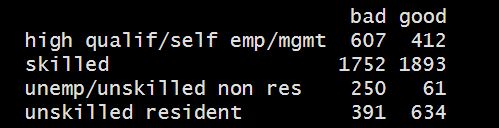
## **3.2 Objective 2: Job and loan commitment on customer class and other features**

### **3.2.1 Examine the distribution of customers based on job categories.**

In this part, we will analyze type of customer categories (good or bad) based on their job roles.

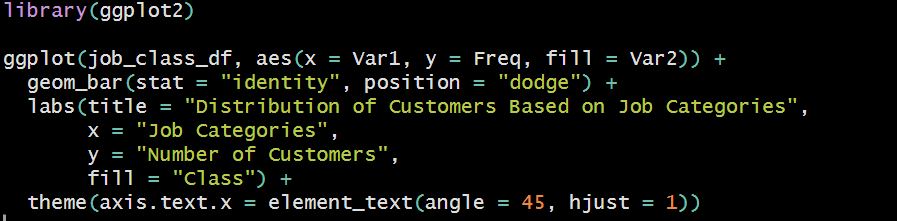


First, we are making frequency table of the columns “job” and “class” and storing the table in job\_class\_distrubution.. When we print the table, it looks like:

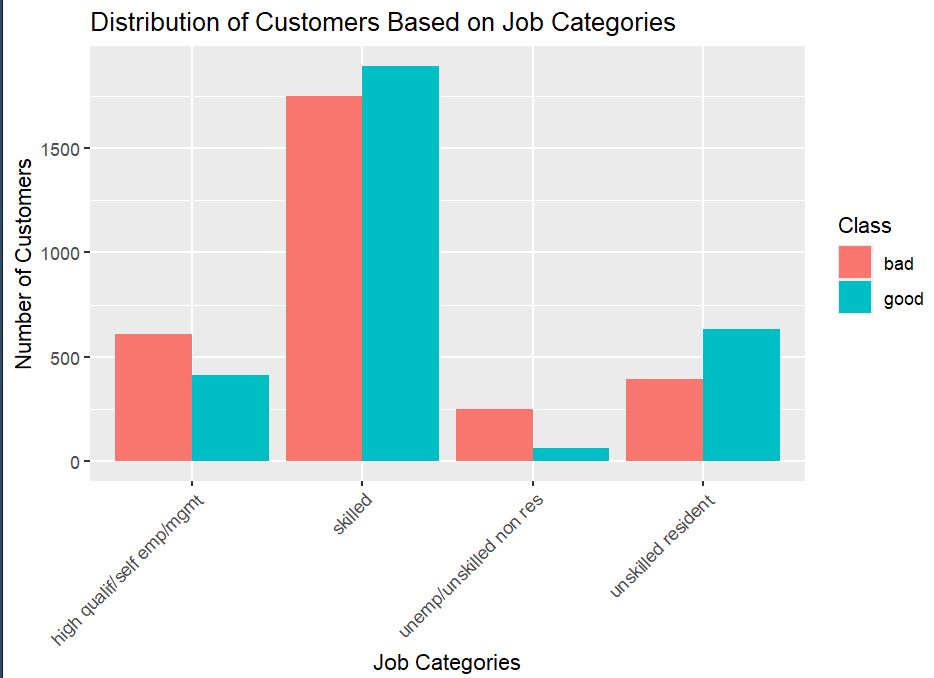


This table shows how number of good and bad customers depending on different jobs roles.

We will then convert this table into more readable dataframe using function as.data.frame( ) Converting it into DataFrame is helpful when plotting graphs.



* We are importing our library ggplot2 for plotting graph.
* x=Var1 is first variable from table that is job roles.
* y=Freq is containing the frequency distribution or number of good or bad customers.
* fill=Var2 this command fill the color of bars on the basis of different class
* Giving title to the graph and also assigning labels to x and y axis



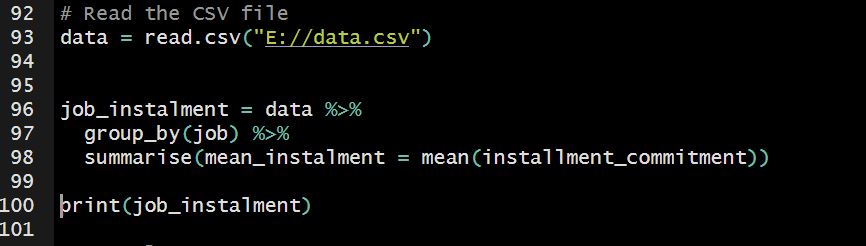
**Results derived after visualizing:**

* 1. It is evident from this graph that the sample data contain most entries from skilled persons. It means majority people who took loan or credits are from this category and least are from unemployed category.
  2. From skilled person category, there difference in the number of good or bad customers is negligible.

### **3.2.2 Assess how job type influences installment commitments.**

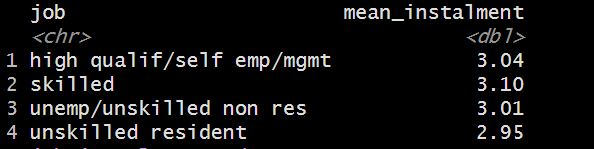
In this part of, we will be assessing how type of job a person is doing is influencing commitment of installments. Note that installment commitment is instalment rate in percentage of disposable income. In simple words it indicates the proportion of a customer's income that is committed to paying loan installments.

Our strategy will be to group the data by job (column) and calculate the mean or median installment commitment for each job type. Then, we will visualize the relationship using a bar chart.

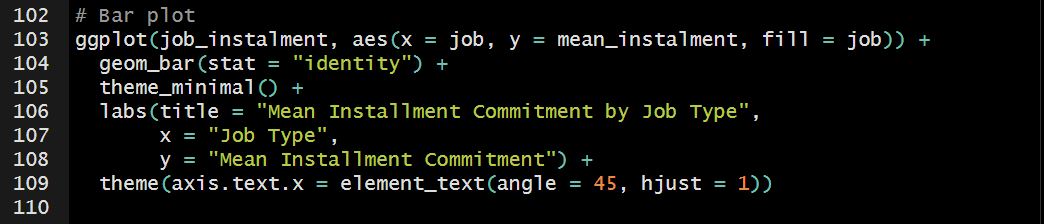


After loading the dataset, we have created and summarised a new column ‘mean\_instalment’ which contain mean of column ‘installment\_commitment’. Note that we have grouped the data by ‘job’. %>% is a method to join multiple operations in readable manner.

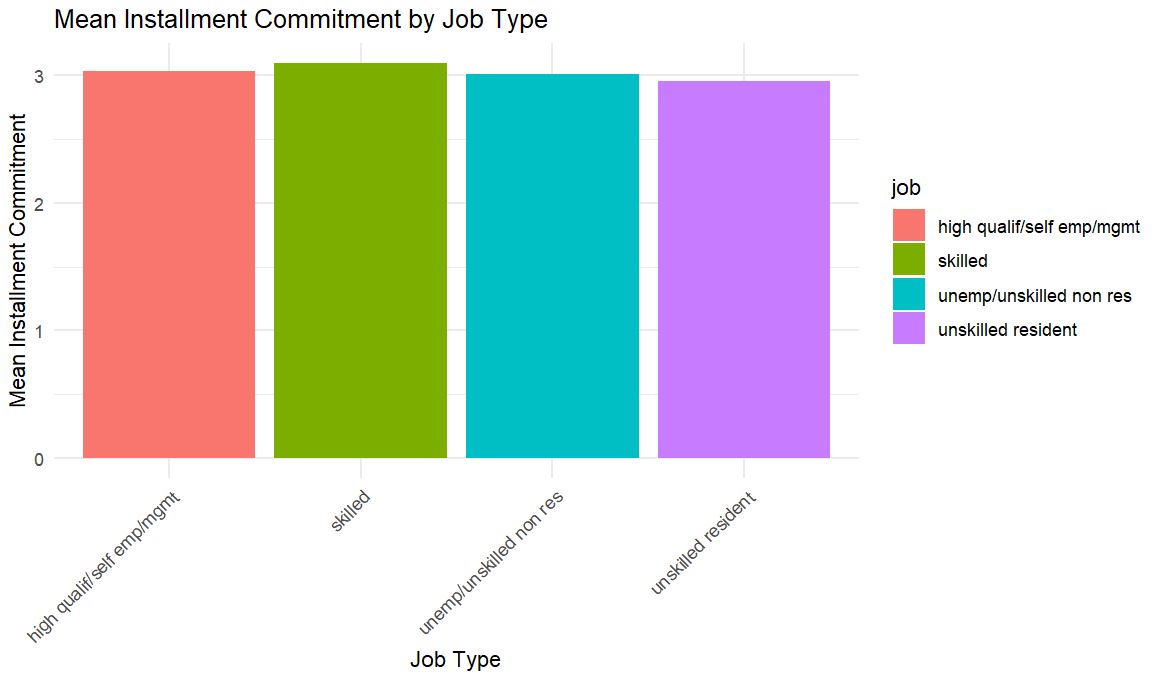
Output of line 100 looks like:



For each job type, we have percentage installment commitment percentage. It means how much percentage one will pay from his/her monthly income. For example High qualified or self-employed persons will be paying 3.04 percent of their monthly income to return the debt.



In these lines we have constructed a bar plot using the concepts discussed in details in previous objective.

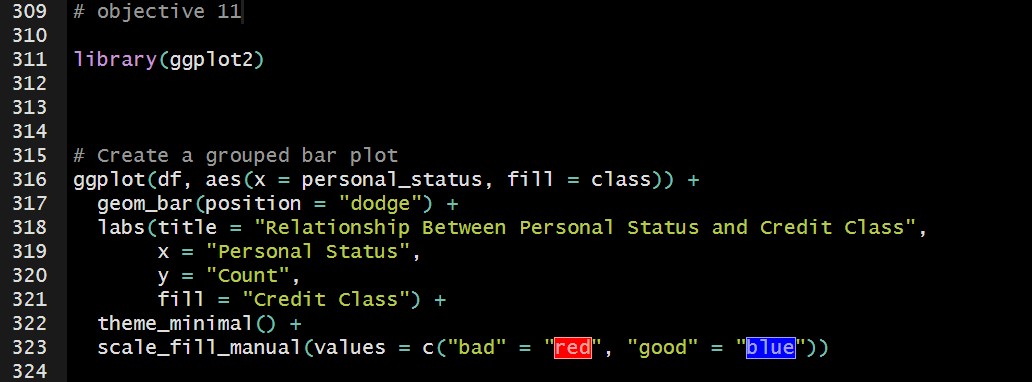


**Results derived after visualizing:**

1. Due to very close average values for each of the job type, information from this image is not clear. From upper image we get more clear idea of the relation between job type and average installment commitment.
2. Skilled individuals are paying most percentage (3.10%) of their monthly income to pay their debt.
3. Unskilled residents are paying least percentage (2.95%) of their monthly income to pay their debt

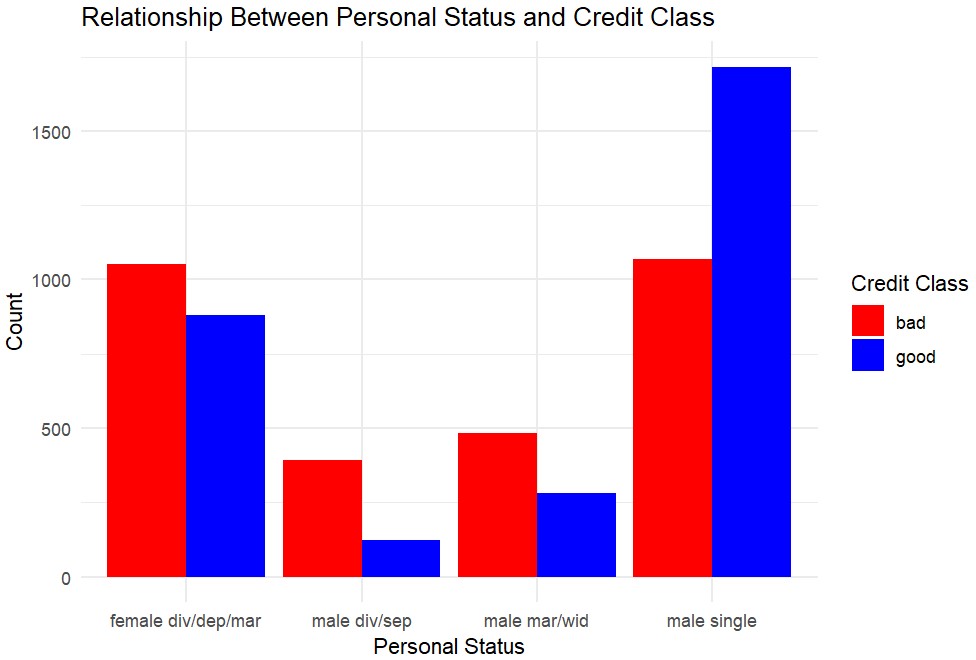
### **3.2.3 Explore the relation between personal status, working type, and credit class.**

In this part we will firstly explore relation between personal status and credit class. For this purpose stacked bar charts can be a good choice.



In this lines we have created a stacked bar chart of personal status and credit class. Note that we have applied color distinction (red for bad and blue for good).

Output:

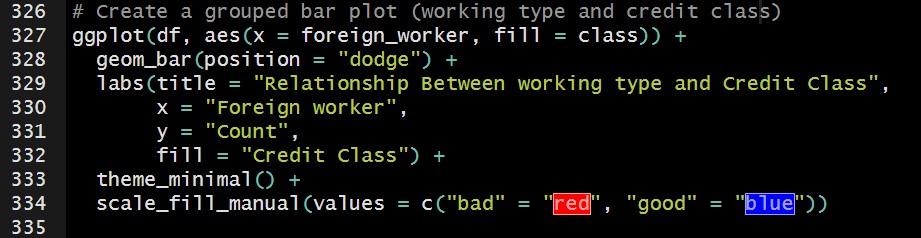


**Results from this graph:**

1. Single males are significantly greater in count classified as ‘good’ class.
2. In all other personal statuses (married male, divorced or separated males, and females) majority is classified as bad customers.

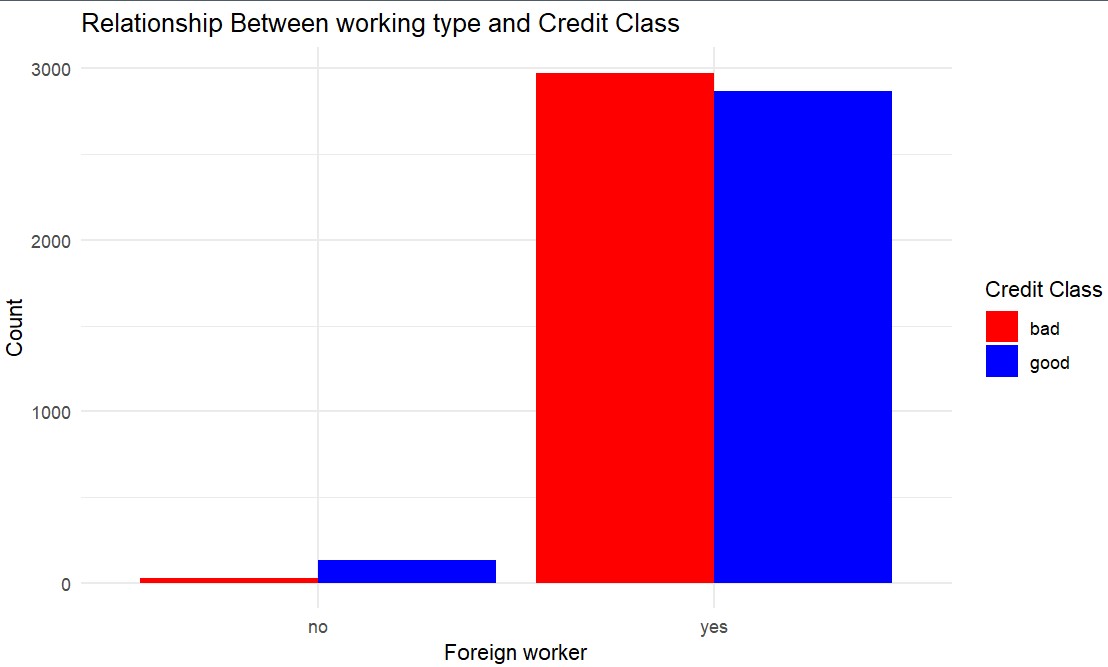
Now let’s analyze such trends in working type of individuals.

Logic will be exactly same as previous graph except for the personal status column will be replaced by working type column. Note that we have two types of workers in dataset (foreign worker and not foreign workers).



Note that we are now plotting foreign\_worker column on x-axis.

Output:



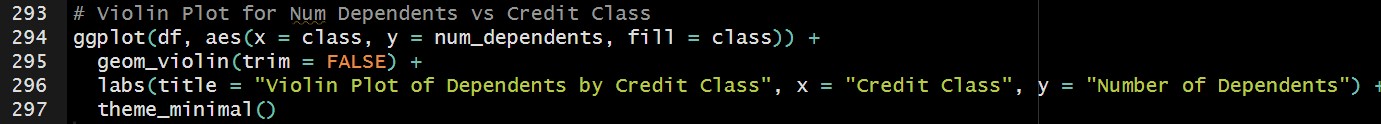
**Results from this graph:**

1. Surprising thing is that, we have more that 90% of the foreign workers in dataset while only small fraction of workers are not foreign workers.
2. There is almost equal distribution of good and bad customers among foreign workers with bad customers slightly greater in number.
3. Out of small fraction of not foreign workers, significantly greater number are labeled as good customers.

### **3.2.4 Explore the relationship between the number of dependents and existing credits in influencing class.**

**How num\_dependants affect credit class**

In analyzing customer credit behavior, one important factor to consider is the number of dependents a person has. The **number of dependents** refers to the number of people financially depending on the customer (such as children or elderly family members). It can influence their financial obligations and, consequently, their ability to manage credit and debt in a good or bad way.

A violin plot is use to shows the distribution of the number of dependents within each credit class while also indicating the density of data points.

* The ggplot library uses our dataset df and aes( ) function for visual mapping.
* As x argument, we pass credit class and for y argument, num\_dependents.
* The violin color will be filled differently for each class.
* Trim=false keeps the entire data in the plot even if there are long tails.
* Giving the title and labels to the x axis according to the requirement.

The graph we get as a result:



**Results from this graph:**

The width of the plot represents the density or concentration of data points. Therefore,

1. Most people have 1 dependant person either they are in good or bad class.
2. A very few customers have 2 dependants from each class.

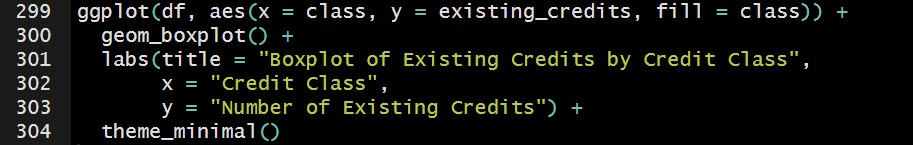
3. Both classes have similar shapes, indicating that the distribution of the number of dependents is quite similar across the "good" and "bad" credit classes.

4. The width is lightly more across 2 dependents around good credit class which is an indication that 2 num\_dependents may lead good credit class but it is not a very reliable measure.

**How existing\_credits affect credit class**

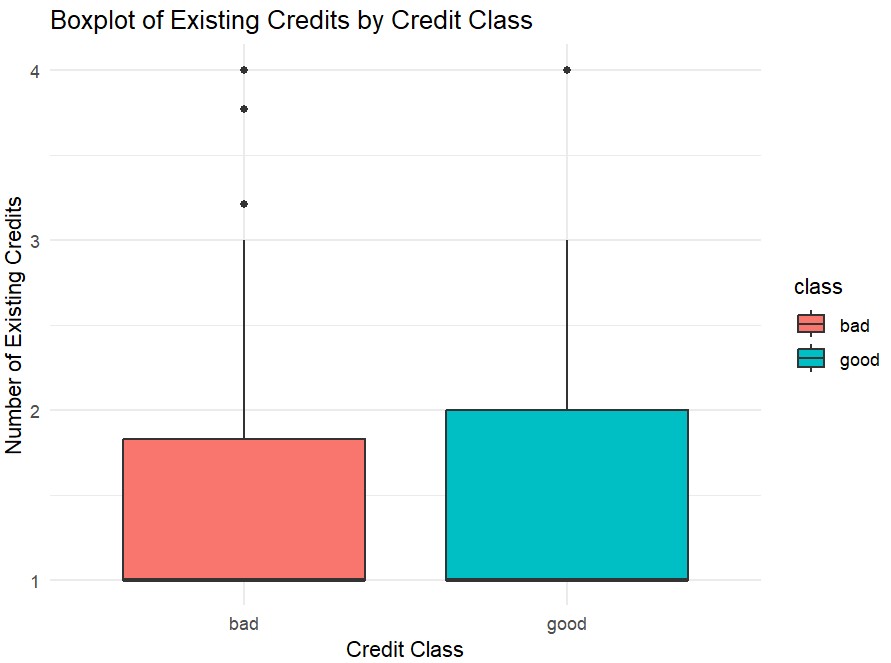
Existing\_credits refer to the pending loans or lines of credits the the customer still has to pay. Studying it impact on credit class is important to check for risk assessment. It means whether a person has yet to pay his loans can be good customer for our bank or no? If it is found that customers with already due loans are bad customers for the bank, it proves as a red flag and the measures can be taken accordingly.

We will visualize how the number of existing\_credits affect credit class using a box plot.



* For boxplot, we will use the same library ggplot.
* geom\_boxplot( ) is the keyword we use to plot boxplot.
* For x, we pass credit class as argument and for y, it is existing number of credits that the customer has to pay.
* labs ( ) is the function we use to give title to our boxplot and labels to our x and y axis.

The graph we get as a result:



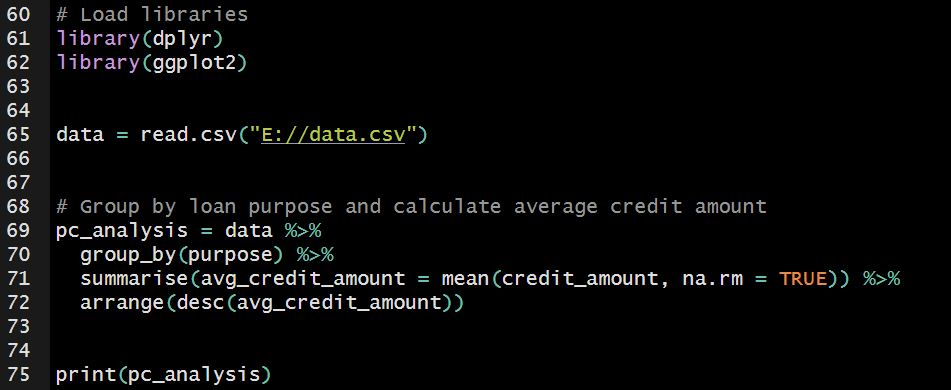
**Results from this graph:**

1. Both the good and bad categories have almost same interquartile range (middle 50% of the data).
2. For both the classes, medians are close to 1.
3. Extended whiskers present that there are customers with 3 number of credits as well.
4. There are few customers in data with even more than 3 credits for both classes (these are outliers). But these customers appear to have majority in bad class. It means that there are few customers in dataset having existing credits more than 3.

## **3.3 Objective 3: Loan Characteristics and Credit Amount**

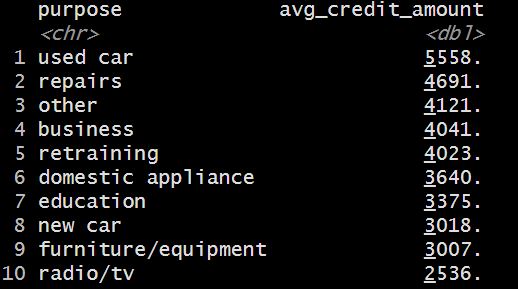
### **3.3.1 Investigate the impact of loan purpose on the credited amount.**

In this part we will interrelate purpose of taking debt and amount of credit somebody have taken as debt.

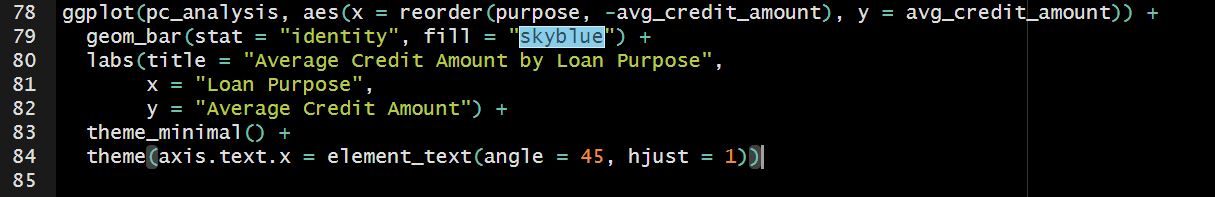


* Firstly we import two libraries (line 61, 62) which will be used later for processing and visualization.
* In line 65 we have loaded the dataset.
* From line 69 – 72, we have done purpose and credit\_amount analysis. Here is brief description:
* **data %>%**: This is the beginning of a pipe operation, which takes the dataset and passes it to the next function. It (operator %>%) is a way to chain multiple operations together in a readable manner.
* **group\_by(purpose)**: This function groups the data by the purpose (column name) variable. This means that any subsequent operations will be performed separately for each unique value of purpose column in our dataset.
* **summarise(avg\_credit\_amount = mean(credit\_amount, na.rm = TRUE))**: This function creates a new summary data frame where:
* avg\_credit\_amount is a new column calculated as the mean of the credit\_amount column for each group defined by purpose.
* The **na.rm = TRUE** argument tells R to ignore any missing values when program will calculate the mean. We don’t have any missing values in our dataset so output will be same without this argument as well.
* **arrange(desc(avg\_credit\_amount))**: This function sorts the resulting summary data frame in descending order based on the avg\_credit\_amount column. This is because we want the highest average credit amounts to appear first.

Output of line 75 will be:



Finally we have average credit amount for each of the reason of taking debt. Let’s plot these for better visualization.



* **ggplot(pc\_analysis, ...)**: This initializes a ggplot object using the pc\_analysis data frame.
* **aes(...)**: Specifies the aesthetic mappings for the plot.
  + **x = reorder(purpose, -avg\_credit\_amount)**: This reorders the purpose factor based on the average credit amounts in descending order (the minus sign indicates descending order).
  + **y = avg\_credit\_amount**: This sets the y-axis to represent the average credit amounts.
* **geom\_bar(...)**: This function adds bar geometries to the plot.
  + **stat = ‘identity’**: Height of each bar will reflect the average credit amount for each loan purpose.
  + **fill = ‘skyblue’**: Specifies the fill color of the bars as sky blue.
* **labs(...)**: This function adds labels to the plot:
* **theme\_minimal()**: This function applies a minimalistic theme to the plot. It provides a clean, modern look, allowing the data to stand out more clearly.
* **theme(...)**: This function allows for further customization of plot elements.
  + **axis.text.x = element\_text(angle = 45, hjust = 1)**: This modifies the appearance of the x-axis text.

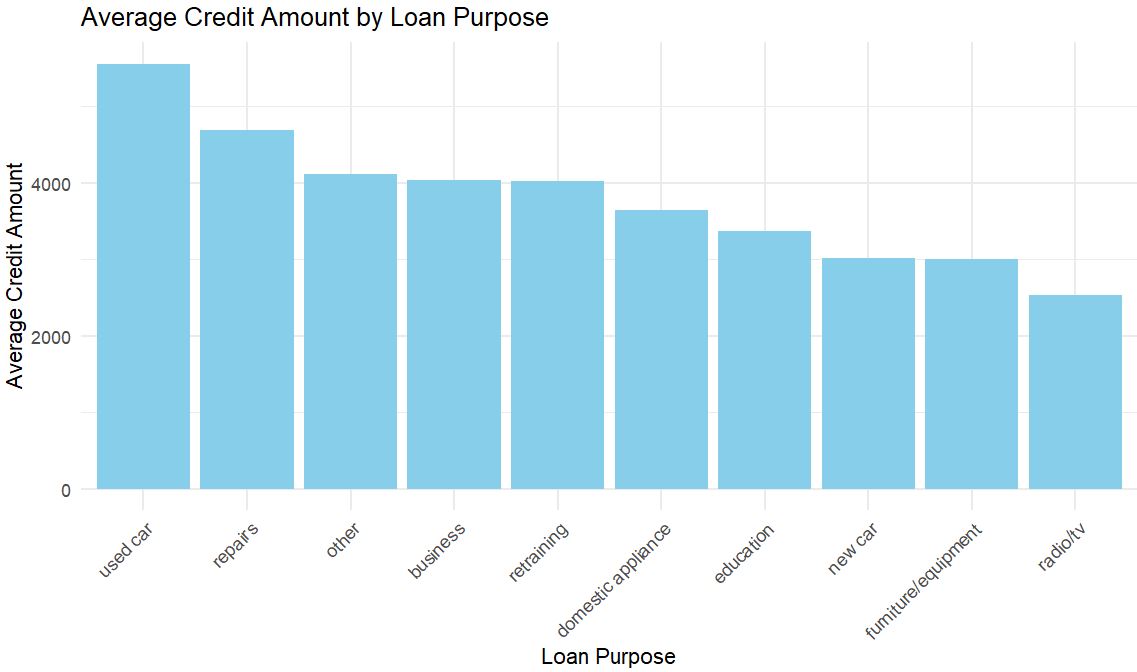
**hjust = 1**: This argument adjusts the horizontal justification of the text.

**Angle = 45**: This argument rotate the text by 45 degrees.

This code creates a bar chart visualizing the average credit amount for different loan purposes. Here’s a concise summary of what each part does:

* **Data Source**: Uses the pc\_analysis data frame.
* **X-Axis**: Displays loan purposes, reordered by their average credit amounts in descending order.
* **Y-Axis**: Represents the average credit amounts.
* **Bar Appearance**: Bars are filled with a sky blue color, with heights corresponding to the average credit amounts.
* **Labels**: A title and labels for the x and y axes are added for clarity.
* **Theme**: A minimalistic theme is applied for a clean look.
* **Axis Text Customization**: X-axis labels are rotated 45 degrees for better readability and aligned appropriately.

Result of the following code is:



**Results derived after visualizing:**

1. It is clear from graph that most of the debt amount is being utilized in purchasing used cars.
2. Least of the debt amount is taken for purchasing a TV/Radio.

### **3.3.2 Explore the correlation between loan duration, credit history, and credit amount in relation to class.**

**Loan duration and customer class**

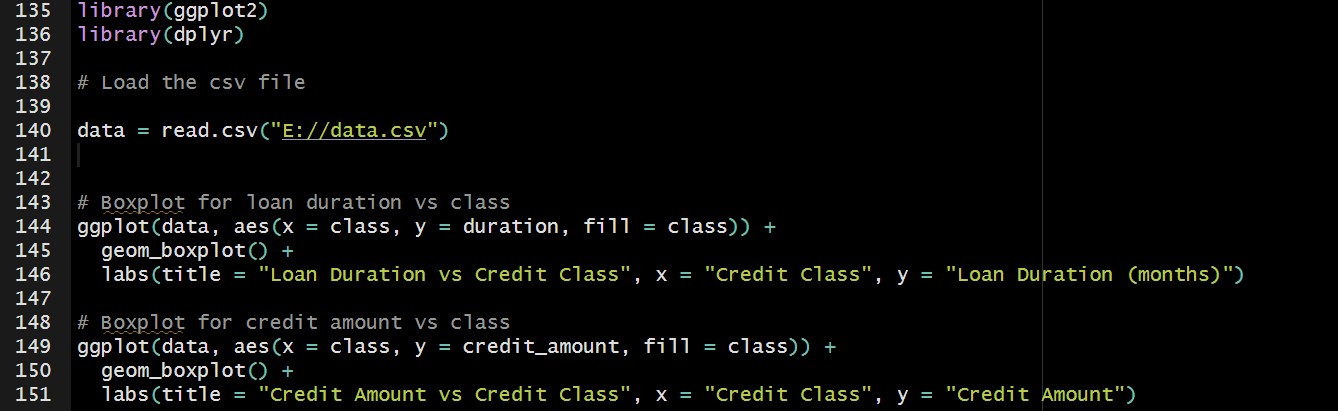
In this part, boxplots will be specifically used for analysis. Boxplots consist of 5 major elements.

**Box:** The main rectangular part of the plot shows the range between the first quartile (Q1) and the third quartile (Q3). This is middle 50% of the data.

**Line inside the box:** This line shows the median or the 50th percentile, dividing the box into two parts.

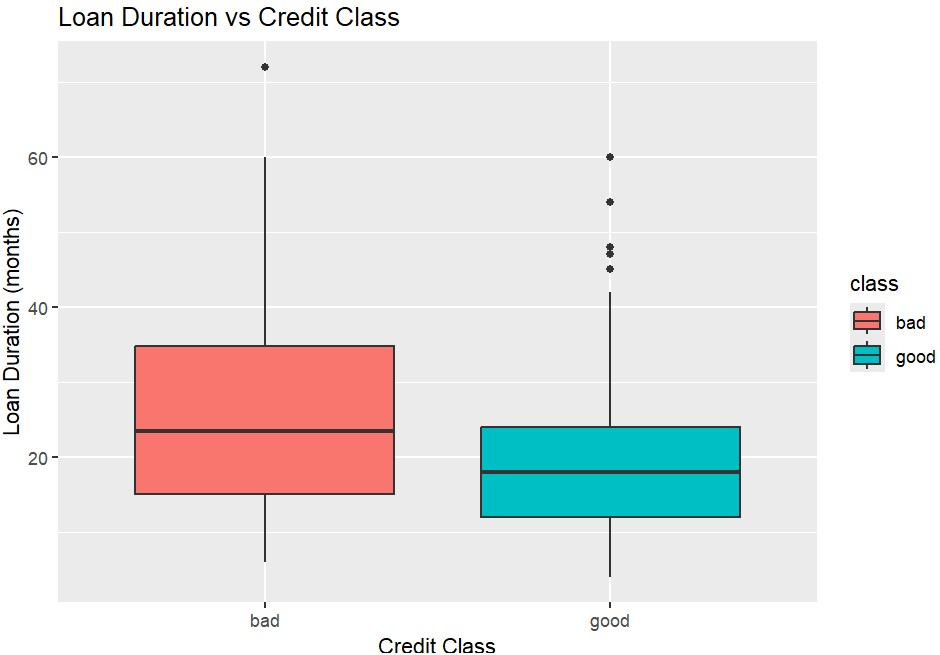
**Whiskers:** These are straight lines extending from the box to the minimum and maximum values. They help show the spread of the data outside the box.

**Outliers:** Individual data points that fall outside the whiskers are considered outliers. They are often as dots or small circles.



After loading the required libraries and dataset, we are making a boxplot of class on x-axis and duration of load on y-axis. Note that this part of code is on lines 143-146. fill = class is used to differ the good and bad class in term of colors.

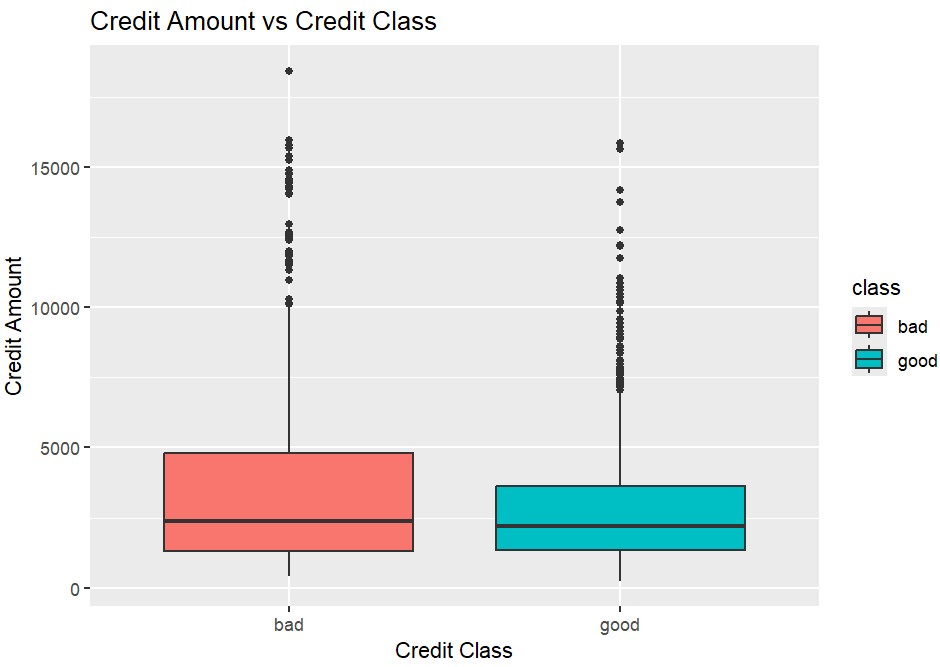
The boxplot will look like:



**Results:**

1. For bad class, we have median duration almost 24 months. It is clear from graph that 50% of the data (for bad customers) have duration of 15 – 36 months. There is one outlier in data above 60 months which indicate one customer who had loan duration of more than 70 months.
2. For good class, we have median duration almost 18 months. 50% of data range from 12 to 24 months. There are few outliers (5) in data between 40 – 60 months, indicating that few customers among good class had loan duration of 40 – 60 months.

From line 149 – 151 in code we have created another boxplot for credit amount and class.



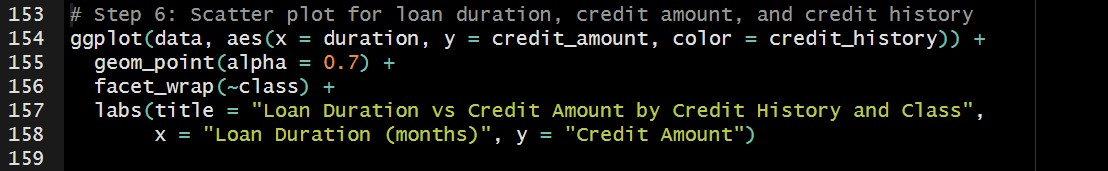
**Results**

1. Bad customers have loan deposit of ranging from almost 1800 – 4900$ with median at 2500$. There are many outliers in data (specifically bad class) who took credit of more than 10,000$.
2. Good customers have loan deposits ranging from 1800 – 3000$ with median at 2400$. It means half of the values in dataset are less than or equal to 2400$. We have many outliers, ranging from loan deposits of 7000 – 16000$.

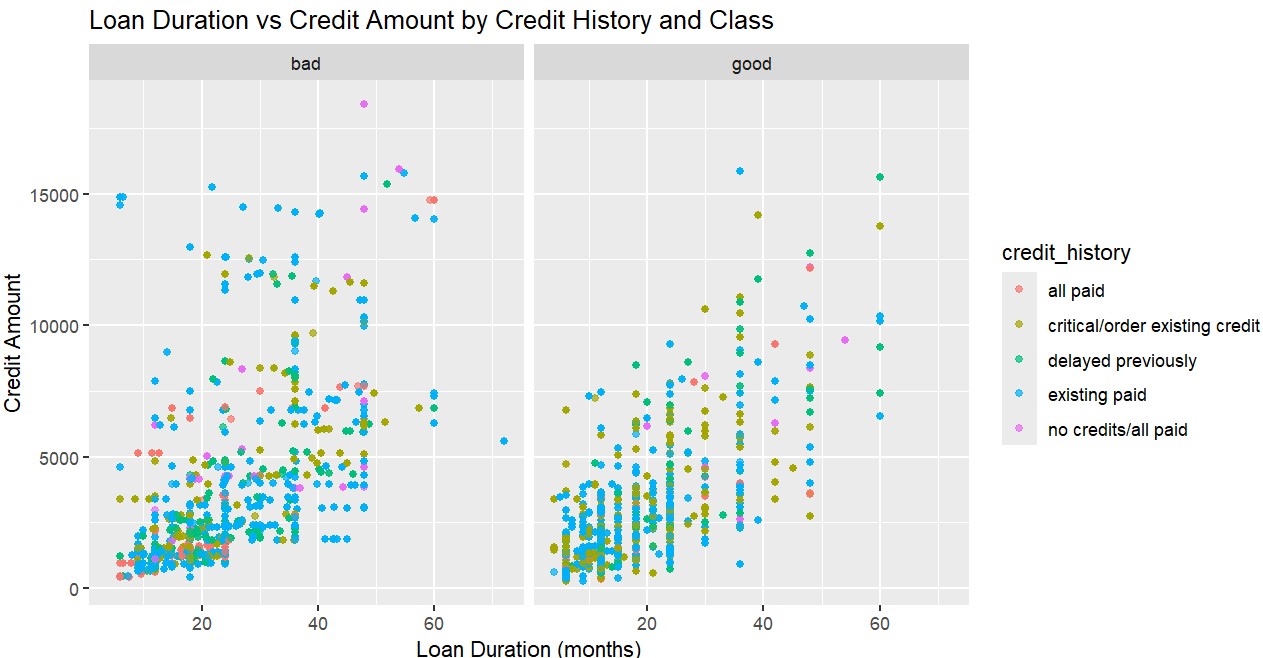
Note: Outliers are data points which are not well suited with our data. We can assume these as exceptions.

**Interrelating loan duration, credit amount for customer class**

In next part we will be plotting loan duration, credit amount and credit history.



This scatter plot presents a l**oan duration (x-axis) and credit amount** (y-axis) analysis, separated by **credit class (bad or good)** and there are different colors based on **credit history**. We will have this graph:



**Results:**

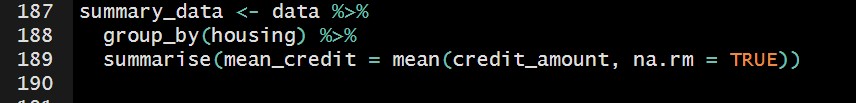
1. Bad class contain greater number of data points having greater credit amounts as compared to good class.
2. Linear behavior can be seen in ‘good’ class, which indicate that higher credit amount is associated with longer duration period in ‘good’ class.

**Results derived after visualizing:**

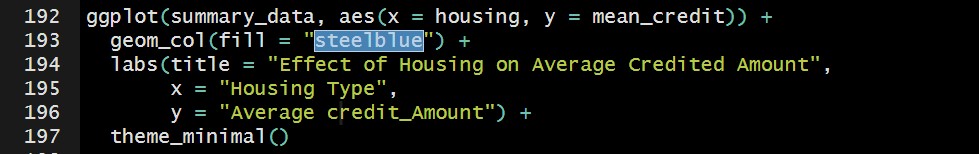
1. For bad class, we have median duration almost 24 months. Average duration in paying loan is 15 – 36 months.
2. For good class, we have median duration almost 18 months. Average duration in paying loan in 12 – 24 months. There are few outliers in data, indicating that few customers among good class had loan duration of 40 – 60 months.
3. Bad customers have credit amount 1800 – 4900$ with median at 2500$.
4. Good customers have credit amounts ranging from 1800 – 3000$ with median at 2400$.
5. In good class, higher credit amount is associated with longer duration period.
6. In ‘bad’ class, customers take greater credit amount as compared to customers of ‘good’ class.

### **3.3.3 Investigate the effect of housing type on the credited amount.**

Studying the impact of housing type on the amount that is credited by the user is important in understanding customer behaviors and financial analysis of data i.e. if a person rents a house or owns the house, how it impact how much money he took as loan? In order to study this we use same ggplot library.

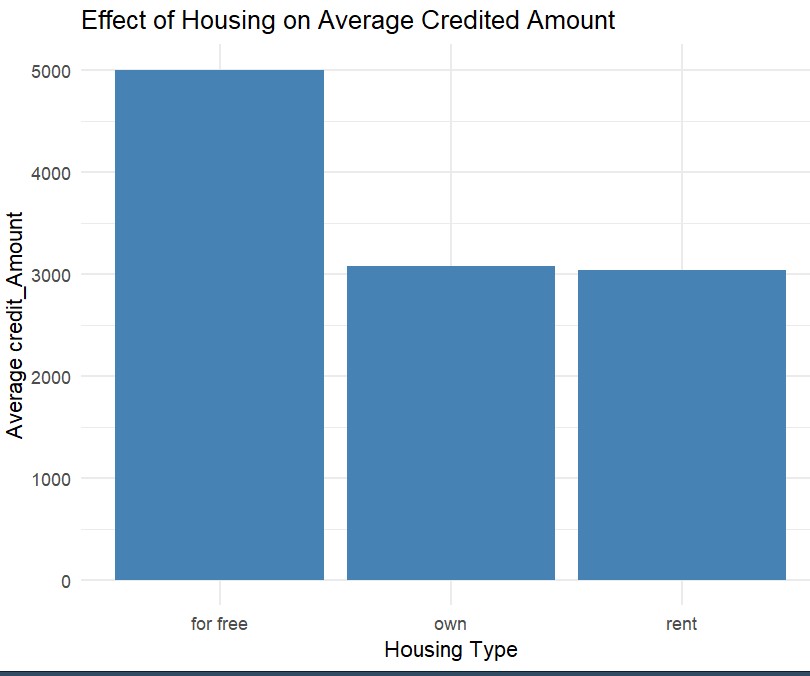


* In line 188, group\_by (housing) this function groups the entire data set into three categories, “rent”, “own”, “for free”.
* mean(credit\_amount) computes the mean for each housing type separately and summaries it.
* na.rm=TRUE ensures that null values are ignores when computing the mean, even thouh our data doesn’t contain any null values.
* This result in a new summary\_data dataframe with two columns.
  + 1. Housing which contain unique values after grouping housing into three categories
    2. Mean\_credit that have mean value of credit amount for each housing group.



* We will create plot by using ggplot library.
* Using the new summary data that we created in line 187-189.
* For x argument, we pass housing (own, rent, for free) and for y argument we pass values of mean credits for each housing category.
* geom\_col( ) function to specify the colors of bards.

The graph we get as result:



What does **‘for free’** mean?

There can be data of customers living in:

* House provided by family or friend.
* House from government (free housing programs).
* House from company.

‘for free’ indicate that person does not have direct financial obligations related to housing.

**Results from this graph:**

1. People owning a house or living on a rent house, both are taking credit amount of 3000$ almost.
2. People living in house for free tend to take greater amount as credit, almost 5000$.

## **3.4 Objective 4: Customer Demographics and Financial Behavior**

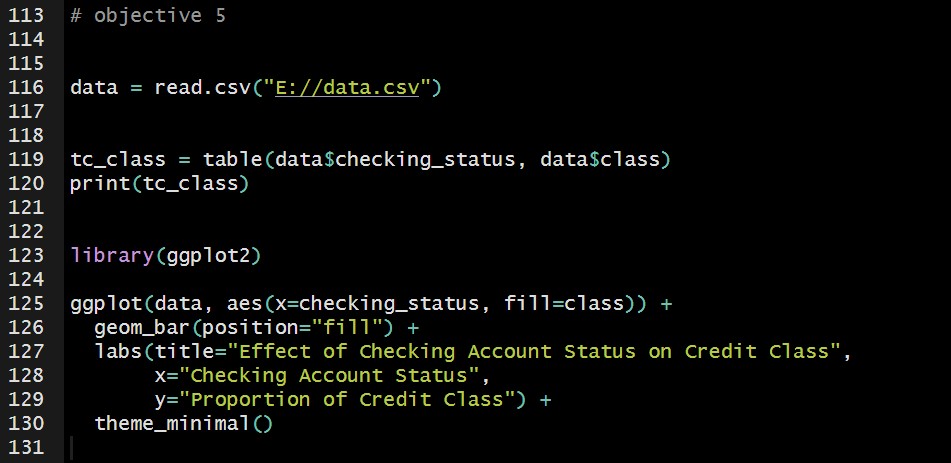
### **3.4.1 Analyze the effect of checking account status on class assignment.**

Now it’s time to analyze type of customer based on the credits they have in their checking account. We have 4 categories of checking account in our dataset. These are:

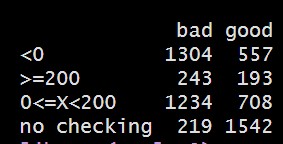
* Customers with no checking account
* Customers having credit less than 0$ (negative).
* Customers having credit between 0 and 200$.
* Customers having credit equal to or more than 200$.

We will analyze type of customers (good or bad) for each of the checking account category.

Firstly we will make table of checking status and customer type using table() keyword. Then using ggplot library, we will plot for better understanding. Note that we also make graph of relation between checking status and customer type without using the table.

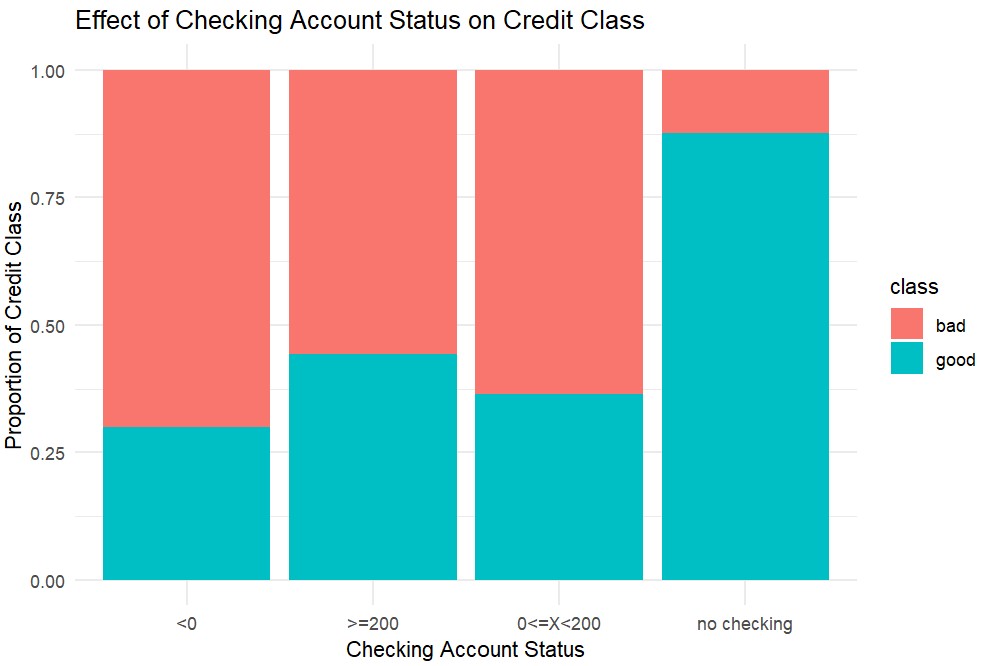


In line 119 we have constructed the required table. Output of line 120 looks like:



This output gives count of good customer and bad customer for each of the checking status category. From line 125 to 130 in code, we have plotted checking status data (on x-axis) and class data (on y-axis). Note that we have two kind of classes, there are will be two kind of bars for each of the checking status.

The plotted graph looks like:

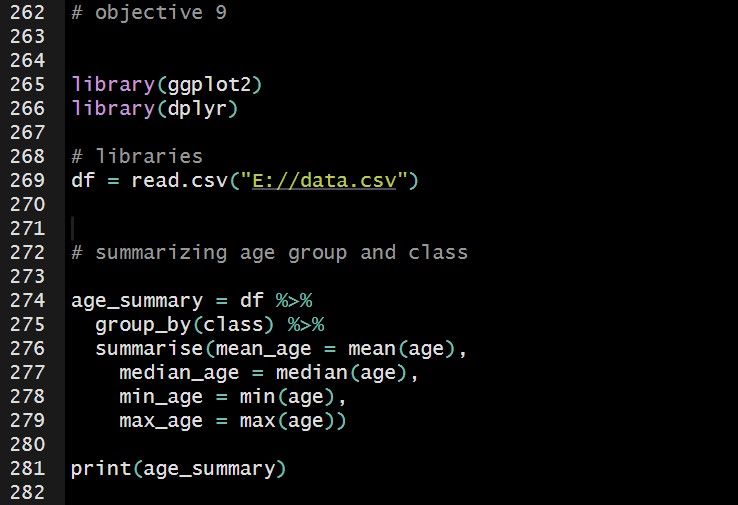


**Results derived after visualizing:**

1. More than 50% of the customers who have checking accounts are bad customers not depending on the amount they have in their checking account.
2. More than 90% customers who don’t have checking account are mentioned as good customers in our dataset.
3. Among the customers having checking account, those who have more than 200$ in there account are comparatively greater in number (for good class) than other two categories (negative balance or balance between 0 and 200$).

### **3.4.2 Explore the influence of age on class classification.**

In this objective we will analyze whether age have influence on class of customer (whether good or bad) or not. If, what is the trend of that effect?



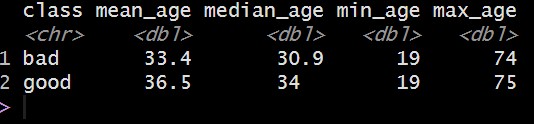
In this part of code we have loaded the required libraries and the dataset.

group\_by(class) function is used to split the data into two groups: one for customers with a "good" class and another for customers with a "bad" class.

summarise() function is applied to get summary statistics for age.

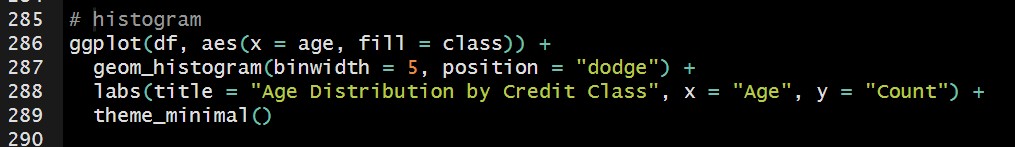
* mean(age) calculate average of the age.
* median(age) calculate median of the age.
* min(age) gives minimum value of age.
* max(age) gives maximum value of age.

Output of line 281 looks like:



As expected, we have mean, median, minimum value and maximum value for both, good and bad class. It looks like a well structured data with not so much deviated values. Note that minimum age of bad and good customer is 19. Maximum age of both the categories are again very close to each other (74 and 75).

Now let’s visualize this data using a histogram.



Output will looks like:



**Results from this graph:**

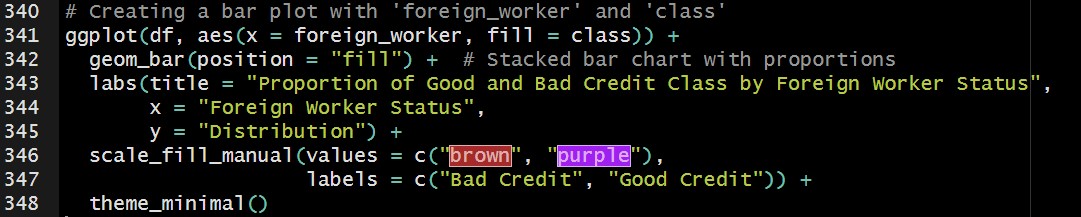
1. Majority of the customers are of age 20 – 40 age.
2. Number of customers gradually decrease from age of 40.
3. There is a significant decrease of customers after age of 60.
4. In younger age groups, there are more customers classified as bad.
5. Mean age of bad customers is 33 while mean age of good customers if 36.
6. 50% of the bad customers are under age of 31, while 50% of the good customers are under age of 34.

### **3.4.3 Explore relation between foreign\_worker and credit class**

Whether the customer of bank is foreign\_worker or not, can help predict the class of customer. This is because foreign workers may have better stability and financial status compared to local workers which can in return have influence on them being good or bad customers for the bank.

In this case, we will plot stacked bar chart to show distribution of “good” or “bad” class for both foreign and local workers.

The R code snippet is as below:



* Line no: 341 of code specifies that we are using df dataset, foreign\_worker will be along x\_axis and stacks will be filled according to class distribution.
* geom\_bar( ) method allows us to make a stack plot where bars are separated on the basis of proportion allowing us to compare good or bad classes.
* scale\_fill\_manual( ) function allows us to customize color schemen according to our requirements or choices.
* Labs ( ) is used to add title and labels to the axes.
* theme\_minimal( ) displays the graph with better visualization.

The graph we get as a result is shown below



**Results from this graph:**

1. The majority of non foreign workers fall in food customer category (more than 75%)
2. Only few local workers (non-foreign) are bad customers.
3. Almost 50% of foreign workers are of good customer category and remaining 50% are from bad customer category.

We conclude that local workers are more likely to be good customers as compared to foreign workers which indicate that local workers have good financial stability.

# **4. Conclusion:**

## **4.1 Summary of Analysis**

After we have conducted detailed analysis of dataset using various techniques and tools of Rstudio and derived useful results, it is important to summarize those results into useful information.

More than 50% of customers with checking accounts are categorized as bad customers regardless of their balance whereas majority of those without checking balance fall into the category of good customers. Younger individuals (less than 30) are more likely to have bad credit whereas those older than 40 tend to have good credit behavior. Single customers are more likely to have good credit while those with other personal statuses tend to have bad credit.

Moreover, the loan durations for bad customers are generally longer across all age groups. Vast majority of people who took load are skilled in their career. The distribution of good or bad credit across foreign and local people are nearly equal. Loan amounts for bad customers are typically larger, with several outliers in both categories. Overall, these findings highlight trends in age, account status, and credit behavior that can help banks identify risks.

## **4.2 Recommendations**

**Focus on young people for risk detection:**

Since, young people are more likely to default and cause loss to bank, Investors and banks should keep check on young people for any kind of risk mitigation.

**Reviewing loan duration:**

People who take loan with duration more than 24 months are at relatively greater risk for causing default. Therefore, proper check and balance should be maintained in such situations.

**Offer financial education:**

Single males and younger individuals (less than 30) should be offered financial education which may help them in managing finances in a better way and manage credit more responsibly.

**Implement strict loan policies:**

The customer which have high existing credits are more prone to default, therefore, stricter policies should be implemented when loan is given for high credits to hold those people accountable.

**Prioritize customers with consistent employment:**

Skilled people make large portion of customers and they show balanced credit performance. Therefore, these individuals with consistent employment should be prioritized by offering favorable policies.

**Limit loan durations for customers over 60:**

Older customers tend to show less defaults when their loan amount is limited and granted for shorter period. Therefore, loan durations should be limited in such cases.

**Offer customized loan products:**

Customer with high number of dependents and large existing credits tend to have moderate impact on default rate and require special attention when loan is being structured. Therefore, they should be offered customized load plans that minimize the default rate.

## **4.3 Limitations and Future Directions**

### **Limitations:**

1. The dataset shows imbalance in certain categories. Such as there is very small fraction of non-foreign workers which limit generalizability of these groups. In addition, single males are over represented in good credit category which may skew conclusions.
2. This dataset does not contain customer further financial information such as as stability of their income which could have offered a clearer picture of credit risk.
3. The classification of customers into “bad” or “good” category is binary which may oversimplify the behavior. A more nuanced classification could have provided better insights.
4. The analysis does not account for external factors like economic conditions or geographic variations that could impact customer creditworthiness.

### **Future Directions:**

1. Future analyses should include time-series data to track customer credit behavior over time, providing a more dynamic and vast understanding of risk.
2. Collecting more detailed information on customer’s assets, financial obligations and liabilities could help in refining risk assessments.
3. Further research should focus on why young individual are more at the risk of having bad credit and how educations programs cam help reduce that risk.
4. If we move beyond binary classification, implementation of machine learning models to make predictions based on various factors can help in better risk assessment.
5. Analyzing personal statuses categories in more depth may help to yield more thoughtful insights.

# **5. Workload matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student name | Student ID | Tasks | Percentage | Signature |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# **6. References**

Daróczi, G. (2022). *pander: An R Pandoc Writer*. From rdocumentation: https://www.rdocumentation.org/packages/pander/versions/0.6.5

*ggplot2*. (n.d.). From ggplot2: https://ggplot2.tidyverse.org/

*pie-charts*. (n.d.). From byjus: https://byjus.com/maths/pie-chart/

*what is a Histogram?* (n.d.). From asq: https://asq.org/quality-resources/histogram

Wickham, H. (n.d.). *dplyr*. From dplyr: https://dplyr.tidyverse.org/

Yi, M. (n.d.). *A complete guide to bar charts*. From ATLASSIAN: https://www.atlassian.com/data/charts/bar-chart-complete-guide

Yi, M. (n.d.). *A complete guide to box plots*. From ATLASSIAN: https://www.atlassian.com/data/charts/box-plot-complete-guide

Yi, M. (n.d.). *A complete guide to scatter plots*. From ATLASSIAN: https://www.atlassian.com/data/charts/what-is-a-scatter-plot

Yi, M. (n.d.). *A Complete Guide to Stacked Bar Charts*. From ATLASSIAN: https://www.atlassian.com/data/charts/stacked-bar-chart-complete-guide