German Credit Risk Analysis

Code

Kailash Baskar - W0790883

Project

This project is about exploring the famous German credit risk data set and provide insights on whether a particular customer's credit is a good one or bad.

Write up

Introduction

The German credit data is the information collected from random 1000 customers about various attributes such as Credit amount, amount which is currently in their savings and checking account, their reason to avail credit from bank, tenure period of the credit and some personal information including age, sex, their occupation and whether they have their own house or not.

With these available information the project is intended to predict and showcase the relationship between the available attributes in the data and to check which customer's credit is at Risk or not.

Also as part of this project using various regression techniques a new model will be created which helps whether a customer with the existing attributes will be a good credit candidate i.e will be paying back the credit on time or is there any risk associated in approving the loan to him.

Analysis:

Loading the data required libraries:

Let's load the data set and explore the variables in it.

Code

Exploratory data analysis

In Exploratory data analysis we shall explore how wide spread the data is and also do the required cleanup and transformation and perform the required analysis.

Dimension of the data set.

Code

```
## [1] 1000   11
```

Let's have a glimpse of data set to ensure if all the variables are needed for our analysis.

Code

```
## Rows: 1,000
## Columns: 11
## $ ...1              <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ Age               <dbl> 67, 22, 49, 45, 53, 35, 53, 35, 61, 28, 25, 24, 22,~
## $ Sex               <chr> "male", "female", "male", "male", "male", "male", "~
## $ Job               <dbl> 2, 2, 1, 2, 2, 1, 2, 3, 1, 3, 2, 2, 2, 1, 2, 1, 2, ~
## $ Housing           <chr> "own", "own", "own", "free", "free", "free", "own",~
## $ `Saving accounts` <chr> NA, "little", "little", "little", "little", NA, "qu~
## $ `Checking account` <chr> "little", "moderate", NA, "little", "little", NA, N~
## $ `Credit amount`   <dbl> 1169, 5951, 2096, 7882, 4870, 9055, 2835, 6948, 305~
## $ Duration          <dbl> 6, 48, 12, 42, 24, 36, 24, 36, 12, 30, 12, 48, 12, ~
## $ Purpose           <chr> "radio/TV", "radio/TV", "education", "furniture/equ~
## $ Risk              <chr> "good", "bad", "good", "good", "bad", "good", "good~
```

Data cleaning and transformation

From the data looks like the first column is just Serial numbers and it would not be useful for our analysis, hence removing the first column in the data set.

Code

…	Sex	…	Housi…	Saving accounts	Checking account	Credit amount	Duration	Purpose
<dbl⌅chr>		<dbl⌅chr>		<chr>	<chr>	<dbl>	<dbl>	<chr>
67	male	2	own	*NA*	little	1169	6	radio/TV
22	female	2	own	little	moderate	5951	48	radio/TV

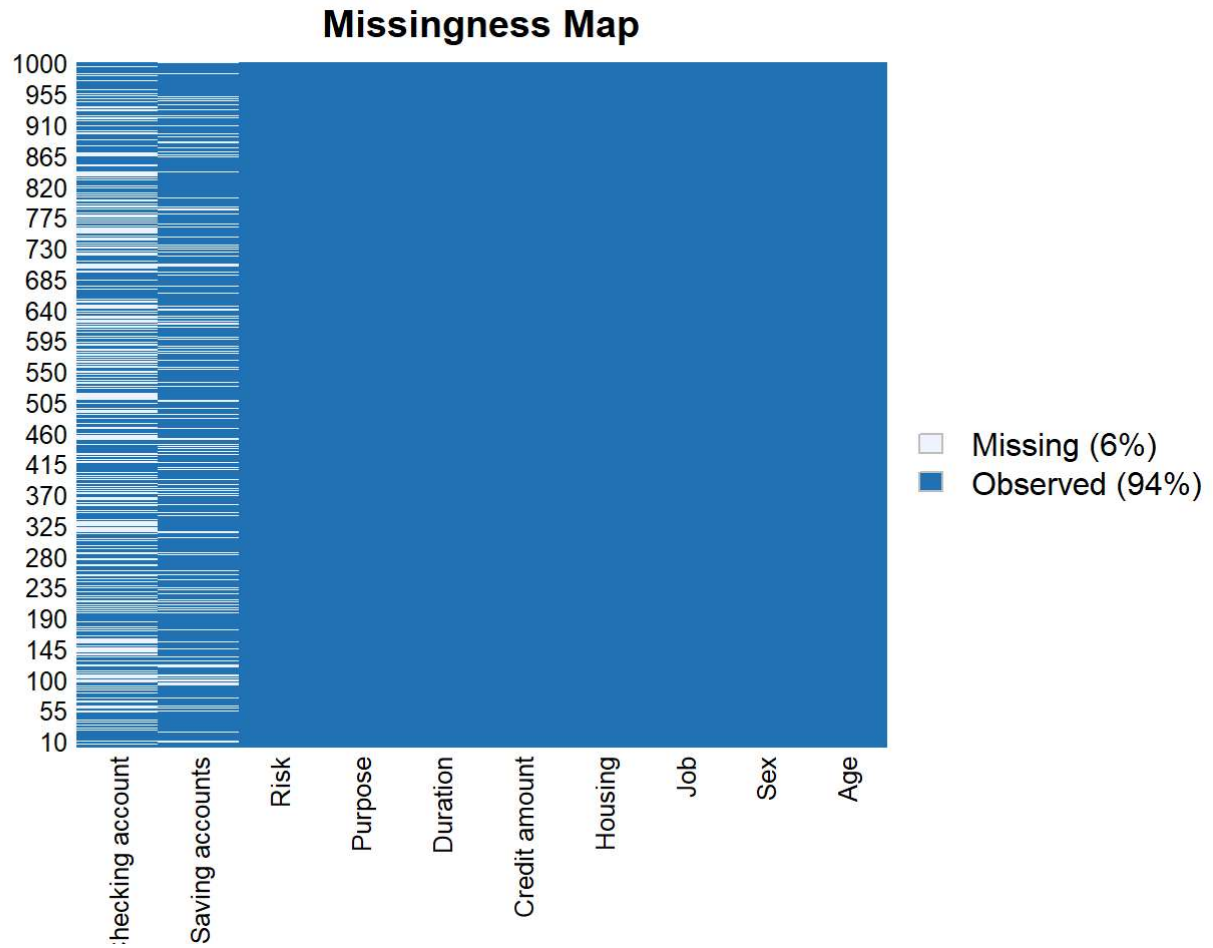
2 rows

Let's check if there are any missing values in any variables in it.

Code

```
##              Age              Sex              Job          Housing
##                0                0                0                0
##  Saving accounts Checking account    Credit amount         Duration
##              183              394                0                0
##          Purpose             Risk
##                0                0
```

Code



From the above graph we could see the variables `Checking account` and `Savings account` has multiple `NA` values in it.

Calculating the percentage of NA values to see if it is less than 5% then we can remove those observations.

Code

```
## [1] "NA count for Saving accounts"
```

Code

n
<dbl>
18.3

1 row

Code

```
## [1] "NA count for Saving accounts"
```

Code

n
<dbl>

	n
	<dbl>
	39.4

1 row

Looks like the NA values contribute to a higher percentage of data. So replacing the `NA` observations to some value to make the data useful for our analysis.

Replacing the value “NO” for the observations which has “NA” in Savings and Checking account variables.

Code

Code

gd$`Saving accounts` <chr>	n <int>
little	603
moderate	103
NO	183
quite rich	63
rich	48

5 rows

Code

gd$`Checking account` <chr>	n <int>
little	274
moderate	269
NO	394
rich	63

4 rows

Replaced the “NA” values to “NO” to indicate there is no savings or checking account for that customer.

To check the correlation of these account types lets map them to a numerical value to find correlation with other variables.

Code

Also mutating other relavent information into numeric values to calculate correlation.

Code

The current data set which we have after cleaning and transforming some of the variables.

Code

...	Sex	...	Housi...	Saving accounts	Checking account	Credit amount	Duration	Purpose
<dbl>	<chr>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
67	male	2	own	NO	little	1169	6	radio/TV
22	female	2	own	little	moderate	5951	48	radio/TV

2 rows | 1-10 of 16 columns

Analysis:

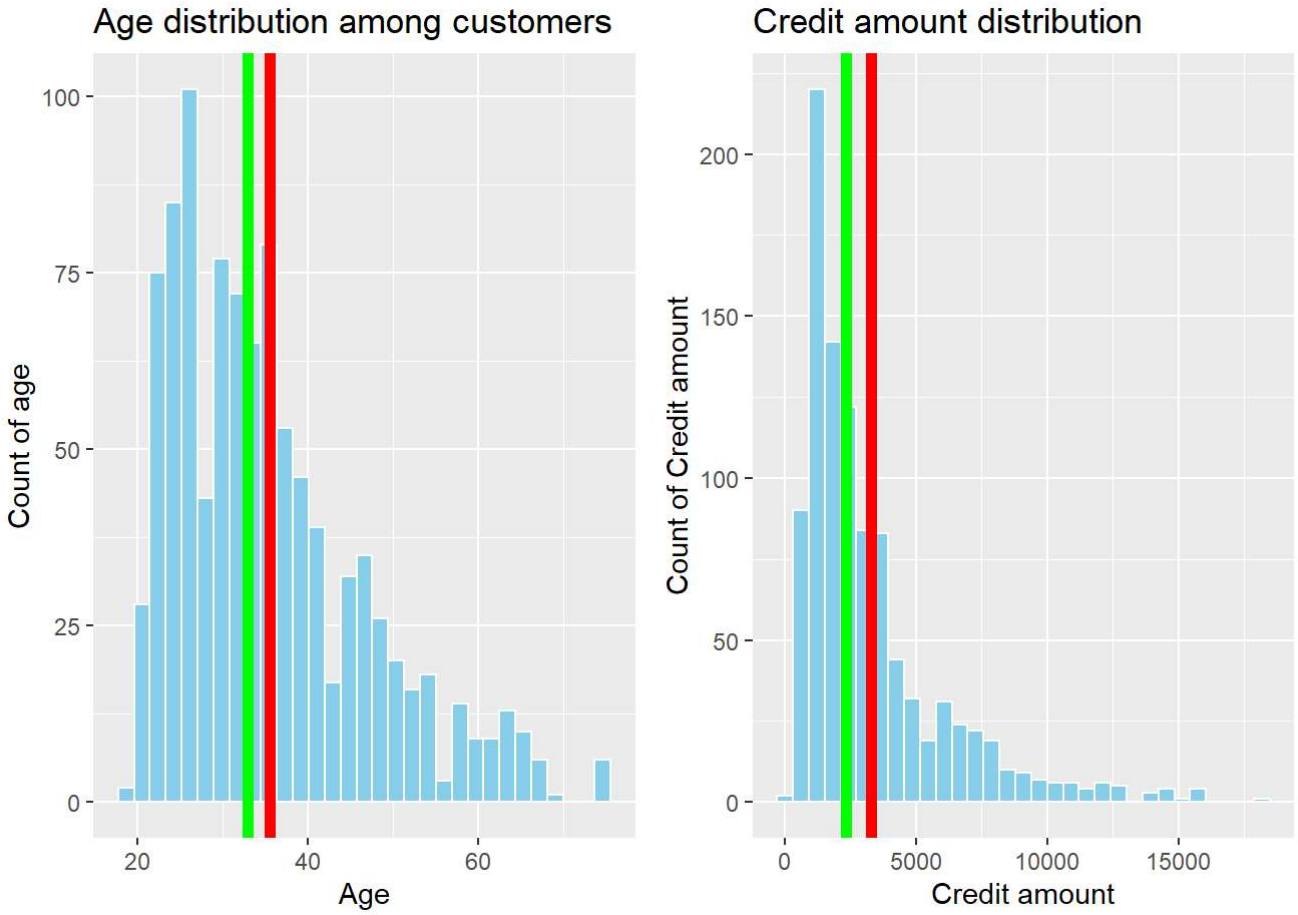
Plotting the variable helps in identification of their measure of spread, central tendency and other important features.

Univariant Analysis:

First let's explore the numerical variables Age and Credit Amount.

Code

Code



The red line indicates the mean and green line indicates the median.

Both the graphs indicate the variables are right skewed and unimodel.

The measure of spread of these variables are displayed below.

Code

```
## [1] "IQR for Age"
```

Code

```
## [1] 15
```

Code

```
## [1] "IQR for credit amount"
```

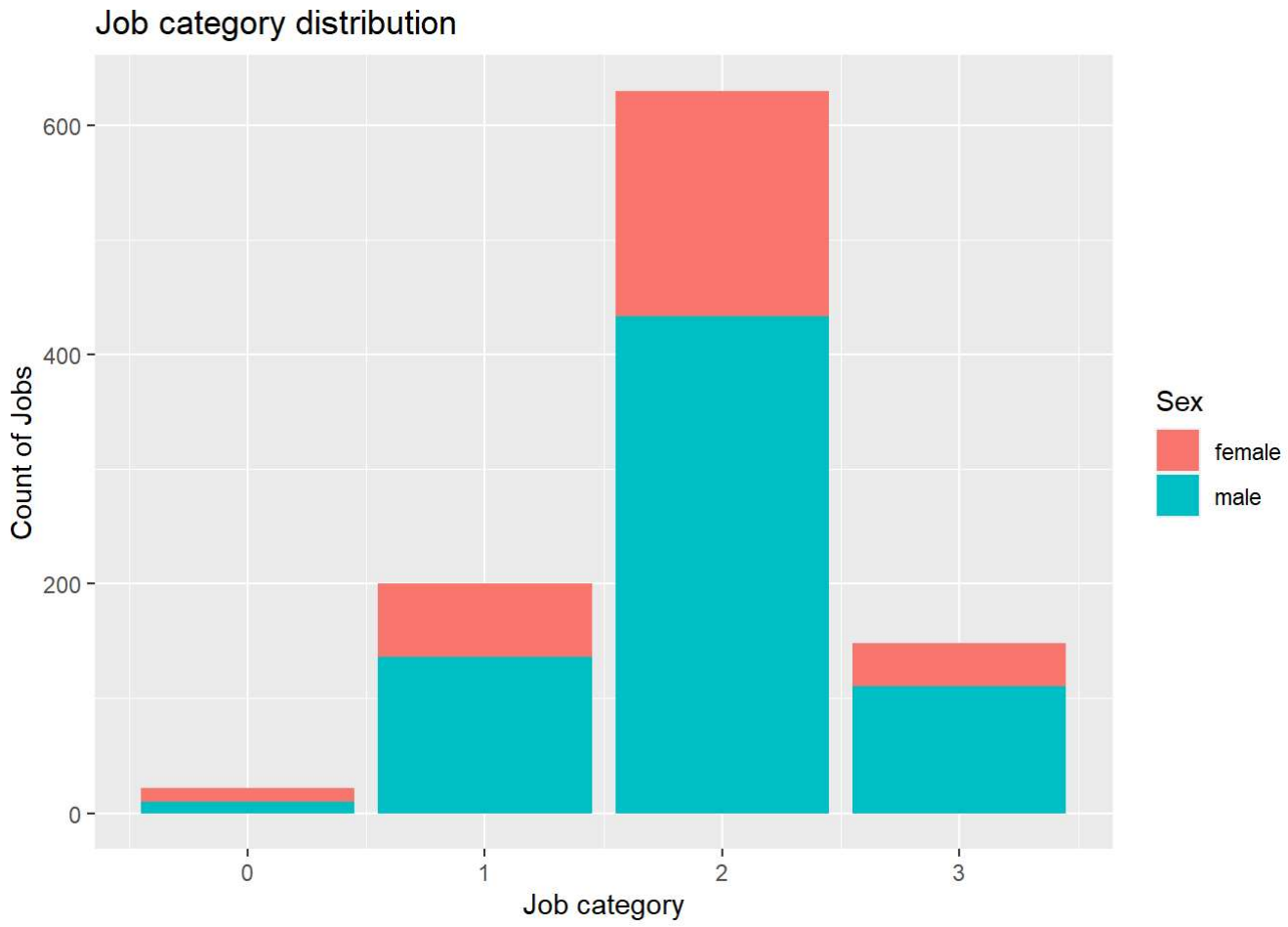
Code

```
## [1] 2606.75
```

Bivariant Analysis:

Exploring other variables to check their distribution. The variable Job indicates the job category, using visualization plotting the bar graph along with Sex as distribution.

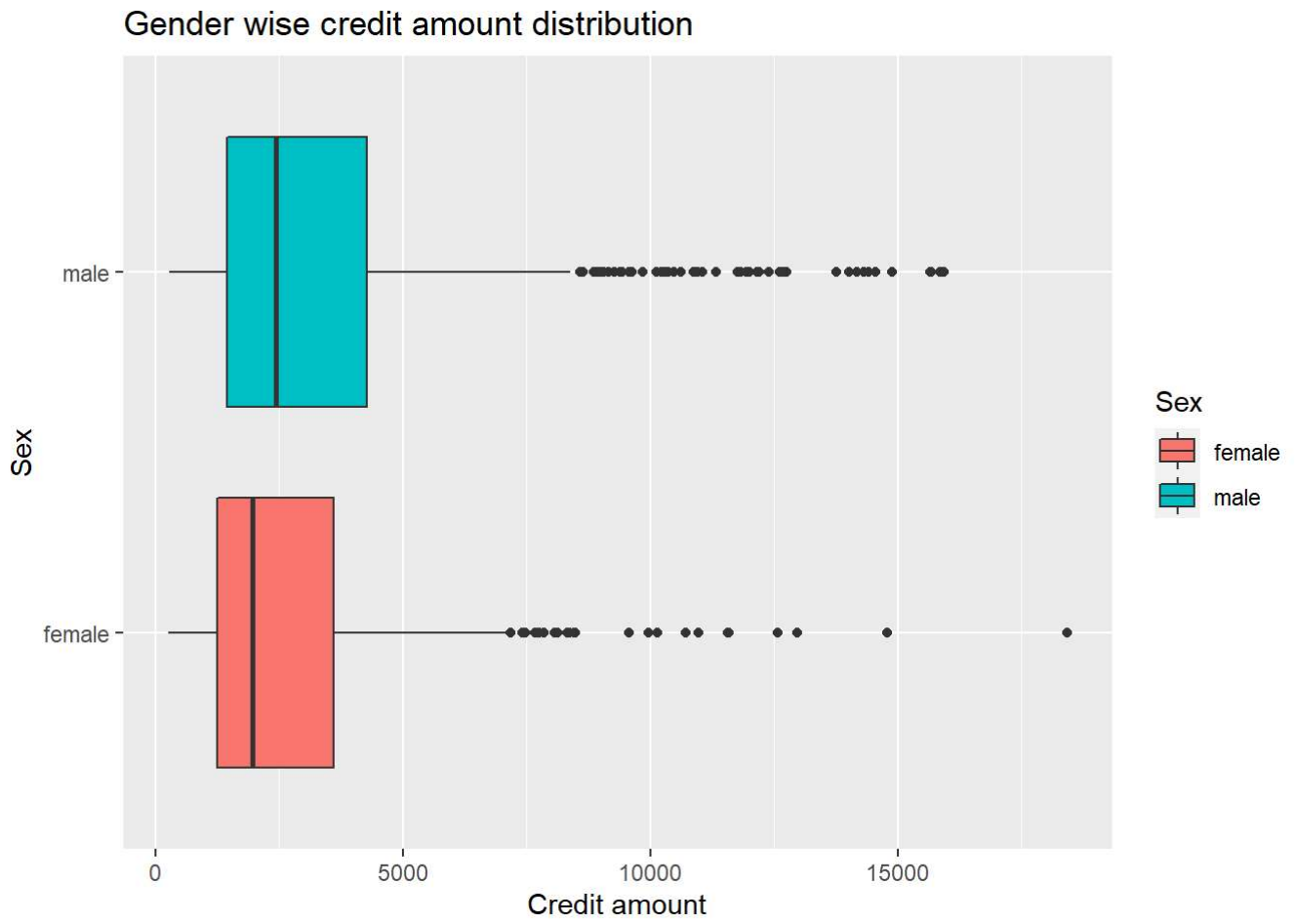
Code



From the graph it looks like there are more customers working in Job Category “2” and in all the categories the male customers are more than female customers.

Finding out which gender has the most credit amount.

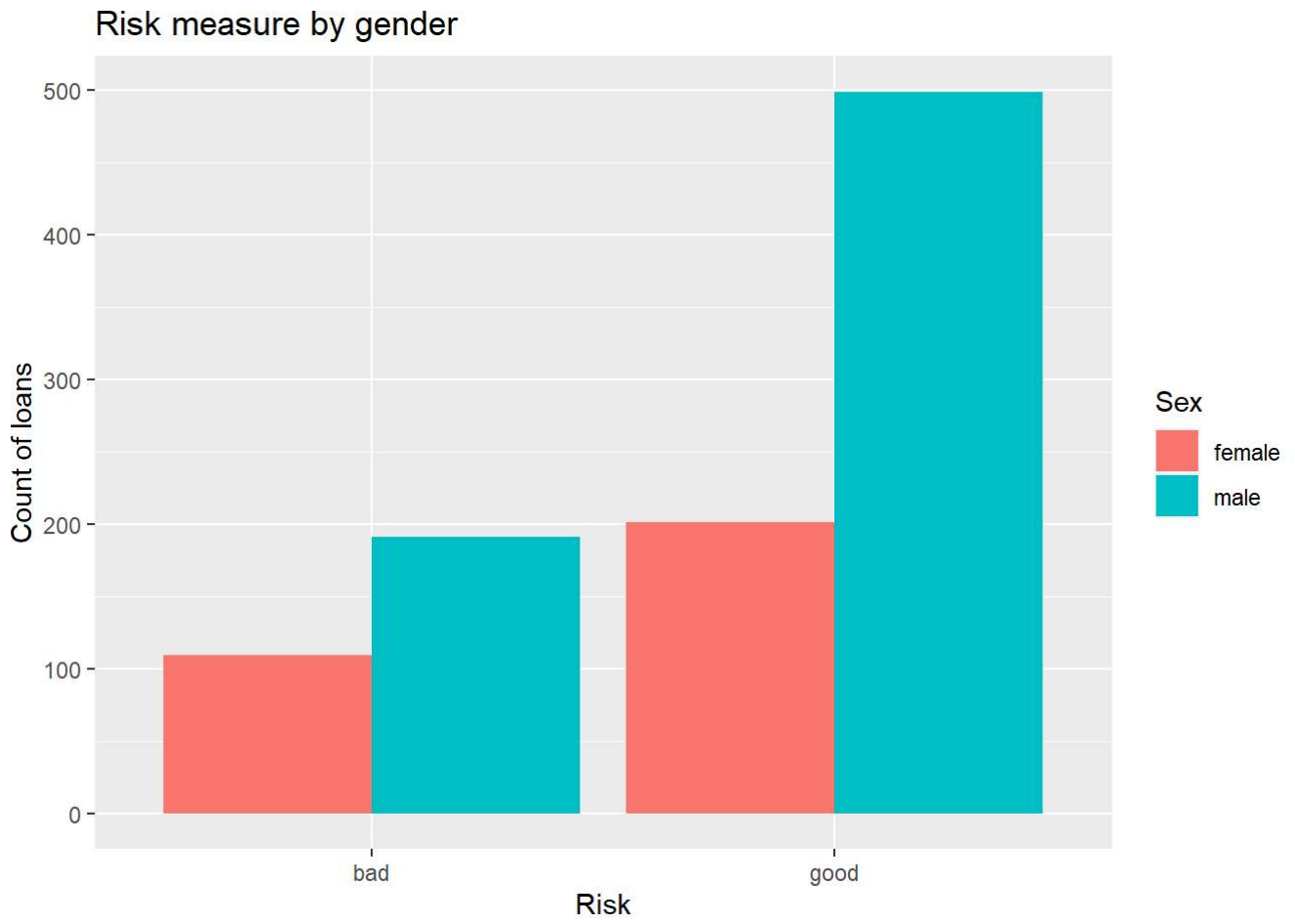
Code



From the graph it looks like though credit amount for females is less than males there are few female customers with high credit amount.

Let’s check the distribution of good and bad loans by sex.

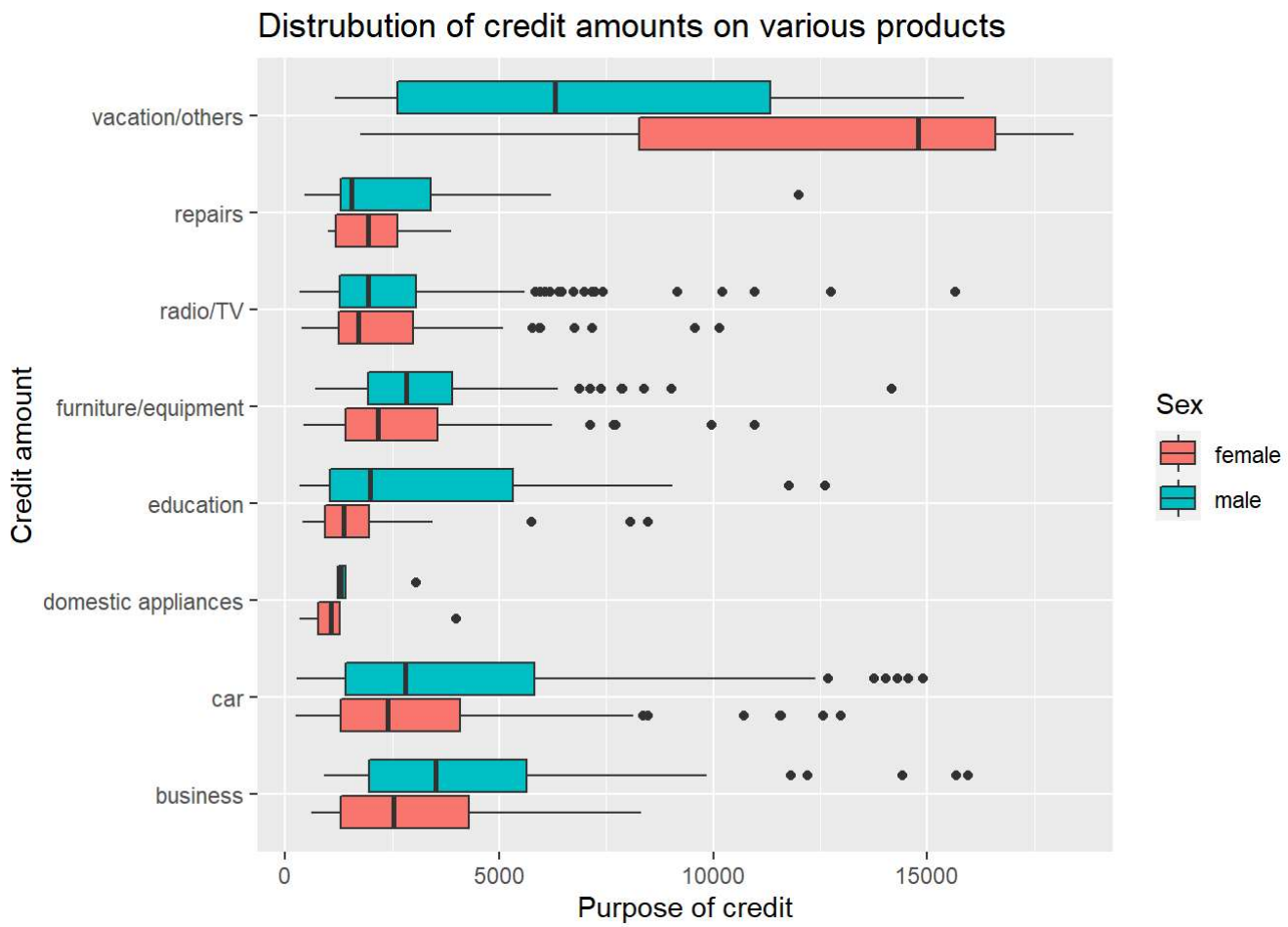
Code



From the graph it looks like more male customers are having a positive good risk score when compared to the females and looks like there are more risk associated with female customers.

A plot to compare credit amount with the purpose of credit.

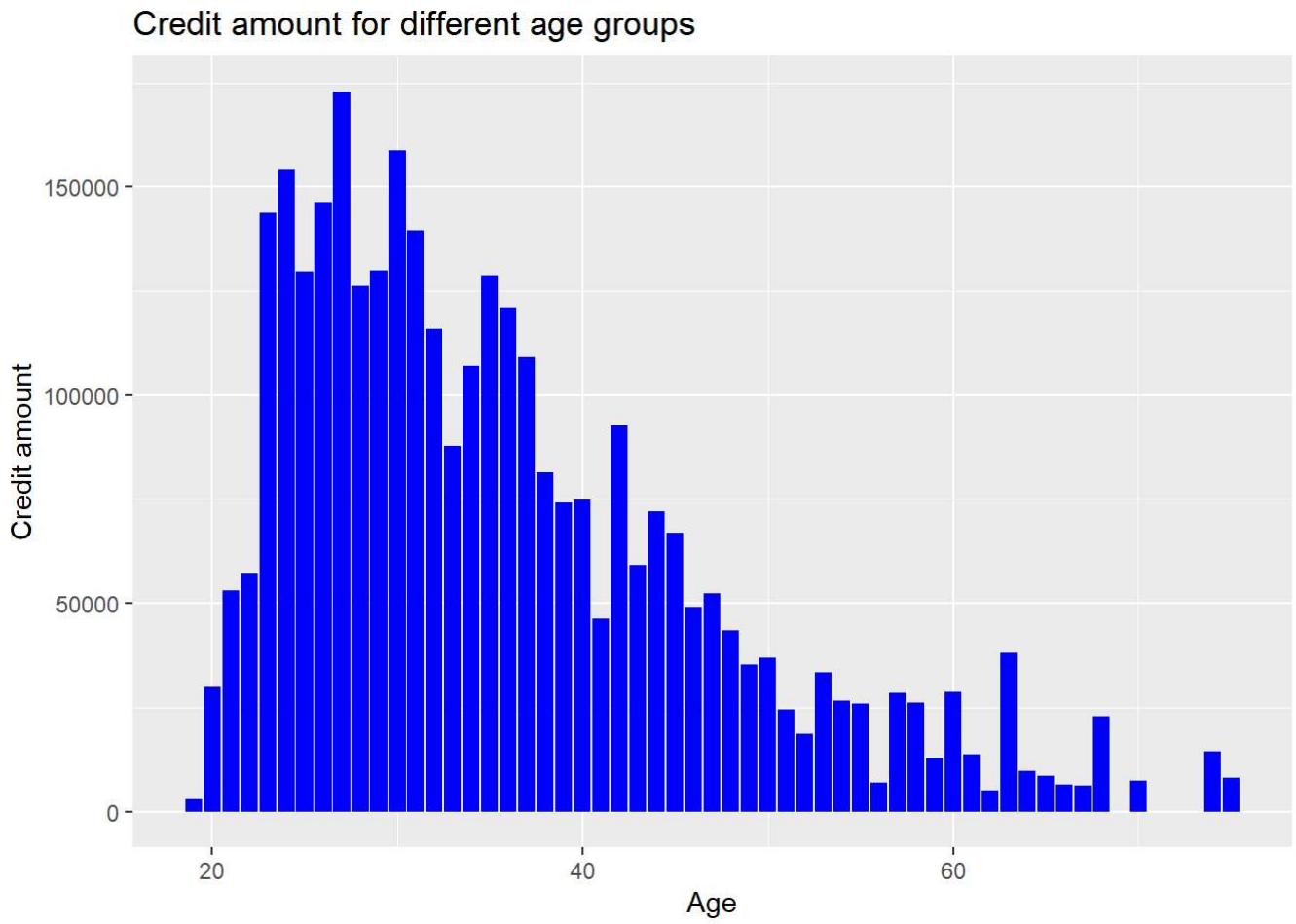
Code



Looks like customers are taking more credit towards vacation and other activities more when compared with the other purposes of credit and to be specific Females take the most loans for vacation when compared with Males.

Let’s find the distribution of customers Age with their credit amount received from bank.

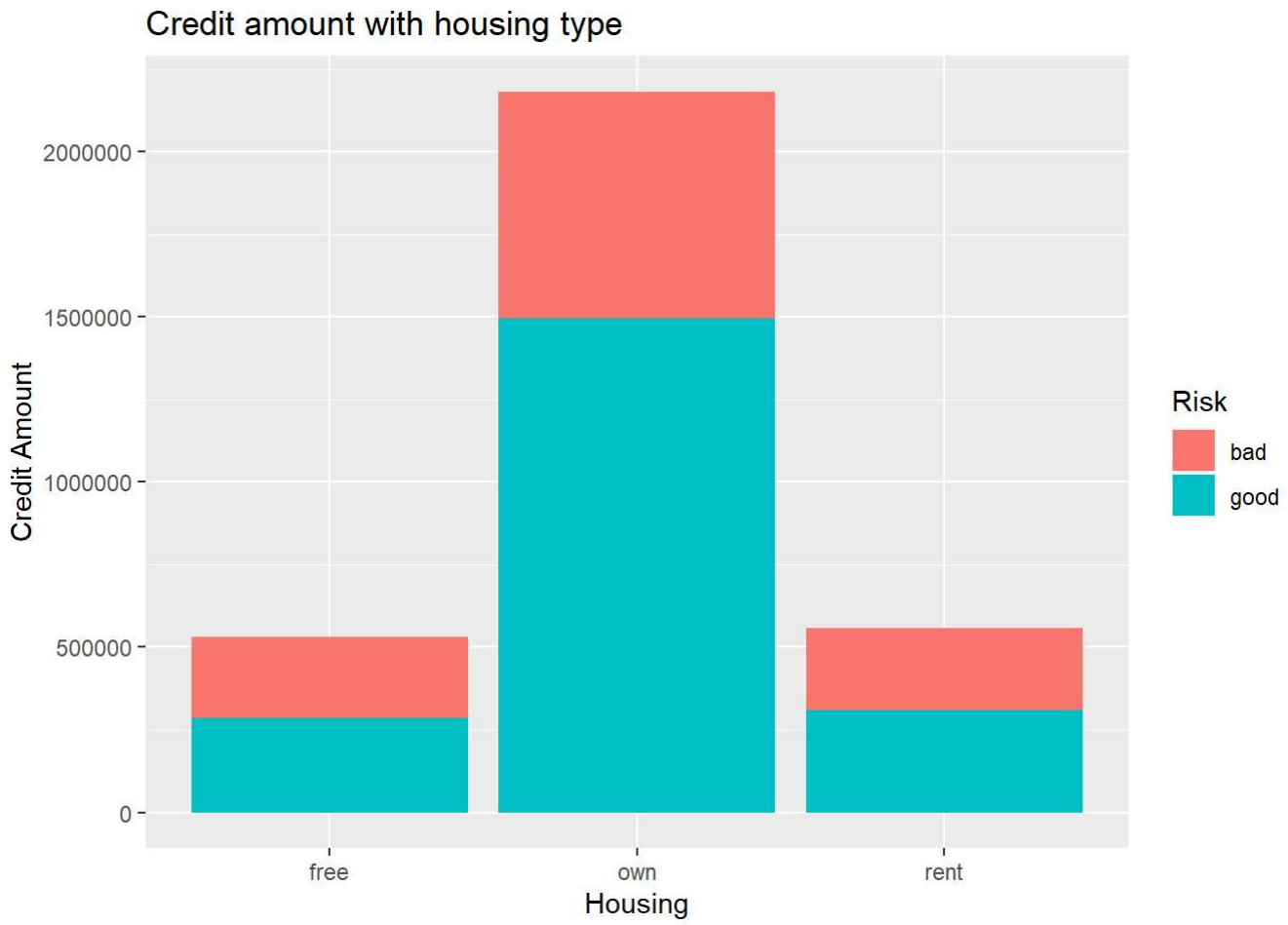
Code



From the graph it looks like customers in the age range of 23-40 has received the most credit amount when compared to the other age people.

Comparing credit amount with the customer's house variable, to ensure if there are more loans given for customers with house or not.

Code

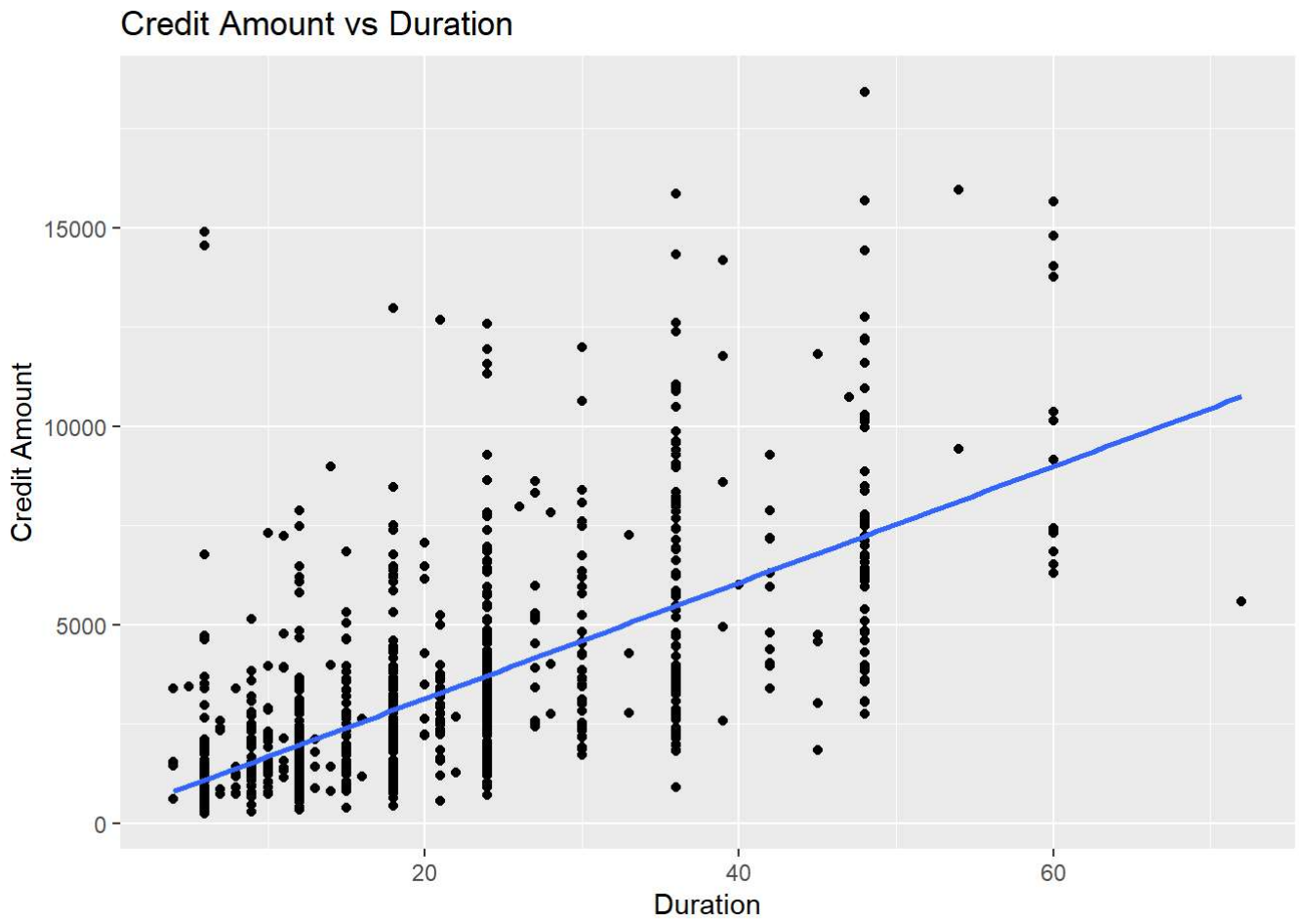


Looks like customers with own house has received the most credit when compared to the other customers and also the risk factor is considerably low for the customer who owns a house.

Lets find the distrubution about how much duration varies with various credit amounts received by the customers.

Code

```
## `geom_smooth()` using formula 'y ~ x'
```



From the above graph it is clear that as the duration increases the credit amount also increases i.e in other words the larger the credir amount the longer the credit duration.

Let's find the correlation between the different variables with Risk to identify which variable correlates to the most risk value.

Code

```
## [1] "Correlation of Age and Risk"
```

Code

```
## [1] 0.09112741
```

Code

```
## [1] "Correlation of Credit Amount and Risk"
```

Code

```
## [1] -0.1547386
```

Code

```
## [1] "Correlation of Duration and Risk"
```

Code

```
## [1] -0.2149267
```

From the correlation values looks like the Duration and Credit amount are negatively correlated to the risk factor. As the duration of the loan tenure increases the credit is supposed to be a good loan and similarly as the credit amount increases the loan is a good loan.

Some statistical calculation to find out which category of customers with various housing and Job type has got the most credit amount from the bank and their average age and duration of the loan.

Code

```
## [1] "Summary of mean grouped by House category"
```

Code

Housing <chr>	mean(`Credit amount`) <dbl>	mean(Age) <dbl>	mean(Duration) <dbl>
free	4906.213	43.81481	27.45370
own	3060.940	35.59327	20.32819
rent	3122.553	30.36872	19.24022

3 rows

Code

```
## [1] "Summary of mean grouped by Job category"
```

Code

Job <dbl>	mean(`Credit amount`) <dbl>	mean(Age) <dbl>	mean(Duration) <dbl>
0	2745.136	40.09091	17.36364
1	2358.520	36.54000	16.53500
2	3070.965	34.25397	21.41111
3	5435.493	39.02703	25.16892

4 rows

From the obtained information it looks like the people who are in free houses has got the most credit from the bank when compared to the Own house owners or rental people. The People in the free house category are also little old when compared to the other category people and they got a larger loan duration as well.

Similarly people who are belong and are working in job category 3 has got the maximul credit amount from the bank and also has the second highest age when compared with others and with the highest loan duration.

Linear model:

Lets create a small linear model comparing Age with the credit amount.

Code

```
##
## Call:
## lm(formula = gd$Risk_num ~ gd$`Credit amount`, data = gd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7713 -0.6040  0.2541  0.2924  0.6163
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.782e+00  2.194e-02  81.220  < 2e-16 ***
## gd$`Credit amount` -2.513e-05  5.080e-06  -4.948  8.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4532 on 998 degrees of freedom
## Multiple R-squared:  0.02394,    Adjusted R-squared:  0.02297
## F-statistic: 24.48 on 1 and 998 DF,  p-value: 8.798e-07
```

So the least square for this regression model can be written as “r = 1.782 + (-2.513e-05) * Credit Amount” So 2.3% of the Risk variability is being explained by Credit Amount

Let’s extend this model for multiple variables to predict the risk.

Creating a model with all the variables to check if they correlate to the Risk factor.

Code

```
##
## Call:
## lm(formula = gd$Risk_num ~ as.numeric(gd$Job) + gd$Age + gd$Duration +
##     gd$Purpose_num + gd$Sex_num + gd$savings_numeric + gd$credit_numeric +
##     gd$Housing_num + gd$`Credit amount`, data = gd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1323 -0.4842  0.1915  0.3154  0.9444
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.480e+00  1.206e-01  12.268  < 2e-16 ***
## as.numeric(gd$Job)   2.041e-03  2.211e-02   0.092 0.926493
## gd$Age               2.377e-03  1.288e-03   1.845 0.065367 .
## gd$Duration         -9.028e-03  1.519e-03  -5.942 3.90e-09 ***
## gd$Purpose_num       1.469e-02  7.070e-03   2.078 0.037935 *
## gd$Sex_num           7.022e-02  3.080e-02   2.280 0.022848 *
## gd$savings_numeric   6.846e-02  1.147e-02   5.971 3.28e-09 ***
## gd$credit_numeric    4.426e-04  1.061e-04   4.172 3.28e-05 ***
## gd$Housing_num      -2.843e-02  2.812e-02  -1.011 0.312160
## gd$`Credit amount` -3.723e-05  9.831e-06  -3.787 0.000161 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4342 on 990 degrees of freedom
## Multiple R-squared:  0.111,  Adjusted R-squared:  0.103
## F-statistic: 13.74 on 9 and 990 DF,  p-value: < 2.2e-16
```

From the above summary statistics looks like the Risk factor prediction can be 11.1% accurate from the model which we have created by summing up all the variables.

Let's create a new model with a product of all these variables.

"summary(lm(gd$Risk_num ~ as.numeric(gd$Job) * gd$Age * gd$Duration * gd$Purpose_num * gd$Sex_num * gd$savings_numeric * gd$credit_numeric * gd$Housing_num + gd$`Credit amount`, data=gd))"

#Turned off the message output for second model as the output is lengthy and takes up more page space.

From the second model which we have created the R squared value is 0.3635 so about 36.3% of the data which will be fed into this model will predict the results accurately and help the bank in identifying a good customer to give credit without any hesitation.

Conclusion

Overall there were multiple parameters in the data set about the customer which indicated some of the customer's personal information as well as banking related information. We have performed extensive exploratory data analysis on continuous and categorical variable both univarent and bivarent analysis and identified some interesting aspects from the data.

We have compared all the variables and looking at their distribution and correlation with other variables using various plots and statistical approaches we could see that customers within age group of 23 - 40 has taken the most amount of credit from the bank. Mostly female customers are at a potentially high credit risk and banks

should review their credit application before proceeding further. In the same context customers with job category 3 has achieved the maximum credit from the bank and on an average their age are in the mid 40. The most credit has been given to vacations and other expenses were females were the most to claim the credit for this purpose.

On the other hand the credit amount with longer tenure has a very low risk factor associated with it. There is also a relation between the duration and amount where higher the credit amount higher the duration and lower the risk factor. From the available data looks like customers who has free house has claimed the most credit amount from the bank when compared to the other customer types.

As part of the project analysis we also calculated the IQR and correlation for various numeric variables and in the data there were many categorical variables which had a key role in calculating the credit risk factor, so converted the categorical data into numeric values and found the correlation with the risk factor. As expected the savings account variable had a very good correlation with the credit risk factor. Similarly the duration of the credit and age had a negative impact to the correlation of the risk factor which opened up new vision to the bank management.

A couple of models to predict the risk factors were also created as part of the project and these model outputs are compared against each other. On comparison the second model where the product of the variables helped to identify the risk factor by 36.35% accuracy and served a better model when compared with the first model where sum of the independent variables are used to predict the dependent variable Risk. Using this model we can inject new customer attributes and predict whether the customer is going to be a good credit risky customer or not.

To conclude, by extensive exploratory data analysis and linear model outputs we were able to find some interesting variables about the bank credits and also created a model to predict the risk factor of the future customers.

References

https://www.rdocumentation.org/packages/Amelia/versions/1.8.0/topics/missmap (https://www.rdocumentation.org/packages/Amelia/versions/1.8.0/topics/missmap)
https://njtierney.github.io/r/missing%20data/rbloggers/2015/12/01/ggplot-missing-data/ (https://njtierney.github.io/r/missing%20data/rbloggers/2015/12/01/ggplot-missing-data/)
[https://www.geeksforgeeks.org/convert-factor-to-numeric-and-numeric-to-factor-in-r-programming/] (https://www.geeksforgeeks.org/convert-factor-to-numeric-and-numeric-to-factor-in-r-programming/DAB501)
DAB501 (https://www.geeksforgeeks.org/convert-factor-to-numeric-and-numeric-to-factor-in-r-programming/) study materials and LAB contents