STAT 412 STATISTICAL DATA ANALYSIS

2023-2024 ACADEMIC YEAR SPRING SEMESTER

INTERIM REPORT

Eren Duralı 2361244

Introduction

This study, Includes Credit score classification data taken from Kaggle (https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data)

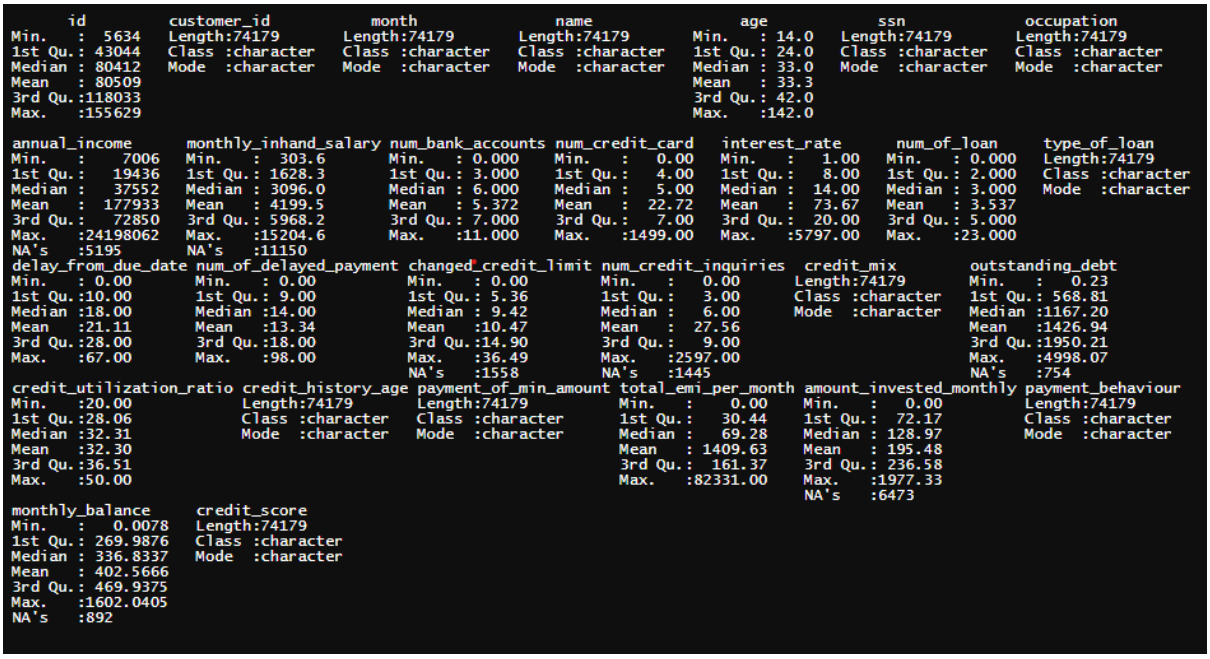
Data includes 28 variables and 100000 observations.

Aim

Our target in this study, classify the credit score correctly by using some statistical methods, and models with machine learning algorithms. Trying to find the best classification model.

Data Cleaning

In the data cleaning part, we deleted some columns that are not our concern for analysis and model building. These columns are Id, Customer_id, Month, Name, SSN, and Occupation.



```
       id           customer_id           month               name                age               ssn              occupation
 Min.   :  5634   Length:74179       Length:74179       Length:74179       Min.   : 14.0   Length:74179       Length:74179
 1st Qu.: 43044   Class :character   Class :character   Class :character   1st Qu.: 24.0   Class :character   Class :character
 Median : 80412   Mode  :character   Mode  :character   Mode  :character   Median : 33.0   Mode  :character   Mode  :character
 Mean   : 80509                                                            Mean   : 33.3
 3rd Qu.:118033                                                            3rd Qu.: 42.0
 Max.   :155629                                                            Max.   :142.0

 annual_income       monthly_inhand_salary num_bank_accounts num_credit_card    interest_rate       num_of_loan      type_of_loan
 Min.   :     7006   Min.   :  303.6       Min.   : 0.000    Min.   :   0.00   Min.   :   1.00   Min.   : 0.000   Length:74179
 1st Qu.:    19436   1st Qu.: 1628.3       1st Qu.: 3.000    1st Qu.:   4.00   1st Qu.:   8.00   1st Qu.: 2.000   Class :character
 Median :    37552   Median : 3096.0       Median : 6.000    Median :   5.00   Median :  14.00   Median : 3.000   Mode  :character
 Mean   :   177933   Mean   : 4199.5       Mean   : 5.372    Mean   :  22.72   Mean   :  73.67   Mean   : 3.537
 3rd Qu.:    72850   3rd Qu.: 5968.2       3rd Qu.: 7.000    3rd Qu.:   7.00   3rd Qu.:  20.00   3rd Qu.: 5.000
 Max.   :24198062   Max.   :15204.6       Max.   :11.000    Max.   :1499.00   Max.   :5797.00   Max.   :23.000
 NA's   :5195        NA's   :11150
 delay_from_due_date num_of_delayed_payment changed_credit_limit num_credit_inquiries  credit_mix         outstanding_debt
 Min.   : 0.00       Min.   : 0.00          Min.   : 0.00        Min.   :   0.00      Length:74179       Min.   :   0.23
 1st Qu.:10.00       1st Qu.: 9.00          1st Qu.: 5.36        1st Qu.:   3.00      Class :character   1st Qu.: 568.81
 Median :18.00       Median :14.00          Median : 9.42        Median :   6.00      Mode  :character   Median :1167.20
 Mean   :21.11       Mean   :13.34          Mean   :10.47        Mean   :  27.56                         Mean   :1426.94
 3rd Qu.:28.00       3rd Qu.:18.00          3rd Qu.:14.90        3rd Qu.:   9.00                         3rd Qu.:1950.21
 Max.   :67.00       Max.   :98.00          Max.   :36.49        Max.   :2597.00                         Max.   :4998.07
                                            NA's   :1558         NA's   :1445                            NA's   :754
 credit_utilization_ratio credit_history_age payment_of_min_amount total_emi_per_month amount_invested_monthly payment_behaviour
 Min.   :20.00            Length:74179       Length:74179           Min.   :    0.00   Min.   :   0.00         Length:74179
 1st Qu.:28.06            Class :character   Class :character       1st Qu.:   30.44   1st Qu.:  72.17         Class :character
 Median :32.31            Mode  :character   Mode  :character       Median :   69.28   Median : 128.97         Mode  :character
 Mean   :32.30                                                      Mean   : 1409.63   Mean   : 195.48
 3rd Qu.:36.51                                                      3rd Qu.:  161.37   3rd Qu.: 236.58
 Max.   :50.00                                                      Max.   :82331.00   Max.   :1977.33
                                                                                       NA's   :6473
 monthly_balance     credit_score
 Min.   :   0.0078   Length:74179
 1st Qu.: 269.9876   Class :character
 Median : 336.8337   Mode  :character
 Mean   : 402.5666
 3rd Qu.: 469.9375
 Max.   :1602.0405
 NA's   :892
```

After that, we look at the summary statistics for detecting unusual moments in the variables.

Figure 1: Summary statistics

We detect some of them and deal with filtering the data. After this process, we were left with 74179 rows. After that, we look at the boxplots for detecting the outliers and distribution of the variables in our data.

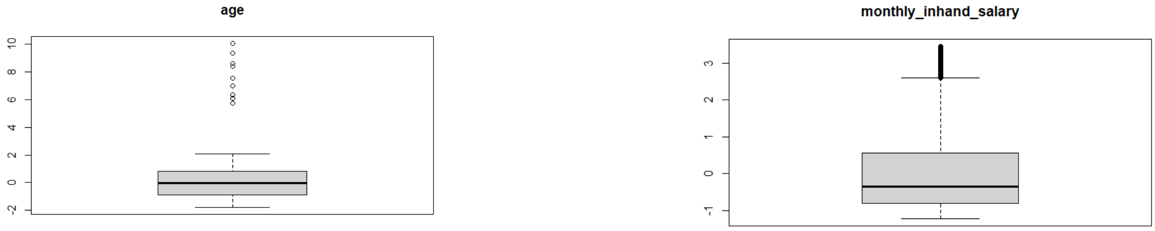


Figure 2: Some examples of boxplots

For the categorical variables, we use bar plots to obtain their distributions.

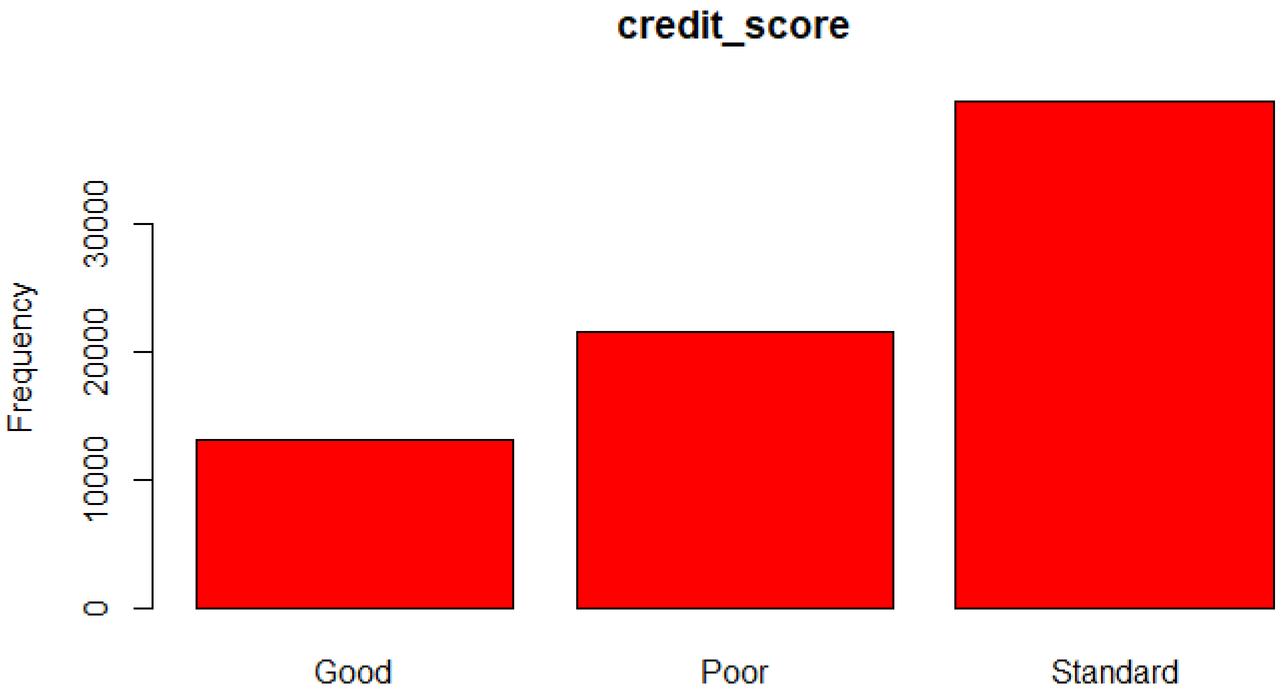
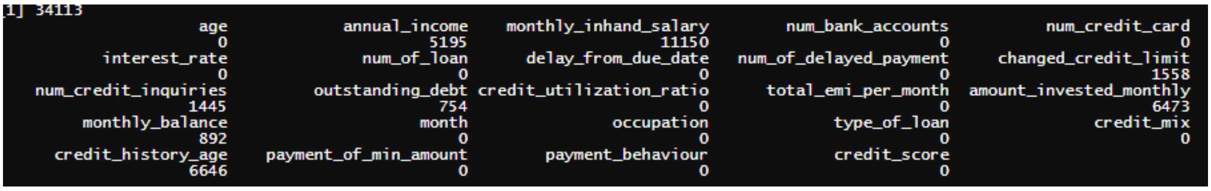


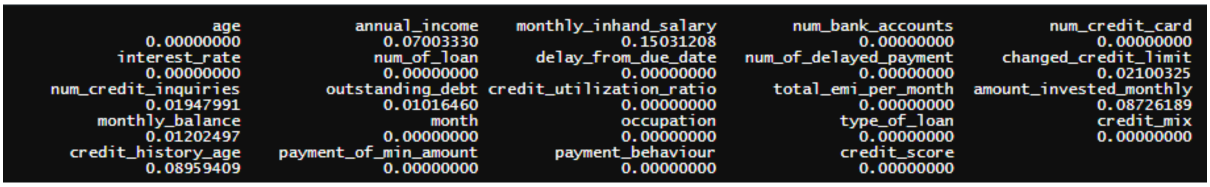
Figure 3: Credit score bar plot

After these processes, we deal with missing values. Firstly, detect the missing values and we have 34113 missing values in our dataset. Monthly in hand salary column has 11150 missing values so we consider that column might be removed from data.



```
1] 34113
                     age            annual_income    monthly_inhand_salary        num_bank_accounts          num_credit_card
                       0                     5195                    11150                        0                        0
           interest_rate              num_of_loan      delay_from_due_date   num_of_delayed_payment     changed_credit_limit
                       0                        0                        0                        0                     1558
    num_credit_inquiries         outstanding_debt credit_utilization_ratio      total_emi_per_month  amount_invested_monthly
                    1445                      754                        0                        0                     6473
         monthly_balance                    month               occupation             type_of_loan               credit_mix
                     892                        0                        0                        0                        0
      credit_history_age    payment_of_min_amount        payment_behaviour             credit_score
                    6646                        0                        0                        0
```

Figure 4: Total Missing Values



```
                     age            annual_income    monthly_inhand_salary        num_bank_accounts          num_credit_card
              0.00000000               0.07003330               0.15031208               0.00000000               0.00000000
           interest_rate              num_of_loan      delay_from_due_date   num_of_delayed_payment     changed_credit_limit
              0.00000000               0.00000000               0.00000000               0.00000000               0.02100325
    num_credit_inquiries         outstanding_debt credit_utilization_ratio      total_emi_per_month  amount_invested_monthly
              0.01947991               0.01016460               0.00000000               0.00000000               0.08726189
         monthly_balance                    month               occupation             type_of_loan               credit_mix
              0.01202497               0.00000000               0.00000000               0.00000000               0.00000000
      credit_history_age    payment_of_min_amount        payment_behaviour             credit_score
              0.08959409               0.00000000               0.00000000               0.00000000
```

Figure 5: Missing Values Percentage

Since, 60% our thresh hold for deleting the column we keep the monthly in hand salary. We use na.approx function from zoo package in R to deal with missing values in numerical columns. This package fills the na values by interpolated values. After dealing with numerical columns, we continue to deal with missing values in categorical columns. We have 6647 missing values in categorical columns and all of them in the credit history age column (Represents the age of credit history of the person). Since this column consist of years and months. We change this column to months and apply na.approx function. When missing values are gone, we try to normalize our data because most of the statistical method assumptions are normality. We use bestNormalize package for that.

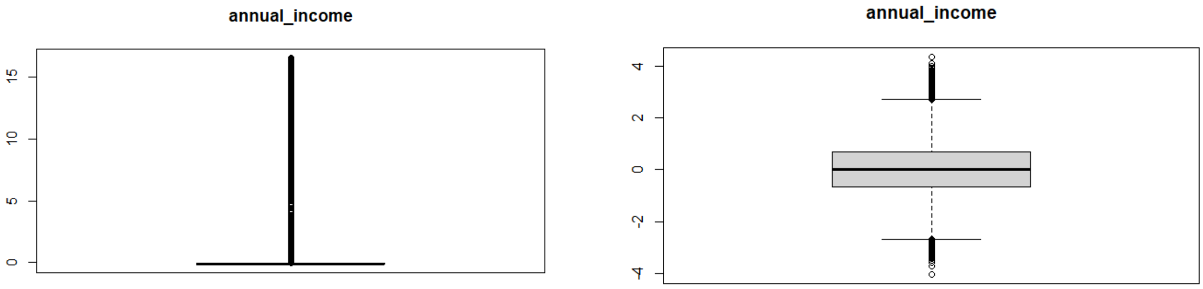


Figure 6: Shows the boxplot before & after normalization

Research Questions

Question1) How does annual income influence the number of bank accounts and credit cards an individual holds?

We want to investigate is annual income has an influence on the number of bank accounts and credit cards and if it has an influence how is it?

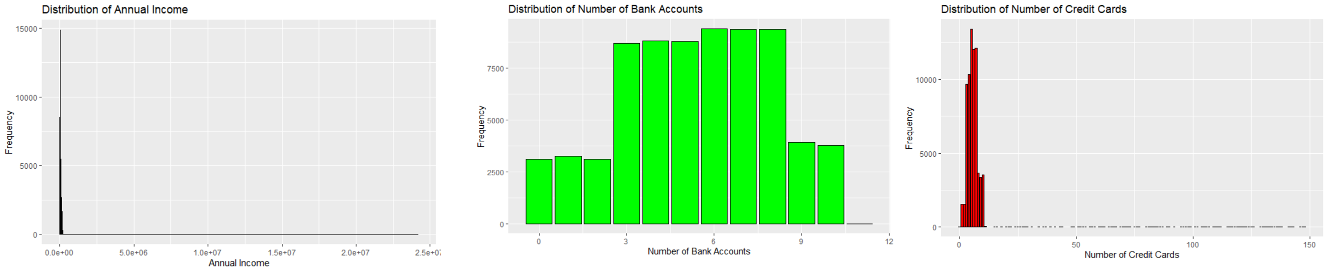


Figure 7: Distribution of Variables

Annual income and number of credit cards has a left skewed distribution which means that people with lower annual income holds most of the population.

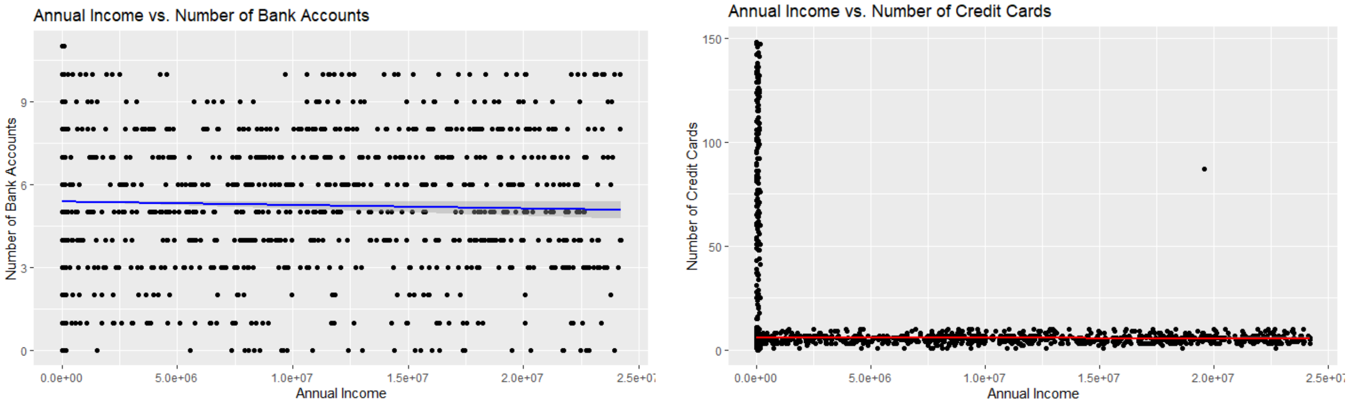
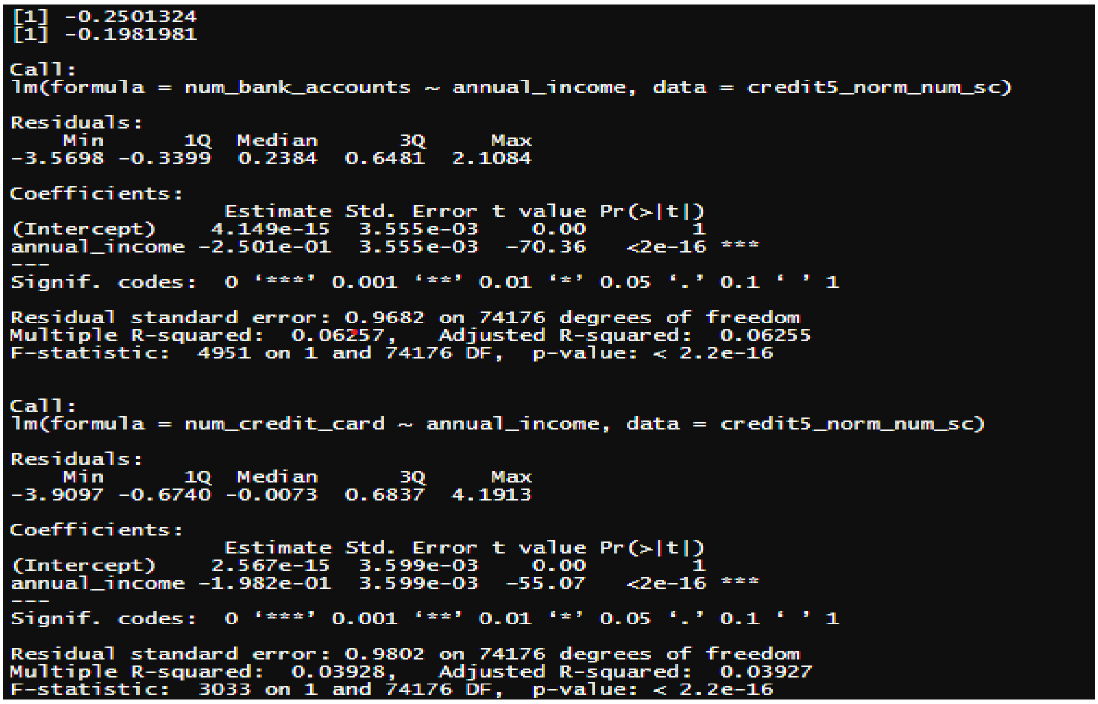


Figure 8: Scatter plot between independent and dependent variables

This scatter plot shows that we don’t have a strong linear relationship between Annual income between number of bank accounts and credit cards. Since it is reasonable because of the number of bank accounts and credit cards values are just countable numbers.



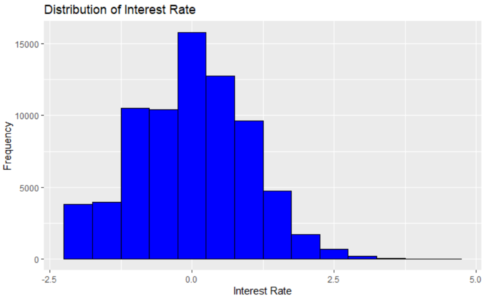
```
[1] -0.2501324
[1] -0.1981981

Call:
lm(formula = num_bank_accounts ~ annual_income, data = credit5_norm_num_sc)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5698 -0.3399  0.2384  0.6481  2.1084 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    4.149e-15  3.555e-03    0.00        1    
annual_income -2.501e-01  3.555e-03  -70.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9682 on 74176 degrees of freedom
Multiple R-squared:  0.06257,	Adjusted R-squared:  0.06255 
F-statistic:  4951 on 1 and 74176 DF,  p-value: < 2.2e-16


Call:
lm(formula = num_credit_card ~ annual_income, data = credit5_norm_num_sc)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9097 -0.6740 -0.0073  0.6837  4.1913 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.567e-15  3.599e-03    0.00        1    
annual_income -1.982e-01  3.599e-03  -55.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9802 on 74176 degrees of freedom
Multiple R-squared:  0.03928,	Adjusted R-squared:  0.03927 
F-statistic:  3033 on 1 and 74176 DF,  p-value: < 2.2e-16
```

Figure 9: Correlation values and Linear regression models

Annual income has a 25% negative correlation between number of bank accounts and 19% negative correlation between number of credit cards. In the linear regression models, Annual income was significant for both models but R-squared values were low. This shows that Annual income explains the variance of the number of bank accounts at 6.2% and the variance of the number of credit cards 3.9%. Residuals were looks good. And it has a negative coefficient as expected.

Q2) What is the impact of interest rate and outstanding debt on the credit score of individuals?

We want to investigate the impact of interest rate and outstanding debt on the our target variable credit score.



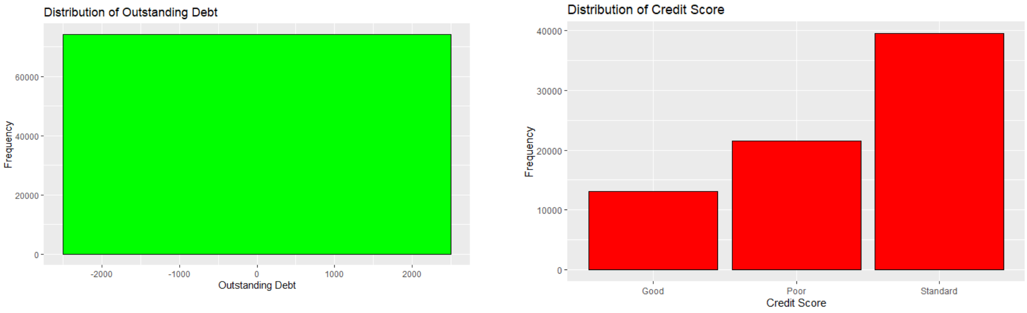
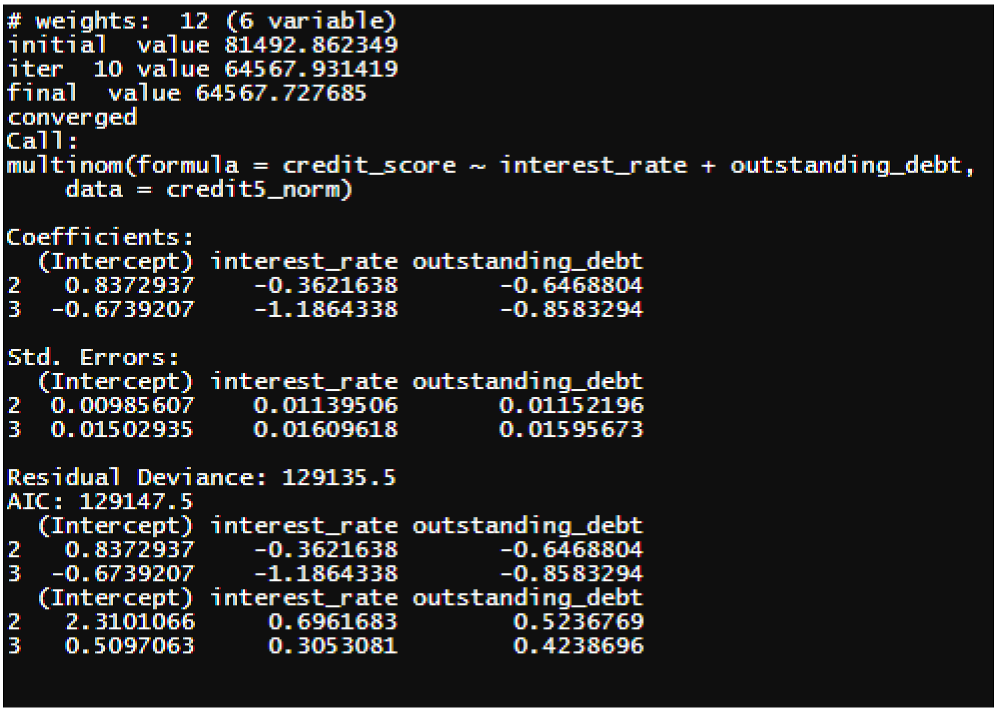


Figure 10: Distributions of Variables

Since our credit score is a categorical variable, we apply a multinominal logistic regression using the “multinom” function in the “nnet” package. Before that we encode the credit score by label encoding.



```
# weights:  12 (6 variable)
initial  value 81492.862349
iter  10 value 64567.931419
final  value 64567.727685
converged
Call:
multinom(formula = credit_score ~ interest_rate + outstanding_debt,
    data = credit5_norm)

Coefficients:
  (Intercept) interest_rate outstanding_debt
2   0.8372937    -0.3621638       -0.6468804
3  -0.6739207    -1.1864338       -0.8583294

Std. Errors:
  (Intercept) interest_rate outstanding_debt
2  0.00985607    0.01139506       0.01152196
3  0.01502935    0.01609618       0.01595673

Residual Deviance: 129135.5
AIC: 129147.5
  (Intercept) interest_rate outstanding_debt
2   0.8372937    -0.3621638       -0.6468804
3  -0.6739207    -1.1864338       -0.8583294
  (Intercept) interest_rate outstanding_debt
2   2.3101066     0.6961683        0.5236769
3   0.5097063     0.3053081        0.4238696
```

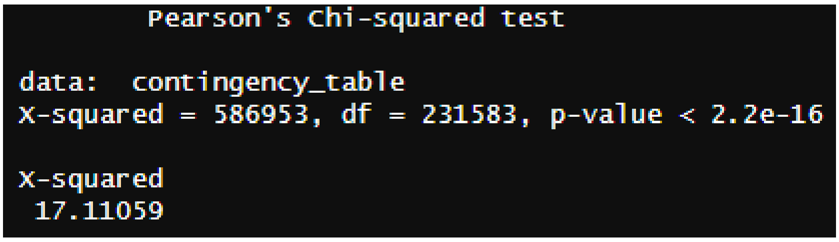
Figure 11: Logistic regression model

In this logistic regression model (Intercept) represents “Poor”, 2 represents “Standard” and 3 represents “Good” credit score. Interest rate and outstanding debt is significant since their p-value is < 0.05 (it showed in the code). For every one unit increase in interest rate; The odds of having a "Standard" credit score versus a "Poor" credit score decrease by approximately 30.38%.The odds of having a "Good" credit score versus a "Poor" credit score decrease by approximately 69.47%.For every one-unit increase in outstanding debt; The odds of having a

"Standard" credit score versus a "Poor" credit score decrease by approximately 47.63%.The odds of having a "Good" credit score versus a "Poor" credit score decrease by approximately 57.61%.

Question 3) Is there a significant association between the type of loan (e.g., mortgage, car loan, personal loan) and the frequency of late payments?

We try to find an association between the type of loan and the frequency of late payments. We conduct a Pearson's Chi-squared test.



```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 586953, df = 231583, p-value < 2.2e-16

X-squared
 17.11059
```

Figure 12: Chi-squared test

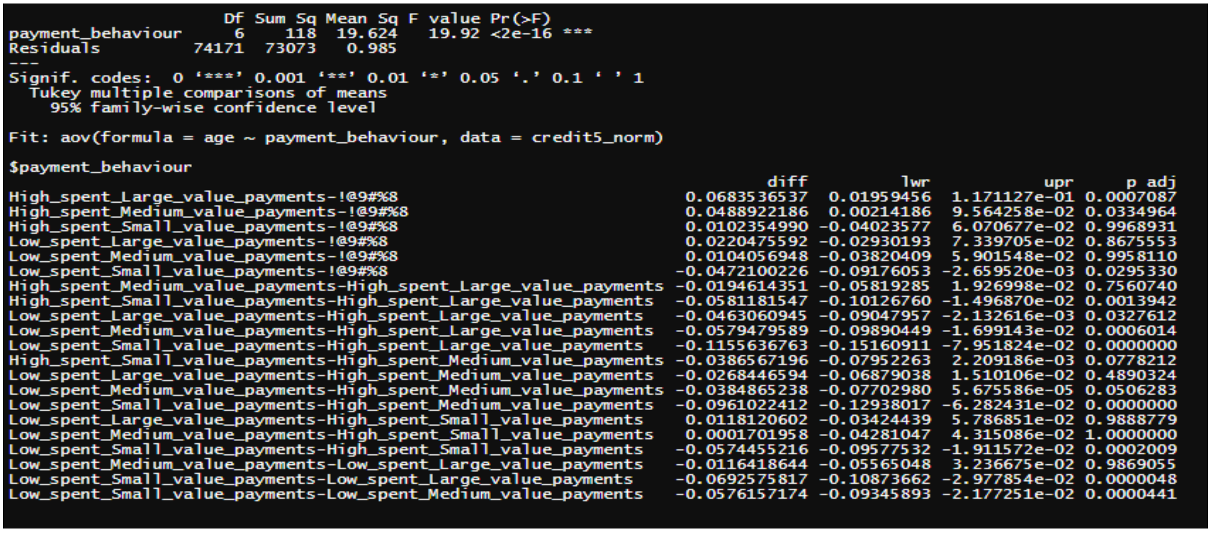
Since Cramer's V = 17.11 type of loan and the frequency of delayed payments has a strong association. This shows that type of loan has a significant impact on the frequency of delayed payments.

Question 4) Does the age of individuals vary significantly across different levels of payment behavior?

We conduct ANOVA to analyze this.

H0: The mean age is the same across all levels of payment behavior.

H1: At least one level of payment behavior has a different mean age compared to others.



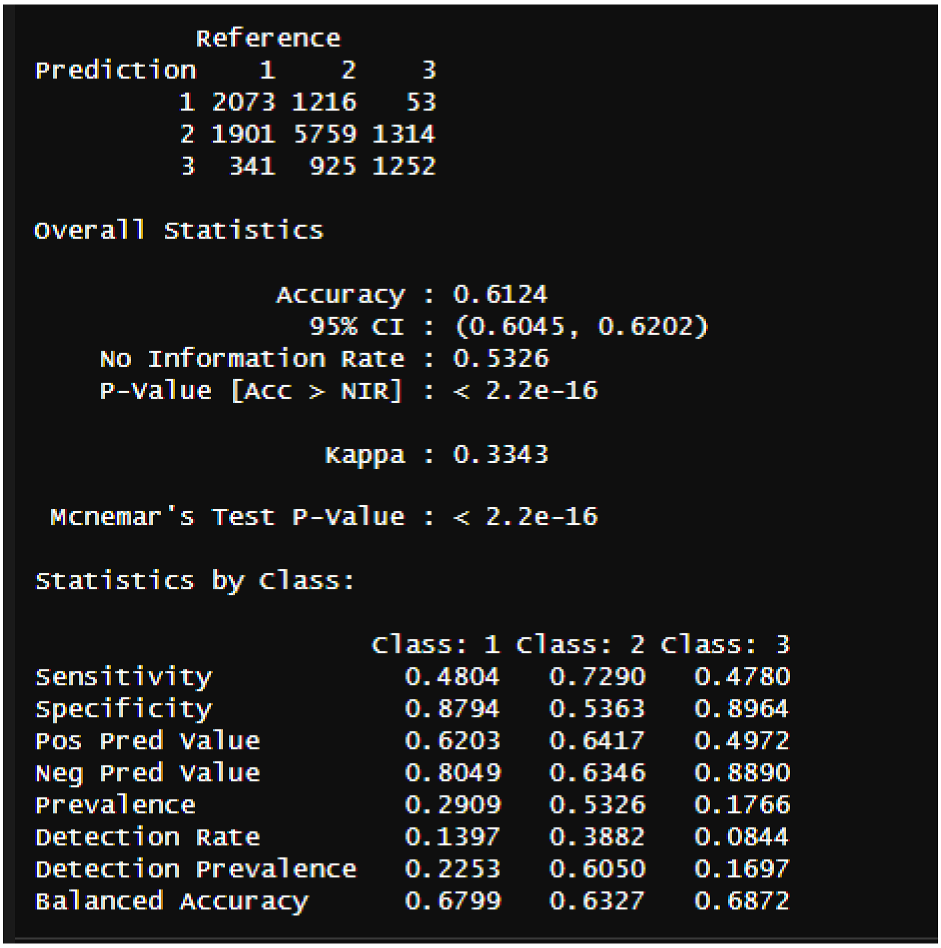
```
                     Df Sum Sq Mean Sq F value Pr(>F)    
payment_behaviour     6    118  19.624   19.92 <2e-16 ***
Residuals         74171  73073   0.985                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = age ~ payment_behaviour, data = credit5_norm)

$payment_behaviour
                                                                  diff         lwr           upr     p adj
High_spent_Large_value_payments-!@9#%8                    0.0683536537  0.01959456  1.171127e-01 0.0007087
High_spent_Medium_value_payments-!@9#%8                   0.0488922186  0.00214186  9.564258e-02 0.0334964
High_spent_Small_value_payments-!@9#%8                    0.0102354990 -0.04023577  6.070677e-02 0.9968931
Low_spent_Large_value_payments-!@9#%8                     0.0220475592 -0.02930193  7.339705e-02 0.8675553
Low_spent_Medium_value_payments-!@9#%8                    0.0104056948 -0.03820409  5.901548e-02 0.9958110
Low_spent_Small_value_payments-!@9#%8                    -0.0472100226 -0.09176053 -2.659520e-03 0.0295330
High_spent_Medium_value_payments-High_spent_Large_value_payments -0.0194614351 -0.05819285  1.926998e-02 0.7560740
High_spent_Small_value_payments-High_spent_Large_value_payments  -0.0581181547 -0.10126760 -1.496870e-02 0.0013942
Low_spent_Large_value_payments-High_spent_Large_value_payments   -0.0463060945 -0.09047957 -2.132616e-03 0.0327612
Low_spent_Medium_value_payments-High_spent_Large_value_payments  -0.0579479589 -0.09890449 -1.699143e-02 0.0006014
Low_spent_Small_value_payments-High_spent_Large_value_payments   -0.1155636763 -0.15160911 -7.951824e-02 0.0000000
High_spent_Small_value_payments-High_spent_Medium_value_payments -0.0386567196 -0.07952263  2.209186e-03 0.0778212
Low_spent_Large_value_payments-High_spent_Medium_value_payments  -0.0268446594 -0.06879038  1.510106e-02 0.4890324
Low_spent_Medium_value_payments-High_spent_Medium_value_payments -0.0384865238 -0.07702980  5.675586e-05 0.0506283
Low_spent_Small_value_payments-High_spent_Medium_value_payments  -0.0961022412 -0.12938017 -6.282431e-02 0.0000000
Low_spent_Large_value_payments-High_spent_Small_value_payments    0.0118120602 -0.03424439  5.786851e-02 0.9888779
Low_spent_Medium_value_payments-High_spent_Small_value_payments   0.0001701958 -0.04281047  4.315086e-02 1.0000000
Low_spent_Small_value_payments-High_spent_Small_value_payments   -0.0574455216 -0.09577532 -1.911572e-02 0.0002009
Low_spent_Medium_value_payments-Low_spent_Large_value_payments   -0.0116418644 -0.05565048  3.236675e-02 0.9869055
Low_spent_Small_value_payments-Low_spent_Large_value_payments    -0.0692575817 -0.10873662 -2.977854e-02 0.0000048
Low_spent_Small_value_payments-Low_spent_Medium_value_payments   -0.0576157174 -0.09345893 -2.177251e-02 0.0000441
```

Figure 13: ANOVA between age and payment behavior

Since P score <2e-16 lower than 0.05 payment behavior is associated with variations on age. Individuals with "High_spent_Large_value_payments" tend to be older compared to those with "High_spent_Medium_value_payments" and "High_spent_Small_value_payments" ($p < 0.05$).There are significant age differences between various combinations of payment behavior categories, such as "Low_spent_Small_value_payments" versus "High_spent_Small_value_payments" ($p < 0.05$).

Question 5) How accurate a multinominal regression model can classify credit score correctly?

Since credit score data aim to build a model which classifies credit scores. We investigate that question. Firstly, we encode all categorical variables for putting in the model. Secondly, we split the data as a train and test data. Thirdly, we use 5-fold cross validation. Lastly, We build our model by using preprocess(“center”, ”scale”, ”pca”) in the multinom function.



```
          Reference
Prediction    1    2    3
         1 2073 1216   53
         2 1901 5759 1314
         3  341  925 1252

Overall Statistics
                                          
               Accuracy : 0.6124          
                 95% CI : (0.6045, 0.6202)
    No Information Rate : 0.5326          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.3343          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.4804   0.7290   0.4780
Specificity            0.8794   0.5363   0.8964
Pos Pred Value         0.6203   0.6417   0.4972
Neg Pred Value         0.8049   0.6346   0.8890
Prevalence             0.2909   0.5326   0.1766
Detection Rate         0.1397   0.3882   0.0844
Detection Prevalence   0.2253   0.6050   0.1697
Balanced Accuracy      0.6799   0.6327   0.6872
```

Figure 14: Prediction of model in the test set and confusion matrix

The confusion matrix shows the counts of correct and incorrect predictions for each class. It looks good but it can be better also the model accuracy is 61.24%. Model accuracy achieved by always predicting the most frequent class around 53.26%. A measure of agreement between actual and predicted classes, adjusted for chance agreement. Here, it's around 33.43%, this means fair agreement. Sensitivity is higher for class 2 as 72.9% but sensitivity in the other class below 50% so we should be deal with that in the following process.