MIDDLE EAST TECHNICAL UNIVERSITY

Department of Statistics

STAT 250 Applied Statistics Term Project

İrem Ekmekçi

1. Abstract

For a loan officer, it is important to determine the best loan program for a customer. It is the key to providing customers great service when a loan officer knows details of applicant income, loan amount etc. of the customers. This study analyzes how loan amount of a customer can be evaluated by various measures. Results and findings part includes the following analysis:

Frequencies of Property Areas

- The Average Coapplicant Income Analysis
- The Average Loan Amount Analysis
- The Applicant Income Analysis For Education Status
- Multiple Linear Regression

2. Introduction

This dataset contains the details of the customers in a bank. Data contains variables such as loan id, gender, marital status, dependents, education status, employment status, applicant income, coapplicant income, loan amount, loan amount term, credit history, and property areas.

- Loan_ID :Unique Loan ID
- Gender :Male/ Female
- Married :Applicant married (Y/N)
- Dependents :Number of dependents
- Education :Applicant Education (Graduate/ Under Graduate)
- Self_Employed :Self employed (Y/N)
- ApplicantIncome :Applicant income
- CoapplicantIncome :Coapplicant income
- LoanAmount : Loan amount in thousands
- Loan_Amount_Term : Term of loan in months
- Credit_History : history meets guidelines
- Property_Area : Urban/ Semi Urban/ Rural

3. Methodology

Bar Plot

Bar plot display frequencies of a categorical variable through vertical or horizontal bars.

Box Plot

A box plot describes the distribution of a continuous variable by plotting its quartiles. It can also display that whether there are outliers or not.

Shapiro-Wilk Test

Shapiro-Wilk test is used to see whether a random sample comes from a normal distribution or not. This method is used before conducting any test test that requires normality assumption.

One Sample T-Test

A one-sample t-test is used to determine whether a sample comes from a population with a specific mean. It is used when sample mean and variance is unknown.

Two Sample T-Test

A two-sample t-test is used when you want to compare two independent groups to see if their means are different. There are two options for estimating the variances for the two sample t-test with independent samples; using pooled variances or using separate variances.

One-Way ANOVA

One-Way ANOVA is used to compare the dependent variable means of two or more groups defined by a categorical grouping factor. The null hypothesis is that all the population group means are equal versus the alternative that at least one of the population means is different.

Bartlett's Test

Bartlett's Test is used for checking homogeneity of variances of samples when there are two or more samples. This method is used to check one of the assumptions of the Two Sample T-Test and One-Way ANOVA.

Sign Test

The sign test is used to compare the sizes of two groups. It is a "distribution free" test, which means the test doesn't assume the data comes from a particular distribution. The sign test is an alternative nonparametric version of a one sample t test or a paired t test.

Mann-Whitney-Wilcoxon Test

Mann-Whitney-Wilcoxon Test is considered as the nonparametric version of two sample t-test. It is used to compare the two population distributions. Also, it does not require to be normally distributed.

Kruskal-Wallis Test

The Kruskal-Wallis test is considered the nonparametric version of the One Way ANOVA. It is used to determine if there are significant differences between two or more groups of an independent variable.

Correlation Matrix

The correlation matrix is used to see the relationship between more than two continuous variables.

Scatter Plot

Scatter plot is used to see the relationship between two continuous variables.

4. Results and Findings

Summary of Data

```
> summary(Bank)
     Loan_ID       Gender    Married   Dependents        Education   Self_Employed ApplicantIncome CoapplicantIncome
 LP001015:  1         : 11   No :134     : 10     Graduate    :283      : 23       Min.   :    0   Min.   :    0
 LP001022:  1   Female: 70   Yes:233   0 :200     Not Graduate: 84   No :307       1st Qu.: 2864   1st Qu.:    0
 LP001031:  1   Male  :286             1 : 58                        Yes: 37       Median : 3786   Median : 1025
 LP001035:  1                          2 : 59                                      Mean   : 4806   Mean   : 1570
 LP001051:  1                          3+: 40                                      3rd Qu.: 5060   3rd Qu.: 2430
 LP001054:  1                                                                      Max.   :72529   Max.   :24000
 (Other) :361
   LoanAmount    Loan_Amount_Term Credit_History     Property_Area
 Min.   : 28.0   Min.   :  6.0    Min.   :0.0000   Rural    :111
 1st Qu.:100.2   1st Qu.:360.0    1st Qu.:1.0000   Semiurban:116
 Median :125.0   Median :360.0    Median :1.0000   Urban    :140
 Mean   :136.1   Mean   :342.5    Mean   :0.8254
 3rd Qu.:158.0   3rd Qu.:360.0    3rd Qu.:1.0000
 Max.   :550.0   Max.   :480.0    Max.   :1.0000
 NA's   :5       NA's   :6        NA's   :29
```

The average applicant income is 4806. The minimum applicant income is 0, while the maximum is 72529. Half of the applicant income is below or above 3786. 25% of the applicant income is below 2864 and above 5060.

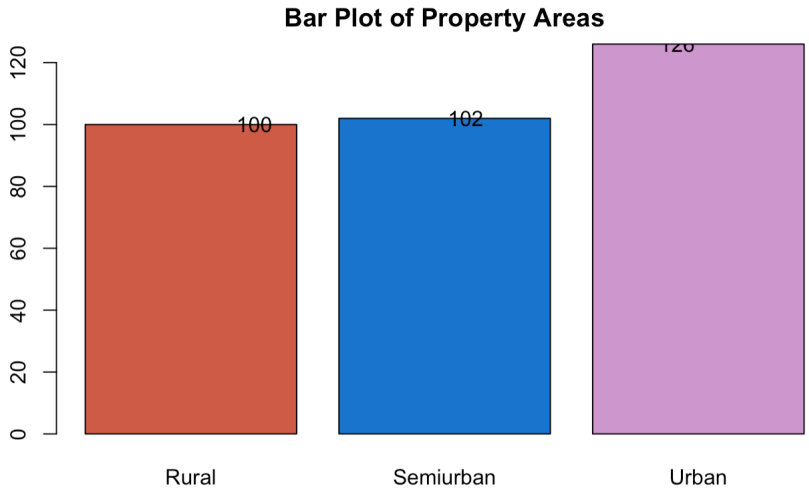
The average coapplicant income is 1570. The minimum coapplicant income is 0, while the maximum is 24000. Half of the coapplicant income is below or above 1025. 25% of the coapplicant income is below 0 and above 2430.

The average loan amount is 136.1. The minimum loan amount is 28, while the maximum is 550. Half of the loan amount is below or above 125. 25% of the loan amount is below 100.2 and above 158.

The average loan amount term is 342.5. The minimum loan amount term is 6, while the maximum is 480. Half of the loan amount term is below or above 360. 25% of the loan amount term is below and above 360.

The variables which have NA terms are LoanAmount, Loan_Amount_Term, and Credit_History. These NA terms are omitted before calculations.

Research Question: *" Which property area has the highest frequency?"*



Bar plot is used when looking at the frequencies of the categorical variables. It can be said that Urban area has the highest frequency in all property areas from the bar plot of property areas.

Research Question: *"Is the average coapplicant income greater than 1500 or not?"*

H0: The mean value of coapplicant income is less than or equal to 1500.

H1: The mean value of coapplicant income is greater than 1500.

Method 1: *One Sample t-test*

Assumptions

- Data points should be independent from each other.
- Data should be normally distributed.
- Data should be randomly selected from a population, where each item has an equal chance of being selected.

```
         One Sample t-test

data:  Bank$CoapplicantIncome
t = 0.068122, df = 327, p-value = 0.4729
alternative hypothesis: true mean is greater than 1500
95 percent confidence interval:
 1299.566      Inf
sample estimates:
mean of x
 1508.634
```

According to One Sample t-test, it can be said that we do not have enough evidence to reject the null hypothesis since the p-value is greater than 0.05. This means that the mean value of coapplicant income is not greater than 1500 at the significance level 0.05.

However, the normality of the variable coapplicant income would be checked.

```
        Shapiro-Wilk normality test

data:  Bank$CoapplicantIncome
W = 0.61876, p-value < 2.2e-16
```

However, the variable coapplicant income is not normally distributed. Hence, One Sample t-test should not be conducted since the normality assumption failed. It is better to use nonparametric approach.

***Method 2**: Sign Test*

```
        One-sample Sign-Test

data:  Bank$CoapplicantIncome
s = 139, p-value = 0.9967
alternative hypothesis: true median is greater than 1500
95 percent confidence interval:
 394.3445      Inf
sample estimates:
median of x
        856

Achieved and Interpolated Confidence Intervals:

                  Conf.Level   L.E.pt U.E.pt
Lower Achieved CI     0.9454 437.0000    Inf
Interpolated CI       0.9500 394.3445    Inf
Upper Achieved CI     0.9566 333.0000    Inf
```

The sign test is the nonparametric version of one sample t-test. Moreover, it does not have the normality assumption.

According to One-sample Sign-Test, it can be said that we do not have enough evidence to reject the null hypothesis since the p-value is greater than 0.05. This means that the mean value of coapplicant income is not greater than 1500 at the significance level 0.05.

Research Question: *"Is the average loan amount equal to 125$?"*

H0: The mean value of loan amount equals to 125$.

H1: The mean value of loan amount does not equal to 125$.

Method 1: *One Sample t-test*

```
        One Sample t-test

data:  Bank$LoanAmount
t = 3.5897, df = 327, p-value = 0.0003818
alternative hypothesis: true mean is not equal to 125
95 percent confidence interval:
 130.2583 143.0100
sample estimates:
mean of x
 136.6341
```

Since the p-value is less than 0.05, we can reject the null hypothesis with 95% confidence. This means that the mean value of the loan amount does not equal 125.

However, the normality of the variable loan amount would be checked.

```
        Shapiro-Wilk normality test

data:  Bank$LoanAmount
W = 0.88253, p-value = 3.701e-15
```

However, the variable loan amount is not normally distributed. Hence, One Sample t-test should not be conducted since the normality assumption failed. It is better to use the nonparametric approach since it does not require normality.

Method 2: *Sign Test*

```
        One-sample Sign-Test

data:  Bank$LoanAmount
s = 163, p-value = 0.6533
alternative hypothesis: true median is not equal to 125
95 percent confidence interval:
 122 131
sample estimates:
median of x
        125

Achieved and Interpolated Confidence Intervals:

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9469    122    131
Interpolated CI       0.9500    122    131
Upper Achieved CI     0.9591    122    131
```

According to One-sample Sign-Test, it can be said that we do not have enough evidence to reject the null hypothesis since the p-value is greater than 0.05. This means that the mean value of loan amount is not equal to 125 at the significance level 0.05.

Research Question: *"Is the average applicant income significantly higher for applicants who are graduate than not graduate?*"

H0: μgrad.≤ μnotgrad.

H1: μgrad.> μnotgrad.

Method 1: *Two Sample t-test*

Assumptions

- Samples should be taken from Normal Distribution randomly with unknown variances.
- Two samples must be independent from each other.

Normality Assumption

```
        Shapiro-Wilk normality test

data:  Graduate
W = 0.41828, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  NotGraduate
W = 0.91205, p-value = 6.244e-05
```

Applicant income for both graduate and not graduate are not normally distributed.

Homogeneity of variances

```
> var.test(Graduate,NotGraduate)

        F test to compare two variances

data:  Graduate and NotGraduate
F = 10.28, num df = 251, denom df = 75, p-value <
2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  6.992886 14.567438
sample estimates:
ratio of variances
          10.27998
```

Since p-value is less than 0.05, it can be said that the variances of two samples are not equal to each other.

Hence, two sample t-test cannot be conducted because both of the assumptions did not meet. However, if it would be conducted it would be like this;

```
> t.test(Graduate,NotGraduate,alternative="greater",conf.level = 0.95)

        Welch Two Sample t-test

data:  Graduate and NotGraduate
t = 4.1736, df = 325.65, p-value = 1.926e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 989.4011      Inf
sample estimates:
mean of x mean of y
 5121.302  3485.303
```

According to Welch Two Sample t-test, we can reject the null hypothesis at significance 0.05 since p-value is less than 0.05. This means that the average applicant income is significantly higher for applicants who are graduate than not graduate. However, it is better to use the nonparametric approach as it does not require normality.

Method 2: *Mann-Whitney-Wilcoxon Test*

```
        Wilcoxon rank sum test with continuity correction

data:  Graduate and NotGraduate
W = 12527, p-value = 2.333e-05
alternative hypothesis: true location shift is greater than 0
```

According to Mann-Whitney-Wilcoxon test, it can be said that the average applicant income is significantly higher for applicants who are graduate than not graduate at significance level 0.05 since the p-value is less than 0.05.

Research Question: *"Is the average loan amount the same for all property areas?"*

H0:μ1=μ2=μ3

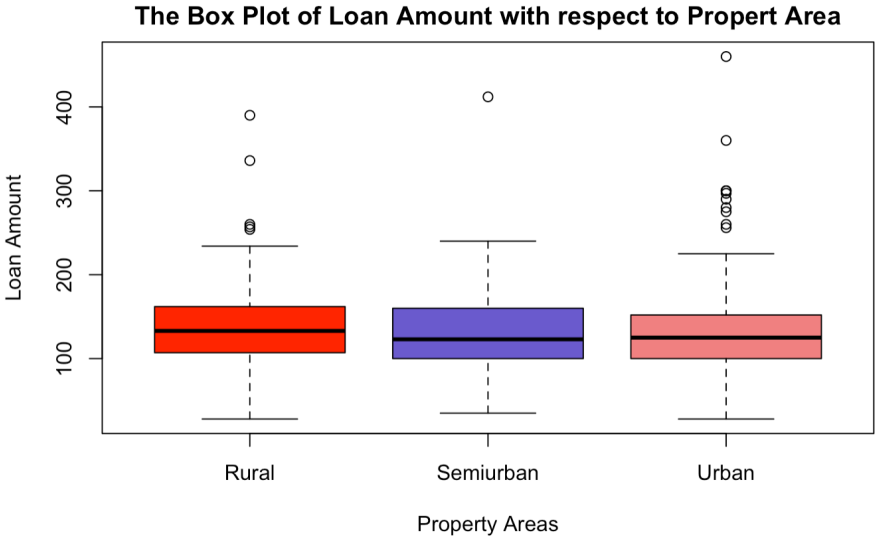
H1: At least one of them is different

Method 1: *One-Way ANOVA*

Assumptions

- Independent samples
- Normally distributed response
- Homogeneity of variances across responses for the groups

Let's first draw the samples with the appropriate plot.



It is seen that the average values of the loan amount of each property area represented by the median are almost same for each case. However, One-Way ANOVA is used to check if this result is true.

Let's start with checking assumption of One-Way ANOVA.

Normality assumption

Normality is checked for each sample with Shapiro-Wilk normality test.

```
        Shapiro-Wilk normality test

data:  Bank$LoanAmount[Bank$Property_Area == "Rural"]
W = 0.92243, p-value = 7.837e-06


        Shapiro-Wilk normality test

data:  Bank$LoanAmount[Bank$Property_Area == "Semiurban"]
W = 0.72784, p-value = 3.035e-13


        Shapiro-Wilk normality test

data:  Bank$LoanAmount[Bank$Property_Area == "Urban"]
W = 0.85994, p-value = 4.159e-10
```

According to the results of Shapiro-Wilk test above, none of the sample are normally distributed.

Homogeneity of variances

H0: All populations have equal variances.

H1: At least one of them has different variance.

```
        Bartlett test of homogeneity of variances

data:  LoanAmount by Property_Area
Bartlett's K-squared = 10.952, df = 2, p-value = 0.004186
```

We can reject the null hypothesis since p-value is less than 0.005. This means that we are 95% confident that at least one of the populations has different variance.

Since normality and homogeneity of variances assumption could not checked, One-Way ANOVA cannot be conducted.

However, if the assumptions were checked, it would be conducted like this;

```
> fit<-aov(LoanAmount~Property_Area,data=Bank)
> summary(fit)
               Df  Sum Sq Mean Sq F value Pr(>F)
Property_Area   2    6336    3168   0.919    0.4
Residuals     325 1120284    3447
```

The null hypothesis of equal means for all property ares cannot be rejected since Pr(>F)=0.4 >0.05. Therefore, there is no need to further analysis such as LSD test, Turkey test and pairwise test.

Method 2: *Kruskal-Wallis Test*

Since the populations are independent from each other, without assuming normality and equality of variances, the populations can be compared with Kruskal-Wallis Test.

Kruskal-Wallis Test is non-parametric version of the One-Way ANOVA.

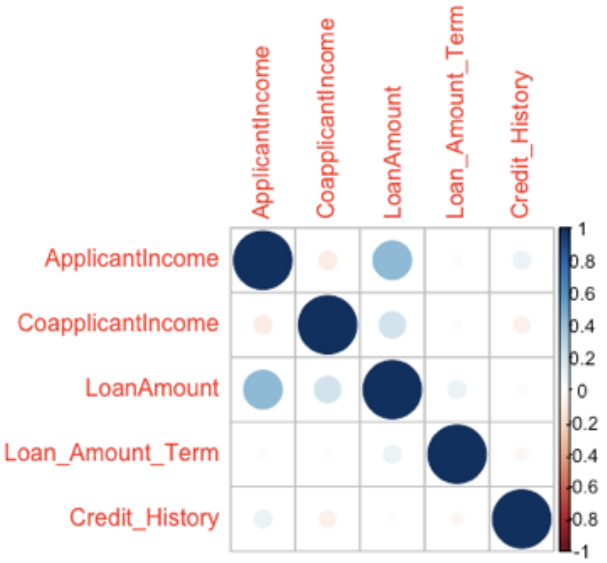
```
  Kruskal-Wallis Test (alpha = 0.05)
------------------------------------------------------------
  data : LoanAmount and Property_Area

  statistic  : 3.387656
  parameter  : 2
  p.value    : 0.1838145

  Result     : Difference is not statistically significant.
------------------------------------------------------------
```

According to Kruskal-Wallis Test result, the difference between the average loan amount for property areas is not statistically significant.

Research Question: *"Is there a linear relationship between loan amount and other variables?"*

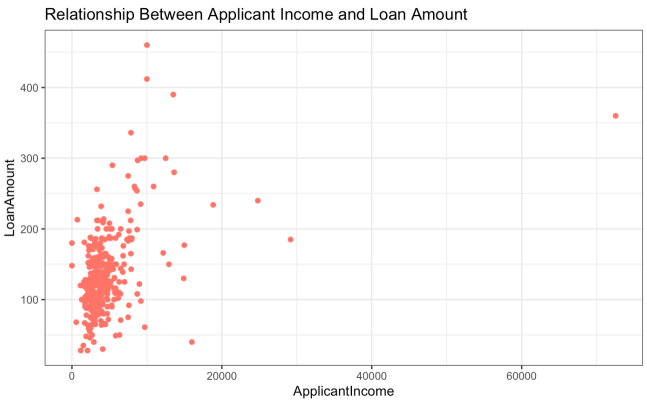
• Dependent Variable: LoanAmount

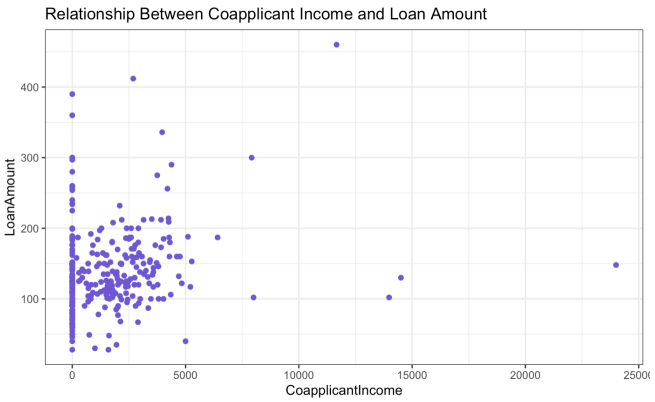
Let's look at the relationship between continuous variables.



The correlations are displayed in blue are positive and the correlations in red color are negative correlations. When the correlation coefficients are higher color and size of the circles intensify. The relationships between loan mount & applicant income and loan amount & coapplicant income are significant.

To see these relationships in more detail, scatter plots are used.





- The relationship between loan amount & applicant income is positive and linear.
- The relationship between loan amount & coapplicant income is positive and linear.

• Independent variables: Gender, Married, Education, Self_Employement, ApplicantIncome, CoapplicantIncome

Firstly, normality of response variable which is LoanAmount should be checked.

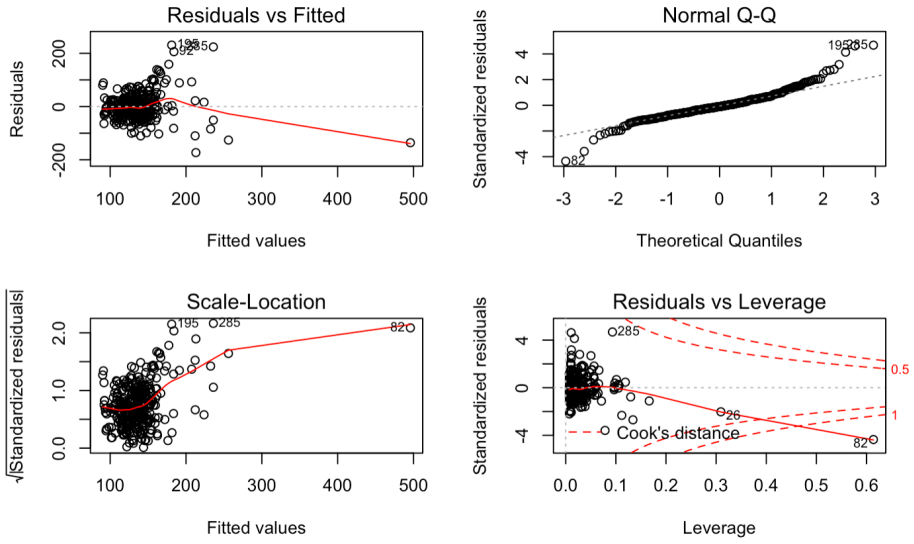
```
        Shapiro-Wilk normality test

data:  Bank$LoanAmount
W = 0.88253, p-value = 3.701e-15
```

According to Shapiro-Wilk test, LoanAmount is not normally distributed. We need a transformation. However, if we assume normality is checked, the model would be conducted like this;

```
Call:
lm(formula = LoanAmount ~ Gender + Married + Education + Self_Employed +
    ApplicantIncome + CoapplicantIncome, data = Bank)

Residuals:
    Min      1Q  Median      3Q     Max
-173.00  -28.12   -4.83   22.87  230.88

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.854e+01  1.963e+01   4.511 9.09e-06 ***
GenderFemale         1.349e+01  1.653e+01   0.816  0.41504
GenderMale           1.082e+01  1.577e+01   0.686  0.49327
MarriedYes           1.987e+01  6.076e+00   3.271  0.00119 **
EducationNot Graduate -1.459e+01  6.680e+00  -2.184  0.02969 *
Self_EmployedNo      -6.602e+00  1.120e+01  -0.590  0.55586
Self_EmployedYes     -5.047e+00  1.370e+01  -0.368  0.71282
ApplicantIncome       5.284e-03  5.832e-04   9.060  < 2e-16 ***
CoapplicantIncome     5.822e-03  1.224e-03   4.757 2.98e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.21 on 319 degrees of freedom
Multiple R-squared:  0.2863,    Adjusted R-squared:  0.2684
F-statistic: 15.99 on 8 and 319 DF,  p-value: < 2.2e-16
```

- p-value(< 2.2e-16) is less than 0.05 and this indicates the model is statistically significant.
- Adjusted R-squared being 0.2684: About 26% of the variation in LoanAmount are explained by Gender, Married, Education, Self_Employed, ApplicantIncome,and CoapplicantIncome. This Adjusted R-squared is quite low.
- Intercept, MarriedYes, EducationNot Graduate,Self_EmployedNo, ApplicantIncome and CoapplicantIncome are significant because their Pr(>|t|) values are less than 0.05.

Whether the residuals are normally distributed or not would be checked.



As a result, it is found that residuals are not normally distributed.

In order to make this model better insignificant variables are removed.

```
Call:
lm(formula = LoanAmount ~ Married + Education + ApplicantIncome +
    CoapplicantIncome, data = Bank)

Residuals:
     Min       1Q   Median       3Q      Max
-166.783  -28.338   -5.147   22.723  230.094

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.357e+01  5.897e+00  15.867  < 2e-16 ***
MarriedYes            2.009e+01  5.760e+00   3.488 0.000554 ***
EducationNot Graduate -1.488e+01  6.627e+00  -2.245 0.025472 *
ApplicantIncome        5.270e-03  5.781e-04   9.117  < 2e-16 ***
CoapplicantIncome      5.777e-03  1.215e-03   4.757 2.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.98 on 323 degrees of freedom
Multiple R-squared:  0.2839,    Adjusted R-squared:  0.2751
F-statistic: 32.02 on 4 and 323 DF,  p-value: < 2.2e-16
```

According to this output

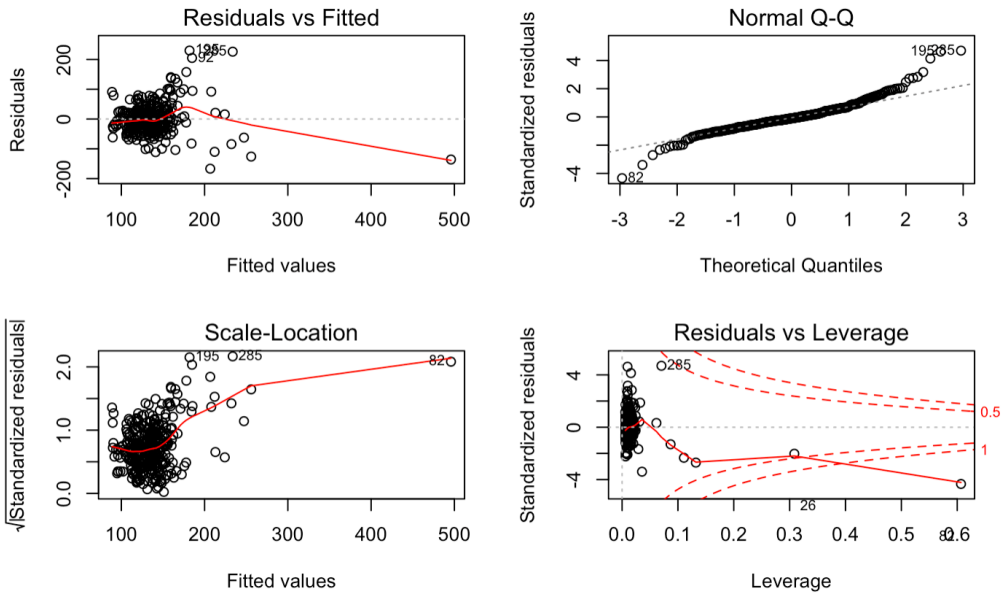
- p-value(< 2.2e-16) is less than 0.05 and this indicates the model is statistically significant.
- Adjusted R-squared being 0.2751: About 27% of the variation in LoanAmount are explained by Gender, Married, Education, Self_Employed, ApplicantIncome,and CoapplicantIncome. This Adjusted R-squared is quite low.
- All of the coefficients are significant because their Pr(>|t|) values are less than 0.05.
- Education coefficient has a negative effect on loan amount while other coefficients have positive effect on loan amount.

Firstly, let’s check whether there is a multicollinearity problem.

```
> vif(fit2)
           MarriedYes EducationNot Graduate       ApplicantIncome     CoapplicantIncome
               1.0073                1.0268                1.0327                1.0177
```

There is no multicollinearity problem since all of the variance inflation factor values are less than 5.

Secondly, let’s do diagnostic checks.



- Residual vs Fitted plot shows a funnel shaped pattern.
- Normality plot shows that some departure from normality.
- Scale-Location plot shows that if the variances of the residuals are constant. To satisfy this assumption the red line should be straight line. So, the variances of the residuals are not constant.

Since diagnostic checks are failed, transformation on response should be applied.

5. Conclusion

In this project, customer informations in a bank is examined with many statistical ways to understand the dataset more effectively. Firstly, the bar plot is used to see frequencies of property areas. From this bar plot, it is found that Urban Area has the highest frequency. Secondly, one sample t-test is used to analyze the average values of coapplicant income and loan amount. However, one sample t-test was not reliable because these two variables are not normally distributed. Hence, the sign test which is nonparametric version of one sample t-test is used since it does not require normality assumption. Thirdly, two sample t-test and Mann-Whitney-Wilcoxon test are used to check if the average applicant income significantly higher for applicants who are graduate than not graduate. It is found that the average applicant income is significantly higher for applicants who are graduate than not graduate. After that, if the average loan amount the same for all property areas is checked with One-Way ANOVA and Kruskal-Wallis test. As a result, the difference between the average loan amount for property areas is not statistically significant. Finally, a multiple linear regression model is conducted with variables LoanAmount, Gender, Married, Education, Self_Employed, ApplicantIncome, CoapplicantIncome however assumptions are failed. Then, the insignificant variables are removed and a new multiple linear regression model is conducted with variables Married, Education, ApplicantIncome, CoapplicantIncome. Although all

of the coefficients are statistically significant, the diagnostic checks are failed. Hence, transformations should be tried on variables to find best fitted model.

6. References

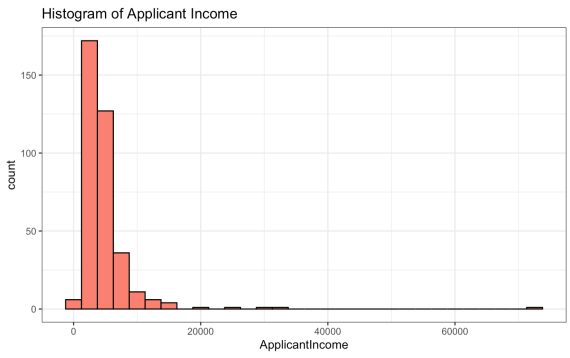
- Kruskal-Wallis H Test using SPSS Statistics [online] Available at: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

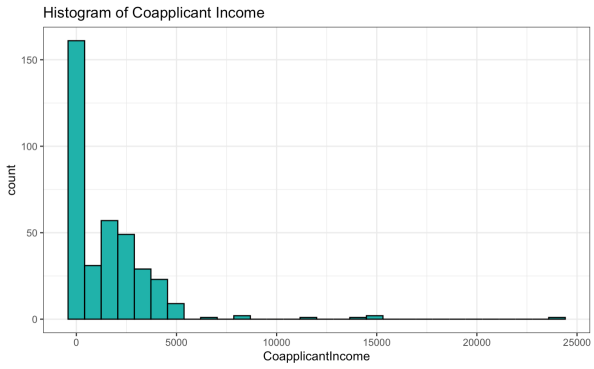
- Sign Test: Step by Step Calculation [online] Available at: https://www.statisticshowto.com/sign-test/

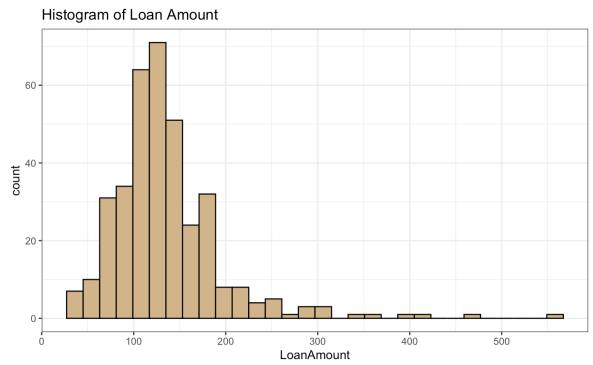
- 5 Strategies Loan Officers Use to Attract More Business [online] https://loanofficerhub.com/5-strategies-loan-officers-use-attract-business/

7. Appendix

"What are the distributions of applicant income, coapplicant income and loan amount?"







Histogram is used to see the distribution of the variable. By looking at the histograms above, it can be said that all of the variables applicant income, copplicant income and loan amount have right-skewed distributions.