ECONOMETRICS PROJECT

ILIYAN TASHINOV

TASHINOV10 (ILIYAN TASHINOV)

INTRODUCTION

This analysis explores the factors influencing Loan Sanction Amounts using a dataset of applicant demographic, financial, and loan-specific variables.

The objective is to develop OLS Linear regression model to identify significant predictors of Loan Sanction Amounts.

№	Variable	Data Type	Unique Values	Continuous
1	Loan.Sanction.Amount..USD.	Numeric	N/A	Yes
2	Gender	Character	['F', 'M', None]	No
3	Age	Integer	N/A	Yes
4	Income..USD.	Numeric	N/A	Yes
5	Income.Stability	Character	['Low', 'High', None]	No
6	Location	Character	['Semi-Urban', 'Rural', 'Urban']	No
7	Loan.Amount.Request..USD.	Numeric	N/A	Yes
8	Current.Loan.Expenses..USD.	Numeric	N/A	Yes
9	Dependents	Integer	N/A	No
10	Credit.Score	Numeric	N/A	Yes
11	No..of.Defaults	Integer	[0, 1]	No
12	Has.Active.Credit.Card	Character	[None, 'Unpossessed', 'Active', 'Inactive']	No
13	Co.Applicant	Integer	[1, 0, -999]	No

Observations: 21 161

Methodology

A. Data Cleaning:
Addressed missing values via imputation or exclusion.
Removed outliers using the IQR method for key variables.

B. Descriptive Analysis:
Explored variable distributions using summary statistics, visualizations, and correlation analysis.

C. Feature Engineering:
Generated dummy variables for categorical predictors.

D. Model Development:
Built and refined OLS multiple regression model, selecting the best based on Adjusted R-squared.
Addressed multicollinearity using VIF analysis.

E. Diagnostics and Validation:
Tested assumptions of normality (Q-Q plot), homoscedasticity (White's tests), and autocorrelation (Durbin-Watson).
Applied transformations and robust standard errors to improve reliability.

A. Data Cleaning | Missing Values

Variable	Action Taken	Reasoning
Gender	Rows with missing values were removed	Small number of missing values (35) and no significant differences in means
Income.Stability	Rows with missing values were removed	Mean values did not differ significantly between categories
Has.Active.Credit.Card	Missing values replaced with 'Unknown'	Distinct mean and quartile ranges for missing values suggested a unique group
Co.Applicant	Rows with -999 values were removed	No significant differences in values; rows deemed unnecessary
Dependents	Missing values replaced with 0	Assumed that missing values implied no dependents
Current.Loan.Expenses..USD.	Values of -999 were replaced with 0	Assumed that -999 represented no loan expenses
Income..USD.	Rows with missing values were removed	Not normally distributed. Could not replace the missing values with the mean

№	Variable	NA Percentage
1	Gender	0.174 %
2	Income..USD.	17.943 %
3	Income.Stability	6.677 %
4	Current.Loan.Expenses..USD.	0.675 %
5	Dependents	9.659 %
6	Credit.Score	6.559 %
7	Has.Active.Credit.Card	5.217 %

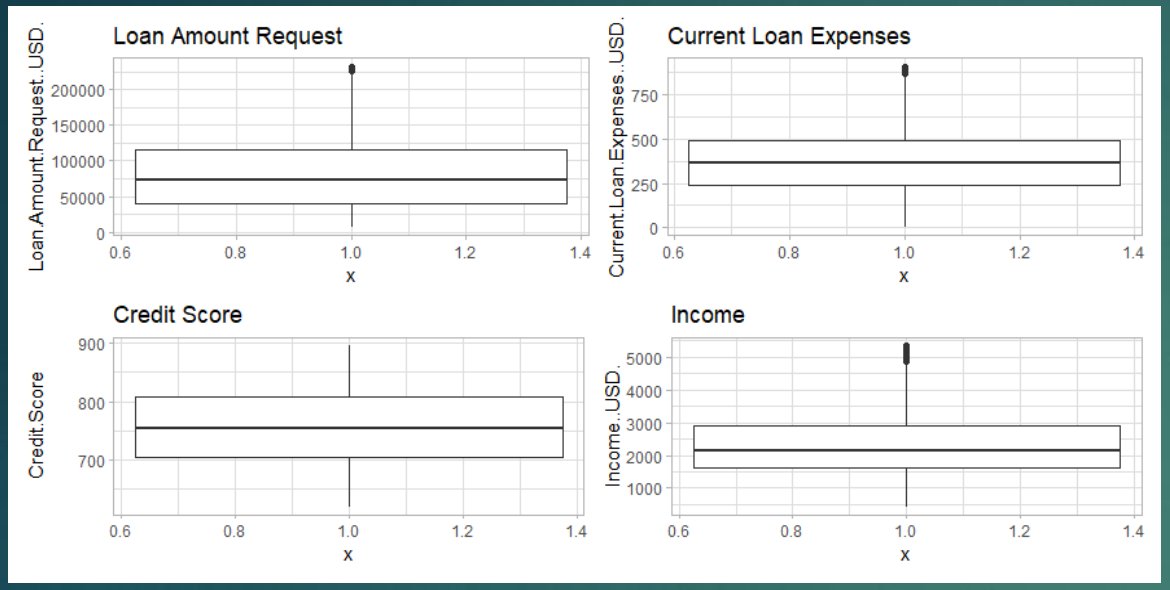
A. Data Cleaning | Extreme Values

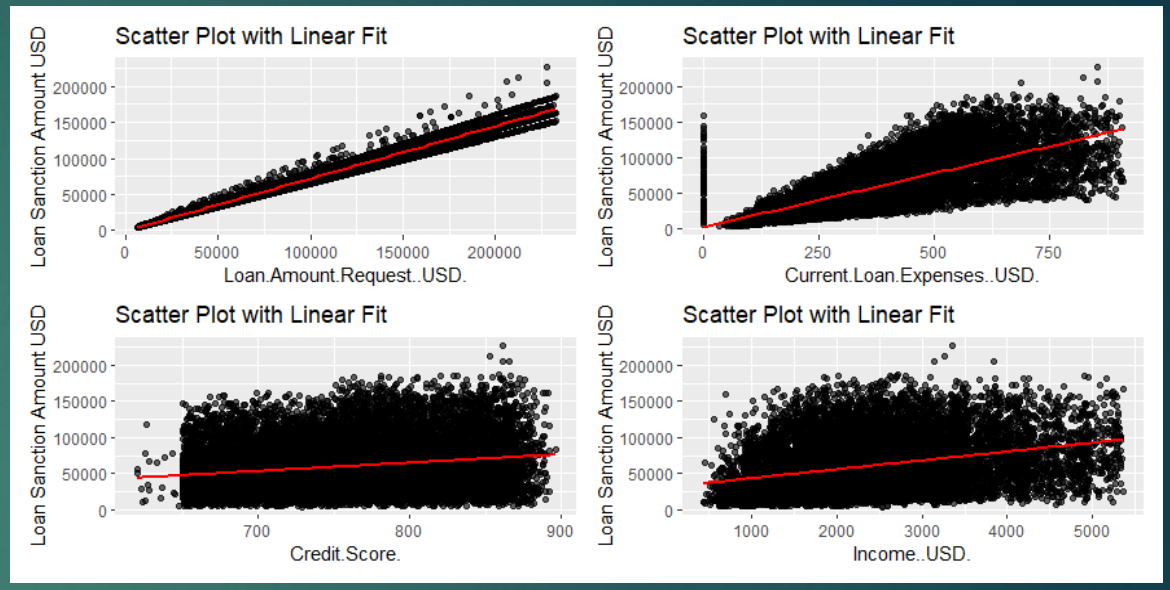
The variables Income..USD., Credit.Score, Current.Loan.Expenses..USD., and Loan.Amount.Request..USD. were adjusted for extreme values by applying the interquartile range (IQR) method.

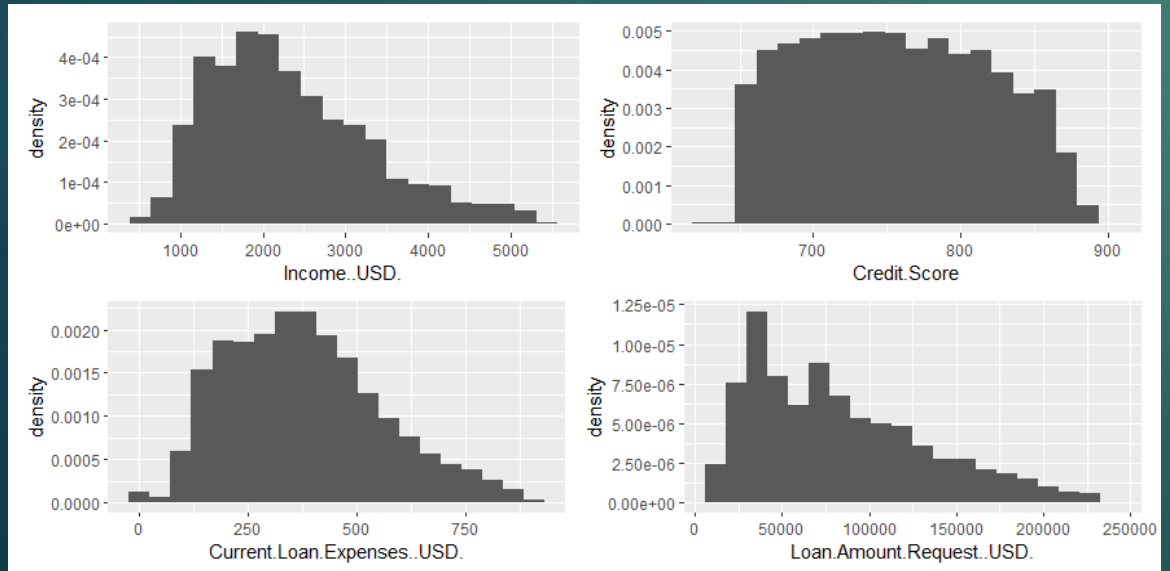
For each variable, observations outside the range of [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] were identified and excluded.

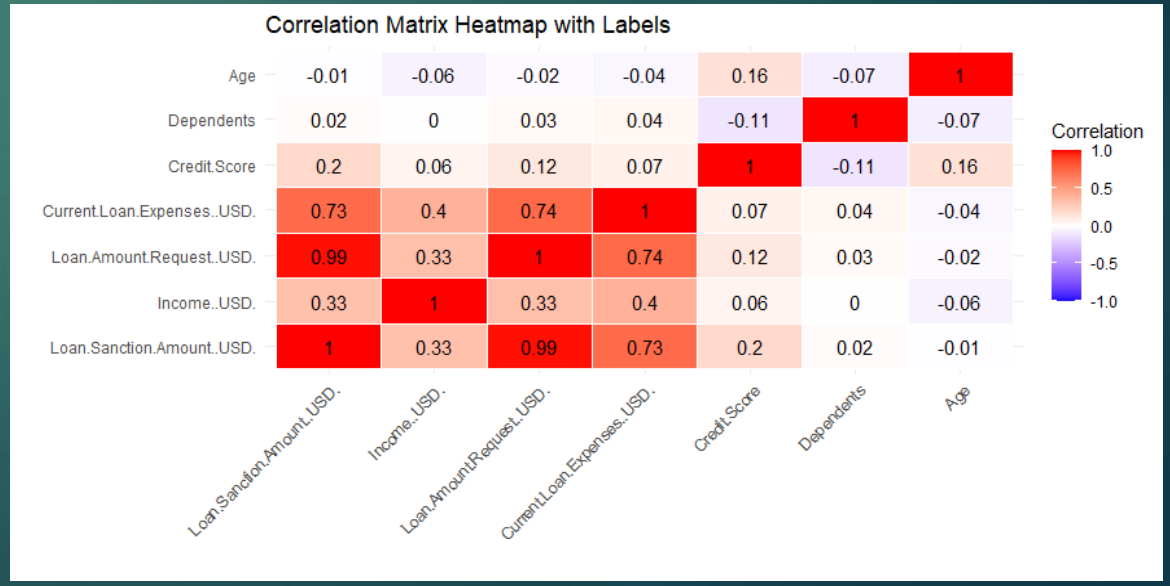
```
q1 <- quantile(loan_data_f1$Credit.Score, 0.25, na.rm = TRUE)
q3 <- quantile(loan_data_f1$Credit.Score, 0.75, na.rm = TRUE)
iqr <- q3 - q1

lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr

cat("Lower Bound:", lower_bound, "\n")
cat("Upper Bound:", upper_bound, "\n")

loan_data_f2 <- loan_data_f1[!is.na(loan_data_f1$Credit.Score) &
                             loan_data_f1$Credit.Score >= lower_bound &
                             loan_data_f1$Credit.Score <= upper_bound, ]
```

B. Descriptive Analysis







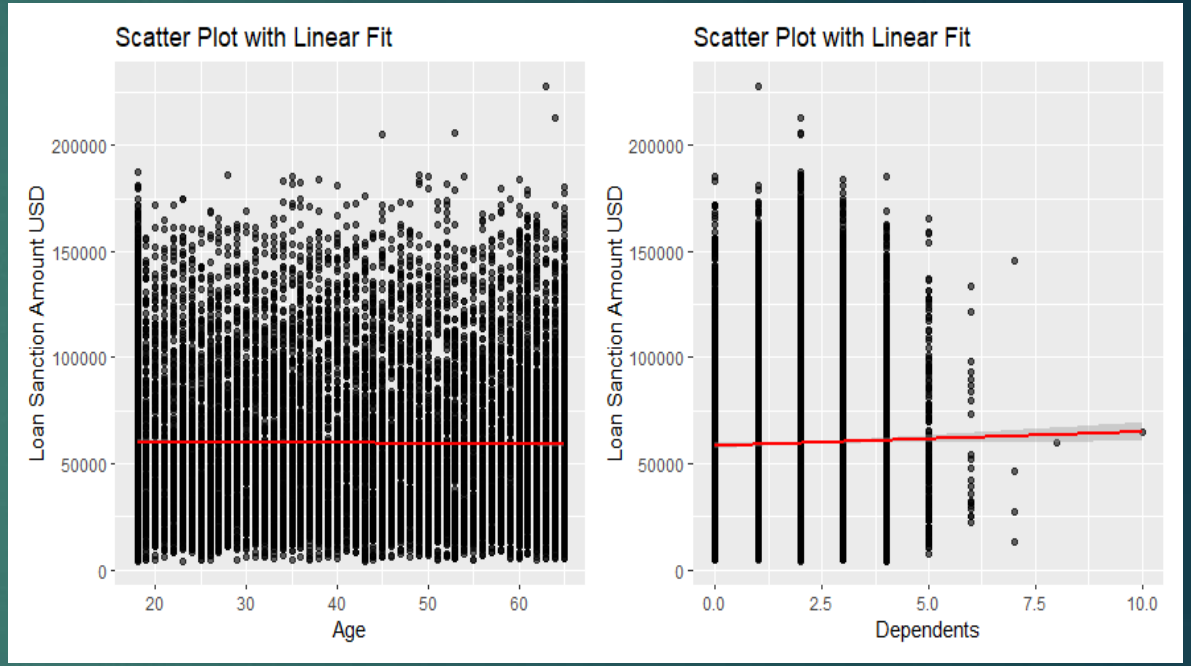


C. Feature Engineering

For the Age variable, it was determined that there was no clear linear relationship between age and the loan sanction amount.

To address this, Age was transformed into bins, to better capture its relationship with the dependent variable:

- **[18-30], [30-40], [40-50], [50-60], [60-70]**



Dependents neither showed evidence for linear relationship with the dependent variable.

It was transformed into categorical variable with two values.

- **Dep_Low** for cases with 0 to 2 dependents and
- **Dep_High** for cases with 3 or more dependents.

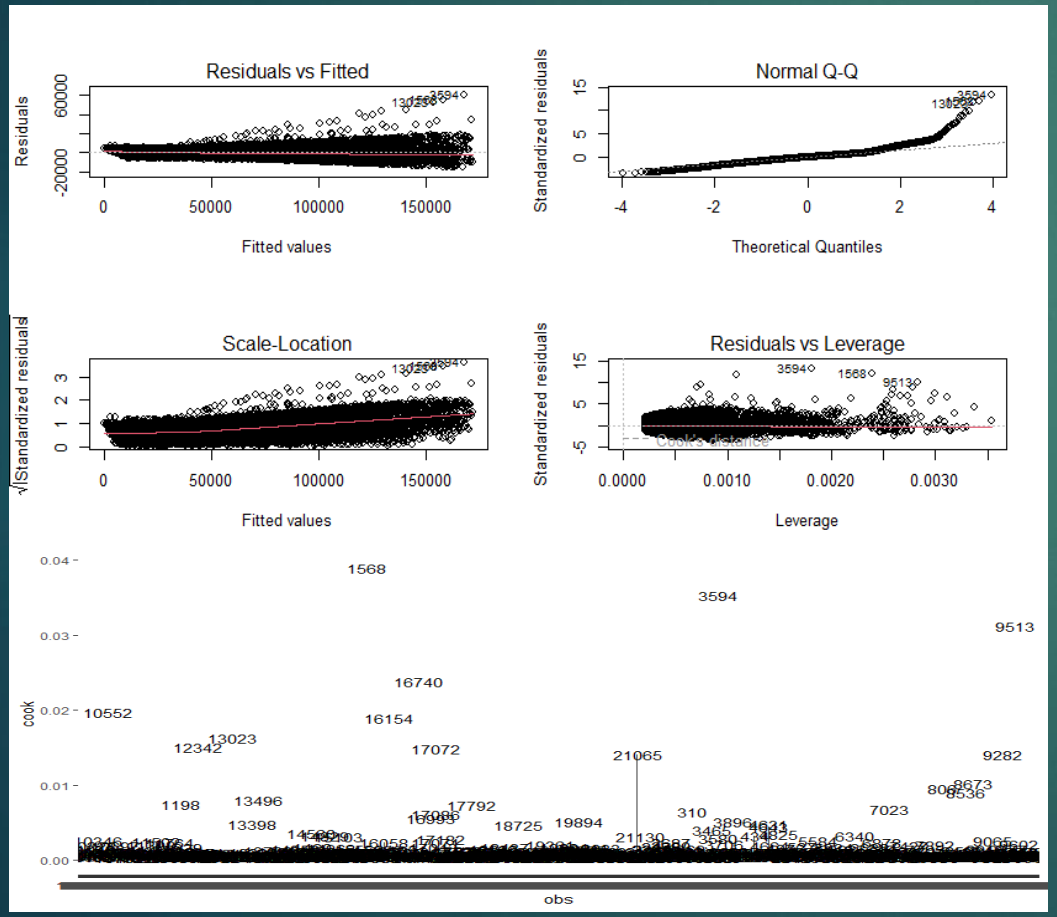
D. Model Development

The primary criterion for selecting a model was its R^2 value. During feature testing, a pattern emerged in the residuals, indicating potential issues. To address possible heteroscedasticity, extreme values were identified and accounted for using Cook's Distance.

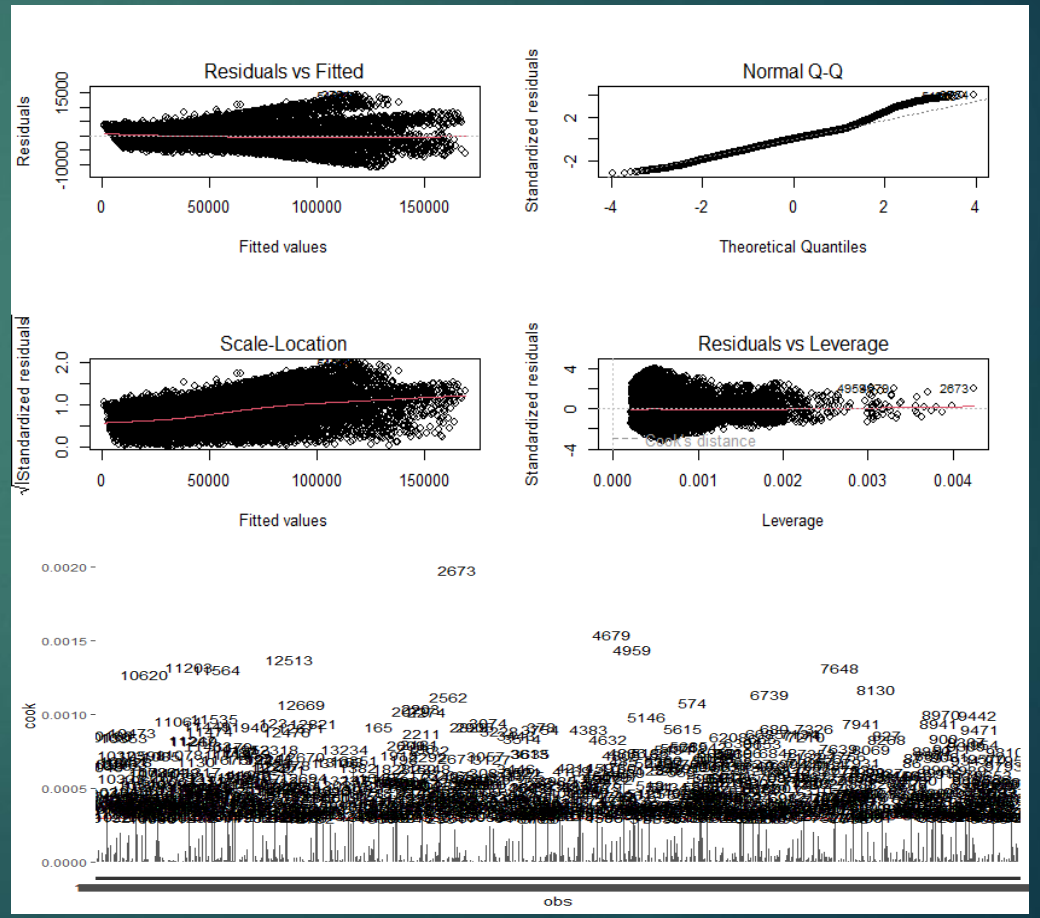
Model	Loan.Amount.Request..USD.	GenderM	LocationSemi-Urban	LocationUrban	Credit.Score	Income..USD.	AgeGroup30-40	AgeGroup40-50	AgeGroup50-60	AgeGroup60-70	Income.StabilityLow	Co.Applicant	No..of.Defaults	Dep_High	R^2
Model 1	<2e-16														0.9785
Model 2	<2e-16	0.0728													0.9785
Model 3	<2e-16	0.0616	3.3E-05	5.7E-13											0.9786
Model 4	<2e-16	0.0446	0.6733	0.696	<2e-16										0.9845
Model 5	<2e-16	0.05014			<2e-16	0.08772	0.79762	0.62038	0.70895	0.00264					0.9846
Model 6	<2e-16	0.0506			<2e-16	0.031	0.7987	0.6358	0.9655	0.5038	6.8E-06				0.9846
Model 7	<2e-16	0.0508			<2e-16	0.0313					1.6E-08				0.9846
Model 8	**<2e-16**	**0.0601**			**<2e-16**	**0.0298**					**<2e-16**	**2.6E-15**	**0.6343**	**0.0327**	**0.9847**

D. Model Development | Cook's Distance

Model 8 before extreme value correction

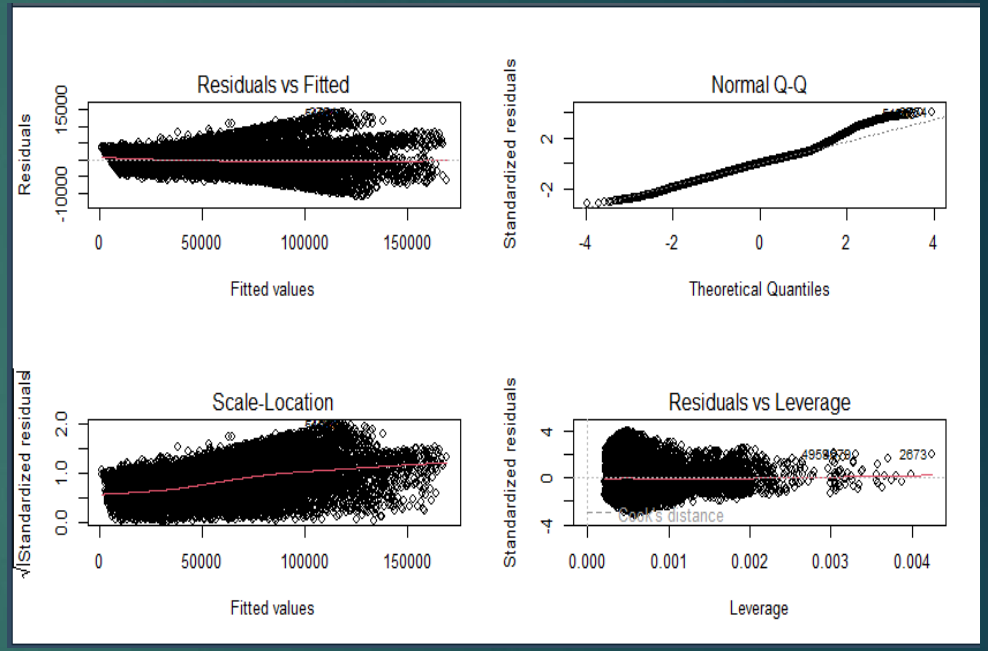


Model 8 after extreme value correction



D. Model Development | model9_c

```
Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)             -3.039e+04  4.477e+02  -67.890  < 2e-16 ***
Loan.Amount.Request..USD.  7.118e-01  6.964e-04 1022.065  < 2e-16 ***
Credit.Score             4.020e+01  5.204e-01   77.237  < 2e-16 ***
Income..USD.            -7.940e-02  3.363e-02   -2.361 0.018241 *
Income.StabilityLow      1.264e+03  1.302e+02    9.705  < 2e-16 ***
Co.Applicant            -6.753e+02  1.813e+02   -3.724 0.000197 ***
Dep_High                -1.891e+02  6.856e+01   -2.757 0.005833 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3522 on 13350 degrees of freedom
Multiple R-squared:  0.989,     Adjusted R-squared:  0.989
F-statistic: 2.001e+05 on 6 and 13350 DF,  p-value: < 2.2e-16
```



E. Diagnostics Validation| Autocorrelation & Multicollinearity

- The Durbin-Watson test was performed on **model9_c** to check for autocorrelation in the residuals. The test statistic is 2.003, which is very close to 2, and the p-value is 0.858. This indicates that there is no significant autocorrelation in the residuals, as we fail to reject the null hypothesis (ρ=0\rho = 0ρ=0). The residuals appear to be independent.

- The Variance Inflation Factor (VIF) values for model9_c indicate no significant multicollinearity among the predictors, as all VIF values are below the commonly accepted threshold of 5. The highest VIF is 1.668 for Income.Stability, which is well within acceptable limits.

```
> vif(model9_c)
Loan.Amount.Request..USD.          Credit.Score          Income..USD.      Income.Stability
                 1.127825              1.132848              1.136708              1.668767
            Co.Applicant              Dep_High
                 1.543202              1.047303
```

E. Diagnostics Validation | Normality & Heteroscedasticity

- The Jarque-Bera test was conducted on the residuals of model9_c to check for normality. The test statistic is 978.93 with a p-value less than 2.2e-16, which is highly significant. The residuals exhibit significant deviations from normality.

- White's Auxiliary Regression confirms evidence of heteroscedasticity in the residuals of the original model. Specific predictors and interactions (Loan.Amount.Request..USD., Credit.Score, and their interactions) contribute to this variance.

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.438e+08  2.585e+07  -5.563 2.71e-08 ***
Loan.Amount.Request..USD.               -1.187e+02  4.765e+01  -2.491  0.01277 *
I(Loan.Amount.Request..USD.^2)          -2.121e-04  7.127e-05  -2.977  0.00292 **
Credit.Score                             3.425e+05  6.835e+04   5.012 5.46e-07 ***
I(Credit.Score^2)                       -2.165e+02  4.531e+01  -4.777 1.79e-06 ***
Income..USD.                             2.241e+03  2.291e+03   0.978  0.32797
I(Income..USD.^2)                       -4.523e-01  1.520e-01  -2.977  0.00292 **
Income.StabilityLow                      3.824e+06  7.153e+05   5.346 9.13e-08 ***
Co.Applicant                             1.677e+05  9.836e+05   0.170  0.86462
Dep_High                                -1.016e+06  3.718e+05  -2.733  0.00628 **
Loan.Amount.Request..USD.:Credit.Score   4.618e-01  6.184e-02   7.468 8.63e-14 ***
Loan.Amount.Request..USD.:Income..USD.   2.848e-03  4.206e-03   0.677  0.49838
Credit.Score:Income..USD.               -5.092e-01  2.925e+00  -0.174  0.86179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19070000 on 13344 degrees of freedom
Multiple R-squared:  0.2202,    Adjusted R-squared:  0.2195
F-statistic: 313.9 on 12 and 13344 DF,  p-value: < 2.2e-16
```

E. Diagnostics Validation | Logarithmisation

After applying logarithmisation to the model, the auxiliary regression results indicate that heteroscedasticity is still present, particularly influenced by Loan.Amount.Request..USD., Credit.Score, and their interactions.

```
Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -3.9870943  0.0400784  -99.482  < 2e-16 ***
log(Loan.Amount.Request..USD.)  0.9981762  0.0007573 1318.156  < 2e-16 ***
log(Credit.Score)              0.5570862  0.0059560   93.534  < 2e-16 ***
log(Income..USD.)             -0.0023658  0.0011757   -2.012   0.0442 *
Income.StabilityLow            0.0131041  0.0019818    6.612 3.93e-11 ***
Co.Applicant                  -0.0164851  0.0027589   -5.975 2.36e-09 ***
Dep_High                      -0.0013559  0.0010431   -1.300   0.1936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05359 on 13350 degrees of freedom
Multiple R-squared:  0.9935,    Adjusted R-squared:  0.9935
F-statistic: 3.403e+05 on 6 and 13350 DF,  p-value: < 2.2e-16
```

Log model*

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.438e+08  2.585e+07  -5.563 2.71e-08 ***
Loan.Amount.Request..USD.               -1.187e+02  4.765e+01  -2.491  0.01277 *
I(Loan.Amount.Request..USD.^2)          -2.121e-04  7.127e-05  -2.977  0.00292 **
Credit.Score                             3.425e+05  6.835e+04   5.012 5.46e-07 ***
I(Credit.Score^2)                       -2.165e+02  4.531e+01  -4.777 1.79e-06 ***
Income..USD.                             2.241e+03  2.291e+03   0.978  0.32797
I(Income..USD.^2)                       -4.523e-01  1.520e-01  -2.977  0.00292 **
Income.StabilityLow                      3.824e+06  7.153e+05   5.346 9.13e-08 ***
Co.Applicant                             1.677e+05  9.836e+05   0.170  0.86462
Dep_High                                -1.016e+06  3.718e+05  -2.733  0.00628 **
Loan.Amount.Request..USD.:Credit.Score   4.618e-01  6.184e-02   7.468 8.63e-14 ***
Loan.Amount.Request..USD.:Income..USD.   2.848e-03  4.206e-03   0.677  0.49838
Credit.Score:Income..USD.               -5.092e-01  2.925e+00  -0.174  0.86179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19070000 on 13344 degrees of freedom
Multiple R-squared:  0.2202,    Adjusted R-squared:  0.2195
F-statistic: 313.9 on 12 and 13344 DF,  p-value: < 2.2e-16
```

Auxiliary model*

Conclusion

After applying logarithmisation to the model, the auxiliary regression results suggest the continued presence of heteroscedasticity, particularly influenced by variables like Loan.Amount.Request..USD. and Credit.Score, as well as their interactions. While robust standard errors could be employed to address heteroscedasticity and improve model reliability, further analysis and alternative model formulations remain viable.

For instance, a logistic regression model could be developed to predict whether a loan sanction exceeds 75% of the requested amount. Such a formulation leverages the high correlation between the loan sanction amount and the loan requested amount. This alternative approach might yield different insights