

**Bank Loan Interest Rate Analysis**

**Project Report**

**DATA 603-Statistical Modelling with Data**

**Group No. 23**

**Submitted to: Dr. Mina Aminghafar**

**Submitted by:**

Warren Wang, UCID: 30260960

Sharlin Kahlon, UCID: 30224889

Gagandeep Thind, UCID:30254133

Cancan Chen, UCID:30270353

Chandanarchutha Namburu, UCID: 30263600

# 1. Introduction

1.1. Motivation  
1.1.1. Context

The banking and financial sector plays a significant role in shaping the global economy. A key function of this sector is providing loans to individuals and businesses, with interest rates being a critical determinant of the cost and accessibility of these loans. According to the Federal Reserve (2020), various factors such as income, credit history, and employment status influence the terms under which loans are approved and the associated interest rates. Understanding how these factors interact can help financial institutions optimize their loan approval processes, reduce risks, and enhance customer satisfaction.

### 1.1.2. Problem

The problem being addressed in this project is understanding the relationship between borrower characteristics and the interest rates charged by bank loans. Specifically, the project aims to explore which features—such as income, employment length, loan amount, and creditworthiness—are the most predictive of interest rate determination.

The analysis seeks to uncover insights that will enable financial institutions to make more informed decisions about loan approval, streamline their processes, and optimize loan offerings. This data-driven approach will help determine interest rates and improve overall decision-making for the client and lending bank.  
1.1.3. Challenges  
The problem is complex due to several challenges commonly encountered in financial data analysis.

First, data quality and cleaning are the major concerns. Financial datasets often contain missing values, outliers, and inconsistencies that can skew analysis (How is Data Science used in Finance? Benefits & Applications, n.d.).

Second, multicollinearity between independent variables can complicate the modeling process by inflating variance and reducing model accuracy (Hayes, 2024)

Furthermore, the relationships between independent variables and interest rates are rarely linear, requiring more sophisticated methods, such as multiple regression, to capture the true nature of these dependencies (Nicholashagemann, 2021).

Lastly, selecting and engineering relevant features is critical to model performance, as irrelevant or redundant features can hinder the predictive power of regression models (Using Feature Selection to Improve Model Performance, 2024).

1.2. Objectives  
1.2.1. Overview

The overall goal of this project is to analyze bank loan data to understand how various borrower characteristics influence the interest rates they are charged. The primary intent is to develop a model that predicts interest rates based on these characteristics, helping financial institutions assess loan applications more efficiently. By focusing on loan amount, employment length, income category and other variables, this project seeks to uncover patterns that can guide future loan decision-making processes.

1.2.2. Goals & Research Questions  
The main objectives of this project are:

* Data Cleaning: Ensure the dataset is free from missing values and outliers, making it ready for accurate modeling.
* Model Building: Implement a multiple regression model to predict interest rates based on the independent variables provided in the dataset. Whether financial institutions can use these insights to streamline their loan approval process, potentially reducing the time spent reviewing applications.

**Research Questions:**

1. **How do the independent variables (e.g., loan amount, income category, employment length, etc.) influence the interest rate?**
2. **Which independent variables are most predictive of the interest rate?**
3. **Can a multiple regression model be developed to accurately predict interest rates, and what level of accuracy can be achieved?**

This project will focus on cleaning the dataset and building a regression model to uncover these relationships. The insights gained will be valuable for financial institutions to determine loan eligibility more efficiently.

2. Methodology  
2.1. Data

2.1.1 Data Introduction

The dataset we obtained is from Kaggle which is an open-source database. Link to the dataset:- <https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank?resource=download>

The dataset has been filtered to include records from 2010. Out of the 30 variables in the dataset, here is a summary of the key variables that will be utilized for the analysis:

The variables that we used in our analysis are listed below:  
**1.) Emp\_length\_int**: employment length, measured in months, quantitative;

2.) **Home\_ownership**: weather the individual owns a house or not, it is a categorical variable, its types are mortgage, rent, other;

3.) **Income\_category**: qualitative variable, measured as low, high and other

4.) **Annual\_inc:** the amount of annual income, measured in dollars- quantitative variable

5.) **Term:** how long the term of loan is, measured in either 36 months or 60 months- quantative;

6.) **Purpose:** qualitative variable, measured in debt\_consolidation, credit\_card and other;

7.) I**nterest\_payments**: qualitative variable, measured in high or low;

8.) **Loan\_amount**: the amount of loan borrowed, measured in dollars- quantitative variable;

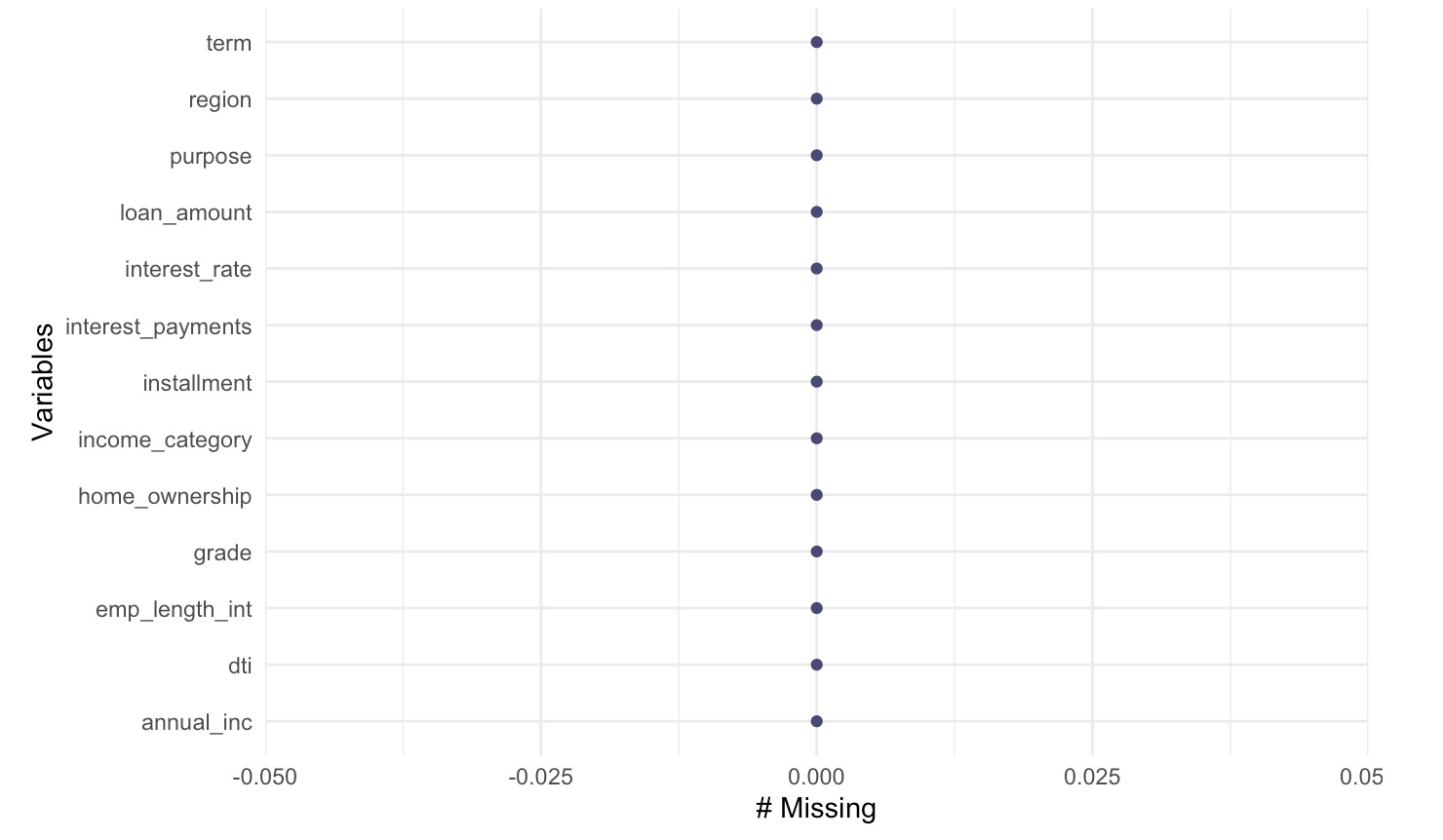
9.) **Grade:** assigned loan grade, measured in B,C or other;

10.) **DTI:** debt-to-income ratio, quantitative variable;

11.) **Interest\_rate**: our response variable, quantitative and continuous variable, its unit is %.

2.1.2 Data Cleaning

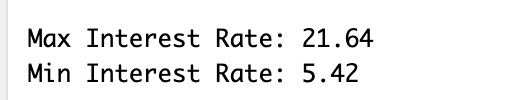
* **Check for Missing Values**



* **Check for Duplicates**

We used the duplicated() method and found that there are 0 duplicate records.

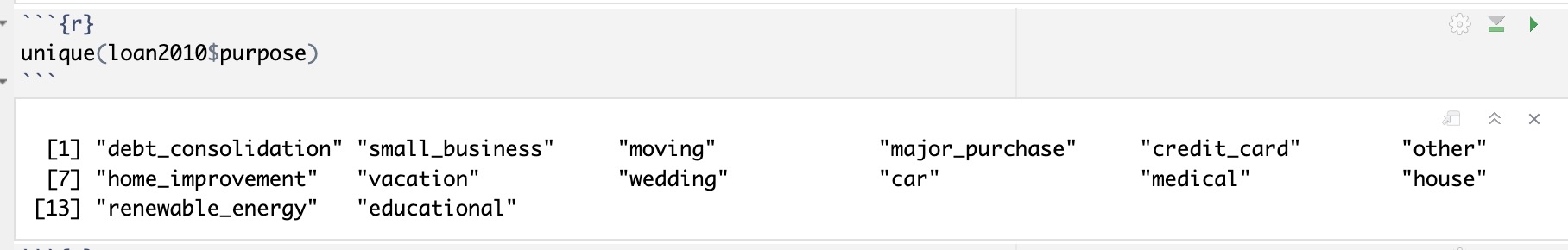
* **Check for outliers in the dependent variable, *interest rate***



From the results, we conclude that the interest rate ranges from 5.42% to 21.64%, which aligns with real-life scenarios.

* **Check for Inconsistent Categorical Values**

We used the unique() function to check the categorical variables. Some of the results are as follows:



After performing checks for missing values, outliers, and inconsistent categorical values, we concluded that our dataset is quite clean, and we can proceed with building the model.

2.2. Approach

In this project, our focus is to find the best model that will determine the rate of interest on a loan and its relationship to the variables mentioned above. Here, we have one response variable and multiple independent variables to be examined, multiple regression will be used as our main method to find the best model that measures the interest rate. This is also why we think that multiple regression will work very well here.

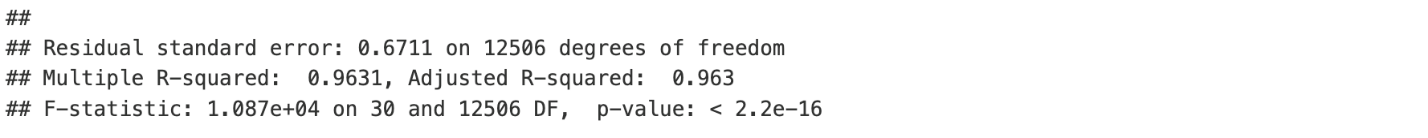
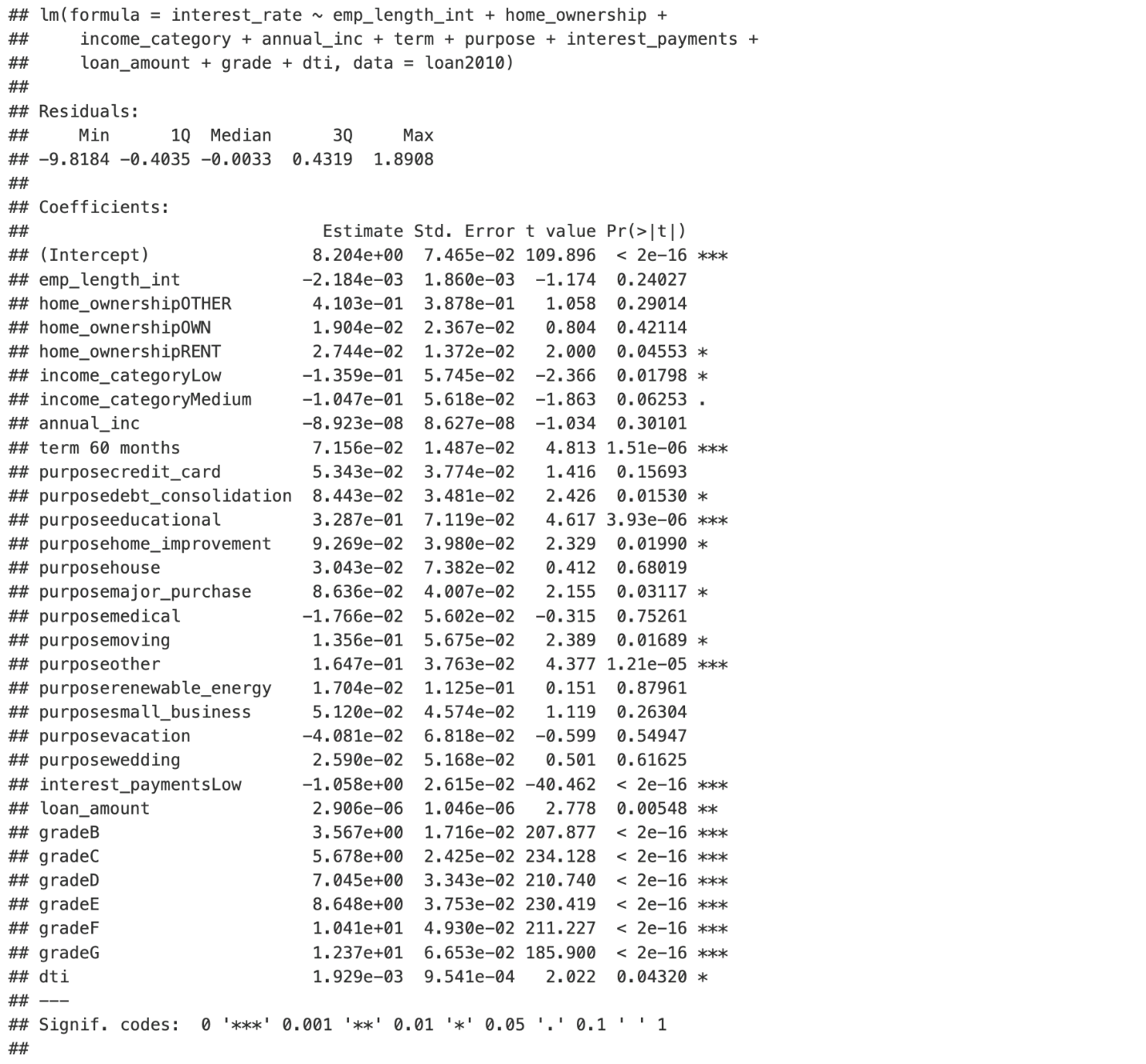
As per usual, our alpha value (level of significance) that we will use in this project will be 0.05.

We also checked the assumptions like linearity, normality, equal variance, independence and multicollinearity with their corresponding graphs like histograms, residuals vs fitted plots etc.

2.3. Workflow

For multiple regression, we will start with the full model, then we perform variable selection to weed out irrelevant variables. We will compare the full model and the reduced model to select the best model and see if any terms should be a higher term. In the end, we will also check necessary assumptions for the model.

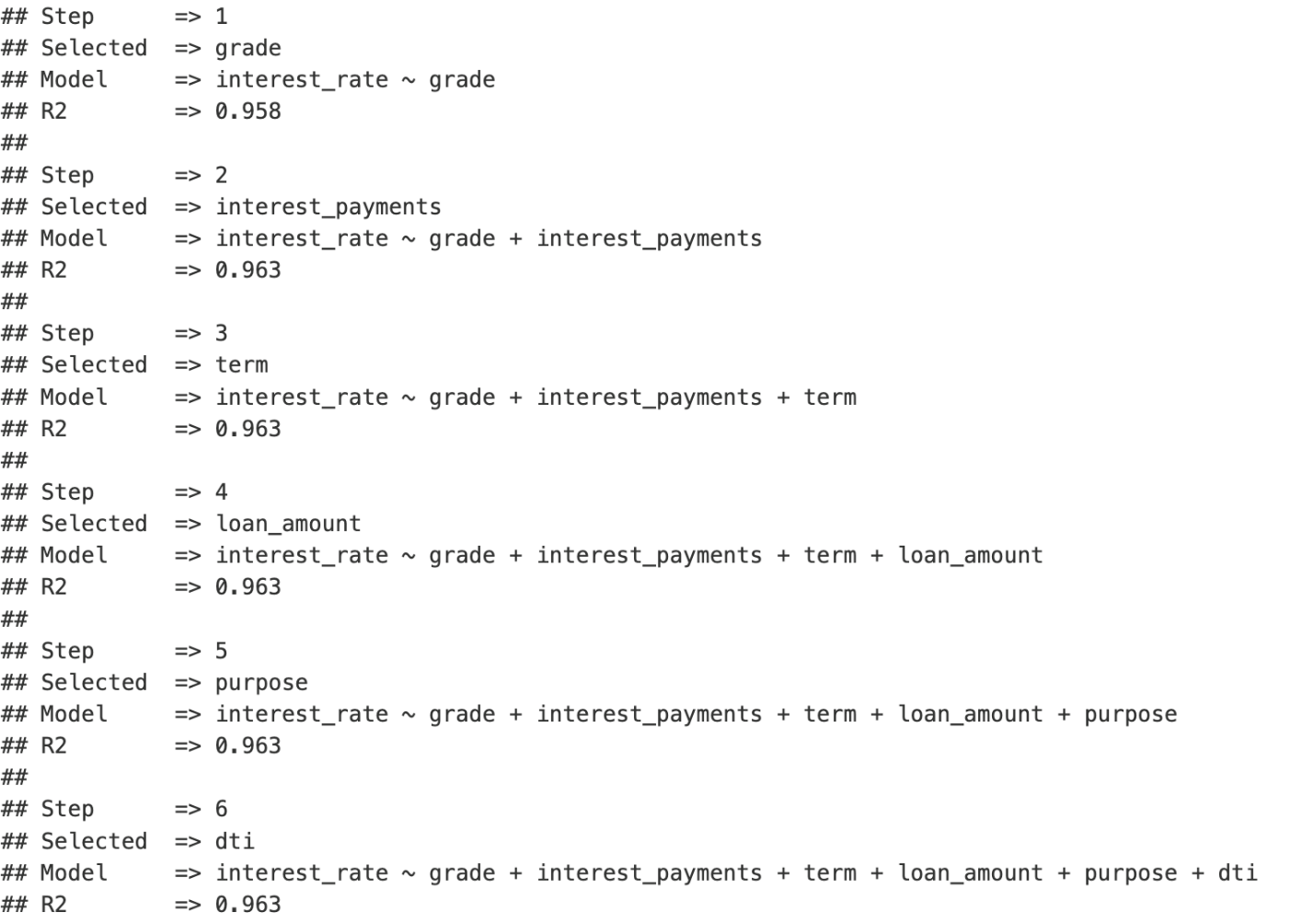
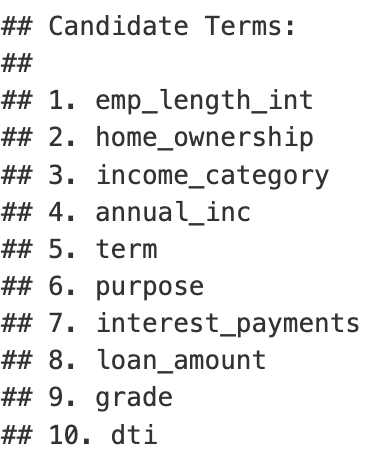
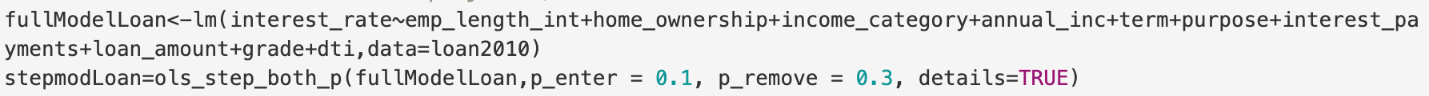
First, we start with the full model:



As can be seen from the output, the model overall had a p-value that is way below the alpha level which provides us with extreme evidence against the null hypothesis that all the coefficients are zero. The adjusted R-squared value is 0.963 with a relatively low RSE of 0.6711. Therefore, this is a good starting point.

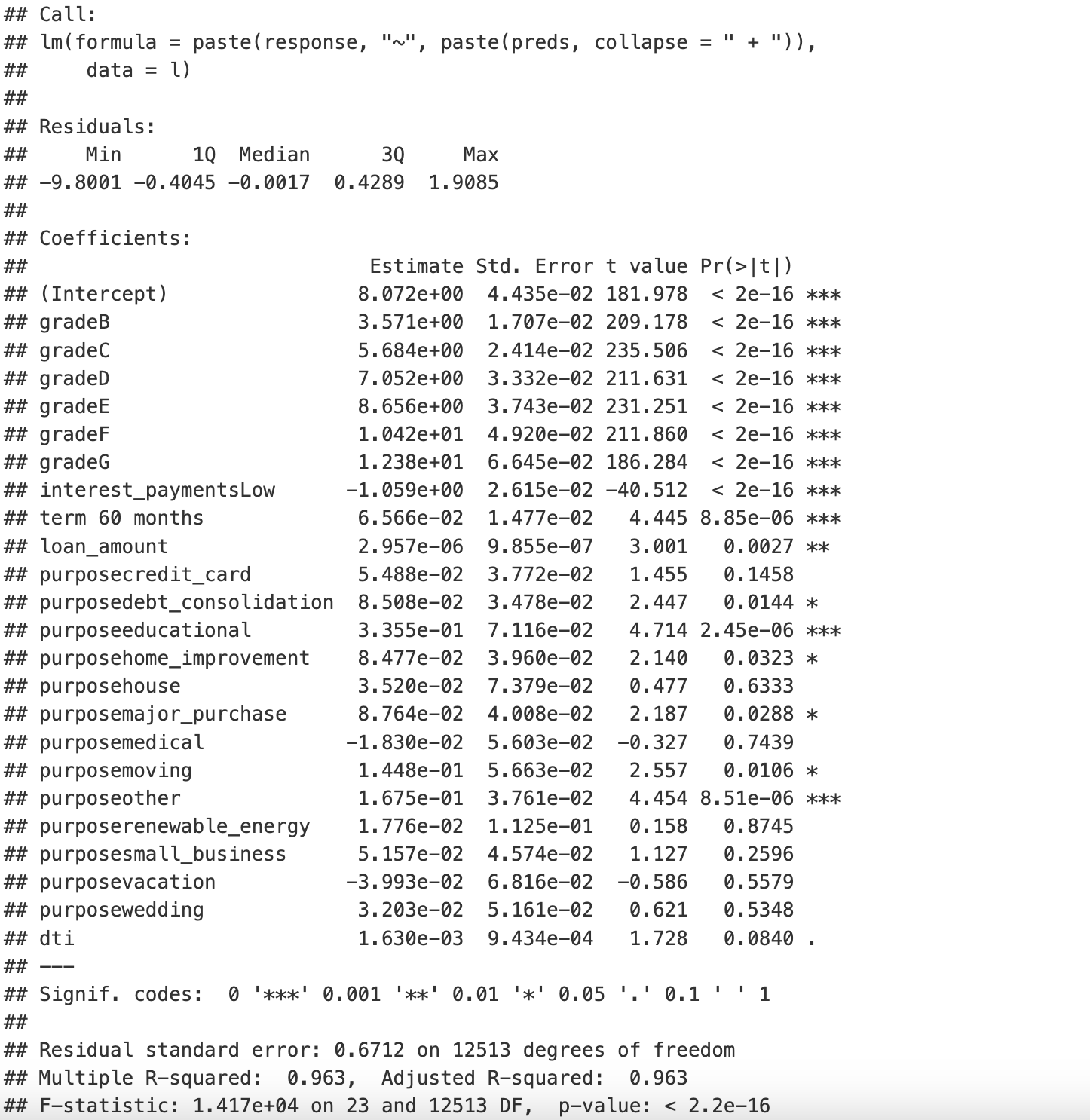
|  |  |
| --- | --- |
| **R-Squared Value** | **RSE** |
| 0.963 | 0.6711 |

Then a variable selection was performed.



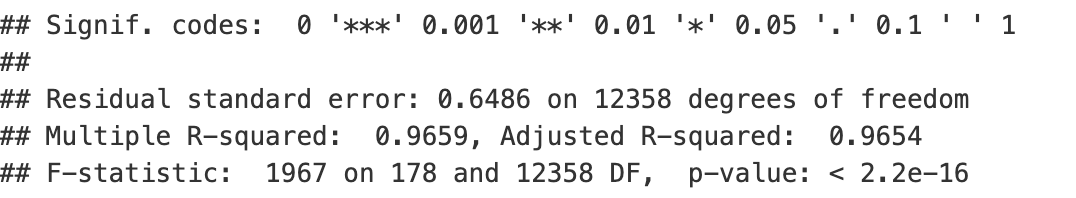
Eventually, there were 6 variables left after the selection. So, reduced model was also constructed, and its summary is below:





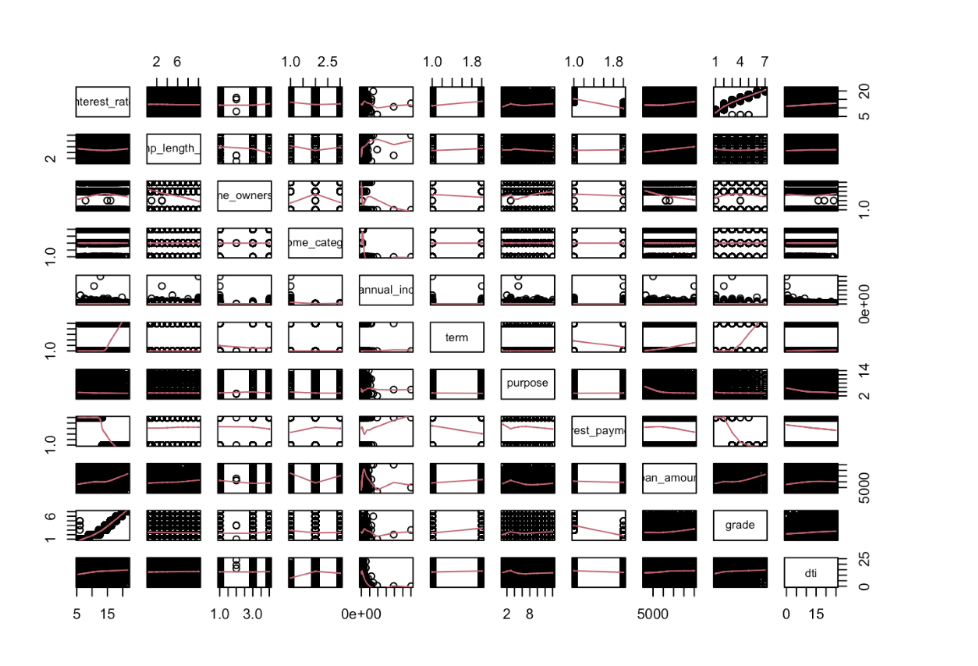
The result was very interesting as we can see the adjusted-R-squared value is the exact same as the full model! However, it has a relatively higher RSE than that of the full model. So far, the full model is a better model.

Then, an interaction model was developed. The results were too long to list here but most of the interaction terms are not significant, and the adjusted-R squared value only improved by 0.24% which is barely a difference. Therefore, it can be concluded that the interaction model is likely overfitting the data. Therefore, we concluded that the full model is the best model. The final part of the interaction model output is attached below.



There is barely any improvement in the performance of the interaction model as can be seen. Therefore, we concluded that the full model is the best model.

We also wanted to see if any of the terms should be a higher order, so we plotted the variables out with one another. It can be seen that interest rate’s relationship (as the response variable on the y-axis) is quite linear with all other variables (on the x-axis) which means no terms should be a higher order in the model.



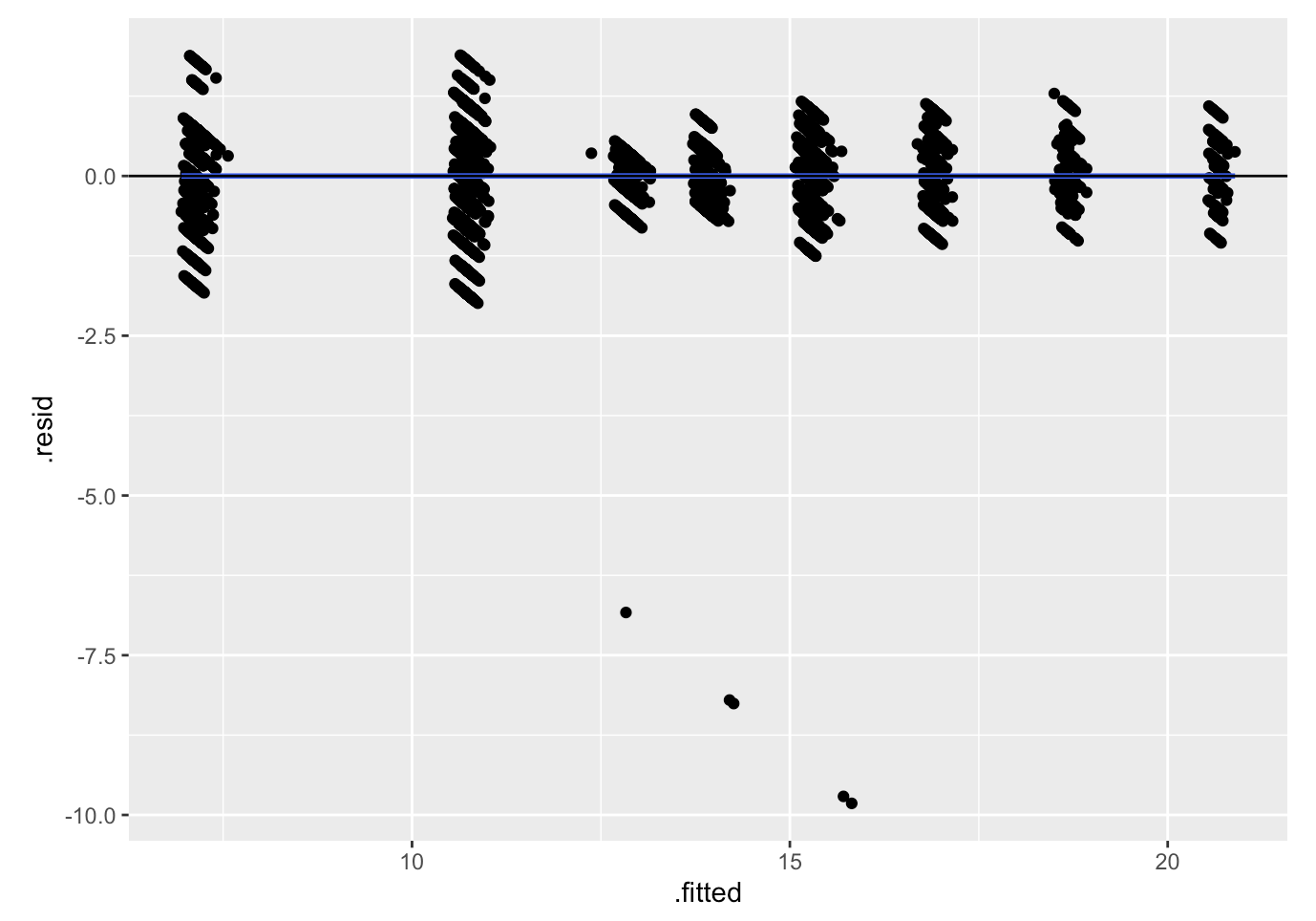
**Assumptions Check:**

We also checked a few assumptions: linearity, normality, equal variance, independence and multicollinearity.

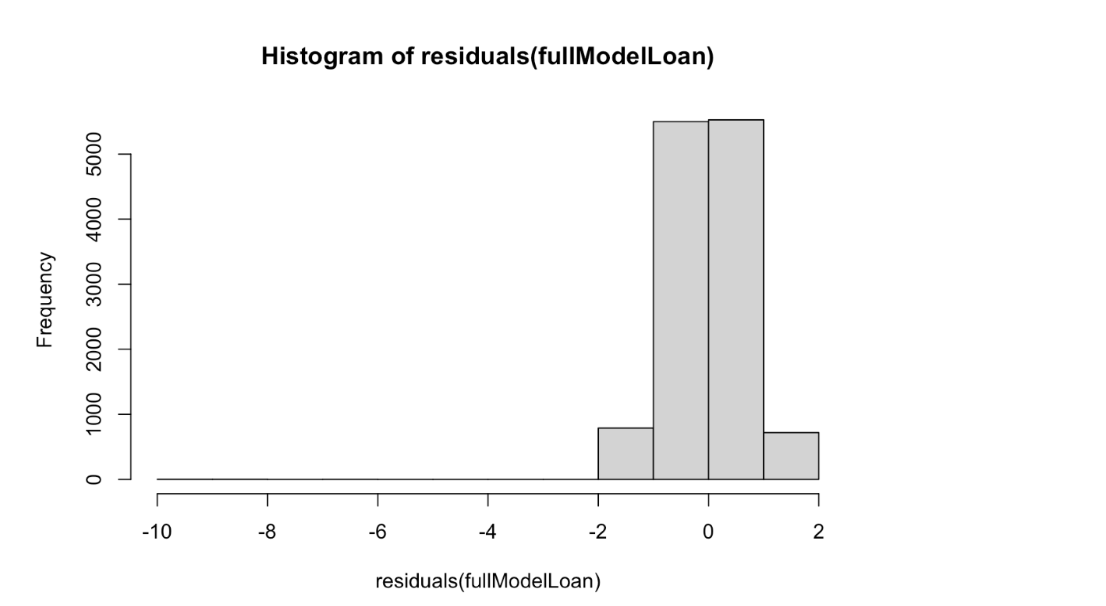
For linearity, if there is no evident pattern in the residual plot, then the assumption is met.

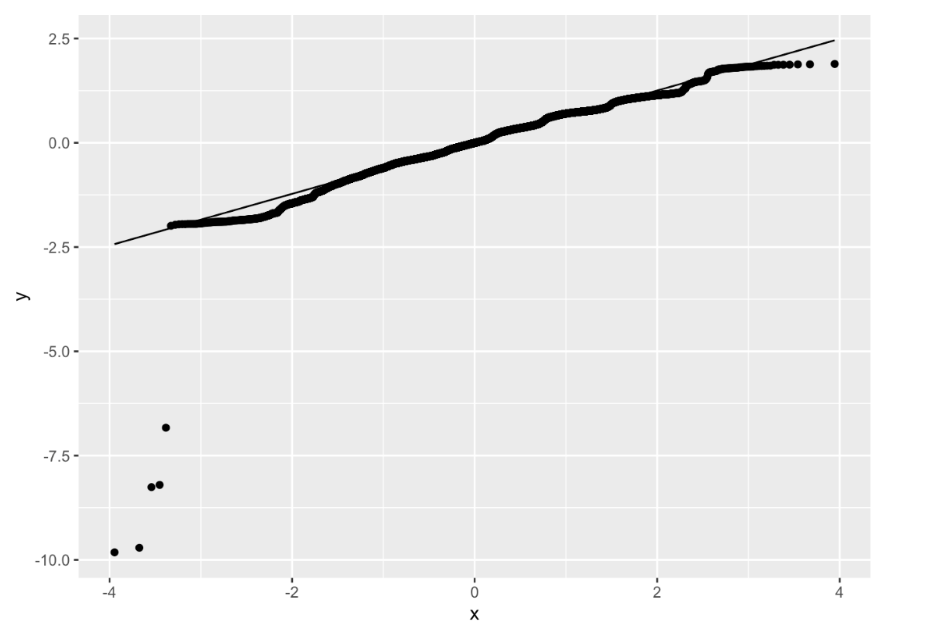
As can be seen from the plot below, the residual of the full model is perfectly straight. Even though there are

a few outliers, but it won’t make a difference in our result as we have more than 12000 data points.



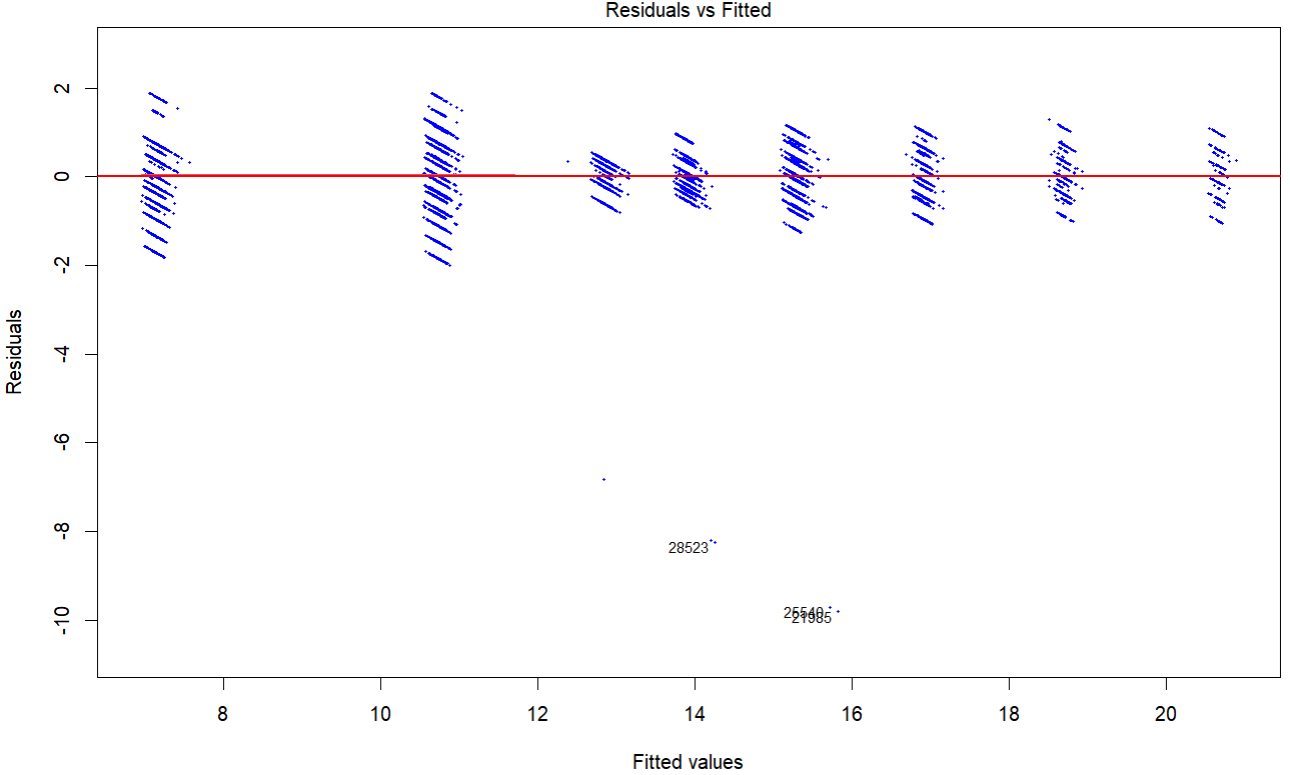
For normality, if the histogram shape is symmetrical and if the line of the QQ-plot is straight, then the assumption is met.





**Result:** As we can see, the normality assumption is very well met in the full model’s residuals.

**Testing for Independence:**

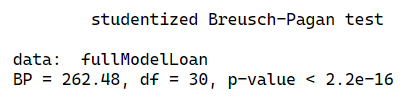


**Result:** In the *Residuals vs. Fitted* plot provided above, the residuals appear to be scattered randomly around the horizontal red line at zero without any clear trend or pattern. The random scatter of points around zero suggests there is no dependency or correlation in the residuals. This is a good sign, indicating that the residuals are independent.

**Breusch-Pagan Test for Consistency of Variance:**

Null Hypothesis (*H0*): The variance of the residuals is constant.

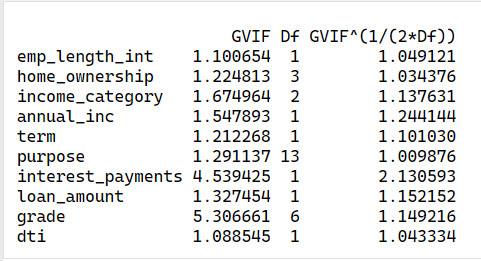
Alternative Hypothesis (*H1*): The variance of the residuals is not constant.



**Result:** Based on the results of the Breusch-Pagan test, the p-value was found to be extremely small (p < 0.05). Therefore, we reject the null hypothesis and conclude that there is strong evidence of heteroscedasticity in the model. While it is possible to apply transformations to stabilize the variance, we opted not to do so for the following reasons:

* **Reliability of Parameter Estimates in Large Datasets**: The central limit theorem ensures that, in models with large datasets, the parameter estimates remain reliable and asymptotically normal, even in the presence of heteroscedasticity.
* **Limited Impact of Heteroscedasticity in Large Samples**: In large samples, small deviations in variance are less likely to significantly affect the validity of the results, reducing the necessity for corrective measures like transformation.

Given these considerations, the model was retained in its current form, as the potential impact of heteroscedasticity was deemed negligible for this analysis.



**Testing for multicollinearity:**

The Variance Inflation Factor (VIF) was calculated to assess multicollinearity among the predictors in the model. Generalized VIF (GVIF) values and their adjusted measures, were reported. The **results** are as follows:

1. **Assessment of GVIF Values**:
2. GVIF values for all predictors are below the commonly accepted threshold of 10, suggxesting that severe multicollinearity is not present in the model.
3. **Adjusted Multicollinearity Measure** :

The adjusted GVIF values are all close to or below 2, further supporting the absence of significant multicollinearity. Notably:

* 1. **Interest Payments (GVIF: 4.54; Adjusted: 2.13)** and **Grade (GVIF: 5.31; Adjusted: 1.15)** show slightly elevated values, but these are still within acceptable limits.
  2. All other predictors exhibit minimal multicollinearity, with adjusted values consistently below 1.25.

1. **Conclusion**:

Multicollinearity does not appear to be a concern for this model. The predictors are sufficiently independent to provide reliable parameter estimates. No corrective actions, such as variable removal or transformation, are necessary at this stage.

2.4. Contributions  
As a team, we all agreed to work on finding the model together and compare our results to avoid the situation where only one member works on finding the model but turns out they made a mistake somewhere which eventually gives us the wrong model and thus the wrong conclusion. For the oral presentation, we will all speak for our assigned part for this report. Sharlin will start with the introduction, Warren will introduce the methodology and Gagandeep will speak for assumption checking. Cancan will interpret the results and state the final model. Finally, Chandanchutha will summarize important findings and discuss their implications.

3. Main Results of The Analysis

## 3.1. Final Linear Model

* The results of the different models are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Number of variables** | **Adjusted R-square** | **RSE** |
| Full model | 10 | 0.963 | 0.6711 |
| Reduced model | 6 | 0.963 | 0.6712 |
| Interaction model |  | 0.9654 | 0.6486 |

* The assumptions of linearity, normality, independence, consistency of variance and multicollinearity have all been satisfied.

From the results table above, although the Adjusted R-squared for the interaction model is the highest and the RSE is the lowest, we have opted for the full model considering its balance between performance and complexity. Therefor we ultimately selected the full model as our final model. The equation is as follows:



From the model coefficients, we can see that the coefficients for **annual\_inc** and **loan\_amount** are very close to 0, indicating a minimal effect on the interest rate. For the sake of accuracy, we have kept them in our model for this report. However, in real-life applications, they could be removed for simplicity.

## 3.2. Results Interpretation

#### **Intercept (β0=8.20)**

#### The intercept represents the interest rate when all predictors are at their reference or zero values. This is the baseline value of the target variable before any predictors are added.

#### Employment Length (β1=−0.002)

* Effect: For each additional year of employment, the predicted outcome decreases by 0.002.
* Implication: Longer employment length slightly decreases the interest rate, but the effect is minimal.

#### **Home Ownership (β2)**

* Mortgage (Reference): Coefficient = 0
  + OTHER: Increases by 0.410.
  + OWN: Increases by 0.019.
  + RENT: Increases by 0.027.
* Implication: Borrowers with "OTHER" home ownership types have a higher interest rate, while "OWN" and "RENT" have minor impacts.

**Income Category (β3)**

* High (Reference): Coefficient = 0
  + Low: Decreases by 0.136.
  + Medium: Decreases by 0.105.
* Implication: Higher-income borrowers are associated with higher interest rates, while lower and medium-income borrowers have negative impacts.

#### **Annual Income (β4=−0.00000008923)**

* Effect: A small decrease in the interest rate with increasing annual income.
* Implication: Though statistically significant, the practical impact is negligible.

**Loan Term (β5)**

* 30 Months (Reference): Coefficient = 0
  + 60 Months: Increases the interest rate by 0.07156.
* Implication: Longer loan terms slightly increase the interest rate.

#### **Purpose of Loan (β6)**

* Car (Reference): Coefficient = 0
  + Educational: Increases by 0.3287 (largest positive effect).
  + Vacation: Decreases by 0.04081.
* Implication: Educational loans are strongly associated with positive outcomes, while vacation loans have a negative effect.

#### **Interest Payments (β7=−1.058)**

* Effect: Higher interest payments significantly decrease the interest rate.
* Implication: Borrowers with higher interest payments face a substantial negative impact.

#### **Loan Amount (β8=0.000002906)**

* Effect: Increasing loan amounts slightly increase the interest rate, but the effect is minimal.
* Implication: Though statistically significant, the practical impact is very small.

#### **Loan Grade (β9)**

* Grade A (Reference): Coefficient = 0
  + B: Increases by 3.567.
  + G: Increases by 12.370.
* Implication: Higher-risk grades are associated with higher predicted outcomes, reflecting higher interest rates.

#### **Debt-to-Income Ratio (DTI) (β10=0.001929)**

* Effect: Higher DTI ratios slightly increase the interest rate.
* Implication: Borrowers with higher DTI may present increased interest rates.

Although the full model is quite complex, it best reflects real-life loan dynamics and provides valuable insights. For example:

1. Key Drivers: Home ownership type, loan grade, purpose, and interest payments have noticeable impacts.
2. Risk Indicators: High DTI and interest payments are potential risk factors.
3. Income Category: Lower and medium-income groups have slightly negative effects compared to high-income groups.
4. Loan Grade Progression: As the grade worsens, the predicted outcome increases, suggesting a need for caution when lending to lower-grade borrowers.

4. Conclusion And Discussion

4.1. Approach

Our objective was to predict interest rates based on various independent variables such as employment length, home ownership, income category, loan amount, and more. After performing the full model, the stepwise model offered a simplified version, which kept the most significant variables like loan amount, grade, and purpose, and removed less significant ones like home ownership and annual income.

Both models had a high adjusted R-squared (96.3%), showing that a significant portion of the variability in interest rates is explained by the predictors in both models. However, the full model had a slightly lower residual standard error (RSE) than the stepwise model, indicating a slightly better fit. This led us to select the full model as our final model.

We then explored whether interactions between variables could improve the model. Although adding interactions slightly increased the adjusted R-squared, most interaction terms were not significant. The increase in model complexity without a significant performance boost suggested that the interaction model could be overfitting the data. For this reason, we rejected the interaction model in favor of the full model.

We checked for potential non-linear relationships by plotting the variables, but we found that the relationship between interest rate and the other variables appeared to be linear, so no higher-order terms were added.

We also checked key assumptions for the model: linearity, normality, independence, and multicollinearity. Given that all the assumptions were satisfied, we can confidently conclude that the model is valid and robust.

4.2. Future Work

While the model is strong in its current form, there are several potential areas for future work. One would be to explore non-linear models such as decision trees or random forests, which could potentially capture more complex relationships between the predictors and interest rates that the linear model might be missing.

Additionally, collecting more detailed data, such as credit score histories or loan repayment performance, could improve the predictive power of the model. This data would help account for customer behavior over time, which might be a crucial factor in how banks assign interest rates.

In conclusion, our analysis found that key factors such as credit grade, loan amount, term length, and purpose of the loan significantly influence the interest rates banks assign to borrowers. The final model explains over 96% of the variance in interest rates, which means it offers strong predictive power. However, improvements could still be made by exploring non-linear methods. This analysis provides actionable insights for banks when determining loan interest rates and for consumers to understand which factors are most influential in the rates they receive.

# 5. References

Nicholashagemann. (2021, 07 30). Predicting Loan Interest Rates. Retrieved from Medium: <https://medium.com/@nicholashagemann0/using-machine-learning-to-predict-interest-rates-and-identify-the-strongest-predictors-aed36ae677c5>

Federal Reserve. (2020, March 17). *How banks decide whether to lend to you*. <https://www.federalreserve.gov/credit.htm>

Hayes, A. (2024, 05 26). *Multicollinearity: Meaning, Examples, and FAQs*. Retrieved from Investopedia: <https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20is%20a%20statistical%20concept,in%20less%20reliable%20statistical%20inferences>.

*Using Feature Selection to Improve Model Performance*. (2024, 05 13). Retrieved from nobledesktop: <https://www.nobledesktop.com/learn/python/using-feature-selection-to-improve-model-performance#:~:text=in%20Machine%20Learning-,Feature%20selection%20is%20a%20critical%20step%20in%20the%20machine%20learning,model%20interpretability%2C%20and%20improve%20ac>

*How is Data Science used in Finance? Benefits & Applications*. (n.d.). Retrieved from <https://aimconsulting.com/insights/data-science-finance-benefits-applications-risks/#:~:text=Financial%20institutions%20can%20leverage%20data,for%20their%20products%20and%20services>.: <https://aimconsulting.com/insights/data-science-finance-benefits-applications-risks/#:~:text=Financial%20institutions%20can%20leverage%20data,for%20their%20products%20and%20services>.