**LOAN ANALYSIS**

**MATH 261A Regression Theory**

**Presented to**

**Dr.** [Martina Bremer](mailto:martina.bremer@sjsu.edu)

**Department of Mathematics**

**San Jose State University**

**In Fulfillment of the Project Report Assignment of the Class**

**MATH 261A**

**By**

**Amelia Le**

**Joshua Chase**

**Tianxiang Chen**

**Xuewei Zhong**

**December 2023**

### Introduction:

Nowadays peer-to-peer financing platforms, such as LendingClub and Prosper Personal Loans, allow for a faster and less stringent way for individuals to qualify for a loan (Treece). These financing platforms use the applicants’ financial history to match them with individuals or companies willing to provide the necessary fundings (Suknanan). Similarly to traditional financial institutions, lenders on peer-to-peer financing platforms will charge a higher interest rate for applicants who possess a volatile financial background as a way to minimize the financial loss in the event that the borrowers default on their loans (Canandaigua National Bank & Trust). This project is inspired by our interest in understanding how financial institutions calculate interest rate of loan based on an applicant’s financial background.

In this project, we will be using the Loans Full Scheme dataset which contains information on 10,000 borrowers who successfully applied for a personal loan through the LendingClub platform from January of 2018 to March of 2018. The dataset can be obtained through either the “OpenIntro” library in R by calling “loans\_full\_scheme” or the OpenIntro website where you can download the csv file (OpenIntro). From the 55 features available for each individual loan, we selected 15 features we believe to have significant influence on determining the interest rate for each loan. Our initial variable selection is based on the different factors lenders commonly use to determine the interest rate (Waugh) and what the LendingClub uses to calculate the APR of a personal loan (LendingClub). We will be using these features to fit a model for the purpose of answering the following questions: 1. Is our fitted linear regression model able to appropriately determine the interest rate for future loan borrowers? and 2. Which features are the most important in determining the interest rates for each borrower? The following report will discuss the methodology and results of our findings.

### Data Exploratory Analysis/Variable Selection

The predictor variables used in this project are categorized into two types, numerical and categorical variables. The numerical predictor variables are annual income, total credit lines, total credit limit, debt-to-income, installment, loan amount, total credit utilized, number total credit card accounts, number of historical failed ot pay, and term, where the predictors total credit lines, terms, total credit card accounts, and number of historical failed to pay were originally ordinal variables that we decided to treat as continuous variables for the simplicity of constructing a linear regression model. In addition, our categorical variables are grade (from A to G), loan purpose (12 different levels), application type (whether the loan is individual or joint), homeownership (mortgage, owner, or renter), and public record bankruptcy (number of bankruptcy listed on the borrower’s history). The two tables below contain the descriptions of each of the predictor variables used in this project.

#### 

Description of Variables:

Quantitative:

| **Variable Name** | **Description** | **Mean**  **(Before Split)** | **Mean**  **(After Split)** |
| --- | --- | --- | --- |
| interest rate (response) | Interest rate of the loan the applicant received (in %) | 12.43 | 12.42 |
| annual\_income | Annual income. (in US dollars) | 79222 | 79762 |
| total\_credit\_lines | Total number of credit lines in this applicant's credit history. | 22.68 | 22.81 |
| total\_credit\_limit | Total available credit, e.g. if only credit cards, then the total of all the credit limits. | 183606 | 184677 |
| debt\_to\_income | Debt-to-income ratio. | 19.31 | 19.3 |
| installment | Monthly payment for the loan the applicant received. (in US dollars) | 476.21 | 476.24 |
| loan\_amount | The amount of the loan the applicant received. (in US dollars) | 16362 | 16368 |
| total\_credit\_utilized | Total credit balance, excluding a mortgage. | 51049 | 51281 |
| num\_total\_cc\_accounts | Total number of credit card accounts in the applicant's history. | 13.03 | 13.08 |
| num\_historical\_failed\_to\_pay | The number of derogatory public records, which roughly means the number of times the applicant failed to pay. | 0.17 | 0.17 |
| term | The number of months of the loan the applicant received. | 43.27 | 43.27 |

Categorical:

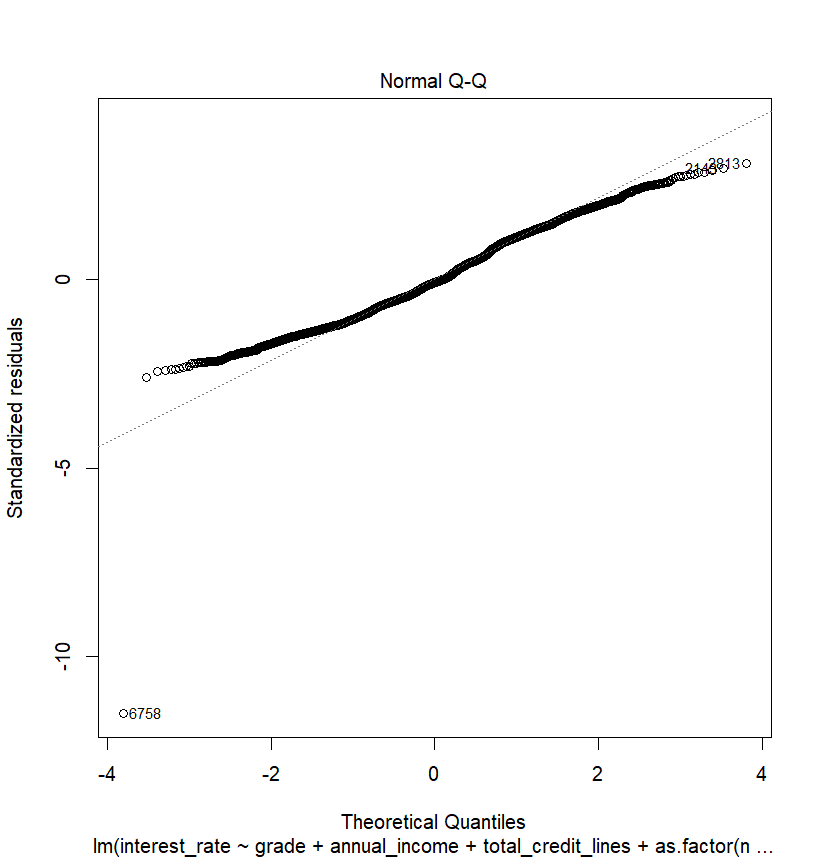
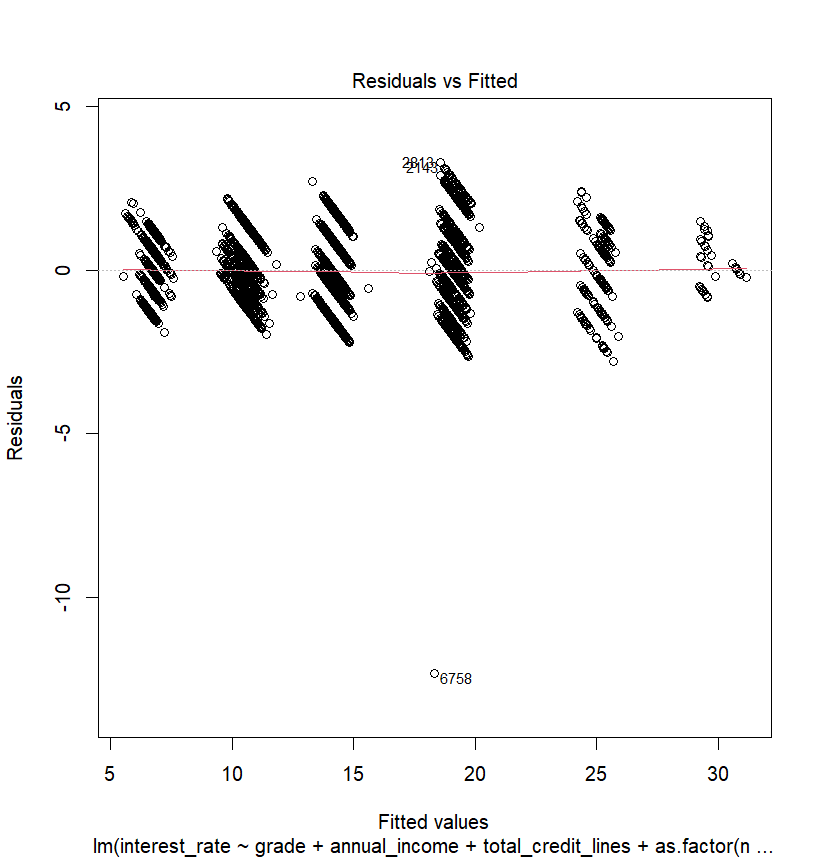
| **Variable Name** | **Description** |
| --- | --- |
| grade | Grade associated with the loan. |
| loan\_purpose | The category for the purpose of the loan. |
| application\_type | The type of application: either individual or joint. |
| homeownership | The ownership status of the applicant's residence. |
| public\_record\_bankrupt | The category for the purpose of the loan. |

#### 2a. Variable Selection

To check for any violation in the linear regression assumptions for the full multiple regression model, fitted with all 15 predictor variables we are interested in observing, we constructed the residual and QQ-plots to observe for any obvious anomaly. The model with the 15 predictor variables is as follow:

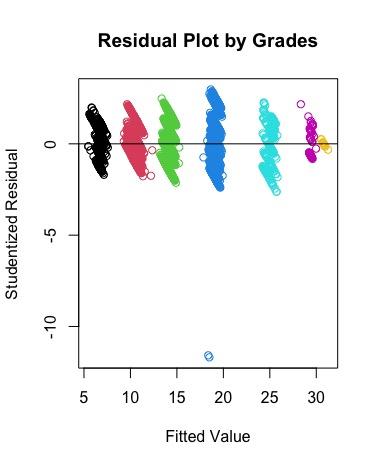
= **0** + **11** - **22** + **33** + **44** + **55** + **66** + **77** + **88** + **99** + **1010** + **1111** + **1212** + **1313** + **1414** + **1515** +

Please refer to Table 1 in Appendix A to find the definition corresponding to the xi’s in the equation above.

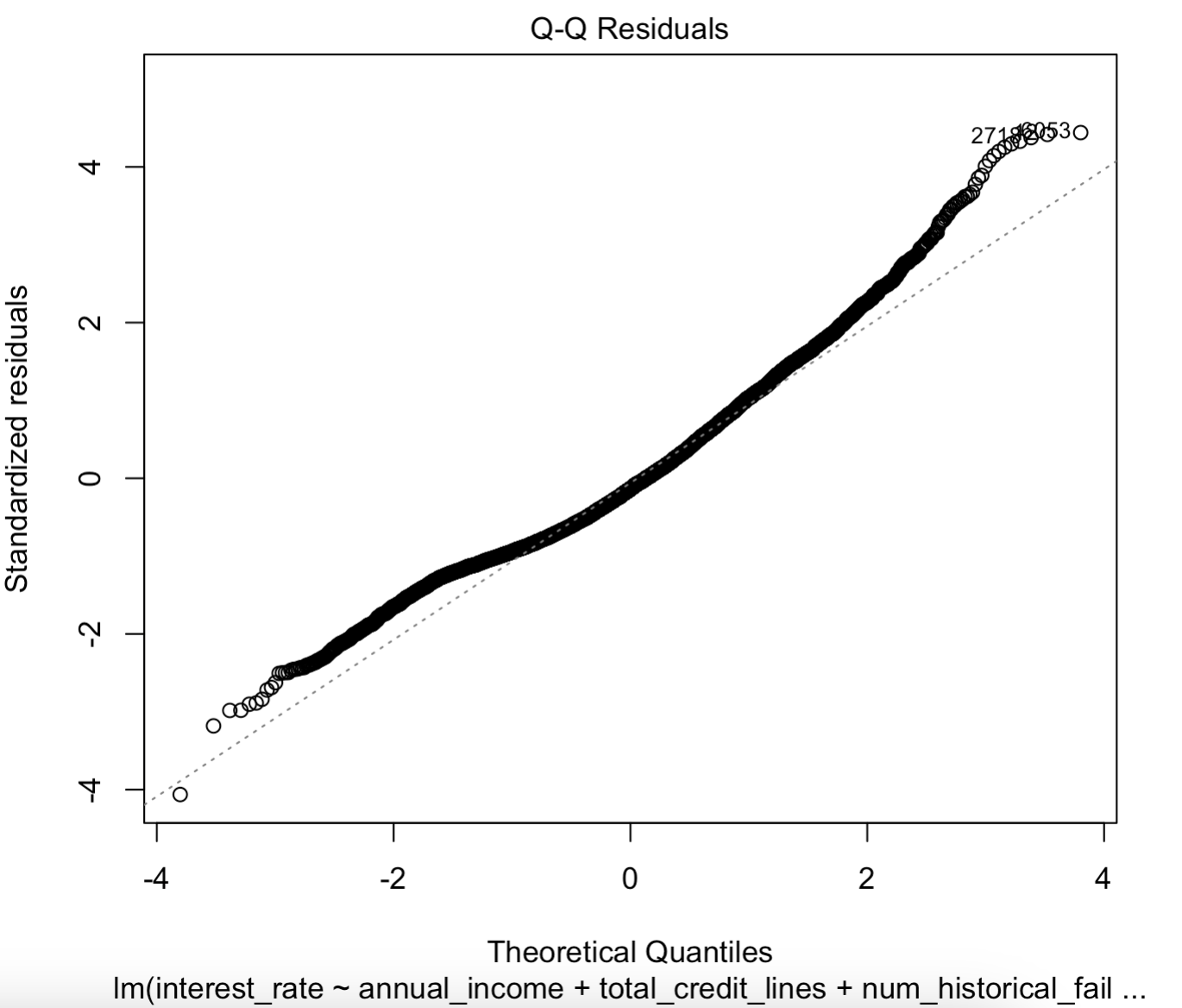
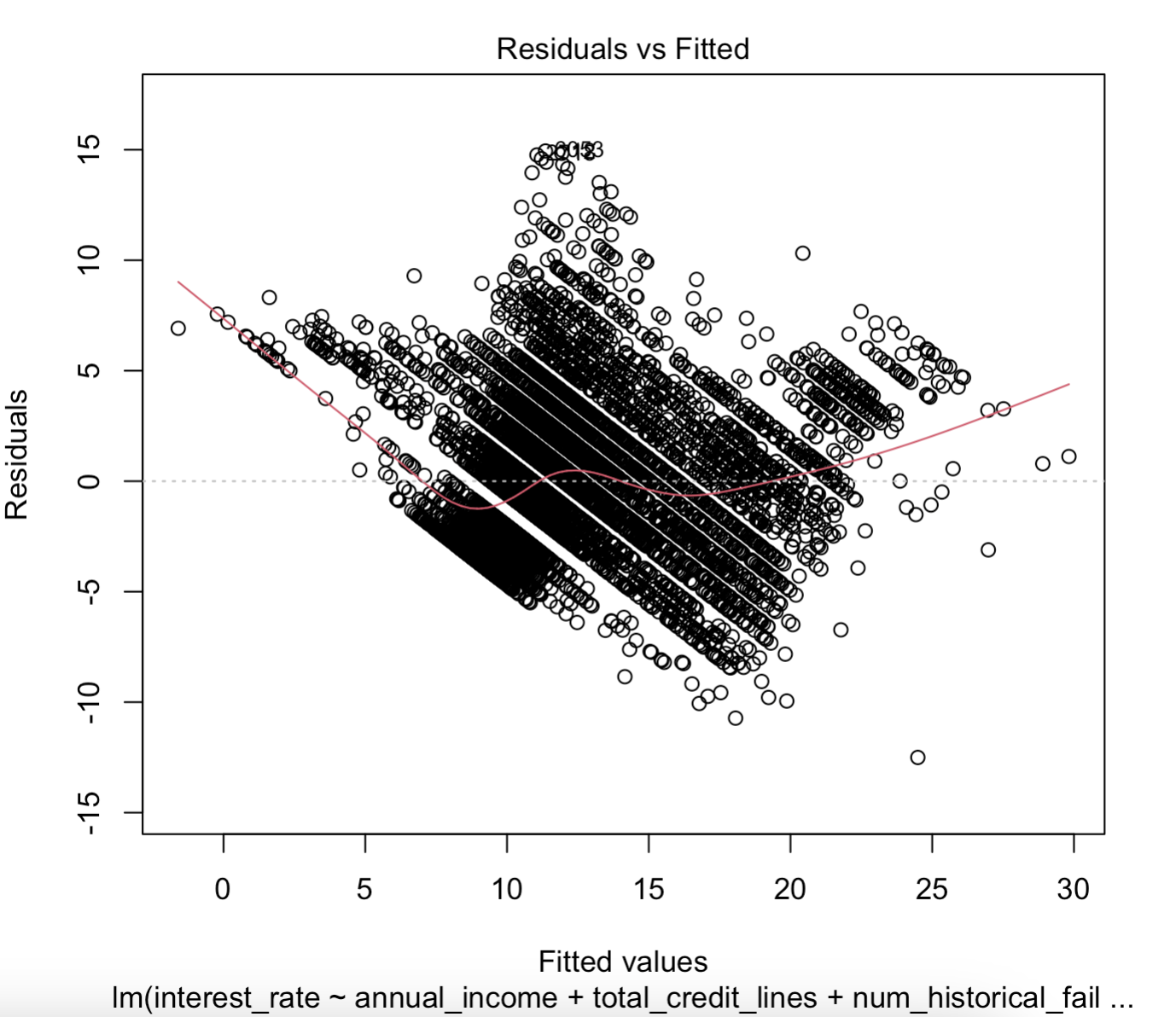


The QQ-plot shows that the data set is not following a normal distribution since the residuals are deviating away from the 45o line at both ends of the graph. The deviation from the line tells us that we should expect our distribution for the response variable to have long tails at both ends.

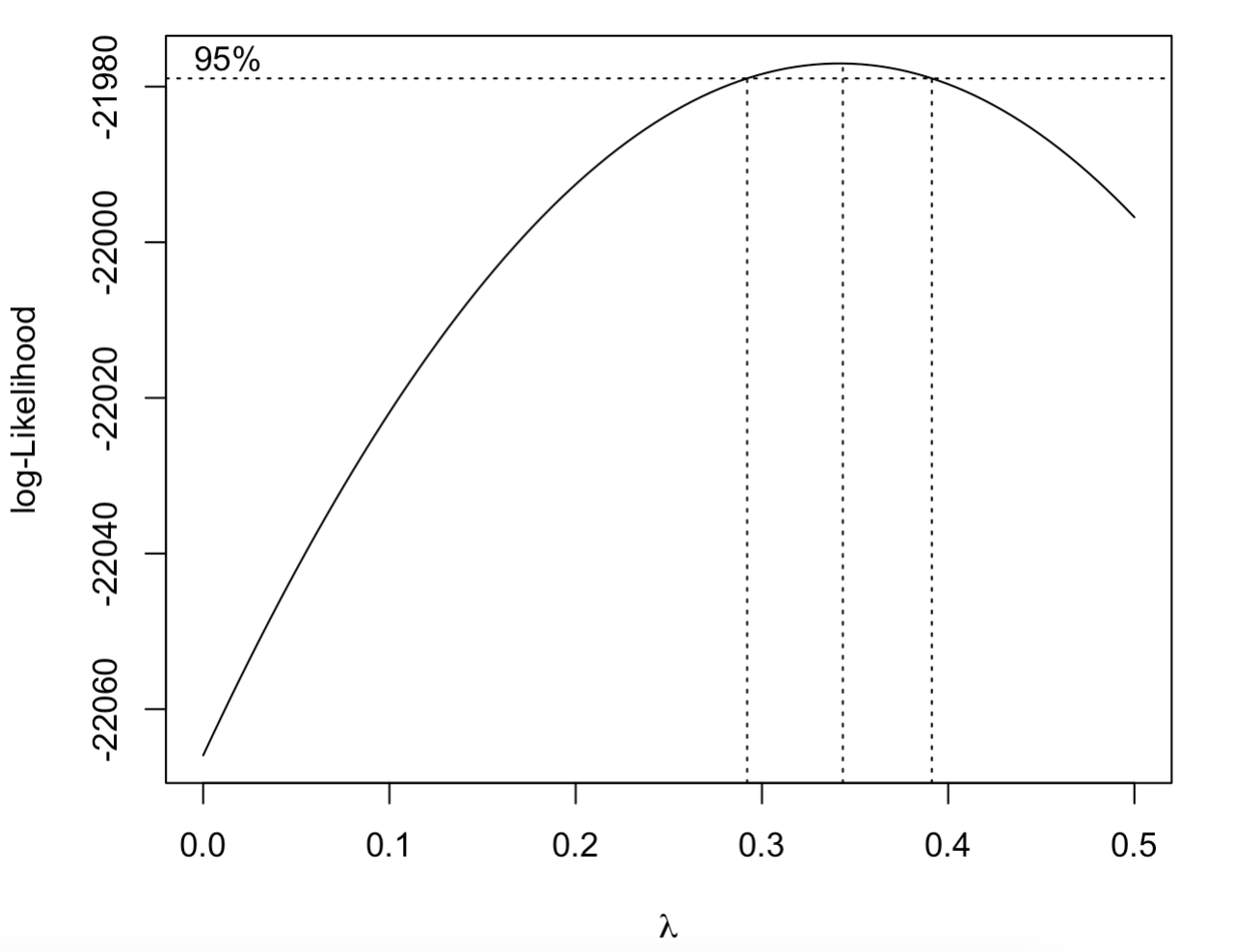
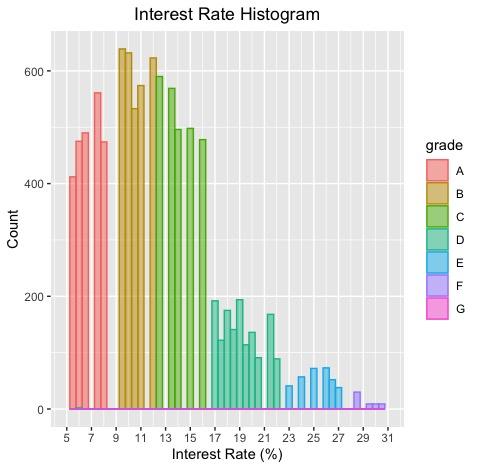
The residual plot shows the residuals forming six distincts groups; this could mean that one of the categorical predictors in the model is separating the response variable into non-overlapping sub-groups. To see which categorical predictor is responsible for this grouping effect, we color coded the residual plot based on the different levels for each of the categorical predictors. As shown in the color-coded residual plot below, the Grade predictor, with 6 levels: A, B, C, D, F, G, and the response variable are highly correlated with each other. The high correlation between the Grade predictor and the response variable means that we could correctly approximate the interest rate of a loan based on the grade assigned to it without needing any other information about the borrower. Due to the strong relationship between Grade and Interest Rate, we decided to remove Grade in order to determine which other predictors are important in determining interest rates.



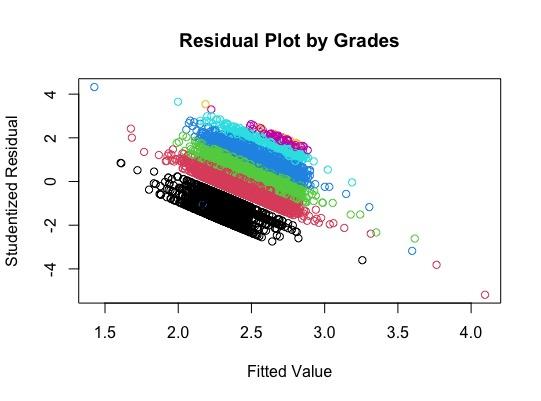
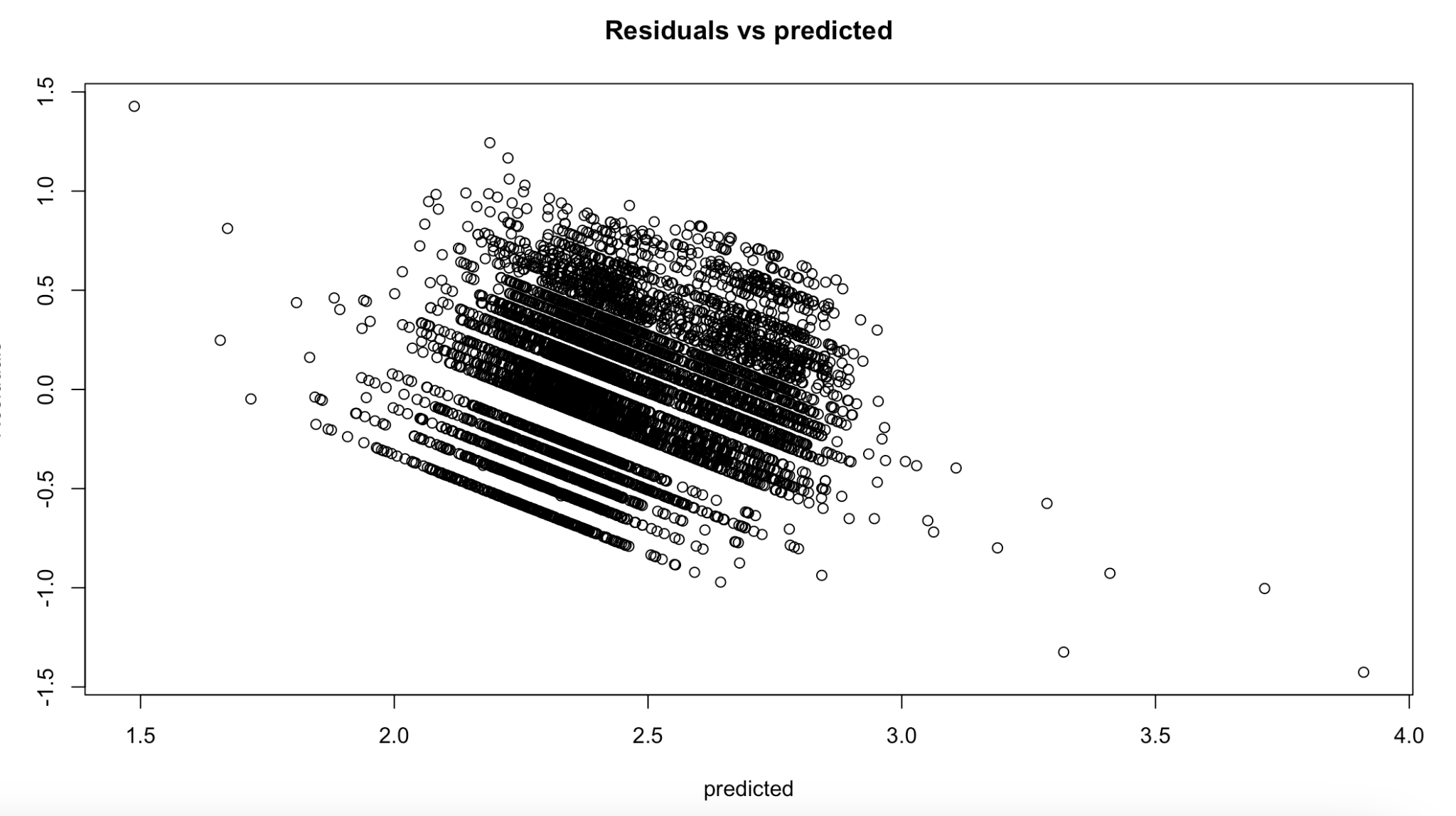
After refitting the model without the Grade predictor, we replotted residual and QQ plots to see if there are any improvements to our constant variance and normality assumptions. The residual plot does show an improvement in the constant variance assumption, as the residuals are no longer grouped into clusters based on the letter grade division. However, the constant variance assumption is still being violated as shown by the funneling patterns the residuals exhibit, with the residual values getting smaller as the fitted values become larger. The QQ-plot shows an improvement in the deviation at the lower end of the plot but also worsening of the deviation from the 45o line at the upper end. This could mean that there is a long tail toward the higher end of the distribution.



To understand where the skewness in the distribution is coming from, we constructed a histogram for the interest rate values and color-coded it by grades. The histogram below shows a skewness to the right in the distribution of interest rate, where the majority of the loans have an interest rate between 5 to16% while the rest are given an interest rate between 17 to 31%. This skewness to the right is affecting our normality assumption as shown in the deviation from the 45o line in the upper-right hand corner of the QQ-plot above. To address the skewness in the distribution of interest rates within the data, we decided to perform a transformation on the response variable to make the data conform more closely to the normal distribution.

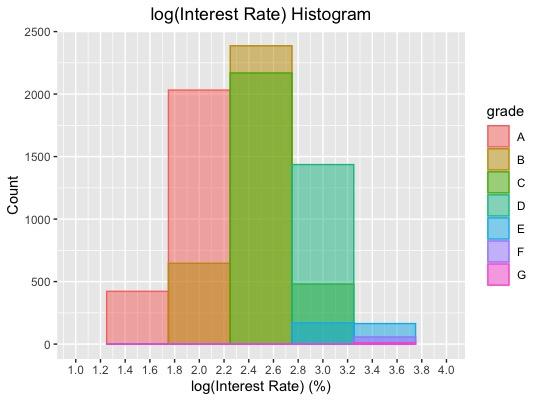
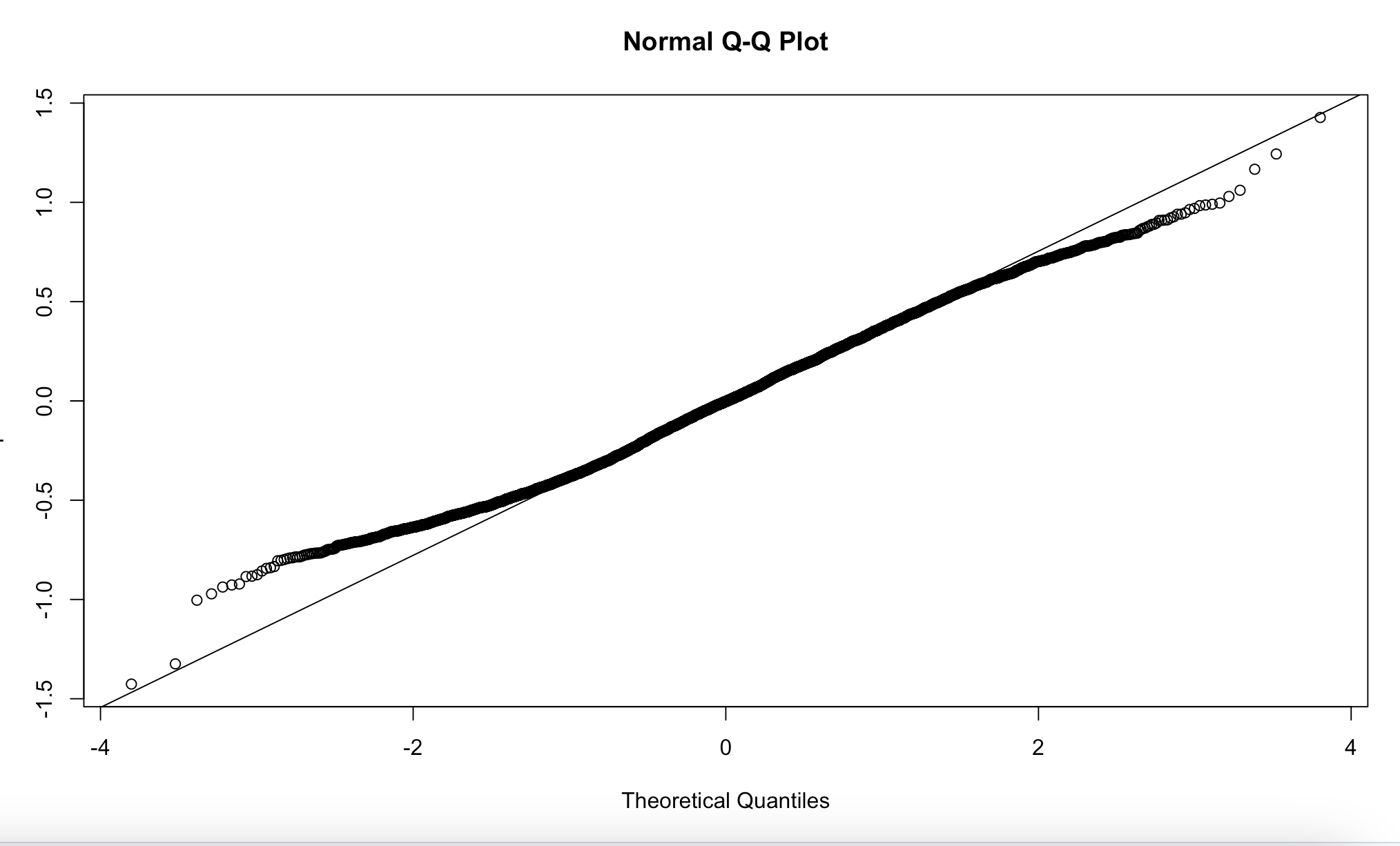


We first used the Box-Cox method to find the optimal transformation of the response variable, but decided against using the result given by the method since the confidence interval for the optimal lambda does not suggest an easily interpretable transformation. However, the Box-Cox plot above does show that the lambda values of 0 and within the vicinity of the 95% confidence interval for lambda, these lambda values are associated with the log and square-root transformation respectively. The log and square-root transformation are often used to reduce the skewness in the response variable (West). We decided that the logarithmic transformation was the “best” transformation since we want to normalize the positive skewness of the response variable.



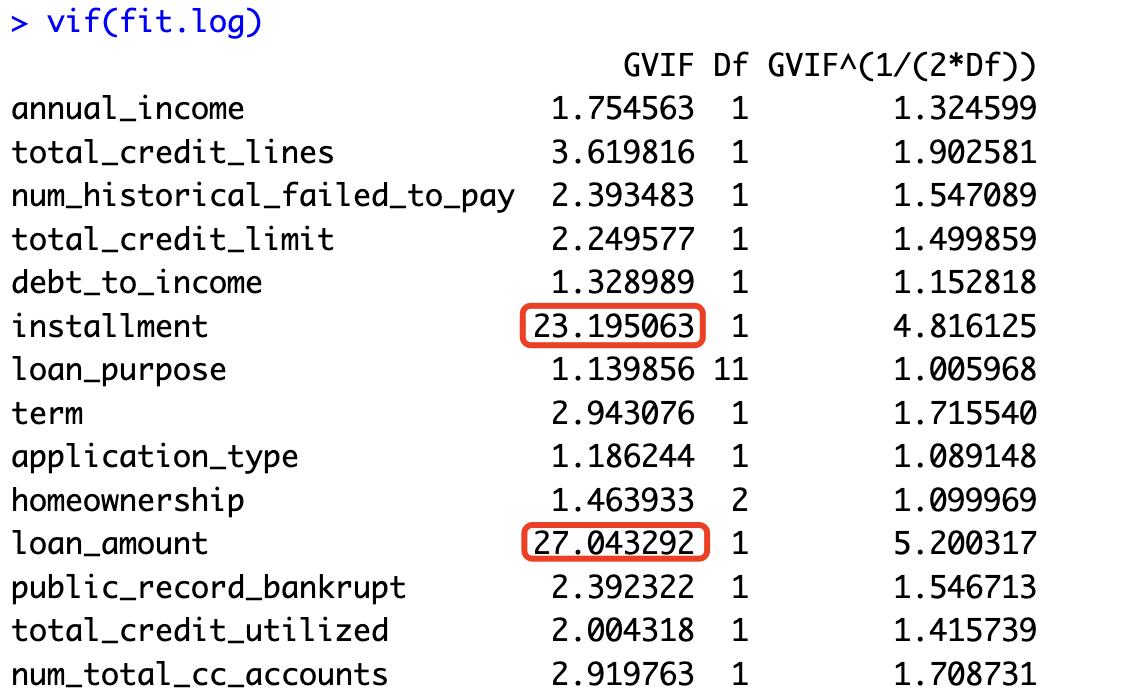
The residual plot of the fitted model for the logarithmic transformed response variables shows an improvement in the constant variance assumptions but the overall residual plot still exhibits the tilting pattern observed in the residual plot of the untransformed response variable. We attempted to address this tilting trend by performing transformations on the predictor variables but found that residual plot for the transformed response and predictor variables to closely resemble that residual plot for the untransformed response variable.

It should be noted that the residual plot color-coded by grade still shows an underlying separation of the residual by grade, since the interest rates calculated by the LendingClub is heavily influenced by the way its borrowers are categorized into 7 different loan grades. With borrowers who are categorized in Grade A to be the most likely to repay their loan and thus receive the lowest interest rate, contrastingly, borrowers who were given a G grading are seen to have the highest odds of defaulting on their loan and this penalized by given a high interest rate for their loan. As a result of this underlying relationship between grades and interest rate, we are unable to fix the tilting pattern seen in the residual plot.

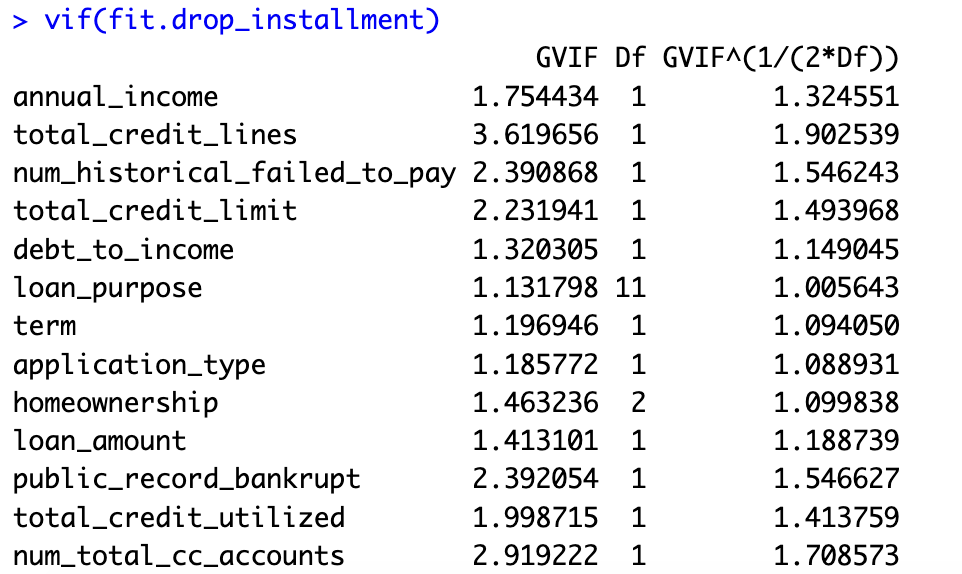


In addition, the QQ-plot for the transformed response variable model shows the residuals are deviating away from the 45o line in the lower-left corner of the plot, signifying that the distribution of interest rate is skewed to the left. However, when we check the histogram of the distribution for the logarithmic transformed interest rate we see that the data now closely follows the normal distribution compared to before. As a result, we decided to move forward with just transforming the response variable for our fitted model.

To check that the independence assumption is being met, we calculated the GVIF values of each of the predictor variables to observe if there is multicollinearity within the data. As shown in the figure below, there is a strong correlation between the predictors installment and loan amount. We decided to remove installment from our model since we know that the amount of installment for a loan is dependent on the loan amount that is being borrowed. In addition, it is observed that the standard error for the loan amount predictor in the “full” model decreased from 2.780e-04 to 8.512e-06.



We recalculated the GVIF values for each of the remaining predictor variables in our fitted model to check if there are still any linearly dependent variables. As seen in the figure below, the remaining predictor variables are now linearly independent of each other, thus satisfying the independence assumption for linear regression.



### 3. Model Selection Process

After performing variance stabilizing transformation on the response variable and doing variable selection on the predictor variables, we proceed forward with the model selection process. We decided to perform 3 different modeling selection methods: forward, backward, and exhaustive stepwise, to determine the “best” model. The following tables and figure shows the results of all three model selection methods:

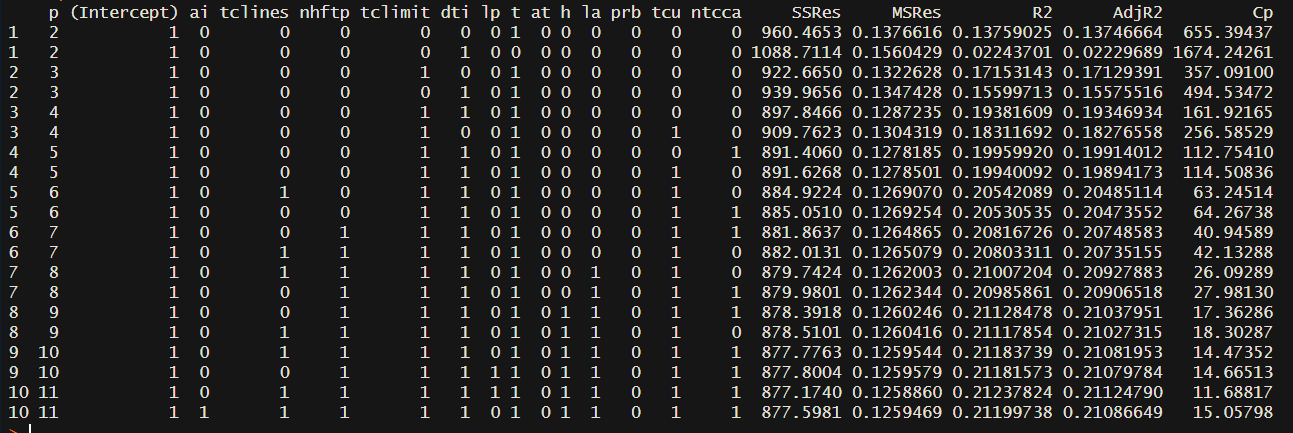
**Table: Forward Selection Steps**

| **Parameter Added** | **F-value** | **P-value** |
| --- | --- | --- |
| term | 1113.1219 | < 2.2e-16 |
| total\_credit\_limit | 285.7968 | < 2.2e-16 |
| debt\_to\_income | 192.8041 | < 2.2e-16 |
| loan\_purpose | 8.5298 | 3.817e-15 |
| total\_credit\_utilized | 51.1356 | 9.499e-13 |
| total\_credit\_lines | 52.0962 | 5.844e-13 |
| loan\_amount | 20.1723 | 7.191e-06 |
| num\_historical\_failed\_to\_pay | 3.8941 | 0.0003048 |
| homeownership | 4.7661 | 0.008541 |
| num\_total\_cc\_accounts | 6.1121 | 0.01345 |

**Table: Backwards Selection Steps**

| **Parameter Removed** | **F-value** | **P-Value** |
| --- | --- | --- |
| public\_record\_bankrupt | 1.6626 | 0.1727979 |
| annual\_income | 1.7606 | 0.1845949 |
| application\_type | 2.0710 | 0.1501637 |

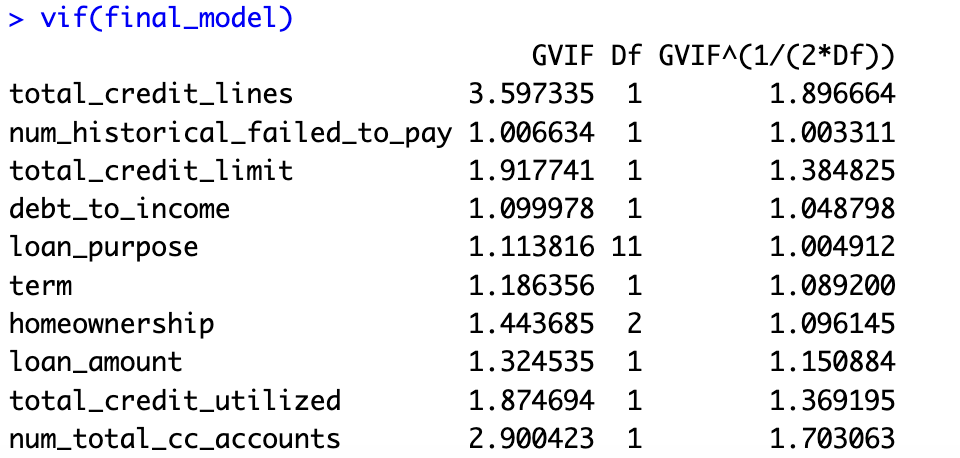
**Figure: Exhaustive Stepwise Method Outputs**



The forwards and backwards stepwise selection methods output the same model, which also happens to be the “best” model shown in the exhaustive stepwise selection method outputs. Thus, we will be using the following model to perform future observations prediction and variable inferences with:

= **0** + **11** - **22** + **33** + **44** + **55** + **66** + **77** + **88** + **99** + **1010** + **1111** + **1212** + **1313** + **1414** + **1515** + **1616** + **1717** + **1818** + **1919** + **2020** + **2121** +

We calculated the GVIF values for the predictor variables for the final model to make sure that none of the predictor variables are linearly dependent on each other, the figure below confirms that there is no multicollinearity problem affecting the final model.



As a result of the model selection process, our final model is given as follow:

= 1.844 + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** - (6.202e-2)**9** + (3.348e-2)**10** - (2.998e-2)**11** - (7.129e-2)**12** - (1.973e-2)**13** - (1.093e-2)**14** + (1.439e-2)**15** + (2.777e-3)**16** + (4.168e-2)**17** + (4.203e-2)**18** + (1.313e-1)**19** + (2.439e-2)**20** + (3.375e-2)**21** +

Please refer to table 2 in the Appendix A to get the definition of xi’s shown in the equation above.

#### 3.2 Outlier Detected

After finalizing our model, we revisited the data to identify potential outliers. Our first focus was on leverage points, identified using the diagonal of the hat matrix, revealing 571 such points in the dataset. To assess their impact, we fitted the model both with and without these leverage points. The comparison, as shown in Table 6, indicates minimal change in the model when excluding these points.

Additionally, we explored the presence of influential points that could significantly bias our model by applying Cook's distance. This analysis confirmed that our dataset does not contain any influential points that would notably skew the model's results.

### 4. Findings

One of our purposes in fitting this model is to determine how accurately we would be able to predict future interest rates with a given set of predictors. To measure the capabilities of our final model to accomplish this, we are looking at the PRESS statistic and the R2 Prediction value. The PRESS statistic for our final model is 873.2984, which is fairly high. Our R2 Prediction value is 0.2158, which is quite low. Neither of these are very surprising, especially after the removal of the predictor Grade, which essentially predicted the interest rate all on its own. The R2 and adjusted R2 for our final model are quite low. Our R2 prediction value shows that there is high variability in our model when attempting to predict new observations, and thus is not very useful to do so.

Our second purpose in this project is to analyze which predictors are the most useful in determining the interest rate for each borrower. Through analyzing residual plots, using model selection processes, and viewing the summary statistics of different models, we have narrowed down the number of predictors. This is, of course, aside from the predictors grade and installment, as grade is very highly correlated to the response and the trends of installment are are also shown in the loan amount. The predictors we determined are the most useful are total\_credit\_lines, num\_historical\_failed\_to\_pay, total\_credit\_limit, debt\_to\_income, loan\_purpose, term, homeownership, loan\_amount, total\_credit\_utilized, and num\_total\_cc\_accounts. Aside from some of the levels of the categorical variables, all of our predictors are significant contributors to interest rate, as shown in Table 7. When comparing the predictors contained in our final model, we find that they are closely related to factors that traditional financial institutions commonly use to determine the interest rate of a loan, such as: type of loans, employment history and income, loan size and term, credit history, and other debt (Waugh). It should also be noted that the final model also contains debt-to-income ratio, credit history, and loan amount, which are predictors that the Lending Club have mentioned on its website that it uses to base its APR calculation on (LendingClub).

Looking at the estimators in the final model, we see that predictors such as terms, number of historical failed to pay, and loan purpose have the most influence in determining the average increase in interest rate of a loan compared to the other predictors. It is observed that the number of historical failed to pay predictors to be the continuous variable that causes the highest average increase in interest rate. Interest rate increases on average by e0.04346 or 1.044% for every additional number of times the borrowers with a mortgage failed to repay their loan given that all other predictors are held constant. When we look at the contrast between the different loan purposes, we found that the biggest difference in the average interest rate happens between a loan taken out for the purpose of repaying a credit card debt and loan taken out to finance a vacation. A borrower with a mortgage who takes out a loan to repay their credit card debt will have on average 2.22% lower in interest rate compared to a borrower who took out a loan to finance their vacation. Overall, the final model was able to give us an overview of the effect and strength that each of the predictors has on determining the interest rate of a loan.

## Appendix A

**Table 1: Full Model Dictionary**

| **Variable Name** | **Variable** | **Variable Name** | **Variable** | **Variable Name** | **Variable** |
| --- | --- | --- | --- | --- | --- |
| grade | x1 | debt to income | x6 | homeownership | x11 |
| annual income | x2 | installment | x7 | Loan amount | x12 |
| total credit lines | x3 | loan purpose | x8 | Public record bankrupt | x13 |
| number historical failed to pay | x4 | term | x9 | Total credit utilized | x14 |
| total credit limit | x5 | application type | x10 | Number of total credit card account | x15 |

**Table 2: Final Model Dictionary**

| **Variable Name** | **Variable** | **Variable Name** | **Variable** | **Variable Name** | **Variable** |
| --- | --- | --- | --- | --- | --- |
| term | x1 | Number of historical failed to pay | x8 | Loan Purpose: moving | x15 |
| Total credit limit | x2 | Loan purpose: credit card | x9 | Loan Purpose: other | x16 |
| Debt-to-Income | x3 | Loan Purpose: debt consolidation | x10 | Loan Purpose: renewable energy | x17 |
| Total credit utilized | x4 | Loan Purpose: home improvement | x11 | Loan Purpose: small business | x18 |
| Total credit lines | x5 | Loan Purpose: house | x12 | Loan Purpose: vacation | x19 |
| Loan amount | x6 | Loan Purpose: major purchase | x13 | Homeownership: Own | x20 |
| Number total credit card account | x7 | Loan Purpose: medical | x14 | Homeownership: Rent | x21 |

**Table 3: Contrast for Homeownership Variable**

| **x20** | **x21** |  |
| --- | --- | --- |
| 0 | 0 | If the borrower has a mortgage |
| 1 | 0 | If the borrower owns a house |
| 0 | 1 | If the borrower is a renter |

**Table 4: Contrast for Loan Purpose variable**

| **x9** | **x10** | **x11** | **x12** | **x13** | **x14** | **x15** | **x16** | **x17** | **x18** | **x19** |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | If loan purpose is wedding |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | If loan purpose is credit card |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | If loan purpose is debt consolidation |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | If loans purpose is home improvement |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | If loan purpose is buy a house |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | If loan purpose is a major purchase |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | If loan purchase is medical reason |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | If loan purchase is moving |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | If loan purchase is other |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | If loan purchase is renewable energy |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | If loan purchase is small business |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | If loan purchase is vacation |

**Table 5: Model Equations for Varying Categorical Variables**

|  | Borrowers With a Mortgage |
| --- | --- |
| Loan Purpose | Equation |
| Wedding | = 1.844 + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Credit Card | = (1.844 - 6.202e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Debt Consolidation | = (1.844 + 3.348e-2 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Home Improvement | = (1.844 -2.998e-2 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Buying a House | = (1.844 -7.129e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Major Purchase | = (1.844 -1.973e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Medical Reason | = (1.844 -1.093e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Moving Expense | = (1.844 + 1.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Other | = (1.844 + 2.777e-3 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Renewable Energy | = (1.844 +4.168e-2 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Small Business | = (1.844 +4.203e-2 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Vacation Expense | = (1.844 + 1.313e-1 ) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |

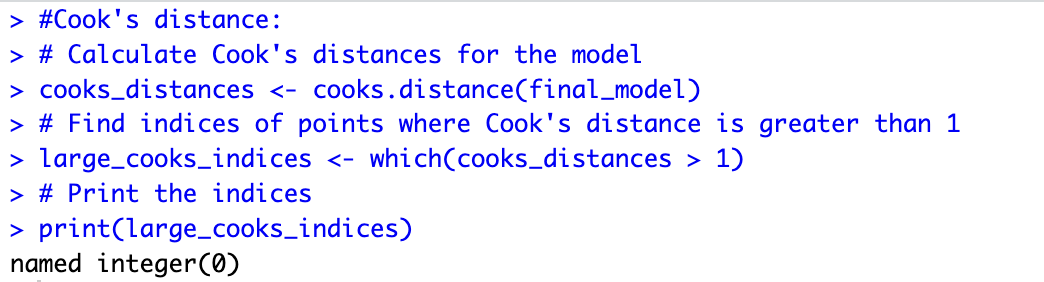
|  | Borrowers that are Homeowners |
| --- | --- |
| Loan Purpose | Equation |
| Wedding | = (1.844 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Credit Card | = (1.844 - 6.202e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Debt Consolidation | = (1.844 + 3.348e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Home Improvement | = (1.844 -2.998e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Buying a House | = (1.844 -7.129e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Major Purchase | = (1.844 -1.973e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Medical Reason | = (1.844 -1.093e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Moving Expense | = (1.844 + 1.439e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Other | = (1.844 + 2.777e-3 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Renewable Energy | = (1.844 +4.168e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Small Business | = (1.844 +4.203e-2 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Vacation Expense | = (1.844 + 1.313e-1 + 2.439e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |

|  | Borrowers that are Renters |
| --- | --- |
| Loan Purpose | Equation |
| Wedding | = (1.844 + 3.375e-) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Credit Card | = (1.844 - 6.202e-2 + 3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Debt Consolidation | = (1.844 + 3.348e-2 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Home Improvement | = (1.844 -2.998e-2 + 3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Buying a House | = (1.844 -7.129e-2 3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Major Purchase | = (1.844 -1.973e-2 + 3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Medical Reason | = (1.844 -1.093e-2 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Moving Expense | = (1.844 + 1.439e-2 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Other | = (1.844 + 2.777e-3 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Renewable Energy | = (1.844 +4.168e-2 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Small Business | = (1.844 +4.203e-2 +3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |
| Vacation Expense | = (1.844 + 1.313e-1 + 3.375e-2) + (1.449e-2 )**1** - (3.964e-7)**2** + (3.847e-3)**3** + (7.740e-7)**4** - (1.470e-3)**5** - (1.995e-6)**6** - (2.240e-3)**7** + (4.346e-2)**8** |

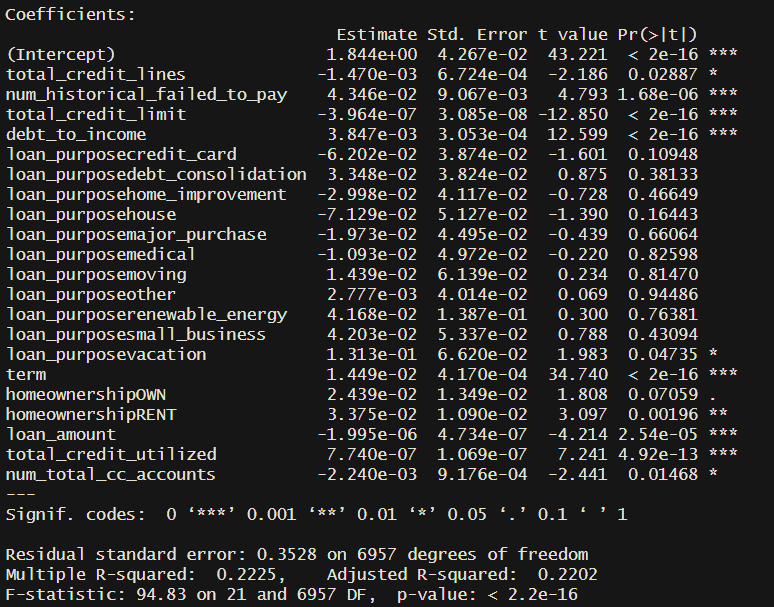
**Table 6: Outlier Analysis**

| Model Comparison: Analyzing the influence of including or excluding certain variables on the model's performance. | | |
| --- | --- | --- |
|
|  | Model without leverage points | Model with leverage points |
| Coefficient: |  |  |
| Intercept | 1.856 | 1.844 |
| term | 1.428e-02 | 1.449e-02 |
| Total credit limit | -4.250e-07 | -3.964e-07 |
| Debt-to-Income | 4.306e-03 | 3.847e-03 |
| Total credit utilized | 7.957e-07 | 7.740e-07 |
| Total credit lines | -2.188e-03 | -1.470e-03 |
| Loan amount | -2.117e-06 | -1.995e-06 |
| Number total credit card account | -1.568e-03 | -2.240e-03 |
| Number of historical failed to pay | 4.338e-02 | 4.346e-02 |
| Loan purpose: credit card | -4.332e-02 | -6.202e-02 |
| Loan Purpose: debt consolidation | 5.090e-02 | 3.348e-02 |
| Loan Purpose: home improvement | -2.185e-02 | -2.998e-02 |
| Loan Purpose: house | -5.306e-02 | -7.129e-02 |
| Loan Purpose: major purchase | 1.545e-02 | -1.973e-02 |
| Loan Purpose: medical | 1.192e-03 | -1.093e-02 |
| Loan Purpose: moving | 5.788e-02 | 1.439e-02 |
| Loan Purpose: other | 2.907e-02 | 2.777e-03 |
| Loan Purpose: renewable energy | 8.898e-02 | 4.168e-02 |
| Loan Purpose: small business | 7.154e-02 | 4.203e-02 |
| Loan Purpose: vacation | 1.639e-01 | 1.313e-01 |
| R^2 | 0.221 | 0.2225 |
| Adjust R^2 | 0.2187 | 0.2202 |
| SE | 0.3526 | 0.3528 |

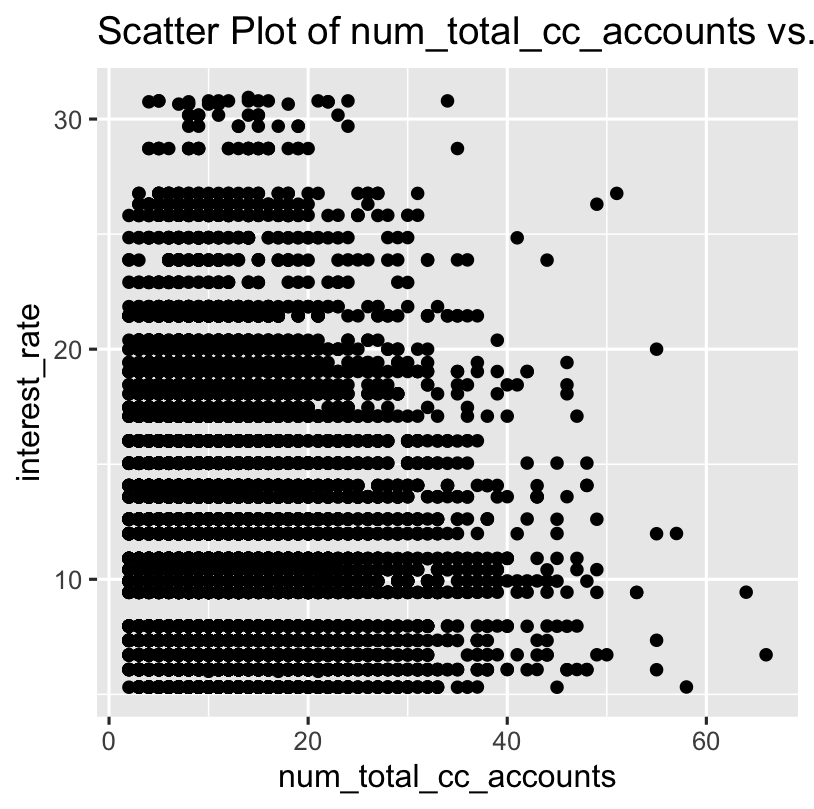
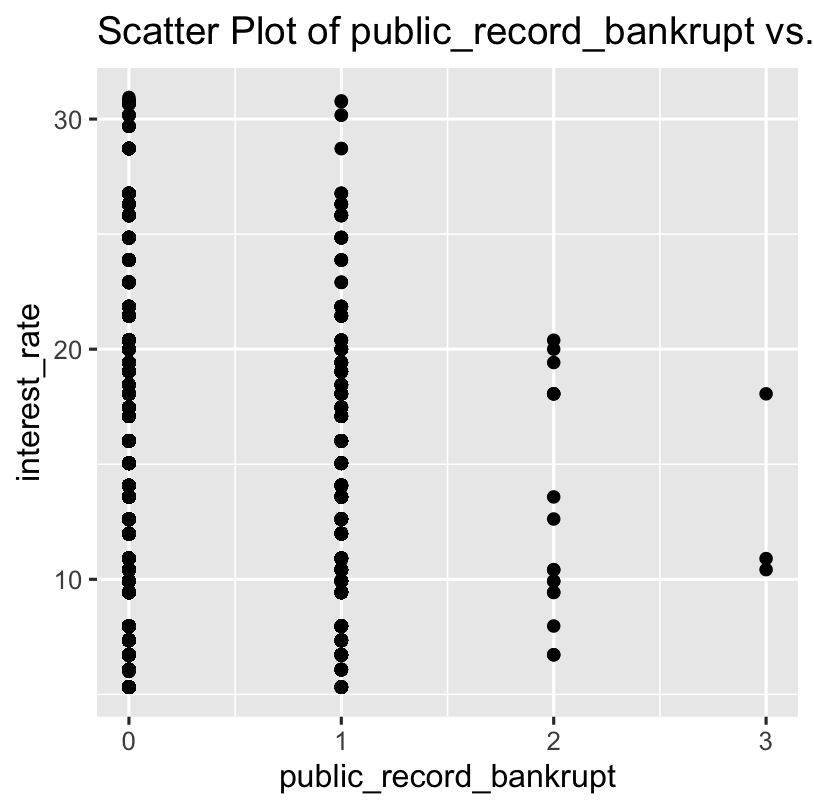
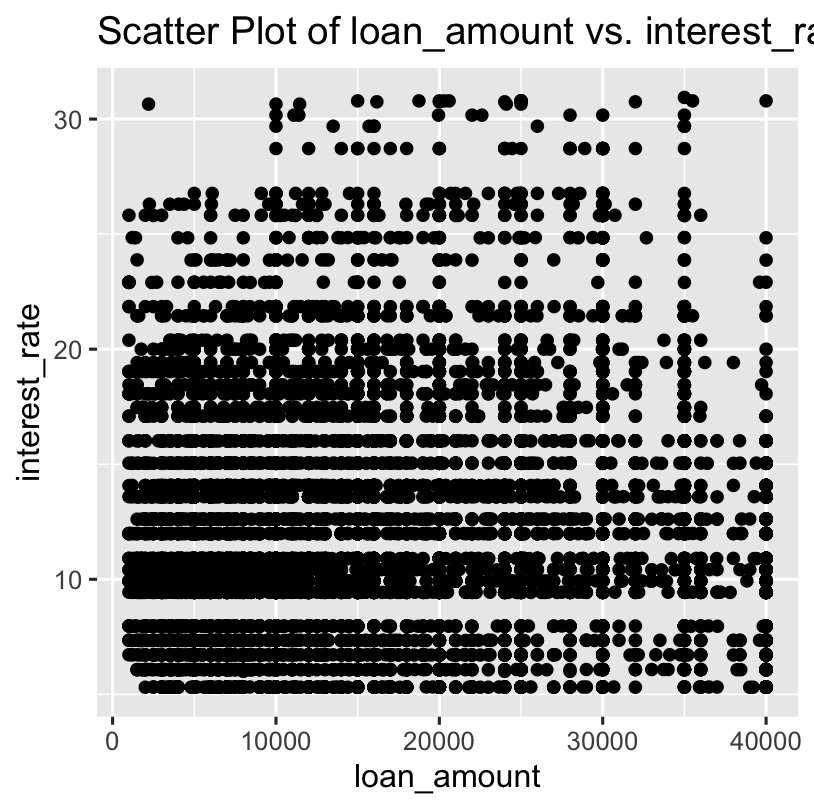
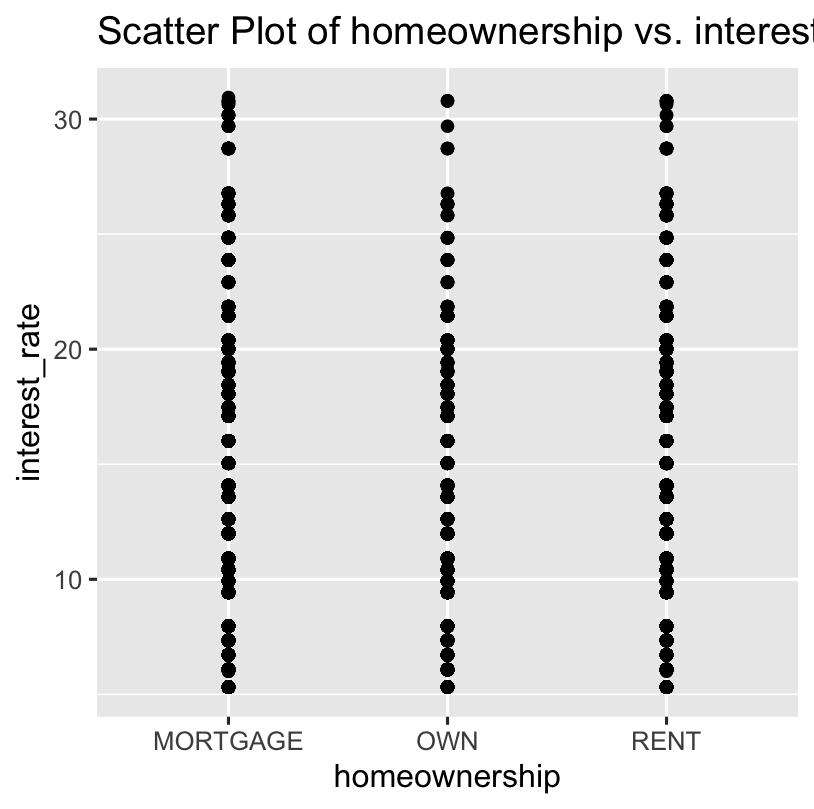
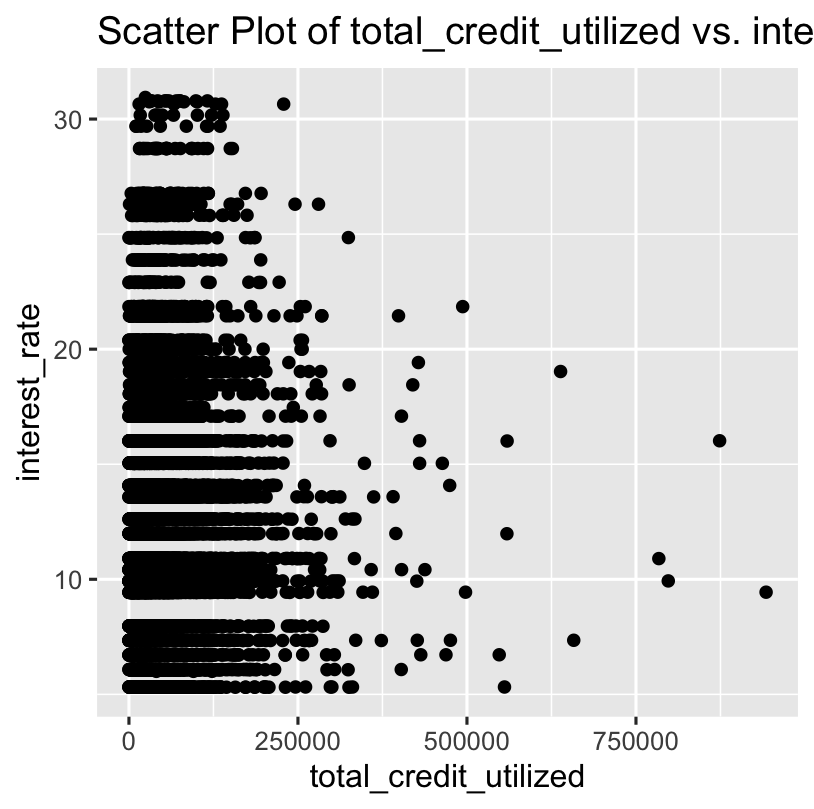
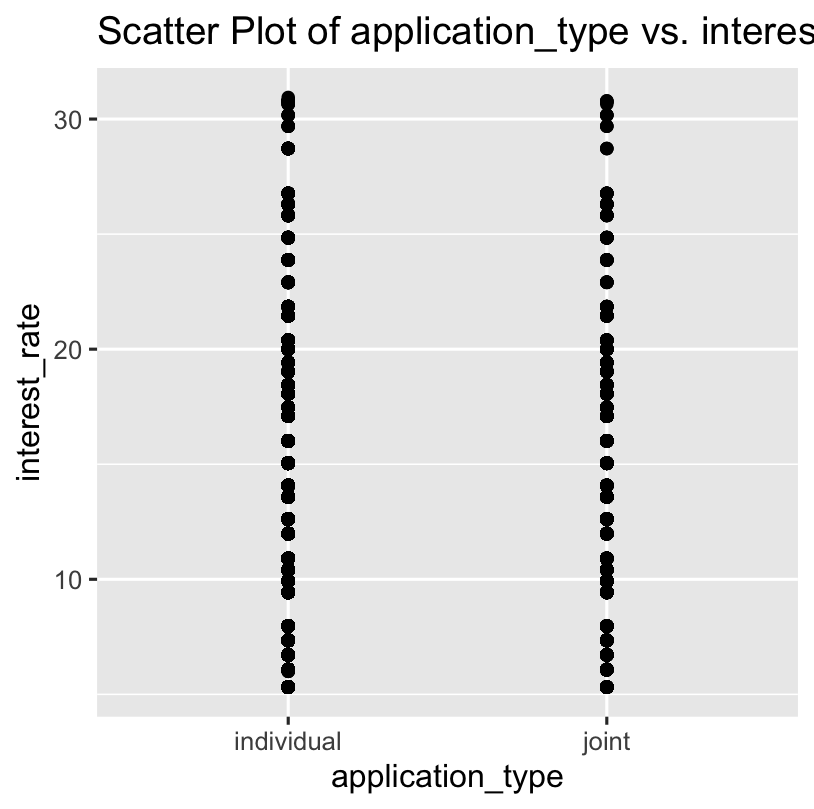
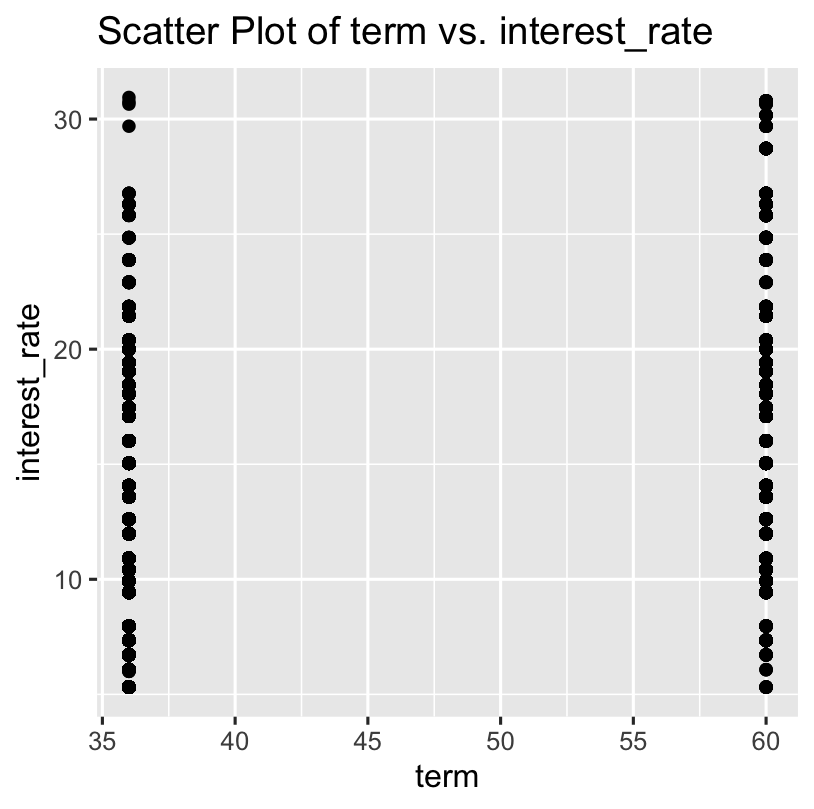
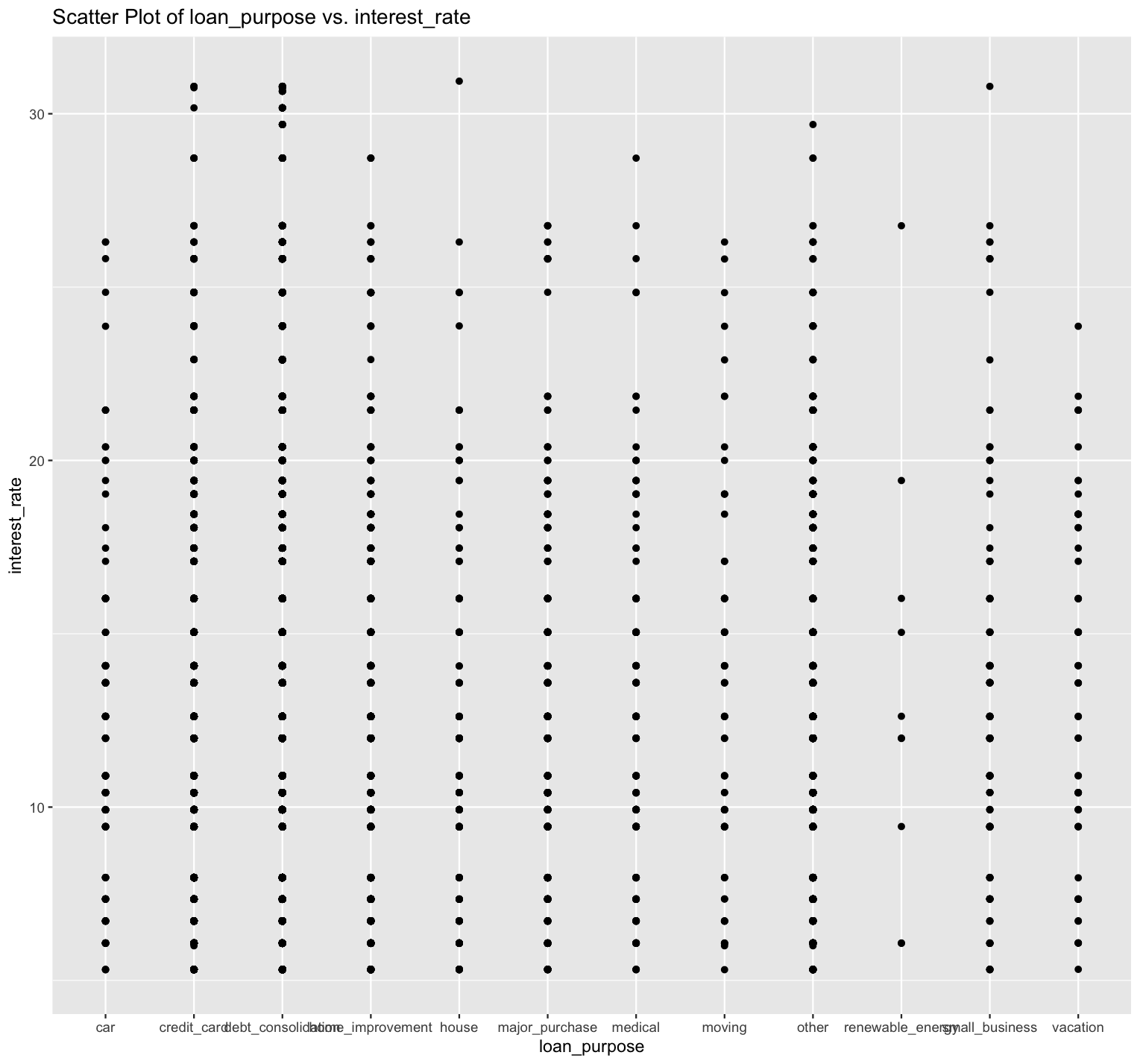
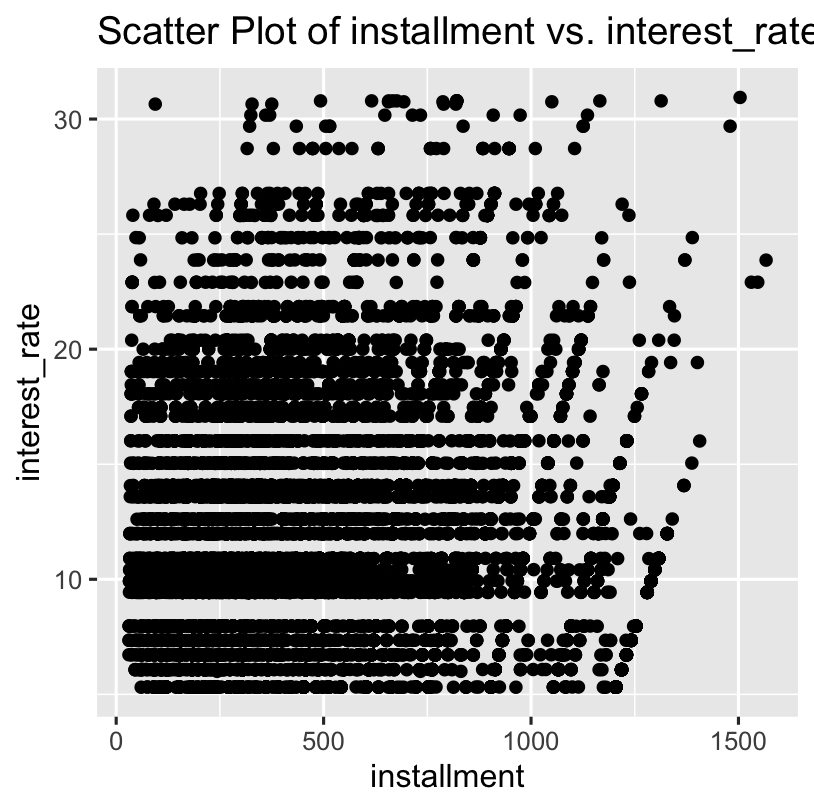
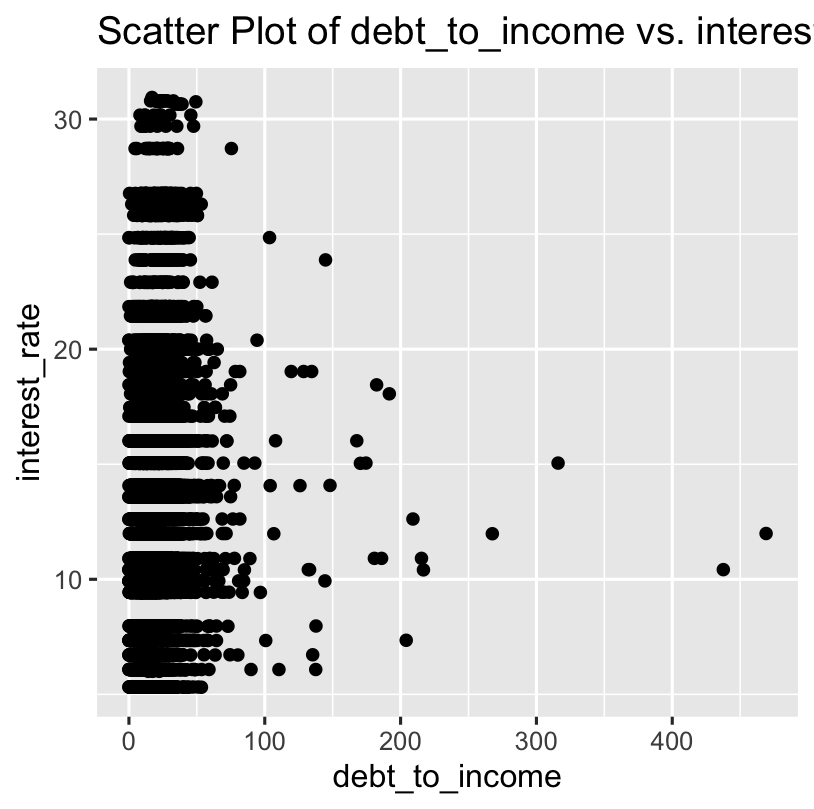
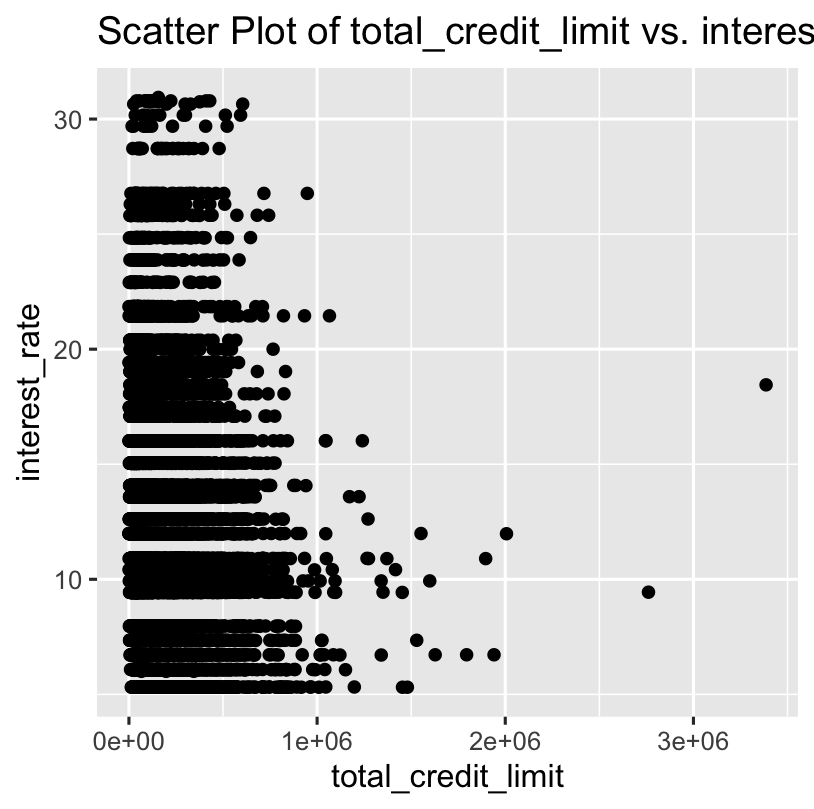
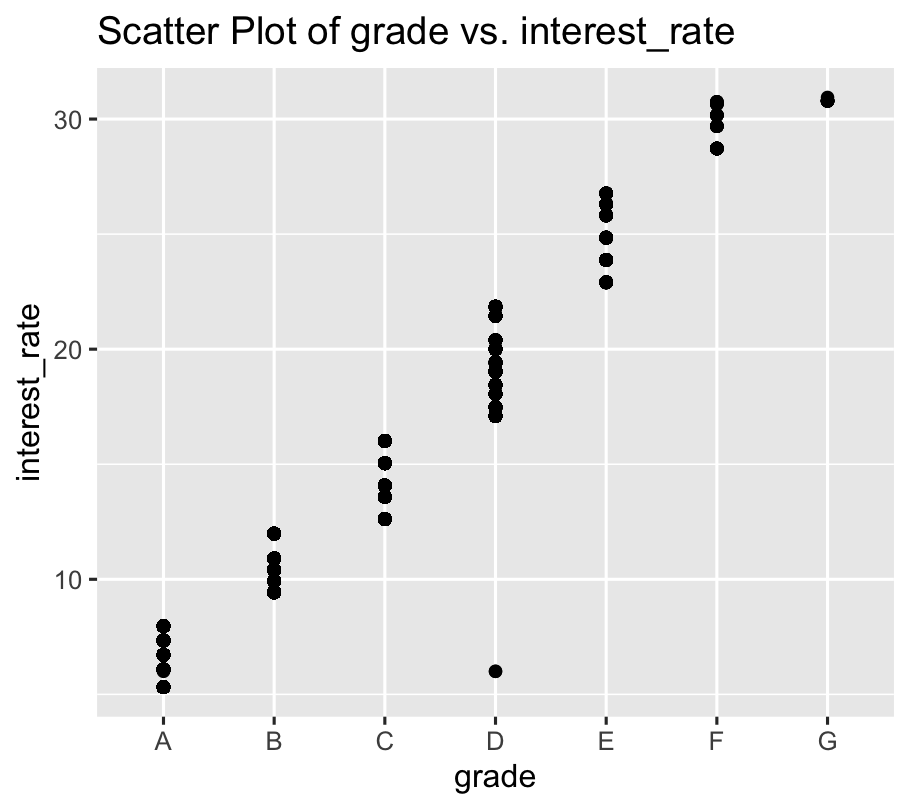
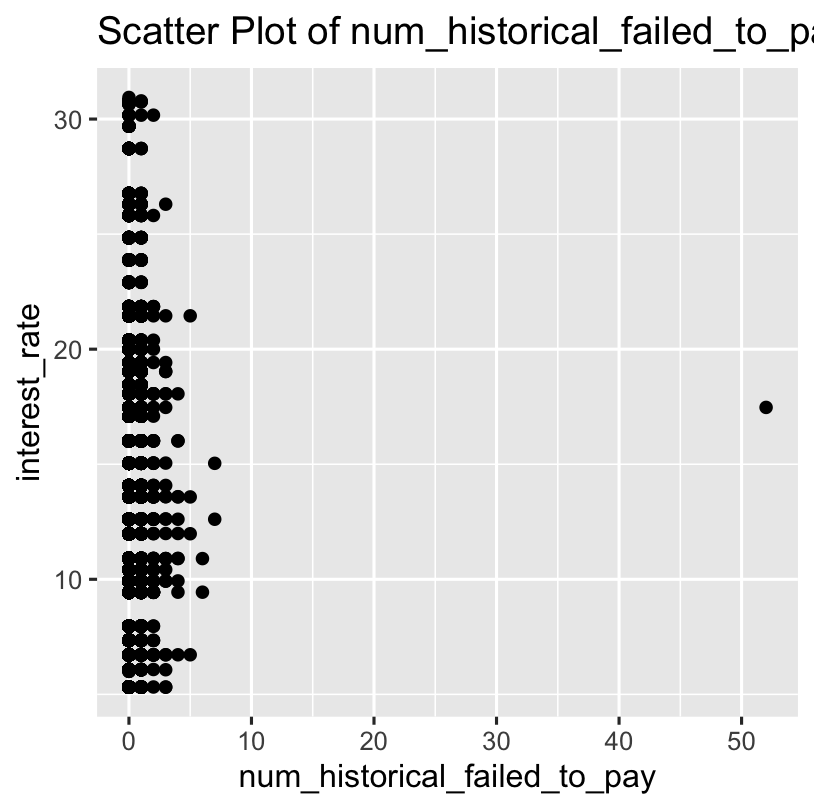
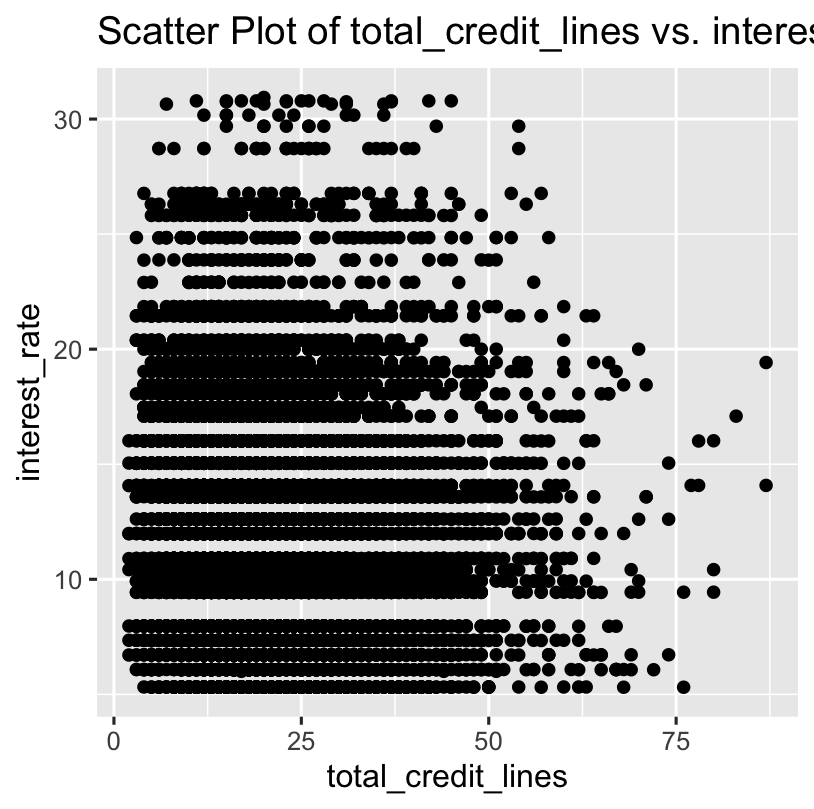
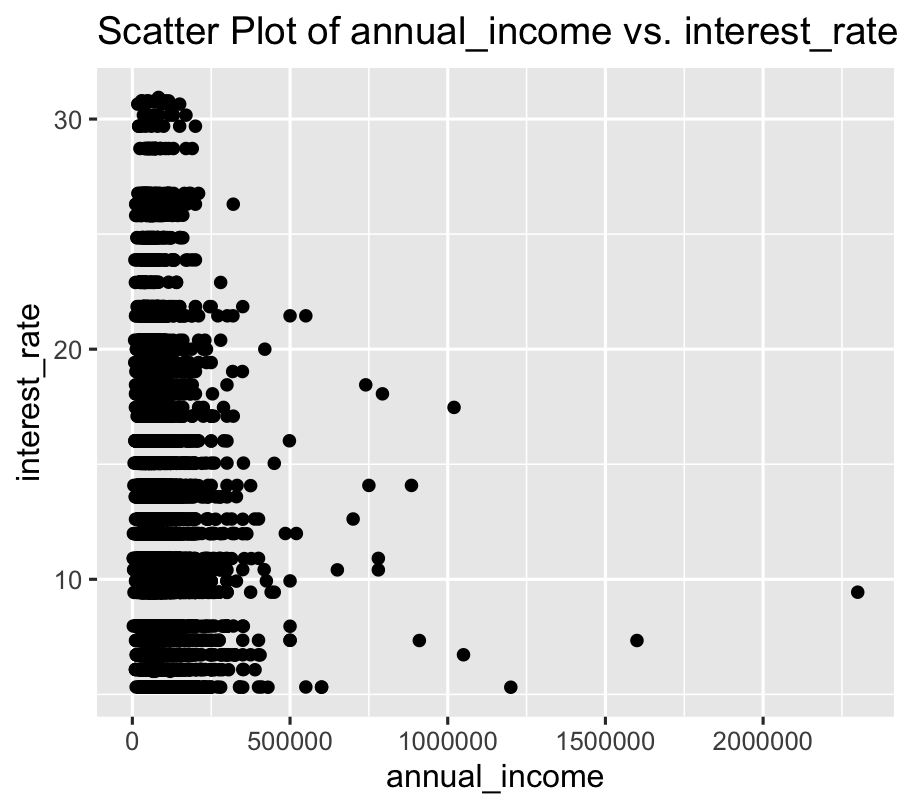
**Table 7: Cook’s distance result**



**Table 8: R Summary Output for Final Model**



**Table 9: Scatterplots of Predictor Variables vs Interest Rate**



Works Cited

Canandaigua National Bank & Trust. *How Banks Limit Risk in Commercial Lending*. 2023. *Canandaigua National Bank & Trust*, https://www.cnbank.com/Your\_Bank/Education\_and\_Advice/CNBU\_Articles/How\_Banks\_Limit\_Risk\_in\_Commercial\_Lending/. Accessed 12 November 2023.

LendingClub. “Personal Loans Rates & Fees.” *Lending Club*, https://www.lendingclub.com/personal-loan/rates-fees#. Accessed 2 December 2023.

OpenIntro. “Loan data from Lending Club.” *OpenIntro*, 2019, https://www.openintro.org/data/index.php?data=loans\_full\_schema. Accessed 12 November 2023.

Suknanan, Jasmin. “3 Best Peer-To-Peer Personal Loans To Consider In 2023.” *CNBC*, 9 June 2023, https://www.cnbc.com/select/best-peer-to-peer-personal-loans/. Accessed 12 November 2023.

Treece, Kiah. “Best Peer-To-Peer Personal Loans Of 2023 – Forbes Advisor.” *Forbes*, 1 November 2023, https://www.forbes.com/advisor/personal-loans/best-peer-to-peer-lending/. Accessed 12 November 2023.

Waugh, Evelyn. “What Factors Do Lenders Consider When Determining My Interest Rate?” *Experian*, 3 October 2022, https://www.experian.com/blogs/ask-experian/what-factors-do-lenders-consider-when-determining-my-interest-rate/. Accessed 2 December 2023.

West, Robert M. “Best practice in statistics: The use of log transformation.” *NCBI*, SAGE, 19 October 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9036143/. Accessed 3 December 2023.