*This is the main submission document****. Save and rename this document filename with your registered full name as Prefix before submission.***

|  |  |
| --- | --- |
| Class / Seminar Grp | Seminar 4 |
| Full Name | Ng Jun Long |
| Matriculation Number | U2110010D |

*\* : Delete and replace as appropriate.*

**Declaration of Academic Integrity**

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University’s Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

*Please insert an “X” within the square brackets below to indicate your selection.*

**[ ] I have read and accept the above.**



Table of Contents

[Answer to Q1: 3](#_Toc117455766)

[Answer to Q1a): 3](#_Toc117455767)

[Answer to Q1b): 5](#_Toc117455768)

[Answer to Q2): 7](#_Toc117455769)

[Answer to Q3: 11](#_Toc117455770)

[Answer to Q3a): 11](#_Toc117455771)

[Answer to Q3b): 11](#_Toc117455772)

[Answer to 3c): 18](#_Toc117455773)

[Answer to Q3d): 19](#_Toc117455774)

[Answer to Q4: 20](#_Toc117455775)

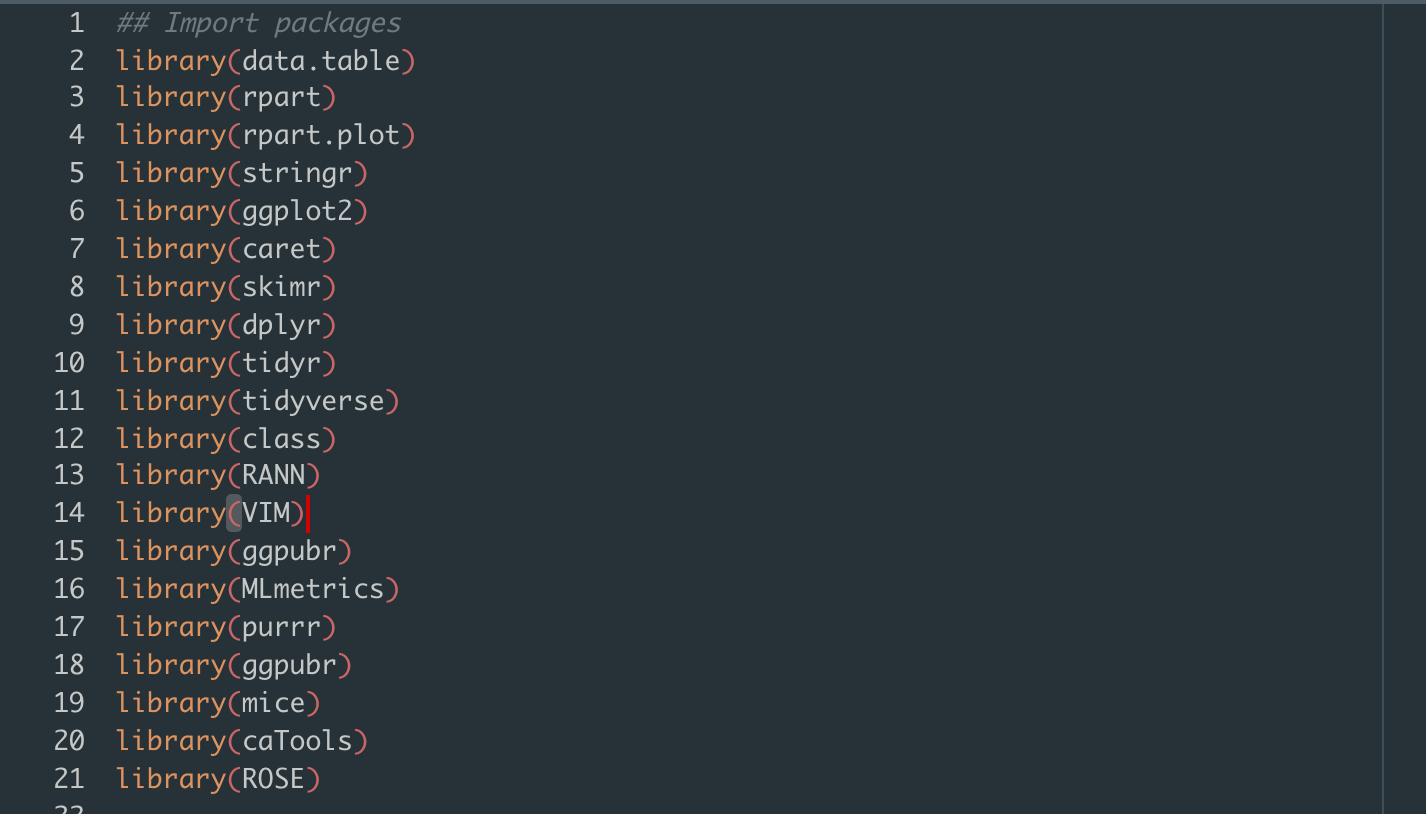
[Answer to Q5: 21](#_Toc117455776)

[Answer to Q6: 26](#_Toc117455777)

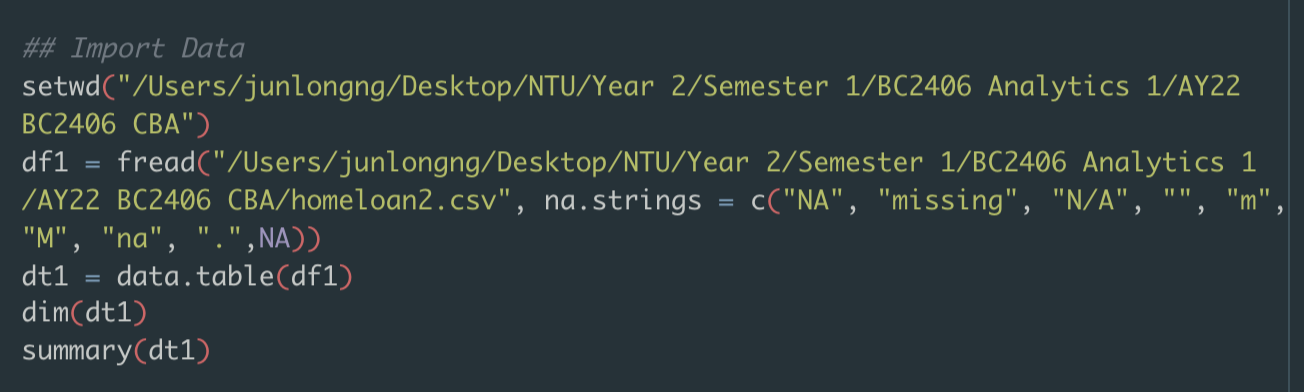
*For each question, please start your answer in a new page.*

# Answer to Q1:

# Answer to Q1a):



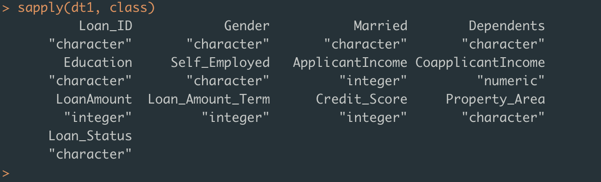
We start by importing all the relevant packages into our environment



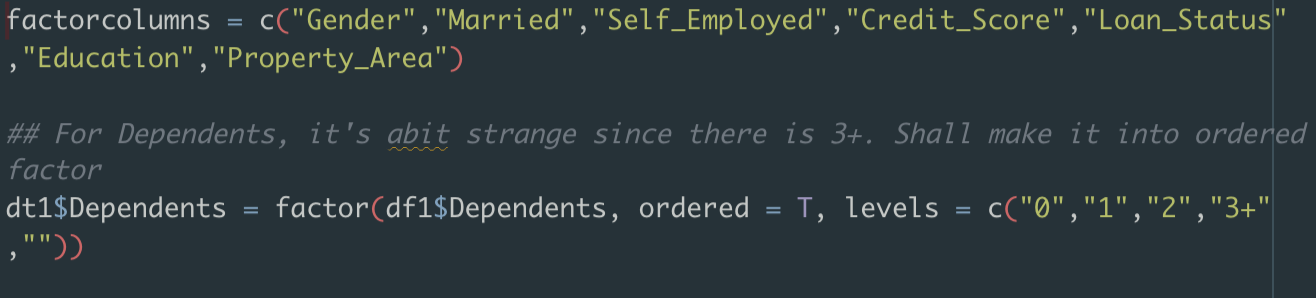
We then import the data and set na.strings = c(….) this allows us to catch all potential values of NA or missing data as NA such that it is easier for analysis later on.

We create a data table using dt1 = data.table(df1)



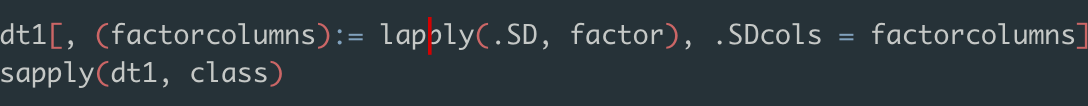


This shows us the class of each data. Most of the data is currently stored as a character and we have to recast them into the appropriate data type for further analysis.

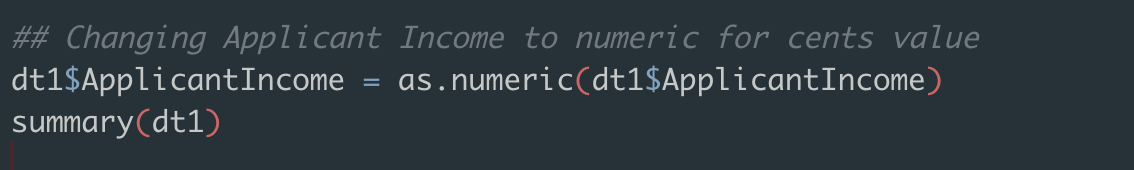


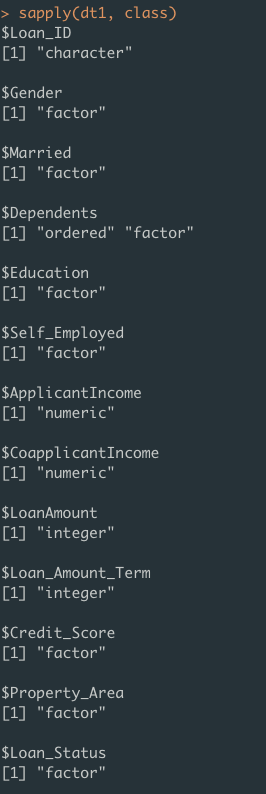
From these two lines of codes, I’ve decided to make Gender, Married, Self\_Employed, Credit\_Score, Loan\_Status, Education and property area into a factor data type

I’ve also made dependents into an ordered factor type.



I then recast them to dt1 using data table functions.

I also changed Applicant Income to a numeric value to allow for decimals.

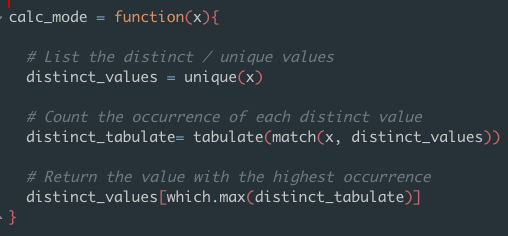


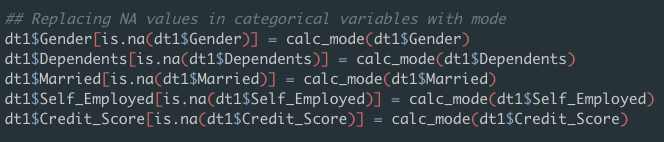
These are the new data types of each column.

# Answer to Q1b):

Based on my observation, will be using Mode to replace all categorical variables with NA values

As such, we have to develop a mode function in R due to a lack of it in the base R package.



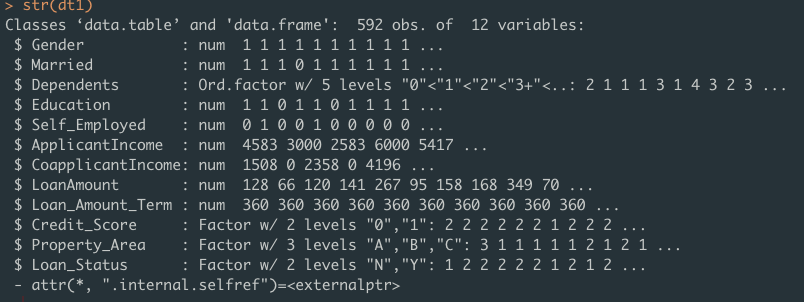


This replaces all Na values in each of the categorical data columns

I’ve decided to use median to replace NA values in the continuous data columns.







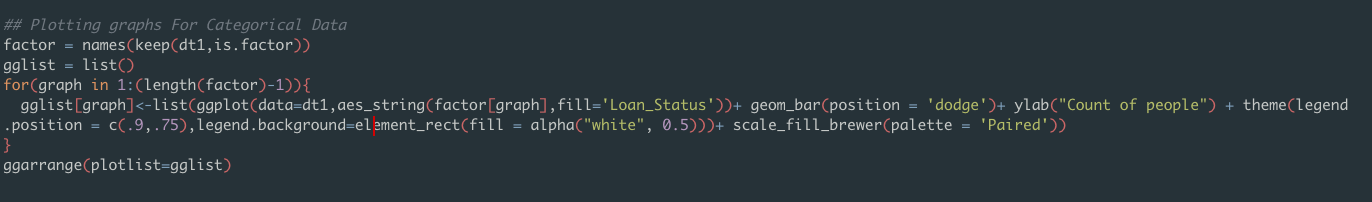
From here we can see the full list of data and their relevant types. We can already see that there are some outliers in Loan\_Amount and Coapplicant Income

# Answer to Q2):

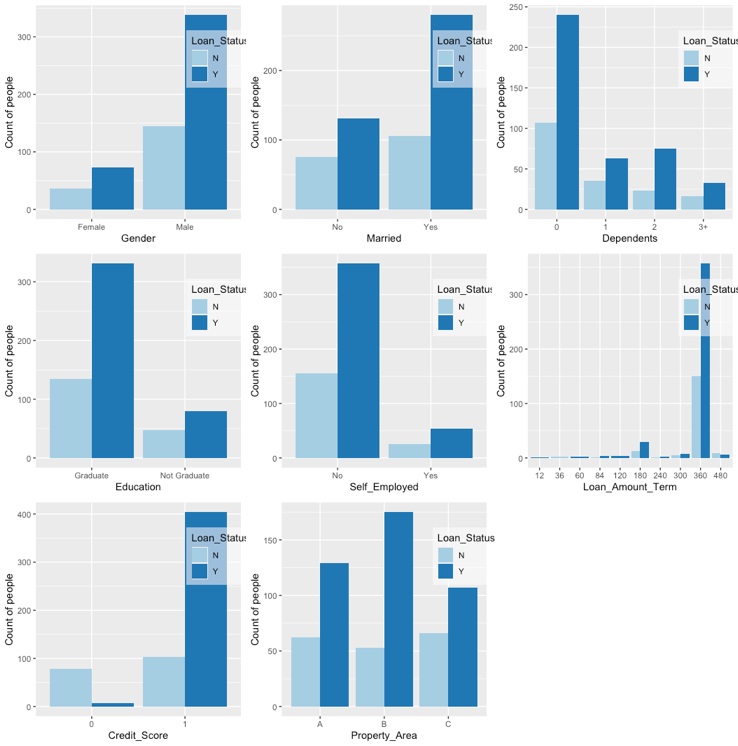
[Start your answers here…]

First step was to remove LoanID which will be explained more later.





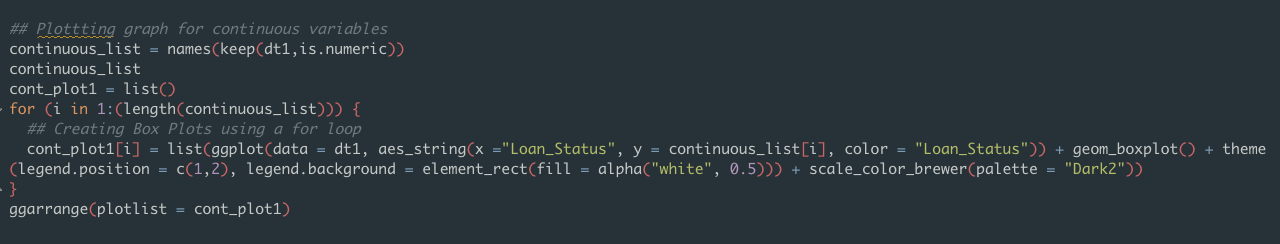
Using a for loop to plot categorical data.

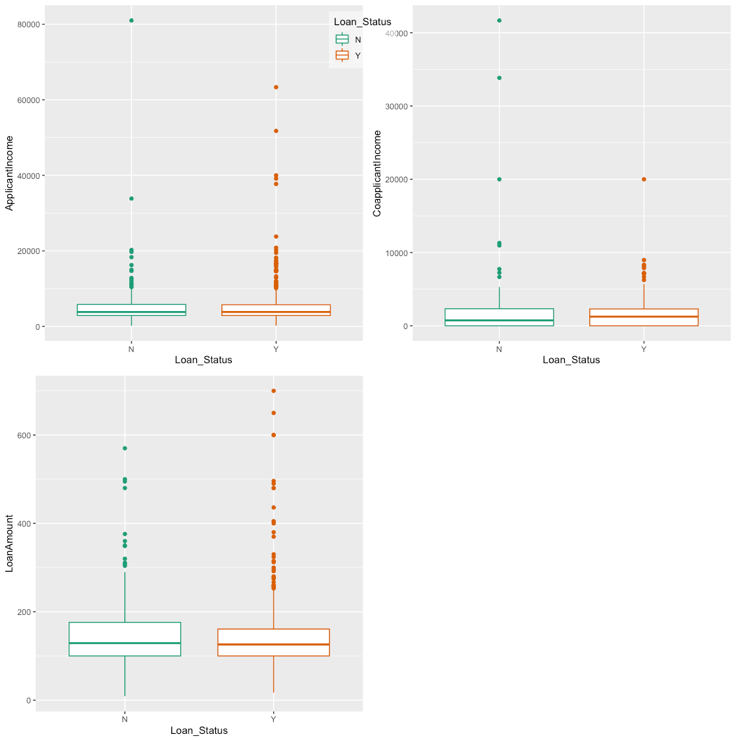


The Categorical graph tells us the following points

1. Gender V Loan Status: On average there are about a lot more males than there are females in our dataset. As a result, more males do get their loans approved. But whether or not there is a gender bias, we do not know yet
2. Married V Loan Status: We can also see that there are about twice more married applicants. Applicants that were married are also more likely to have their loans approved
3. Dependents V Loan Status: Most of our applicants do not have any dependents. Applicants with 2 applicants have a slightly higher chance of their loan being approved in terms of ratio.
4. Education V Loan Status: Applicants with “Graduated” as education seems to be correlated to a loan approval.
5. Self\_Employed V Loan Status: We have more employed people applying for loans as compared to people who are self-employed. Based on the preliminary analysis, there seems to be no bias for employment status yet.
6. Property Area V Loan\_Status: People in property area B have the highest amount of approvals for their loans.
7. Credit\_Score V Loan Status: Most Applicants have a credit score of 1. There seems to be a strong relation with having a loan and credit score of 1.
8. Loan\_Amount\_Term V Loan Status: Seems that most of the loans are for 360 months.

Plotting Continuous Graphs

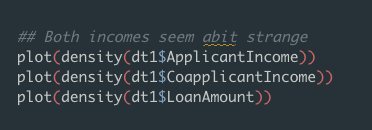


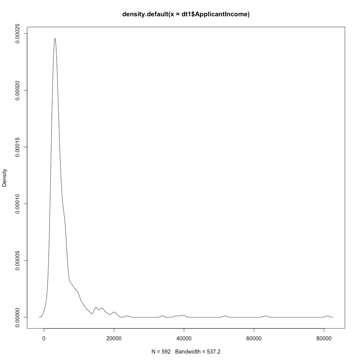


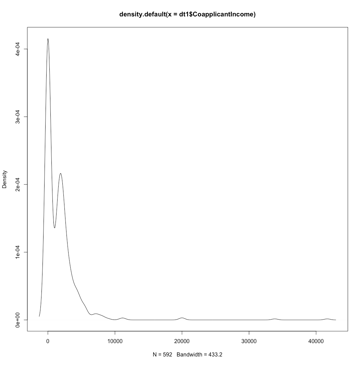
From the continuous table graph, we can see these points:

1. Applicant Income V Loan Status: Of those that were approved, there were more outliers in this column of data.
2. CoApplicant Income V Loan Status: There is a lower median for this column of data. Could suggest that coapplicant income is not significant. Need to investigate more.
3. Loan Amount V Loan Status: This reveals that the median for loan amount being approved or not is quite close. Suggests that perhaps loan amount has no effect on loan status.

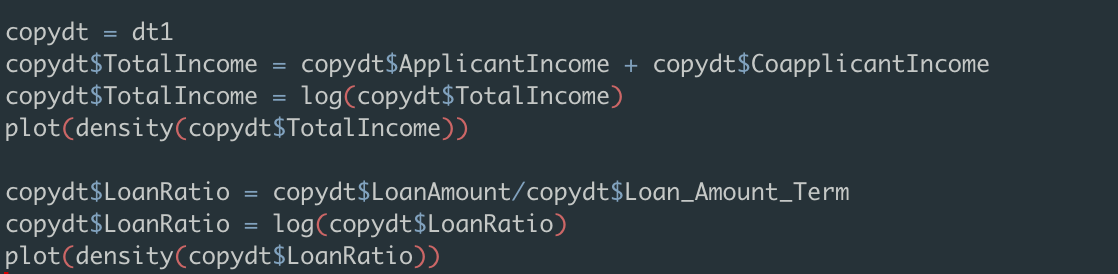
We need to take a closer look into income for applicant and co-applicant as it seems like there are various outliers



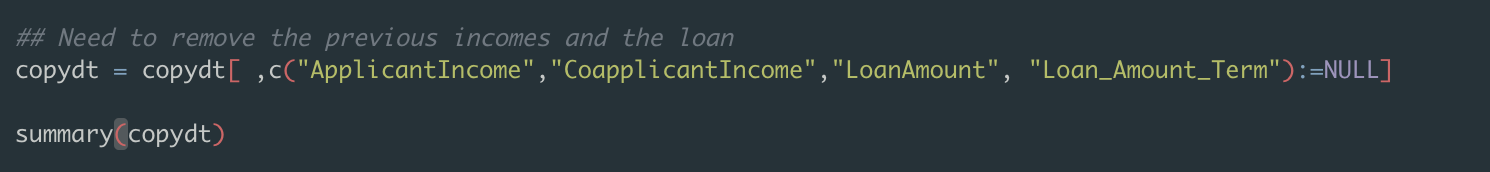


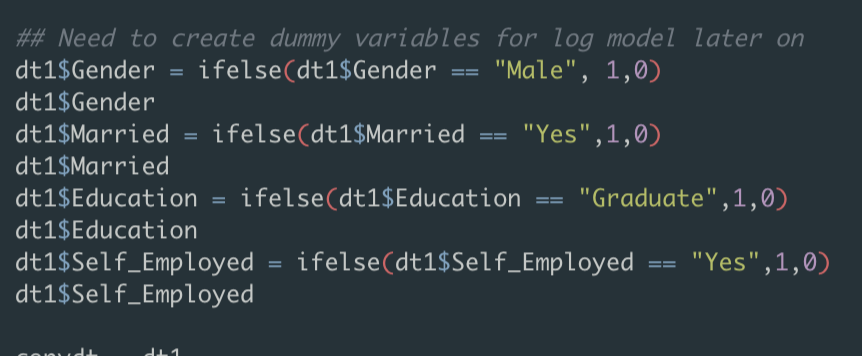


This shows us that the two data are extremely left skewed and needs to be handled.

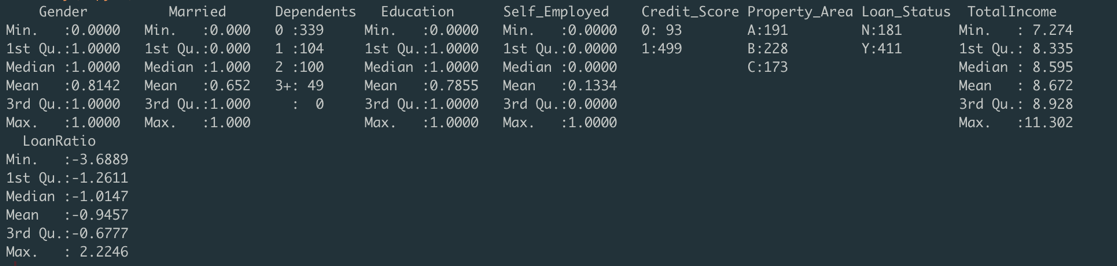


Decided to experiment with combining both applicants’ income together as well as creating a ratio of loan amount divided by loan amount terms to balance out the datasets.

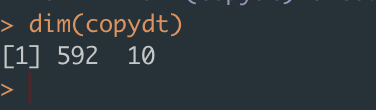


Removing the previous variables if not there will be a multi-collinearity effect and the effect of combining It would bring no extra value to the analysis model. I also created dummy variables by transforming the following variables to binary split for further analysis.

These are the final variables that I will be analysing later on.



For my final dataset, I will have 592 rows of 10 columns.



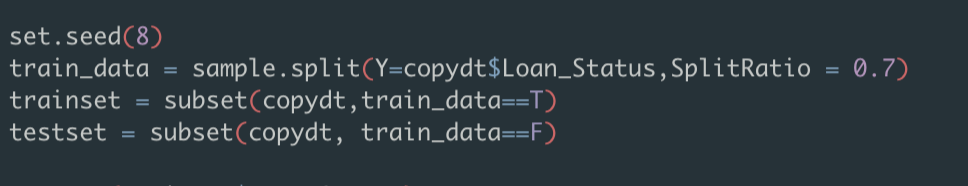
# Answer to Q3:

# Answer to Q3a):

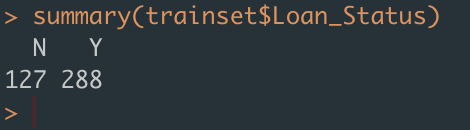
No, Loan\_ID should not be used as a predictor variable. This is because it is a unique identifier value that showcases each person's ID, and its function as a variable is to identify the person based on their existence. As a result, LoanID does not support any classification capabilities and should therefore not be a predictor X variable.

# Answer to Q3b):

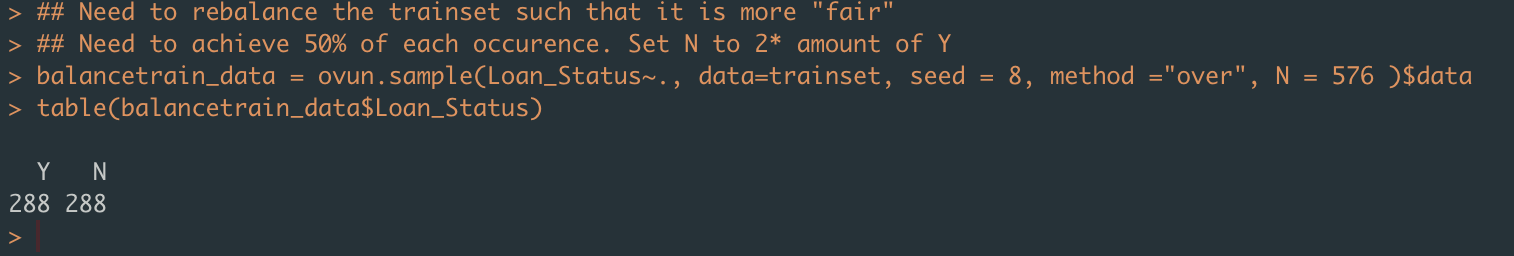
Setting the seed as 8 and creating a train test split



There seems to be an overwhelming amount of Yes, this might cause an imbalance in the trainset. Hence, we will need to perform an over sampling of No cases of under sampling of Yes cases.

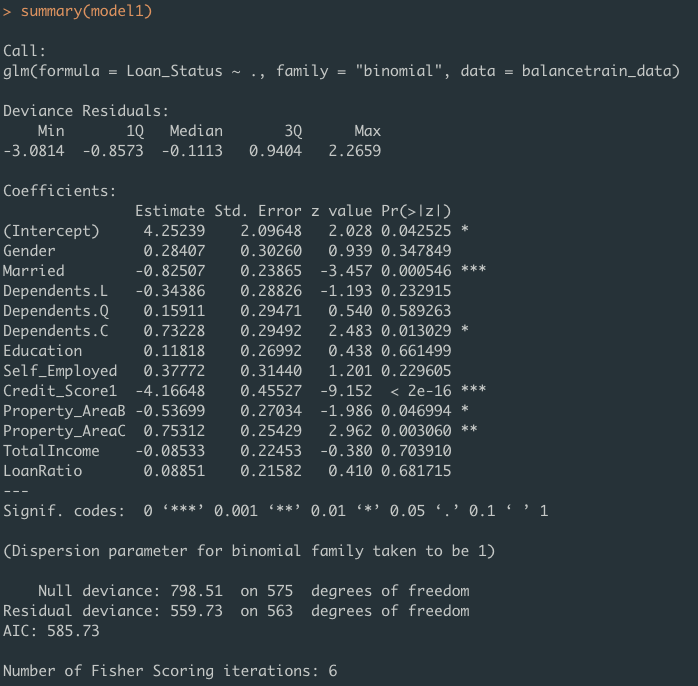


Decided to oversample No values, this will create a larger sample size for greater analysis.



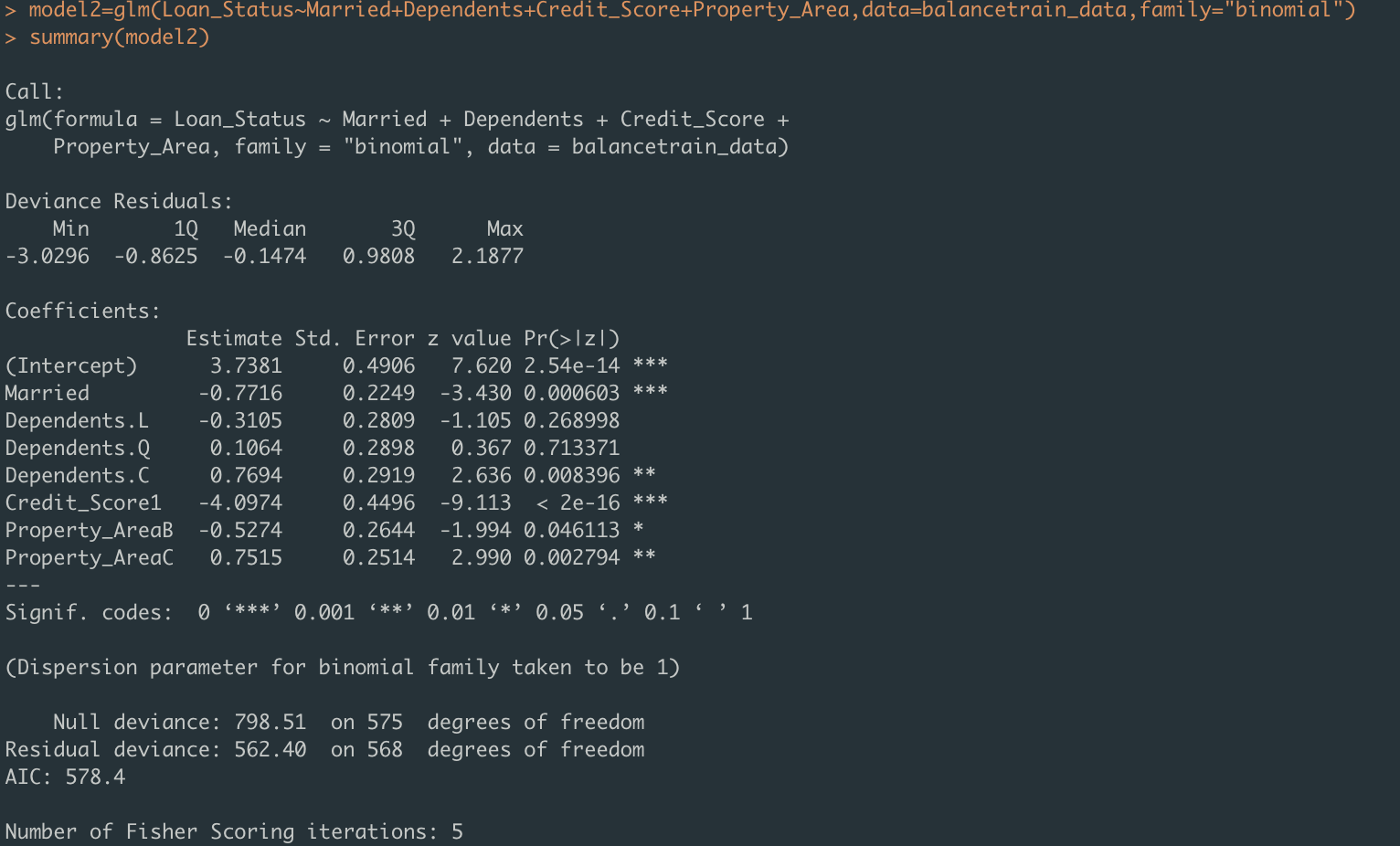
Performing a logistic regression on Categorical data “Loan\_Status”

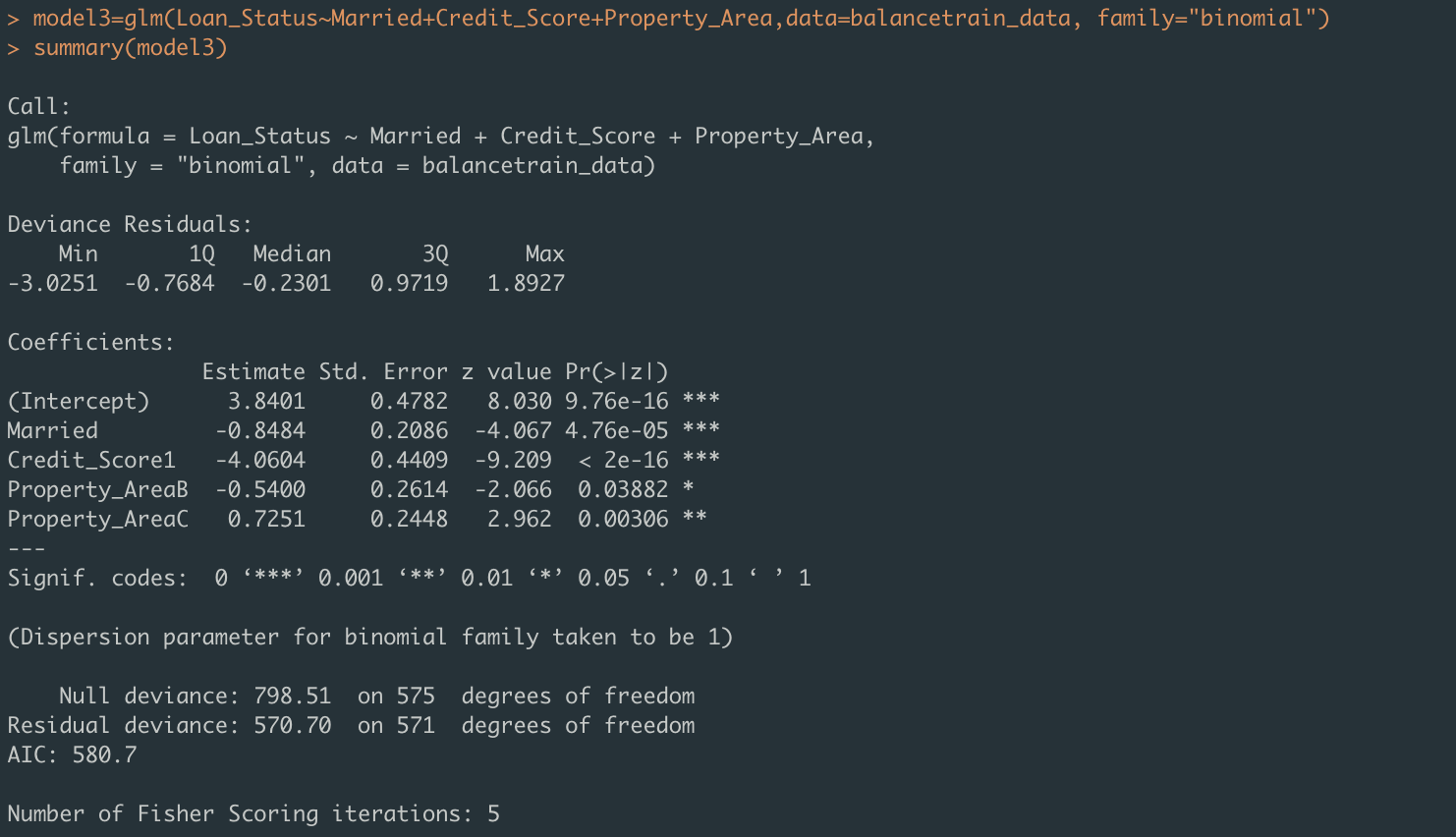


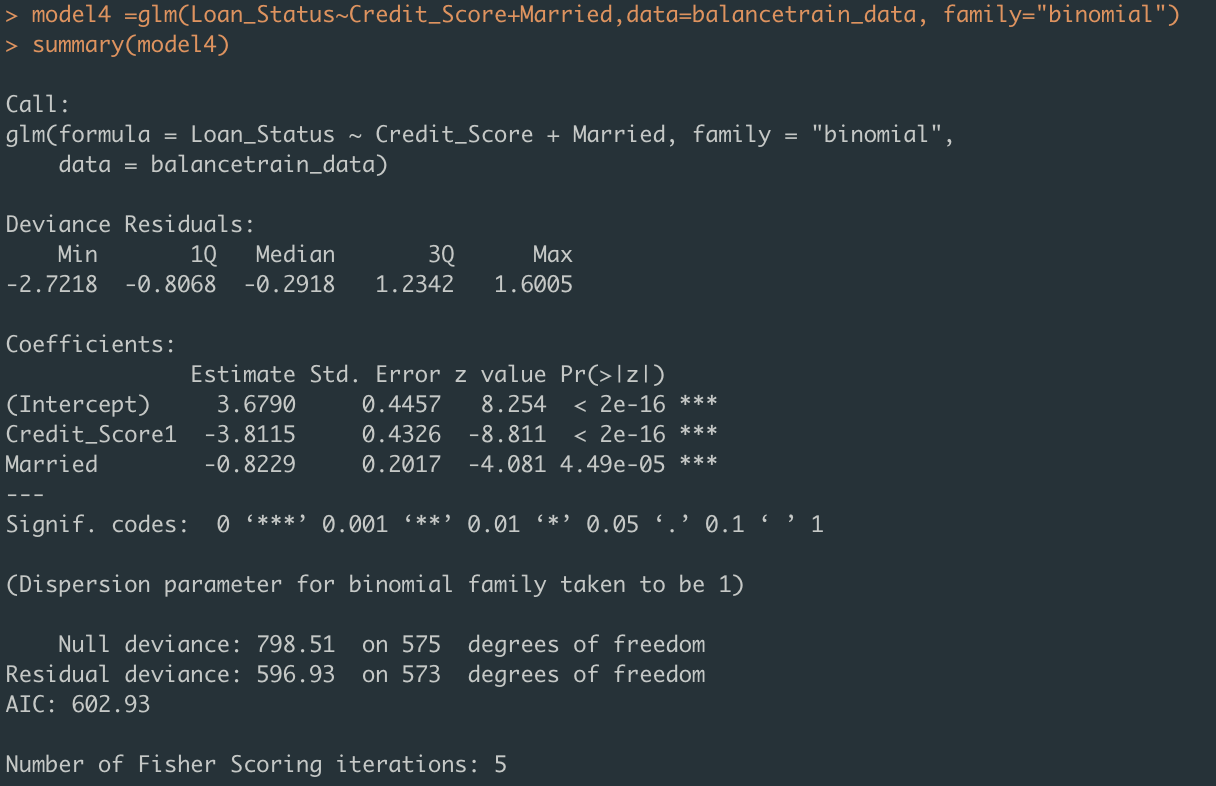


From this, we can see that marriage, Credit\_score, Property Area and dependents are significant

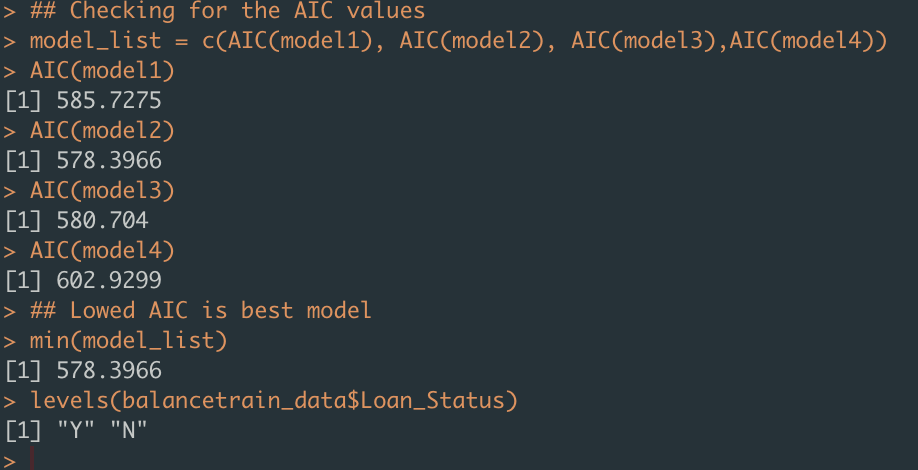
Using backwards elimination to figure out the best logistic regression model for predicting Loan Status



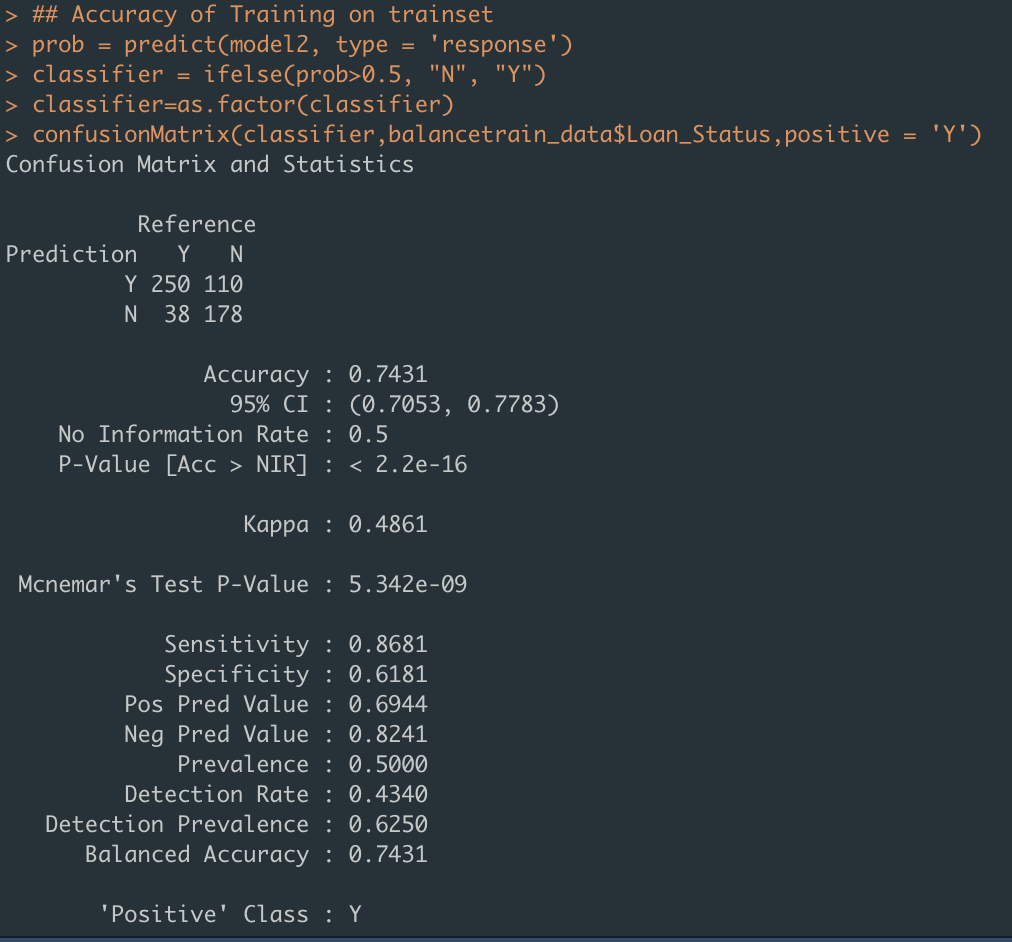




Using AIC to find the best logistic regression to use to predict Loan Status.

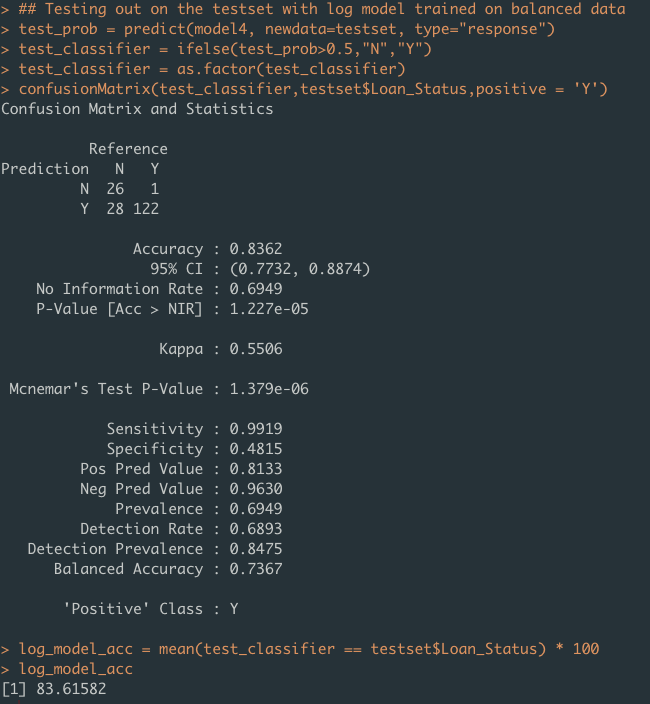


From here, we can see that model 2 has the lowest AIC value which means that it is the most accurate in predicting loan status through logistic regression. We also use levels function to discover the base reference value for loan\_status for further analysis.



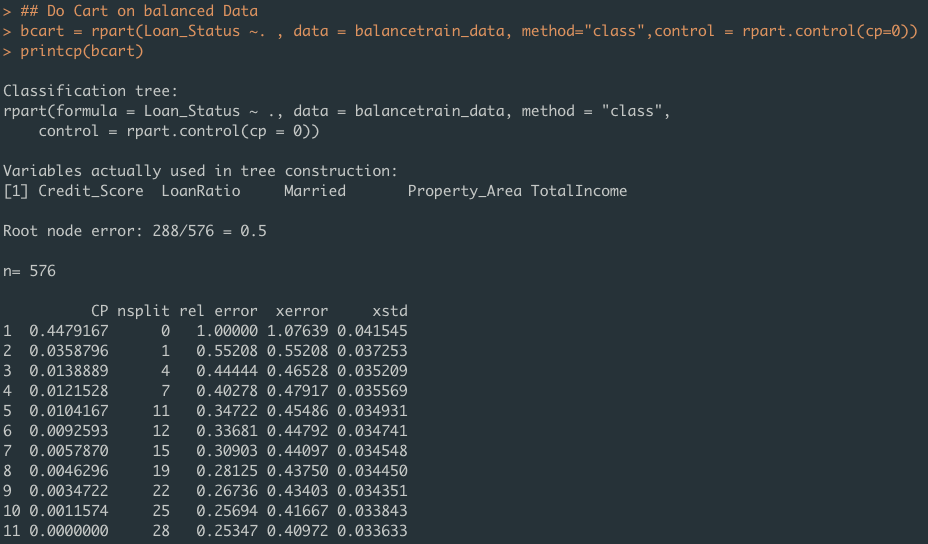
Using model 2, we are able to have a 74% accuracy in predicting Loan Status through logistic regression on the trainset data.

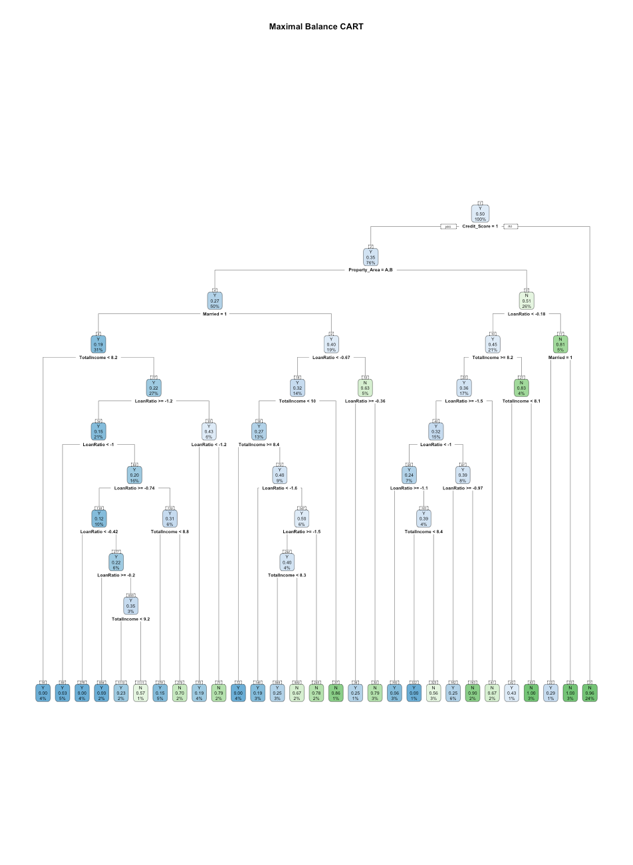
We will now use it on the testsetdata



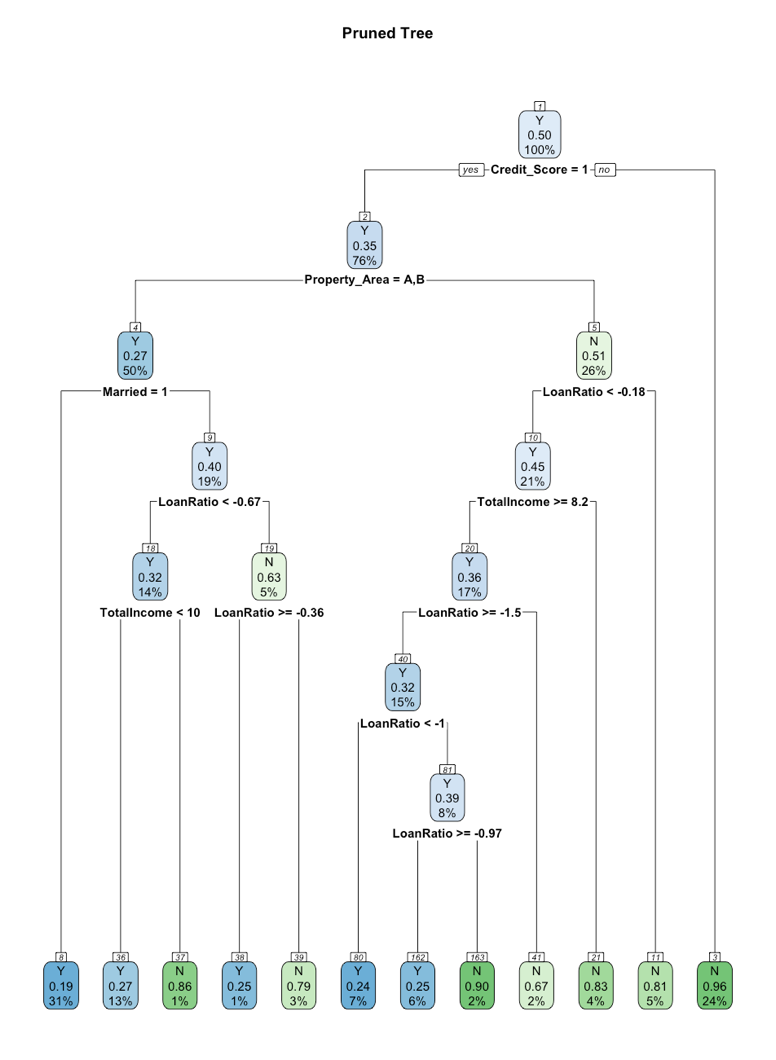
The logistic model on testset has an accuracy of 83%

We weill now begin building a CART model for the categorical variable Loan\_Status using the balanced train set data.

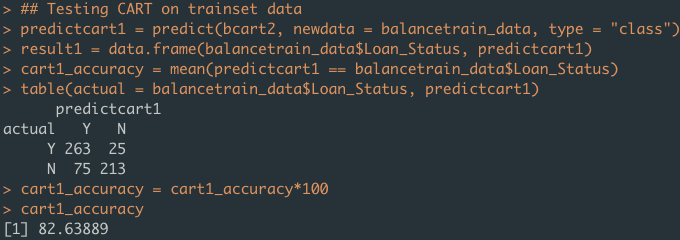




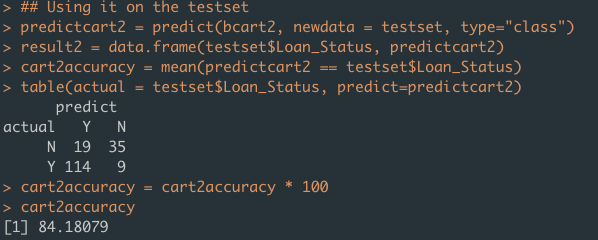
Once we’ve grown the full CART tree, we will need to prune the tree and subsequently plot it again.



Now we shall begin calculating the accuracy of the CART model based on the balanced train set data!

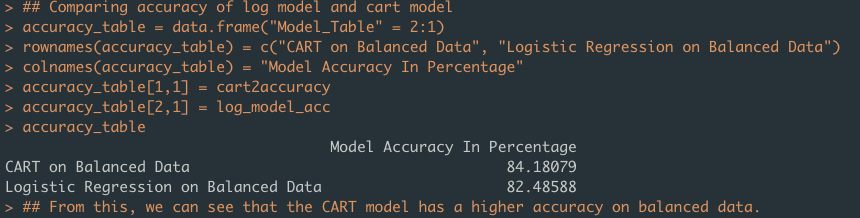


This gives us a CART model accuracy of 82.6% for the balanced train set data. We will now begin testing on the test set data!

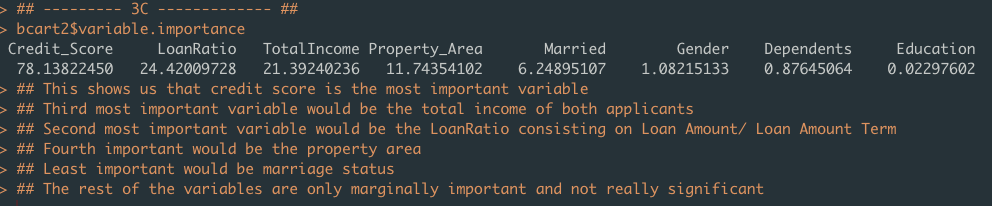


The CART model on test set accuracy is rounded up as 84.2%

Presenting it in a table format.

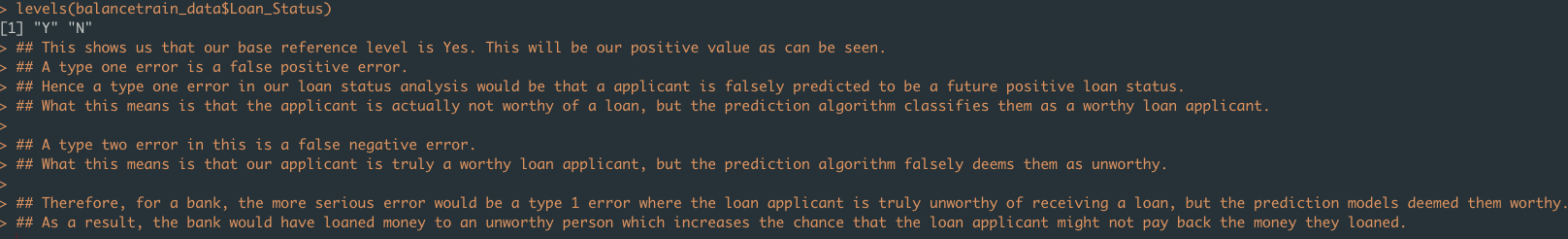


# Answer to 3c):



# Answer to Q3d):

To answer this question, we have to look at the base reference level for our analysis.



Our base level is Yes. This will be our positive value for type 1 and type 2 error comparisons.

A type one error is a false positive error.

Hence a type one error in our loan status analysis would be that an applicant is falsely predicted to be a future positive loan status. What this means is that the applicant is actually not worthy of a loan, but the prediction algorithm classifies them as a worthy loan applicant.

A type two error in this is a false negative error.

What this means is that our applicant is truly a worthy loan applicant, but the prediction algorithm falsely deems them as unworthy. The cost of a type 2 error in this case would be lost potential revenue if the applicant is actually worthy of paying loan interest and the principal amount back

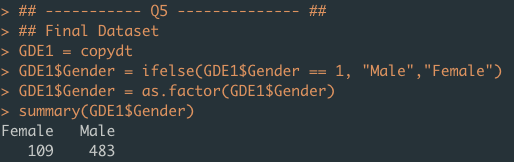
Therefore, for a bank, **the more serious error would be a type 1 error** where the loan applicant is truly unworthy of receiving a loan, but the prediction models deemed them worthy. As a result, the bank would have loaned money to an unworthy person which increases the chance that the loan applicant might not pay back the money they loaned. This error suggests that the bank will lose money instantly if the applicant defaults on the loan and is unable to pay the loan back. Making it the more serious error.

# Answer to Q4:

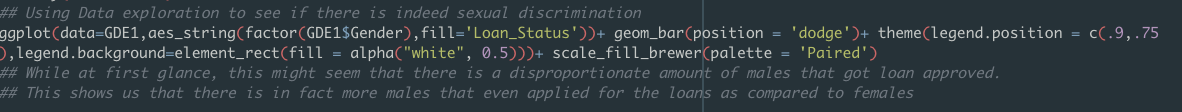
To reduce the chance of a type 1 error, the bank could run hypothesis testing with a smaller significance level. However, this method required some fine tuning as lowering the significance level by too much would mean that the bank’s researcher has failed to capture the true parameter of the hypothesis tests.

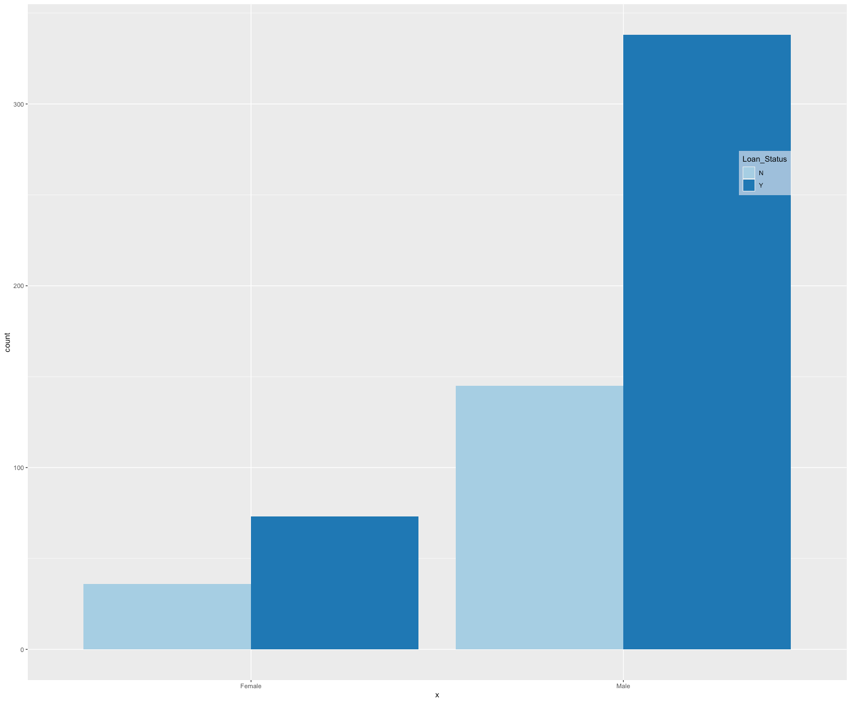
# Answer to Q5:

We will start by analysing the final dataset. From the code below, we can see that there are slightly more than 4 times the number of males that applied for a loan with the bank.



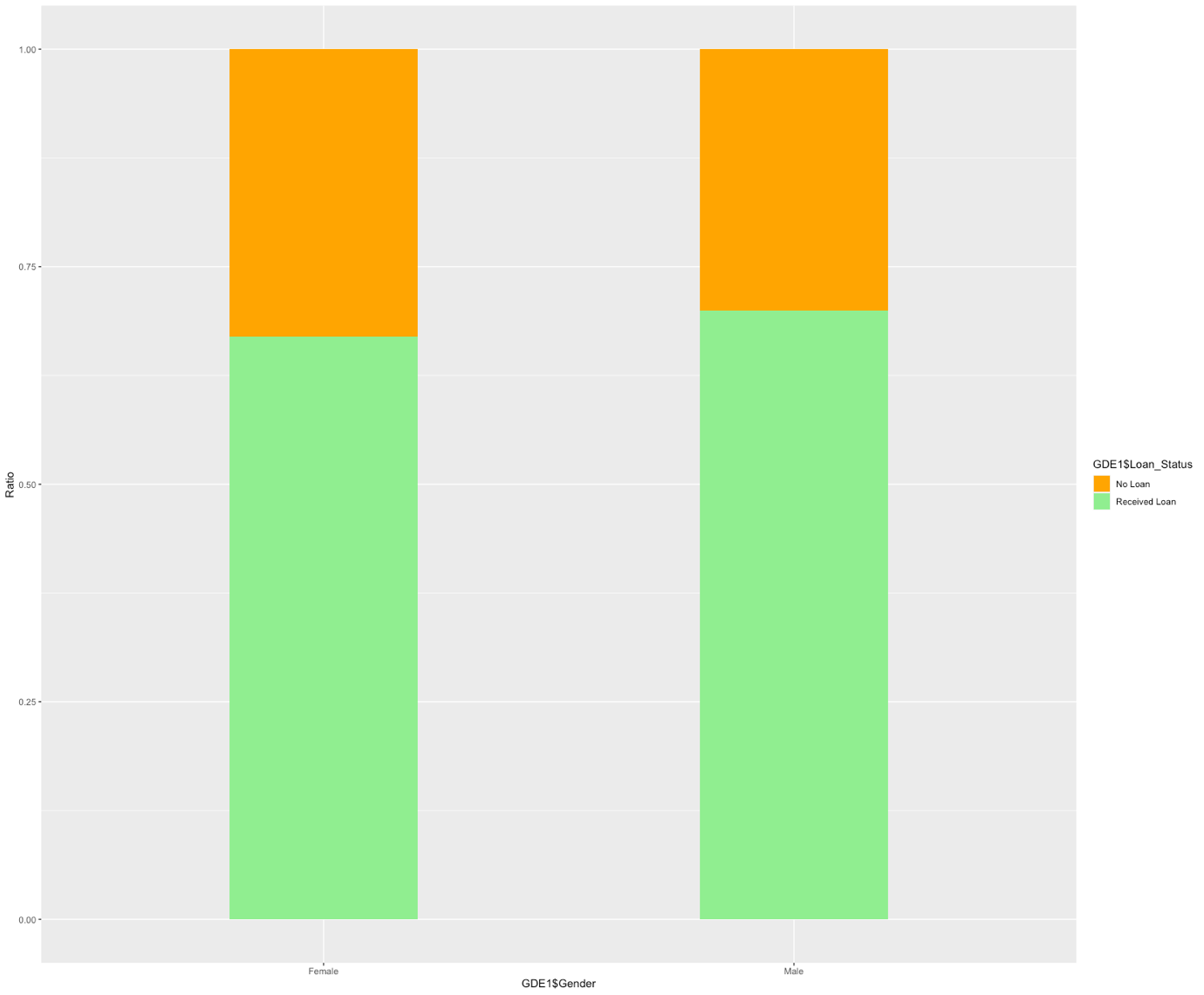
We’ve also plotted out the graph to showcase the proportion of males and females who managed to successfully receive a loan from the bank.



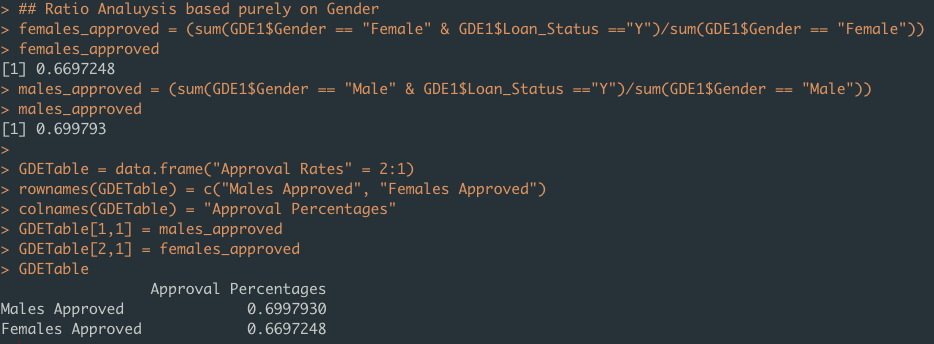


We will now look into the specific comparison of proportion between males and females that received a loan.



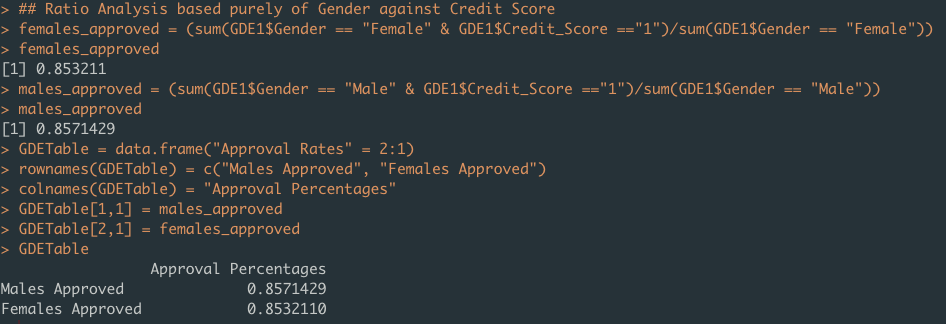


From the graph, we can see that the amount of proportion of females to males differ only slightly. This potentially suggests that there isn’t a gender bias but rather just that more males applied for the loan. I’ve developed a separate table to analyse the ratio of this observation.

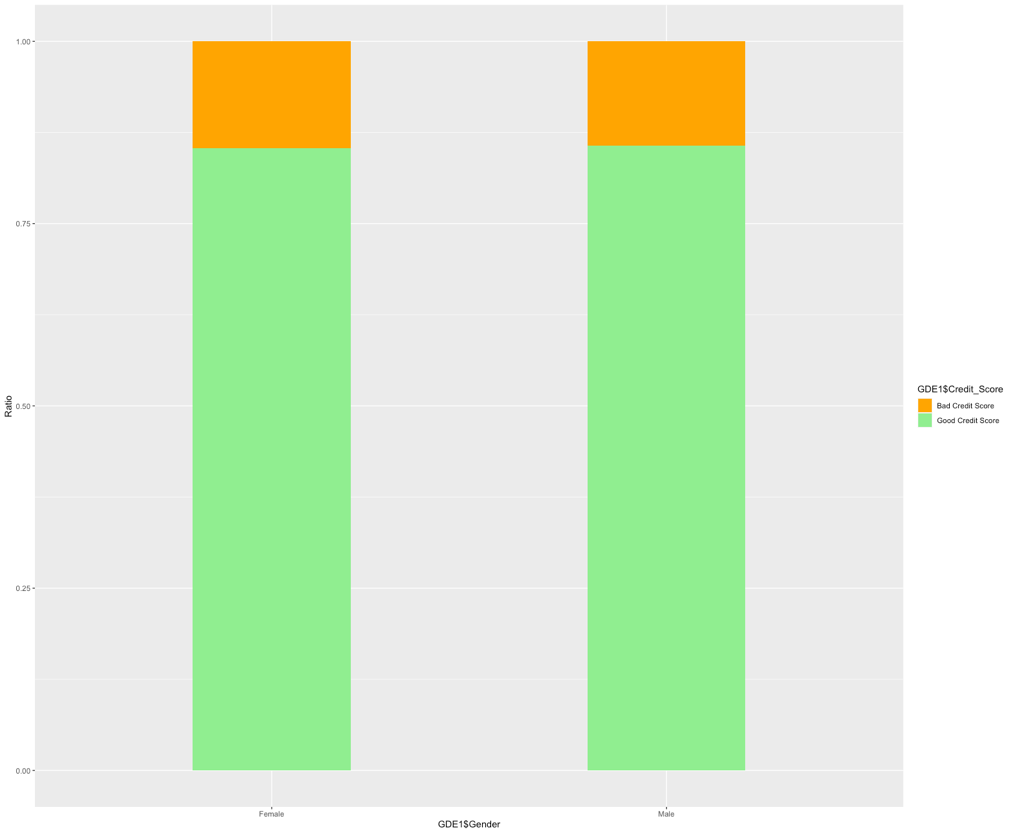


It seems that there is a very similar approval percentage for males and females based purely on their gender. This further disproves that there is a gender bias in the loan approval process.

Looking back at variable importance, the strongest variable that is correlated to loan status is credit score. Hence, one might argue that perhaps more males receive a good credit score as compared to females. As such, we will build similar ratio analysis and plot similar graphs to showcase the results of this hypothesis.

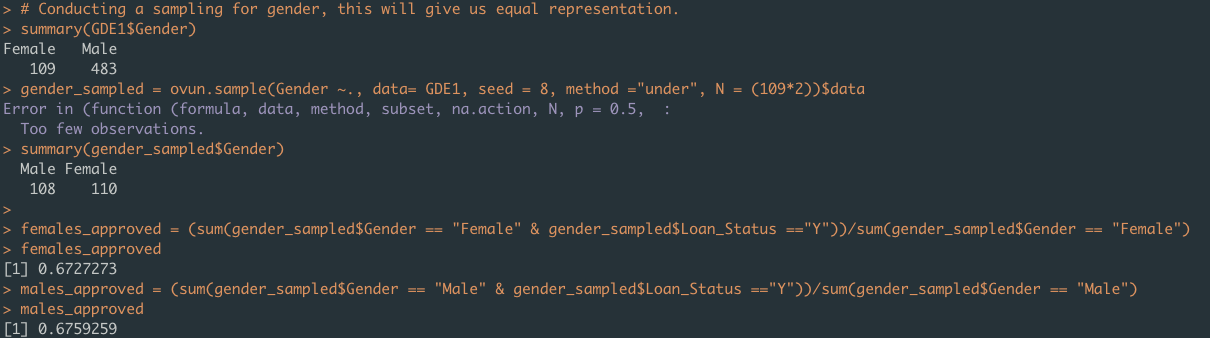


Similarly, there are similar percentages for both males and females to receive a positive credit score.



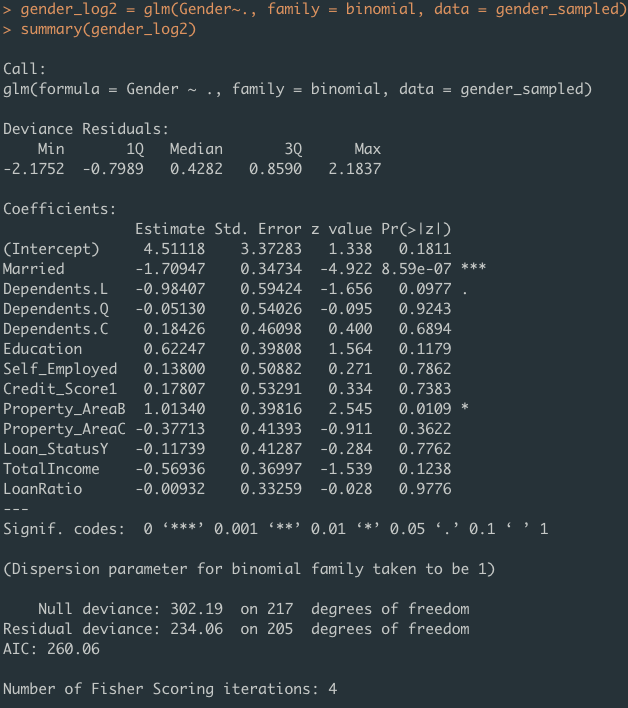
Graphically, it seems like the hypothesis is also disproved when both males and females receive the same credit score. Hence there is no gender bias even in the variable of credit score.

We will now attempt the same methods but with a balance of genders in our analysis. Which we start by creating a balance of males and females in our test data.



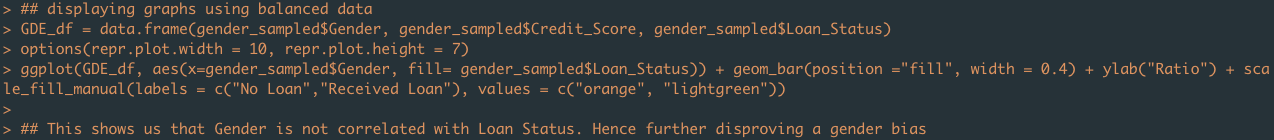
By removing the inflated number of males that applied, we can see the pure ratio of equal number of males and females. Similarly, we can see that males and females that are in our balanced data receive generally the same ratio of approval. Further disproving the gender bias theory.

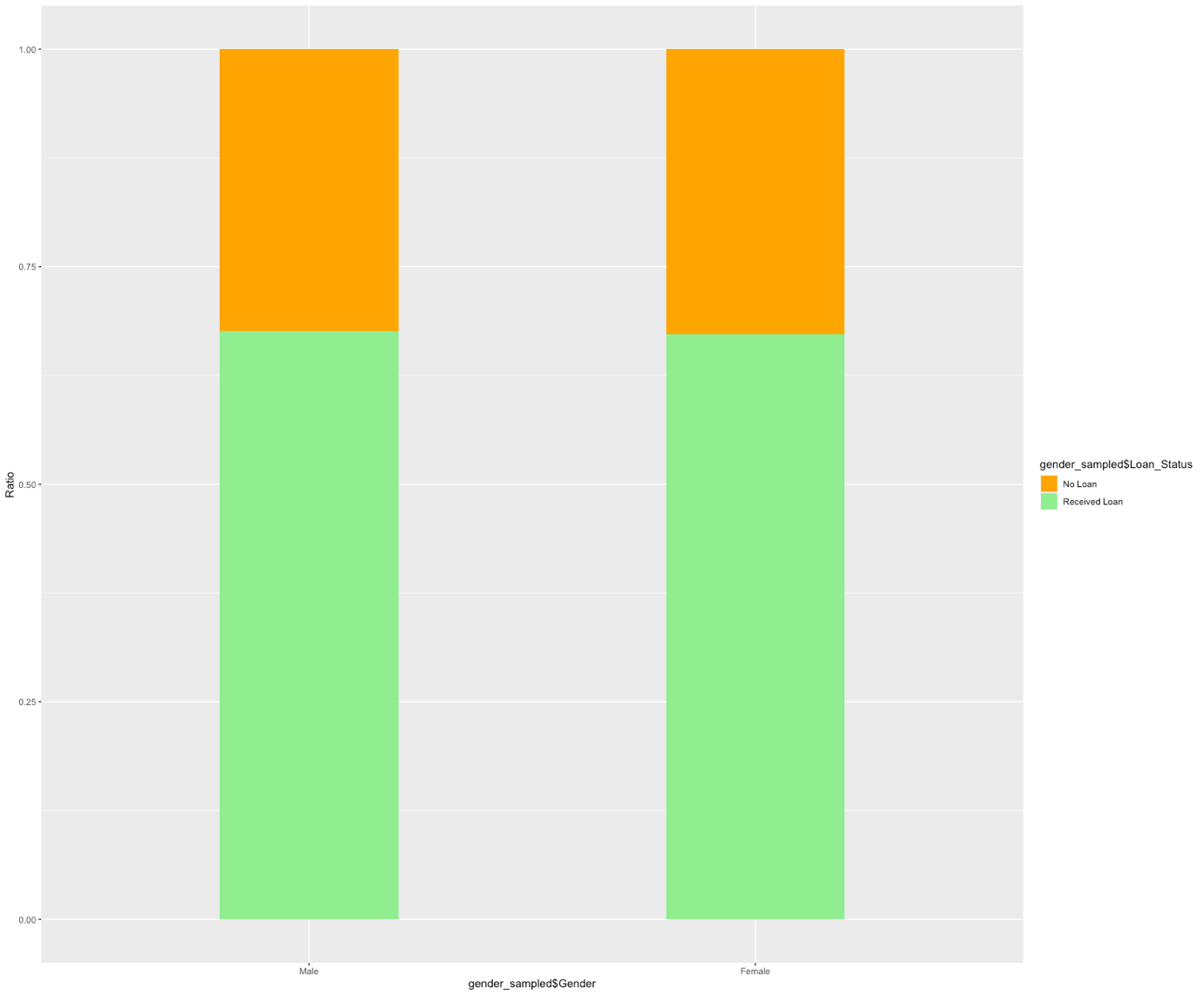
We will now run a logistic regression to see if Gender is correlated with Loan Status



From this, we can see that Loan Status has no significant correlation with Gender.

We will lastly end of with calling a graph to visually depict this result.





Therefore, through our analysis, we have checked on the unbalanced data and it reveals that there isn’t a gender discrimination through the loan process. We then created a balanced dataset which shares the same results that there is no gender discrimination. We’ve also analysed gender on credit score and there isn’t a strong correlation between gender and credit score as well. Hence, through our analysis, we are unable to detect gender discrimination.

# Answer to Q6:

One way that the bank can increase the accuracy of the analytics would be to deploy an ensemble model that combines base learners into a final model of a meta learner to reduce generalization errors within the bank.

However, this method relies on the assumption that each base learner model will yield a different aspect of the data and capture differing results. To do this, the bank should develop and train individual models independently of one another using various training and test splits.

With regards to data quality, the bank could also ensure that only processed data are fed into the analysis pipeline and that only quality data are recorded. This includes the enforcement of information fields that cannot be NA as can be seen in the current dataset where certain important fields such as Gender were left as NA. This increases the quality of the data and will be of greater benefit to the bank's analysis of data later on.

Another way to increase the accuracy of this model is to allow it to train on more data perhaps from cross collaboration with other banks or other branches. This allows the model to predict and train on a greater number of samples.