*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: 2](#_Toc87111201)

[Answer to Q2: 3](#_Toc87111202)

[Answer to Q3: 4](#_Toc87111203)

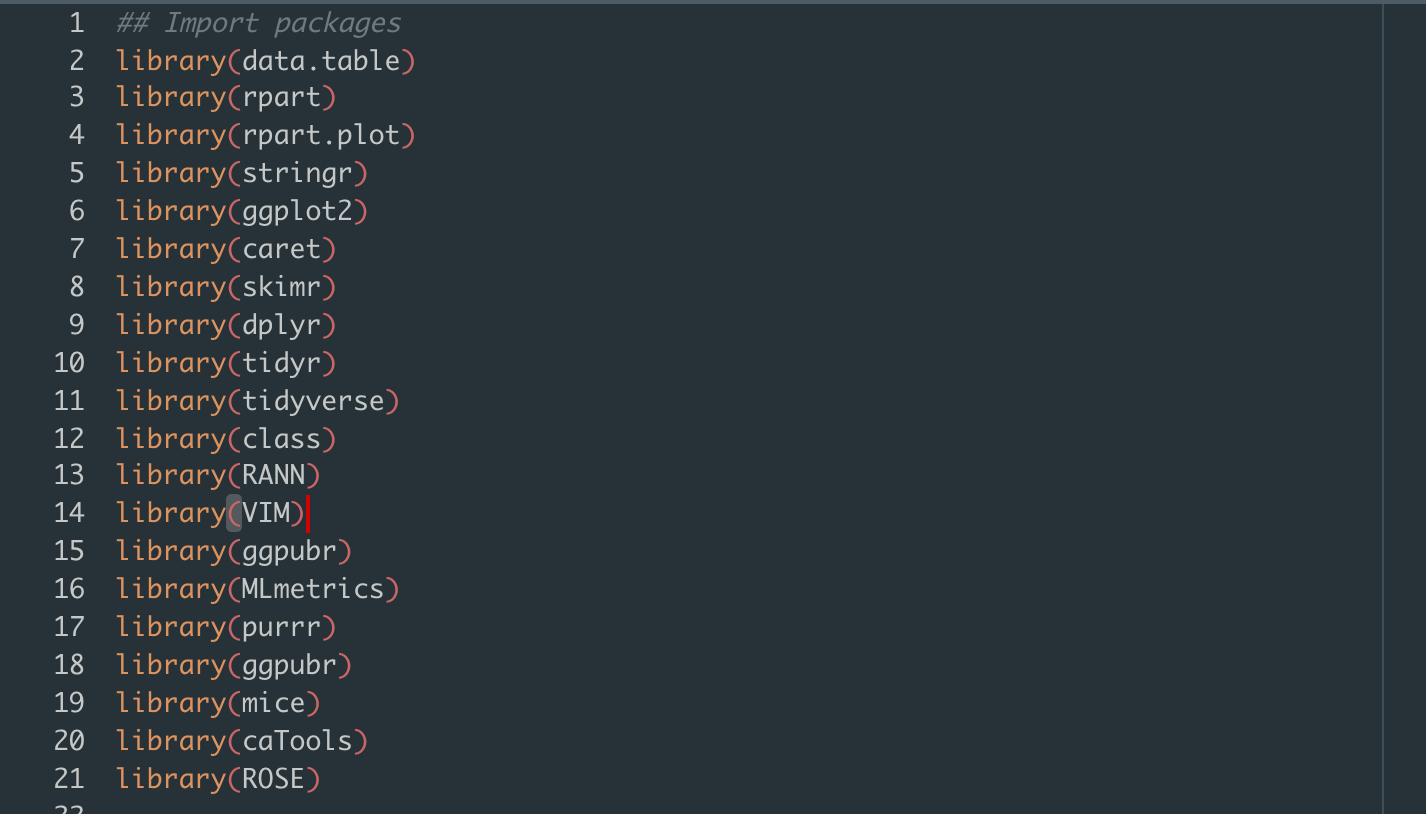
[Answer to Q4: 5](#_Toc87111204)

[Answer to Q5: 6](#_Toc87111205)

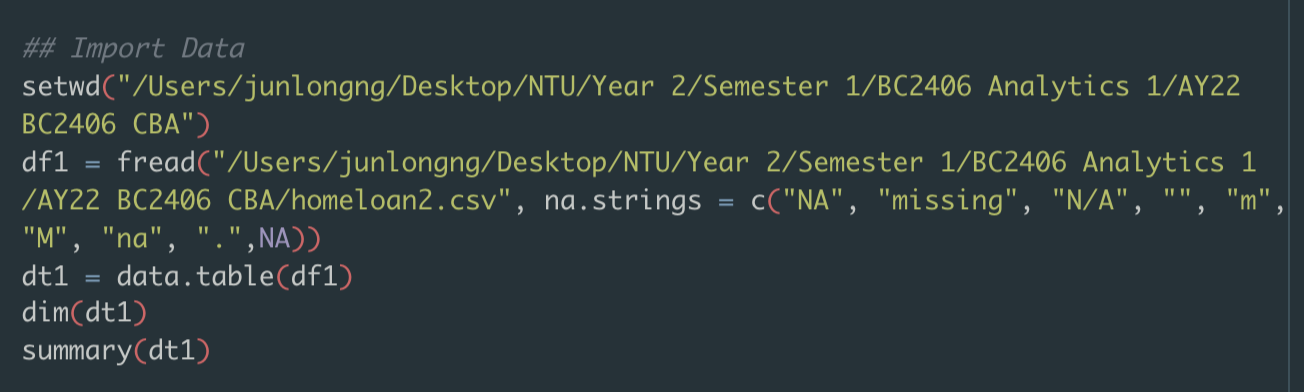
[Answer to Q6: 7](#_Toc87111206)

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

# Answer to Q1:



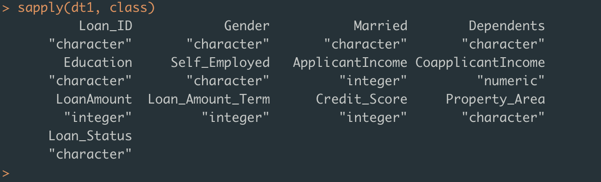
We start by importing all the relevant packages into our environment



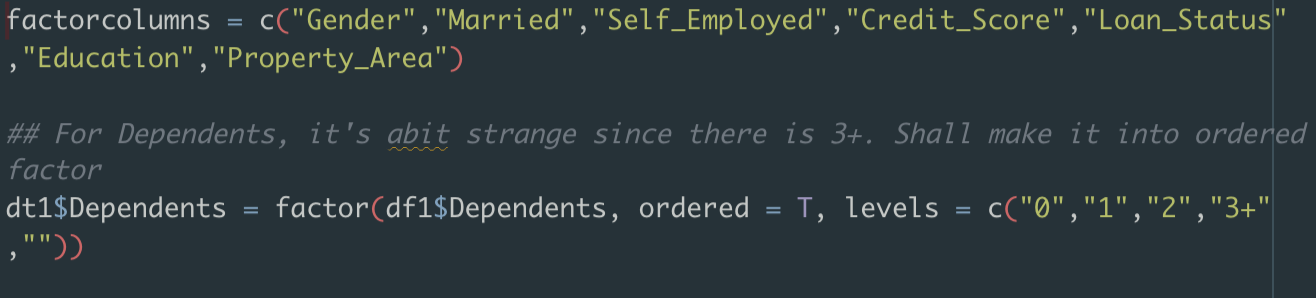
We then import the data and set na.strings = c(….) this allows us to catch all potential values of NA or missing data as NA such that it is easier for analysis later on.

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



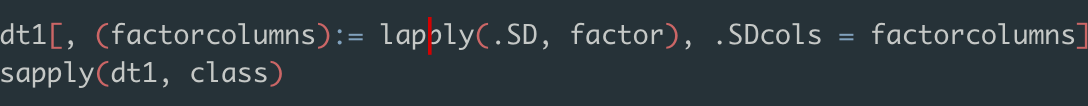


This shows us the class of each data. Most of the data is currently stored as a character and we have to recast them into the appropriate data type for further analysis.

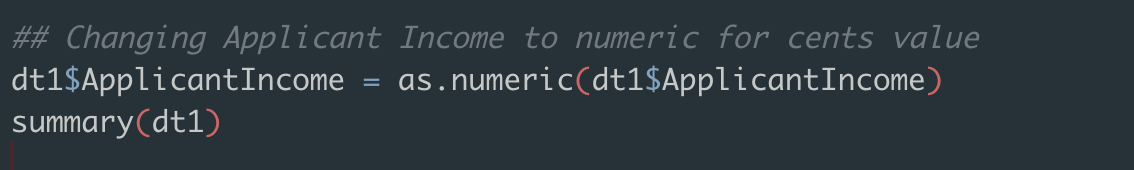


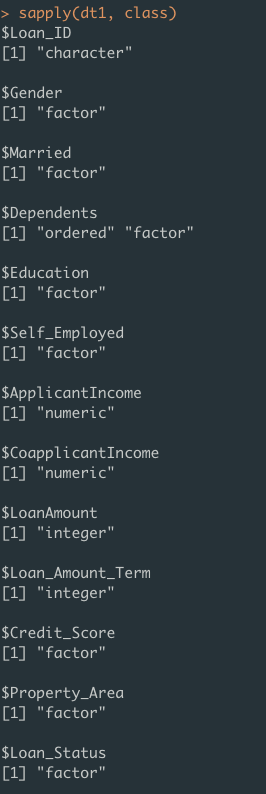
From these two lines of codes, I’ve decided to make Gender, Married, Self\_Employed, Credit\_Score, Loan\_Status, Education and property area into a factor data type

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.

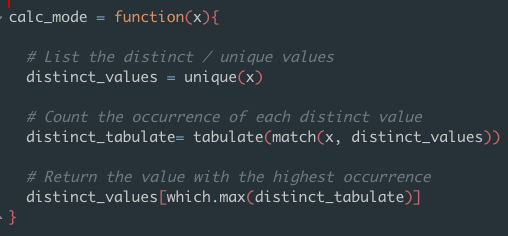


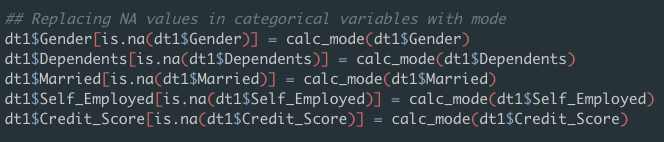
These are the new data types of each column.

# Answer to Q1b:

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

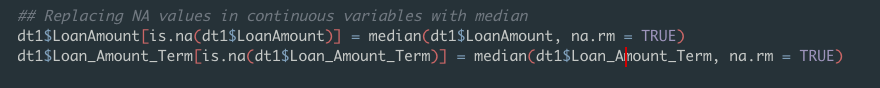
As such, we have to develop a mode function in R due to a lack of it in the base R package.





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.



# Answer to Q2:

[Start your answers here…]

# Answer to Q3:

[Start your answers here…]

# Answer to Q4:

[Start your answers here…]

# Answer to Q5:

[Start your answers here…]

# Answer to Q6:

[Start your answers here…]