Class / Seminar Grp	1
Full Name	SAREEN YOGYA HRIDEY
NTU Email Address	N2203421C@e.ntu.edu.sg

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.

[X] I have read and accept the above.

Table of Contents

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

Answer to Q1:

Variables	Data Type Before Correction	Data Type After Correction
Loan_ID	Character	Character
Gender	Character	Categorical
Married	Character	Categorical
Dependants	Character	Categorical
Education	Character	Categorical
Self_Employed	Character	Categorical
ApplicantIncome	Integer	Integer
CoappplicantIncome	Number	Integer
LoanAmount	Integer	Integer
Loan_Amount_Term	Integer	Integer
Credit_Score	Integer	Categorical
Property_Area	Character	Categorical
Loan_Status	Character	Categorical

Variable	Number Of Missing Values
Loan_ID	0
Gender	13
Married	13
Dependants	13
Education	0
Self_Employed	31
ApplicantIncome	0
CoappplicantIncome	2
LoanAmount	0
Loan_Amount_Term	14
Credit_Score	49
Property_Area	0
Loan_Status	0

We handle the missing values by deleting all the rows with missing values in the dataset. This helps us ensure that all the entries in the dataset are complete and therefore help in analysing the data in a better way.

```
  Loan_ID            Gender    Married    Dependents        Education   Self_Employed
Length:592         Female:109   No  :206   0   :334   Graduate    :465   No  :482
Class :character   Male  :470   Yes :384   1   : 98   Not Graduate:127   Yes : 79
Mode  :character   NA's  : 13   NA's:  2   2   : 98                      NA's: 31
                                          3+  : 49
                                          NA's: 13


ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term Credit_Score
Min.   :  150   Min.   :    0     Min.   :  9.0   Min.   : 12.0    0   : 85
1st Qu.: 2887   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360.0    1   :458
Median : 3806   Median : 1240     Median :128.0   Median :360.0    NA's: 49
Mean   : 5404   Mean   : 1646     Mean   :146.4   Mean   :342.1
3rd Qu.: 5754   3rd Qu.: 2324     3rd Qu.:168.0   3rd Qu.:360.0
Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480.0
                NA's   :2                         NA's   :14
Property_Area Loan_Status
A:191         N:181
B:228         Y:411
C:173
```

1: Before Data Cleaning

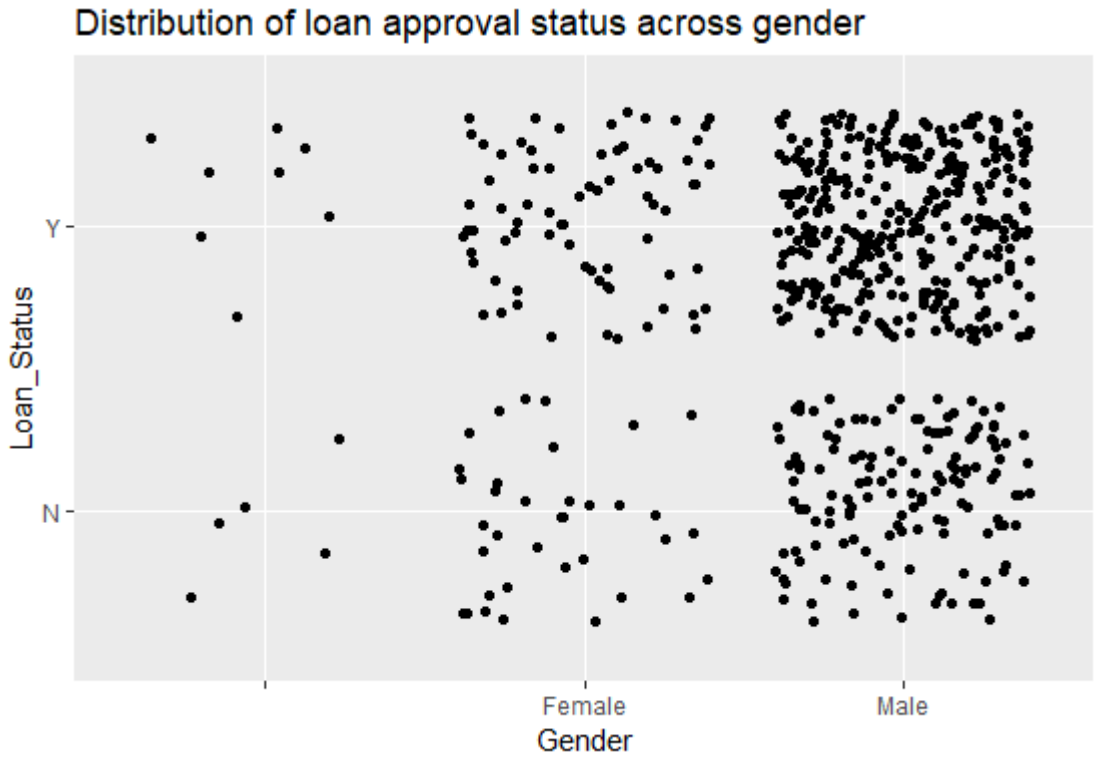
```
  Loan_ID            Gender    Married   Dependents        Education   Self_Employed
Length:478         Female: 86   No :169   0 :273     Graduate    :381   No :412
Class :character   Male  :392   Yes:309   1 : 80     Not Graduate: 97   Yes: 66
Mode  :character                         2 : 84
                                         3+: 41


ApplicantIncome CoapplicantIncome   LoanAmount  Loan_Amount_Term Credit_Score Property_Area
Min.   :  150   Min.   :    0     Min.   :  9   Min.   : 36.0    0: 70        A:149
1st Qu.: 2904   1st Qu.:    0     1st Qu.:100   1st Qu.:360.0    1:408        B:191
Median : 3863   Median : 1106     Median :128   Median :360.0                 C:138
Mean   : 5376   Mean   : 1586     Mean   :145   Mean   :342.4
3rd Qu.: 5900   3rd Qu.: 2254     3rd Qu.:170   3rd Qu.:360.0
Max.   :81000   Max.   :33837     Max.   :600   Max.   :480.0
Loan_Status
N:148
Y:330
```

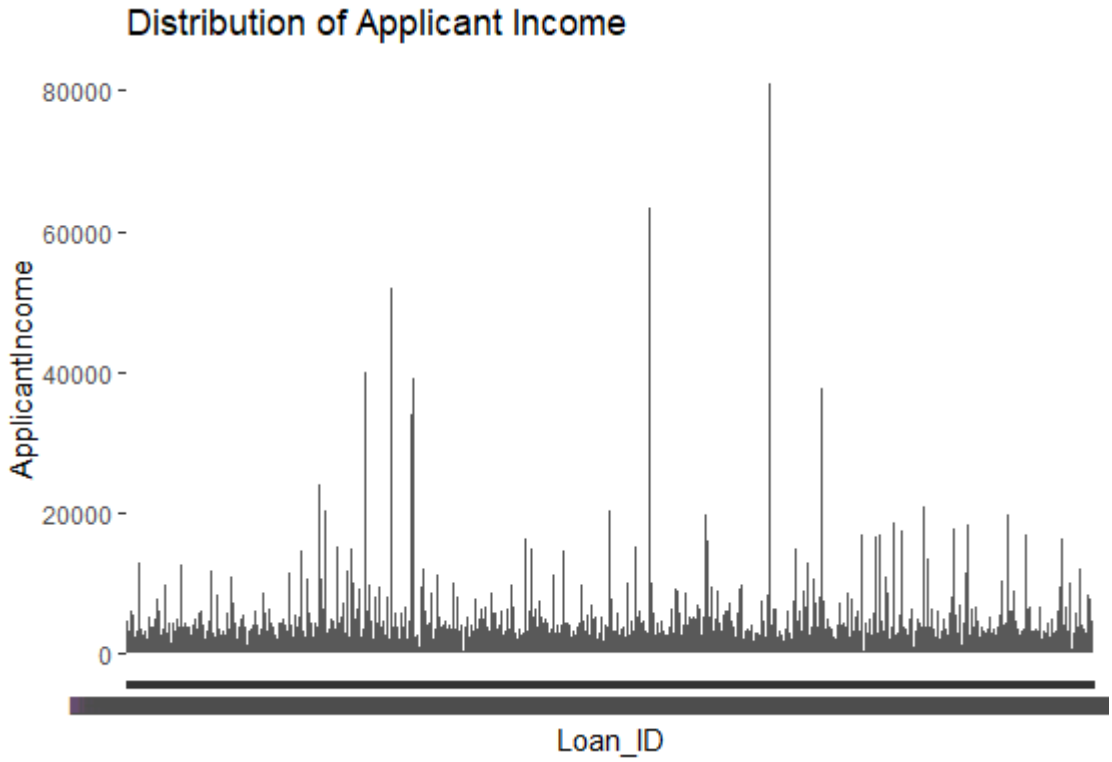
2: After Data Cleaning

In this summary, we can see that there exist missing values in almost every column of the data set. During data cleaning, we delete all the rows with missing values from the data set so that all observations in our data set are complete. This helps in better analysis of the data as the system gets complete observations.

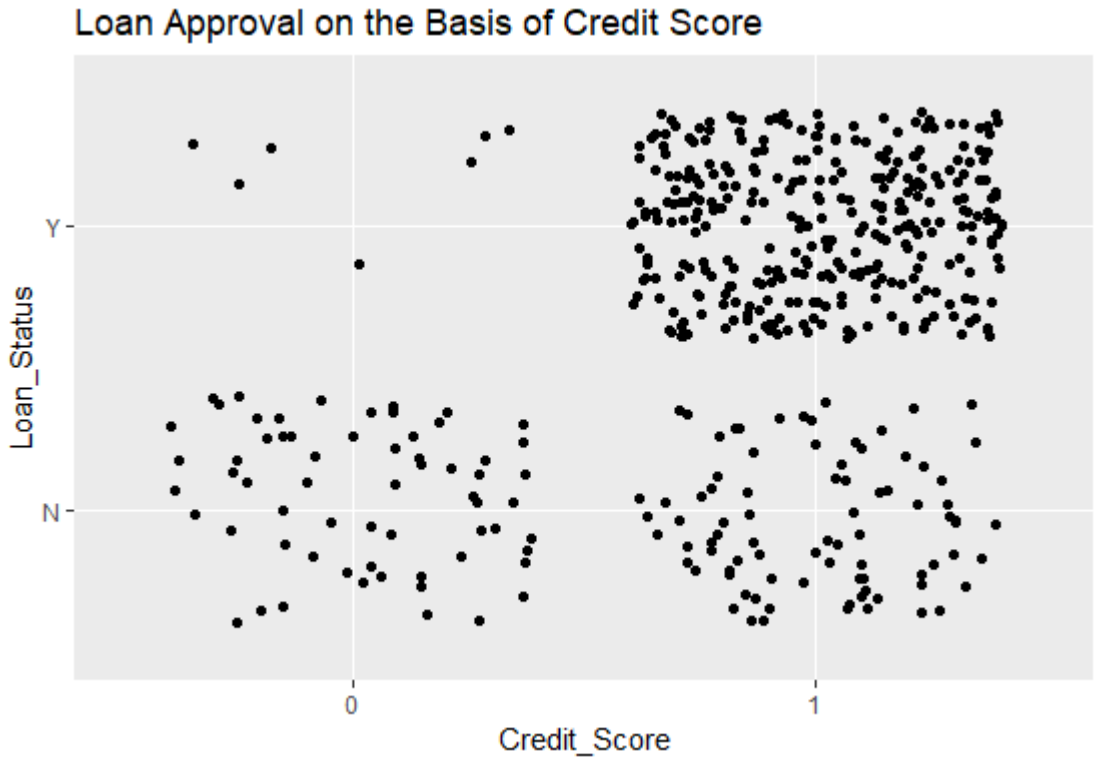
Answer to Q2:



We can see that the number of men applying for loans is much higher than that of women.



We can see that the income of applicants ranges from less than 10000 to more than 80000.



It is also clearly visible that having a credit score is very beneficial when applying for a loan, since the number of people with loans approved is much higher for those with a credit socre.

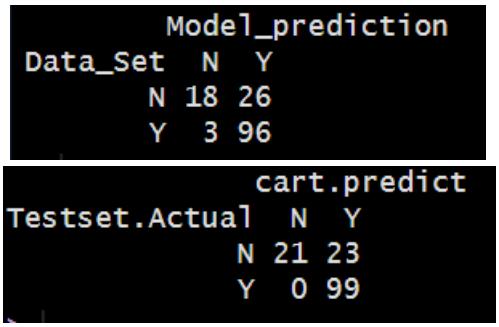
Answer to Q3:

a) Loan ID should **not be used** as one of the predictor X variables. This is because Loan ID is a reference number used to keep a record of the person applying for the role. Using this might lead to overfitting.

b) The model accuracy is-

   a. Logistical Regression =
   (18 + 96) / (18 + 96 + 3 +26) = **0.79 or 79%**

   b. CART =
   (21 + 99) / (21 + 99 + 23) = **0.83 or 83%**



According to this, the **best model found is the CART Model**, since it has a higher accuracy.

Model	Accuracy
Logistical Regression	79%
CART	83%

c) According to the **Logistical Regression Model**, the most significant factor in determining the loan status of a person is their **credit scores**. Other important factors are their marital status, co-applicant's income, and whether they live in Property Area B.

In the **CART Model**, the most important variable is still **credit score**.

d) In this case, **false negatives** (Type II errors) often more serious. This is because they directly impact the business's growth. We can also see that these are the errors that are less prominent.

Answer to Q4:

To further reduce serious prediction error, we should aim to reduce the number of false positives as well as the number of false negatives. This can be done through various ways such as –

- Ensuring proper data collection and reducing the number of missing values. This would help us in having a bigger and more precise data set which would in turn help us train and test the model with even more accuracy.

- Implementing proper data collection techniques, which would help in reducing the number of inaccuracies in the data set.

- Adjusting the threshold of loan approval. This would reduce the number of either false positives, or false negatives, depending on whether you increase

Answer to Q5:

There are several ways you can improve the success of analytics such as-

1. Using better and more advanced analytics and machine learning models. Some of such models are-
    a. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions through voting or averaging.
    b. Gradient Boosting Machines (GBM): Another ensemble learning technique that builds multiple weak learners sequentially, with each learner trying to correct the errors of its predecessors.
    c. Support Vector Machines (SVM): A powerful and versatile supervised learning algorithm that can handle both linear and non-linear classification tasks.
    d. Neural Networks: Deep learning models like Convolutional Neural Networks (CNNs) for image recognition, Recurrent Neural Networks (RNNs) for sequence data, and Transformer models for natural language processing have achieved state-of-the-art performance in various domains.
    e. XGBoost and LightGBM: Gradient boosting libraries that are faster and more memory-efficient than traditional GBM.

2. Having regular checks on the model to ensure that it indeed working in the manner that you want it to and is not encountering any errors or issues.

3. Establish a better communication network between the data analysts / scientists and the domain experts so that we ensure that the work put in aligns with the goals of every project. This could be done by hiring more business analysts who can serve as a middleman to the domain experts and analysts, with their knowledge of the domain and experience in analytics work.

4. As already mentioned in the previous question, ensuring proper data collection as well as data cleaning would also help improve the success rate of analytic models which would in turn help the bank.

References

- Class notes
- Homework Solutions
- List of better analytics techniques taken by ChatGPT