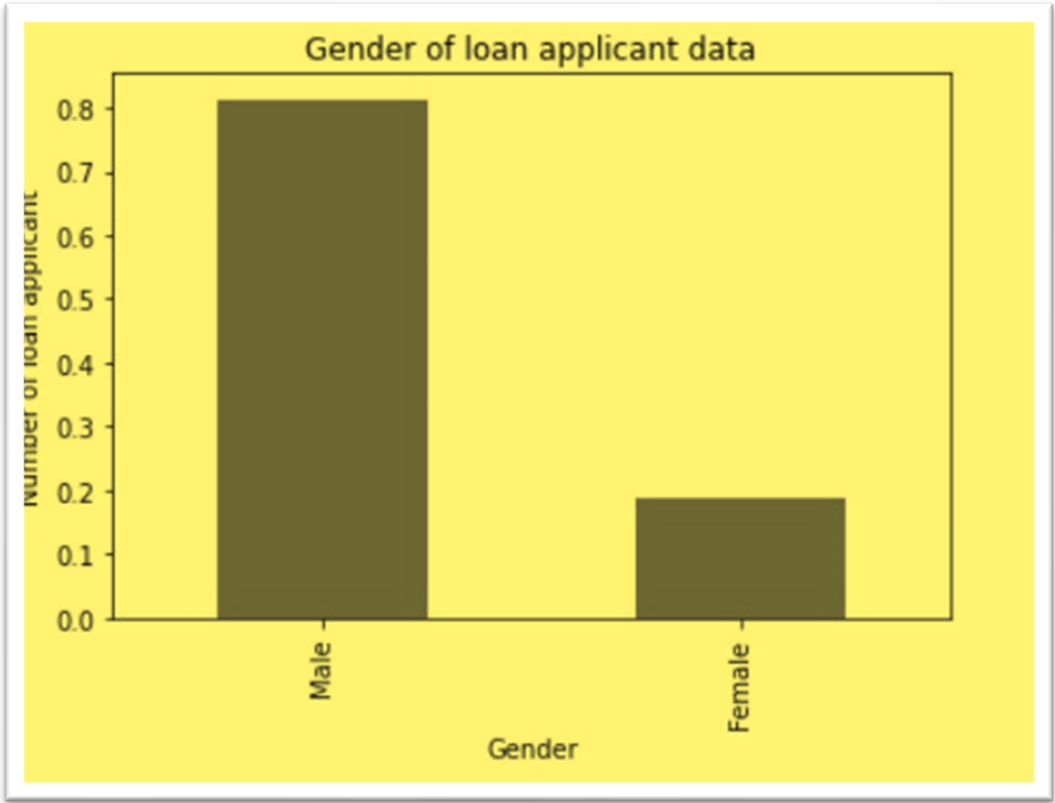


## Project 2:

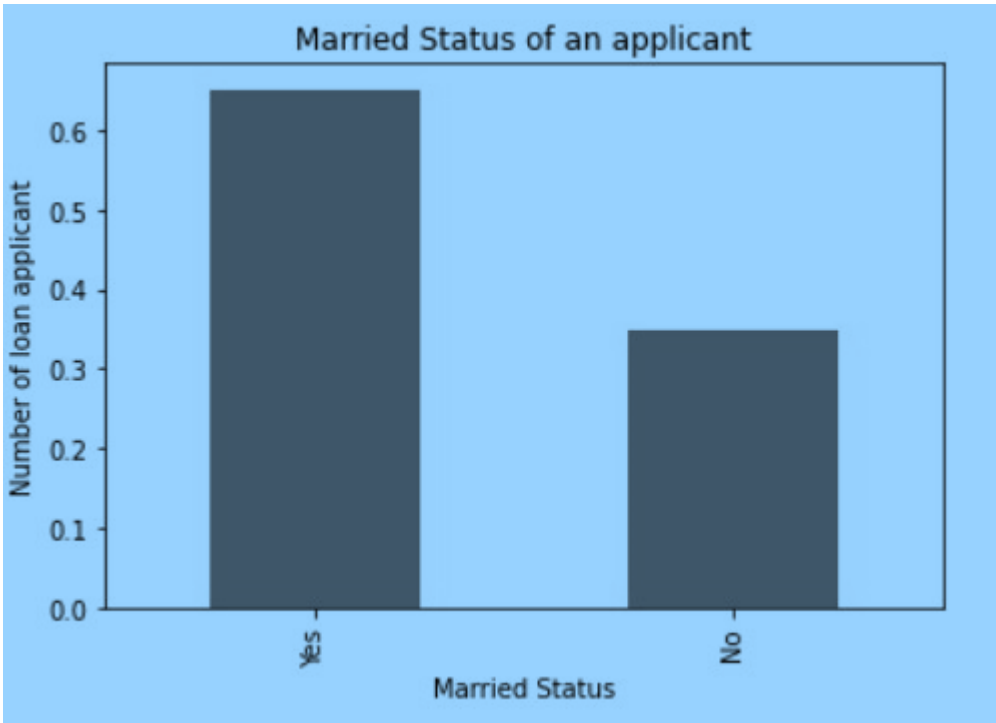
### Section A: Exploratory Data Analysis

1. Let us analyze and visualize the categorical attribute of the given train dataset using single variable.

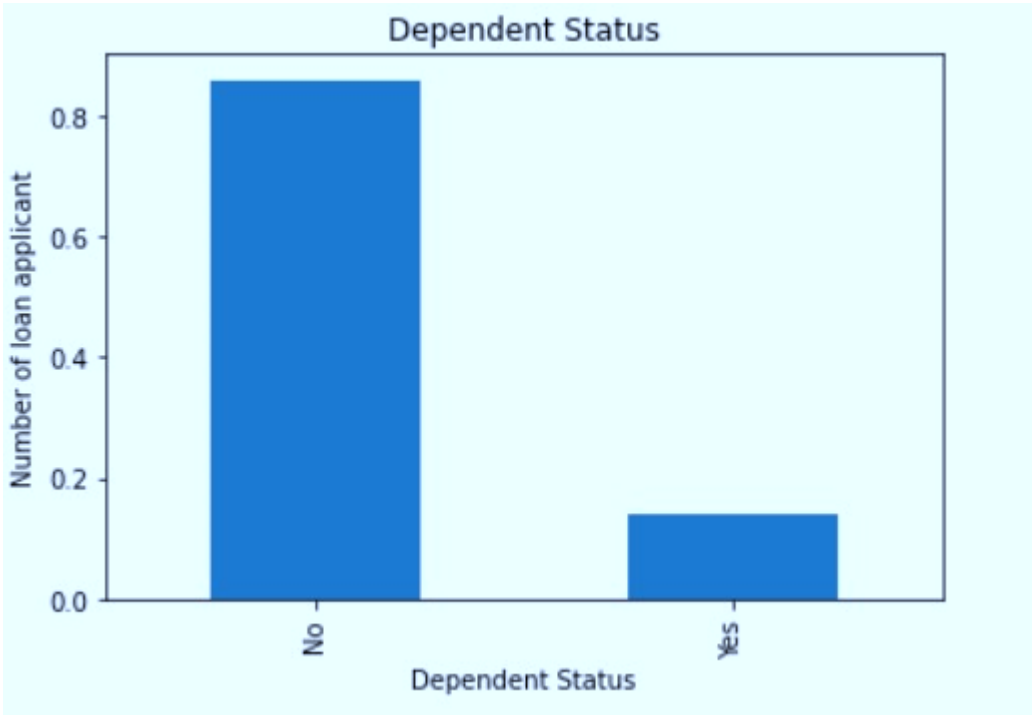
a. Find out the number of male and female in loan applicants' data.



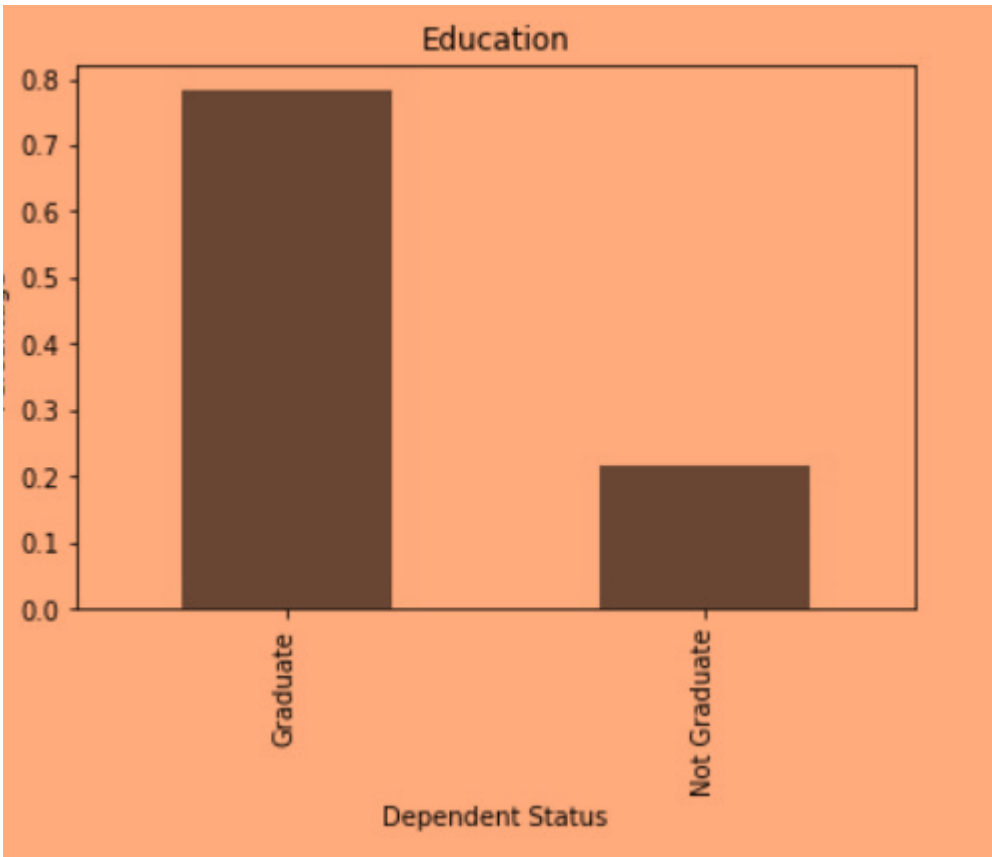
b. Find out the number of married and unmarried loan applicants.



c. Find out the overall dependent status in the dataset.



d. Find the count how many loan applicants are graduate and non-graduate.



e. Find out the count how many loan applicants property lies in urban, rural and semi-urban areas.



2. What conclusions are you derived from the single variable analysis?

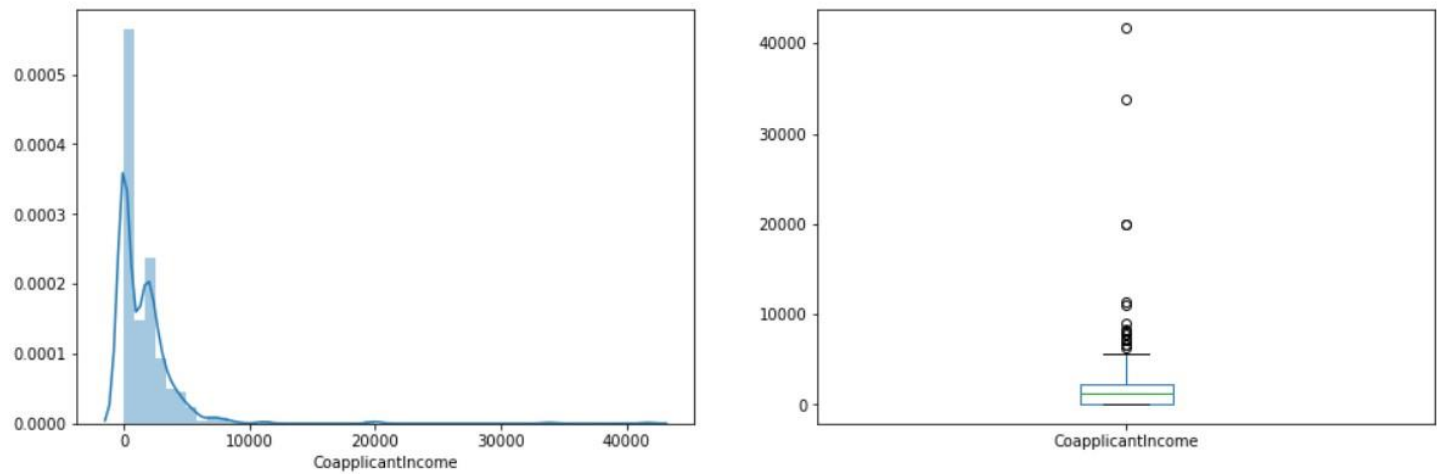
Based on my analysis, I may conclude the following points:

- a) Among 582 people only 14% are self-employed and rest of the 86% are not self-employed.
- b) Around 84% applicants have repaid their debts.
- c) Around 58% people have 0 dependent.
- d) Around 17% people have 1 dependent.
- e) Around 17% people have 2 dependents.
- f) Around 8% people have 3+ dependents.
- g) The loan of 422(around 69%) people out of 614 was approved.
- h) There are male (81%) and female (19%) have applied for loan.
- i) 65 % of married people and 35 % of unmarried people have applied for loan.
- j) 78 % people are graduated and 22% are not graduated.
- k) Around 38%, 33% and 29% are from semi-urban, urban and rural area respectively.

3. To visualize and plot the distribution plot of all numerical attributes of the given train dataset i.e. ApplicantIncome, CoApplicantIncome and LoanAmount.

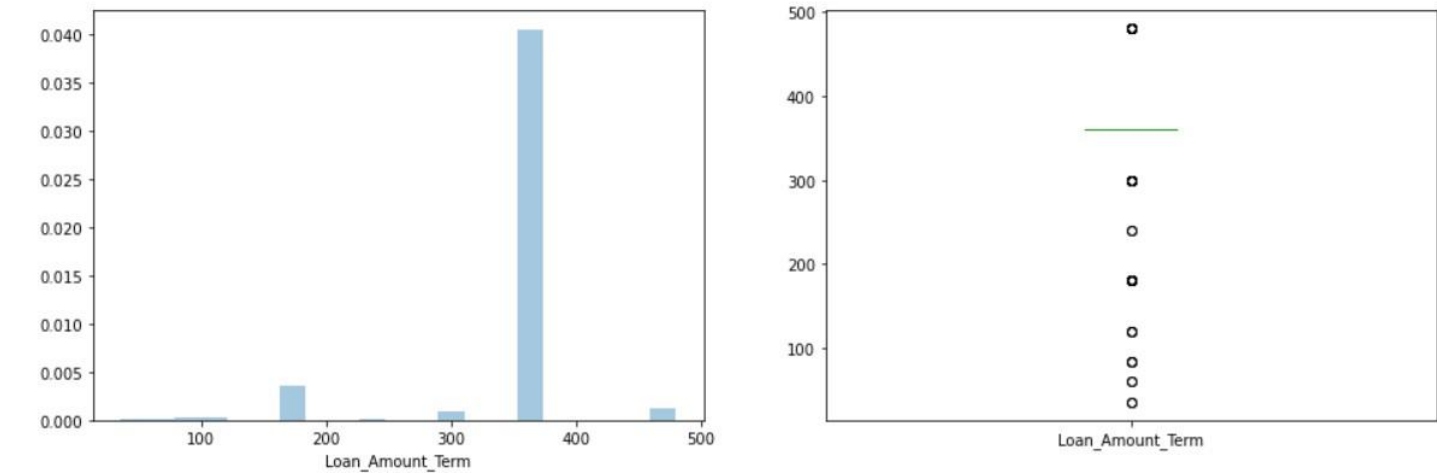
a.

Coapplicant Income distribution



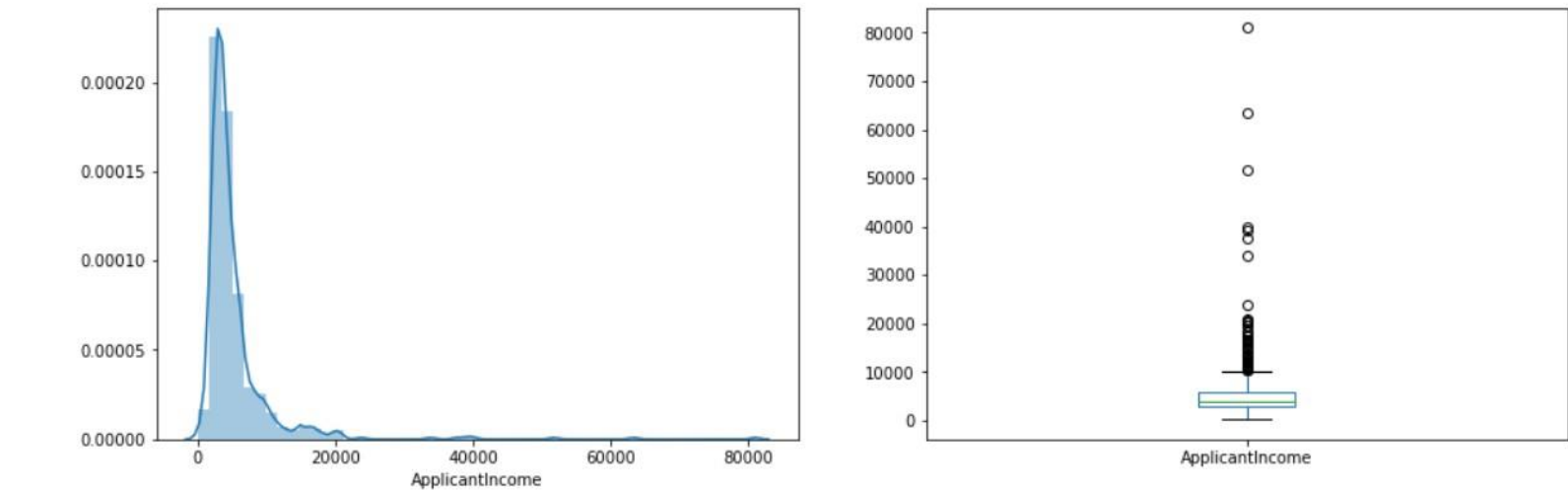
b.

Loan Amount Term Distribution



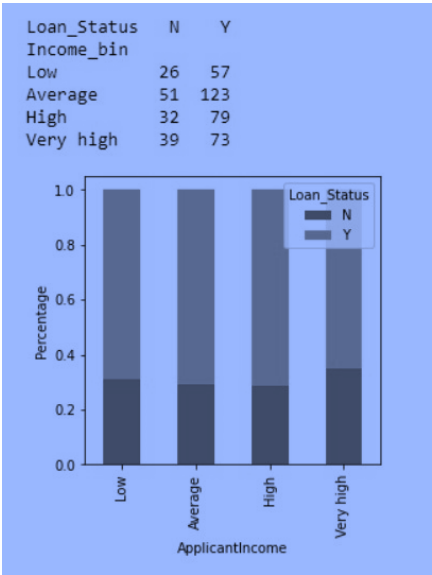
c.

ApplicantIncome distribution

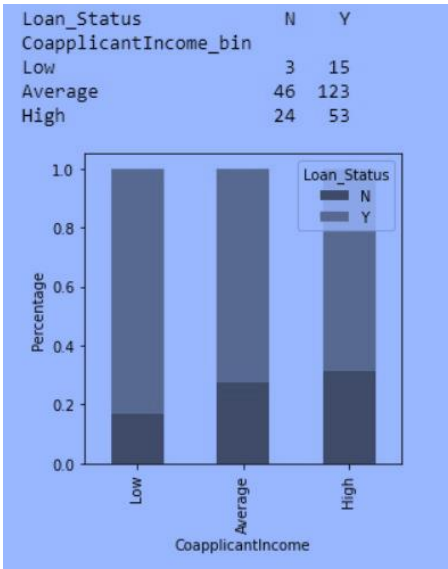


4. Also visualize and plot the Question-1 based on Loan\_status of loan applicant (Multi variable analysis).

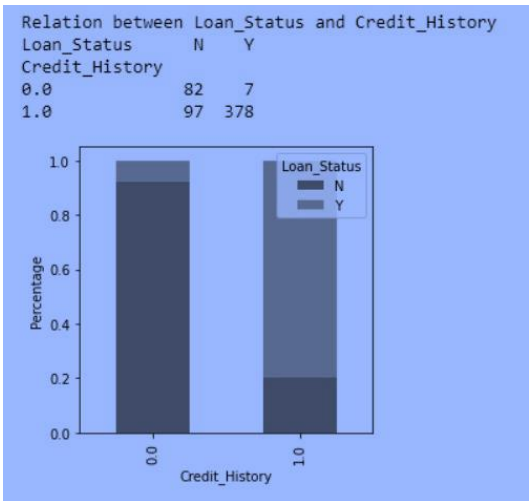
a. Loan Status VS Applicant Income



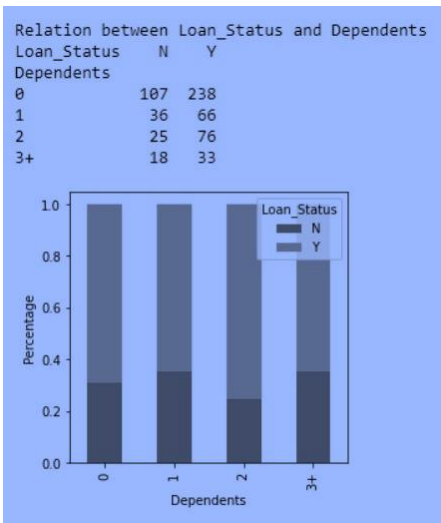
b. Loan Status VS Co-applicant Income



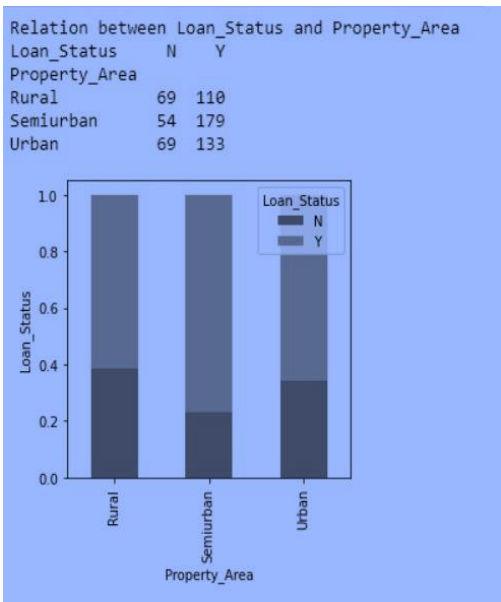
c. Loan Status and Credit History



d. Loan Status VS Dependents



e. Loan Status VS Property Area



5. What conclusions are you derived from the multi variable analysis?

Based on my Analysis, I can conclude the following points

- a) Proportion of male applicants is higher for the approved loans.
- b) Proportion of married applicants is higher for the approved loans.

- c) Proportion of graduate applicants is higher for the approved loans.
- d) People with credit history as 1 are more likely to get their loans approved.
- e) Proportion of loans getting approved in semi-urban area is higher as compared to that in rural and urban areas.
- f) Proportion of loans getting approved for applicants having low total income is very less as compared to that of applicants with average, high and very high income.
- g) Proportion of approved loans is higher for low and average loan amount as compared to that of high loan amount.

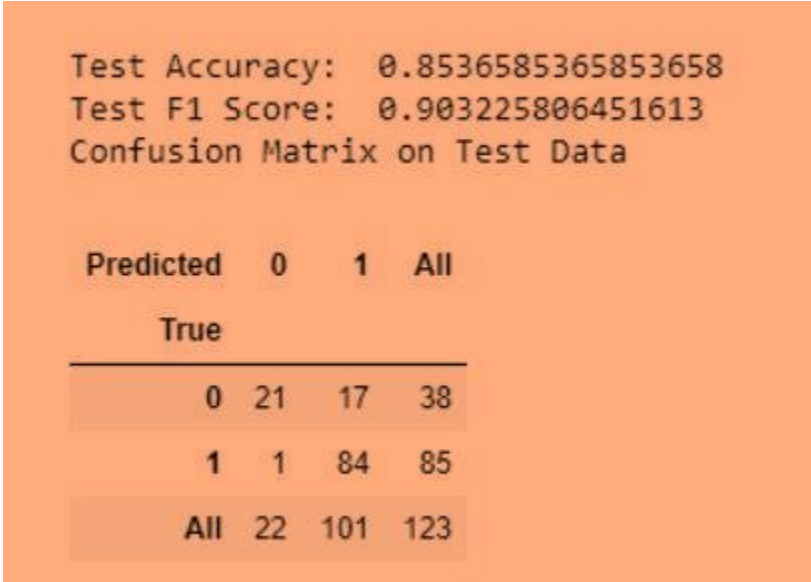
[https://github.com/pursharthgujral/170029\\_Big-Data-W8\\_Activity2/blob/main/SectionA.py](https://github.com/pursharthgujral/170029_Big-Data-W8_Activity2/blob/main/SectionA.py)

Section B: Market Basket Analysis

For question from 1 to 8, refer the following link:

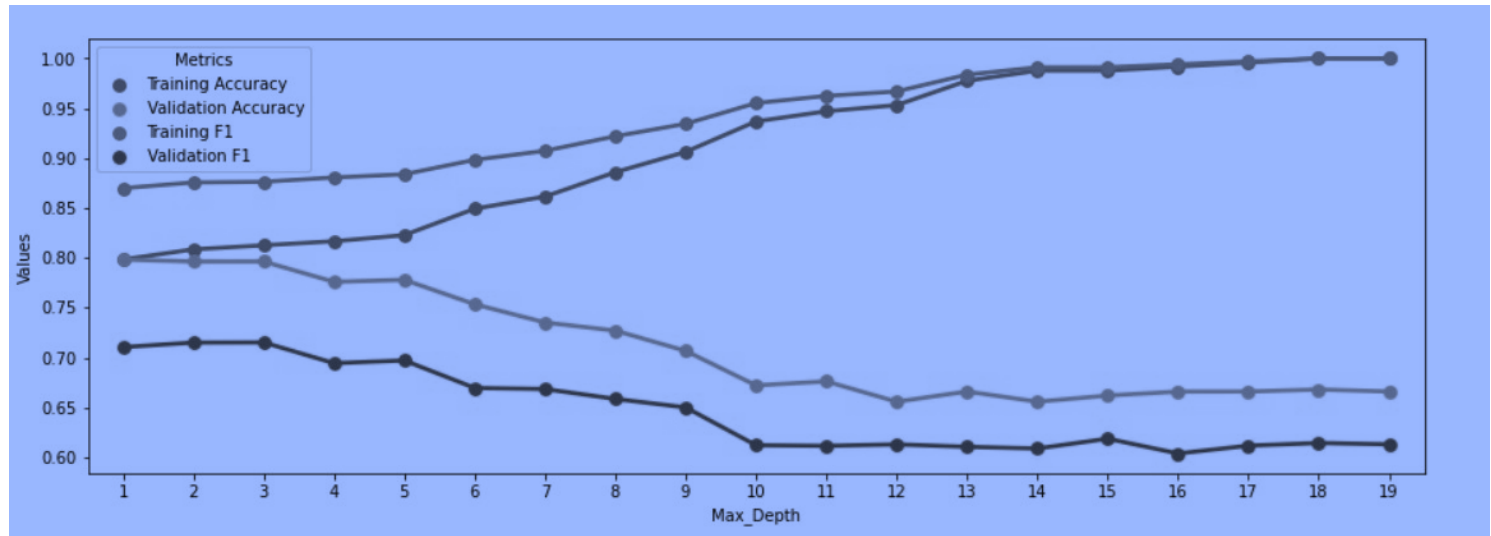
[https://github.com/pursharthgujral/170029\\_Big-Data-W8\\_Activity2/blob/main/SectionB.py](https://github.com/pursharthgujral/170029_Big-Data-W8_Activity2/blob/main/SectionB.py)

IX. Calculate the accuracy of the model using the accuracy score function.

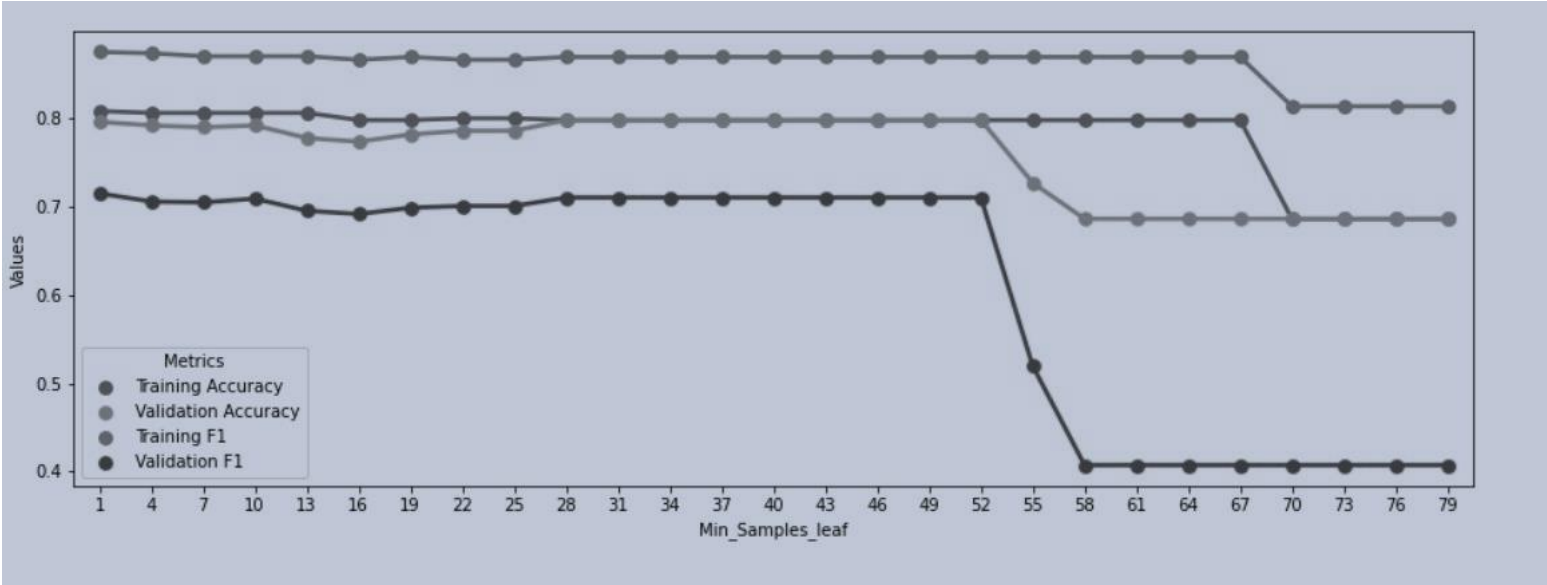


X. Visualizing Decision Trees

- Tree max Depth



- Tree Min Sample Leaf



- Tree PNG

