Loan Approval Prediction

22nd September 2021

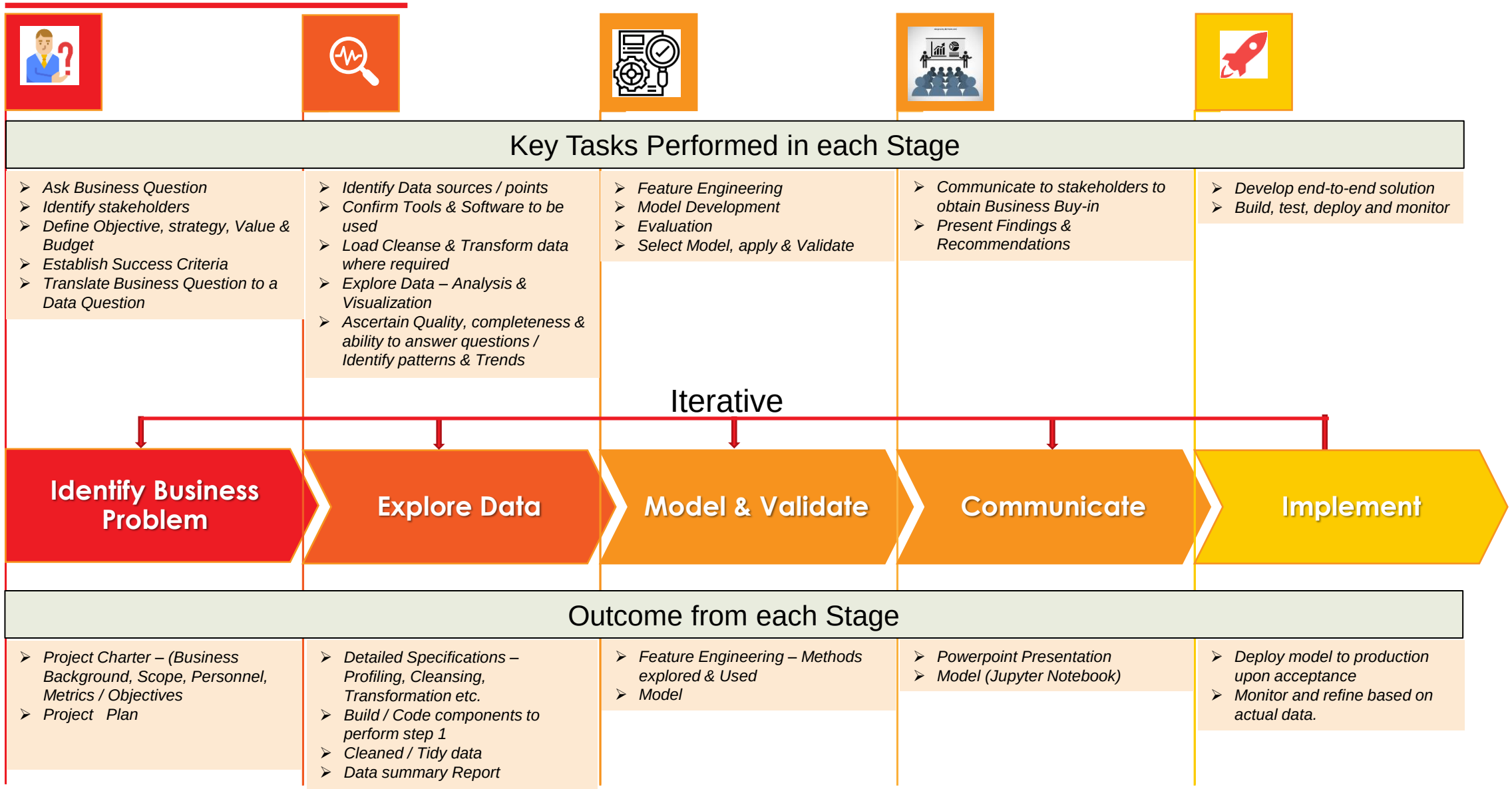
Table of Contents

Engagement Background

A Leading Bank wants a Model to be built to predict the Loan approval for a client based on a set of features.

The Organization has engaged the services of Sai Science Pte Ltd for the same.

Data Science Process



Identify Business Problem

A **Leading Bank** is expanding its **retail segment** specifically in **consumer Loans** and hence needs a Robust Model to predict whether the consumer should be provided a Loan based on a set of features. The Organization has engaged the services of Sai Science Pte Ltd and needs the following done

- 1-Exploratory Data Analysis (EDA)
- 2-Data Pre-processing
- 3-Model Training, Development and Evaluation
- 4-Prediction using the model on Test Data

Data Problem

This is a **binary classification problem** where we must predict whether a loan will be approved or not.

The dependent variable or target variable is the Loan approval Status, while the rest are independent variable or features. We need to develop a model using the features to predict the target variable.

Identify Business Problem – Stakeholders

Key Client Stakeholders	Vendor Stakeholders
Client Engagement Director	Engagement Director
Client Project Manager	Project Manager
Client BA / SME	Lead Data Scientist
Business Sponsor – Head of New Business	Data Architect
Technology Sponsor – Head of Technology	Developers

Key Assumptions

- The client team would make themselves available to clarify any questions on the data set. (2 sessions of 2 hours each have been planned to tackle such questions)
- If there is a change in the # of Features, the model needs to be re-trained & validated.
- As discussed, and agreed upfront, this is a **3-weeks** engagement
- The data set is complete and a significant representation.
- The output will be the Model **(Jupyter notebook) & a Powerpoint Presentation**

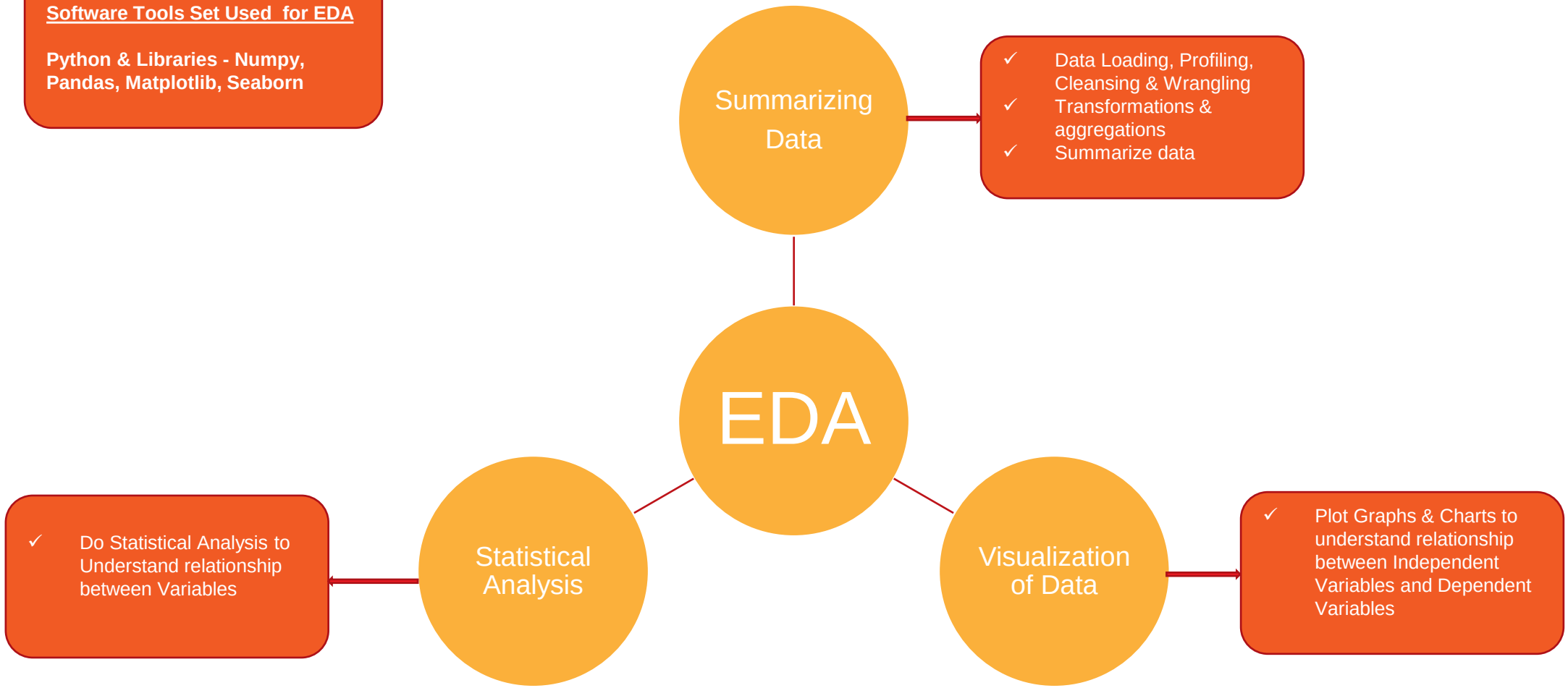
Understanding the data

- ✓ **Data Source –** CSV file (Comma separated Values)
- ✓ **2 sets –**
  - ✓ **1 train.csv - For Training and Developing the Model.**
  - ✓ **2 Test – For Predicting the loan approval status based on the trained model.**
- ✓ **# of records**
  - ✓ **Training Data set – 614**
  - ✓ **Testing Data set - 367**
- ✓ **# of Features / Variables – 11**
- ✓ **Type of Data – Loan approval data –** To predict the whether Loan application of a customer will be approved or not

Explore Data – Key Components of EDA Considered

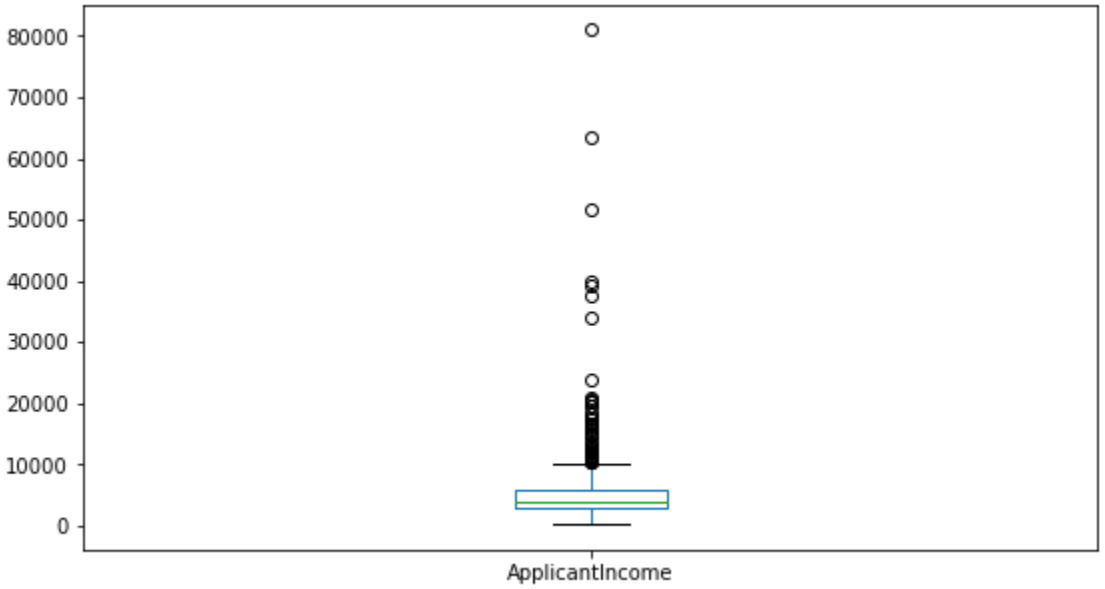
Software Tools Set Used for EDA

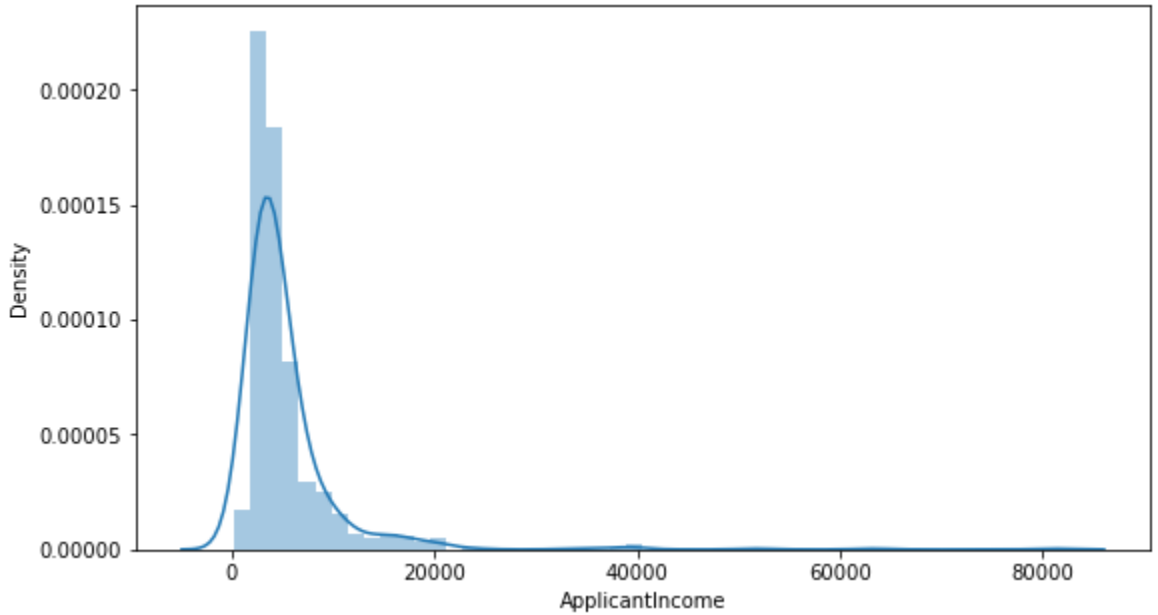
Python & Libraries - Numpy, Pandas, Matplotlib, Seaborn



Explore Data – Key Numerical Data (Visualization)

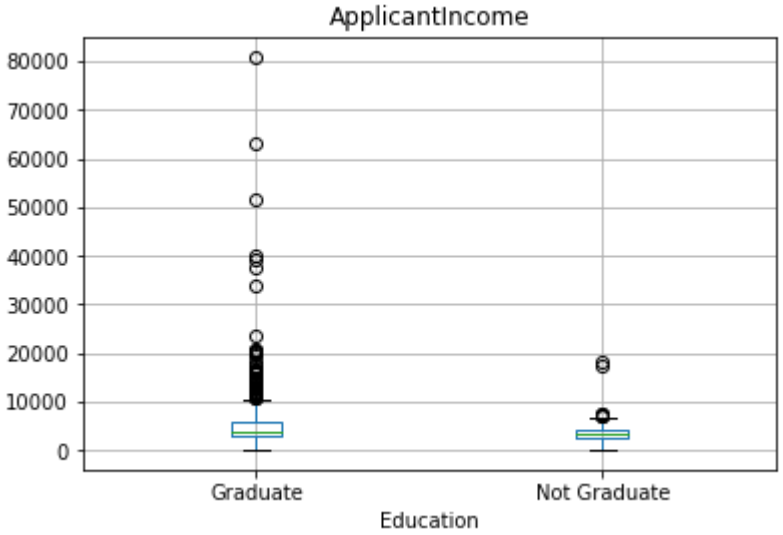
Applicant Income Distribution





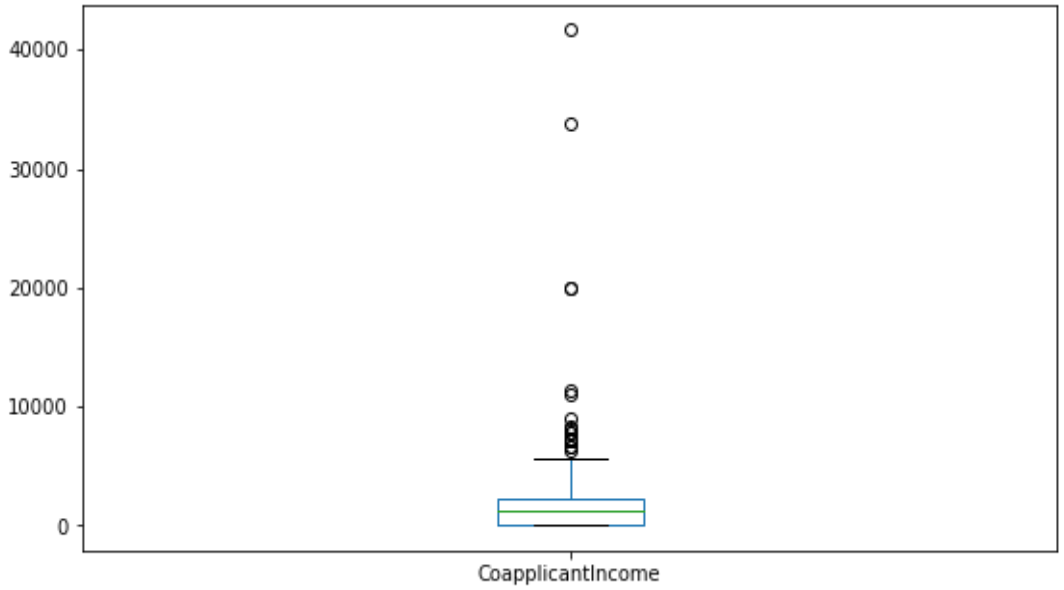
- Distribution of Data is more Towards Left, Distribution is Right Skewed. (Positive Skewness)
- Algorithm Works Better if the Data is Normally Distributed.
- The Boxplots Represents the Presence of Outliers Values, Data contains many Outliers.

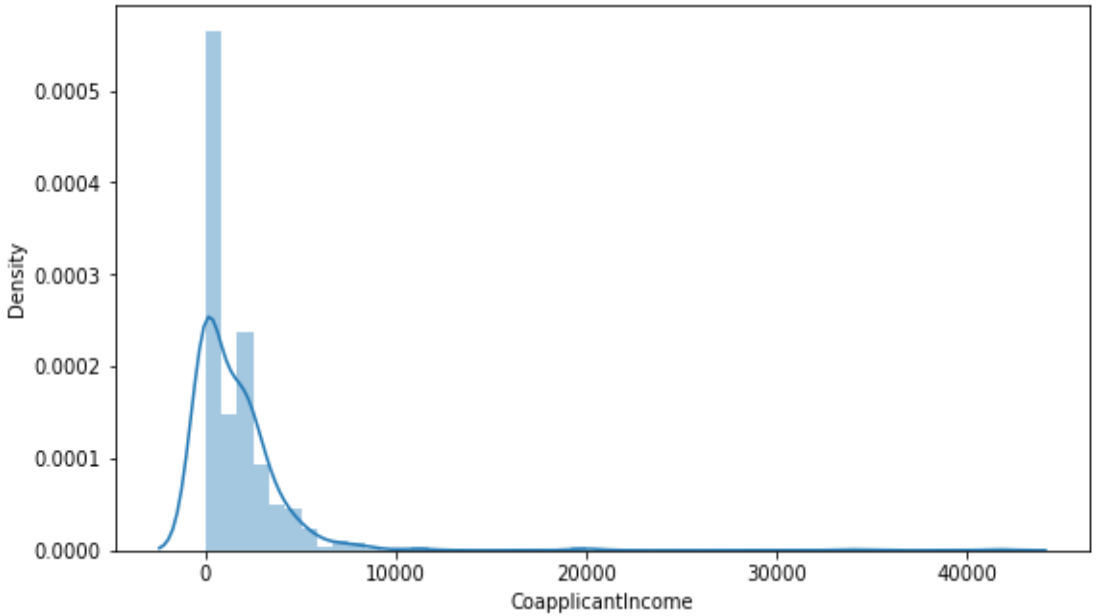
Applicant Income By Education Level



- Graduates Have higher Income

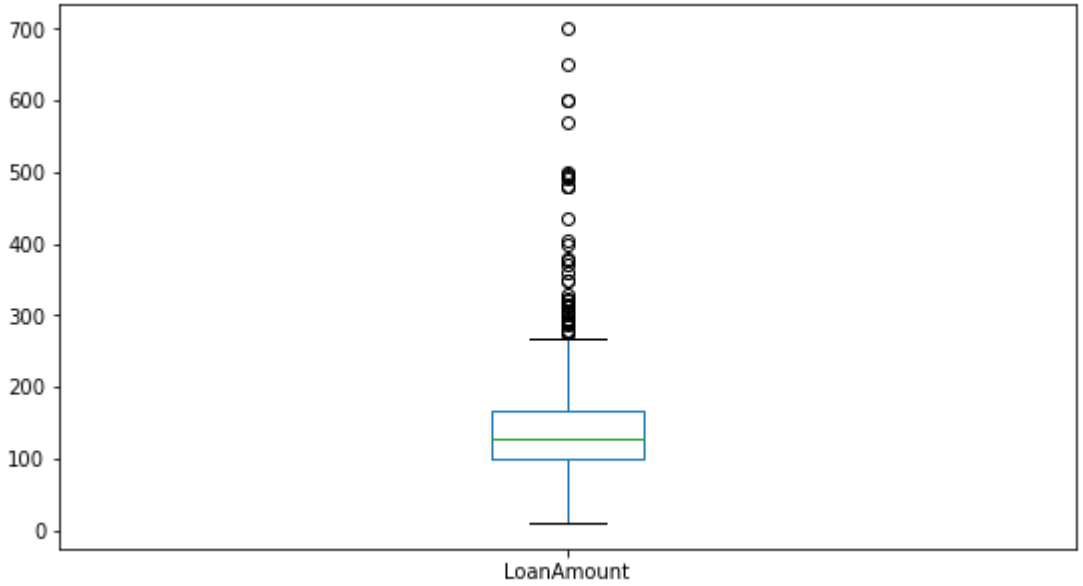
Co-Applicant Income Distribution

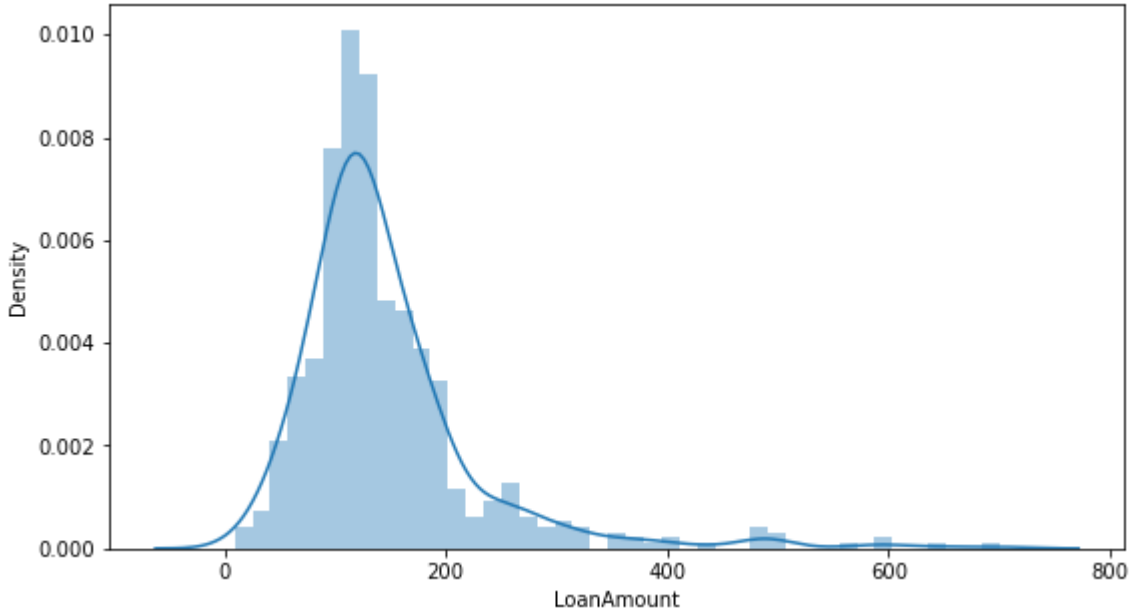




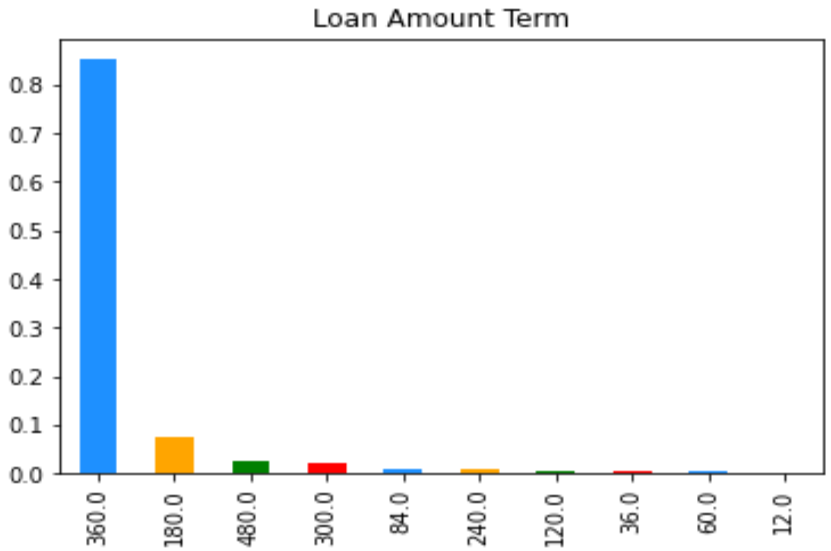
- Co-applicant Income is Right Skewed and consists of lots of Outliers.

Plot of Loan Amount





Loan Amount Term



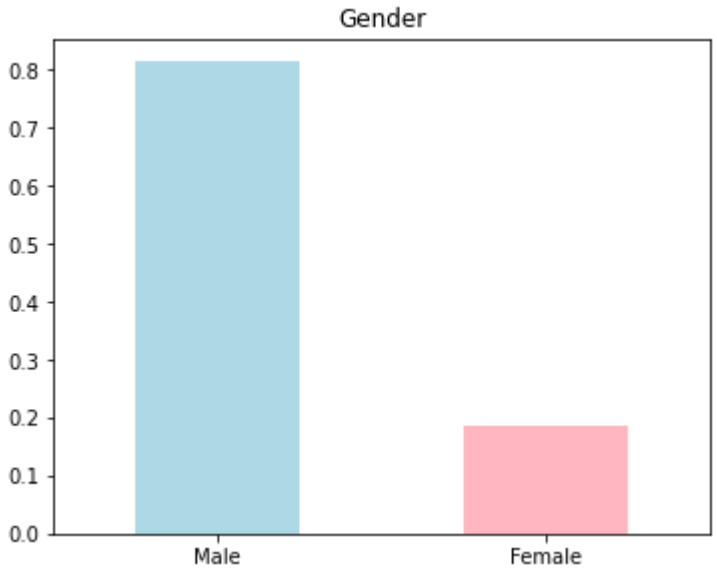
- Around 85% of Loans are of 360 Months (30 Years)

Explore Data – Key Categorical Data (Visualization)

Plot of Loan Approval Status

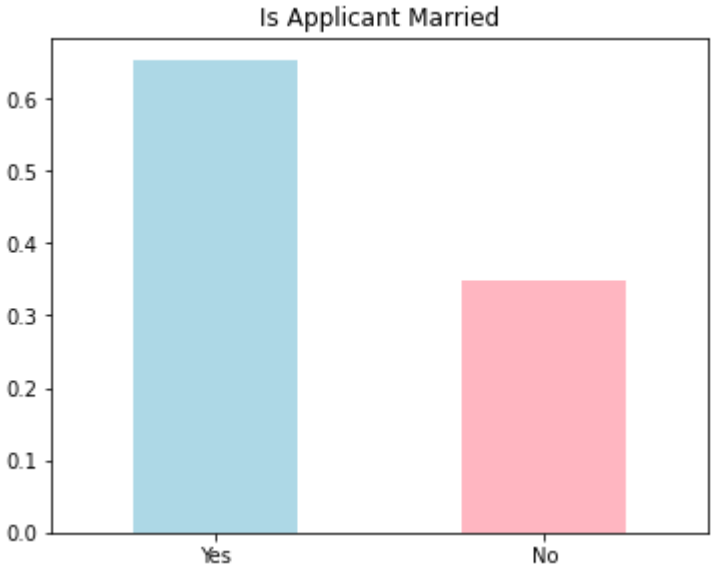
- Around 69% of Loan applications are approved

Applicant Gender Plot



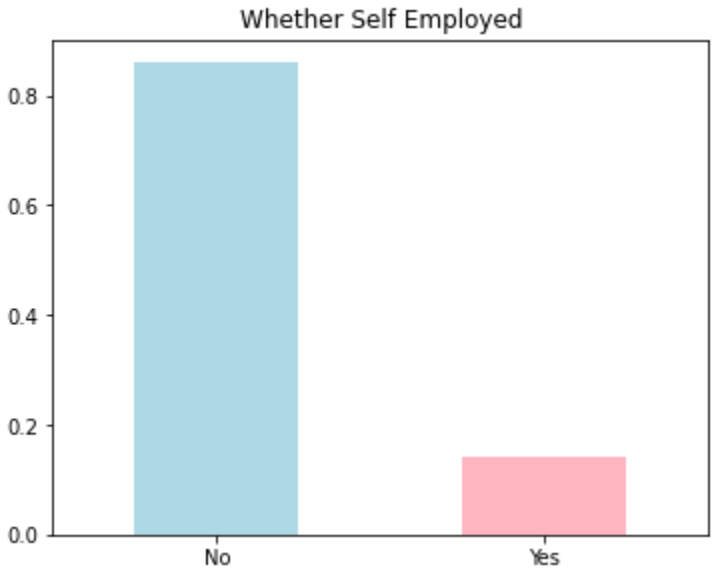
- Around 80% of Loan applicants are Males

Applicant Marital Status Plot



- Close to 65% applicants are married

Applicant Employment Status

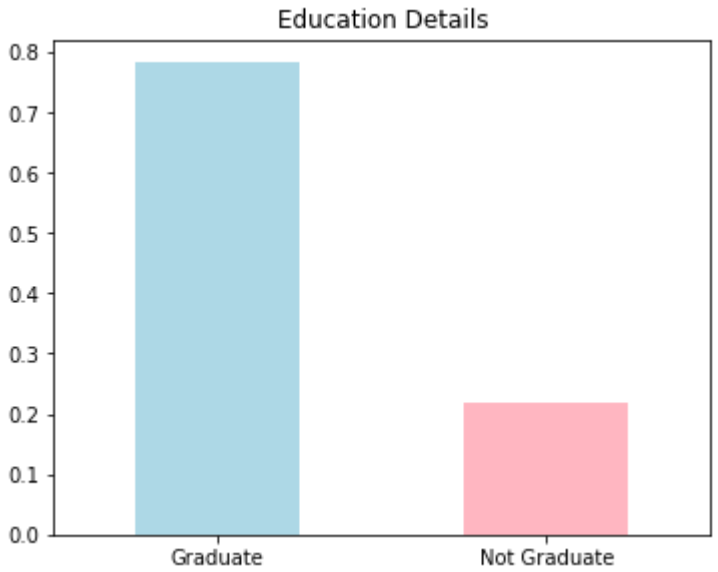


- Close to 20% applicants are self-employed

Applicant Credit Score

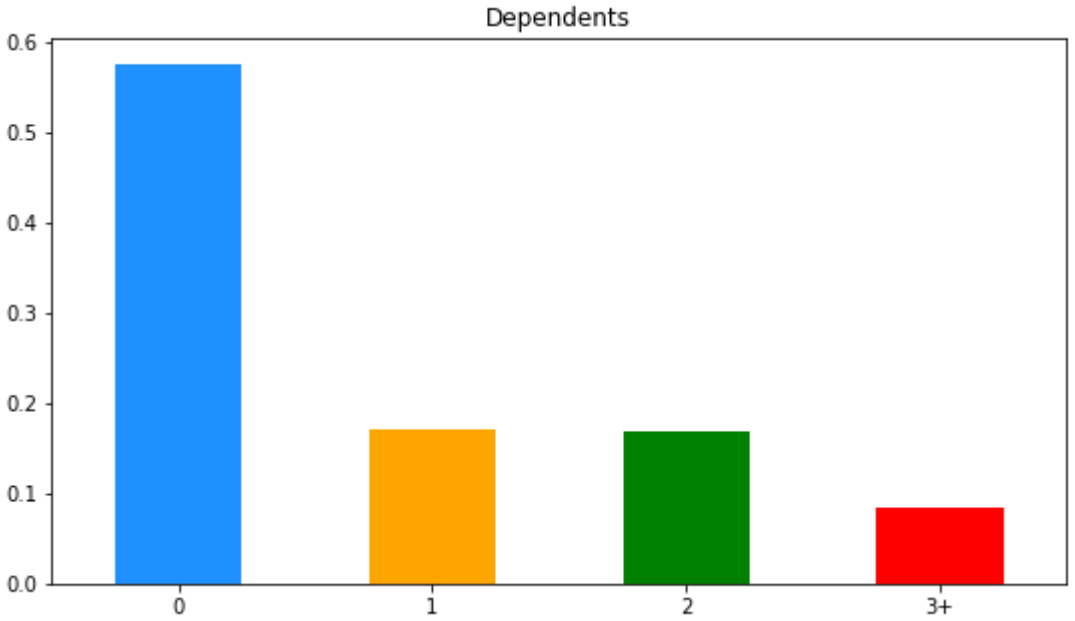
- Close to 85% of the applicants have a good credit score

Applicant Education Details



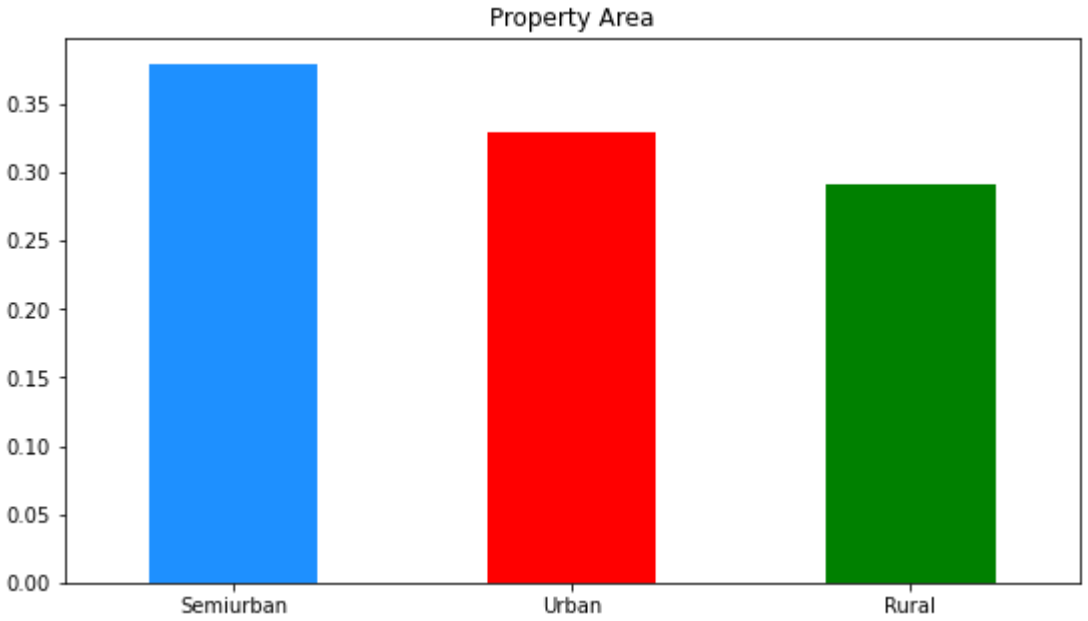
- Close to 80% of the applicants are Graduates

Plot of Applicants by # of Dependents



- Majority of Applicants don’t have dependents

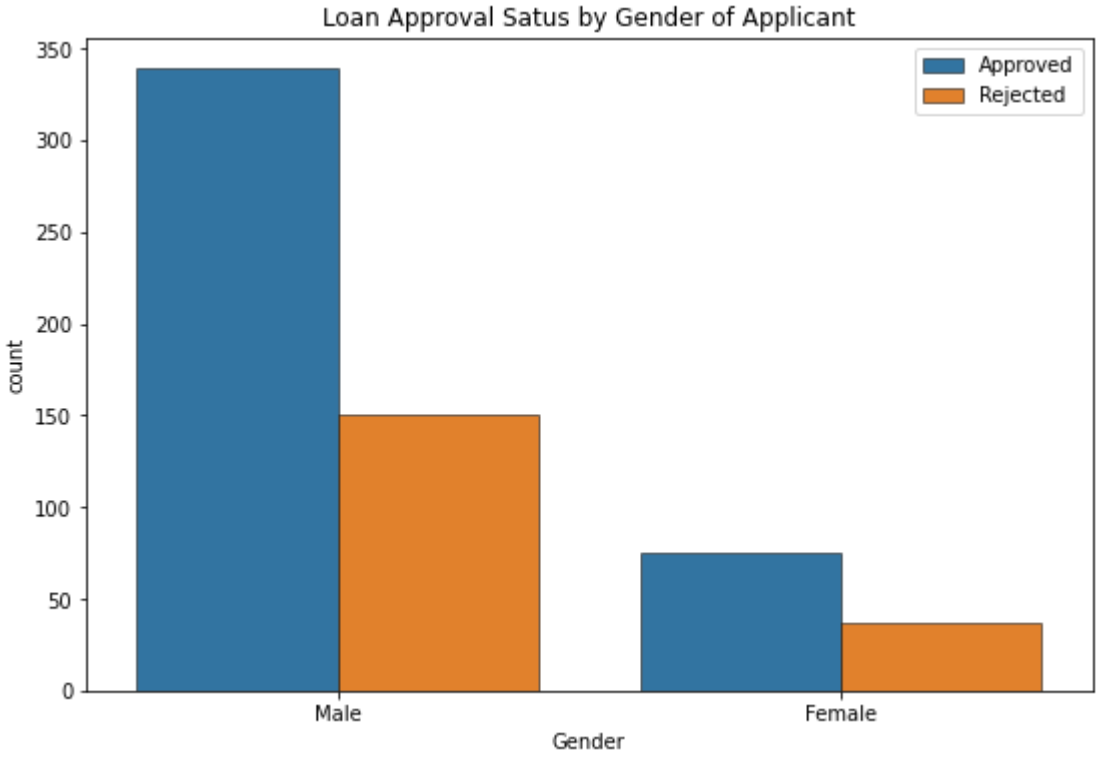
Plot of Applicants by Property Area



- Majority of Applicants are from Semi-Urban & Urban Areas

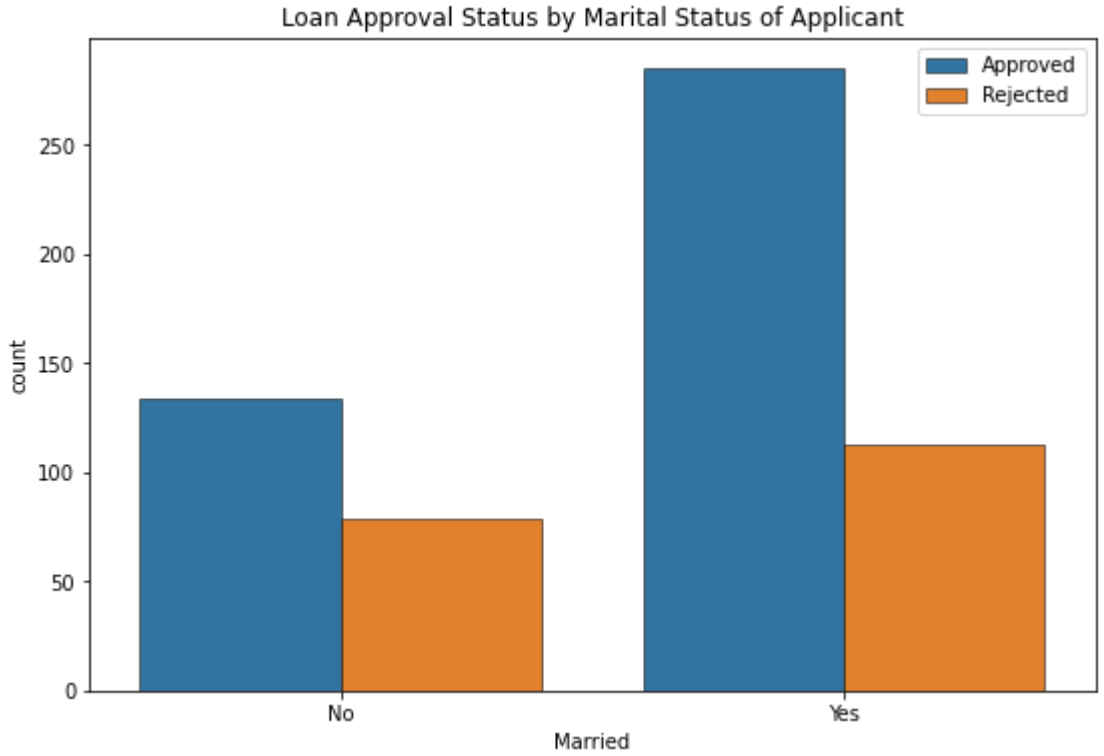
Explore Data – Key Categorical Data vs Independent Variable (Visualization)

Loan Approval Status by Gender of Applicant



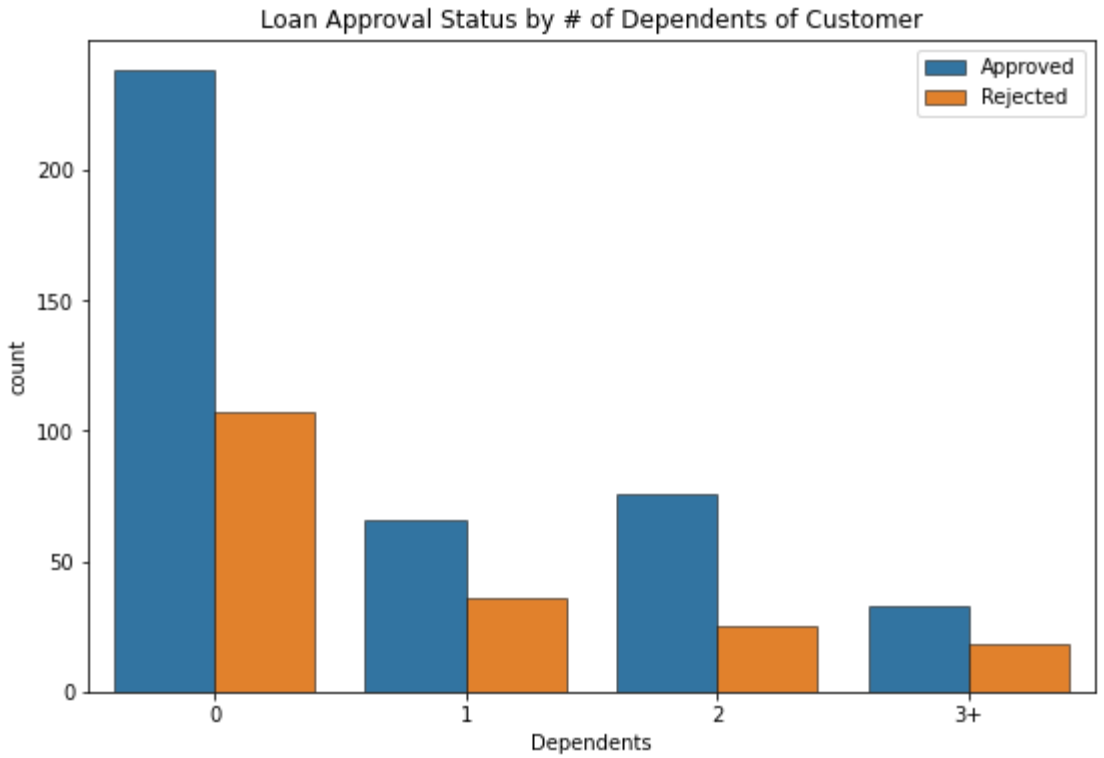
- Male Applicants have a higher rate of Approval

Loan Approval Status by Marital Status of Applicant



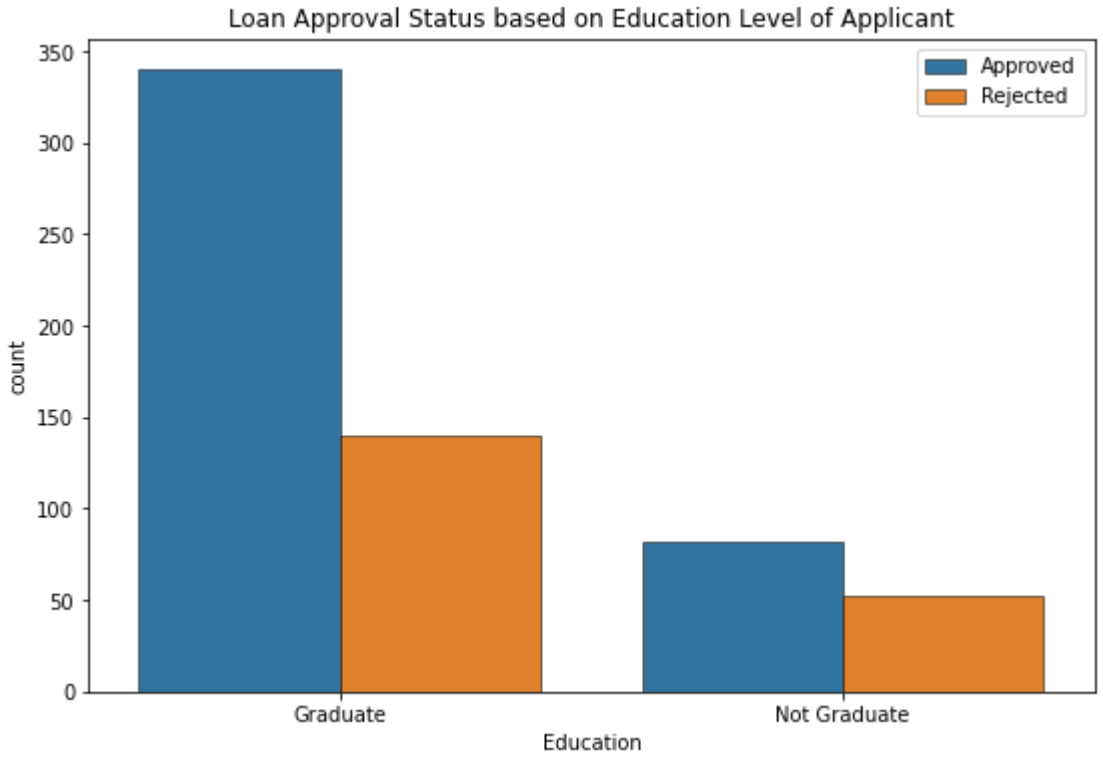
- Married People have a higher Approval Rate

Loan Approval Status by # of Dependents of Applicant



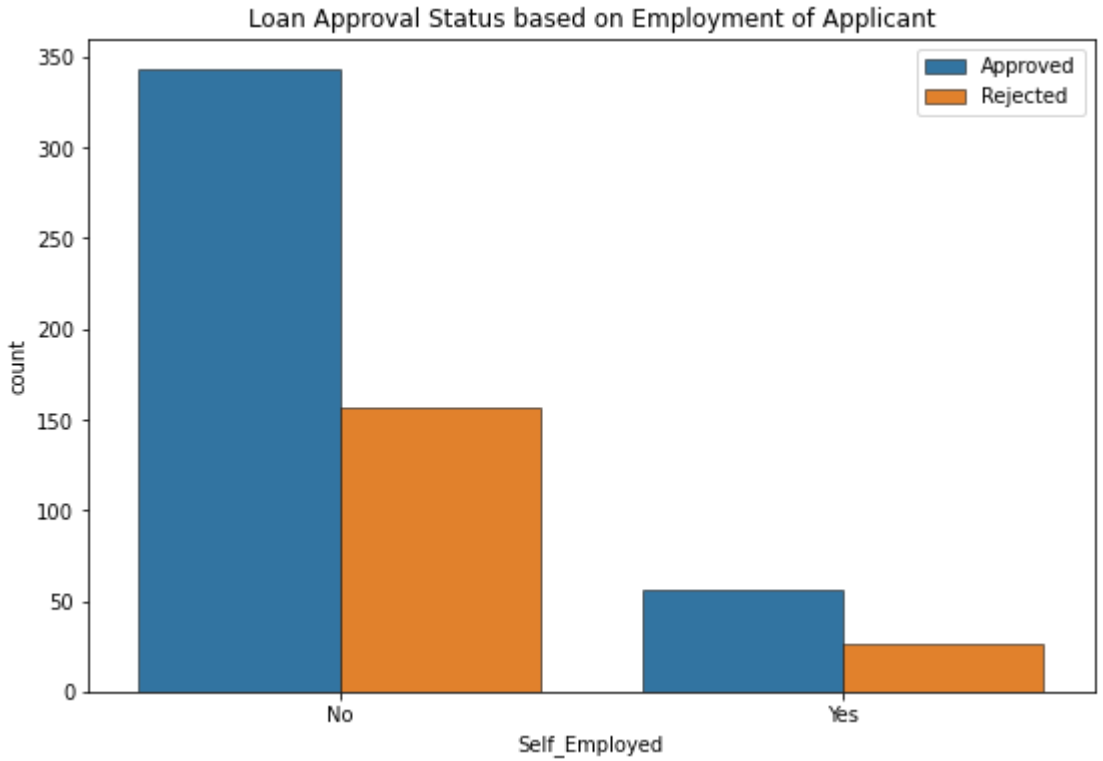
- Applicants without Dependents have a higher Approval Rate

Loan Approval Status by Education Level



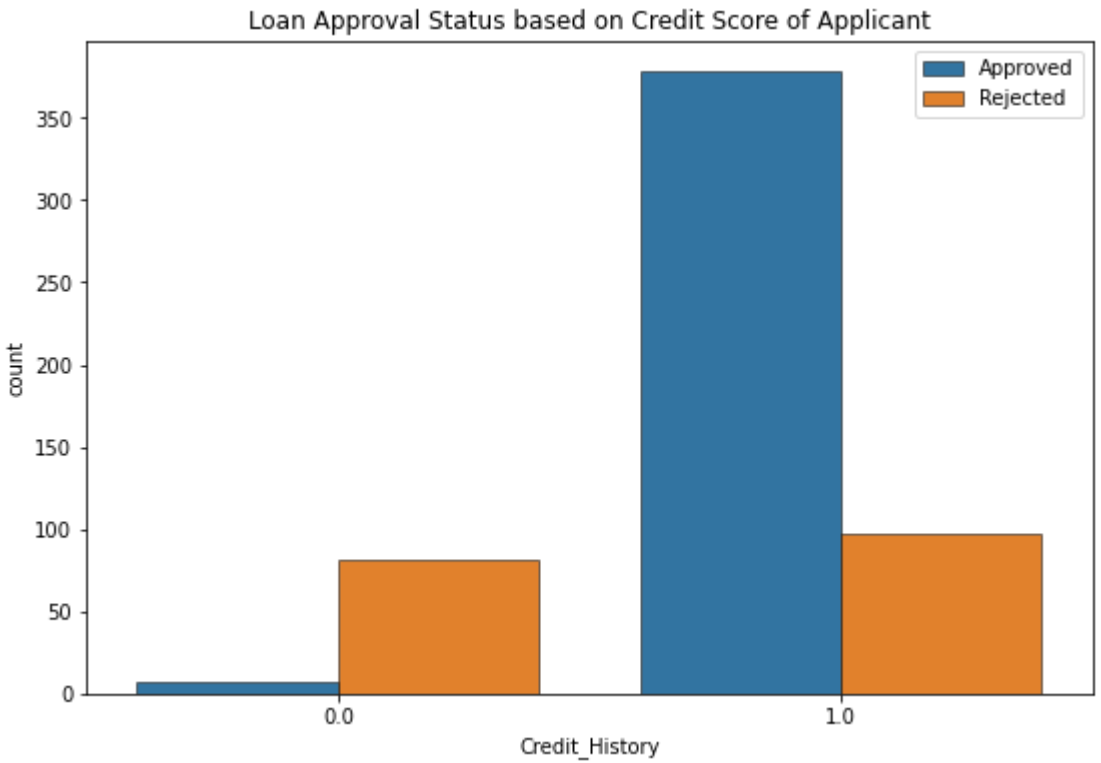
- Graduate Applicants have a higher Approval Rate

Loan Approval Status by Employment



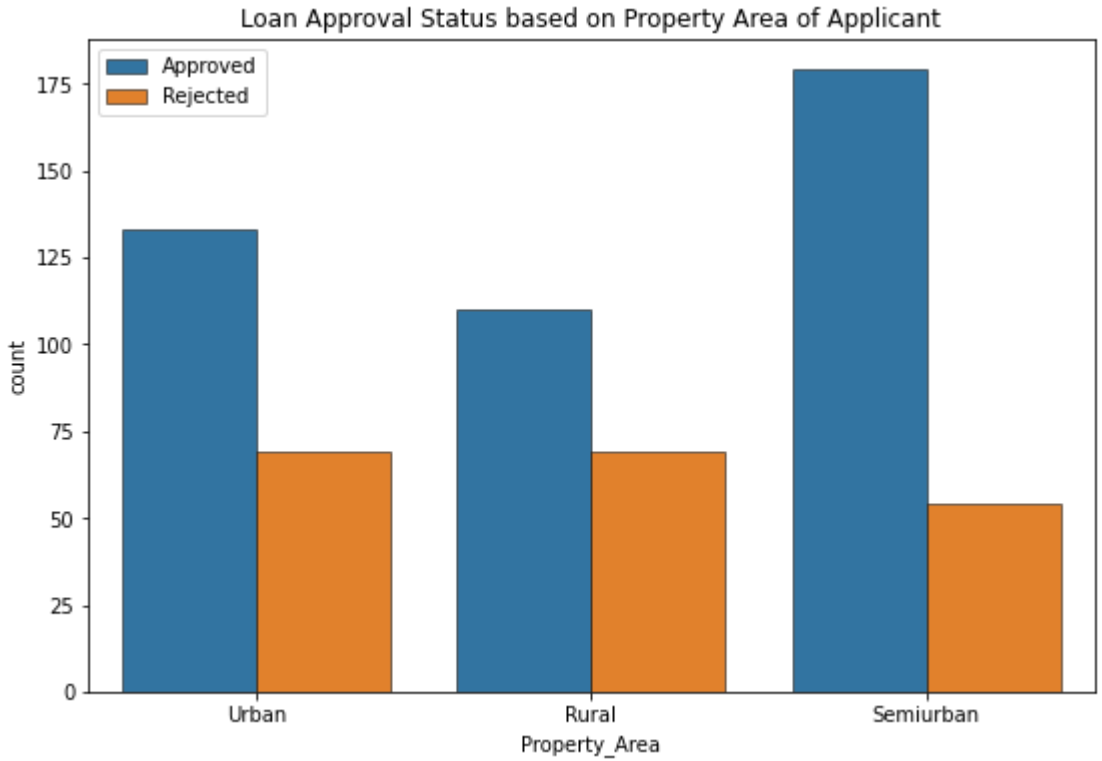
- Self Employed Applicants have a lower Approval Rate

Loan Approval Status by Credit Score



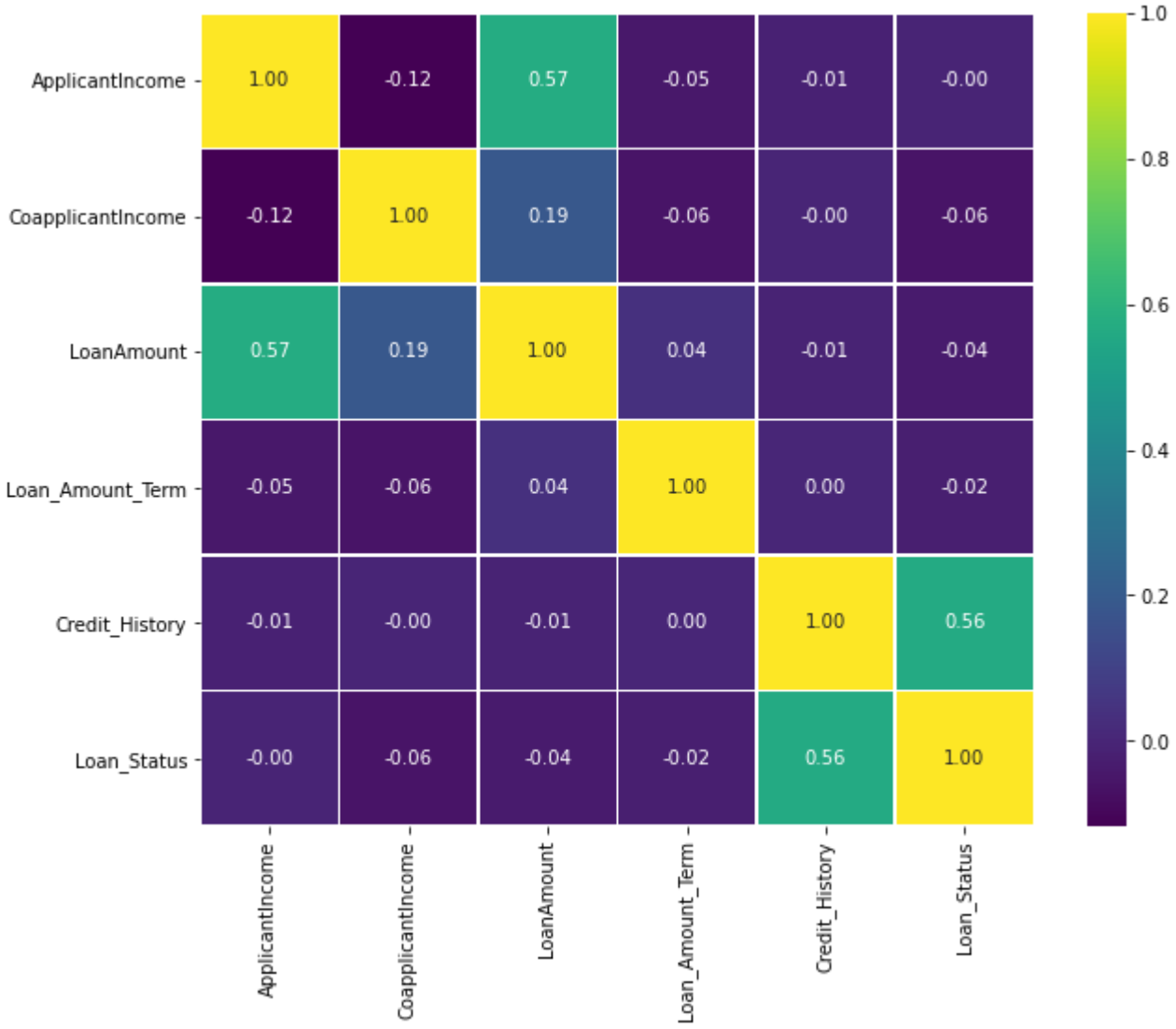
- Applications with a Good Credit Score have a higher Rate of Approval

Loan Approval Status by Property Area



- Applications with Property in Urban or Semi Urban areas have a higher approval Rate

Heat Map for Checking Correlation



Better Correlations :

1) Applicant Income and Loan Amount.
2) Credit History and Loan Status.

Data Pre-Processing

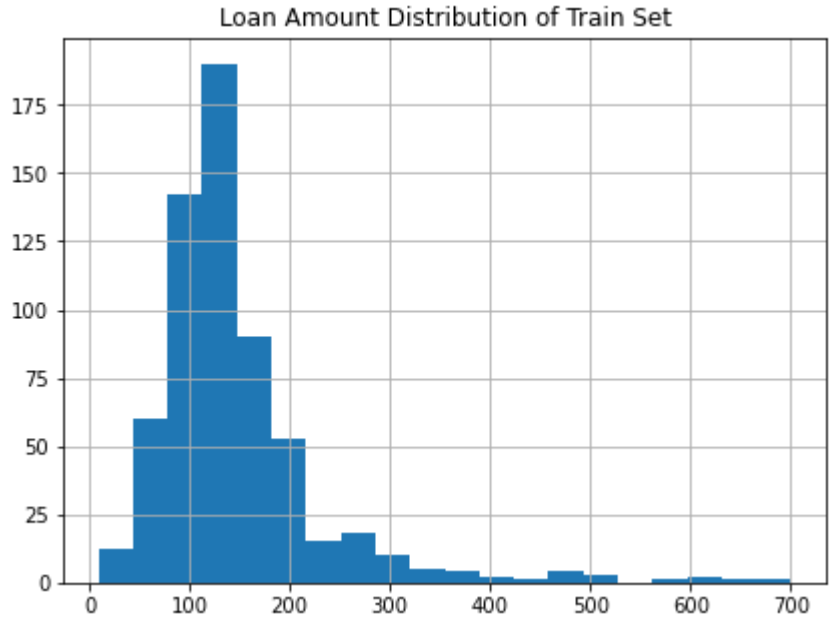
Null Values Handling

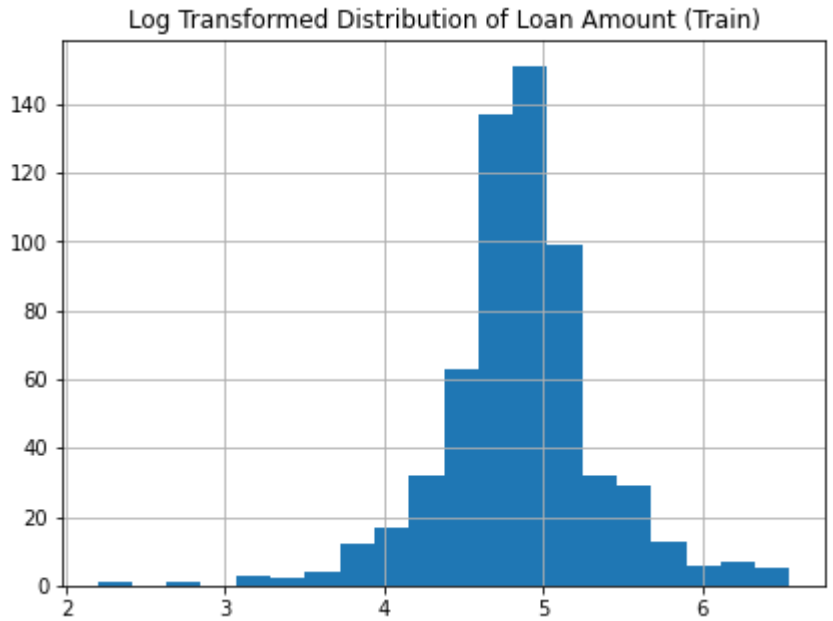
Based on the assessment of the missing values in the dataset, We will make the following changes to the data:

- ✓ If "Gender" is missing for a given row, we will impute with Male (most common answer).
- ✓ If "Married" is missing for a given row, we will impute with yes (most common answer).
- ✓ If "Dependents" is missing for a given row, we will impute with 0 (most common answer).
- ✓ If "Self_Employed" is missing for a given row, we will impute with no (most common answer).
- ✓ If "LoanAmount" is missing for a given row, we will impute with mean of data.
- ✓ If "Loan_Amount_Term" is missing for a given row, we will impute with 360 (most common answer).
- ✓ If "Credit_History" is missing for a given row, we will impute with 1.0 (most common answer).

Outlier Treatment

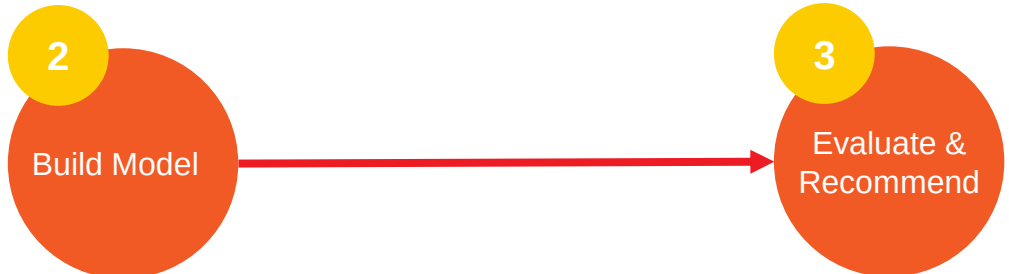
- Major Outliers were observed in the Loan Amount
- Outliers in the Data Set Often Affects the Mean and Standard Deviation by affecting the Distribution of Data.
- More Data is Present on Left and Long Tail is on Right. (Right Skewed : Positive Skewed)
- One Way to Remove Skewness is to Perform Log Transformation.
- Log Transformation does not Affect the Smaller Values but Reduces the Larger Values, so we get Similar to Normal Distribution.





Model Evaluation & Recommendation

- **70% of Data is used for Training Model**
- **30% of Data is used for Validation**



- Logistic Regression
- Decision Tree
- Gradient Boost
- Random Forest
- K-Nearest Neighbours
- SVM

- **Recommended Model is**
- **Random Forest**

Model Recommendation Basis

- **Cross Validation Score** - It will be used to perform the evaluation, taking the dataset and cross-validation configuration and returning a list of scores calculated for each fold.
- **Accuracy Score** – Model Accuracy
- **Precision** – Precision score is a useful measure of success of prediction when the classes are imbalanced. It represents the model's ability to correctly predict the positives out of all the positive prediction it made.

	Model	Cross Validation Score (Classification Performance)	Accuracy Score	Precision
0	Logistic Regression	0.77	0.76	0.73
1	Decision Tree	0.72	0.70	0.74
2	Gradient Boost	0.75	0.76	0.74
3	Random Forest	0.80	0.78	0.75
4	KNN	0.62	0.56	0.62
5	SVM	0.65	0.65	0.65

- *Since we Use Cross Validation Score, Accuracy Score & Precision as the metrics for comparison,* ***Random Forest*** *emerges as the Recommended Model.*

Thank You