Real Estate Loan Approval Prediction

SWE4

Abdelrahman Tamer Adel	**20191702042**
Abdelwahab Mohamed Nabil	**20191702045**
Seif Hossam ElDeen	**20191702038**
Omar Mohamed Mohamed	**20191702050**
Mai Mansour Mohamed	**20191702071**
Sema Saeed Saad	**20191702034**

- Project Description
    - The dataset is generated from Dream Housing company, it deals with all home loans in urban, semiurban, and rural areas.
    - The Real Estate Loan Approval Prediction project involves collecting and analysing data from past loan applications and developing a machine learning model to predict loan approval outcomes.

- Dataset and variables description
    - Loan_ID: Unique Loan ID
    - Gender: Male/ Female
    - Married: Applicant married (Y/N)
    - Dependents: Number of dependents or people responsible from the applicant
    - Education: Applicant Education (Graduate/ Undergraduate)
    - Self_Employed: Self-employed (Y / N)
    - ApplicantIncome: Applicant income
    - CoapplicantIncome: Coapplicant income
    - LoanAmount: Loan amount took by applicant in thousands
    - Loan_Amount_Term: Term of the loan in months
    - Credit_History: report about applicant credit history
    - Property_Area: Urban/ Semi-Urban/ Rural
    - Loan_Status: (Target) Loan approved? (Y/N)

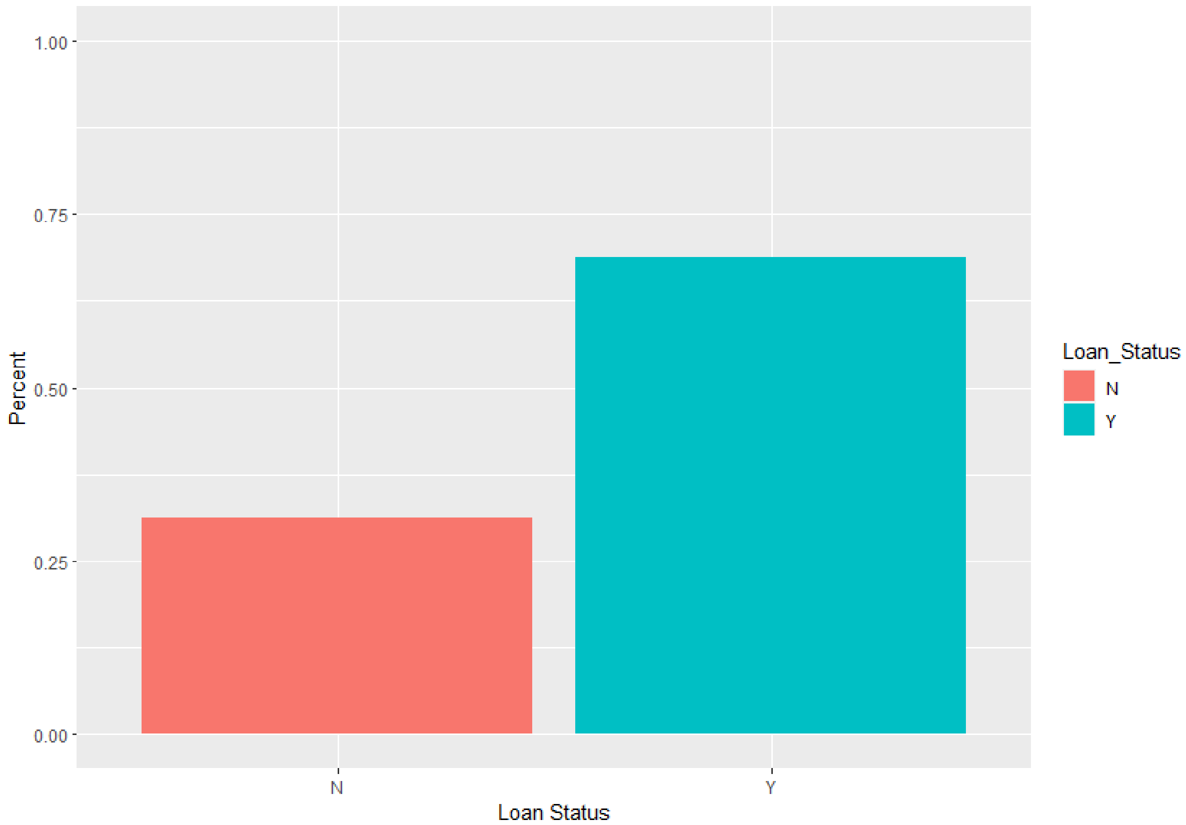
- Problem Definition
    - The Real Estate Loan Approval Prediction project addresses the problem of lengthy and complex loan approval processes for real estate loans.

- Problem Objectives
    - The project aims to automate the loan approval process by developing a machine learning model that can predict loan approval outcomes based on historical loan application data.

- Data Visualisations

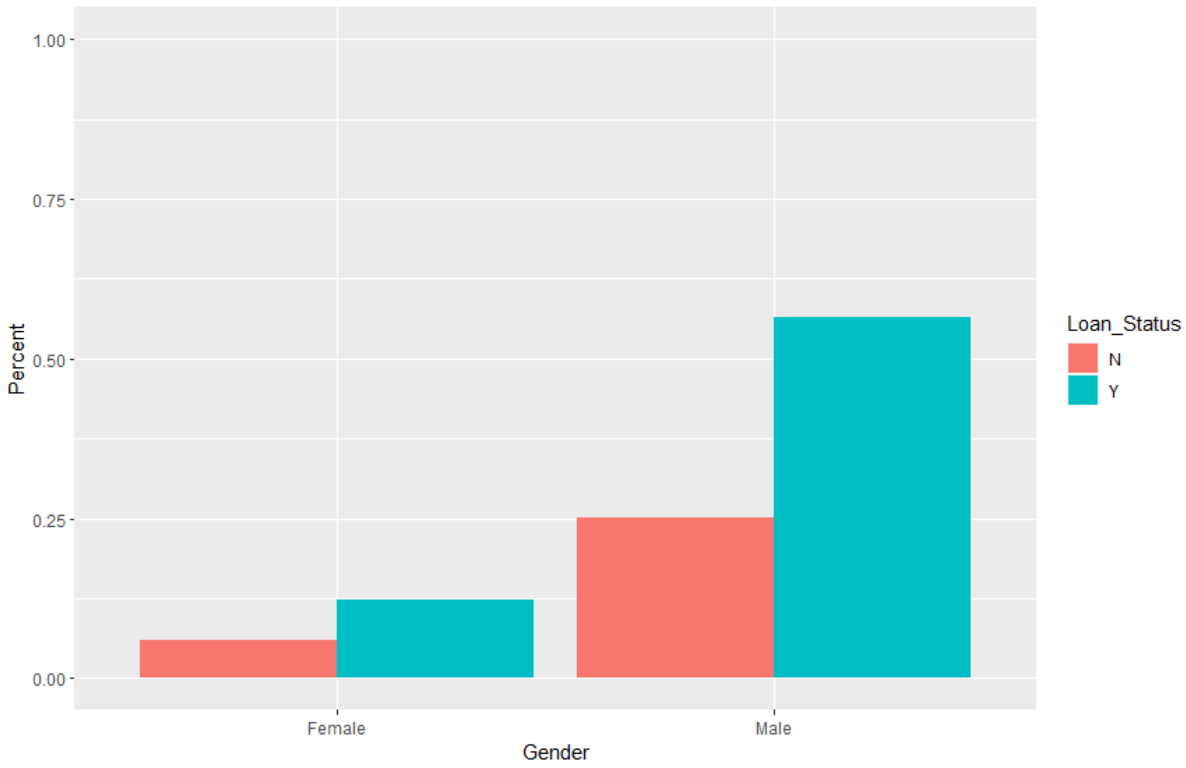
Plot for percentage of each Loan Status.

According to the insight it is clear that the percentage of accepted loans is much higher than the rejected ones (Y ~ 70%, N ~ 30%).



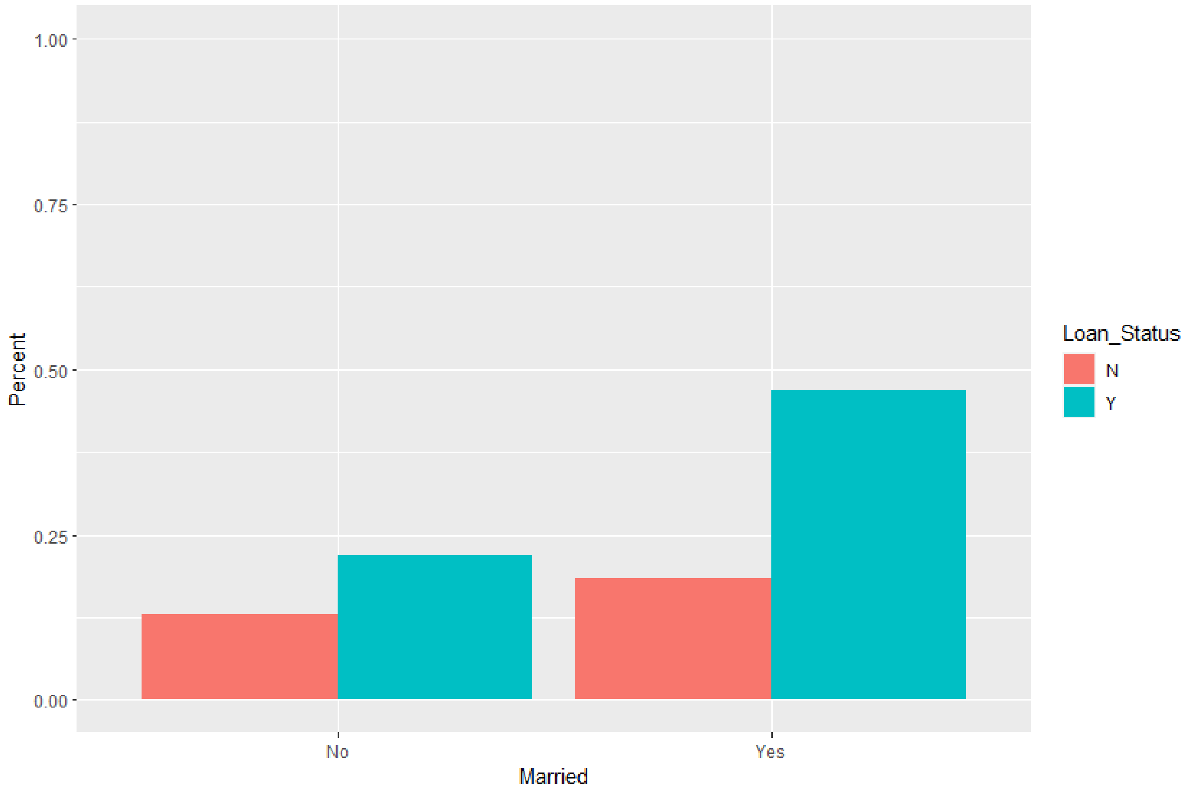
Gender by Loan Status.

According to the insight, there are more men in the population than women. They are about 3x the number of women.



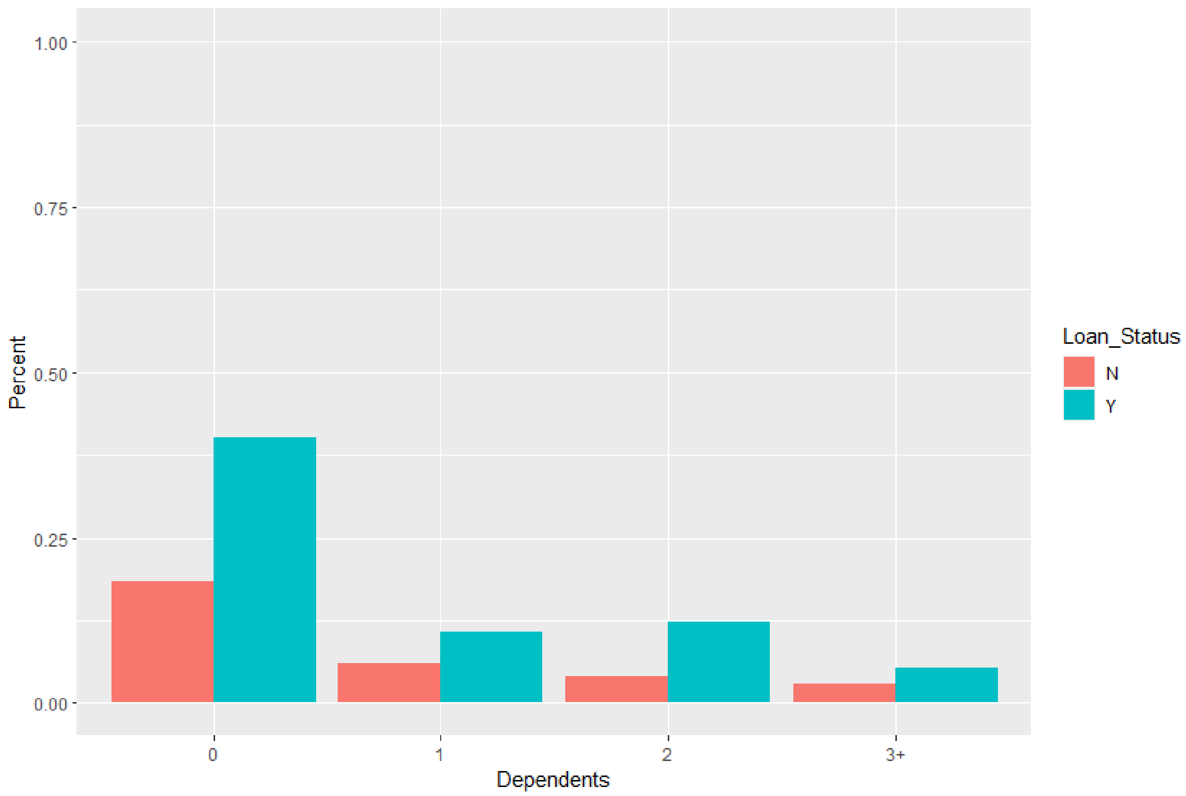
Marital Status by Loan Status.

According to this insight, married applicants are more likely to apply for loans than the not married ones. This may be because married people can't afford to pay for a house without taking loans.



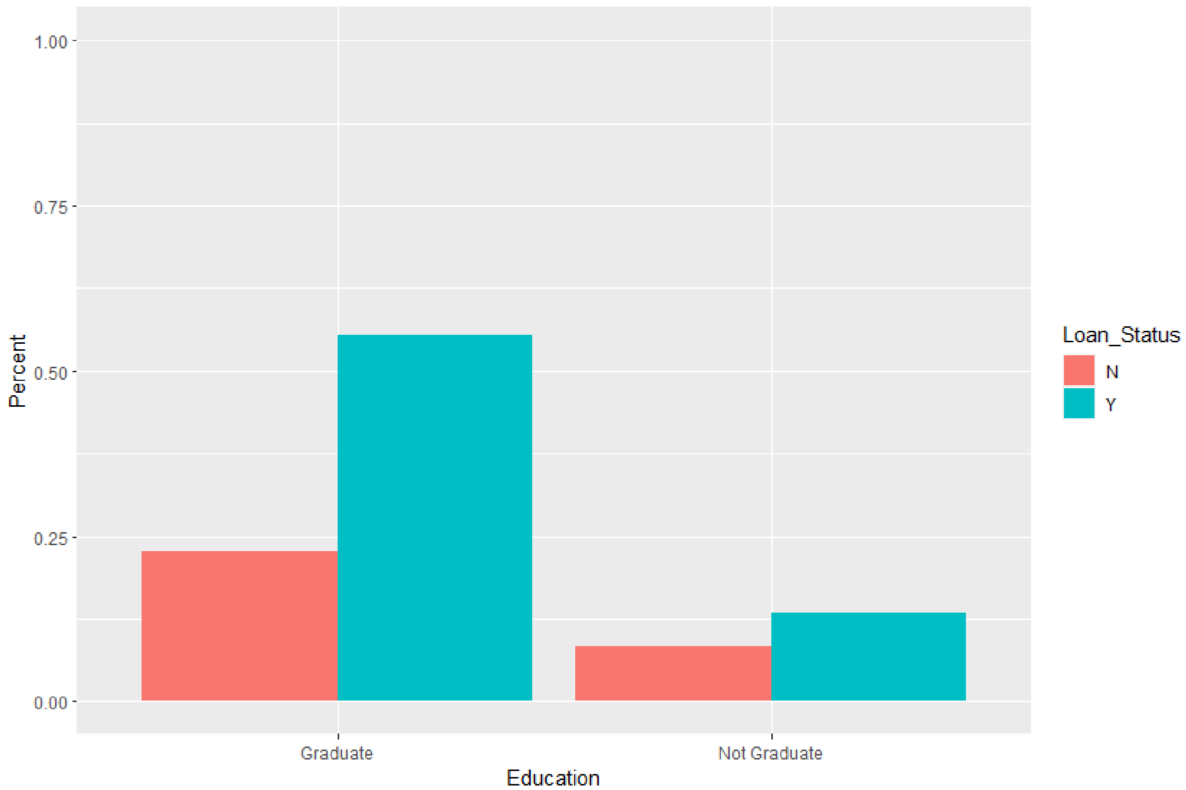
Dependents By Loan Status.

According to this insight, the majority of the population has zero dependents and are likely to get accepted for loans.



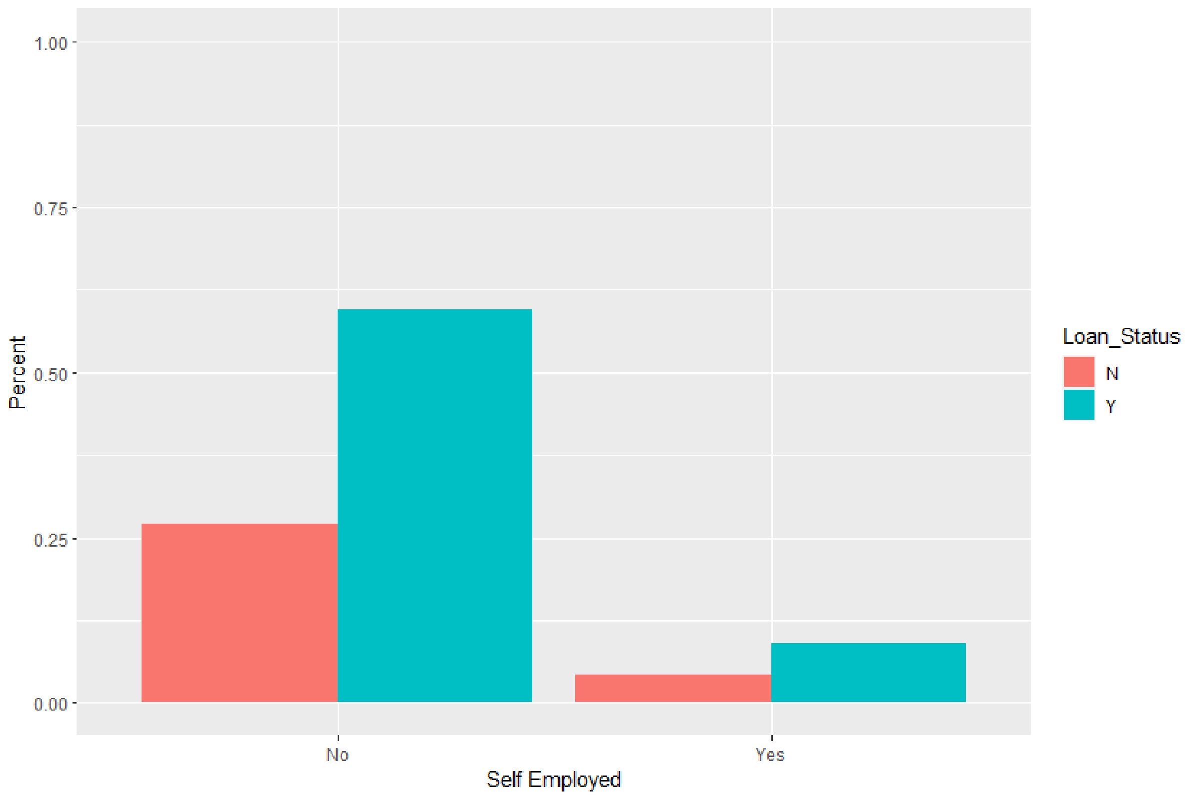
Education by Loan Status.

According to the insight, most of the applicants are graduates and are very likely to get accepted for loans.



Self Employed by Loan Status.

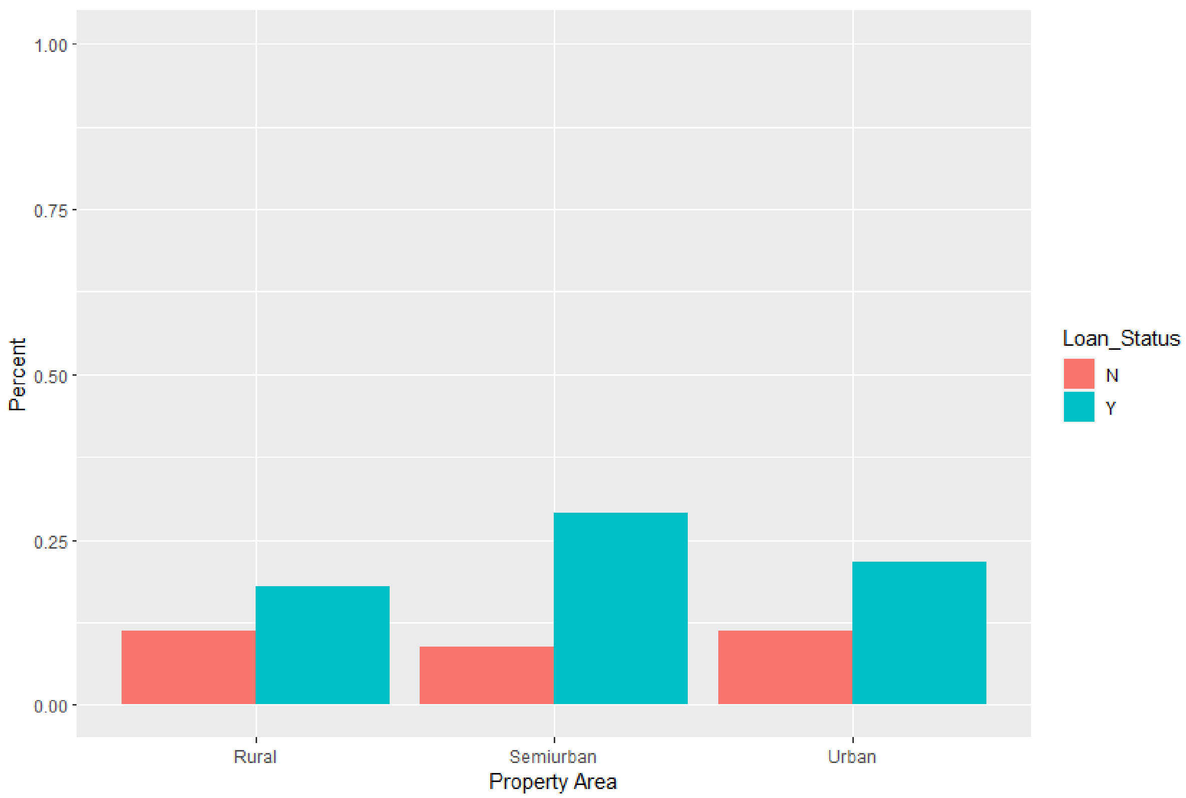
According to the insight, most of the applicants are not self employed, maybe because their income is more stable than the others so they are more flexible with taking loans.



Property Area by Loan Status.

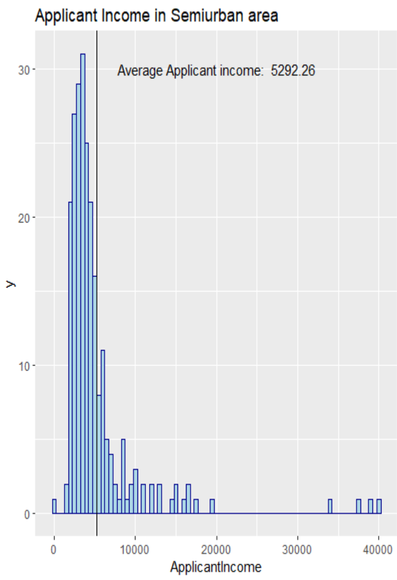
According to the insight, more applicants are taking loans for properties in semi urban areas. This could happen because of multiple reasons:

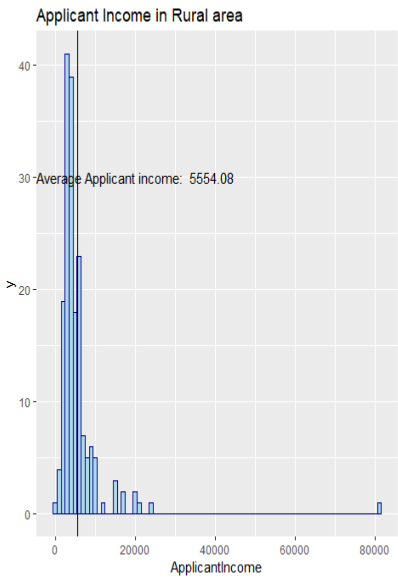
- The property price in a semi urban areas is much higher than urban and rural areas.
- The income of applicants in semi urban areas is lower than that of urban and rural areas.

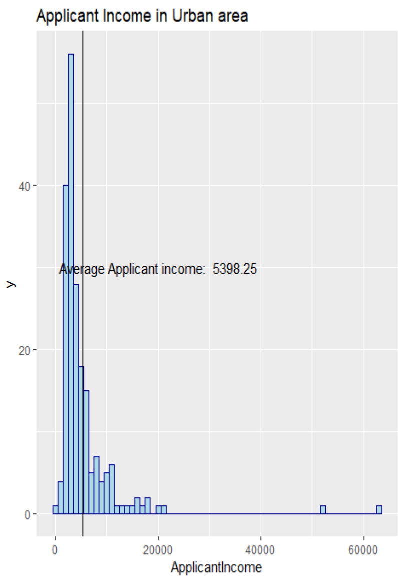


Applicant income in different areas.

According to the insight, the average income in the different areas doesn't vary that much. So the number of applicants for semi urban properties doesn't really relate to the income.

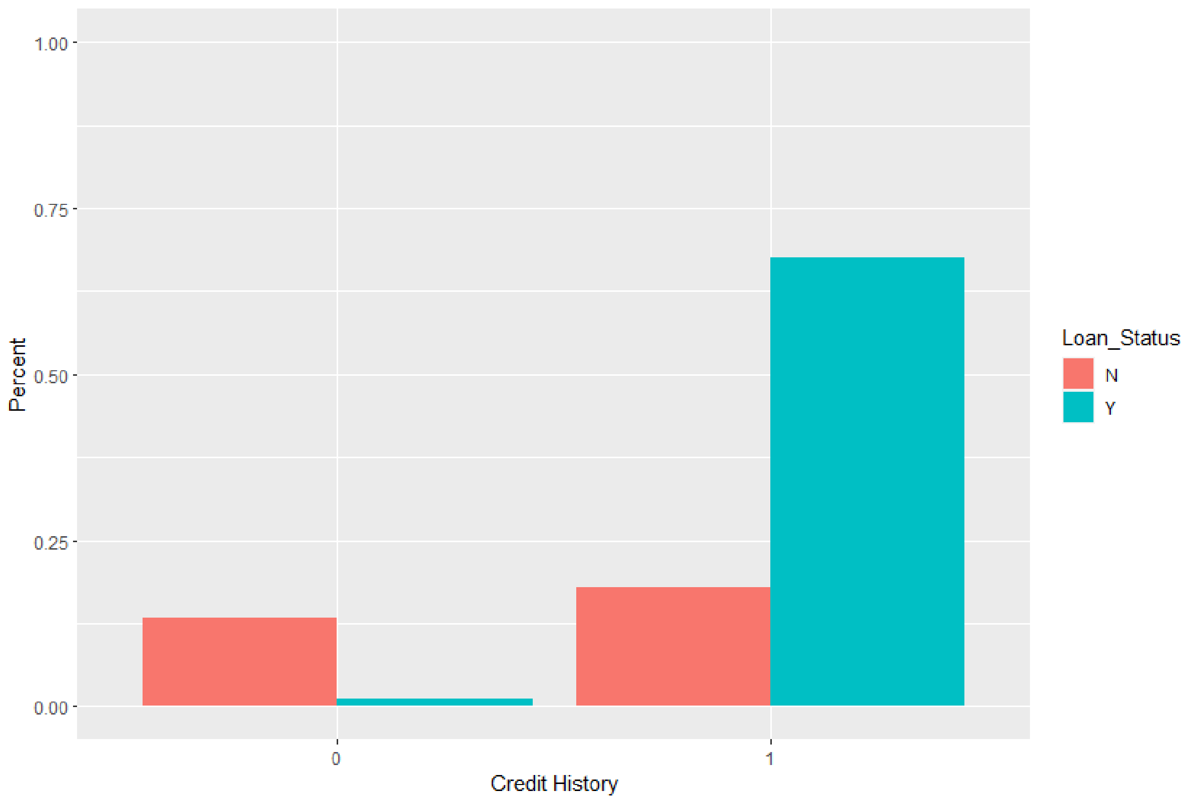






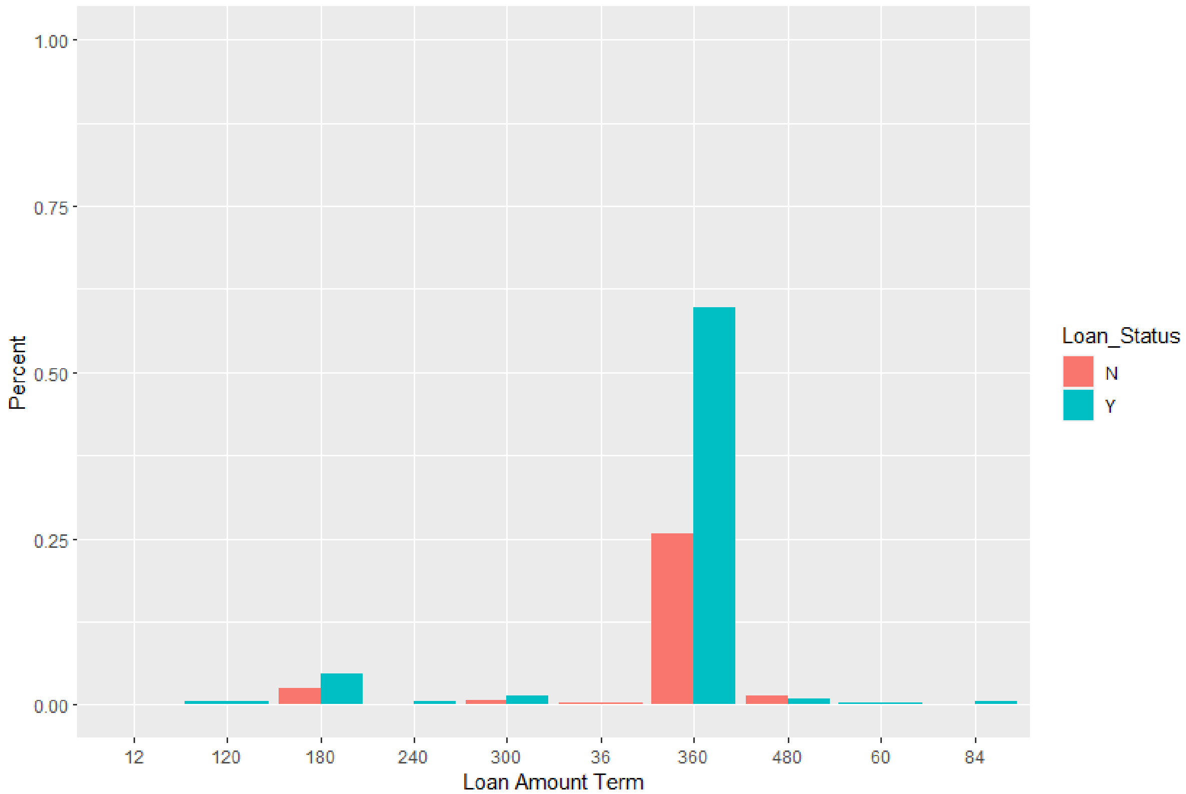
Credit History by Loan Status.

According to the insight, applicants with credit history are very likely to get approved on loans, very few percent of applicants that doesn't have credit history got accepted loans.



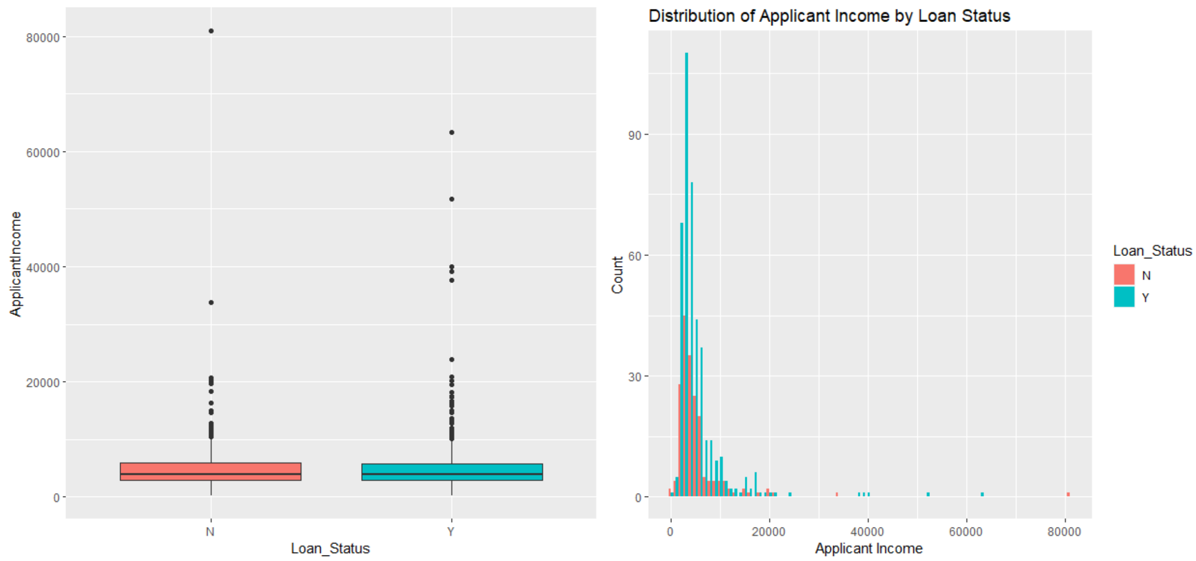
Loan Amount Term by Loan Status.

According to the insight, most applicants pay their loans in 360 months. This could be that this is the most convenient period for applicants to pay their loans.

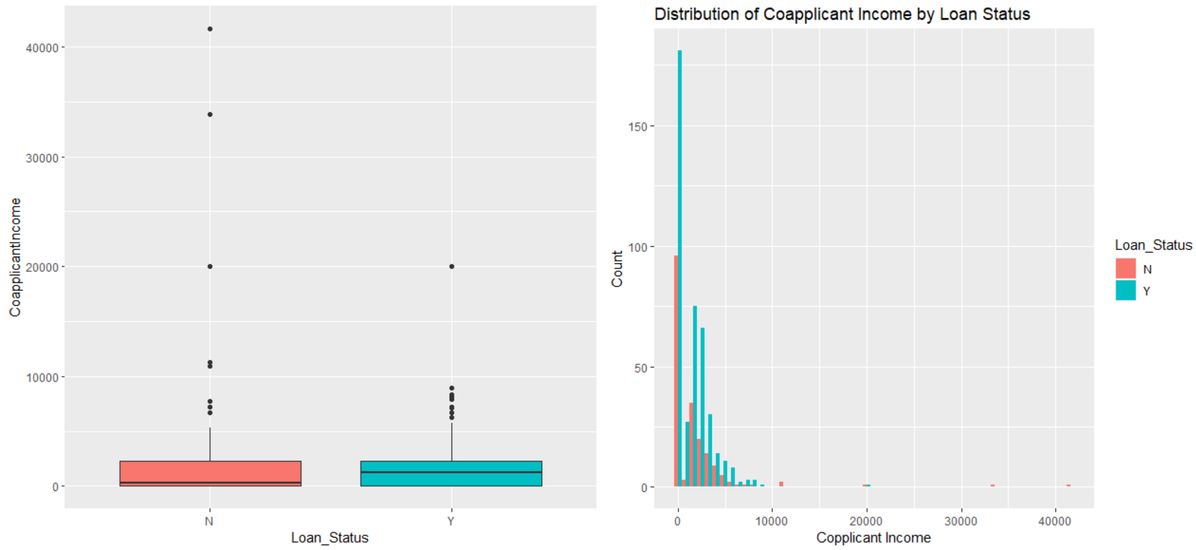


Applicant Income.

Applicant income column is right skewed and contains a lot of outliers.



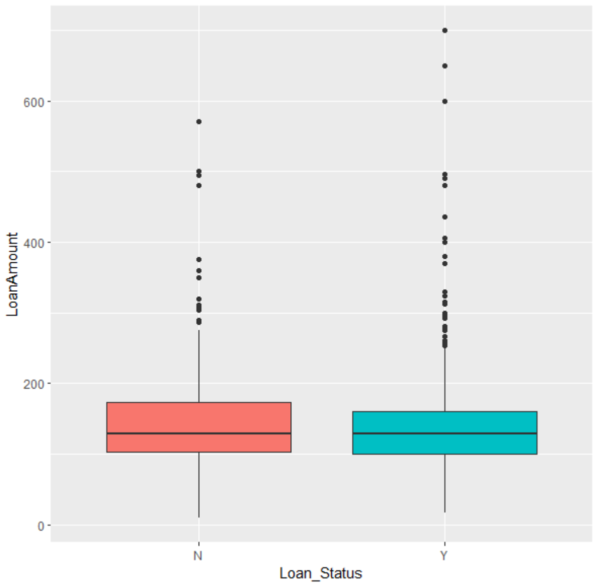
Co Applicant Income.

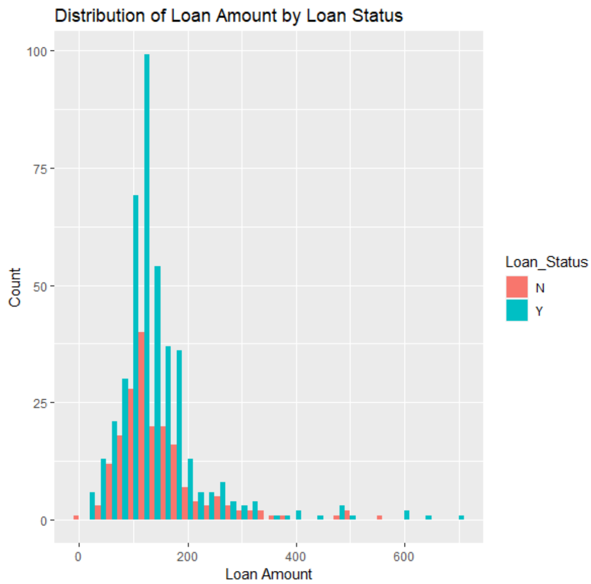


The Co Applicant income column is right skewed and contains a lot of outliers.

Loan Amount.

The data in the loan amount column is more normal than the applicant income and co applicant income columns, but it also contains outliers.





- Data Cleansing
  - Dropping Loan ID column
  - Removing nulls from columns (Loan Amount, Loan Amount Term, Credit History)
  - Removing wrong entries in columns (Gender, Married, Dependents, Self Employed)
  - Normalizing columns (Applicant Income, Co Applicant Income, Loan Amount) using log transformation

- Dataset Preparation in terms of ML
  - Transforming columns of type Character to be of type Numeric for this columns:
    - Gender, Married, Dependents, Education, Self-Employed, Credit History, Property Area and Loan Status
  - Scaling these columns: Applicant Income, CoApplicant Income, Loan Amount, Loan Amount Term
  - Splitting the Data to 80% training set and 20% testing set.

- Data Analytics Techniques
  - Logistic Regression
  - Support Vector Machine
  - Naive Bayes
  - Random Forest

- ## Performance Measures and Evaluation
    - Logistic Regression Accuracy: 88.52 %, with confusion matrix:
        - ```
          Confusion Matrix and Statistics

                    Reference
          Prediction  0  1
                   0 17  1
                   1 13 91

                         Accuracy : 0.8852
                           95% CI : (0.815, 0.9358)
              No Information Rate : 0.7541
              P-Value [Acc > NIR] : 0.0002344
          ```
    - SVM Accuracy: 88.52 %
    - Naive Bayes Accuracy: 88.52 %
    - Random Forest Accuracy: 88.52 %
    - ```
      > print(paste("Logistic Regression Accuracy:", round(logit_acc * 100, 2), "%"))
      [1] "Logistic Regression Accuracy: 88.52 %"
      > print(paste("SVM Accuracy:", round(svm_acc * 100, 2), "%"))
      [1] "SVM Accuracy: 88.52 %"
      > print(paste("Naive Accuracy:", round(naive_acc * 100, 2), "%"))
      [1] "Naive Accuracy: 88.52 %"
      > print(paste("Random Forest Accuracy:", round(rf_acc * 100, 2), "%"))
      [1] "Random Forest Accuracy: 86.89 %"
      >
      ```
- ## Discussion/Quantification for relevant project findings for your project.

Most relevant projects focus on the same type of analysis and conclude the same observations as in our project.

1. **Importance of different features:** Many projects revealed that the Credit History feature is one of the most important features in the dataset that could affect decision making and the Machine Learning models.
2. **Analyzing loan approval rates:** comparing loan approval rates for different borrowers in the dataset.
3. **Factors affecting loan amount:** exploring factors that could affect loan amount approved for borrowers, this could include variables like income, credit score, and the employment status.
4. **Loan approval rates:** gives a sense about how difficult it is to get a loan.