REPORT – MINI PROJECT 1: STRUCTURED DATA

Anvitha Koojugodu Shashidhar | aks100@syr.edu

Data And Source

The dataset for the project has been downloaded from Kaggle. It consists of home loan approval data from the Dream Housing Finance company website. They have a presence across all urban, semi-urban and rural areas. The customer first applies for a home loan after that company validates the customer's eligibility for a loan.
The customer fills out details such as: Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others.

Data Exploration and Data Cleaning

To explore the dataset, we used head method to return the first 5 rows of the dataset to get an idea about the data stored.
Info method from the Pandas library was used to print information. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).
We used dataframe.isnull().sum() function to return the number of missing values in the dataset. There were missing values in the dataset that were dropped using the dropna method of Pandas. describe function is used to calculate the summary statistics in Python.

Comparison Questions with Unit of Analysis

The following questions were selected to analyze the dataset in a view different from what is available in the dataset.

1. How does the credit history of loan applicants affect the loan approval rate?
   The unit of analysis is individual loan applications, and the comparison values are the number of approved and rejected loans for each credit history status. The approval rate can be computed by dividing the number of approved loans by the total number of loans for each credit history status.
2. For each property area, what is the average loan amount of the people of that type?
   The Unit of analysis is property area, and the comparison values are average loan amount. It is computed by grouping the data by property area and then calculating the

mean of loan amount for each property area group.

3. How many applicants based on gender are self-employed or not?
   Unit of analysis is Loan applicants, and the comparison values are Number of applicants based on their gender and self-employment status. The computation is done as follows: The dataset is grouped by gender and self-employment status, and then the count of loan applicants is calculated for each group using the count() function on the Loan_ID column.

4. How does the credit history of loan applicants affect the loan approval rate?
   Unit of analysis is the individual loan applicants, and the comparison value is loan approval rate based on credit history. The number of loan applicants with credit history 0 or 1 and loan approval status Y or N are counted, and the loan approval rate is computed as the ratio of the number of approved loans to the total number of loans for each credit history group.

5. How income differs based on sex?
   Unit of analysis is the individual loan applicants, and the comparison value is income based on sex. The mean income of male and female loan applicants is calculated separately and then compared to determine how income differs based on sex.

6. How does marriage and dependents affect property area?
   Unit of analysis is Loan applicants, and the comparison values are marriage, dependents, and property area.
   For each combination of marital status and property area we count the number of loan applicants. For each combination of number of dependents and property area we count the number of loan applicants.
   Finally the counts can be used to analyze how the distribution of loan applicants differs by marital status and dependents across different property areas.

Description of the Program

The program has been implemented in Jupyter Notebook mainly using the Pandas framework, CSV framework and OS modules.

The programs starts with importing the necessary modules and then the dataset from the CSV file downloaded from Kaggle. After the prompt data exploration and cleaning, we create multiple dataframe objects for storage of counts of loan status based on the combinations like Education, Gender, Employment status, and Credit history. We see that features Gender, Education and Self_Employed do not have significant impact on the Loan Status. However, Credit history has a significant impact on the approval status of a loan.

For the next visualization, we group the data by Gender and Loan_Status, and count the number of observations in each group. We can interpret the following: The number of males who got the loans approved are 278 whereas 116 were rejected. The number of females who got the loans approved are 54 whereas 32 were rejected. Males have a higher loan approval rate as compared to females.

Next, we review the relationship between Applicant Income and Loan Amount. The scatter plot shows whether there is any correlation between Applicant Income and Loan Amount. We see how the loan status varies across different values of the variables.

We move on to answer our analysis questions.
For each property area, we check the average loan amount based on the area of residence.
We then compare the count of loan applicants for each combination of gender and self-employment status. The output shows that there are 340 male applicants who are not self-employed, 12 female applicants who are self-employed, and 15 female applicants who are not self-employed.
Males have higher average income as compared to females by 481.998
Credit history affects the loan approval rates. In our case, there are 70 loan applicants with bad credit history(0) out of which 63 were not approved and 7 were approved. However, there are 410 applicants with good credit history and 325 were approved for the loan and 85 were not approved due to other reasons. We can clearly understand the relationship between credit history and loan approval status. The loan applicants with good credit score are more likely to be approved for a loan than those with bad credit score.
Finally, we review the distribution of loan applicants based on their marital status, number of dependents, and property area, which can be helpful in identifying trends or patterns in the data.

Description of the result of the analysis

In our first analysis we see that people from Rural have the highest average loan amount and the people from urban have the lowest.
In our second analysis, we cannot draw any conclusion about loan approval or rejection from the output.
In our third analysis, we understand that males have higher average income as compared to that of females.
In our fourth analysis, we conclude that credit history affects the loan approval rates. The loan applicants with good credit score are more likely to be approved for a loan than those with bad credit score
In our last analysis, we can say that there are 57 loan applicants who are not married, have 0 dependents, and live in a semi-urban property area. The third row shows that there are 43 loan applicants who are not married, have 0 dependents, and live in an urban property area, and so on.