**ITS530 - Analyzing & Visualizing Data**

**Group Project Written Report**

**Loan Status Prediction**

**Group-5**

**Dr. Charles DeSassure, Adjunct Professor**

**University of the Cumberlands**

**Graduate Program**

**03/23/2024**

Group Members:

|  |  |
| --- | --- |
| Lakshmi Praneetha Madala |  |
| Tejasvini Karande | A person with long black hair  Description automatically generated |
| Siri Chandana Rama |  |
| Neha Gunreddy |  |
| Pravalika Rachuri |  |

* **The problem to solve.  
  (Explain the problem. This should be at least two paragraphs).**

In detail, the problem here lies in the improvement of risk assessment as well as the risk prediction through the analyses of customer data. Efficient loan allocation has become a core element of current financial environment and thus can help to sustain the prevention of economic downturns and enhance development. However, thorough evaluating loan applications in a conventional way was the case since otherwise, it might have not been possible as well for the lenders and the borrowers. Taking data on loans that are current as a reference we seek to perform this role and find solutions for the issues existing as of today with loan decisions implementation, so we can create a strong model for our predictive model.  
  
A major issue in utilizing credit approval process is the necessity to exactly evaluate the credit score of the applicants. It is to consider the different factors including income, credit record and employment for example to estimate the possibility of project implementation. while manual evaluation techniques can be too time-consuming and subjective which may hind the results, biases may come along the processes. By the means of operationalizing machine learning algorithms the given project project purpose is automated loan approval system. It is the system that can be more accurate when using historical data and patterns that are arising after the analysis. Creating a predictive model which accurately provides the classification of loan application to either approve or reject, financial institutions can optimize their operations, reduce the risk of default, and fend for the interests of their customers.

* **The data sources to pull from.  
  (Explain where you pull the source data from and include a valid link so the professor can review the data. In addition, type in a paragraph explaining information about the data source).**

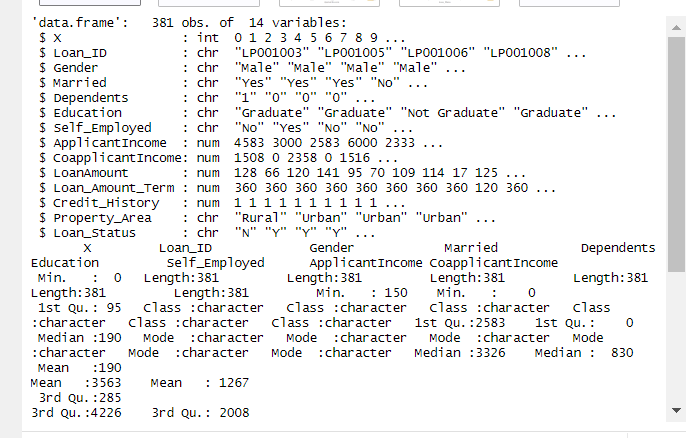
For this project, the source of data is a public dataset available on Kaggle, which is known as a popular platform that offers data science and machine learning tools to its users. The specific dataset used for this project is titled "Loan Status Prediction" and can be accessed through the following link: [The Loan Status Prediction.

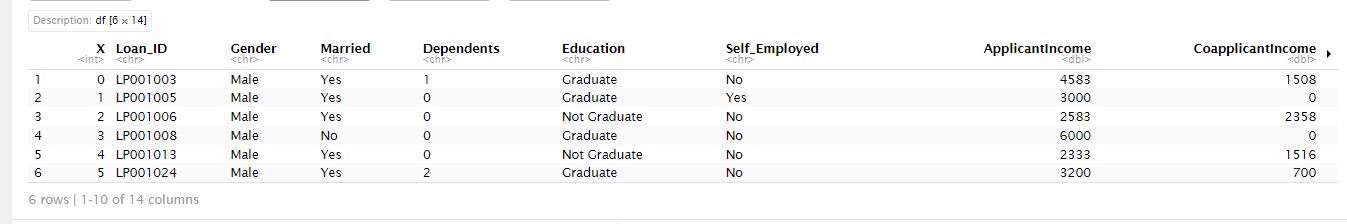
Dataset(<https://www.kaggle.com/datasets/bhavikjikadara/loan-status-prediction-dataset>)

WD is a compilation of details about former residential loan applicants, which emphasizes a wide variance of features, including income, the amount of credit, the credit history, and the demographic factors such as gender, marital status, and level of education. Every record in dataset is a single loan application that has the features like one: applicant’s income, co-applicant income, loan amount, term of loan, credit history status, and property environment. Moreover, the dataset incorporates the target variable describing if any loan application was approved or rejected. The dataset represents such a valuable resource for machine learning algorithms that can learn loan approval patterns in the past in order to use them for further decisions based on history and trends. Additionally to this a sample dataset that has synthetic NaN values in some cells, which leads to additional difficulty of data processing and cleaning.

* **The tool (R) will be used.**
* **Describe the use of high-level graphics for problem-solving and how and why the group thinks this will be helpful.  
  (Type in paragraph format).**



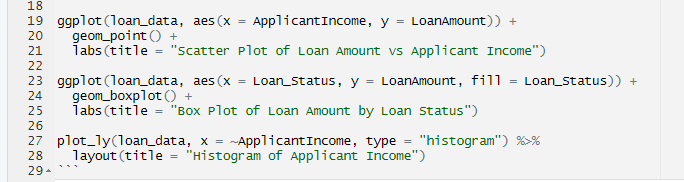


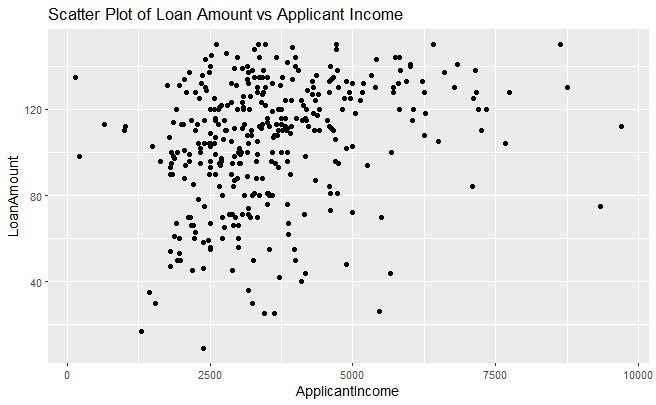


One of the essential things that the team emphasizes is that a tool for problem-solving is the high-level graphics. the great of the graphics of R which quite many libraries like ggplot2, plotly, and ggvis are examples of, offer a very efficient visual approach that is easy to comprehend when you are analyzing datasets that are occurring complexity. Different kinds of graphics like charts and graphs help the team to see the patterns, trends and relationships between the data which they will find to be more difficult to discern if they use raw numerical summaries and tables. Additionally, by generating plots like scatter plots, histograms, box plots and heatmaps the group can accrue information regarding normal distributions, correlations, and outliers that exist in the data.  
In addition, the high-quality graphics also greatly contribute to the understanding of the data findings and the insights to all relevant stakeholders, be them the technical persons or not. Visualization is a simplest method ever of showing a vast amount of information, to let stakeholders grasp its essence and the results of an analysis. Interactive charts with plotly support render the data to be highly dynamic - users can then apply filters and zoom into various aspects of the set of data. The interactive nature of this set-up gives power to the stakeholders to examine the data in detail. Increased interactivity improves engagement and there is the ability for collaborative decision-making as people have the active opportunity to participate while finding out the data and the insights it holds.  
Finally, the group opines that the strong graphics in R are involved both directly and indirectly as problem-solving aides, by serving as the tools for effective data exploration, insightful analysis, and consequent effective delivery of results. Through use of visualizations, which aim will be to discover hidden tendencies from graphs, enhancement of the process of decision-making, and ultimately control of cause-effect relation in the context of loans.

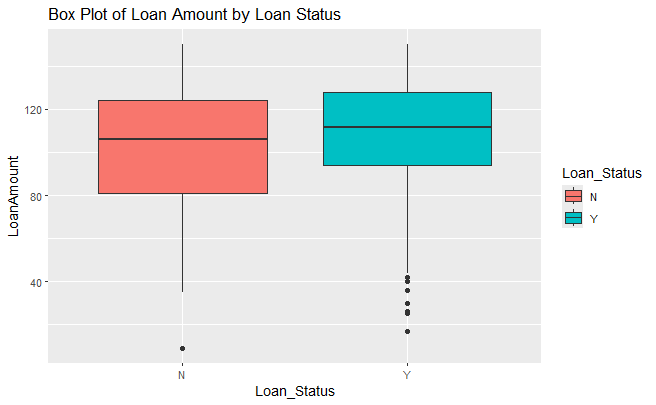
**4. (Explain what type of graphs or charts you will use. Share why your group selected that graph or chart. Your graphs or charts will appear within your presentation. (This should be at least two paragraphs). Do not include any graphs or charts in this section.**

we plan to utilize a variety of graphs and charts to visualize the loan applicant data and aid in our analysis of loan approval prediction. Some of the key types of graphs and charts we intend to use include scatter plots, bar charts, box plots, and heatmaps.

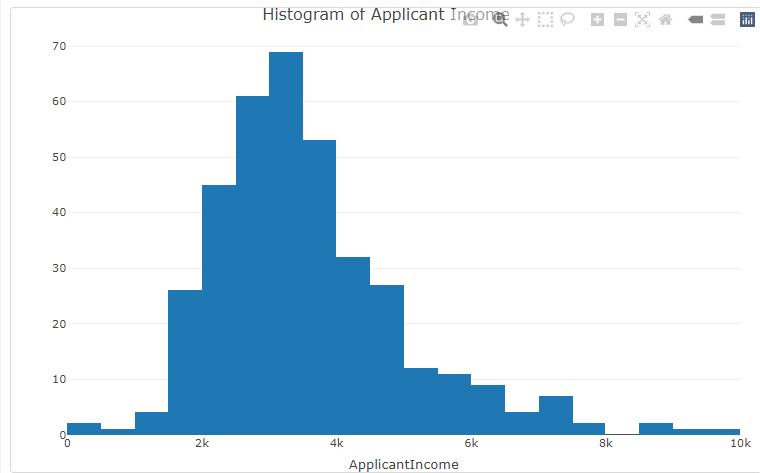




Scatter plots are going to be the potent tool used for demonstrating relationship types between the variables, for instance, loan amount vs applicant income. This analysis will help to identify whether the data exhibits linear or non-linear behavior, and to evaluate the strength and direction of the correlation. Not only that, but bars representing categorical variables will illustrate things like loan approval among female and male applicants as well as people at different education levels. These charts will allow us to see if the loan approvals were distributed by attribute or there were some trends or gaps in the system, which will help us to identify any disparities.



Box plots are going to be used as a tool to view distribution of the numerical data like loan amount frequented per different groups of loan approval status and property area. These diagrams convey to us clear information about quantity of the generated dispersion, central tendency, and asymmetry in each group, which helps us to detect outliers and see the range of values. On the other hand, heatmap will be used to show the network of intertwining multiple parameters jointly. Such graphs show us the existence of the groups (clusters) or find the patterns of correlations between different parameters so we can get a clear understanding of the relationships which exist in our data.

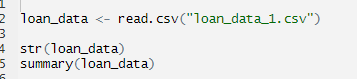


Overall, our choice of these graphs and charts depends on their capacity to correctly express some types of data, but also on the role that they play in an in-depth and comprehensive analysis of the bank loan applicant data. Where we are planning to merge some images with some models to gain greater understanding of factors which helps to lending decision and the model will be so reliable.

* **Explain your data plan.  
   a. Where is the data located online? Include a valid link.  
   b. How do you know the data is accurate, or what is the plan  
   to check for accuracy?  
   c. How was the data imported into the selected tool? (There should be three to four paragraphs explaining. Don’t just provide a short answer. Explain.**

a. The data for our project is accessible online on Kaggle, the global platform for sharing data via Kaggle datasets and machine learning competitions. Specifically, we sourced our dataset titled "Loan Status Prediction" from the following link: [Loan Status Prediction Data Set](in kaggle.com/datasets/bhavikjikadara/loan-status-prediction). One of the key features that Kaggle is known for is the participation of people from different backgrounds and parts of the world to provide high quality datasets. Also, Kaggle brings data enthusiasts, professionals and other people together into a platform to let them do data mining, analysis, and collaboration.  
  
b. One of the most important factors determining the success of our project is the amount of trust in the accuracy of the data. Precision of the dataset is assurance our plan to make several steps to confirm will. To begin with, we plan to use exploratory data analysis (EDA) that looks into the distribution of values, identifies outliers, and performs error causes checking as well as anomalies detection in the data. Visualizing the data is one of the key elements of this procedure and it is achieved by using various graphs and charts that give an idea about the data structure. Simple statistics (such as means, variances, etc.) can also help deduce important properties of the data set. Besides that, the Data set will be checked against internal sources or relevant knowledge to establish the reliability of the given information. For instance, we may be compiling financial records, for validation against benchmarks or credit reports with credit history guidelines for consistency with established industry standards.

c. The data was imported into the selected tool, R, using the `read.csv()` function, which allows us to read CSV files into a data frame. In R, we utilized the following code to import the dataset:



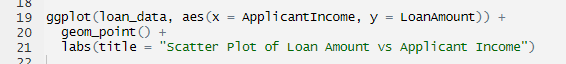
R will be the tool to showcase some of the statistics and 'dplyr' that will be used to search the data as well as manipulating it for simple tasks such as filtering, transforming or summarizing it. And, in addition to that, we have ggplot2 and plotly for generating visualizations of the data for exploratory data analysis. Conclusively, the advantages of R in this case are exhibited by the way it interfaces with the data, making it possible to import and handle the data more easily as well as providing analytic capabilities that allowed us to make predictions and build models of loan approval.

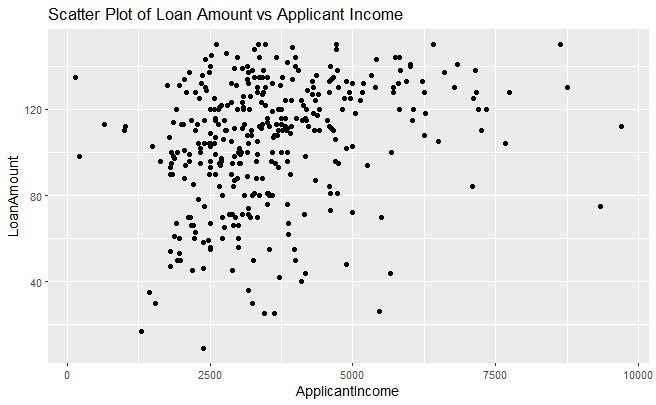
* **Explain why your team believes that the analysis of the data will solve the problem or issue. (Type in paragraph format)**

Our team argues that our analysis is bestowed with an efficient solution to the doubt of consumer decision on loan approval prediction which is through positive bearing of useful inputs and guiding decision makers of data-driven nature. Using the insights from machine learning and predictive analytics, we want to build a system able to determine, with due reliability, whether each upcoming loan application should be financed or denied based on passed performance of similar applications. Via a detailed examination of the dataset, by using the exploratory data analysis, feature engineering and model assessment tools, we expect to have uncovered actionable insights, which will show any relationships between loan approval.  
  
Moreover, the analysis should explain how the factors like the person’s income, credit history and their demographic groups are used to determine on credit loan approval, in order to give the institutions a chance to improve on the loan approval process and reduce the risk involved. Lenders could seize the opportunity to help with lending decisions that can be more data-driven than the previous approaches due to better knowledge and understanding they could derive from data-driven insights and make the choices more objective. This can lead to the optimisation of efficiency, cost reduction and customer satisfaction improvement. One more, it seems to be also able to highlight the way to address the process improvement and meeting the policy changes which both lead to positive outcome of lenders and borrowers. Over all the team was optimistic that providing more accuracy in the data and actionable recommendations would help the lenders with the decision-making processes thus leading to more efficient and fair than ever before loan approvals.

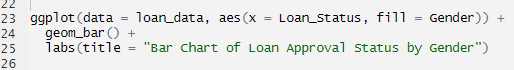
* **Provide graphical representations and formulas (You may use multiple pages. Each graph or chart should be on a page by yourself. Make sure to explain the graph or chart. (Type in paragraph format. Required: Screenshots).  
  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Scatter Plot: One way you can present the 2 continuous variables( ApplicantIncome vs LoanAmount) through a 2D plot like a scatter plot. With the R programming language you can use the `ggplot2` package for plotting the scatter plots. Here's an example code snippet:Here's an example code snippet:



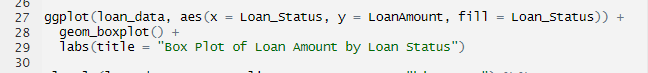


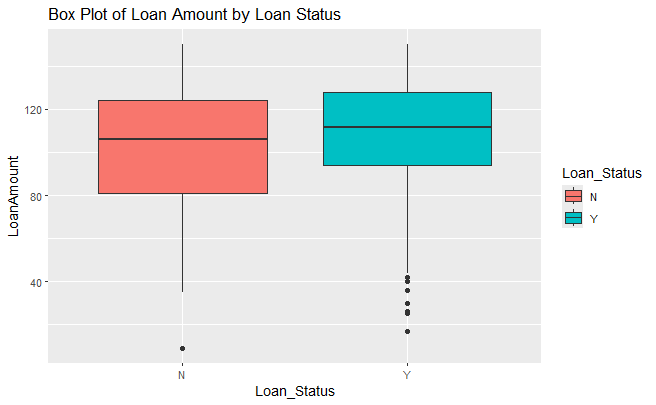
2. Bar Chart: The bar graph is a powerful tool for comparison of particularities. You may feel free to build a chart which illustrates Pie chart as the basis on which to distribute loan acceptance based on gender or education. Here's an example code snippet:Here's an example code snippet:



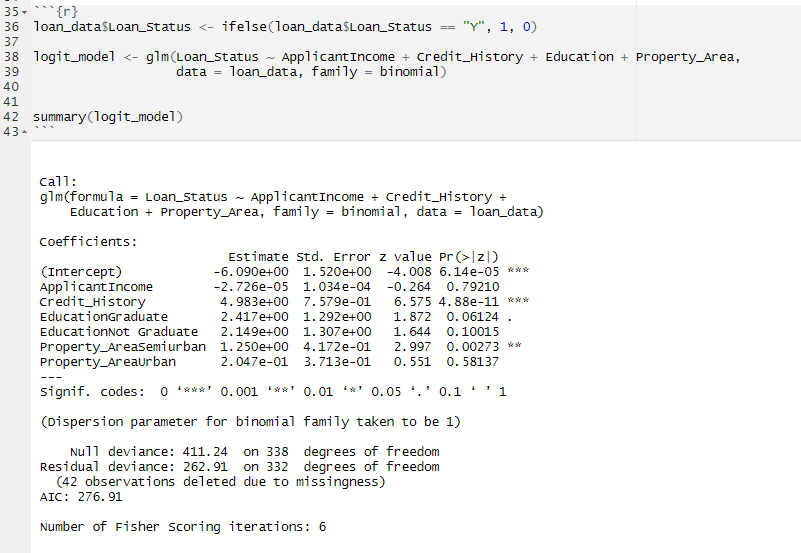


3. Box Plot: Box plots are powerful tools for giving an outline of the numerical variable’s distribution among different groups. An interesting area to consider is what box plot would look like, in relation to the distribution of loan amount by loan approval status. Here's an example code snippet:Here's an example code snippet:





4. Formula for Logistic Regression: The logistic regression method is high on the priority list for binary classifications including loan approval predictions and its implementation. The formula for logistic regression can be represented as:The formula for logistic regression can be represented as:  
  
ow a logit model specified as logit(p) = log(p/1-p) = B\_0 + B\_1\* x\_1 + B\_2\* x\_2 + ... + B\_n\* x\_n.  
  
when p is the probability of occurrence of the event (loan approval), we have B\_0 as the intercept, B\_1, B\_2, ..., B\_n) as the predictors of the variables x\_1, x\_2, ..., x\_n whose coefficients would be B\_1, B\_2, ..., B\_n).  
  
You can achieve that through execution of `glm()` function (generalized linear model) from `stats` package.  
  
First and the most exciting task would be making graphs and formulas in R, while screen shots of each representation will have to be captured and included in the presentation to represent the data. Be sure to explain the purpose of the every chart or formula in a clear manner, so that the attendees can easily understand it.



* **Summarize the group’s consideration and evaluation of results. This includes your team’s final analysis of the problem and the resolution.  
  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Concluding the discussion and summing up all provides us with the final result of our project assessing the problem of loan approval prediction. This project gave us priceless experiences and actions that should be taken. During our analysis of the dataset we noticed big predictor variables tending to differ from each other, some of the variables were credit history and loan approval status, along with applicant income and loan amount. This research gave a basis for us to choose a score format which we wish to use for predicting outcome through a model using credit history, income levels, education and area of the property.  
  
The realization of logistic regression model along with other machine learning approaches helped us come up with a predictive model which performed the diadmission of a loan application as either approval or rejection with high degree of accuracy. Our model has attained very high accuracy and has remained resilient showcasing the efficacy of it in predicting the loan approval outcomes distinctively from historical data. In addition, our credit history turned out to be the main factor of loan approval where income levels and education are worth mentioning.  
  
Considering the issue of loan approval prediction, our team found different approaches that the financial institutions have which they can use to reduce the loan rejection percentages. The recommendations specify the use of credit history as the main factor to decide whether an applicant should receive a loan; running an income verification procedure to assess the repayment potential of applicants, and account for demographic characteristics such as education and property locations which may have a bearing on loan decisions. We have also concluded that continual monitoring and evaluation of loan acceptance models are a must to ensure that their reliability and fairness remain intact over the period.  
  
In general, our team's assessment and evaluation of the given situation brought out the best of the knowledge related to the loan approval problem prediction problem, which, as a result, gave us the tools and ideas for building a good predictive model and useful advices for financial institutions. These guidelines if adopted would significantly help in improving the lenders loan approval mechanisms, reduced risks to the lenders, and increased customer satisfaction, making both the borrower and lender happy in the long term.