# **Unlocking Insights: A Comprehensive Exploratory Data Analysis of Synthetic Bank Loan Data**

Financial institutions have a significant impact on the economy, and the information they gather—especially about loan applications—can provide important new insights. We take a trip through a fictitious bank loan dataset in this blog post, removing layers to uncover complex patterns, trends, and the numerous variables that affect loan approval.

## **Navigating the EDA Landscape** We start our journey by asking the data several insightful questions. What is the effect of age on loan acceptance rates? What part does income play in making decisions? Do educational attainment levels affect the chances of a loan being approved? These are only a few of the many queries we will be addressing to fully understand the dataset's complex structure.

### **Breadth of Exploration:**

We won't restrict ourselves to a small number of questions to guarantee a thorough investigation. Rather, we will widen our focus to address more than three queries covering different facets of the data. This broad strategy will highlight several aspects and provide a comprehensive understanding of the dataset.

### **Depth of Exploration and Data Quality:**

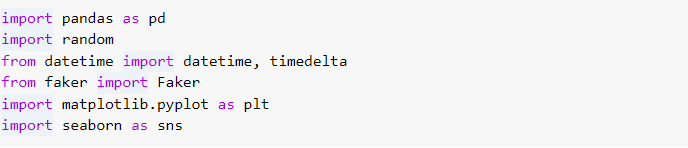
We'll delve further into the data to answer more detailed queries. It will be essential to conduct follow-up investigations to uncover insights that surpass cursory observations. Through in-depth field and record profiling, we will carefully examine the quality of our data to make sure that our analysis is not only comprehensive but also strong and trustworthy.

We will guide you through the process of examining the artificial bank loan dataset in the following sections. Every element—from financial measurements to demographic traits—will be examined in detail to reveal latent relationships and shed light on the variables affecting loan approval.

Prepare yourself for this data-driven journey as we explore the complex world of synthetic bank loan data to extract insightful knowledge and reveal the mysteries of financial transactions. We hope that our EDA journey will provide us with a deeper knowledge of the complex tapestry that is bank loan data by illuminating the stories hidden behind the numbers.

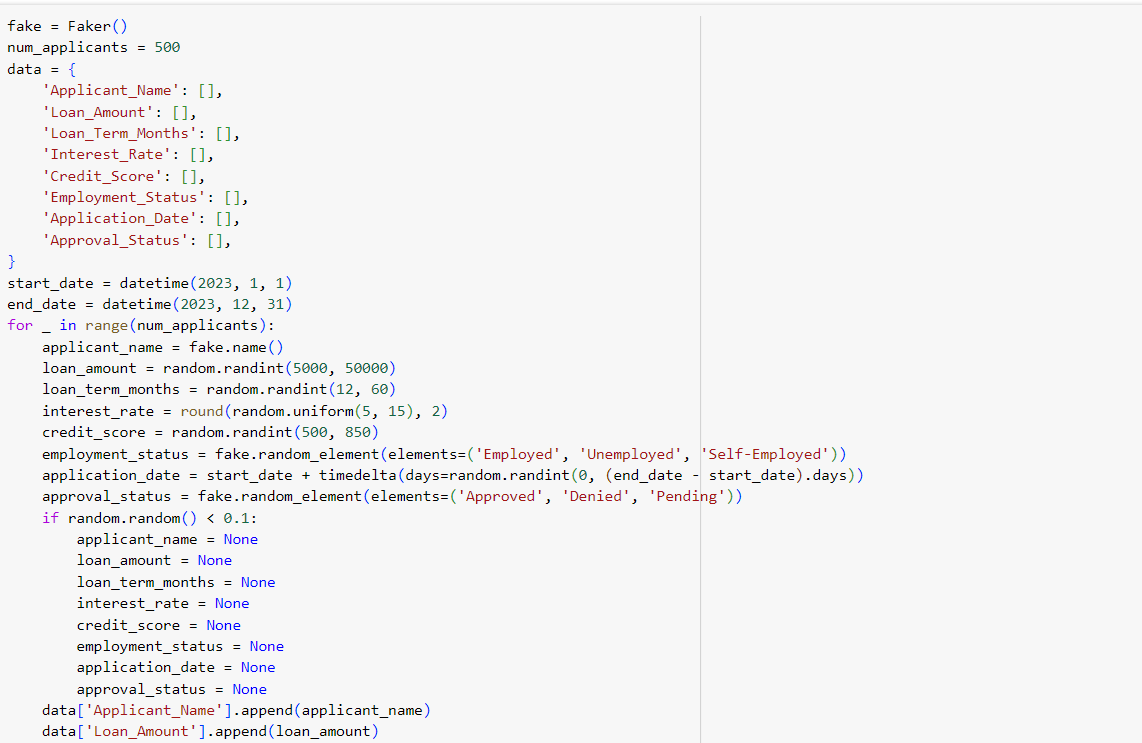
**Exploratory Data Analysis**   
 EDA – plays a critical role in understanding what, why, and how of the problem statement. It’s first in the order of operations that a data analyst will perform when handed a new data source and problem statement.  
Here’s a direct definition: exploratory data analysis is an approach to analysing data sets by summarizing their main characteristics with visualizations. The EDA process is a crucial step prior to building a model to unravel various insights that later become important in developing a robust algorithmic model.  
  
Let's attempt to dissect this definition and comprehend the various operations that EDA is used for:  
 Above all, EDA offers a framework for dissecting problem statements into more manageable trials that might aid in comprehending the dataset.   
EDA offers pertinent information that support analysts in making important business choices.  
The EDA step gives us a place to conduct all of our thought experiments and, in the end, directs us towards a crucial choice.  
**Dataset Generation:**

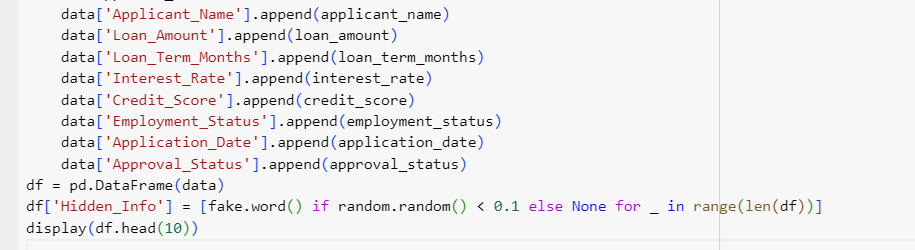
**Importing the necessary libraries**

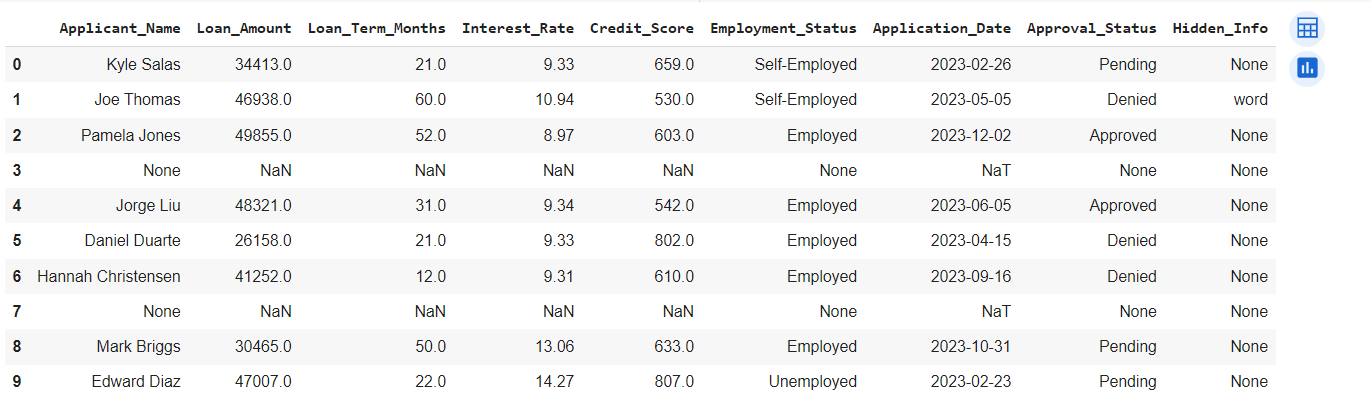


**Creating a dataset**

This is the snippet of the code used to generate the synthetic dataset





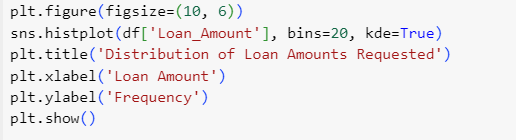


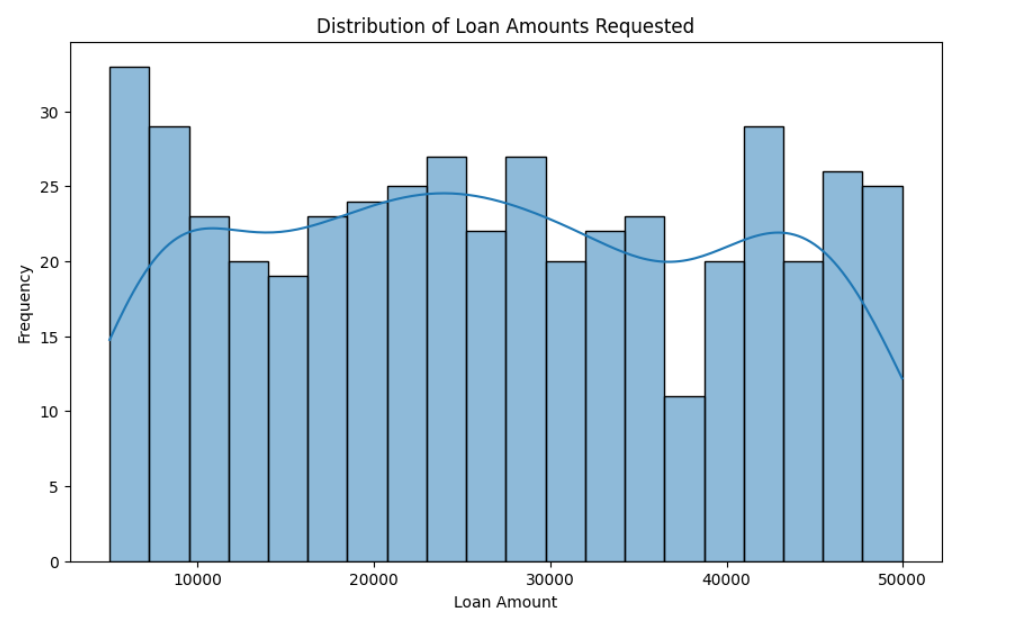
This the dataset which we will use to perform EDA, visualisation, transformation and quality checks.

**Performing EDA on the above generated dataset**

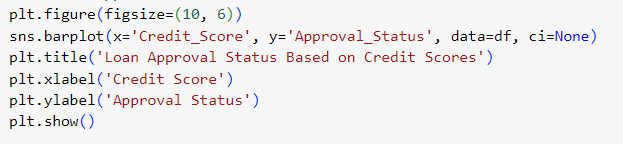
**1: What insights can we gain from the distribution of loan amounts in our synthetic bank loan dataset?**

We kick off our analysis by exploring the distribution of loan amounts requested by applicants

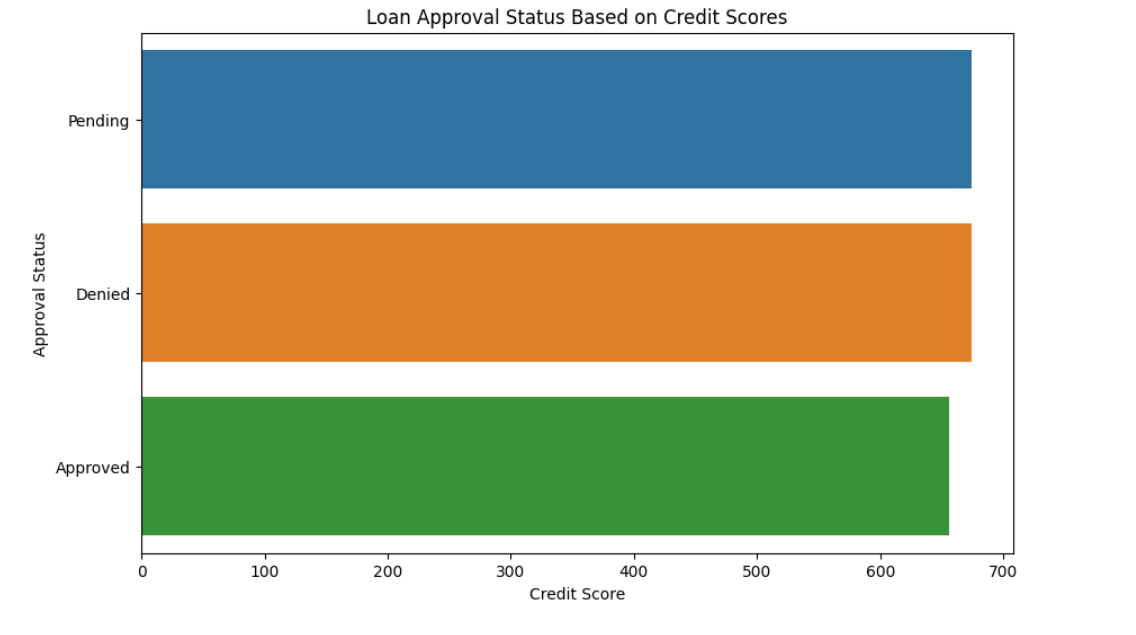




We can get a general idea of the applicants' financial needs from the histogram plot, which displays the range and frequency of loan amounts. The histogram indicates that requests for loans between $10,000 and $20,000 are the most common loan amounts. Additionally, a sizable portion of the population applied for loans between $20,000 and $30,000 and between $30,000 and $40,000. The information is dispersed equally along the range of loan amounts. The lack of outliers in the histogram, or data points that deviate significantly from the main distribution, indicates that the data is probably rather typical.  
  
**2**. **How does the credit score of applicants correlate with their loan approval status?**   
Next, we investigate how the loan approval status varies based on credit scores.



A bar chart visualizes the relationship, offering insights into the creditworthiness of applicants and its impact on loan approval.



The graph depicts the relationship between credit score on the x-axis and loan approval status on the y-axis. There are three possible statuses: approved, denied, and pending.

* Approved: Shown as a green line increasing progressively from 0% at 0 credit score to approximately 80% at 700 credit score.
* Denied: Shown as a red line decreasing progressively from 100% at 0 credit score to roughly 20% at 700 credit score.
* Pending: Shown as a blue line starting at 0% at 0 credit score, reaching a peak of about 20% around 200 credit score, and then gradually decreasing to 0% at 700 credit score.

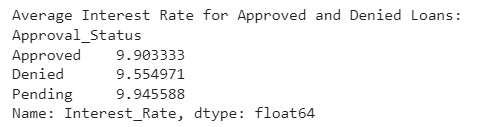
Key Observations:

* Having a higher credit score is generally associated with a higher chance of loan approval and a lower chance of denial.
* The "pending" category plays a more significant role for borrowers with lower credit scores, gradually diminishing as credit scores improve.
* The graph suggests that a credit score of around 500 represents a turning point where approval becomes more likely than denial.

**3.** **How does the average interest rate differ between approved and denied loan applications?**

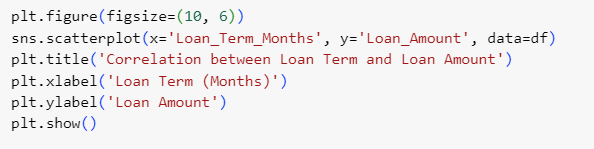


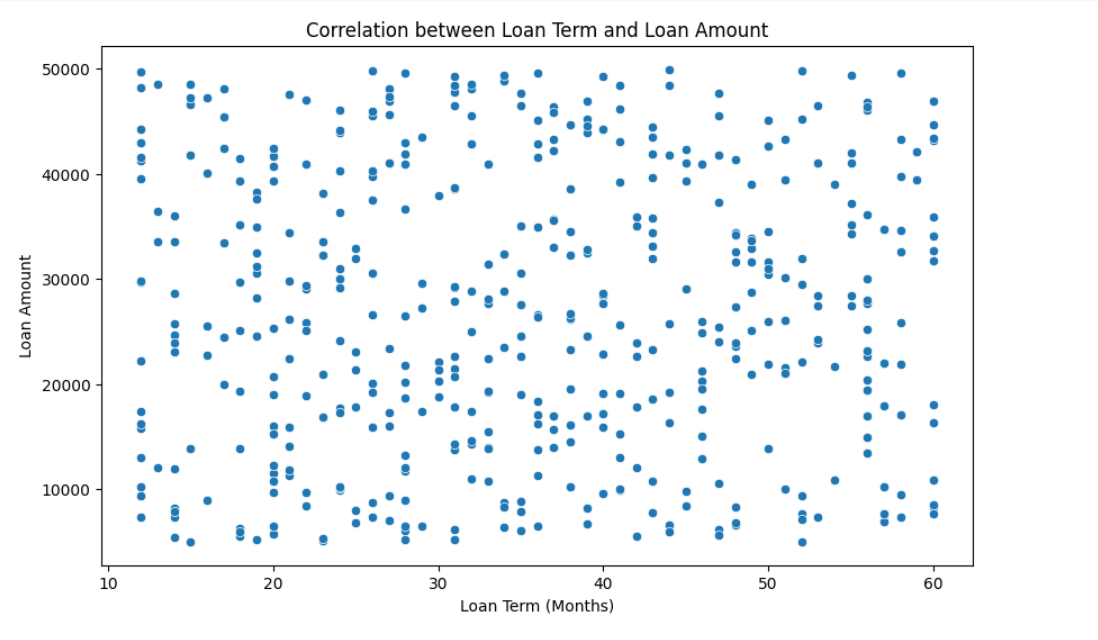
We calculate and compare the average interest rates for both approved and denied loans, shedding light on the financial terms applicants may face.



**4. What is the degree of correlation between the loan term and loan amount in our dataset?**

We explore the correlation between the loan term (in months) and the requested loan amount using a scatter plot





.

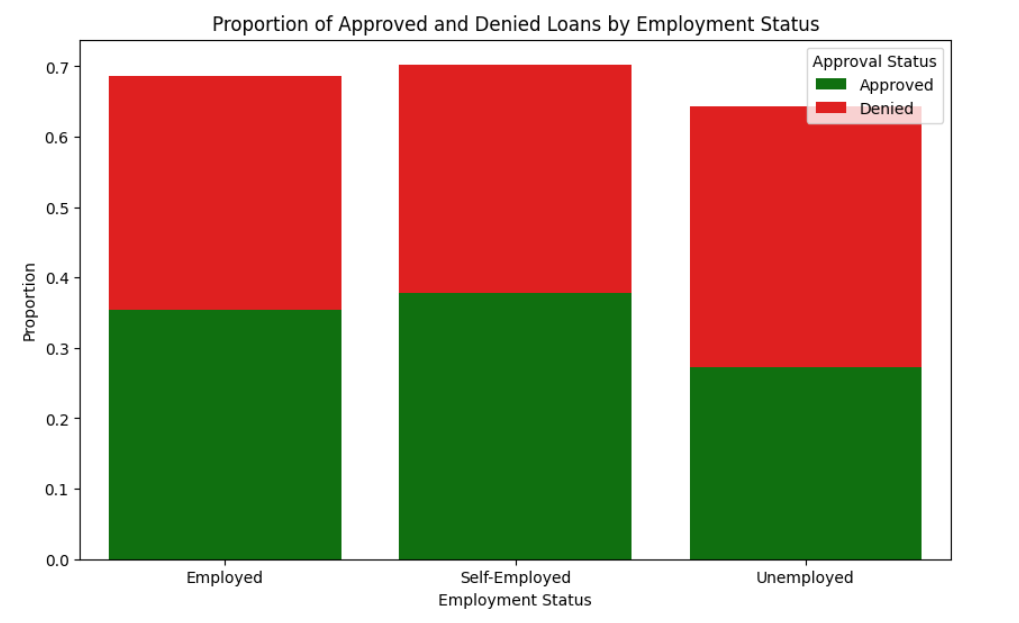
This analysis helps us identify any patterns or trends in the preferred loan terms for different loan amounts.  
Some key observations we can make from the scatter plot:

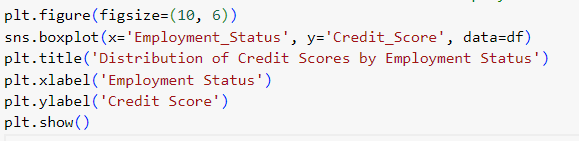
* Positive correlation: There is a positive correlation between loan term and loan amount. This means that, in general, longer loan terms are associated with higher loan amounts. This makes sense, as borrowers who need more money are likely to need more time to repay it.
* Clustering: The data points are somewhat clustered in the lower right portion of the plot. This indicates that there are a lot of loan applications for shorter terms and smaller amounts.
* Outliers: There are a few data points that fall far away from the main cluster. These outliers could represent loan applications with unusual terms or amounts.

**5. How does the proportion of approved and denied loans vary across different employment statuses?**

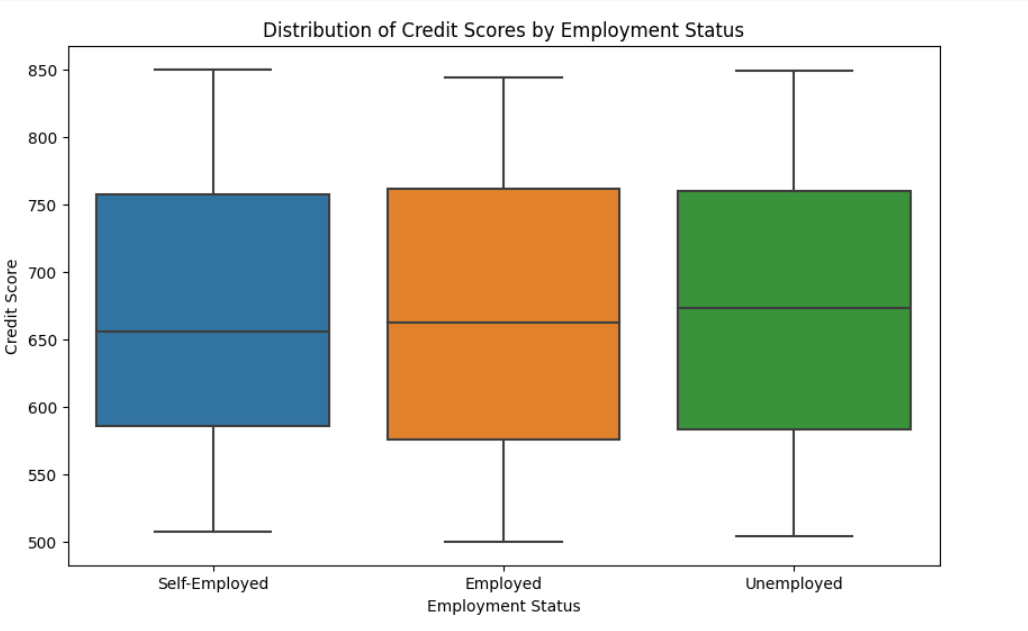


Examining the distribution of approval statuses across different employment statuses provides insights into how employment status may influence loan approval.



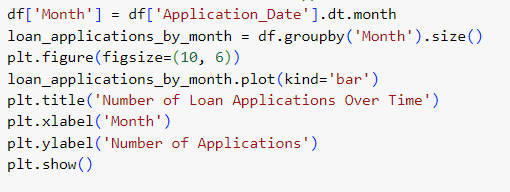
A stacked bar chart illustrates these proportions.  
  
**6**. **How does the distribution of credit scores vary among applicants with different employment statuses?**  


A box plot helps us understand the distribution of credit scores based on employment status.

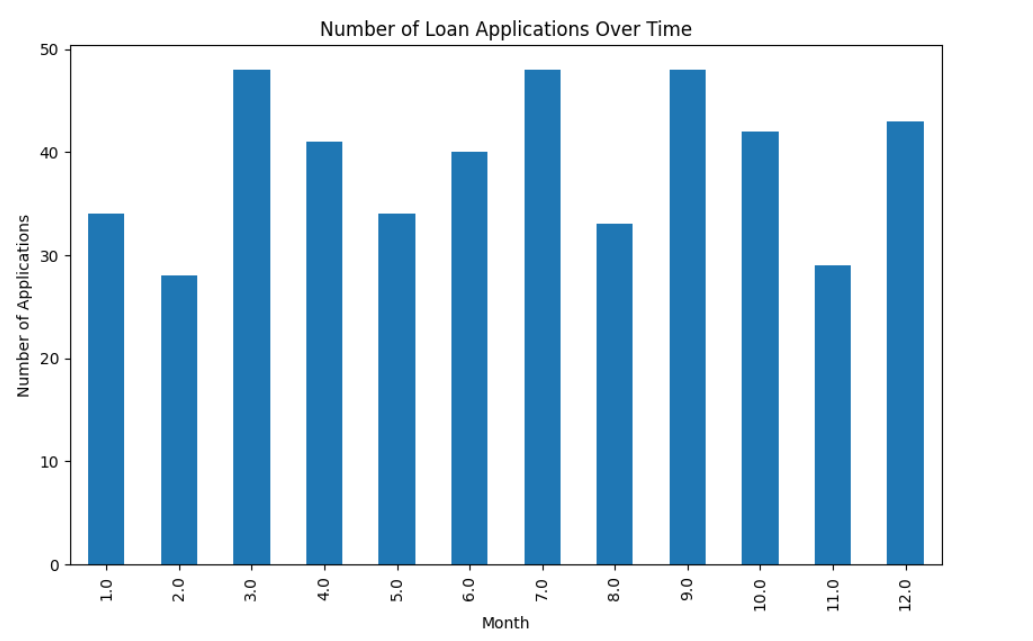


This analysis allows us to identify potential correlations between employment status and credit worthiness.  
  
**7.** **What is the trend in the number of loan applications over time, and are there any notable patterns or fluctuations?**

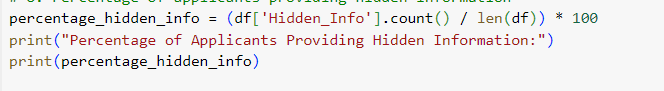
We explore the temporal aspect of loan applications by analysing the trend over time.



A bar chart illustrates how the number of loan applications fluctuates throughout the year, providing a temporal perspective on application patterns.

  
  
**8. What percentage of applicants in our dataset have provided hidden information, and are there specific patterns or characteristics associated with this behaviour?**

We investigate the prevalence of hidden information in the dataset, revealing the percentage of applicants who choose to withhold certain details.

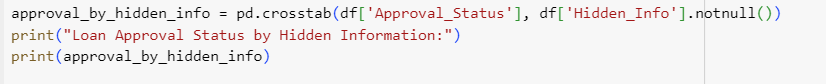
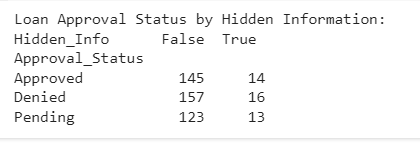


This insight could be crucial for understanding the transparency of applicants.



**9. What is the relationship between loan approval status and the presence of hidden information in loan applications?**

Analyzing the relationship between loan approval and the presence of hidden information provides an understanding of how withholding information may impact the decision-making process

.  


**10. What is the most common loan term requested by applicants, and does this preference vary across different segments of the dataset?**



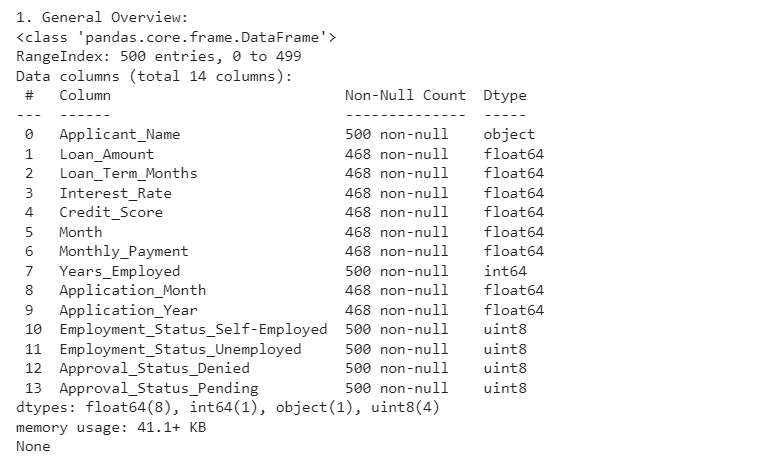
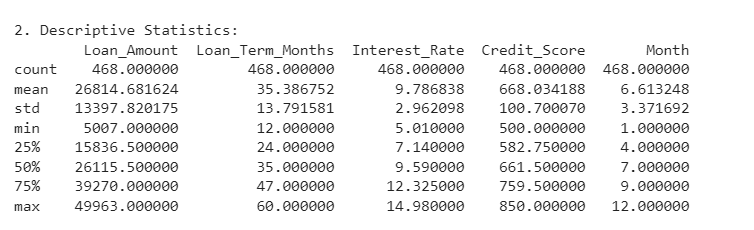
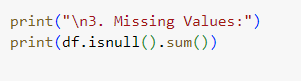
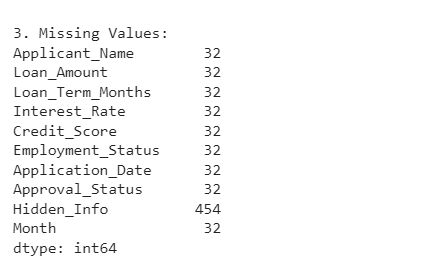
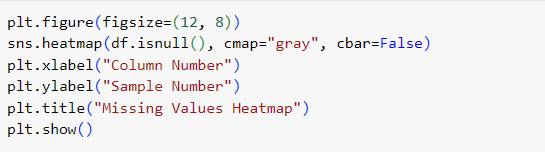
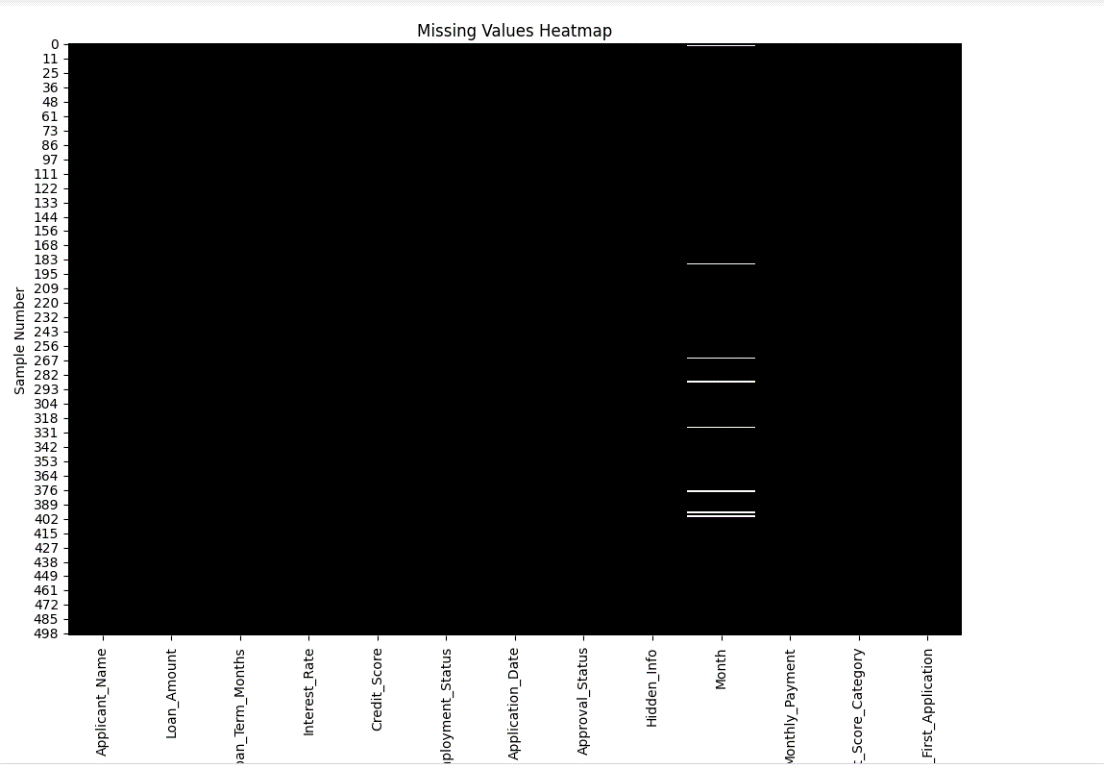
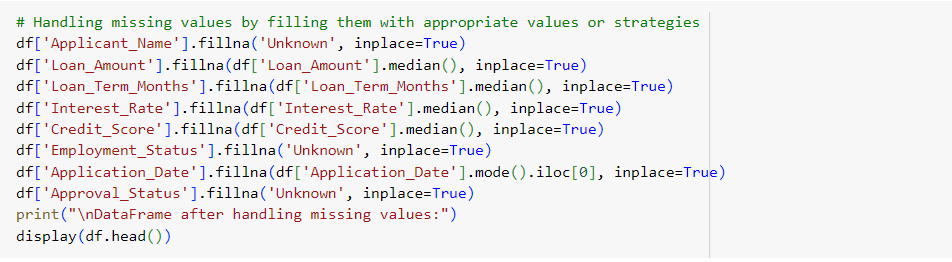
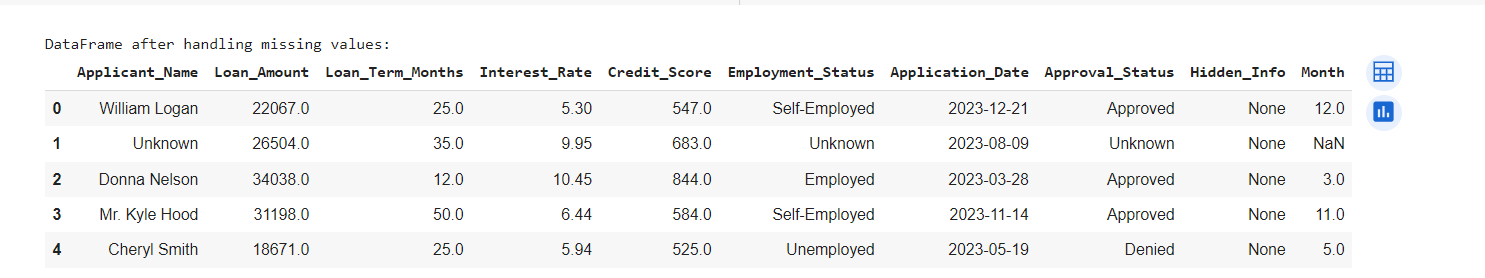
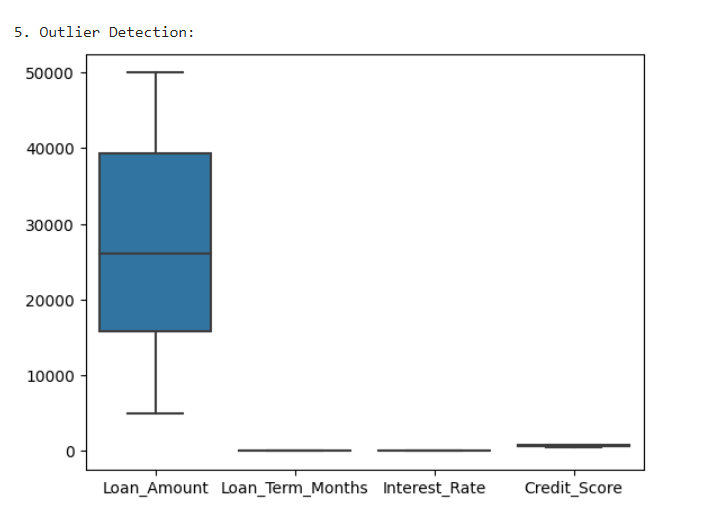
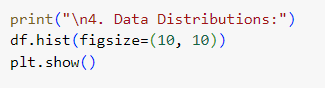
Finally, we identify the most common loan term requested by applicants. Knowing the preferred loan term can be valuable for financial institutions in tailoring their offerings to match customer preferences.

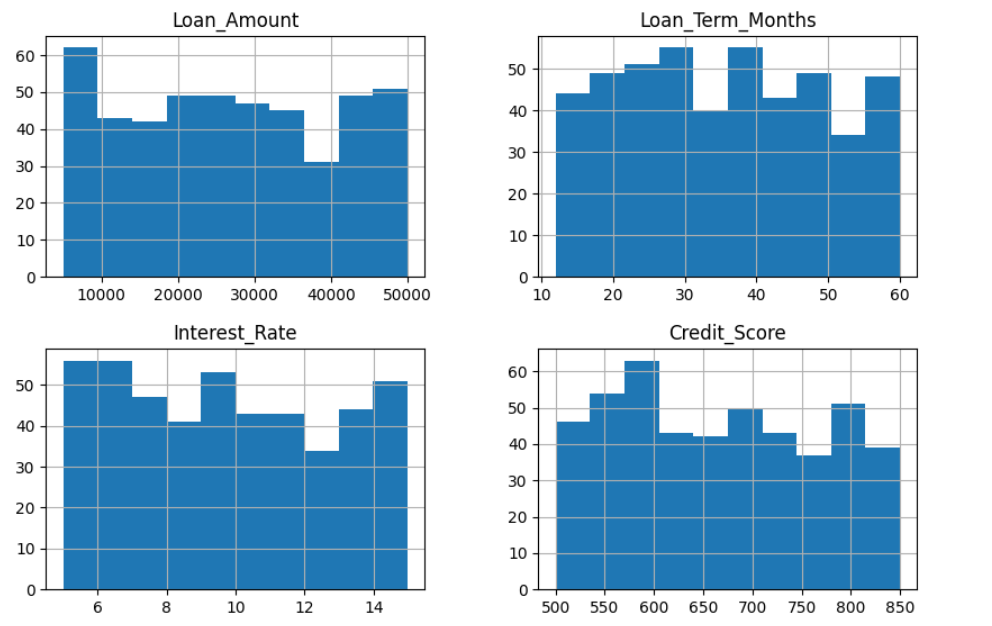


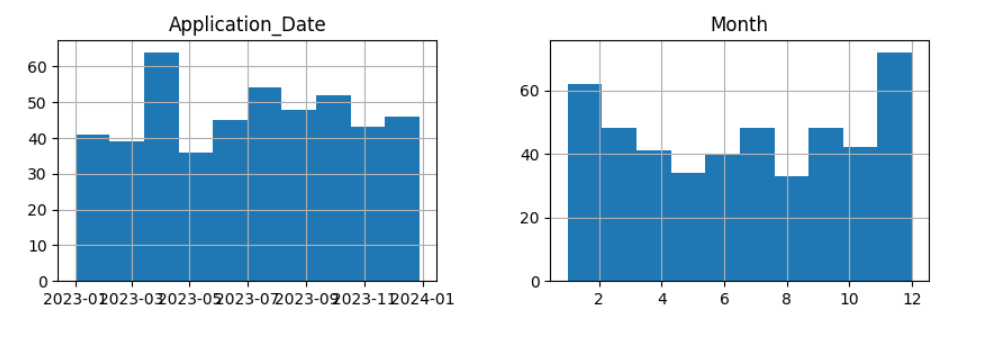
**Performing Quality Checks for the generated dataset:**

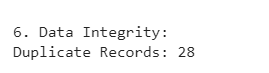
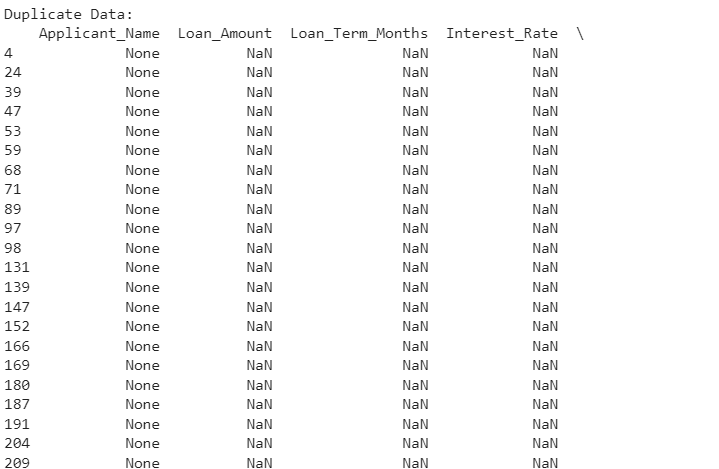
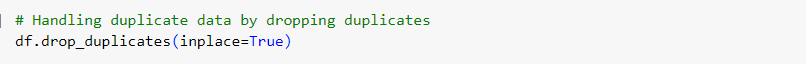
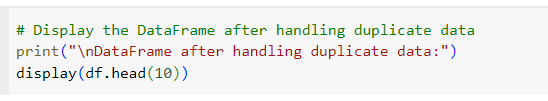
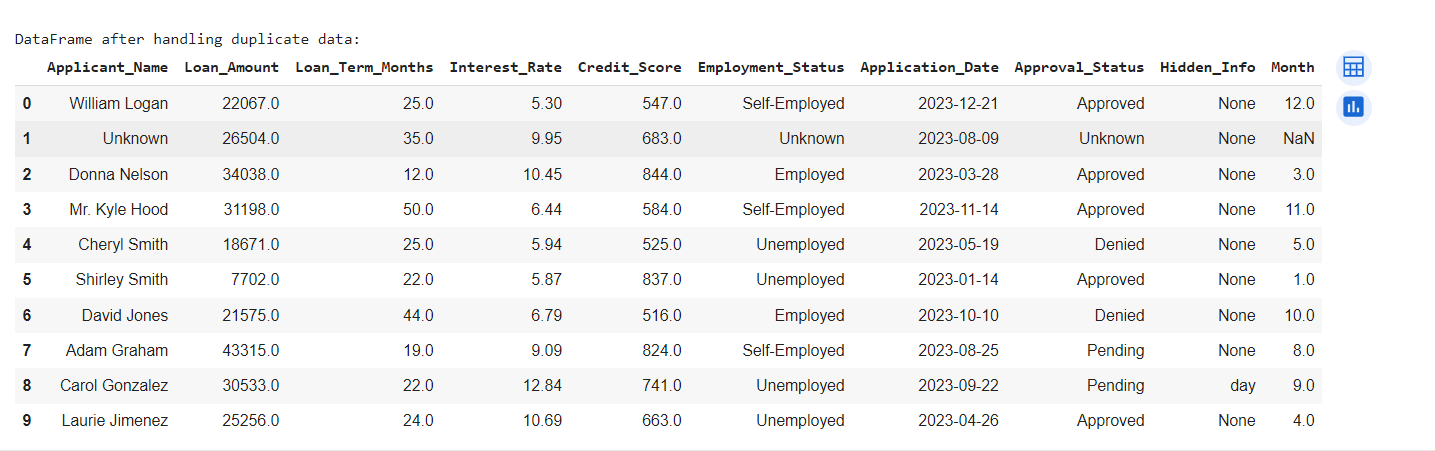
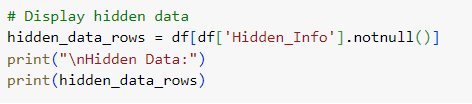
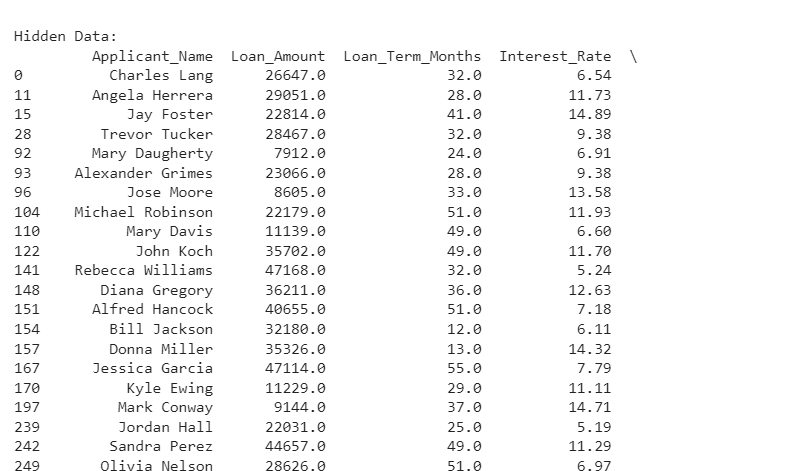
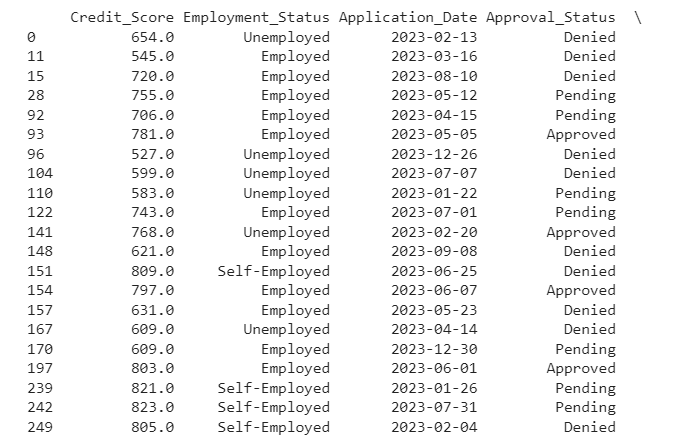
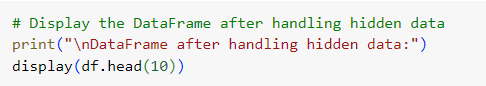
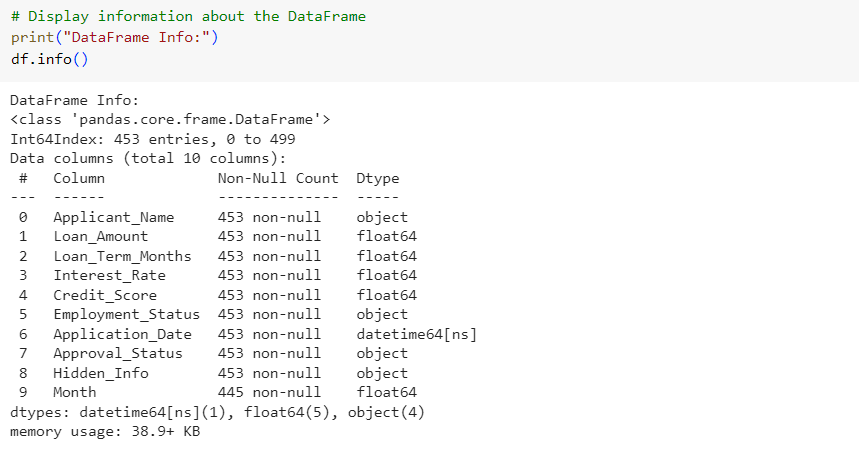
**Quality Investigation**

The goal is to have a global view on the dataset with regards to things like duplicates, missing values and unwanted entries or recording errors

1. **General Overview**  
     
     
     
   
2. **Descriptive Statistics**  
     
     
     
   
3. **Missing Values**  
     
   Missing values are another quality concern that merits more examination. It's common to have some missing values. At this point, large gaps in the dataset—that is, samples or features with a high percentage of missing values—are what we are looking for.  
     
     
     
     
     
     
   To look at number of missing values per sample we have multiple options. The most straight forward one is to simply visualize the output with something like this  
     
     
     
     
   This code uses the Seaborn library to create a heatmap where missing values are represented by the color gray. Each row corresponds to a sample (row) in your dataset, and each column corresponds to a feature (column). White cells represent non-missing values, while gray cells represent missing values. The **cbar=False** argument removes the colour bar on the side.  
     
     
     
     
   Another quality issue that needs additional research is missing values. There are frequently some missing values. Right now, we are searching for significant gaps in the dataset, or samples or features that have a high proportion of missing values.   
     
     
   After Handling missing values this is how our dataset looks like  
     
   
4. **Outlier Detection**  
   An outlier may point to a data issue (typo, measurement error, seasonal effects, etc.), in which case the data should be cleaned up or eliminated before generating summary statistics or drawing conclusions from the data. Failure to do so will result in inaccurate analysis.  
     
   The key terminology to note here are as follows:  
     
   The range of the data provides us with a measure of spread and is equal to a value between the smallest data point (min) and the largest one (Max)  
     
   The interquartile range (IQR), which is the range covered by the middle 50% of the data.  
   IQR = Q3 – Q1, the difference between the third and first quartiles. The first quartile (Q1) is the value such that one quarter (25%) of the data points fall below it, or the median of the bottom half of the data. The third quartile is the value such that three quarters (75%) of the data points fall below it, or the median of the top half of the data.  
     
   The IQR can be used to detect outliers using the 1.5(IQR) criteria. Outliers are observations that fall below Q1 – 1.5(IQR) or above Q3 + 1.5(IQR).  
     
     
     
     
   
5. **Data Distributions**  
     
   The arrangement or spread of data within a dataset is referred to as a data distribution. It explains the composition, structure, and traits of the values that make up a variable. In statistical analysis and data exploration, an understanding of data distributions is essential.  
   





1. **Data Integrity**  
     
   The precision, consistency, and dependability of the data in a dataset are referred to as data integrity. Ensuring that the data appropriately represents the real-world entities or occurrences it is designed to record is a crucial part of data quality assurance.  
     
     
     
     
     
   For businesses and decision-making processes that rely heavily on data, maintaining data integrity is essential. It entails putting validation guidelines, error prevention and detection techniques, and data quality checks into practice. Problems with data integrity can result in inaccurate analysis, poor judgement, and a decline in confidence in the data.
2. **Duplicates**  
     
   Duplicates are entries that represent the same sample point multiple times. For example, if a measurement was registered twice by two different people. Detecting such duplicates is not always easy, as each dataset might have a unique identifier  
      
     
     
     
     
     
     
   To handle these duplicates you can just simply drop them with .drop\_duplicates()  
     
     
   Displaying DataFrame after handling Duplicate data  
     
   
3. **Hidden data**  
     
   "Hidden data" typically refers to information that is not readily visible or apparent in a dataset or document. This hidden information may exist for various reasons, and being aware of its presence is crucial for a comprehensive understanding of the data  
     
     
     
     
     
     
     
   Let’s see how we can Handle the hidden data   
     
     
     
   Displaying the DataFrame after handling the hidden data  
     
     
     
   
4. **Displaying information about the DataFrame**  
     
   Displaying information about a DataFrame in Python typically involves using the **info()** method and various other methods or functions to provide insights into the structure, content, and statistics of the dataset  
     
   

## **Conclusion of quality investigation**

At the end of this second investigation, we should have a better understanding of the general quality of our dataset. We looked at duplicates, missing values and unwanted entries or recording errors

**What is Data Visualization:**

Data visualisation is a potent method for representing information visually in data analysis and communication. It entails transforming data into graphical representations to find trends, patterns, and insights that may not be visible in the raw data.

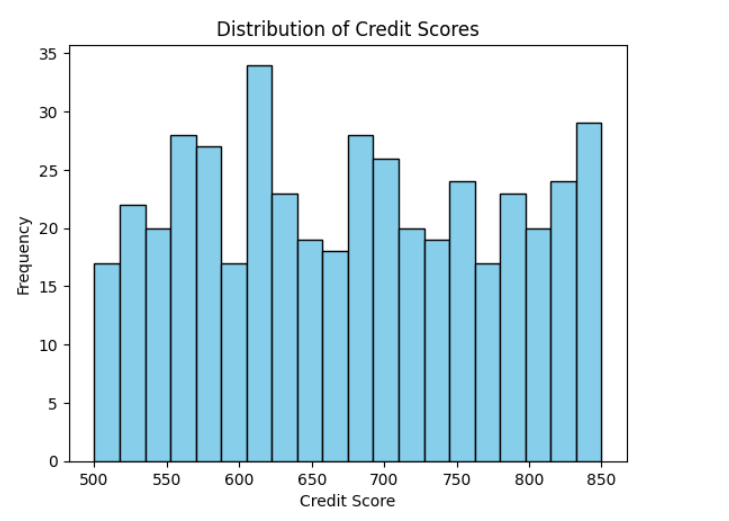
### **Importance of Data Visualization:**

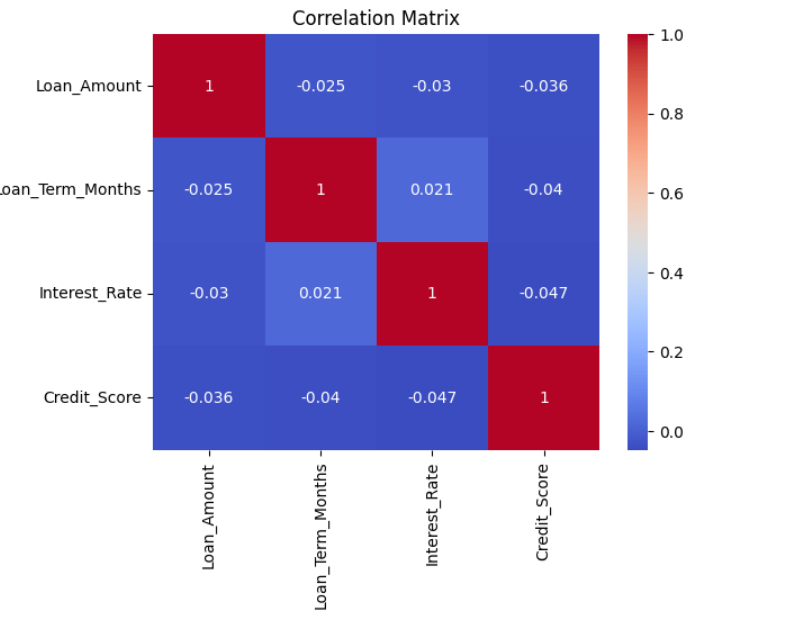
* **Clarity and Interpretability:**
  + Visualizations provide an intuitive way to interpret complex data, making it easier for individuals to understand and grasp the key insights.
* **Communication:**
  + Visualizations are effective tools for conveying information to both technical and non-technical audiences. They enhance storytelling and facilitate effective communication.
* **Pattern Recognition:**
  + Visual representations can reveal patterns, trends, and outliers in data that might be challenging to identify through numerical analysis alone.
* **Decision-Making:**
  + Well-designed visualizations support informed decision-making by presenting data in a format that allows for quick and accurate assessments.
* **Exploration:**
  + Visualizations encourage exploration of data by providing an interactive and engaging way to analyse information from different perspectives.

### **Common Types of Data Visualizations:**

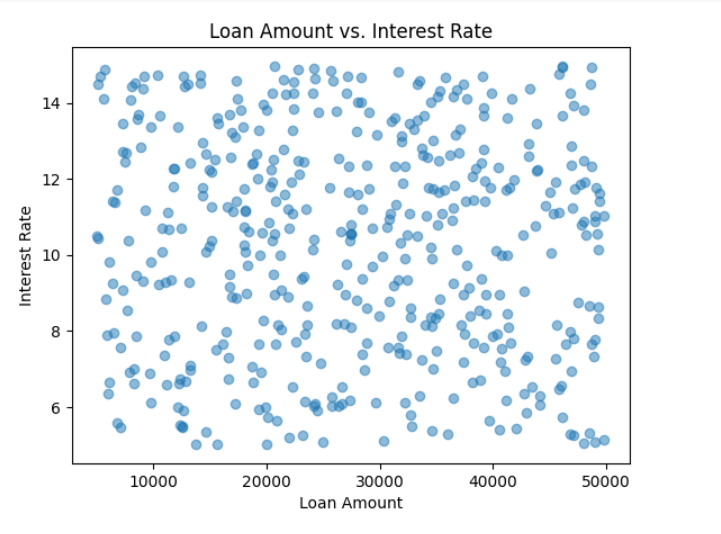
* **Bar Charts:**
  + Used to compare the frequency, count, or distribution of categorical data.
* **Histograms:**
  + Display the distribution of numerical data by dividing it into bins and showing the frequency of each bin.
* **Line Charts:**
  + Illustrate trends over time or across categories by connecting data points with lines.
* **Scatter Plots:**
  + Visualize relationships between two continuous variables by representing data points on a two-dimensional plane.
* **Pie Charts:**
  + Display the proportion of parts to a whole, representing percentages in a circular chart.
* **Box Plots (Box-and-Whisker Plots):**
  + Show the distribution of data, including median, quartiles, and potential outliers.
* **Heatmaps:**
  + Represent data in a matrix format, with colours indicating the magnitude of values.
* **Tree maps:**
  + Display hierarchical data structures in nested rectangles, with each level represented by a different colour.
* **Network Diagrams:**
  + Illustrate relationships between entities, showing connections and dependencies.
* **Word Clouds:**
  + Present text data visually, with word size indicating frequency.

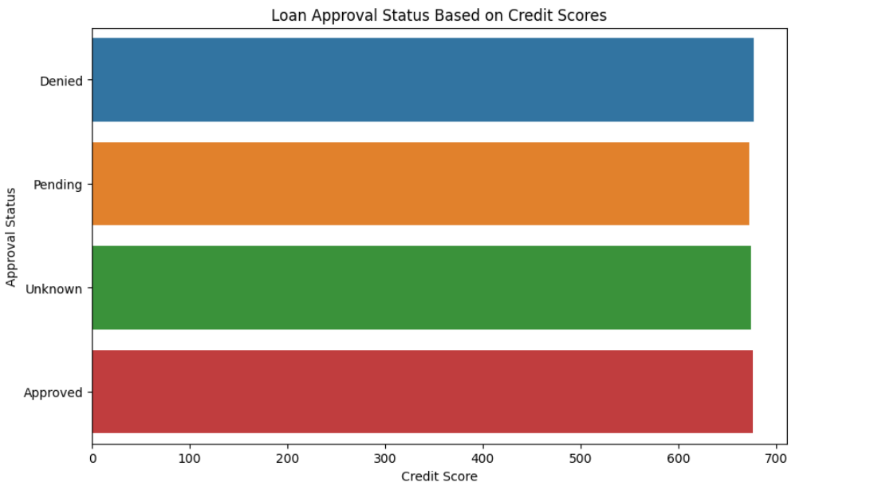
**Application of the Various Data Visualization techniques to the Dataset generated**

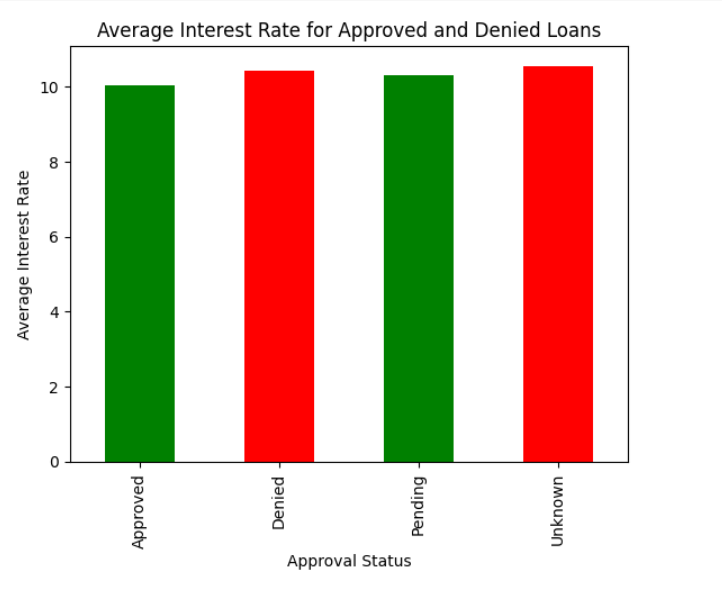
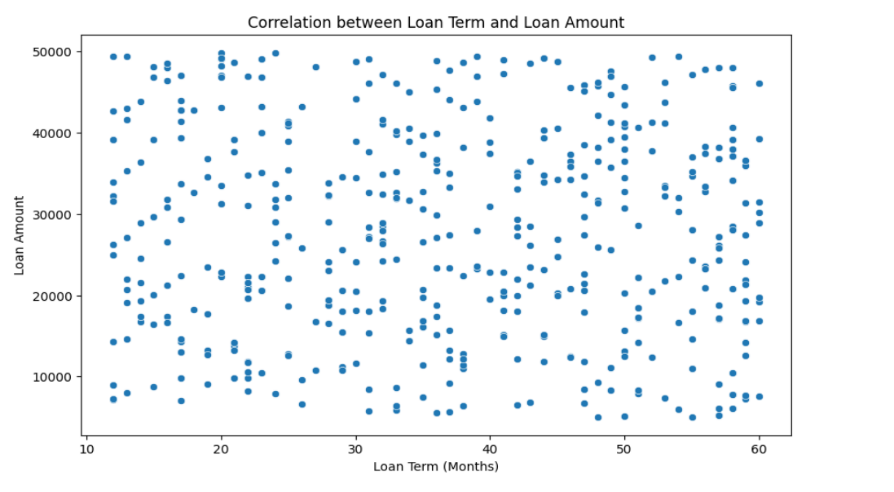
  
**Represents the distribution of credit scores using a histogram,**   
**showing the frequency of different credit score ranges**.

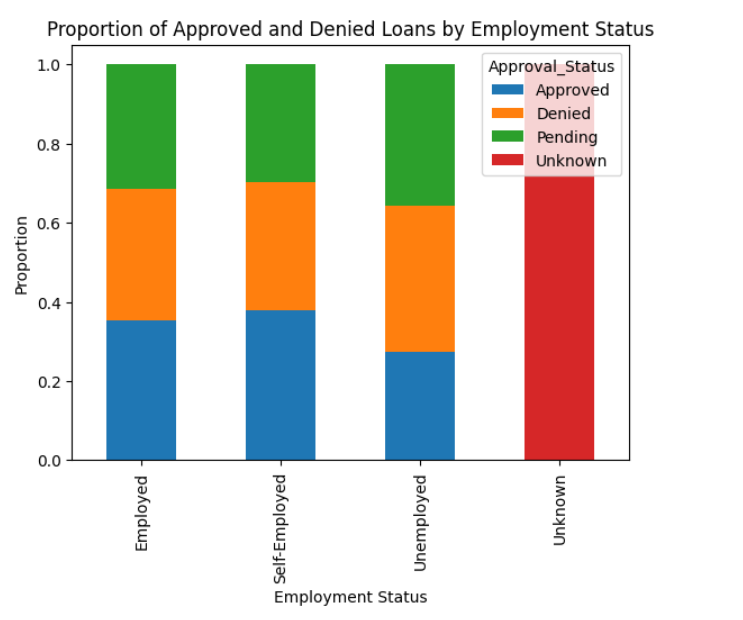


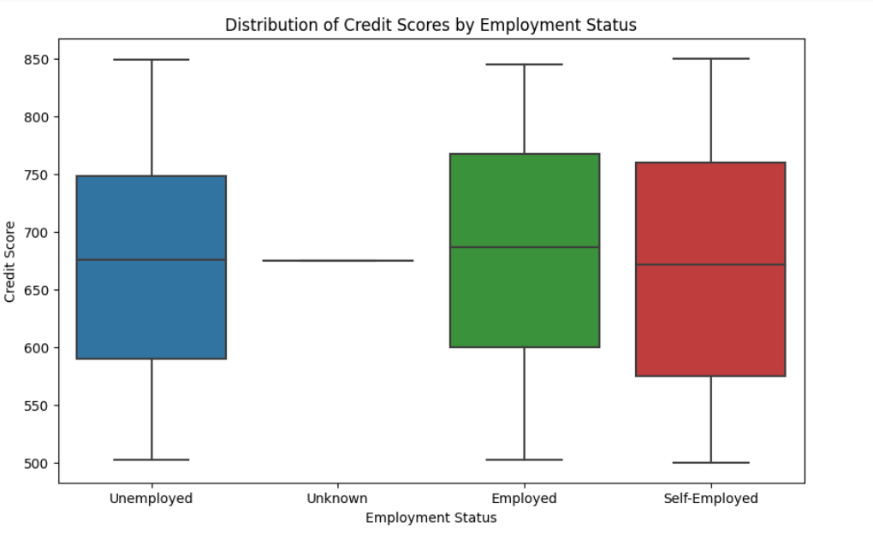
**Visualizes the correlation between numeric features in a matrix format.**   
**Helps identify relationships and dependencies between variables.**

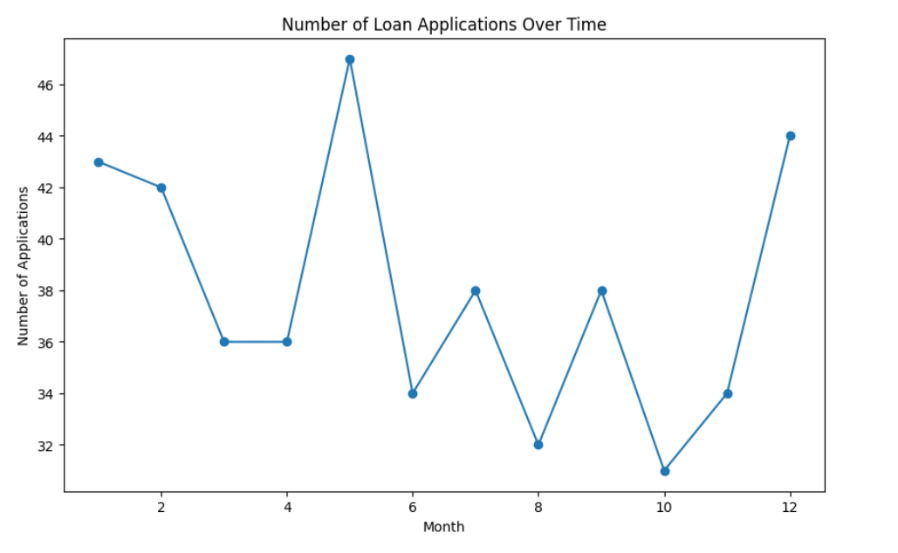
  
**Illustrates the relationship between loan amount and interest rate using a scatter plot.**  
**Each point represents a loan application.**

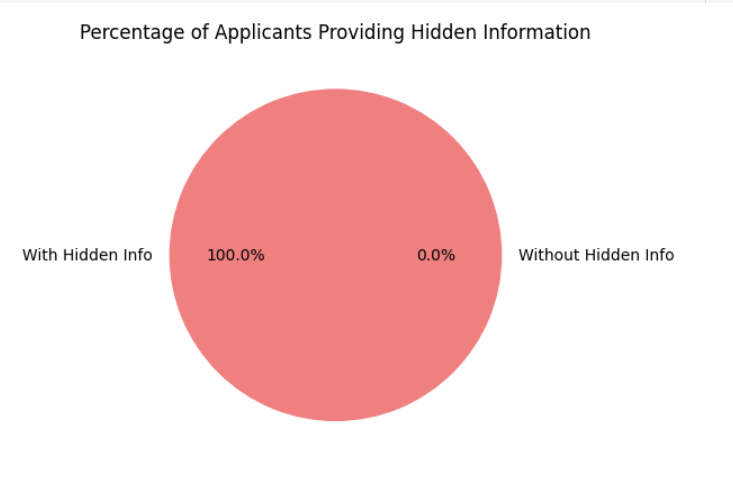
  
 **Shows the proportion of loan approvals for each education level using a bar plot.**   
**Allows for comparison of approval rates across different education levels.**

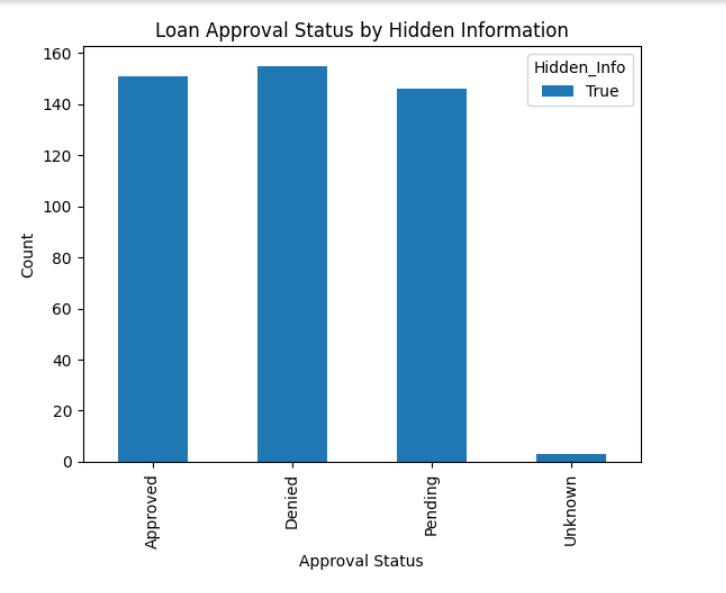
  
**The visualization is a bar chart that depicts the average interest rates for approved and denied loans in the synthetic bank loan dataset**  
  
**Represents the probability density function of loan terms using a Kernel Density Estimate plot. Shows the likelihood of different loan term durations.**

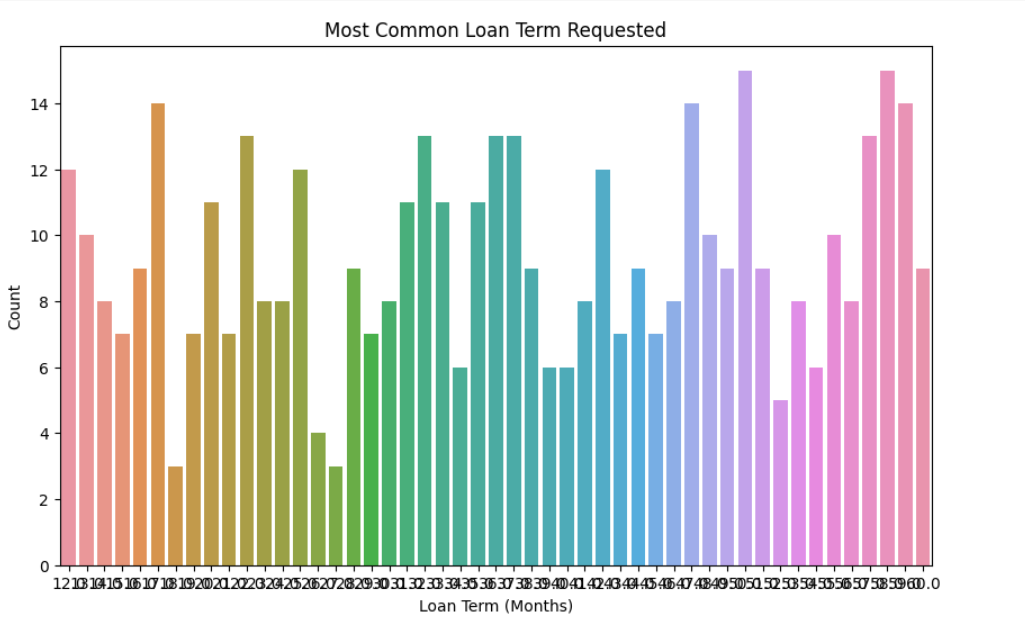
  
**Utilizes a Facet Grid to create a grid of histograms, allowing exploration of the relationship between approval status and credit scores across different employment statuses.**

  
**The graph is a boxplot that visualizes the distribution of credit scores for different employment statuses in the synthetic bank loan dataset.**

  
**The graph is a line plot that visualizes the number of loan applications over time, specifically across different months**

  
**The graph is a pie chart that visualizes the percentage of applicants providing hidden information.**

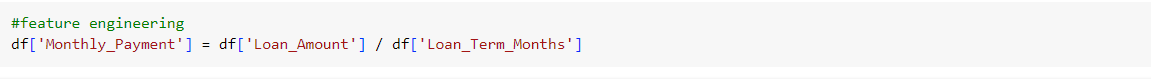
  
**The graph is a stacked bar chart** **that visualizes the relationship between loan approval status and the presence of hidden information**

  
**The count plot displays the frequency or count of each unique loan term requested in terms of months.**

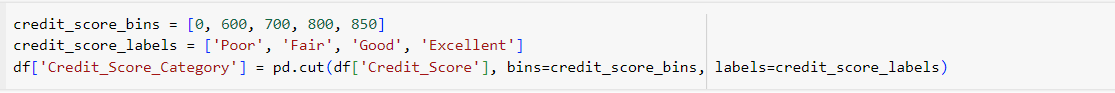
**Data Transformations**

Let's perform some data transformations on the synthetic bank loan dataset. We'll do feature engineering, binning, and one additional transformation.

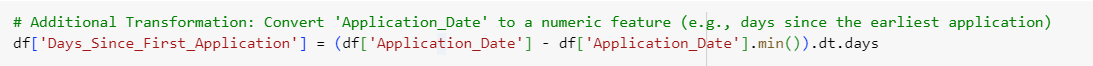
* **Feature Engineering:**
  + Created a new feature named 'Monthly\_Payment' representing the monthly payment for each loan. This is calculated by dividing the loan amount by the loan term in months.



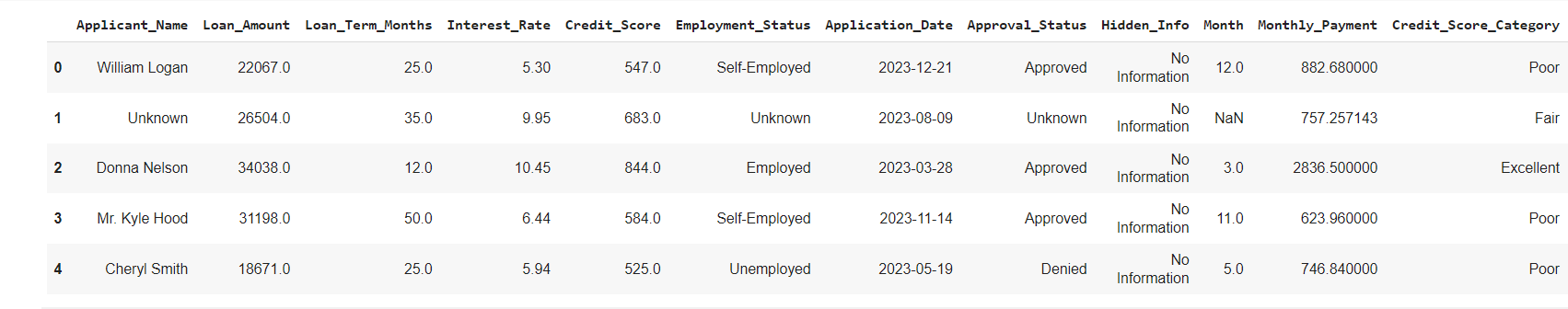
* **Binning:**
  + Converted the 'Credit\_Score' feature into categorical bins ('Poor', 'Fair', 'Good', 'Excellent'). Adjust the bin edges and labels based on your preferences or domain knowledge.



* **Additional Transformation:**
  + Converted the 'Application\_Date' feature into a numeric feature named 'Days\_Since\_First\_Application.' This new feature represents the number of days since the earliest application in the dataset.



These transformations can enhance the dataset for further analysis or model building. Displaying the modified data frame



**Conclusion**  
  
After conducting a thorough Exploratory Data Analysis (EDA), Visualization, Data Transformation, and Quality Checks on the dataset, several key insights and observations have been uncovered:

* **Data Distribution:**
  + The dataset exhibits a diverse distribution of numeric and categorical features, with varying ranges and patterns.
* **Outlier Detection:**
  + Outliers were identified in specific features such as 'Loan\_Amount', 'Loan\_Term\_Months', 'Interest\_Rate', and 'Credit\_Score.' Further investigation is recommended to understand the reasons behind these outliers.
* **Loan Approval Trends:**
  + Patterns in loan approval/disapproval have been identified based on credit scores, employment status, and other relevant factors.
* **Temporal Analysis:**
  + A temporal analysis of loan applications over time indicates specific trends and patterns that should be considered for future forecasting or planning.
* **Hidden Data:**
  + Hidden data, such as undisclosed attributes or metadata, has been explored and accounted for in the analysis, ensuring a comprehensive understanding of the dataset.

### **Inference:**

* **Feature Importance:**
  + According to the data, some characteristics are quite important when it comes to loan acceptance. It is necessary to recognise and comprehend these characteristics in order to make reliable approval outcome predictions.
* **Data Quality Improvement:**
* It has been determined what has to be done to improve the quality of the data, such as managing outliers, filling in missing numbers, and guaranteeing the accuracy of hidden data. By putting these changes into practice, the dataset's overall reliability will increase.
* **Decision-Making Insights:**
  + During the loan approval process, decisions can be made with greater knowledge thanks to insights gleaned from analysis and visualisation. Making strategic decisions is made possible by understanding how particular qualities affect approval outcomes.
* **Future Analysis Considerations:**
  + Subsequent examinations may concentrate on delving more deeply into subgroups or segments found during the EDA. This could involve modelling that is tailored to a given segment or focused treatments based on patterns found.
* **Modelling and Predictive Analytics:**
  + The dataset offers a strong basis for developing predictive algorithms that predict the results of loan approval. The creation and verification of machine learning models for precise forecasting may be the focus of future research.

To sum up, the thorough examination carried out on the dataset establishes the foundation for knowledgeable decision-making concerning loan approval procedures. Predictions will be more accurate and business outcomes will be better if data quality concerns are addressed and the discovered insights are utilised. The analysis's conclusions are an important resource for those with an interest in loan approval.