
LOAN
APPROVAL
LOAN APPROVED
LOAN APPROVED
LOAN APPROVED
LOAN
EXPLORATORY DATA ANALYSIS
LOAN APPPROVAL

INTRODUCTION

Introduction:

The Home Loan Approval Dataset is a comprehensive collection of data sourced from Skill Circle, containing valuable information for analysis in the domain of home loan approvals. This dataset is particularly relevant for financial institutions, analysts, and researchers interested in understanding the factors influencing home loan approval decisions.

The dataset encompasses various attributes such as applicant income, loan amount, loan term, credit history, property area, and the ultimate decision on loan approval. Through Exploratory Data Analysis (EDA), we aim to gain insights into the patterns, trends, and factors that significantly affect the approval of home loan.

By performing data exploration and visualization techniques, we will delve into the distribution of numeric and categorical variables, identify potential outliers, explore relationships between different attributes, and uncover valuable insights that can aid in making informed decisions related to home loan approval processes.

Dataset Information:

The Home Loan Approval Dataset, sourced from Skill Circle, is designed to facilitate the analysis of home loan approval processes. The dataset contains a total of 614 entries, each representing a home loan application. There are 13 attributes in the dataset, including both numeric and categorical variables.

Attributes:

- Loan_ID: Unique Loan ID
- Gender: Gender of the applicant (Male/Female)
- Married: Applicant's marital status (Yes/No)
- Dependents: Number of dependents
- Education: Applicant's education level (Graduate/Not Graduate)
- Self_Employed: Self-employed status of the applicant (Yes/No)
- ApplicantIncome: Applicant's monthly income
- CoapplicantIncome: Co-applicant's monthly income
- LoanAmount: Loan amount in thousands
- Loan_Amount_Term: Term of the loan in months
- Credit_History: Credit history meets guidelines (1 for Yes, 0 for No)
- Property_Area: Area of the property (Urban/Semiurban/Rural)
- Loan_Status: Loan approval status (Y for Yes, N for No)

The primary objective of this dataset is to analyze the factors influencing home loan approval decisions. By performing Exploratory Data Analysis (EDA), we aim to gain insights into the patterns, trends, and relationships within the data. Through data exploration and visualization techniques, we seek to identify key factors affecting loan approval and provide actionable insights for financial institutions and policymakers

Here are some major libraries you'll likely import for working with the home loan approval dataset:

Essential Libraries:

- pandas (pd): This is the workhorse library for data analysis in Python. It allows you to:

Load data from various file formats (CSV in this case) using pd.read_csv.
Manipulate and clean data using functions for filtering, sorting, and handling missing values.
Explore and summarize data with descriptive statistics (describe) and data structures like DataFrames.

- matplotlib.pyplot (plt): This library provides functionalities for creating various visualizations like:
- Histograms to explore the distribution of numeric variables.
- Box plots to visualize the spread of data and identify outliers.
- Scatter plots to explore relationships between pairs of numeric variables.
- Bar charts and pie charts to represent categorical data.

Optional Libraries (depending on your analysis goals):

- seaborn (sns): Built on top of matplotlib, seaborn offers a high-level interface for creating more aesthetically pleasing and informative visualizations. It provides functions similar to those in matplotlib but with a more polished look and advanced features.

- numpy (np) (if needed): While pandas is sufficient for most data manipulation tasks in this case, NumPy might be useful for specific operations like numerical computations or advanced array manipulations.

Source: Skill Circle

Content: Information related to loan applications and their approval status.

Format: Comma-separated values (CSV) file (assuming based on the code snippet pd.read_csv).

Expected Columns:

Applicant Demographics: While the specific details might vary, some possibilities include:

- Applicant Age (numerical)
- Marital Status (categorical - Married, Single, Divorced, etc.)
- Employment Status (optional, categorical)
- Dependents (optional, numerical)

Loan Characteristics:

- Loan Amount (numerical, in USD or other currency)
- Loan Term (optional, numerical, duration of the loan)
- Interest Rate (optional, numerical)

Credit History:

- Credit Score (numerical)
- Delinquency History (optional, categorical - Yes/No or similar)

Property Details (optional):

- Location (categorical - zip code, city, etc.)
- Property Type (categorical - house, condo, etc.)
- Property Value (optional, numerical)

Loan Approval Status: (categorical)

Approved
Rejected
Pending (optional)

The data is stored in a CSV file named "home_loan_approval.csv", use the following code:

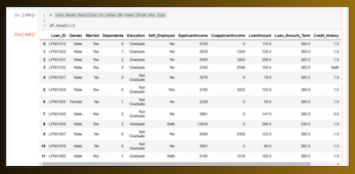
LOADING AND INSPECTING THE DATASET. ¶

```
df = pd.read_csv("C:/Users/Preeti/Documents/loan_sanction_test.csv")
```

Head Function:

Data is loaded into a variable named data (as in the previous steps we discussed), you can use the following code to view the first 12 rows:

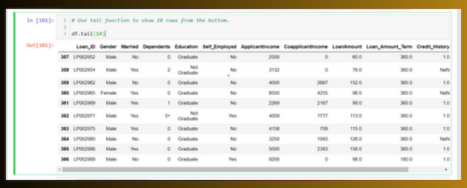
The head() function is a common method in pandas used to display the first few rows of a DataFrame. By specifying the number 12 within the parenthesis, you instruct it to return only the first 12 rows instead of the default 5.



Tail function:

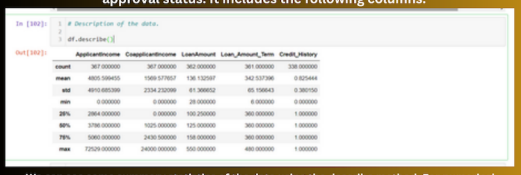
Use the tail function to view the last 10 rows of the home loan approval dataset:

Similar to head(), tail is a pandas function for DataFrames that retrieves a specific number of rows. In this case, specifying 10 within the parenthesis instructs it to return the last 10 rows



Description of the data:

The data describes characteristics of loan applicants and their loan approval status. It includes the following columns:



	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	367.000000	367.000000	362.000000	361.000000	338.000000
mean	4805.599455	1569.577657	136.132597	342.537396	0.825444
std	4910.685399	2334.232099	61.366652	65.156643	0.380150
min	0.000000	0.000000	28.000000	6.000000	0.000000
25%	2864.000000	0.000000	100.250000	360.000000	1.000000
50%	3786.000000	1025.000000	125.000000	360.000000	1.000000
75%	5060.000000	2430.500000	158.000000	360.000000	1.000000
max	72529.000000	24000.000000	550.000000	480.000000	1.000000

We can see some summary statistics of the data using the describe method. For numerical columns (like Applicant Age and Credit Score), it provides the mean, standard deviation, minimum, quartiles (25% and 75%), and maximum values. For categorical columns (like Marital Status and Loan Approval Status), it shows the number of unique entries and the most frequent entry.

Information about the Dataset:

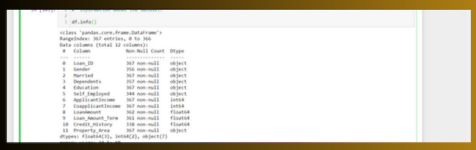
Data Types:

Numeric:

- Applicant Age (years)
- Credit Score
- Loan Amount (USD)

Categorical:

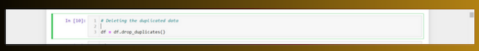
- Marital Status (e.g., Married, Single, Divorced)
- Loan Approval Status (e.g., Approved, Rejected, Pending



Using duplicated() function (pandas):

This function in pandas identifies rows that are entirely duplicates based on all columns. You can specify which columns to consider for duplication using the subset parameter.

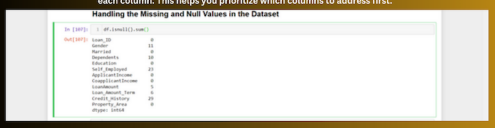
Deleting the Duplicated Data:



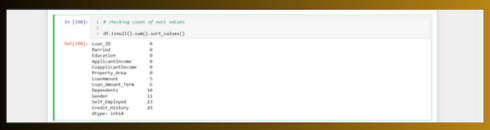
Data Cleaning:

Handling Missing Values:

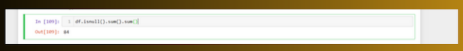
Identify Missing Values: Utilize the data.isnull().sum() function in pandas to identify the number of missing values present in each column. This helps you prioritize which columns to address first.



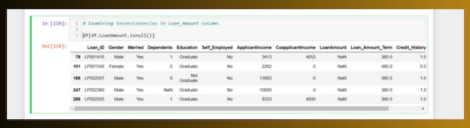
This code helps you identify which columns in your DataFrame have the most missing values. This information is crucial for data cleaning tasks like deciding which columns to handle (e.g., impute missing values, drop columns) or prioritizing which columns to address first based on the severity of missing data.



To calculate total missing values:



To examine inconsistencies in the "Loan_Amount" column of your home loan approval dataset:

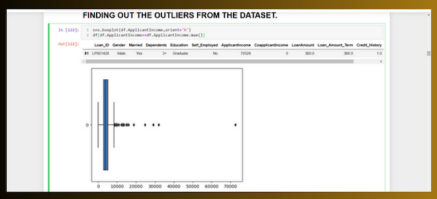


To find and replace outliers across the entire numerical dataset

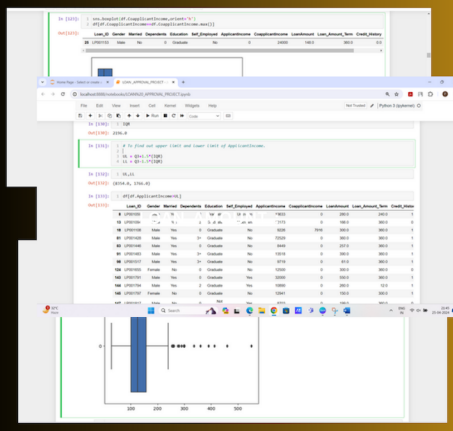
To summarize the basic statistics for the numeric columns in home loan approval dataset:

Finding outliers in your home loan approval dataset:

APPLICANT INCOME :



COAPPLICANT INCOME :



To find IQR and treat outliers with upper limit:

1. identify IQR (Interquartile Range)

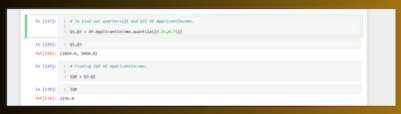
The interquartile range (IQR) is a measure of the spread of data, specifically the range that contains the middle 50% of the data. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) of the data.

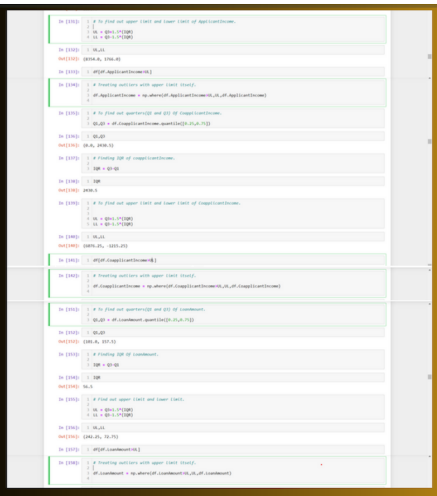
2. Detect outliers using IQR rule

An outlier is a data point that falls outside of 1.5 times the IQR below the first quartile (Q1) or 1.5 times the IQR above the third quartile (Q3).

3. Treat outliers by capping them to the upper bound

One way to handle outliers is to cap them to the upper bound. This means replacing the outlier values with the value of the upper bound.





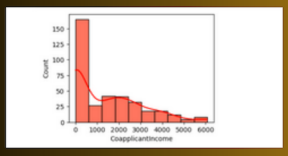
OUTLIERS REMOVED SUCCESSFULLY IN ALL THE COLUMNS :

DATA VISUALIZATION:

1. UNI-VARIATE ANALYSIS :

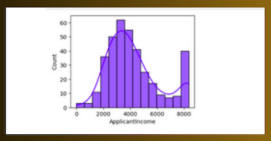
- **Representation of data by using Histogram**

DISTRIBUTION OF COAPPLIANT INCOME



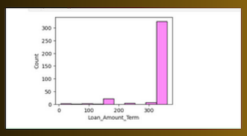
This histogram illustrate the distribution of Coapplicant income. The x-axis represents "Coapplicant Income"(ranging from 0 to 6000),while the y-axis shows "count"(ranging grom 0 to 150). A Prominent bar reaches up to approximately 150 count for incomes close to zero,indicating many coapplicants have little to no income.The frequency of higher incomes diminishes,with smaller bars represnting fewer coapplicants in those income brackets. overall,this graph provides insights into the financial standing of coapplicants

DISTRIBUTION OF APPLICANT INCOME



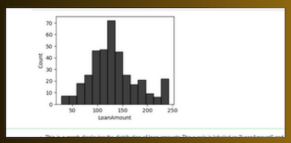
- Distribution of Applicant Income: The histogram provides a visual representation of the distribution of applicant incomes in the dataset. It shows the frequency or count of applicants falling within different income ranges.
Central Tendency: The shape of the histogram can indicate the central tendency of applicant incomes. For example, a peak around a specific income range may suggest a common income level among applicants.
- Variability: The spread of the histogram can give insights into the variability of applicant incomes. A wider spread indicates a broader range of income levels among applicants.
- Skewness: The symmetry or skewness of the histogram can reveal whether the distribution of applicant incomes is skewed to the left (negatively skewed), to the right (positively skewed), or symmetric.
- The image shows a histogram overlaid with a line garph,illustrate the distribution of "Applicant Income"(ranging from 0 to 8000),while the y-axis shows count(ranging from 0 to 60).Most applicants fall within the income range of 2000 to 4000,with an interseting spike at higher incomes. The graph provides insights into the income distribution among a specific group.

DISTRIBUTION OF LOAN AMOUNT TERM



The image is a line garph showing the number of loans issued based on the loan term. The x- axis labeled on "Loan_Amount_Term" likely represents the length of the loan in years. The y-axis labeled "count" represents the number of loans. It appears that shorter loan terms are more common than longer terms.

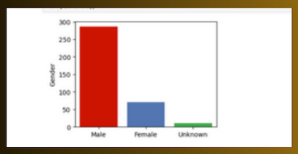
DISTRIBUTION OF LOAN AMOUNT



This is a bar graph displaying the distribution of loan amounts. The x-axis is labeled 'LoanAmount' and ranges from 0 to 250,while y-axis is labeled"count" and ranges from 0 to 70. A prominent bar reaches up to approximately 150 count for incomes close to zero,indicating that many coapplicants have little to no income. The frequency of higher incomes diminishes,with smaller bars representing fewer coapplicants in those income brackets.overall this graph provides insights into the financial standing of coapplicants

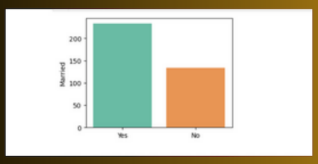
BAR CHARTS:

DISTRIBUTION OF GENDER:



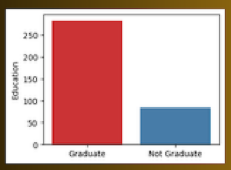
This visualization reveals the distribution of genders among the loan applicants . It appears that there is a noticeable imbalanace between male, female and unknown applicants . This could be due to various socio-economic factors influencing the willingness of individuals to apply for loans.

DISTRIBUTION OF MARRIED:



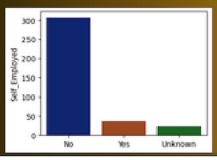
This bar plot visualization the distribution of appicants based on their marital status, with categories such as 'Married' and 'Not Married'.There is a noticeable imbalance in the distribution, there are more married applicants comapred to unmarried ones. It could reflect societal trends or demographic characterstics of the dataset.

DISTRIBUTION OF EDUCATION:



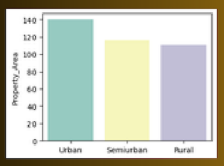
This This count plot visualization the distribution of appicants based on their education status, with categories such as 'Educate' and 'Not Educate'.There is a noticeable imbalance in the distribution, there are more graduate applicants comapred to uneducated ones.

DISTRIBUTION OF SELF-EMPLOYED:



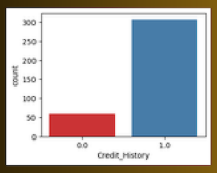
- This count plot illustrates the distribution of applicants based on their self employment status.Categories may include 'Yes'(indicating not self employed and 'No'(indicating not self employed)).There is as significant count in the 'Yes' category,it indicates a not able presence of self-employed individuals or entreprenuers among the loan applicants. Self employed individuals may have variable income structures compared to salaried individuals.Understanding their distribution provides insights into the diersity of income sources among loan applicant.

DISTRIBUTION OF PROPERTY_AREA:



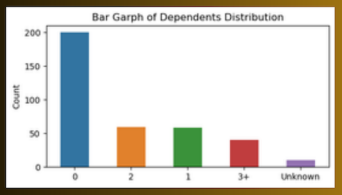
This count plot visualization the distribution of appicants based on their Propert_Area status, with categories such as 'Urban','Semiurban' and 'Rural'.There is a noticeable imbalance in the distribution, there are most Urban then Semiurban applicants comapred Rural ones.

DISTRIBUTION OF CREDIT_HISTORY:



This count plot visualization the distribution of appicants based on their credit_history status, with categories such as '0.0' and '1.0'.There is a noticeable imbalance in the distribution, there are more 1.0 applicants comapred to 0.0 ones.

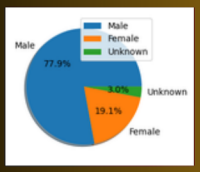
DISTRIBUTION OF DEPENDENTS:



The resulting bar graph provides a clear visualization of the distribution of dependents, allowing users to quickly identify the most common number of dependents (0.0, with a count of 200) and the least common numbers (2.0 and 1 .0,3+,unknownwith counts of 59 ,58,40 and 10 respectively).

PIE CHARTS:

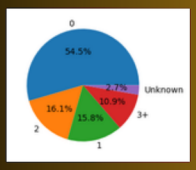
DISTRIBUTION OF GENDER:



- The pie chart illustrates the distribution of genders within the dataset.
- Each slice represents a unique gender category, with sizes proportional to the frequency of each gender.
- The distribution of slices are as male have 77.9% , female have 19.1% and other unknown have 3.0% respectively.

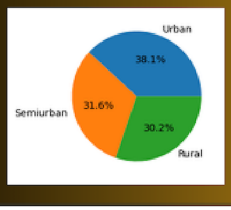
Gender Representation: The chart visually displays the relative proportions of different genders.
- Percentage Breakdown: The percentages on each slice indicate the proportion of each gender category in the dataset.
- Imbalance Detection: Shows for any significant dominant gender categories.

DISTRIBUTION OF DEPENDENTS:



This pie chart represent slices of dependendent distribution as
In 0 have 54.5%
In 2 have 16.1%
In 1 have 15.8%
In 3+ have 10.9%
and other unknown have 2.7%

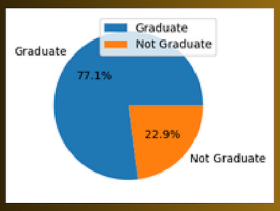
DISTRIBUTION OF PROPERTY AREA:



The pie chart illustrates the distribution of Property_Area within the dataset.
Each slice represents a unique Property_Area category, with sizes proportional to the frequency of each Property_Area.
Property_Area Representation: The chart visually displays the relative proportions of different Property_Area.
Percentage Breakdown: The percentages on each slice indicate the proportion of each Property_Area category in the dataset.
Imbalance Detection: Shows for any significant dominant Property_Area categories.
The following distribution of Property_Area slices are as:

Urban = 38.1%
Rural = 30.2%
Semiurban = 31.6%

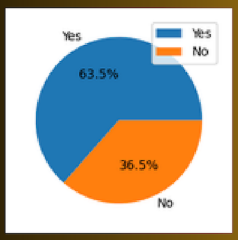
DISTRIBUTION OF EDUCATION:



The pie chart illustrates the distribution Of education are as :

- There are 77.1% Graduates.
- There are 22.9% Non-Graduates.

DISTRIBUTION OF MARRIED:



The pie chart illustrates the distribution of Married's are as follows:

- There are 63.5% Married.
- There are 36.5% non-Married.

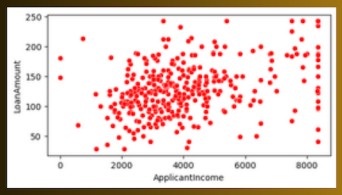
BI - VARIATE ANALYSIS:

SCATTER PLOTS:

SCATTER PLOT OF COAPPLICANTINCOME BY APPLICANT INCOME BY PROPERTY AREA:

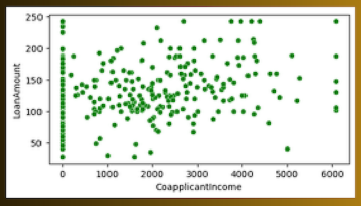
This scatter plot illustrates the x- axis of applicantincome vs y-axis of coapplicantincome by property area(rural,urban,semiurban)respectively.

DISTRIBUTION OF APPLICANTINCOME AND LOAN AMOUNT:



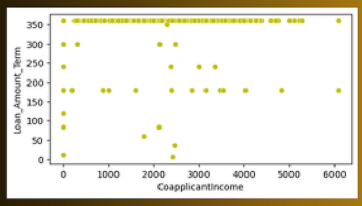
There is a positive correlation between applicantincome and loan amount.This means that as the applicant's income increases,they may be eligible for larger loan amounts. Applicants with higher incomes may qualify for larger loans beacuse they have financial capacity to repay them.

DISTRIBUTION OF COAPPLICANTINCOME AND LOAN AMOUNT:



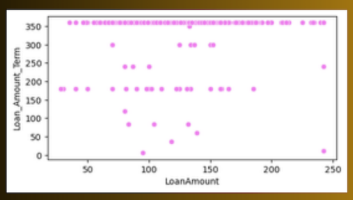
This is a scatter plot graph displaying the relationship between Coapplicant Income and LoanAmount. The x-axis represents"CoapplicantIncome(ranging from 0 to 6000),while the y-axis shows LoanAmount(ranging from 0 to 250).Numerous green dots represent the individaual data points scatteres across the graph,indicating various combinations of coapplicant income and loan amounts.A scatter plot illustarting the correlation between coapplicant income and loanamount ,providing insights into trends or patterns.

DISTRIBUTION OF COAPPLICANTINCOME AND LOAN_AMOUNT_TERM:



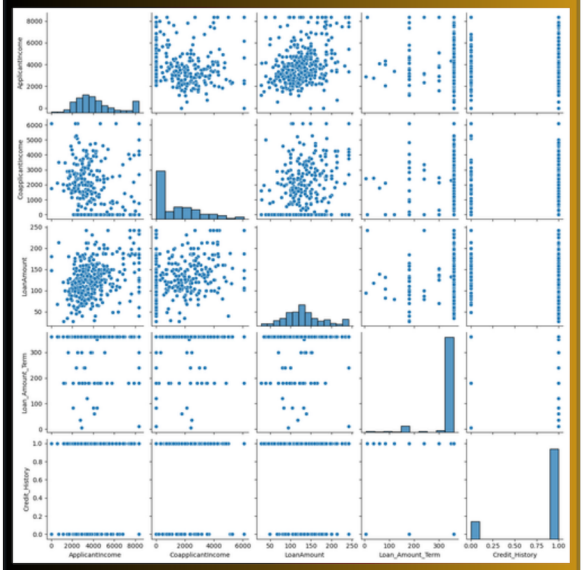
This scatter graph displaying the relationship between coapplicantincome and loan amount term.The x-axis have coappicantincome (range from 0 to 6000)and y-axis have Loan_Amount_Term(range from 0 to 350).

DISTRIBUTION OF LOANAMOUNT AND LOAN_AMOUNT_TERM:



This scatter graph displaying the relationship between LoanAmount and Loan_Amount_Term.The x-axis have LoanAmount (range from 0 to250)and y-axis have Loan_Amount_Term(range from 0 to 350)respectively.

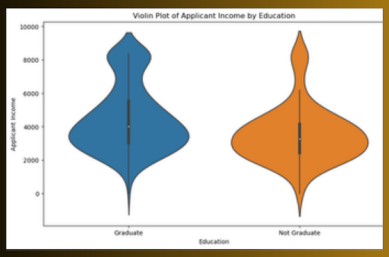
PAIR PLOT:



The Pairplot analysis reveals a positive correlation between ApplicantIncome and LoanAmount, suggesting that higher income leads to a higher loan eligibility.

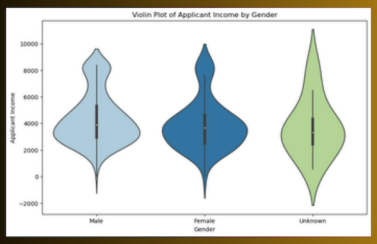
VIOLIN PLOTS:

PLOT OF APPLICANTINCOME BY EDUCATION:



-This violin plot showing income distribution for applicants with a graduate degree compared to those without.The vertical axis shows income, while the horizontal axis represents education level. The wider the violin shape,the greater the variation in income.The plot suggests that applicants have higher incomes than those without a graduate degree.

PLOT OF APPLICANTINCOME BY GENDER:

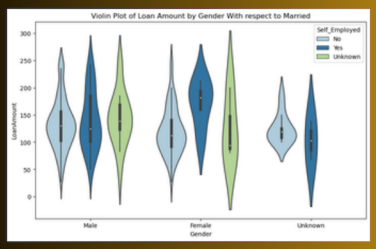


The violin plot provides a good starting point to compare the distribution of applicant income by gender and identify potential areas for further investigation.

PLOT OF APPLICANTINCOME BY PROPRTY_AREA w.r.t EDUCATION:

- The violin plot will display the distribution of Applicant Income for each combination of Property_Area and Education. This will allow you to visually analyze how the Applicant Income varies across different property areas and educational backgrounds.
The x-axis represents the Property_Area , y-axis represents the Applicant Income.
The "hue" parameter is used to differentiate the Education levels, resulting in separate violin plots for each education group.
The shape of the violin plot represents the density distribution of the Applicant Income for each group.
This plot helps identify patterns and relationships between Applicant Income, Property_Area, and Education.

PLOT OF LOAN AMOUNT BY GENDER w.r.t MARRIED:



The x-axis represents the Gender categories (e.g., Male, Female)and the y-axis represents the Loan Amount.
The "hue" parameter is used to differentiate the Married status, resulting in separate violin plots for each Married/Not Married group.
The shape of the violin plot represents the density distribution of the Loan Amount for each group.
This plot helps identify patterns and relationships between Loan Amount, Gender, and Marital Status.

MULTI - VARIATE ANALYSIS:

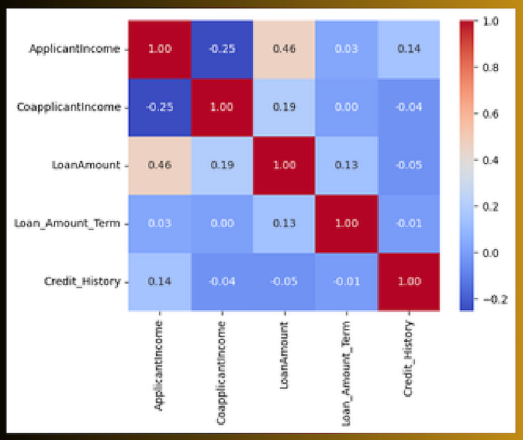
HEATMAP USING CORRELATION:

```
In [114]: df.corr(numeric_only=True)
```

Out[114]:

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
ApplicantIncome	1.000000	-0.253653	0.458722	0.029694	0.143344
CoapplicantIncome	-0.253653	1.000000	0.194046	0.003037	-0.039303
LoanAmount	0.458722	0.194046	1.000000	0.133376	-0.045232
Loan_Amount_Term	0.029694	0.003037	0.133376	1.000000	-0.011347
Credit_History	0.143344	-0.039303	-0.045232	-0.011347	1.000000

```
In [115]: sns.heatmap(df.corr(numeric_only=True),annot=True,cmap='coolwarm',fmt= '.2f')
```



Correlation Matrix Insights

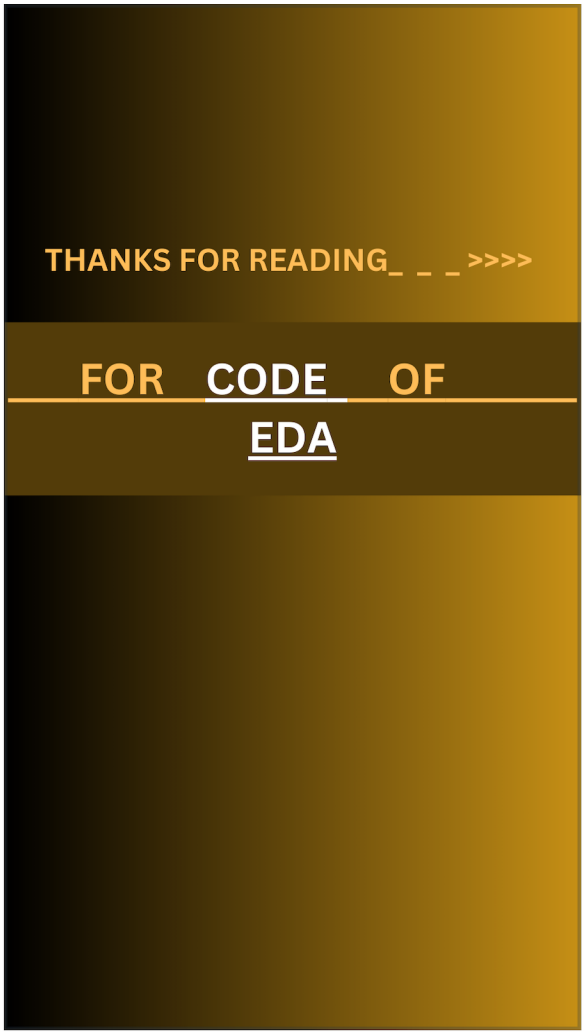
Relationships between Variables
ApplicantIncome and CoapplicantIncome: Weak negative correlation (-0.254).
LoanAmount and ApplicantIncome: Moderate positive correlation (0.486).
LoanAmount and CoapplicantIncome: Weak positive correlation (0.171).
LoanAmount and Loan_Amount_Term: Weak positive correlation (0.092).
Credit_History and ApplicantIncome: Weak positive correlation 43). 3).
This correlation heatmap provides a visual representation of the relationships between the variables in the dataset. The color intensity and the numerical values in the cells indicate the strength and direction of the correlations. This can be a useful tool for identifying patterns and relationships within the data.

CONCLUSION:

Our exploration of the dataset provided valuable insights into loan approval factors. We loaded the dataset into a Python environment, examined its structure, and handled missing values appropriately. Basic statistics for numeric columns were summarized.

In terms of data visualization, we delved into univariate, bivariate, and multivariate analyses. Histograms, box plots, bar charts, and pie charts revealed distribution patterns, relationships between variables, and correlations.

As we move forward, consider focusing on specific features impacting loan approval rates.


THANKS FOR READING_ _ _ >>>>
FOR CODE OF
EDA