Name: Alok Yadav Div: D15C Roll No.: 59

EXPERIMENT 2

Aim: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

1. Create bar graph, contingency table using any 2 features.
2. Plot Scatter plot, box plot, Heatmap using seaborn.
3. Create histogram and normalized Histogram.
4. Describe what this graph and table indicates.
5. Handle outlier using box plot and Inter quartile range.

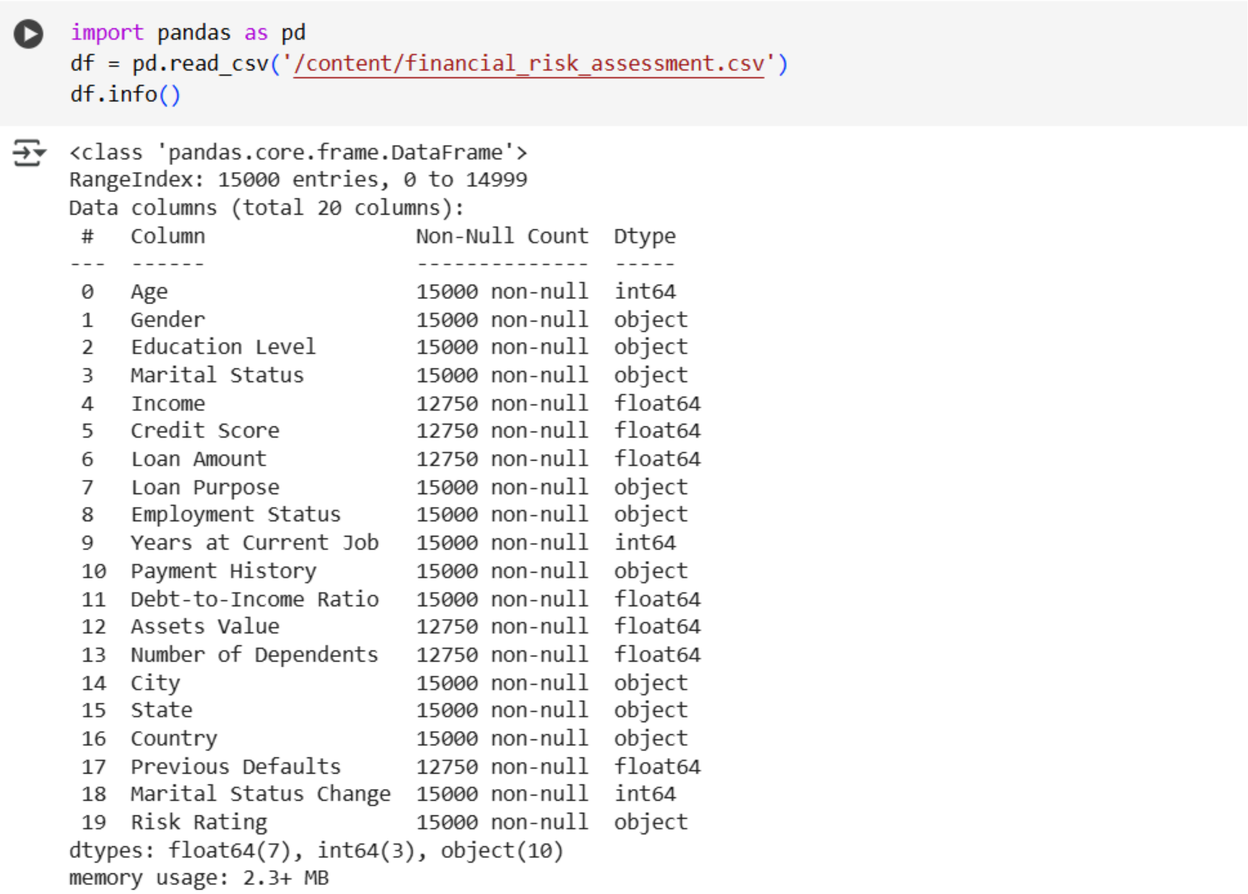
Steps:

Dataset: Financial Risk Assessment

Link: https://www.kaggle.com/datasets/preethamgouda/financial-risk

1) Loading the dataset\

Loading the dataset as pandas DataFrame and using df.info() to get information about the features and attributes of the dataset.



```
import pandas as pd
df = pd.read_csv('/content/financial_risk_assessment.csv')
df.info()
```

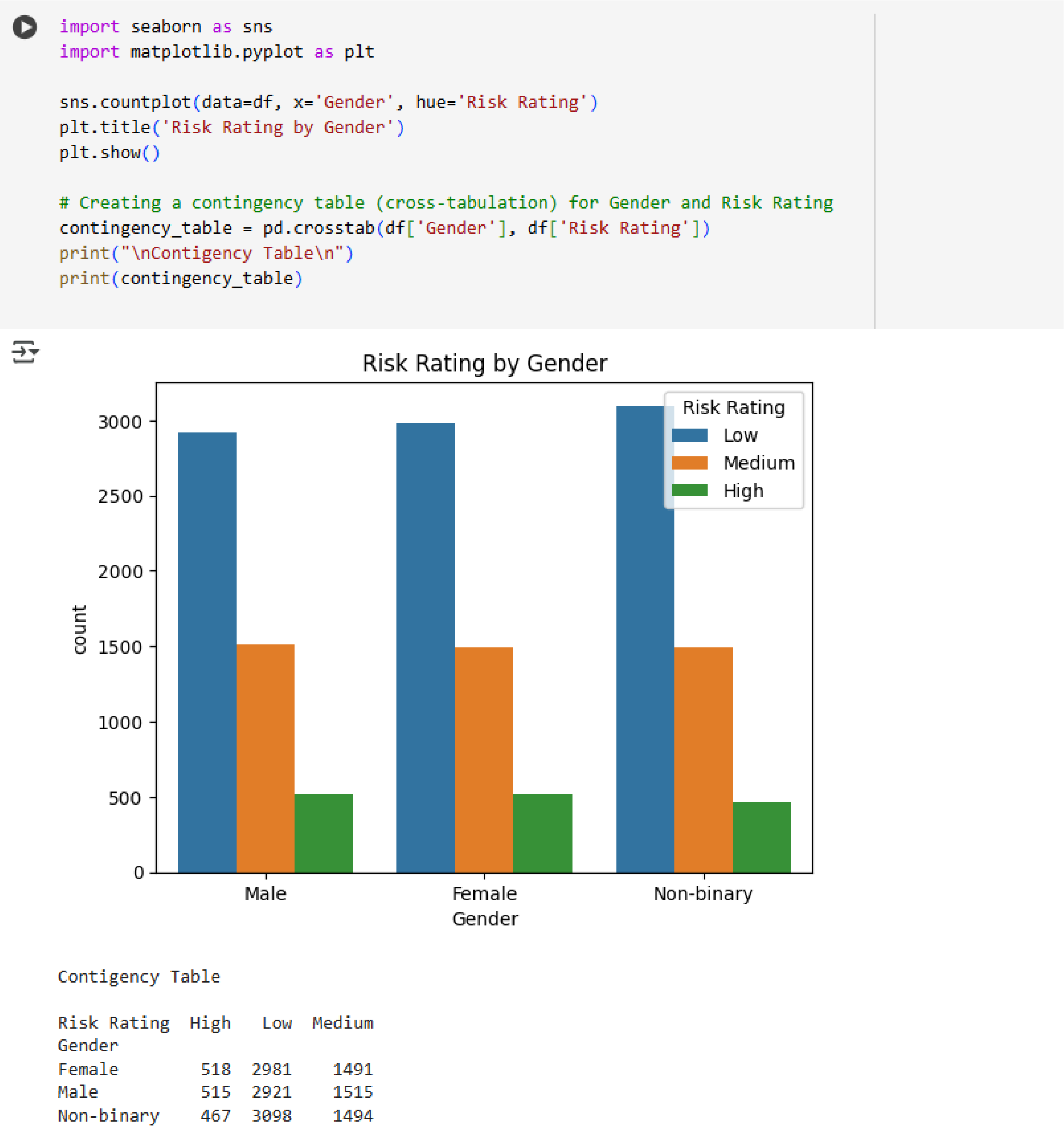
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15000 entries, 0 to 14999
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Age                    15000 non-null  int64  
 1   Gender                 15000 non-null  object 
 2   Education Level        15000 non-null  object 
 3   Marital Status         15000 non-null  object 
 4   Income                 12750 non-null  float64
 5   Credit Score           12750 non-null  float64
 6   Loan Amount            12750 non-null  float64
 7   Loan Purpose           15000 non-null  object 
 8   Employment Status      15000 non-null  object 
 9   Years at Current Job   15000 non-null  int64  
 10  Payment History        15000 non-null  object 
 11  Debt-to-Income Ratio   15000 non-null  float64
 12  Assets Value           12750 non-null  float64
 13  Number of Dependents   12750 non-null  float64
 14  City                   15000 non-null  object 
 15  State                  15000 non-null  object 
 16  Country                15000 non-null  object 
 17  Previous Defaults      12750 non-null  float64
 18  Marital Status Change  15000 non-null  int64  
 19  Risk Rating            15000 non-null  object 
dtypes: float64(7), int64(3), object(10)
memory usage: 2.3+ MB
```

2) Creating bar graph and Contingency table

Importing seaborn library which is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

Plotting the bar graph using the columns - Gender and Risk Rating.

Creating contingency table using the following command:
The pd.crosstab() function in Pandas is used to create a cross-tabulation (contingency table) of two or more categorical variables. It helps analyze relationships between variables by showing the frequency of their occurrences.



```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(data=df, x='Gender', hue='Risk Rating')
plt.title('Risk Rating by Gender')
plt.show()

# Creating a contingency table (cross-tabulation) for Gender and Risk Rating
contingency_table = pd.crosstab(df['Gender'], df['Risk Rating'])
print("\nContigency Table\n")
print(contingency_table)
```

```
Contigency Table

Risk Rating  High   Low  Medium
Gender
Female        518  2981    1491
Male          515  2921    1515
Non-binary    467  3098    1494
```

The contingency table shows that most individuals fall into the low-risk category, with non-binary individuals having the highest count (3,098), followed by females (2,981) and males (2,921). Medium risk is fairly balanced across genders, while high risk is the least common, slightly higher in females (518) and males (515) than non-binary individuals (467). Overall, financial risk perception appears similar across genders, with a predominant trend toward low risk.

3) Plot Scatter plot, box plot, Heatmap using seaborn.

Scatter plot:

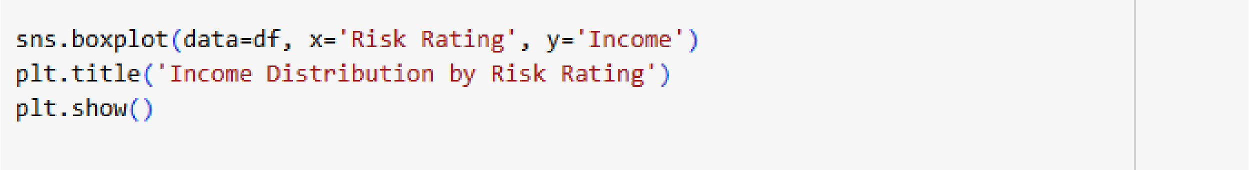
A scatter plot is a graphical representation used to visualize the relationship between two numerical variables. Each point on the plot represents an observation. Scatter plots help identify patterns such as positive or negative correlations, clusters, and outliers. If the points form an upward trend, it suggests a positive correlation, whereas a downward trend indicates a negative correlation.

Box plot:

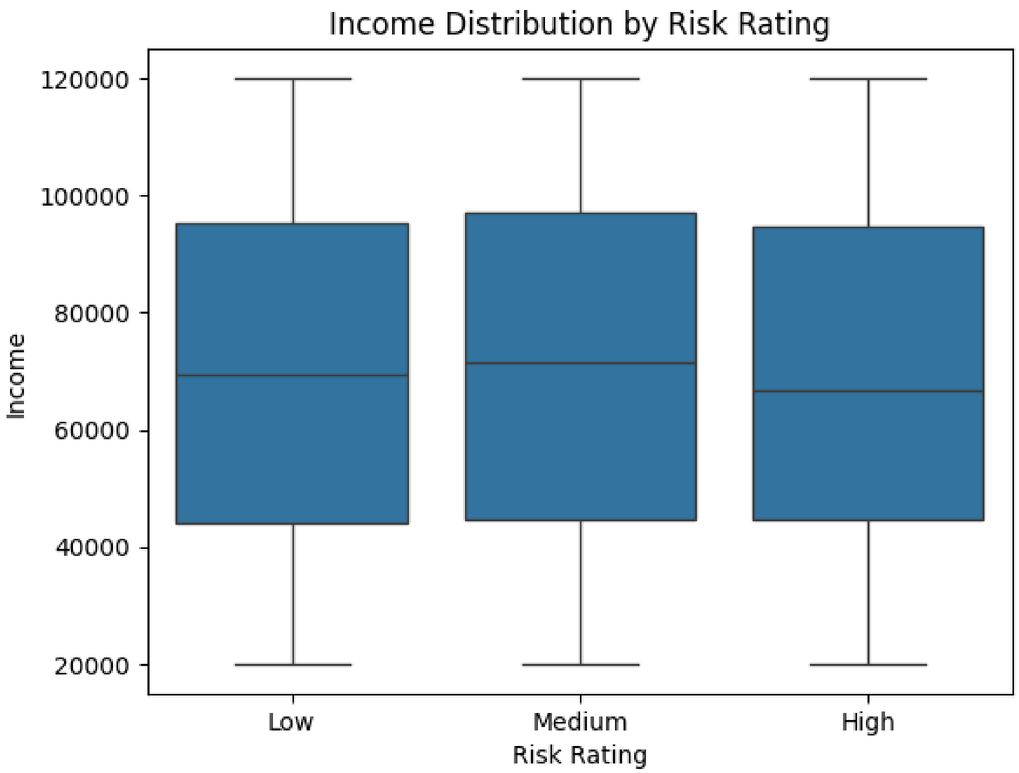
A **box plot** (also known as a box-and-whisker plot) is a statistical visualization that summarizes the distribution of a dataset using five key measures: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box represents the interquartile range (IQR), with the median marked inside it, while "whiskers" extend to show variability outside the quartiles.

Heatmap:

A **heatmap** is a data visualization technique that represents values in a matrix format using color intensity. It is often used to display correlations between variables, frequency distributions, or hierarchical clustering results. Darker or more intense colors indicate higher values, while lighter colors represent lower values.

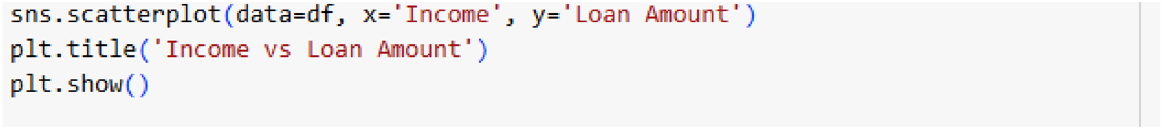


```
sns.boxplot(data=df, x='Risk Rating', y='Income')
plt.title('Income Distribution by Risk Rating')
plt.show()
```

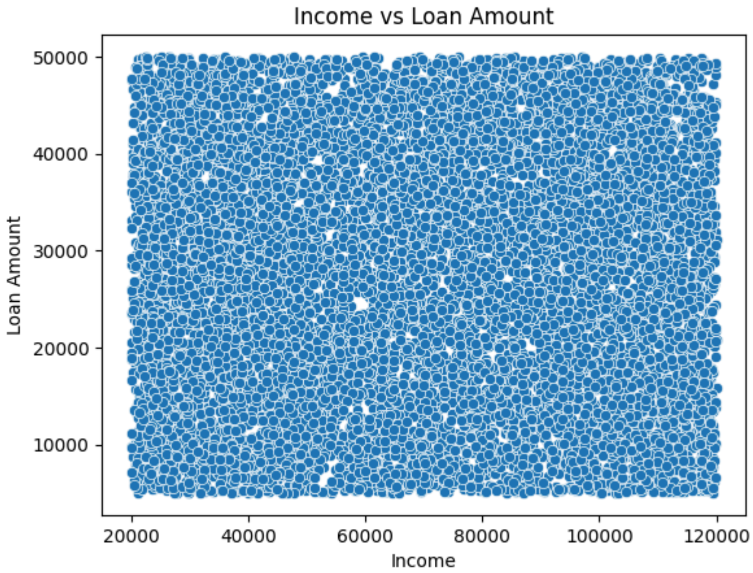


The box plot illustrates the distribution of income across different financial risk ratings (Low, Medium, High). The median income is similar across all risk categories, suggesting that income levels do not strongly differentiate financial risk ratings. The interquartile ranges (IQR) and overall spread of incomes are also comparable, indicating a consistent income distribution across risk groups. Since there are no significant outliers or deviations, it implies that factors other than income may play a key role in determining financial risk ratings.

Scatter Plot

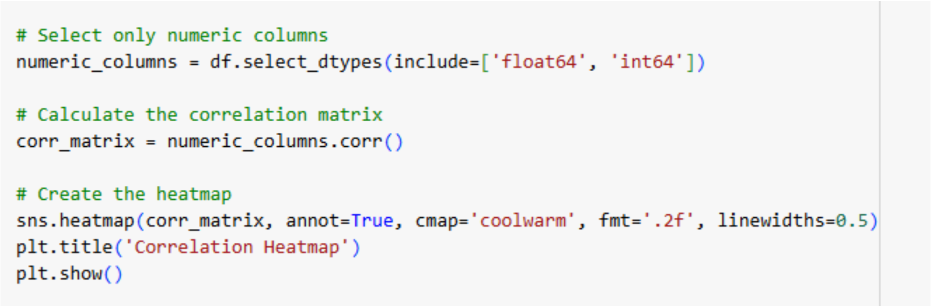


```
sns.scatterplot(data=df, x='Income', y='Loan Amount')
plt.title('Income vs Loan Amount')
plt.show()
```

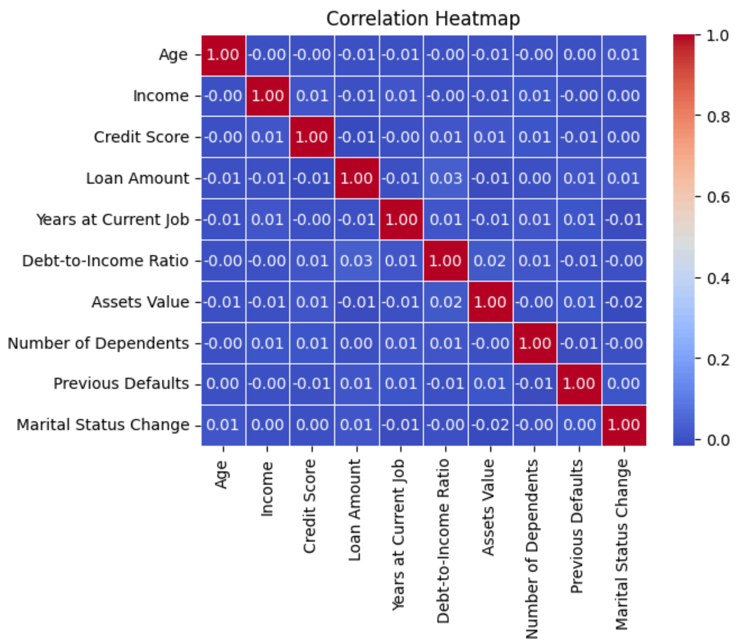


The scatter plot illustrates the relationship between income and loan amount. The data appears densely packed and evenly distributed across the entire range, suggesting no clear correlation between income and loan amount. The loan amounts vary widely regardless of income levels, indicating that factors other than income may play a significant role in determining loan amounts. The lack of an apparent trend suggests that loan approvals or allocations are influenced by additional factors beyond just income.

Heatmap



```
# Select only numeric columns
numeric_columns = df.select_dtypes(include=['float64', 'int64'])

# Calculate the correlation matrix
corr_matrix = numeric_columns.corr()

# Create the heatmap
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



The correlation heatmap shows the relationships between various financial factors. Most correlations are close to zero, indicating weak or no linear relationships between the variables. Strong correlations (value of 1.00) appear only along the diagonal, which represents self-correlation. Overall, no significant relationships exist among the variables, implying that financial risk assessment may depend on a combination of multiple independent factors rather than direct correlations.

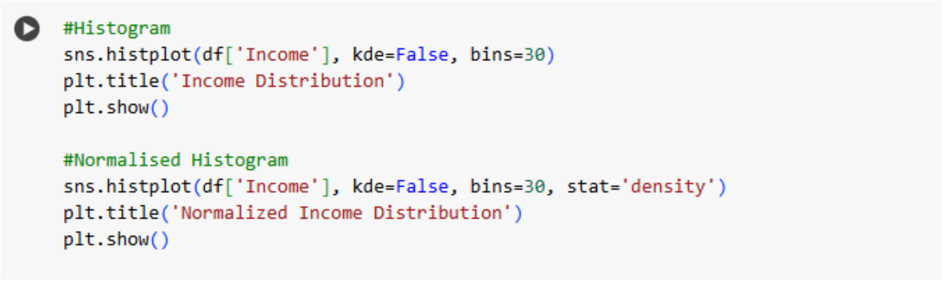
4) Create histogram and normalized Histogram

Histogram:

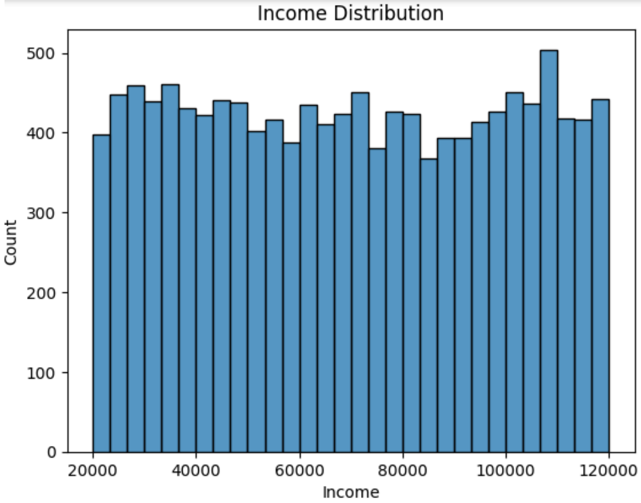
Histograms help visualize the shape, spread, and central tendency of data, making it easier to identify patterns such as skewness, modality (unimodal, bimodal), and outliers.

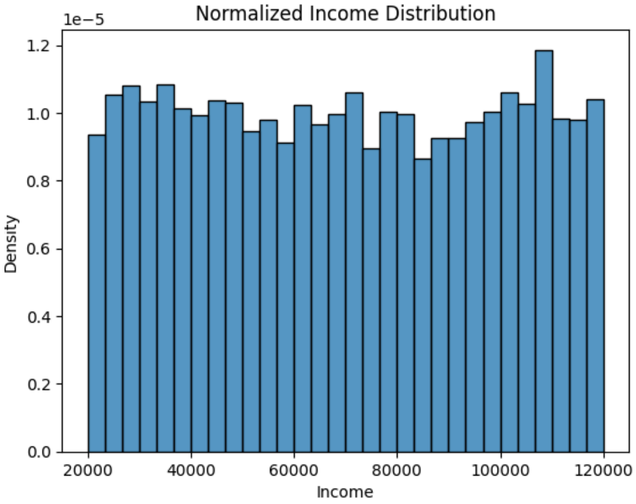
Normalized Histogram:

A normalized histogram is a variation of a histogram where the frequencies are scaled to represent relative frequencies instead of raw counts. This is done by dividing each bin's frequency by the total number of observations or by ensuring the total area under the histogram sums to 1.



```
#Histogram
sns.histplot(df['Income'], kde=False, bins=30)
plt.title('Income Distribution')
plt.show()

#Normalised Histogram
sns.histplot(df['Income'], kde=False, bins=30, stat='density')
plt.title('Normalized Income Distribution')
plt.show()
```





The income distribution histograms show a relatively uniform spread of income levels, with no strong skewness or extreme outliers. The first plot represents the raw count of individuals across income ranges, while the second normalizes the distribution to show density. Both indicate a fairly even income distribution, with slight fluctuations but no significant peaks or gaps. This suggests that income levels are well-distributed across the dataset, without any dominant income group.

5) Describe what this graph and table indicates.

Bar Graph (Risk Rating by Gender):

This bar graph will show the distribution of different risk ratings across the genders, helping us understand if there's a significant difference in risk ratings between male and female applicants.

Contingency Table:

The contingency table will show the exact count of observations for each combination of Gender and Risk Rating. This allows you to see how many male and female applicants fall into each risk rating category.

Scatter Plot (Income vs Loan Amount):
The scatter plot will reveal the relationship between income and loan amount. This could show if higher incomes tend to have higher loan amounts, or if there's any clustering or pattern.

Box Plot (Income by Risk Rating):
The box plot will give an idea of how income is distributed across different risk ratings. For example, if risk ratings are low for high-income individuals or vice versa, it will be evident.

Heatmap (Correlation Matrix):
The heatmap will help you quickly identify correlations between numeric features. For example, you might notice a strong correlation between Income and Loan Amount, or low correlation between Debt-to-Income Ratio and Number of Dependents.

6) Handle outlier using box plot and Inter quartile range

Outliers can be handled using the Box Plot and Interquartile Range (IQR) method. A box plot visually identifies outliers as points beyond the "whiskers," which represent data within 1.5 times the IQR.

Steps to Handle Outliers:

1. Calculate the IQR

IQR = Q3 - Q1

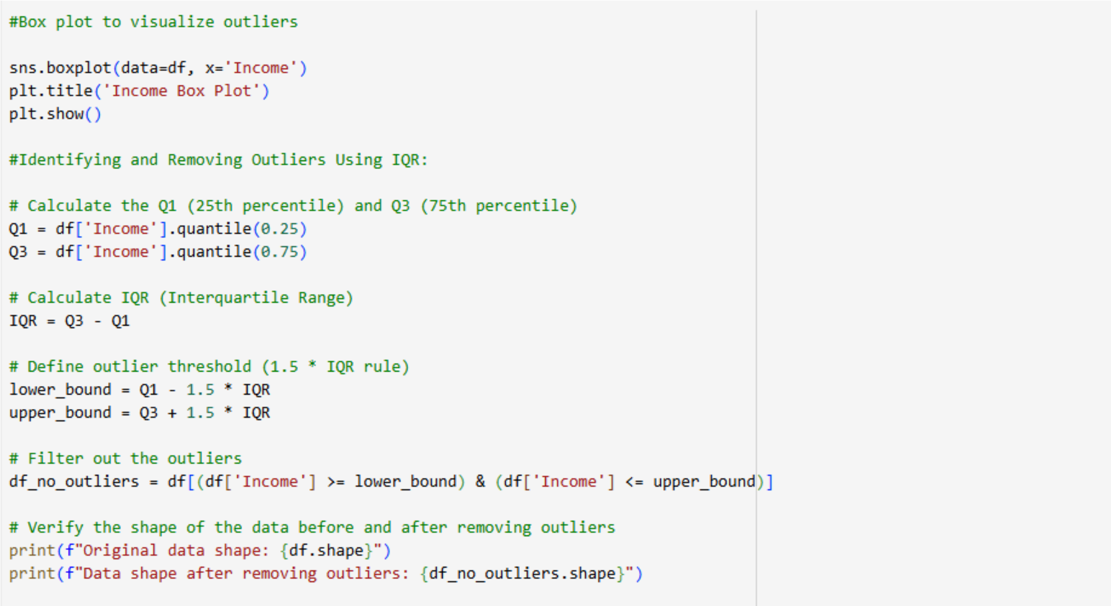
where Q1 (25th percentile and Q3 (75th percentile) are the lower and upper quartiles.

2. Determine the Outlier Thresholds:

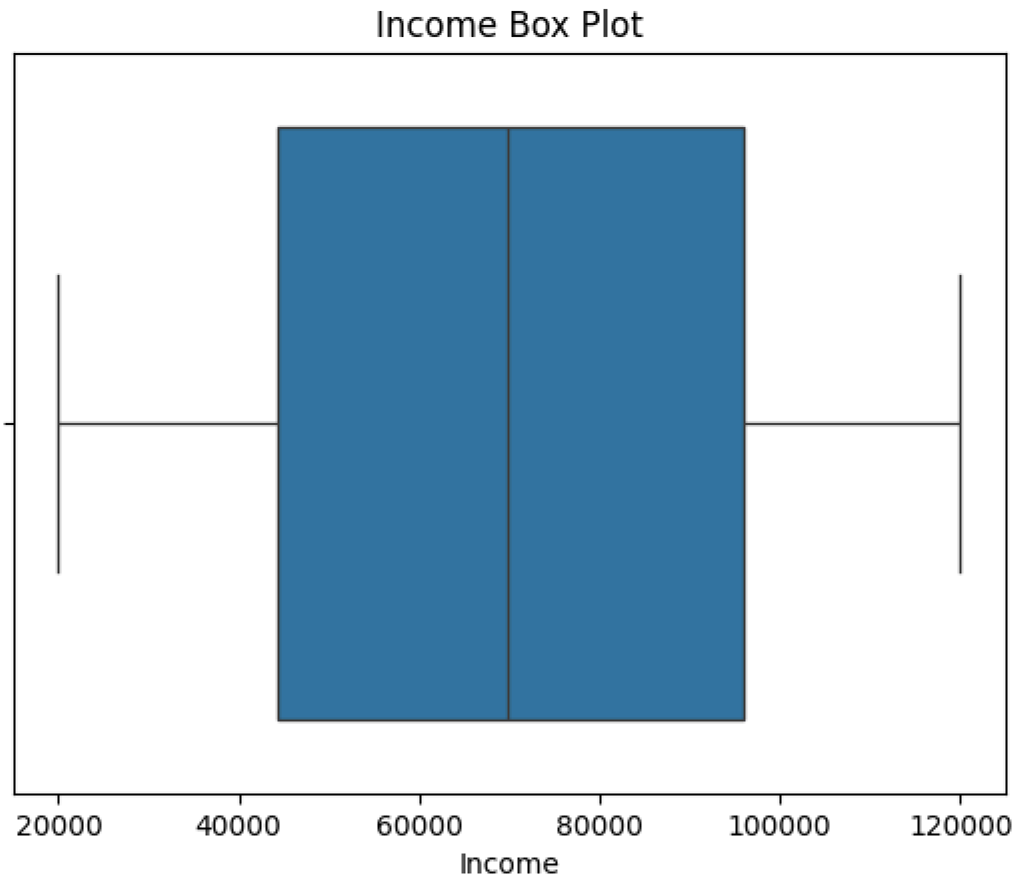
- Lower Bound = Q1 - 1.5 × IQR
- Upper Bound = Q3 + 1.5 × IQR

3. Remove or Adjust Outliers:

- Drop outliers (if errors or extreme values).
- Cap them at the lower or upper bound (winsorization).
- Transform the data (log, square root) to reduce the impact.



```
#Box plot to visualize outliers

sns.boxplot(data=df, x='Income')
plt.title('Income Box Plot')
plt.show()

#Identifying and Removing Outliers Using IQR:

# Calculate the Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df['Income'].quantile(0.25)
Q3 = df['Income'].quantile(0.75)

# Calculate IQR (Interquartile Range)
IQR = Q3 - Q1

# Define outlier threshold (1.5 * IQR rule)
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter out the outliers
df_no_outliers = df[(df['Income'] >= lower_bound) & (df['Income'] <= upper_bound)]

# Verify the shape of the data before and after removing outliers
print(f"Original data shape: {df.shape}")
print(f"Data shape after removing outliers: {df_no_outliers.shape}")
```



```
Original data shape: (15000, 20)
Data shape after removing outliers: (12750, 20)
```

The box plot represents the distribution of income after removing outliers. The interquartile range (IQR) spans from around 30,000 to 100,000, with the median income close to 70,000. The whiskers extend from approximately 20,000 to 120,000, indicating the range of non-outlier values. The removal of outliers reduced the dataset size from 15,000 to 12,750, suggesting that around 2,250 data points were considered extreme values. The cleaned data now provides a more concentrated view of the central distribution, reducing the influence of extreme income variations.

Conclusion:

In this experiment, the analysis reveals that financial risk perception is largely similar across genders, with most individuals falling into the low-risk category. Box plots indicate that income levels do not significantly influence financial risk ratings, as distributions remain consistent across risk groups. The scatter plot shows no clear correlation between income and loan amount, suggesting that loan allocation depends on other factors. The correlation heatmap confirms weak or no linear relationships among financial variables, implying that financial risk assessment is influenced by multiple independent factors. Histograms demonstrate a relatively uniform income distribution with no dominant income group. After outlier removal, the dataset size reduced from 15,000 to 12,750, offering a more refined view of income distribution and reducing the impact of extreme values.