Credit Behaviour Analysis and Risk Assessment

Overview: This case study analyzes client data to improve loan approval processes, risk assessment, and customer engagement strategies in the financial industry. By examining credit behavior, loan applications, and client segmentation, it aims to uncover insights that enhance decision-making, optimize loan approvals, and manage risks effectively. Key activities include data cleaning, feature engineering, and profiling clients based on credit activity to support strategic decision-making.

The case study is structured into multiple stages, each designed to address specific questions relevant to its context, ensuring a logical flow in data exploration. The given expected output format serves as a reference to guide you in achieving similar results; however, actual numbers may vary based on the methodology used. The expected output format may not always contain all rows, and you are encouraged to explore alternative methods where applicable to enhance the analysis. Additionally, validating all results is crucial to ensure consistency and accuracy.

Datasets: Link

The following are the stages in which the case study questions are aligned:

1. Understanding the Data

Before diving into analysis, it is essential to explore the data. This phase focuses on:

- **Client Coverage & Uniqueness Checks** - Ensure every client exists across all datasets and appears only once per file to maintain consistency and avoid duplication issues.
- **Data Quality & Completeness Validation** - Identify missing values in critical columns and verify that all categorical fields (like Applied, Approved) contain expected values.
- **Logical & Statistical Consistency Checks** - Confirm that tradeline and enquiry counts increase logically over time, detect outliers in credit behavior using IQR, and ensure that approved loans do not exceed applications.

Understanding the data ensures accuracy, consistency, and reliability in analysis by identifying gaps, anomalies, and logical inconsistencies early in the case study.

2. Data Cleaning

Once the data is explored, the next step is to handle the inconsistencies and outliers. This phase focuses on:

- **Fix Missing & Invalid Values:** Handled nulls in key fields like Applied, Approved, Total_Tradelines_6m_In, and Total_Enquiry_6m_In by using logic-based imputation or removal, and corrected invalid (e.g., negative) entries for financial consistency.
- **Ensure Type & Logical Consistency:** Converted numeric fields to proper formats and aligned data relationships (e.g., ensuring 6-month values ≤ 1-year values) for coherent credit history tracking.
- **Manage Outliers via Winsorization:** Applied the 99th percentile cap on outlier-heavy columns to reduce skew and stabilize future analysis.

- **Final Integrity Check:** Verified cleaned data for completeness, consistency, and ensured minimal client data loss post-cleaning.

Data cleaning is a crucial step to ensure reliability in downstream analysis and modeling.

3. Data Analysis - Univariate & Bivariate:

Understanding distribution of different features and the relationships between variables is key to identifying business trends and patterns. This stage examines:

- **Analyze distributions of credit inquiries, tradelines, and loan applications -** Analyze the distribution of variables such as Total_Tradelines_6m_In, Total_Enquiry_6m_In, Total_Tradelines_1y_In, and Total_Enquiry_1y_In to understand the patterns, outliers, and central tendencies.
- **Identify spending and loan approval trends -** Examine Max_Trade_3m_Out and loan approval data (Applied, Approved) to understand spending behavior and approval rates, identifying any skewness or trends.
- **Investigate relationships between credit behavior and loan application success -** Explore correlations between credit activity (tradelines, inquiries) and loan application success (approved/rejected), including examining spending behavior (Max_Trade_3m_Out) and its influence on loan applications.

4. Feature Engineering & Advanced Analysis:

To enhance analytical capabilities, new features are created based on existing data. These include:

- **Credit Behavior & Utilization Patterns -** This focuses on how clients are using their credit - analyzing credit utilization rates, recent enquiry patterns, and overall credit activity to understand financial behavior.
- **Loan Applications & Approval Trends -** Here, we assess how frequently clients apply for loans, their approval rates, and how effective or selective the loan approval process is across different segments.
- **Risk Profiling & Client Segmentation -** Clients are categorized into low, medium, and high-risk groups based on credit activity, spending habits, and enquiry levels to enable more strategic risk management.
- **Client Engagement & Conversion Analysis -** This part identifies how engaged clients are with credit services, and whether their behavior (like enquiries and tradelines) translates into successful loan approvals.

Importance of this Project

By leveraging structured data analysis and risk assessment techniques, this case study will help improve loan decision accuracy, minimize financial risks, and enhance customer profiling. The findings will support data-driven decision-making in lending strategies, fraud detection, and customer relationship management. This case study enables:

- **Improved Data Accuracy:** Identify and rectify inconsistencies, missing values, and errors in credit records.
- **Enhanced Risk Assessment:** Identify high-risk clients and credit-seeking patterns for better decision-making.
- **Strategic Customer Segmentation:** Categorize clients based on spending behavior, credit activity, and loan application frequency.
- **Actionable Insights for Lending Decisions:** Understand approval trends, spending habits, and credit behavior to optimize loan offerings and reduce financial risk.

Project Questions

Stage 1. Understanding the data

Q1: Validate Client Coverage Across All Datasets.

Before performing any analysis, it is crucial to check whether the same clients exist across all three datasets: tradeline_data, enquiry_data, and base_file. If clients are missing from any dataset, it could impact future analysis.

Steps to follow:

- Count the total number of unique Client_IDs in each dataset.
- Identify clients present in tradeline_data but missing in base_file or enquiry_data.
- Identify clients present in enquiry_data but missing in base_file.
- Calculate the percentage of clients missing in at least one of the three datasets.

*(**Hint:** Calculates the percentage of missing clients in at least one of the three datasets out of the total unique clients across all datasets. Finally, round off the percentage up to 2 decimal places.)*

Expected Output Format:

```
Total Unique Clients in tradeline_data: 25000
Total Unique Clients in enquiry_data: 24500
Total Unique Clients in base_file: 24800
Clients missing in at least one dataset: 696
Percentage of inconsistent clients: 2.78%
```

Q2: Investigate Duplicate Client Entries.

Now that we have checked for missing clients, the next step is to ensure that no client appears multiple times within a single dataset. If a client appears more than once, it could indicate either data duplication or an issue with how tradelines and enquiries are being recorded.

Steps to follow:

- Identify Client_IDs that appear more than once in each dataset.

- Check whether duplicate rows have different values in the 'Total_Tradelines_6m_In' (tradelines), 'Total_Enquiry_6m_In' (enquiries), and 'Applied' (base file) columns.
- Compute the total number of duplicate clients and analyze how many have different values.

Expected Output Format:

```
Dataset: tradeline_data, Duplicate Clients: 0, Cases with Different Values: 0
Dataset: enquiry_data, Duplicate Clients: 0, Cases with Different Values: 0
Dataset: base_file, Duplicate Clients: 0, Cases with Different Values: 0
```

Q3: Identify Clients with Missing Key Data.

Having confirmed the uniqueness of clients, the next step is to check for missing values. If critical data points are missing, it may impact downstream analysis and model performance.

Steps to follow:

- Check for missing values in essential columns (Total_Tradelines_6m_In, Total_Enquiry_6m_In, Applied, Approved).
- Compute the percentage of missing critical data for each dataset and display it by rounding off the values up to 2 decimal places.

Expected Output Format:

```
Missing critical data in tradeline_data: 0.80%
Missing critical data in enquiry_data: 1.50%
Missing critical data in base_file: 0.20%
```

Q4: Validate Tradeline Data Consistency Over Time.

After checking for missing data, we need to ensure that tradelines follow a logical pattern. The total number of tradelines over a shorter period (e.g., 6 months) should never exceed the number over a longer period (e.g., 1 year or 2 years). If inconsistencies exist, it could indicate a data entry issue.

Steps to follow:

- For each client, check whether Total_Tradelines_6m_In is always less than or equal to Total_Tradelines_1y_In, and Total_Tradelines_1y_In is less than or equal to Total_Tradelines_2y_In.
- Identify how many clients have decreasing tradeline values over time and what percentage does it contribute (round off the % up to 2 decimal places).

Expected Output Format:

```
Clients with Decreasing Tradelines: 0
Percentage of Inconsistent Clients: 0.00%
```

Q5: Validate Enquiry Data Consistency.

Similar to tradelines, enquiries should also be consistent over time. If Total_Enquiry_6m_In exceeds Total_Enquiry_1y_In, it suggests an issue with how data has been recorded.

Steps to follow:

- Ensure that for each client, Total_Enquiry_6m_In ≤ Total_Enquiry_1y_In and Total_Enquiry_1y_In ≤ Total_Enquiry_2y_In.
- Identify clients who violate this rule and what percentage does it contribute (round off the % up to 2 decimal places).

Expected Output Format:

```
Clients with Inconsistent Enquiries: 0
Percentage of Inconsistent Clients: 0.00%
```

Q6: Detect Outliers in Tradeline Counts.

Outliers in tradeline counts may indicate unusual credit behavior, such as an individual acquiring an excessive number of tradelines in a short period. To detect such cases, we will calculate the interquartile range (IQR) and find clients who exceed the upper bound.

Steps to follow:

- Compute Q1 (25th percentile) and Q3 (75th percentile) for Total_Tradelines_6m_In.
- Determine the upper bound using Q3 + 1.5 * IQR.
- Identify clients exceeding this threshold.
- Display Q1, Q2, IQR, upper bound and the number of outliers.

Expected Output Format:

```
Q1: 5.0, Q3: 14.0, IQR: 9.0, Upper Bound: 27.5
Clients exceeding threshold: 0
```

Q7: Ensure Loan Approval Does Not Exceed Applications.

An approved loan count should never be greater than the number of applications made. If such cases exist, it suggests an error in data entry or processing.

Steps to follow:

- Identify cases where Approved > Applied.
- Compute the count and percentage of such cases in the dataset (round off the % up to 2 decimal places).

Expected Output Format:

```
Clients with Approved > Applied: 0
Percentage: 0.00%
```

Q8: Check Unique Value Distribution for Key Categorical Columns.

Finally, before concluding data exploration, we need to verify that categorical columns contain expected values. Unexpected values could indicate errors in data collection or encoding.

Steps to follow:

- Extract unique values for categorical columns (Applied, Approved).
- Ensure all values fall within expected ranges.

Expected Output Format:

```
Unique Values in Applied: [ 1.  3.  0.  4.  2. nan]
Unique Values in Approved: [1 0 2 3 4]
```

Stage 2. Data Cleaning

Q1: Handle Missing Values in Critical Columns.

We know that missing values can cause serious issues during analysis, especially in critical columns like Total_Tradelines_6m_In, Total_Enquiry_6m_In, and Applied. Handling these missing values correctly is essential to ensure valid and consistent results in the later analysis phase.

Steps to follow:

- In Stage 1: Q3, we identified the missing percentage of critical columns. If the missing values in Total_Tradelines_6m_In and Total_Enquiry_6m_In are found to be less than 2% of the dataset, we fill them using the median value.
- For Applied, check if there is any missing value. If Applied is missing and Approved is available, set Applied = Approved. If both are missing, remove these records.
- Verify that no missing values remain in these 4 columns after cleaning.

Expected Output Format:

```
Missing values in 'Total_Tradelines_6m_In': 0
Missing values in 'Total_Enquiry_6m_In': 0
Missing values in 'Applied': 0
Missing values in 'Approved': 0
```

Q2: Standardize Data Types for Numeric Columns.

In some cases, numeric data might be stored incorrectly as strings (e.g., object data type in pandas). This can lead to errors when performing calculations. It is important to ensure that all numeric columns are properly formatted as integers or floats.

Steps to follow:

- Check the data types of all columns, particularly those with numeric values like Total_Tradelines_6m_In, Total_Enquiry_6m_In, Applied and Approved.
- If any numeric columns are stored as object (string) types, convert them to int64 or float64 as appropriate.
- If conversion fails, check the column for any non-numeric characters and clean the data before converting.
- Display the correct datatypes after standardization.

Expected Output Format:

```
Data type of 'Total_Tradelines_6m_In': float64
Data type of 'Total_Enquiry_6m_In': float64
Data type of 'Applied': float64
Data type of 'Approved': int64
```

Q3: Remove Records with Invalid or Out-of-Range Values.

It is important to clean records with values that do not make sense in the context of the data, such as negative values in Total_Tradelines_6m_In, Total_Enquiry_6m_In, or Max_Trade_3m_Out. Negative values are likely data entry errors and should be either corrected or removed.

Steps to follow:

- For negative values in Total_Tradelines_6m_In and Total_Enquiry_6m_In, replace them with 0.
- If there are invalid negative values in Max_Trade_3m_Out, assess the cause and remove the affected rows.
- Count all rows where Total_Tradelines_6m_In, Total_Enquiry_6m_In, or Max_Trade_3m_Out have negative values individually.

Expected Output Format:

```
Negative values in 'Total_Tradelines_6m_In': 0
Negative values in 'Total_Enquiry_6m_In': 0
Negative values in 'Max_Trade_3m_Out': 0
```

Q4: Correct Approved Values Based on Applied.

In some cases, the Approved column may be missing, especially when Applied is present. This could indicate incomplete loan application data. It's important to apply the correct logic to ensure that missing values in Approved are handled properly based on the value in Applied.

Steps to follow:

- Identify clients where Approved is missing but Applied is present.
- For those clients, if Applied = 0, set Approved = 0. If Applied > 0, mark Approved as 0.
- Verify that no Approved values remain missing for clients who have Applied > 0.

Expected Output Format:

```
Missing values in 'Approved' after cleaning: 0
```

Q5: Verify and Correct Inconsistent Tradeline and Enquiry Counts.

Tradeline and enquiry counts over time should not decrease. If a client has Total_Tradelines_6m_In > Total_Tradelines_1y_In, there is likely an error in the data entry. Such cases need to be corrected to ensure the data follows a logical pattern.

Steps to follow:

- Identify clients where Total_Tradelines_6m_In > Total_Tradelines_1y_In or Total_Enquiry_6m_In > Total_Enquiry_1y_In.
- For these cases, either set the 6m count equal to the 1y count.
- Recheck the consistency of the data after cleaning by verifying that no records remain which fall under the above conditions.

Expected Output Format:

```
Inconsistent Tradelines: 0
Inconsistent Enquiries: 0
```

Q6: Apply Winsorization to Remove Extreme Outliers.

Winsorization (Data Transformation Technique): Instead of removing the outliers completely, Winsorization caps the extreme values to a certain percentile.

Some clients may have an unusually high number of tradelines or enquiries, which can skew statistical analysis. We will apply Winsorization to cap extreme values at a reasonable threshold (e.g., the 99th percentile).

Steps to follow:

- Calculate the 99th percentile for Total_Tradelines_6m_In and Total_Enquiry_6m_In and display them.
- Cap all values exceeding this percentile to the maximum value within the 99th percentile.
- Ensure that no value exceeds the capped threshold after cleaning.

Expected Output Format:

```
99th Percentile for Total_Tradelines_6m_In: 19.0
99th Percentile for Total_Enquiry_6m_In: 14.0
```

Q7: Conduct Final Integrity Check.

Once all data cleaning steps are completed, perform a final integrity check to ensure the dataset is consistent and ready for analysis.

Steps to follow:

- Check that no missing values remain in critical columns (Applied, Approved, Total_Tradelines_6m_In, Total_Enquiry_6m_In).
- Ensure there are no rows with invalid or inconsistent data (e.g., negative values in numeric fields) in the critical columns: 'Total_Tradelines_6m_In', 'Max_Trade_3m_Out' and 'Total_Enquiry_6m_In'.
- Verify that the number of clients has not decreased significantly unless removed for data issues. This can be verified by validating the total number of remaining rows in all the datasets.

Expected Output Format:

```
Missing values in critical columns after cleaning: 0
Negative values in numeric columns after cleaning: 0
Total number of records after cleaning: 74300
```

Stage 3. Data Analysis - Univariate Questions

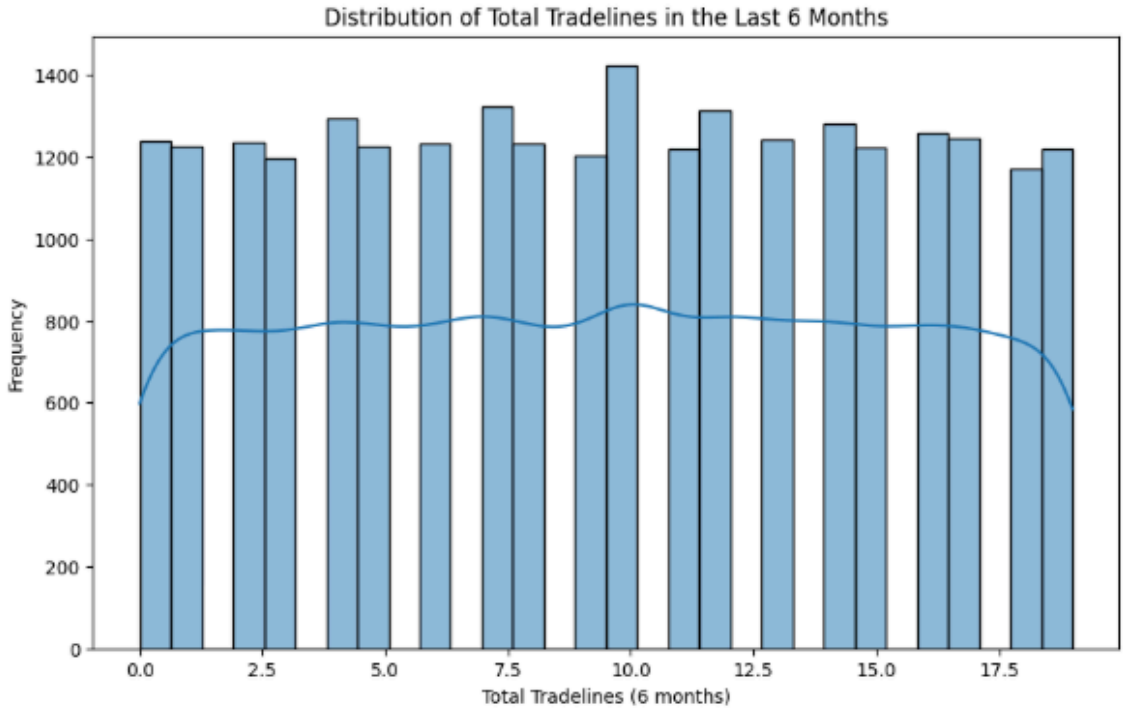
Q1: Analyze the Distribution of Total_Tradelines_6m_In.

We need to understand the distribution of Total_Tradelines_6m_In to see if it follows a normal distribution, is skewed, or has any outliers. This analysis will help us understand the central tendency and variability of the data.

Steps to follow:

- Calculate the mean, median and standard deviation values for Total_Tradelines_6m_In.
- Visualize the distribution of Total_Tradelines_6m_In using a histogram (bin size = 30).
- Check if the data is normally distributed (e.g., by comparing the mean and median, and looking for skewness).
- Identify any outliers or extreme values using the IQR (Interquartile Range) method.
- Display the following: mean, median, std deviation and the number of outliers.

Expected Output Format:




```
Mean: 9.48996, Median: 10.0, Std Dev: 5.720602084564762
Outliers detected: 0
```

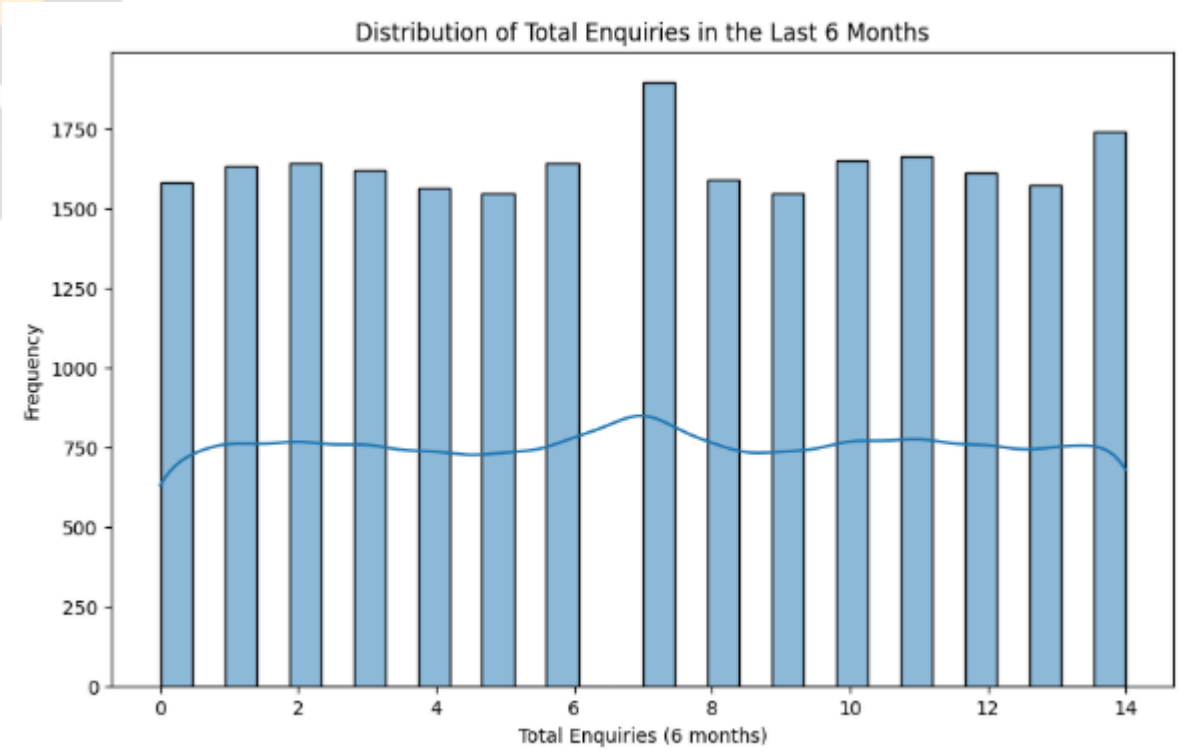
Q2: Analyze the Distribution of Total_Enquiry_6m_In.

Similarly, it is important to analyze how Total_Enquiry_6m_In is distributed to identify if there are any patterns, skewness, or outliers.

Steps to follow:

- Calculate summary statistics (mean, median, standard deviation) for Total_Enquiry_6m_In.
- Plot the histogram of Total_Enquiry_6m_In to visualize its shape and distribution.
- Compare the mean and median to identify skewness in the data.
- Check for any outliers using IQR and finally display the key metrics: mean, median, std deviation and number of outliers.

Expected Output Format:




```
Mean: 7.041469387755102, Median: 7.0, Std Dev: 4.3161514237643885
Outliers detected: 0
```

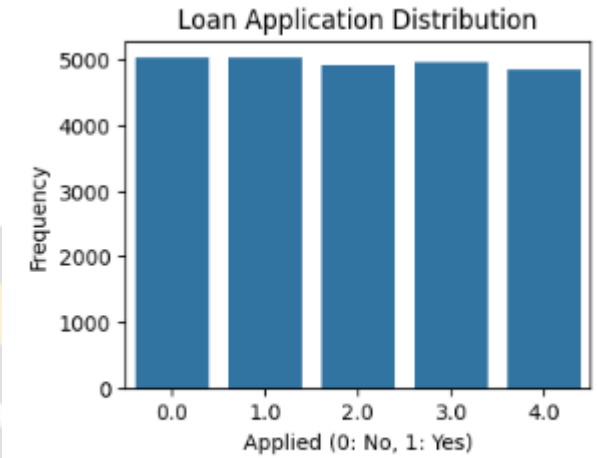
Q3: Distribution of Loan Applications (Applied).

Understanding how many clients apply for loans can help us identify any potential patterns or biases in the application process.

Steps to follow:

- Calculate the percentage of clients who have applied for loans (Applied > 0) versus those who haven't (Applied = 0).
- Display the percentage by rounding off the values up to 2 decimal places.
- Plot the distribution (count plot) of the Applied column, highlighting the proportion of applications versus non-applications.
- Insight: Identify if there is any concentration of loan applications among a specific group (e.g., many clients apply only once).

Expected Output Format:




```
Percentage of clients who applied: 79.70%
Percentage of clients who did not apply: 20.30%
```

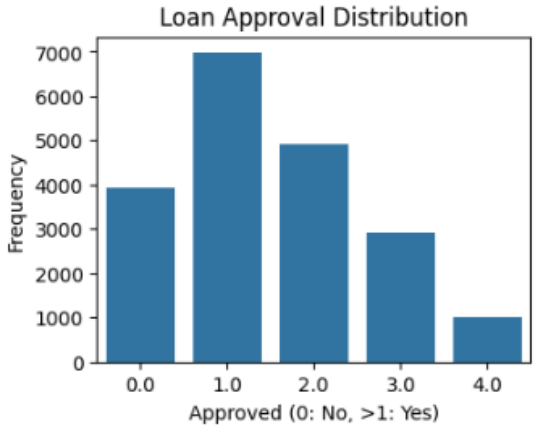
Q4: Analyze the Loan Approval Rate.

The Approved column represents the number of clients who got their loans approved. Analyzing this column will help assess approval patterns.

Steps to follow:

- Calculate the approval rate (Approved / Applied) if applied > 0 (round up to 2 decimal places).
- Extract the data where applied > 0 to plot the distribution.
- Plot the distribution (count plot) of the Approved column to understand how many clients are actually getting their loans approved.

Expected Output Format:




```
Loan Approval Rate: 0.60
```

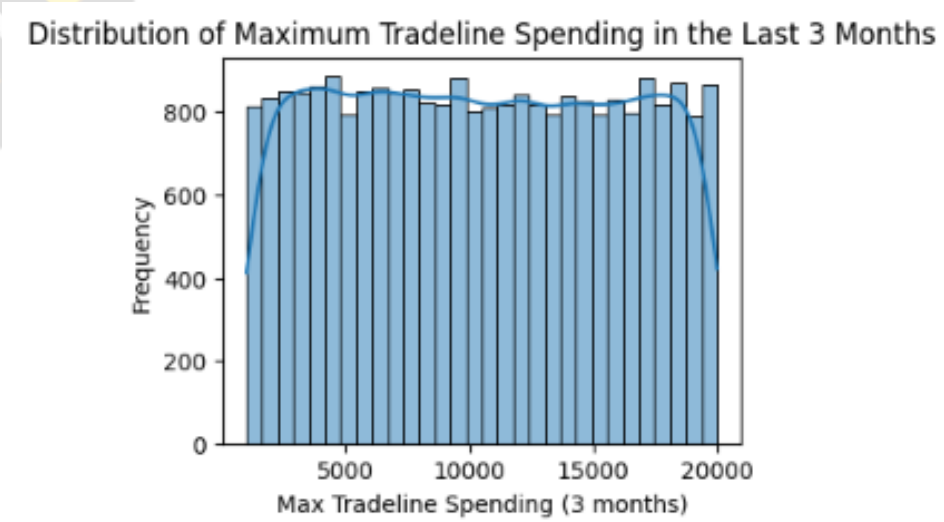
Q5: Examine the Distribution of Max_Trade_3m_Out.

The Max_Trade_3m_Out column represents the maximum tradeline amount for a client over the past 3 months. It is crucial to understand how these values are spread out to detect any unusual spending or extreme behavior.

Steps to follow:

- Calculate summary statistics (mean, median and standard deviation) for Max_Trade_3m_Out.
- Plot a histogram of Max_Trade_3m_Out to visualize its distribution and detect any skewness or outliers.
- Check for extreme values or potential outliers using IQR.
- Finally, display the mean, median, std deviation and the number of outliers.

Expected Output Format:




```
Mean: 10471.49164, Median: 10429.0, Std Dev: 5498.743149376866
Outliers detected: 0
```

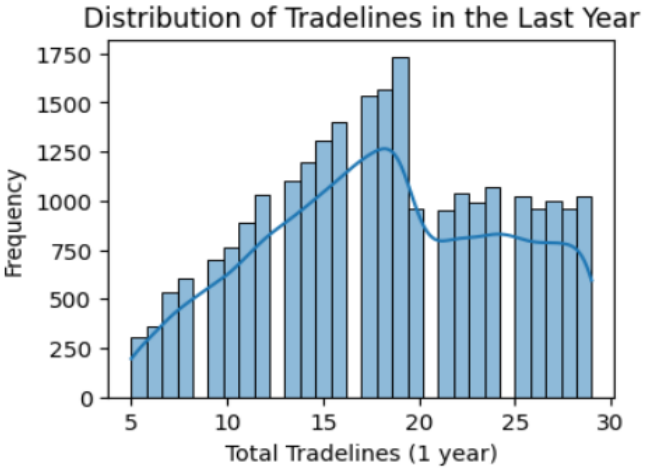
Q6: Identify the Spread of Total_Tradelines_1y_In.

Analyzing the distribution of Total_Tradelines_1y_In will help us understand the average number of tradelines that clients have over a 1-year period. This is essential for assessing the typical credit usage behavior.

Steps to follow:

- Calculate the summary statistics (mean, median, and standard deviation) for Total_Tradelines_1y_In.
- Plot a histogram to visualize the spread of Total_Tradelines_1y_In (bin size = 30).
- Identify whether the data is skewed, and check for any unusual concentrations or outliers.
- Finally, display the mean, median and standard deviation.

Expected Output:




```
Mean: 18.13032, Median: 18.0, Std Dev: 6.279251058633691
```

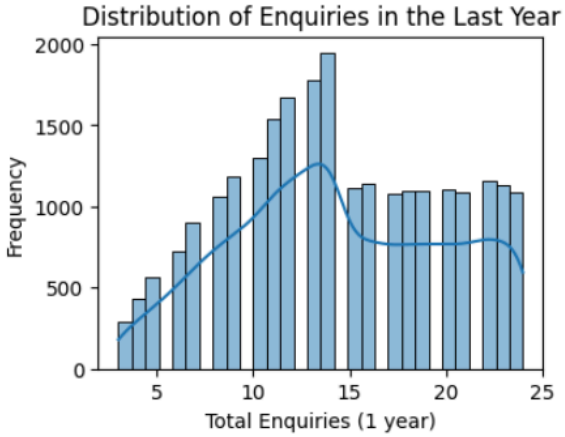
Q7: Distribution of Total_Enquiry_1y_In.

Understanding the distribution of Total_Enquiry_1y_In allows us to assess the overall credit inquiry activity of clients over a one-year period. It's important to detect any skewness or high concentration in a specific range of values.

Steps to follow:

- Compute the summary statistics (mean, median, std dev) for Total_Enquiry_1y_In and display them.
- Plot a histogram to visualize the distribution and look for any signs of skewness or high concentrations.
- Insight: Identify any outliers or extreme values that might indicate clients with unusually high inquiry activity.

Expected Output:




```
Mean: 14.372897959183673, Median: 14.0, Std Dev: 5.531962071561176
```

Stage 4. Data Exploration and Bivariate Analysis

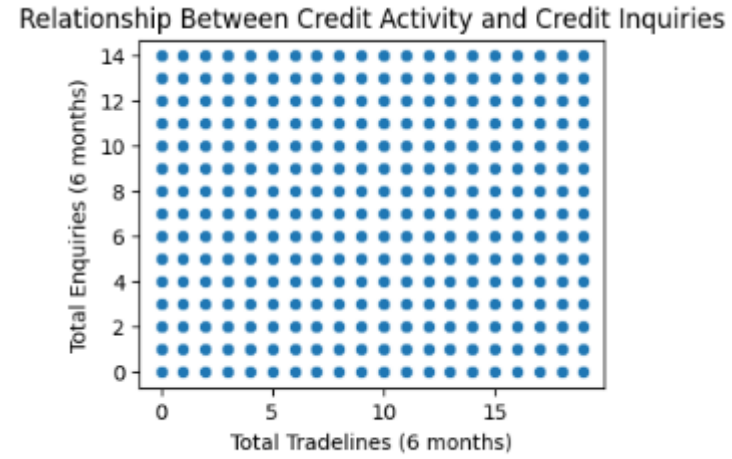
Q1: Investigating Credit Behavior: Does Higher Credit Activity Lead to More Enquiries?

In order to understand the correlation between credit activity and customer engagement with credit providers, we need to explore whether clients with higher tradeline activity tend to inquire about credit more frequently. By examining Total_Tradelines_6m_In (tradelines in the last 6 months) and Total_Enquiry_6m_In (credit enquiries in the same period), we can uncover whether a pattern exists between a client's credit activity and the frequency of their inquiries.

Steps to follow:

- Perform inner join between tradeline_data and enquiry_data on Client ID.
- Calculate the correlation between Total_Tradelines_6m_In and Total_Enquiry_6m_In.
- Visualize the relationship between the two using a scatter plot.
- Interpret the results to determine if increased credit activity is associated with more credit inquiries.

Expected Output Format:




```
Correlation between Credit Activity and Credit Inquiries: 0.01
```

Q2: Exploring Loan Application Behavior: Do Clients with More Tradelines Apply More for Loans?

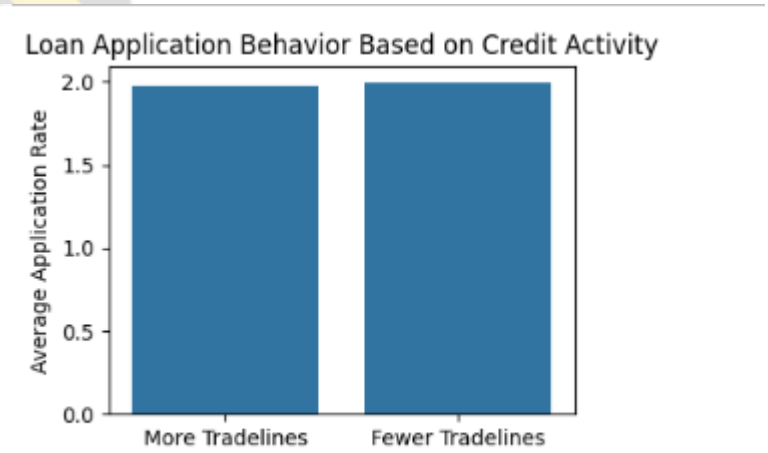
Understanding how credit behavior influences loan application patterns can provide insights into the types of clients who are more likely to apply for loans. By examining the relationship between Total_Tradelines_6m_In and Applied, we can investigate whether clients with a larger number of tradelines are more inclined to apply for loans.

Steps to follow:

- Perform inner join between tradeline_data and base_file on Client ID.

- Calculate the average Applied rate for clients with more than the median number of tradelines (Total_Tradelines_6m_In).
- Compare it to the average Applied rate for clients with fewer tradelines.
- Visualize the differences using a bar plot.
- Also, display the below metrics by rounding off up to 2 decimal places.
    - ➔ Average application rate for clients with more tradelines: {}
    - ➔ Average application rate for clients with fewer tradelines: {}

Expected Output Format:




```
Average application rate for clients with more tradelines: 1.97
Average application rate for clients with fewer tradelines: 1.99
```

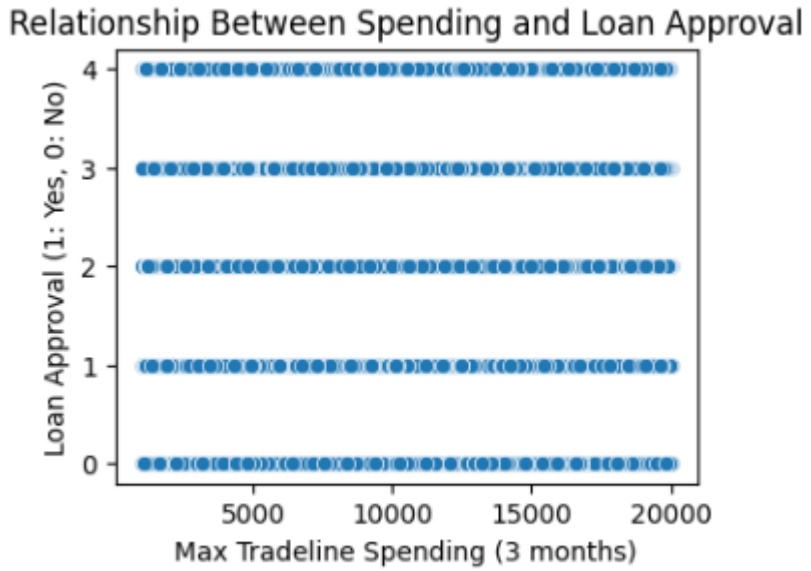
Q3: Understanding Loan Approvals: Does Higher Spending Lead to More Approvals?

To gain a deeper understanding of the approval process, we should investigate whether clients who spend more (Max_Trade_3m_Out) are more likely to have their loans approved. This analysis helps to understand if spending behavior influences the likelihood of loan approval, potentially identifying high-risk applicants.

Steps to follow:

- Perform inner join between base_file and tradeline_data on Client ID.
- Extract the data where Applied > 0.
- Calculate the correlation between Max_Trade_3m_Out (maximum spending in the last 3 months) and Approved.
- Create a scatter plot to visualize the relationship between spending and loan approvals.
- Analyze the results to identify trends in loan approvals for clients with higher spending.
- At last, display the correlation by rounding off the value up to 2 decimal places.

Expected Output Format:




```
Correlation between Spending and Loan Approval: -0.00
```

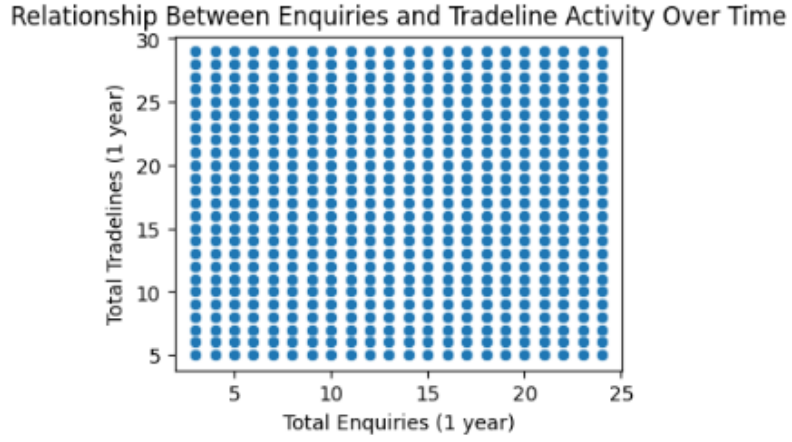
Q4: Exploring Enquiries and Tradeline Behavior Over Time: Do More Enquiries Lead to More Tradelines?

It's important to explore whether clients who make more credit inquiries (Total_Enquiry_1y_In) are also likely to have more tradelines (Total_Tradelines_1y_In) over time. This can help us understand whether frequent inquiries correlate with increased credit utilization and expansion of tradelines, or if it's unrelated.

Steps to follow:

- Perform inner join between enquiry_data and tradeline_data on Client ID.
- Calculate the correlation between Total_Enquiry_1y_In and Total_Tradelines_1y_In.
- Visualize the relationship using a scatter plot.
- Determine whether higher enquiry activity over the last year is associated with an increase in tradeline activity during the same period.
- At last, display the correlation by rounding off the value up to 2 decimal places.

Expected Output Format (Sample):




```
Correlation between Enquiries and Tradeline Activity: 0.00
```

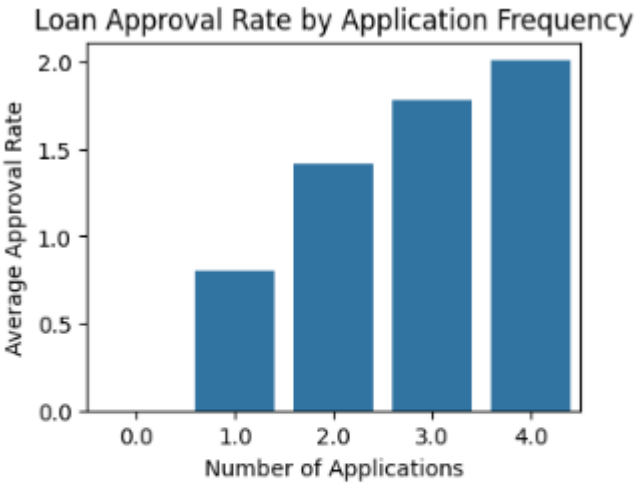
Q5: Analyzing Loan Approvals: Are Clients with More Applications More Likely to Be Approved?

Clients who apply for loans frequently may have different approval rates compared to those who apply less often. By exploring the relationship between loan applications (Applied) and loan approvals (Approved), we can identify whether more frequent applicants have a higher chance of loan approval or if approvals are more likely for one-time applicants.

Steps to follow:

- Calculate the application frequency per Client ID.
- Calculate the average approvals based on application frequency.
- Visualize the comparison using a bar plot.
- Display the average approval rate by application frequency.

Expected Output Format (Sample):




```
Approval rates by application frequency:
Application_Frequency
0.0    0.000000
1.0    0.799960
2.0    1.415497
3.0    1.786318
4.0    2.013803
Name: Approved, dtype: float64
```

Q6: Investigating Spending and Loan Applications: Does Higher Spending Encourage Loan Applications?

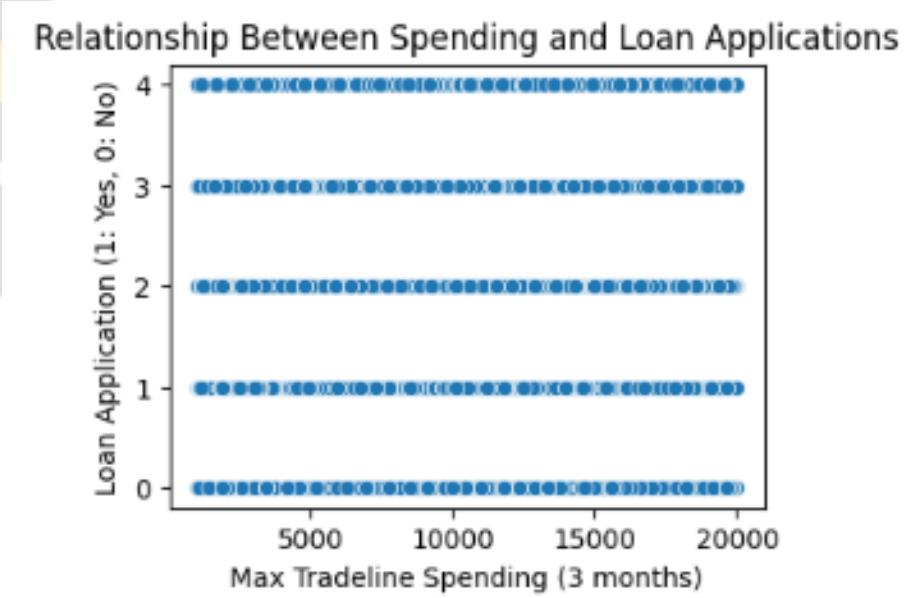
Clients who engage in more spending behavior might have higher credit needs and, consequently, might apply for loans more often. To explore this, we need to analyze the relationship between Max_Trade_3m_Out and Applied.

Steps to follow:

- Perform inner join between tradeline_data and base_file on Client ID.
- Calculate the correlation between Max_Trade_3m_Out and Applied.

- Visualize the relationship using a scatter plot.
- Interpret whether higher spending corresponds to an increased likelihood of applying for loans.
- At last, display the correlation by rounding off the value up to 2 decimal places.

Expected Output Format:




```
Correlation between Spending and Loan Applications: 0.00
```

Stage 5. Feature Engineering

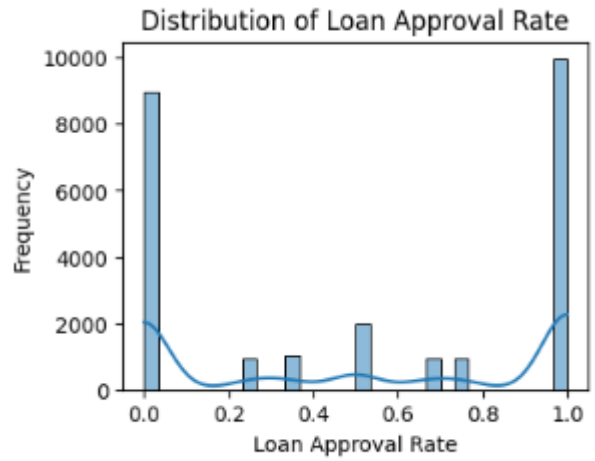
Q1: Assessing the Loan Approval Process: Exploring the Loan Approval Rate.

The loan approval process can be assessed by looking at the proportion of loan applications that result in approval. By calculating the loan approval rate, we can determine the effectiveness of the approval system.

Steps to follow:

- Calculate the loan approval rate by dividing Approved by Applied.
- After that, handle any potential cases where Applied is zero to avoid division by zero errors. Try replacing invalid values (np.nan, np.inf, -np.inf) with value 0.
- Analyze how this loan approval rate varies across clients using histogram and whether it provides insight into the approval process (bin size = 30).
- Calculate the mean approval rate and round off the value up to 2 decimal places.

Expected Output Format:




```
Mean Loan Approval Rate: 0.52
```

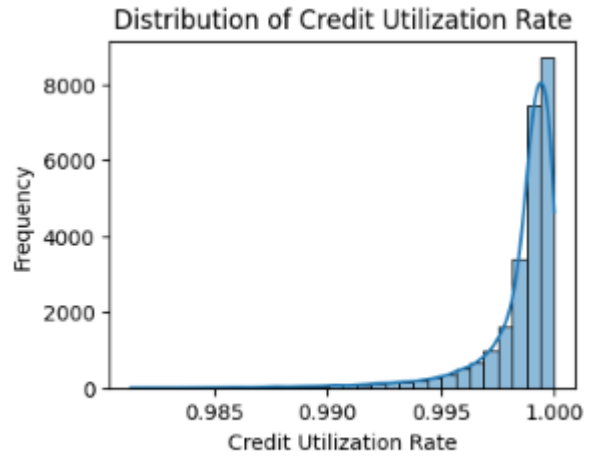
Q2: Understanding Credit Utilization: Evaluating How Clients Use Their Credit.

Clients' credit utilization behavior plays an important role in understanding their financial health. By examining the proportion of credit utilized (Max_Trade_3m_Out) relative to their available tradelines (Total_Tradelines_6m_In), we can understand their credit usage patterns.

Steps to follow:

- Calculate the credit utilization rate by dividing Max_Trade_3m_Out by the sum of Total_Tradelines_6m_In and Max_Trade_3m_Out.
- Analyze the distribution of credit utilization across clients using a histogram.
- Insight: Investigate how this behavior correlates with loan approval or credit inquiries.
- Calculate the average credit utilization rate by rounding off the value up to 2 decimal places.

Expected Output Format:




```
Average Credit Utilization Rate: 1.00
```

Q3: Segmenting Clients Based on Credit Activity.

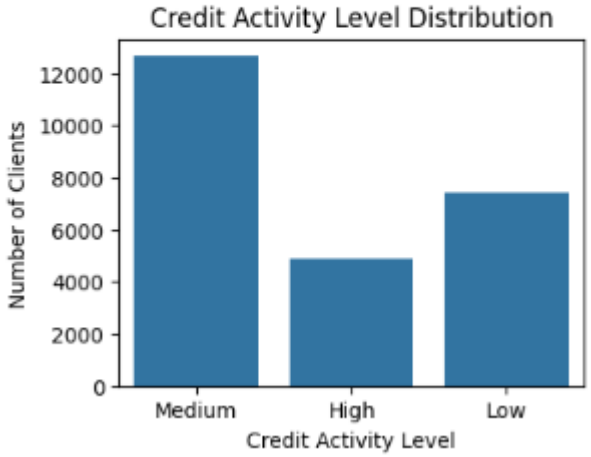
To better understand how different levels of credit activity affect client behavior, we can classify clients into different categories based on their credit usage. These categories can help identify clients with low, medium, or high credit engagement.

Steps to follow:

- Classify clients into three groups based on their Total_Tradelines_6m_In:
  - Low activity (0-5 tradelines)
  - Medium activity (6-15 tradelines)
  - High activity (16+ tradelines)
- Count the number of clients in each segment and display the result.
- Analyze how these groups differ by the number of clients by plotting a count plot.

Expected Output Format:

```
Segment distribution:
Credit_Activity_Level
Medium    12689
Low        7419
High       4892
Name: count, dtype: int64
```



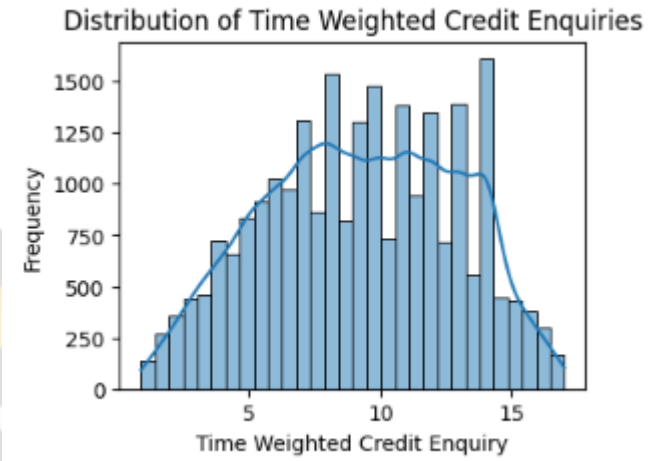
Q4: Capturing Recent Credit Seeking Behavior.

Recent credit enquiries are a strong indicator of a client's current credit-seeking behavior. We can assess the recency and intensity of clients' credit-seeking activity by giving more weight to recent enquiries (Total_Enquiry_6m_In) compared to those from earlier periods.

Steps to follow:

- Calculate a weighted score for credit enquiries, giving more weight to recent enquiries (e.g., Total_Enquiry_6m_In * 0.7 + Total_Enquiry_1y_In * 0.3).
- Plot the distribution of the weighted enquiry using histogram with bin size = 30.
- Calculate the average enquiry score by rounding off the value up to 2 decimal places.

Expected Output Format:




```
Average Time Weighted Enquiry: 9.24
```

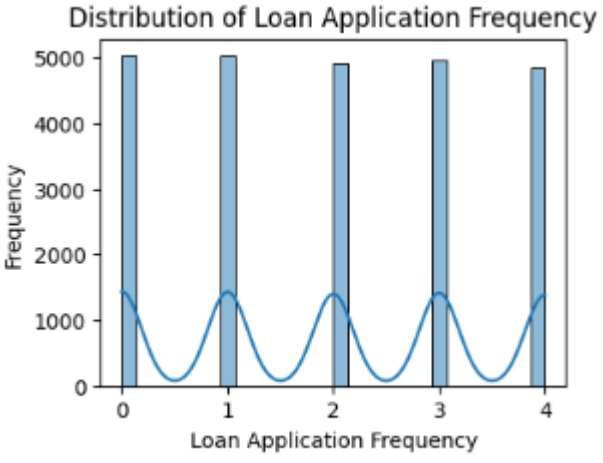
Q5: Examining Loan Application Frequency.

Frequent loan applications might indicate clients with specific financial needs or behaviors. By examining how often clients apply for loans, we can identify patterns that suggest higher or lower credit risk.

Steps to follow:

- Calculate the loan application frequency for each client
- Plot the distribution of loan application frequency using histogram with bin size of 30.
- Calculate the average application frequency by rounding off the value up to 2 decimal places.

Expected Output Format:




```
Average Application Frequency: 1.98
```

Q6: Evaluating Risk Based on Credit Behavior.

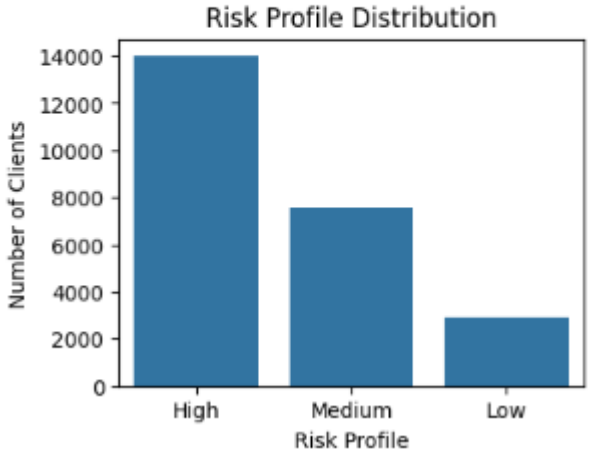
Clients with high credit activity and frequent credit inquiries might be considered higher risk. By combining these behaviors, we can develop a risk profile that categorizes clients based on their tradeline activity and credit inquiries.

Steps to follow:

- Create a risk profile by categorizing clients into three groups based on their Total_Tradelines_6m_In and Total_Enquiry_6m_In (merge suitable datasets):
  - Low risk: Total_Tradelines_6m_In and Total_Enquiry_6m_In are less than or equal to 5.
  - Medium risk: Total_Tradelines_6m_In and Total_Enquiry_6m_In are less than or equal to 15.
  - High risk: otherwise, marked as High.
- Count the number of clients in each risk profile and display them.
- Plot the distribution of Risk Profile using the count plot.

Expected Output Format:

```
Risk Profile distribution:
Risk_Profile
High      14020
Medium     7586
Low        2894
Name: count, dtype: int64
```



Stage 6. Advanced Analysis

Q1: Identify High-Risk Clients Based on Credit Behavior and Spending.

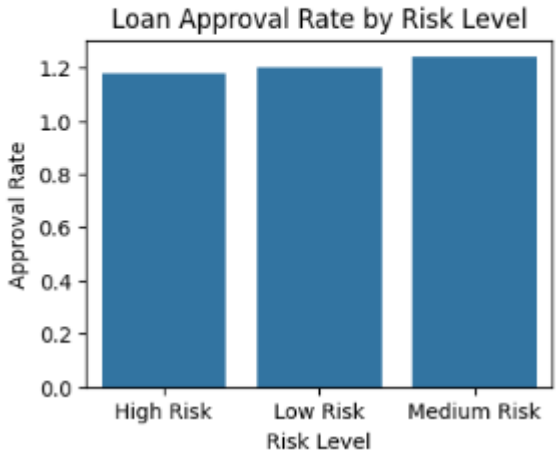
Clients with high credit usage and frequent spending could pose a higher risk. We'll define high-risk clients based on the number of tradelines and spending, segmenting them into low, medium, and high-risk categories.

Steps to follow:

- Define high-risk clients as those who have high numbers of both tradelines (Total_Tradelines_6m_In) and spending (Max_Trade_3m_Out).
- Create a risk profile based on these two criteria:
    - High Risk: Total_Tradelines_6m_In is greater than 10 and Max_Trade_3m_Out is greater than 5000.
    - Medium Risk: Total_Tradelines_6m_In is between 5 (included) and 10 (included) and Max_Trade_3m_Out is between 2000 (included) and 5000 (included).
    - Low Risk: otherwise, consider them as low risk.
- Calculate the average loan approval rate by risk level and display the results.
- Visualize the approval rate by risk level using bar plot.

Expected Output Format:

```
Approval rate by risk level:
Risk_Level
High Risk      1.177218
Low Risk       1.201465
Medium Risk    1.241379
Name: Approved, dtype: float64
```



Q2: Understand Loan Application Patterns Based on Credit Activity.

Clients who have a higher number of tradelines may have different patterns when it comes to loan applications. We need to explore whether clients with more tradelines tend to apply for loans more frequently, and if so, how their approval rates compare.

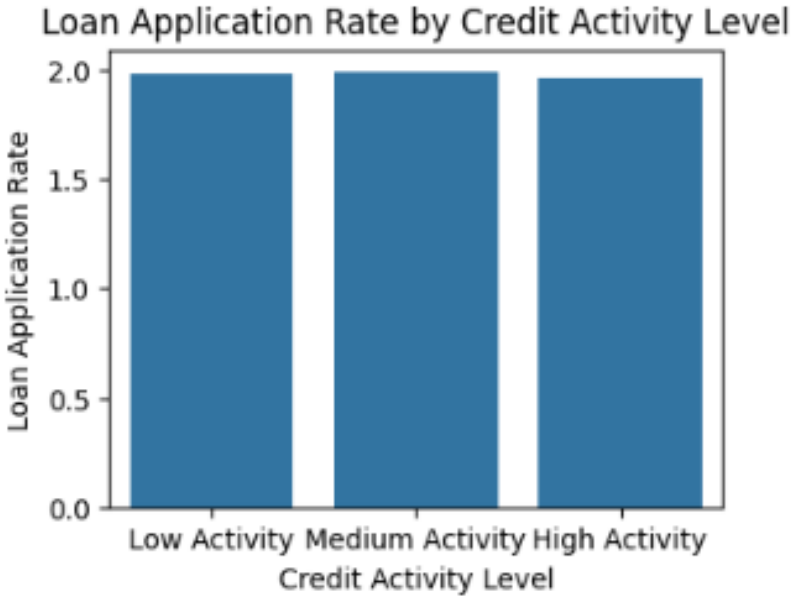
Steps to follow:

- Group clients by their credit activity (Total_Tradelines_6m_In), segmenting them into low, medium, and high activity.
    - **Low Risk:** Total_Tradelines_6m_In is less than or equal to 5.
    - **Medium Risk:** Total_Tradelines_6m_In is greater than 5 and less than or equal to 15.
    - **High Risk:** Total_Tradelines_6m_In is greater than 15.
- Calculate the average loan application rate for each group and display the results up to 2 decimal places.

- Investigate whether clients with higher tradeline activity tend to apply for loans more often by plotting a bar plot.

Expected Output Format:

```
Average loan application rate for low activity: 1.98
Average loan application rate for medium activity: 1.99
Average loan application rate for high activity: 1.97
```



Q3: Explore the Impact of Spending Behavior on Loan Approval.

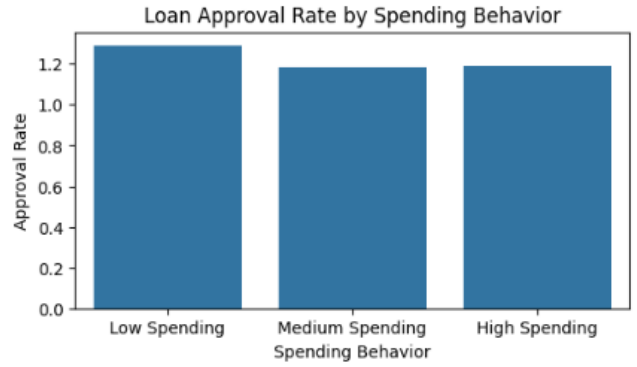
By comparing spending behavior (Max_Trade_3m_Out) with loan approval rates (Approved), we can understand if clients who spend more are more likely to get their loans approved.

Steps to follow:

- Group clients by spending behavior (Low, Medium, High spending)
  - **Low Risk:** Max_Trade_3m_Out is less than or equal to 2000.
  - **Medium Risk:** Max_Trade_3m_Out is greater than 2000 and less than or equal to 5000.
  - **High Risk:** Max_Trade_3m_Out is greater than 5000.
- Calculate the average approval rate for each group and display the results by rounding off the value up to 2 decimal places.
- Analyze the approval rate across different spending thresholds by plotting the bar plot.

Expected Output Format:

```
Approval rate for low spending clients: 1.29
Approval rate for medium spending clients: 1.18
Approval rate for high spending clients: 1.19
```



Q4: Identifying Client Segments with High Loan Application Frequency.

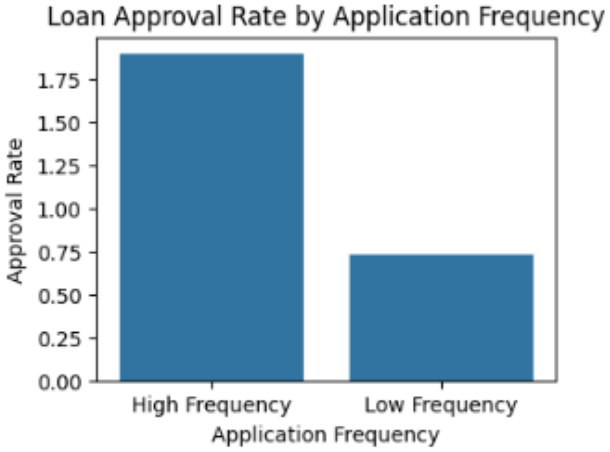
Understanding the behavior of clients who frequently apply for loans will help us assess which clients may need more support or tailored services. This question will explore how often clients apply for loans and identify segments of frequent applicants.

Steps to follow:

- Segment clients into high and low application frequency.
    - **High Application Frequency**: If Application_Frequency is greater than median of Application_Frequency.
    - **Low Application Frequency**: If Application_Frequency is less than or equal to median of Application_Frequency.
- Calculate the average approval rate for each group and display the results by rounding off the value up to 2 decimal places.
- Investigate whether frequent applicants are more likely to be approved or face more rejections by plotting the visualization via bar plot.
  Note: *Application_frequency column is already created in Stage 5: Q5.*

Expected Output Format:

```
Approval rate for high application frequency clients: 1.90
Approval rate for low application frequency clients: 0.73
```



Q5: Investigating the Link Between Loan Approval and Client Engagement.

Clients with higher levels of engagement, as indicated by tradelines and inquiries, may exhibit different approval patterns. We will explore if clients who are more engaged in managing their credit are more likely to get their loans approved.

Steps to follow:

- Merge tradeline_data with enquiry_data and finally combine them with the base_file dataset.
- Segment clients into engaged and less engaged groups:
    - **Engaged Clients**: If Total_Tradelines_6m_In is greater than 5 and Total_Enquiry_6m_In is greater than 5.
    - **Less Engaged Clients**: If Total_Tradelines_6m_In is less than or equal to 5 and Total_Enquiry_6m_In is less than or equal to 5.
- Compare average loan approval rates between engaged and less engaged clients and display the results by rounding off the value up to 2 decimal places.
- Explore how engagement affects the likelihood of loan approval by visualizing barplot.

Expected Output Format:

```
Approval rate for engaged clients: 1.20
Approval rate for less engaged clients: 1.20
```