Exploratory Data Analysis (EDA) for Credit

Name: Kartik Ukhalkar

Content

Problem Statement

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Data Sourcing

Data Sourcing involves gathering and acquiring data from various sources. This step is crucial as the quality and relevance of the data can significantly impact the analysis outcomes.

Identifying Sources: We have been provided 3 csv file containing banking data in csv format to analyze.

We then import the files in notebook and do further steps in EDA.

Data Cleaning

- ## Data types

1

Identify Data Types

Ensure all variables are properly categorized as numeric, categorical, or date/time.

2

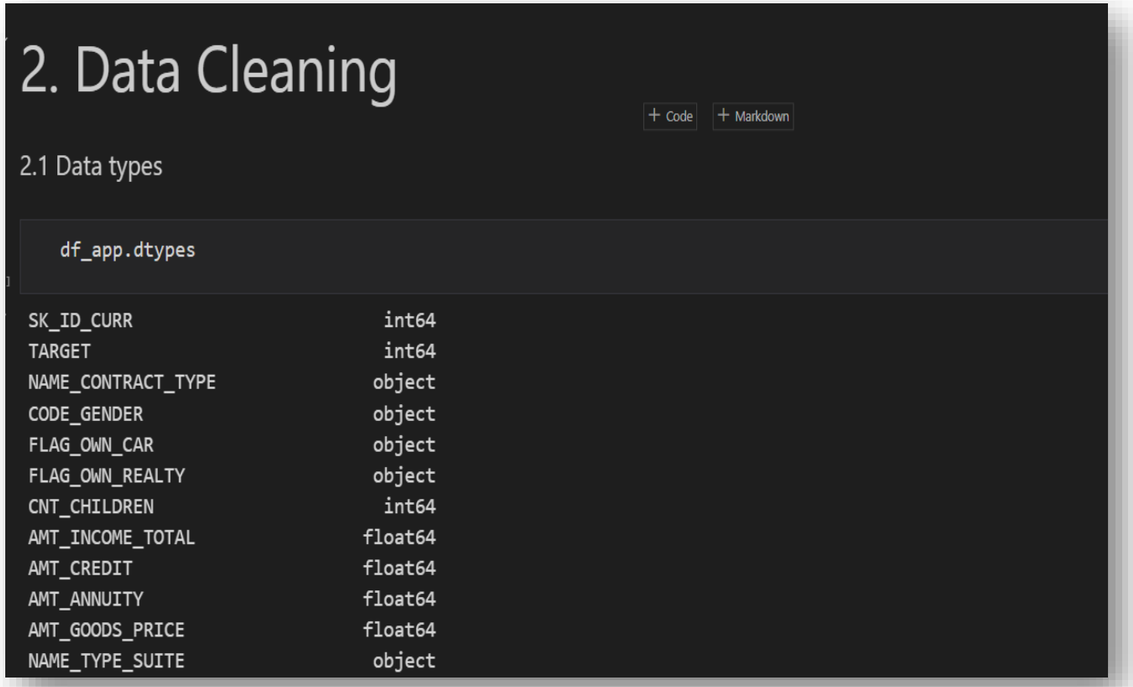
Handle Inconsistencies

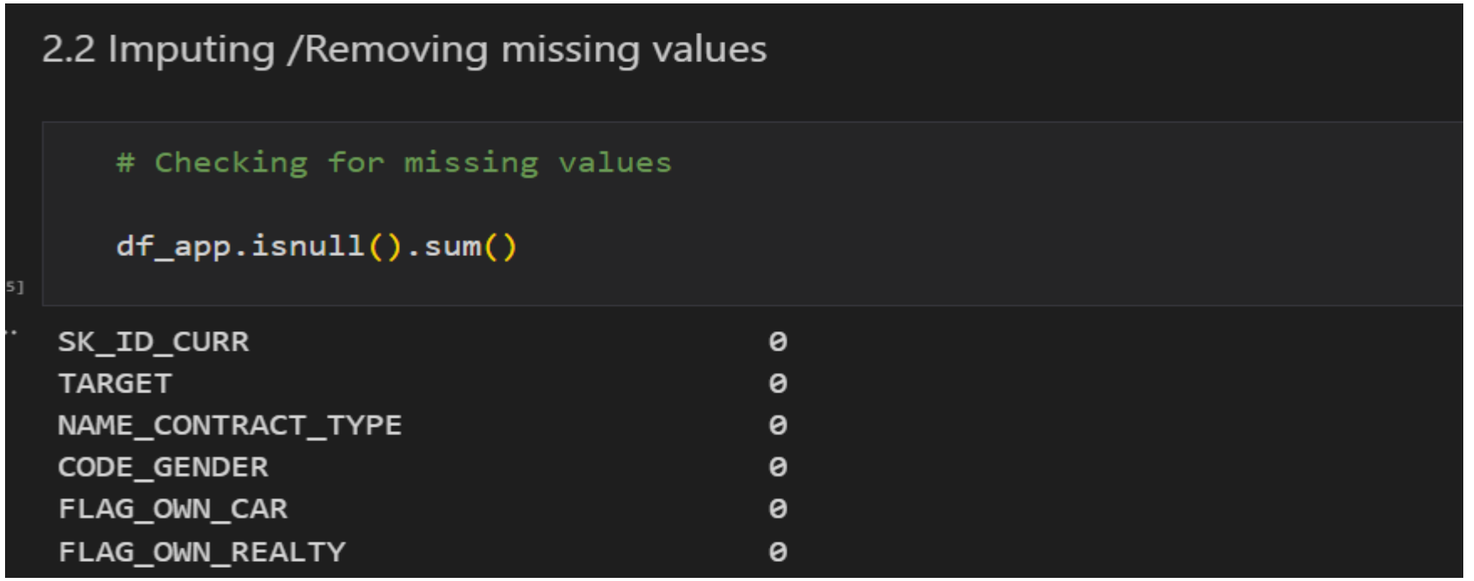
Address any inconsistencies in how data is represented and recorded.

3

Standardize Formats

Ensure all values are in a consistent, usable format across the dataset.





• Imputing /Removing missing values

1 **Identify Missing Values**

Thoroughly audit your dataset to find and quantify any missing data.

2 **Imputation Strategies**

Leverage techniques like mean/median imputation, K-nearest neighbors, or more sophisticated ML-based imputation.

3 **Remove Unreliable Data**

If imputation is not feasible, carefully remove rows or columns with excessive missing values.

• Fixing rows and columns

Rename Columns

Ensure column names are clear, concise, and consistent.

Reorder Columns

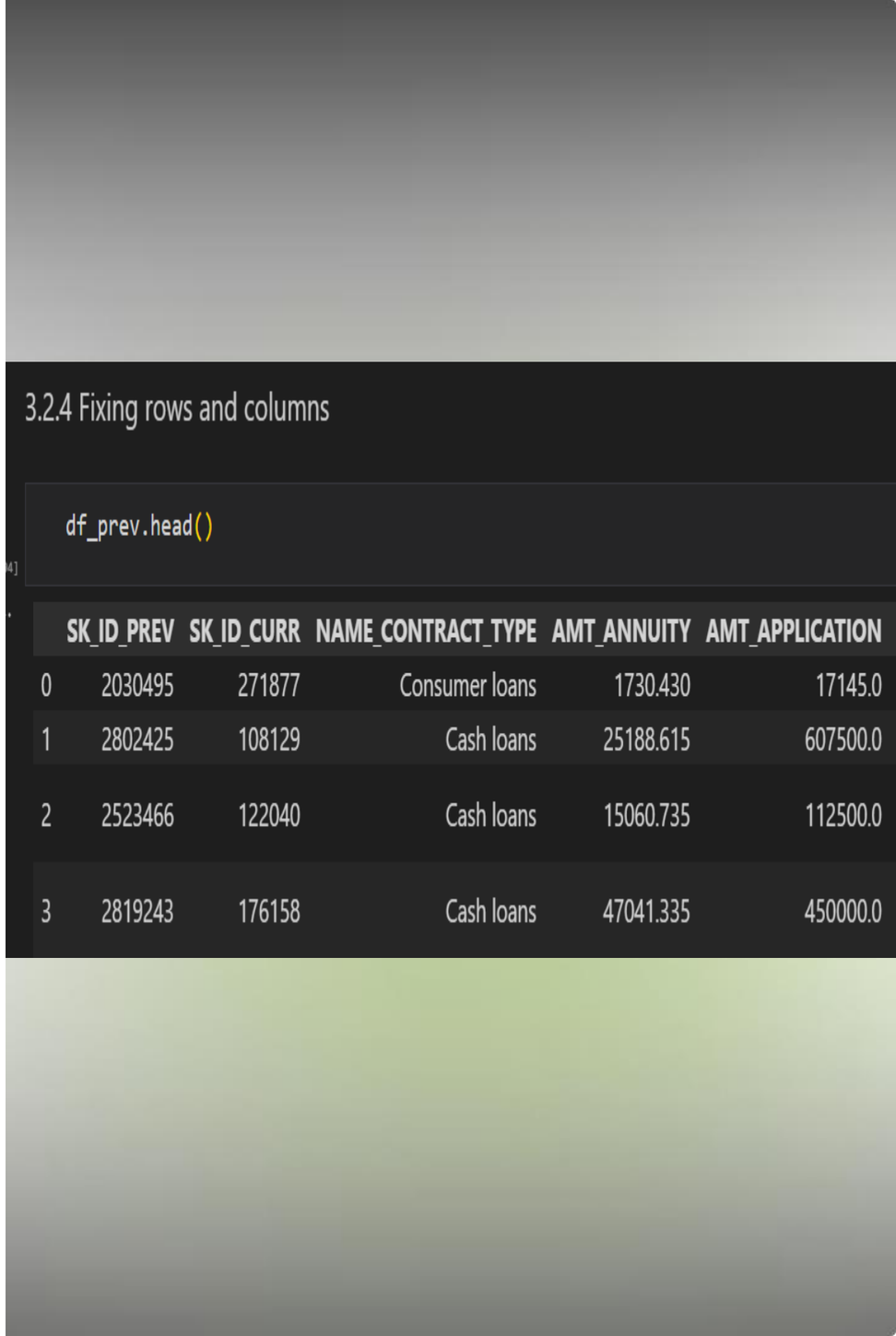
Arrange columns in a logical order to facilitate analysis.

Reshape Data

Pivot, melt, or transpose the data as needed for a tidy structure.

Handle Duplicates

Identify and address any duplicate rows in the dataset.



	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION
0	2030495	271877	Consumer loans	1730.430	17145.0
1	2802425	108129	Cash loans	25188.615	607500.0
2	2523466	122040	Cash loans	15060.735	112500.0
3	2819243	176158	Cash loans	47041.335	450000.0

• Handling outliers

Identify

Use statistical techniques like Z-scores or box plots to pinpoint outliers.

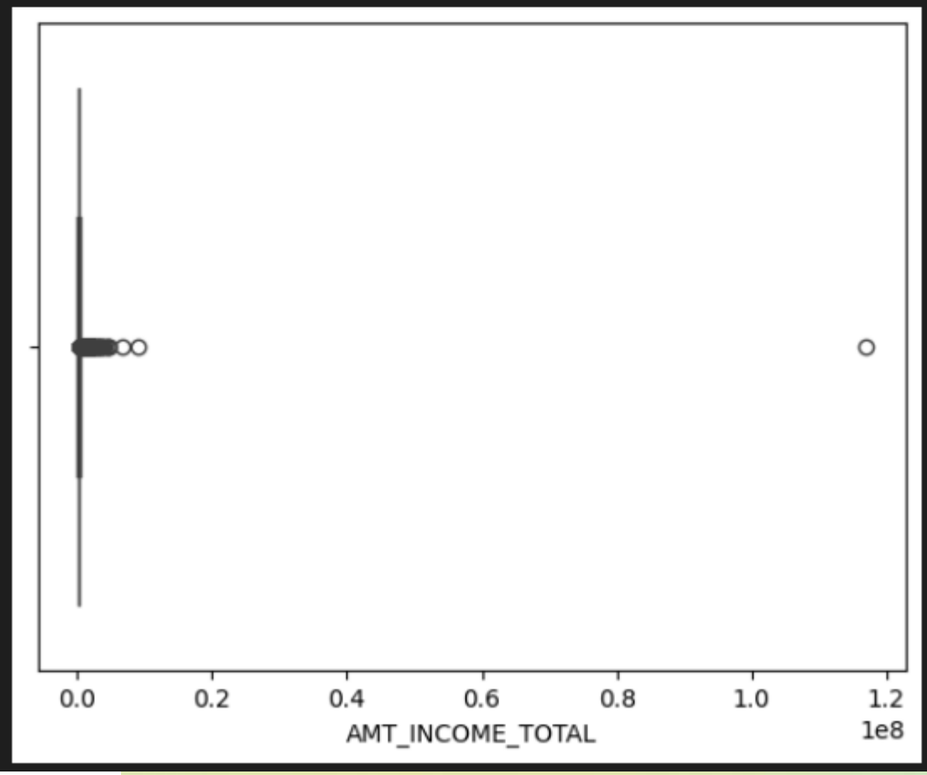
Remove

Carefully remove outliers, ensuring you don't lose valuable information.

Transform

Consider transforming outliers to reduce their influence on the data.

```
sns.boxplot(x = df_app.AMT_INCOME_TOTAL)
plt.show()
```



• Standardising values

1

Min-Max Scaling

Rescale variables to a common range, often 0 to 1.

2

Z-Score Normalization

Standardize variables by subtracting the mean and dividing by the standard deviation.

3

Robust Scaling

Use the median and median absolute deviation to create more outlier-resistant scaling.

3. Univariate Analysis

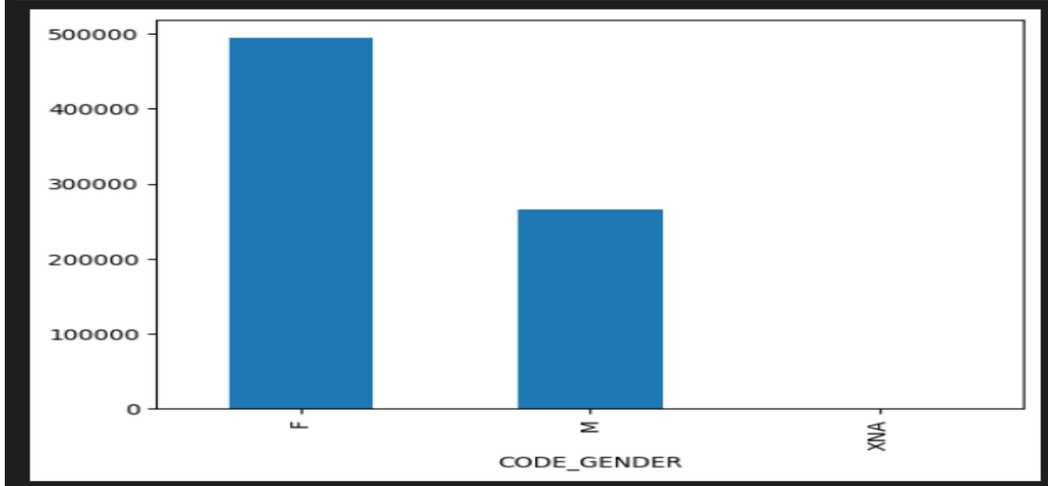
Univariate Analysis involves examining each variable in isolation to understand its distribution and basic properties.

3.1 Categorical univariate analysis

Categorical Univariate Analysis focuses on analyzing and summarizing a single categorical variable to understand its distribution and characteristics. This type of analysis helps you to gain insights into the frequency and pattern of categories within the variable.

```
# Plotting for Income range across various genders

df_merge[df_merge.TARGET==0].CODE_GENDER.value_counts().plot(kind='bar')
plt.show()
```



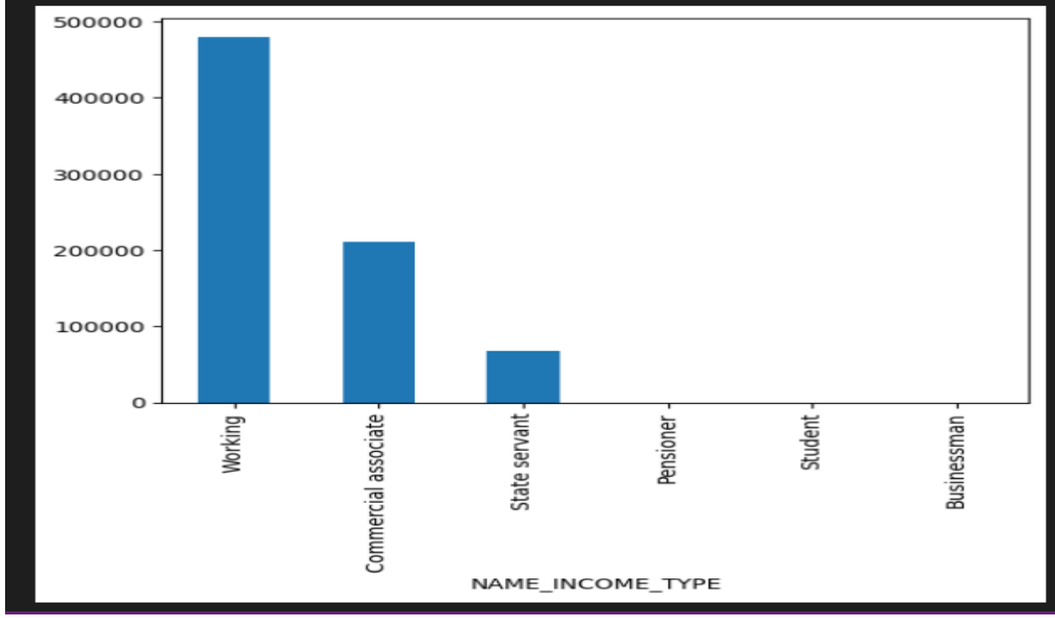
3.2 Numerical univariate analysis

Numerical Univariate Analysis focuses on analyzing and summarizing a single numerical (or quantitative) variable to understand its distribution, central tendencies, and variability. This analysis helps you gain insights into the characteristics of the data, such as its shape, spread, and central value.

Numerical univariate analysis provides foundational insights into a dataset, making it easier to perform more advanced analyses and make informed decisions.

```
# Plotting for various Income types

df_merge[df_merge.TARGET==0].NAME_INCOME_TYPE.value_counts().plot(kind='bar')
plt.show()
```

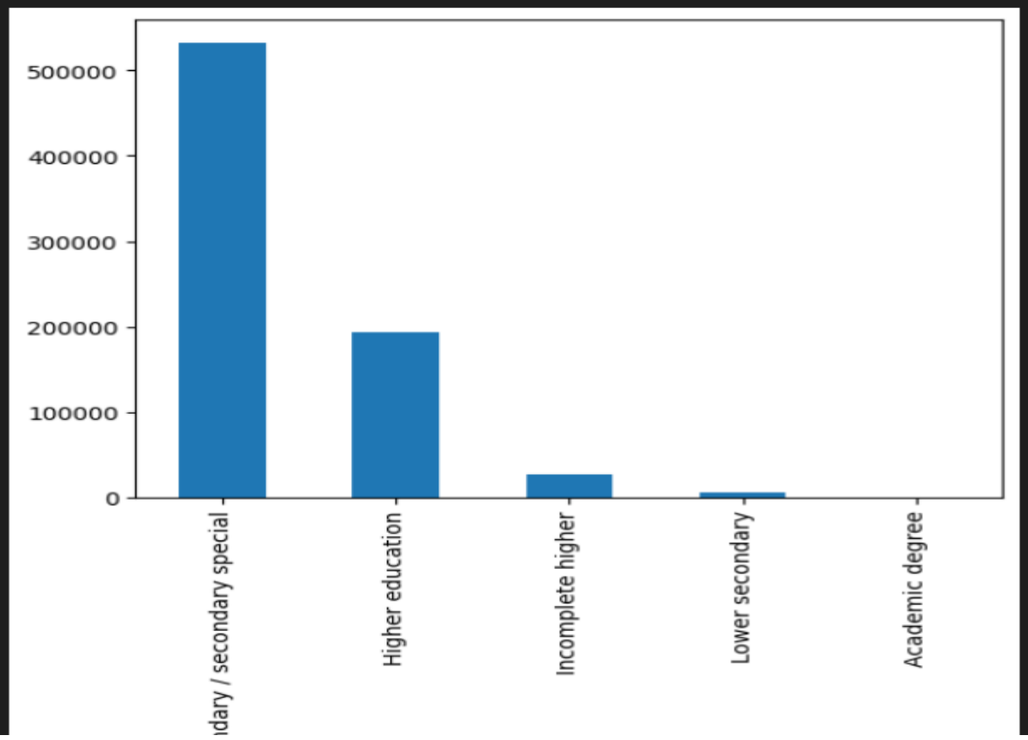


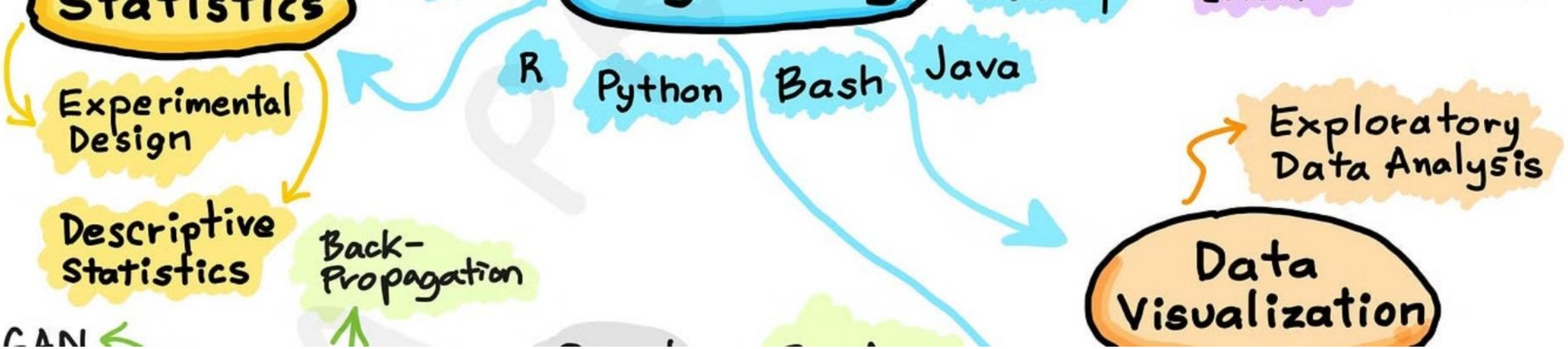
Bivariate Analysis

Bivariate Analysis is the statistical analysis of two variables to understand the relationship or association between them. This type of analysis helps determine how one variable may influence or correlate with another and is a crucial step in data exploration and modeling.

```
# Box plotting for the TARGET = 0, Education

df_merge[df_merge.TARGET==0].NAME_EDUCATION_TYPE.value_counts().plot(kind='bar')
plt.show()
```





Conclusion

In conclusion, a comprehensive framework for EDA in credit risk assessment is essential for effective decision-making. By following the outlined steps, organizations can enhance their understanding of credit risks and improve their overall risk management strategies.