Akshay Panchal

Bank Loan Case Study

Description:

In this case study, we will explore the application of Exploratory Data Analysis (EDA) in the context of risk analytics in the banking and financial services industry. The objective is to understand how data can be utilized to minimize the risk of financial loss when lending money to customers.

The primary challenge faced by loan providing companies is the difficulty of granting loans to individuals with insufficient or nonexistent credit history. This creates an opportunity for some consumers to take advantage of the situation by intentionally defaulting on their loans.

To address this issue, let's consider the scenario where we work for a consumer finance company specializing in offering various types of loans to urban customers. By applying EDA techniques, we can analyze the patterns and insights present in the data. This analysis will enable us to identify applicants who are capable of repaying the loan, thereby minimizing the rejection of potentially eligible applicants.

When the company receives a loan application, it needs to make a decision regarding loan approval based on the applicant's profile. This decision involves two types of risks:

1. **Risk of Not Approving a Loan:** If the applicant is likely to repay the loan, rejecting the loan application would result in a loss of business for the company. EDA can help identify reliable indicators and patterns that indicate the applicant's creditworthiness, allowing the company to make informed decisions.

2. **Risk of Default:** If the applicant is unlikely to repay the loan, approving the loan could lead to a financial loss for the company. EDA can help identify potential factors or red flags associated with defaulters, enabling the company to minimize the risk of approving loans to individuals who are likely to default.

In the loan application process, there are four possible decisions that can be made by the client or the lending company:

1. **Approval:** The lending company approves the loan application, indicating that the client is eligible and meets the necessary criteria for borrowing.
2. **Cancellation:** The client cancels the loan application after it has been approved or during the approval process. This can occur if the client changes their mind about the loan or if, in some cases, they are offered unfavourable terms due to a higher perceived risk.
3. **Refusal:** The lending company rejects the loan application because the client does not meet their requirements or fails to satisfy certain criteria.
4. **Unused Offer:** The loan is approved but ultimately cancelled by the client at different stages of the process.

In this case study, the goal is to utilize Exploratory Data Analysis (EDA) to gain insights into how consumer attributes and loan attributes impact the likelihood of default. By analysing these factors, we can better understand the patterns and relationships that influence the probability of a borrower defaulting on their loan. This information will enable the lending company to make more informed decisions and develop strategies to minimize the risk of defaults.

In this case study, our objectives are as follows:

1. **Identify Patterns of Payment Difficulty:** We aim to identify patterns in the data that indicate whether a client is likely to have difficulty paying their loan installments. This information can be used to take appropriate actions, such as denying the loan, reducing the loan amount, or offering loans to risky applicants at higher interest rates.
2. **Determine Driving Factors for Loan Default:** We want to identify the key variables that strongly indicate the likelihood of loan default. These variables act as drivers and play a significant role in predicting the probability of a borrower defaulting on their loan.
3. **Presenting the Data Analysis Approach:** We will outline the overall approach for analyzing the data, including steps such as data cleaning, identifying outliers, addressing data imbalance, and performing various types of analysis, such as univariate, segmented univariate, and bivariate analysis. These steps will help us gain insights into the data and understand the relationships between variables.
4. **Top 10 Correlations for Clients with Payment Difficulties:** We will specifically explore the correlations between variables for clients who have experienced payment difficulties, which is our target variable. By identifying the top 10 correlations, we can highlight the relationships that have the strongest impact on the likelihood of payment difficulties.

Overall, this case study aims to leverage EDA techniques to uncover meaningful insights about loan default risks and develop a comprehensive understanding of the data to inform decision-making in the lending process.

- **Tech Stack Utilized for Data**
  To approach the given tasks efficiently, you have utilized MS-Excel as the primary tool for schema exploration and data analysis. By thoroughly examining the structure of the dataset and understanding the relationships between different columns, you can execute the required tasks effectively.
  By leveraging the capabilities of MS-Excel, you can perform various operations such as data cleaning, column manipulation, filtering, and calculations to extract the necessary insights. MS-Excel provides a user-friendly interface and a wide range of functions and formulas that facilitate data analysis and manipulation

- **Approach:**
  Analyzing Patterns of Payment Difficulty using EDA

- **Problem Statement:**
  The objective of this case study is to identify patterns that indicate whether a client will have difficulty paying their loan installments. This analysis aims to support decision-making processes such as loan approval, loan amount adjustment, or offering loans to risky applicants at higher interest rates. The goal is to ensure that eligible borrowers capable of repaying the loan are not rejected. By leveraging Exploratory Data Analysis (EDA), we aim to identify applicants who are more likely to experience payment difficulties and understand the influence of consumer attributes and loan attributes on the tendency of default.

- **Analysis Approach:**
    1. Dataset Overview: We start by examining the dataset provided by the client, which contains information about loan applications at the time of applying for the loan. The dataset is categorized into two scenarios:
        a. Clients with payment difficulties: These are individuals who had late payments exceeding a specified threshold (X days) on at least one of the first Y loan installments.
        b. All other cases: This category includes individuals whose loan payments were made on time.
    2. Exploratory Data Analysis (EDA): We will employ EDA techniques to gain insights into the data and understand the relationships between consumer attributes, loan attributes, and the tendency of default. The EDA process will involve:
        a. Data Cleaning: Handling missing values, outliers, and data inconsistencies to ensure data quality.
        b. Data Visualization: Creating visual representations such as histograms, box plots, and scatter plots to identify patterns, distributions, and potential relationships.
        c. Univariate Analysis: Analyzing individual variables to understand their distributions, central tendencies, and potential outliers.
        d. Segmented Univariate Analysis: Exploring variables by segmenting the dataset based on relevant factors to identify differences in payment difficulties.
        e. Bivariate Analysis: Examining the relationships between different variables to identify correlations, associations, and potential drivers of loan default.

By following this approach, we aim to gain a comprehensive understanding of the data, uncover meaningful patterns, and identify key factors that contribute to the likelihood of payment difficulties. These insights will aid in developing strategies for risk assessment and decision-making processes in the lending industry.

- **Analysis Approach:**

1. **Importing Libraries:**
   The first step in the analysis process involves importing essential Python libraries such as NumPy, pandas, matplotlib, and seaborn. These libraries provide the necessary tools for data manipulation, analysis, visualization, and statistical calculations.

2. **Importing Datasets:**
   The analysis utilizes two datasets: "Application_Data" and "Previous_Application." These datasets contain relevant information about loan applications and previous loan history.

3. **Identification:**
   This step focuses on understanding the data and determining the approach for analysis. It involves exploring the dataset's structure, identifying missing data, and formulating a plan to handle it effectively. By gaining a comprehensive understanding of the data, we can proceed with further analysis.

4. **Outlier Analysis:**
   Outliers, if present in the dataset, can significantly impact the analysis results. This step involves identifying and examining outliers in the data. By analyzing their influence, we can determine whether to remove them or handle them through appropriate techniques.
5. **Data Imbalance:**
   Understanding the data imbalance is crucial, especially when dealing with classification problems. This step involves assessing the ratio of imbalance between the different classes in the dataset. By

understanding the data distribution, we can make informed decisions regarding sampling techniques or applying appropriate algorithms to handle the imbalance.

6. **Correlation Analysis:**
   Correlation analysis helps uncover relationships between variables and the target variable. This step involves calculating correlations between different variables in the dataset, specifically focusing on their relationship with the target variable. The top three correlations are identified to understand the variables' impact on the target.

7. **Data Visualization:**
   Visualizing the data using charts and graphs is an effective way to gain insights and present findings. This step involves creating visual representations such as histograms, bar plots, scatter plots, and heatmaps to showcase the data patterns, distributions, and relationships between variables.

By following these steps, we can conduct a comprehensive analysis of the data, identify key insights, and effectively communicate the findings through data visualizations.

- **Dealing with Missing Data:**

1. **Identification of Columns with Missing Values:**
   We observe that in the "Application_Data" dataset, there are 49 columns and in the "Previous_Application" dataset, there are 11 columns that have missing values greater than 40%.

2. **Correlation Analysis:**
   Further analysis reveals that the columns "EXT_SOURCE_2" and "EXT_SOURCE_3" have no correlation with the "TARGET" column, which is the loan repayment status.

3. **Relationship with Loan Repayment Status:**
   Upon investigating the relationship between the "FLAG_DOCUMENT_X" columns and loan repayment status, we find that clients applying for loans only submitted the "FLAG_DOCUMENT_3" column.

4. **No Correlation with Loan Repayment Status:**
   Several columns such as "FLAG_MOBIL", "FLAG_EMP_PHONE", "FLAG_WORK_PHONE", "FLAG_CONT_MOBILE", "FLAG_PHONE", and "FLAG_EMAIL" show almost no correlation with the "TARGET" column.

5. **Unnecessary Columns:**
   The columns "WEEKDAY_APPR_PROCESS_START", "HOUR_APPR_PROCESS_START", "FLAG_LAST_APPL_PER_CONTRACT", and "NFLAG_LAST_APPL_IN_DAY" in the "Previous_Application" dataset are not required for the analysis.

6. **Dropping Columns:**
   We will drop the above-mentioned columns from both datasets. In total, 76 columns will be dropped from the "Application_Data" dataset and 15 columns from the "Previous_Application" dataset.

7. **Converting Negative Days:**
   For columns representing days, we will convert negative values to positive values.

8. **Imputing Null Values:**
   We will impute the remaining null values in the columns needed for data analysis. For numerical data, we will use mean or median, and for categorical data, we will use mode.

9. **Imputing Categorical Variables:**
   The categorical variable "NAME_TYPE_SUITE" will be imputed using the mode, and "OCCUPATION_TYPE" will have an additional category of 'Unknown' for missing values.

10. **Imputing Numerical Variables:**
    Numerical variables such as "AMT_REQ_CREDIT_BUREAU_HOUR", "AMT_REQ_CREDIT_BUREAU_DAY", "AMT_REQ_CREDIT_BUREAU_WEEK", "AMT_REQ_CREDIT_BUREAU_MON", "AMT_REQ_CREDIT_BUREAU_QRT", and "AMT_REQ_CREDIT_BUREAU_YEAR" will be imputed using the median.

11. **Imputing Remaining Variables:**
    "AMT_ANNUITY" will be imputed with the median, "AMT_GOODS_PRICE" with the mode, and "CNT_PAYMENT" with 0, as the "NAME_CONTRACT_STATUS" indicates that most of these loans were not started.

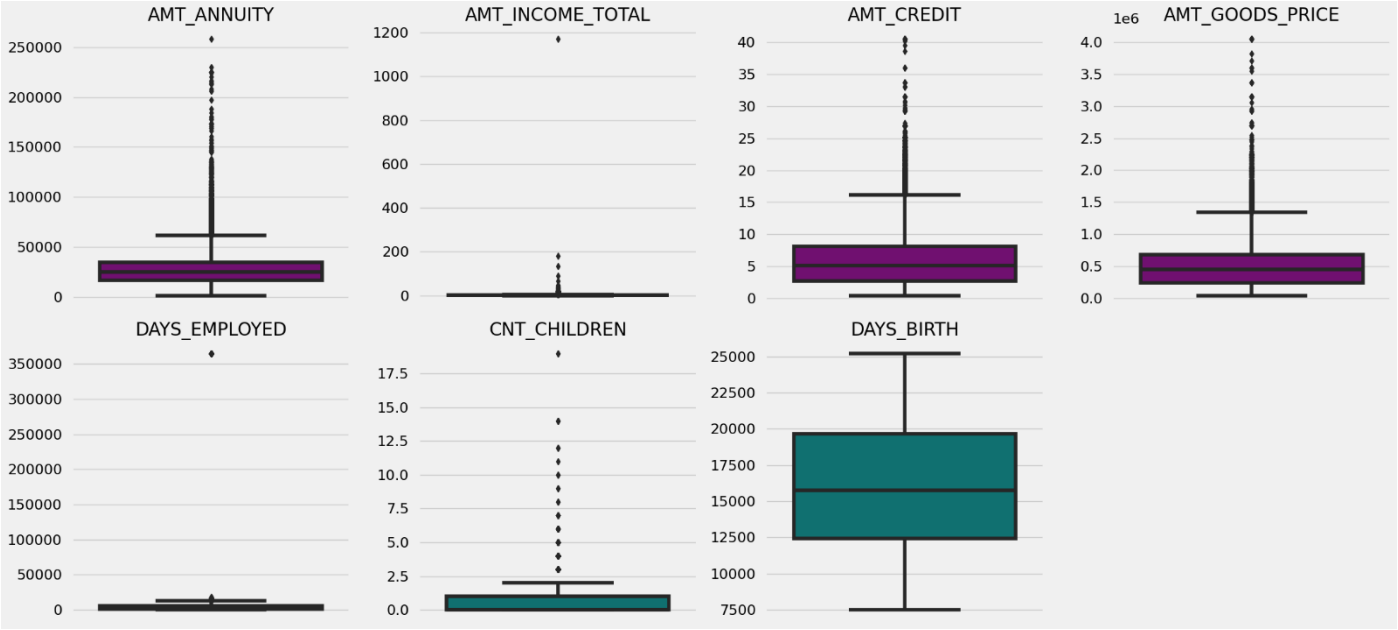
- **Identification of Outliers:**
  Outliers are observations that deviate significantly from other values in a dataset. They can be identified through various statistical methods, including the use of box plots. In a box plot, outliers are typically represented as individual data points lying above the upper whisker (maximum) or below the lower whisker (minimum).

  To identify outliers in the dataset, we can create box plots for numerical variables and visually inspect any data points lying outside the whiskers. If a data point is located far away from the central distribution of values, it can be considered an outlier.

Gained Insights:

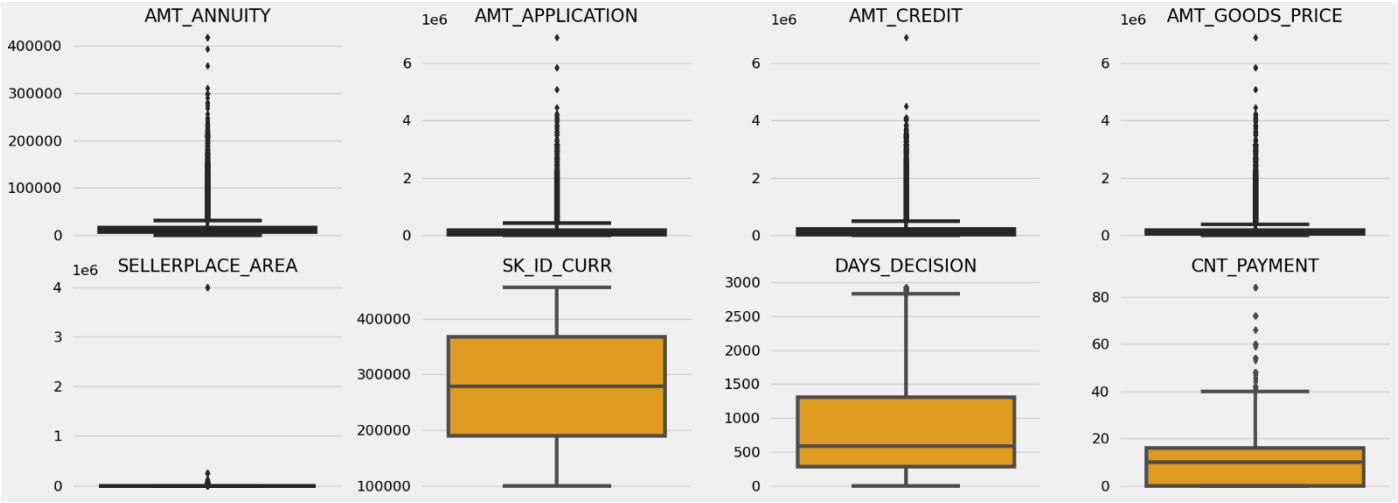
A. Application_Data:

1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
2. AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income compared to the others.
3. DAYS_BIRTH has no outliers which means the data available is reliable.
4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



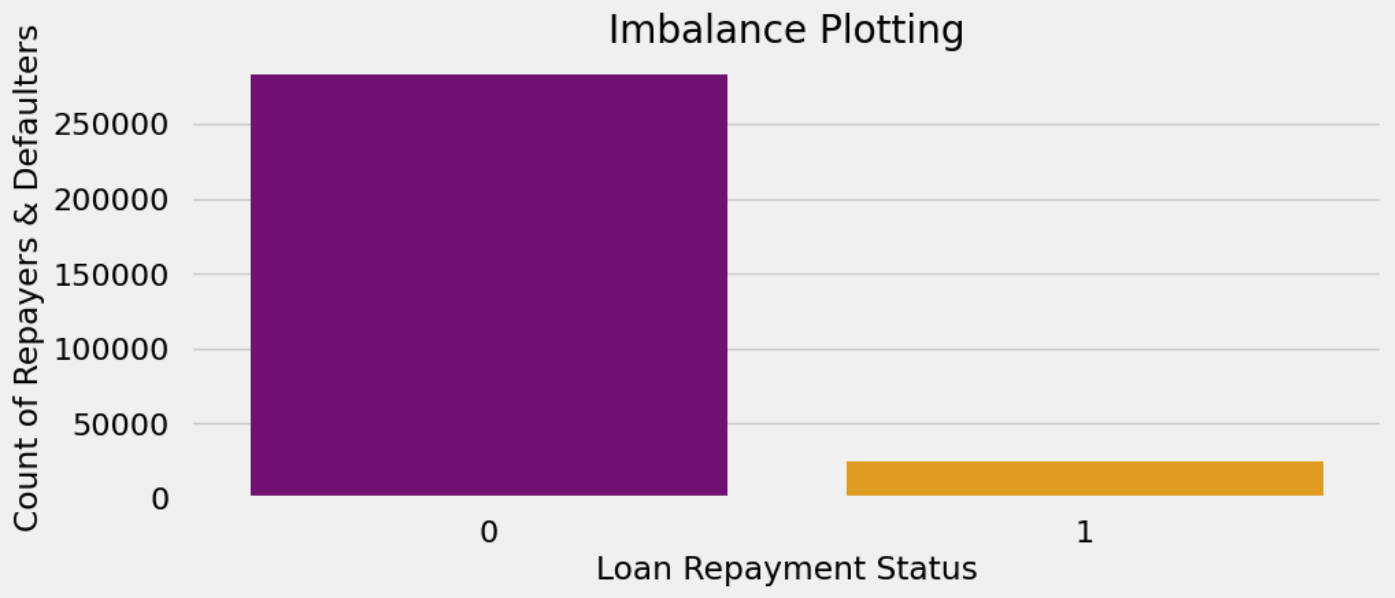
B. Previous_Application:

1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
2. CNT_PAYMENT has few outlier values.
3. SK_ID_CURR is an ID column and hence no outliers.
4. DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken
long back.



- Identify if there is data imbalance in the data. Find the ratio of data imbalance.

1. This data is highly imbalanced as number of defaulters is very less in total population.
2. Data Imbalance Ratio with respect to Repayment and Default: 11.39 : 1 (approx.)



- **Results of Univariate, Segmented Univariate, and Bivariate Analysis in Business Terms:**

1. **Gender:**
   Males have a higher chance of defaulting on loans compared to females, with a default rate of approximately 10% for males and 7% for females.
2. **Car Ownership:**
   There is no correlation between car ownership and loan repayment. Both clients who own a car and those who don't have a similar default percentage.
3. **Real Estate Ownership:**
   Owning real estate does not have a significant impact on loan default. The default rates for clients who own real estate and those who don't are approximately the same.
4. **Housing Type:**
   The majority of people live in houses/apartments. This information provides an understanding of the living arrangements of the clients.
5. **Living in Office Apartments:**
   Clients living in office apartments have the lowest default rate, indicating a lower risk of loan default for this specific living arrangement.
6. **Living with Parents and Rented Apartments:**
   Clients living with parents and in rented apartments have a higher probability of defaulting on loans compared to other housing types.
7. **Marital Status:**
   Most loan applicants are married, followed by single/not married and civil marriage. This information provides insights into the marital status distribution among loan applicants.
8. **Marital Status and Loan Default**:
   Civil marriage has the highest percentage of loan default (10%), while widow has the lowest default rate.
9. **Education Level:**
   The majority of clients have secondary/secondary special education, followed by higher education. Only a small number of clients have an academic degree.
10. **Education Level and Loan Default:**
    Clients with lower secondary education have the highest default rate (11%), while clients with academic degrees have a default rate of less than 2%.

11. **Income Type:**
    Most loan applicants have working income, followed by commercial associates, pensioners, and state servants.
12. **Income Type and Loan Default:**
    Applicants with maternity leave income and unemployed status have the highest default rates, while other income types have default rates below the average of 10%.
13. **Student and Businessmen:**
    Students and businessmen have no default records, indicating that these two categories are the safest for providing loans.
14. **Region Rating:**
    Most loan applicants are from Region Rating 2.
15. Region Rating and Loan Default:
    Region Rating 3 has the highest default rate (11%), indicating higher risk in this region.
16. **Occupation Type:**
    The majority of loans are taken by laborers, followed by sales staff. IT staff members take the lowest amount of loans.
17. **Occupation Type and Loan Default:**
    Low-skill laborers, drivers, waiters/barmen staff, security staff, and cooking staff have the highest percentages of non-repayment. Self-employed individuals also have a relatively high defaulting rate.
18. **Organization Type:**
    Organizations in the transport industry (type 3), industry (type 13 and 8), and restaurants have higher default rates. Providing loans to self-employed people in these industries may carry higher risks.
19. **Organization Type and Loan Default:**
    Business Entity Type 3 has the highest number of loan applications.
20. **Document Submission:**
    There is no significant correlation between document submission and loan default. Both applicants who submitted document 3 and those who didn't have similar default percentages.
21. **Age Group:**
    Clients in the age group of 20-30 have a higher probability of defaulting, while clients above age 50 have a lower probability of defaulting.
22. **Employment Length:**
    Clients with 0-5 years of employment have the highest default rate (10%), while clients with 40+ years of experience have a default rate of less than 1%.
23. **Loan Amount:**
    The majority of loans provided are for amounts less than 900,000. Clients who receive loans for 300-600k tend to default more than others.
24. **Annual Income:**
    90% of the loan applicants have an annual income less than 300,000. Applicants with lower incomes have a higher probability of defaulting, while those with incomes above 700,000 are less likely to default.
25. **Children:**
    Clients with more than 4 children have a very high default rate, with child counts of 9 and 11 showing a 100% default rate. Having more children increases the risk of loan default.
26. **Family Members:**
    Similar to the number of children, having more family members increases the risk of loan default.
27. **Businessman's Income:**
    Businessmen have the highest income, with an estimated range indicating income levels between approximately 4 lakhs and slightly above 10 lakhs. This information provides insights into the income range of businessmen.

Overall, these analysis results help businesses understand the relationships between various factors and loan default. They can inform risk assessment strategies, loan approval decisions, and the development of targeted marketing and intervention initiatives to minimize loan default risks and optimize lending practices.

- **Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable).**

The top 10 correlation for the Client with repayment:

1. Credit amount is highly correlated with amount of goods price, loan annuity, total income
2. We can also see that repayment have high correlation in number of days employed.

In [101]:
```
# Getting the top 10 correlation for the Repayers data
corr_repayer = Repayer_df.corr()
corr_repayer = corr_repayer.where(np.triu(np.ones(corr_repayer.shape),k=1).astype(np.bool))
corr_df_repayer = corr_repayer.unstack().reset_index()
corr_df_repayer.columns =['VAR1','VAR2','Correlation']
corr_df_repayer.dropna(subset = ["Correlation"], inplace = True)
corr_df_repayer["Correlation"]=corr_df_repayer["Correlation"].abs()
corr_df_repayer.sort_values(by='Correlation', ascending=False, inplace=True)
corr_df_repayer.head(10)
```

Out[101]:

	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
71	AMT_ANNUITY	AMT_CREDIT	0.771309
167	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
70	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
47	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
138	DAYS_BIRTH	CNT_CHILDREN	0.336966
190	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

- **The top 10 correlation for the Client with default:**

1. Credit amount is highly correlated with amount of goods price which is same as repayments.
2. But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayment(0.77).
3. We can also see that repayment have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
4. There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayment.
5. Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayment.
6. There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayment (0.254).

In [103]:
```
# Getting the top 10 correlation for the Defaulter data
corr_Defaulter = Defaulter_df.corr()
corr_Defaulter = corr_Defaulter.where(np.triu(np.ones(corr_Defaulter.shape),k=1).astype(np.bool))
corr_df_Defaulter = corr_Defaulter.unstack().reset_index()
corr_df_Defaulter.columns =['VAR1','VAR2','Correlation']
corr_df_Defaulter.dropna(subset = ["Correlation"], inplace = True)
corr_df_Defaulter["Correlation"]=corr_df_Defaulter["Correlation"].abs()
corr_df_Defaulter.sort_values(by='Correlation', ascending=False, inplace=True)
corr_df_Defaulter.head(10)
```

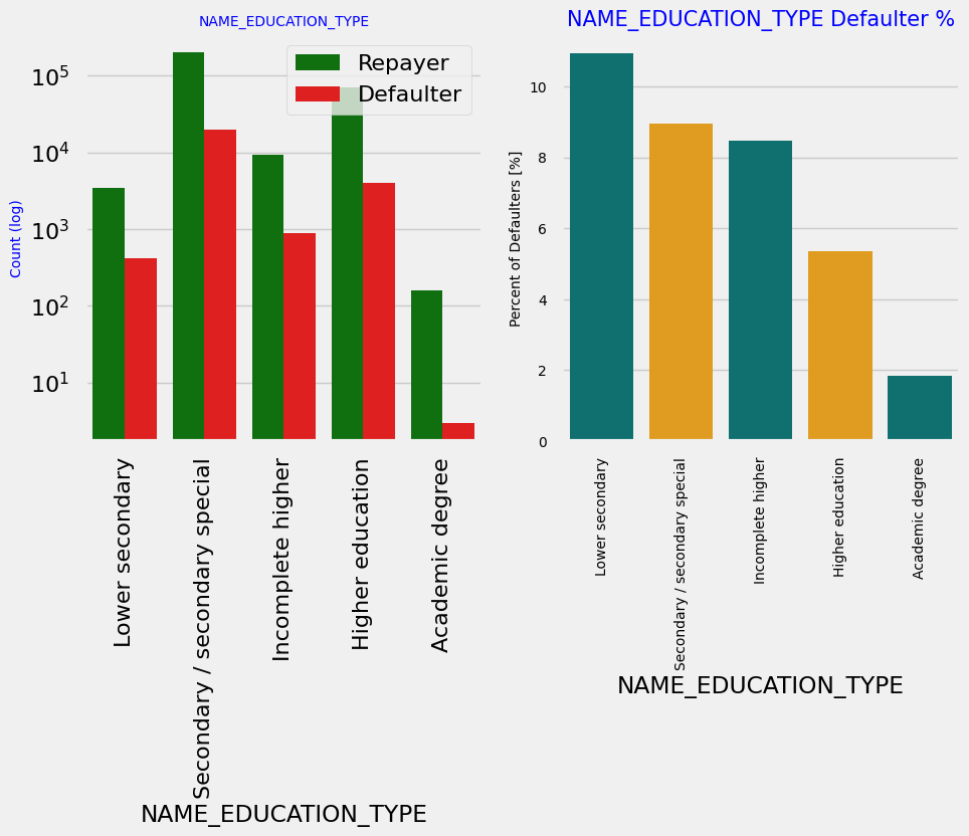
Out[103]:

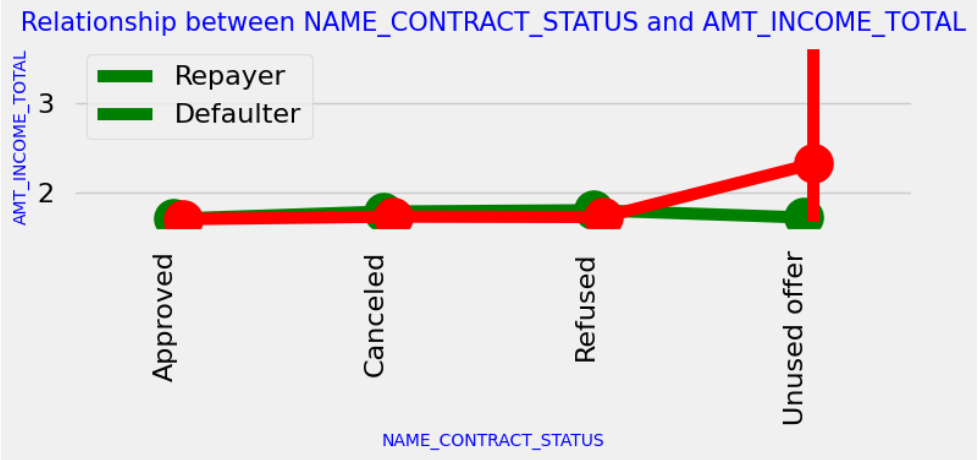
	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
71	AMT_ANNUITY	AMT_CREDIT	0.752195
167	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
190	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
375	FLAG_DOCUMENT_3	DAYS_EMPLOYED	0.272169
335	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
138	DAYS_BIRTH	CNT_CHILDREN	0.259109
213	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863

- **Analysis Insights and Key Findings on Loan Default Risk:**

Akshay Panchal

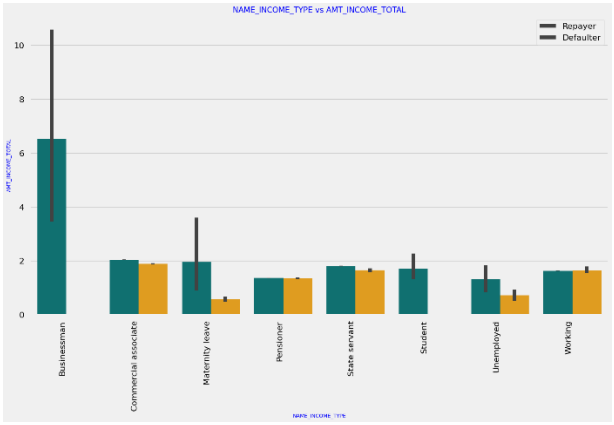
TRAINITY FINAL PROJECT-02

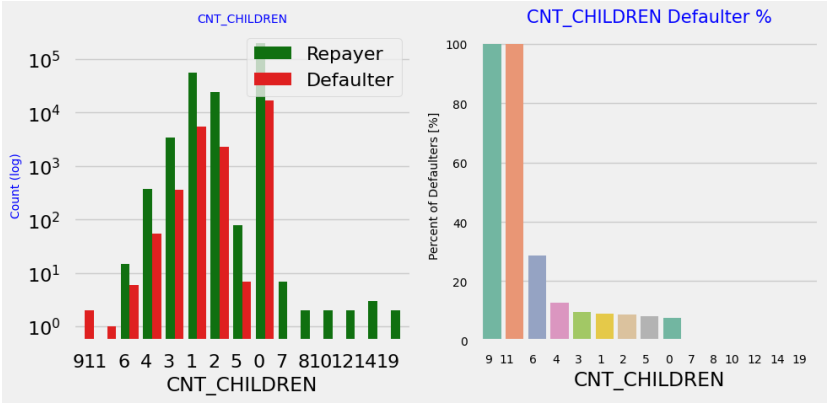


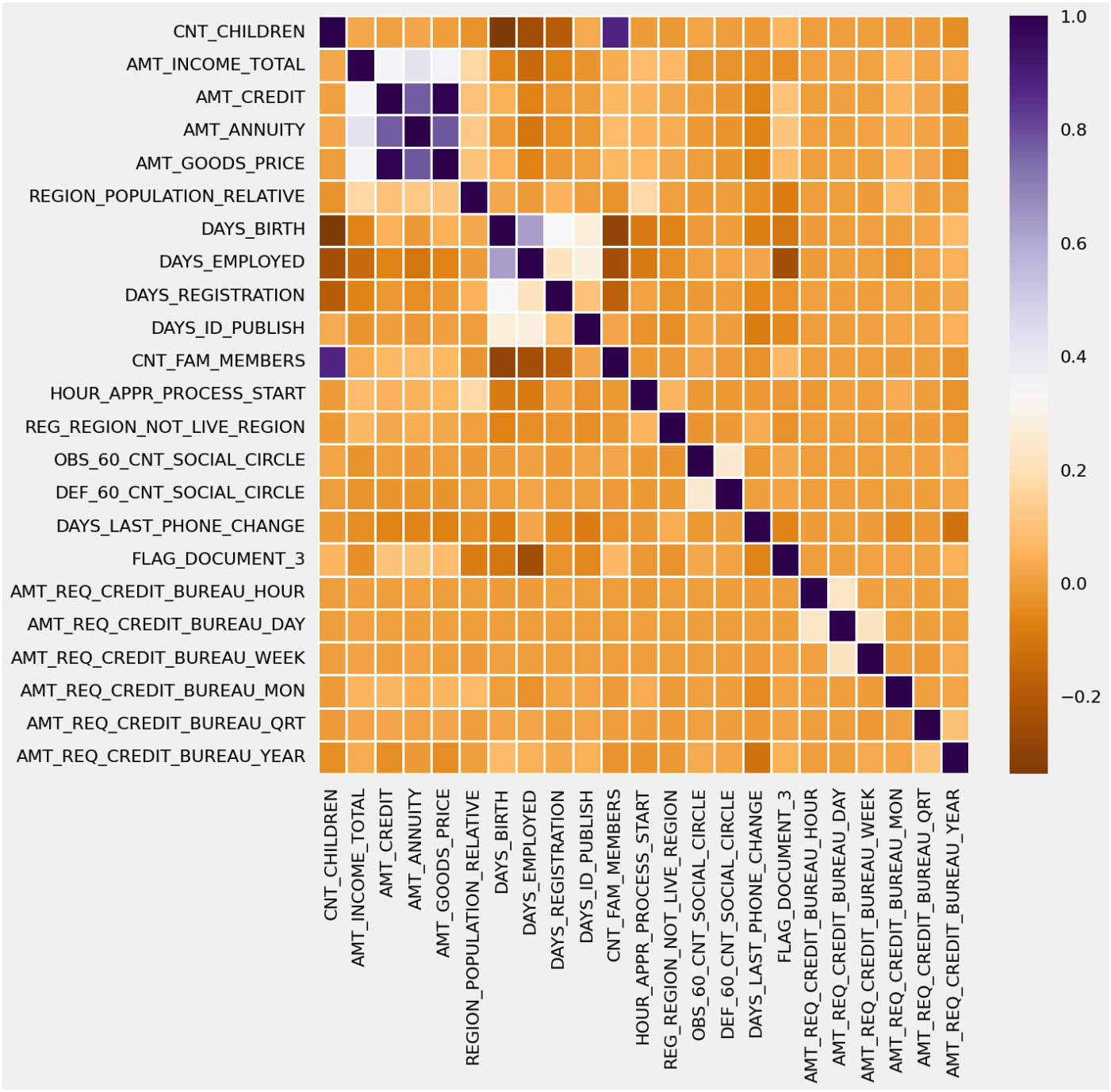


Akshay Panchal

TRAINITY FINAL PROJECT-02







- **Insights:**

Decisive Factors whether an applicant will Repay:

1. NAME_EDUCATION_TYPE: Applicants with an academic degree have lower default rates, indicating higher financial stability and repayment capability.
2. NAME_INCOME_TYPE: Students and businessmen have no default records, suggesting they are reliable borrowers.
3. REGION_RATING_CLIENT: Region Rating 1 is associated with lower default rates, making it a safer region for loan approval.
4. ORGANIZATION_TYPE: Clients working in Trade Type 4 and 5, and Industry Type 8 organizations have default rates below 3%, indicating their financial stability.
5. DAYS_BIRTH: Applicants above the age of 50 have a lower probability of defaulting, suggesting they have more stable financial situations.
6. DAYS_EMPLOYED: Clients with 40+ years of work experience have a default rate of less than 1%, indicating their financial reliability.
7. AMT_INCOME_TOTAL: Applicants with an annual income exceeding 700,000 are less likely to default, indicating higher financial capacity.
8. NAME_CASH_LOAN_PURPOSE: Loans taken for hobbies or buying garages have a higher likelihood of being repaid, indicating a stronger commitment to repayment.
9. CNT_CHILDREN: Clients with zero to two children have a higher probability of loan repayment, while those with more children have a higher default rate, suggesting additional financial responsibilities may impact repayment ability.

These insights can assist in refining loan approval criteria, identifying lower-risk borrowers, and developing strategies to minimize default rates.

- **Decisive Factors whether an applicant will Default:**

Decisive Factors for Loan Default:

1. Gender:
   Male applicants have a relatively higher default rate compared to females.
2. Marital Status:
   Applicants with civil marriage or single status have a higher likelihood of defaulting.
3. Education:
   Applicants with lower secondary and secondary education levels are more likely to default.
4. Income Type:
   Clients on maternity leave or unemployed have a higher default rate.
5. Regional Rating:
   Applicants living in regions with Rating 3 have the highest default rates.
6. Occupation Type:
   Avoid approving loans for low-skill laborers, drivers, waiters/barmen staff, security staff, laborers, and cooking staff due to their high default rates.
7. Organization Type:
   Organizations in the transport (type 3), industry (type 13 and 8), and restaurant sectors have high default rates. Caution is advised when approving loans for self-employed individuals, and higher interest rates can be considered to mitigate the risk of defaulting.
8. Age:
   Young applicants in the age group of 20-40 have a higher probability of defaulting.
9. Employment Duration:
   Applicants with less than 5 years of employment experience have a higher default rate.
10. Number of Children and Family Members:
    Clients with 9 or more children have a 100% default rate and should have their applications rejected.
11. Loan Amount:
    Default rates increase when the credit amount exceeds 3 million.

These factors can help identify applicants who are at higher risk of defaulting and inform decision-making processes, such as setting stricter criteria or offering loans at higher interest rates to mitigate the risk associated with these factors.

- **Summary:**

This case study focused on the application of Exploratory Data Analysis (EDA) in a real business scenario within the banking and financial services industry. By utilizing EDA techniques, I gained valuable insights into the risk analytics involved in lending and minimizing the risk of financial loss.

Through this project, I developed a solid understanding of how data analysis is used to assess and mitigate risks associated with lending to customers. I successfully applied various techniques such as correlation analysis, identification of data imbalance, and detection of outliers to extract meaningful insights from a large dataset.

One of the key challenges in this project was handling the complexity of the dataset and summarizing the results in a concise manner. By studying the correlations between different variables, I was able to extract and present the most crucial insights for the clients' decision-making processes.

Overall, this case study provided a valuable learning experience in analysing and visualizing large datasets, identifying decisive factors, and summarizing important results for business stakeholders. It enhanced my skills in risk analytics and provided hands-on experience in tackling real-world challenges in the financial domain.

GitHub:

Excel Workbook: Bank Loan Case Study.xlsx