Credit Eda Assignment Portfolio

Business Objective:
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

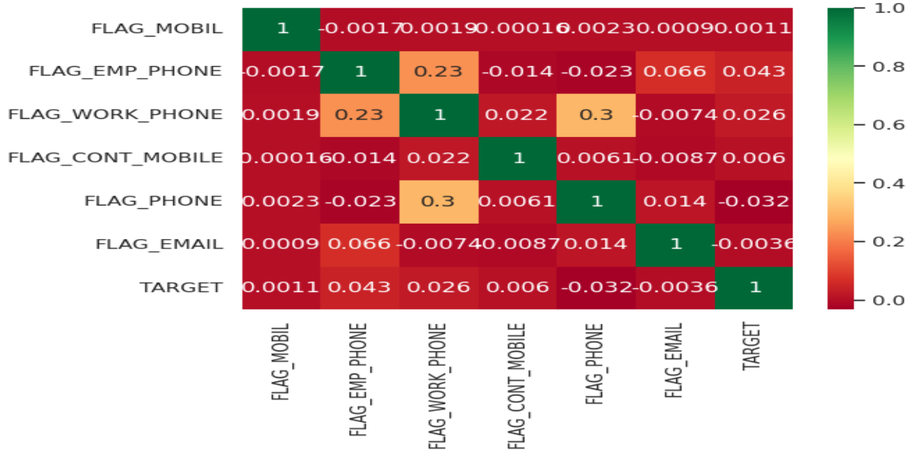
- **Approach and Methodology**
- Approach:
- Data Preprocessing:
- Performing data cleaning by handling missing values, outliers, and inconsistencies in the dataset.
- Transform and format the data as necessary to ensure it is suitable for analysis.

- Exploratory Data Analysis (EDA):
- In this we have to conduct a comprehensive exploration of the loan application data to gain insights and understand patterns.
- Analyze the distribution and statistical properties of variables using descriptive statistics.
- Visualize the data using charts, graphs, and plots to identify trends and relationships between the variables.
- Exploring the relationships between variables to understand potential factors influencing loan defaults.
- Variable Selection:
- Identify key variables that are strong indicators of loan default.
- Consider variables related to applicant profiles, loan characteristics, and payment history.

- Methodology:
- Descriptive Analysis:
- Analyze the distribution of variables such as loan amount, income, age, and education level.
- Calculate summary statistics, including mean, median, standard deviation, and percentiles.
- Identify any outliers or extreme values that may impact the analysis.
- Feature Engineering:
- Create new variables that capture important information for predicting defaults which can be done
- by binning, or creating interaction variables to enhance the quality of our analysis.

- Visualization Techniques:
- Performing Univariate analysis by utilizing visualizations like distplot, box plots for numerical columns and countplot for categorical columns to understand the distribution of variables.
- Performing Bivariate analysis between dependent
- and independent or Target variable using visualization techniques such as Scatter plot, Pair plot, HeatMap to understand relationship between variables.
- Explore visual patterns to identify potential factors associated with loan defaults.

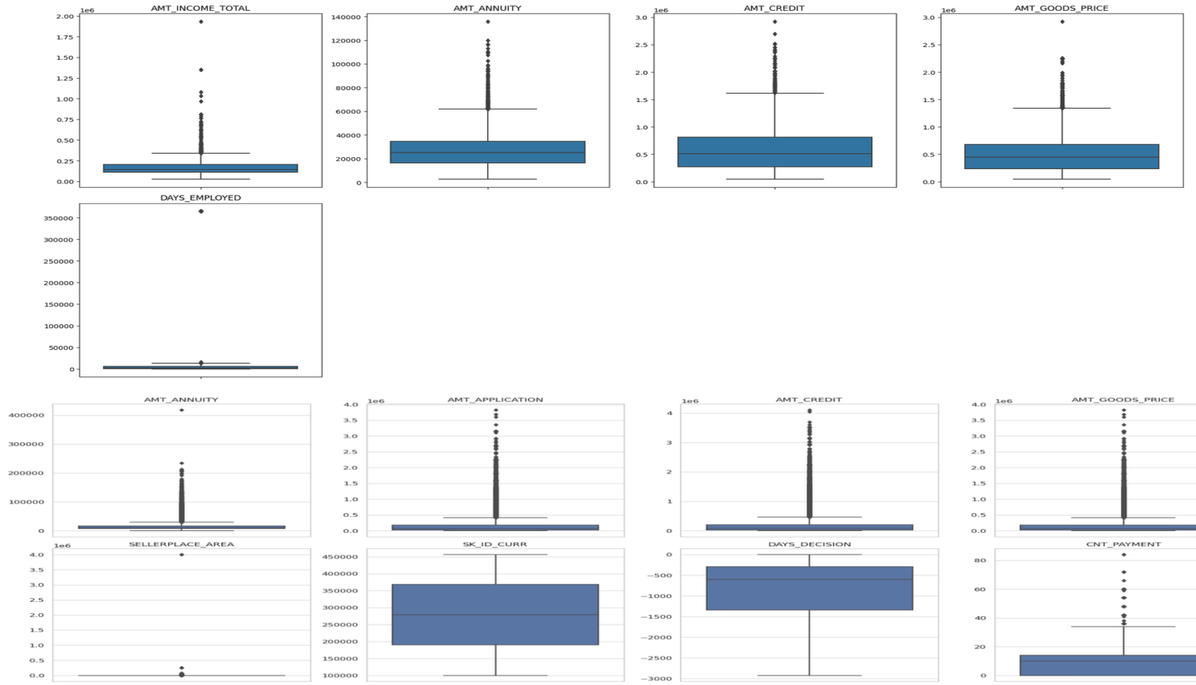
- Correlation Analysis:
- Calculate correlation to measure the strength and direction of relationships between variables.
- Identify variables strongly correlated with loan defaults. we can then use heatmap to understand correlation between variables.
- Merging the two datasets and performing analysis with important variables.

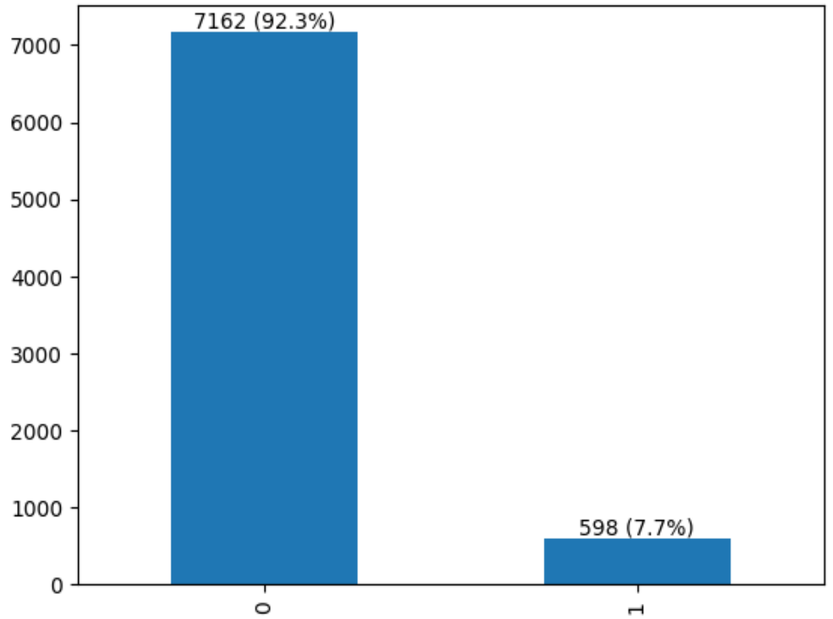


1.Since EXT_SOURCE_1,EXT_SOURCE_2 ,EXT_SOURCE_3 have no correlation with target we will delete these col.

2.From the above graph we can understand that client who has applied for loan has not submitted the document except flag document 3 so except flag document 3 we will delete all the other flag documents.

3.There is no correlation between contact columns and loan repayment so we can delete them.





1. Insight:

. It can be seen that in application data

AMT_INCOME_TOTAL has huge number of outliers which indicate that some of the loan applicants have high income compared to the others.

. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.

. DAYS_BIRTH has no outliers

.DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

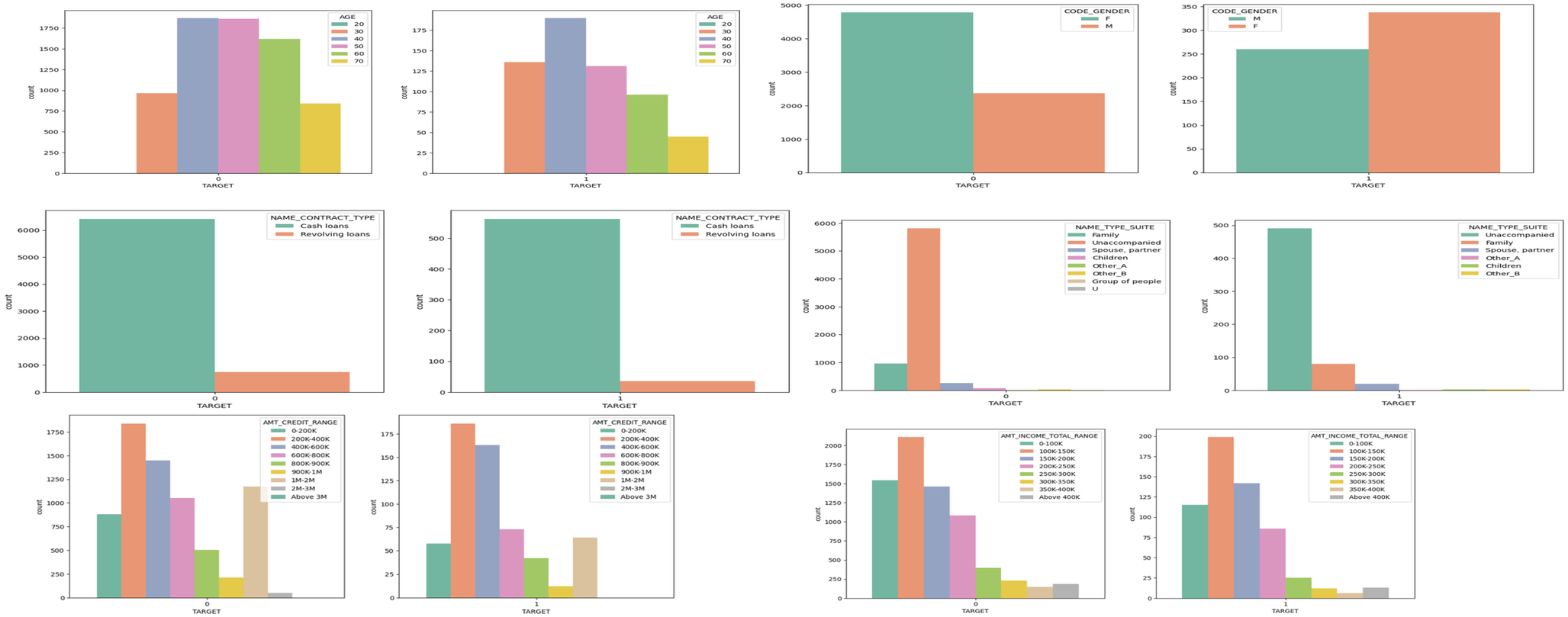
2. It can be seen in previous application data that:

. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.

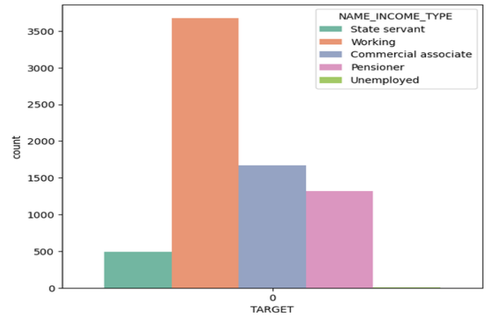
. SK_ID_CURR has no outliers.

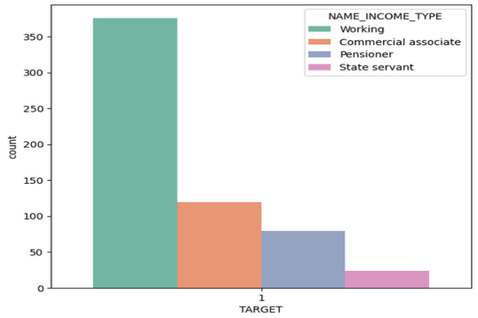
. CNT_PAYMENT and DAYS_DECISION has few outliers.

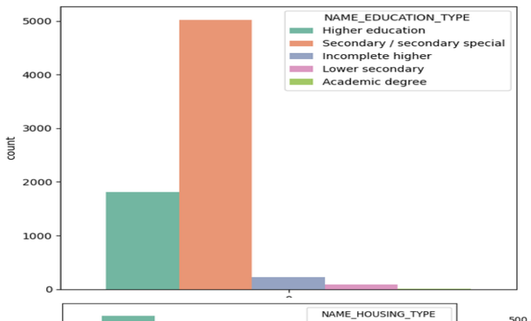
3. Fig 3 shows imbalance value and percentage of Target variable which shows that 92.3% of applicants are Repayers and 7.7% of applicants are defaulters.

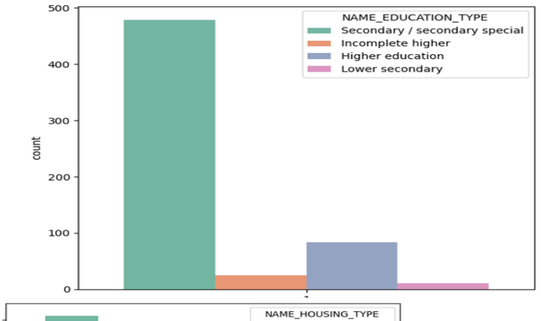


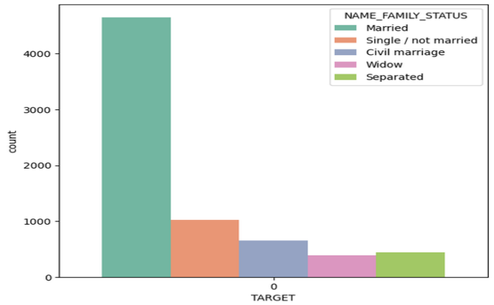
1. People of age 40 seems to apply higher than any other age people for loan in case of defaulter as well as repayer.
2. Female seems to apply loan more than male in case of defaulter as well as repayer.
3. Cash loan are applied by applicants in case of defaulter as well as repayer.
4. In case of repayer Unaccompanied type of applicant are most likely to apply for loan while in case of defaulter family type of people are most likely to apply for loan.
5. In case of repayer as well as defaulter amount credit for loan are most likely ranging between 200k to 400k.
6. In case of repayer as well as defaulter income amount of applicants most likely ranges between 100k to 150k.

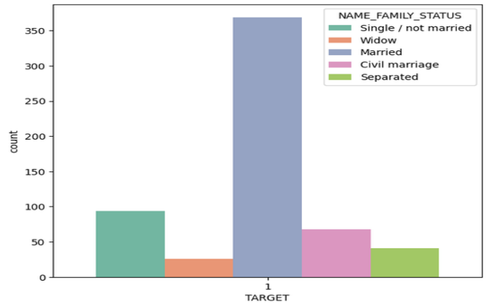


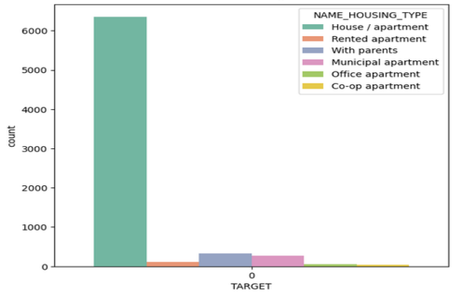


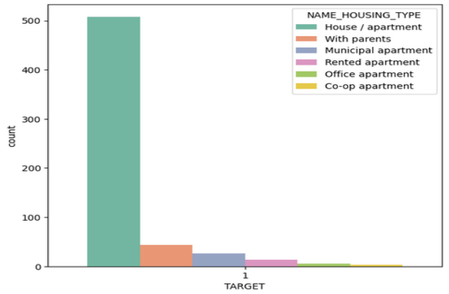


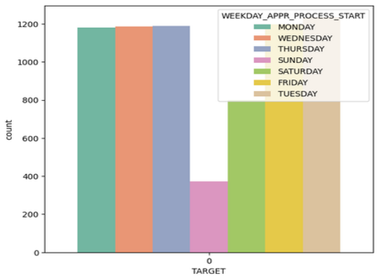


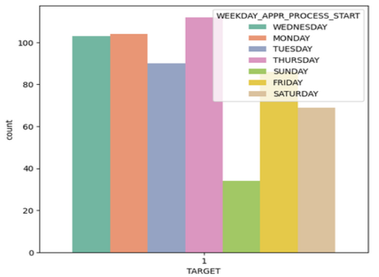












1. Name Income Type

Working, Commercial associate and pensioner are most likely to apply for loan.

working people are most likely to repay the loan and they are also most likely to become defaulter.

2. Name Education Type

Customers who are most likely to repay the loan as well as become defaulter has secondary/secondary special education qualification level.

3. Name Family Status

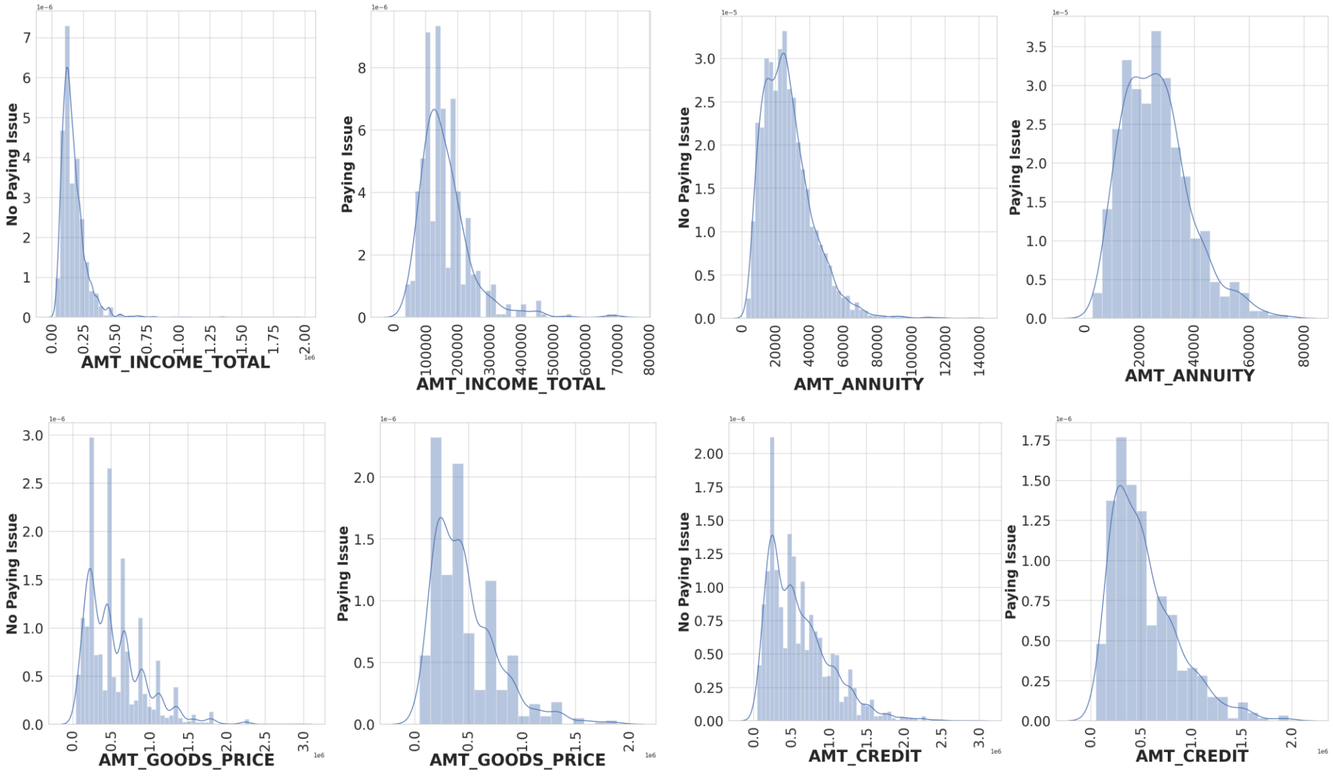
Married people seems to apply most loans in both cases Repayer as well as defaulter.

widow customers seems to have minimal risk to become defaulter.

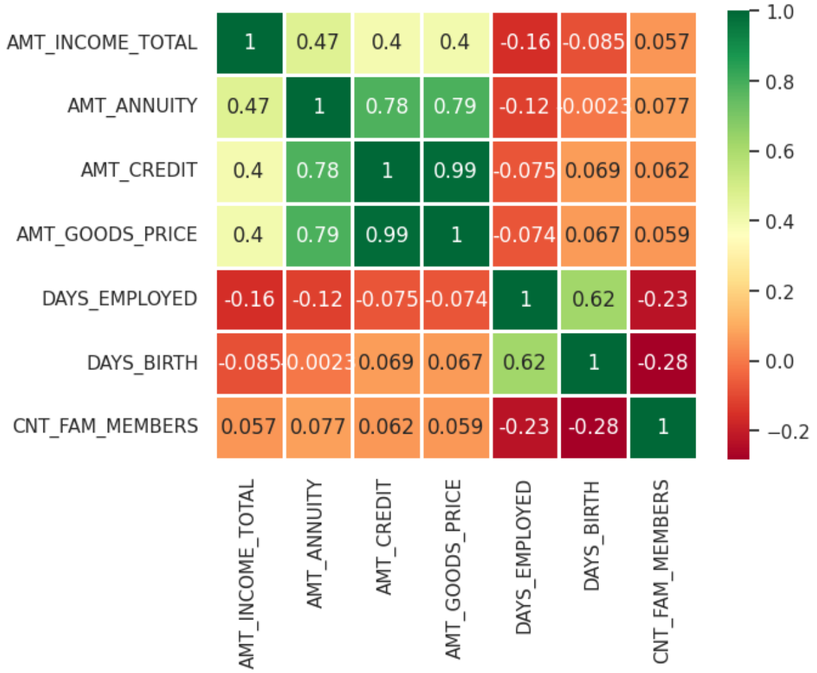
4. Name Housing Type

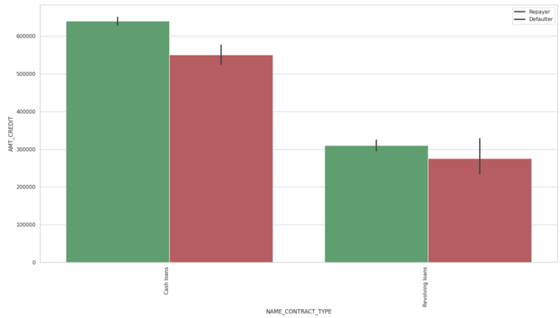
Most of customers who has applied for loan owns house/apartment in case of both Repayer and defaulter.

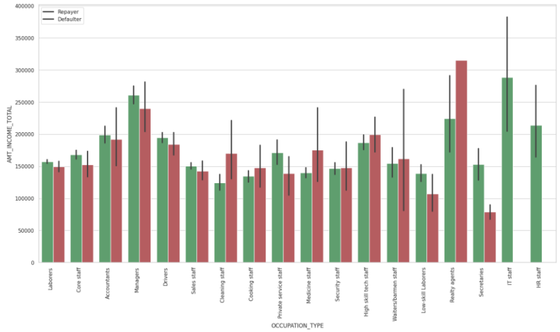
5. There is no major difference in days for both repayer and defaulter.

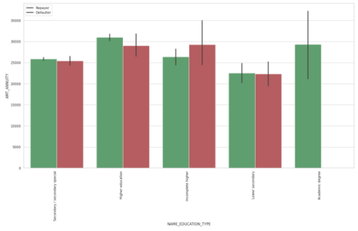


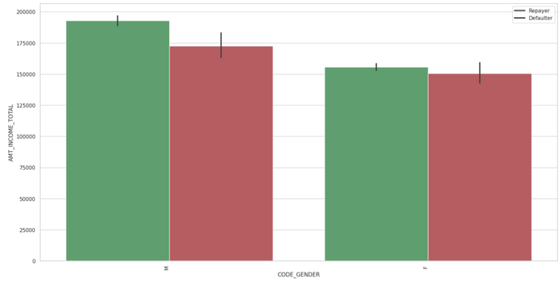
- People having no payment issue has high income compared to people having payment issue.
- People with no paying issue has taken high amount credit for the loan as compared people with paying issue.
- The dataset of amount annuity for people with no paying issue are much widely spread than dataset of paying issue.
- The shape of Amount goods price are quite equal for both paying issue and non paying issue people.











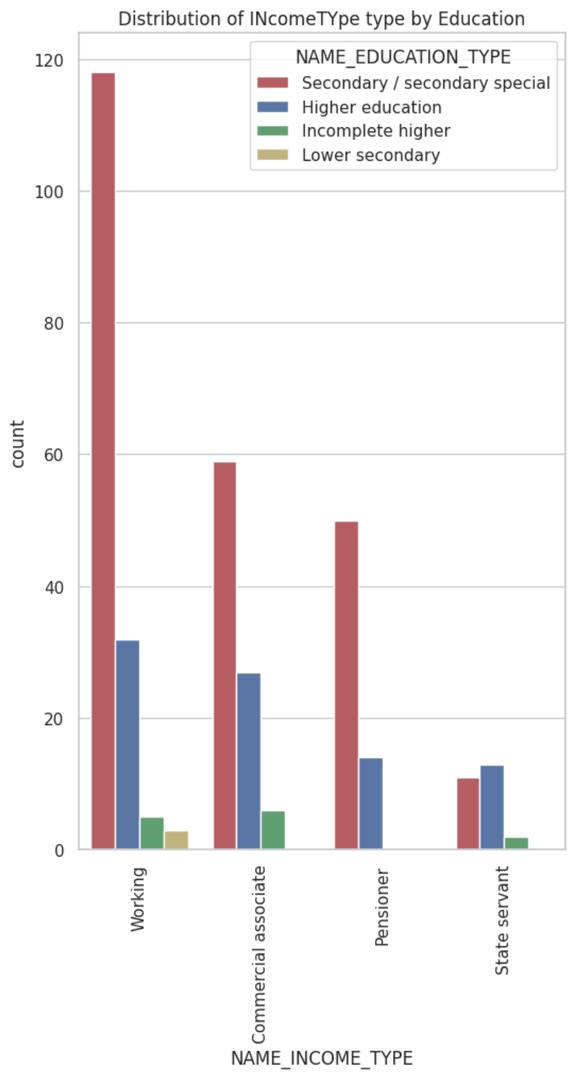
- Correlation among Repayers: Credit Amount is highly correlated with Amount Annuity, Amount Goods Price and Income of Repayers. We can see that repayers have high correlation count family members and number of days employed.
- Correlation among defaulters: Credit Amount is highly correlated with Amount Annuity which is little less than Repayer Amount Goods Price and Income is also less correlated with amount credit compared to repayers.Here days employed is more correlated than repayers.

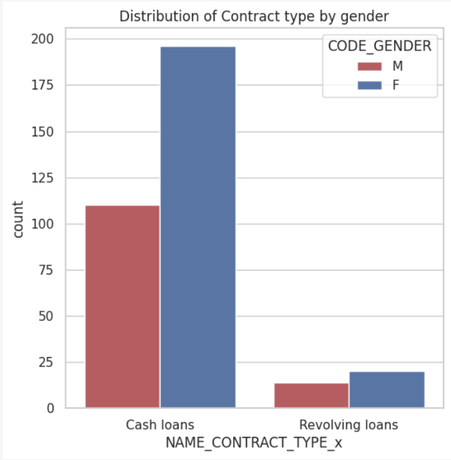
• 3. It can be seen that amount credited for cash loan is higher than revolving loan for both repayer and defaulter.

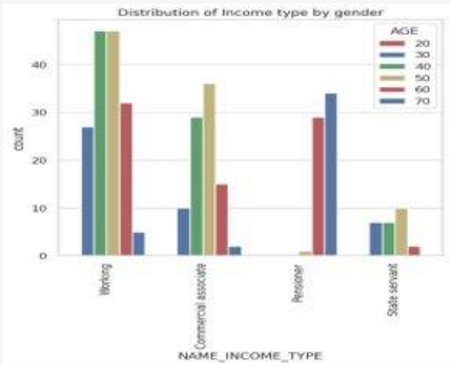
• 4. It can be seen that males applying for loan have higher income than female for both repayer and defaulter.

• 5. It can be seen that It employer applying for loan have higher income than other applicants and there are not single IT staff who are defaulter where as Realty agents have high default rate than any other employee.

• 6. People having academic degree have highest amount annuity than other education level group of people.







- Working People with age ranging between 40-50 have highest income than other applicants.
- Females are likely to apply for cash as well as revolving loan as compared to males.
- Working People having Secondary/Secondary special are likely to take loan compared to others

Conclusion

In this assignment, we applied Exploratory Data Analysis (EDA) techniques to analyze loan application data in the context of risk analytics in the banking and financial services industry. The objective was to identify patterns and variables that could serve as strong indicators of loan defaults, allowing us to make informed decisions and minimize the risk of losing money while lending to customers.

Through our EDA, we gained valuable insights into the loan application process and the associated risks. We discovered that the lack of credit history poses a challenge for loan providers, as it becomes difficult to assess the repayment capability of applicants. This leads to the advantage some consumers take by intentionally becoming defaulters.

Our analysis focused on distinguishing between clients with payment difficulties and those who make timely payments. By leveraging EDA techniques, we successfully identified key variables that significantly impact loan defaults. We conducted descriptive analysis, visualization techniques, comparative analysis, and correlation analysis to uncover important insights. Variables such as applicant profiles, loan characteristics, and payment behavior emerged as strong indicators of loan defaults.

Our findings can be utilized by the lending company to make informed decisions and minimize the risk of defaults. By incorporating the identified variables and patterns into their risk assessment and portfolio management strategies, they can effectively screen loan applicants and tailor loan terms to mitigate potential losses.

In conclusion, this assignment demonstrates the power of EDA in understanding risk analytics in the banking and financial services sector. By analyzing loan application data and identifying relevant variables, we have provided valuable insights to aid in decision-making and improve risk assessment practices.

Thank you for giving me the opportunity to showcase my creativity and knowledge through this project!