CREDIT EDA ASSIGNMENT

JALPA VATALIYA

Problem Statement

- Loan providers encounters challenges in providing loans to individuals who lack or have insufficient credit history, which can result in some borrowers defaulting on their loans
- The risks associated with loan approval decisions include the loss of potential business if a likely-to-repay applicant is dropped and the likelihood of financial loss if an unlikely-to-repay applicant is accepted
- A consumer finance company specializing in offering different types of loans to urban clients aims to analyze loan application data using EDA to detect patterns that can help conclude applicants capable of repaying the loan while avoiding defaults.
- The loan application data consist of information of payment difficulties, payment timing, and four types of decisions taken by the clients and the company, comprising approved, canceled, refused, and unused offers
- The primary **objective** of this case study is to determine the **influence of consumer and loan attributes on the likelihood of defaulting** and utilize the insights gained through EDA to identify critical factors that can enhance the loan approval process and reduce defaults
- Here, we have two types of datasets available:
    - **Application data**: The data contains information about the clients at the time of application
    - **Previous Application data**: The data provides details about their previous loan applications
- The analysis aims to determine the influence of consumer and loan attributes on the likelihood of defaulting, with the ultimate goal of enhancing the loan approval process and reducing defaults

Objective

- This case study aims to use EDA to find patterns that indicate if a client may have difficulty paying their loan instalments, which can be used to take appropriate actions such as denying the loan, reducing the loan amount, or lending to risky applicants at a higher interest rate
- The primary goal of this analysis is to ensure that applicants who are capable of repaying the loan are not rejected
- Through this analysis, the company aims to identify the key driving factors or variables behind loan default, which can be utilized for portfolio and risk assessment
- By recognizing strong indicators of default, the company can better evaluate risk and make informed lending decisions, ultimately reducing the likelihood of loan defaults

Assumptions

- The loan providing companies are struggling to assess the creditworthiness of loan applicants, especially those with insufficient or non-existent credit history, which leads to a higher risk of loan defaults
- To address this challenge, we will identify the patterns that can predict the likelihood of loan default and to ensure that qualified applicants are not rejected
- Here, we will focus more on to current application data more as compared to the previous application data assuming that the new or applicants with less credit history can be potential defaulters than other group
- The company aims to identify key variables or factors that are strongly associated with loan default, which can help with portfolio and risk assessment
- By identifying strong predictors of default, the company can make informed lending decisions and manage risk more effectively, ultimately reducing the likelihood of loan defaults

Approach [1/3]

- Majorly we have utilized Application data to analyze various variables or combination of variables to identify clients who may have difficulty paying their loan instalments
- We will use previous application data to a lower extent
- Below are the steps we have followed during this exercise:
  1. **Importing libraries and Reading out files**
  2. **Data cleaning**:
     I. ***Missing value treatment***:
        - Removal of all the columns which have more than 40% null values e.g., OWN_CAR_AGE column has approx. 66% missing values hence it was removed
        - The columns which have less than 40% missing values were treated differently based on the data provided
        - The columns with higher missing values (below 40% null values) and can skew the data by imputing it with the new data, were created as different categories e.g., OCCUPATION_TYPE columns has 96391 null data entries and it can skew the data if we replace with the mode hence categorized it with "Others“
        - Some numerical and categorical data columns were treated with mean, median and mode based on the data provided. Some columns left untreated and decisions have been taken depending upon the analysis, with the reasons mentioned in the python notebook

Approach [2/3]

II. ***Outlier treatment***:
   - The identification of outliers were done through quantile range and boxplots
   - The columns which have numerical data were mostly treated with mean, median and mode depending upon the situation e.g. YEARS_EMPLOYED columns having applicants with 1000 years of employment
   - While the categorical columns were treated with mode e.g., CODE_GENDER has “XNA” which was imputed with the mode

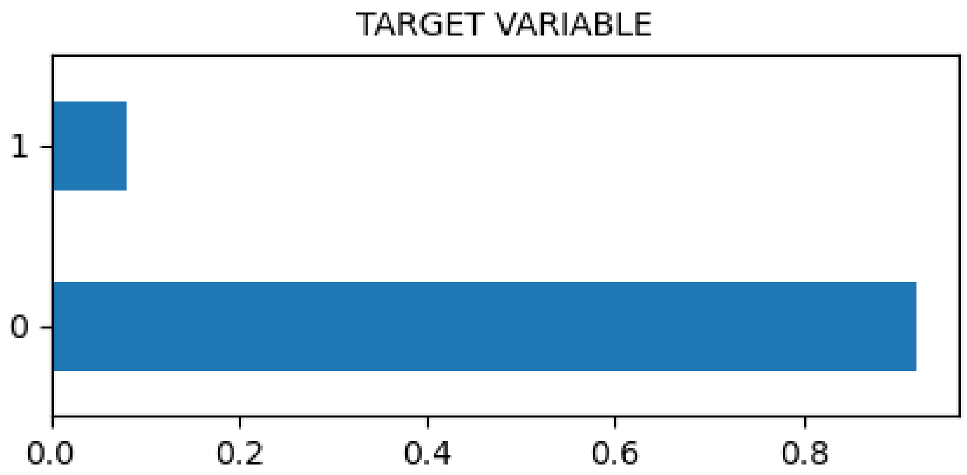
3. **Standardization of data**:
   - The numerical columns which were standardized based on the requirement e.g., DAYS_BIRTH and all the DAYS were converted to Years for easier analysis
   - The categorical data which were provided as flags were standardized to analyze better during bivariate or multivariate analysis. E.g., FLAG_OWN_CAR and FLAG_OWN_REALTY columns’ data were changed into 0 and 1 from “N” and “Y” respectively
4. **Dropping extraneous columns**:
   - The columns which had low relevance during the analysis were removed from the data. E.g., 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5', 'FLAG_DOCUMENT_6’, etc.
5. **Segmentation of data**:
   - At first, imbalance in the dataset was identified. Then, the segmentation of data was done based on target variables i.e. Applicants with payment difficulties and Applicants with payments on time

Approach [3/3]

6. **Analysis of data**:
   - **Univariate, bivariate and segmented univariate/bivariate** analyses were performed on loan application data to identify variables and combinations of variables that are associated with payment difficulties, and to examine the correlation between different variables
   - Exploratory data analysis was conducted on the combined application and previous application data to investigate how previous loan history affects the likelihood of loan defaults

Imbalance of Target Variable

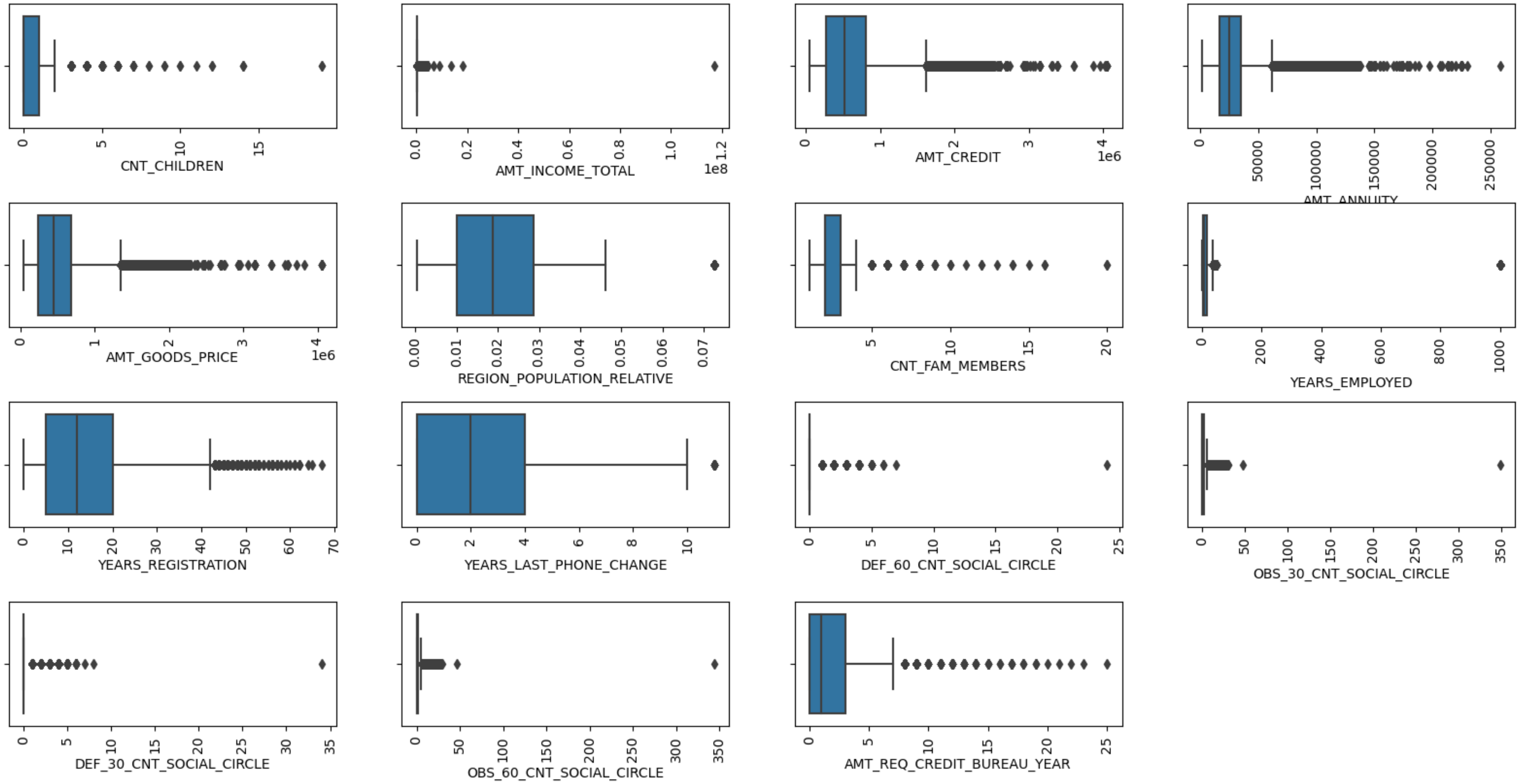
- Data imbalance was observed for Target variable
- We can see the same in the below graph;



- Based on the bar plot, we can determine that only 8% applicants with payment difficulties (he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample) while 92% applicants are the other cases
- There is a great imbalance in the Target variable

Outlier treatment [1/2]

- Based on the 95th, 99th quantile and boxplots, we can identify the outliers for the numerical cols as mentioned below;



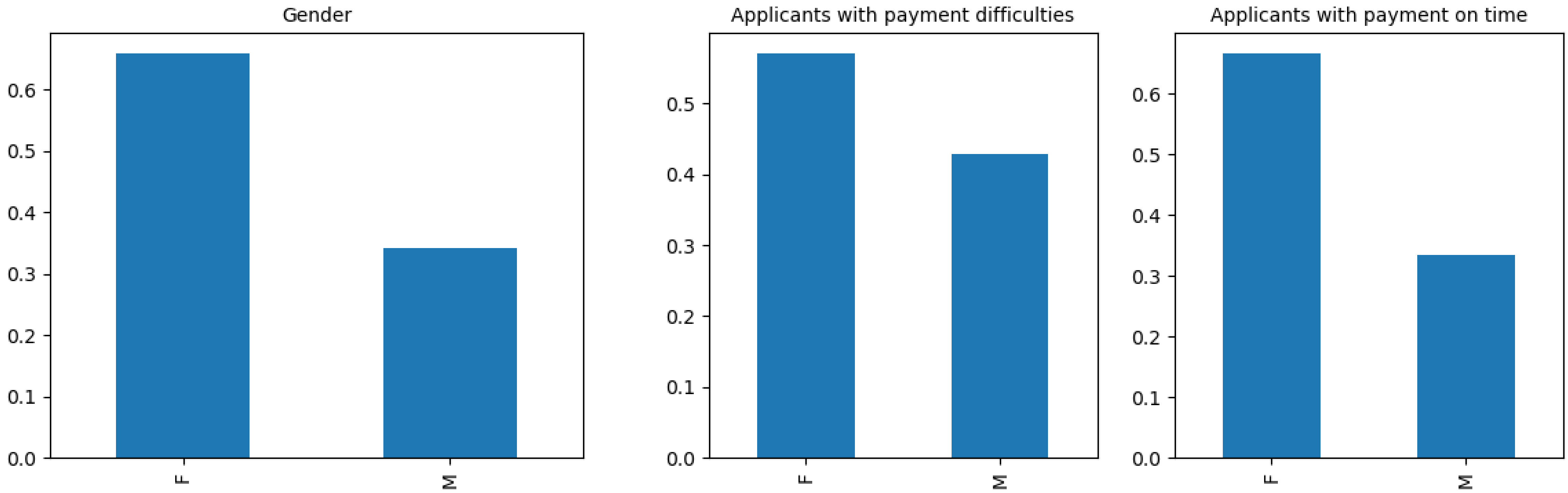
Outlier treatment [2/2]

1. **CNT_CHILDREN** : Majority of the applicants having less than 3 children while only 0.18% applicants is having children more 3 which are the outliers. Although, we do not have to treat those outliers further as people can have more than 3 children
2. **AMT_INCOME_TOTAL** : Majority of the applicants have less than 337K total income which seems normal. Only 4.5% people are having above 337K income which considered as outliers but no treatment is needed as they may have more income
3. **AMT_CREDIT** : Majority of applicants have credited loan amount less than 1350K. Only 4.56% applicants have credited more than 1350K amount which considered as outliers. There is no need to take any action as people can credit more loan amount
4. **AMT_ANNUITY** : Approx 95% applicants have loan amount annuity below 53K. Applicants with above 53K amount annuity are considered as outliers as they may have higher loan amount
5. **AMT_GOODS_PRICE** : Applicants who have loan amount higher than 1305K are the outliers. Although, no further actions are required as they may have higher loan to pay
6. **CNT_FAM_MEMBERS** : Only 1.3% applicants have more than 4 family members which are considered as outliers. Although, there is no action needed as some applicants may have larger family
7. **YEARS_EMPLOYED** : 55352 Pensioners and 22 unemployed were having more than equals to 1000 years, we can replace the Pensioners years of employment with the median of Pensioners years of employment less than 1000 years, and we can consider zero years of employment for unemployed as they are smaller in number
8. **AMT_REQ_CREDIT_BUREAU Cols** : We can observe the outliers especially at Yearly col although we can ignore the outliers as some applicants may have more inquiries based on the loans taken

- REGION_POPULATION_RELATIVE, YEARS_REGISTRATION, YEARS_LAST_PHONE_CHANGE are the cols do not require additional outlier analysis in terms of the distribution of data. We can standardize the rest cols which have higher spread across the data

Univariate Analysis: Overall and segment level [1/9]

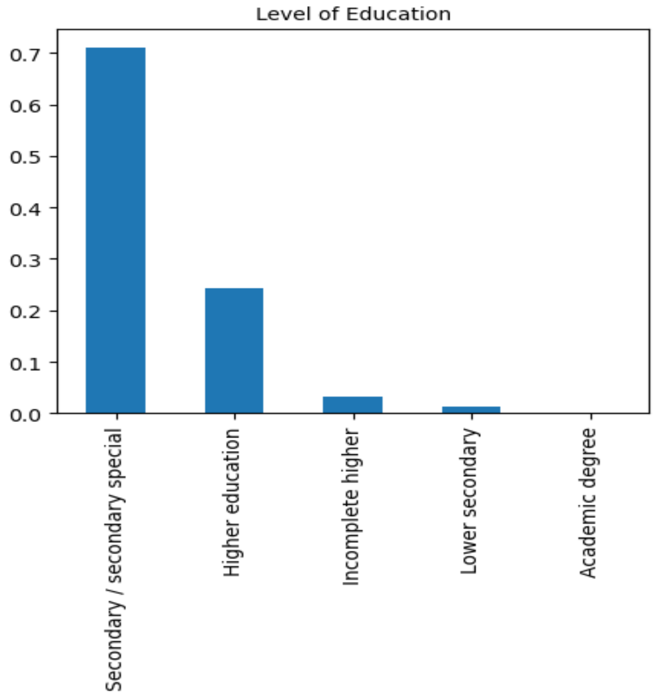
Gender:

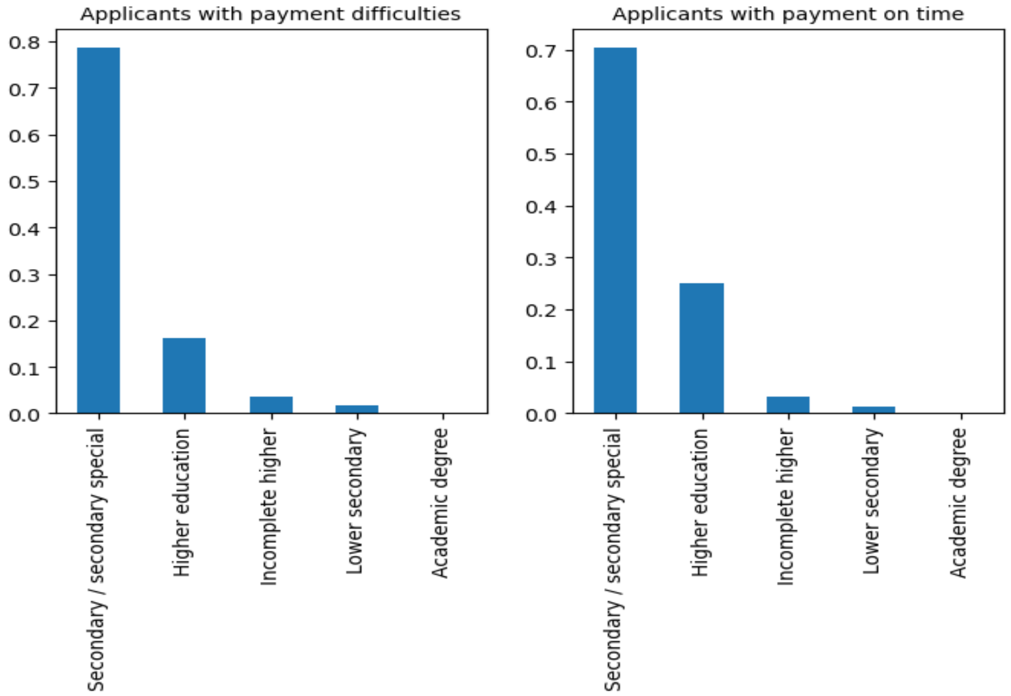


- Majority of the applicants (approx. 65%) are female at overall level
- Majority of the female applicants are regular on payments while male applicants finding difficulties with payments

Univariate Analysis: Overall and segment level [2/9]

Education:

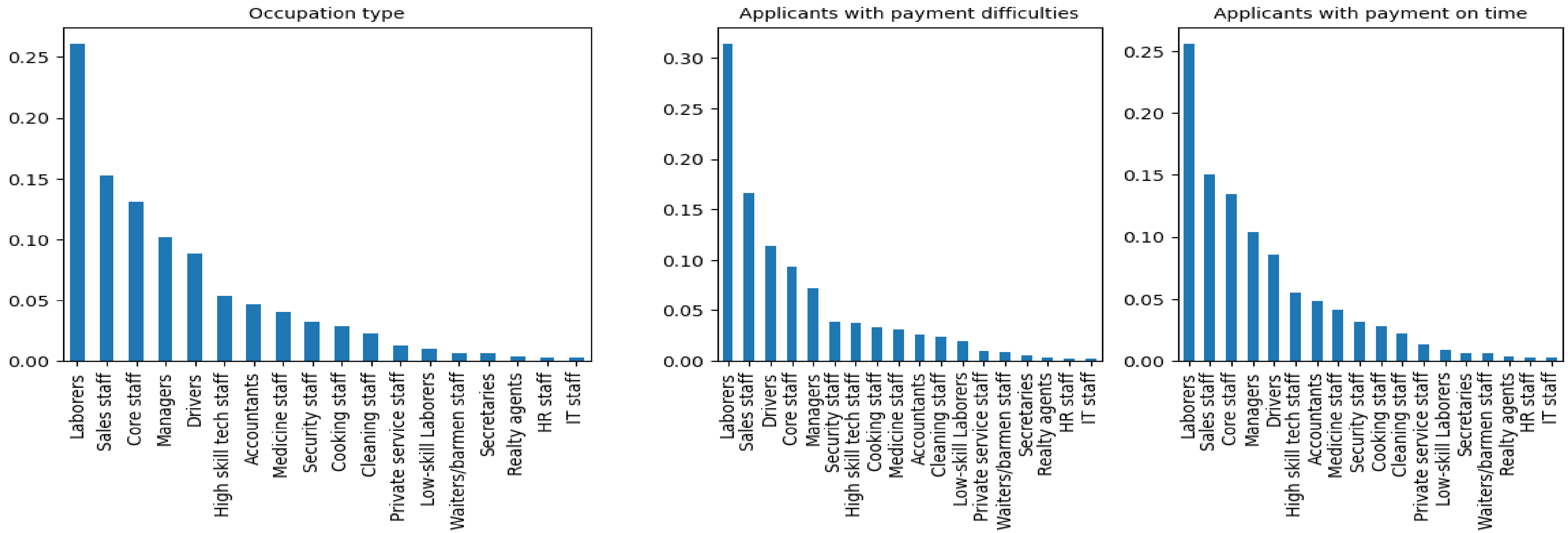




- Majority of the applicants achieved Secondary / secondary special level of education
- Majority of the working applicants who achieved Secondary / secondary special level of education facing difficulties with payments while applicants with Higher education level are able to pay on time

Univariate Analysis: Overall and segment level [3/9]

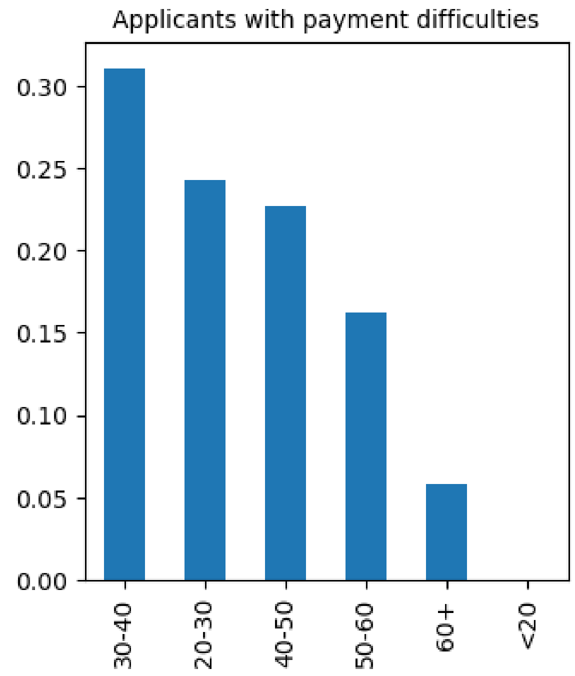
Occupation Type:

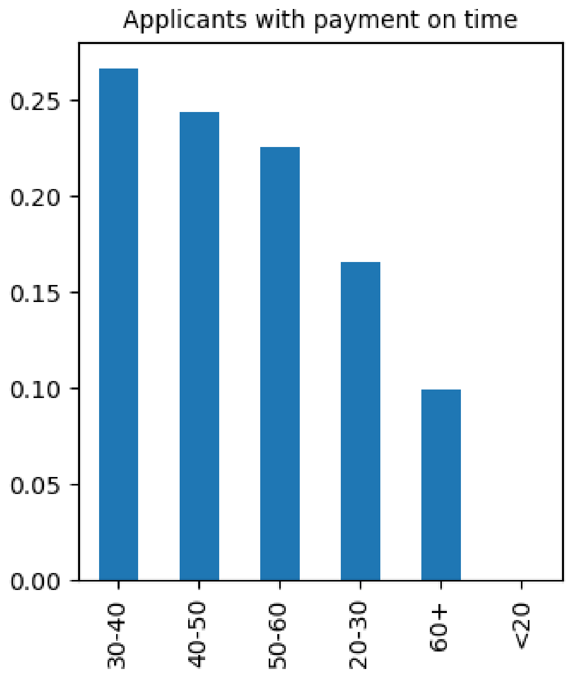


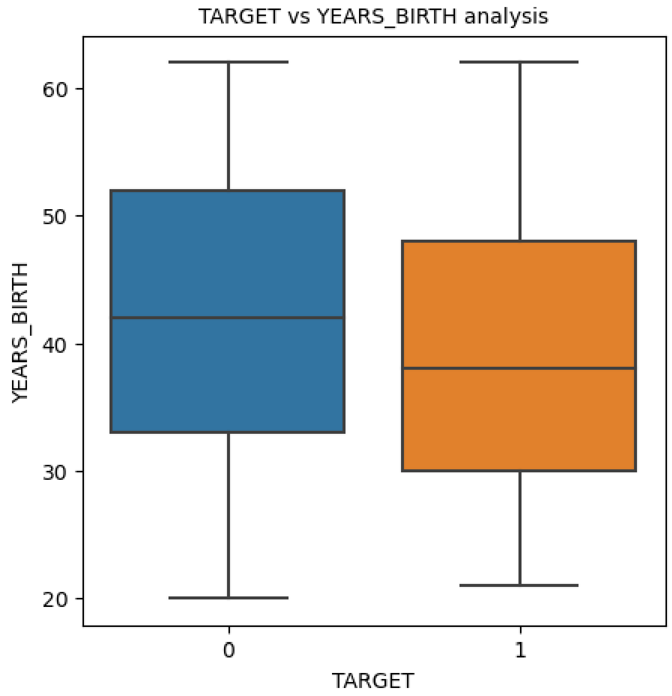
- Majority of applicants are Laborers followed by Sales staff and Core staff
- Laborers and Sales staff applicants are facing slightly more difficulties with payments than all the other cases

Univariate Analysis: Overall and segment level [4/9]

Age:


Applicants with payment difficulties

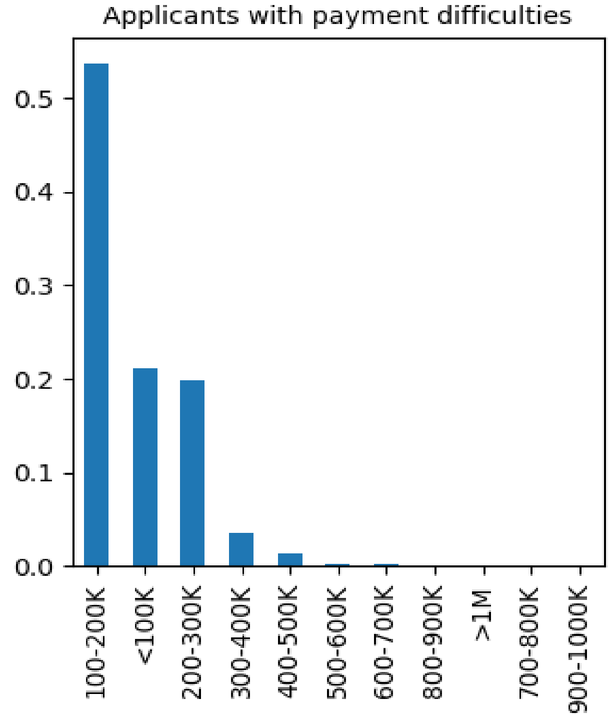

Applicants with payment on time

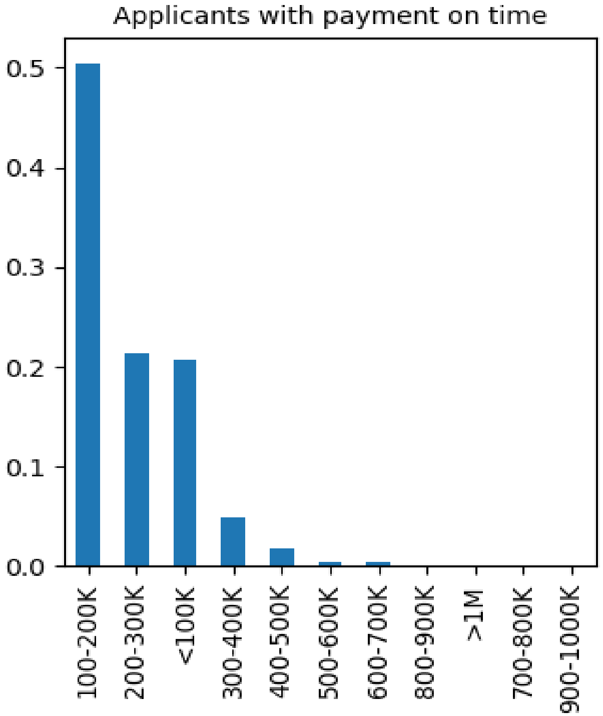

TARGET vs YEARS_BIRTH analysis

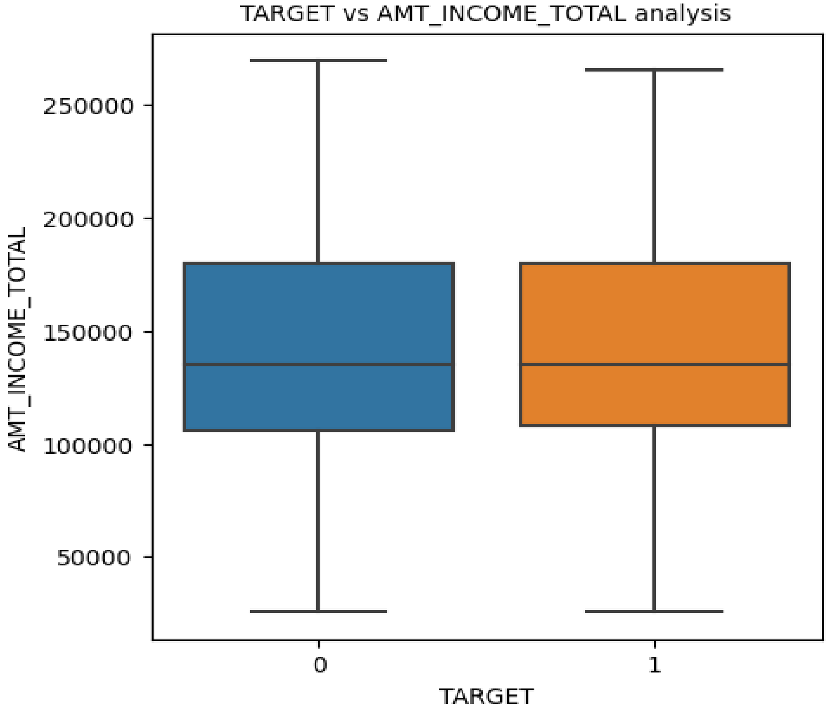
- Majority of the applicants are in 30 to 40 age group; Applicants who are in the age group of 30 to 40 are facing difficulty in payment
- There is a low median age of applicants for applicants with payment difficulties vs applicants who are able to pay on time

Univariate Analysis: Overall and segment level [5/9]

Income:



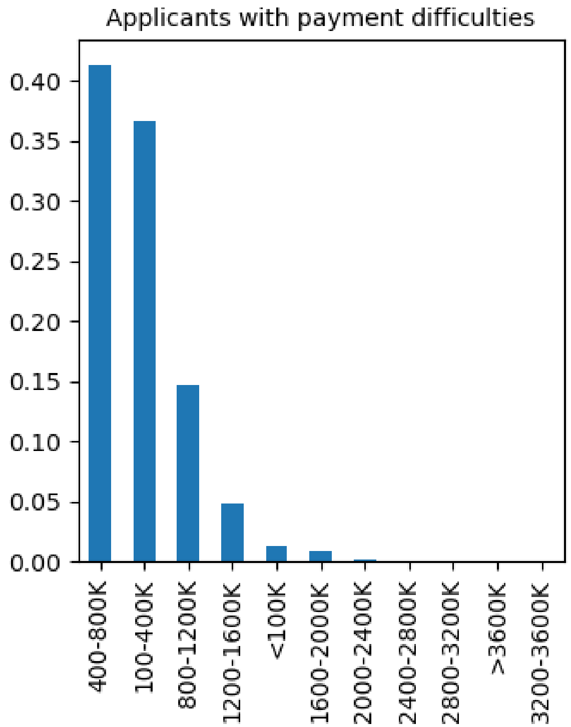


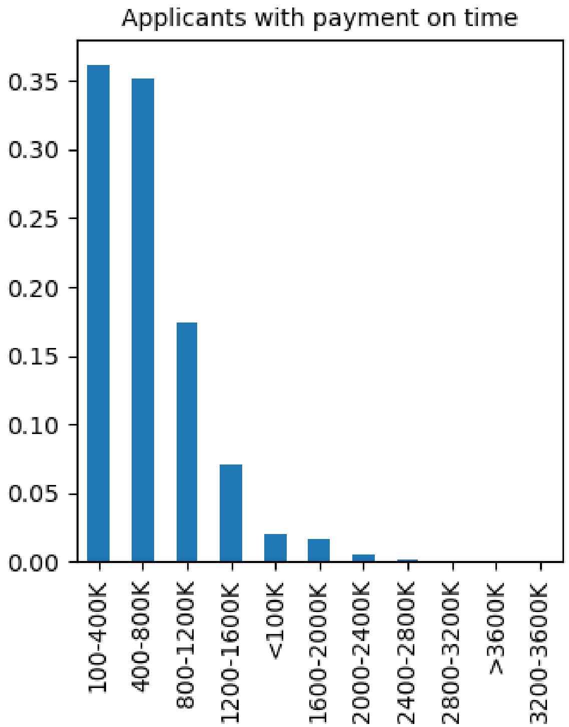


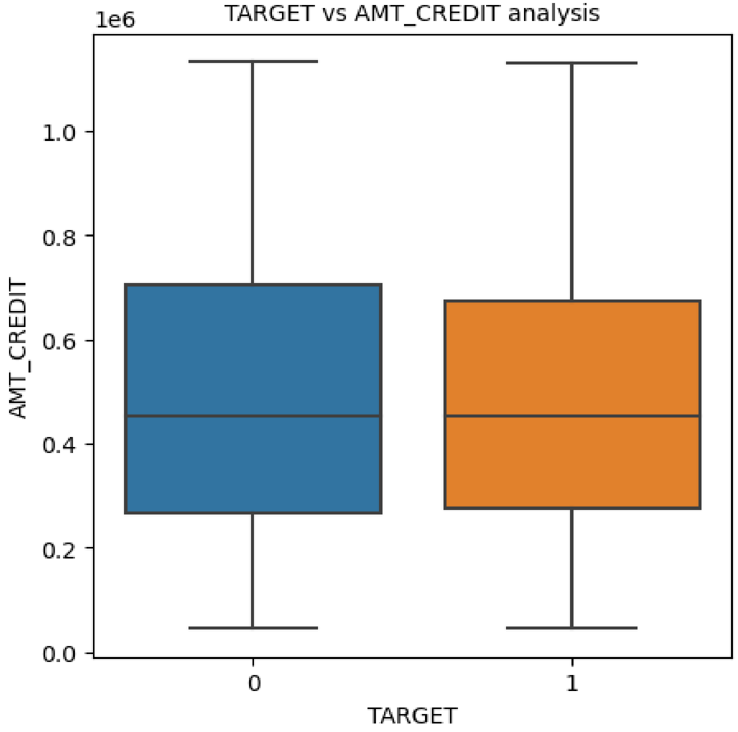
- Majority of the applicants have income in range of 100 to 200K; no significant difference has been observed in Target vs total income groups
- Majority of the applicants who have below 200K income are facing slightly more difficulty in payment than all the other cases. Applicants with 200 to 300K income are able to pay on time

Univariate Analysis: Overall and segment level [6/9]

Credit loan amount:



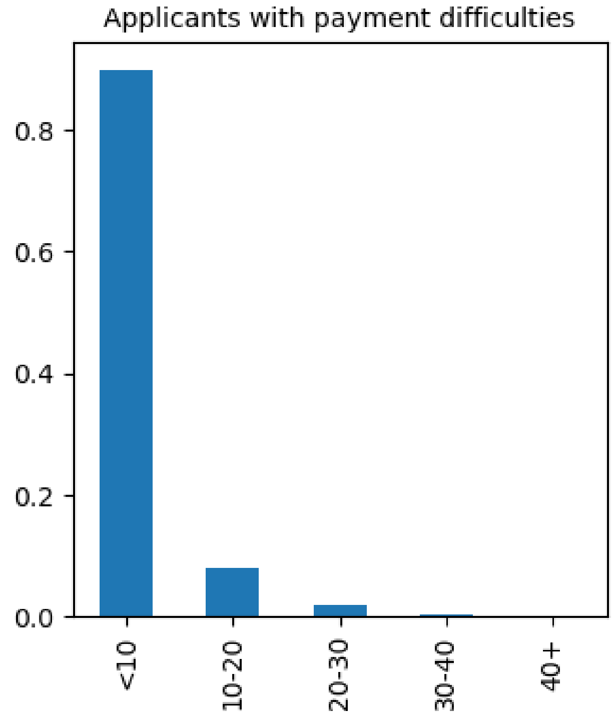


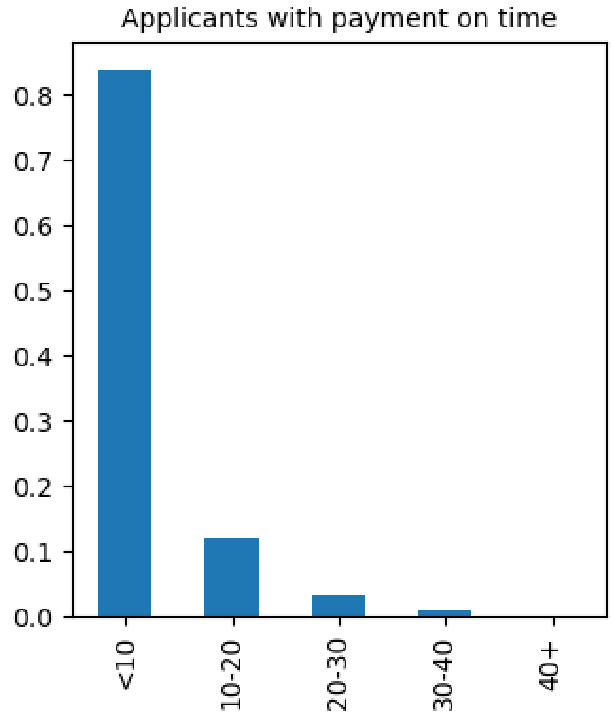


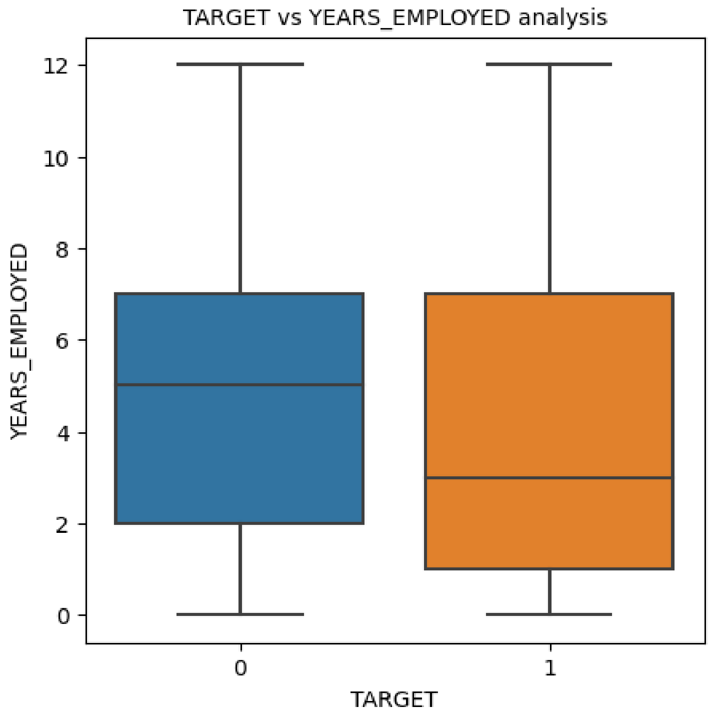
- Majority of the applicants have credited the loan amount in range of 100 to 400K
- Majority of the applicants who have credited the amount between 400K-800K are facing slightly more difficulty in payment than all the other cases. Applicants with 800 to 1200K credited amount are able to pay on time. There is no significant difference observed for the group of applicant who have credited 100 to 400K

Univariate Analysis: Overall and segment level [7/9]

Employment history:



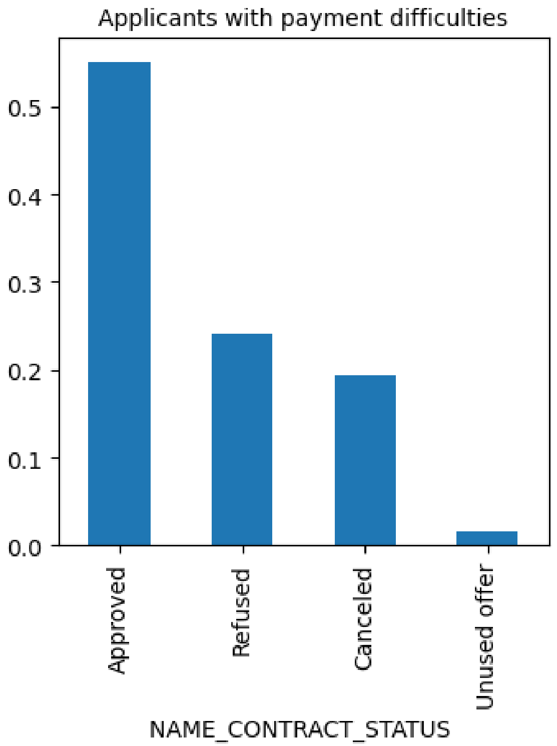


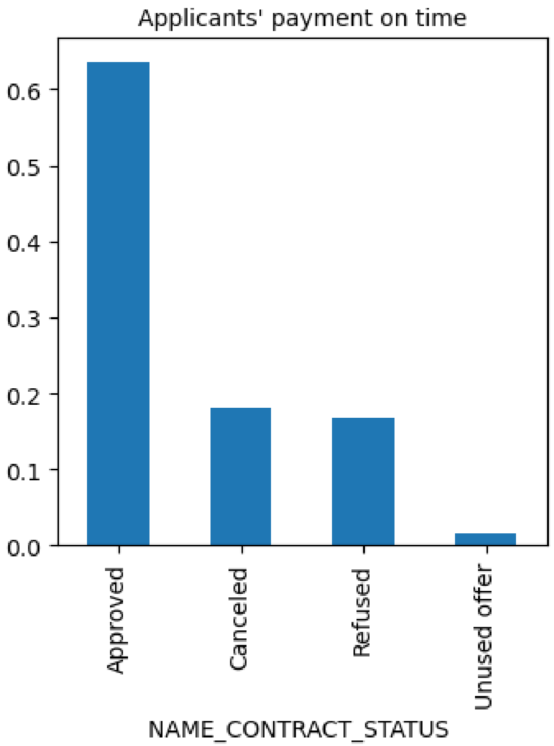


- Majority of the applicants are falls below 10 years of employment history
- Applicants who have less than 10 years of history of employment are facing difficulties in payment. Applicants with 10 to 20 years of history of employment are able to pay on time; low median employment history for applicants with payment difficulties vs applicants who are able to pay on time

Univariate Analysis: Overall and segment level [8/9]

Previous loan contract status:

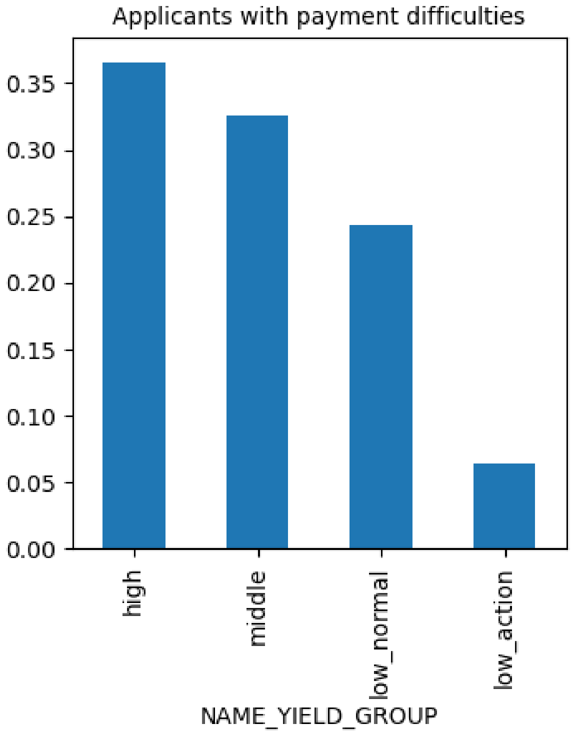

Applicants with payment difficulties

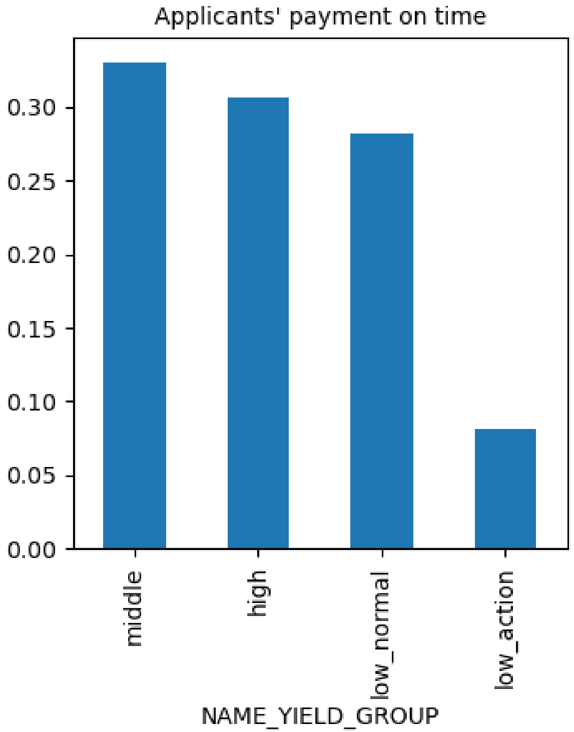

Applicants' payment on time

- Majority of applicants are approved on their previous application able to pay on time vs applicants with payment difficulties
- However, There is Approved applicants on their previous application are higher as compared to others groups
- The applicants who are refused on their previous application are having difficulties in payments with current application

Univariate Analysis: Overall and segment level [9/9]

Previous loan Yield group:

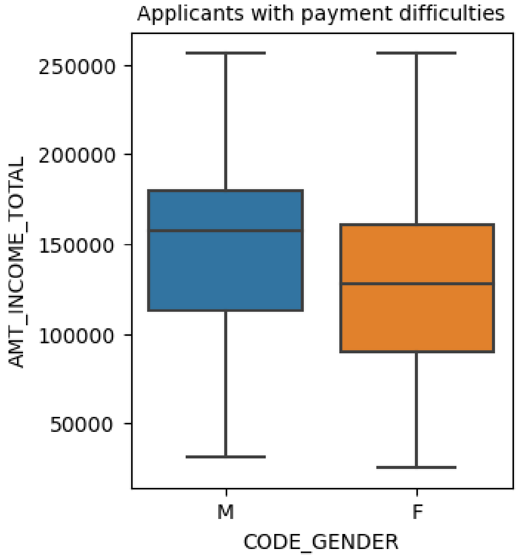


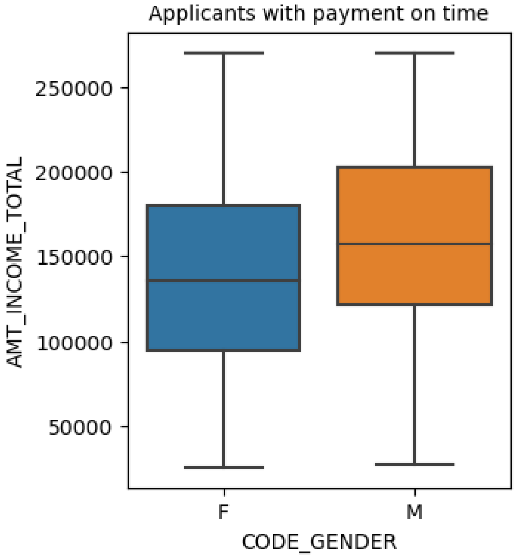


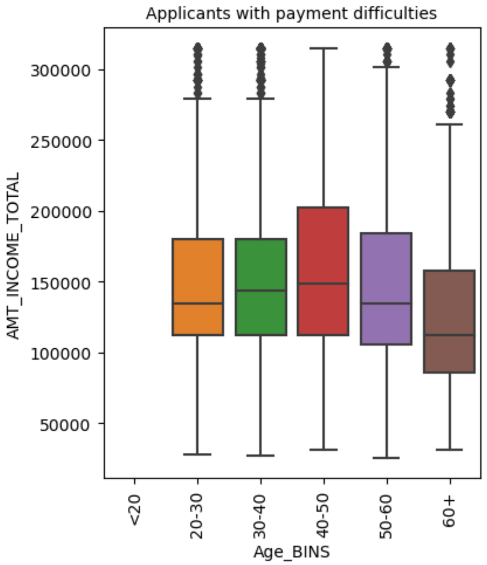
- Majority of applicants are having group interest rates with high interest rates of their previous application
- Applicants with high interest rates on their previous application are having payment difficulties vs applicants who are paying on time. The applicants who are having low normal interest rates of their previous application are able to pay on time vs applicants with payment difficulties

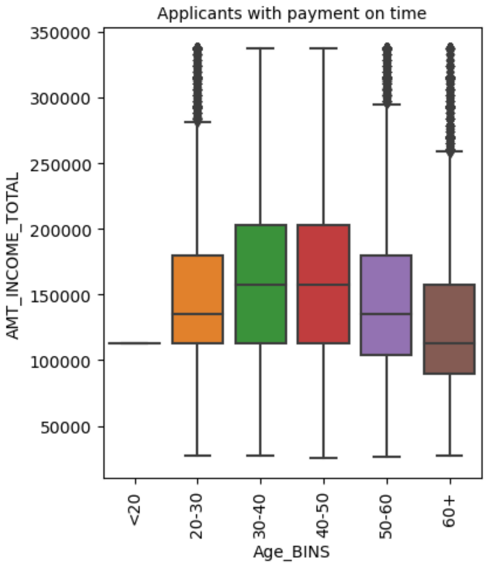
Bivariate Analysis: Overall and segment level [1/4]

Income vs Gender and Age:


Applicants with payment difficulties


Applicants with payment on time

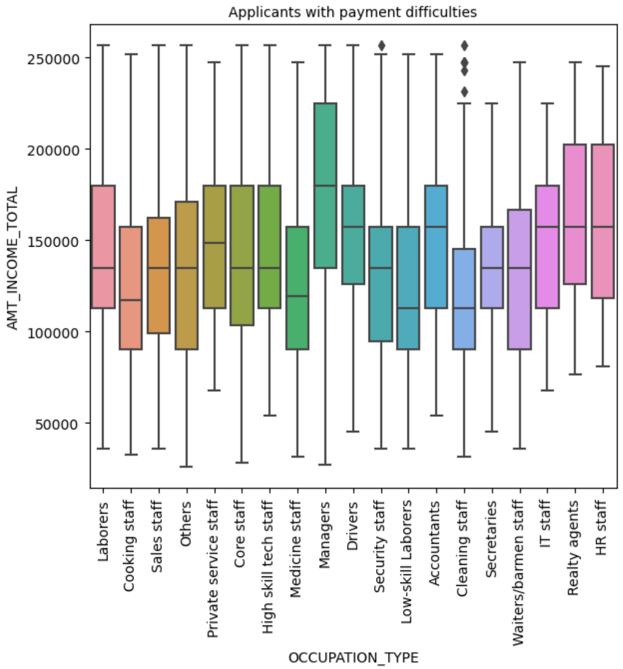

Applicants with payment difficulties

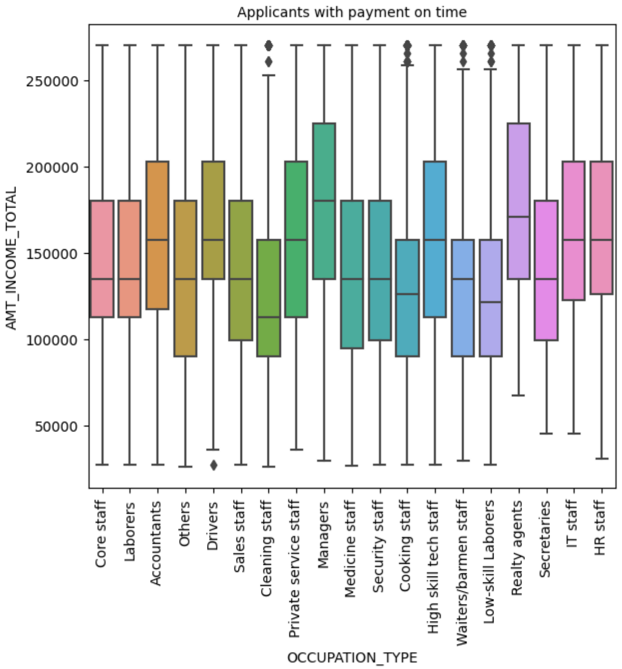

Applicants with payment on time

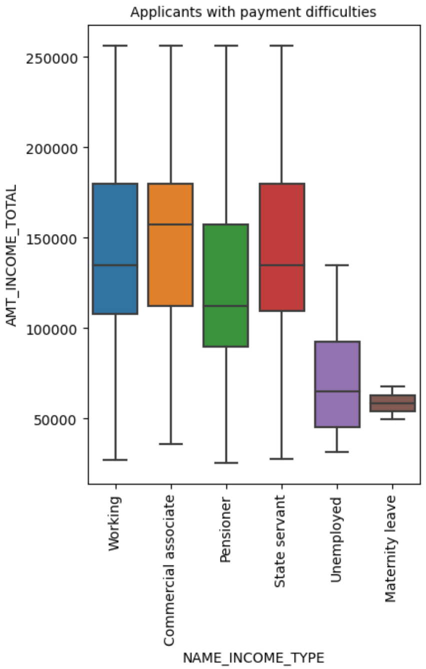
- Male applicants are on the higher income than females irrespective of Payment difficulties/ payment on time
- Applicants from 30 to 50 age group are able to pay on time and have higher median income while the 40 to 50 age group is also having payment difficulties as compared to other groups
- 60+ applicants are the lowest income group which are hvaing difficulties in payment vs other groups

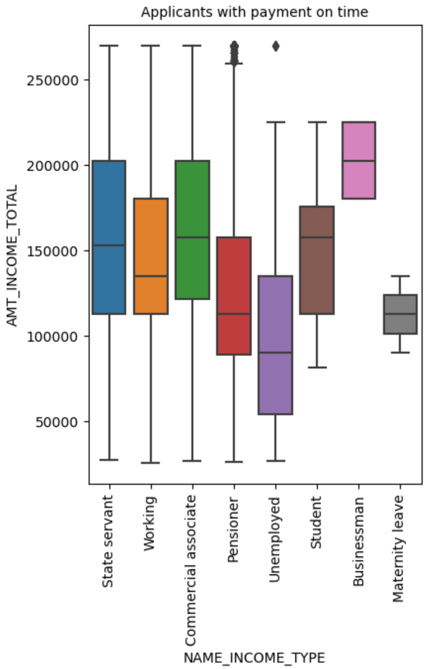
Bivariate Analysis: Overall and segment level [2/4]

Income vs Occupation and Income type:





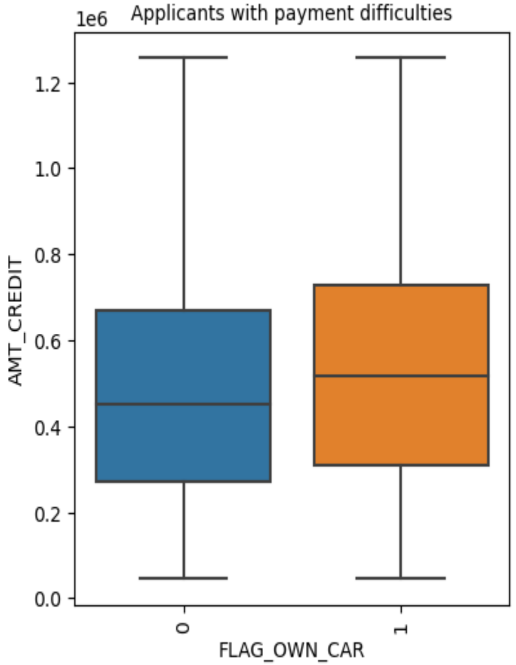


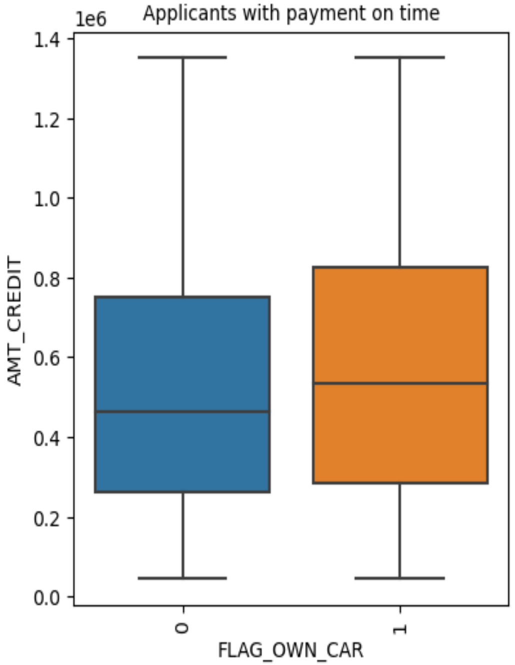


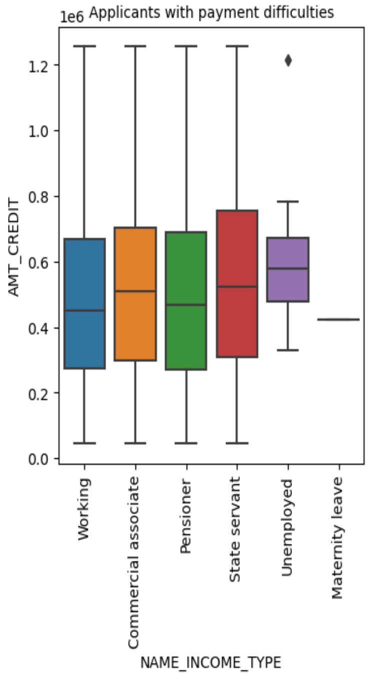
- Analyzing the occupation type, Managers, IT & HR staff, and Realty agents are having the higher median income as compared to other groups. Also, they have payment on time as compared to other applicants
- Applicants who are low skilled laborers and cleaning staff are having the difficulties with payment, and are the low income groups among others
- Analyzing the income type, applicants who are unemployed and on maternity leave are having the difficulties with payment, and are the low income groups among others

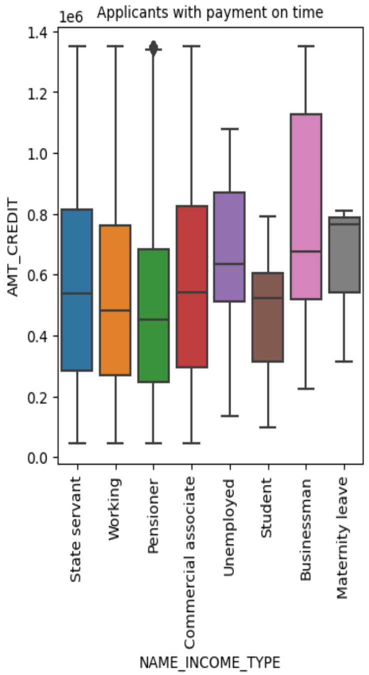
Bivariate Analysis: Overall and segment level [3/4]

Credit vs car status and income type:





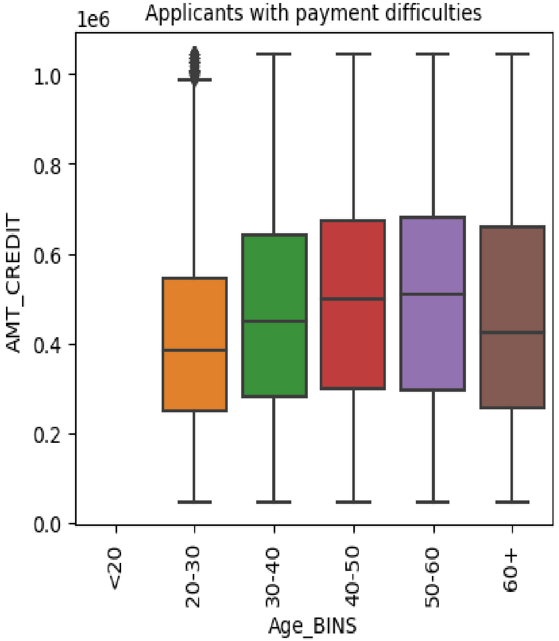


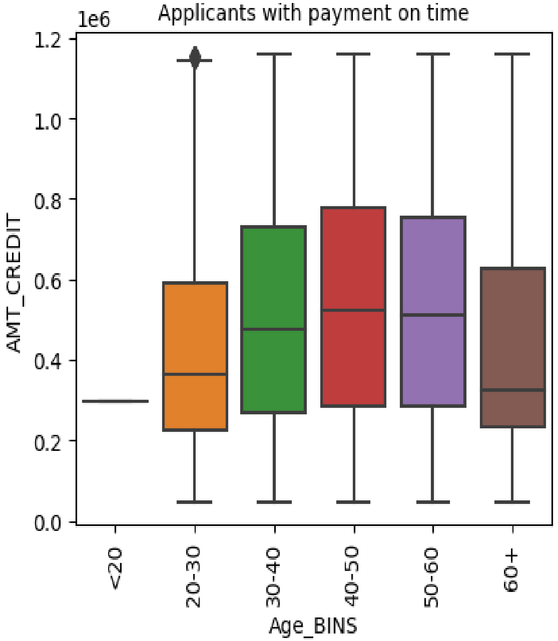


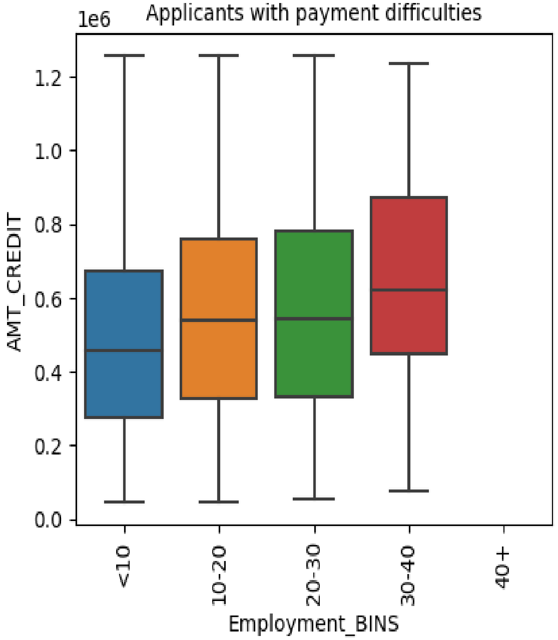
- Applicants who own car are having a higher median value for credit amount for a loan. Also, they are having difficulties in payment as compared to applicants who do not own a car and also with applicants who own a car and paying own time
- Businessman with larger credit amount (loans) are somehow able to pay on time as compared to other groups
- Maternity leave applicants are having the higher median credit loans and able to pay on time as compared to other applicants

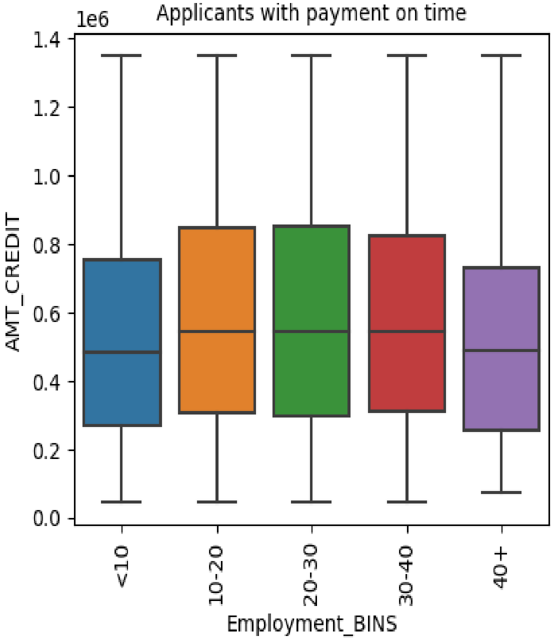
Bivariate Analysis: Overall and segment level [4/4]

Credit vs Age and Employment history:



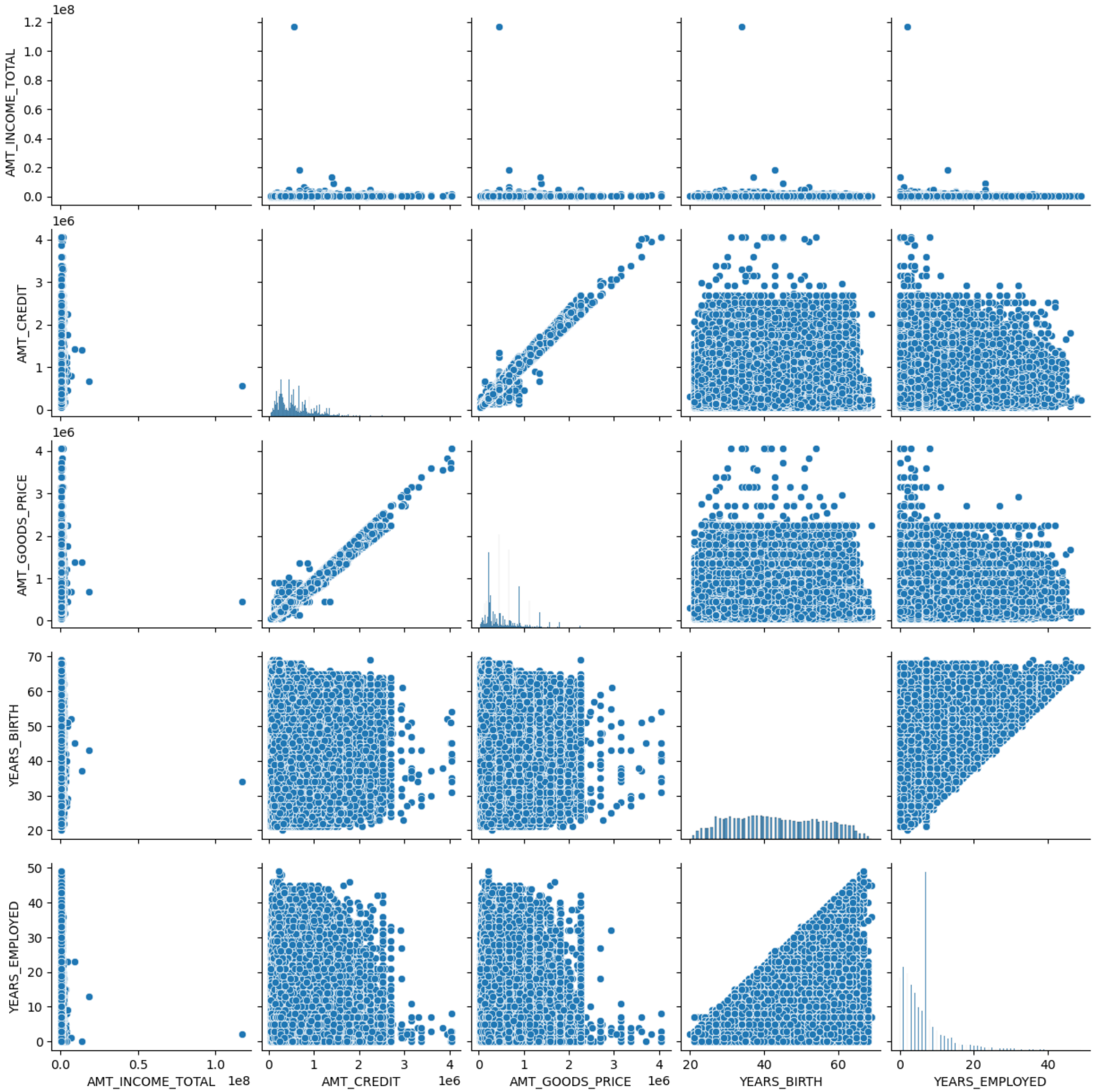


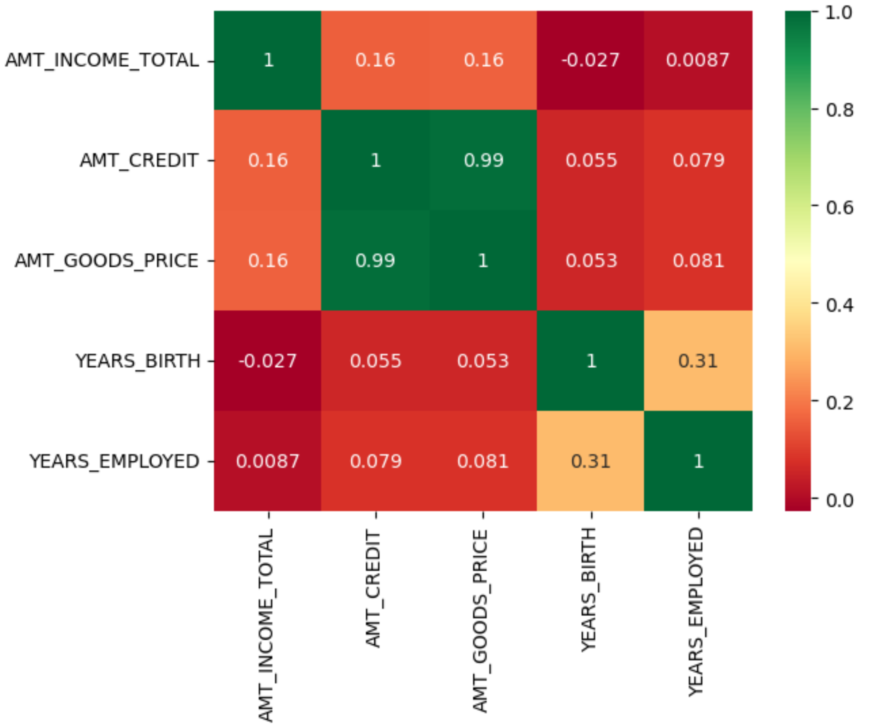




- Applicants from 40 to 60 age group are having higher median credit loan amount and also facing payment difficulties as compared to other groups. There is no significant difference in both with and without payment difficulties groups except 60+ applicants have low credit loan amount vs other groups
- Applicants with 30 to 40 years employment history are having a higher median credit loan and facing trouble in payments vs other groups. The applicants who have 10 to 40 years of employment history are able to pay on time and have higher median values vs others

Correlation: Overall and segment level





- There is a strong positive corelation between AMT_GOODS_PRICE and
- There is a weak or no relation between AMT_GOODS_PRICE and YEARS_BIRTH, YEARS_EMPLOYED with AMT_CREDIT and AMT_GOODS_PRICE.

Conclusion

Based on the several combinations of analysis on variables, we can recommend the potential defaulters as mentioned below;

1. **Gender**: Male applicants find difficulties with payments
2. **Education**: Applicants who achieved Secondary / secondary special level of education having less chances of paying the loans hence company should target applicants with Higher education level who are able to pay on time
3. **Occupation**: Laborers and Sales staff applicants are having higher chances of defaulters
4. **Age**: Applicants who are in the age group of 30 to 40 are facing difficulty in payment hence can be a loan defaulters
5. **Income amount**: Applicants with less than 100K income has more chances of default rates while applicants with more than 200K of income are able to pay on time
6. **Credit amount**: Applicants who have credit loan between 400 to 800K having more default rates
7. **Employment history**: Applicants with lower employment history are likely to be defaulters while applicants with more employment history are the best target for the loan
8. **Previous loan history**:
   - Majority of applicants are approved on their previous application able to pay on time while applicants who are refused on their previous application are having difficulties in payments with current application hence company should reconsider this group before considering for a loan
   - Applicants with high interest rates on their previous application are having high chances of defaulting
   - The applicants who are having low normal interest rates of their previous application are able to pay on time and can be targeted for the loan

THANK YOU!