Credit EDA Assignment

Arsalan Khursheed

DS C52

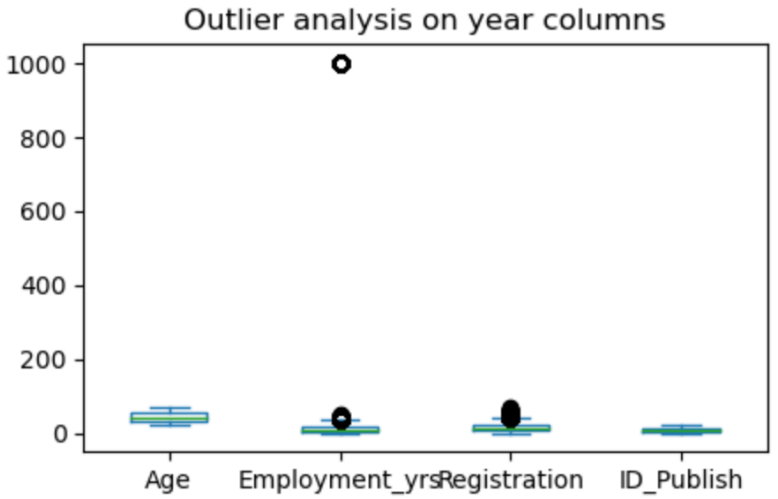
Business Understanding & Objectives

- Perform EDA to analyze the data for loan applications and draw insights from patterns present in the data
  - Applicants capable of repaying the loan should not be rejected (This results in loss of business if rejected)
  - Applicants not likely to repay the loan should be rejected (This results in financial loss if approved)
- 4 types of decisions can be taken on an application
  - Approved, Cancelled, Refused, Unused Offer
- Identify the driving variables using the 2 datasets provided which lead to a default

Application data analysis

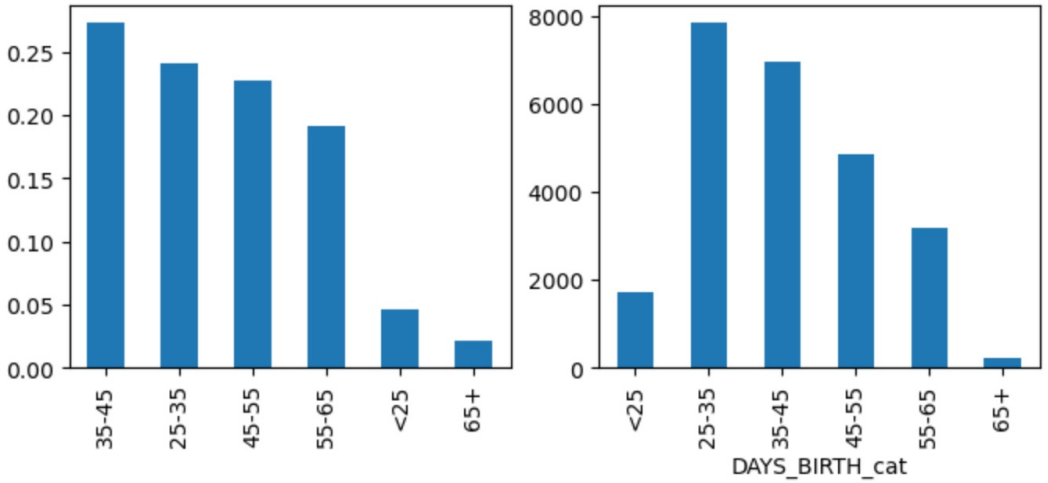
- Looking for missing values in all columns
  - Drop columns with more than 40% missing values as they will not add value to our analysis
- Based on understanding of data we should drop the flag columns
- We may drop the miscellaneous columns which we feel will not add any value to our analysis
- After analysis of a few data points visually we see that there are many “XNA” values which seem to be missing values. We can replace these with NaN
- We find that there are few columns in days which we can convert to years for ease of our analysis.
  - Few are incorrectly negative as well which means we should use absolute
- We can now do outlier analysis on our dataframe
- We should then do univariate and segmented univariate analysis
- We should also do bivariate and multivariate analysis on the data to derive correlations and insights

Outlier analysis on the days columns after converting to years

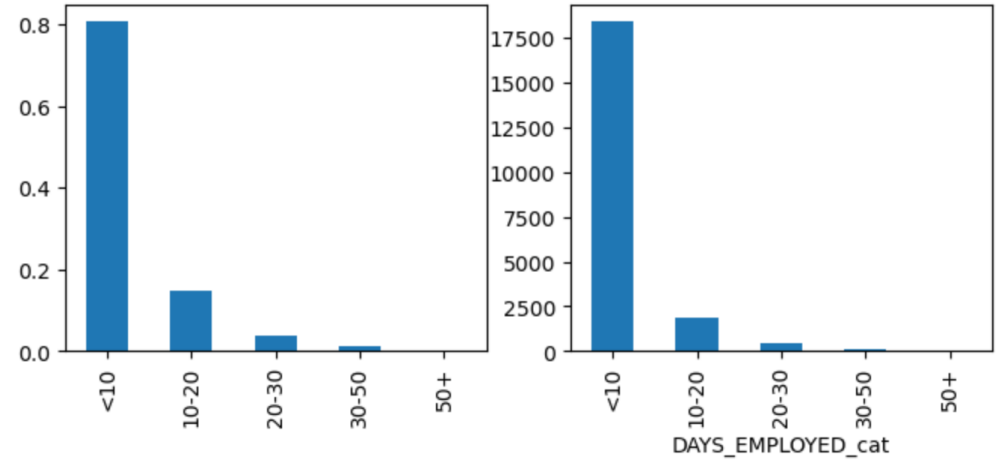


- 18% of the employment column has values which are more than 100yrs.
  - We can treat these as missing values an continue
- We should convert our year columns to categorical for ease of visualization and finding insights

Univariate analysis on Age & Employment columns



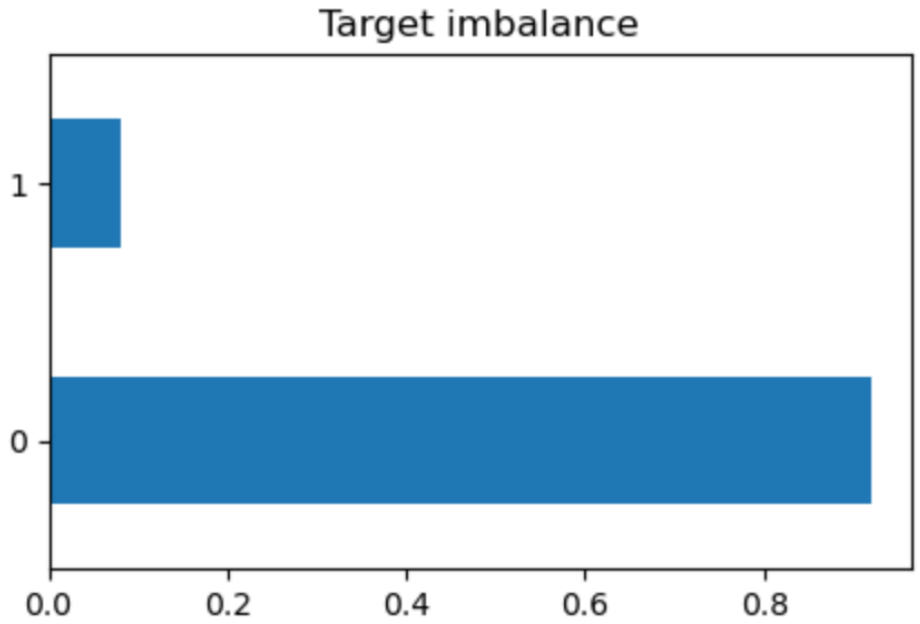
Most applications are from 35-45 age group and most defaults are from 25-35 followed by 35-45



Most applications and defualts are from people with less than 10 yrs of work experience

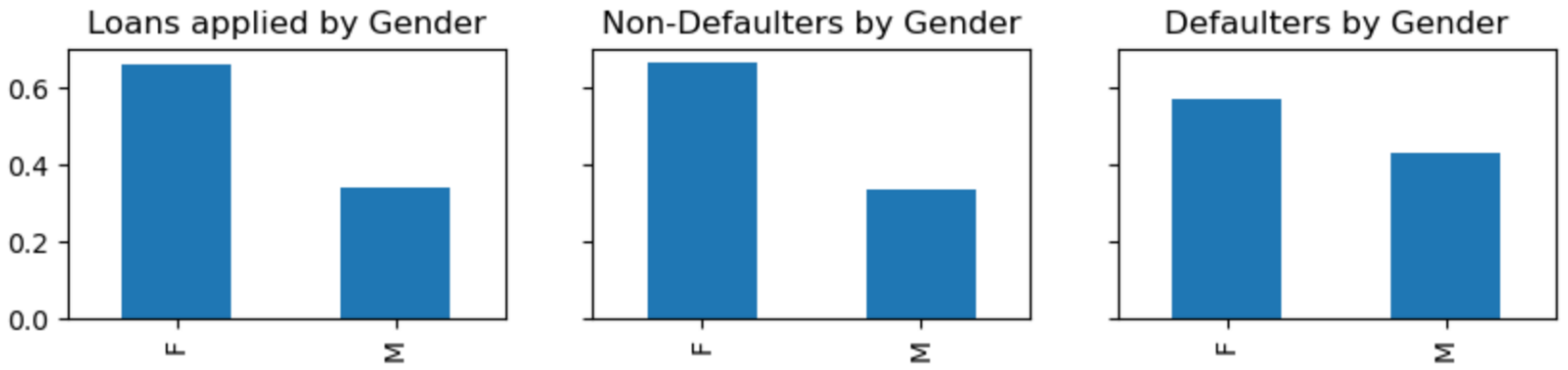
- We find that most applicants are from the 35-45 age group followed by 25-35.
  - Most defaults are from the 25-35 age group followed by the 35-45 age group
- 80% applications are from less than 10 yrs employed group which is also the highest default group

Data imbalance check of Target



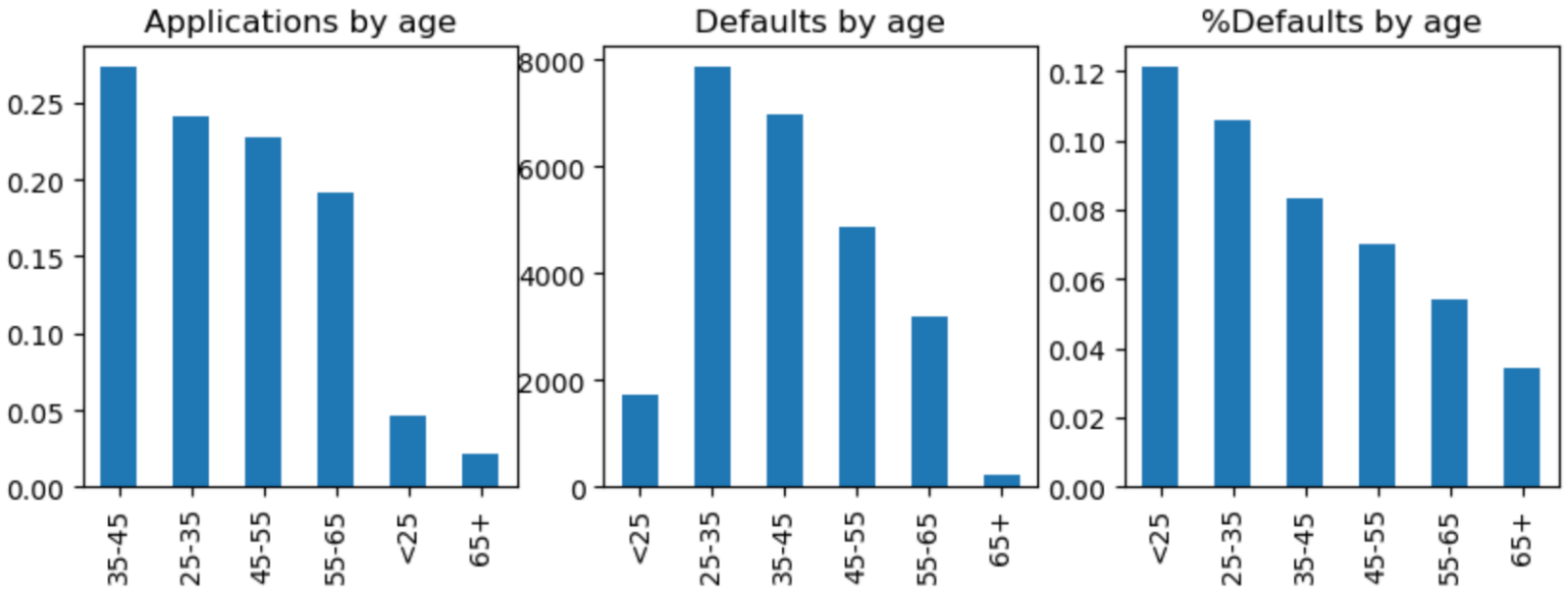
- More than 90% of the Target column has the value 0 which means that the column is highly imbalanced
  - We can divide our dataset into 2 dataframes for further analysis and to identify trends among the defaulters

Bivariate analysis of Gender column



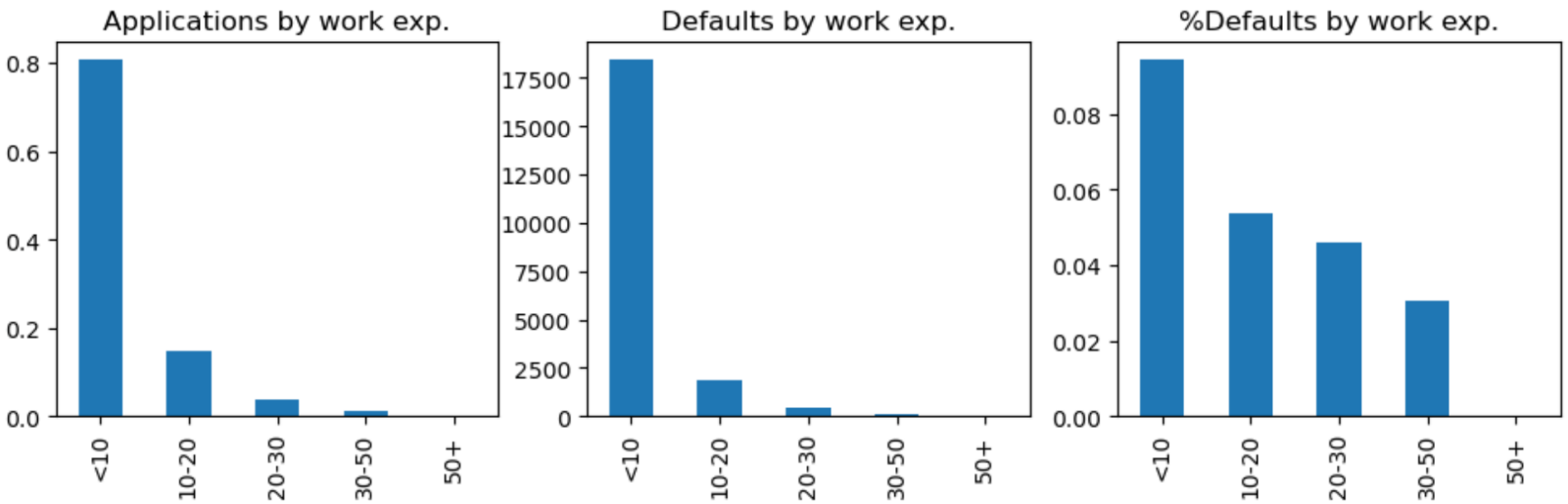
- Female applications are almost double than that of males
- % of males defaulting on loans is more compared to that of females when we analyze them against the number of applications. This means that males have higher chances of defaulting compared to females

Bivariate analysis of Age groups



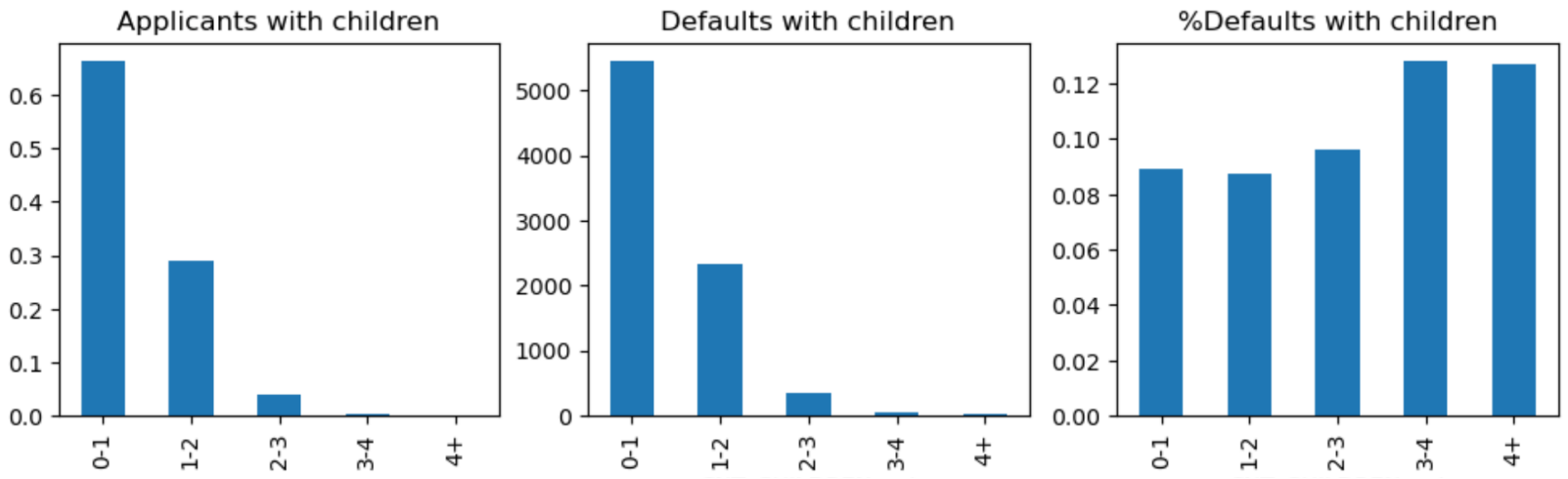
- Most applications are from 35-45 age group and most defaults are from 25-35 followed by 35-45
- %Defaults with respect to applications is mostly in the <25 age group

Bivariate analysis of Employment tenure



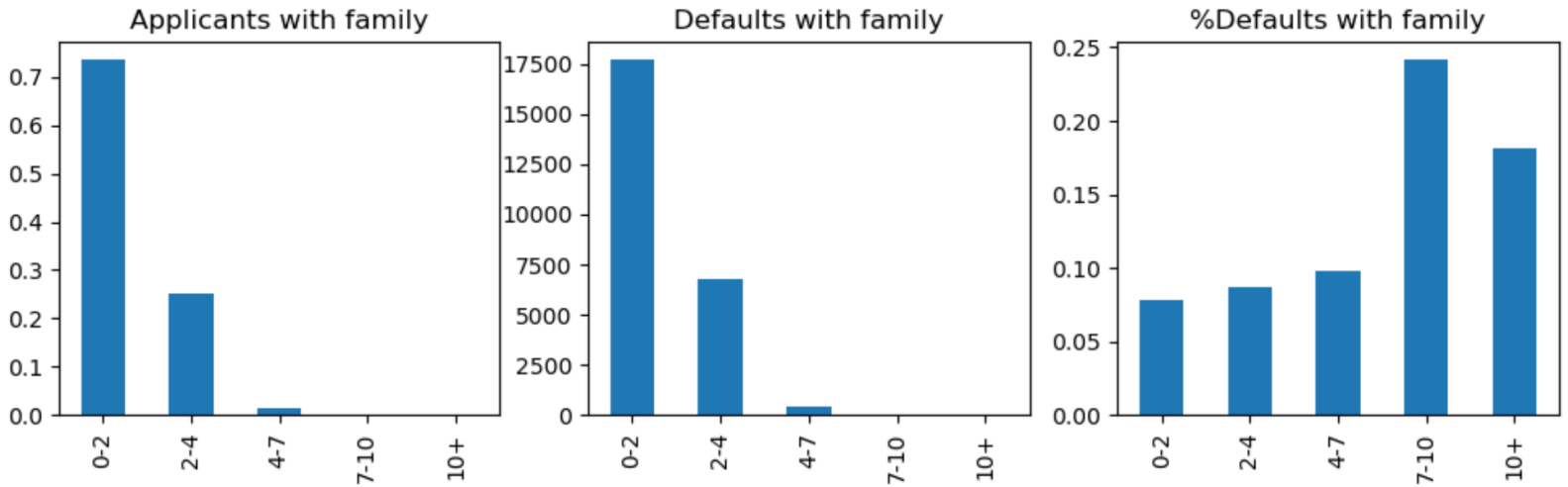
- Most applications and defaults are from people with less than 10 yrs of work experience
- %Defaults is also highest in applications with less than 10 yrs of work experience

Bivariate analysis of Applicants with children



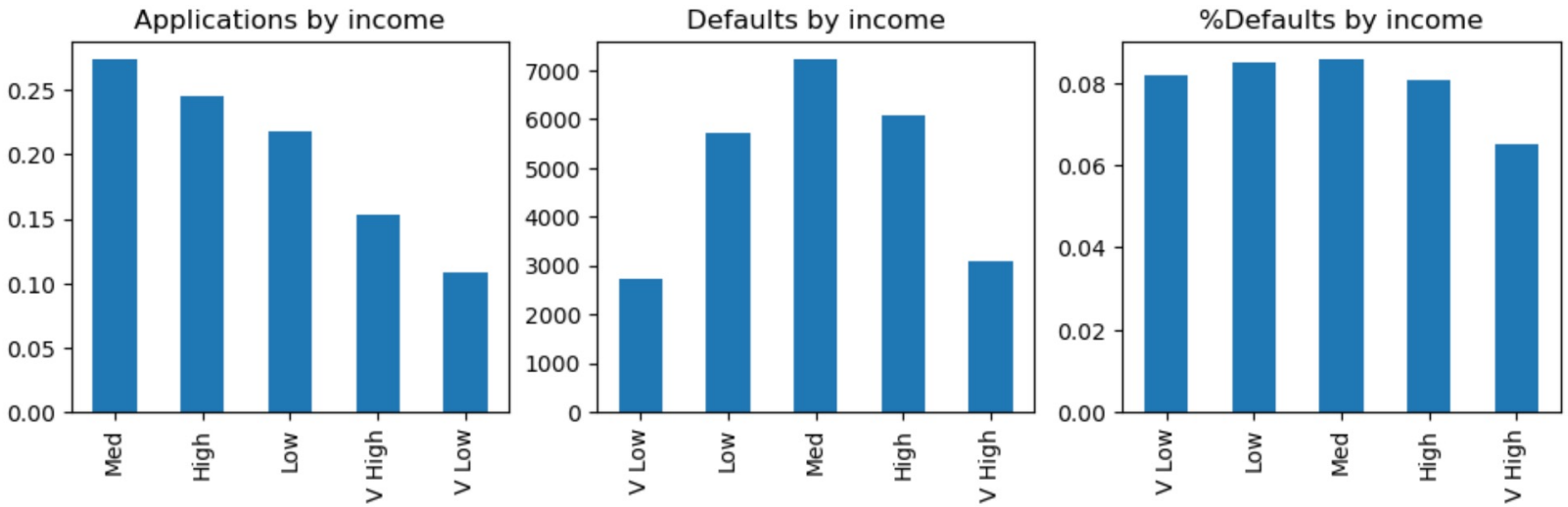
- Most applications and defaults are from people with 1 child followed by 2
- Most %Defaults are from applicants with 3 or more children which means applicants with 3 or more children have a higher chance to default

Bivariate analysis of Applicants with family members



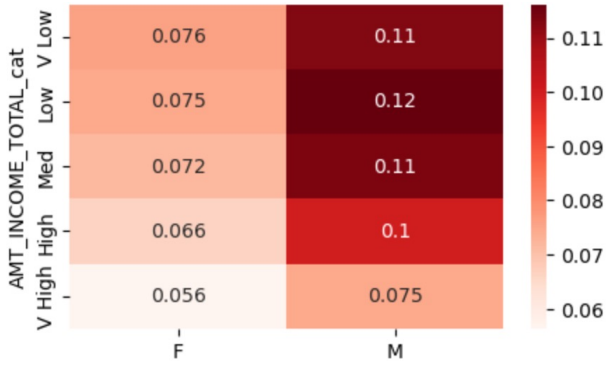
- Most applications and defaults are from people with up to 2 family members followed by 2-4
- Most %Defaults are from applicants with 7-10 family members followed by more than 10 which means applicants with 7 or more family members have a higher chance to default

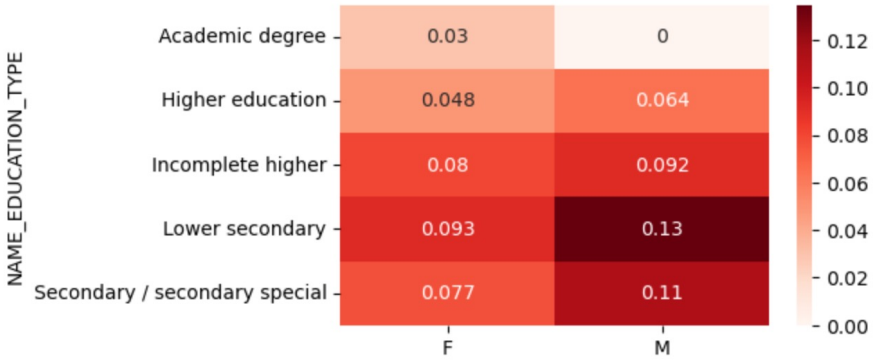
Bivariate analysis of Applicants with income

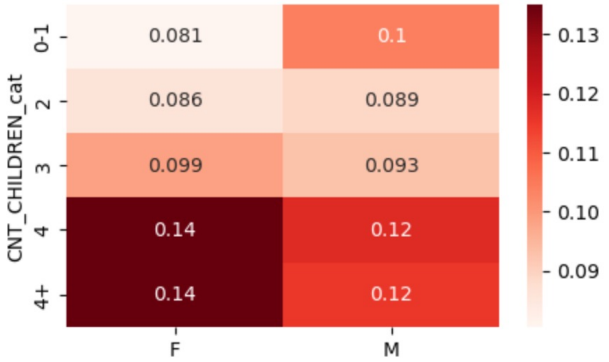


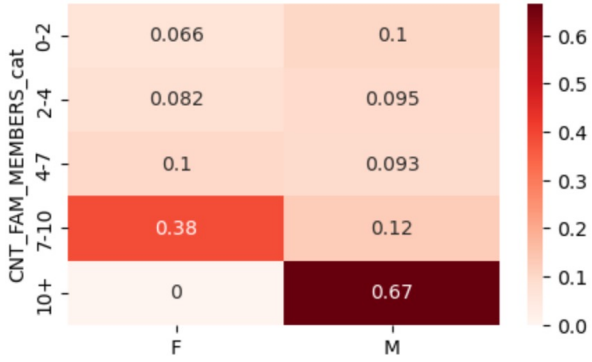
- Most applications and defaults are from people within the medium income group
- However, the %defaults is most in Low & Med group followed by Very Low & High

Multivariate analysis of Applicants by Gender



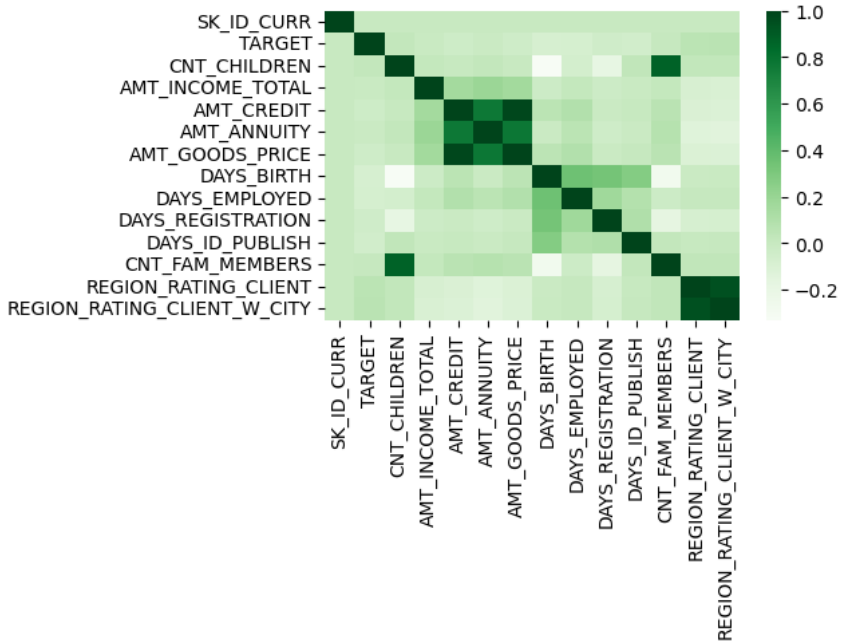






- Males with 10 or more family members tend to default the most followed by females with 7 or more
- More than 4 children parents have higher chance to default with females defaulting more than men
- Males who do not have Higher education have high chances of defaulting compared to females
- Males with Very Low - Medium income have high chances of defaulting compared to females

Top Correlations

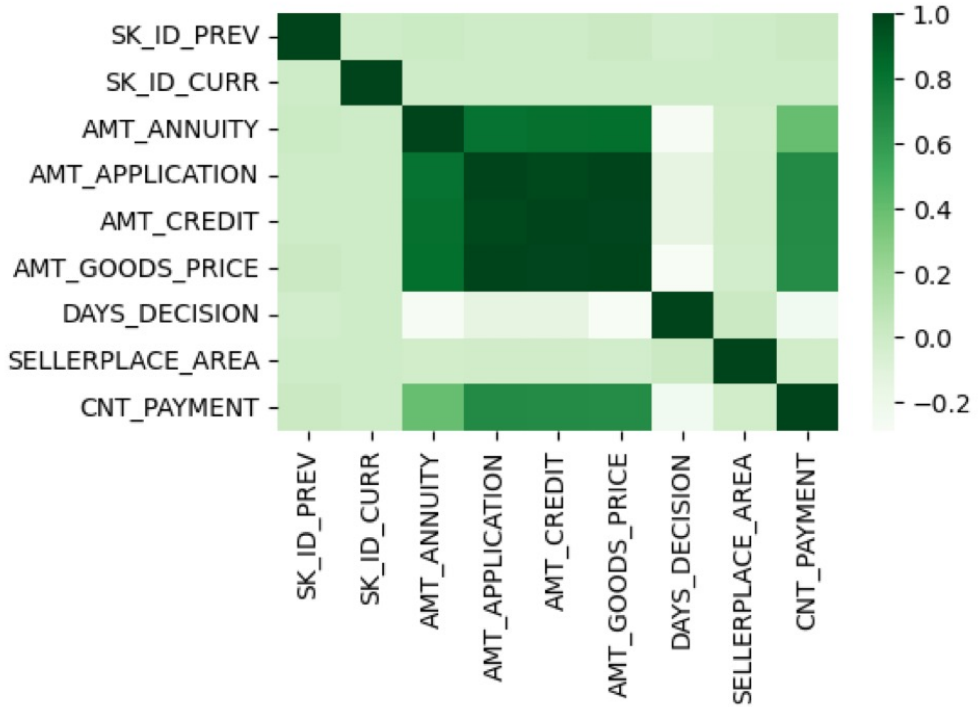


- AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE are correlated.
- CNT_FAM_MEMBERS, CNT_CHILDREN are correlated.
- REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY are correlated.

Previous Application data analysis

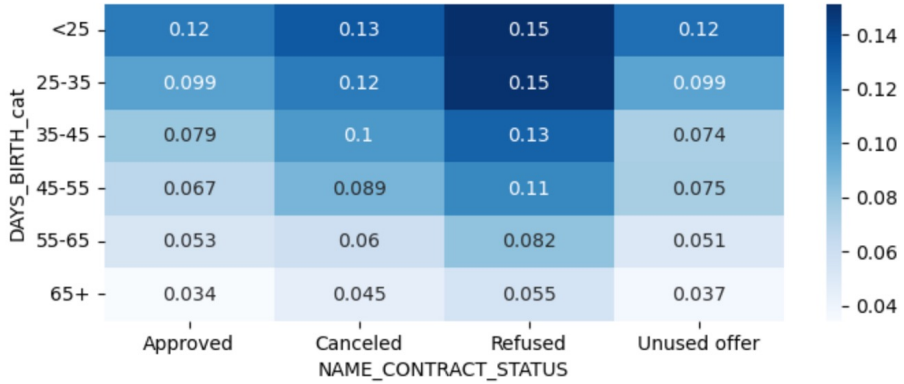
- Looking for missing values in all columns
  - Drop columns with more than 40% missing values as they will not add value to our analysis
- Based on understanding of data we can drop the flag columns & APPR_PROCESS_START columns
- We find that there is a columns in days which we can convert to years for ease of our analysis.
- We can now do outlier analysis on our dataframe
- We should then merge the 2 dataframes and continue with our analysis for further findings if possible. We should do left join as the target variable is present in application dataset
- NAME_CONTRACT_STATUS is an important column to do further analysis with application dataset so that we may understand the approved/refused/cancelled/unused offers

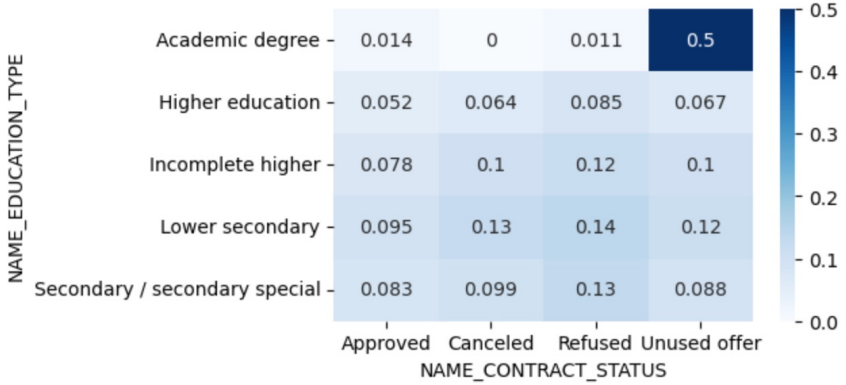
Top Correlations

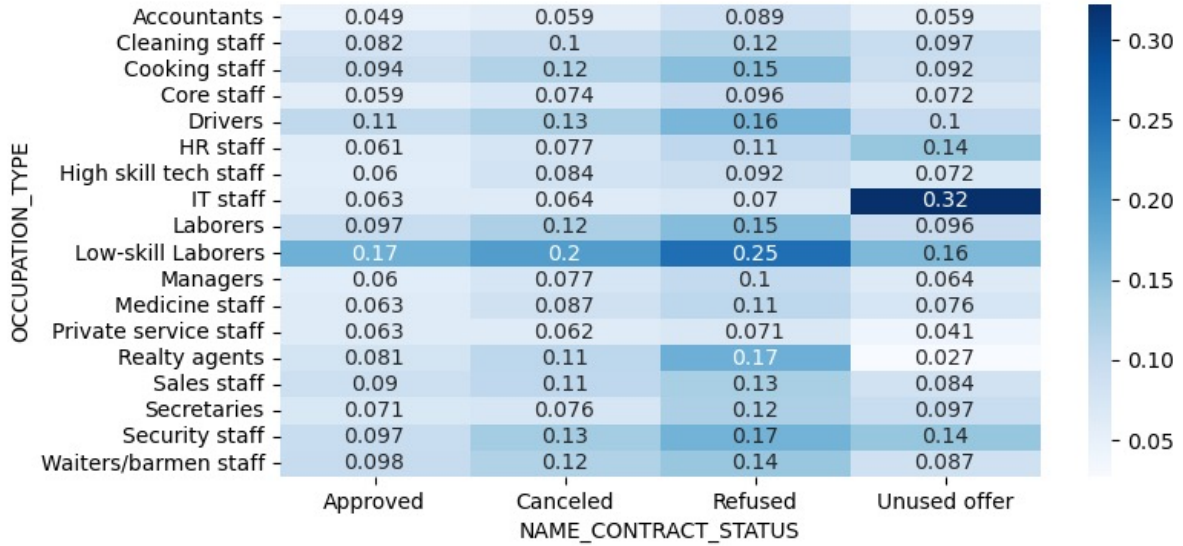


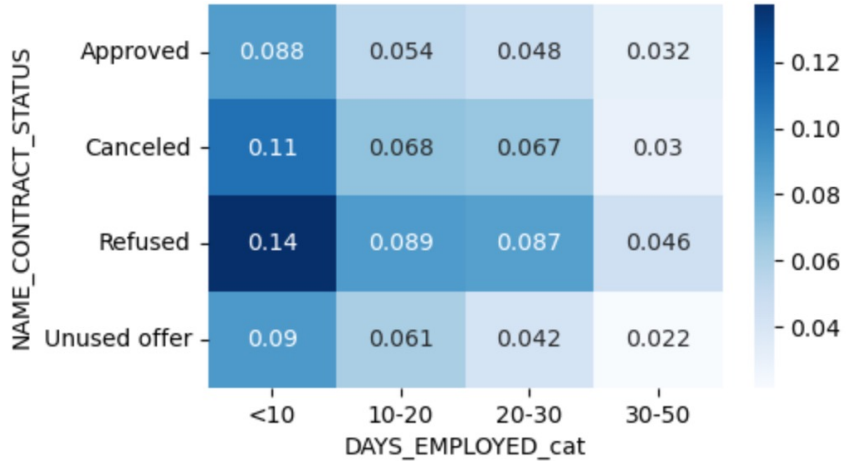
- AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE,AMT_APPLICATION, CNT_PAYMENT are correlated.

Multivariate analysis of Contract status by Attributes

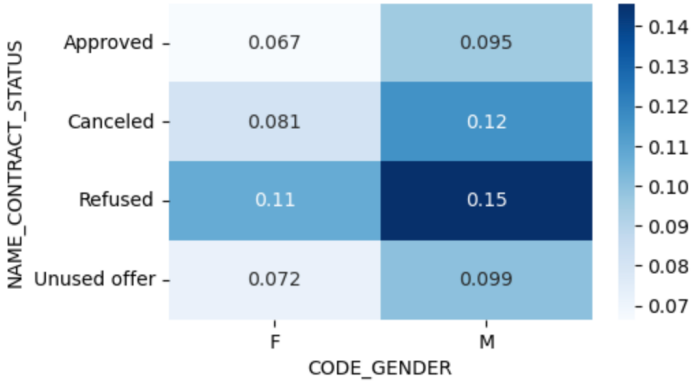


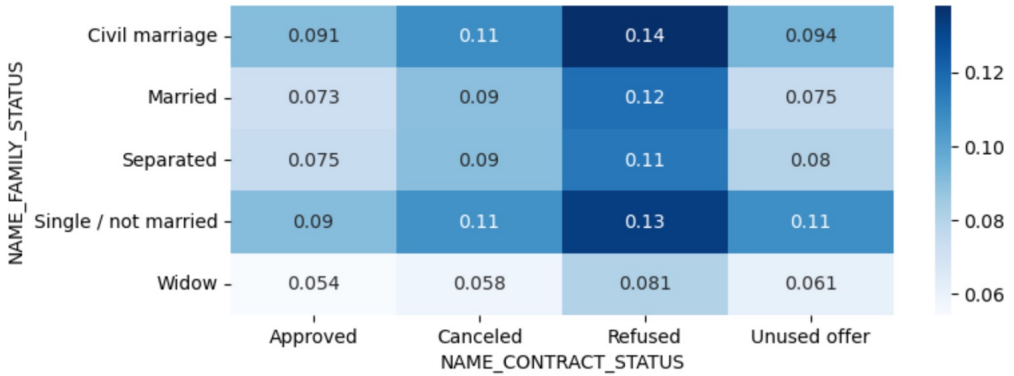






Insights based on merged dataframe





- Most refusals are from <25 and 25-35 age group followed by 35-45 age group which is in line with our analysis
- Most refusals are from the <10 years work experience which is in line with our analysis
- People with academic degrees have the highest unused offers which may mean that they have a decent income and living thus not in need of a loan
- IT staff has the highest unused offers and Laborers & Low-skill Laborers have the highest refusals
- Males have more refusals compared to females which is in line of analysis on %defaults by gender