Credit EDA Assignment – Case Study

To understand the driving factors (or driver variables) behind loan default, i.e., the variables which are strong indicators of default.

- Sharod Chand Dey

1 Data Understanding, Cleaning and Manipulation

In this section we will report the overall approach taken to analyse the case study and the steps taken to mitigate the data quality issues present in the data set.

The most important step would be to clean the data properly which includes handling of missing data, treating outliers, and dropping irrelevant columns and focusing on the variables which make logical and business sense.

Once this is completed, we can start to dig into the data using bivariate and multivariate analysis to draw inferences about clients who might be more likely to default on loans. We will use different kinds of plots like bar charts, pie charts, boxplots and heatmaps to gain an understanding of the correlation between different variables and how it affects the loan defaulters.

1.1 Data Understanding

- Based on the data dictionary and my own business understanding some of the most important columns I have identified for this analysis are:

Column Name	Reason Identified
SK_ID_CURR -	Will serve as primary key when merging
TARGET	Target variable to analyse, 1 means likely defaulters
NAME_CONTRACT_TYPE	Type of loan taken cash or revolving
CODE_GENDER	Gender of client
AMT_INCOME_TOTAL	Total income of client, will be important to determined default
AMT_CREDIT	Credit amount for loan, might be an important factor
AMT_ANNUITY	Loan annuity, could be an important factor
AMT_GOODS_PRICE	Price of goods for which the loan is being taken
NAME_INCOME_TYPE	Type of income, will be useful when analysed with total income
NAME_EDUCATION_TYPE	Level of education, will be important to determine defaults
NAME_FAMILY_STATUS	Marital status, could be a driving factor
NAME_HOUSING_TYPE	Type of housing for the client
DAYS_BIRTH	Age of client in days, will have to converted to years
OCCUPATION_TYPE	Occupation type of the client
CNT_FAM_MEMBERS	Total no of family members
NAME_CONTRACT_STATUS	Contract status (approved, cancelled, ...) of previous application
CODE_REJECT_REASON	Reason for previous application rejection

- I believe that conducting analysis using these variables may give us a better understanding of what could be the major driving factors for loan default.
- Now all the irrelevant columns are dropped. Dropping and cleaning of the columns has been spread out over the Jupyter notebook. This is primarily due to the flow of my analysis.

1.2 Data Cleaning and Manipulation

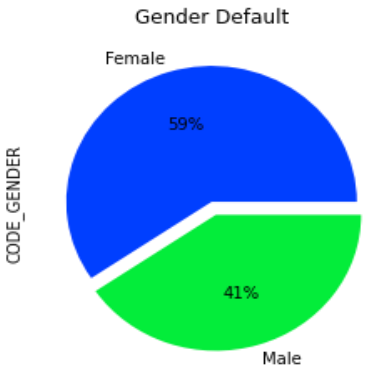
- Some values had to be **standardized** for better visual representation. E.g.- CODE_GENDER column was converted from single letters (M & F) to 'Male' and 'Female'
- **Missing values** in this column were treated by imputing the mode of the column. This was done since it is a categorical variable
- Some columns had negative values which had to be converted into positive, such as DAYS_BIRTH. This was handled by **deriving** a new column 'YEARS_BIRTH' and converting the age of the client into years while also making it a positive number
- **Percentage null values** in the data frame was found. Columns containing low % of null values were analysed for outliers but the null values were left alone. An **alternative approach** could be to **impute** these null values with the median/mode since this is a numerical variable
- Some of these numerical columns were also **segmented** to convert them into a categorical variable. E.g.- AMT_ANNUITY was segmented into 4 buckets of appropriate size and a new column was derived AMT_ANNUITY_GROUP
- The same binning treatment was done for AMT_INCOME_TOTAL by creating AMT_INCOME_TOTAL_GROUP. Two other columns were also segmented this way.
- **Outliers** were also identified and treated in parallel. 5 columns were treated for outliers:
    - AMT_ANNUITY
    - AMT_CREDIT
    - OWN_CAR_AGE
    - AMT_INCOME_TOTAL
    - AMT_GOODS_PRICE
- Two different approaches were taken here in treating outliers:
    - The difference between the 99th and 100th quantile was calculated. Since it was found to be very large it was clear that the max value point was an outlier. These outliers were dropped. E.g.- Approach taken for AMT_ANNUITY
    - Capping the outliers at a specific value. E.g.- AMT_GOODS_PRICE was capped at 3.5M
- Data imbalance was found to exists in the TARGET variable which had a very high proportion of zeros. There are two approaches to address this:
    - Segment the data frame into two parts w.r.t the TARGET variable and plot charts for each data frame separately
    - Take the mean of the TARGET variable and it will be equal to the proportion of ones in the column. Since we are only concerned about the driver variables for default I have personally chosen this approach

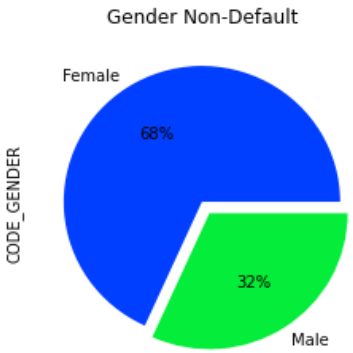
2 Data Analysis

In this section we will look at some of the important plots created wile conducting univariate and multivariate analysis. We will also try to find some insights into how these factors will play a role in loan defaults.

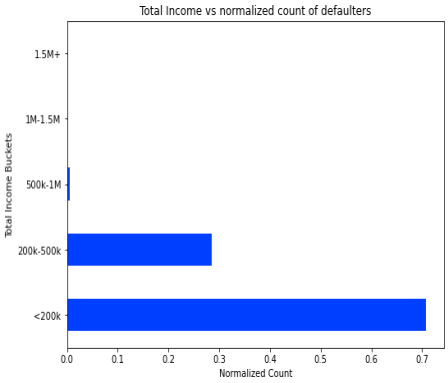
2.1 Univariate and Segmented Univariate Analysis

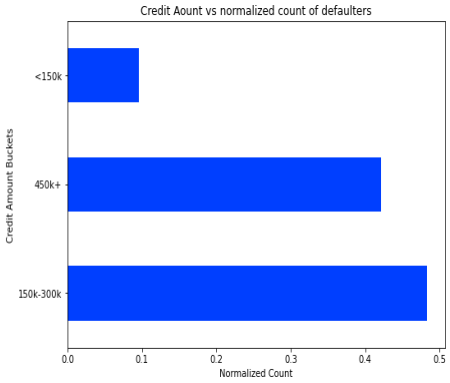
- The first thing we notice is the total number of clients applying for loans are Females but the difference in genders for payment difficulties is not very high, as seen below.

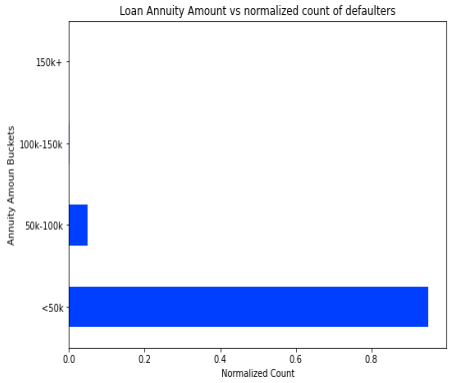




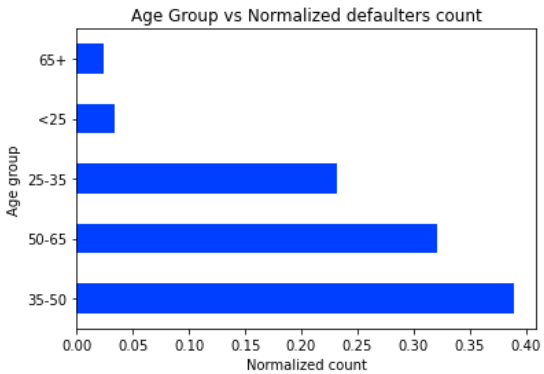
- Next, we look at the total income group, credit group and annuity group and infer the following:
    - Most people who have payment difficulties earn below 200 thousand
    - The credit amount for loans extended to majority of applicants falls between 150-300 thousand
    - The amount annuity for majority of applicants is less than 50 thousand

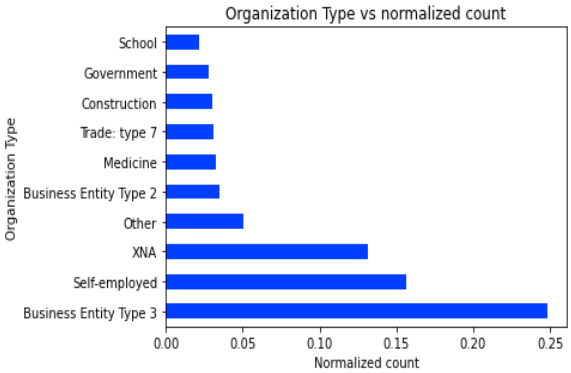






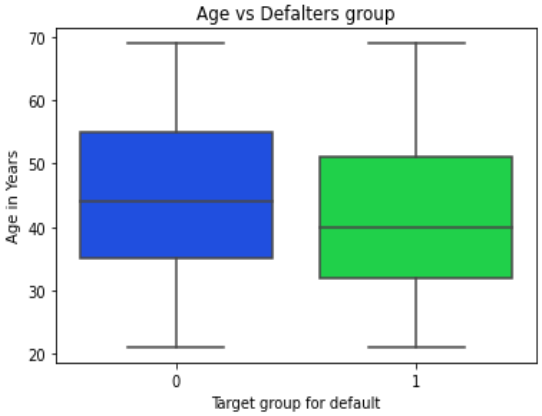
- Since these are plotted for clients who face payment difficulties, we can surmise that clients who have less than 200k income, credit amount between 150k-450k and have annuity less than 50k are more likely to default on their loans.
- For the sake of brevity, the plots on cell 112 gives us the following information: Clients who are most likely to default on their loans are ‘Married, working people with secondary education who are living in a house or apartment’
- Another interesting observation we can make is that clients aged between 35-65 who are either self-employed or work in type 3 business entities are more likely to default as seen below

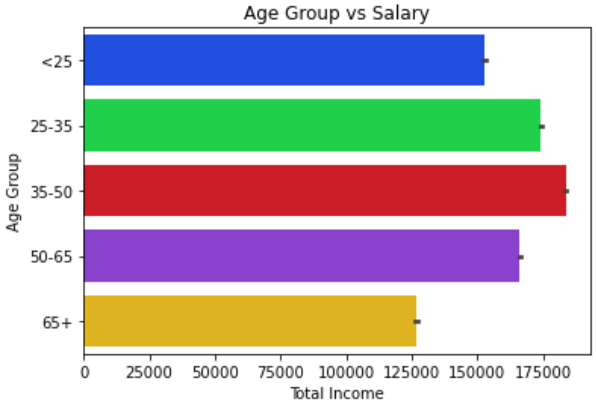




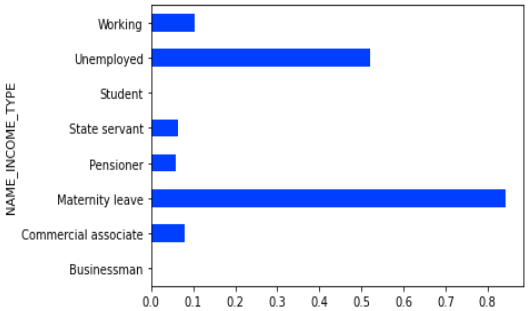
2.2 Bivariate Analysis

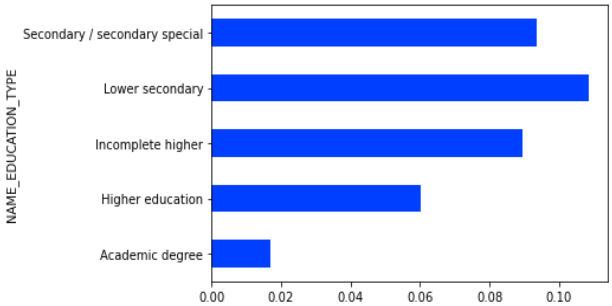
- Here we see that the age does not play a significant role in the target variable. We can infer that as the age increases the likelihood of loan repayment is increased. This could be attributed to higher incomes.
- This is further reinforced by the plot between income and ag

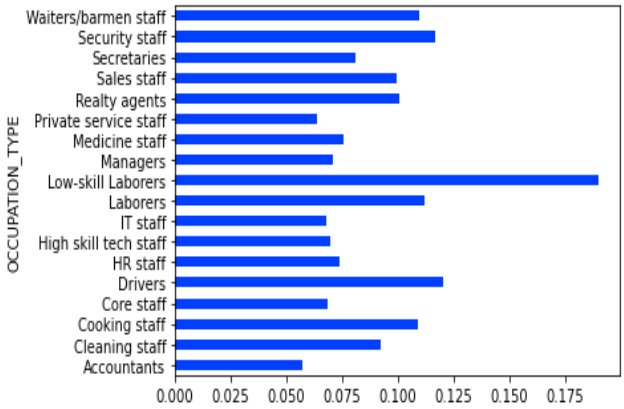


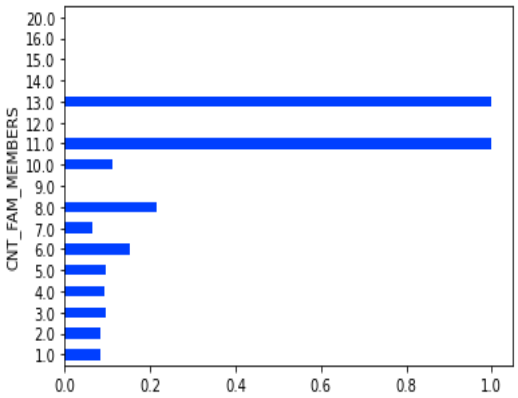


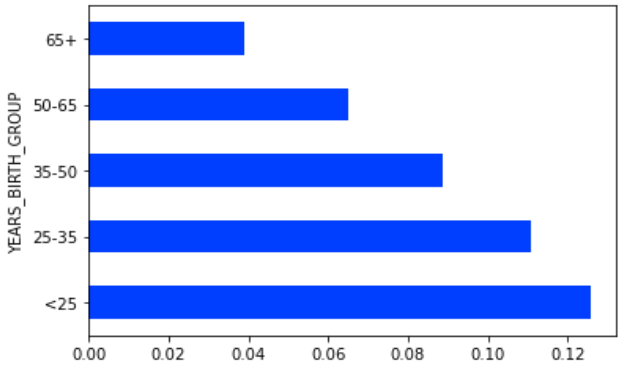
- In this next section we run a loop over certain columns by grouping them and plotting against the Target variable. This tell us that which client attributes are more likely to default.
    - We see that clients on maternity leave and clients who are unemployed who also have lower/secondary education have a high propensity to default on loans
    - Clients who are single or are in civil marriages are only slightly more likely to default than separated or married
    - Clients who are low skill labourers and clients living with parents or rented apartments are likely to default
    - Clients with more than 8 family members are likely to default
    - Clients who have more than 1.5M income are more likely to default. This is an interesting observation and will have to be explored further
    - Clients less than the age of 35 are more likely to default, which is to be expected
    - Clients who take our loans for good priced higher than 2.5M are more likely to default
- Some of the graphs plotted are shown below

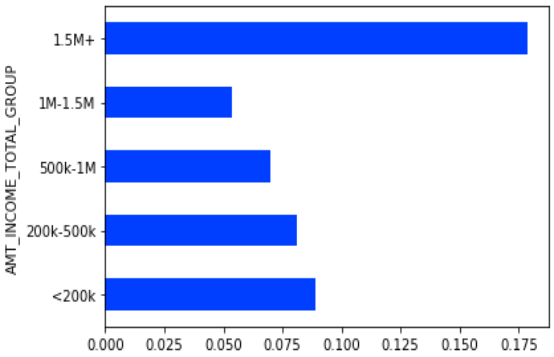












2.3 Multivariate Analysis

In this section we have several heatmaps with multiple indexes and columns to give us a thorough understanding of which factor form string correlations, and this will help us identify the driving variables for loan default along with our bivariate analysis.

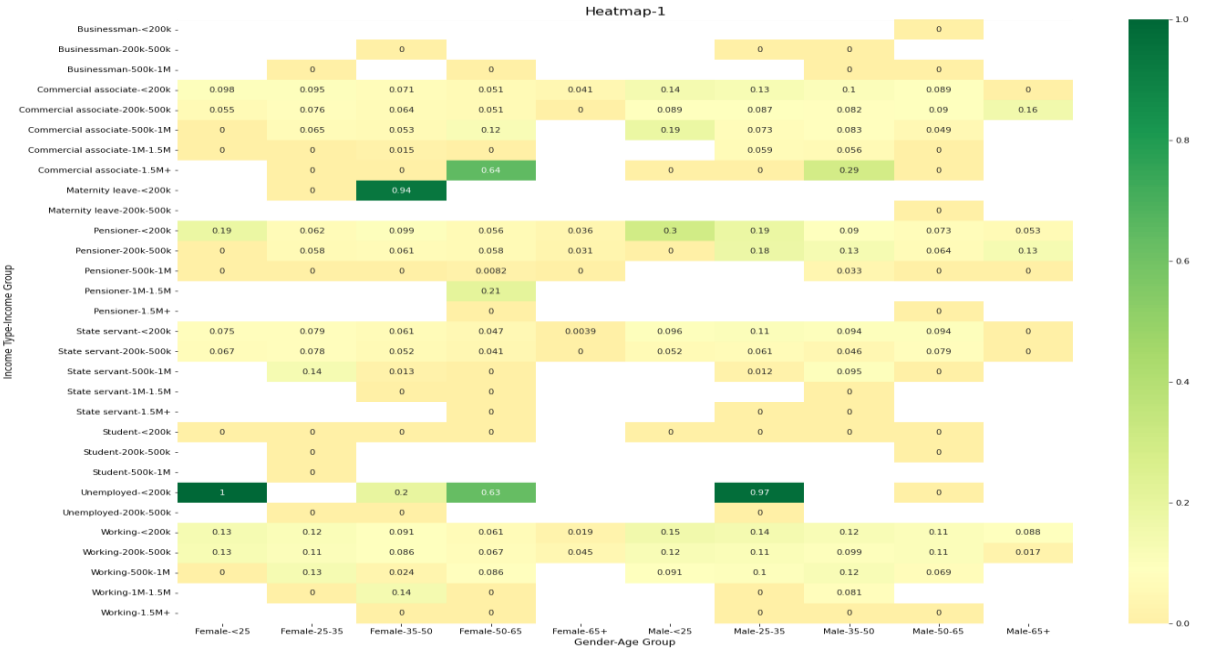
- Not all the heat maps will be presented in this report for the sake of conciseness, but all the inferences drawn from the various heatmaps will be discussed
- Some of the heat maps are given below. We can draw the following inferences:

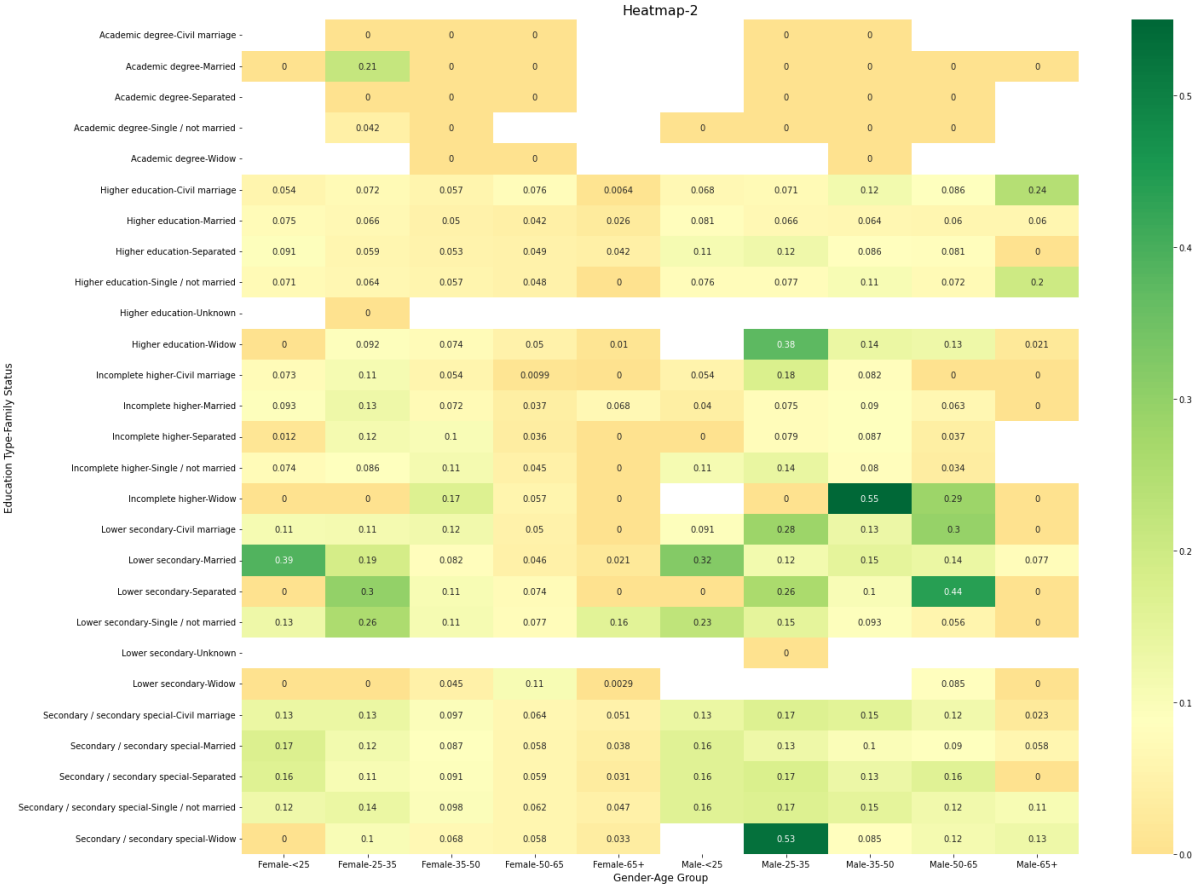
Correlations Heatmap -1	Unemployed<200k	Maternity Leave <200k	Commercial associate 1.5M+
Female<25	Very High		
Female 35-50		Very High	
Female 50-65	High		High
Male 25-35	Very High		

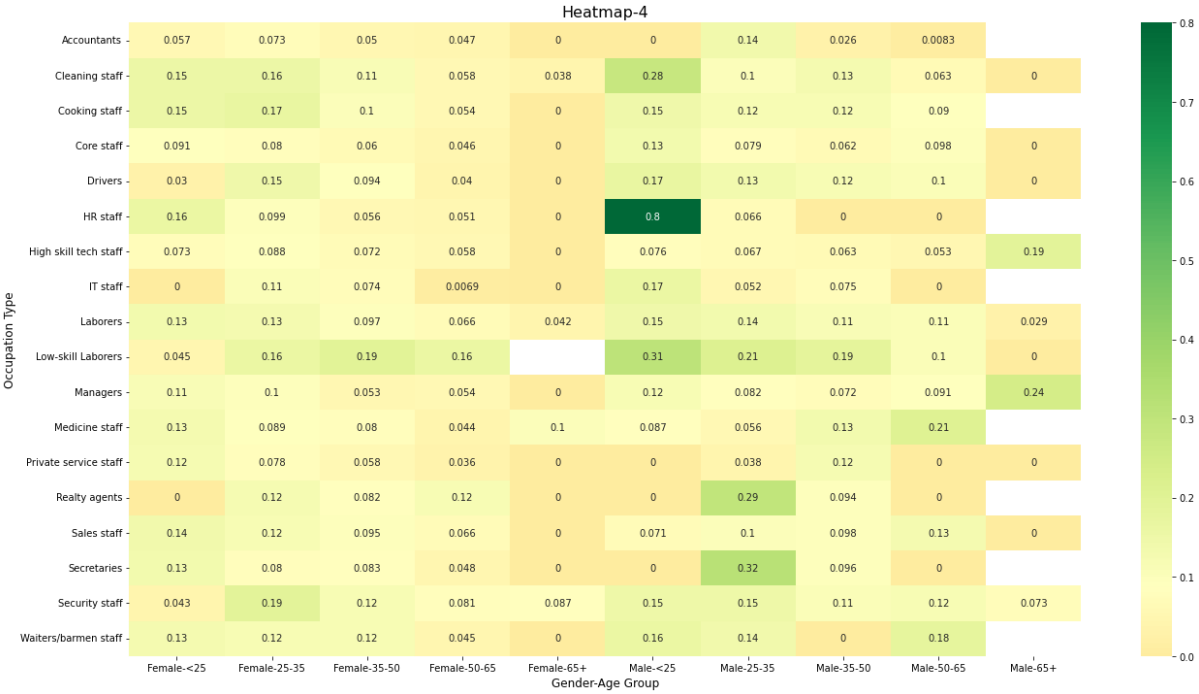
Correlations Heatmap -2	Lower Secondary – Married	Lower Secondary-Separated	Lower Secondary-Single	Secondary -Widow	Higher education-Widow	Incomplete higher -Widow
Female<25	Very High					
Female 35-50		High	High			
Female 50-65	High		High			

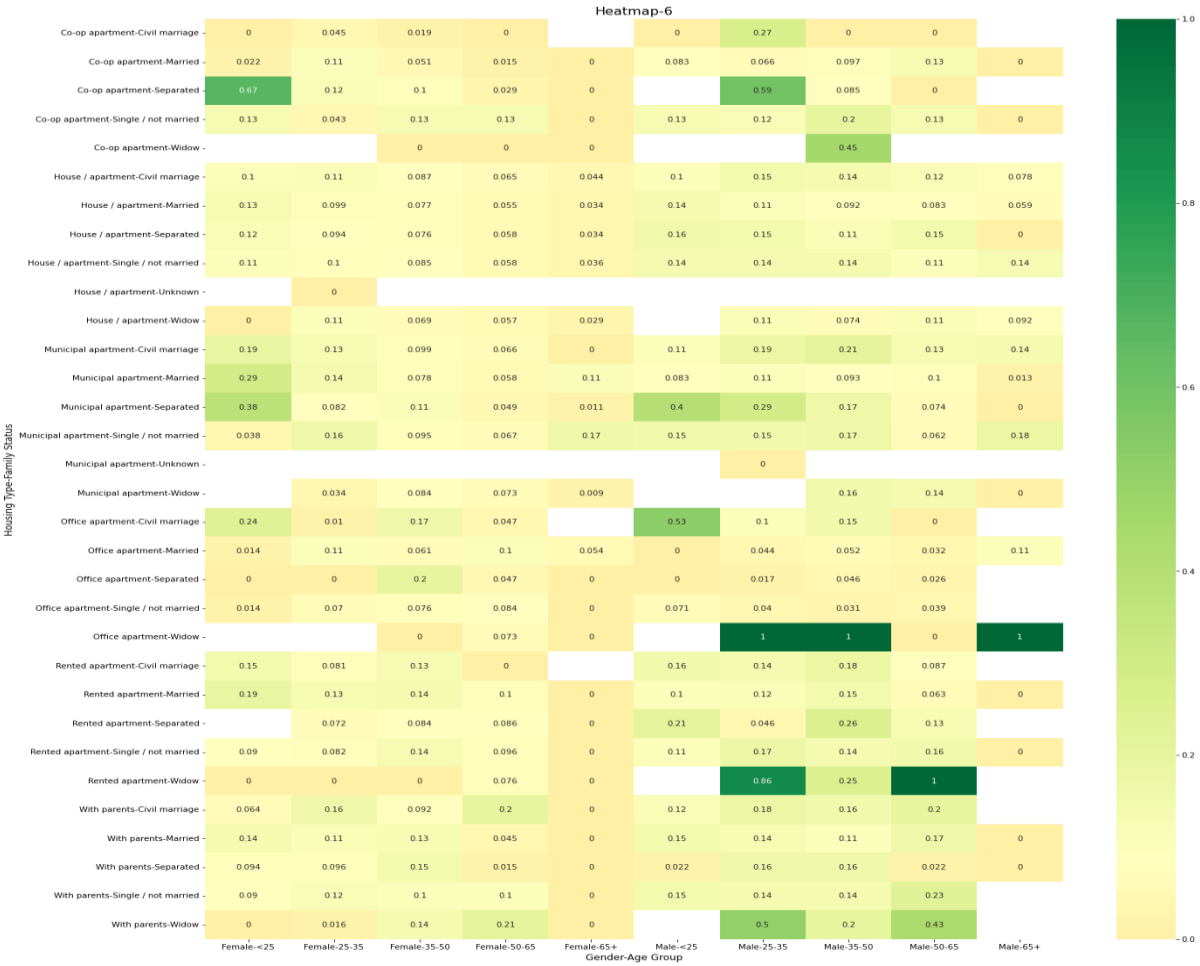
Male 25-35		High		Very High	Very High	
Male <25	High					
Male 35-50						Very High
Male 50-65		Very High				High

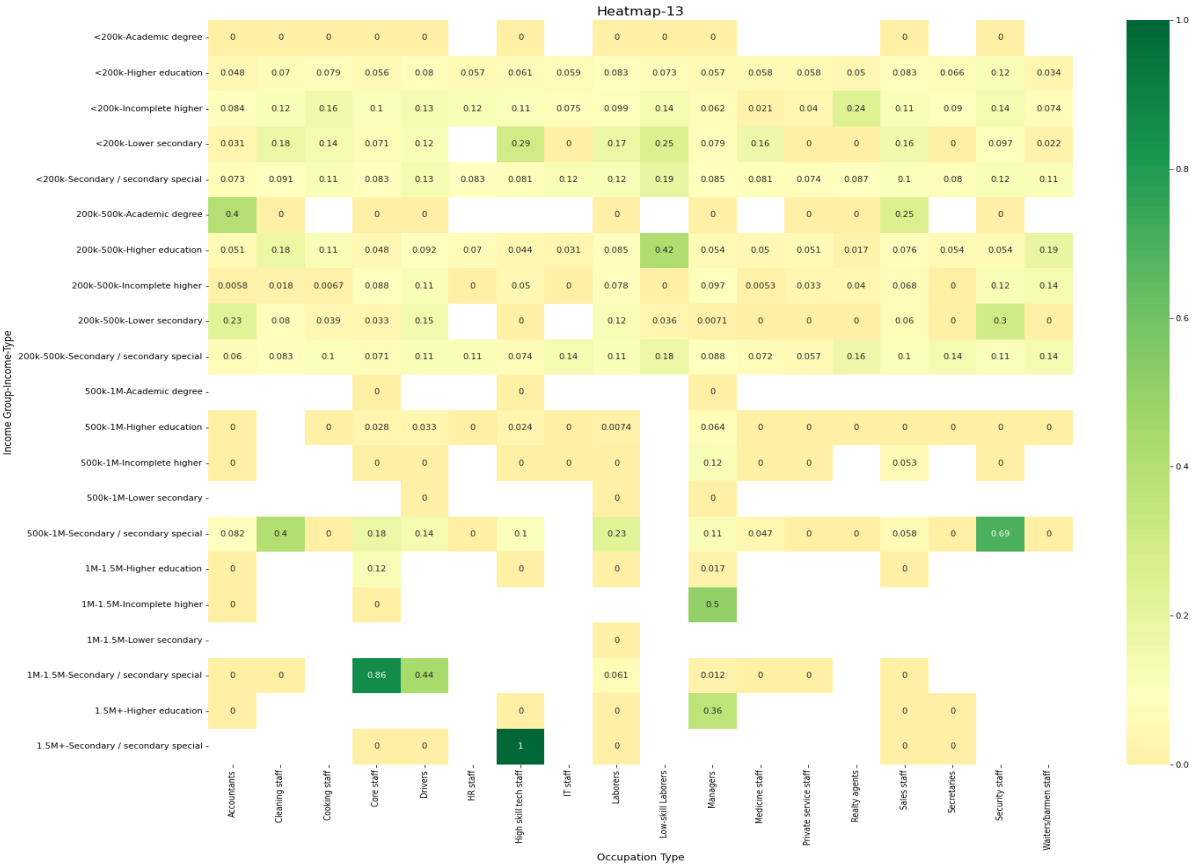
- For heatmap 3 we can see that both males and females who have more than 8 family members (married or not married) have very likely to default on their loans.
    - We can also see from the same map that males above 65 year who are widowed or separated are highly likely to default
- From heatmap 4 we infer males less than 25 years working as HR staff are very likely to have payment issues which can lead to default. Once can assume that income levels for young HR professions is low
    - From the same map we see that males below 25 years have the most difficulty in payments especially those working in low-income occupations such as cleaning staff, low skill labour, etc.
- From heatmap 6 we can see that both females below 25 and males between 25-35 living in co-op apartments who are also separated are highly likely to default
    - Similarly, males ranging from 25-65 who are widows either living in office apartments or rented apartments are very likely to default
    - This is further reinforced by heatmap 7
- Heatmap 8 gives us an excellent idea of income levels and income types:
    - We see a very high correlation between cleaning staff earning between 500k-1M, likely to default
    - Core staff/Laborers who earn less than 200k and are on maternity leave are also likely to face difficulty in payments
- From heatmap 10 we can see that males between 25-35 years who are widowed and any status of previous application for loan are very likely to have payment difficulties
- Finally, from heatmap 12 we can see that likely defaulters will be older and have large difference between their income level and the price of goods for which they are applying the loan




Heatmap-2
Education Type-Family Status
Academic degree-Civil marriage 0 0 0 0 0
Academic degree-Married 0 0.21 0 0 0 0 0
Academic degree-Separated 0 0 0 0 0
Academic degree-Single / not married 0.042 0 0 0 0
Academic degree-Widow 0 0
Higher education-Civil marriage 0.054 0.072 0.057 0.076 0.0064 0.068 0.071 0.12 0.086 0.24
Higher education-Married 0.075 0.066 0.05 0.042 0.026 0.081 0.066 0.064 0.06 0.06
Higher education-Separated 0.091 0.059 0.053 0.049 0.042 0.11 0.12 0.086 0.081 0
Higher education-Single / not married 0.071 0.064 0.057 0.048 0 0.076 0.077 0.11 0.072 0.2
Higher education-Unknown 0
Higher education-Widow 0 0.092 0.074 0.05 0.01 0.38 0.14 0.13 0.021
Incomplete higher-Civil marriage 0.073 0.11 0.054 0.0099 0 0.054 0.18 0.082 0 0
Incomplete higher-Married 0.093 0.13 0.072 0.037 0.068 0.04 0.075 0.09 0.063 0
Incomplete higher-Separated 0.012 0.12 0.1 0.036 0 0 0.079 0.087 0.037
Incomplete higher-Single / not married 0.074 0.086 0.11 0.045 0 0.11 0.14 0.08 0.034
Incomplete higher-Widow 0 0 0.17 0.057 0 0 0.55 0.29 0
Lower secondary-Civil marriage 0.11 0.11 0.12 0.05 0 0.091 0.28 0.13 0.3 0
Lower secondary-Married 0.39 0.19 0.082 0.046 0.021 0.32 0.12 0.15 0.14 0.077
Lower secondary-Separated 0 0.3 0.11 0.074 0 0 0.26 0.1 0.44 0
Lower secondary-Single / not married 0.13 0.26 0.11 0.077 0.16 0.23 0.15 0.093 0.056 0
Lower secondary-Unknown 0
Lower secondary-Widow 0 0 0.045 0.11 0.0029 0.085 0
Secondary / secondary special-Civil marriage 0.13 0.13 0.097 0.064 0.051 0.13 0.17 0.15 0.12 0.023
Secondary / secondary special-Married 0.17 0.12 0.087 0.058 0.038 0.16 0.13 0.1 0.09 0.058
Secondary / secondary special-Separated 0.16 0.11 0.091 0.059 0.031 0.16 0.17 0.13 0.16 0
Secondary / secondary special-Single / not married 0.12 0.14 0.098 0.062 0.047 0.16 0.17 0.15 0.12 0.11
Secondary / secondary special-Widow 0 0.1 0.068 0.058 0.033 0.53 0.085 0.12 0.13
Female-<25 Female-25-35 Female-35-50 Female-50-65 Female-65+ Male-<25 Male-25-35 Male-35-50 Male-50-65 Male-65+
Gender-Age Group
0.0 0.1 0.2 0.3 0.4 0.5


Heatmap-4
Occupation Type
Accountants 0.057 0.073 0.05 0.047 0 0 0.14 0.026 0.0083
Cleaning staff 0.15 0.16 0.11 0.058 0.038 0.28 0.1 0.13 0.063 0
Cooking staff 0.15 0.17 0.1 0.054 0 0.15 0.12 0.12 0.09
Core staff 0.091 0.08 0.06 0.046 0 0.13 0.079 0.062 0.098 0
Drivers 0.03 0.15 0.094 0.04 0 0.17 0.13 0.12 0.1 0
HR staff 0.16 0.099 0.056 0.051 0 0.8 0.066 0 0
High skill tech staff 0.073 0.088 0.072 0.058 0 0.076 0.067 0.063 0.053 0.19
IT staff 0 0.11 0.074 0.0069 0 0.17 0.052 0.075 0
Laborers 0.13 0.13 0.097 0.066 0.042 0.15 0.14 0.11 0.11 0.029
Low-skill Laborers 0.045 0.16 0.19 0.16 0.31 0.21 0.19 0.1 0
Managers 0.11 0.1 0.053 0.054 0 0.12 0.082 0.072 0.091 0.24
Medicine staff 0.13 0.089 0.08 0.044 0.1 0.087 0.056 0.13 0.21
Private service staff 0.12 0.078 0.058 0.036 0 0 0.038 0.12 0 0
Realty agents 0 0.12 0.082 0.12 0 0 0.29 0.094 0
Sales staff 0.14 0.12 0.095 0.066 0 0.071 0.1 0.098 0.13 0
Secretaries 0.13 0.08 0.083 0.048 0 0 0.32 0.096 0
Security staff 0.043 0.19 0.12 0.081 0.087 0.15 0.15 0.11 0.12 0.073
Waiters/barmen staff 0.13 0.12 0.12 0.045 0 0.16 0.14 0 0.18
Female-<25 Female-25-35 Female-35-50 Female-50-65 Female-65+ Male-<25 Male-25-35 Male-35-50 Male-50-65 Male-65+
Gender-Age Group
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8


Heatmap-6
Housing Type-Family Status
Co-op apartment-Civil marriage
Co-op apartment-Married
Co-op apartment-Separated
Co-op apartment-Single / not married
Co-op apartment-Widow
House / apartment-Civil marriage
House / apartment-Married
House / apartment-Separated
House / apartment-Single / not married
House / apartment-Unknown
House / apartment-Widow
Municipal apartment-Civil marriage
Municipal apartment-Married
Municipal apartment-Separated
Municipal apartment-Single / not married
Municipal apartment-Unknown
Municipal apartment-Widow
Office apartment-Civil marriage
Office apartment-Married
Office apartment-Separated
Office apartment-Single / not married
Office apartment-Widow
Rented apartment-Civil marriage
Rented apartment-Married
Rented apartment-Separated
Rented apartment-Single / not married
Rented apartment-Widow
With parents-Civil marriage
With parents-Married
With parents-Separated
With parents-Single / not married
With parents-Widow
Female-<25
Female-25-35
Female-35-50
Female-50-65
Female-65+
Male-<25
Male-25-35
Male-35-50
Male-50-65
Male-65+
Gender-Age Group
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap-13
Income Group-Income-Type
<200k-Academic degree
<200k-Higher education
<200k-Incomplete higher
<200k-Lower secondary
<200k-Secondary / secondary special
200k-500k-Academic degree
200k-500k-Higher education
200k-500k-Incomplete higher
200k-500k-Lower secondary
200k-500k-Secondary / secondary special
500k-1M-Academic degree
500k-1M-Higher education
500k-1M-Incomplete higher
500k-1M-Lower secondary
500k-1M-Secondary / secondary special
1M-1.5M-Higher education
1M-1.5M-Incomplete higher
1M-1.5M-Lower secondary
1M-1.5M-Secondary / secondary special
1.5M+-Higher education
1.5M+-Secondary / secondary special
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
HR staff
High skill tech staff
IT staff
Laborers
Low-skill Laborers
Managers
Medicine staff
Private service staff
Realty agents
Sales staff
Secretaries
Security staff
Waiters/barmen staff
Occupation Type
1.0
0.8
0.6
0.4
0.2
0.0

2.4 Recommendations

Based on the various types of analysis performed on the data I have come to the following conclusions:

- The top 5 driver variables for consideration when assessing loan applications which are likely to default are:
    - Gender-Age Group
    - Family/Marital Status
    - Occupation type
    - Education type
    - Income group/Income type
- To identify the top 3 correlations, it is important to make certain assumptions:
    - Since there are many variables involved in loan default it will be prudent to consider a pair of variables as a single entity while making decision. E.g. – Gender and age group should be considered as a single entity
    - Not all clients who are facing payment difficulties will result in default. Once the likelihood of default is ascertained a deeper investigation will have to be done to determine the appropriate action, such as rejecting application, reducing the loan amount, higher interest rates etc.
- Keeping these things in mind, I conclude that the top 3 correlations are:
    - **Gender – Age Group --- Income Type – Income Group**
    - **Gender – Age Group --- Education Type – Family Status**
    - **Occupation --- Income Group – Education Type**
- These variable groups have been chosen as I believe they make the most logical and business sense in this scenario. These variables are easily visible and simple to analyse withing the loan application.