PURPOSE OF CASE STUDY:

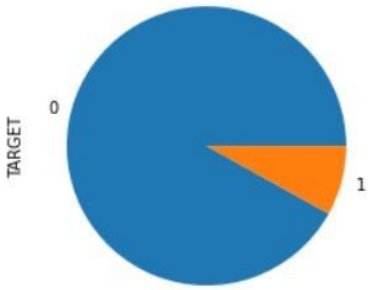
This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable 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.

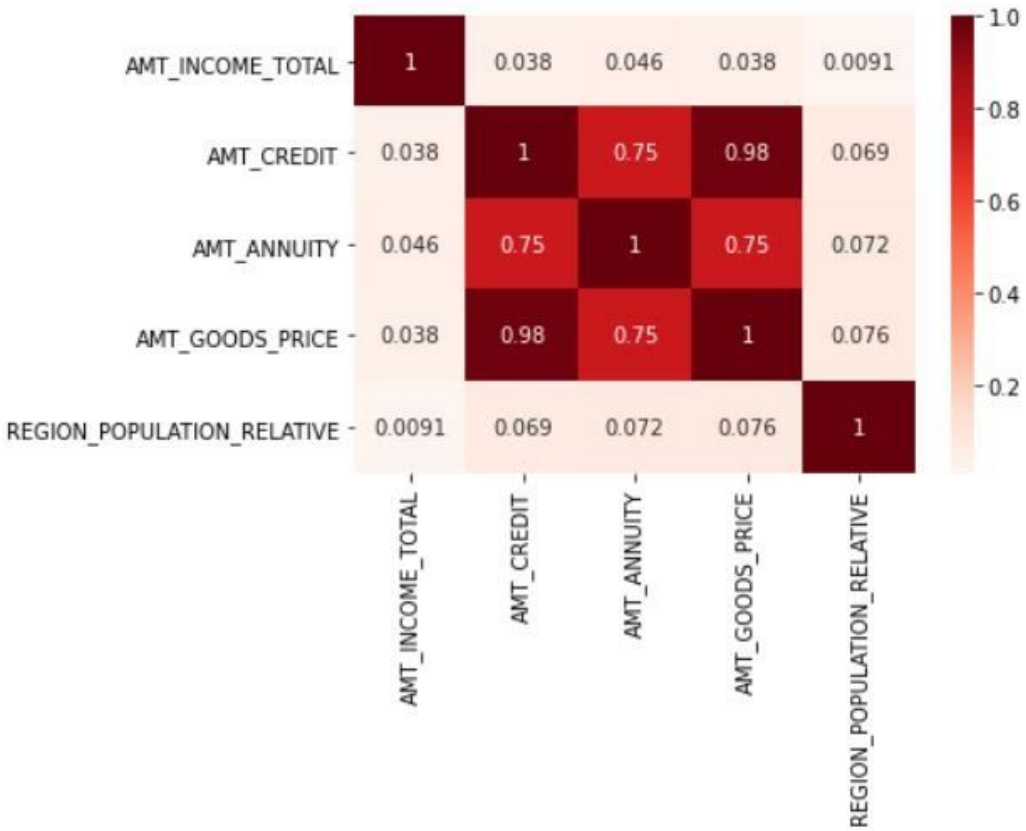
STEPS INVOLVED:

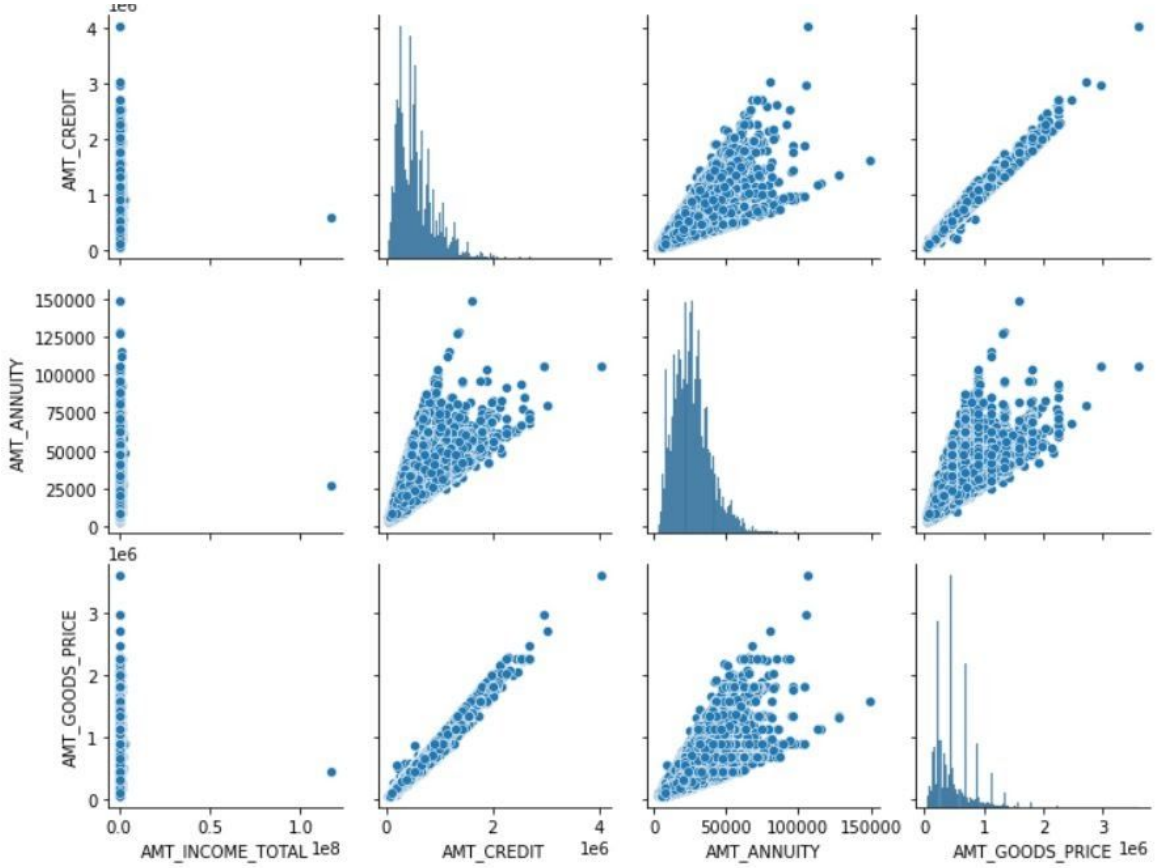
1. Data Understanding and Preparation:
    1.1. Check the structure of the data, read the csv file and use python commands to check the structure of data.
    1.2. Data quality check: Check whether all columns represent their respective data types. If not, perform operations to get them in desired data type format
2. Data Cleaning and Manipulation:
    2.1. Missing Value Treatment and Outlier Reporting. This is done for both ‘Application Data and Previous Application data sets
    2.2. Here in application data, all columns with more than 45% null values are dropped and for previous application data more than 40% null values are dropped.
    2.3. Binning of Continuous numerical variable to categorical variable is done. Eg: ‘Age’ and ‘Annual Income’ etc
    2.4. Binning of certain categorical variables which has redundant categories is done.
3. Data Analysis:
    3.1. Process of Univariate Analysis on Application Data.
    3.2. Process of Bivariate Analysis on Application Data.
    3.3. Process of Data cleaning and Manipulation on Previous Application Data
    3.4. Process of Univariate Analysis on merged data of Application Data and Previous Application Data
    3.5. Process of Bi and Multivariate Analysis on merged data of Application Data and Previous Application Data

SCREENSHOTS AND INFERENCES:



Inference: We see that around 8% from the data set have defaulted on loans and 92% have not defaulted



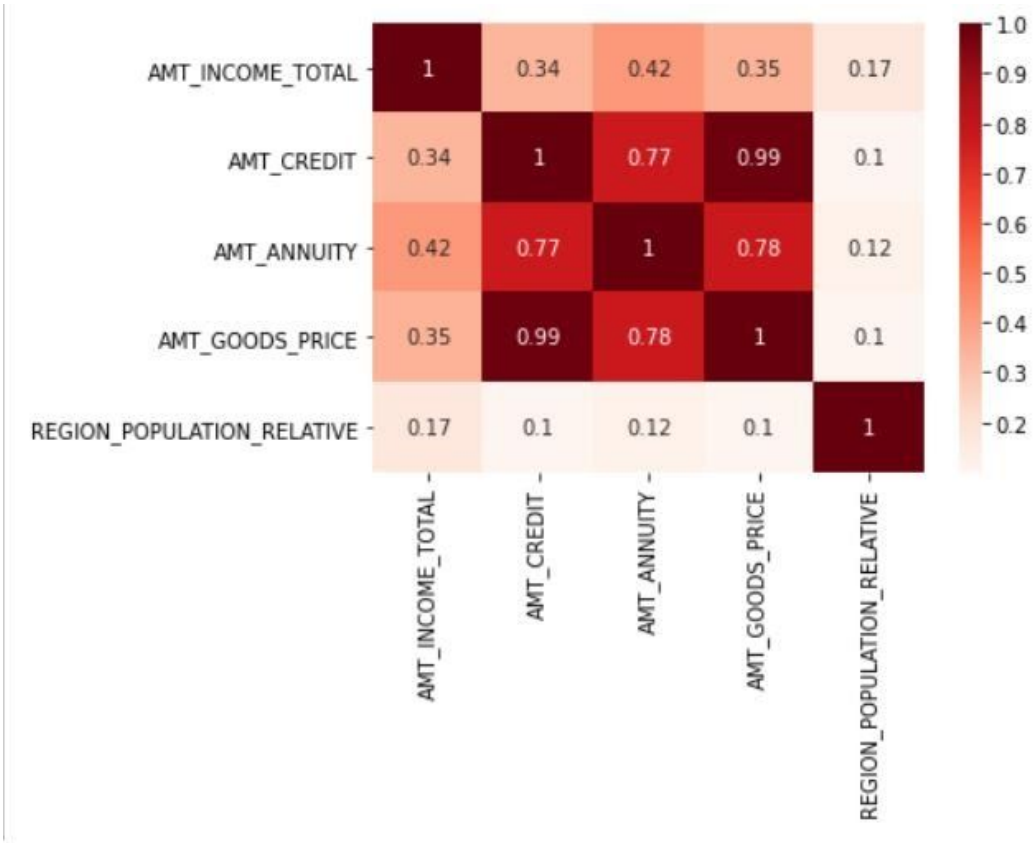


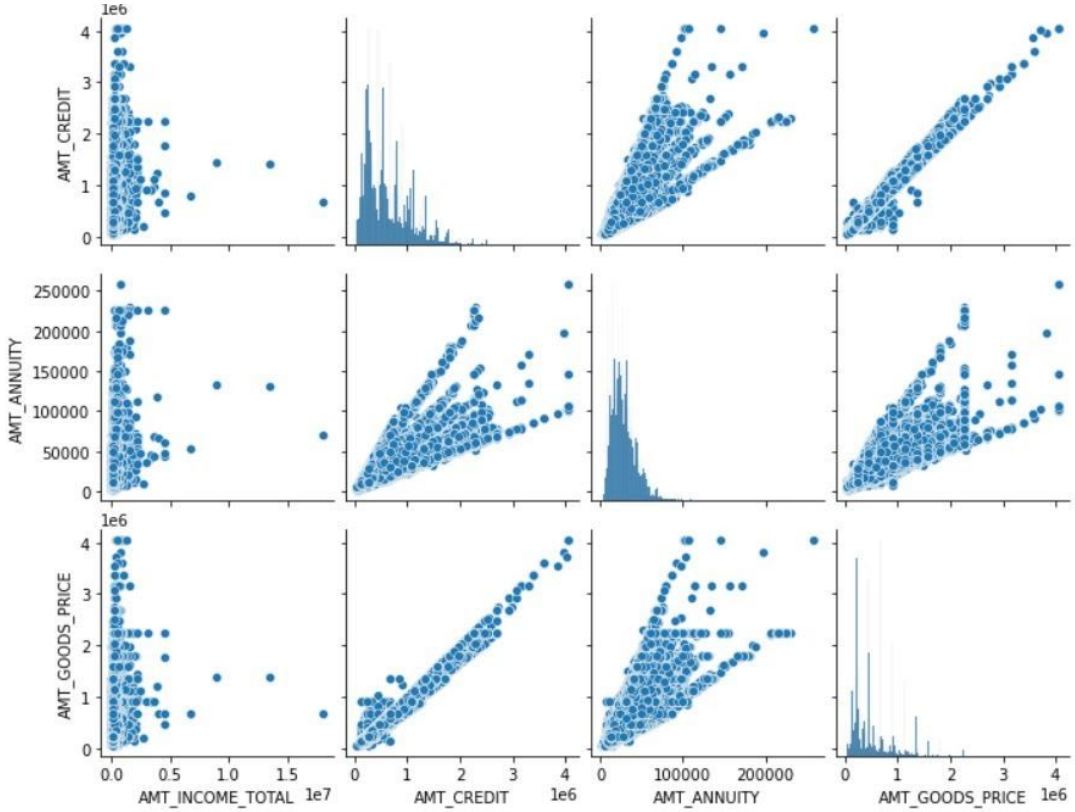
FROM ABOVE PLOTS INFERENCE For Defaulters:

From the above Heatmap and Pairplots we can draw the top 2 correlations. They are:

1. As AMT_CREDIT increases , the capacity to go for higher consumer loans(AMT_GOODS_PRICE) also increases.

2. As AMT_CREDIT increases, the 'AMT_ANNUITY' the capacity to go for higher installments in loan payback also increases.


AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
1
0.34
0.42
0.35
0.17
0.34
1
0.77
0.99
0.1
0.42
0.77
1
0.78
0.12
0.35
0.99
0.78
1
0.1
0.17
0.1
0.12
0.1
1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2



FROM ABOVE PLOTS IS INFERENCE For Non Defaulters:

From the above Heatmap and Pairplots we can draw the top 3 correlations. They are:

1. As AMT_CREDIT increases , the capacity to go for higher consumer loans(AMT_GOODS_PRICE) also increases.

2. As AMT_CREDIT increases, the 'AMT_ANNUITY' the capacity to go for higher installments in loan payback also increases.

3. As AMT_INCOME_TOTAL increases, 'AMT_ANNUITY' and 'AMT_CREDIT' also increases.

We can see that correlation matrix values for both defaulters and non defaulters is the "Same"

Inference for Outliers using Box Plots in Application Data set- and Reporting Them

1. The Following columns have outliers and need to dealt with accordingly

1. CNT_CHILDREN

2. AMT_INCOME_TOTAL

3. AMT_CREDIT

4. AMT_ANNUITY

5. AMT_GOODS_PRICE

6. DAYS_EMPLOYED

7. DAYS_REGISTRATION

8. CNT_FAM_MEMBERS

9. HOUR_APPR_PROCESS_START

10. OBS_30_CNT_SOCIAL_CIRCLE

11. DEF_30_SNT_SOCIAL_CIRCLE

12. OBS_60_CNT_SOCIAL_CIRCLE

13. DEF_60_CNT_SOCIAO_CIRCLE

14. DAYS_LAST_PHONE_CHANGE

15. AMT_REQ_CREDIT_BUREAU_HOUR

16. AMT_REQ_CREDIT_BUREAU_DAY

17. AMT_REQ_CREDIT_BUREAU_WEEK

18. AMT_REQ_CREDIT_BUREAU_MON

19. AMT_REQ_CREDIT_BUREAU_QRT

20. AMT_REQ_CREDIT_BUREAU_YEAR

Out of the Above columns below are the most important columns where we can drop rows with high outliers and further streamline our dataset for analysis. Such columns are:(atleast five colums to be reported)

1. CNT_CHILDREN

2. AMT_INCOME_TOTAL

3. AMT_CREDIT

4. AMT_ANNUITY

5. AMT_GOODS_PRICE

6. DAYS_EMPLOYED

7. CNT_FAM_MEMBERS

8. HOUR_APPR_PROCESS_START

Inference for Outliers using Box Plots In Previous Data Set- and Reporting Them

1. The Following columns have outliers and need to dealt with accordingly:

1. AMT_ANNUITY

2. AMT_APPLICATION

3. AMT_CREDIT

4. AMT_GOODS_PRICE

6. HOUR_APPR_PROCESS_START

7. DAYS_DECISION

8. SELLERPLACE_AREA

9. CNT_PAYMENT

Out of the Above columns below are the most important columns where we can drop rows with high outliers and further streamline our dataset for analysis. Such columns are:(atleast five colums to be reported):

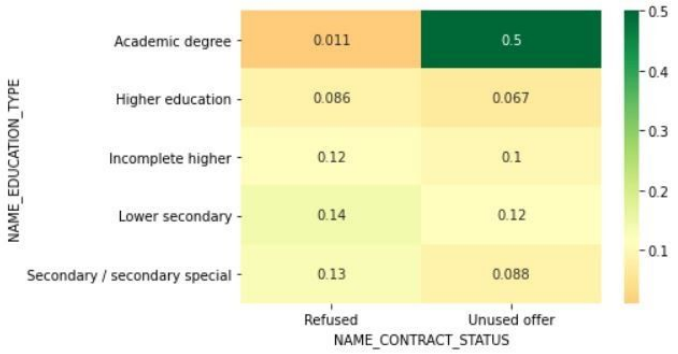
1. AMT_ANNUITY

2. AMT_APPLICATION

3. AMT_CREDIT

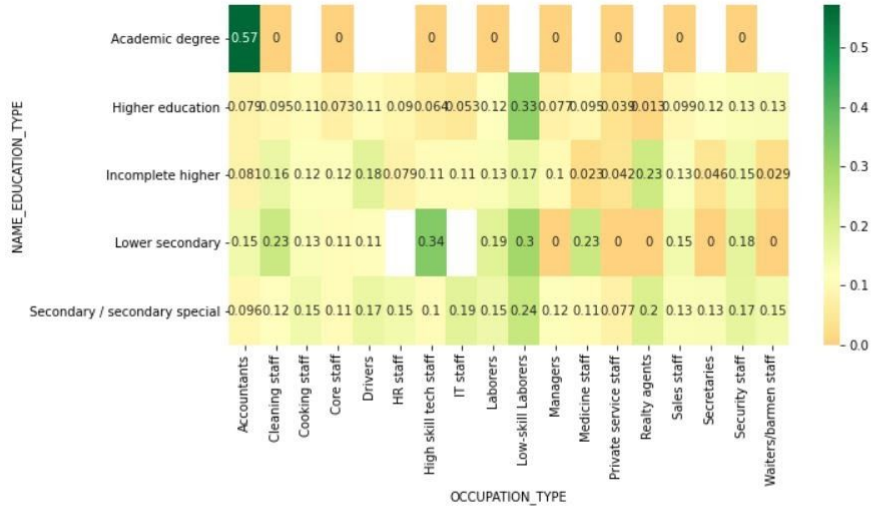
4. AMT_GOODS_PRICE

5. CNT_PAYMENT



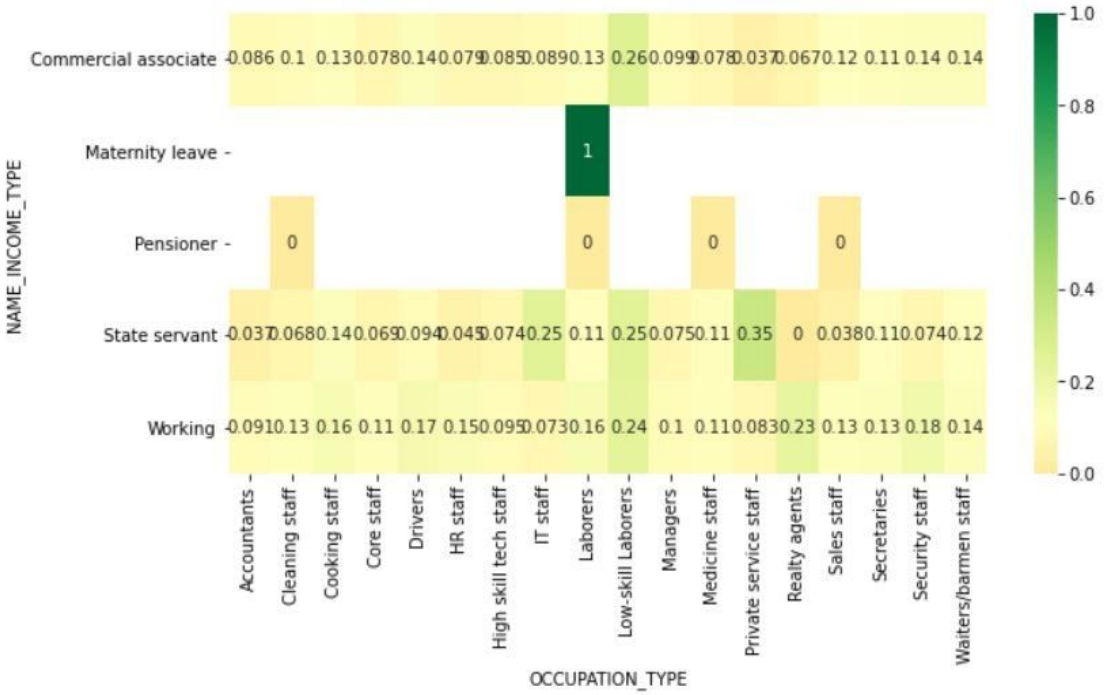
Inference from above plot:

It can be seen that people who have not taken the loan in the previous appliation and with an academic degree are the ones defaulting on loan.



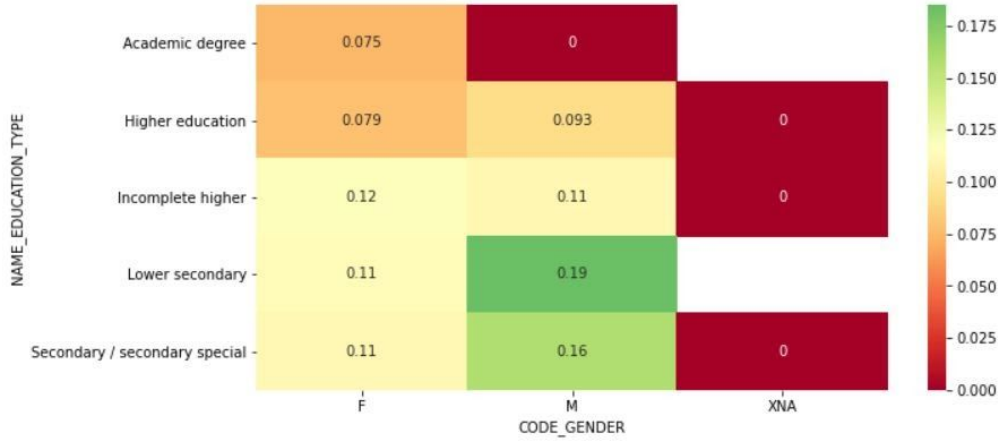
Inference from above plot:

Can see that Accountants will Academic degree are the highest people who have defaulted, Next High skill tech staff with lower secondary Education have defaulted



Inference from above plot:

Very much evident that 'Labourers' who are not working and on maternity leave have defaulted



Inference from above plot:

Very much evident that 'Males' with Lower Secondary and Secondary/Secondary Special Education types have defaulted