CREDIT EDA Case Study

Prepared and Submitted By:

- Harsh Sahajwani (harsh.sahajwani@gmail.com)
- Ashni Damaria (ashni.damaria@gmail.com)

Table of Contents

Problem Statement 1

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.

Data Understanding

- This dataset has 3 files as explained below:
  - 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
  - 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
  - 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Problem Statement 2

Results Expected by Learners

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
- Identify the missing data and use the appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.
- You need to submit one Jupyter notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

Solution

Solution Approach

- Working with Application Data
  - Data Cleaning
    - Finding & Treating Missing Data
    - Discarding Columns, not useful for Analysis
    - Deriving New Columns
    - Data Type Treatment
  - Data Analysis
    - Imbalance and Split
    - Univariate Analysis
      - Numerical
      - Categorical
      - Business Results
    - Bivariate Analysis
      - Numerical to Numerical & Business Results
      - Categorical to Categorical & Business Results
    - Multivariate Analysis
      - Numerical to Numerical & Business Results
      - Categorical to Categorical & Business Results
      - Categorical to Numerical & Business Results
- Working with Previous Data
  - Merging and Splitting
- Working with Merged Data
  - Data Analysis
    - Numerical to Numerical & Business Results
    - Categorical to Categorical & Business Results
    - Categorical to Numerical & Business Results
- Results in Business Terms

Working with Application Data

Data Cleaning

Data Cleaning

- Find and Treating Missing Data
  - Drop unnecessary columns with very high null percentages (>40%)
    - 49 columns dropped
    - 73 columns remain.
  - Analyzing columns that still have null percentage > 10%
  - Final decisions taken for all these columns:

COLUMN	DECISION
OCCUPATION_TYPE	Create a new occupation_type as 'Others' and impute the null values.
EXT_SOURCE_3	Fill this column with mean (0.51) for null values.
AMT_REQ_CREDIT_BUREAU_HOUR	Ignore this column.
AMT_REQ_CREDIT_BUREAU_DAY	Ignore this column.
AMT_REQ_CREDIT_BUREAU_WEEK	Ignore this column.
AMT_REQ_CREDIT_BUREAU_MON	Fill this column with median (0.0) for null values.
AMT_REQ_CREDIT_BUREAU_QRT	Fill this column with median (0.0) for null values.
AMT_REQ_CREDIT_BUREAU_YEAR	Impute null values with median (1.0)

Data Cleaning

- Discarding Columns not useful for Analysis
  - 46 columns dropped
  - 27 columns remain.
- Derived New Columns
  - 1 column added, 1 removed
    - AGE added, derived from DAYS_BIRTH
    - DAYS_BIRTH dropped.
  - 1 new column added
    - NET_INCOME derived from AMT_INCOME_TOTAL and AMT_ANNUITY
  - So now total columns are 28
- Data Type Treatment
  - Data type of CNT_FAM_MEMBERS changed from float64 to int64
  - Data type of AMT_REQ_CREDIT_BUREAU_YEAR changed from float64 to int64

Working with Application Data

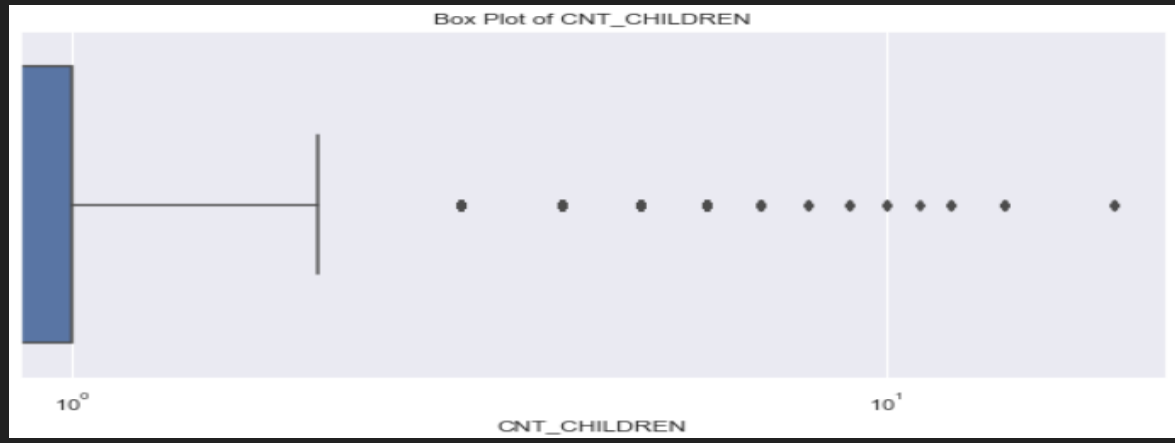
Data Analysis

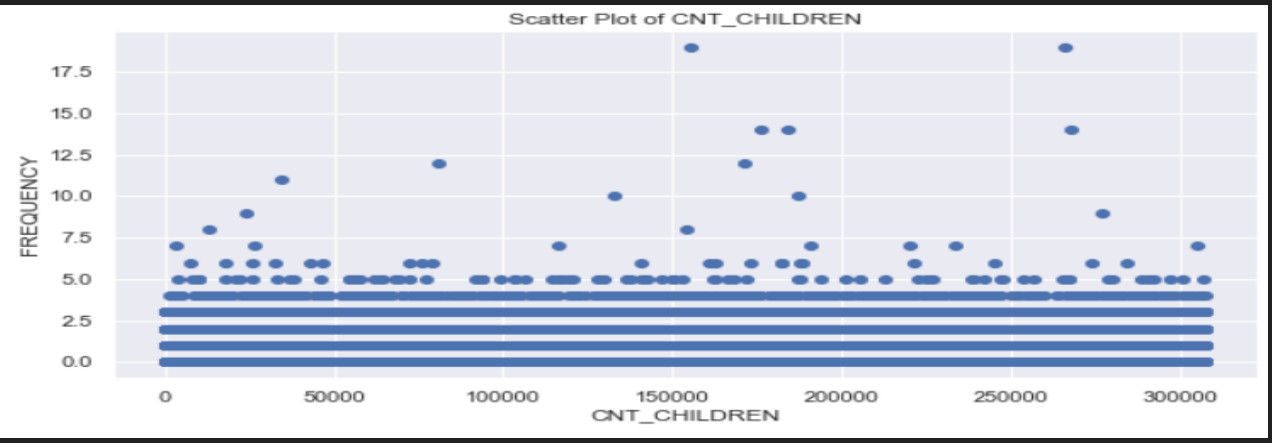
Imbalance and Split

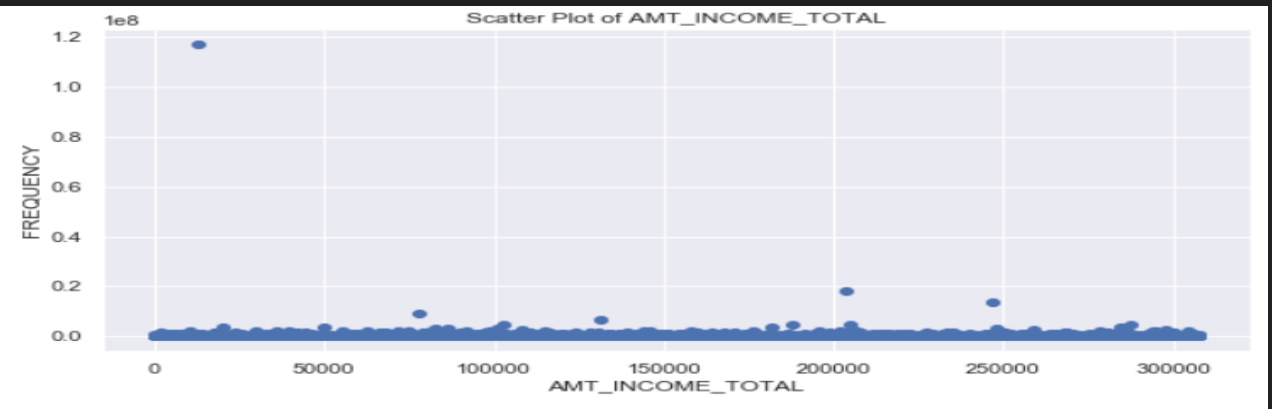
- Ratio of Imbalance
  - 8.07% represents 'Clients with Payment Difficulties'.
  - 91.03% represents 'Other Clients'.
  - Ratio of Imbalance is 11.39
  - 'Other Clients' are 11.39 times more than 'Clients with Payment Difficulties'.
- Split Data
  - Split application dataframe into 2 dataframes on the basis of TARGET value – target_1 and target_0
  - target_0 has 282684 rows and 28 columns
  - target_1 has 24825 rows and 28 columns

Univariate - Numerical

- OUTLIERS
  - Used box and scatter plots to visualize outliers
  - Some Examples:





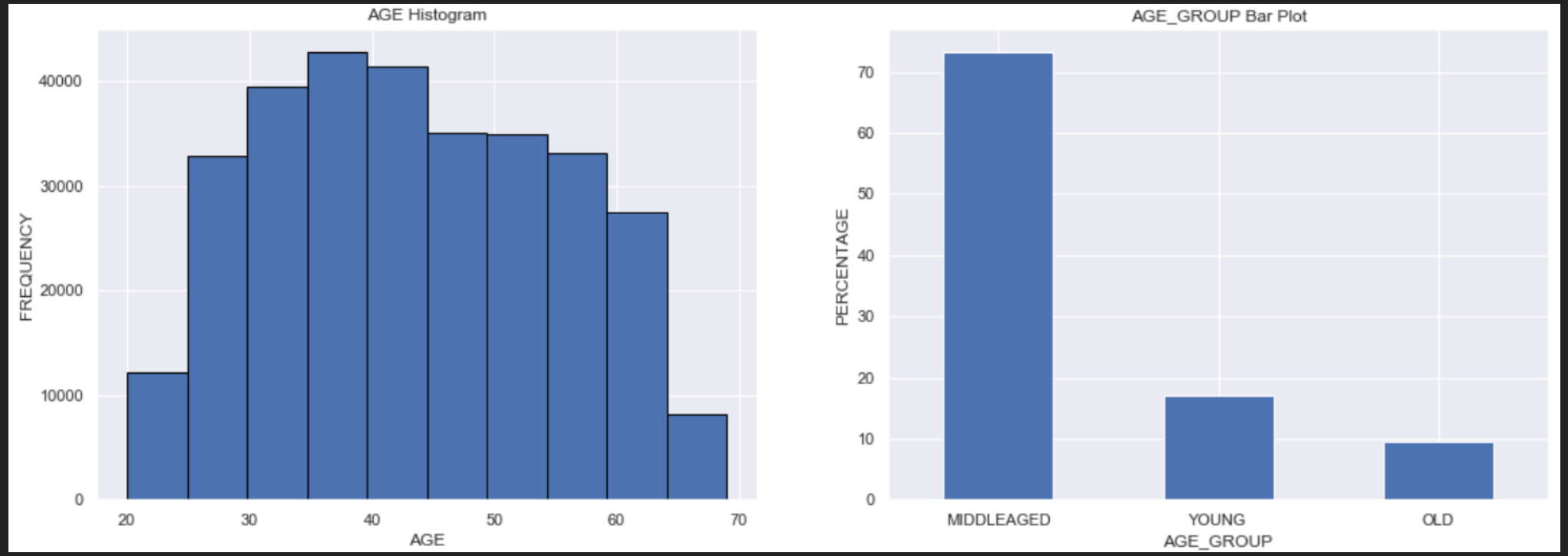


Univariate - Numerical

- OUTLIERS
  - Found Outliers using Box and Scatter Plots in 3 columns:
    - CNT_CHILDREN
    - AMT_INCOME_TOTAL
    - DAYS_EMPLOYED
  - Treated Outliers:
    - CNT_CHILDREN: Decided to not treat these outliers.
    - AMT_INCOME_TOTAL: Decided to not treat these outliers.
    - DAYS_EMPLOYED: Replaced outliers with median, which is 2219.0

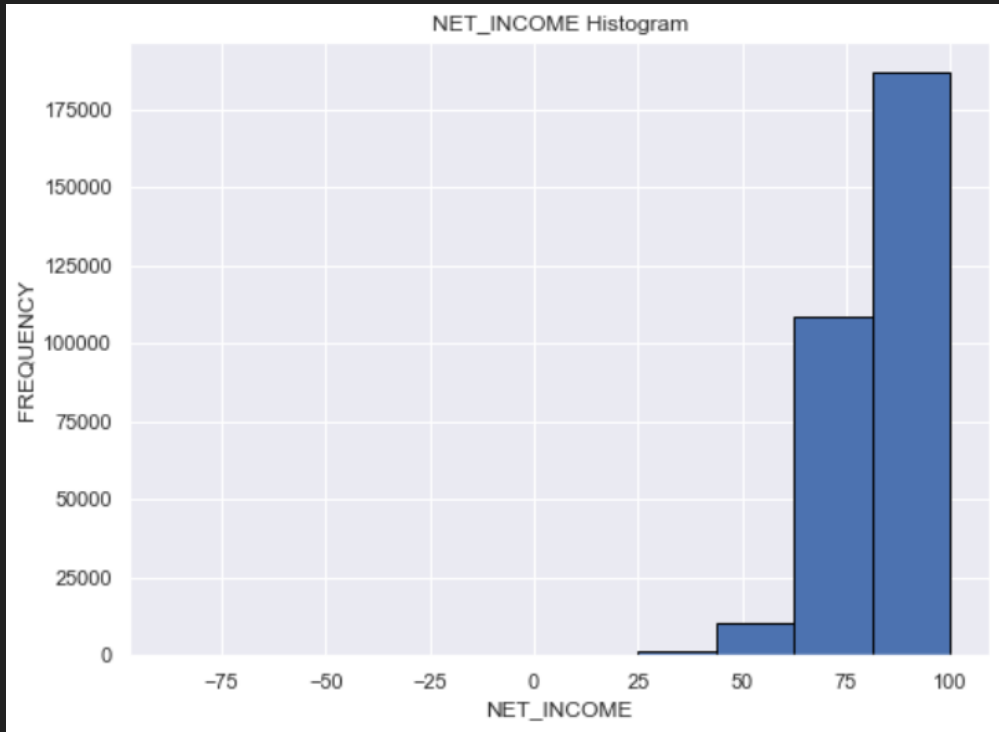
Univariate - Numerical

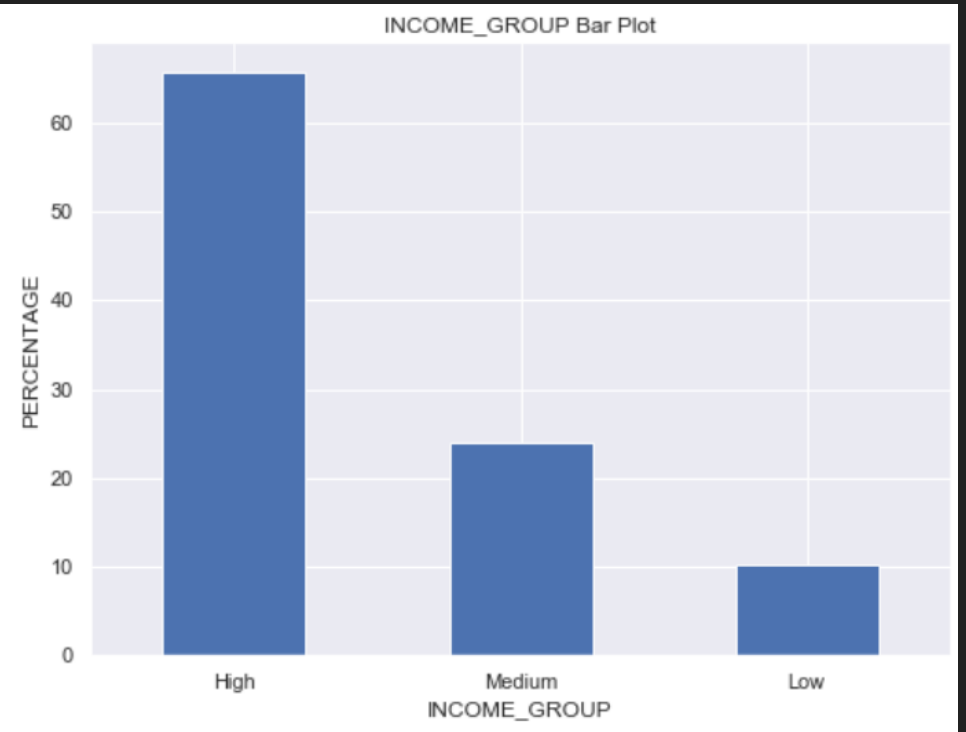
- Binning Continuous Variables
  - AGE binned to create AGE_GROUP with 3 categories
    - 0 to 30 – YOUNG
    - 30 TO 60 – MIDDLEAGED
    - 60+ - OLD



Univariate - Numerical

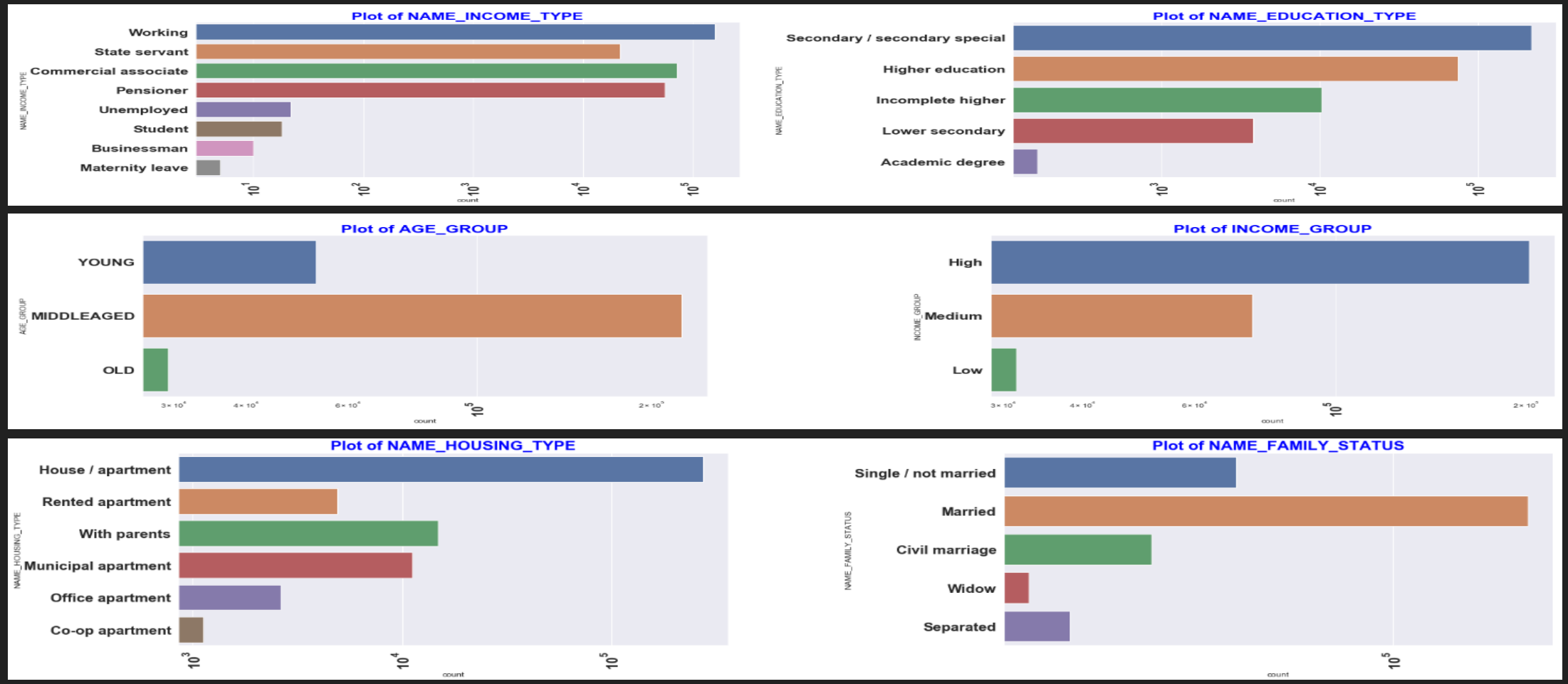
- Binning Continuous Variables
  - NET_INCOME binned to create INCOME_GROUP with 3 categories
    - UP TO 70% – LOW
    - 70% TO 80 % – MEDIUM
    - 80%+ - HIGH





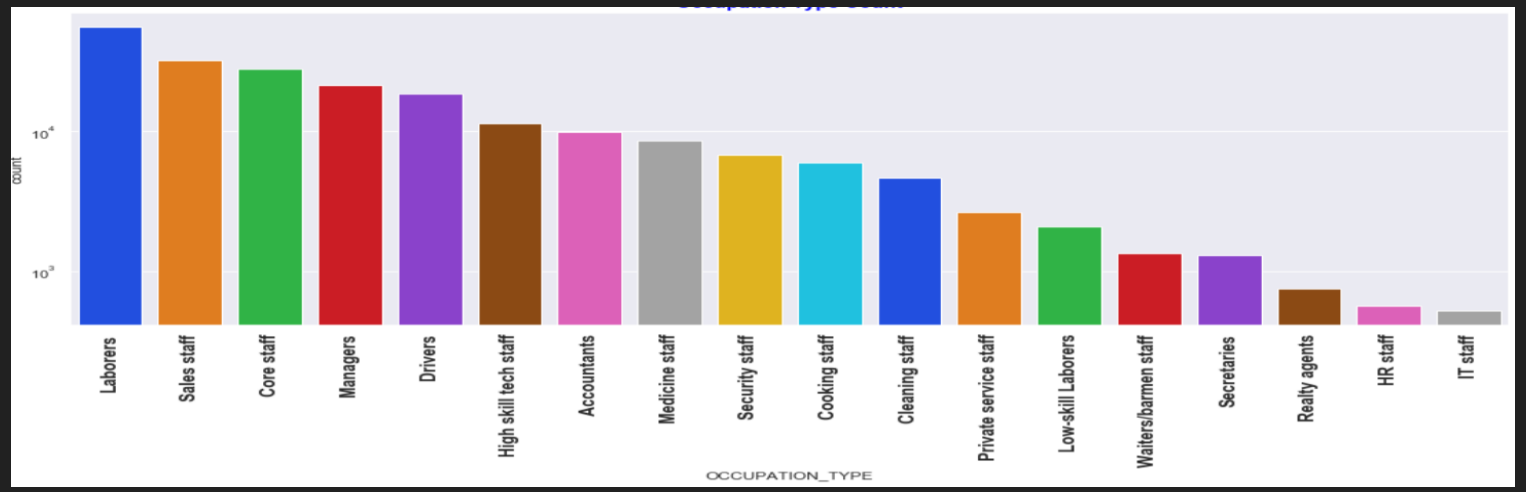
Univariate - Categorical

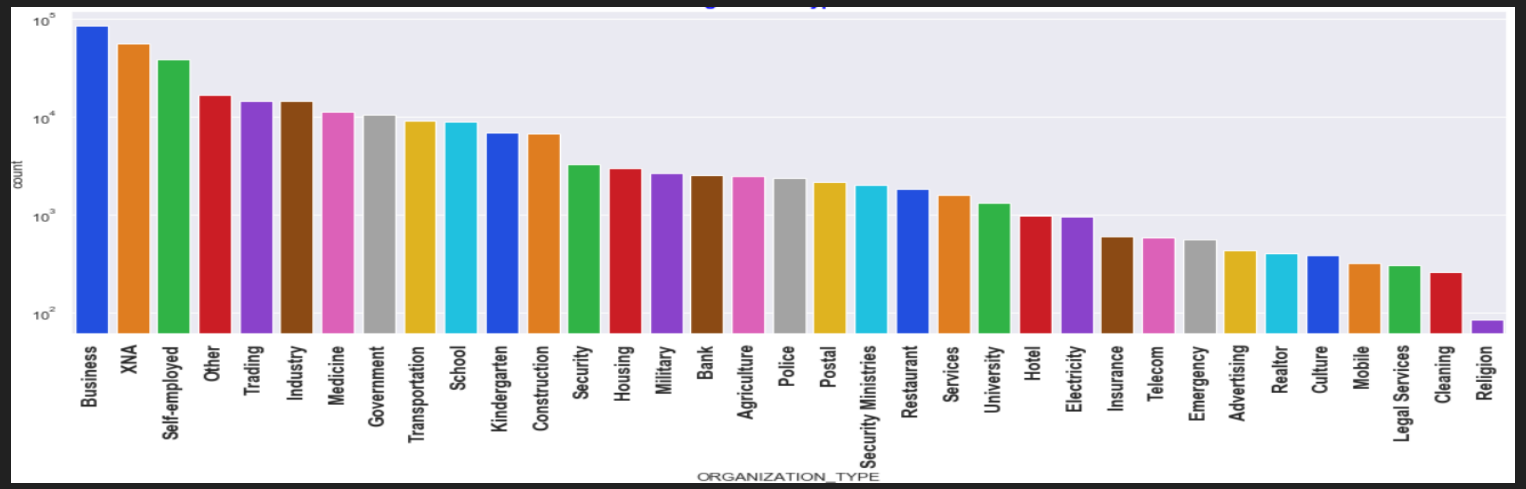
- Used Seaborn Countplots in for loop for categorical columns
- Some examples:



Univariate - Categorical

- Used Seaborn Countplots in for loop for categorical columns
- Some examples:



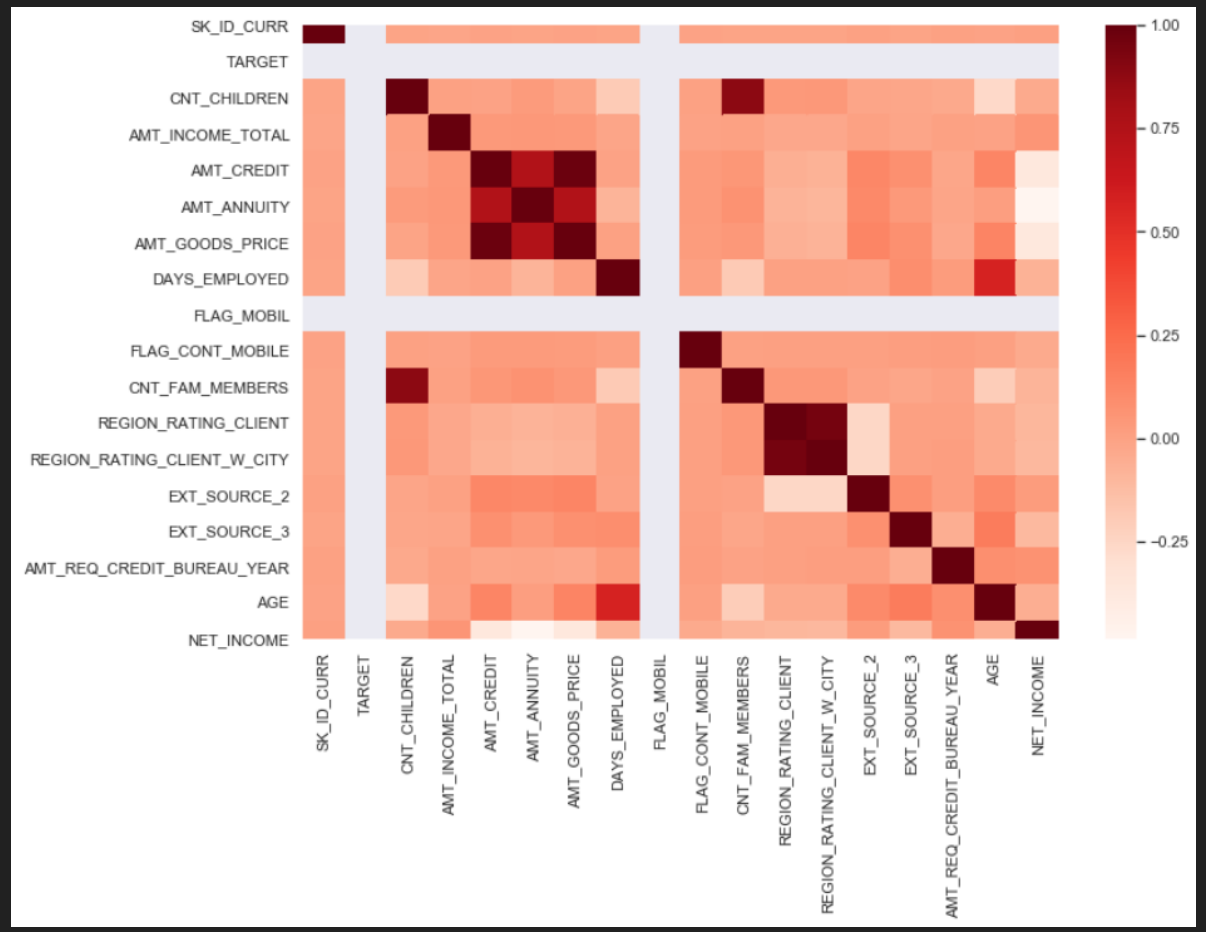


Results in Business Terms - Univariate

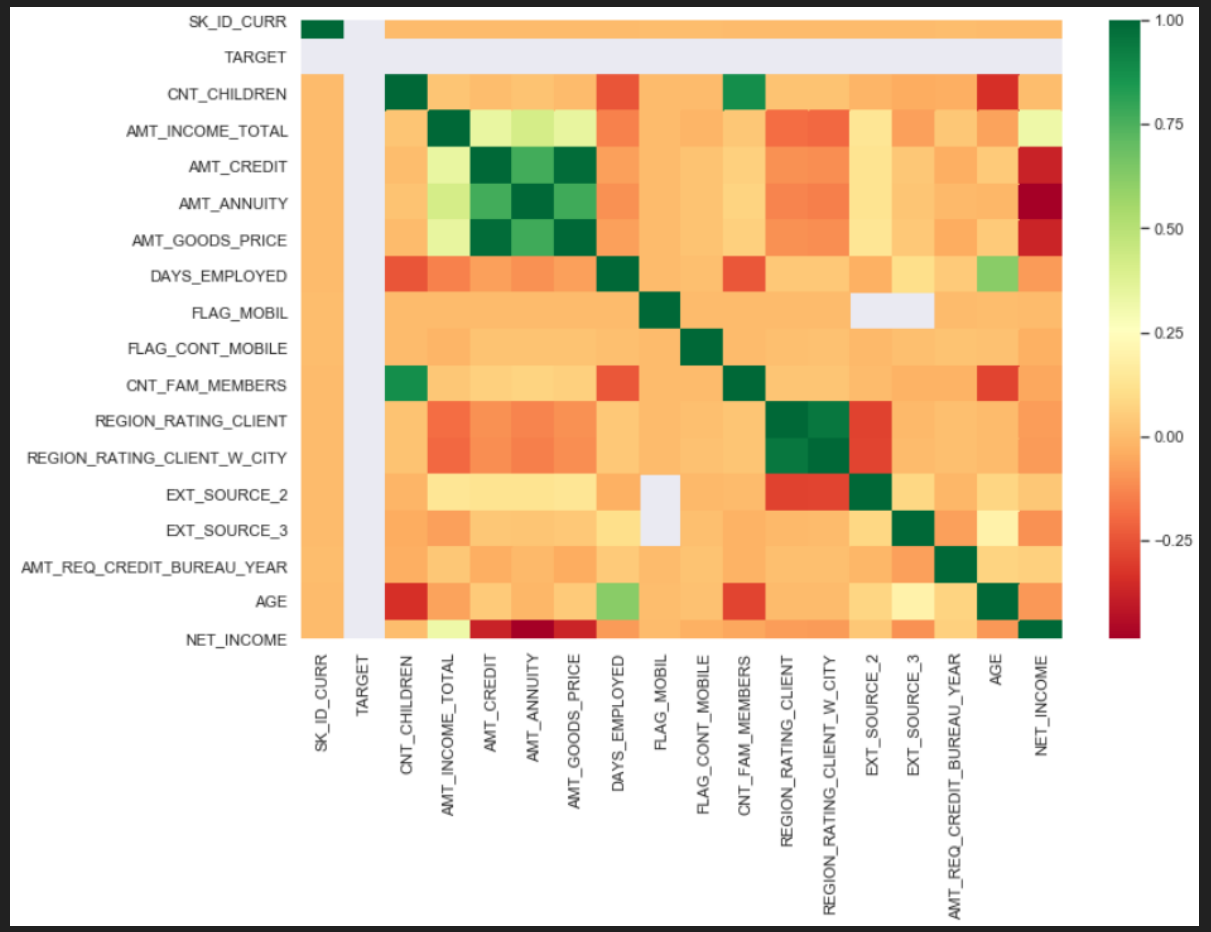
- Univariate Analysis Results
  - NAME_CONTRACT_TYPE: Most of the clients have opted for cash loans.
  - CODE_GENDER: There are more number of female clients as compared to male.
  - NAME_INCOME_TYPE: The count of Working clients is highest, followed by Commercial Associate and then by pensioners.
  - NAME_EDUCATION_TYPE: Mojority of clients have completed secondary / secondary special education, followed by 'Higher Education' and then by 'Incomplete higher'.
  - FLAG_OWN_CAR : Majority of the clients do not own a car.
  - FLAG_OWN_REALTY : Majority of the clients own a flat / appartment.
  - AGE_GROUP : Majority of the clients are middle aged (30 - 60).
  - INCOME_GROUP : Majority of the clients have high income (more than 80%).
  - NAME_HOUSING_TYPE: Majority of the clients live in House / apartment, followed by clients living with parents, followed by clients living in municipal apartment.
  - NAME_FAMILY_STATUS: Most of the clients are married followed by singles or not married and then followed by civil marriage.
  - OCCUPATION_TYPE: The top three OCCUPATION_TYPE in descending order are Laborers, Sales Staff, Core Staff.
  - ORGANIZATION_TYPE: Ignoring XNA, the top three ORGANIZATION_TYPE in descending order are Business, Self-Employed and Other.

Bivariate – Numerical to Numerical

- Used TARGET = 1 (Clients with Payment Difficulties) and TARGET = 0 (Other Clients)
- Plotted heat maps for numerical columns



TARGET = 1



TARGET = 0

Results in Business Terms – Bivariate Numerical to Numerical

- Used TARGET = 1 (Clients with Payment Difficulties) and TARGET = 0 (Other Clients)
- Found Top 10 pairs of columns with strong correlation
- There are different pairs of columns that impact clients in terms of payment difficulties and no difficulties

Sr. No.	COLUMN 1	COLUMN 2	CORRELATION
1	AMT_CREDIT	AMT_GOODS_PRICE	0.983
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.957
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885
4	AMT_ANNUITY	AMT_GOODS_PRICE	0.753
5	AMT_CREDIT	AMT_ANNUITY	0.752
6	AGE	DAYS_EMPLOED	0.286
7	AGE	EXT_SOURCE_3	0.172
8	AGE	AMT_GOODS_PRICE	0.136
9	AGE	AMT_CREDIT	0.135
10	AMT_GOODS_PRICE	EXT_SOURCE_2	0.131

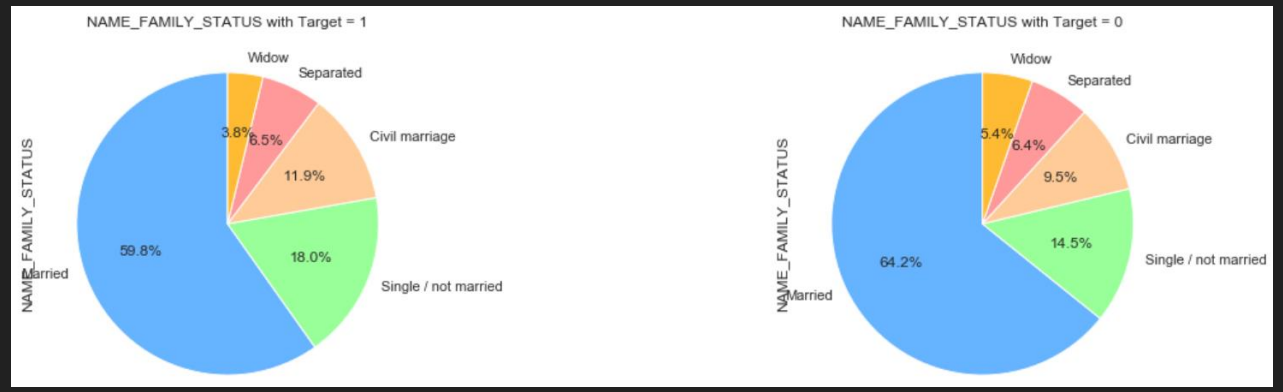
TARGET = 1

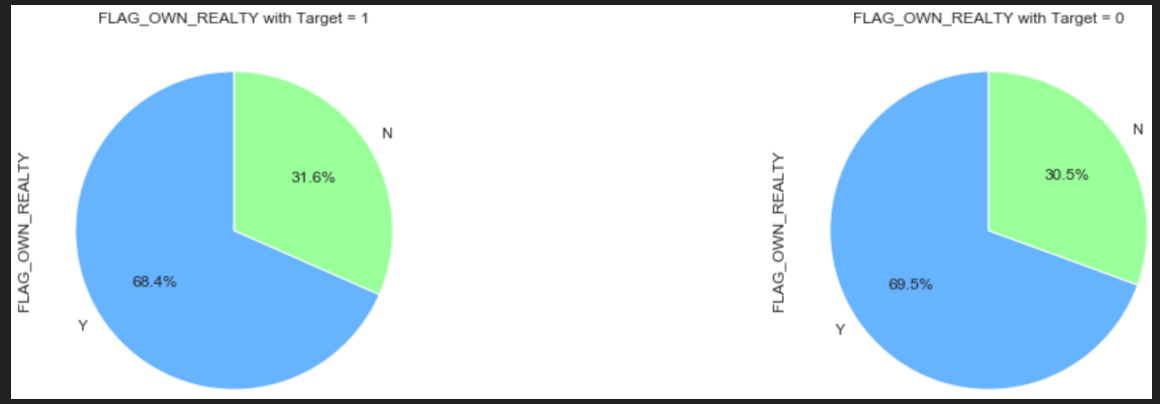
Sr. No.	COLUMN 1	COLUMN 2	CORRELATION
1	AMT_CREDIT	AMT_GOODS_PRICE	0.987
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879
4	AMT_ANNUITY	AMT_GOODS_PRICE	0.777
5	AMT_CREDIT	AMT_ANNUITY	0.771
6	AMT_INCOME_TOTAL	AMT_ANNUITY	0.419
7	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.349
8	AMT_INCOME_TOTAL	AMT_CREDIT	0.343
9	AGE	DAYS_EMPLOYED	0.242
10	AGE	EXT_SOURCE_3	0.197

TARGET = 0

Bivariate – Categorical to Categorical

- Used Matplotlib Pie Plots in for loop for categorical columns
- Some examples:



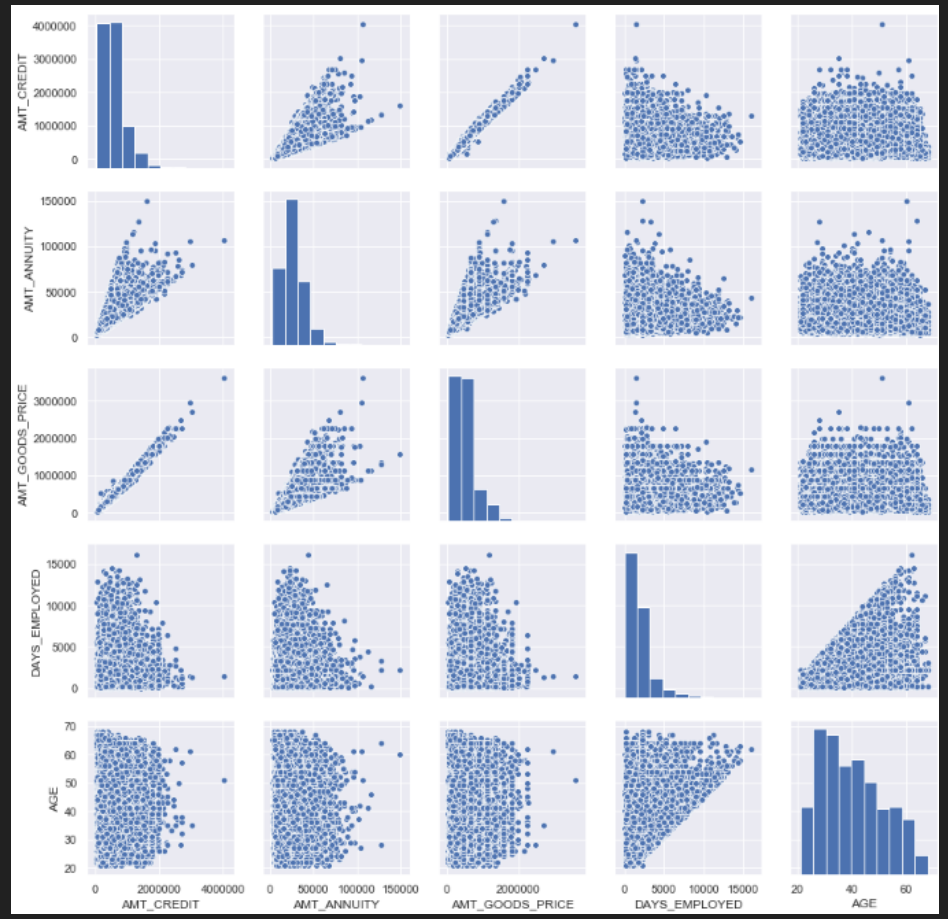


Results in Business Terms – Bivariate Categorical to Categorical

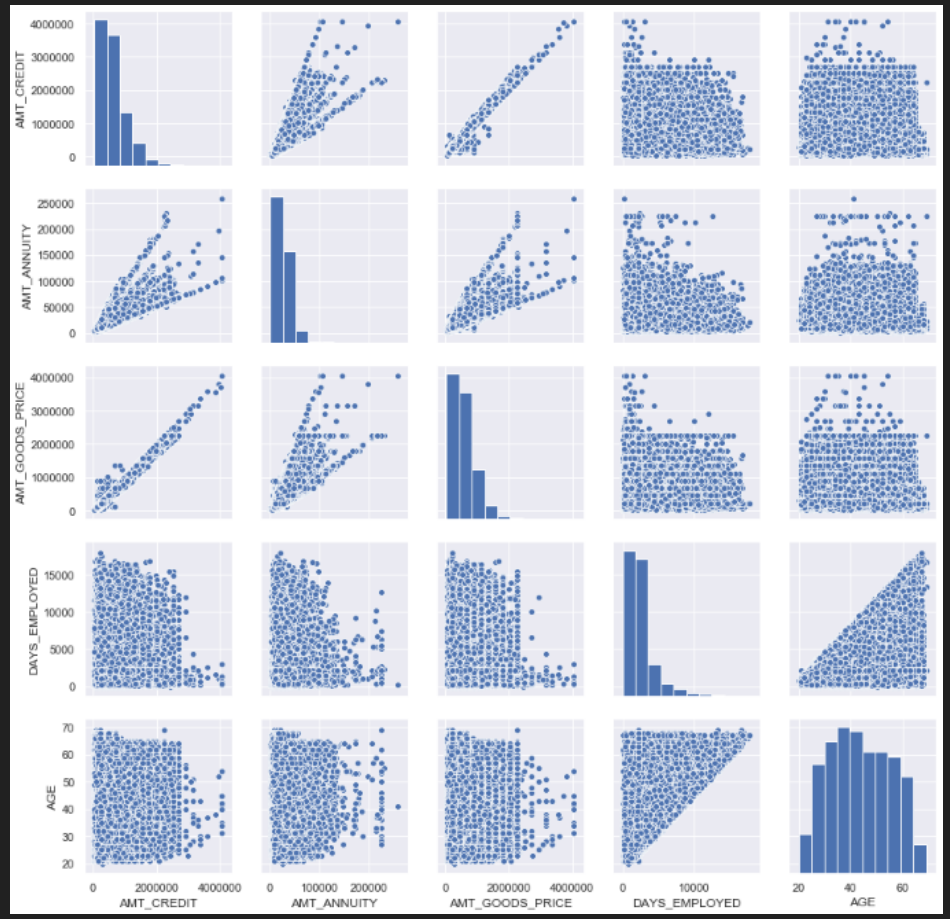
- Bivariate Categorical to Categorical
  - CODE_GENDER: More number of female clients as compared to male clients are likely to face payment difficulties.
  - FLAG_OWN_CAR: Clients who do not own a car are more likely to face payment difficulties.
  - FLAG_OWN_REALTY: Clients who own Realty are more likely to face payment difficulties.
  - CNT_CHILDREN : Clients with no children are less likely to face payment difficulties.
  - NAME_FAMILY_STATUS : Clients who are married are more likely to face payment difficulties.

Multivariate – Numerical to Numerical

- Used TARGET = 1 (Clients with Payment Difficulties) and TARGET = 0 (Other Clients)
- Used Seaborn Pair Plots



TARGET = 1



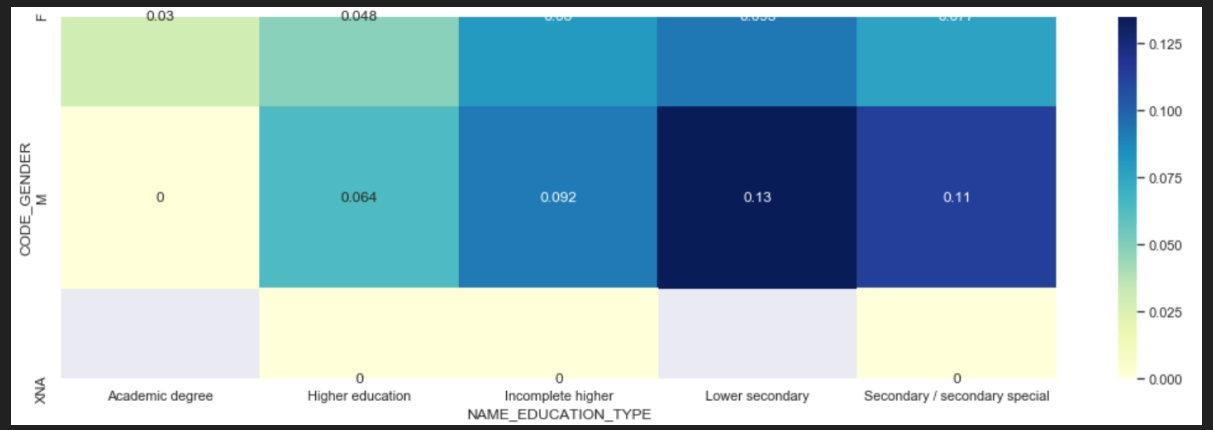
TARGET = 0

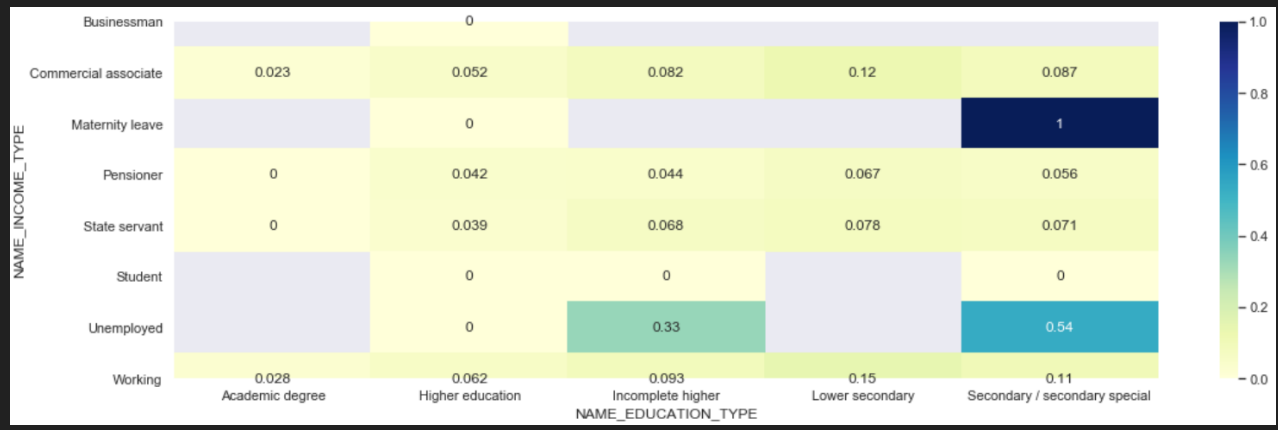
Results in Business Terms – Multivariate Numerical to Numerical

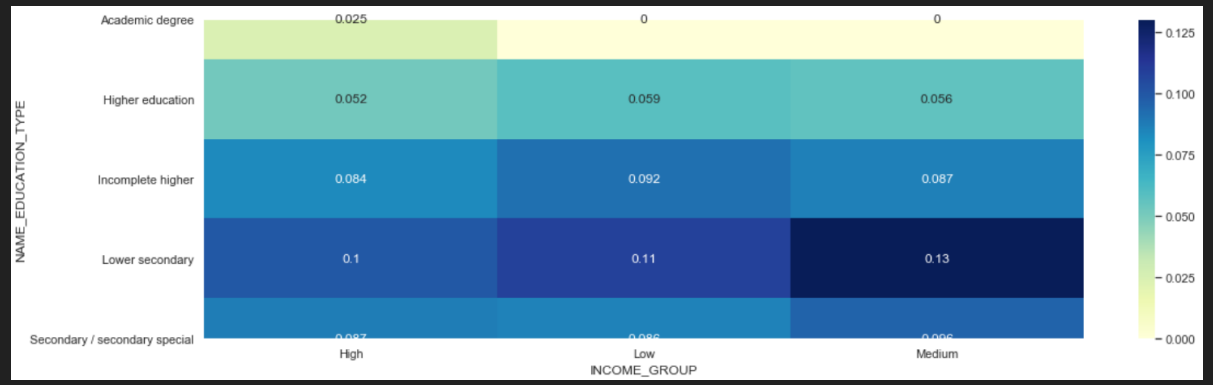
- Multivariate Numerical to Numerical
  - Linear Relationship between following pairs of columns
    - AMT_CREDIT and AMT_ANNUITY
    - AMT_CREDIT and AMT_GOODS_PRICE
    - DAYS_EMPLOYED and AGE
  - The above inference holds true for both TARGET = 0 and TARGET = 1

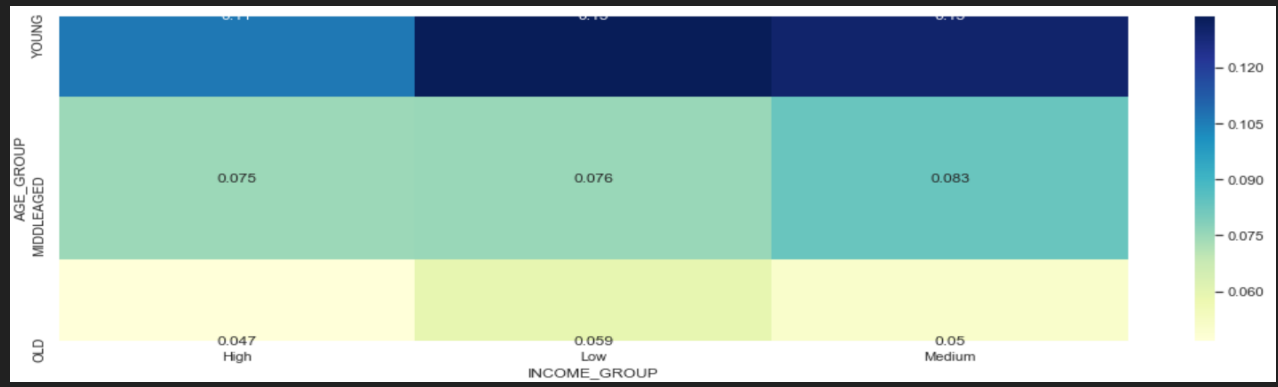
Multivariate – Categorical to Categorical

- Used Seaborn Heat Maps in for loop for categorical columns
- Some examples:







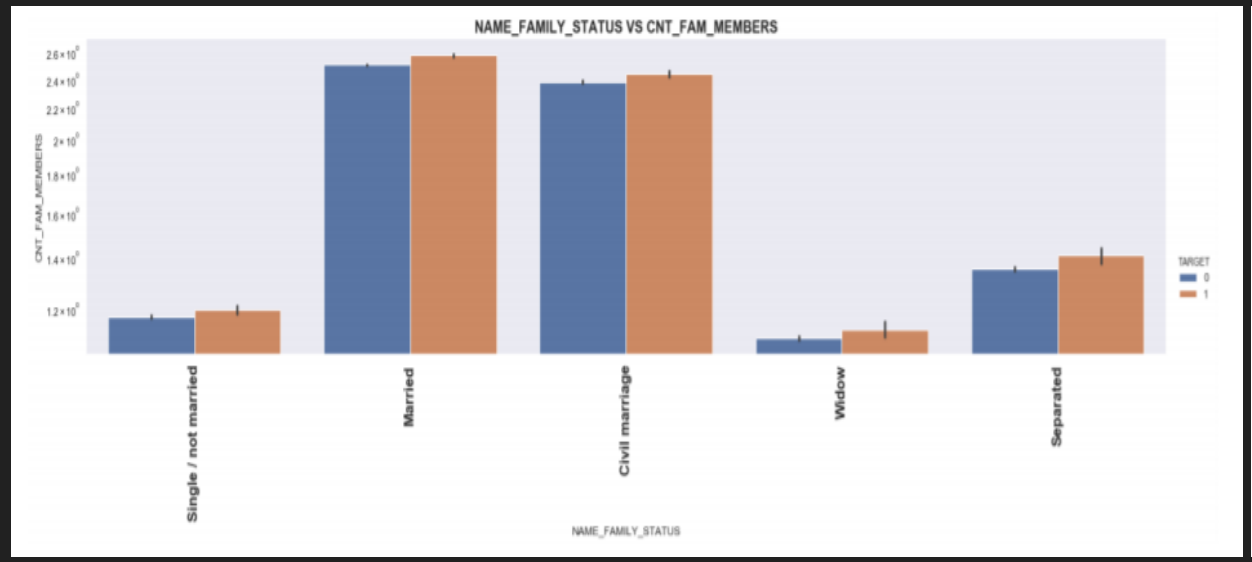


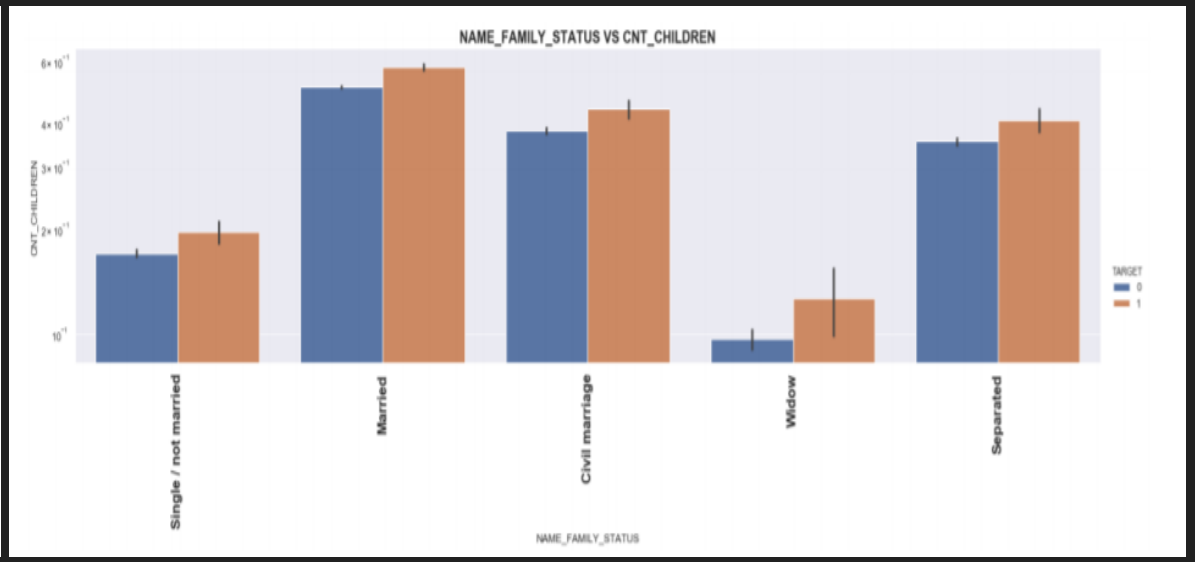
Results in Business Terms - Multivariate – Categorical to Categorical

- Multivariate – Categorical to Categorical
  - Clients who are more likely to have payment difficulties:
    - Unemployed and Old (age more than 60)
    - Female clients with Maternity Leave as Income Type
    - Male clients with Lower Secondary as Education Type
    - Male clients who are young with less than 30 years as age
    - Male clients who are in Medium Income Group (70% to 80%)
    - Clients who have Maternity Leave as Income Type and Secondary / Secondary special as Education Type
    - Clients who are Young (age less than 30) and have Lower Secondary as Education Type
    - Clients with Lower Secondary as Education Type in Medium Income Group (70% to 80%)
    - Clients who are Young (age less than 30) in Low Income Group (less than 70%)
  - Clients who are less likely to have payment difficulties:
    - Students, Pensioners and Businessman
    - Male clients with Academic Degree as Education Type
    - Female clients who are old with more than 60 years as age
    - Female clients who are in High Income Group (more than 80%)
    - Clients with Academic Degree as Education Type
    - Clients who are Old (age more than 60) in High Income Group (more than 80%)

Multivariate – Categorical to Numerical

- For Categorical to Numerical Analysis, used seaborn cat plots.
  For each Categorical column vs numerical columns plotting bar graphs for Target 0 and Target 1 dataframes as hue.
  - Target 0 = Blue color
  - Target 1 = Orange color
- Some Examples:





Results in Business Terms - Multivariate – Categorical to Numerical

- Multivariate – Categorical to Numerical
  - Clients who are more likely to have payment difficulties: Clients who are married and have big families or more number of children
  - Clients who are less likely to have payment difficulties: Clients who are widows and have small families
  - Clients who are more likely to have payment difficulties : are married and have more number of children
  - Clients who are less likely to have payment difficulties: are Single / Not Married and have low number of days employed

Working with Previous Data

Merging and Splitting

Merging and Splitting

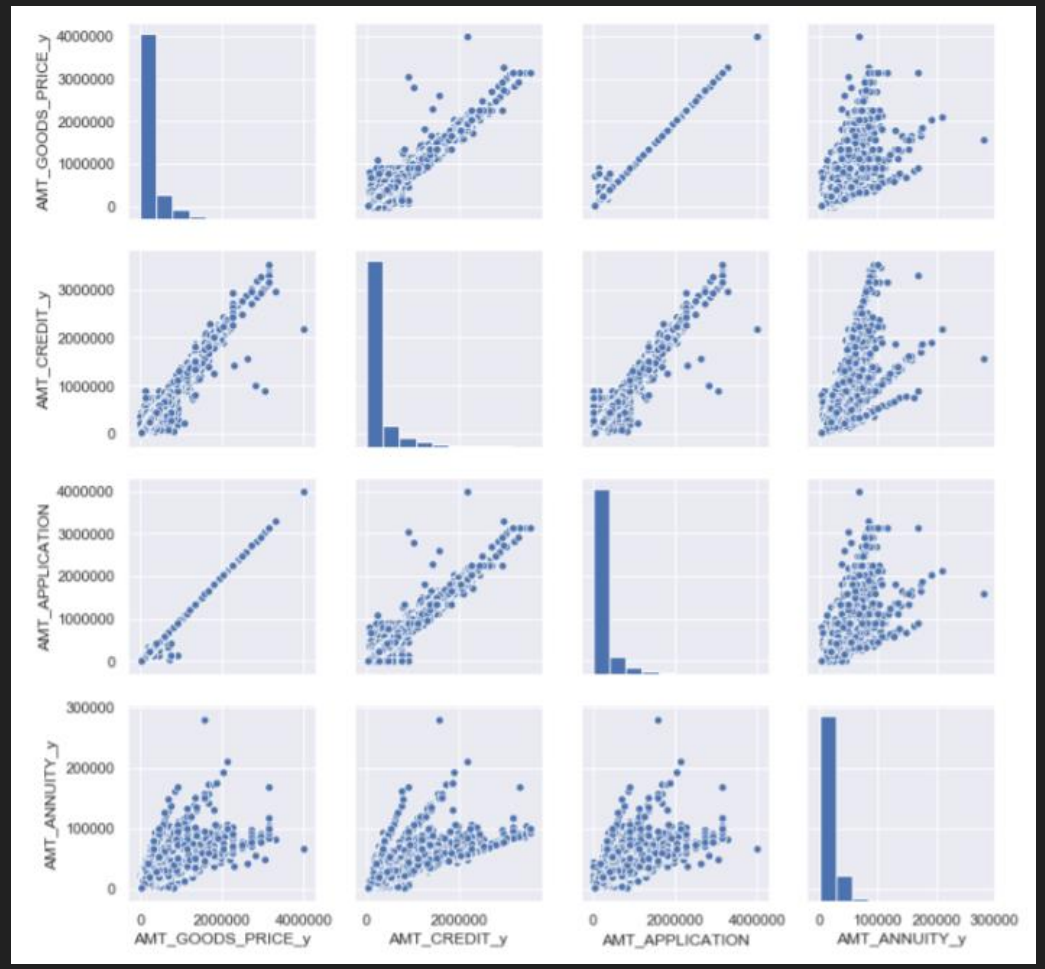
- Read the previous application data
- Removed columns having high number of null values [>40%]
- Merged the application data set with the previous dataset using pandas inner join.
- Splitting merged data frame again into Target 0 and Target 1

Working with Merged Data

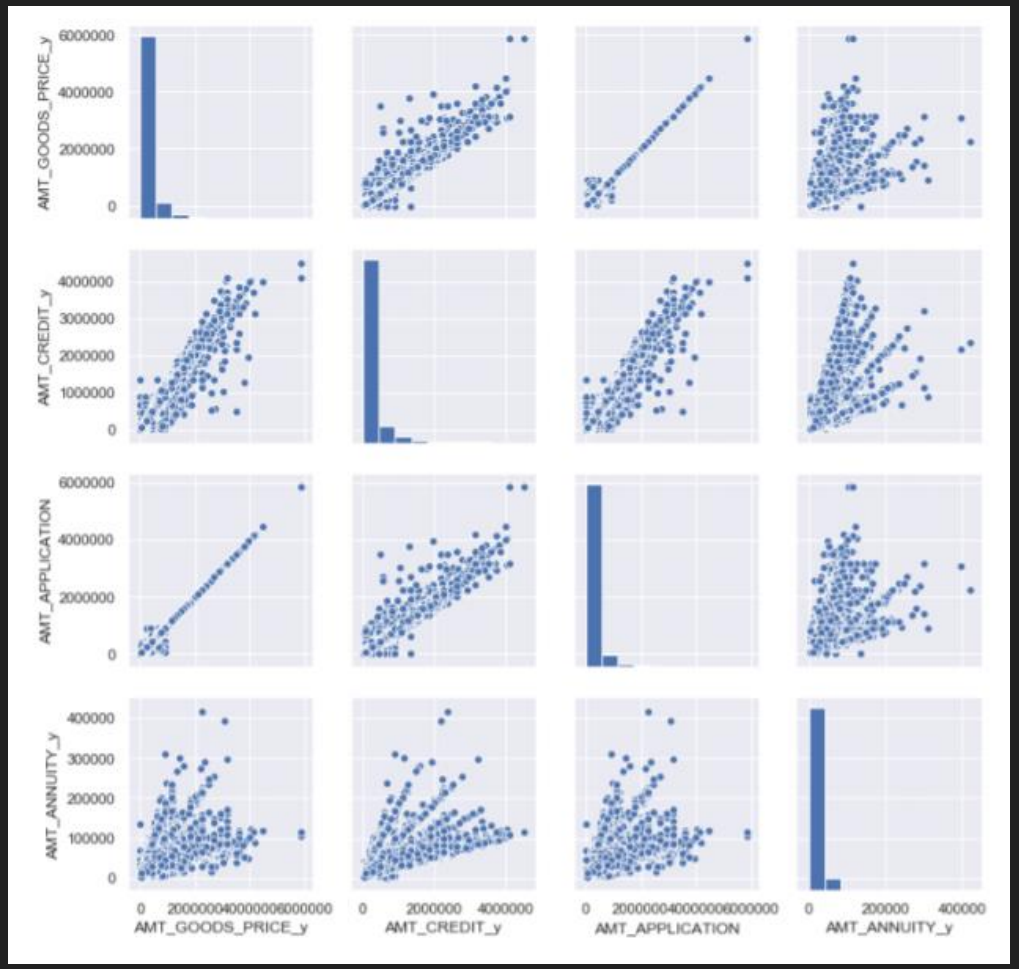
Data Analysis

Multivariate – Numerical to Numerical

- Used TARGET = 1 (Clients with Payment Difficulties) and TARGET = 0 (Other Clients)
- Used Seaborn Pair Plots



TARGET = 1



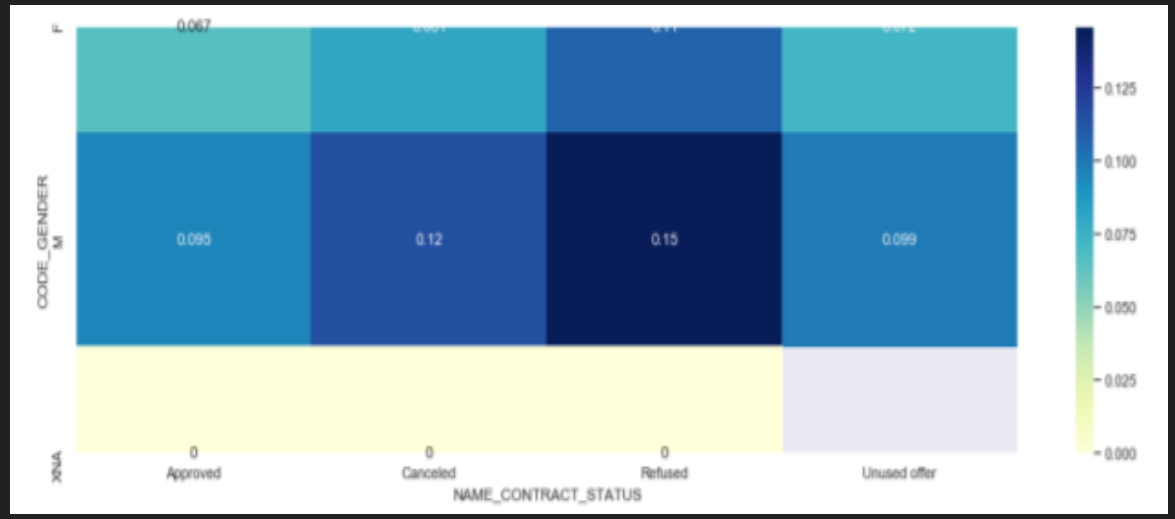
TARGET = 0

Results in Business Terms - Multivariate – Numerical to Numerical

- It is observed that there is a linear relationship between following pairs of columns for both target 0 and target 1 dataframes
  - AMT_CREDIT_y and AMT_GOODS_PRICE_y
  - AMT_APPLICATION and AMT_GOODS_PRICE_y
  - AMT_CREDIT_y and AMT_APPLICATION
  - AMT_ANNUITY_y and AMT_CREDIT_y
  - AMT_ANNUITY_y and AMT_GOODS_PRICE_y
  - AMT_ANNUITY_y and AMT_APPLICATION

Multivariate – Categorical to Categorical

- Used Seaborn Heat Maps in for loop for categorical columns
- Used pivot table for column NAME_CONTRACT_STATUS
- Some examples:

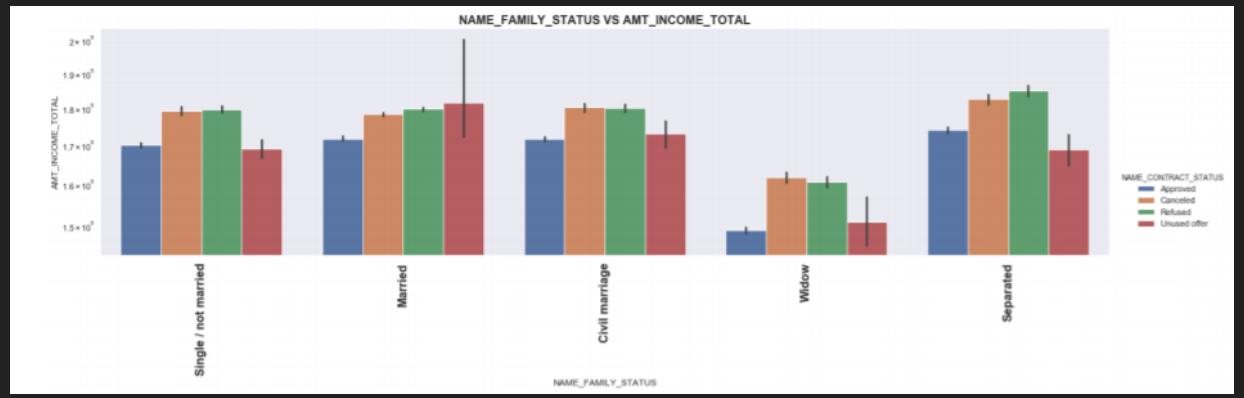


Results in Business Terms - Multivariate – Categorical to Categorical

- Multivariate – Categorical to Categorical
  - Clients who are more likely to have payment difficulties: Male clients having previous Contract status as refused
  - Clients who are less likely to have payment difficulties: Female clients whose previous application was approved.
  - Clients who are more likely to have payment difficulties: with academic degree as Education type and previous contract status as unused offer.
  - Clients who are less likely to have payment difficulties: with academic degree as Education type and previous Contract status as Cancelled.

Multivariate – Categorical to Numerical

- Used NAME_CONTRACT_TYPE as hue color to plot numerical and categorical columns
  - Blue – Approved
  - Orange - Cancelled
  - Green – Refused
  - Red – Unused offer
- Used Seaborn Cat Plots



Results in Business Terms - Multivariate – Categorical to Numerical

- Multivariate – Categorical to Numerical
  - Clients who have higher chances of getting loan approved based on previous application data: with family status as separated with high income.
  - Clients who have lower chances of getting loan approved based on previous application data: with family status as widow with low income.
  - Clients who have higher chances of getting loan approved based on previous application data: with family status as widow with high number of days employed.
  - Clients who have lower chances of getting loan approved based on previous application data: with family status as single/ not married with low number of days employed.

Results in Business Terms

Results in Business Terms

- NAME_CONTRACT_TYPE: Most of the clients have opted for cash loans.
- CODE_GENDER: There are more number of female clients as compared to male.
- NAME_INCOME_TYPE: : The count of Working clients is highest, followed by Commercial Associate and then by pensioners.
- NAME_EDUCATION_TYPE: : Mojority of clients have completed secondary / secondary special education, followed by 'Higher Education' and then by 'Incomplete higher'.
- FLAG_OWN_CAR : Majority of the clients do not own a car.
- FLAG_OWN_REALTY : Majority of the clients own a flat / appartment.
- AGE_GROUP : Majority of the clients are middle aged (30 - 60).
- INCOME_GROUP : Majority of the clients have high income (more than 80%).
- NAME_HOUSING_TYPE: Majority of the clients live in House / apartment, followed by clients living with parents, followed by clients living in municipal apartment.
- NAME_FAMILY_STATUS: Most of the clients are married followed by singles or not married and then followed by civil marriage.
- OCCUPATION_TYPE: The top three OCCUPATION_TYPE in descending order are Laborers, Sales Staff, Core Staff.
- ORGANIZATION_TYPE: Ignoring XNA, the top three ORGANIZATION_TYPE in descending order are Business, Self-Employed and Other.
- CODE_GENDER: More number of female clients as compared to male clients are likely to face payment difficulties.
- FLAG_OWN_CAR: Clients who do not own a car are more likely to face payment difficulties.
- FLAG_OWN_REALTY: Clients who own Realty are more likely to face payment difficulties.
- CNT_CHILDREN : Clients with no children are less likely to face payment difficulties.
- NAME_FAMILY_STATUS : Clients who are married are more likely to face payment difficulties.

Results in Business Terms

- Linear relationship between following pairs of columns for both target 0 and target 1 dataframes:
  - AMT_CREDIT and AMT_ANNUITY
  - AMT_CREDIT and AMT_GOODS_PRICE
  - DAYS_EMPLOYED and AGE
  - AMT_CREDIT_y and AMT_GOODS_PRICE_y
  - AMT_APPLICATION and AMT_GOODS_PRICE_y
  - AMT_CREDIT_y and AMT_APPLICATION
  - AMT_ANNUITY_y and AMT_CREDIT_y
  - AMT_ANNUITY_y and AMT_GOODS_PRICE_y
  - AMT_ANNUITY_y and AMT_APPLICATION

Results in Business Terms

- Clients who are more likely to have payment difficulties:
  - Unemployed and Old (age more than 60)
  - Female clients with Maternity Leave as Income Type
  - Male clients with Lower Secondary as Education Type
  - Male clients who are young with less than 30 years as age
  - Male clients who are in Medium Income Group (70% to 80%)
  - Clients who have Maternity Leave as Income Type and Secondary / Secondary special as Education Type
  - Clients who are Young (age less than 30) and have Lower Secondary as Education Type
  - Clients with Lower Secondary as Education Type in Medium Income Group (70% to 80%)
  - Clients who are Young (age less than 30) in Low Income Group (less than 70%)
  - Male clients having previous Contract status as refused
  - Clients having academic degree as Education type having previous Contract status as unused
  - Clients who are young (less than 30 years) and have previous contract status as refused
  - Clients with low income group (Less than 70%) and previous contract status as refused
  - Clients with reject reason as score and previous contract status as refused
  - Clients who are new applicants and have previous contract status as cancelled
  - Clients who have 'Seller Industry' as Jewelry and previous contract status as cancelled
  - Clients who have channel type as AP + (Cash loans) and previous contract status type as refused

Results in Business Terms

- Clients who are more likely to have payment difficulties:
  - Clients who are married and have
    - big families or
    - more number of children
  - Clients who are widows and have high number of days employed
  - Clients who have Maternity Leave as Income Type and have
    - big families or
    - more number of children or
    - more days employed
  - Clients with Academic Degree as Education Type and have
    - big families
    - more number of children
  - Clients with high number of days employed and Education Type as
    - Secondary / Secondary special or
    - Higher Education or
    - Lower Secondary

Results in Business Terms

- Clients who are less likely to have payment difficulties:
  - Students, Pensioners and Businessman
  - Male clients with Academic Degree as Education Type
  - Female clients who are old with more than 60 years as age
  - Female clients who are in High Income Group (more than 80%)
  - Clients with Academic Degree as Education Type
  - Clients who are Old (age more than 60) in High Income Group (more than 80%)
  - Clients who are widows and have small families
  - Clients who are Single / Not Married and have low number of days employed
  - Clients with low number of days employed and Education Type as Incomplete Higher
  - Females whose previous application was approved
  - CLients having academic degree as Education type having previous Contract status as approved.
  - Clients who are old (more than 60 years) and have previous contract status as unused.
  - Clients with low income group (Less than 70%) and previous contract status as unused
  - Clients with client type as 'Refreshed' and previous contract status as approved

Results in Business Terms

- Clients who are less likely to have payment difficulties:
  - Students and Businessman
    - across all type of families
    - irrespective of number of children
    - days employed
  - Clients with Incomplete Higher as Education Type and have
    - small families or
    - less number of children

Thank You !