CREDIT EDA CASE STUDY

BY

S.Anil & Somyaranjan Das

Problem Statement

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.

Identifying the missing data & cleaning

- Step 1 – The data analysis can go wrong if we have null values or if any data is missing
- Step 2 – Identified the null values by using isnull function and also by using the condition of columns having null values which are greater than 30%
- Step 3 – Removed all the null values by using drop function
- Step 4 – Identified the columns which are having less null value rows to impute the missing values
- Step 5 – Identified that 'AMT_ANNUITY' column is having an outlier which is very large it will be inappropriate to fill those missing values with mean, Hence we will fill those missing banks with median value.

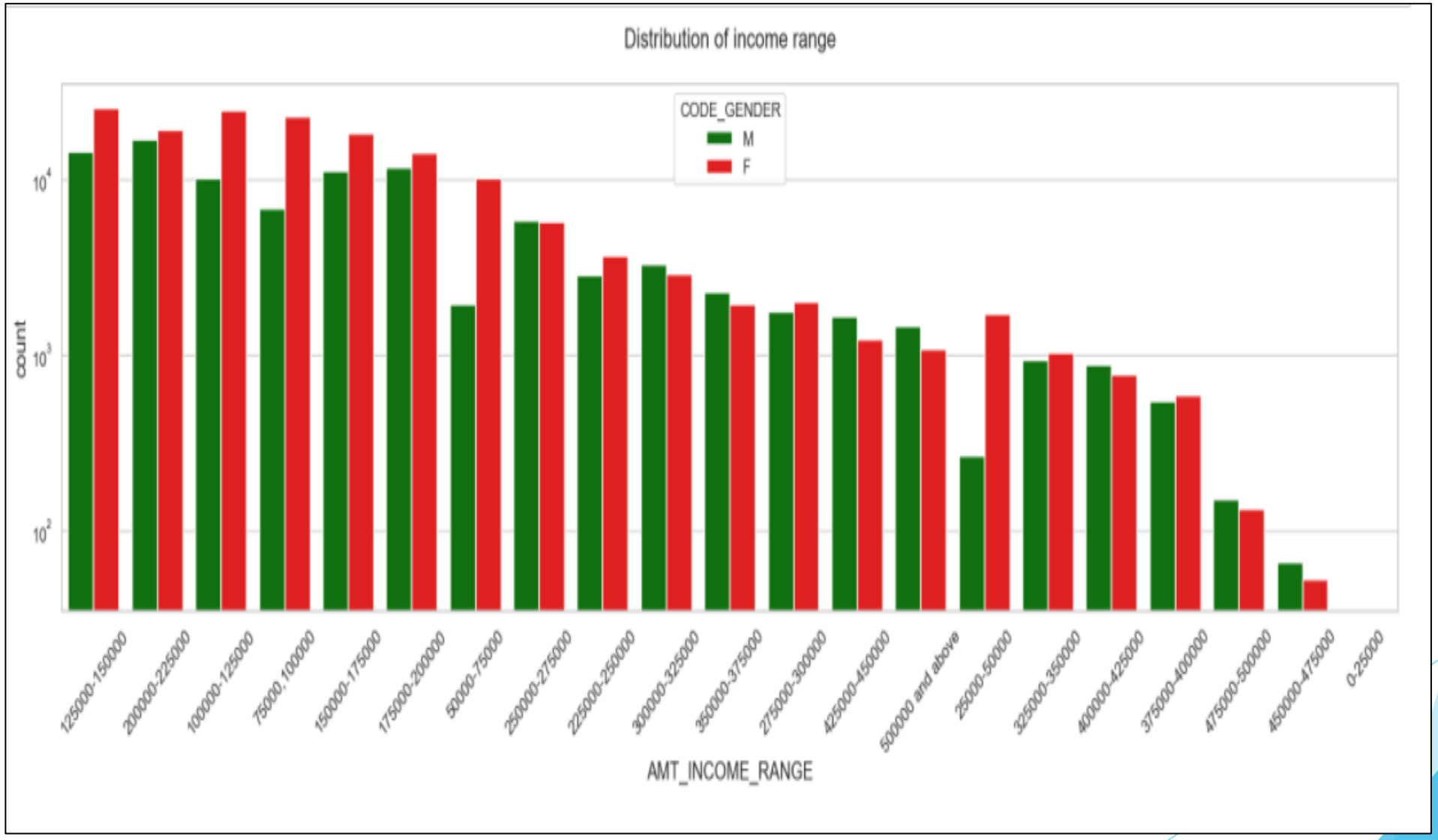
Dropping the unwanted columns from the dataset

- The unwanted columns will alter the data analysis so they need to be removed
- We observed that there are some columns where the value is mentioned as 'XNA' which means 'Not Available'. So we need to find the number of rows and columns having "XNA" and implement suitable techniques on them to fill those missing values or to delete them
- Identified XNA values from gender and organizational columns
- We have observed that there are 4 rows from Gender column and 55974 rows from the Organization column
- As we observe that Females are greater in number so we can replace the "XNA" value with F as will have negligible affect on the data.
- As we observe for column 'ORGANIZATION_TYPE', we have total count of 307511 rows of which 55374 rows are having 'XNA' values. Which means of about 18% of the column is having this values. Hence if we drop the rows of total 55374, will not have any major impact on our dataset.

Data Imbalance in the Data

- Created bins for continuous variable categories column 'AMT_INCOME_TOTAL' and 'AMT_CREDIT’
- Divided the dataset into two dataset of Client=1(client with payment difficulties) and Client=0(all other)
- Calculated the imbalance percentage with ratio of client 0 to client 1 as majority is client 0 and minority is client 1

Univariate Analysis

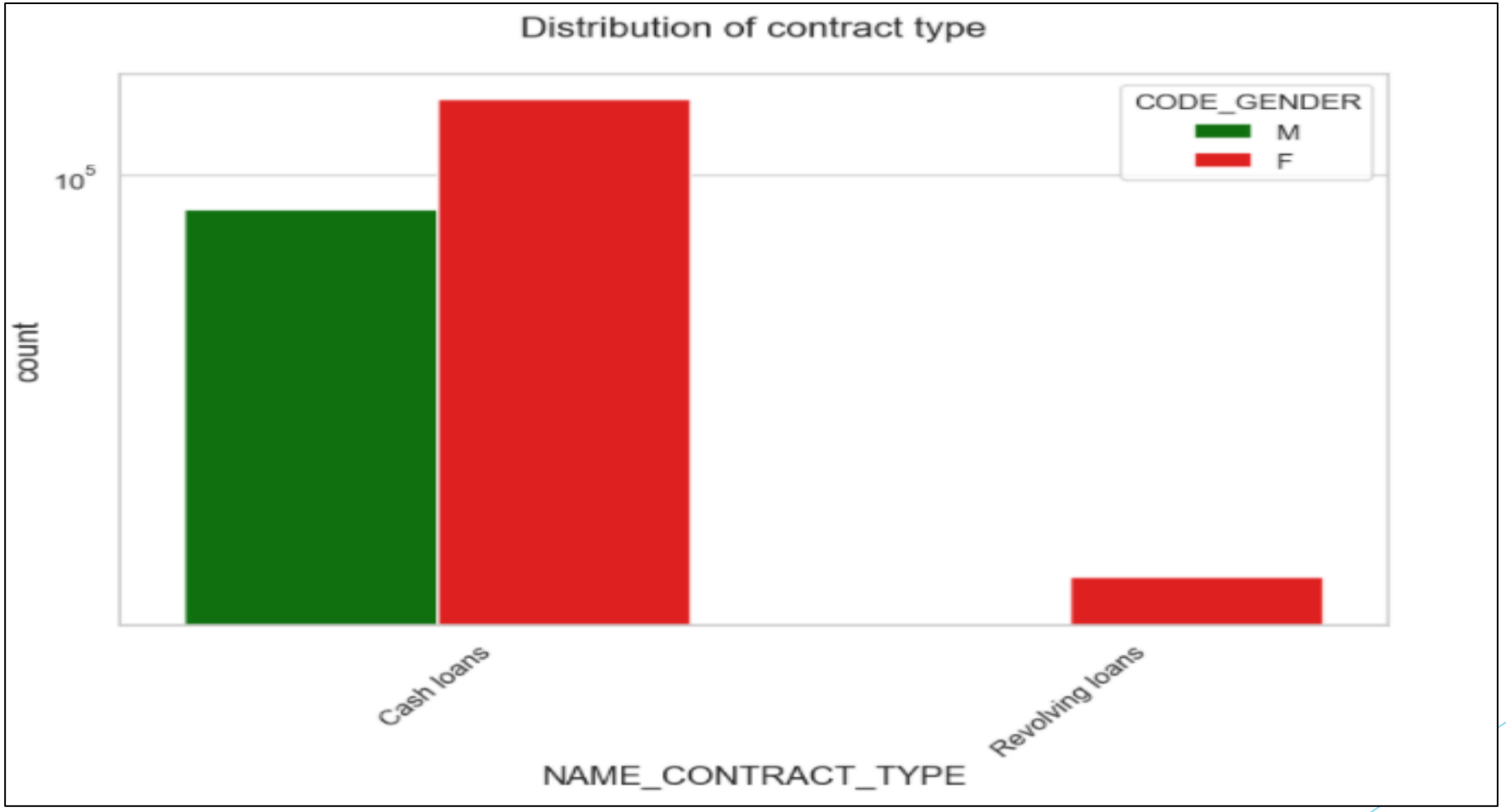


Univariate Analysis

- Analysis of before slide
- Categorical Univariate Analysis performed in logarithmic scale for client=0(client with no payment difficulties)
- Conclusion from the above graph:-
- Counts of female are higher as compared to male. More number of credits are with Income range from 100000 to 200000. This graph show that females are more than male in having credits for that range. Very less count for income range 400000 and above.

.

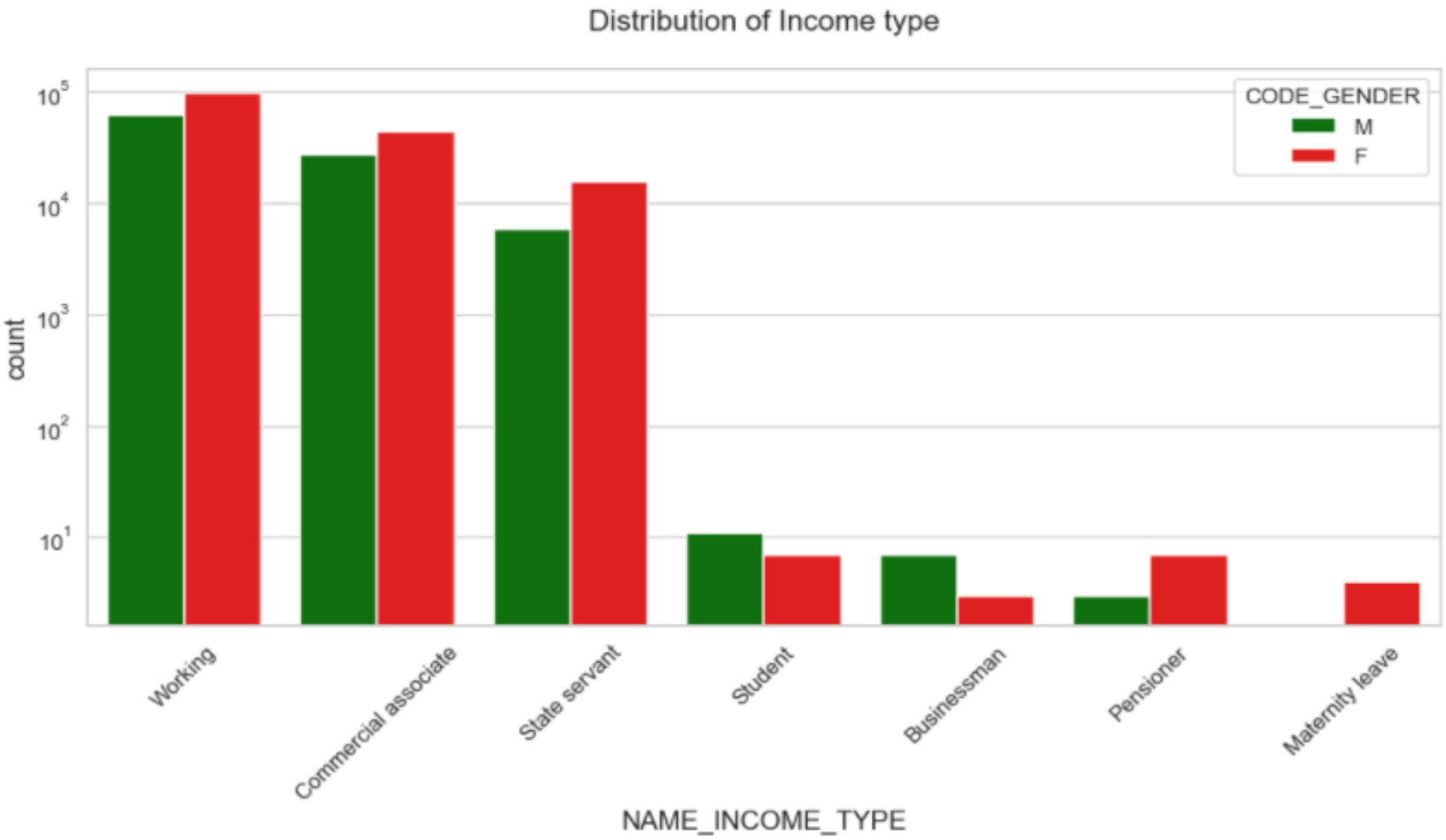
Contract type plot for client 0



Contract type plot - Conclusion

- Conclusion from the above graph:-
- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type. For this also Female is leading for applying credits.

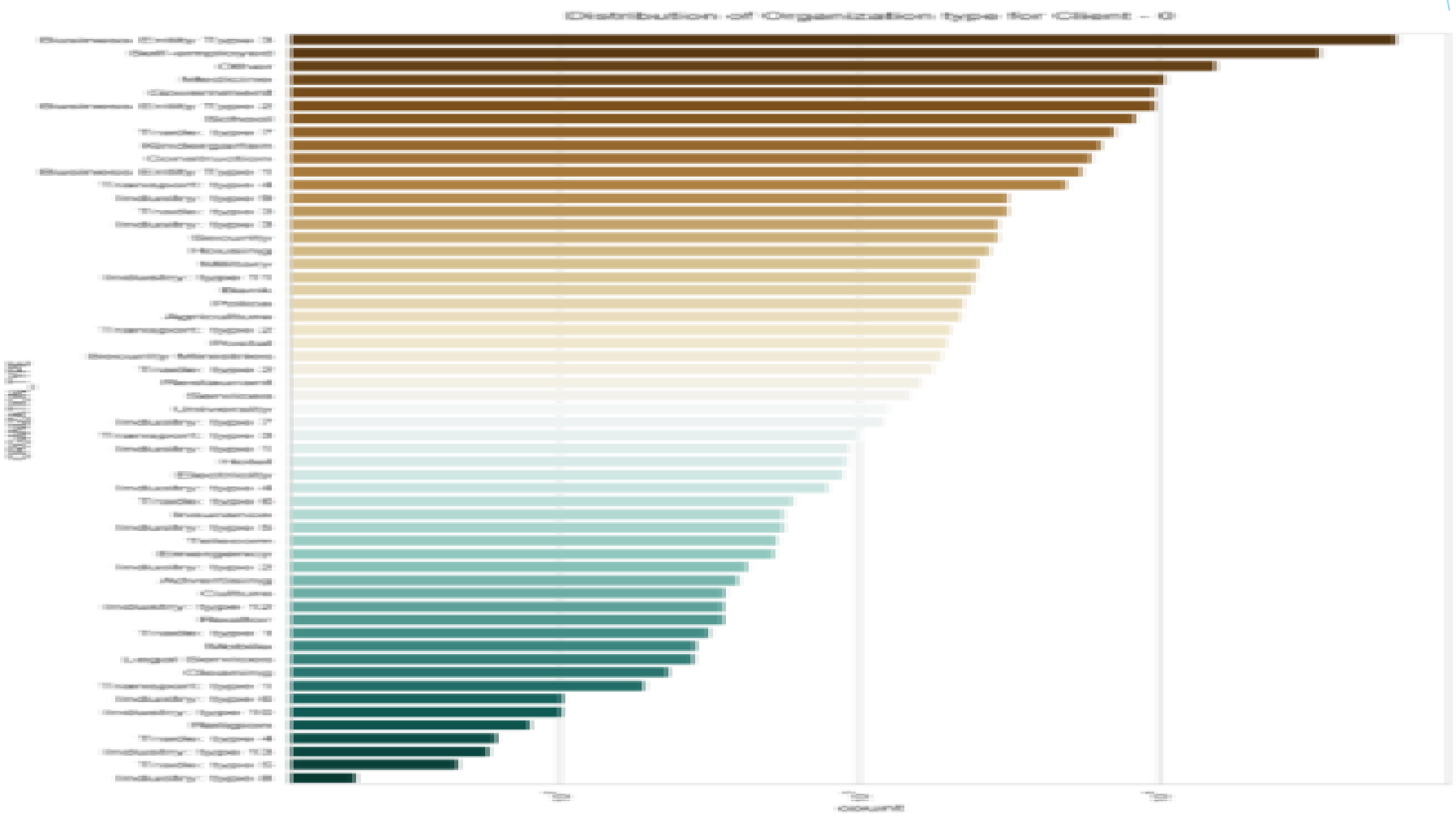
Income type plot for client 0



Income type plot - Conclusion

- Conclusion from the above graph:-
- For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than others. For this Females are having more number of credits than male. Less number of credits for income type ‘student’ ,’pensioner’, ‘Businessman’ and ‘Maternity leave’.

Organization type plot for client 0



Organization type plot - Conclusion

- Conclusion from the above graph:-
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’, ‘Other’ , ‘Medicine’ and ‘Government’. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4

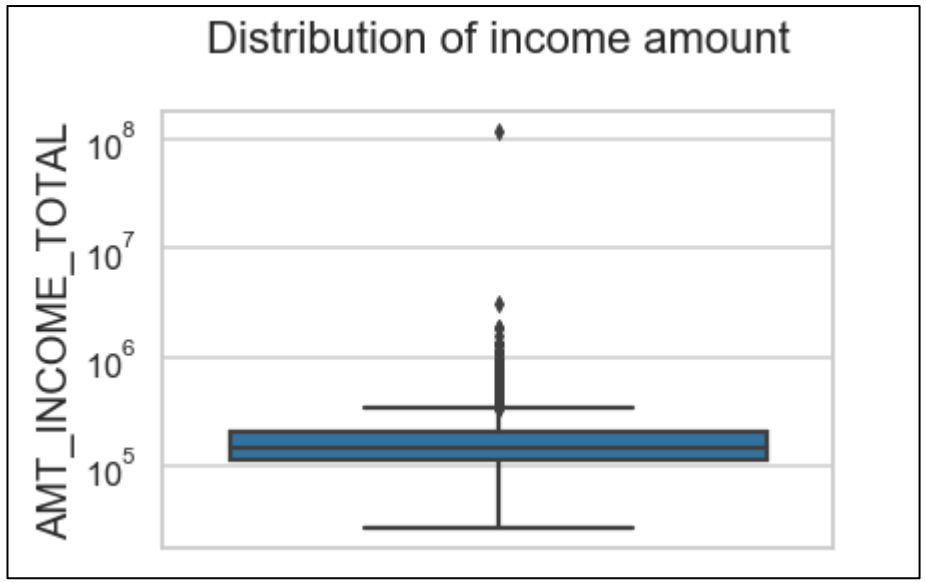
Categorical Univariate Analysis in logarithmic scale for client 1 - Conclusion

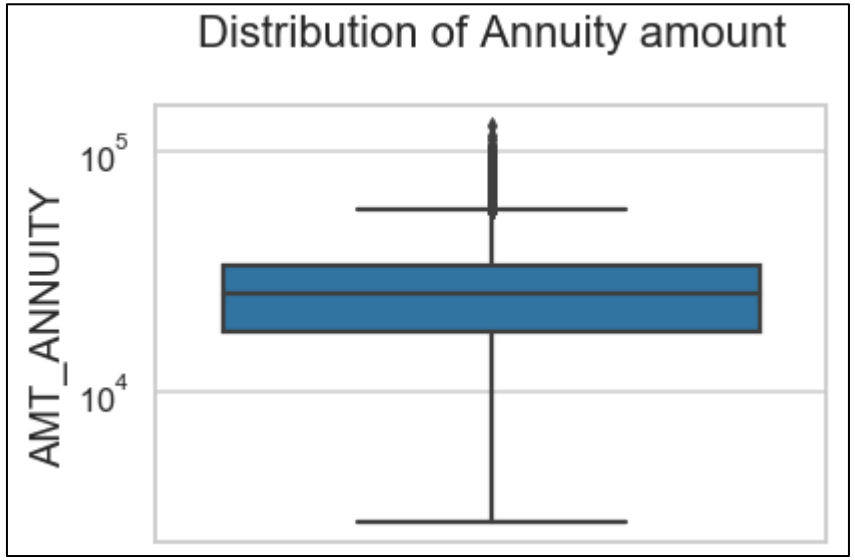
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4. Same as type 0 in distribution of organization type.
- **From Distribution of Income Range**
- Male counts are higher than female. Income range from 100000 to 200000 is having more number of credits. This graph show that males are more than female in having credits for that range. Very less count for income range 400000 and above.
- **From Distribution of Income Type**
- For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave. For this Females are having more number of credits than male. Less number of credits for income type ‘Maternity leave’. For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.
- **From Distribution of Contract Type**
- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type. For this also Female is leading for applying credits. For type 1 : there is only Female Revolving loans.

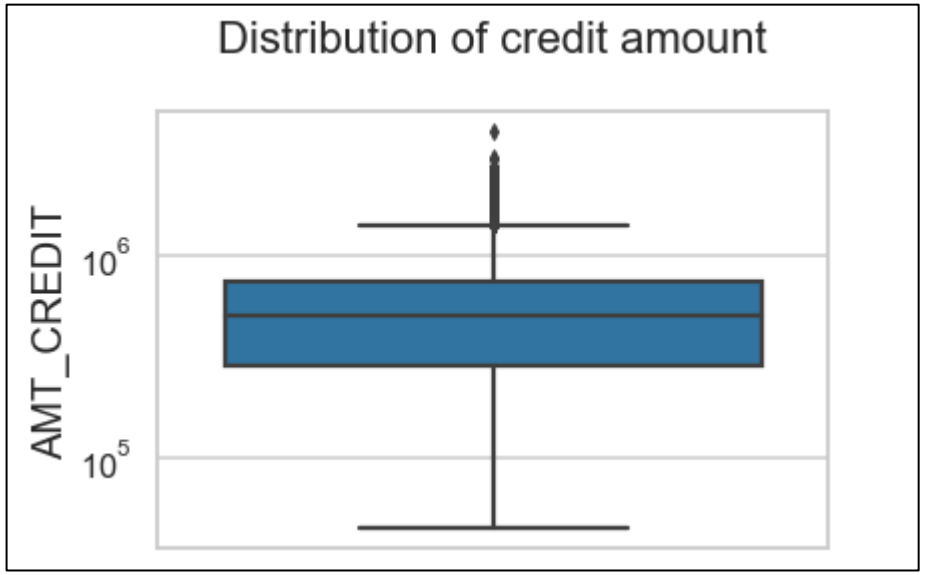
Corelation for client 0 & 1 using heat maps

- **For Client 0**
- As we can see from above correlation heatmap, There are number of observation we can point out
- Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa. Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa. Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa. less children client have in densely populated area. Credit amount is higher to densely populated area. The income is also higher in densely populated area.
- **For Client 1**
- This heat map for client 1 is also having quite a same observation just like client 0. But for few points are different. They are listed below.
- The client's permanent address does not match contact address are having less children and vice-versa the client's permanent address does not match work address are having less children and vice-versa

Univariate analysis for variables – Outliers for Client 0







Univariate analysis for variables – Outliers for Client 0 - Conclusion

Conclusion from above distribution

Income Account

Some outliers are noticed in income amount. The third quartiles is very slim for income amount.

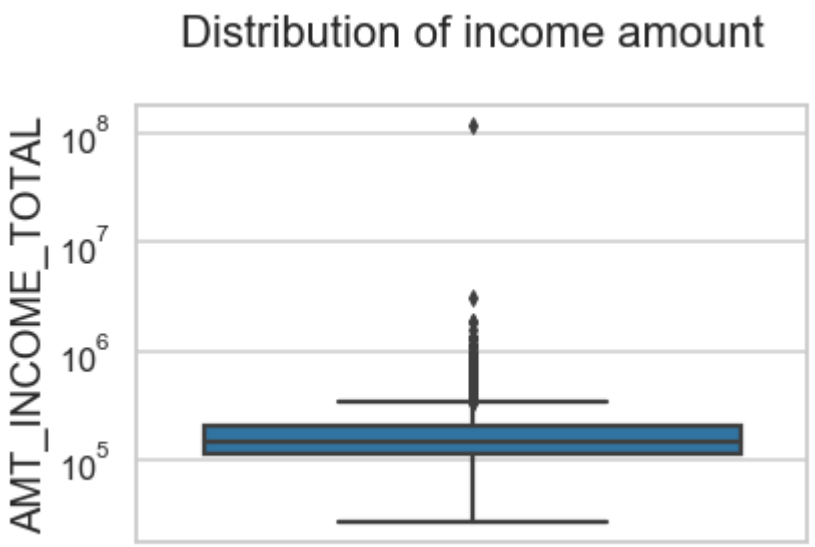
Annuity Amount

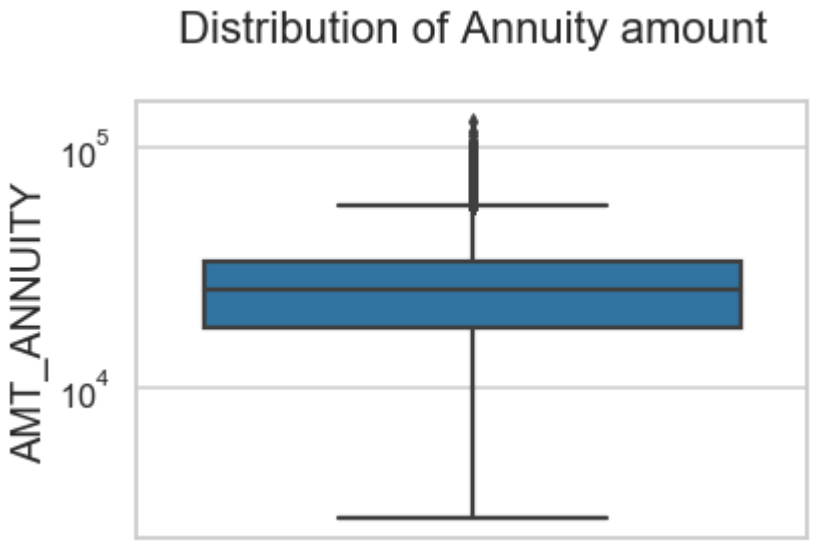
Some outliers are noticed in annuity amount. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

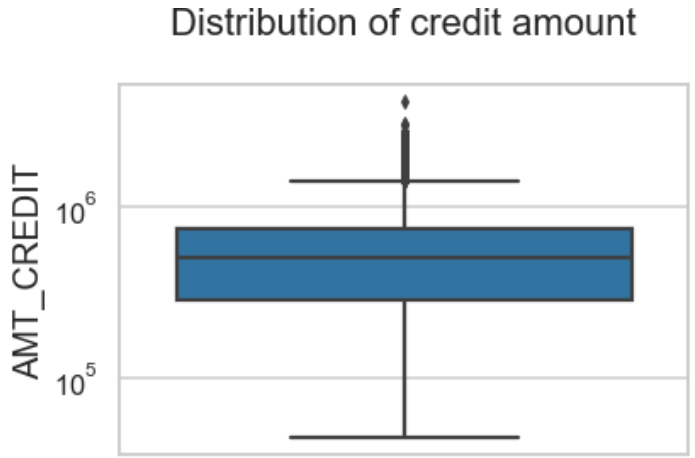
Credit Amount

Some outliers are noticed in credit amount. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile

Univariate analysis for variables – Outliers for Client 1







Univariate analysis for variables – Outliers for Client 0 - Conclusion

Conclusion from above distribution

Income Account

Some outliers are noticed in income amount. The third quartiles is very slim for income amount. Most of the clients of income are present in first quartile.

Annuity Amount

Some outliers are noticed in annuity amount. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

Credit Amount

Some outliers are noticed in credit amount. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

Conclusion for bivariate analysis – client 0

Conclusion from the Bivariate analysis for client 0

Income Amount

From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers. Less outlier are having for Academic degree but there income amount is little higher that Higher education. Lower secondary of civil marriage family status are have less income amount than others.

Credit Amount

From the above box plot we can conclude that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

Conclusion for bivariate analysis – client 1

Conclusion from the Bivariate analysis for client 1

Credit Amount

Quite similar with client 0 From the above box plot we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

Income Amount

Have some similarity with client0, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. Less outlier are having for Academic degree but there income amount is little higher that Higher education. Lower secondary are have less income amount than others.

Conclusion for univariate analysis after merging application data with previous application

Conclusion from Univariate Analysis

For Contract Status

Loan purposes with 'Repairs' are facing more difficulties in payment on time. There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties.

For Contract Status in logarithmic scale

Most rejection of loans came from purpose 'repairs'. For education purposes we have equal number of approves and rejection Paying other loans and buying a new car is having significant higher rejection than approves.

Conclusion for bivariate analysis after merging application data with previous application

Conclusion from Bivariate Analysis

For Credit amount in logarithmic scale

The credit amount of Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher. Income type of state servants have a significant amount of credit applied Money for third person or a Hobby is having less credits applied for.

For Credit amount prev vs Housing type in logarithmic scale

Here for Housing type, office apartment is having higher credit of client 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

Conclusion for Problem Statement

The bank should get "with parents" client for housing type as they are the most trusted clients leading to least number of unsuccessful payments. The bank must avoid to emphasize on "working" income type as they cause maximum number of unsuccessful payments. Loan purpose - "Repair" have higher number of unsuccessful payments on time. Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.