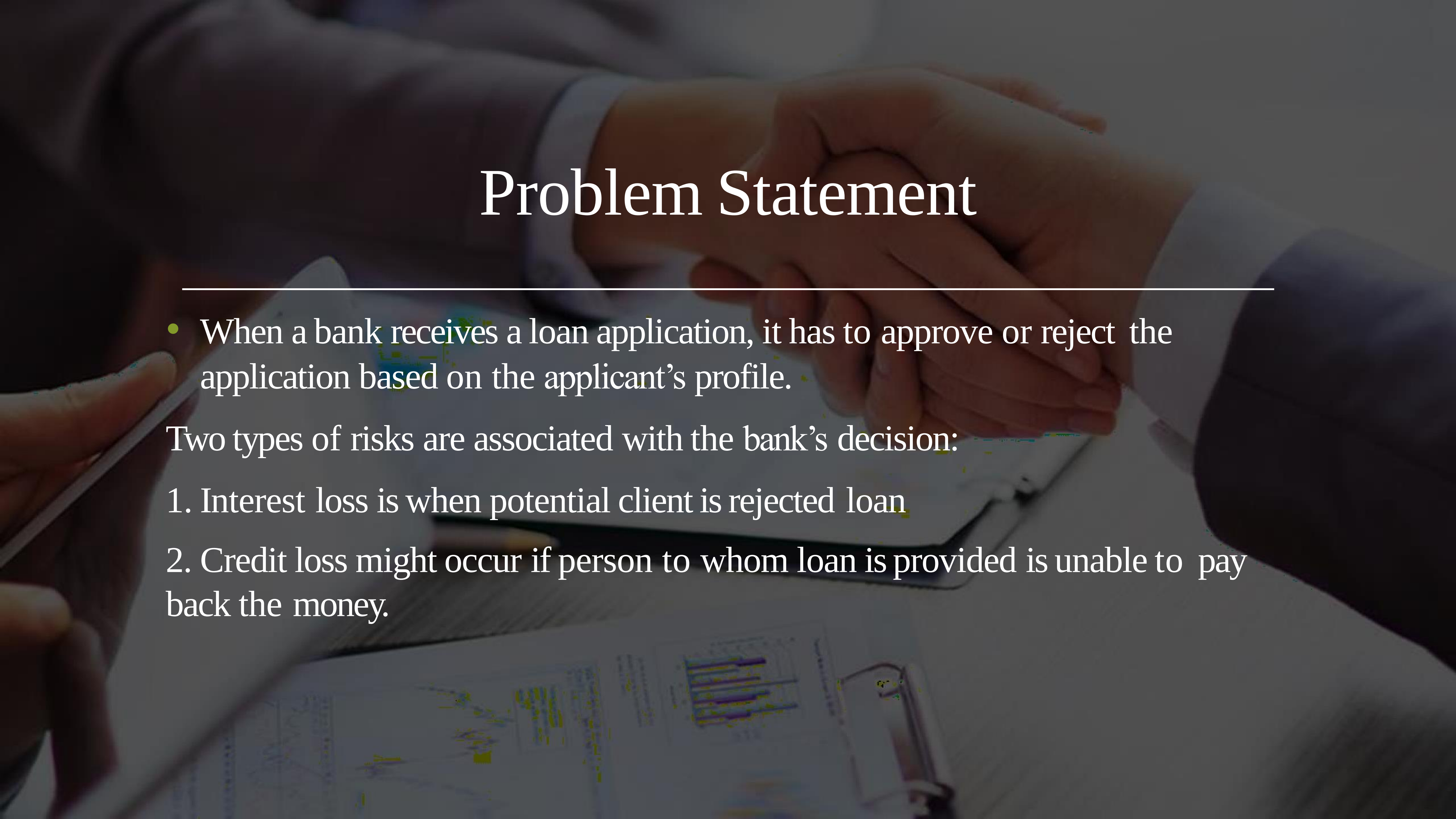
Bank Client Credit Risk Analysis EDA Case Study



Problem Statement

- When a bank receives a loan application, it has to approve or reject the application based on the applicant’s profile.

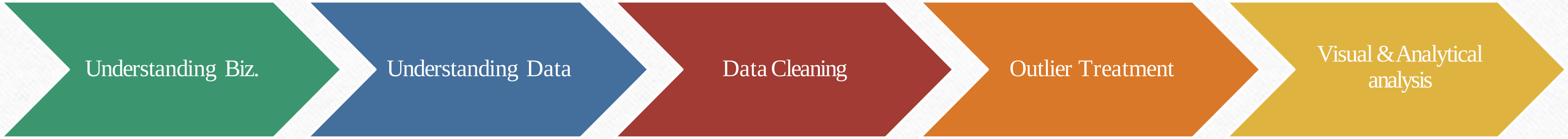
Two types of risks are associated with the bank’s decision:

1. Interest loss is when potential client is rejected loan

2. Credit loss might occur if person to whom loan is provided is unable to pay back the money.

Objective & Approach

- Firstly, to identify the key client attributes and credit attributes which drive interest or credit loss.
- Secondly, to identify top 10 variables in terms of correlation to define relationships between them.
- Visualization of data post EDA process using Univariate, Bivariate and Multivariate analysis.



Understanding Biz.

Understanding Data

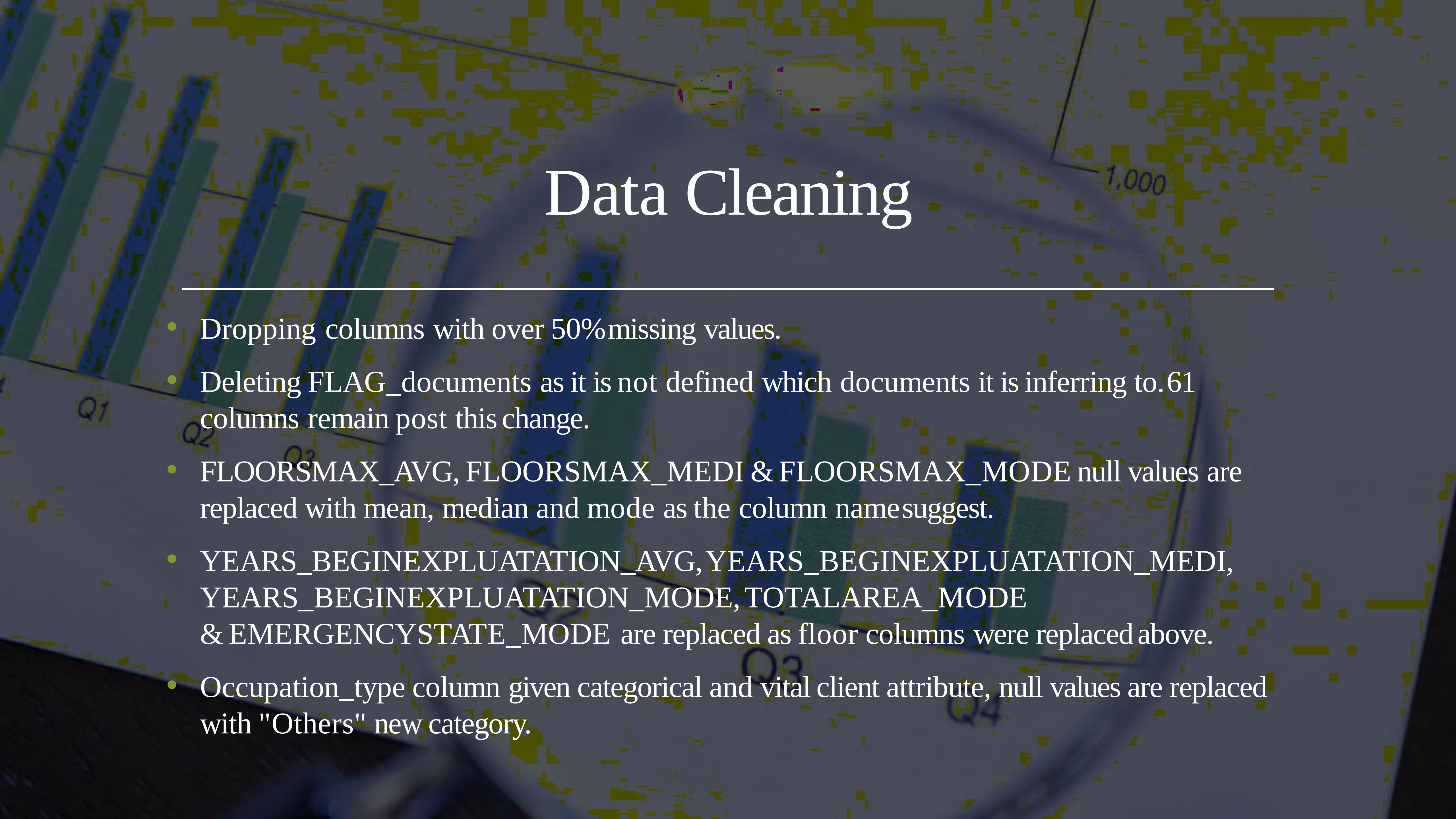
Data Cleaning

Outlier Treatment

Visual & Analytical analysis

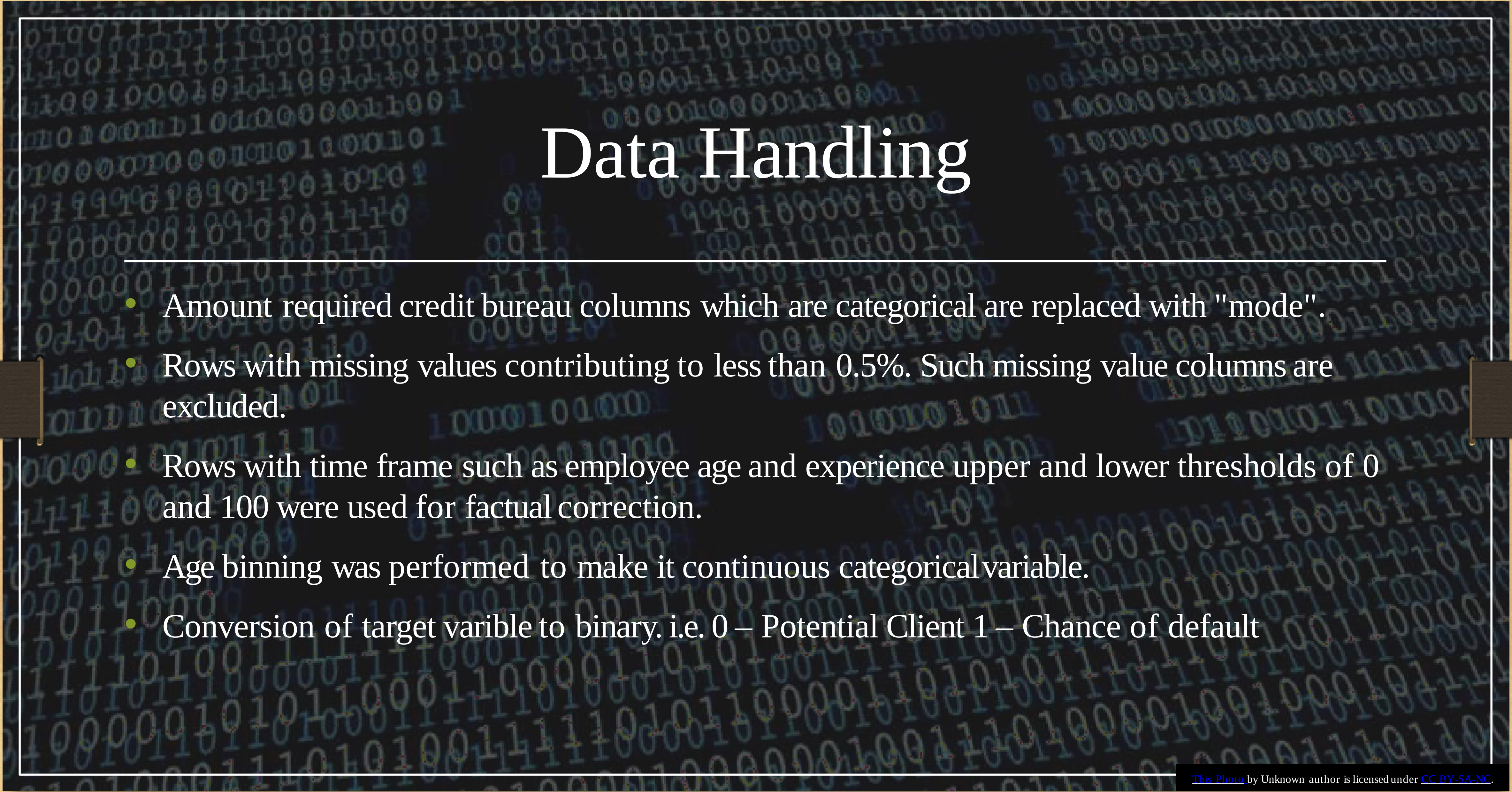
Understanding Data

- application_data
  - This data set represents new loan application details upon which primary EDA objectives are focused upon.
  - 162 MB | 3075211 Rows | 122 Columns | 2D | float64(65), int64(41), object(16)
- previous_application
  - This data set represents historical loan application details which support our EDA objectives and provide us with historical scenarios.
  - 395 MB | 1670214 Rows | 37 Columns | 2D | float64(15), int64(06), object(16)



Data Cleaning

- Dropping columns with over 50%missing values.
- Deleting FLAG_documents as it is not defined which documents it is inferring to.61 columns remain post this change.
- FLOORSMAX_AVG, FLOORSMAX_MEDI & FLOORSMAX_MODE null values are replaced with mean, median and mode as the column namesuggest.
- YEARS_BEGINEXPLUATATION_AVG, YEARS_BEGINEXPLUATATION_MEDI, YEARS_BEGINEXPLUATATION_MODE, TOTALAREA_MODE & EMERGENCYSTATE_MODE are replaced as floor columns were replaced above.
- Occupation_type column given categorical and vital client attribute, null values are replaced with "Others" new category.

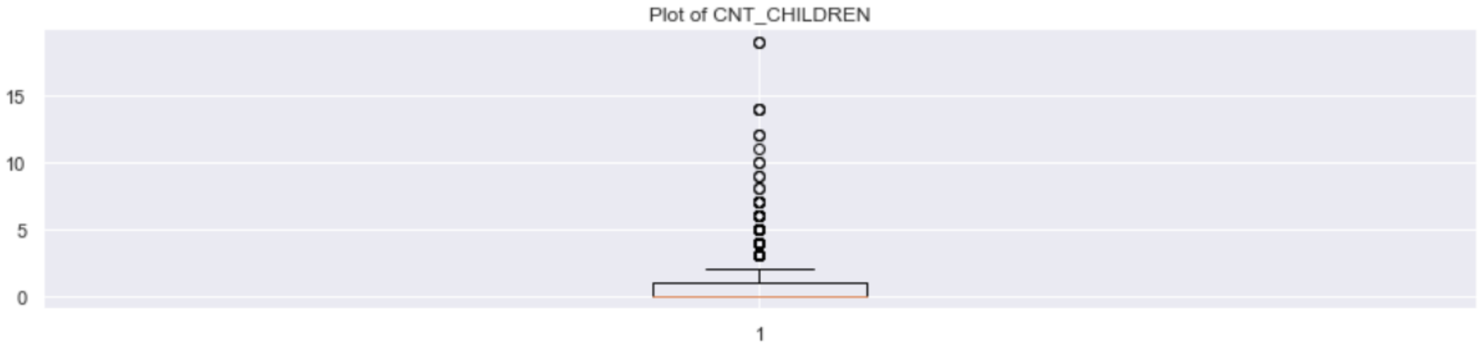


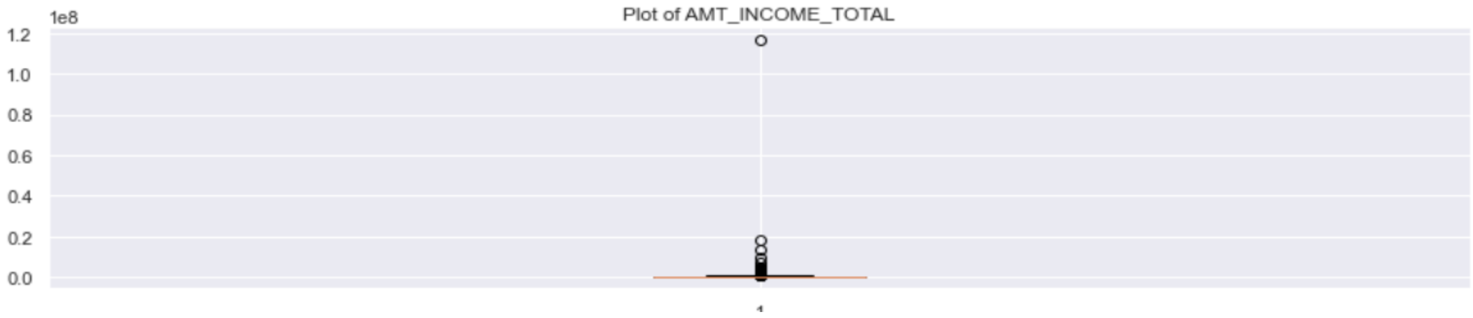
Data Handling

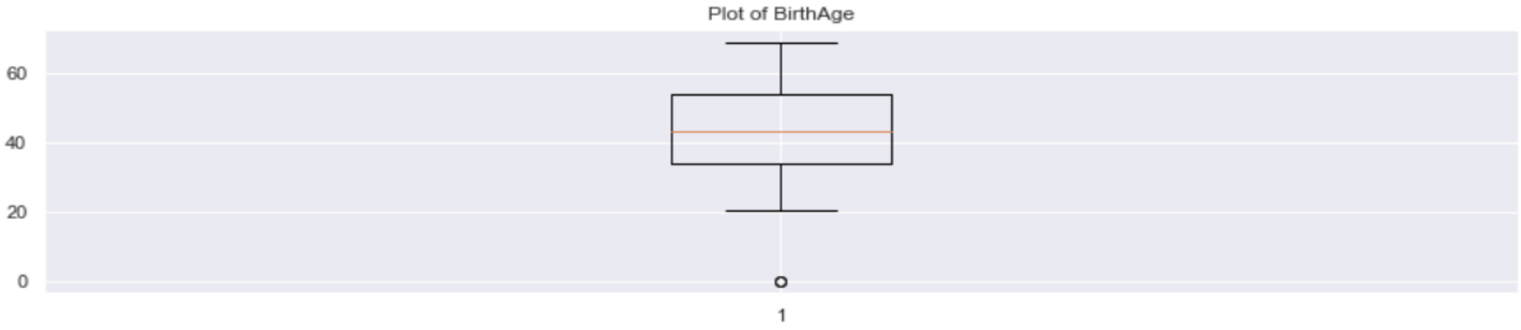
- Amount required credit bureau columns which are categorical are replaced with "mode".
- Rows with missing values contributing to less than 0.5%. Such missing value columns are excluded.
- Rows with time frame such as employee age and experience upper and lower thresholds of 0 and 100 were used for factual correction.
- Age binning was performed to make it continuous categoricalvariable.
- Conversion of target varible to binary. i.e. 0 – Potential Client 1 – Chance of default

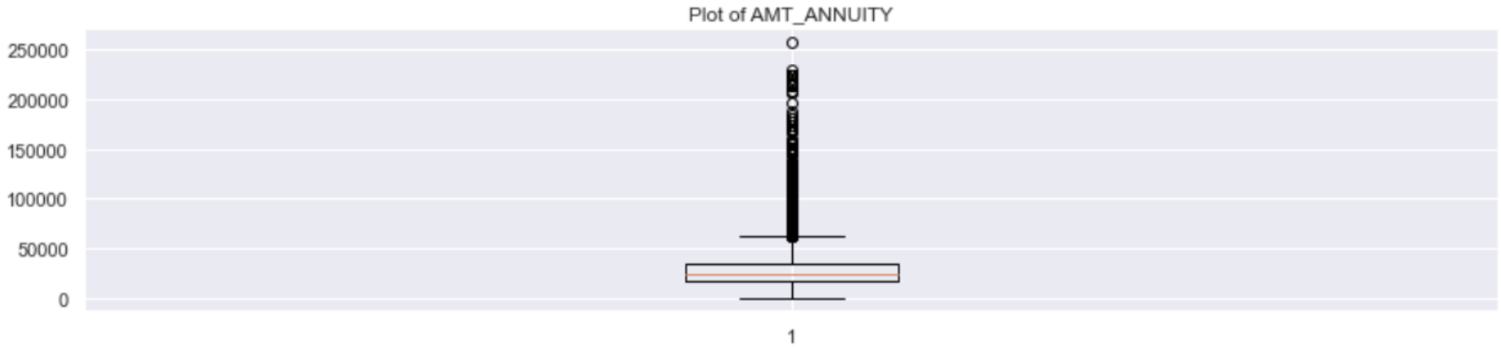
Insights after plotting box plots for some Numerical columns:

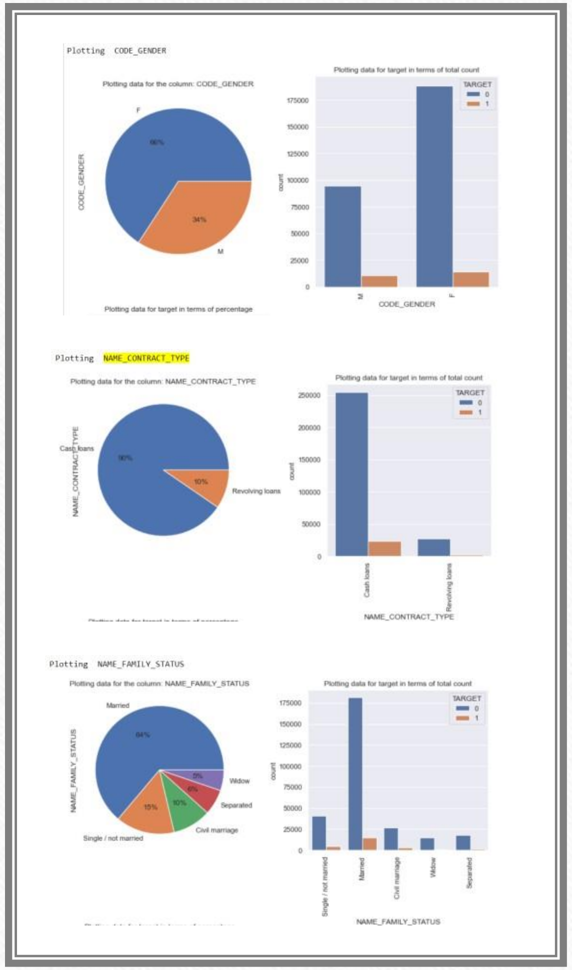
- CNT_CHILDREN as many outliers with average is towards 0
- AMT_INCOME_TOTAL has one major outlier which has a large income. We can delete this row as it can affect the average og the income
- Birth Age has average value of around 42 years. One outlier is in negative. Which can be deleted
- Years employed have outliers where years employed is 1000 years, which is clearly not possible. We can ignore these rows too.
- 'AMT_ANNUITY' has an average around 25000









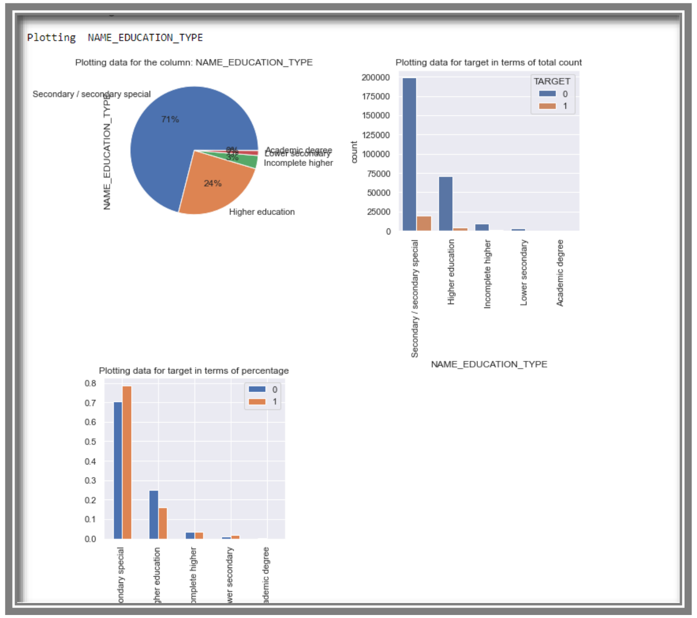


```
Plotting  CODE_GENDER
```

```
Plotting  NAME_CONTRACT_TYPE
```

```
Plotting  NAME_FAMILY_STATUS
```

- Percentage of people taking Cash loans is higher than Revolving loans.
- Percentage of Females taking Loans is higher than that of males and Females defaulting to pay back loan on time is higher than Males.
- Percentage of people defaulting to pay back the loan, do not own a car but it is opposite in case of reality, people having Reality are more likely to pay back the loan
- People who are working are taking more no of loans followed by commercial associates. Working people are more likely to default than any other income type.

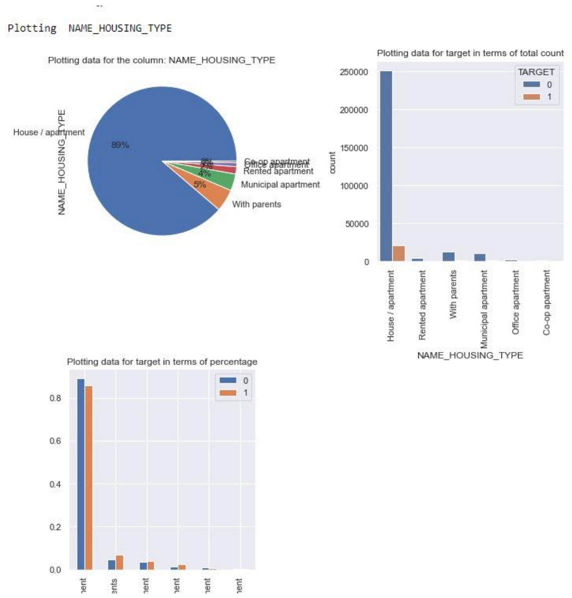


Plotting NAME_EDUCATION_TYPE

Plotting data for the column: NAME_EDUCATION_TYPE

Plotting data for target in terms of total count

Plotting data for target in terms of percentage



Plotting NAME_HOUSING_TYPE

Plotting data for the column: NAME_HOUSING_TYPE

Plotting data for target in terms of total count

Plotting data for target in terms of percentage

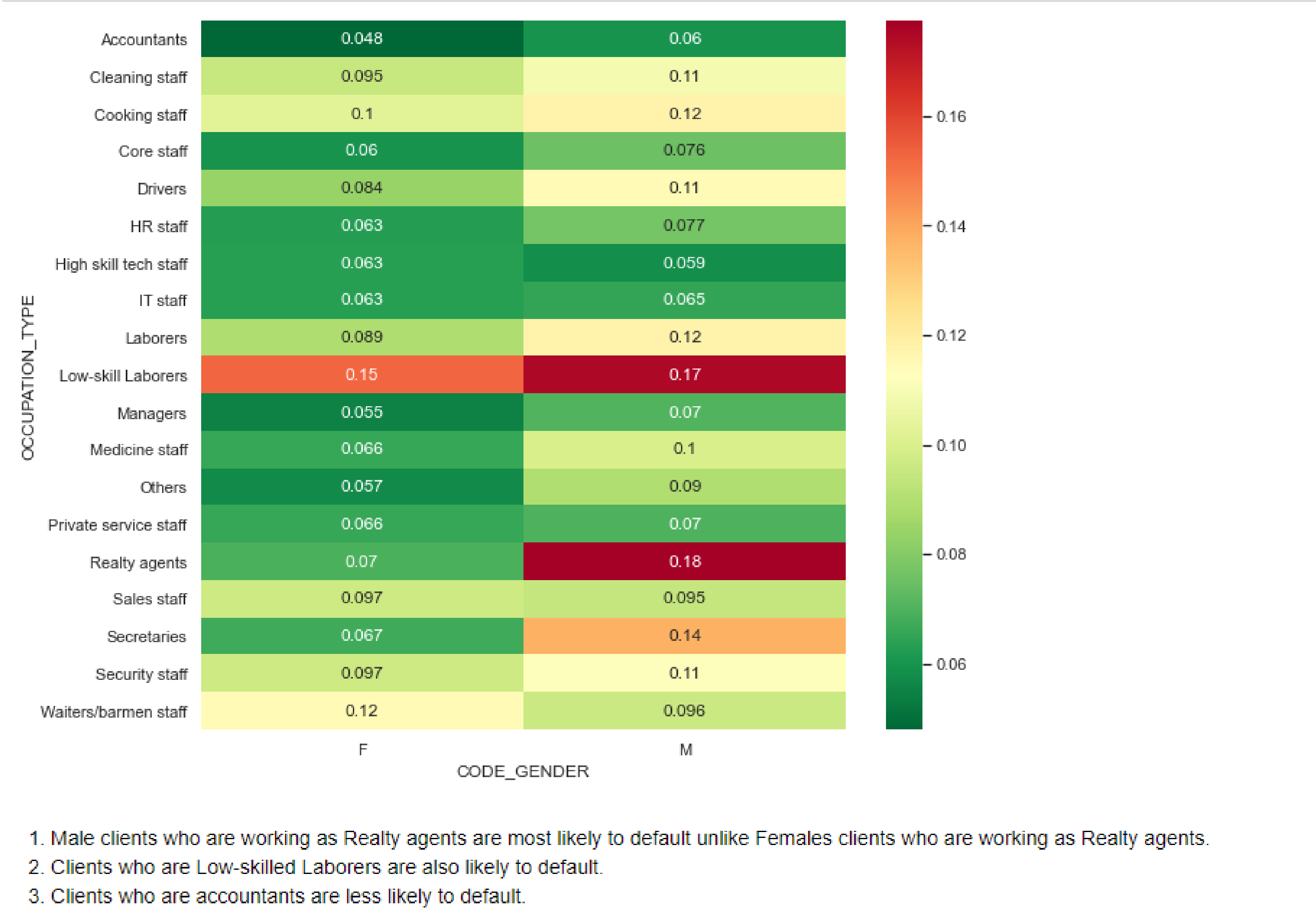
People with Secondary education are more likely to default, while people with Academic Degree are very less likely to default.

People who are Single/ Not married or with civil marriage are more likely to default than pay the loan

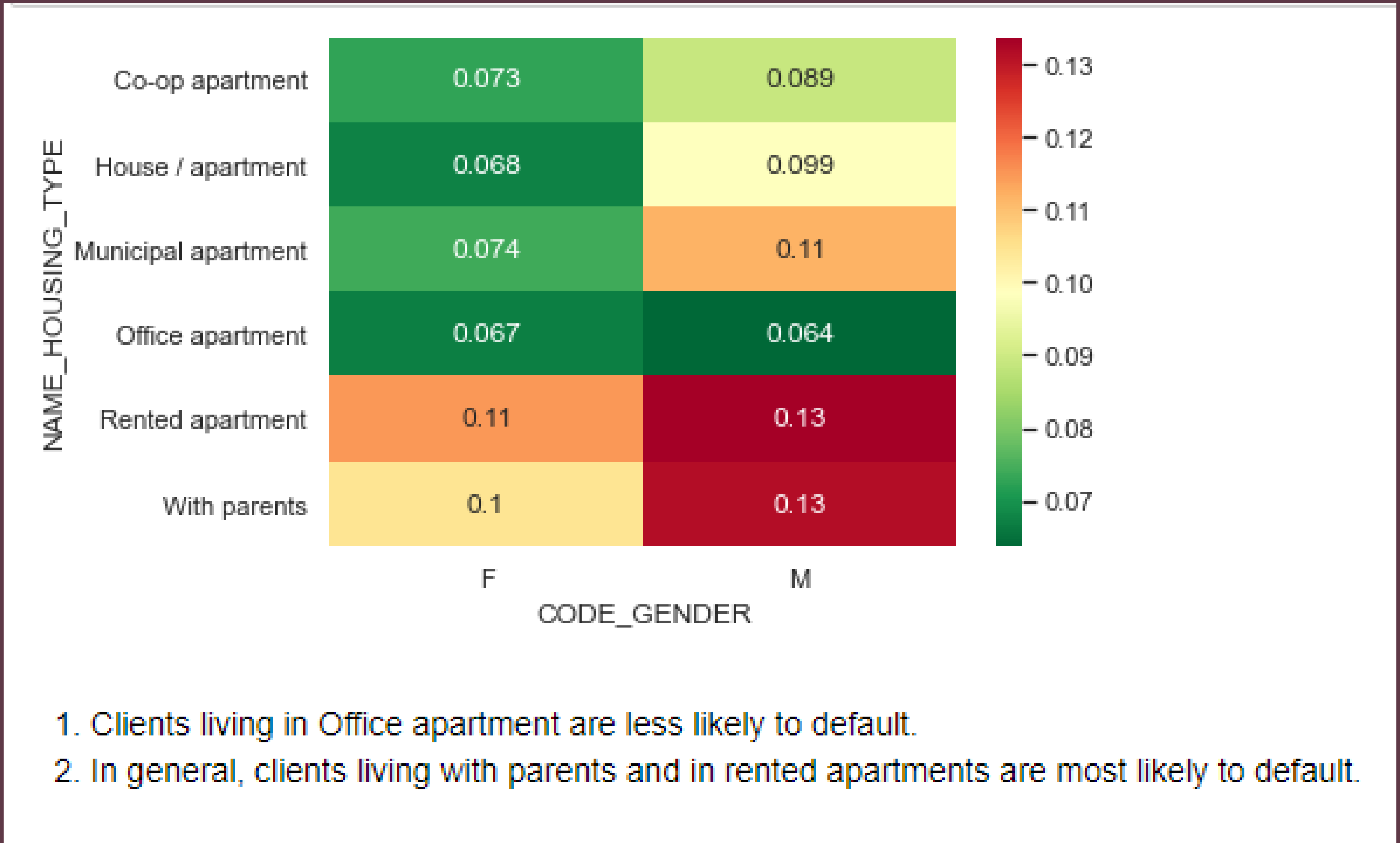
Most People taking the loan are living in Housing/Appartments. while people living with parents are likely to default more.

Top Correlation Factors

Case of potential client	Case of default
• Price of the product for which loan was applied for and amount of credit provided have strong correlation • There is a strong correlation between employment of the client and client having work phone. • Price of the product for which loan was applied for and annuity amount provided have strong correlation	Where Client is likely to default, • AMT_ANNUITY AMT_GOODS_PRICE • AMT_CREDIT AMT_GOODS_PRICE Correlation between above mentioned attributes is lower as compared to who are not likely to default.



1. Male clients who are working as Realty agents are most likely to default unlike Females clients who are working as Realty agents.
2. Clients who are Low-skilled Laborers are also likely to default.
3. Clients who are accountants are less likely to default.



NAME_HOUSING_TYPE	F	M
Co-op apartment	0.073	0.089
House / apartment	0.068	0.099
Municipal apartment	0.074	0.11
Office apartment	0.067	0.064
Rented apartment	0.11	0.13
With parents	0.1	0.13

1. Clients living in Office apartment are less likely to default.
2. In general, clients living with parents and in rented apartments are most likely to default.

Recommendations

- The bank can use these identified 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.

- The bank can make a checklist or Rule book to identify a possible “loan default” i.e. appearance of strong indicators of default. This will help them to reduce the risk of Credit Loss and Interest loss.

Conclusion

In the Credit EDA Case study we were exposed to a real

bank case scenario. We gained an understanding of risk analytics in banking and financial services.

The insights show us how banks can use the data analytics to minimize the risk of losing money while lending to customers.