



UPGRAD EDA ASSIGNMENT MODULE 11

By Chinmay Kumar Sahu



AGENDA



- Business Understanding.
- Problem Statement & Purpose
- EDA approach (*“application_data.csv”*)
- Analysis (*“application_data.csv”*)
- EDA approach (*“previous_application.csv”*)
- Analysis (*“previous_application.csv”*)
- Conclusion
- Credits



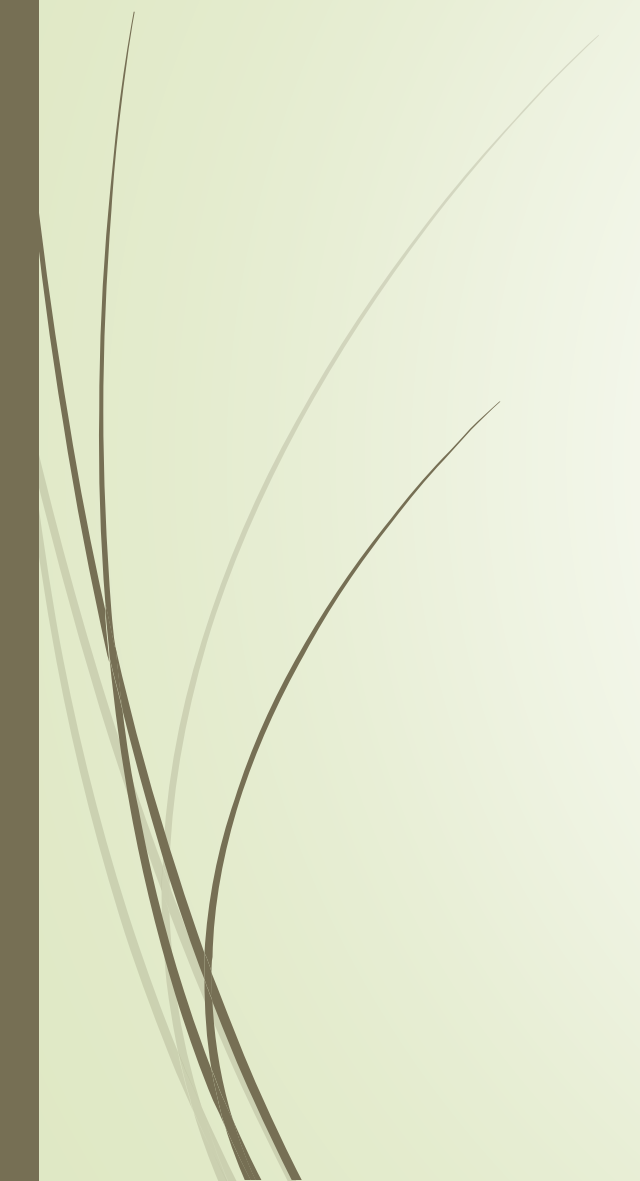
Business Objectives



- 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 the loan, lending (to risky applicants) at a higher interest rate, etc.



Problem Statement & Purpose

- 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 utilize this knowledge for its portfolio and risk assessment.
 - This can help the company to avoid major losses to select best candidate to provide the loan & ultimately increasing profit
- 



EDA Approach (*"application_data.scv"*)

1. Importing Necessary libraries for EDA.
2. Dataset Loading
3. Dataset Understanding
4. Dataset Preparation



EDA Approach (*"application_data.scv"*)

1. Importing Necessary libraries for EDA.

- Import the libraries which will help for analysis such as Numpy, Pandas, Seaborn, matplotlib & Warnings.

2. Dataset Loading

- Upload the given datasets.
- *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.



EDA Approach (*"application_data.scv"*)

3. Data Understanding.


- Shape (Rows:307511, Features:122)
- Using .describe() function to get the statistical information & distribution
- Data types available:
 - float64(65)
 - int64(41)
 - object(16)



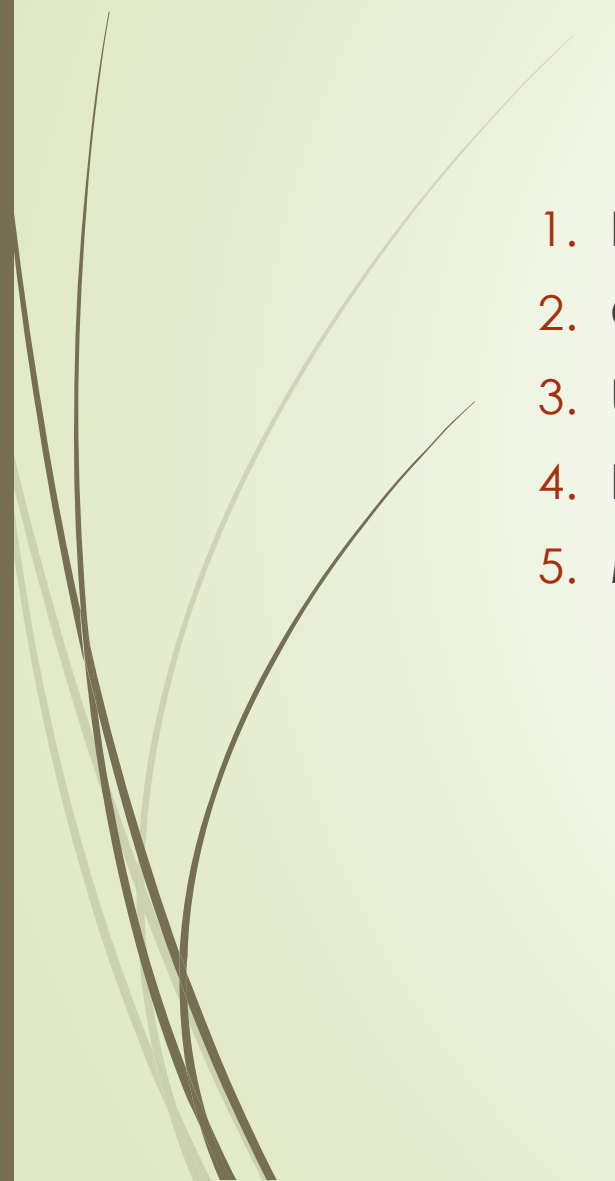
EDA Approach (*"application_data.scv"*)

4. Data Preparation

- Data has been inspected & all sanity check has been performed.
- Handling of **Null/ NA values** has been done on the data set.
- Upon further analysis, It was found that **49 features has more than 40% data missing** form the dataset.
- Hence all these **49 features has been dropped** to make the dataset viable. (The number 40 % was chosen by doing trails on the data)
- From the remaining 73 features again **irrelevant data(20 features) has been dropped**. This was achieved by using the *"columns_description.csv"* file.
- For the reaming 53 features again **Null values were checked & were imputed** based on the type of data [**numerical (MODE) & categorical (MEDIAN)**].
- Upon Observation Certain features in data had **Negative values** which were taken care.
- Also **"XNA"** type of data was also corrected.
- Based on the data given further **segmentation** was performed for numerical features.

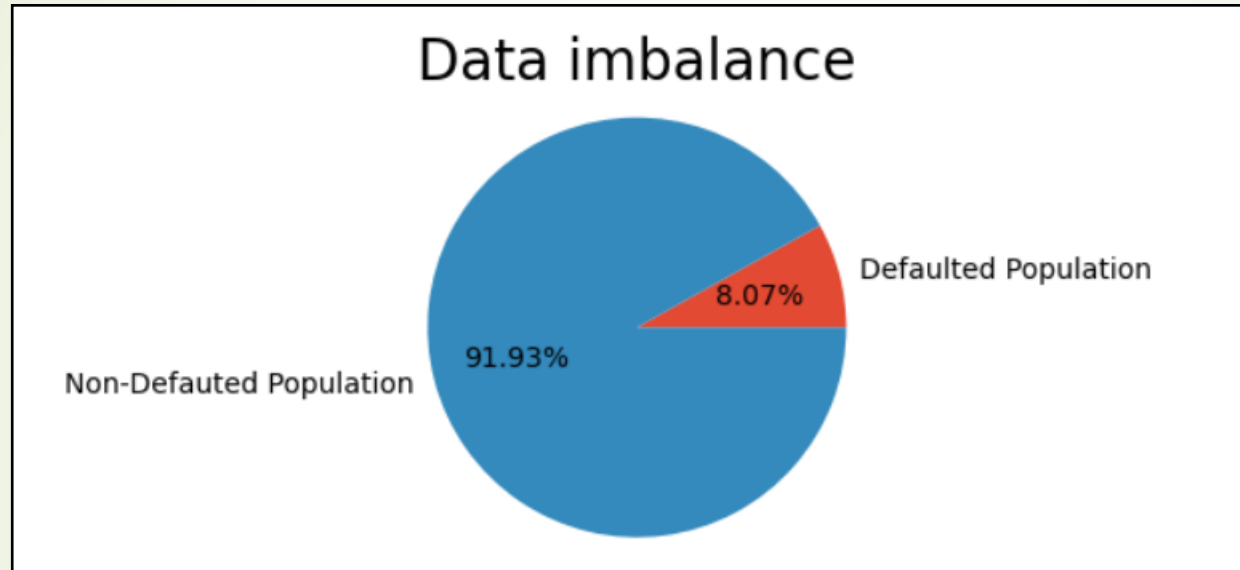


Analysis (*"application_data.scv"*)

1. Data Imbalance
 2. Outlier Analysis
 3. Univariate Analysis
 4. Bivariate Analysis
 5. Multivariate Analysis
- 

Analysis (*"application_data.scv"*)

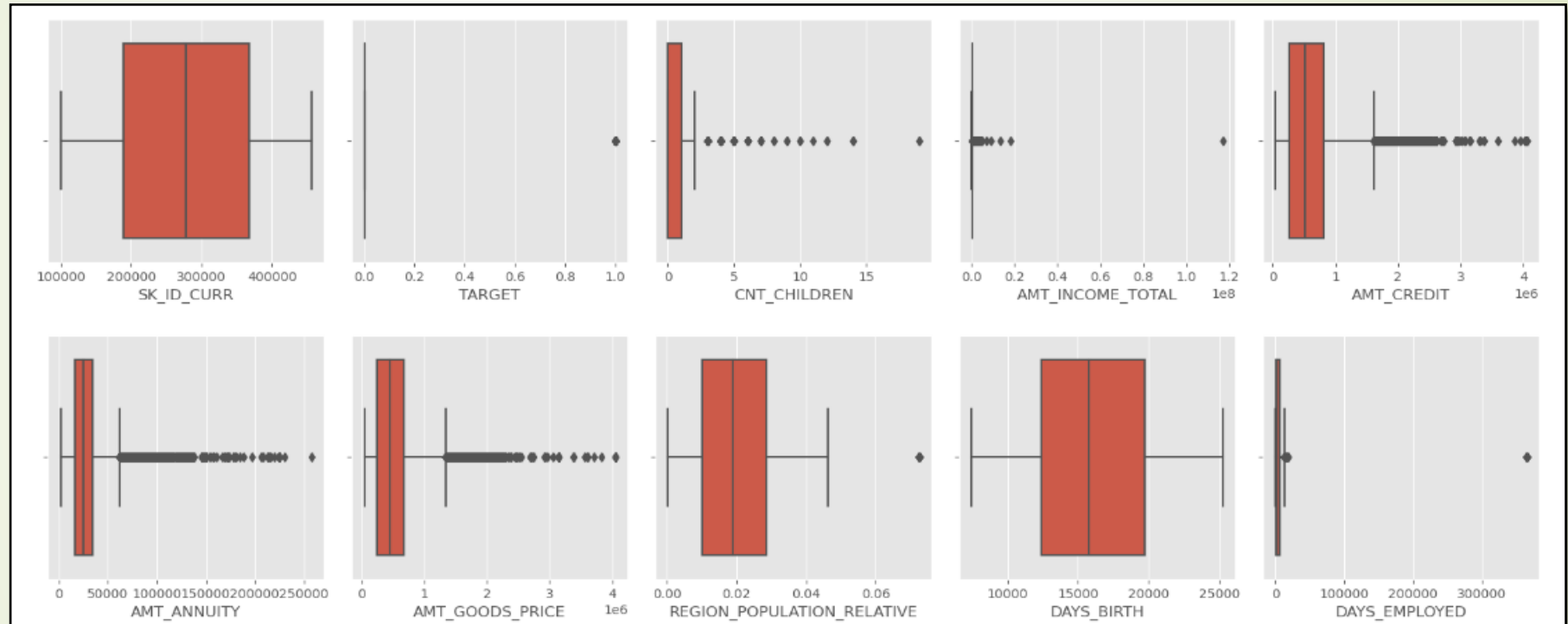
1. Data Imbalance



- As per the pie chart the given **data is highly imbalanced**.
- Defaulted Population is 8.07%
- Non-Defaulted Population is 91.93%
- **Ratio of Imbalance is 11.39%.**

Analysis ("application_data.scv")

2. Outlier Analysis



- CNT_CHILDREN has outliers highest up to 19 children.
- AMT_CREDIT, AMT_ANNUITY & AMT_GOODS_PRICE has outliers. Majority of the points are present in the Q3 to MAX range for all 3 features.
- Namely around 22 features has outliers present.(NOTE: no capping & flooring performed for these outliers)

Analysis (*"application_data.scv"*)

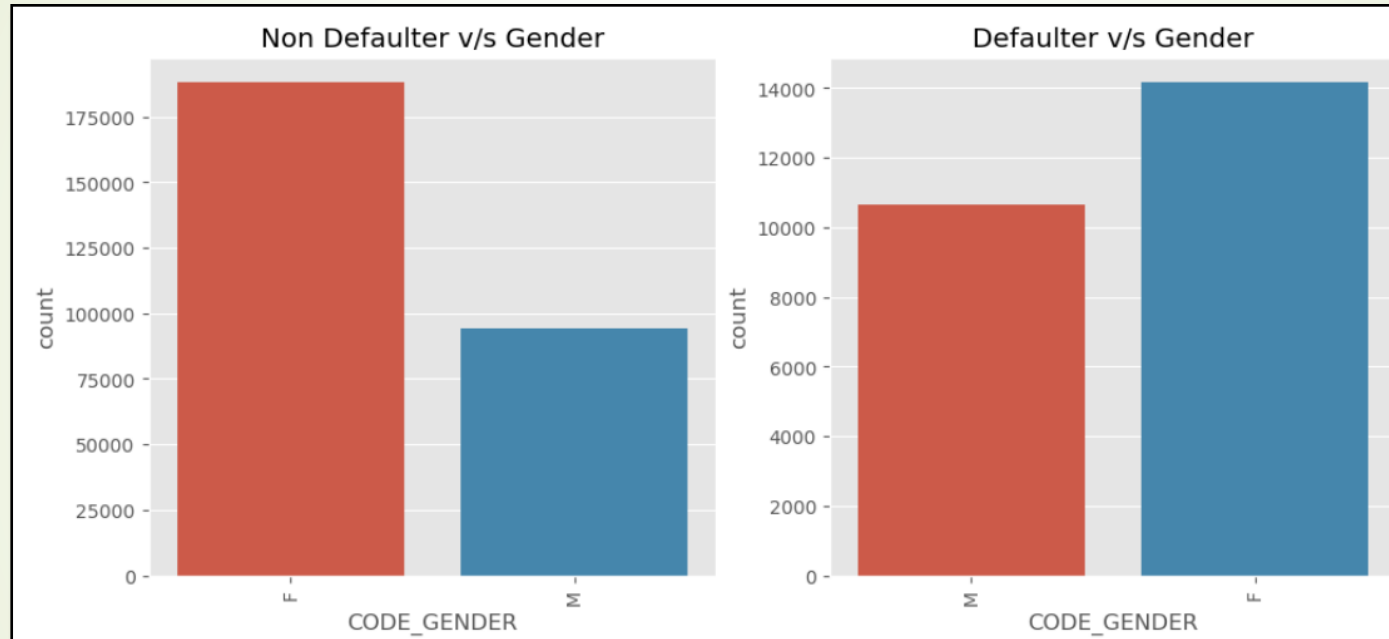
3. Univariate Analysis



- Age group of **40-60 has been facing difficulties** has the highest percentage of applying for loan.
- Age group of **40-60 has been facing difficulties** to repay the loans.
- The most reliable **age group not having difficulties** to pay loans is **above 60 years**.

Analysis (*"application_data.scv"*)

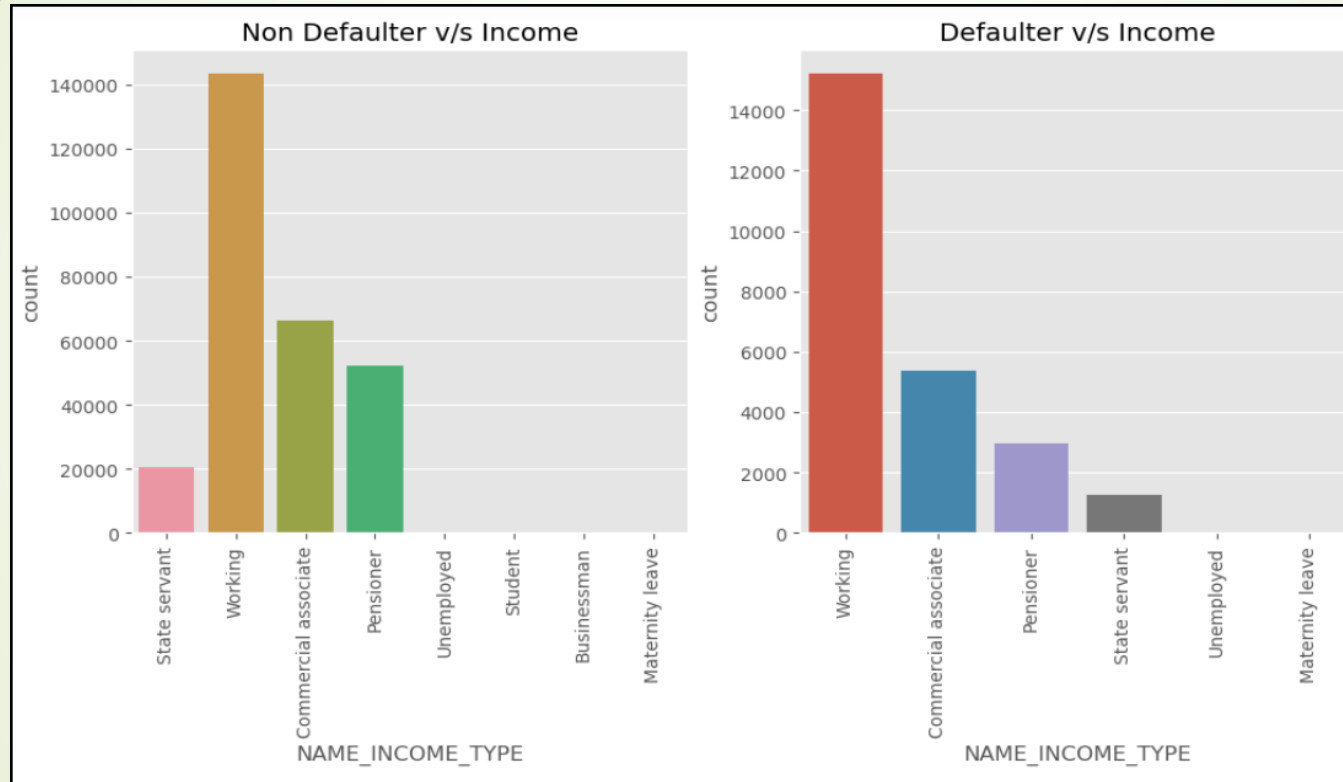
3. Univariate Analysis



- As per the graph Female clients has higher application rate than male clients for loan.
- **66.6% Female** clients are **non-defaulters** while **33.4% male** clients are **non-defaulters**.
- **57% Female** clients are defaulters while **42% male** clients are **defaulters**.

Analysis (*"application_data.scv"*)

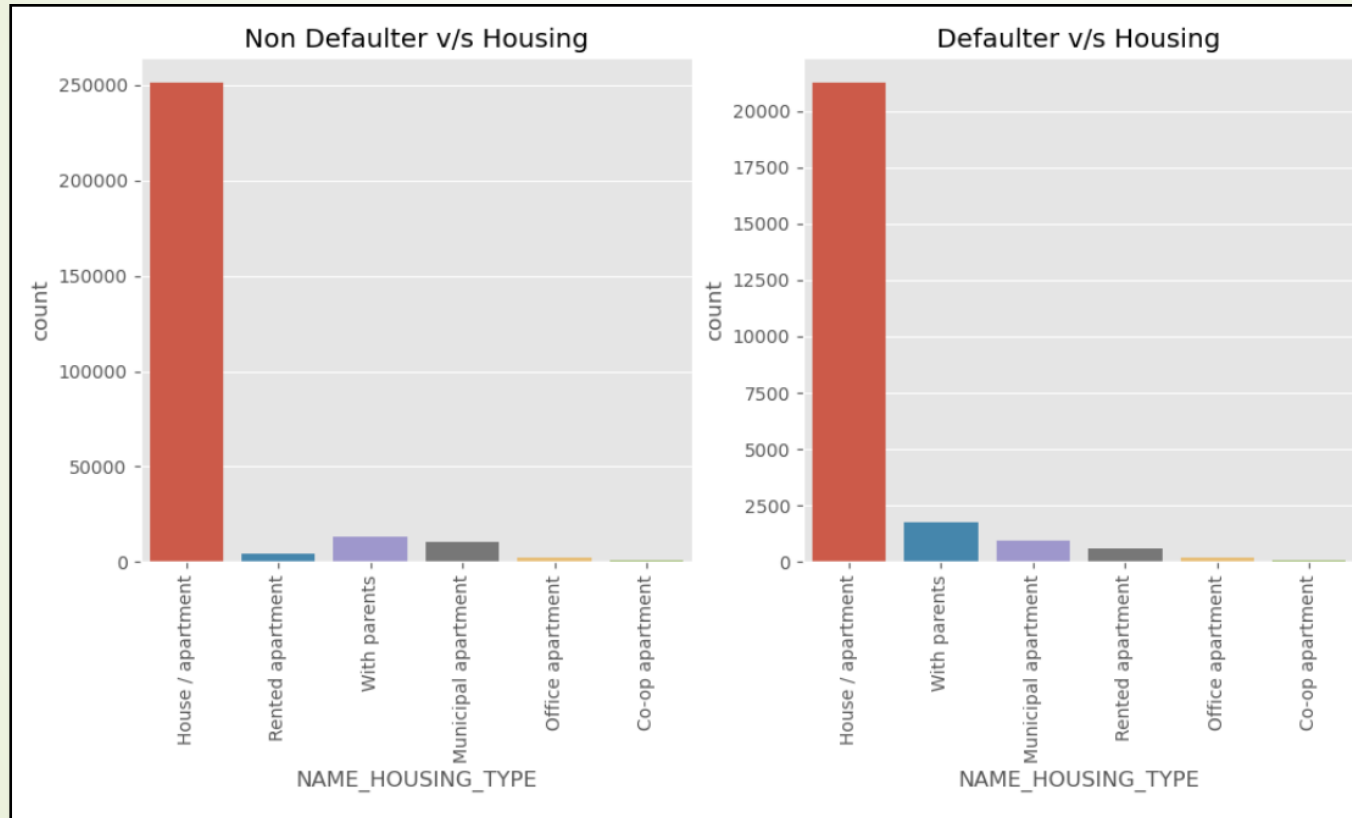
3. Univariate Analysis



- As per the graph working category has the high risk of not paying the loans.
- State Servant being an reliable category to provide the loan.

Analysis (*"application_data.scv"*)

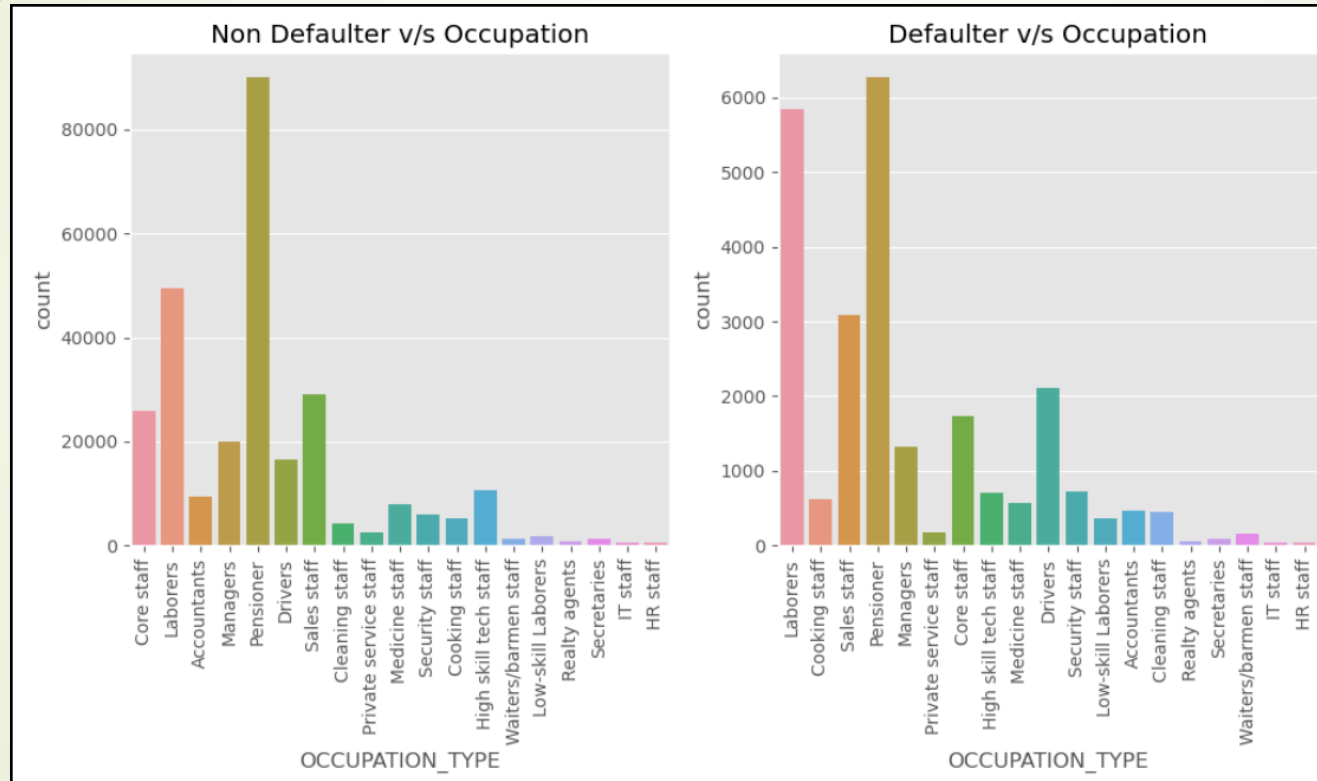
3. Univariate Analysis



- As per the graph Clients living with parents have difficulties to repay loan.
- While Clients having House/Apartments are dominant in both the sections

Analysis (*"application_data.scv"*)

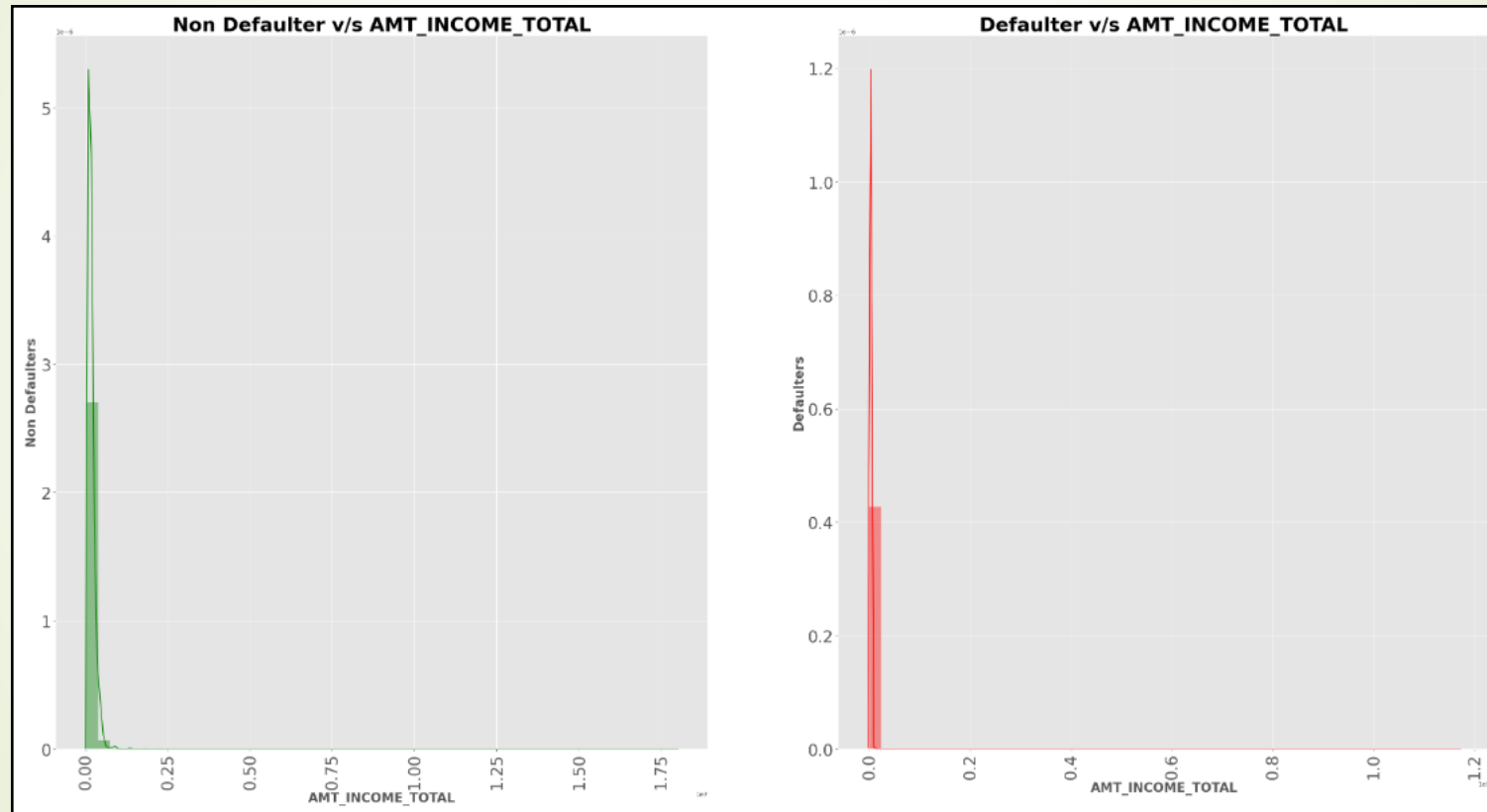
3. Univariate Analysis



- Laborers has a high default rate as per the above graph.
- Pensioners has the highest percentage of loan applications & are dominant in both category.

Analysis (*"application_data.scv"*)

3. Univariate Analysis



- Defaulter has less spread in the INCOME graph , where as Non Defaulters has significant spread.

Analysis (*"application_data.scv"*)

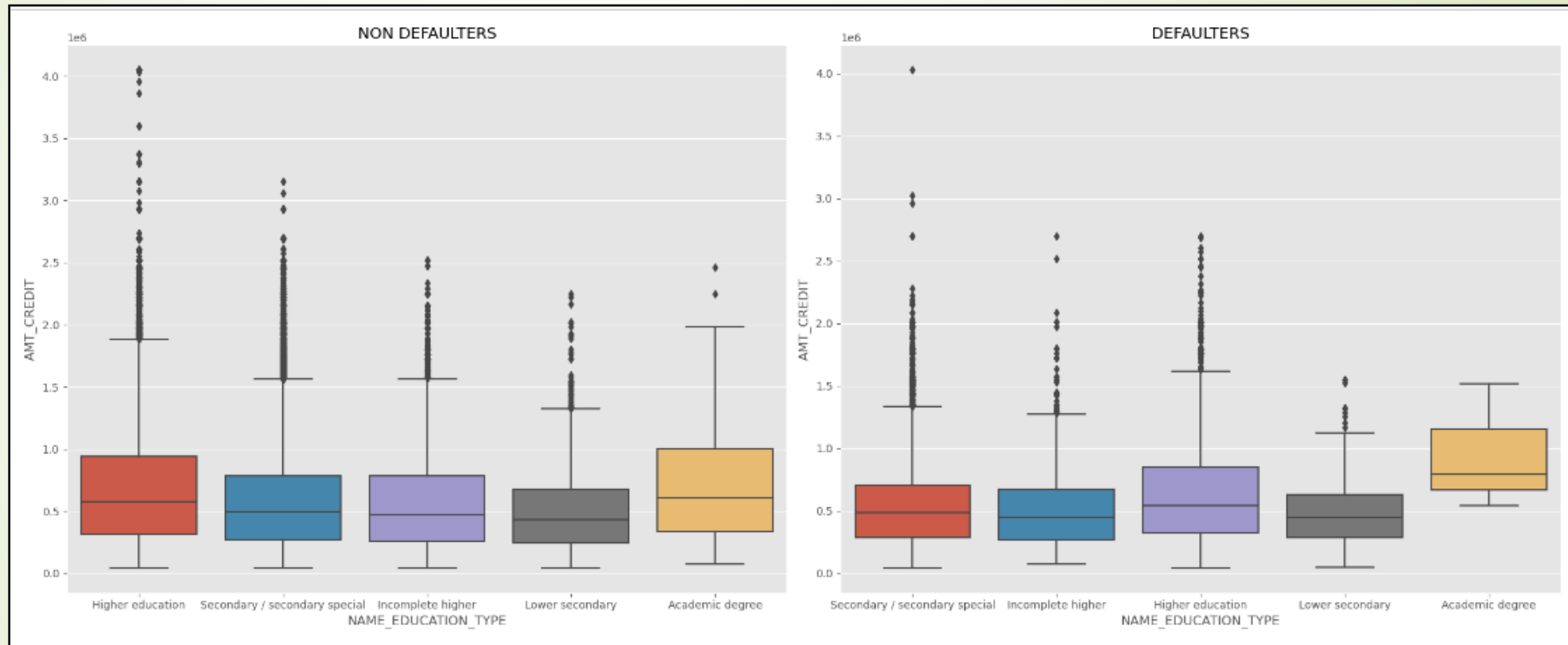
4. Bivariate Analysis



- The graphs shows almost a linear relation ship between AMT_CREDIT & AMT_GOODS_PRICE.

Analysis ("application_data.scv")

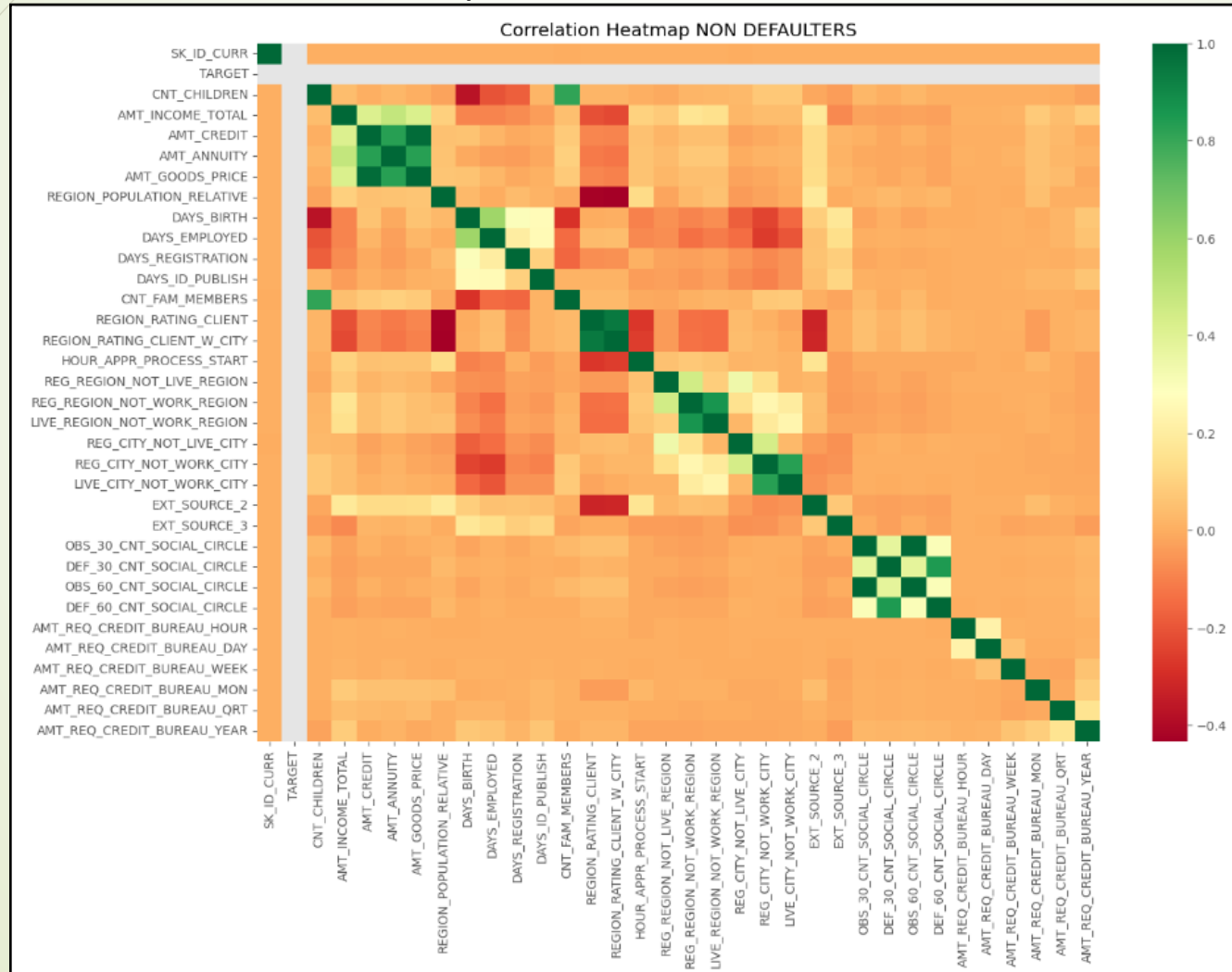
4. Bivariate Analysis



- The Higher Education category seems to have less difficulties in repaying loan.
- Clients with Academic Degree have high default rate.
- Clients with Lower secondary, Incomplete Higher seems to be safer after academic degree.

Analysis ("application_data.scv")

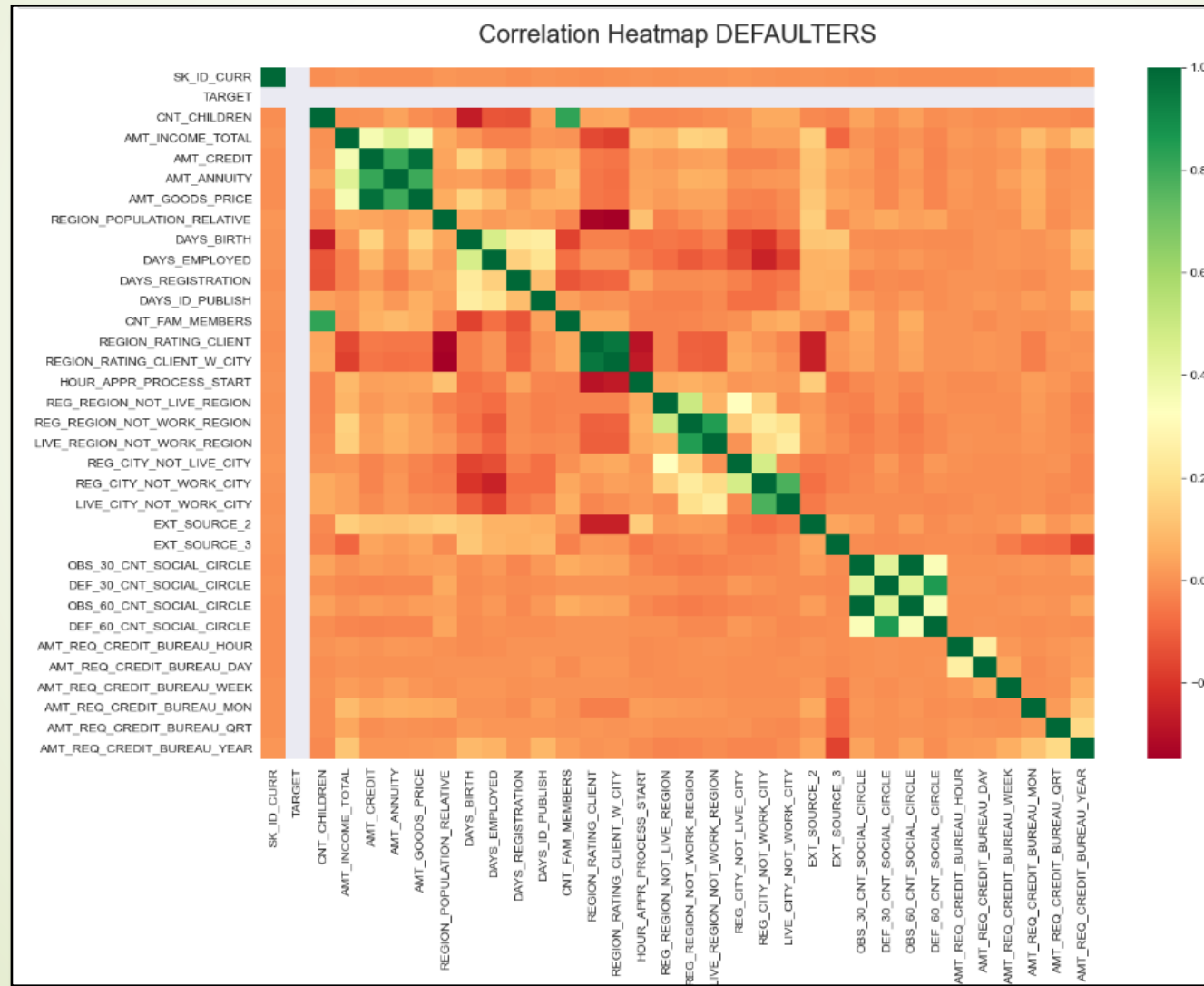
4. Multivariate Analysis




- AMT_GOODS_PRICE & AMT_CREDIT is highly correlated
- AMT_CREDIT, AMT_GOODS_PRICE, AMT_ANNUITY area is highly dense in terms of relations.

Analysis ("application_data.scv")


4. Multivariate Analysis




- AMT_GOODS_PRICE & AMT_CREDIT is highly correlated
- AMT_CREDIT, AMT_GOODS_PRICE, AMT_ANNUITY area is highly dense in terms of relations.
- Considering the fact that AMT relations are high in both categories it may not be a good factor to detect the defaulters.



EDA Approach (*“previous_application.scv”*)

1. Dataset Loading
 2. Dataset Understanding
 3. Dataset Preparation.
 4. Dataset merging with *“application_data.csv”*
- 




EDA Approach (*“previous_application.scv”*)

1. Dataset Loading

➤ Upload the given datasets.

- *previous_application.csv*: contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.



EDA Approach (*“previous_application.scv”*)

2. Dataset Understanding

- Shape (Rows: 1670214, Features: 37)
- Using .describe() function to get the statistical information & distribution
- Data types available:
 - float64(15)
 - int64(6)
 - object(16)



EDA Approach (*“previous_application.scv”*)

3. Dataset Preparation

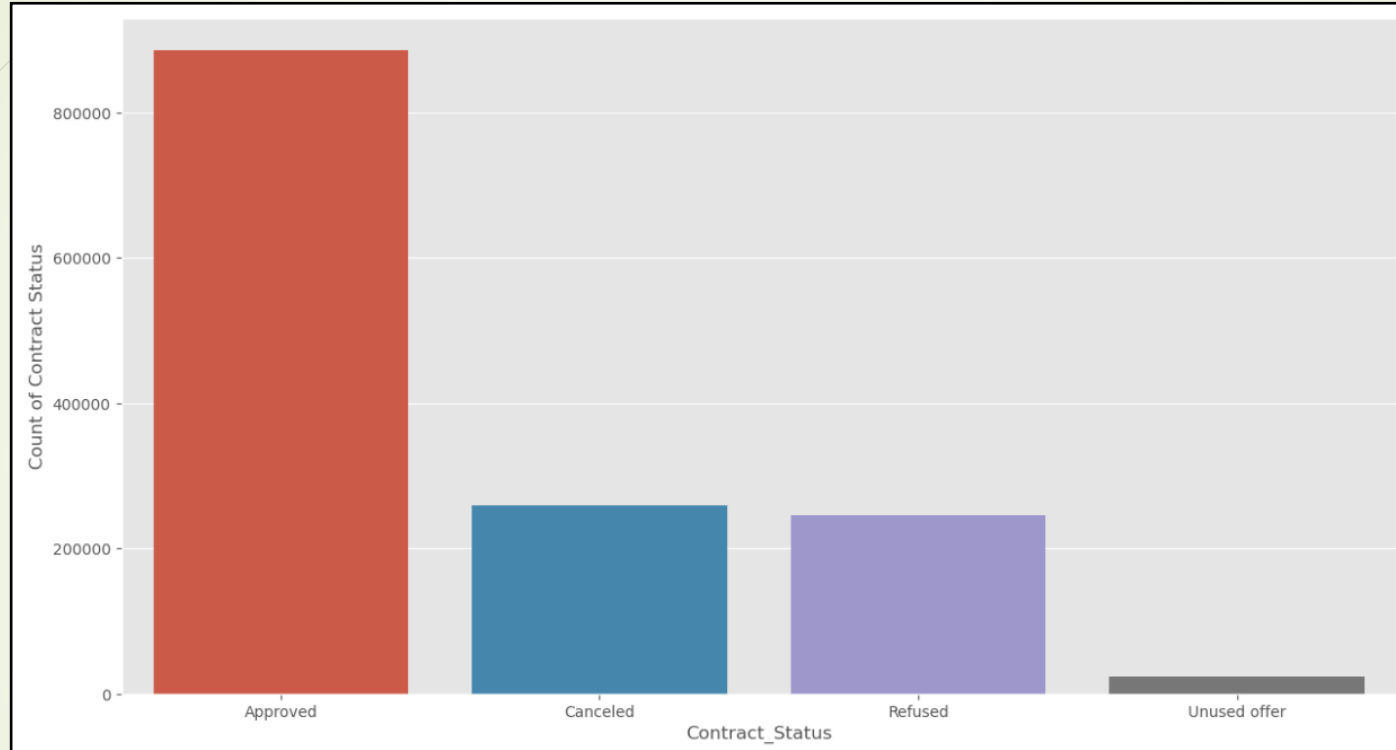
- Data has been inspected & all sanity check has been performed.
- Handling of **Null/ NA values** has been done on the data set.
- Upon further analysis, It was found that **11 features has more than 40% data missing** form the dataset.
- Hence all these **11 features has been dropped** to make the dataset viable. (The number 40 % was chosen by doing trails on the data)
- For the reaming 53 features again **Null values were checked & were imputed** based on the type of data [**numerical (MODE) & categorical (MEDIAN)**].
- Also **“XNA”** type of data was also corrected.
- Based on the data given further **segmentation** was performed for numerical features.

EDA Approach (*"previous_application.scv"*)

1. Dataset merging with *"application_data.csv"*

- `merged_df = pd.merge(left=df1_drop2, right=df2_drop1, how='inner', on='SK_ID_CURR')`
as *"SK_ID_CURR"* is the common feature between these two datasets.
- Shape (Rows: 1413701,, Features: 72)
- Using `.describe()` function to get the statistical information & distribution.
- 80.65% clients were repeaters applying for loan.
- 14.47% clients are new applying for the loan.
- Data types available:
 - float64(24)
 - int64(19)
 - int32(1)
 - object(26)
 - Category(2)

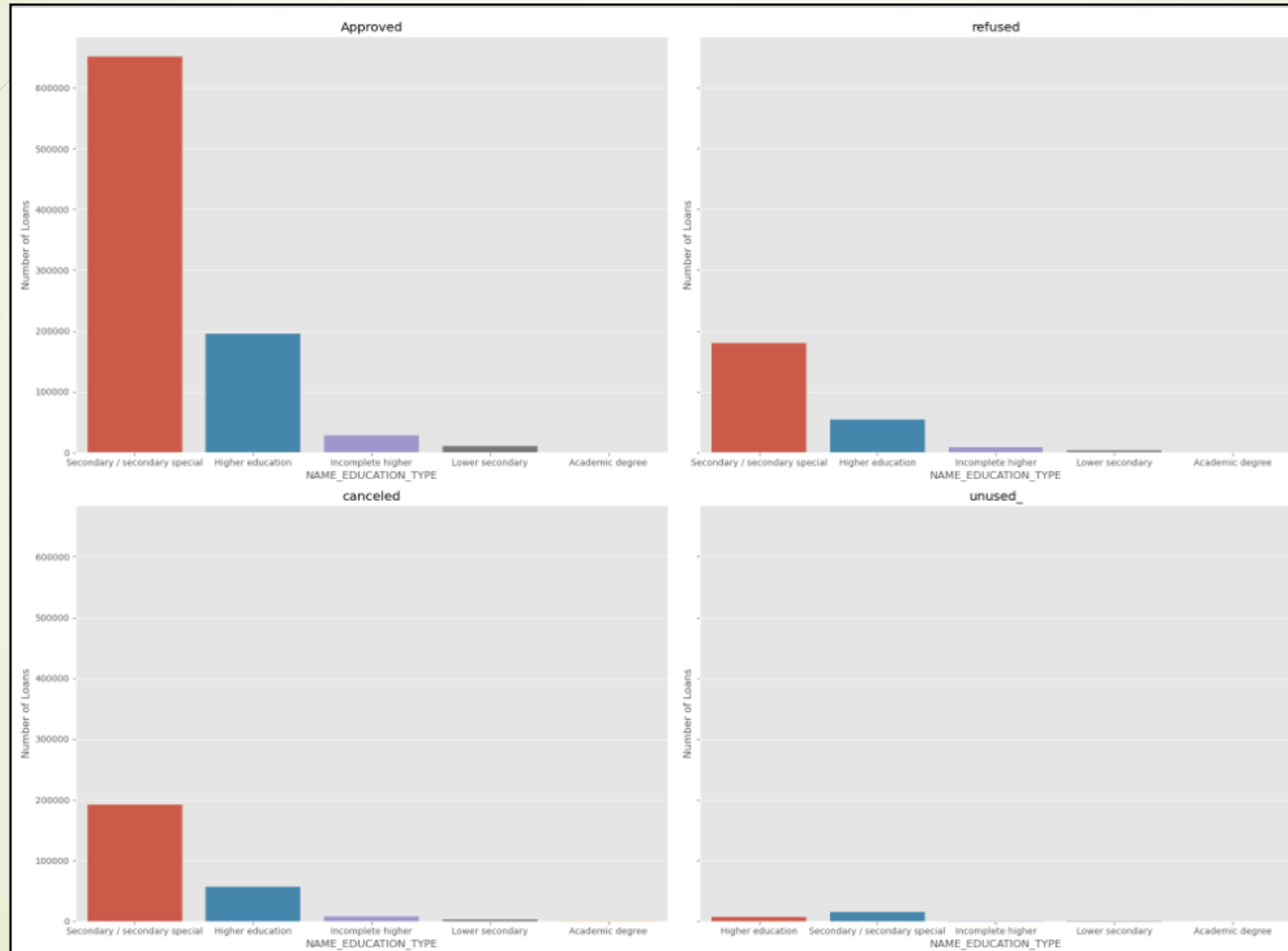
Analysis (*"previous_application.csv"*)



Percentage of contracts approved or not in previous applications

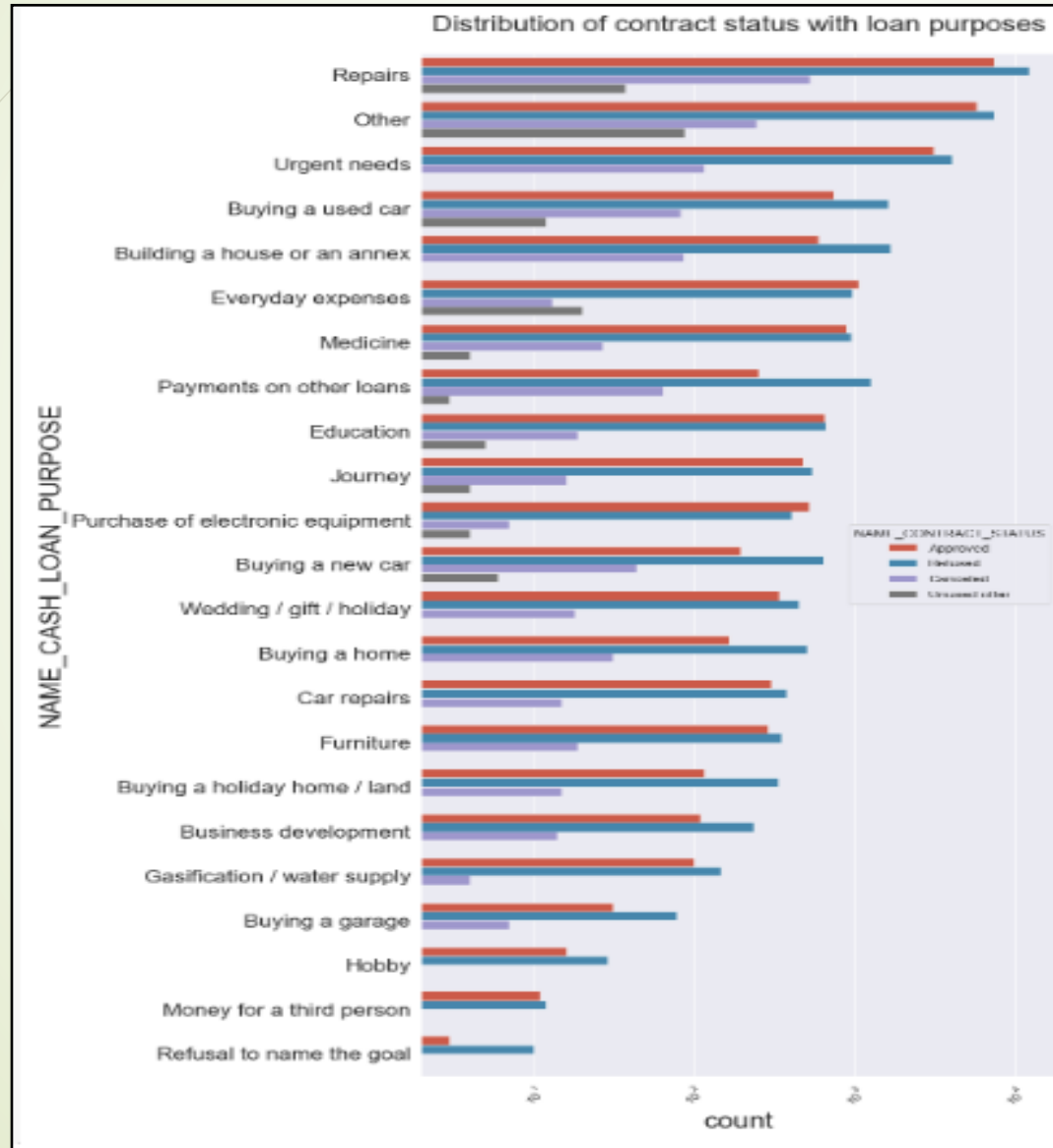
- Approved :- 38.8%
- Refused :- 58.5%
- Canceled :- 2.3%
- Unused offer :- 0.31%

Analysis ("previous_application.scv")



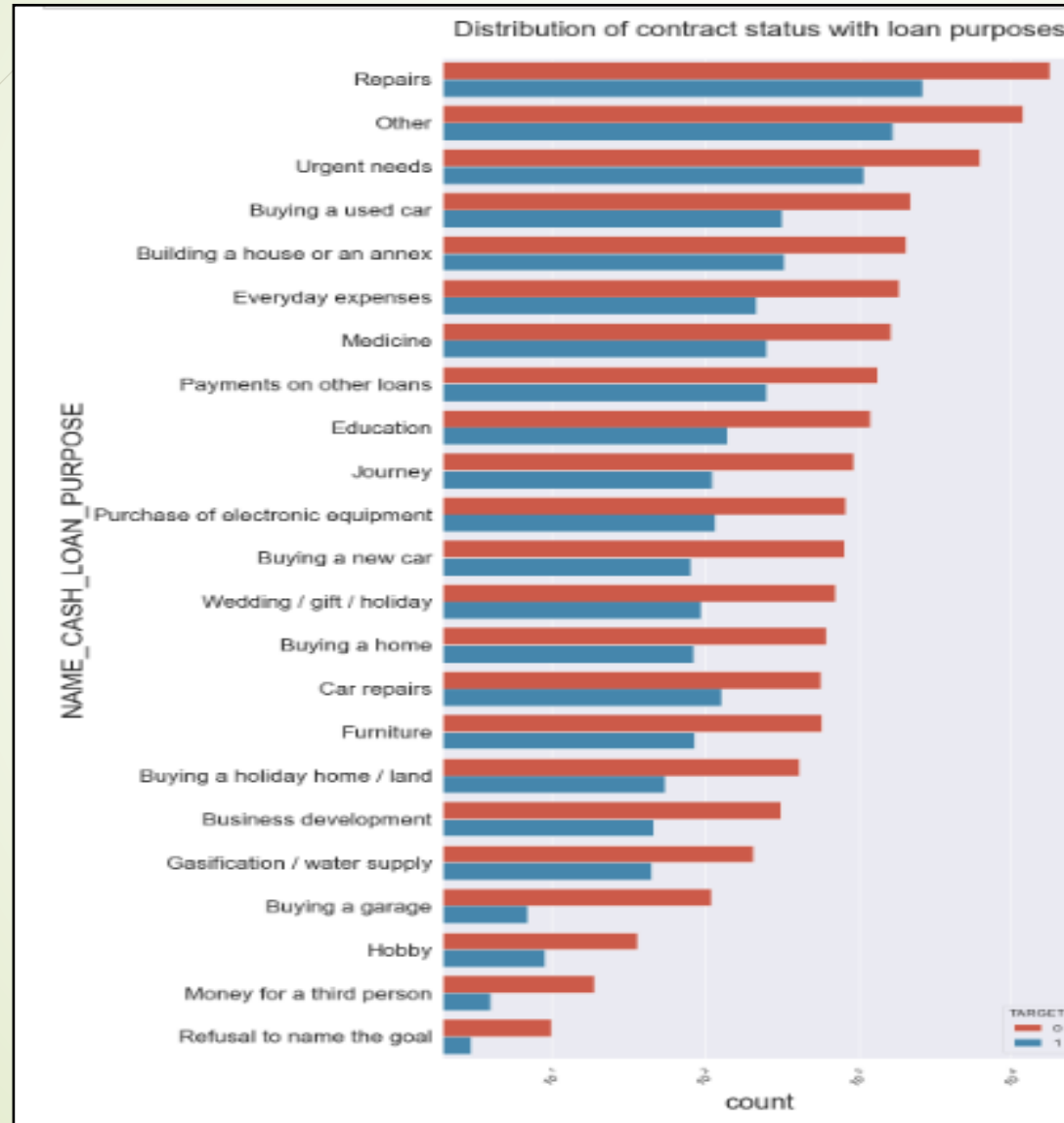
- Approved has the highest percentage of Secondary/Secondary special category.

Analysis ("previous_application.scv")



- Most rejection of loans came from purpose "Repairs".
- For "Education" & "Medicine" purposes we have almost equal number of approves and rejections
- "Paying other loans", "Buying a used Car", "Buying a new car" is having significant higher rejection than approvals.

Analysis ("previous_application.scv")



- "Repairs" again for both the defaulter & defaulters has same trend as in previous
- "Buying used Car" and Building purpose client having difficulties in payment have equal ratio of defaulting & non defaulting



Conclusion

- Ratio of Imbalance is **11.39%**.
- Most of the clients targeted should be
 1. **State Servant** in Income category type
 2. Clients with good academic degree.
 3. Pensioners have less difficulties in repaying loans.
 4. Clients having age **above 60 years**.
- Clients through that bank can face loss:
 1. **Laborers** in occupation type.
 2. Clients are in **with parents/ rented apartments**
 3. **unemployed, Maternity Leave, student clients**
 4. **Repair** is having higher number of unsuccessful payments on time.



CREDITS



- [UPGRAD](#)
- <https://www.tibco.com/reference-center/what-is-risk-analytics>
“getting to know what is risk analytics.”
- <https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>
- <https://s4be.cochrane.org/blog/2015/07/24/nominal-ordinal-numerical-variables/>
- <https://seaborn.pydata.org/index.html>
- <https://docs.python.org/3/>



THANK YOU