Credit EDA Case Study

Risk Analytics in Banking and Financial Services

–Suryansh Bhardwaj

Problem Statement

- There are 2 types of risks associated with the bank’s decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

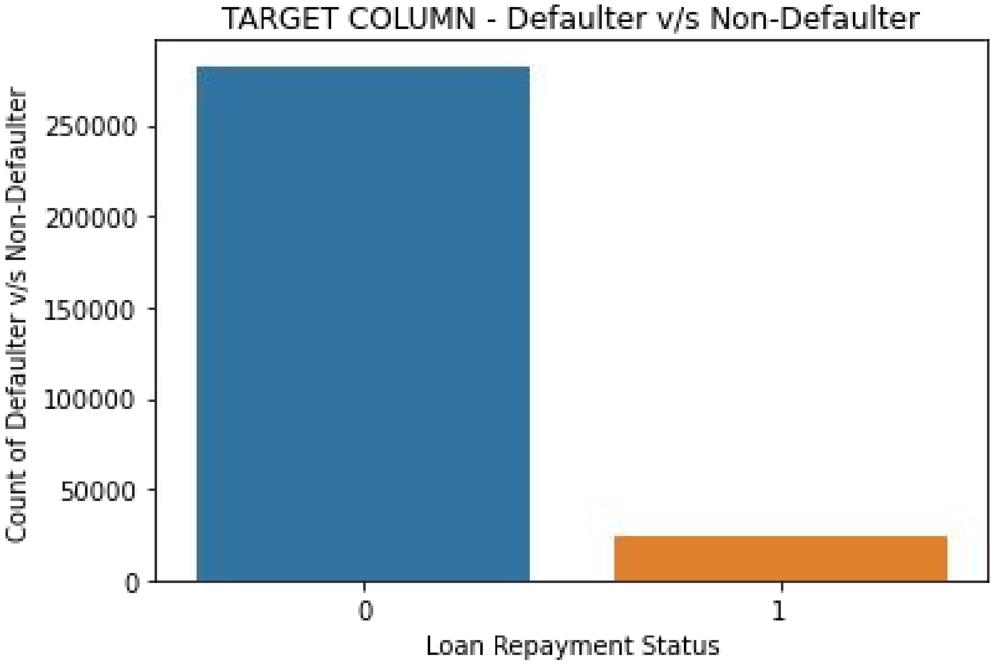
Analysis Basic Approach

- Steps:

1. Finding missing values and making a decision of which values to handle and how to handle those missing values.
2. Checking Outliers in the data provided.
3. Checking Data Imbalance and the ratio of imbalance.
4. Finding Top 10 correlations for the Clients with payment difficulties for both application data and previous application data.
5. Finding interesting Insights and correlations between features.

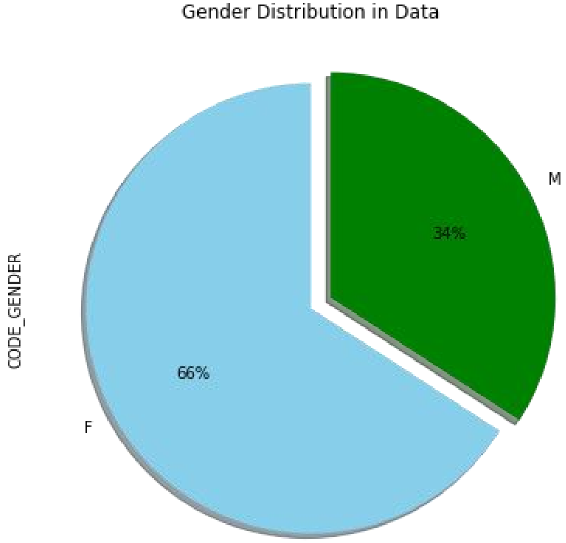
PROPORTION OF DEFAULTERS IN THE DATASET.

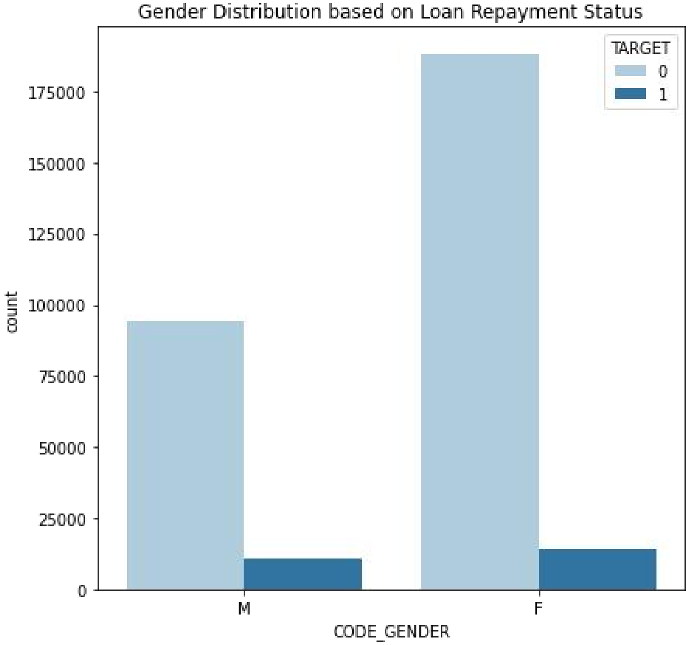
There is huge data imbalance between number of defaulters v/s number of non-defaulters.



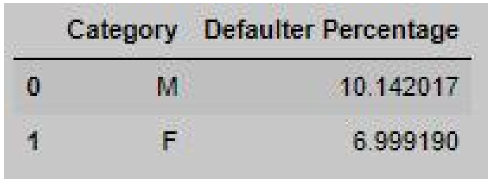
Percentage of non-defaulters = 91.93%
Percentage of defaulters = 8.07%
Data Imbalance Ratio => approx. 8:92 = 2:23

Gender Imbalance in Application Data



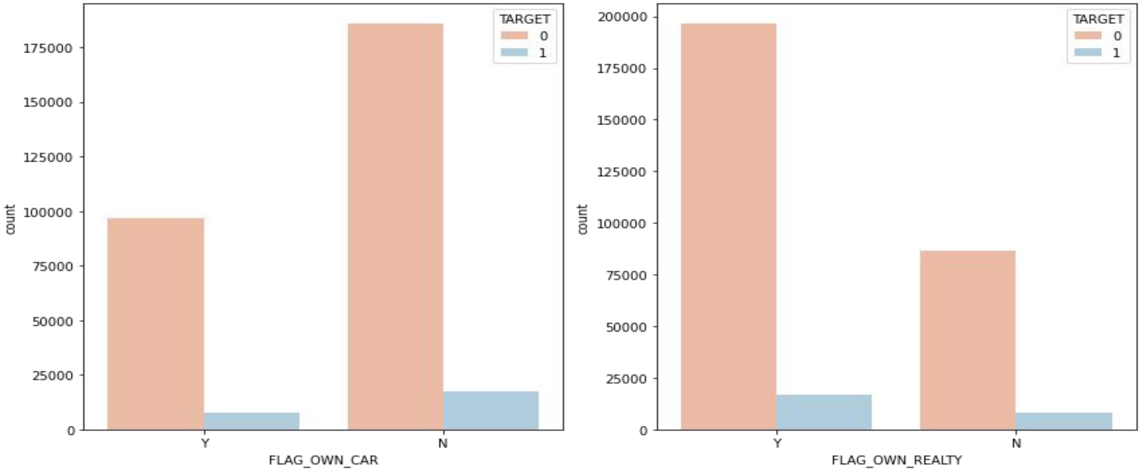


Percentage of Females in data = 66%
Percentage of Males in data = 34%



	Category	Defaulter Percentage
0	M	10.142017
1	F	6.999190

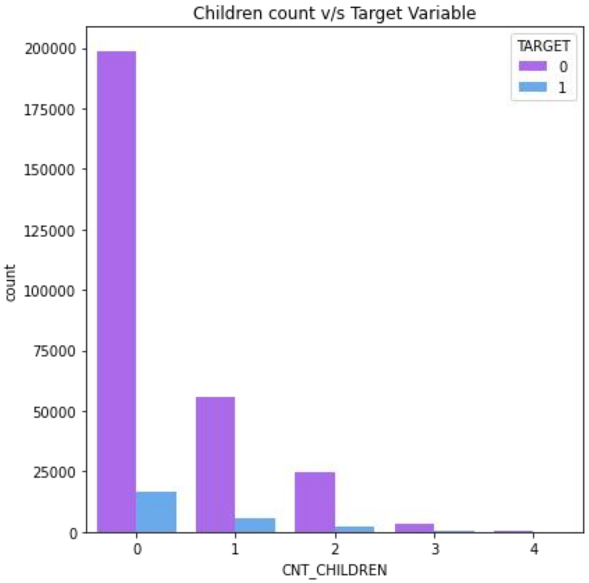
ASSET DETAILS OF APPLICANTS (HOUSE AND CAR)

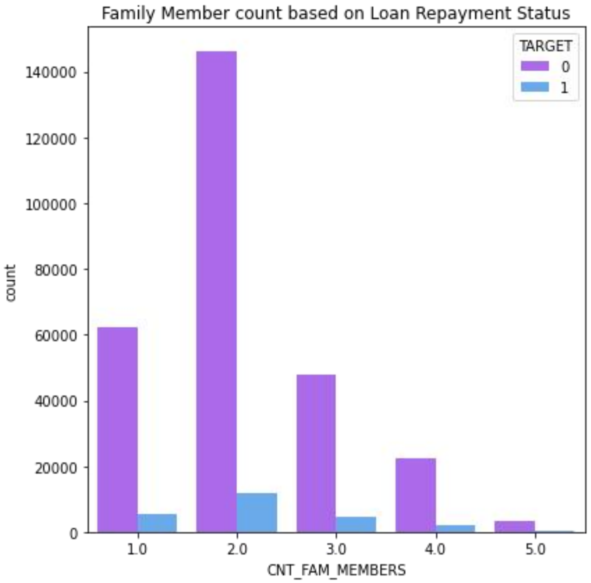


Insights:

- Number of applicants who own a car are much less than the applicants who don't own a car.
- Number of applicants who own a realty are much more than the number of applicants who don't own a realty.
- According to defaulter percentage, applicants who don't own a car and realty are much more likely to be defaulters than the ones who have their own cars and realty.

PROPORTION OF DEFAULTERS BY NUMBER OF KIDS

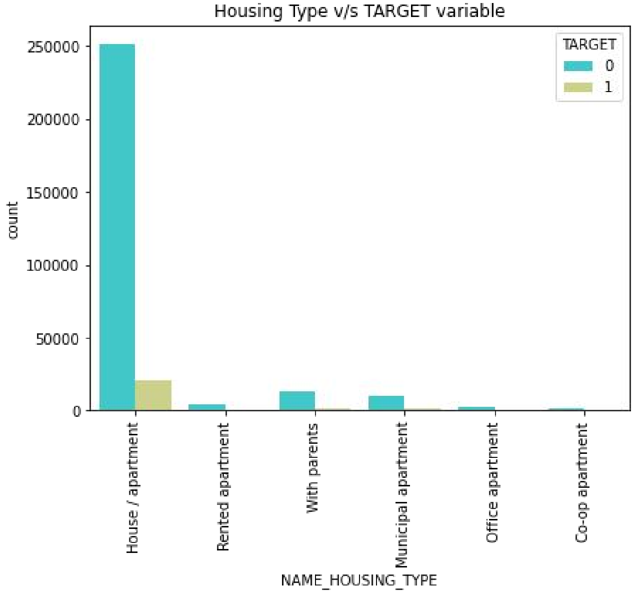


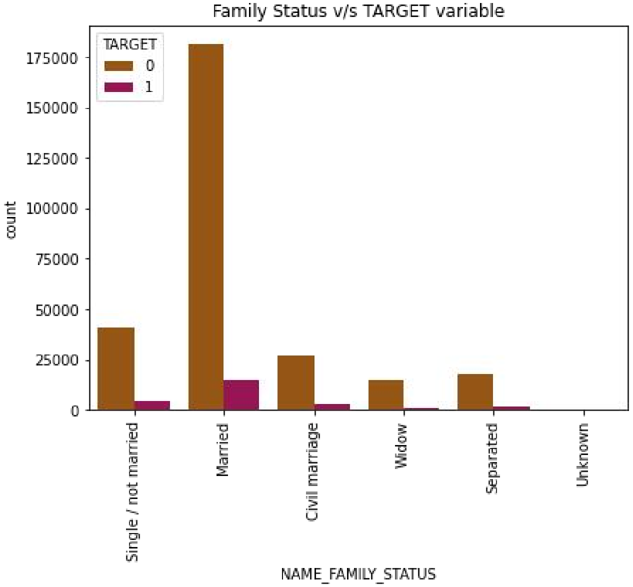


Insights:

- Applicants with very large number of children/family member count and defaulter percentage also high or even low,cannot be used for providing insights since the frequency of such applicants is very low.
- Applicants with 0 children have a defaulter percentage of 7.71%, so they are likely to be non-defaulters/repayers.
- Applicants with 2 family members have a defaulter percentage of 7.58%, so they are likely to be non-defaulters/repayers and should be approved for there loan.

Proportion of Defaulters by Housing Type and Family Status

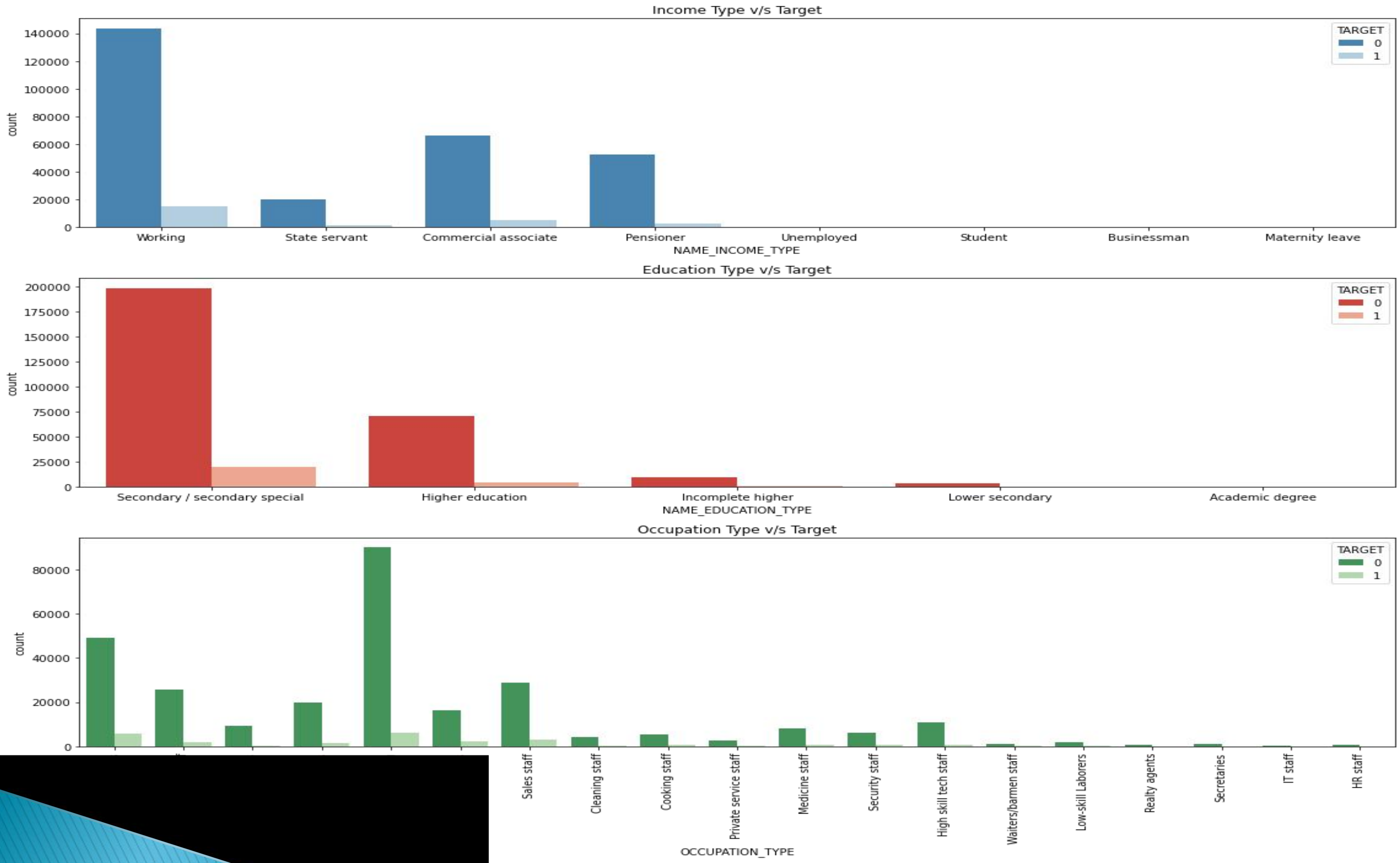




Insights from above 2 count plots:

- A lot of applicants live in House/Apartment
- Applicants with Family Status as "with parents" or "Rented apartment" have much higher rate of default.
- Most of the applicants are married and still with 7.55% defaulter percentage, which is really good

INSIGHTS FOR PROFESSIONAL LIFE OF APPLICANT

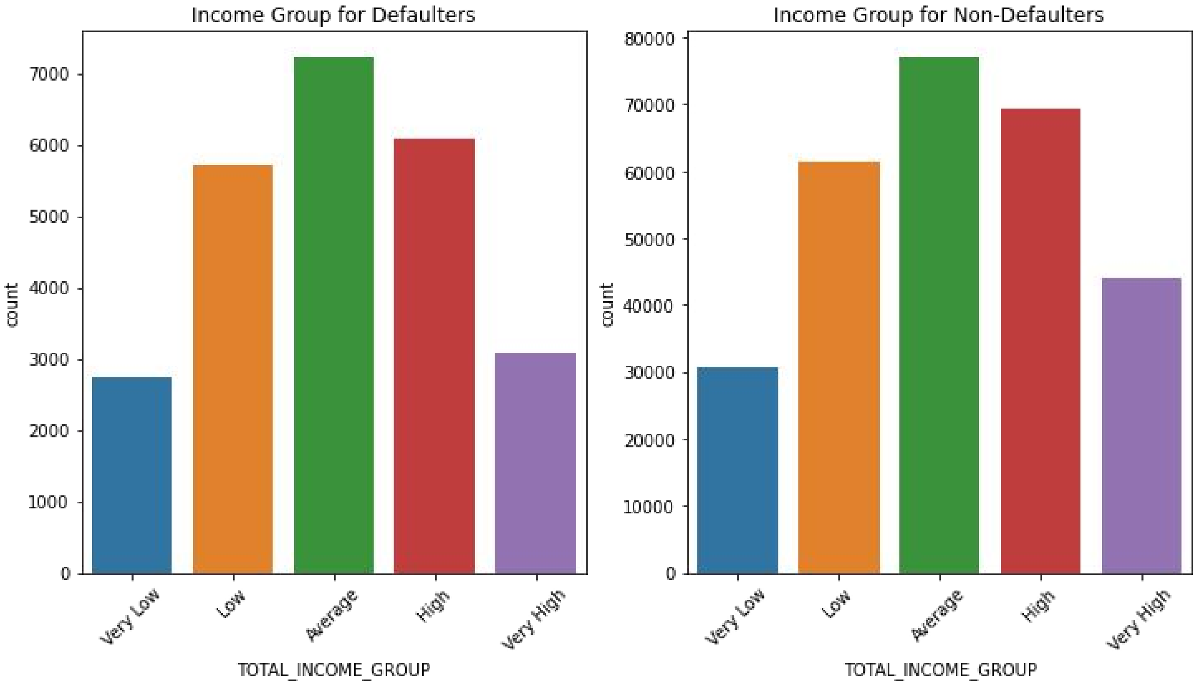


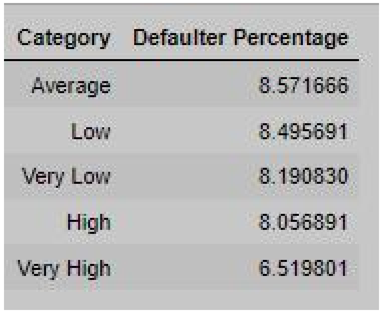
Continue.. (Professional Life)

Insights:

- Applicants on Maternity leave or if they are unemployed have the highest chances of defaulting.
- Applicants who are working, or Pensioner or Commerical associate are good since the defaulting percentage is not much.
- Applicants in ('Unemployed','Student','Businessman','Maternity leave') don't contribute much to the analysis as there is not much data available for these types.
- Applicants with 'Academic degree' even with defaulter percentage only 1.83 won't contribute to the analysis because there is not much data.
- There is a clear observation that applicants of education type 'Lower Secondary' have high defaulting chance.
- Most of the applicants data have Occupation type as "Unknown/Missing" and defaulter percentage is 6.5%.
- Low-skill laborers are the applicants with highest defaulter percentage.

PROPORTION OF DEFAULTERS VARIED IN INCOME GROUPS.



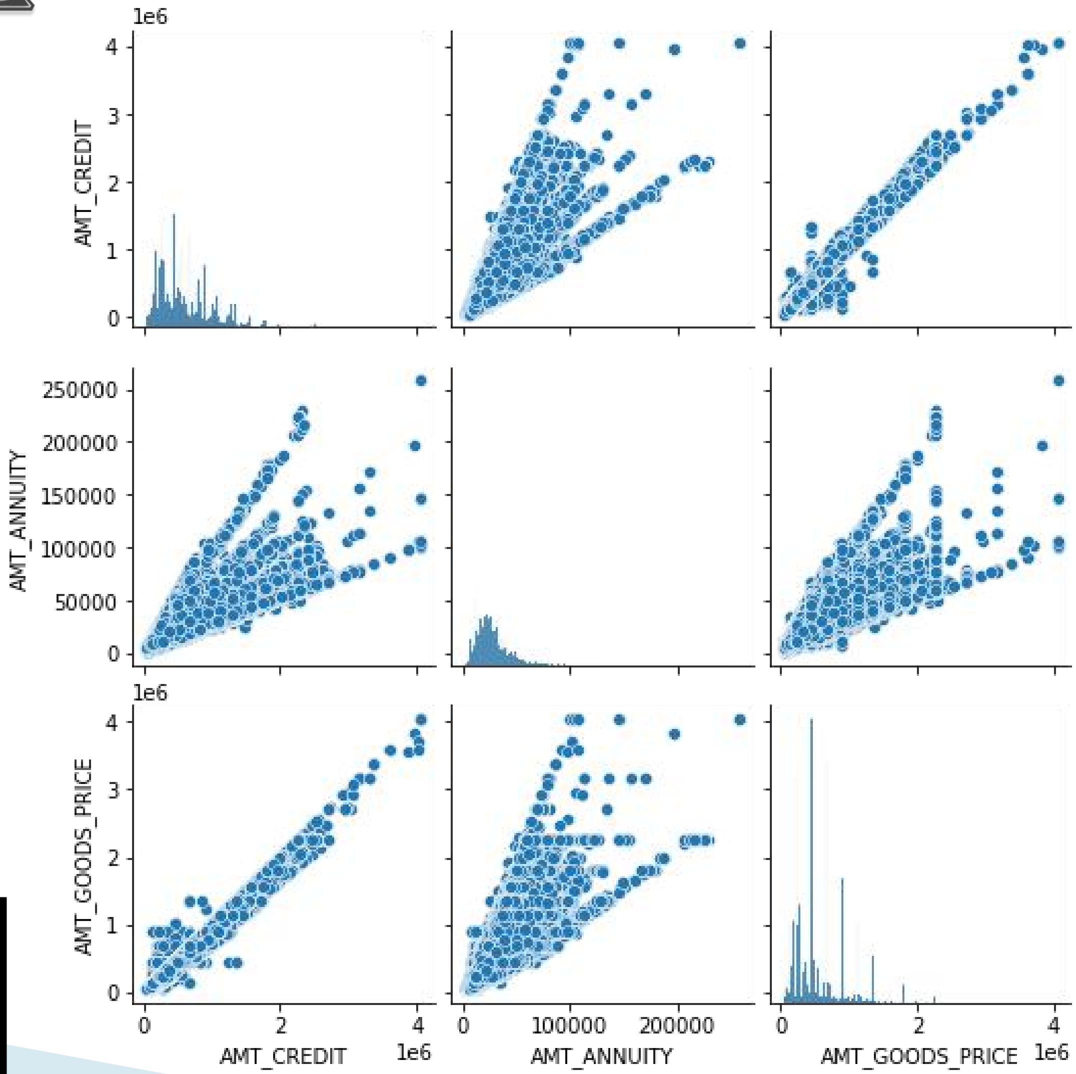


Category	Defaulter Percentage
Average	8.571666
Low	8.495691
Very Low	8.190830
High	8.056891
Very High	6.519801

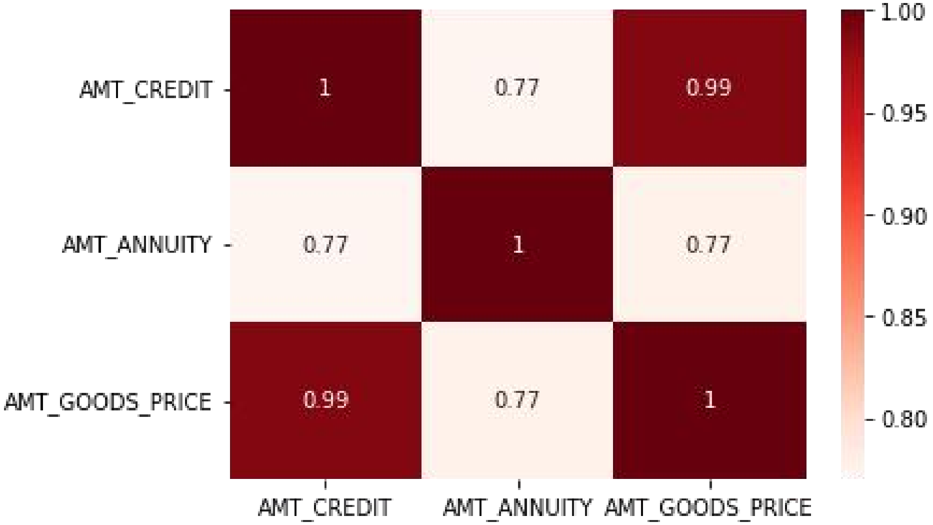
Insights:

- **Applicants in average income group have more defaulter percentage than other groups.**
- **Applicants with very high income tend to default less often compared to other groups.**

Correlation between 'AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE'



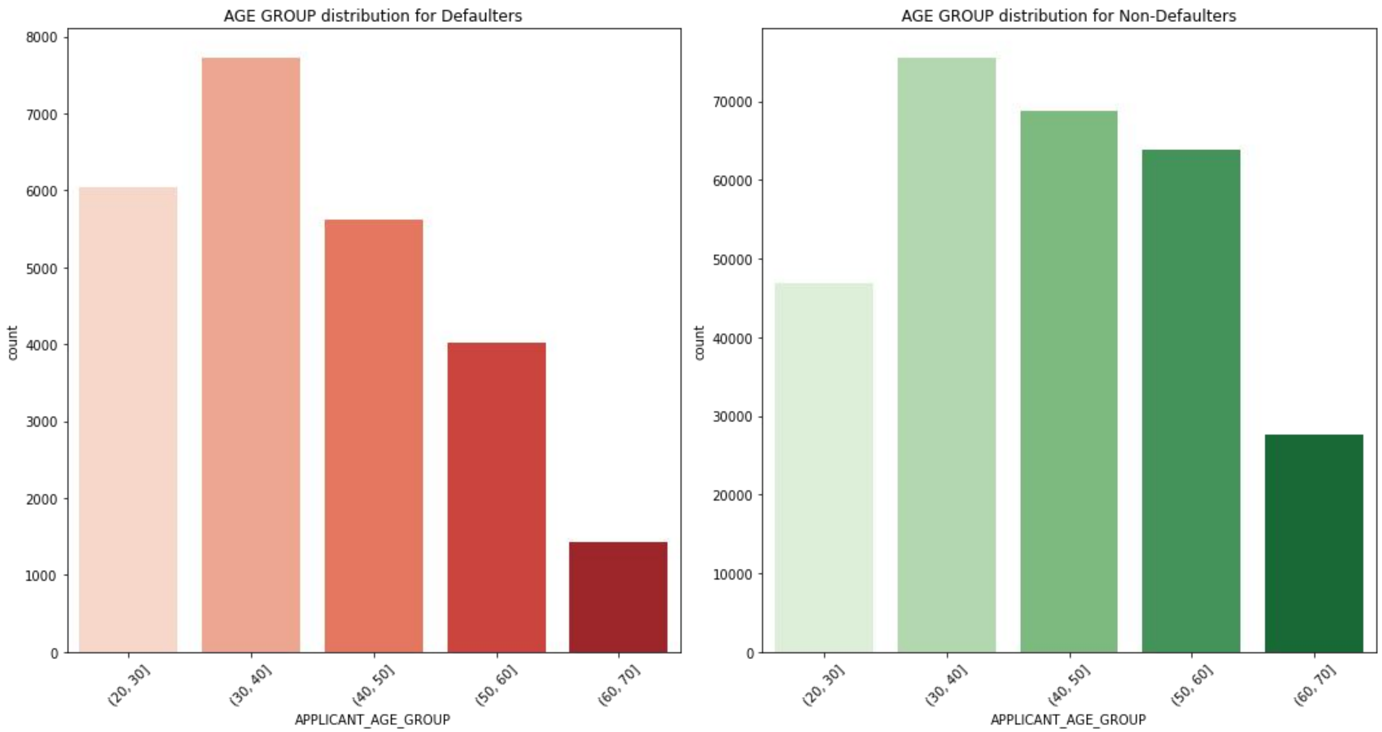
CONTINUE…



Insight from both pair plot and heat map:

- **AMT_CREDIT and AMT_GOODS_PRICE are very strongly correlated.**

AGE GROUP DISTRIBUTION



Insights:

- **Applicants of age group (30,40] are much more likely to be defaulters compared to other age groups.**
- **As the age group is increasing, applicants tend to default less often (starting from age:30). This could be because people get married and earn more, so they are able to pay their loan.**

TOP 10 CORRELATIONS FOR THE CLIENTS WITH PAYMENT DIFFICULTIES (DEFAULTERS)

- FLAG_EMP_PHONE and DAYS_EMPLOYED
- DAYS_BIRTH and APPLICANT_AGE
- AMT_GOODS_PRICE and AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT
- CNT_CHILDREN and CNT_FAM_MEMBERS
- LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION
- LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY
- AMT_GOODS_PRICE and AMT_ANNUITY
- AMT_CREDIT and AMT_ANNUITY
- FLAG_DOCUMENT_6 and FLAG_EMP_PHONE

Top 10 Correlations for Non-Defaulters

- FLAG_EMP_PHONE and DAYS_EMPLOYED
- DAYS_BIRTH and APPLICANT_AGE
- AMT_GOODS_PRICE and AMT_CREDIT
- REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT
- CNT_CHILDREN and CNT_FAM_MEMBERS
- LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION
- LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY
- AMT_GOODS_PRICE and AMT_ANNUITY
- AMT_CREDIT and AMT_ANNUITY
- FLAG_DOCUMENT_6 and FLAG_EMP_PHONE

IMPORTANT DRIVING FEATURES/COLUMNS

- Gender Biased: Female applicants are more than male applicants and still Defauter percentage is higher for male applicants.
- Percentage of Females in data = 66%, Defaulter Percentage = 7%
  Percentage of Males in data = 34%, Defaulter Percentage = 10%
- Family Info: (Important driving features : 'CNT_FAM_MEMBERS', 'CNT_CHILDREN')
- Applicants with very large number of children/family member count and defaulter percentage also high or even low,cannot be used for providing insights since the frequency of such applicants is very low.
  Applicants with 0 children have a defaulter percentage of 7.71%, so they are likely to be non-defaulters/repayers.
  Applicants with 2 family members have a defaulter percentage of 7.58%, so they are likely to be non-defaulters/repayers and should be approved for there loan.

- Education and Occupation Info: (Important driving features :'NAME_INCOME_TYPE', 'OCCUPATION_TYPE')
- Applicants on Maternity leave or if they are unemployed have the highest chances of defaulting. Applicants who are working, or Pensioner or Commerical associate are good since the defaulting percentage is not much.
  Applicants in ('Unemployed','Student','Businessman','Maternity leave') don't contribute much to the analysis as there is not much data available for these types.
  Occupation: Low-skill laborers are the applicants with highest defaulter percentage.

- A derived column 'APPLICANT_AGE' from 'DAYS_BIRTH' gave useful information.Applicants of age group (30,40] are much more likely to be defaulters compared to other age groups.
  As the age group is increasing, applicants tend to default less often (starting from age:30). This could be because people get married and earn more, so they are able to pay their loan.

CONCLUSION

- This data is highly imbalanced as number of defaulter is very less in total population.
- Percentage of non-defaulters = 91.93%
  Percentage of defaulters = 8.07%
  Data Imbalance Ratio => approx. 8:92 = 2:23

- Asset details of an applicant
- Number of applicants who own a car are much less than the applicants who don't own a car.
  Number of applicants who own a realty are much more than the number of applicants who don't own a realty.
  According to defaulter percentage, applicants who don't own a car and realty are much more likely to be defaulters than the ones who have their own cars and realty.

- 'CNT_FAM_MEMBERS', 'CNT_CHILDREN','NAME_INCOME_TYPE', 'OCCUPATION_TYPE',CODE_GENDER, are some of the important driving factors.
- Documents : Considered features 'FLAG_DOCUMENT_2','FLAG_DOCUMENT_3',...,'FLAG_DOCUMENT_21' for this segment. Majority of the applicants did not submit any documents apart from DOCUMENT_3. FLAG_DOCUMENT_3 has similar impact on defaulters and non-defaulters. Hence these columns can be dropped.None of the FLAG_DOCUMENTS are linearly correlated with TARGET Variable.
- Housing:
- All of the features which have very high (47-70%) missing data percentage can be dropped.
  Plot of 'NAME_HOUSING_TYPE' vs 'TARGET' shows that a lot of applicants live in House/Apartment.
  Applicants with Family Status as "with parents" or "Rented apartment" have much higher rate of default.

- Social Circle Info: The features show similar trend for defaulters and nondefalters, can be dropped.
- Income Group, derived from AMT_INCOME_TOTAL after BINNING
- Applicants in average income group have more defaulter percentage than other groups.
  Applicants with very high income tend to default less often compared to other groups.