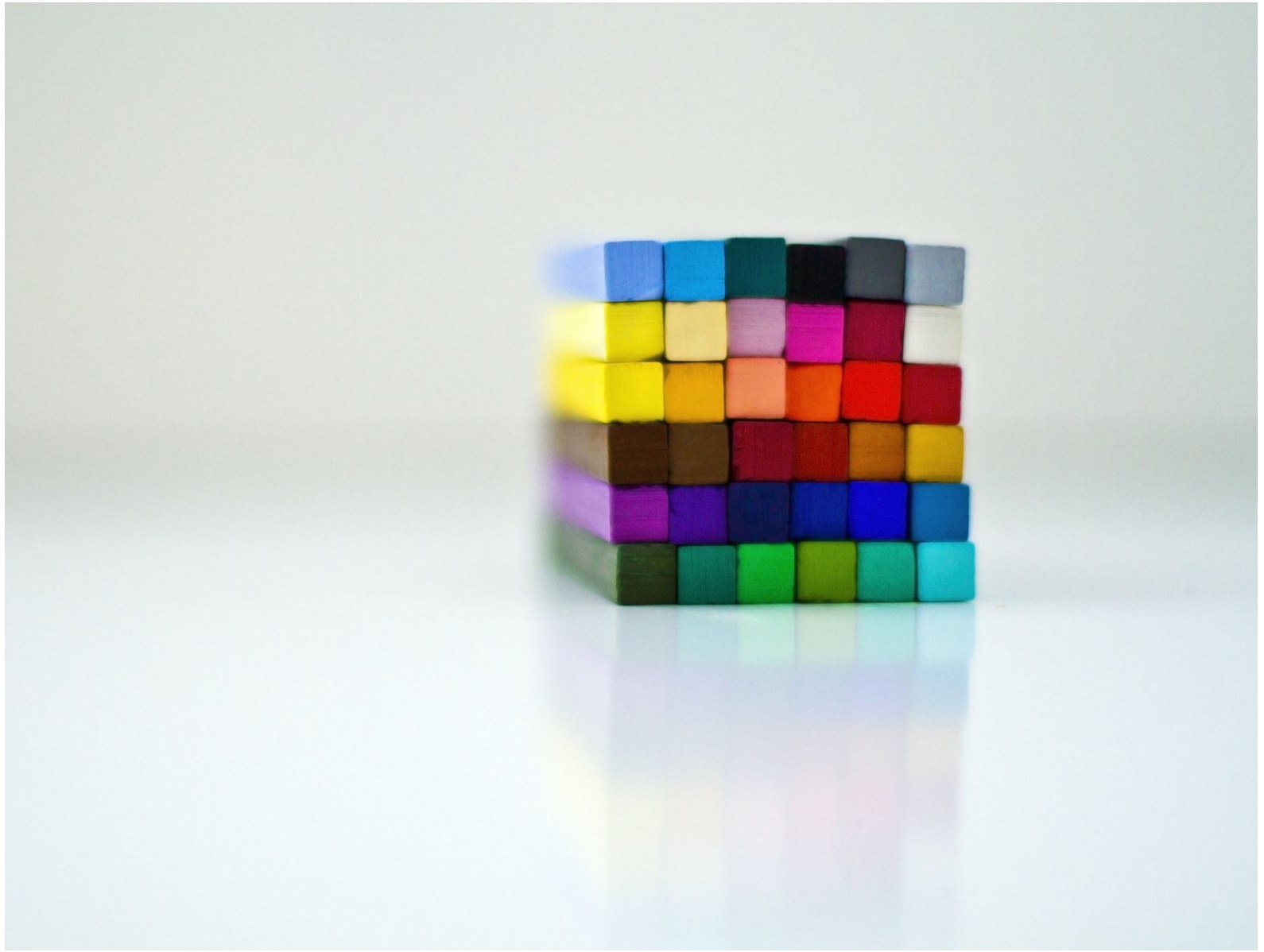


EXPLORATORY DATA ANALYSIS

SAI GEETHA M N

PROBLEM STATEMENT

Given 2 data sets of a Bank

- Current Application for Loans
- Previous Applications for loans
- And a Target variable that indicate which clients have defaulted and which have not

Need to find

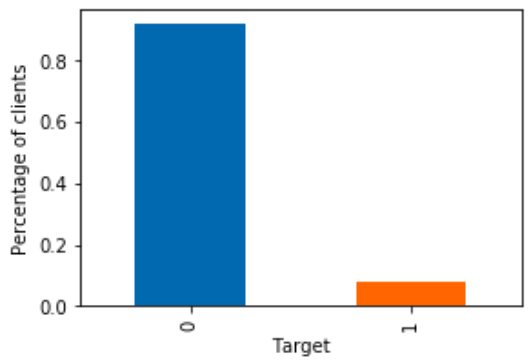
- Variables that influence Defaulters from the current application data
- Provide any insight from previous application data that shows patterns related to default/non-default

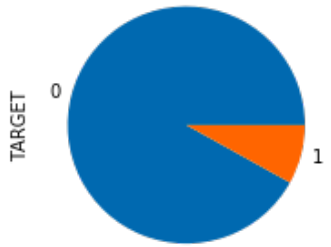
FINDINGS BASED ON CURRENT APPLICANT'S DATA

INCLUDES

- UNIVARIATE ANALYSIS
- BIVARIATE ANALYSIS
- MULTI-VARIATE ANALYSIS
- COORELATION

CHECK THE IMBALANCE IN DATA





- Target 1 – Defaulters
- Target 0 – Non-Defaulters

As expected, this is an imbalanced data with a large percentage of non-defaulters

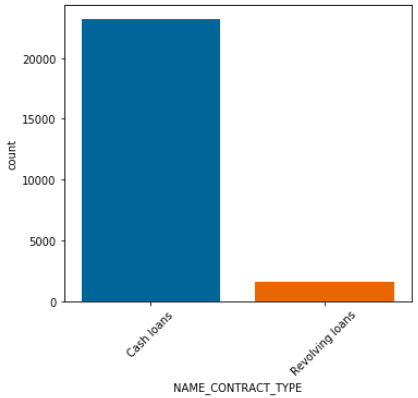
UNIVARIATE ANALYSIS – CURRENT APPLICATION

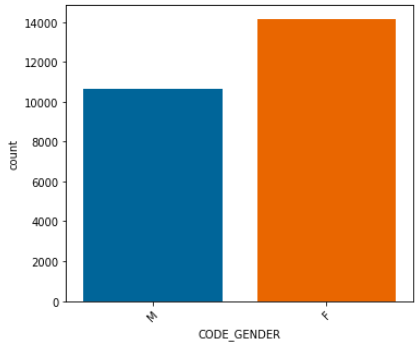
A STUDY OF THE INDIVIDUAL VARIABLES FOR BOTH DEFAULTERS AND NON-DEFAULTERS

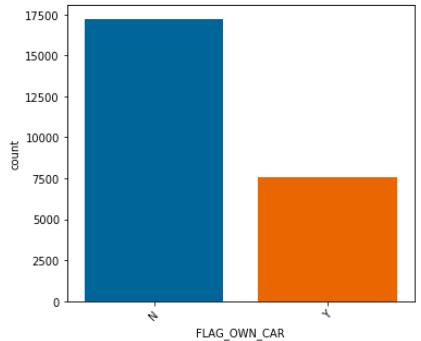
DEFAULTERS: UNIVARIATE ANALYSIS - FOR CATEGORICAL VARIABLES

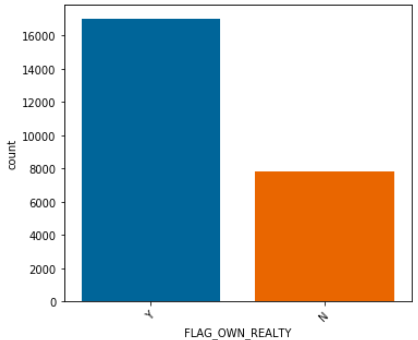
The following category of people were higher or highest in defaults:

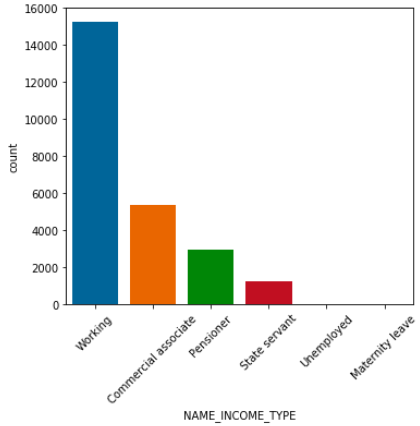
- Cash loans more than revolving loans
- Females compared to males
- People who did not have their own car
- Surprisingly, people who had their own Reality
- **Working professional were the highest defaulters out of 6 categories**
- People with only Secondary / secondary special education were the highest
- Married people - highest
- People living in their own house/Apartment
- Labourers by occupation

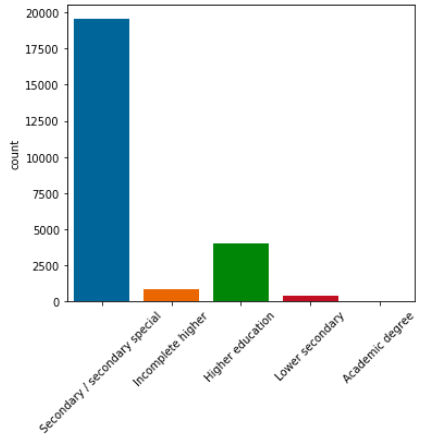


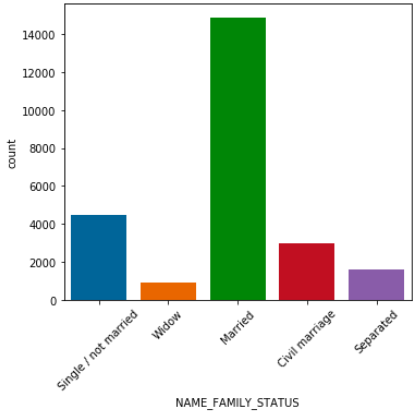


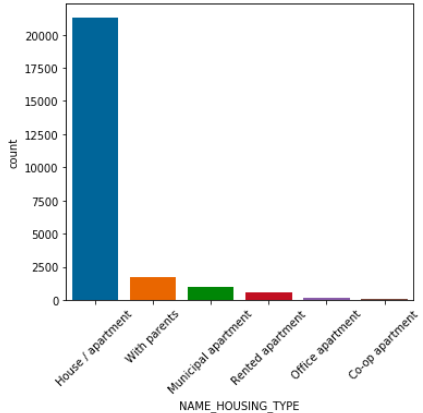


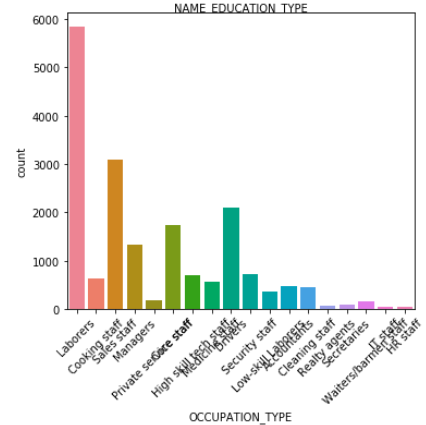








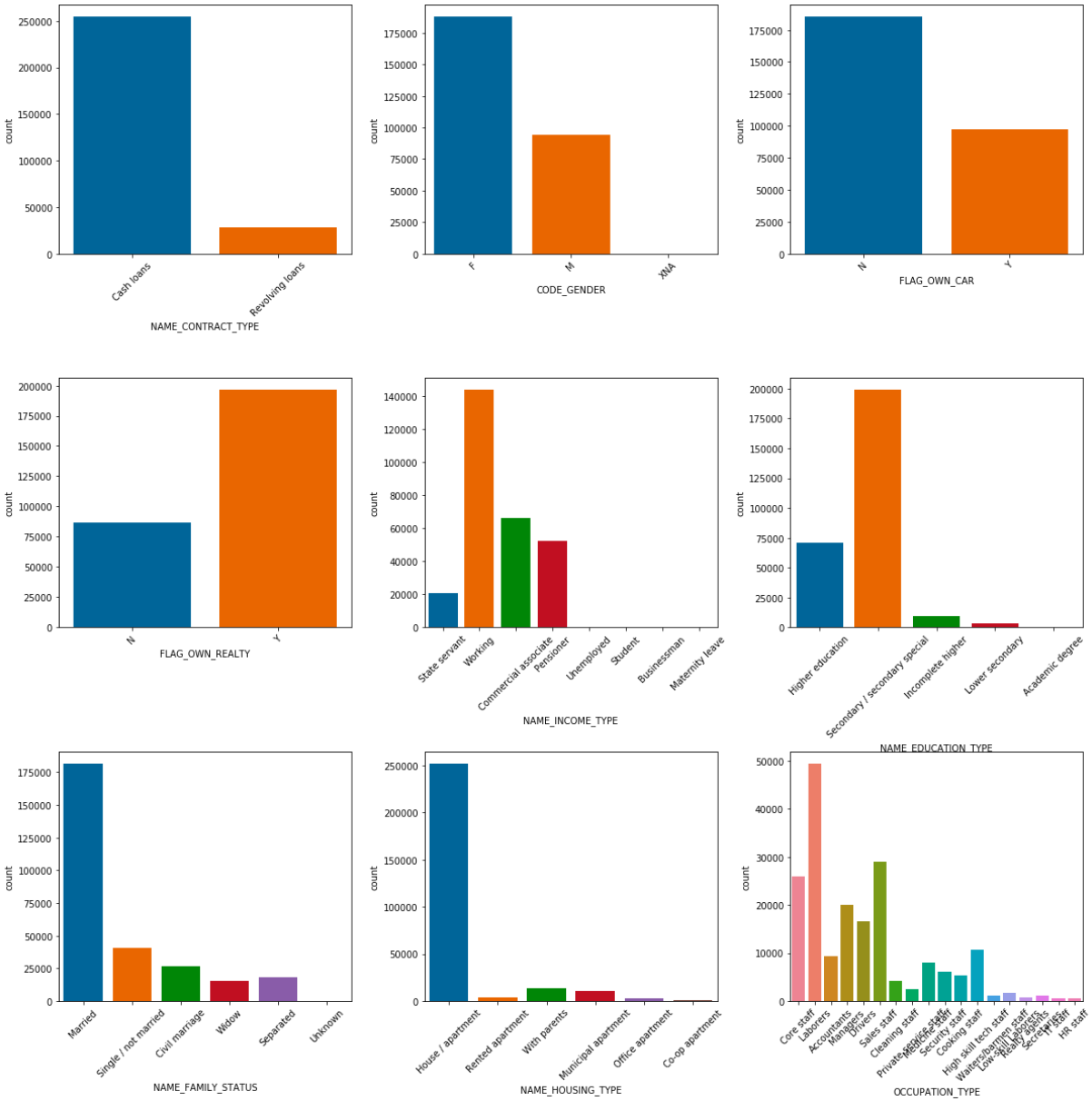




NON-DEFAULTERS: UNIVARIATE ANALYSIS - FOR CATEGORICAL VARIABLES

The following category of people were higher or highest in non-defaults:

- The Type of loans not-defaulted too were more of Cash loans than revolving loans
- Females compared to males
- People who did not have their own car
- People who had their own Realty
- **State-servants were the highest non-defaulters out of 6 categories**
- Again people with only Secondary / secondary special education were the highest non-defaulters
- Again Married people - highest non-defaulters
- Again People living in their own house/Apartment
- Again Labourers by occupation



CONCLUSIONS FOR UNIVARIATE – CATEGORICAL VARIABLES

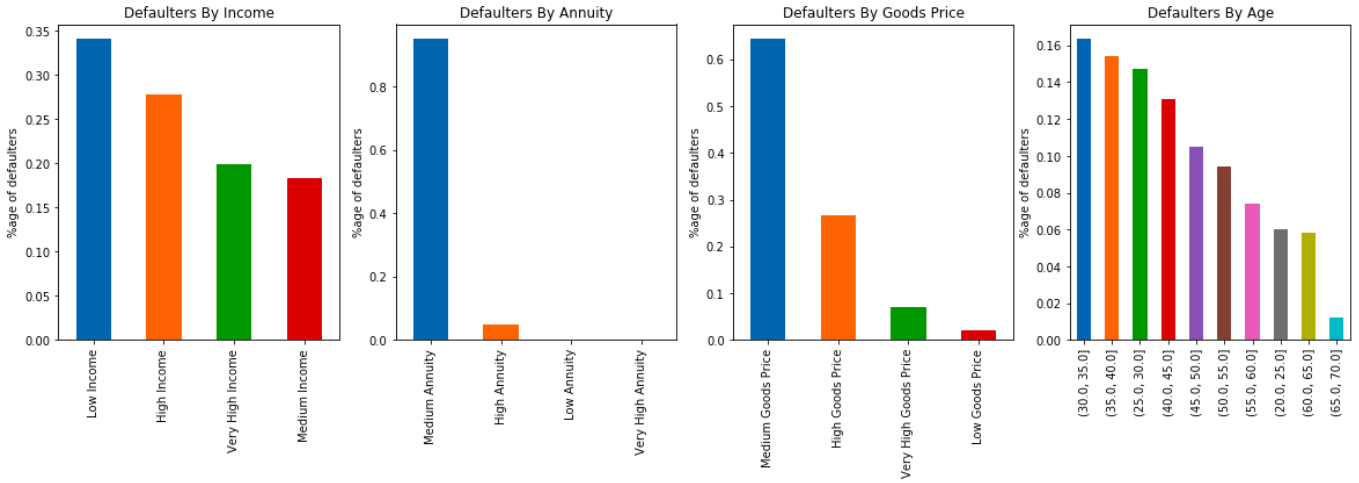
There is not too much of an insight from separately analysis defaulters and non-defaulters except that Working professionals topped defaulters and State-Servants topped non-defaulters

Rest of the categories being highest in both the categories just implies that they are the largest category of borrowers e.g. labourers, married people, people with Secondary education.

DEFAULTERS:
UNIVARIATE ANALYSIS - CONTINUOUS VARIABLES BY BINNING

Defaulters by bins:

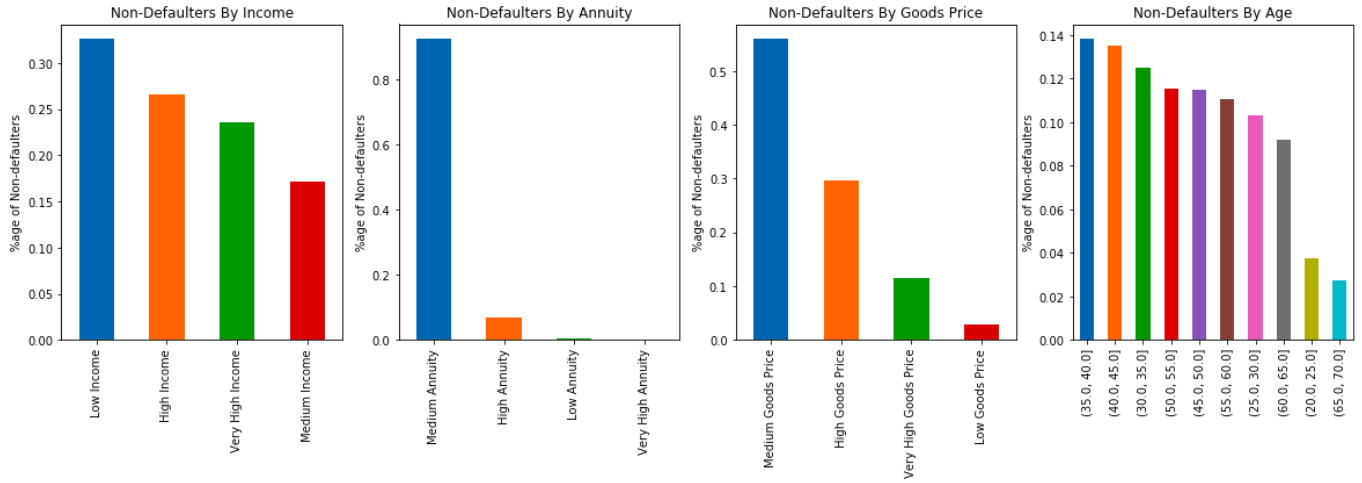
- Max Defaulters - close to 35% are from the low income bracket (0 to 25650)
- More than 85% of defaulters has a Medium Annuity (5000 to 50000)
- Close to 65% of the price of goods for which default are in the Medium range (5000 to 50000)
- The lower age groups from 25 to 40 years are the highest defaulters.



NON-DEFAULTERS: UNIVARIATE ANALYSIS - CONTINUOUS VARIABLES BY BINNING

Non-Defaulters by bins:

- Nearly 35% of low income group do not default
- Medium annuity is highly non-defaulted - Close to 85% of non-defaulters has a Medium Annuity (5000 to 50000)
- Close to 55% of the price of goods for which non-default happened were in the Medium range (5000 to 50000)
- 35 to 40 age group has the large non-defaulters



ORDERED CATEGORICAL VARIABLES - CONCLUSION

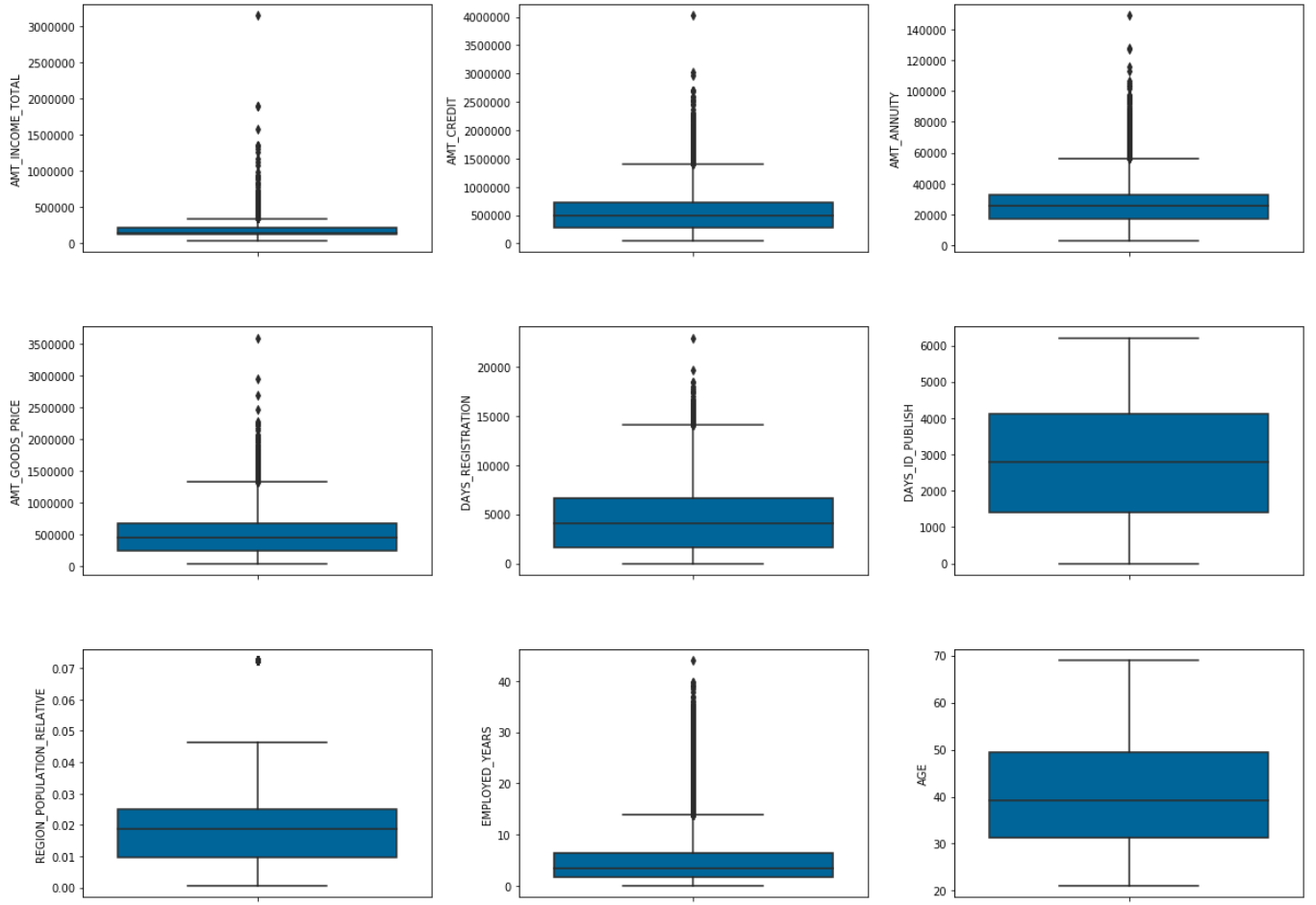
The only significant insight we get from ordered categorical variables is by binning ages. It is very clear that the lower age groups from 20 to 30 default a lot more and are the riskiest age groups

30 to 45 age group has maximum percentage of non-defaulters and hence a safe group to lend to

The other bins show that the maximum demand for loans is more in the medium range and also more from lower income groups

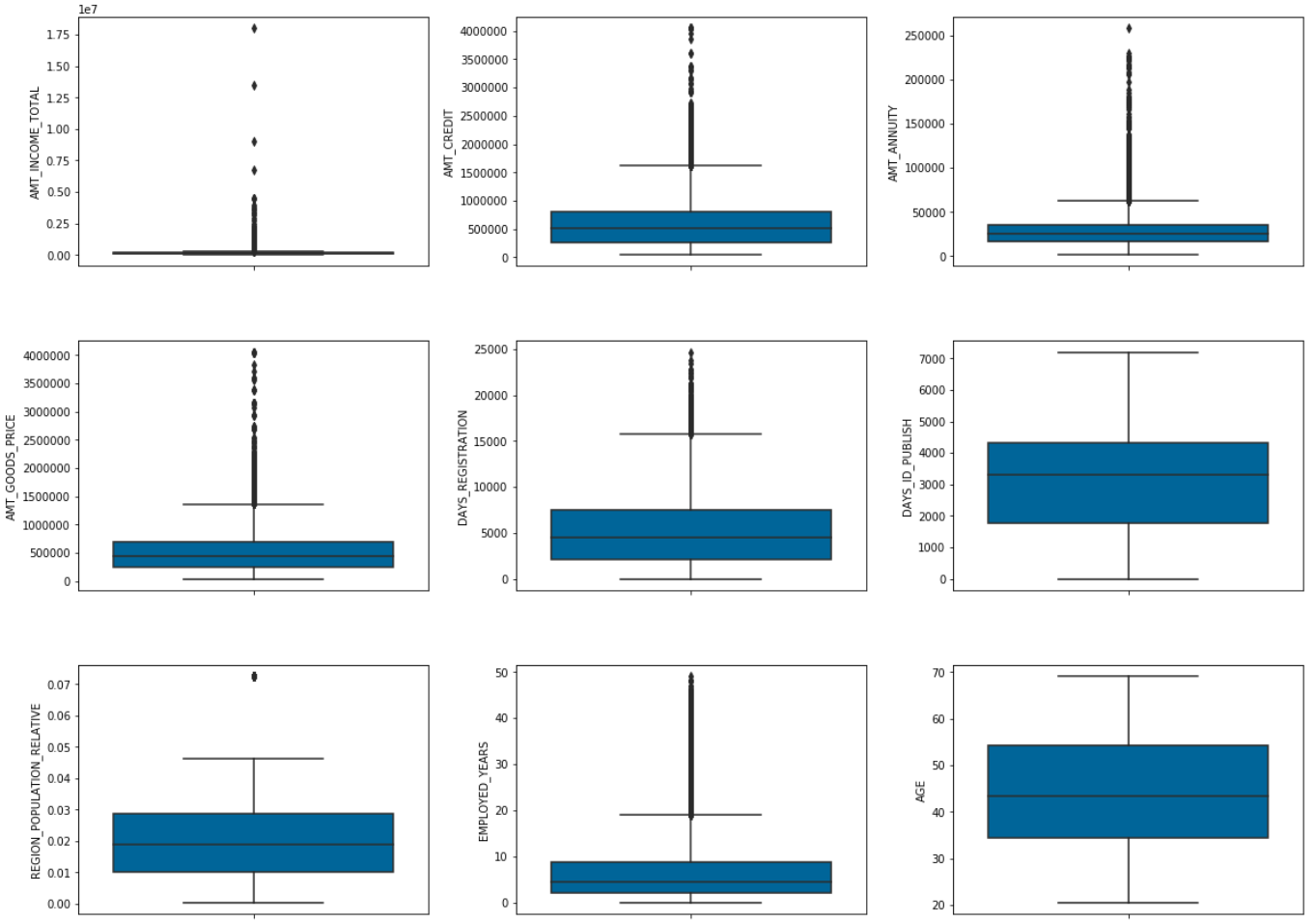
DEFAULTERS: UNIVARIATE ANALYSIS -– CONTINUOUS VARIABLES

- The spread of income is very narrow and is largely below 5 lakh
- Credit, annuity and goods price go in tandem - with the credit being below 15 lakh and annuity being below 60000
- Days since registration has a median of less than 5000 days but has a max of 15000 days!
- Days since ID change seems to be evenly spread with a median of 3000 and max of 6000 days
- The region's relative population is having a median of 0.02
- Employed years has a long tail but the median less than 5 years
- The median age of defaulters is around 40 years



NON-DEFAULTERS: UNIVARIATE ANALYSIS -– CONTINUOUS VARIABLES

- Most of the income is very low, hovering around 5 lakh
- Credit and goods price have a median of about 5 lakh
- Median annuity is about 30000
- Days since registration and days since ID change almost show the same behaviour as defaulters
- The median of employed years seems slightly higher
- The median age of non-defaulters is more hovering around 42 years



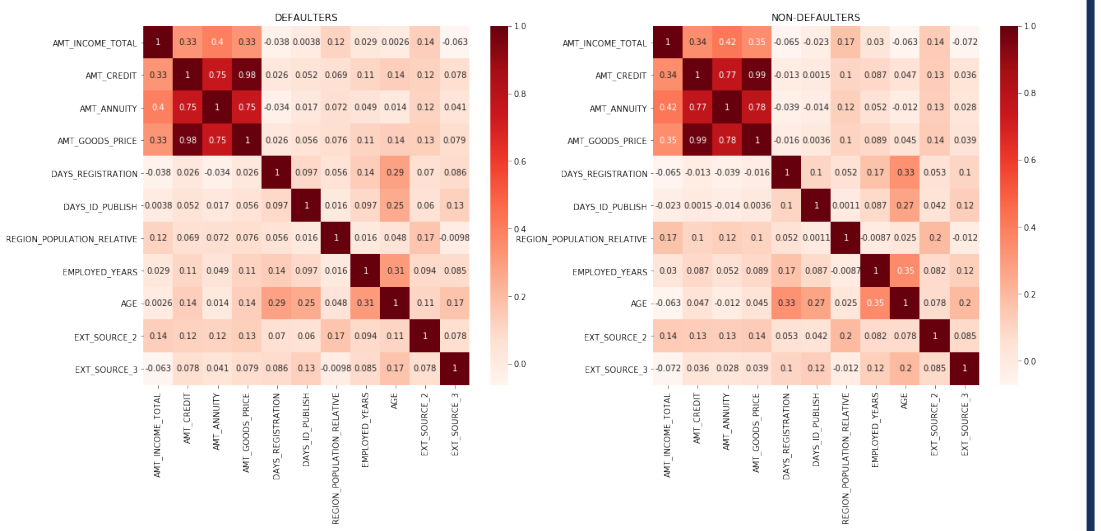
CONCLUSIONS FROM CONTINUOUS VARIABLES – UNIVARIATE ANALYSIS

- Median Age group for defaulters is smaller than the median age group of non-defaulters
- Median years of employment is also smaller for defaulters

BIVARIATE ANALYSIS – CURRENT APPLICATION

A STUDY OF CORRELATION BETWEEN 2 VARIABLES FOR BOTH DEFAULTERS AND NON-DEFAULTERS

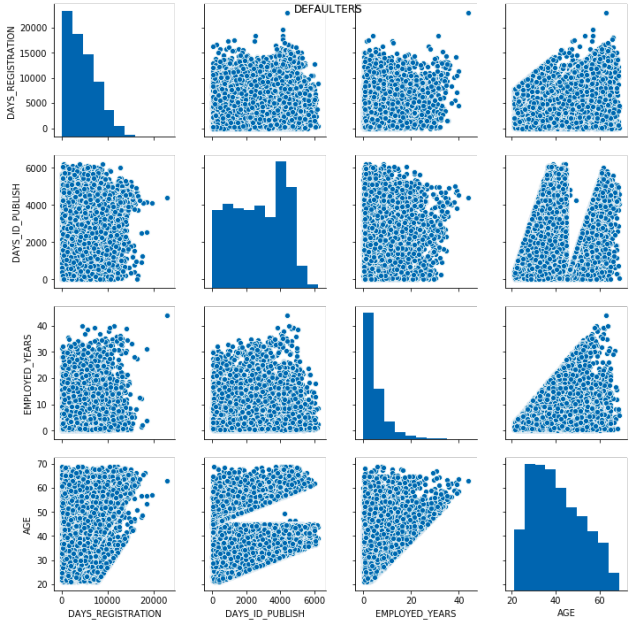
FINDING THE CLOSELY RELATED VARIABLES

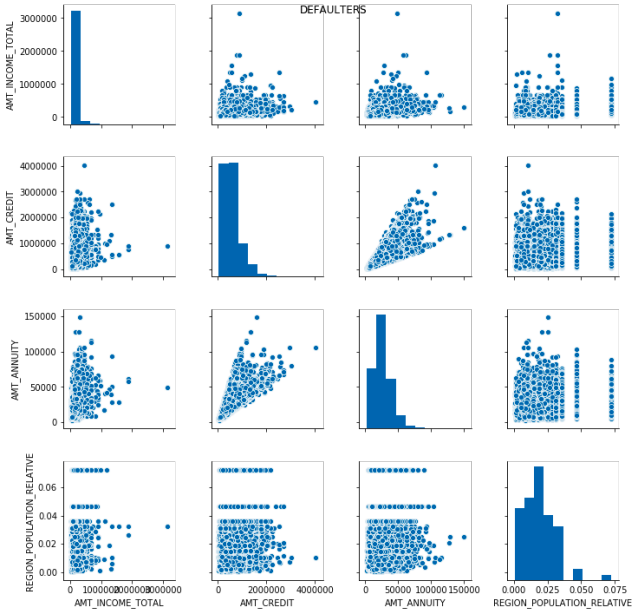


- **Observations are not too different for defaulters or non-defaulters**
- Income, credit, annuity, goods price are highly corelated. The Region population relative is slightly correlated to income. Hence these 4 variables can be plotted together in a pair plot to view visually.
- Also, even though it is not a very strong correlation, there is a co-relation between age and employed years, days of id changed before loan application and days since registration. These 4 variables can be plotted together in a pair plot to see their relation ships

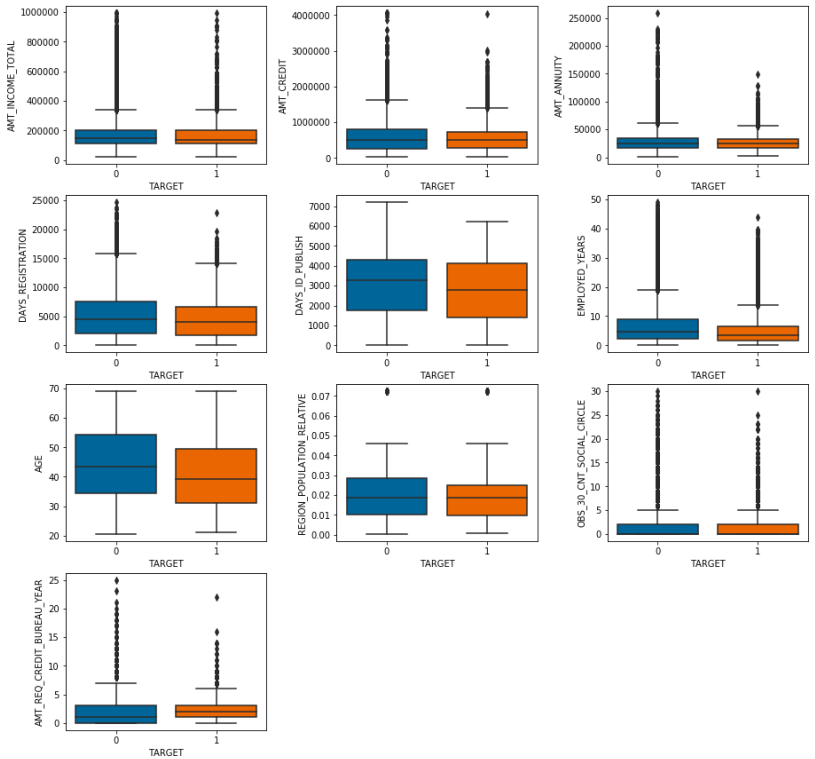
DEFAULTERS: NUMERICAL-NUMERICAL ANALYSIS

- There is trend visible that the higher the income, the higher the credit - though not very strong co-relation
- Obviously, the higher the credit, the higher the annuity
- With age, number of employed years increase
- The clusters in ID changes seem distinct. At 20, people seem to have changed or created their new IDs for the 1st time. Until 45, they do not seem to change that much. Again at 45, they change the IDs and then keep the same till 70.
- Even the number days since registration are fewer at 10 and increase with age





CATEGORICAL-NUMERICAL: CURRENT APPLICATION

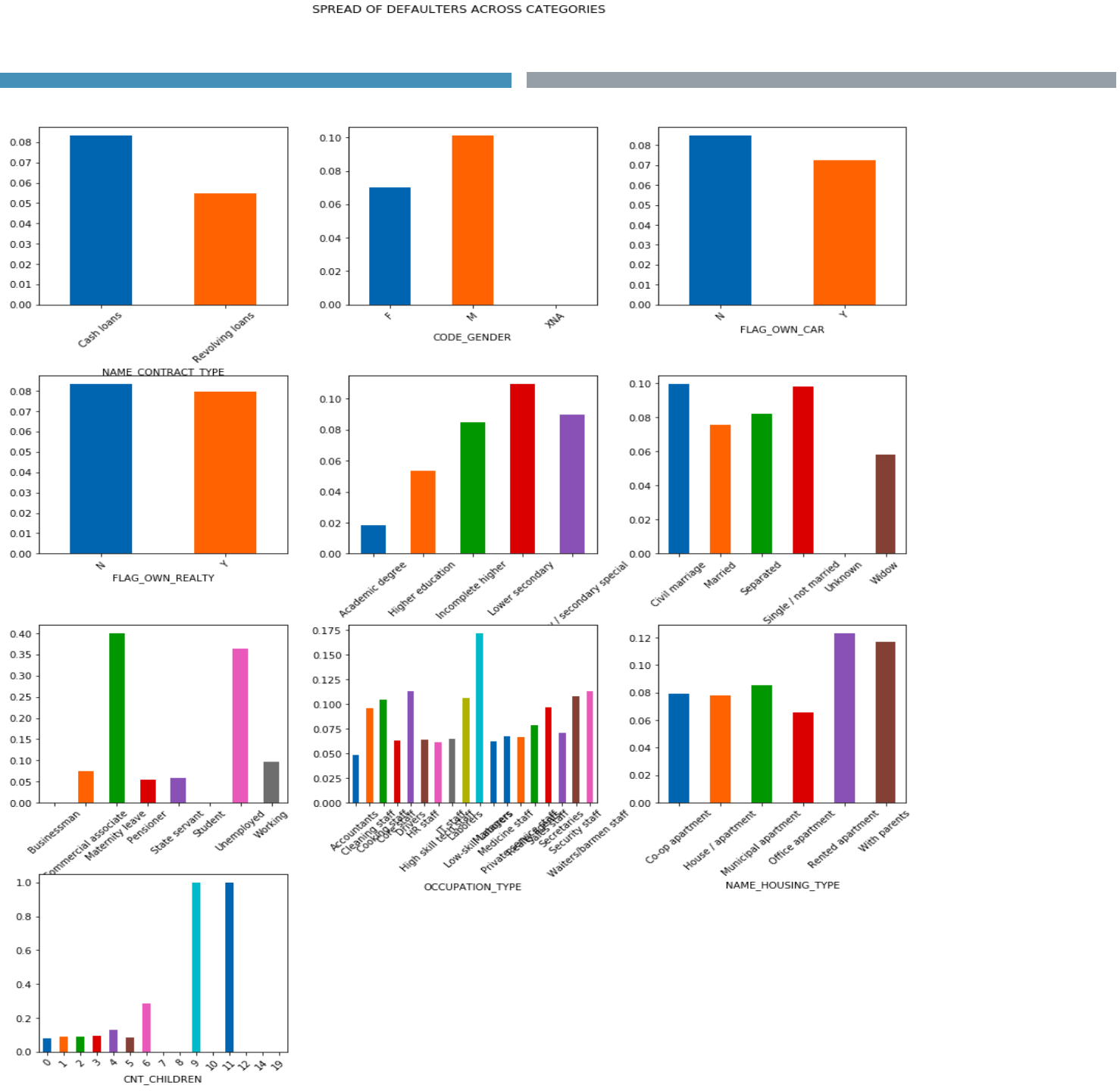


- The median income of defaulters is lower - 135000.0 versus 148500.0
- There is not too much of a difference in the annuity and credit between the 2 categories
- The spread and the median of 'days since registration' is higher for non-defaulters (Median: 4544.0 versus 4056.0) and (Spread: 5505 vs 4975)
- 'Days since ID change' is also lower median for defaulters - 2797.0 versus 3295.0
- Median number of employed years is also lower for defaulters 3.36 versus 4.63 years
- Median Age is also lower for defaulters - 39.12 versus 43.49 years
- The 75th percentile for the relative region population is lower for defaulters and so is the spread - 0.025164 vs 0.028663
- The median number of enquiries with credit bureau in the last year is higher for defaulters - 2 versus 1

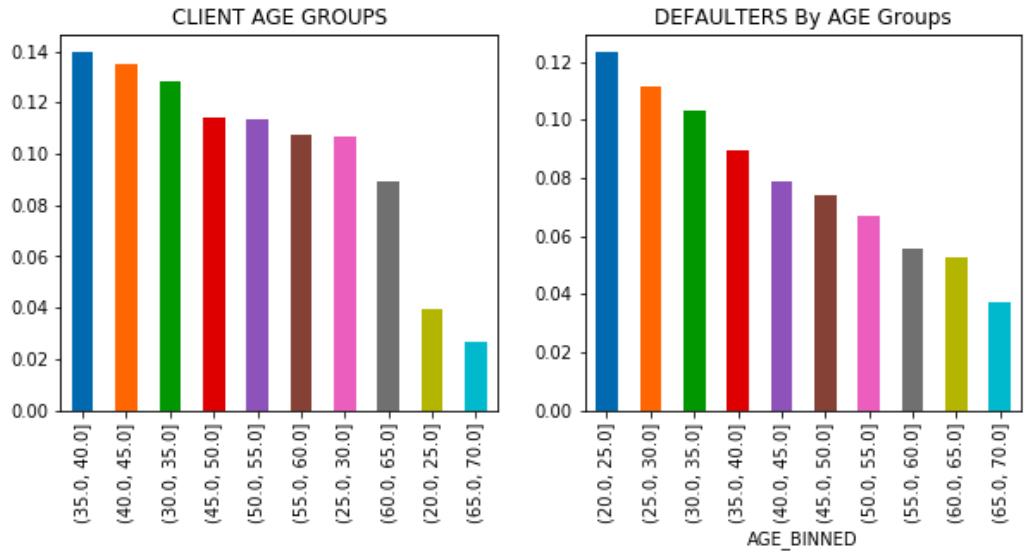
CATEGORICAL-CATEGORICAL ANALYSIS

The highest of defaults belong to the following category and hence need to be careful about lending to them:

- Lower Secondary Educated
- Commercial Associates
- Labourers
- People living in rented apartments or with parents
- People having > 5 children



SPECIFIC MENTION OF AGE GROUP CO-RELATION



- You can see that even tough t he age group 25 to 30 is a small number of clients, the default rate is very high. Something to really note

HIGHLY CORELATED VARIABLES TO LOAN-DEFAULTS

Top Positively Correlated

- REGION_RATING_CLIENT_W_CITY
- DAYS_LAST_PHONE_CHANGE
- REG_CITY_NOT_WORK_CITY
- DEF_30_CNT_SOCIAL_CIRCLE
- LIVE_CITY_NOT_WORK_CITY

Top Negatively Correlated

- EXT_SOURCE_3
- EXT_SOURCE_2
- AGE
- EMPLOYED_YEARS
- DAYS_ID_PUBLISH
- DAYS_REGISTRATION
- AMT_GOODS_PRICE

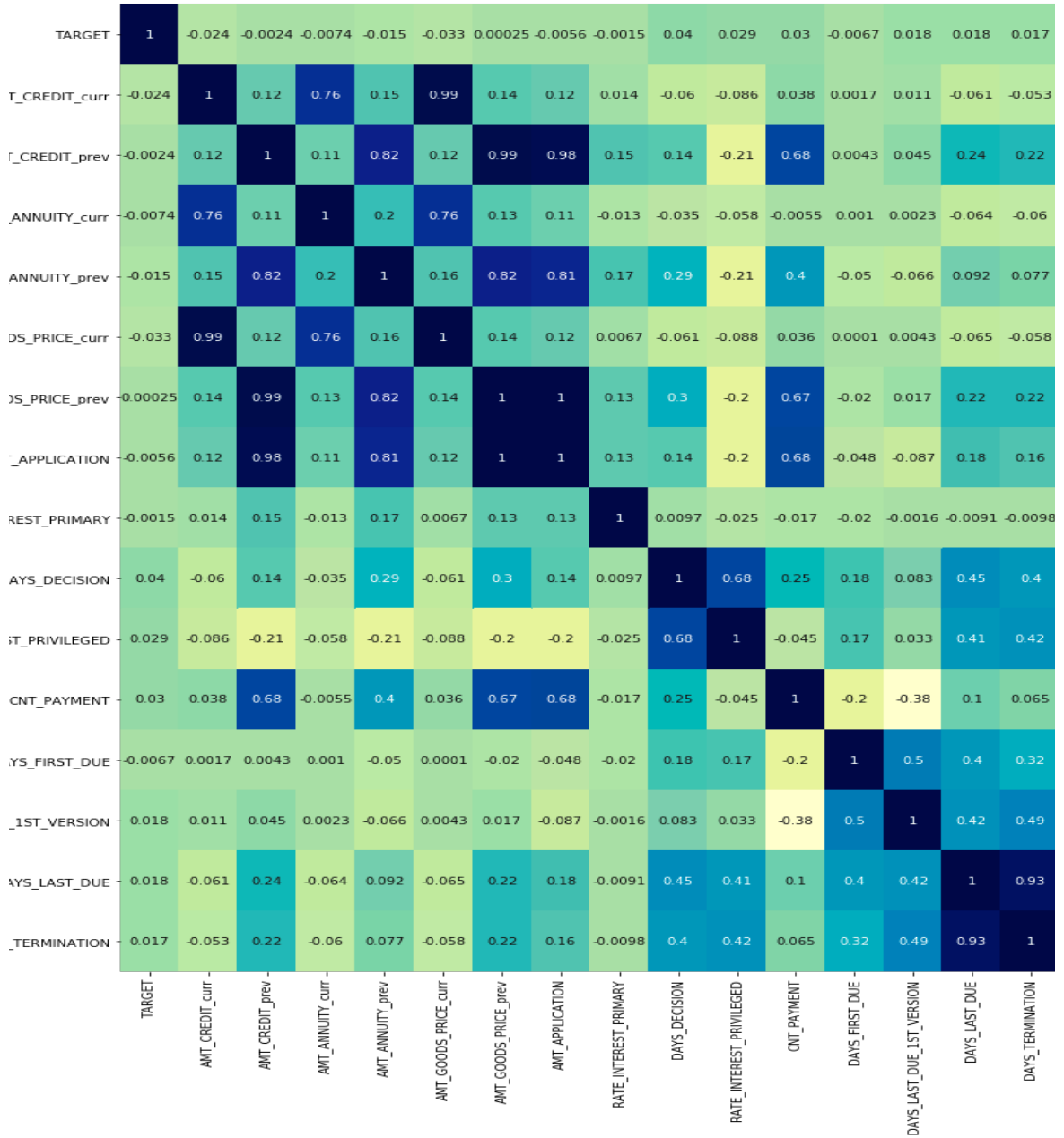
PREVIOUS APPLICATION ANALYSIS

PREVIOUS APPLICATIONS CO-RELATION MATRIX

ONLY OBVIOUS CO-RELATIONS

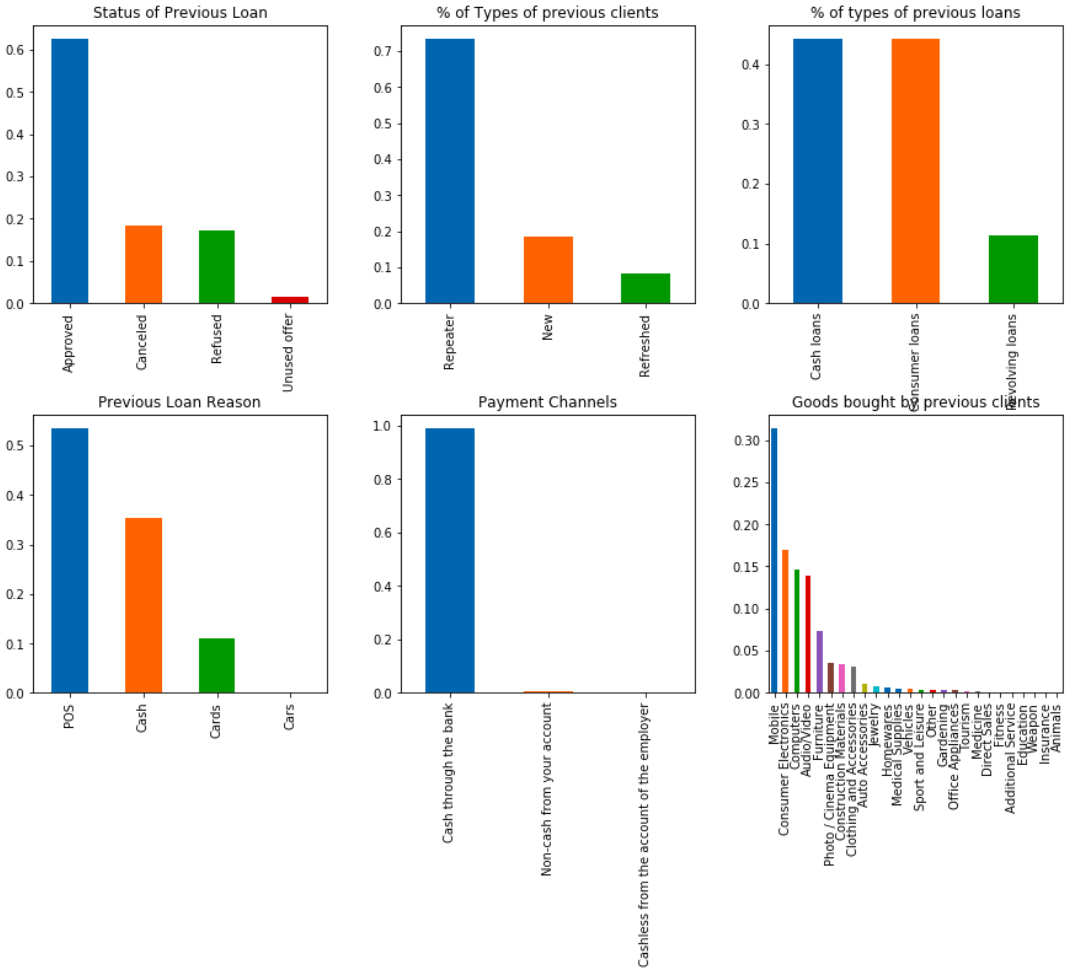
ARE SHOWN HERE. NO NEW

INSIGHTS



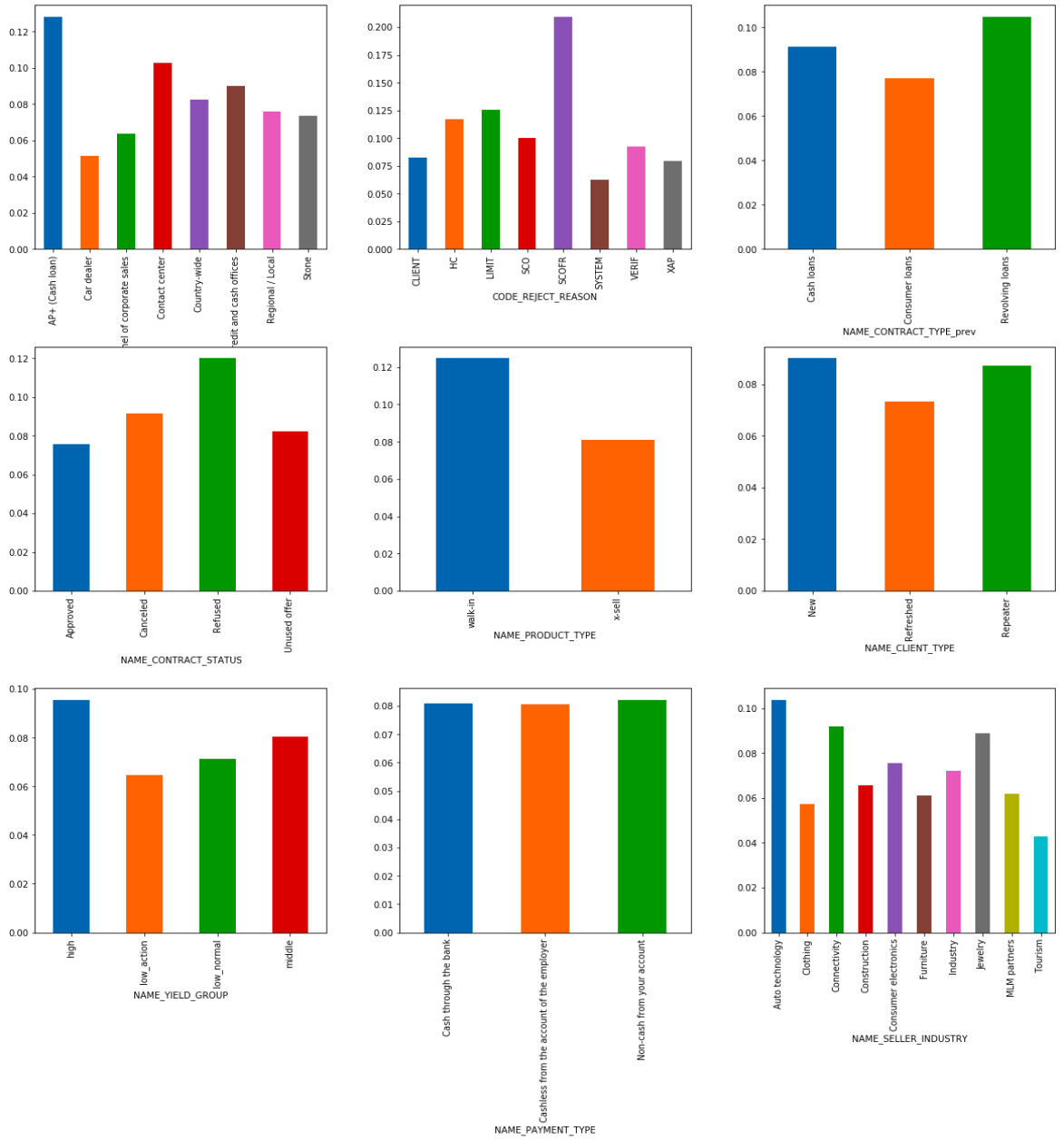
PREVIOUS APP - UNIVARIATE

- 1. More than 60% of current loan applications have been approved earlier, with 20% each being either refused or cancelled
- 2. More than 70% are repeating clients, close to 20% are new and about 10% are refreshed
- 3. Previous loans were mostly cash loans (45%), consumer loans(45%) and only 10% were revolving loans
- 4. Most of the previous types of loans were of the type POCC
- 5. Most of the previous loans had a payment channel as "Cash through the bank"
- 6. The largest goods bought by previous loans were mobiles - 30%



PREVIOUS APPLICATIONS – CAT-CAT

- Clients from AP+ (Cash loan) channel are high defaulters
- SCOFR reject reason in previous application - are highest defaulters
- Apart from no-data available, Revolving loans previous seem to be slightly high on defaults
- Previously "Refused" applicants are highest on default
- All types of payment have an equal amount of defaulters
- No of new applicants seem to default more than repeaters or refreshed clients
- People with highest yield are the highest defaulters
- Walk-in customers default more
- Loans sold by Auto-technology are defaulted more



FINAL RECOMMENDATIONS

Safer to give loans to

- The 30-45 age group
- Recommended by external sources 2 and 3
- Higher income group
- More loyal employees (longer duration at current employment)
- Clients with the number of enquiries with credit bureau in the past year around 1
- loans being given through car dealers
- Clients with previously unused offers as well as previously approved loans
- Clients whose previous loan was a cross-sell

Riskier to give loans to

- Low income category (<25000)
- Age group < 30 years
- More recent ID change
- More recent registration
- Client with registered city being different from work city. Or work city being different from living city
- Loans given through the AP + Cash loan channel. Reduce exposure here
- Previously refused clients
- Walk-in loans are also a bit risky. Need to assess more stringently
- The high-yield group
- auto-technology or connectivity