# Group Case Study 1
# on
# CREDIT EDA


**By :**

1. **Malathi Ashok (Reg. Email ID : malathiashok99@gmail.com)**
2. **Yogalakshmi Pasupathy (Reg. Email ID : laku1511@gmail.com)**

**From :**

**DS C24 Sept 2020 Batch of upGrad**

# Table of Contents

# Problem Statement

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - 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.

- 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

- 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.

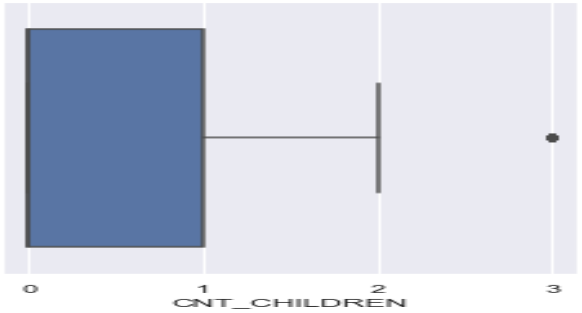
# Data Loading, Inspection & Cleaning

- For this Credit EDA Analysis, two excel sheets are provided
  1. Application data
  2. Previous Application data

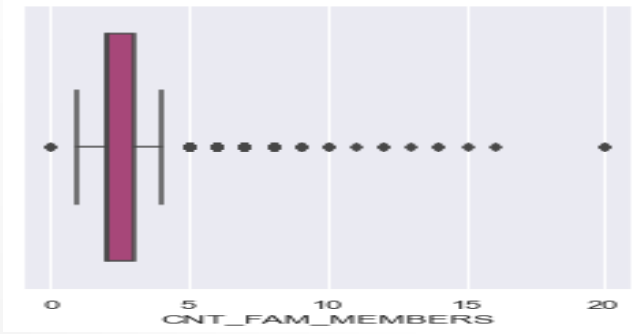
- To begin with let us start with Application data
  1. There are 307511 rows and 122 columns
  2. There are 3 data types (Float with 65 columns, Integer with 41 columns and Object with 16 columns)
  3. On performing Null value analysis, more number of columns are identified
     a) Decision is taken to drop the columns from the Dataframe as there is no connection or clear perception of the column or very limited information
        1. Columns having more than 48% of Null values
        2. Columns having more than 13% of Null values (Except Occupation Type)
        3. Others columns which are not needed for Analysis
     b) Impute the other Null value columns having less percentage with Mean/Median/Mode/Zeroes based on the datatype
  4. Convert the Days columns to Years
  5. Segment the Amount and Year variables for better understanding

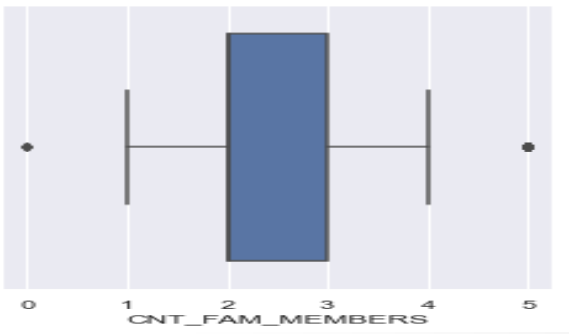
# Outlier Analysis

- Outlier Analysis has been done on the below columns
  1. Number of Children – Found that the some values were lying beyond Upper Fence/Whisker limit. Hence updated with the rounded of Upper Fence/Whisker limit value

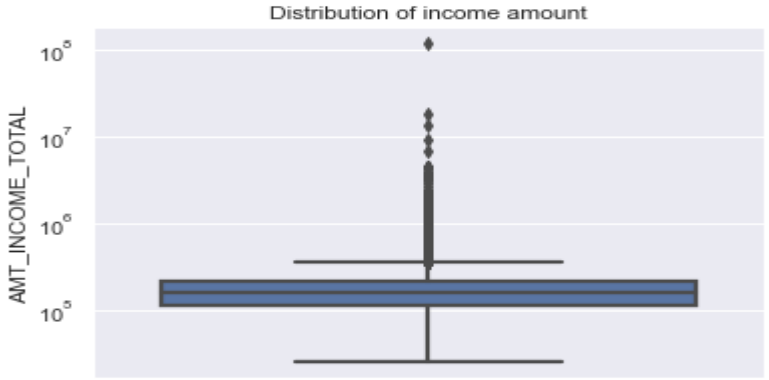

2. Number of Family members – Found that the some values were lying beyond Upper Fence/Whisker limit. Hence updated with the rounded of Upper Fence/Whisker limit value



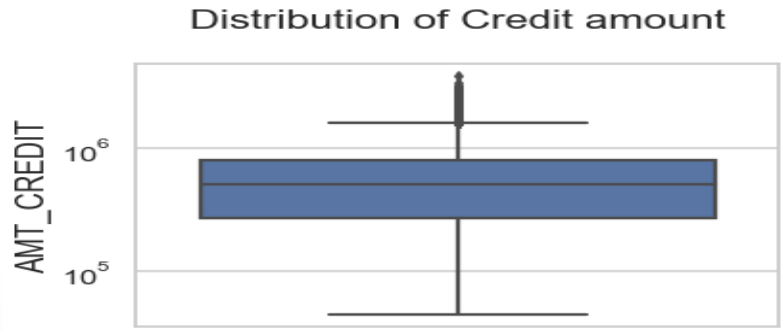



# Outlier Analysis – contd.

3. Income – Found that the some values were lying beyond Upper Fence/Whisker limit. Hence deleted the record having value greater than Upper Fence/Whisker limit value. Also third 3rd Quartile is slimmer as compared to the first quartile

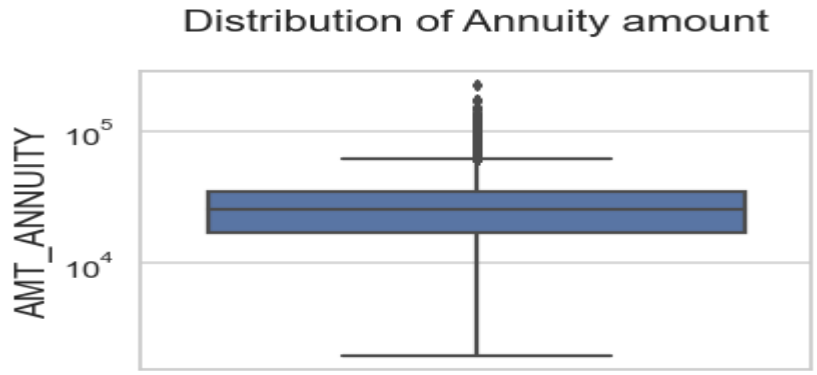



4. Credit– Found that the some values were lying beyond Upper Fence/Whisker limit. Third 3rd Quartile is slimmer as compared to the first quartile



# Outlier Analysis – contd.

5. Annuity – Found that the some values were lying beyond Upper Fence/Whisker limit. Third 3rd Quartile is slimmer as compared to the first quartile

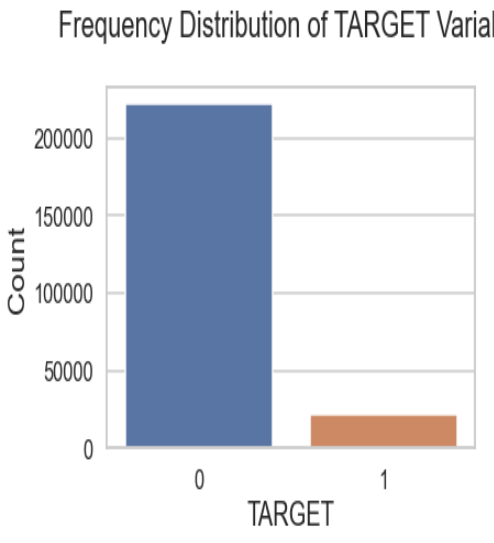


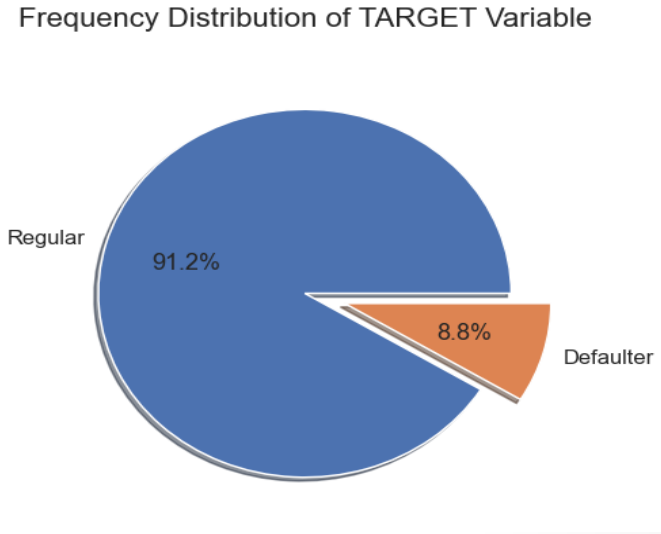

6. Goods Piece – Found that the some values were lying beyond Upper Fence/Whisker limit. Third 3rd Quartile is slimmer as compared to the first quartile

# Outlier Analysis – contd.

7. Years Employed – Found that the some values were lying beyond Upper Fence/Whisker limit. Hence updated with the rounded of Upper Fence/Whisker limit value


# Distribution of the Target variable

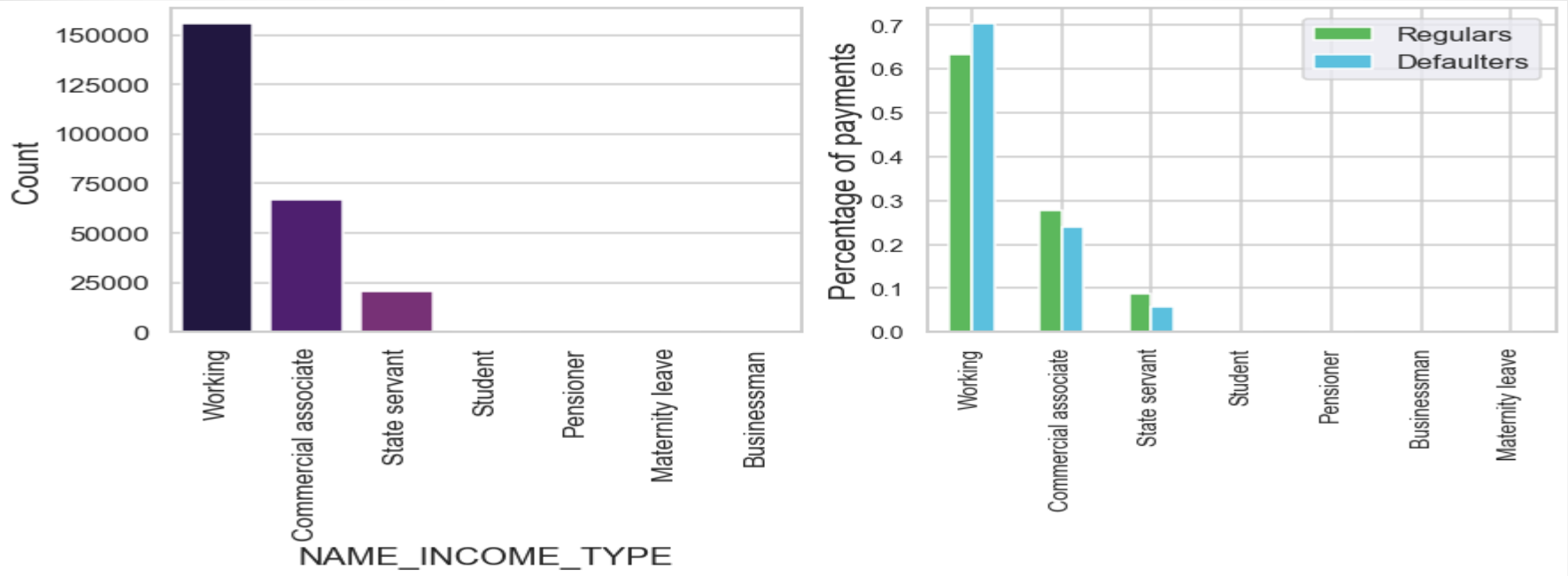



- 91.2% of Clients are Regular Payers
- 8.8% of Clients are Defaulters
- There is huge imbalance between the data which is approx. 10.4%

# Uni-variate Analysis

- **Income Type**



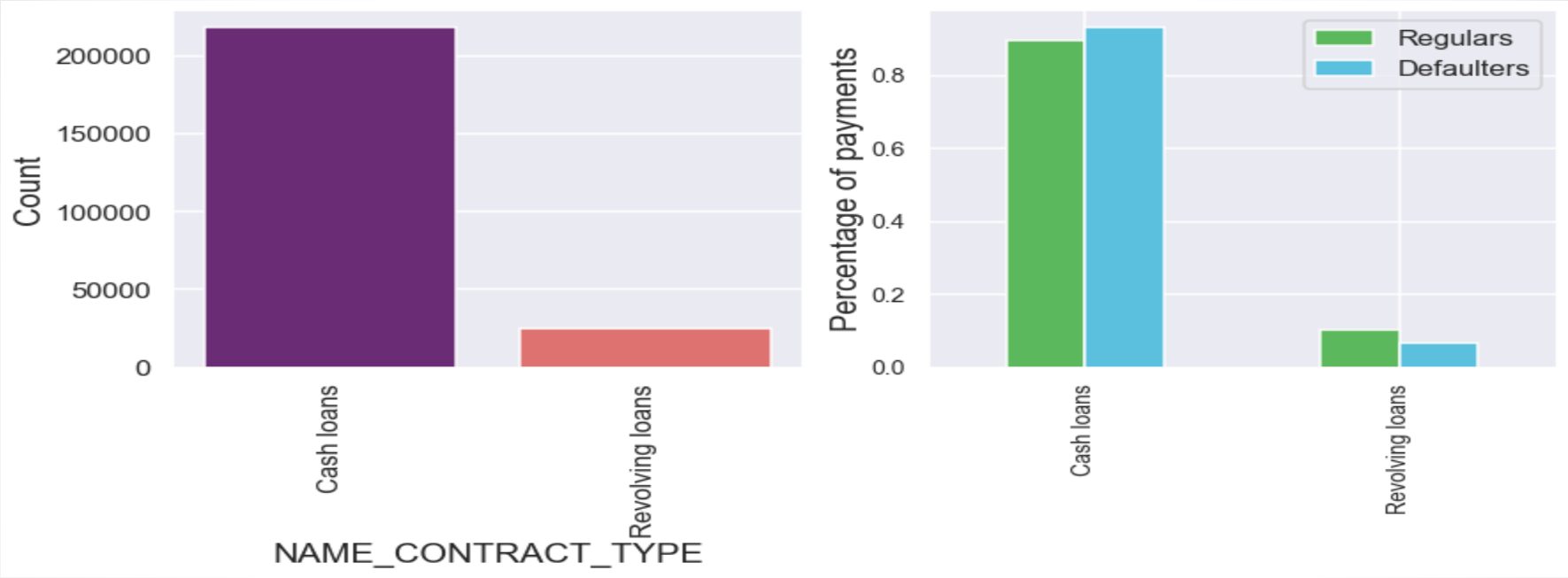

1) Clients applying for loans are from Working, Commercial Associate and State Servant Categories

2) Defaulters are also mainly from working, Commercial Associate and state servant categories

3) There are more defaulters in the working categories than in the other categories Have to be careful about approving loan for the working Category

# Uni-variate Analysis – contd.

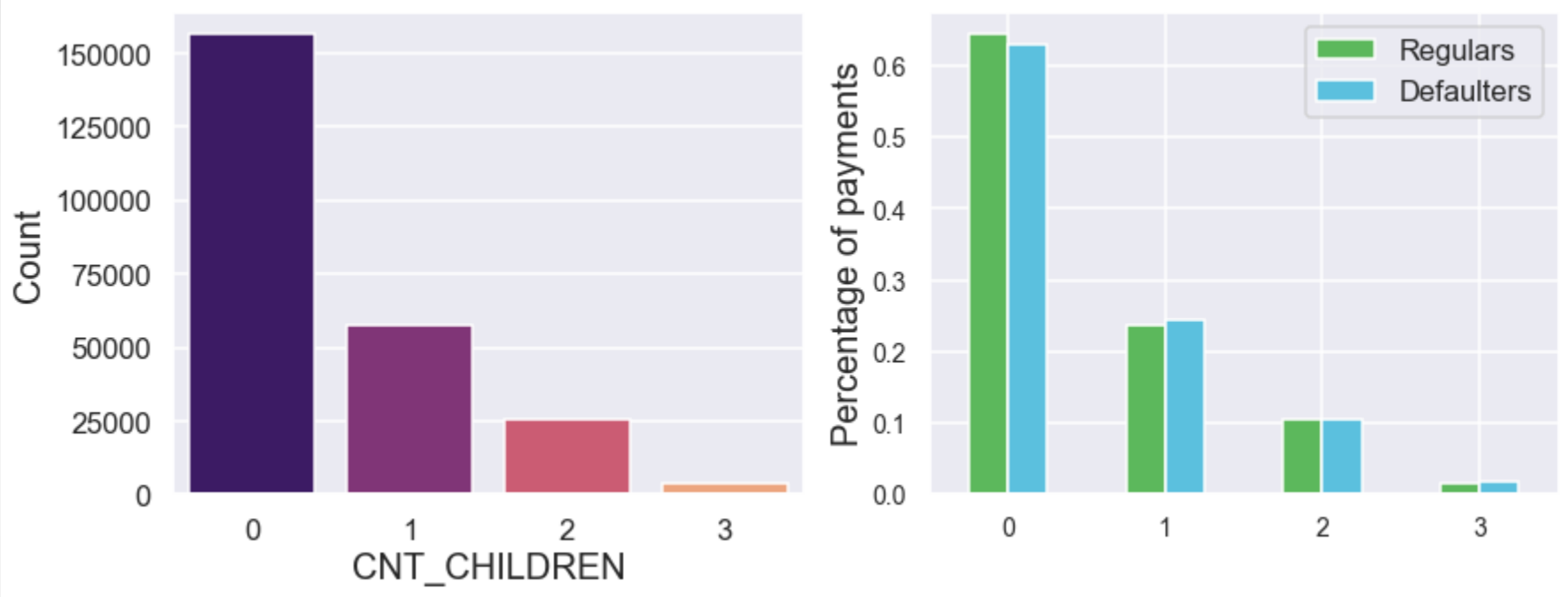
- **Contract Type**




1) The Revolving loans are small compared to Cash loans
2) No of defaulters in cash loans are higher than in revolving loans
3) Need to be careful and do thorough due diligence while approving a cash loan

# Uni-variate Analysis – contd.

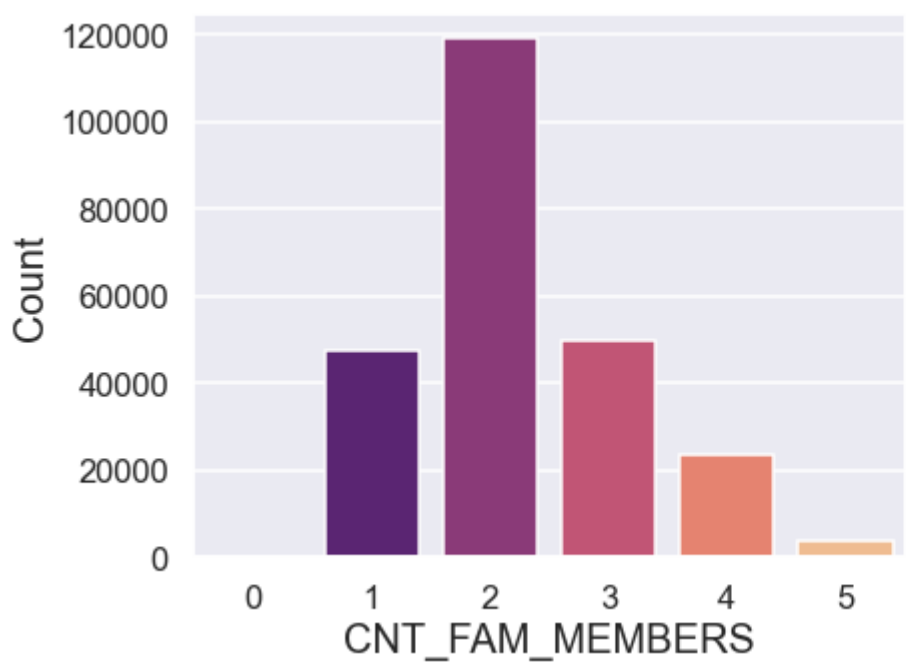
- **Number of Children**

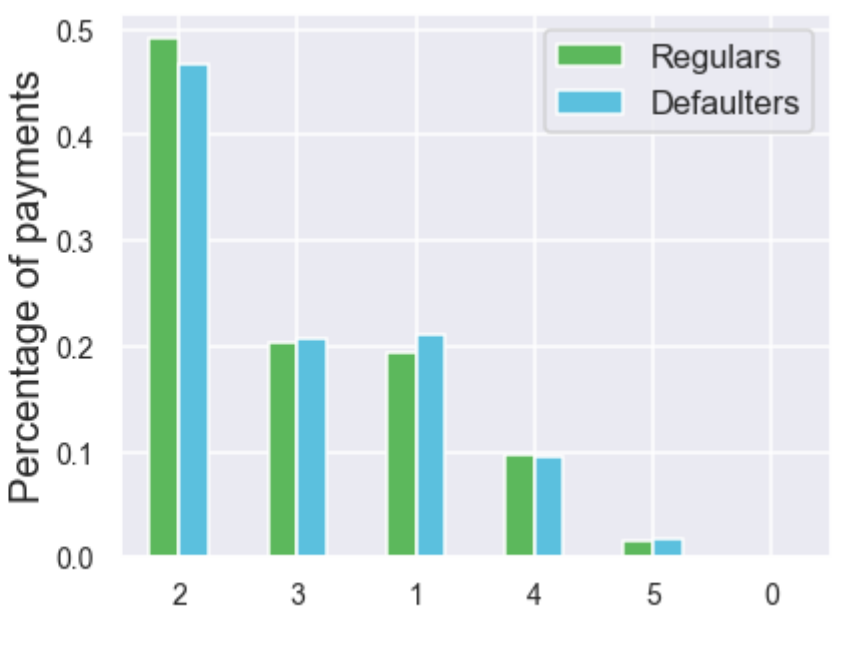



1) Maximum no. of loan applicant are from people who don't have children
2) Maximum % of Applicants who are defaulters are the ones with 1 or more Children

# Uni-variate Analysis – contd.

- **Number of Family members**

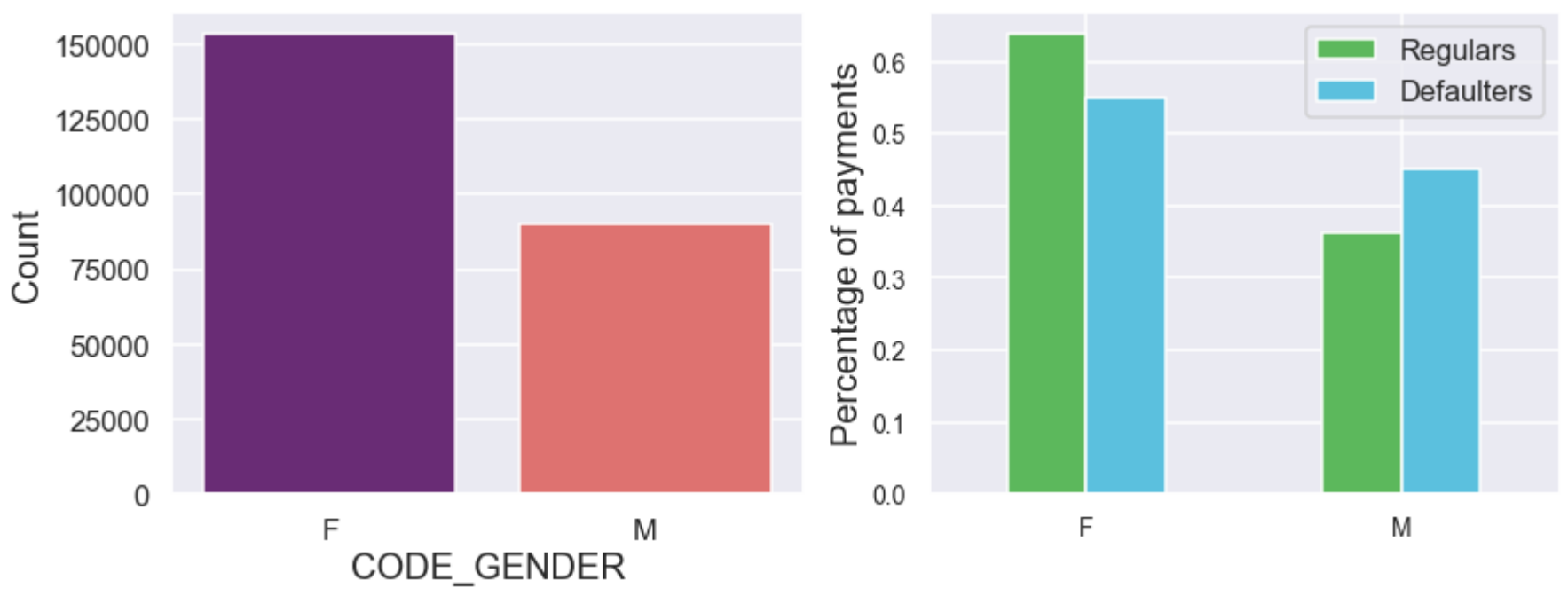



1) Most of the applicants have a family of 2.
2) Maximum % of Applicants who are defaulters are the ones with 2 or more family members

# Uni-variate Analysis – contd.

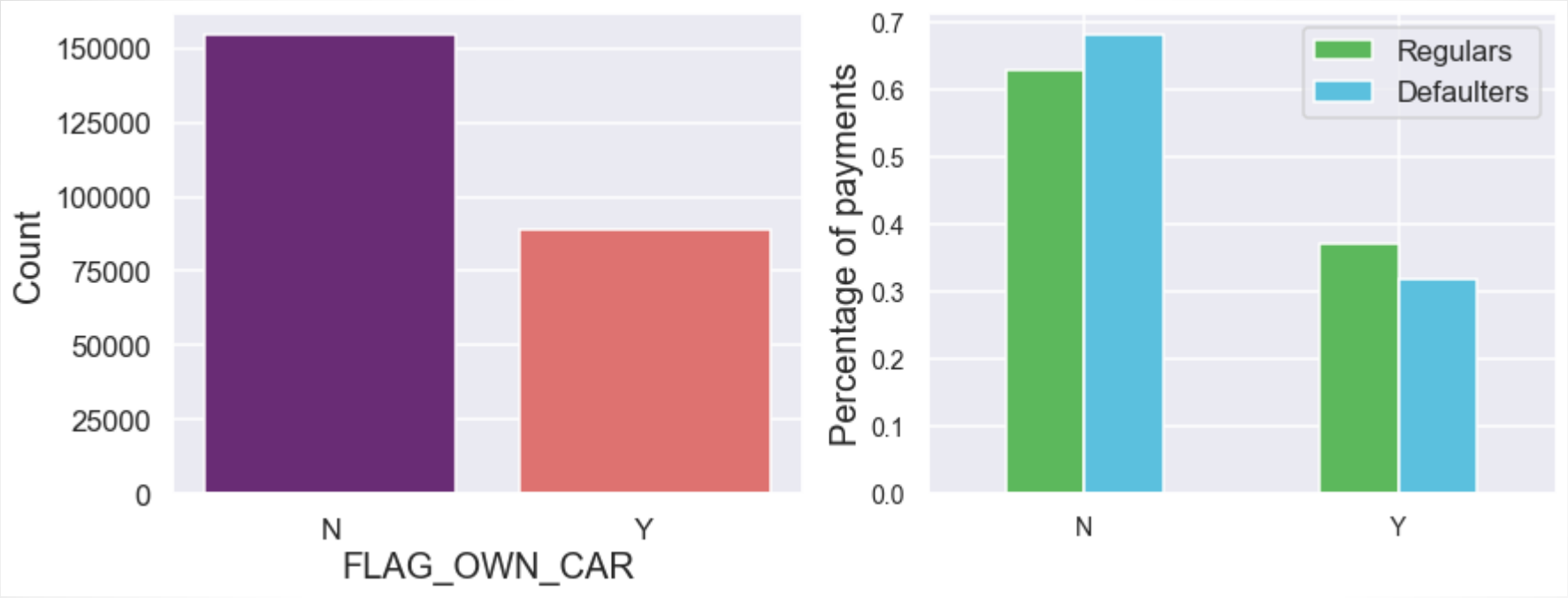
- **Gender**




1) Female Applicants are more than male Applicants
2) Men seem to be maximum defaulters when compared to females

# Uni-variate Analysis – contd.

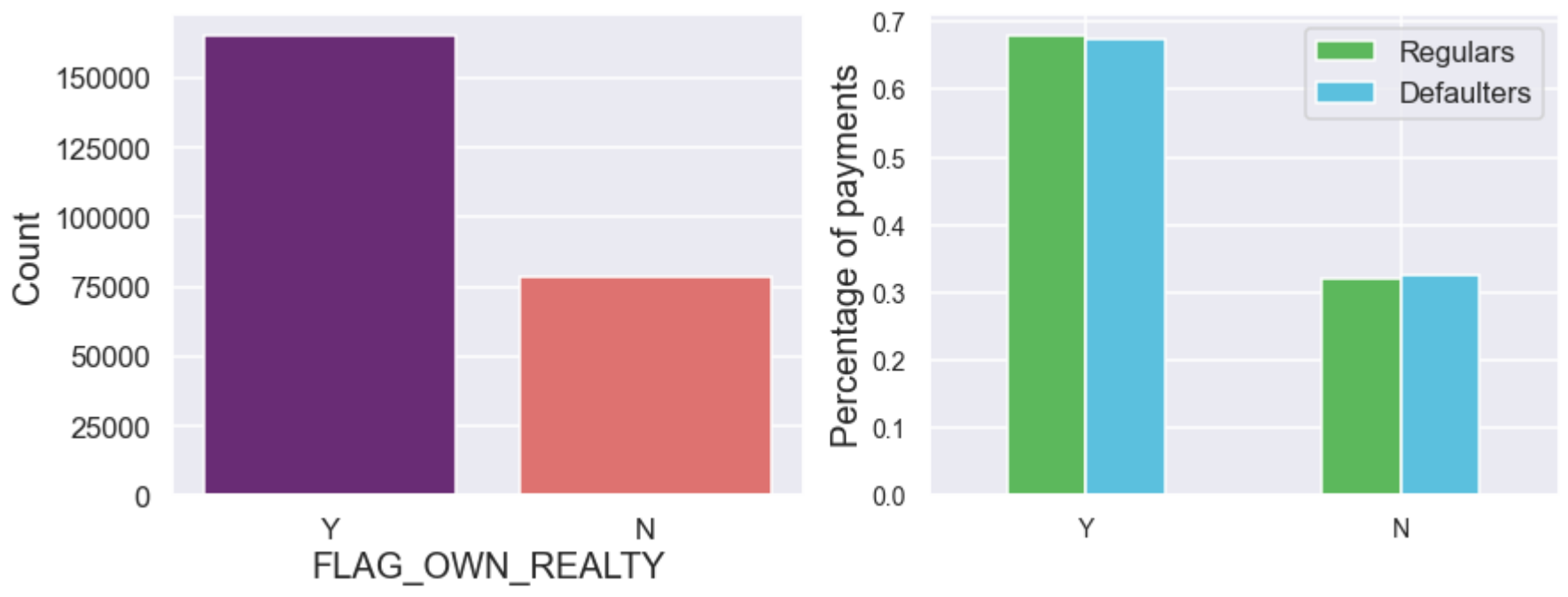
- **Owning Car**




1) Applicants without a car are more in number than applicants with a car
2) Clients without a car seem to be defaulting more when compared to people with cars

# Uni-variate Analysis – contd.

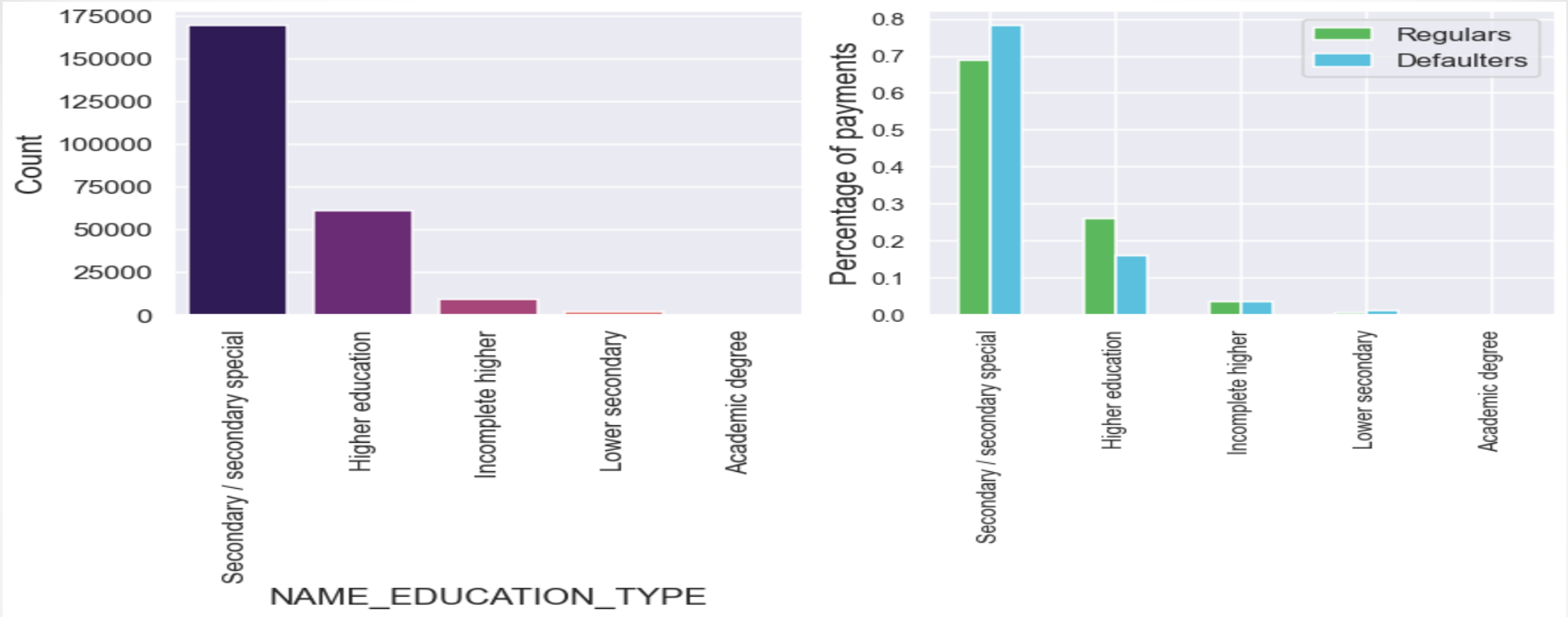
- **Owning House**




1) Applicants with a house are borrowing more than applicants without a house
2) Defaulters with a house are more than defaulters without a house

# Uni-variate Analysis – contd.

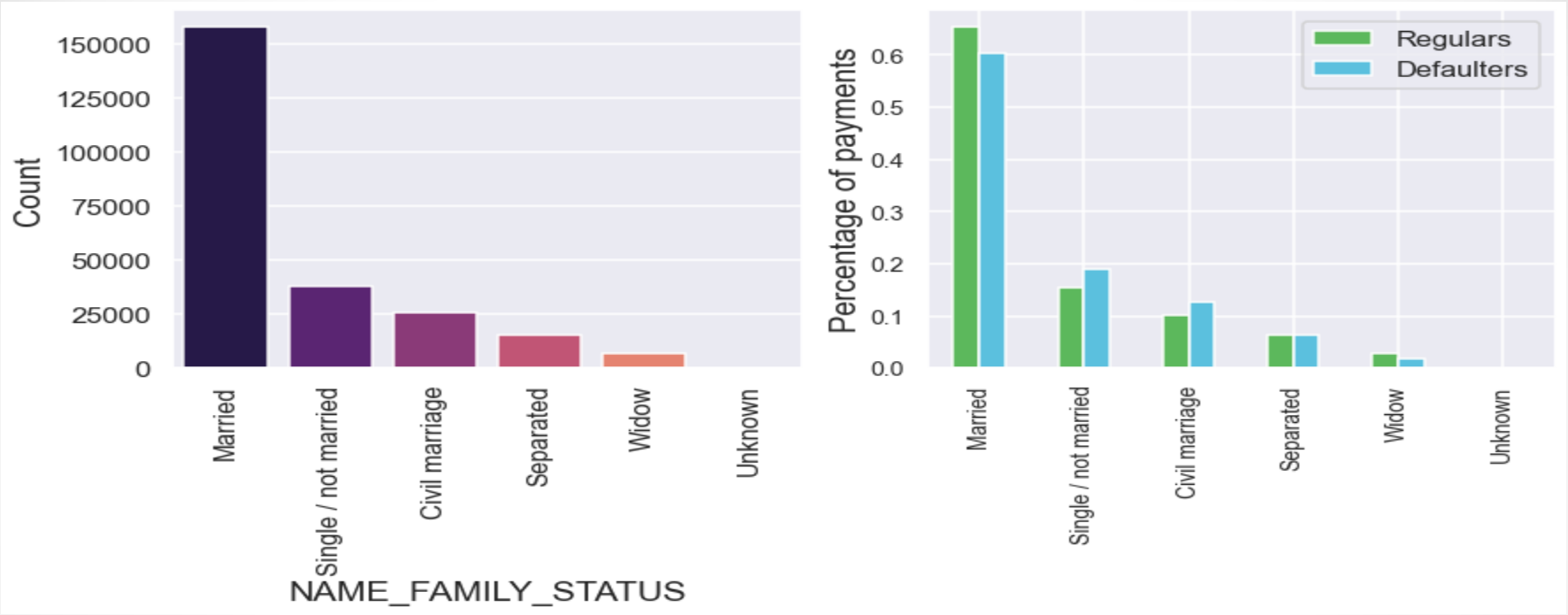
- **Education Type**




1) The Education Background of most of the loan applicants is Secondary Education
2) Maximum defaults are also from that category

# Uni-variate Analysis – contd.

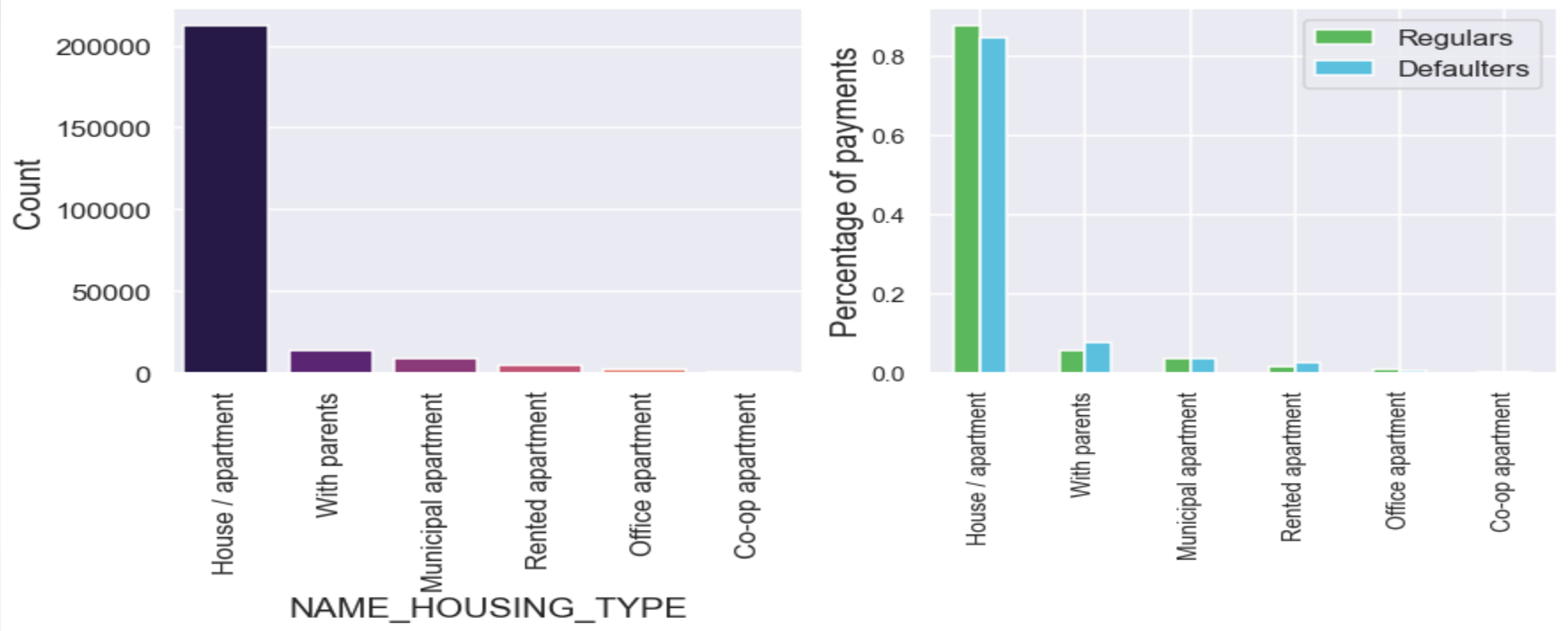
- **Family Status**




1) Married people are the maximum borrowers
2) While Defaulters are mostly from all types of family status

# Uni-variate Analysis – contd.

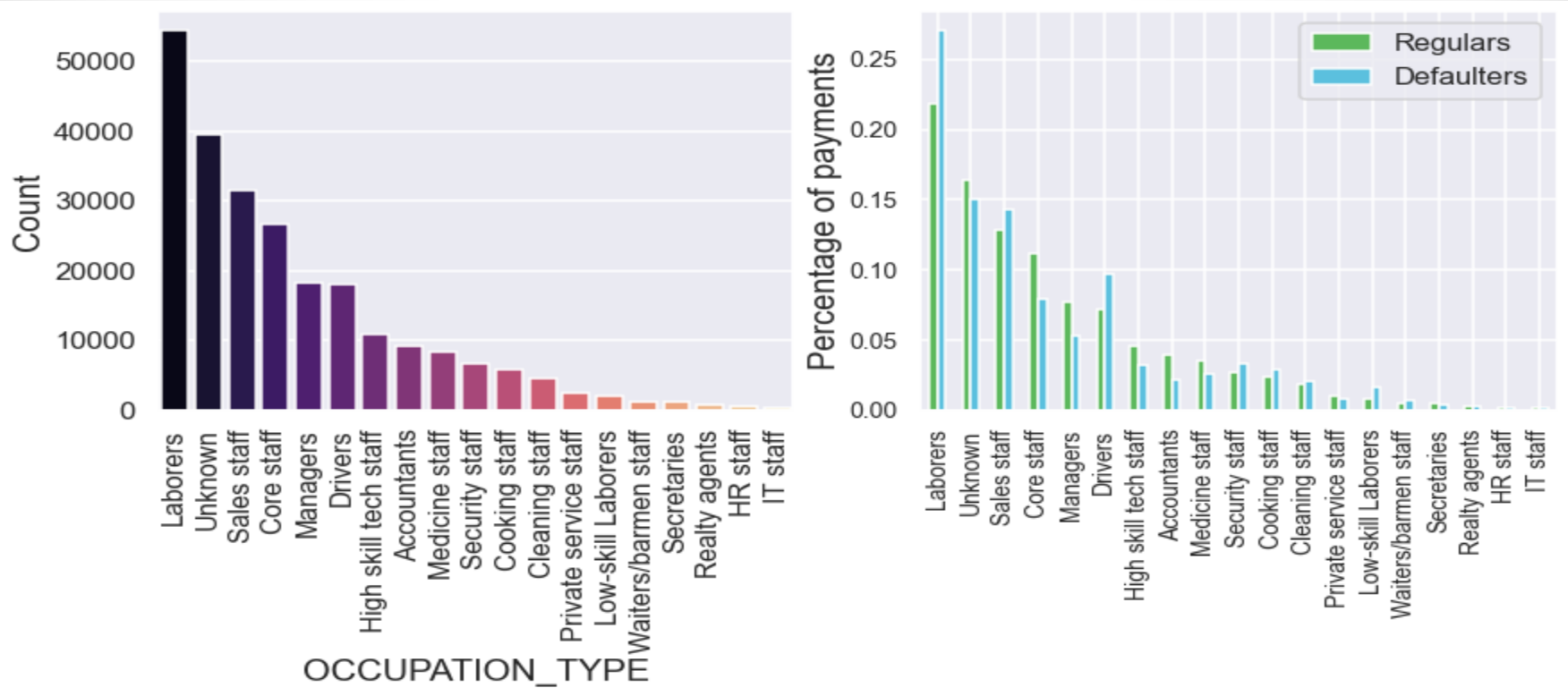
- **Housing Type**




1) Clients who are living in a house or an apartment apply the most for the loans
2) These are the same category of people who are the maximum defaulters
3) % of clients living with parents default more

# Uni-variate Analysis – contd.

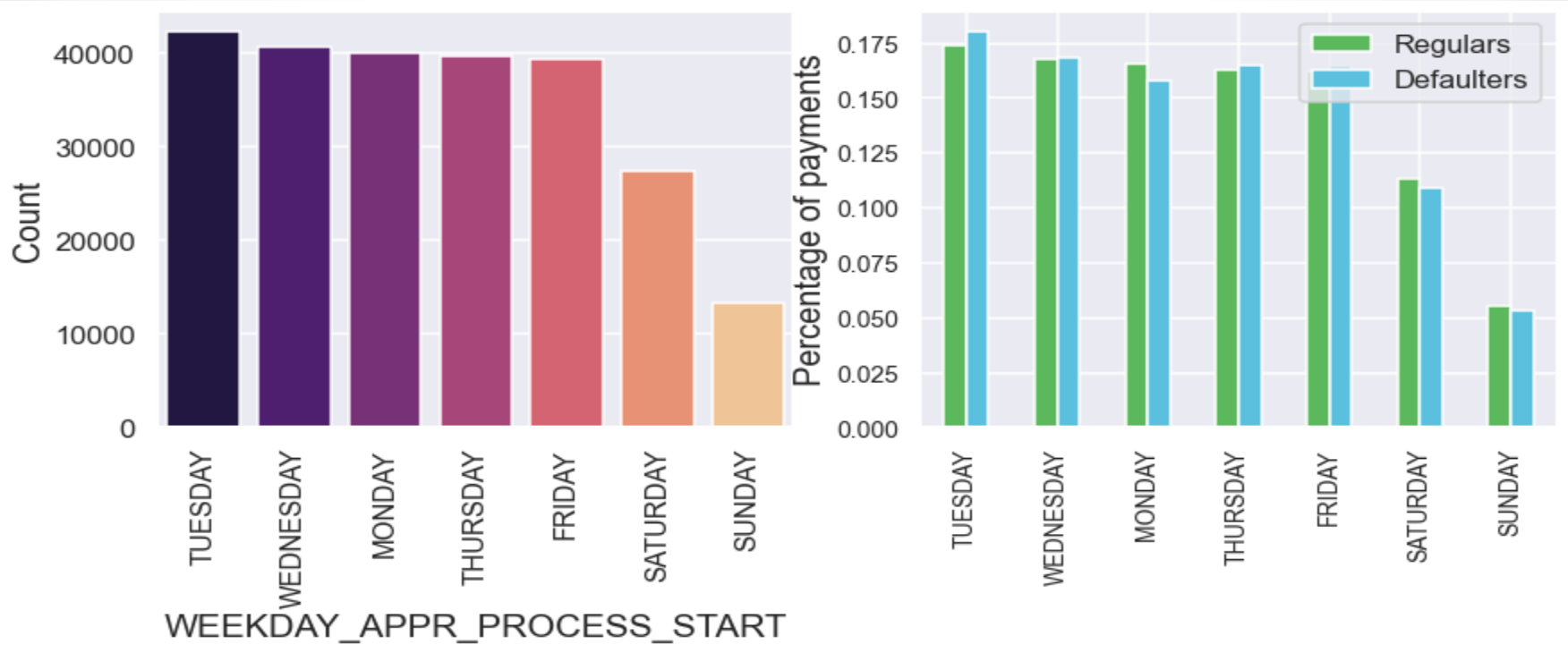
- **Occupation Type**




1) Maximum defaulters are labourers, Sales Staff, Core Staff and Drivers

# Uni-variate Analysis – contd.

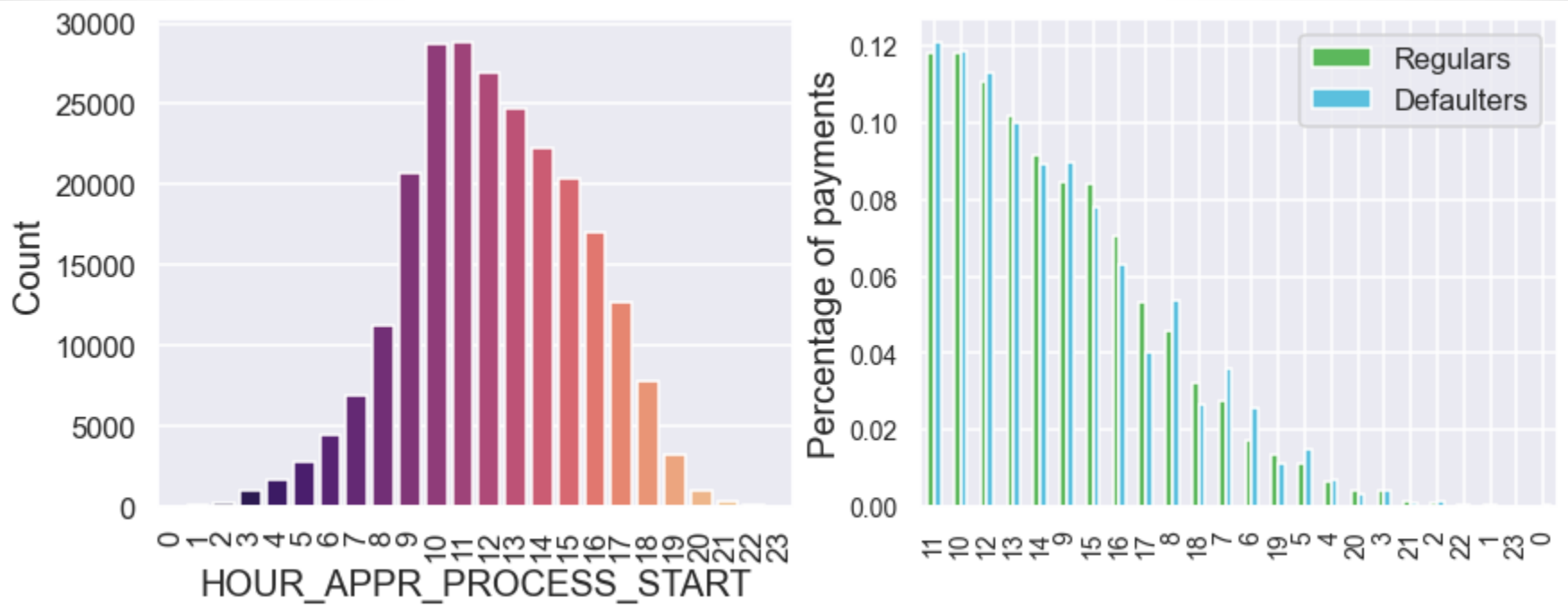
- **Week day process**




1) Tuesdays is the most busiest in application processing compared to all other days.
2) Most of the defaulters applications was processed on a Tuesday

# Uni-variate Analysis – contd.

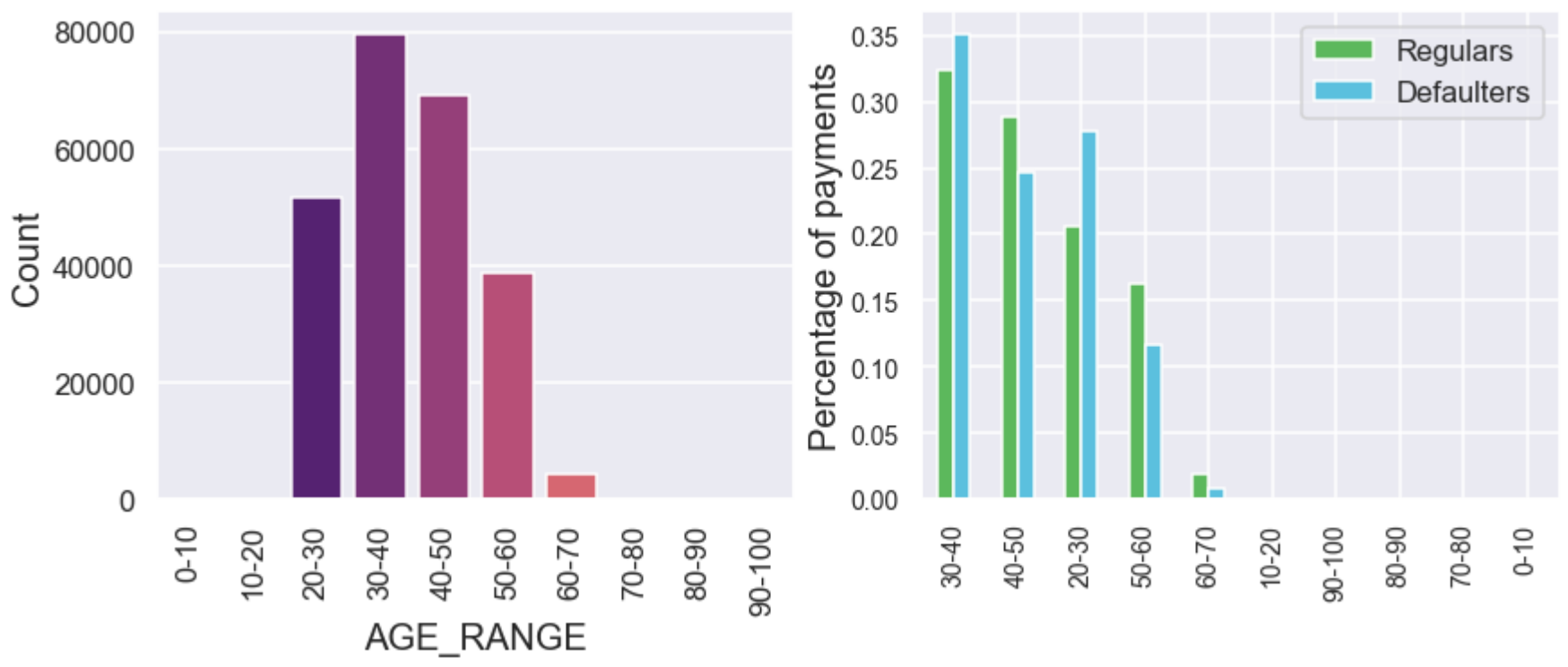
- **Hour process**




1) The busiest time of the day for application processing is between 10 – 11 am
2) Most Default Applications are processed during the same time

# Uni-variate Analysis – contd.

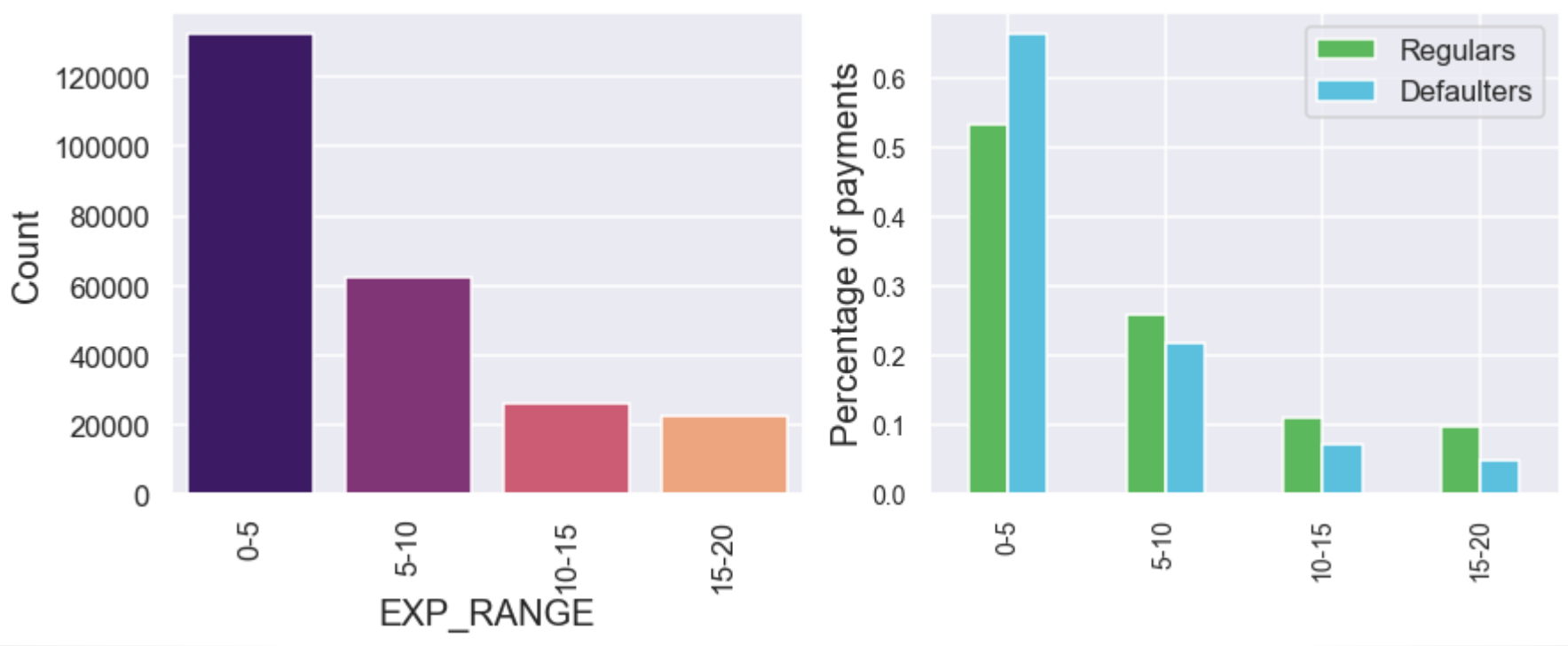
- **Age Range**




1) Most of the Applicants fall under the age group 30-40, 40-50 and 20-30 years, 30-40 being the maximum category
2) Maximum defaulters are in the age range 30-40 years and 20-30 years

# Uni-variate Analysis – contd.

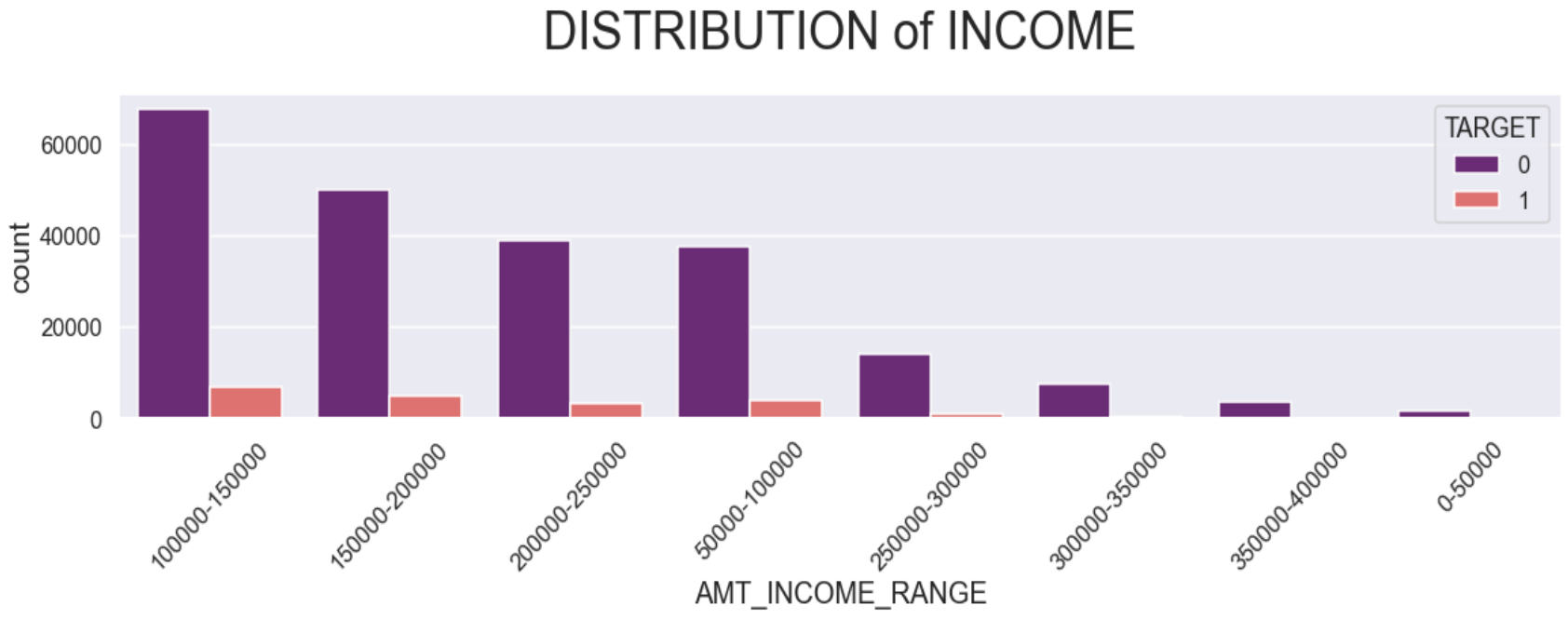
- **Experience Range**




1) Most of the Applicants have below 5 years experience with next category 5-10 years experience
2) Maximum defaulters have below 5 years experience

# Uni-variate Analysis – contd.

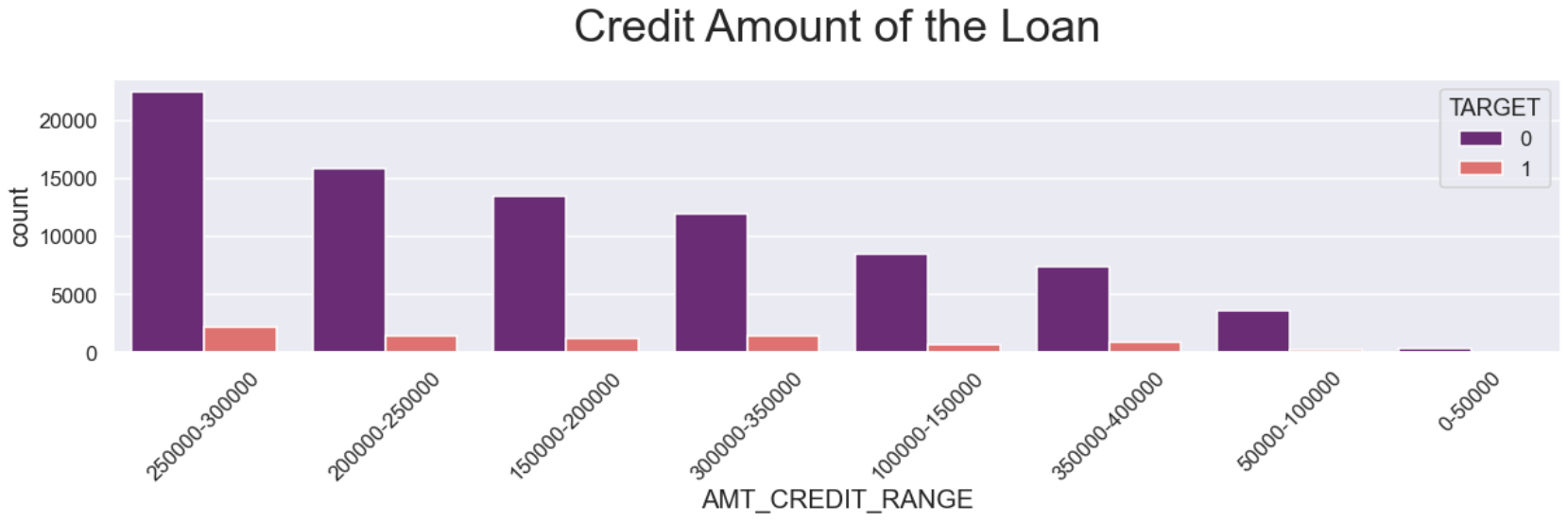
- **Income Distribution**




1) Most of the applicants Income range is around 50k -250k
2) Most of the defaulters have income in the same range, maximum defaulters being in the income range 100k-150k

# Uni-variate Analysis – contd.

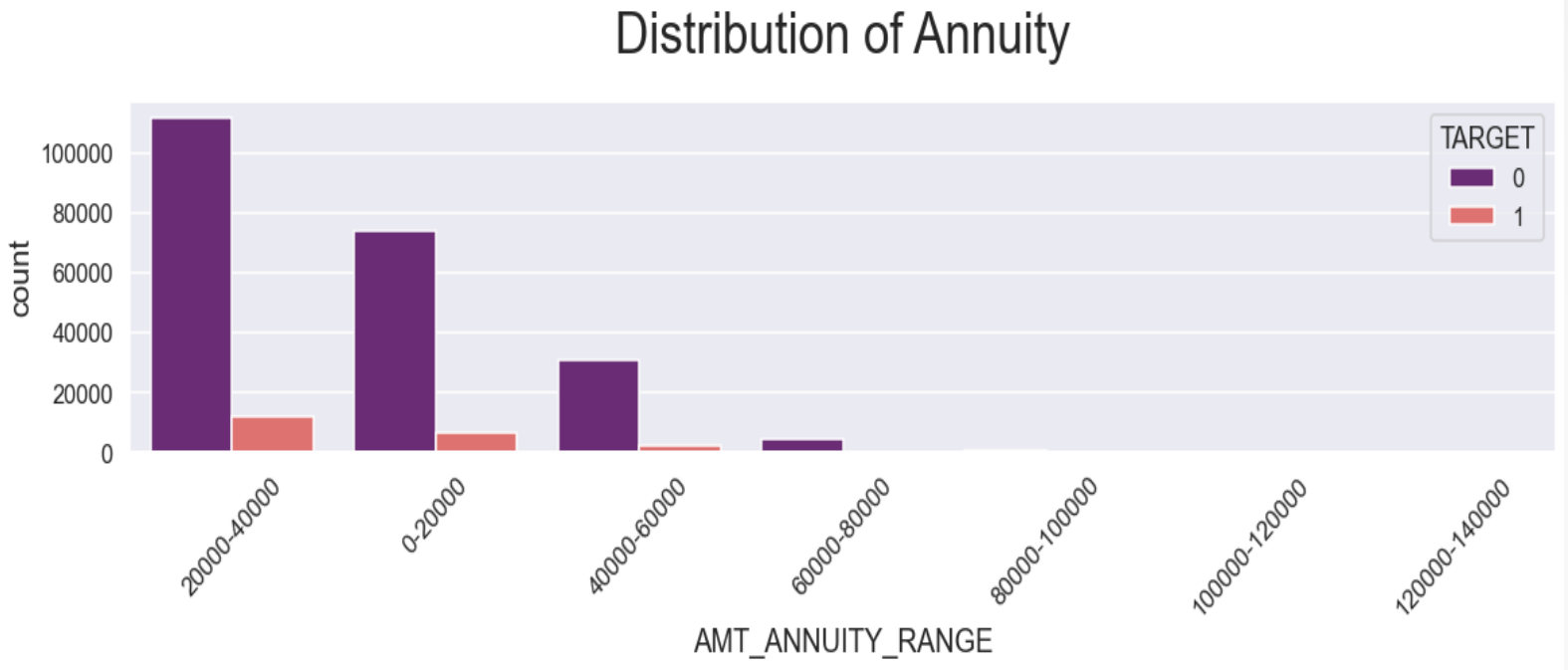
- **Credit Distribution**




1) Maximum Loan applications are for Credit amount 250-300k
2) Maximum Defaulters are the ones who are seeking the same credit amount

# Uni-variate Analysis – contd.

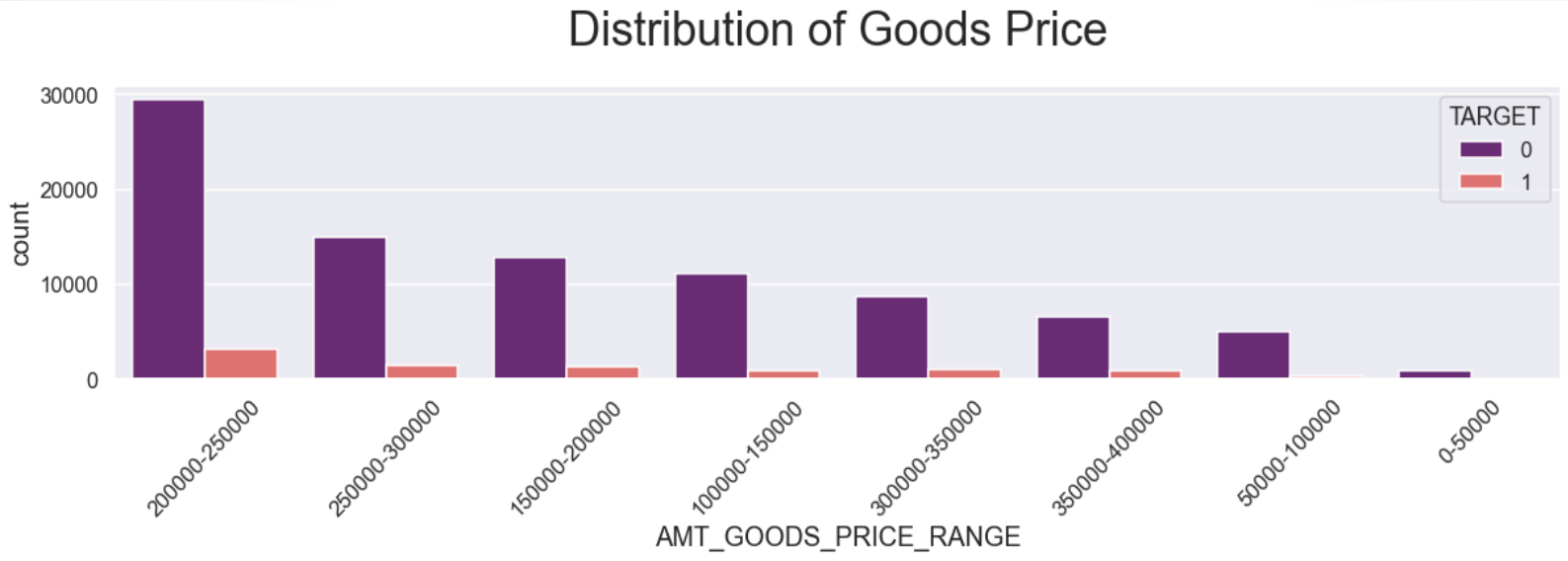
- **Annuity Distribution**




1) Maximum Applicants are paying an annuity of 20-40k
2) People paying an annuity of 20-40k are the ones who are defaulting a lot
3) People paying an annuity of 60k or more are hardly defaulting

# Uni-variate Analysis – contd.

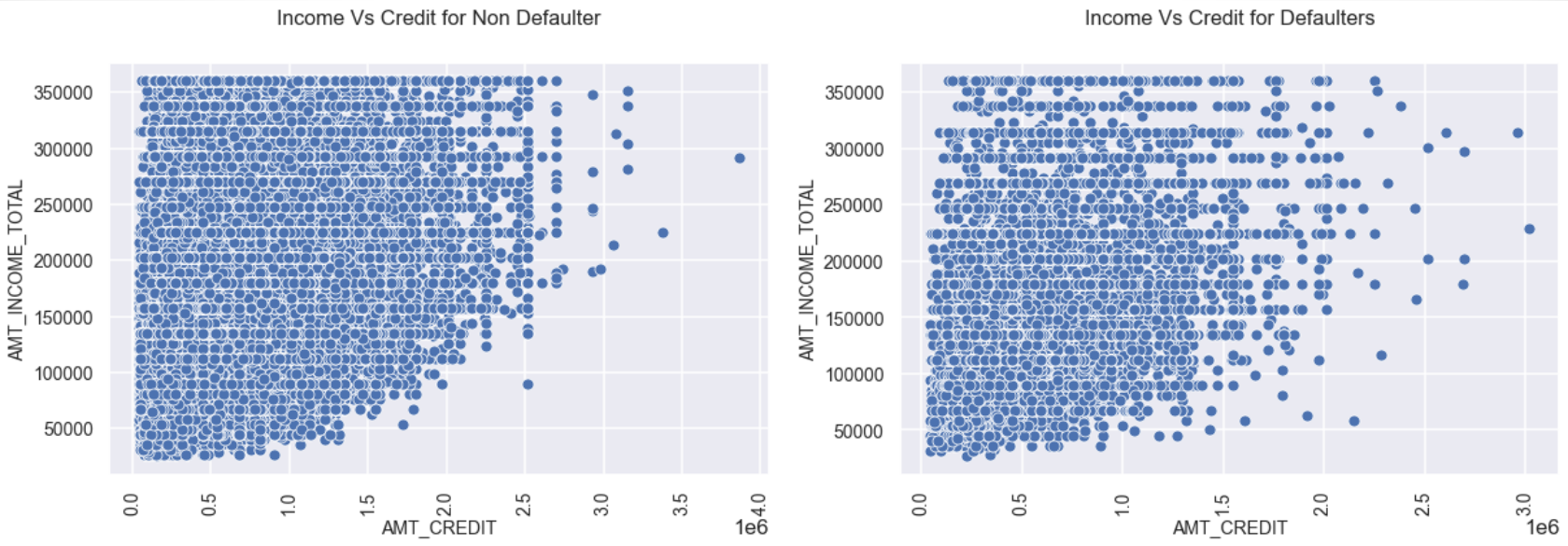
- **Goods Price Distribution**




1) The value of asset for which loan is sought maximum is in the range 200k-250k
2) That is also the category in which there are maximum defaulters

# Bi-variate Analysis

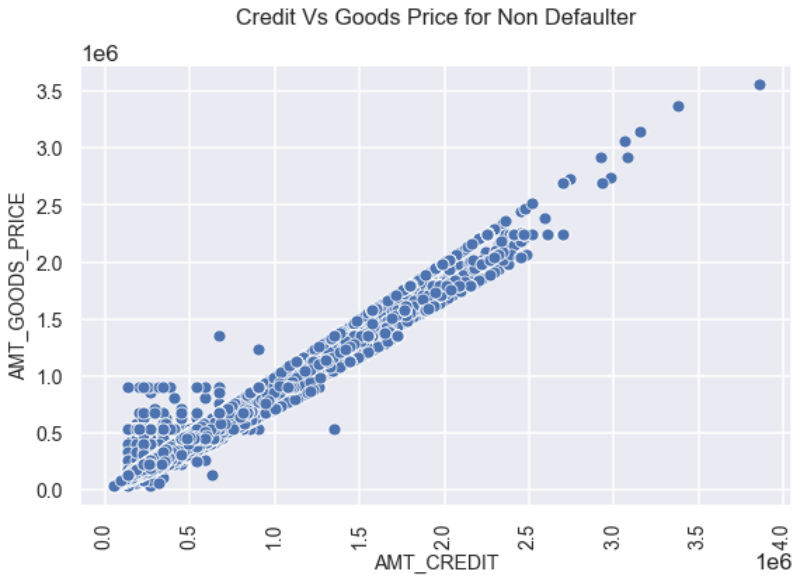
## ➢ Income Vs Credit

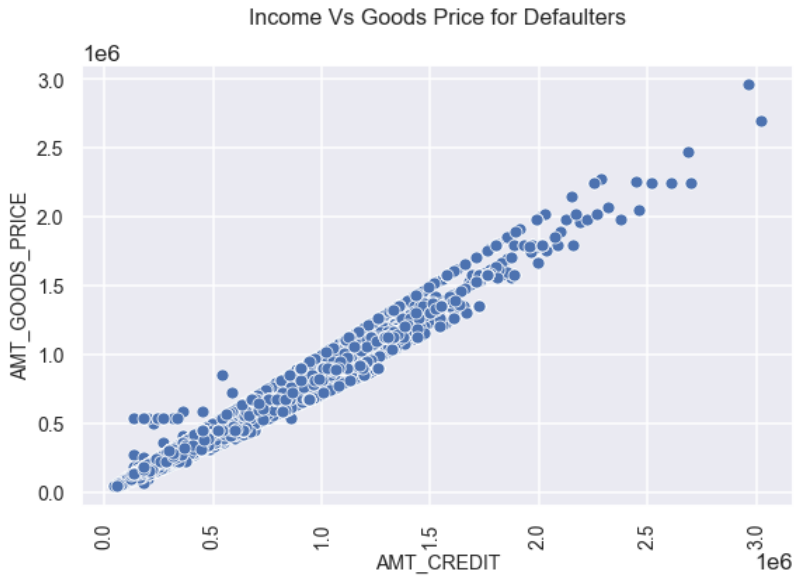



1) Lower density of defaults where income is higher than 300k
2) More defaults are in the range 0-150k

# Bi-variate Analysis – contd.

## ➢ Credit Vs Goods Price

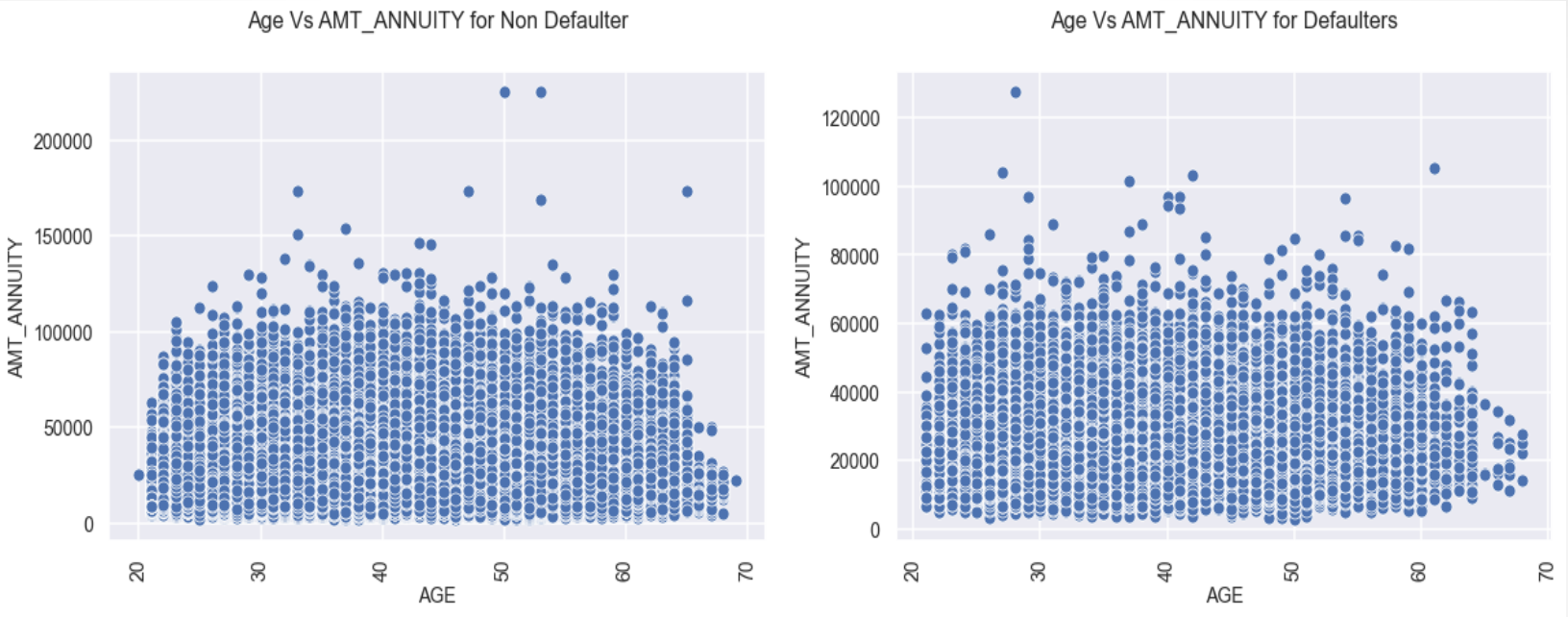



1) Most of the defaulters are defaulting for the credit amount in the range 0 -200k

# Bi-variate Analysis – contd.

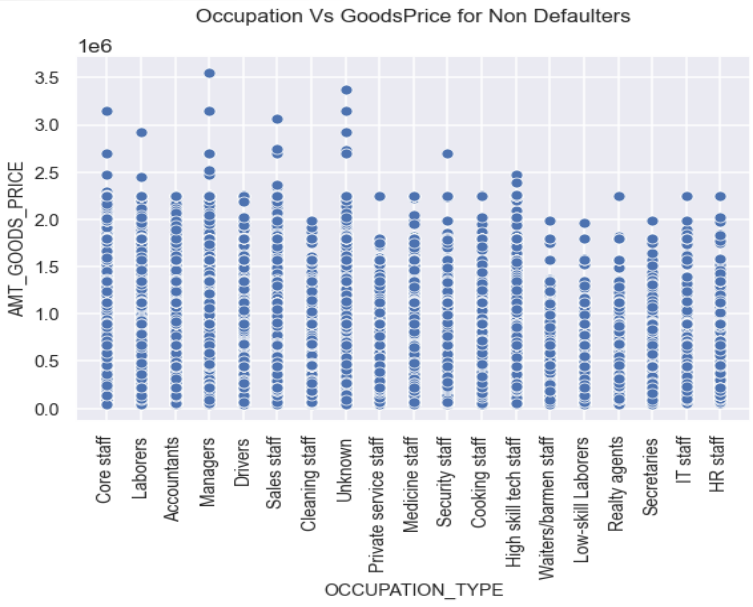
- **Age Vs Annuity**

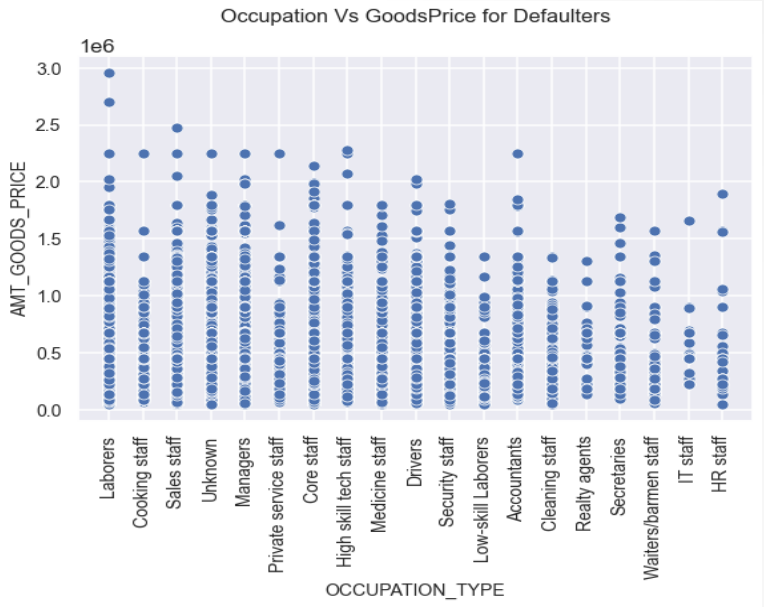



1) Most of the defaulters are defaulting for the Annuity amount in the range 0 -60k

# Bi-variate Analysis – contd.

## ➢ Occupation Vs Goods Price

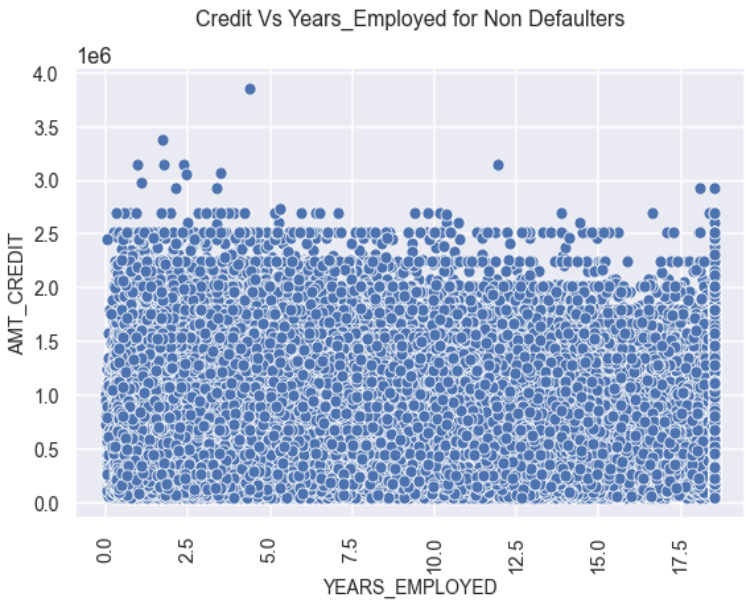


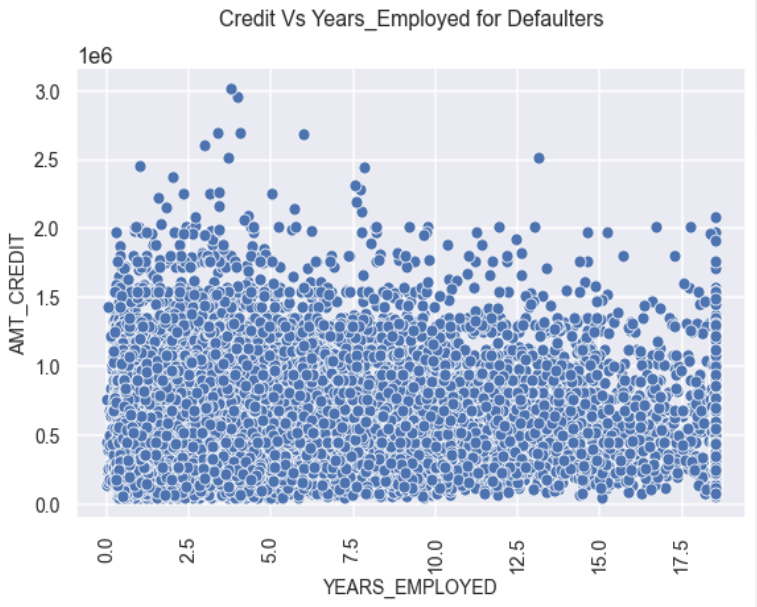




1) Most of the defaulters are labourers, Core Staff , Sales staff ,Managers & Medicine for the Asset Price in the range of 0 – 150k

# Bi-variate Analysis – contd.

- **Credit Vs Years_Employed**

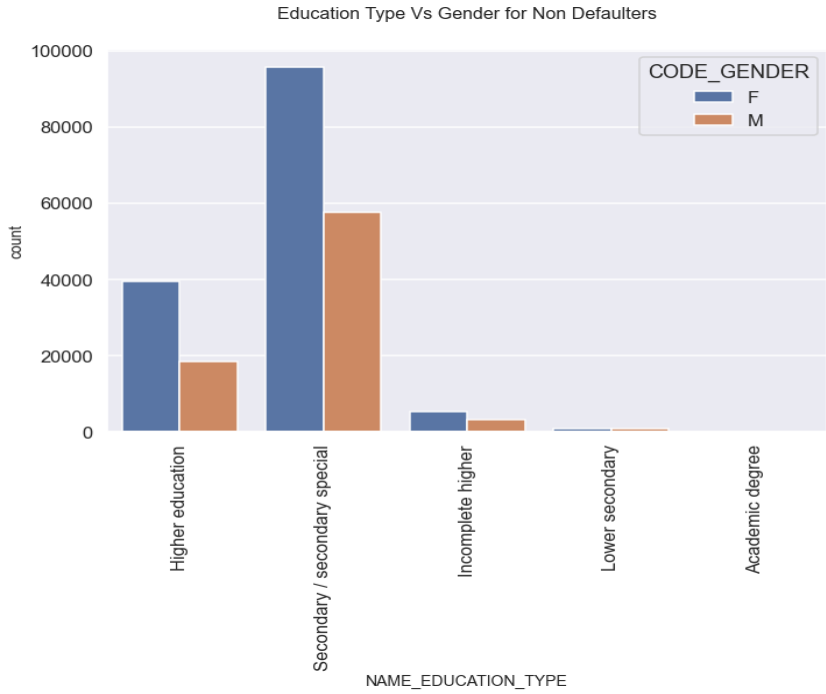


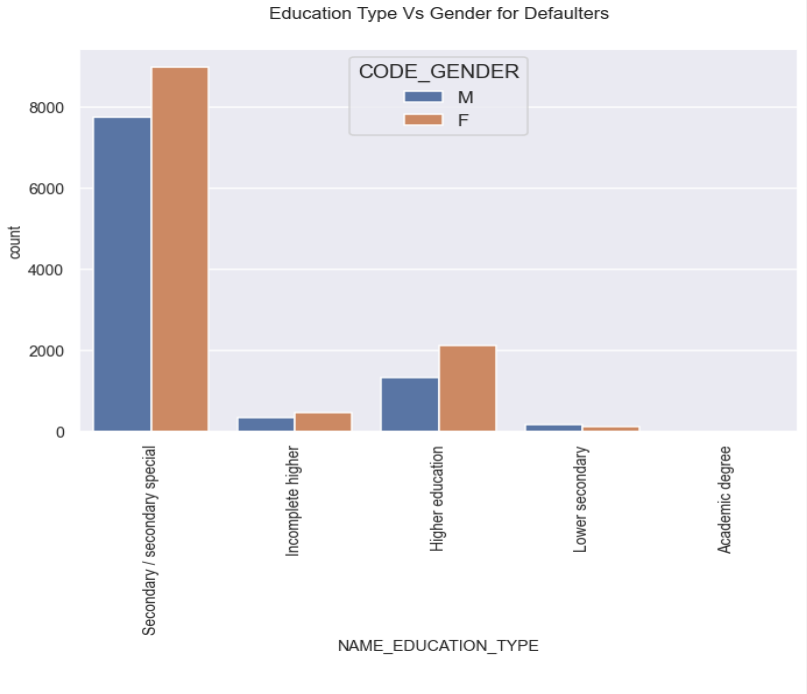




1) Most of the defaulters fall in the credit limit range between 0-1500k
2) These defaulters have an experience around 0-15 years

# Bi-variate Analysis – contd.

## ➢ Education Type Vs Gender

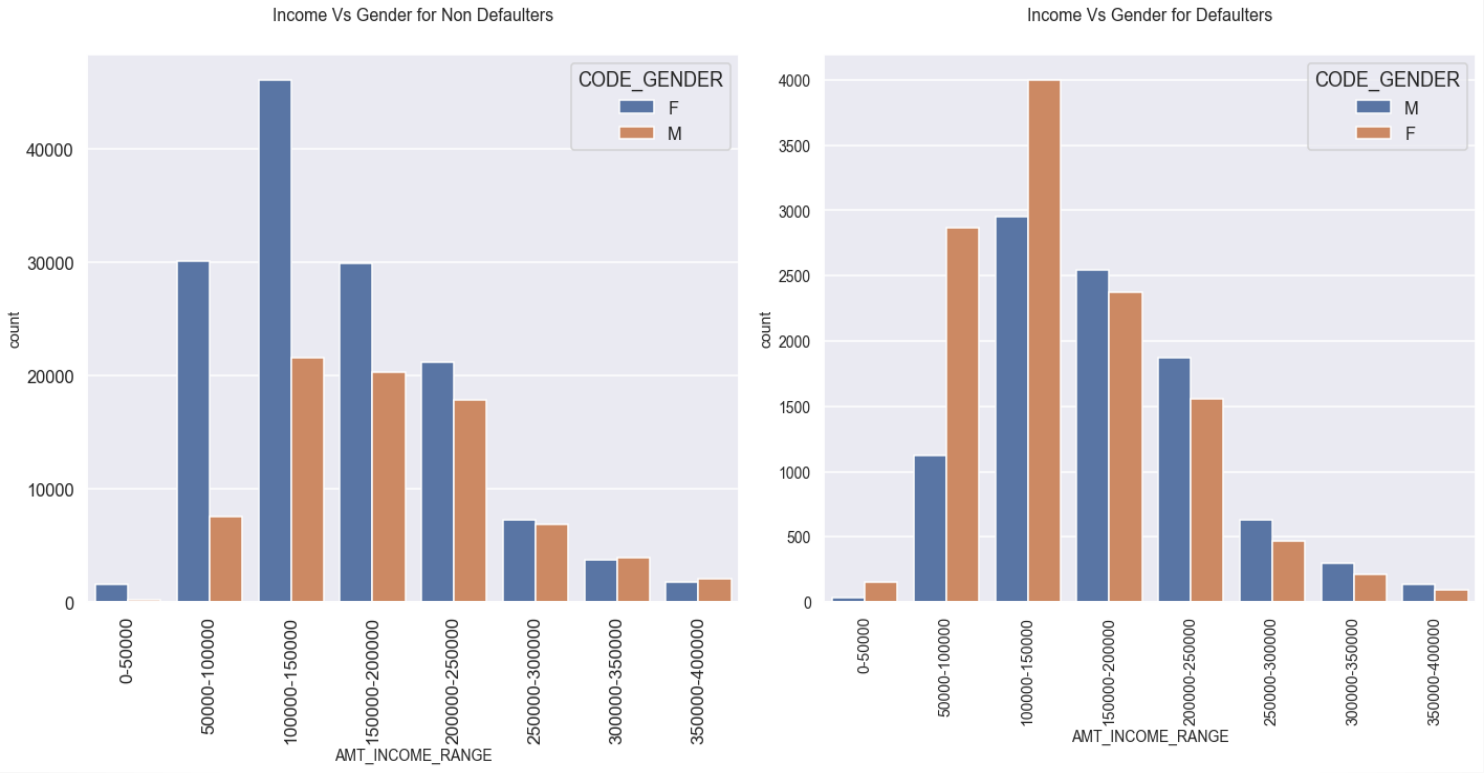



1) Most of the defaulters are females with education levels of Secondary and Higher Education

# Bi-variate Analysis – contd.

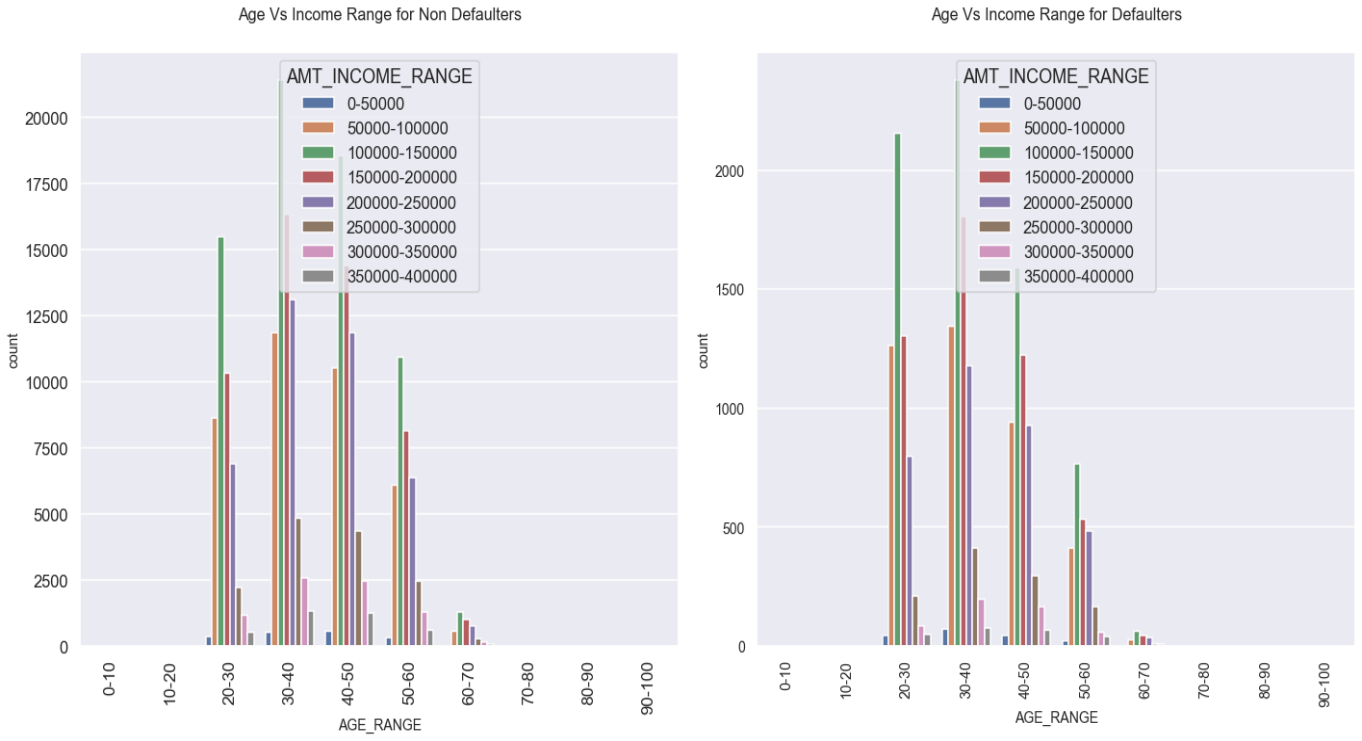
## ➢ Income Vs Gender




1) Most of the defaulters are females with income range 50k-150k

# Bi-variate Analysis – contd.

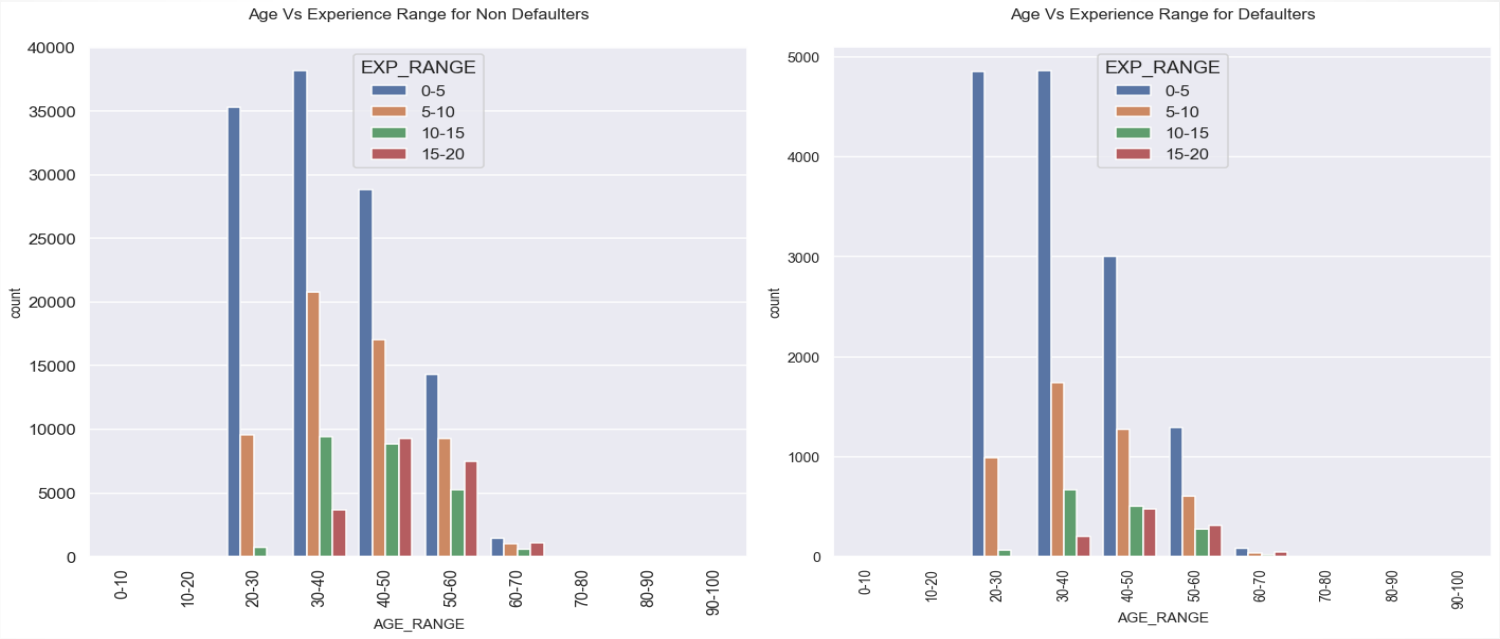
- **Age Vs Income Range**




1) Most of the defaulters belong to the income group 100-150k
2) Defaulters are less in the age group 60-70 years

# Bi-variate Analysis – contd.

## ➢ Age Vs Experience Range



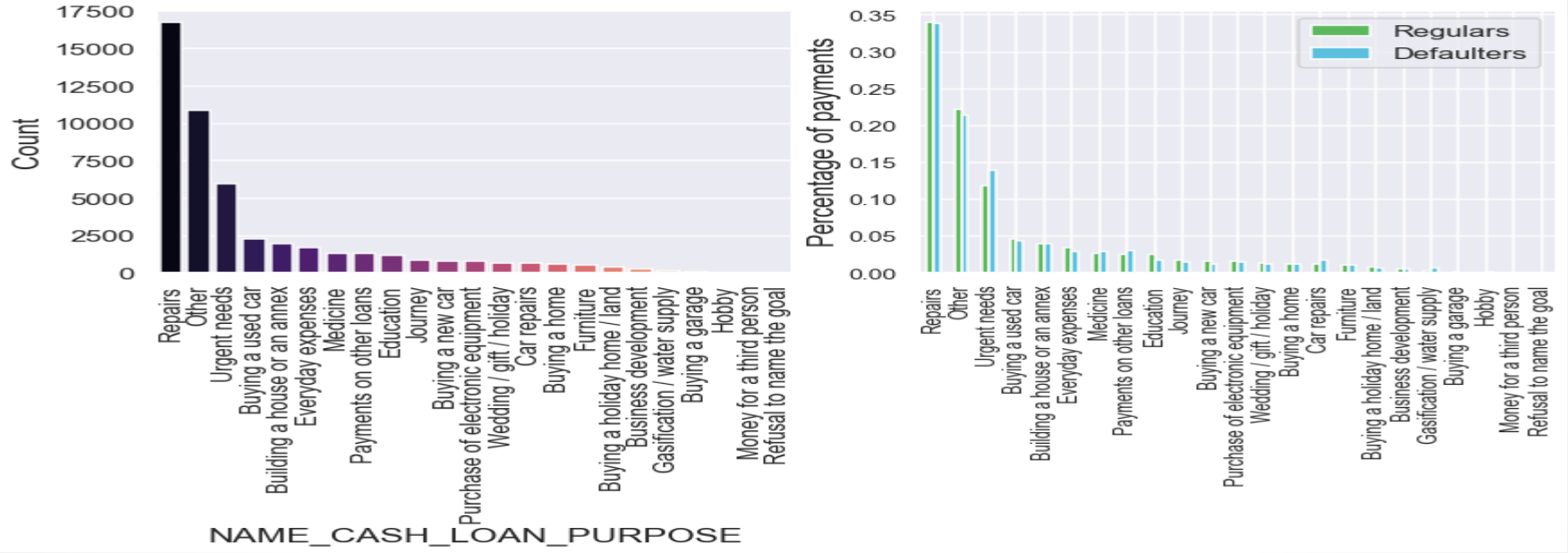

1) Most of the defaulters have an experience of 0-5 years and they fall into the age category 20-30, 30-40 and 40-50 yrs

# Data Loading, Inspection & Cleaning on Previous Application data

1. There are 1670214 rows and 37 columns
2. On performing Null value analysis, more number of columns are identified
   a) Decision is taken to drop the columns from the Data frame as there is no connection or clear perception of the column or very limited information
      1. Columns having more than 40% of Null values
   b) Impute the other Null value columns having less percentage with Mode
3. Then merge Application and Previous Data

# Uni-variate Analysis on Previous data

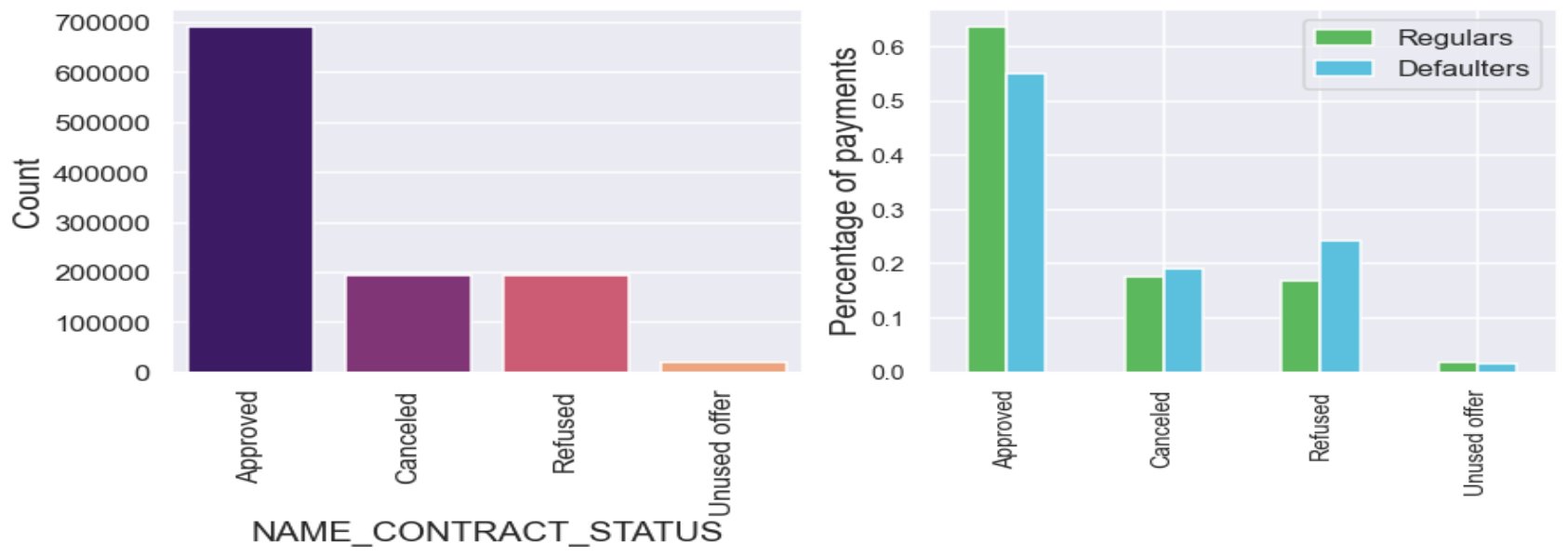
- **Cash Loan Purpose**




1) The Reasons for applying for a loan are mostly for Repairs, Other, Urgent Needs & Buying a used car
2) Defaulters of loan are mainly from the category 'Repairs, Other and Urgent Needs' and who borrowed for investing in a hobby

# Uni-variate Analysis on Previous data – contd.

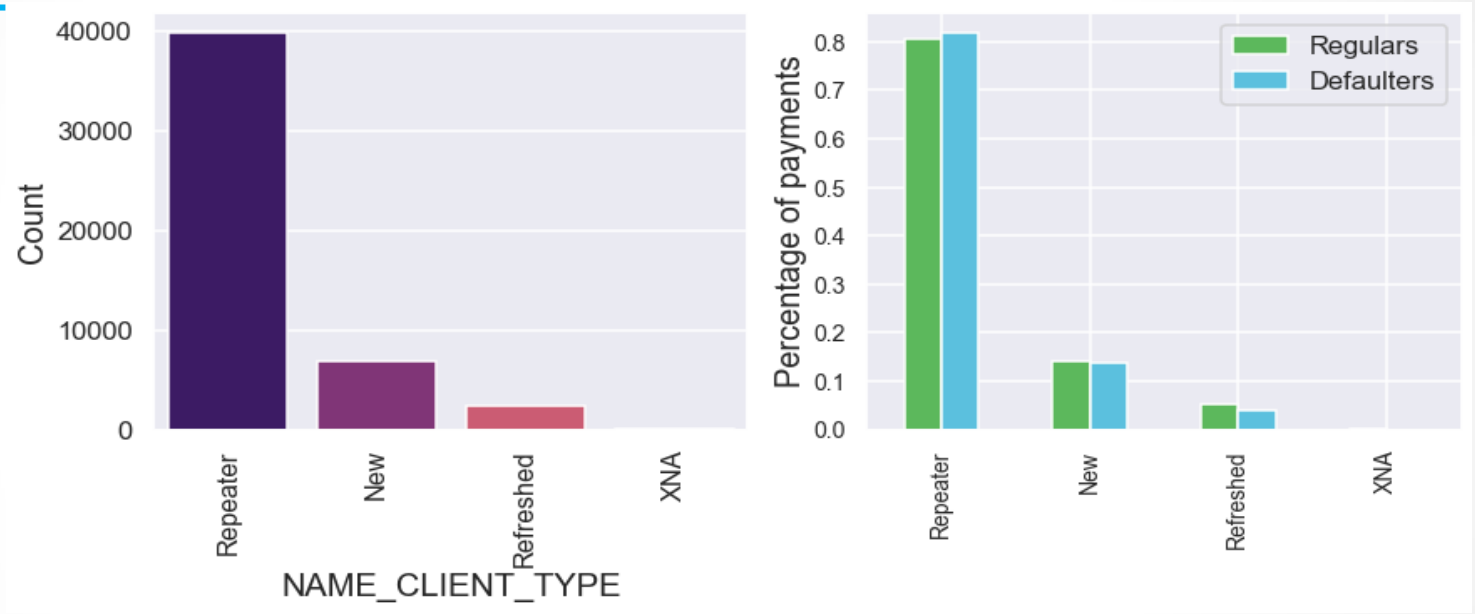
- **Contract Status**




1) Of all the Loans applied, most of them have been approved. Some have been Cancelled or Refused. Some have been unused
2) Most of the defaults are from those loans which have been approved

# Uni-variate Analysis on Previous data – contd.

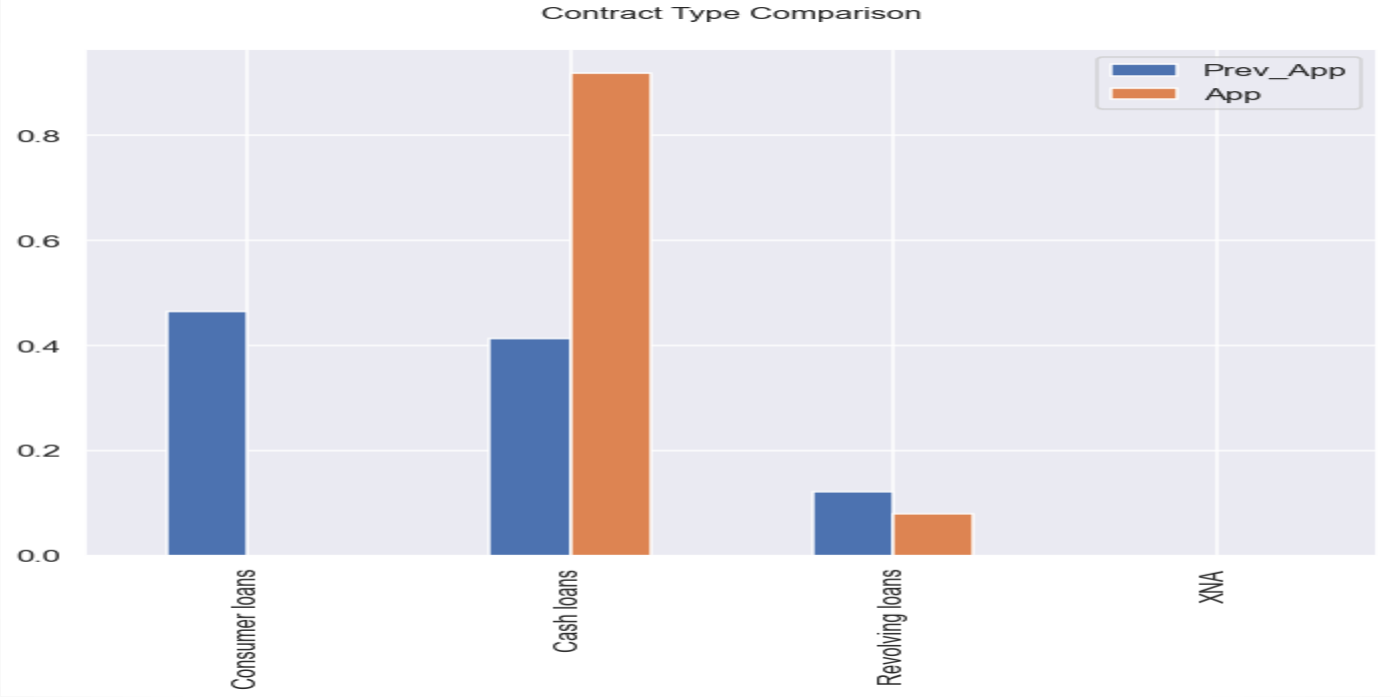
- **Client T**




1) Repeat Customers are also defaulting on the loans

# Comparison between Application and Previous Application

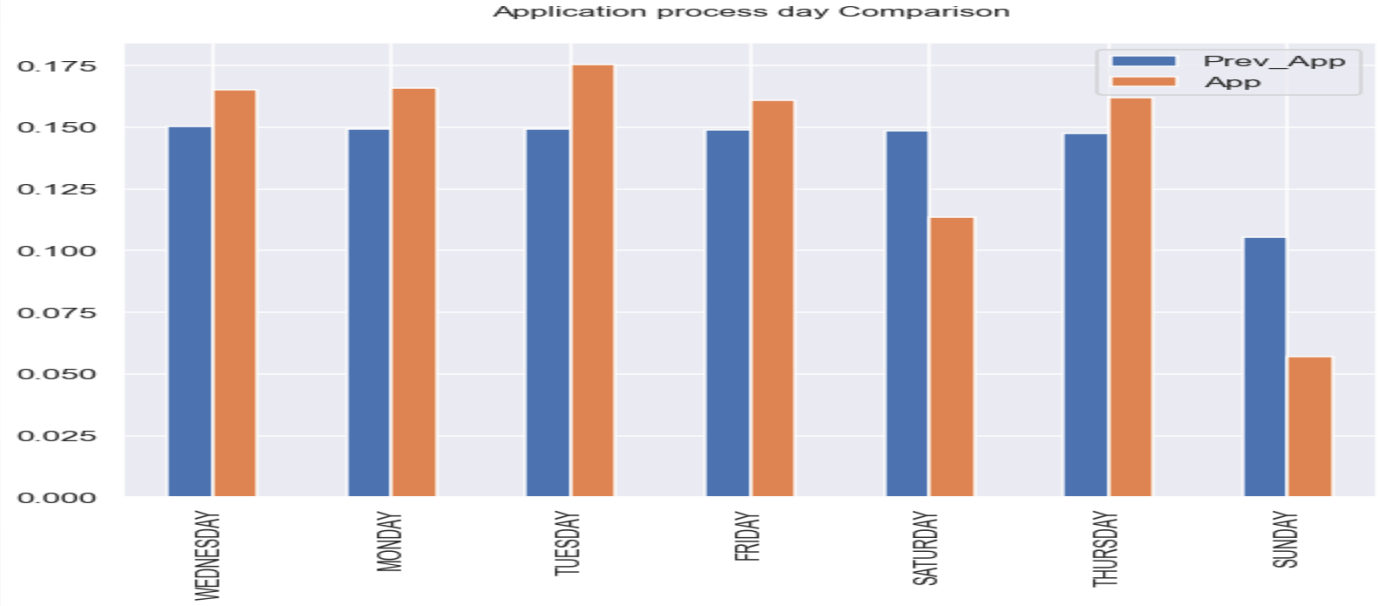
- **Contract Type**




1) Previous Application Data also has Consumer loans along with Cash & Revolving Loans

# Comparison between Application and Previous Application – contd.

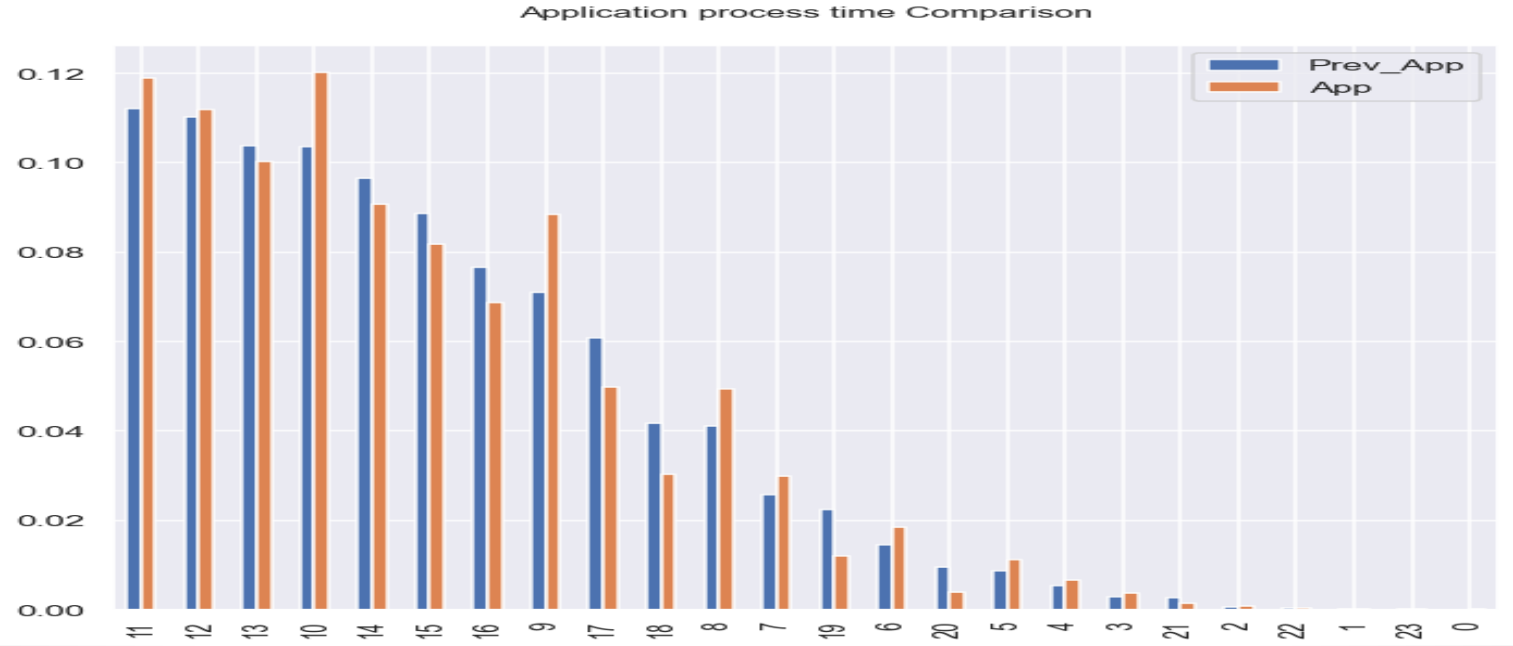
- **Week Day process**




1) According the Previous Application Data, All Week days are engaged in processing loan Applications equally except Sundays where as in current Application Data, Tuesdays are when most applications are processed

# Comparison between Application and Previous Application – contd.

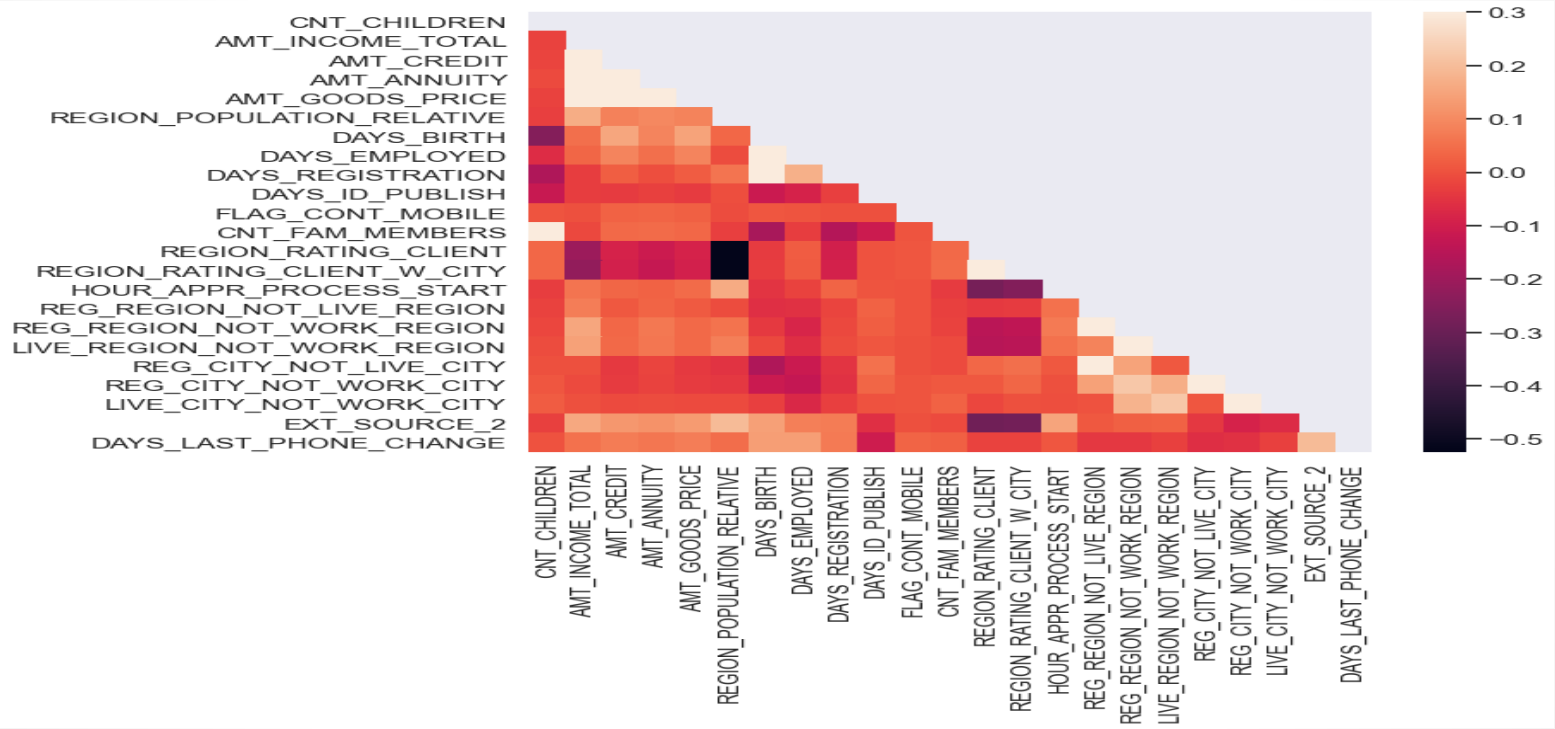
- **Hours process**




1) According to Previous Application Data, the time interval of processing Loan Applications the most is 11-12 am while in Current Application Data, it is 10-11 am

# Top 10 Correlations

- **Regular Payers – Heat Map**



# Top 10 Correlations – contd.

- **Regular Payers – Correlation**

| | | |
|---|---|---|
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.820155 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.820155 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.860388 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.860388 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.890078 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.890078 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.949275 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.949275 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.985697 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.985697 |

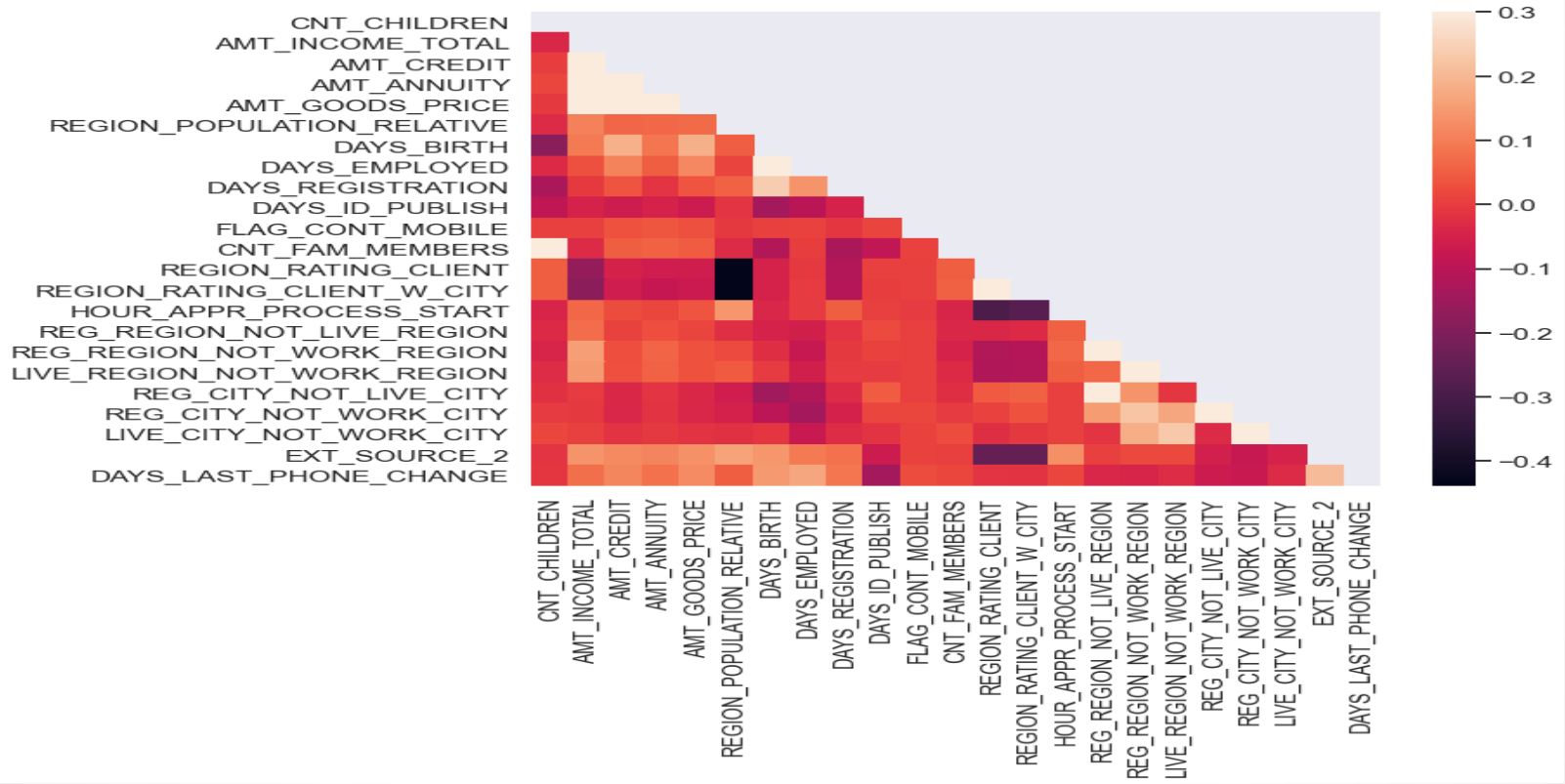
dtype: float64

- Regular Payers have high correlation between below columns

1. AMT_CREDIT & AMT_GOODS_PRICE
2. AMT_GOODS_PRICE & AMT_CREDIT
3. REGION_RATING_CLIENT_W_CITY & REGION_RATING_CLIENT

# Top 10 Correlations – contd.

- **Defaulters – Heat Map**



# Top 10 Correlations – contd.

- **Defaulters – Correlation**

| | | |
|---|---|---|
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.767435 |
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.767435 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.848419 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.848419 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.889049 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.889049 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.957783 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.957783 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.981571 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.981571 |

dtype: float64

- Defaulters have high correlation between below columns

1. AMT_CREDIT & AMT_GOODS_PRICE
2. AMT_GOODS_PRICE & AMT_CREDIT
3. REGION_RATING_CLIENT_W_CITY & REGION_RATING_CLIENT

# Top 10 Correlations – contd.

## Observations

- Credit amount is inversely proportional to Age, which means Credit amount is higher for low age.
- Credit amount is inversely proportional to the number of Children (i.e) Credit amount is higher for Client with less no of Children.
- Income amount is inversely proportional to the number of Children (i.e) Clients who tend to have more income have lesser no of Children
- Credit amount is higher in densely populated area.
- The income is also higher in densely populated area.

# Recommendations

- The percentage of Regular Payers are 91.2% and Defaulters are 8.8%
- Bank can try opting for Revolving loans than Cash Loans as default percentage is more in Cash Loans
- Bank lends more loans to Females over Males as they have more Income
- Most of the applicants were from Income group  Working, Commercial Associate and State Servant Categories. Bank can lend loan for Commercial Associate and State Servant Categories as they are less likely to default
- Most of the applicants belong to Secondary and Higher Education category. Bank can lend loan for Higher Education category as they are less likely to default
- Most of the applicants  are Married, Single or Civil marriage. Bank can lend loan for Married clients as they are less likely to default
- Most of the applicants have their own House / Apartments and live with their parents. Bank can lend loan for clients with own House / Apartments as they are less likely to default
- Most of the applicants are from 30-50 Age group. Bank can lend loan for clients having Age more than 40 years as they are less likely to default
- Bank can lend loan for clients having Experience more than 5 years as they are less likely to default
- Bank can lend loan for clients having Income more than 250k as they are less likely to default
- Bank can lend loan for clients having Credit more than 300k as they are less likely to default
- Bank can lend loan for clients having Annuity less than 60k as they are less likely to default
- Bank can lend loan for clients having Goods Price less than 100k as they are less likely to default
- Bank can lend loan for clients for Buying a home, new or used cars, Education, Medicine as they are less likely to default
- Bank can lend loan for clients having Approved status by comparing with Previous history
- Bank can lend loan for clients who are repeatedly getting loans by analysing their previous history