# Credit EDA Case Study

Team members:

- Adnan Hassan
- Avez Shariq

# Introduction

- This case study aims to explore the data sets given using various Exploratory Data Analysis methods in python, the libraries used in this case study are:
  - Pandas
  - Matplotlib
  - Seaborn

- The given data sets are:
  - “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**
  - “previous_data.csv”: Contains information about the client’s previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer**

# “application_data.csv”: Target Variable

The file has data related to current application

Going through the feature list we observe the target variable is “TARGET” which is binary:

1 - For the people who default

0- For the people who don’t default

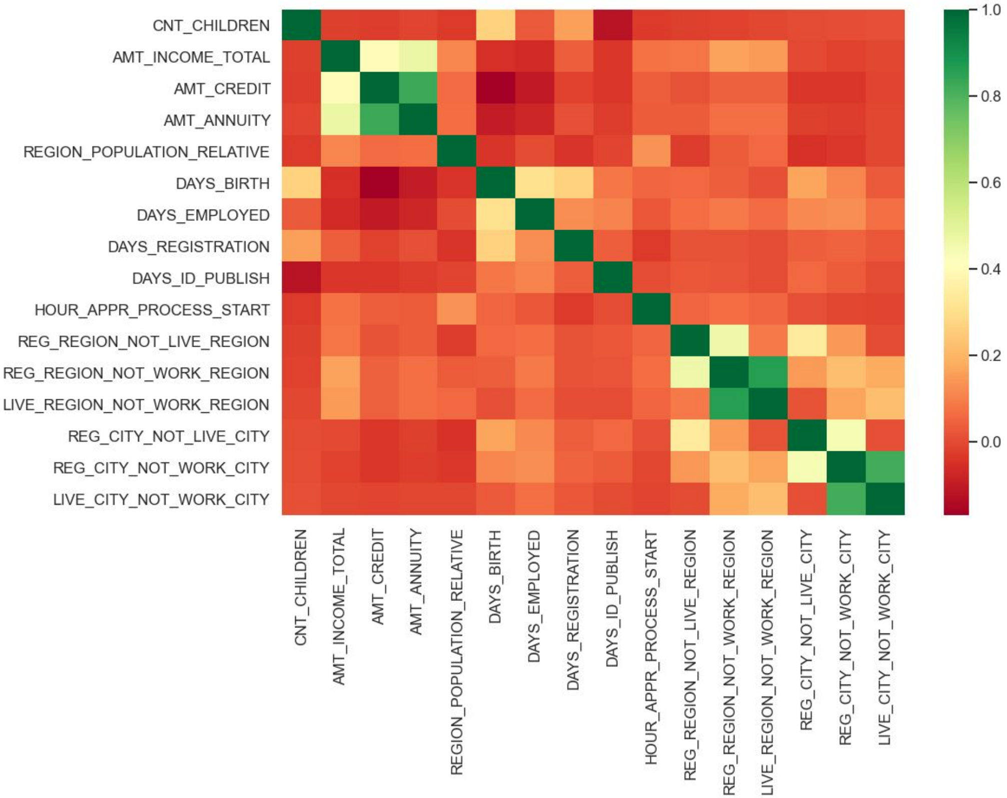
# “application_data.csv”: Target Variable

Going through the feature list we observe the target variable is “TARGET” which is binary:

1 - For the people who default

0- For the people who don’t default

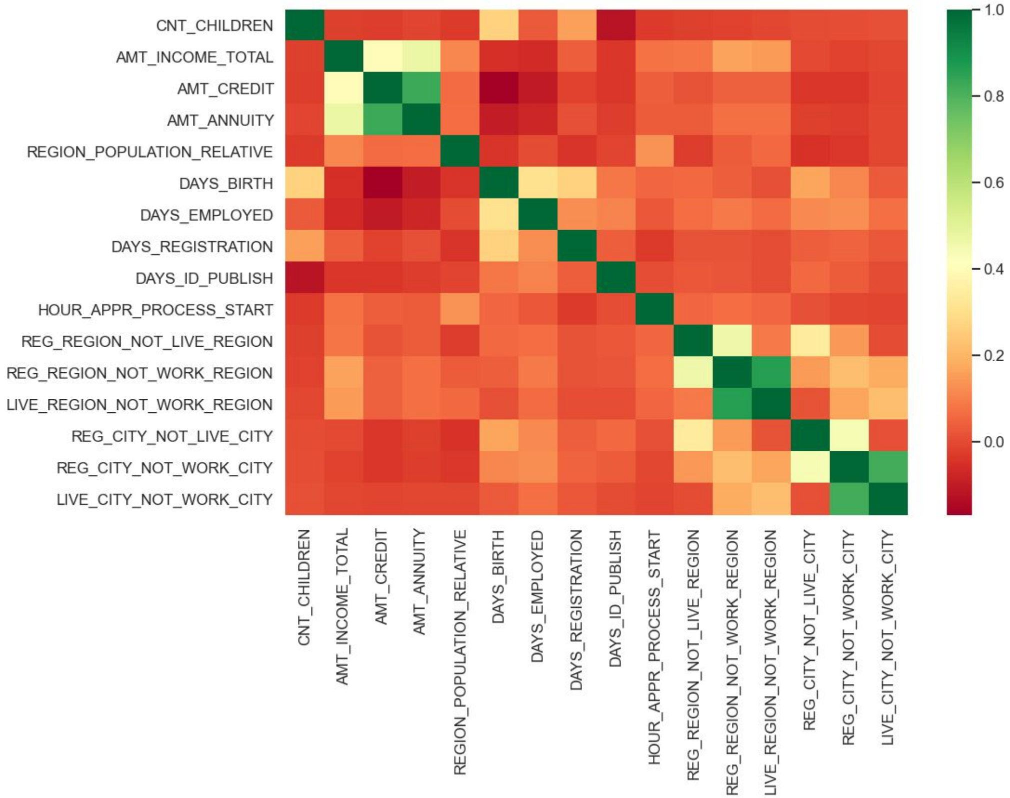
# Correlation for Target 0



Conclusions from the correlation of non defaulters:

1. Credit is positively correlated to annuity amount
2. Credit is negatively correlated to age of person
3. Credit is negatively correlated to count of children
4. Income Total is slightly positively correlated with annuity amount and credit

# Correlation for Target 1

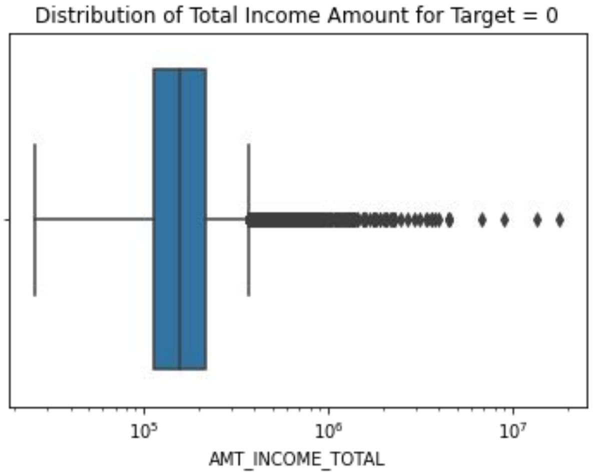


Both the correlations yield similar conclusions

# Univariate Analysis on Application Data (Numerical Variables)

A box plot for the total income amount plotted for the non defaulters, we observe that:

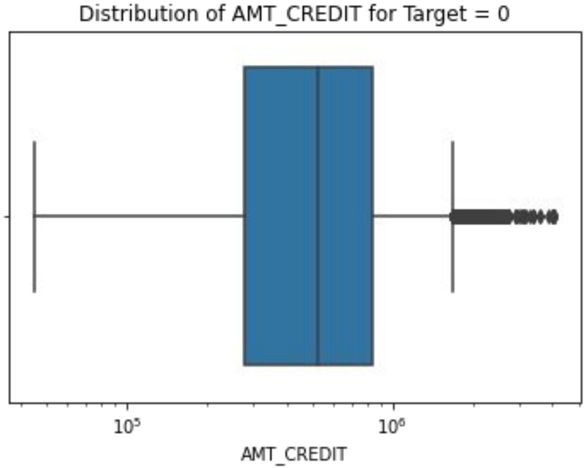
- This plot indicates that the majority of loan seekers are lower in terms of amount income
- Second quartile is smaller than first quartile



# Univariate Analysis on Application Data (Numerical Variables) cont.

A box plot for credit amount for non defaulters, here we observe that:

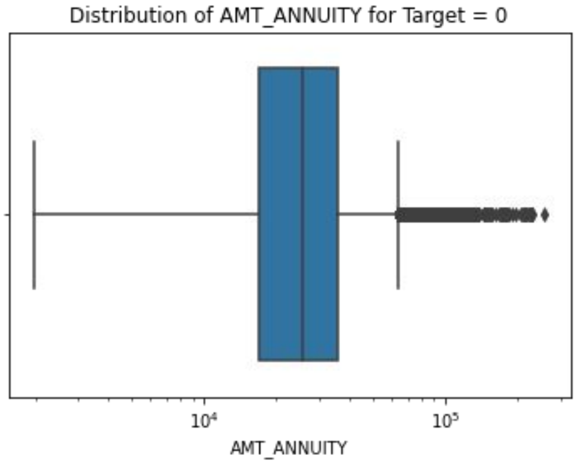
- Some outliers present
- Most credits in the first quartile
- The second quartile is larger as compared to AMT_INCOME and AMT_ANNUITY
- Slight positive skewed distribution



# Univariate Analysis on Application Data (Numerical Variables) cont.

A box plot for annuity amount for non defaulters, here we observe that:

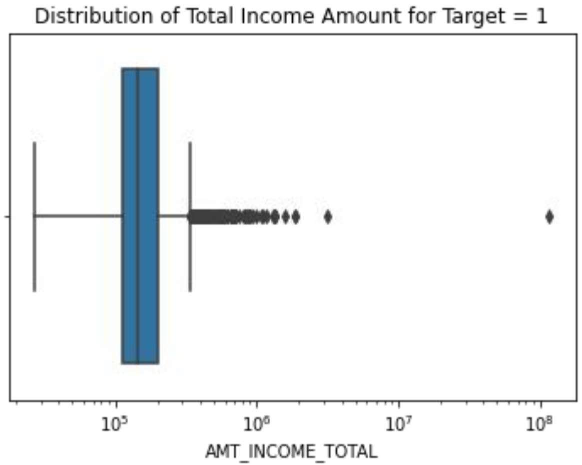
- Some outliers present
- Most clients are present in the first quartile
- The Second quartile is smaller than that of AMT_CREDIT but larger than AMT_INCOME
- Slight positive skewed distribution



# Univariate Analysis on Application Data (Numerical Variables)

A box plot for the total income amount plotted for the defaulters, we observe that:

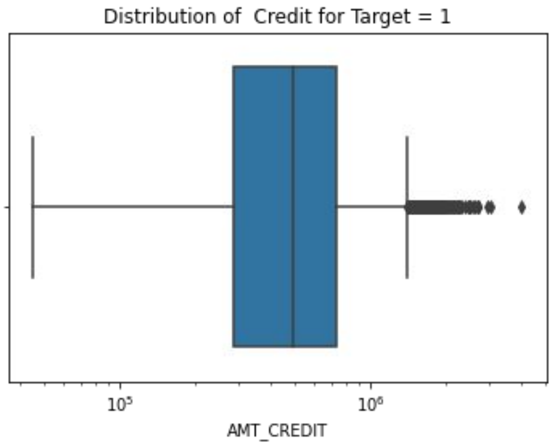
- Third quartile constitutes very few clients
- The median of this plot is lower than that of non defaulters
- This plot indicates that the majority of loan seekers are lower in terms of amount income



# Univariate Analysis on Application Data (Numerical Variables) cont.

A box plot for credit amount for defaulters, here we observe that:

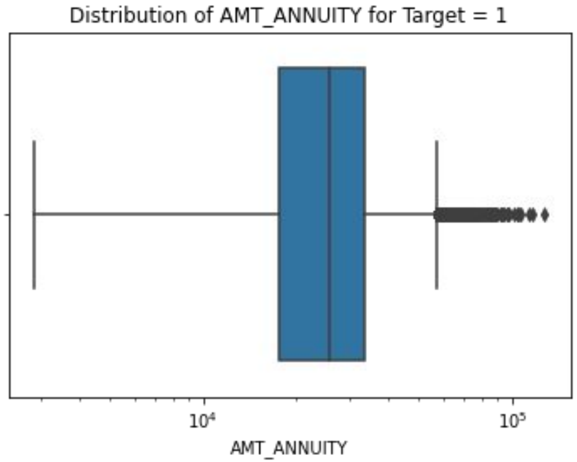
- Some outliers present
- Most credits in the first quartile
- The second quartile is larger as compared to AMT_INCOME and AMT_ANNUITY
- Almost similar to credit distribution of defaulters
- Slightly positively skewed



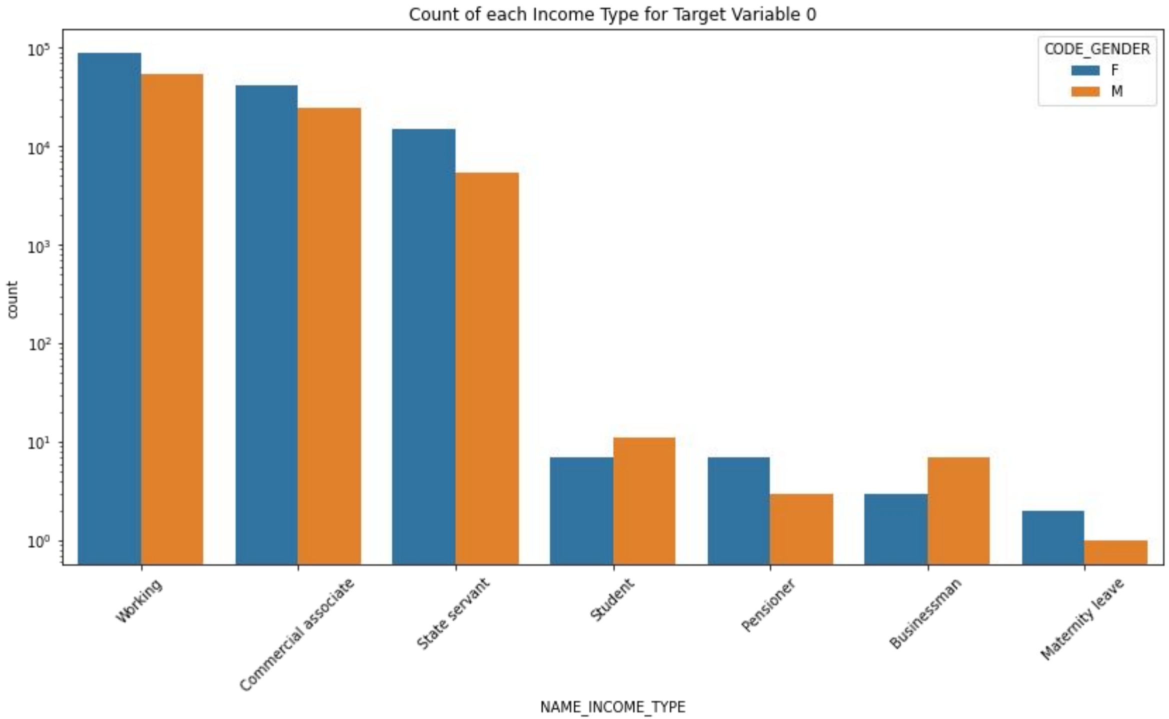
# Univariate Analysis on Application Data (Numerical Variables) cont.

A box plot for annuity amount for defaulters, here we observe that:

- Some outliers present
- The third quartile is smaller than that of AMT_CREDIT but larger than AMT_INCOME
- The amount annuity is more wide-spread as compared to that of non defaulters, i.e. lower whisker is lower and upper whisker is higher
- Positively skewed median



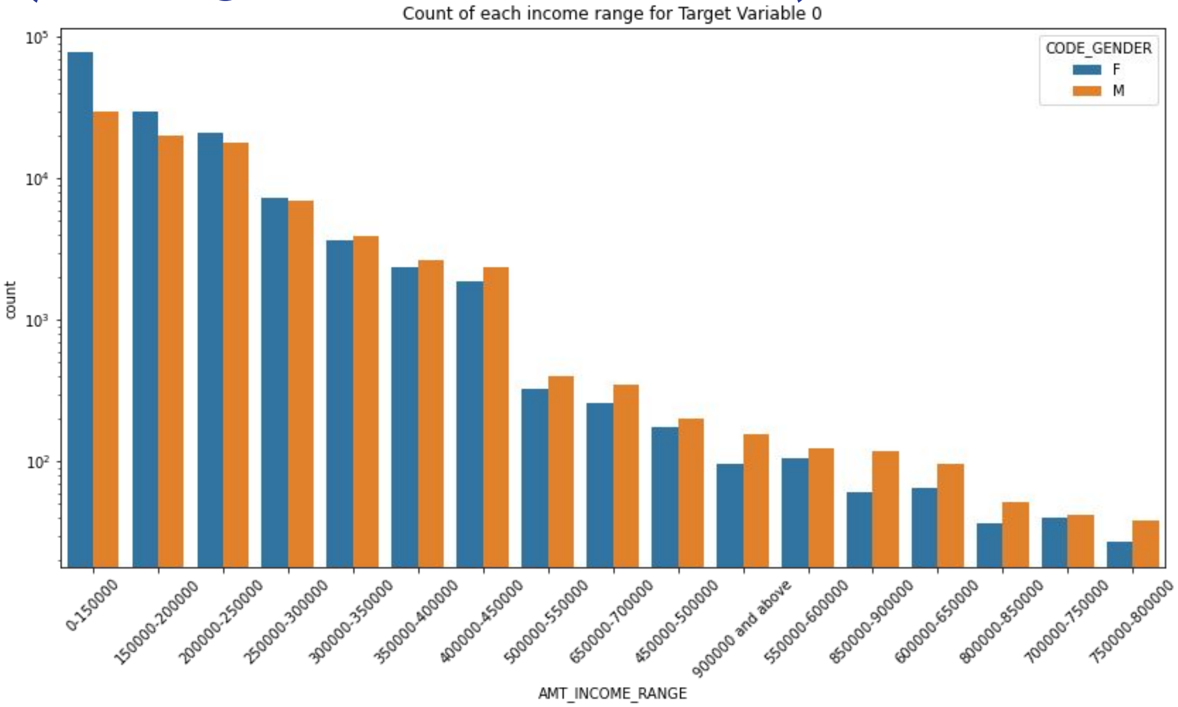
# Univariate Analysis on Application Data (Categorical Variables) cont.



Conclusions from countplot of Income type for non defaulters:

1. Working, Commercial associate, state servant categories have the most credit
2. Females have more credit than males for the above cases
3. Males have more credit for Student and Businessman categories
4. Student, Pensioner, Businessman have lower credit
5. Maternity leave has the least credit

# Univariate Analysis on Application Data (Categorical Variables) cont.



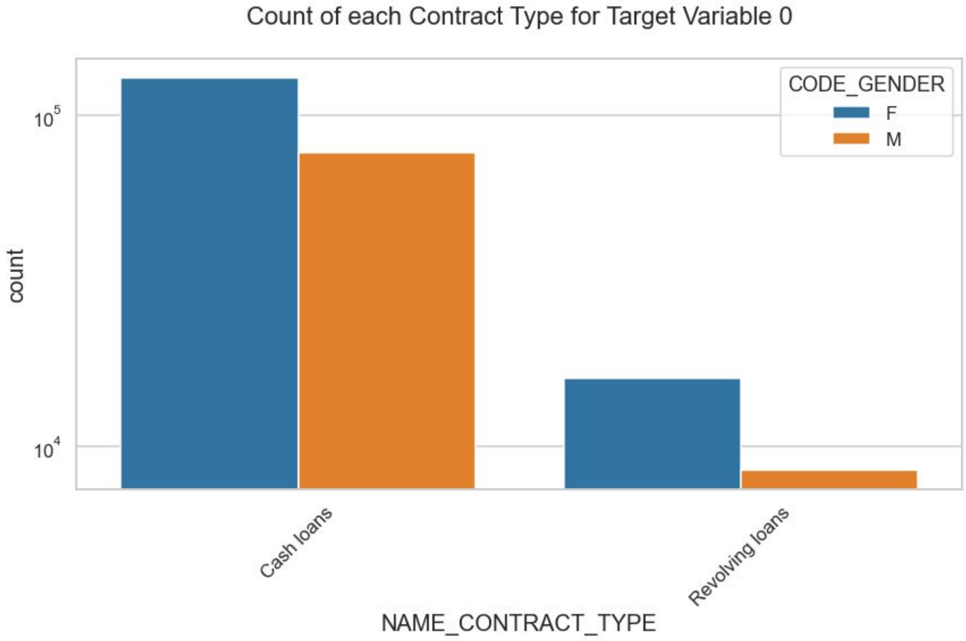
Conclusions from countplot of income range for non defaulters:

1. Females in general have more credit than males

2. Most people with income range 0-150000

3. Less counts for income range 450000 and above

# Univariate Analysis on Application Data (Categorical Variables) cont.

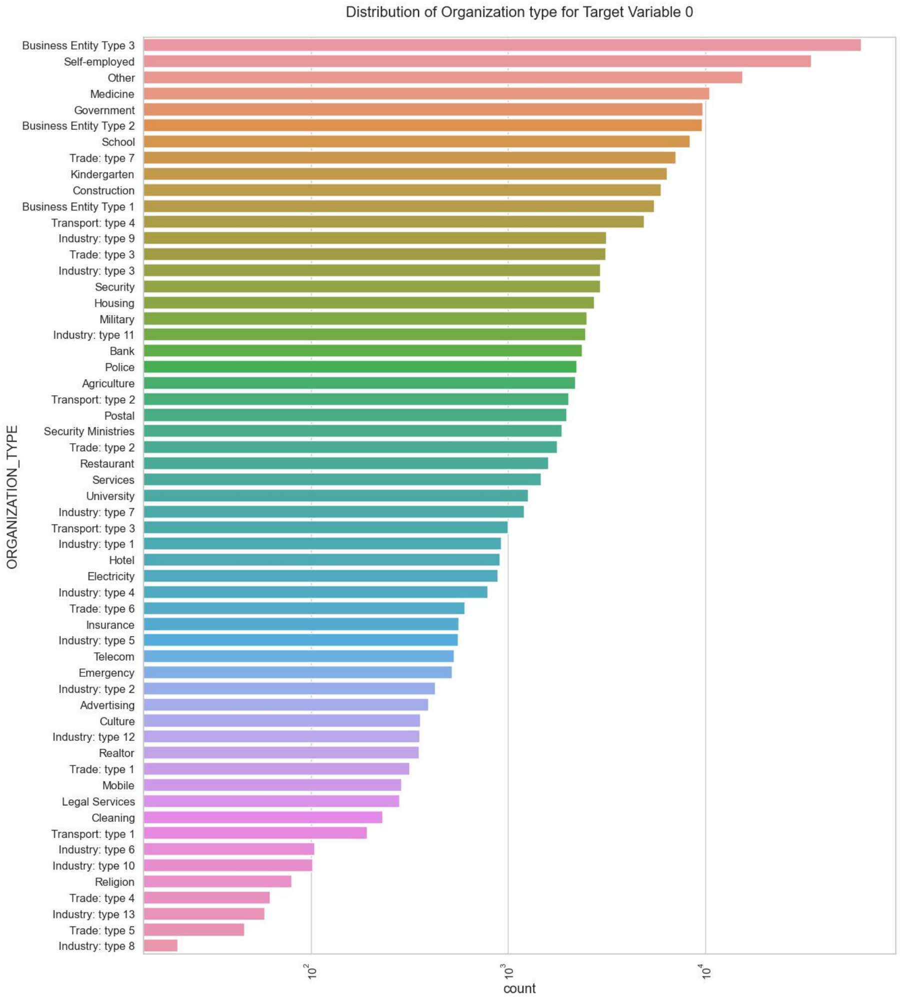


Conclusions from countplot of contract type non defaulters:

1. Cash Loans has more credits than Revolving loans

2. Males have negligible credit in case of revolving loans

3. Both cases Females are more who applied in case of Not Defaulting



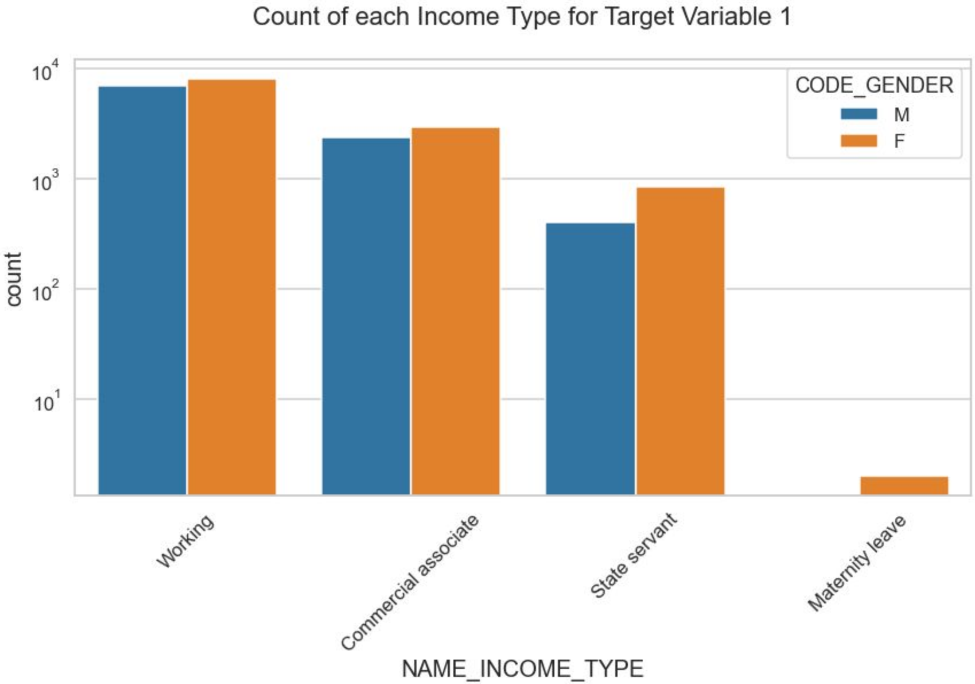
# Univariate Analysis on Application Data (Categorical Variables) cont.

Conclusions from count plot of Organization Type for non defaulters plot:

1. 'Business Entity Type 3', 'Self-Employed', and 'Other' are constitute top 3

2. Less clients from type 8, type 5 and type 13.

# Univariate Analysis on Application Data (Categorical Variables) cont.



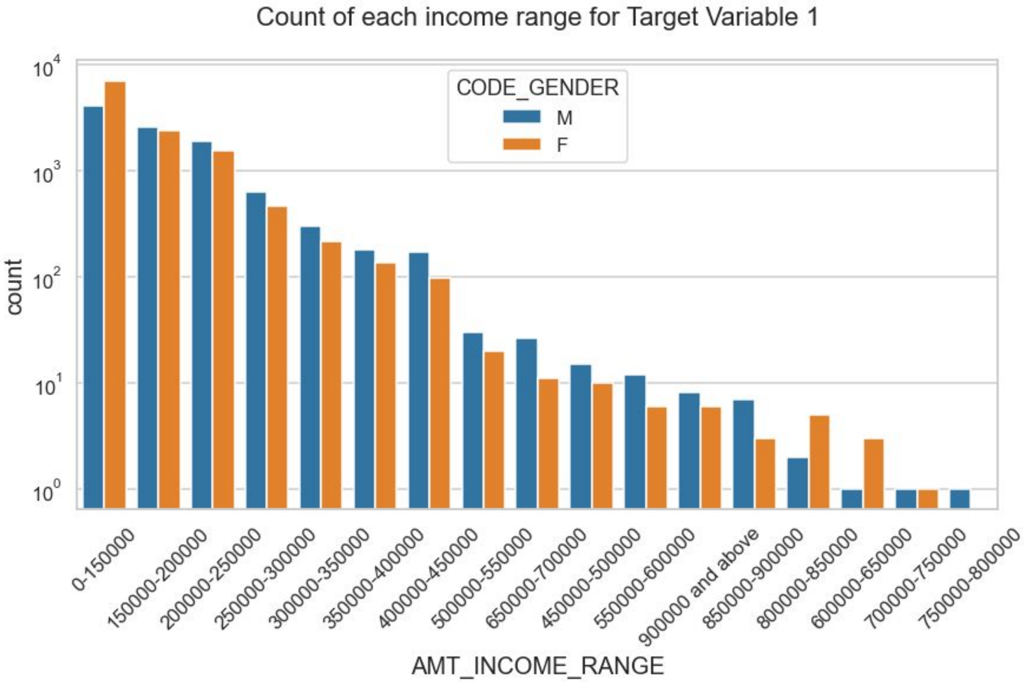
Conclusions from countplot of Income type for defaulters:

1. Working have the most credit

2. Students, Businessmen and pensioners aren't present

3. There are more females across the board

# Univariate Analysis on Application Data (Categorical Variables) cont.



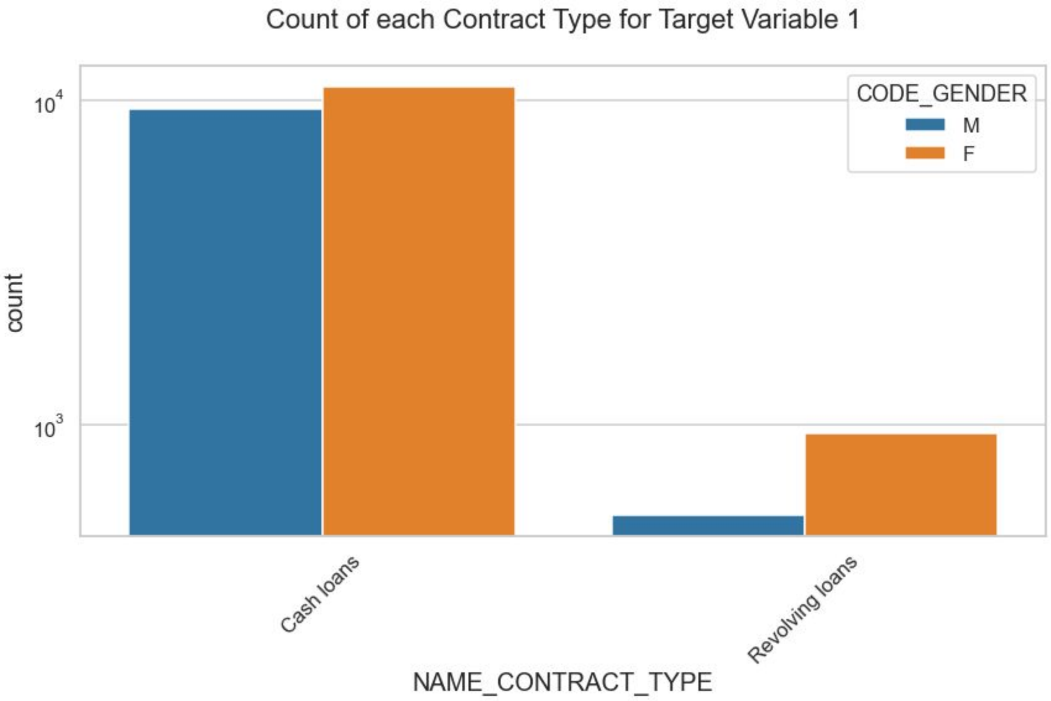
Conclusions from countplot of income range for defaulters:

1. Males in general have more credit than Females here

2. Most people with income range 0-150000

3. Less people with income range 600000 and above

# Univariate Analysis on Application Data (Categorical Variables) cont.



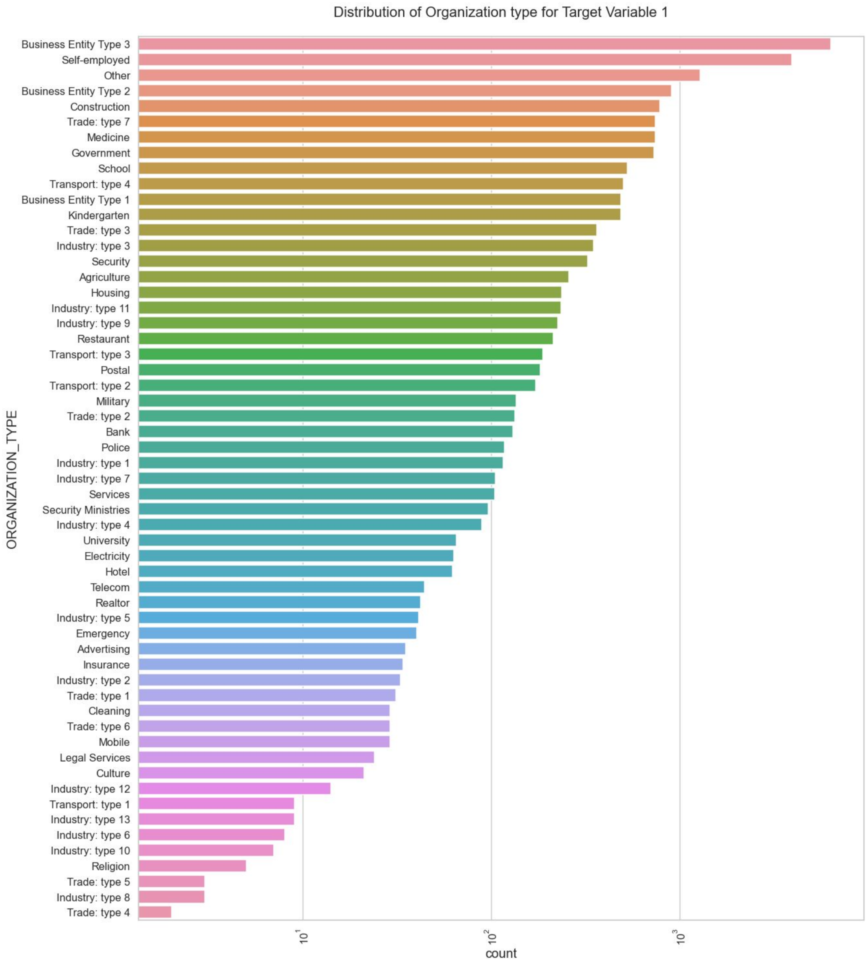
Conclusions from countplot of contract type defaulters:

1. Cash Loans has more credits than Revolving loans

2. Males are negligible in case of Revolving loans

3. Both cases Females are more who applied in case of Defaulting

# Univariate Analysis on Application Data (Categorical Variables) cont.

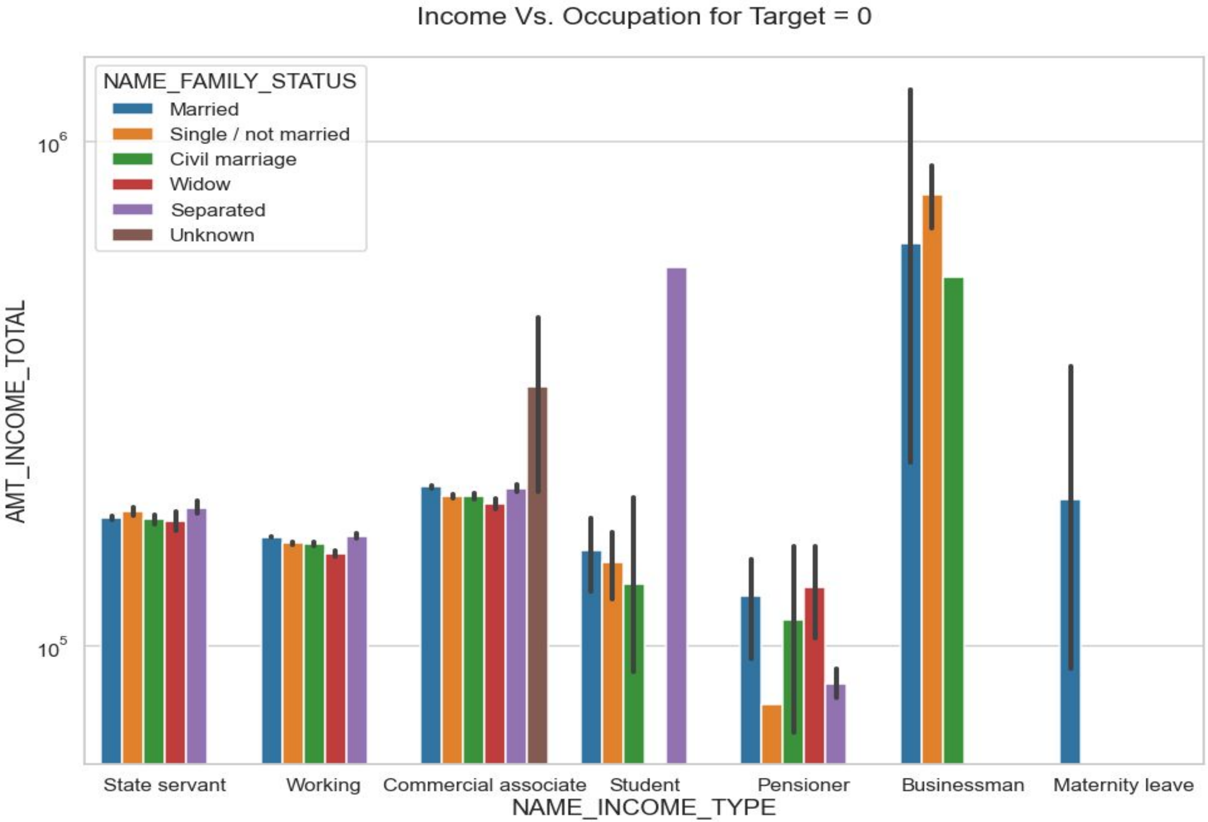


Conclusions from count plot of Organization Type for defaulters plot:

1. 'Business Entity Type 3', 'Self-Employed', and 'Other' are constitute top 3

2. Less client from type 8, type 5 and type 4 and religion

# Bivariate Analysis on Application Data



Conclusion from barplot for Income Vs. Occupation of non defaulters:

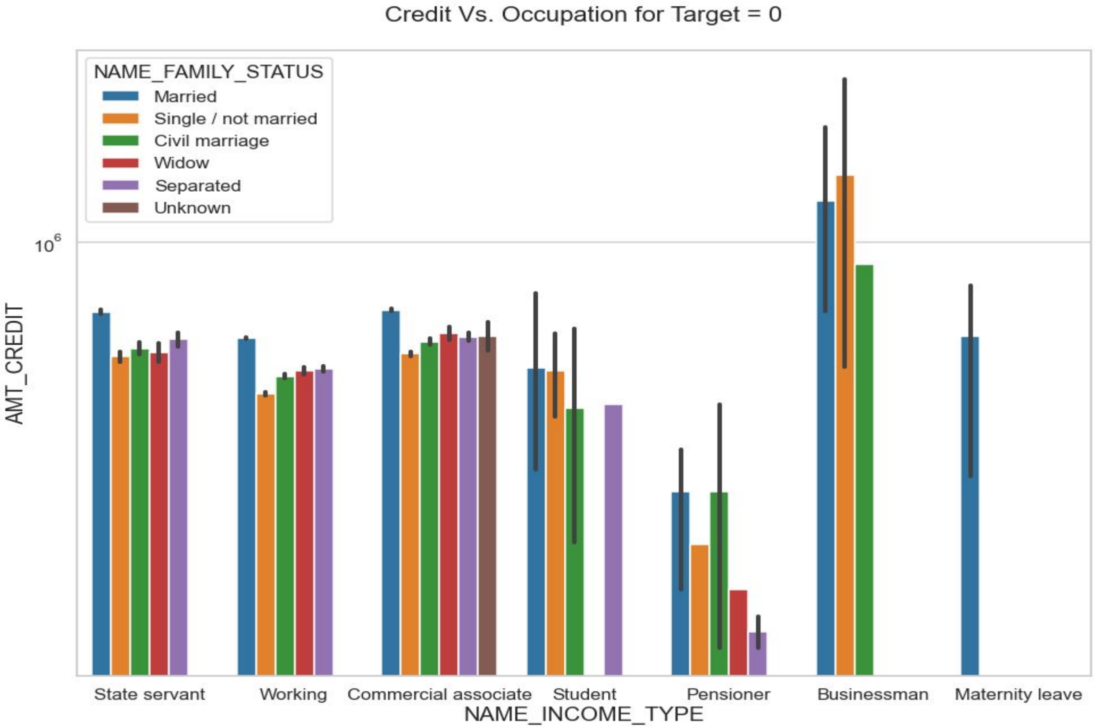
1. In most cases of occupation Married category has most income

2. Businessmen have the most income total

3. Lot of missing data regarding family status for commercial associate

4. Pensioners have the least income

# Bivariate Analysis on Application Data cont.



Conclusion from barplot for Credit Vs. Occupation of non defaulters:

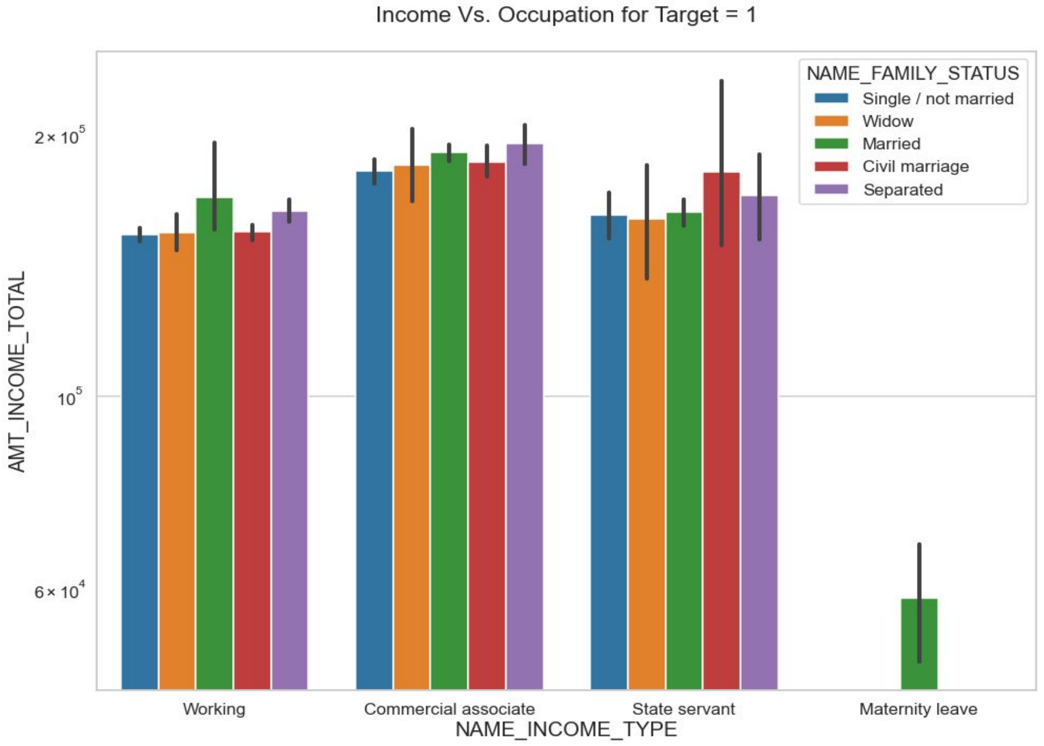
1. In most cases of occupation Married category has most credit

2. Businessmen have the most credit followed by Commercial associate and state servant

3. Lot of missing data regarding family status for commercial associate

4. Pensioners have the least credit

# Bivariate Analysis on Application Data cont.



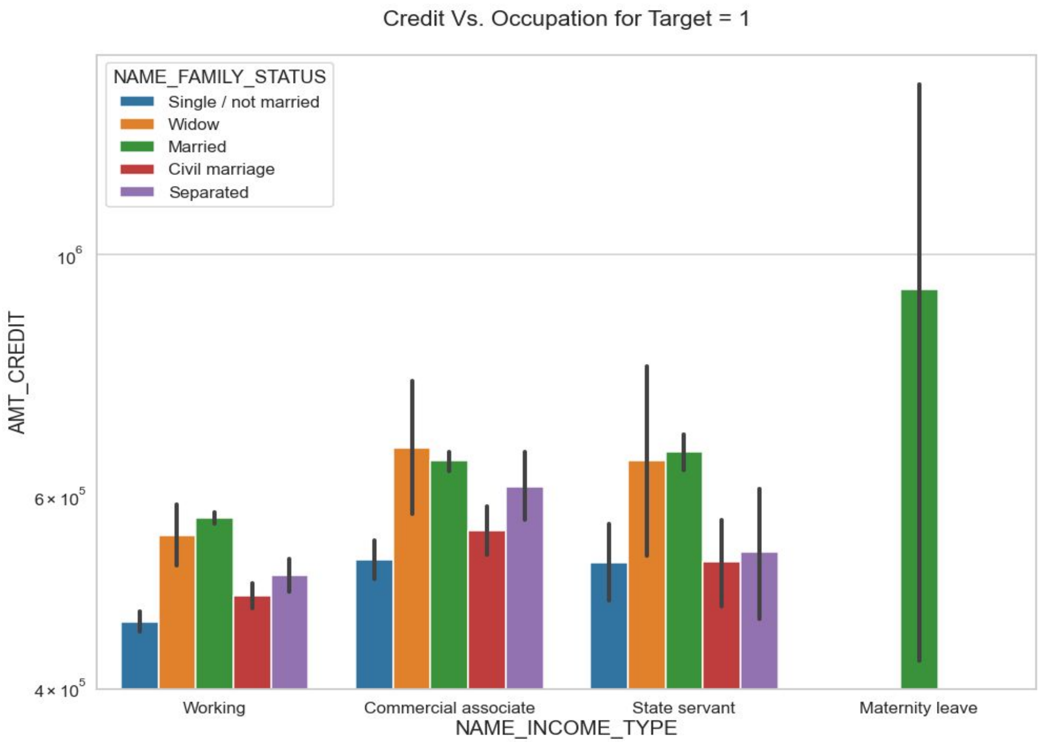
Conclusion from barplot for Income Vs. Occupation of defaulters:

1. There are no students, businessmen and pensioner

2. Married have most income across majority of family status

3. Commercial associate is the most earning group

# Bivariate Analysis on Application Data cont.



Conclusion from barplot for Credit Vs. Occupation of defaulters:

1. Maternity leave category has most credit in case of being defaulter

2. Working have least in credit in case of being defaulter

# Conclusion of Application Data

From all the visualizations we can conclude that:

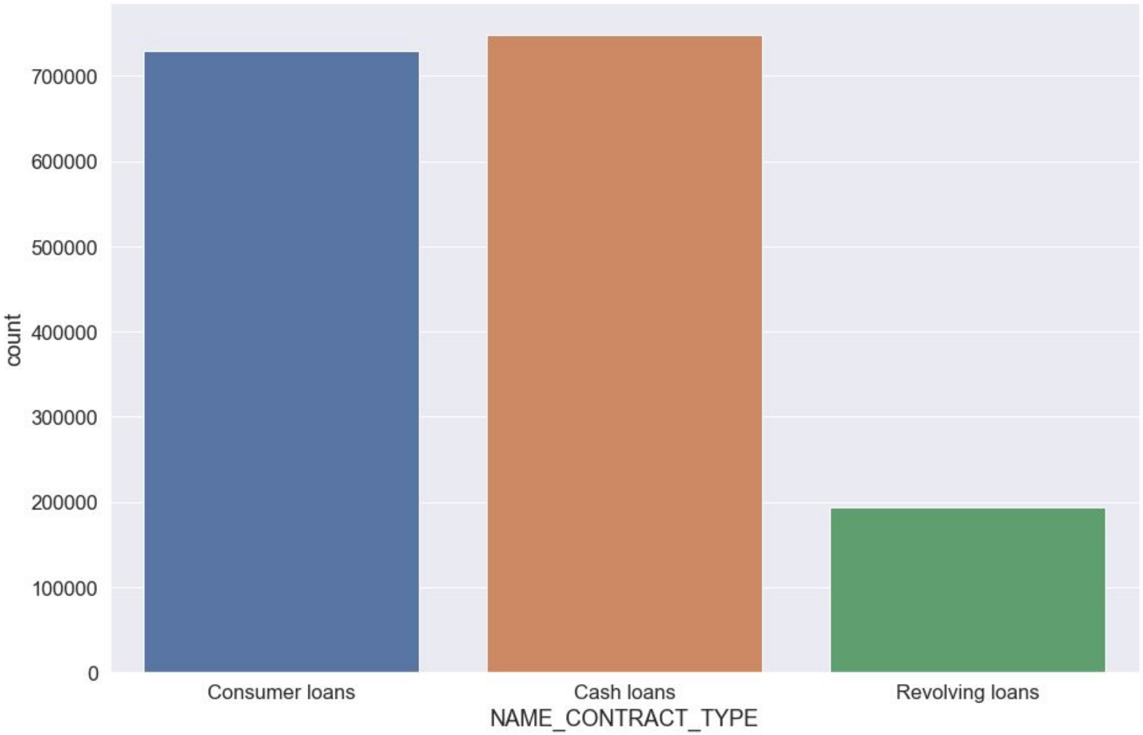
- Banks should trust the following income types:
  - Businessmen
  - Pensioners
  - Students
- Working income type has least successful payments
- The Commercial associates don’t tend to share family details
- Males tend to have more defaulters, hence rigorous credit checks must be put to place for the particular category

# Previous Applications CSV file

The file has data related to previous applications of the same customers

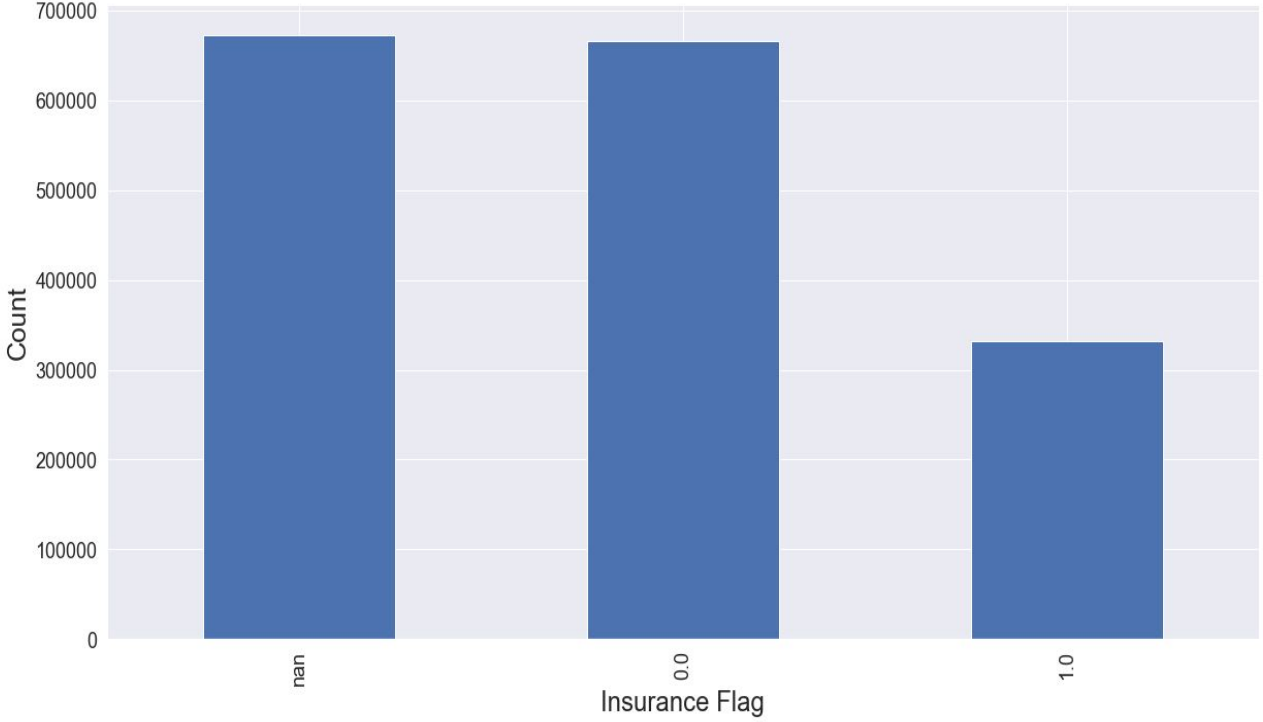
EDA was also performed on previous_applications.csv file and the following observations were made

# What type of loans are preferred?



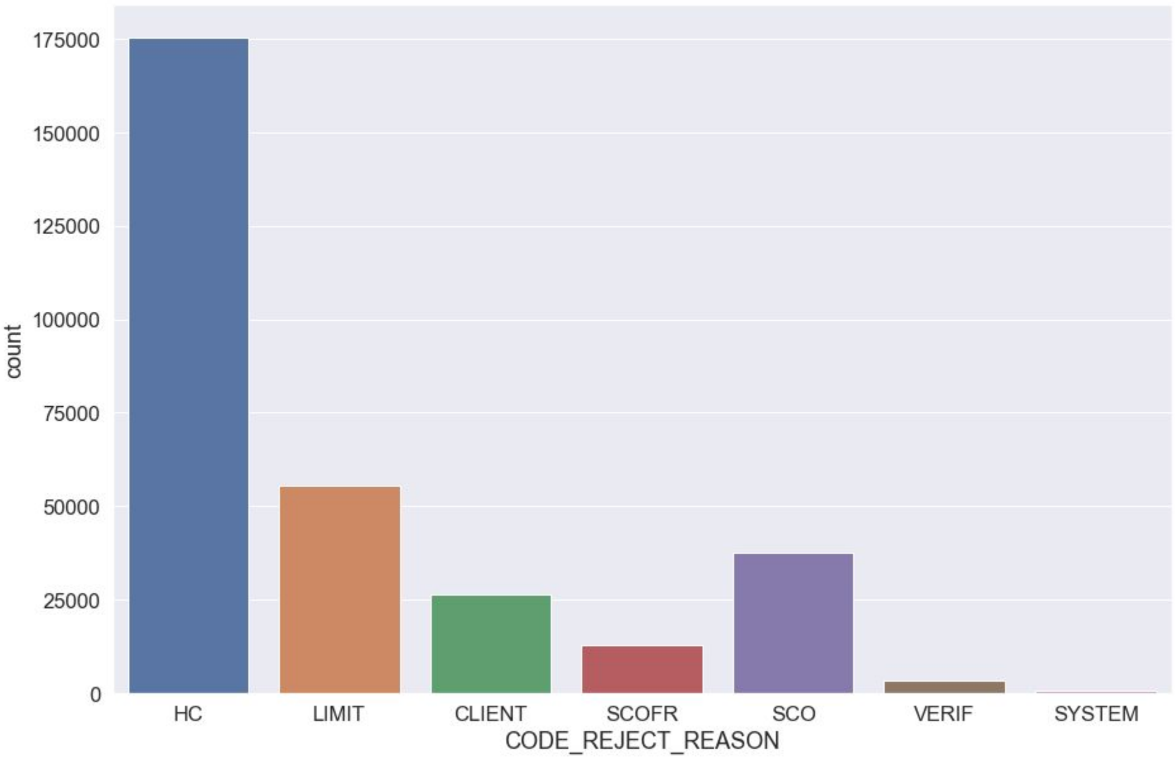
Customers do not prefer ‘Revolving loans’. This can be seen from the countplot. Revolving loans could be attributed to a particular purpose/business. But no such data was found to support this theory

# Do customers prefer Insurance?



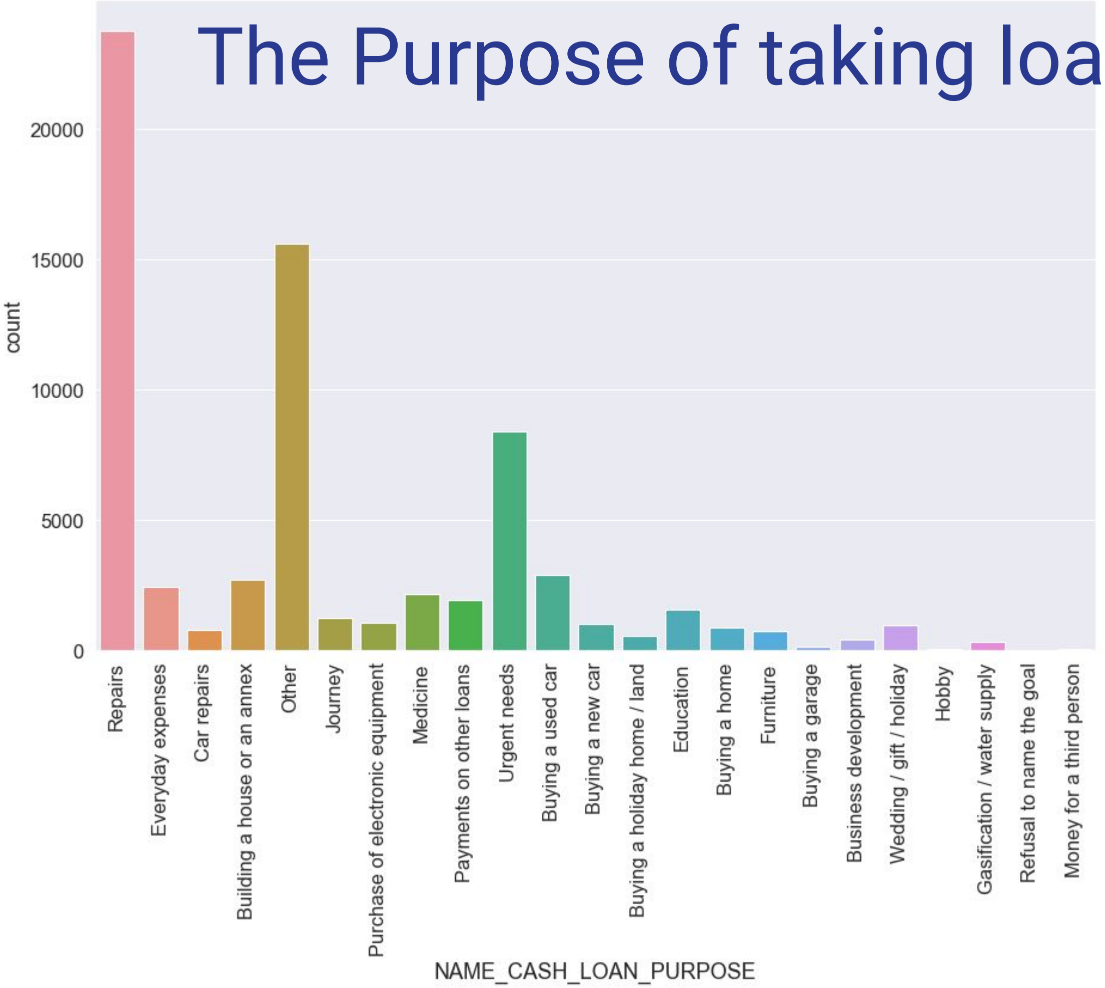
It is observed that Customers do not prefer Insurance from this countplot. However, there are too many 'NULL values' (shown as **nan** in the plot) in the data. This makes the observation probabilistic (not deterministic) because in reality the customers who did not give details may later decide to get insurance

# Why does a loan get rejected?



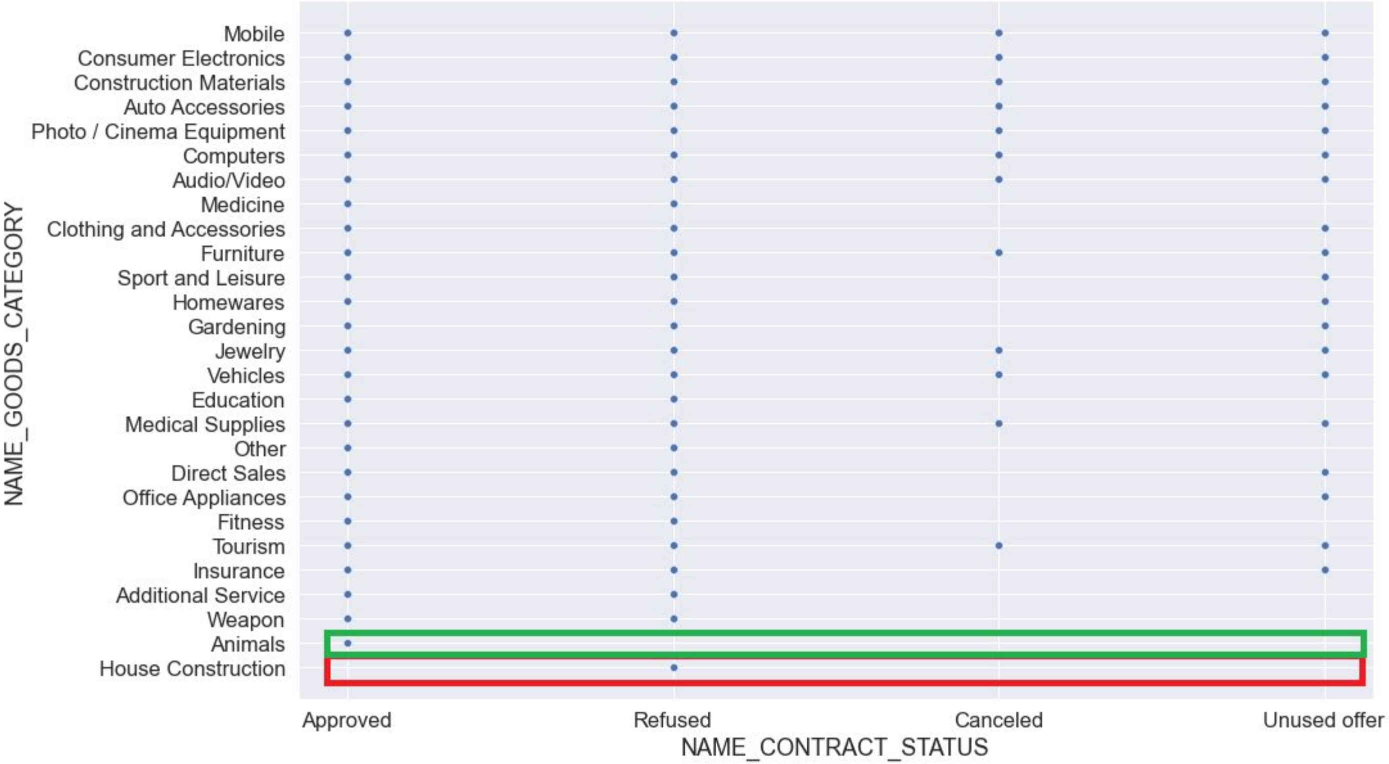
The most common reason was found to be **XAP**. However. Assuming that it is a form of NULL value, the next highest cause of loan rejection is **HC** which can be seen from the countplot. The margin by which **HC** dominates the next reason **LIMIT** is large enough.

# The Purpose of taking loan?



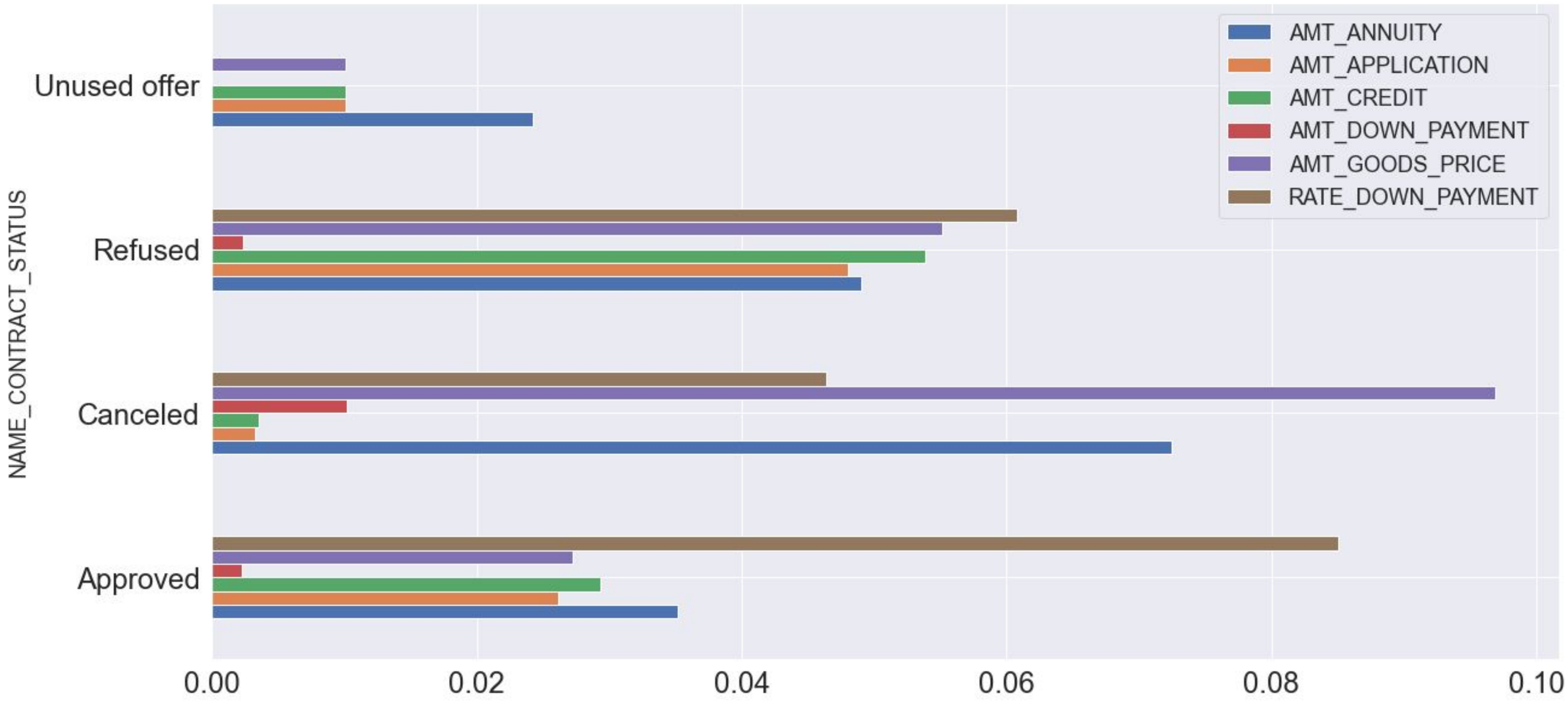
The most common purpose of taking loans is **Repairs** and **Urgent Needs**. It can also be seen that **others** is also used as a common reason. It is possible that there is a frequent reason that is not present in the bank form. If so, the bank needs to find out and make suitable schemes for that purpose. Else, if **others** means miscellaneous, then it is alright

# Goods category and Loan Approval



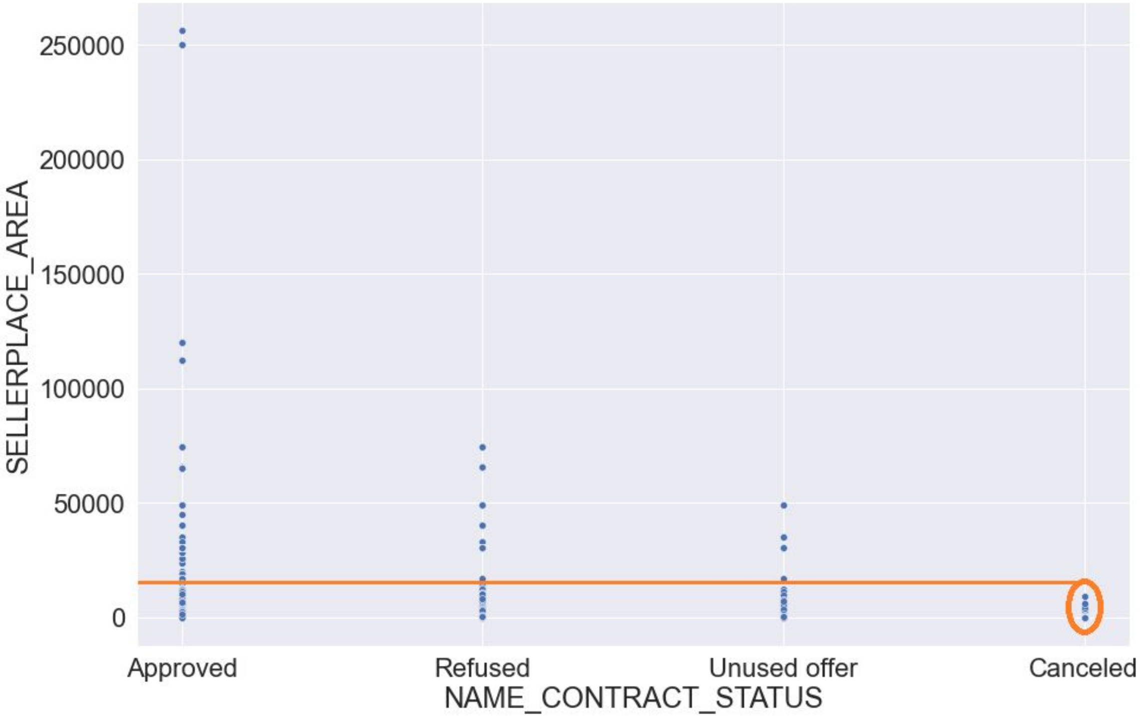
Loans for **House construction** were always refused as highlighted red and loans for **Animals** maintenance were always approved. This could be due to a government schemes that encourage loans on animals in rural areas

# For what price loan gets cancelled?



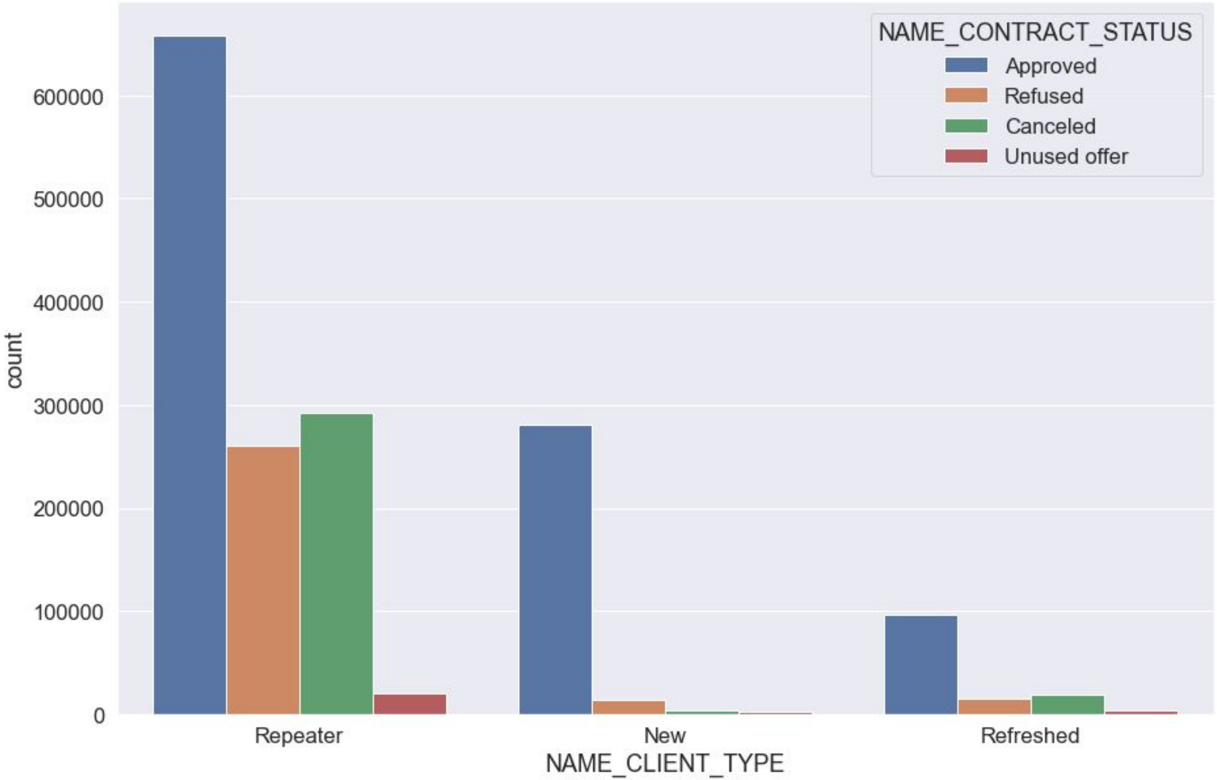
Customers getting **very low amount of credit** (indicated in Green) for **very high Goods Price** (indicated in Purple) tend to cancel their loan. Banks tend to refuse loan when credit amount is high.

# Effect of Sellerplace area



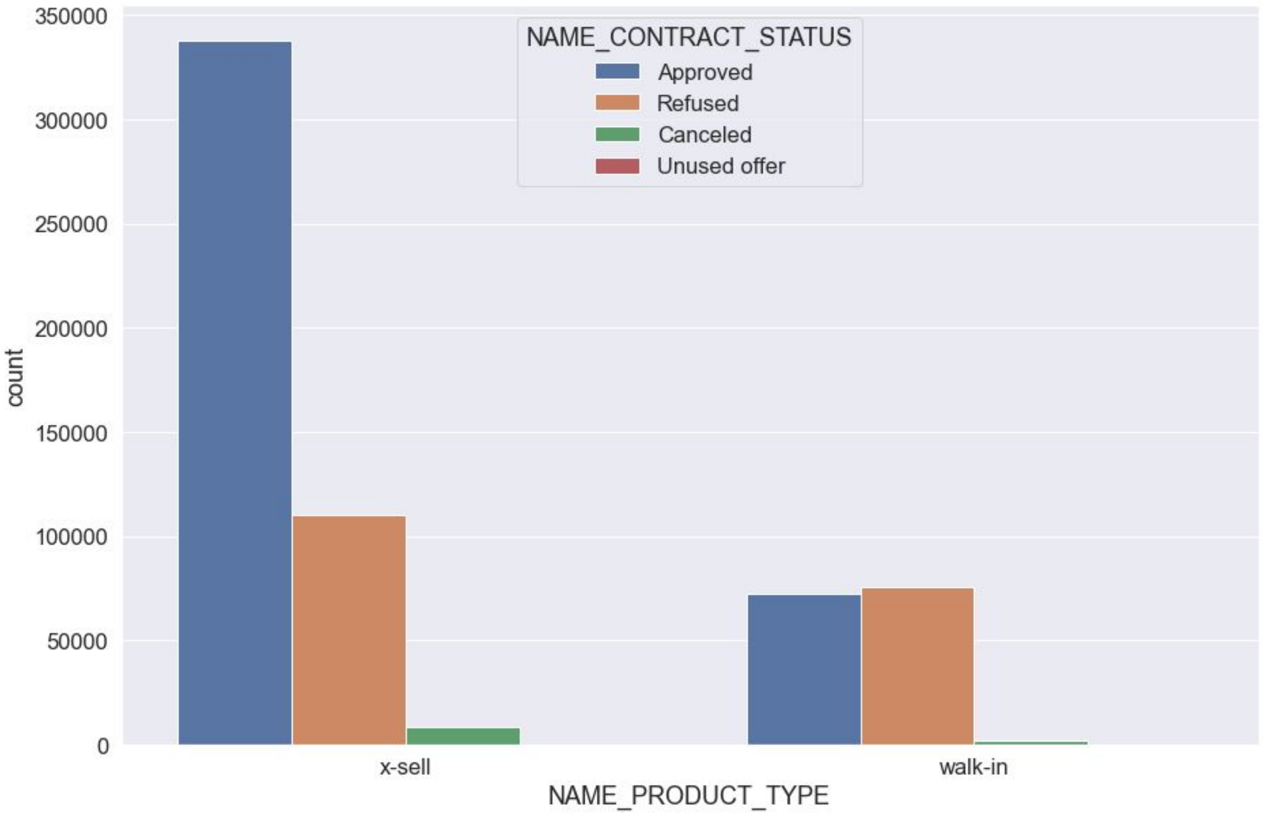
In this plot, 5 banks with very large sizes were excluded, to see the plot properly. From this plot it can be inferred that Customers tend to cancel their loans only in smaller sized banks.

# Are repeater customers profitable?



Repeating customers may not be profitable. It can be seen that they have high approval rate but also have significant cancellation and refusal rate. Instead, **new customers** have good approval rates and almost no cancellation and refusal rates. So, they are more profitable to us

# Product type?



X-sell is more beneficial because they have a better approval rate for the cancellations that happen, when compared to a walk-in customer

# The overall conclusions

- Customers do not prefer **Revolving loans** and do not prefer to get Insurance on their loans
- Financial institutions have rejected many loans on the basis of criteria **HC**
- Many customers take loans for **Repairs** and **Urgent needs**
- Customers do not cancel their loans in larger financial institutions
- Customers cancel their loans if the ratio of **Goods amount** to **Credit amount** is high
- **New customers** are more profitable than repeatable customers
- **X-sell** is better than a Walk-In customer