# EDA Case study

Aditya Anand (adityaanand2701@gmail.com)

Amol Kumar (amol.mee@gmail.com)

# Agenda

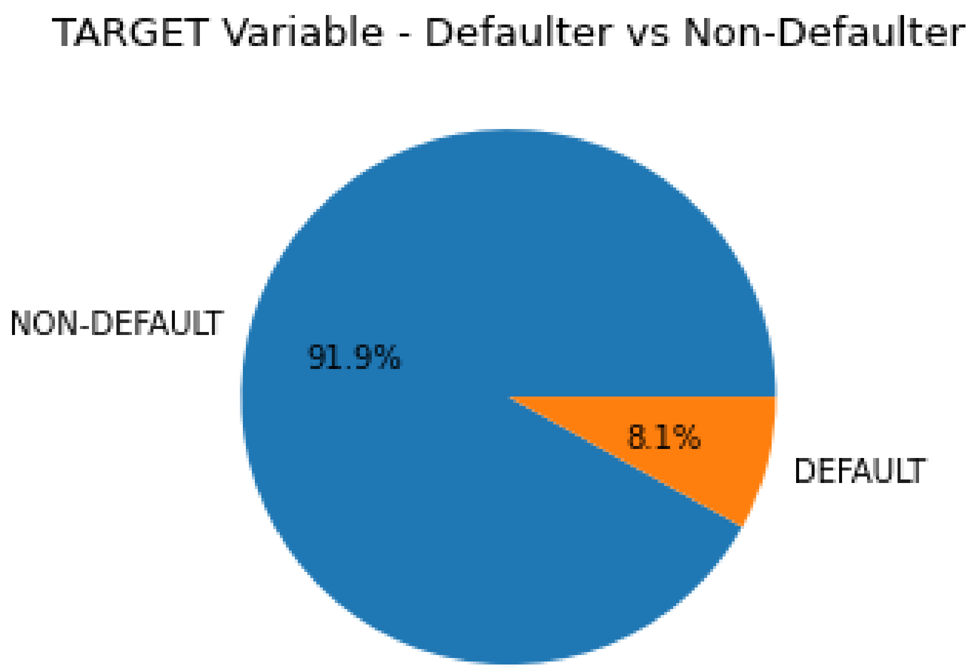
- Brief Description of Cleaning & preparing data – App_data
- Uni-Variate & Bi-Variate Analysis of App_data
- Brief Description of Cleaning & preparing data – Prev_app
- Uni-Variate & Bi-Variate Analysis of Prev_app
- Merging two DataFrame and getting insights
- Final Insights

# Brief Description of Cleaning & preparing data – App_data

1. Dropped all columns which has more than 40% null values
2. Impute the missing values in remaining column
    1. Replaced the null values with mode in case of categorical column
    2. Replaced the null values with mean in case of numerical column with no outliers (outliers are identified using box plot)
    3. Replaced the null values with median in case of numerical column with outliers (outliers are identified using box plot)
3. Dropped the rows which has XNA values in Gender
4. Converted all DAYS related column to Year
5. Created Bins for Income and Age
6. Splitted the dataframe based on ‘TARGET’ column

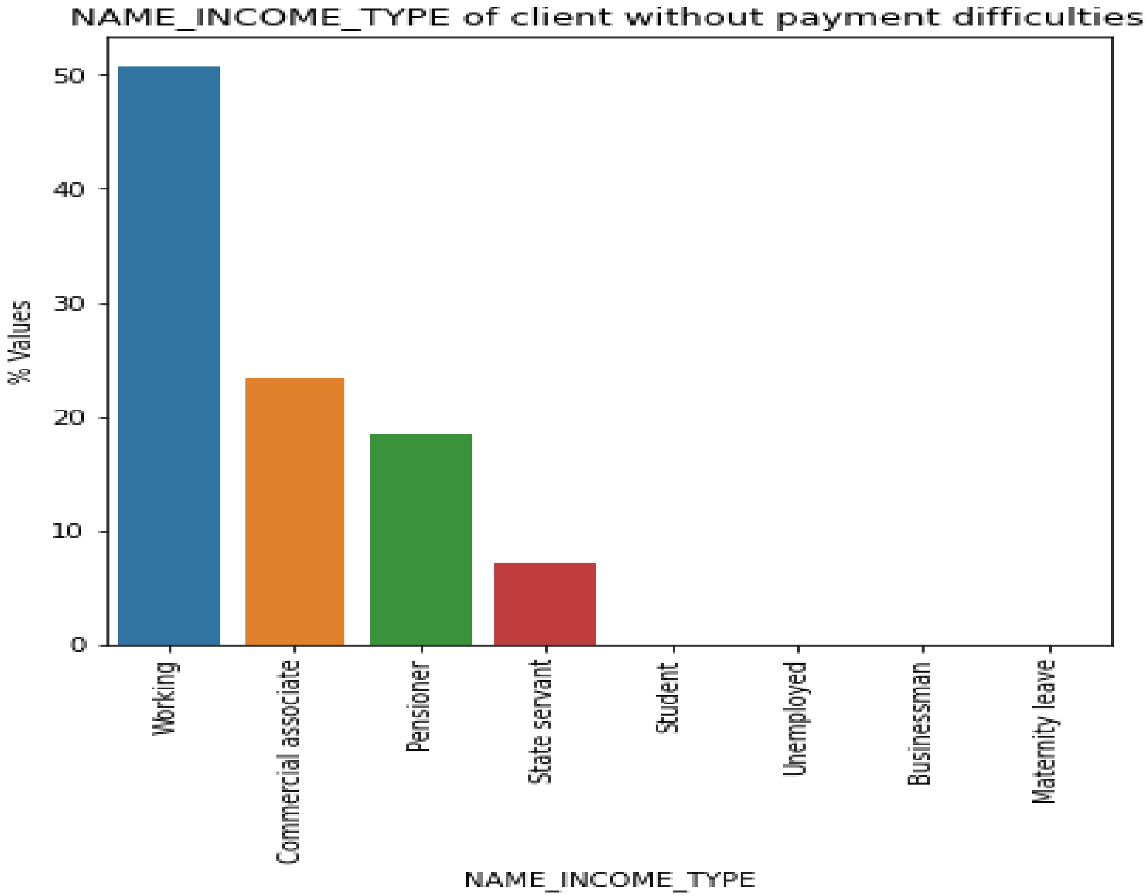
# Uni-Variate & Bi-Variate Analysis of App_data

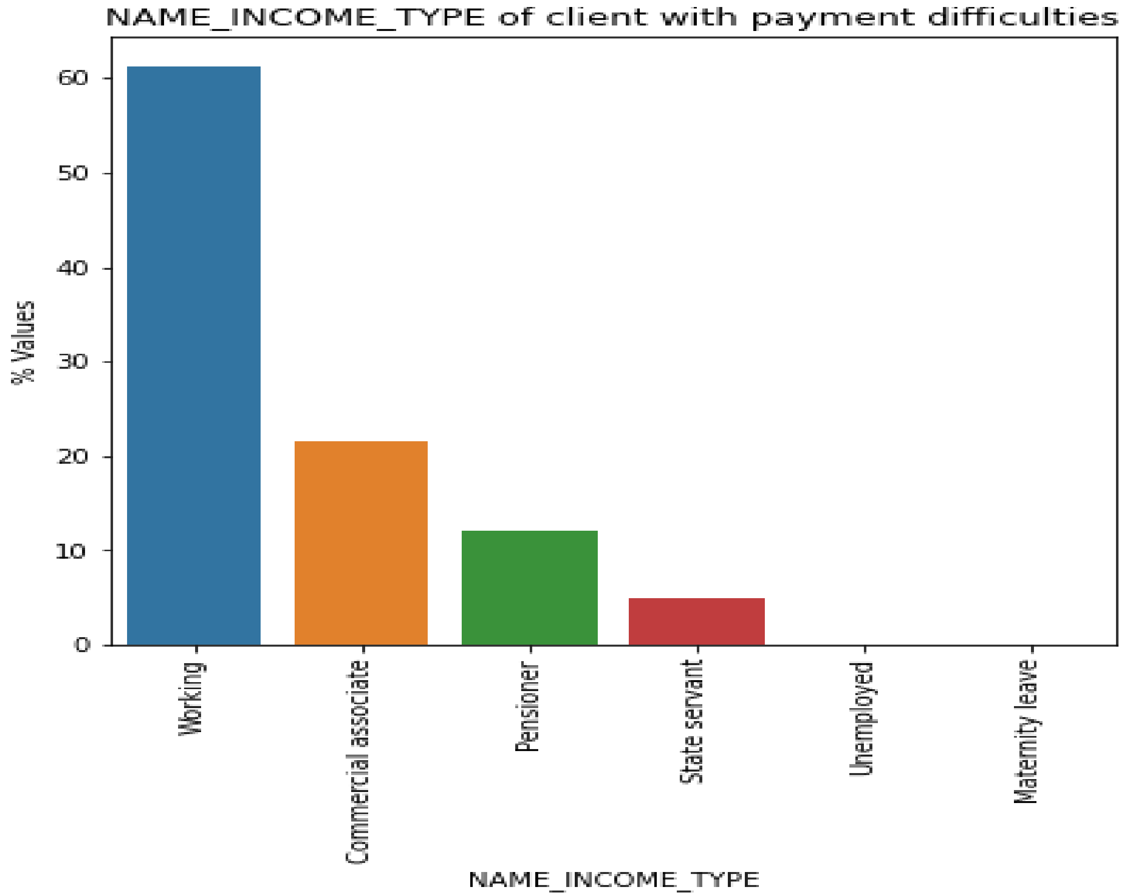
| Checking Imbalance in data | **Observation:**<br>Approx. 92 % of clients are non defaulter |
|---|---|



# Uni-Variate & Bi-Variate Analysis of App_data

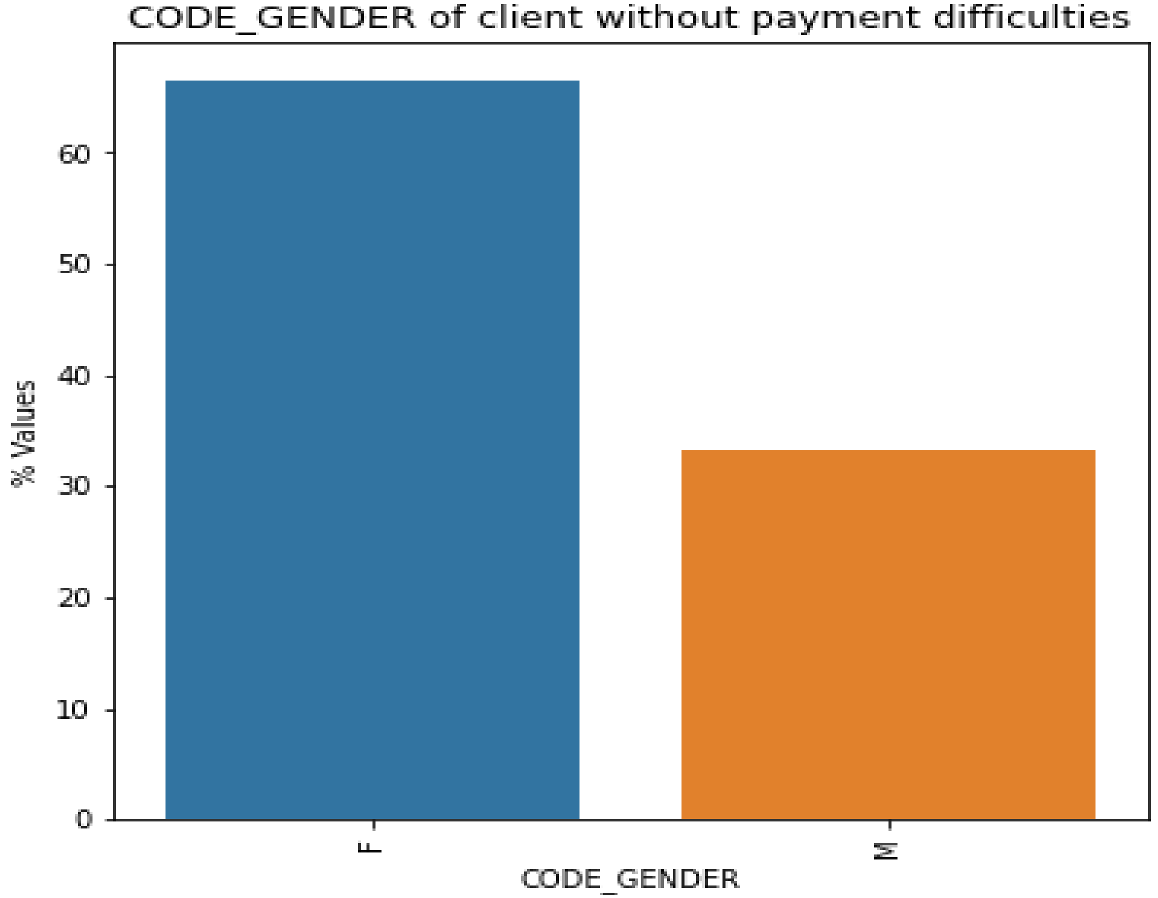
| **Checking Income Type of different target group** | **Observation:**<br>Students & Businessman never default<br>Working class clients are more in % with payment difficulties as compared to non payment difficulties so chance of defaulting is more<br>Pensioners clients chance of defaulting is less |
|---|---|



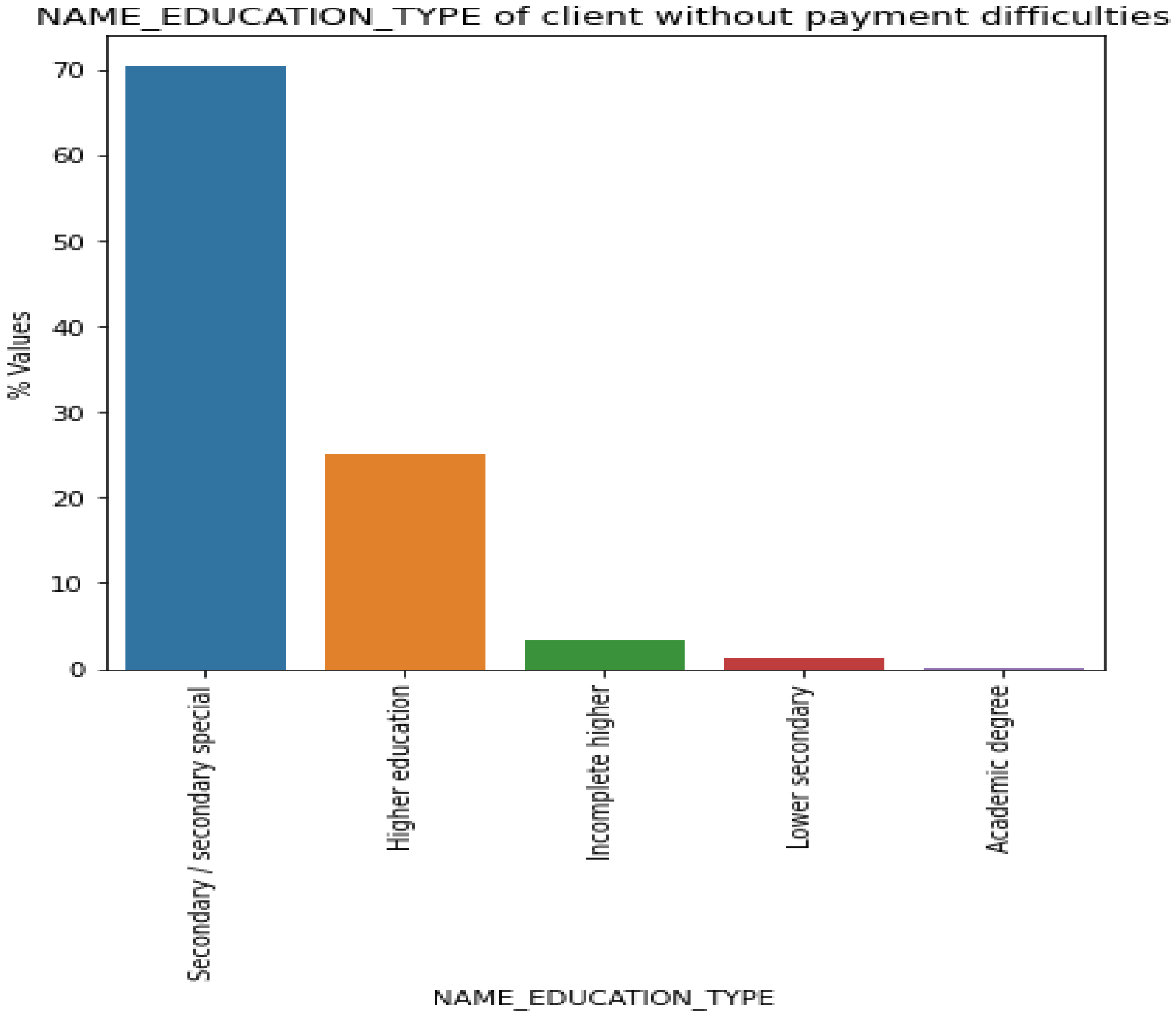
# Uni-Variate & Bi-Variate Analysis of App_data

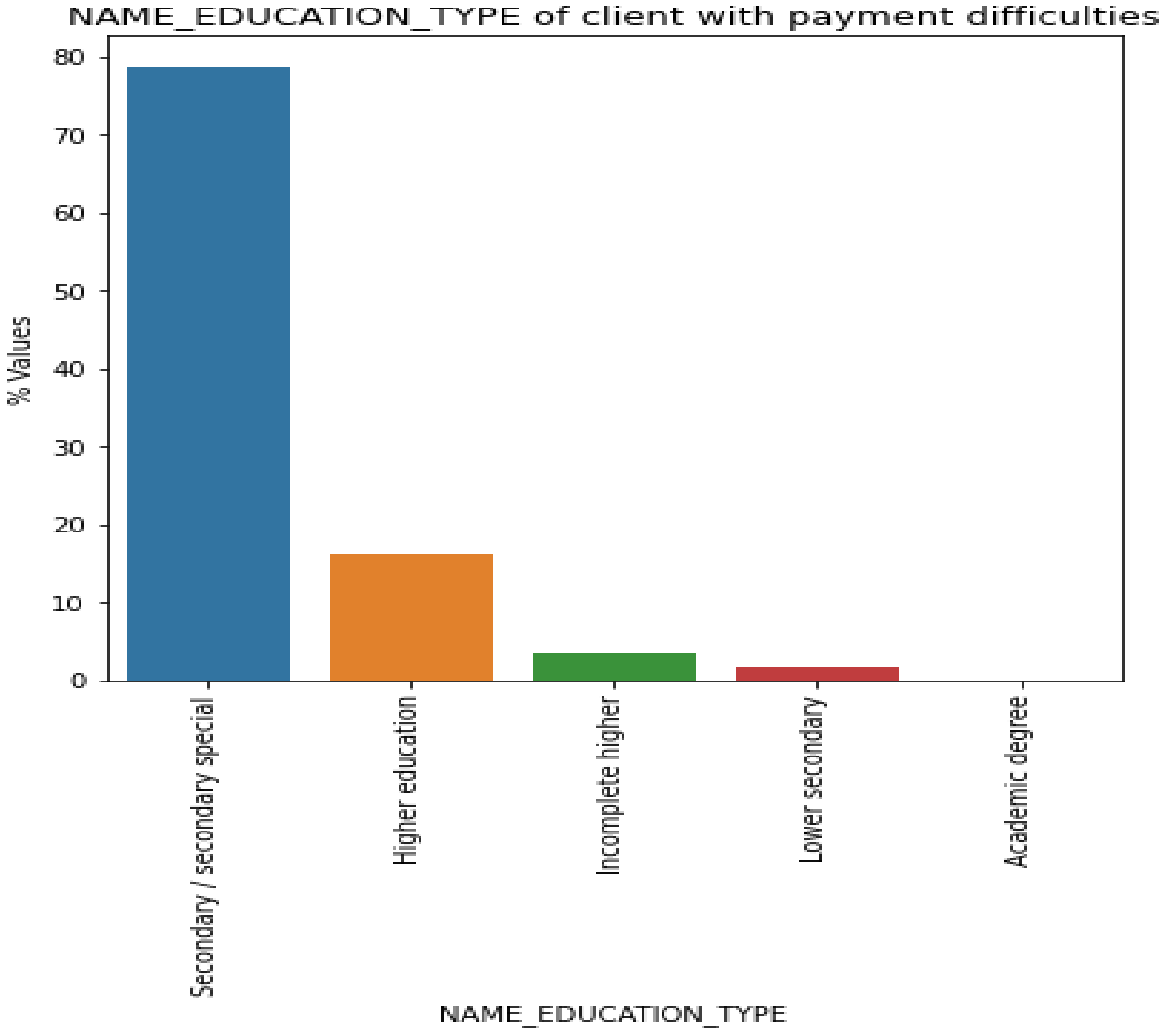
| | |
|---|---|
| **Checking Gender of different target group** | **Observation:**<br>Female applied for loan more than male<br>Increase in % of payment difficulties for male client and decrease in payment difficulties for female client so chances for male to default is more |




# Uni-Variate & Bi-Variate Analysis of App_data

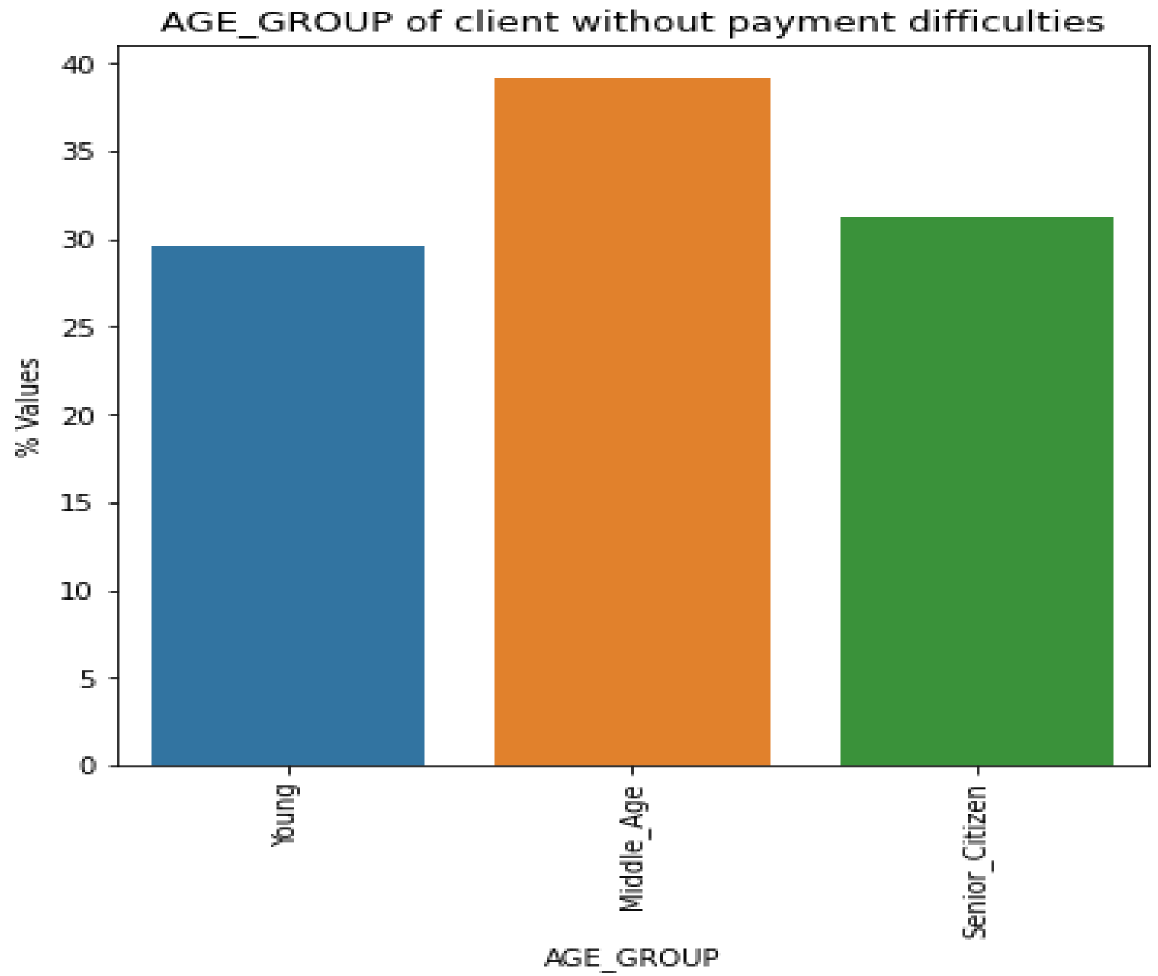
| **Checking Education of different target group** | **Observation:**<br>Increase in payment difficulties for secondary educated people->chance of defaulting more<br>Decrease in payment difficulty for higher educated people -> chance of defaulting is less |
|---|---|

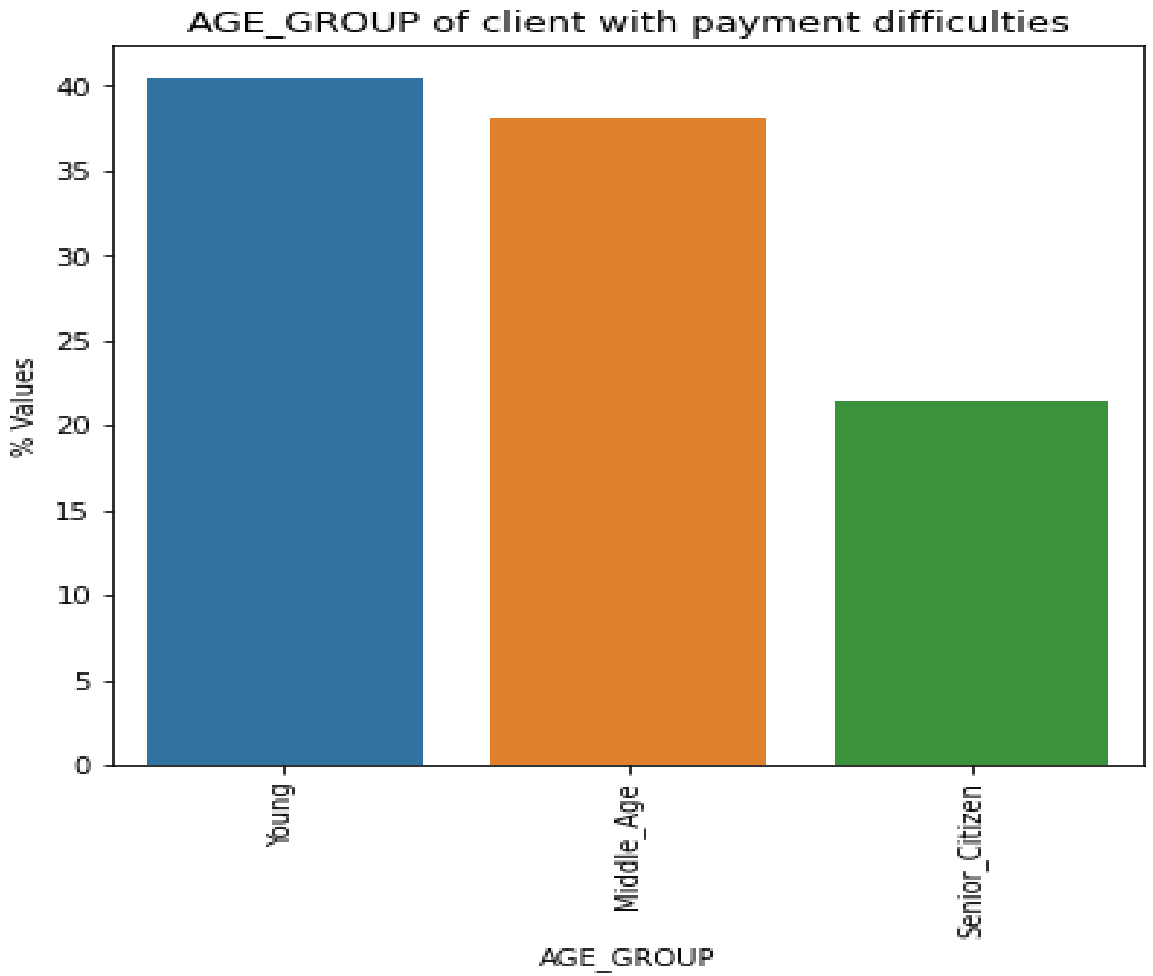


# Uni-Variate & Bi-Variate Analysis of App_data

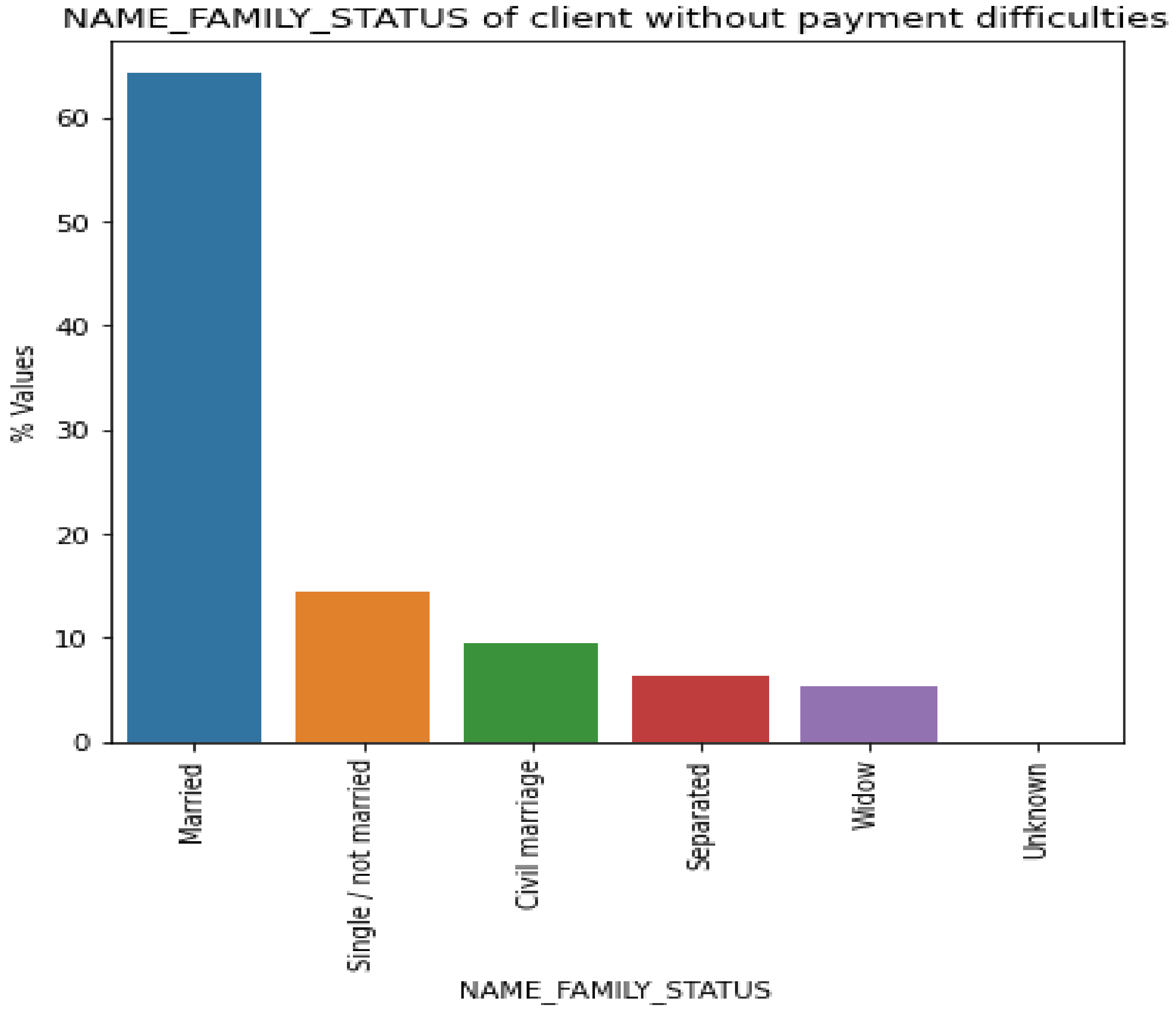
| | |
|---|---|
| **Checking Age Group of different target group** | **Observation:**<br>Decrease in % of payment difficulties for Senior citizen-> chance of defaulting less<br>Increase in % of payment difficulties for young clients-> chance of defaulting more |

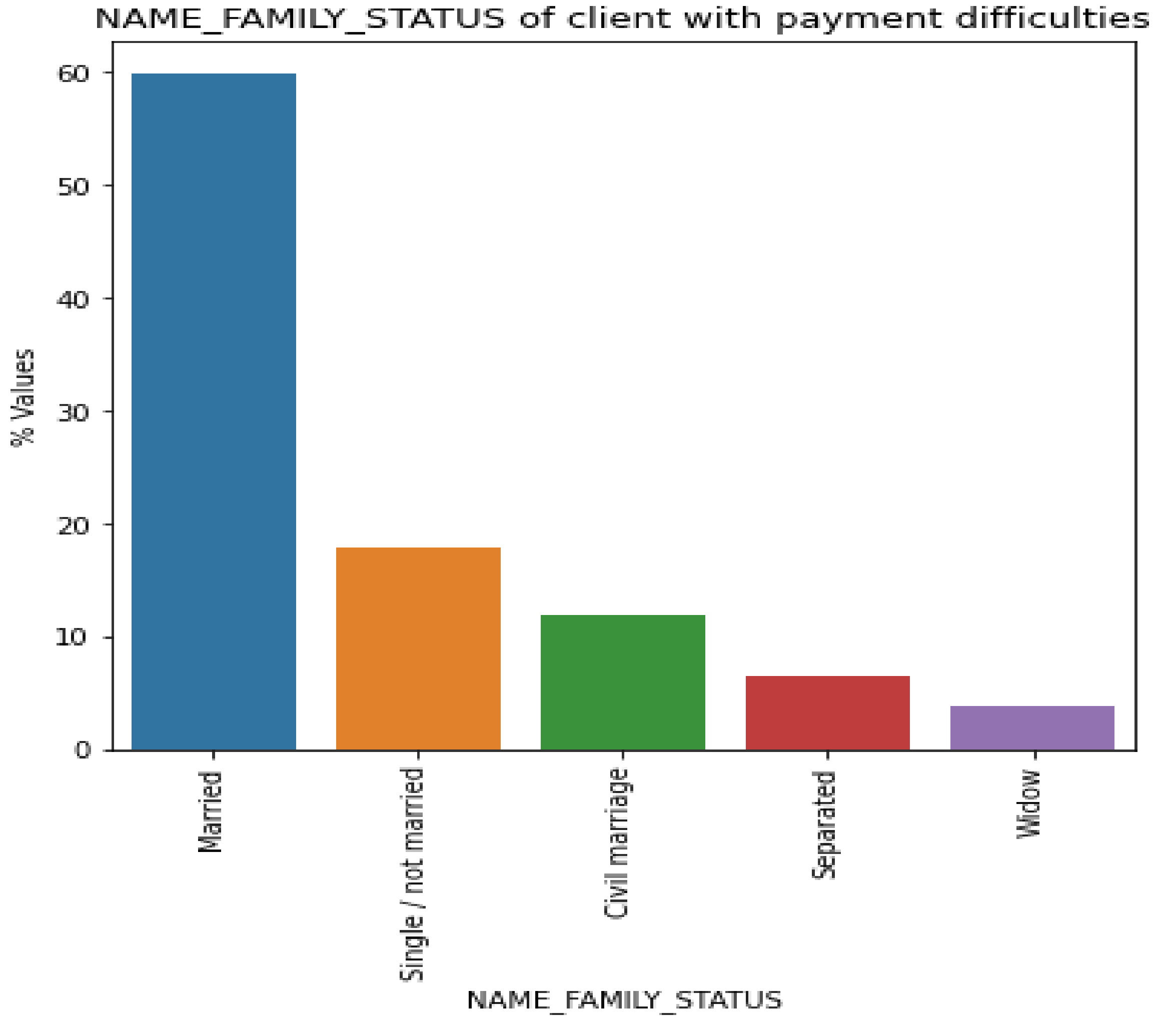


# Uni-Variate & Bi-Variate Analysis of App_data

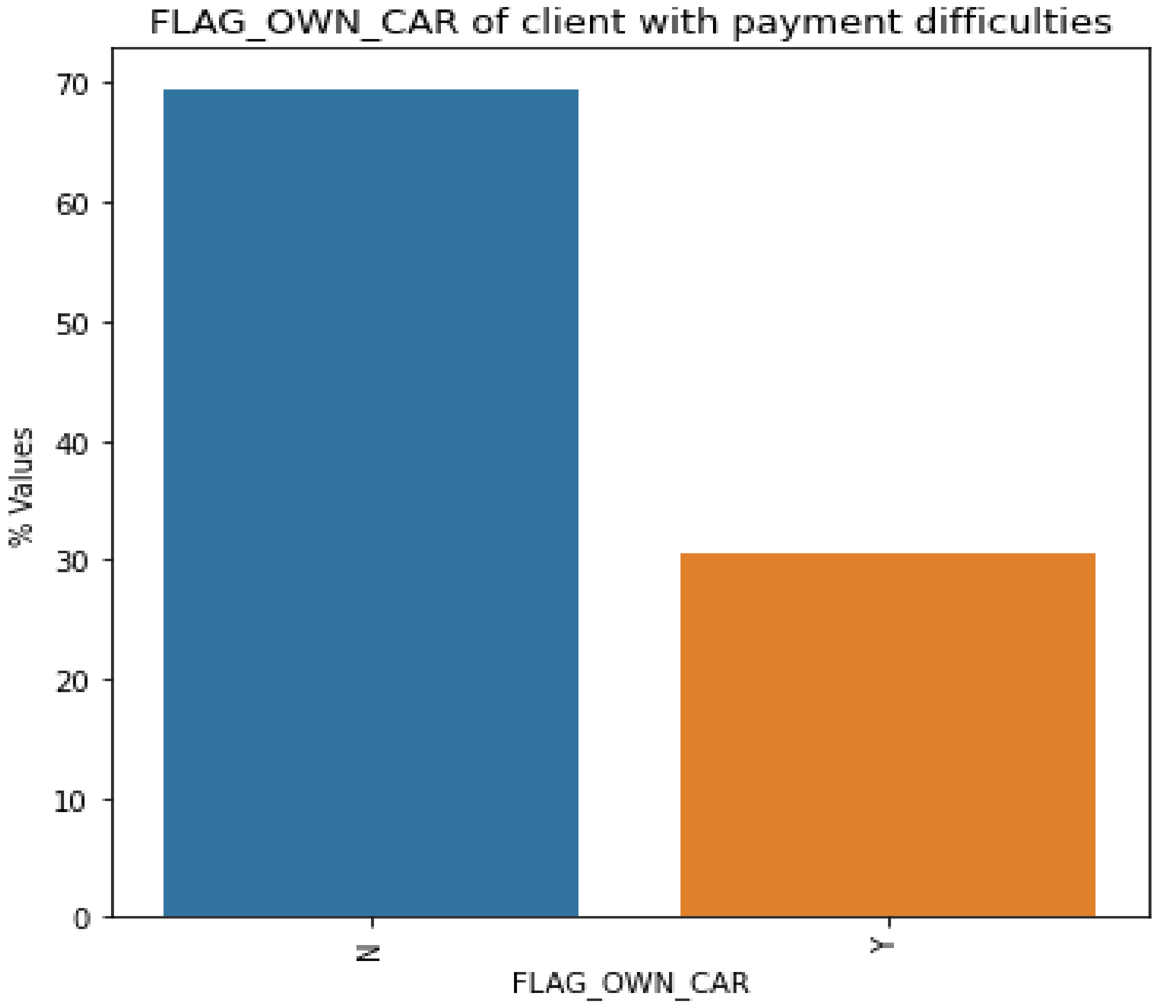
| Checking Family Status of different target group | Observation:<br>Decrease in payment difficulties for married one-> chance of defaulting less<br>Increase in payment difficulty with single & civil marriage-> chances are more |
|---|---|



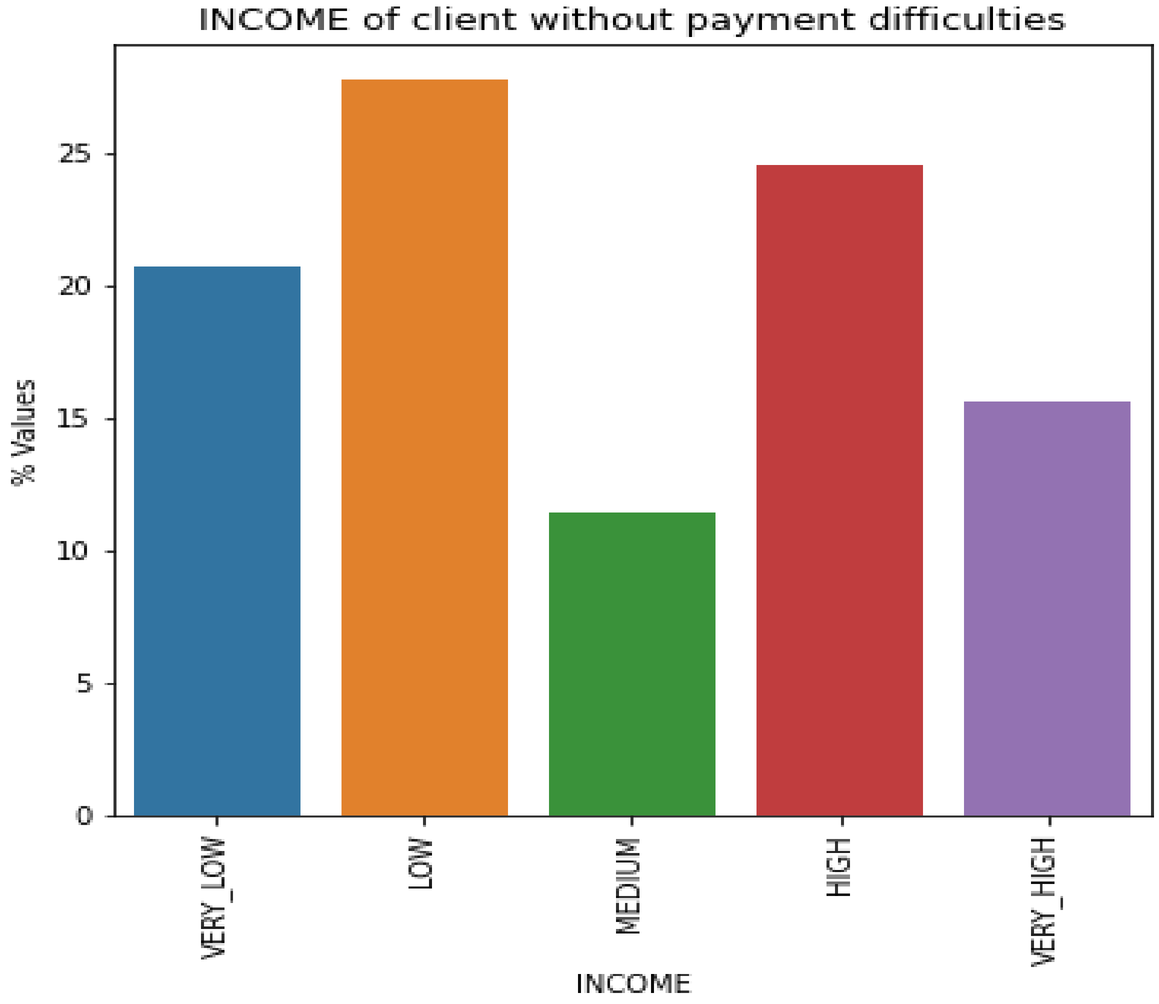
# Uni-Variate & Bi-Variate Analysis of App_data

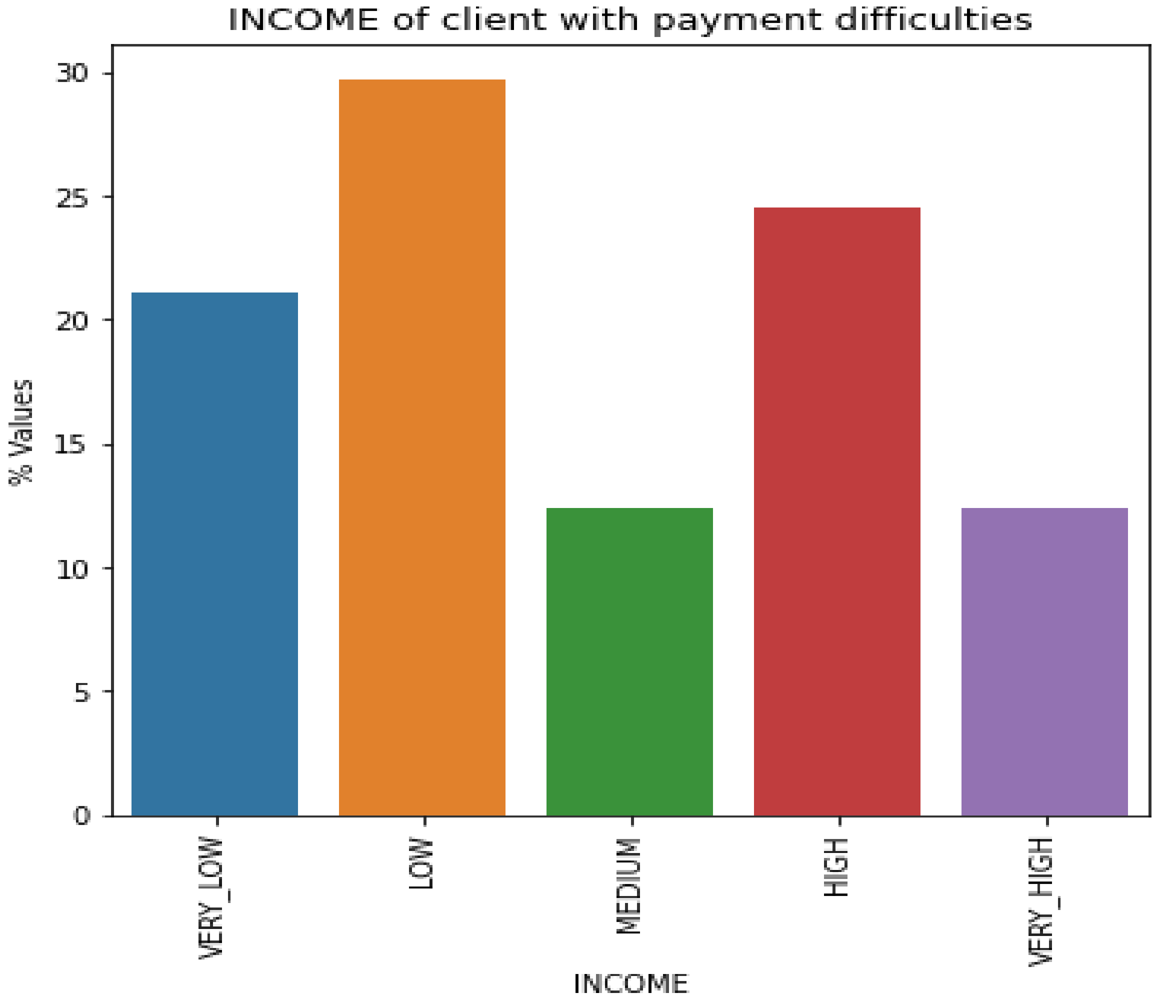
| **Checking Car Flag of different target group** | **Observation:**<br>Decrease in payment difficulty people having car -> chance of defaulting less |
| --- | --- |

# Uni-Variate & Bi-Variate Analysis of App_data

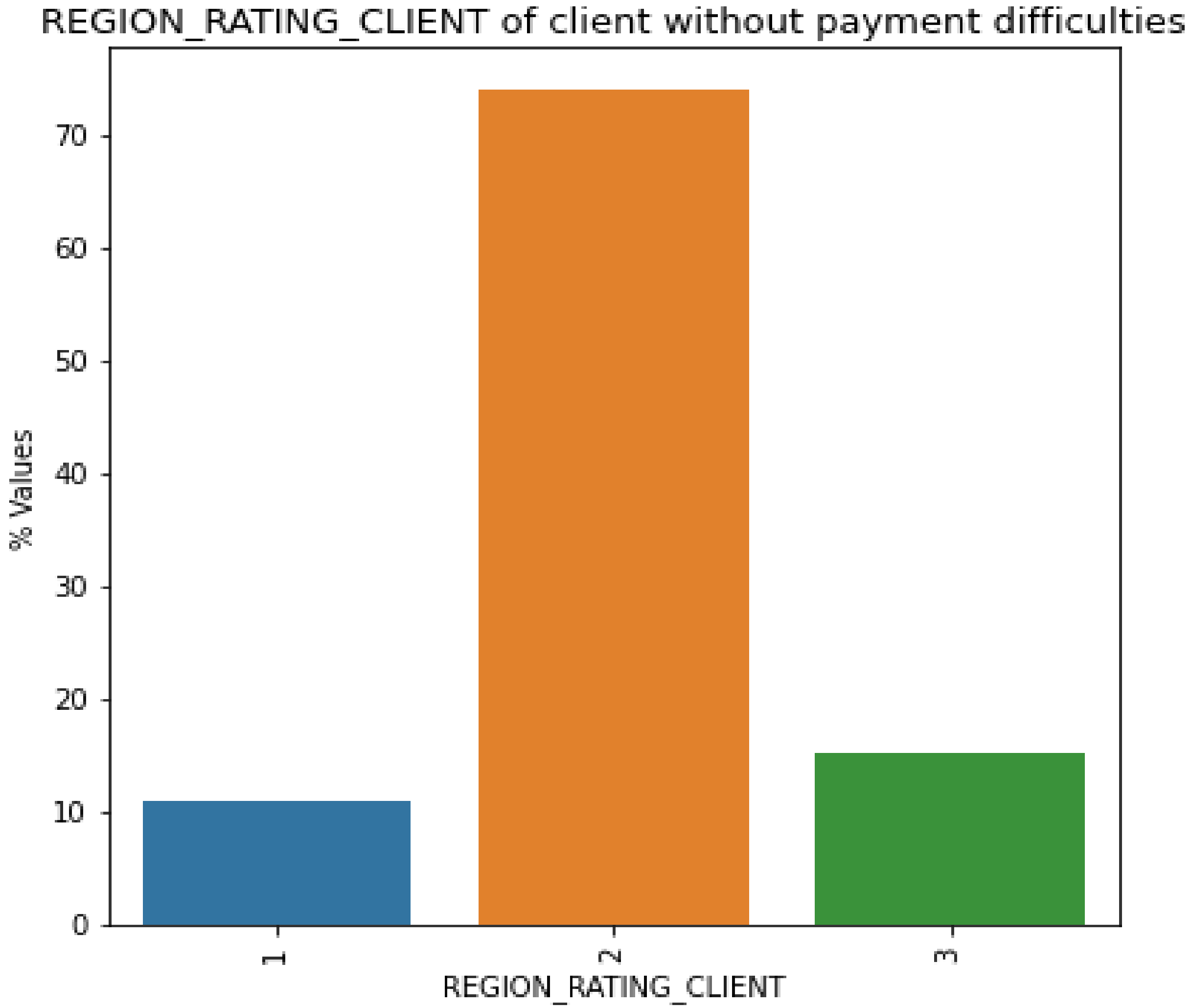
| **Checking Income Group of different target group** | **Observation:**<br>Increase in % of payment difficulties for low range income people-> chances are more<br>Decrease in % of payment difficulties for high range income people-> less chance |
| --- | --- |

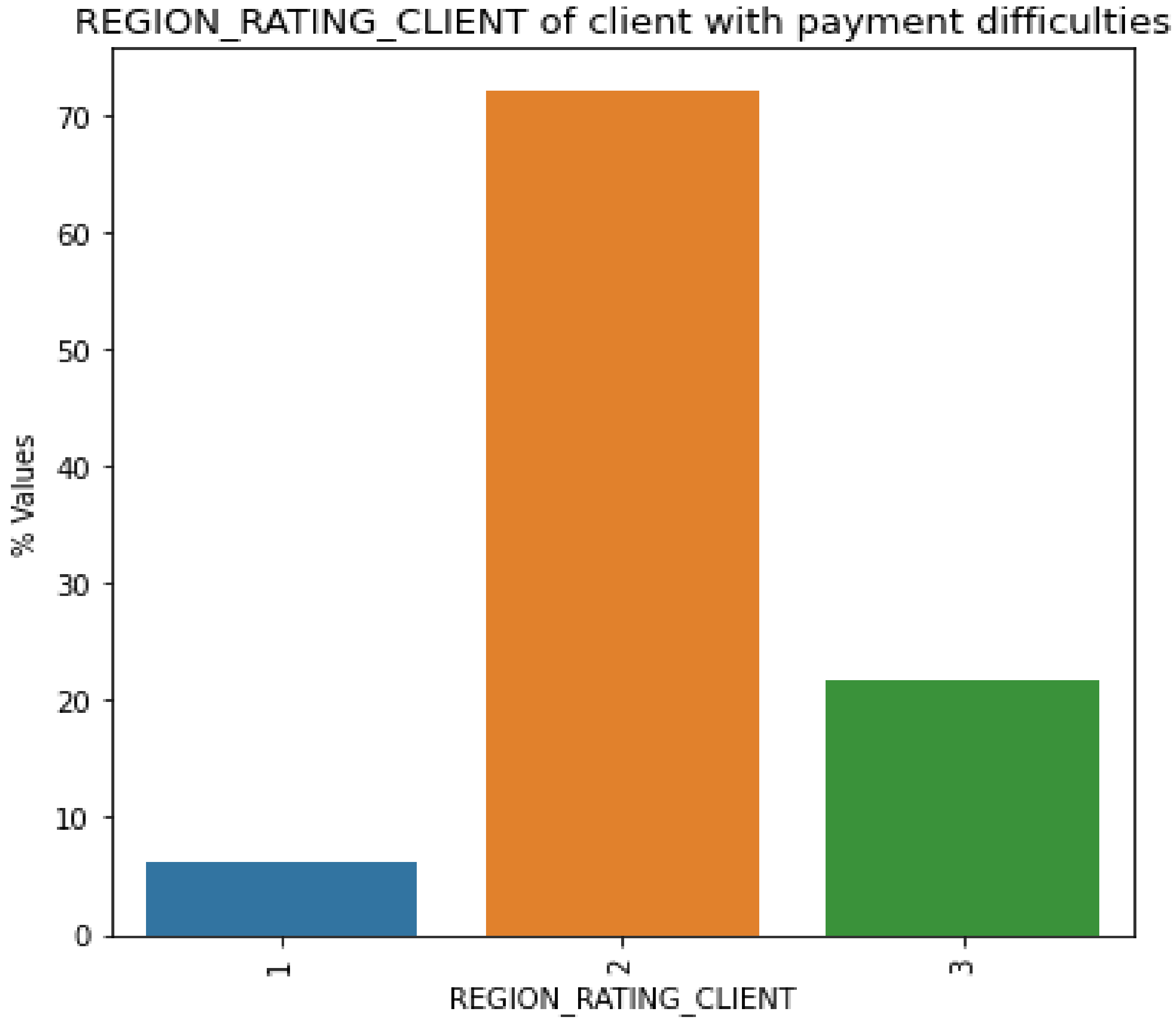


# Uni-Variate & Bi-Variate Analysis of App_data

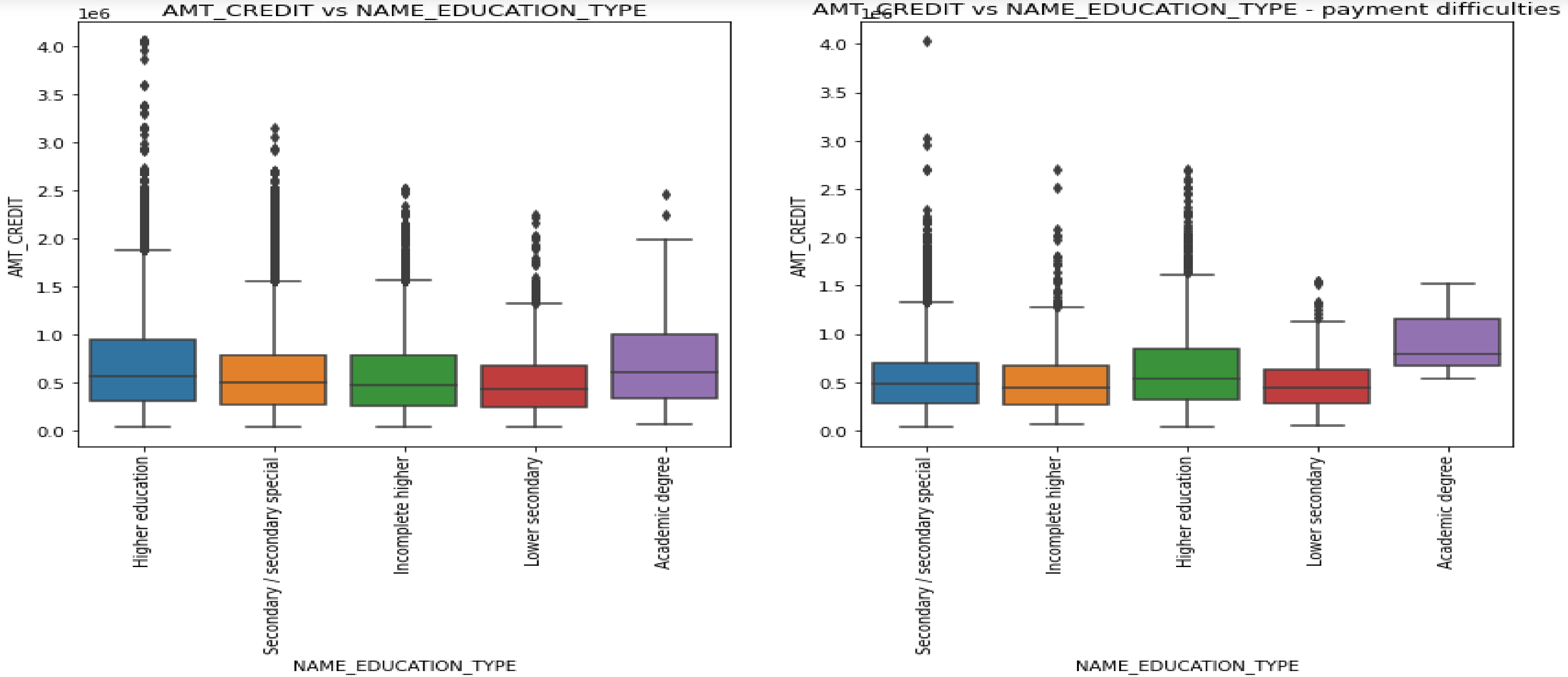
| | |
|---|---|
| **Checking Region Rating of different target group** | **Observation:**<br>People living in 2 rating apply for loan more than others<br>Increase payment difficulties for those living in rating 3->Chances are more to default |



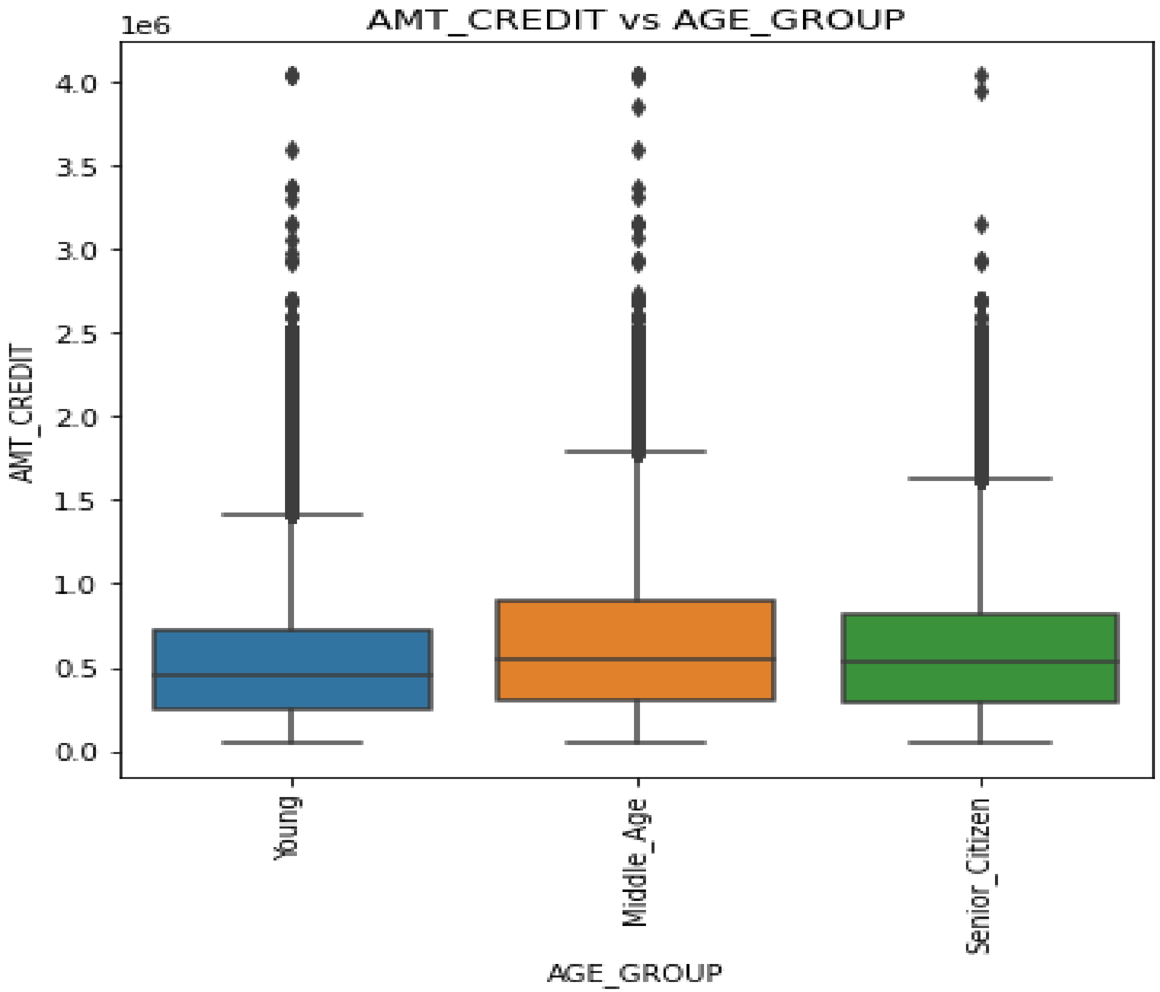
# Uni-Variate & Bi-Variate Analysis of App_data

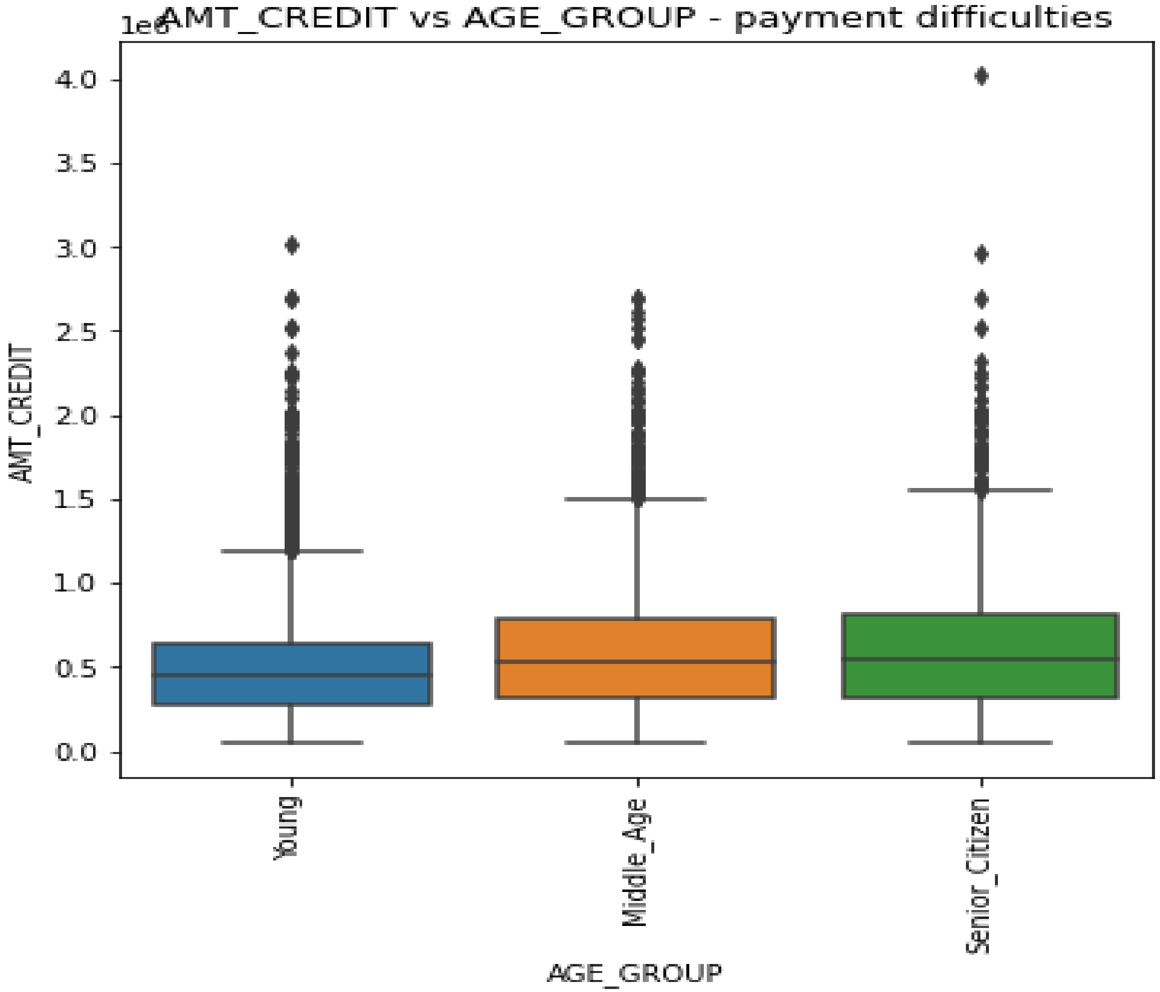
| **AMT_CREDIT vs Education Type** | **Observation:**<br>Higher education people having more outliers. People with academic degree has more amt_credit as compared with other |
|---|---|



# Uni-Variate & Bi-Variate Analysis of App_data

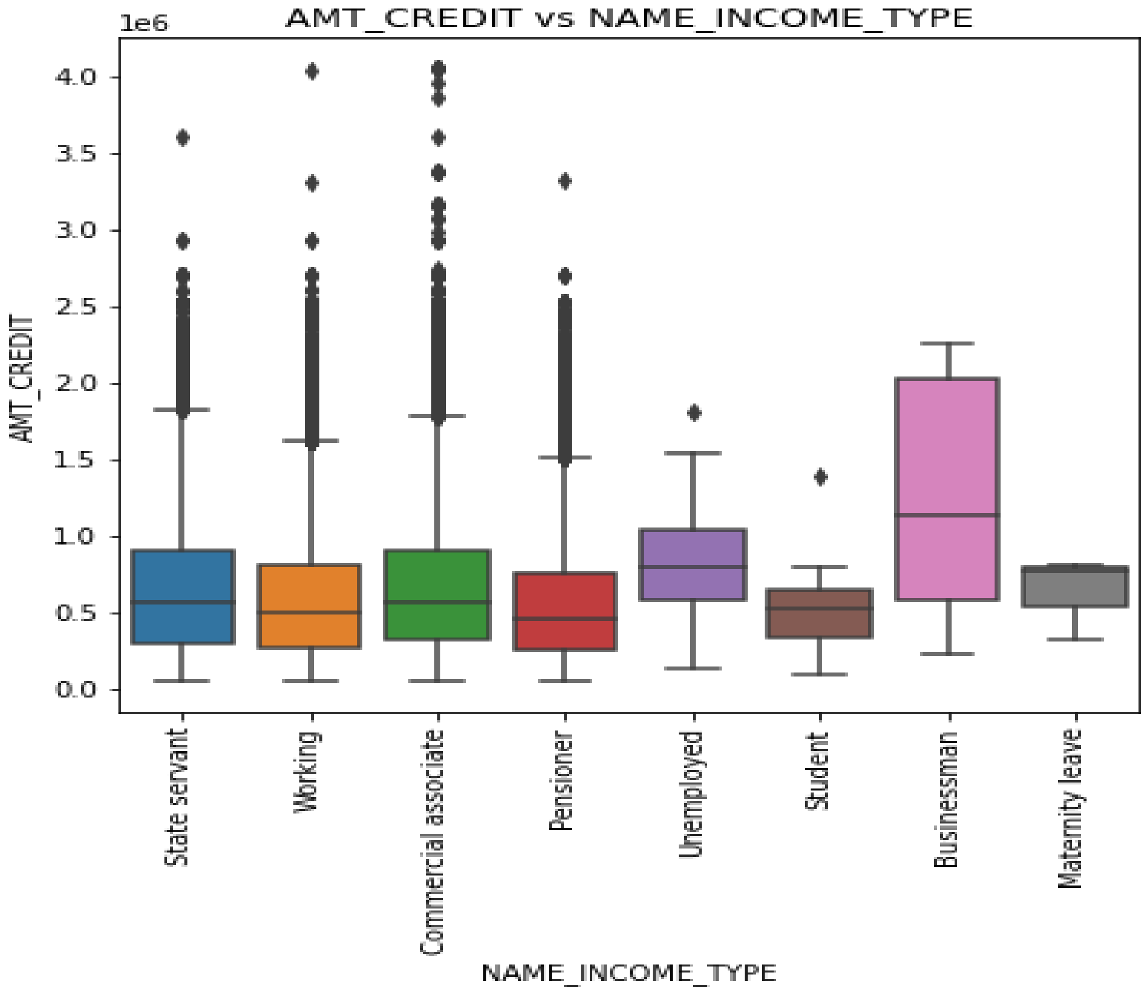
| **AMT_CREDIT vs Age Group** | **Observation:** Middle age & senior citizen has more credit than other |
| --- | --- |

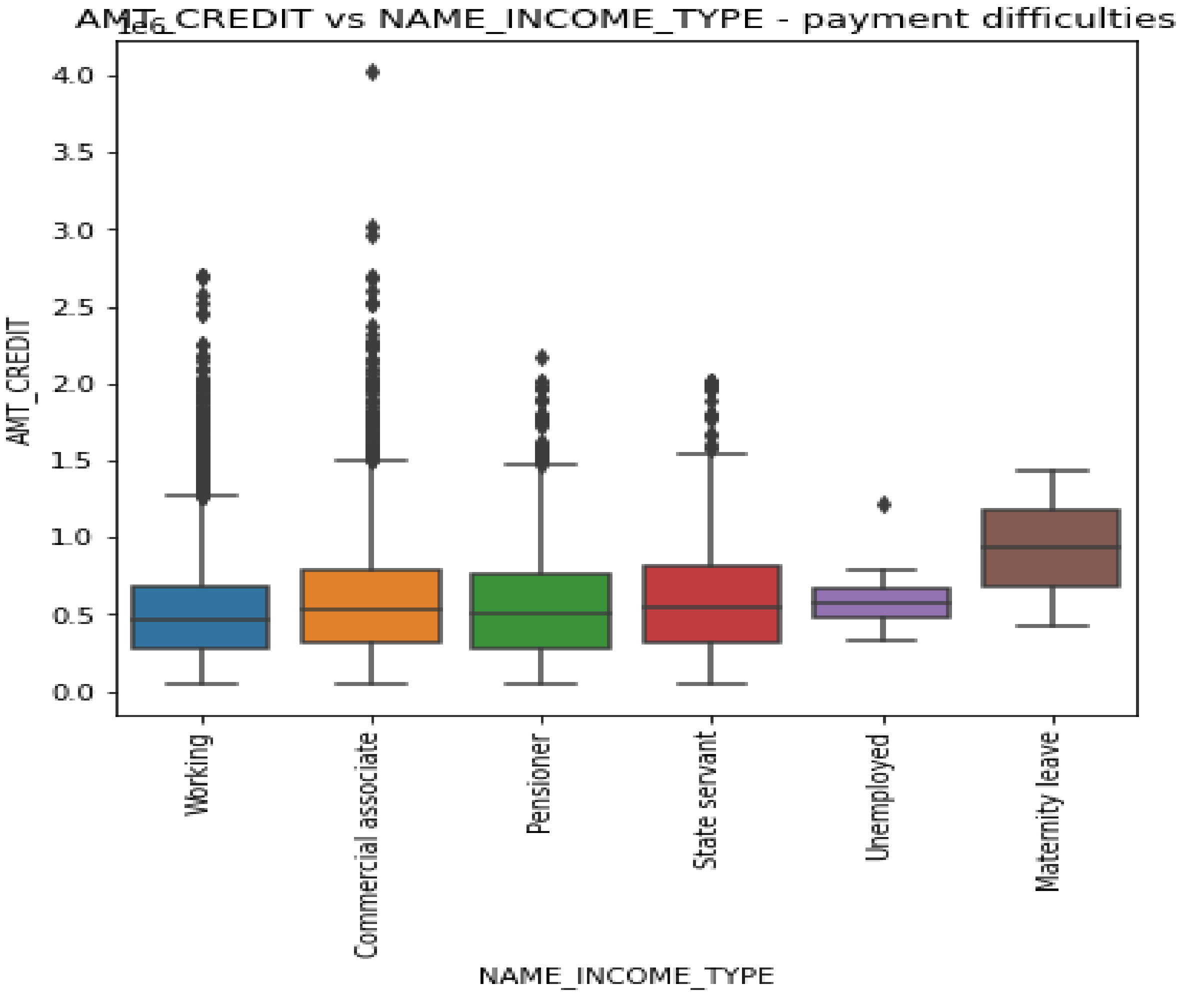


# Uni-Variate & Bi-Variate Analysis of App_data

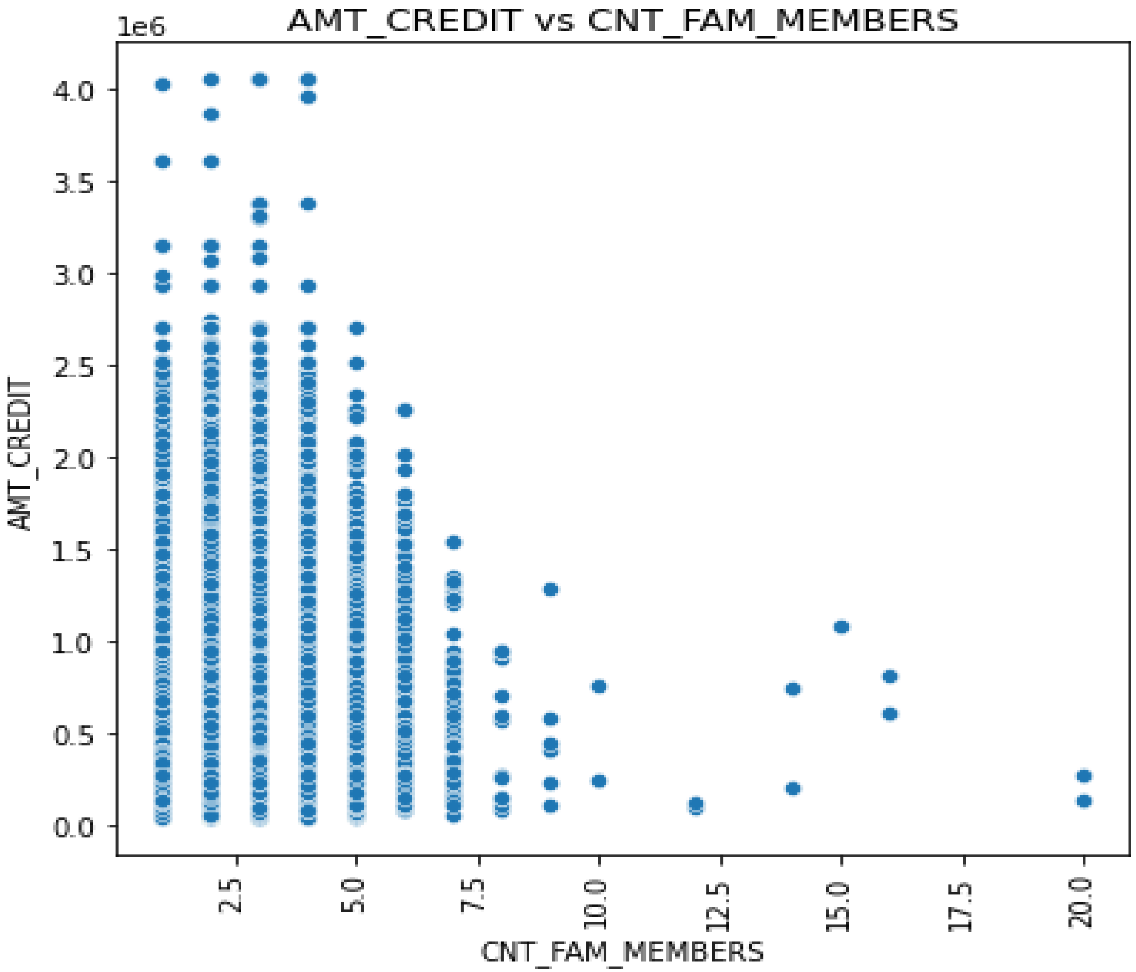
| **AMT_CREDIT vs Income Type** | **Observation:**<br>Comercial associate & state servant with payment difficulties have higher number of credit |
|---|---|

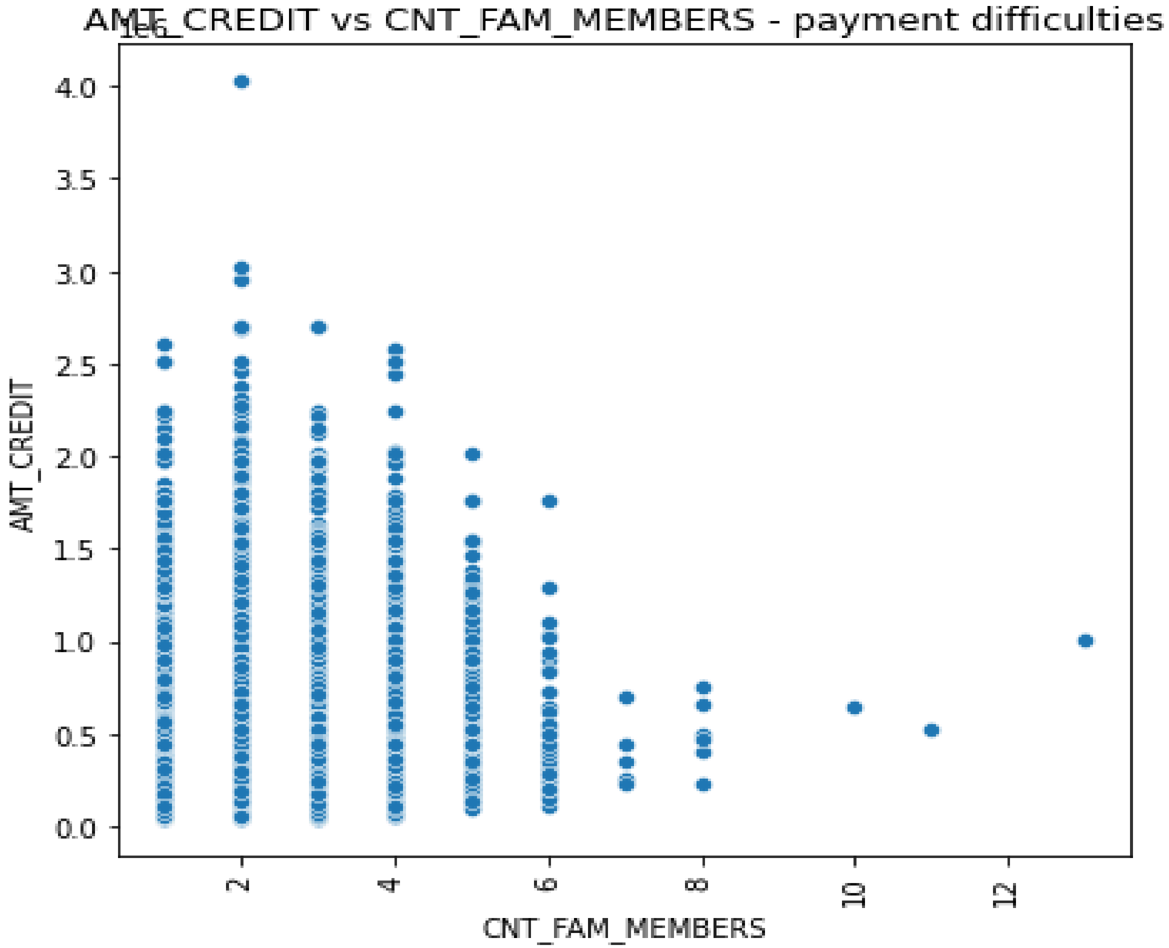


# Uni-Variate & Bi-Variate Analysis of App_data

| **AMT_CREDIT vs Family Count** | **Observation:**<br>People with family cnt less and the AMT_CREDIT is low are having more chances with payment difficulties and people with large family cnt and with larger AMT_CREDIT are having less chances with payment difficulties |
|---|---|



# Uni-Variate & Bi-Variate Analysis of App_data

| **AMT_CREDIT vs Goods Price** | **Observation:**<br>Credit linearly increases with good price |
| --- | --- |


# Uni-Variate & Bi-Variate Analysis of App_data

**Top 10 Correlation for client with payment difficulties**

| | | |
|---|---|---|
| AMT_CREDIT | AMT_ANNUITY | 0.752195 |
| AMT_ANNUITY | AMT_CREDIT | 0.752195 |
| AMT_GOODS_PRICE | AMT_ANNUITY | 0.752295 |
| AMT_ANNUITY | AMT_GOODS_PRICE | 0.752295 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.778540 |
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.778540 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.847885 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.847885 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.869016 |
| DEF_30_CNT_SOCIAL_CIRCLE | DEF_60_CNT_SOCIAL_CIRCLE | 0.869016 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.885484 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.885484 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.956637 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.956637 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.982783 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.982783 |
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998270 |
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 0.998270 |
| FLAG_EMP_PHONE | DAYS_EMPLOYED | 0.999701 |
| DAYS_EMPLOYED | FLAG_EMP_PHONE | 0.999701 |

# Uni-Variate & Bi-Variate Analysis of App_data

**Top 10 Correlation for client without payment difficulties**

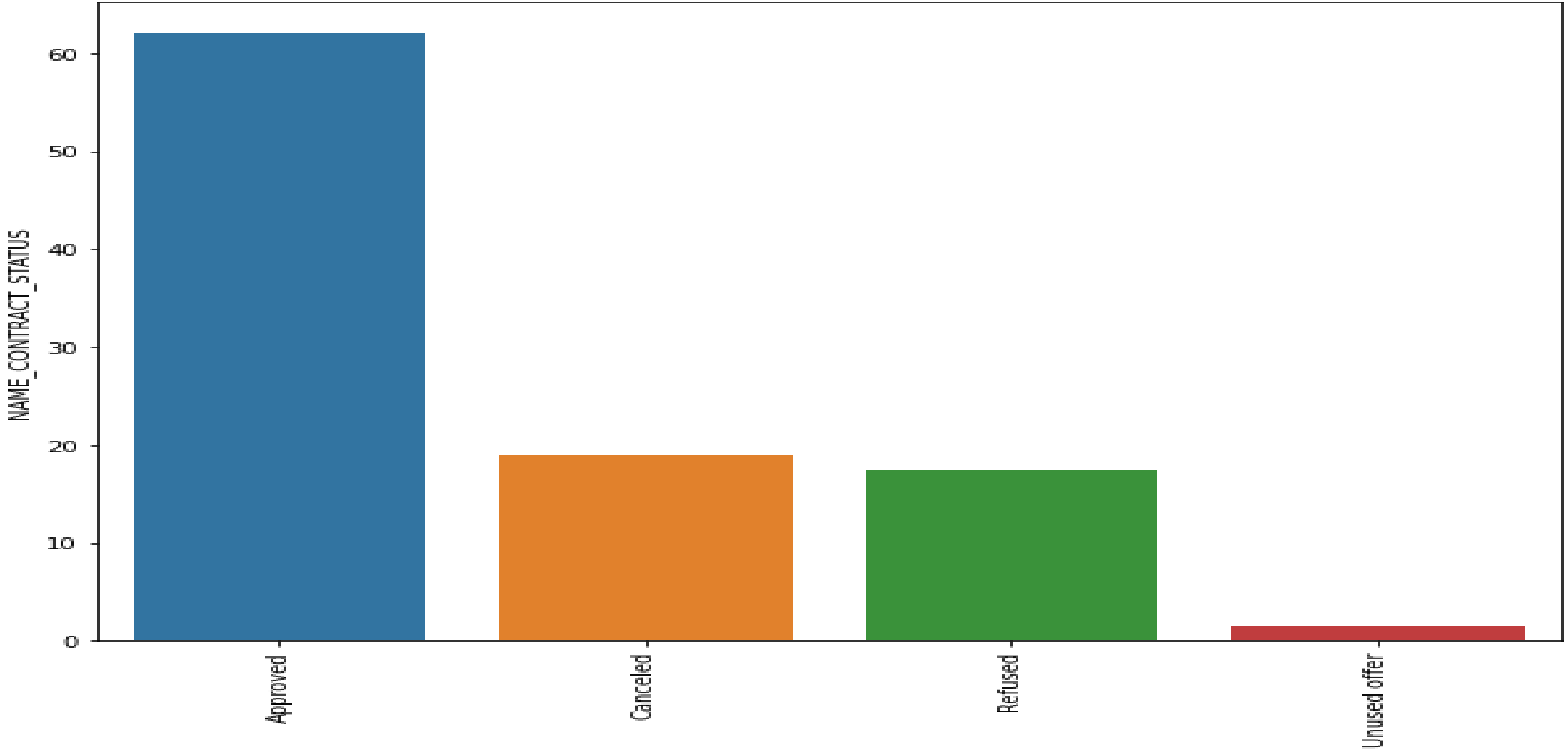
| | | |
|---|---|---|
| AMT_ANNUITY | AMT_CREDIT | 0.771296 |
| AMT_CREDIT | AMT_ANNUITY | 0.771296 |
| AMT_GOODS_PRICE | AMT_ANNUITY | 0.776421 |
| AMT_ANNUITY | AMT_GOODS_PRICE | 0.776421 |
| LIVE_CITY_NOT_WORK_CITY | REG_CITY_NOT_WORK_CITY | 0.830381 |
| REG_CITY_NOT_WORK_CITY | LIVE_CITY_NOT_WORK_CITY | 0.830381 |
| DEF_60_CNT_SOCIAL_CIRCLE | DEF_30_CNT_SOCIAL_CIRCLE | 0.859328 |
| DEF_30_CNT_SOCIAL_CIRCLE | DEF_60_CNT_SOCIAL_CIRCLE | 0.859328 |
| REG_REGION_NOT_WORK_REGION | LIVE_REGION_NOT_WORK_REGION | 0.861861 |
| LIVE_REGION_NOT_WORK_REGION | REG_REGION_NOT_WORK_REGION | 0.861861 |
| CNT_CHILDREN | CNT_FAM_MEMBERS | 0.878569 |
| CNT_FAM_MEMBERS | CNT_CHILDREN | 0.878569 |
| REGION_RATING_CLIENT_W_CITY | REGION_RATING_CLIENT | 0.950148 |
| REGION_RATING_CLIENT | REGION_RATING_CLIENT_W_CITY | 0.950148 |
| AMT_GOODS_PRICE | AMT_CREDIT | 0.987024 |
| AMT_CREDIT | AMT_GOODS_PRICE | 0.987024 |
| OBS_30_CNT_SOCIAL_CIRCLE | OBS_60_CNT_SOCIAL_CIRCLE | 0.998510 |
| OBS_60_CNT_SOCIAL_CIRCLE | OBS_30_CNT_SOCIAL_CIRCLE | 0.998510 |
| FLAG_EMP_PHONE | DAYS_EMPLOYED | 0.999758 |
| DAYS_EMPLOYED | FLAG_EMP_PHONE | 0.999758 |

# Brief Description of Cleaning & preparing data – Prev_app

1. Dropped all columns which has more than 40% null values
2. Impute the missing values in remaining column
   1. Replaced the null values with mode in case of categorical column
   2. Replaced the null values with mean in case of numerical column with no outliers (outliers are identified using box plot)
   3. Replaced the null values with median in case of numerical column with outliers (outliers are identified using box plot)
3. Replacing XNA & XAP Values with NaN values

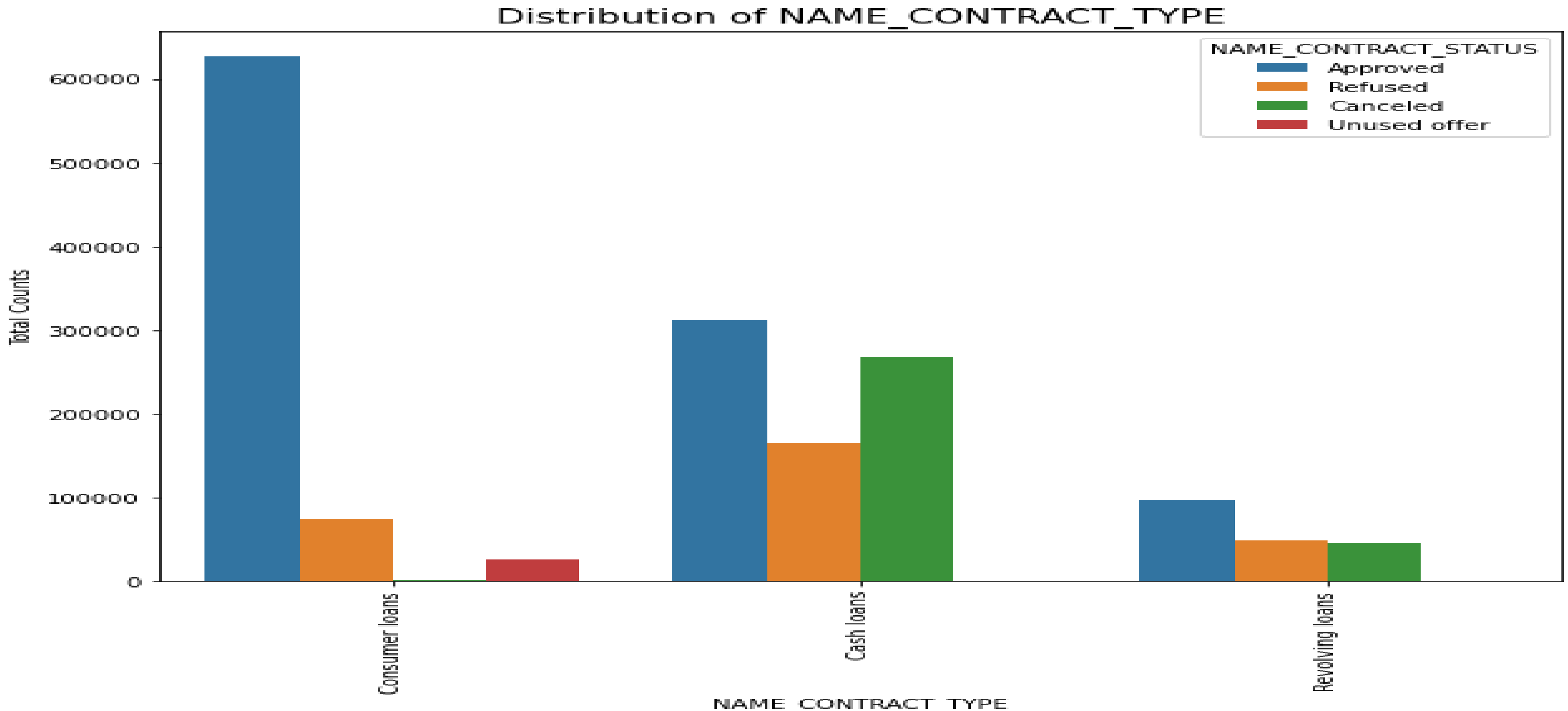
# Uni-Variate & Bi-Variate Analysis of Prev_app

| | |
|---|---|
| **Checking Contract Status** | **Observation:**<br>Most of the loan are approved by the bank |



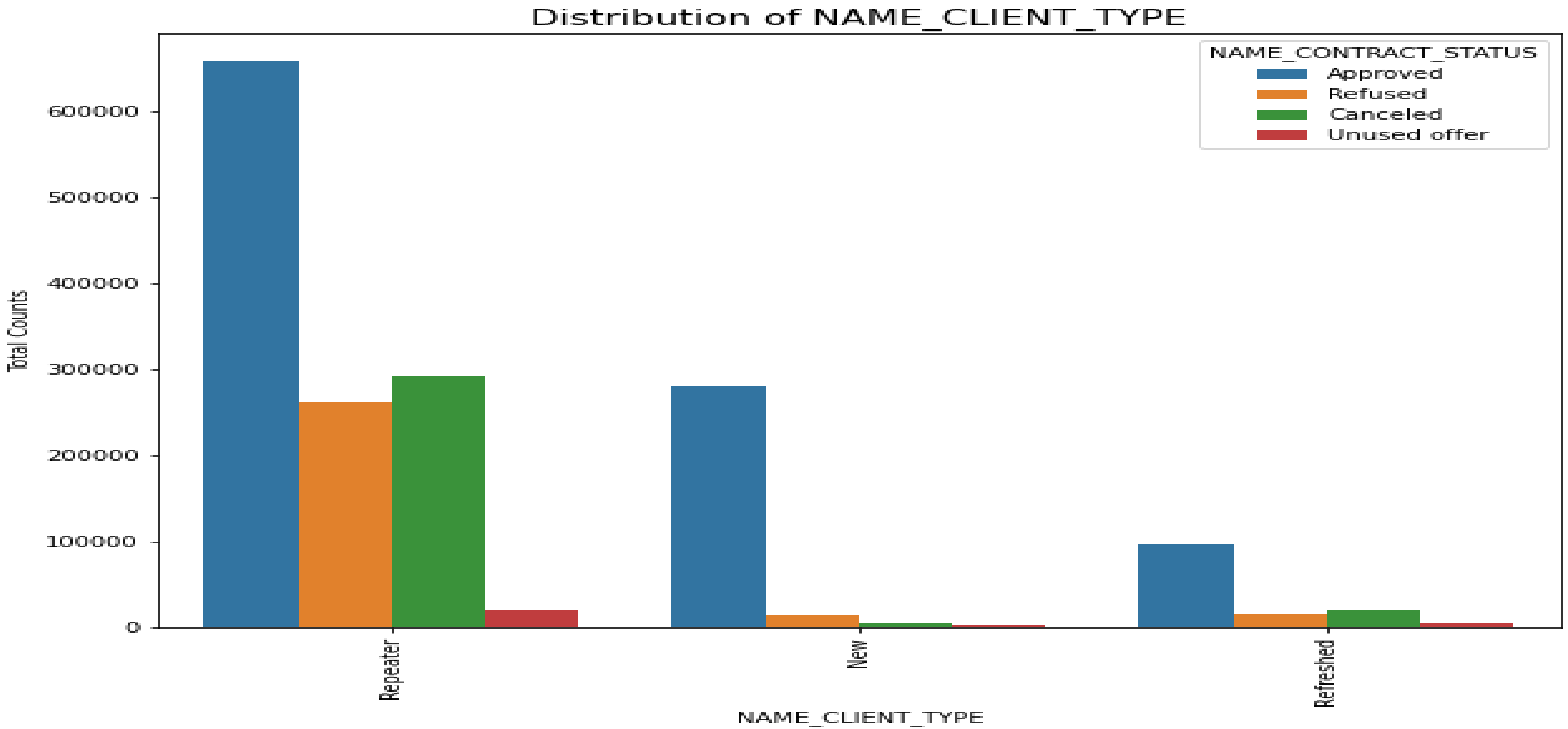
# Uni-Variate & Bi-Variate Analysis of Prev_app

| **Contract Status vs Contract Type** | **Observation:**<br>Number of approved consumer loans are much higher than any other and also number of unused consumer loan are higher |
|---|---|



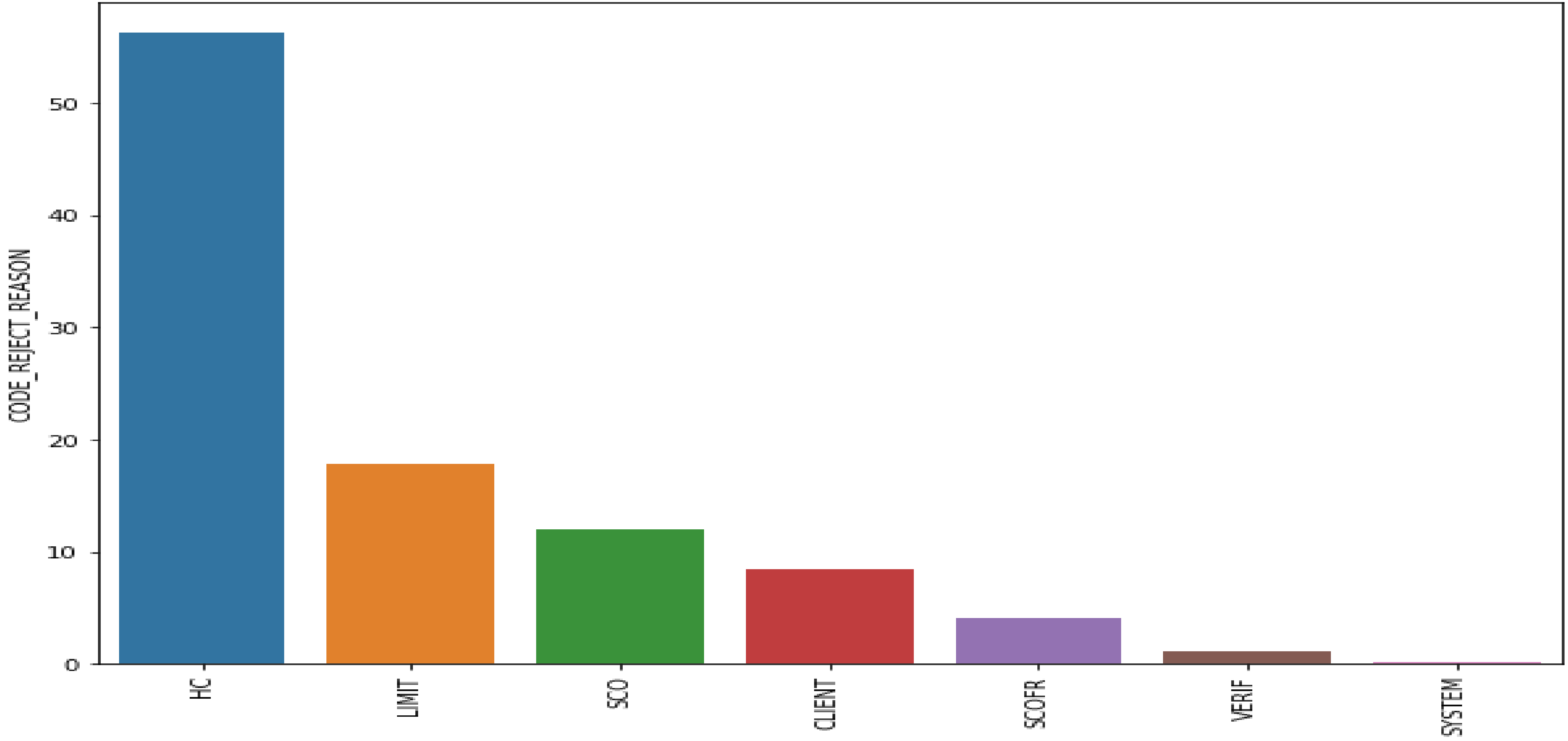
# Uni-Variate & Bi-Variate Analysis of Prev_app

| **Contract Status vs Client Type** | **Observation:**<br>All contract status of repeater are higher than other |
| --- | --- |



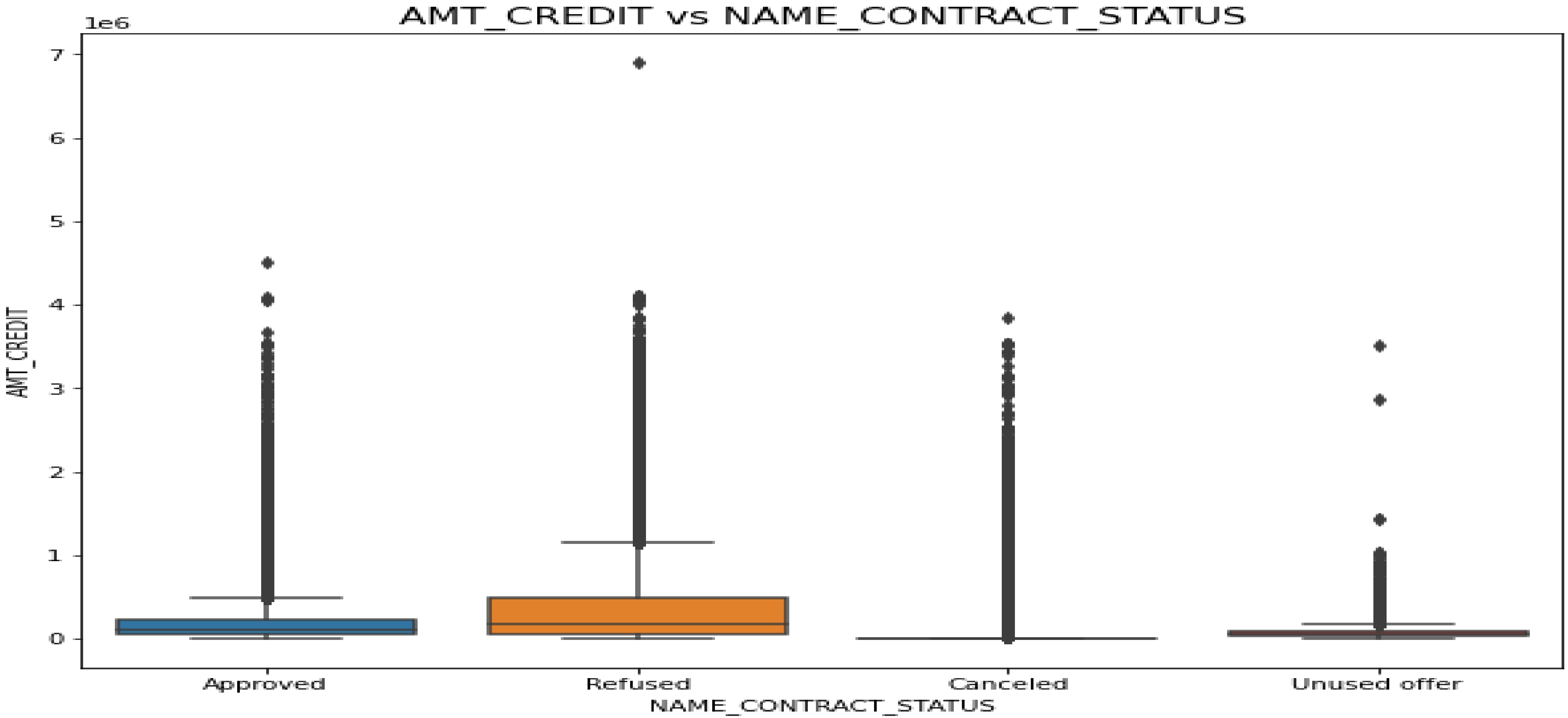
# Uni-Variate & Bi-Variate Analysis of Prev_app

| Checking Reject Reason | **Observation:**<br>HC & limit are the important reason of rejecting the previous application |
|---|---|



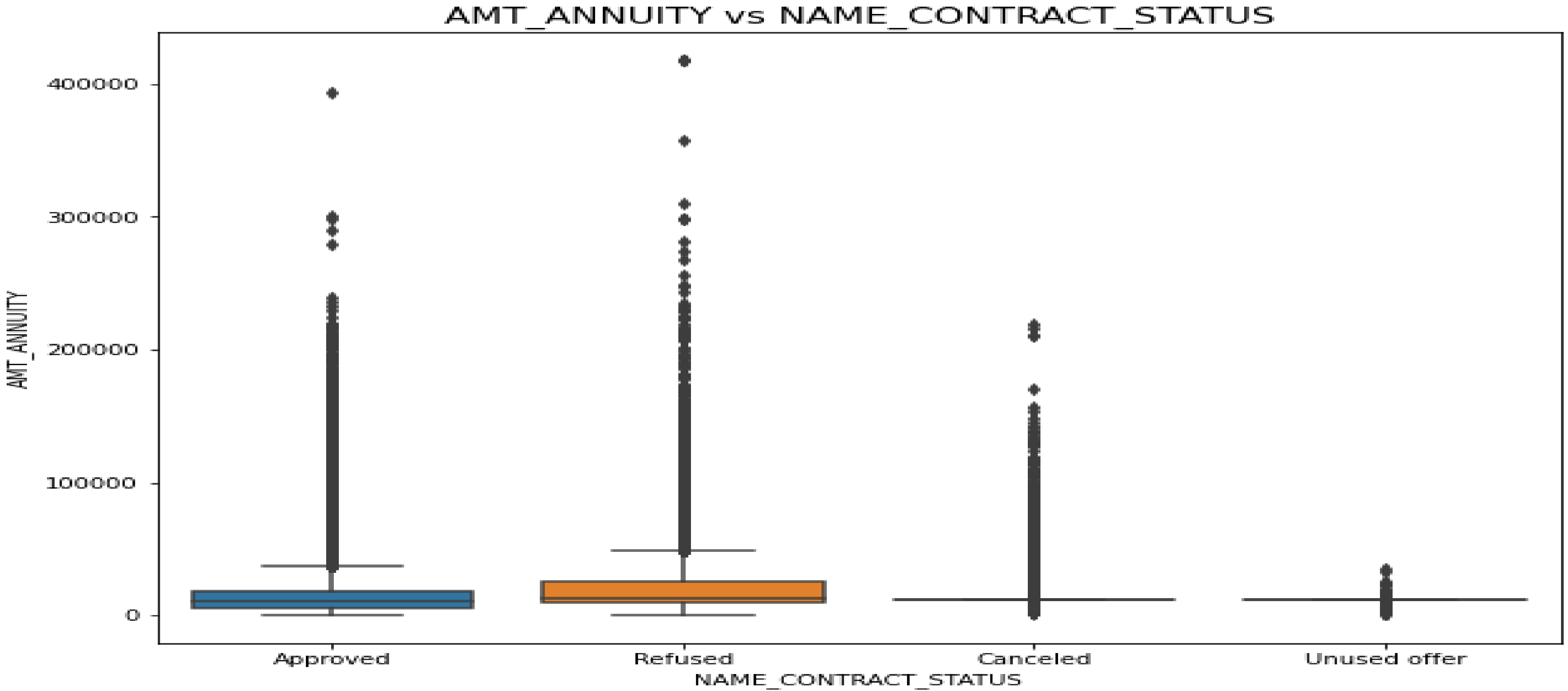
# Uni-Variate & Bi-Variate Analysis of Prev_app

| | |
|---|---|
| **AMT_Credit vs Contract Status** | **Observation:**<br>When AMT_CREDIT is low more chance of loan to be cancelled & unused |



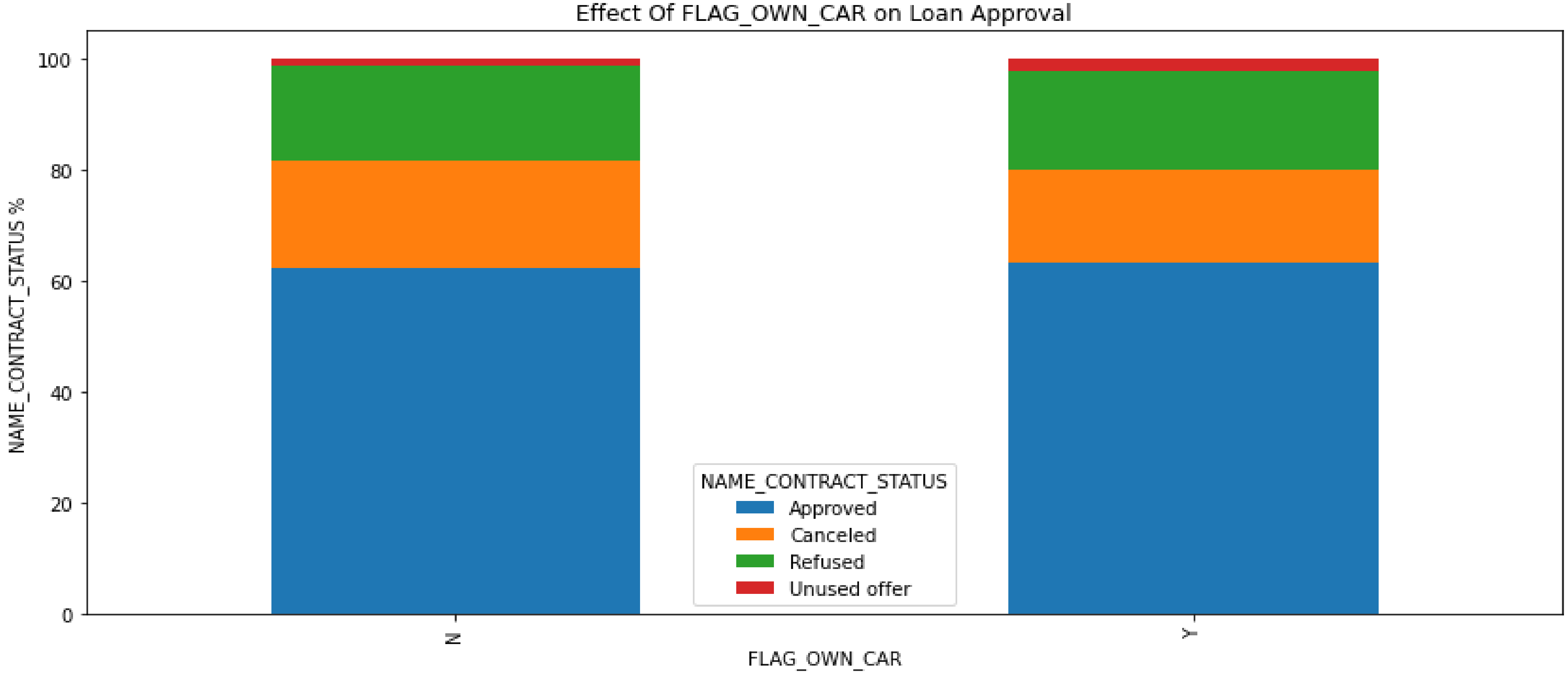
# Uni-Variate & Bi-Variate Analysis of Prev_app

| **AMT_Annuity vs Contract Status** | **Observation:**<br>Loan application for people with lower AMT_ANNUITY gets cancelled or Unused more and application with high AMT ANNUITY also got refused more |
|---|---|



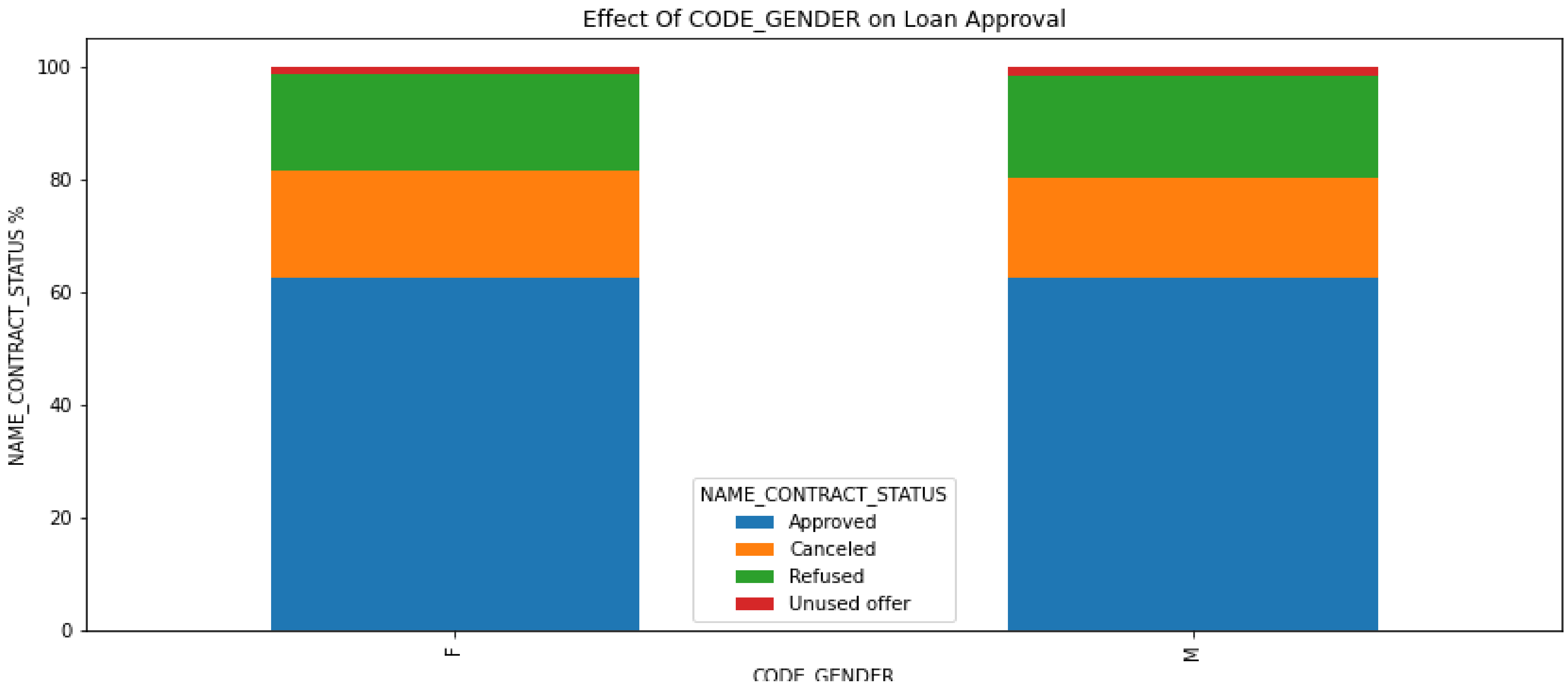
# Merging two DataFrame and getting insights

| **Effect Of Own Car on Loan Approval** | **Observation:**<br>People with car has less chance of default. The bank can add more weightage to car ownership while approving a loan amount |
| --- | --- |



# Merging two DataFrame and getting insights

| **Effect Of Gender on Loan Approval** | **Observation:** Female have less chance of default than man. The bank can add more weightage to female while approving a loan amount. |
| --- | --- |



# Final Insights

- Less Chances to be a defaulter
  - State servant clients
  - Senior citizen
  - High Income clients
  - Female clients
  - Higher education clients (female)
  - Clients who’s previous loan status was approved
- More chances to be a defaulter
  - Civil marriage clients (male)
  - Previously refused loan clients
  - Lower secondary education clients

# Thank You