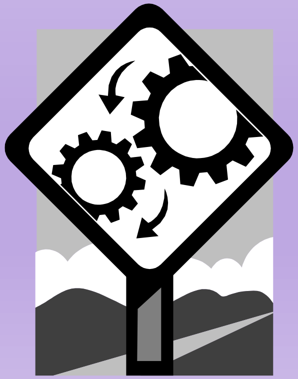
EDA CREDIT CASE STUDY

Sandesh Sawant & Anubhav Mukherjee



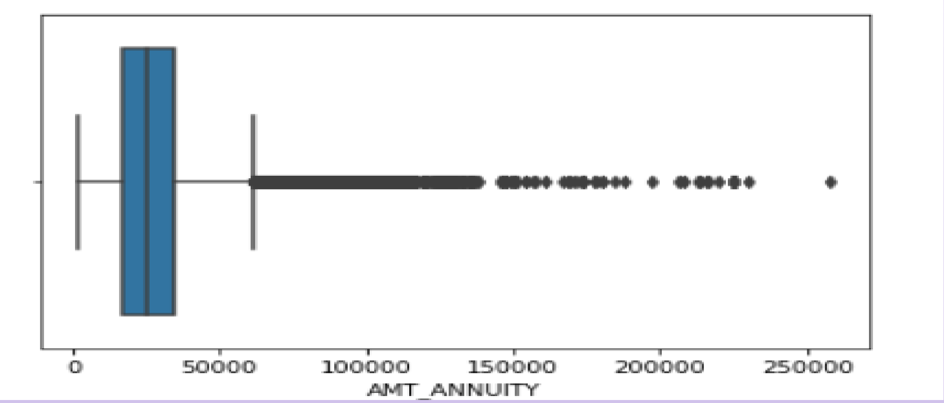
Data Quality and Missing Values Check

From the describe method and the boxplot, it is observed in column "AMT_ANNUITY", a huge difference between 3rd Quartile and the max value, leading to the conclusion of more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

Thus, it would be appropriate to fill those missing values with Median values instead of Mean because Mean will shift as per the imputed value. Hence Median comes to rescue for this and we will fill those missing banks with Median value.

Also, the boxplot is comparatively short, suggesting that the values are evenly concentrated around the median

```
# Checking for Outliers using Boxplot
sns.boxplot(x=df['AMT_ANNUITY'])
```



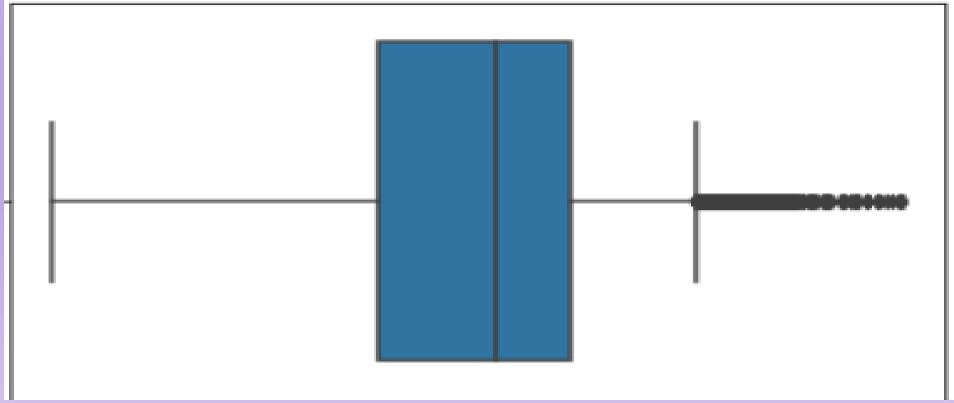
From the describe method and boxplot, it is observed in column "AMT_GOODS_PRICE", there seems to be a huge difference between 3rd Quartile and the max value, leading to the conclusion of values concentrated near 4th quartile. Thus, any values greater than 1.5(IQR) are considered as outliers.

Thus, it would be appropriate to fill those missing values with Median values instead of Mean because Mean will shift as per the imputation. Hence Median comes to rescue for this and we will fill those missing banks with Median value.

Also, the longer lower whisker in the figure means that goods price is varied amongst the lower quartiles.

Additionally, values are concentrated more below the median value in lower quartiles

```
# Checking for Outliers using Boxplot
sns.boxplot(x=df['AMT_GOODS_PRICE'])
plt.xscale('log')
plt.show())
```



Categorical , Negative Values & XNA Check

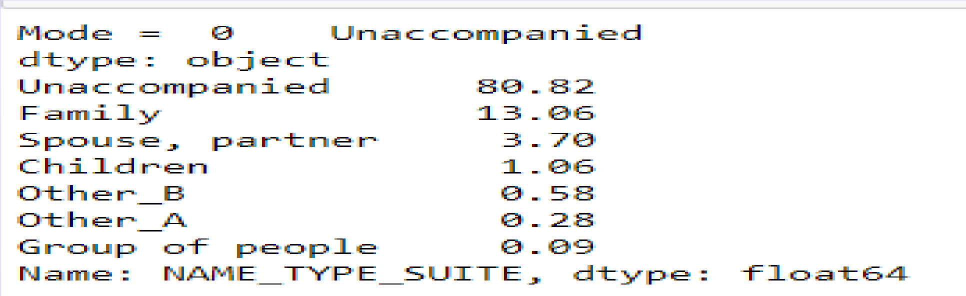
Recommending to impute the value of 'Unaccompanied' to the Null values as it holds to 81% of the column's values rescue for this and we will fill those missing banks with Median value.

Also, the boxplot is comparatively short, suggesting that the values are evenly concentrated around the median

# As this is a Categorical variable, checking the Mode of the "NAME_TYPE_SUITE" Column

```
print("Mode = ", df['NAME_TYPE_SUITE'].mode())

# Finding the percentage of all occurences of specific values
print(round(df.NAME_TYPE_SUITE.value_counts()/len(df)*100,2))
```



```
Mode =   0     Unaccompanied
dtype: object
Unaccompanied         80.82
Family                13.06
Spouse, partner        3.70
Children               1.06
Other_B                0.58
Other_A                0.28
Group of people        0.09
Name: NAME_TYPE_SUITE, dtype: float64
```

VALIDATION FOR NEGATIVE VALUES

```
#  validating for negative values within in columns and converting to positive values

df['DAYS_BIRTH'] = df['DAYS_BIRTH'].abs()
df['DAYS_EMPLOYED'] = df['DAYS_EMPLOYED'].abs()
df['DAYS_REGISTRATION'] =
df['DAYS_REGISTRATION'].abs()
df['DAYS_ID_PUBLISH'] = df['DAYS_ID_PUBLISH'].abs()
df['DAYS_LAST_PHONE_CHANGE'] =
df['DAYS_LAST_PHONE_CHANGE'].abs()
```

VALIDATION of 'XNA' values

```
df.isin(['XNA']).any()
```

For 'ORGANIZATION_TYPE', we have total count of 307511 rows of which 55374 rows are having 'XNA' values, which means 18% of the values. Hence if we drop the 55374 rows, it will not have any major impact on our dataset.

4 rows from CODE_GENDER column were having 'XNA' values
However, for column 'CODE_GENDER', we have total count of 307511 rows of which 4 rows are having 'XNA' values. Thus, we can update those columns with Gender 'F' as there will be no impact on the dataset.

Correlation between Target 0 and Target 1

1) There is perfect correlation(0.99) between AMT_INCOME_TOTAL and AMT_GOODS_PRICE, wherein increase in one variable tends to increase in other variable as well. Here, more is the price on the goods earned by the consumer, his income increases as well, and he is in a better position to pay his loan.

2) Similarly, there is a strong positive correlation(0.99) between AMT_ANNUITY & AMT_CREDIT as well. Here, higher are the loan annuity payments made by the client, more is the credit available and he would be able to borrow more from the bank.

3) Also, there is a strong correlation(0.75) between AMT_ANNUITY & AMT_INCOME_TOTAL, wherein more is the income of the client, more possibility of the client making regular payments on the loan to the bank.

4) There seems to be weak correlation(0.24) between AMT_ANNUITY & CNT_CHILDREN. The no of children the client has, does not bear a great impact on his loan annuity payments.

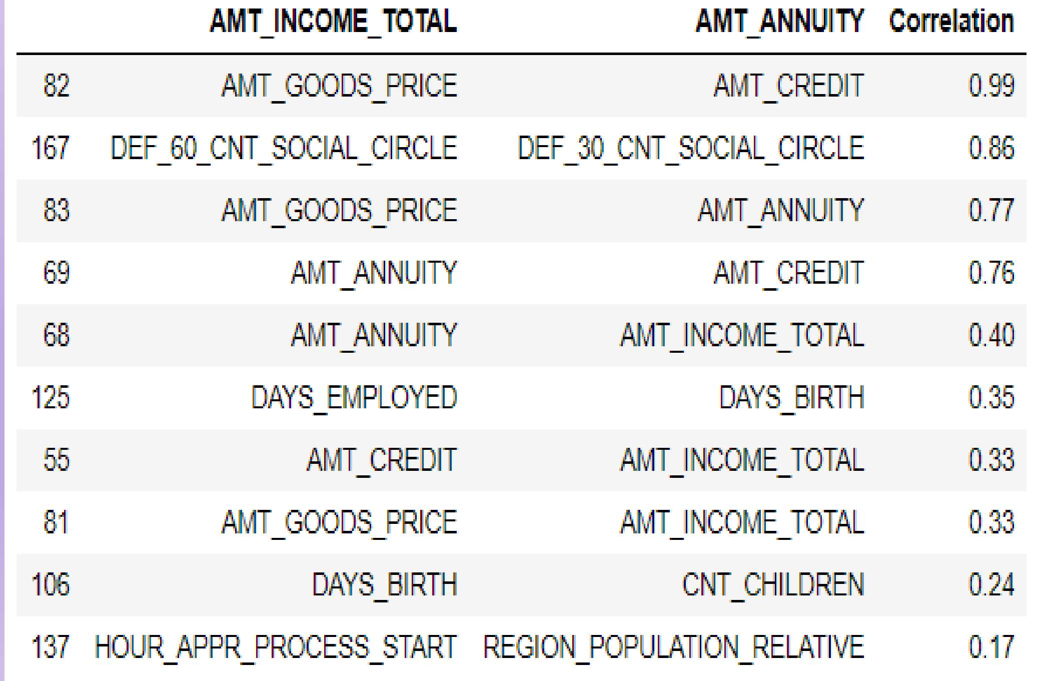
5) Additionally, there is a weak correlation(0.3) between AMT_INCOME_TOTAL & DAYS_EMPLOYED, wherein the employment of the consumer is not greatly impacting his loan application.

6) Weak correlation(0.17) implies customers residing in densely populated regions have less likely to receive loan facility

```
# For Target=0 (client with payment difficulties)
corr0 = target0.corr()
corr0 = corr0.where(np.triu(np.ones(corr0.shape), k=1).astype(np.bool))
corr0

corrdf0 = corr0.unstack().reset_index()
corrdf0.columns = ['AMT_INCOME_TOTAL', 'AMT_ANNUITY',
'Correlation']
corrdf0.head()

corrdf0.dropna(subset = ['Correlation'], inplace = True)
corrdf0['Correlation'] = round(corrdf0['Correlation'], 2)
```



	AMT_INCOME_TOTAL	AMT_ANNUITY	Correlation
82	AMT_GOODS_PRICE	AMT_CREDIT	0.99
167	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
83	AMT_GOODS_PRICE	AMT_ANNUITY	0.77
69	AMT_ANNUITY	AMT_CREDIT	0.76
68	AMT_ANNUITY	AMT_INCOME_TOTAL	0.40
125	DAYS_EMPLOYED	DAYS_BIRTH	0.35
55	AMT_CREDIT	AMT_INCOME_TOTAL	0.33
81	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.33
106	DAYS_BIRTH	CNT_CHILDREN	0.24
137	HOUR_APPR_PROCESS_START	REGION_POPULATION_RELATIVE	0.17

Correlation between Target 0 and Target 1

1) There is perfect correlation(0.98) between AMT_INCOME_TOTAL and AMT_GOODS_PRICE, wherein increase in one variable tends to increase in other variable as well. Here, more is the price on the goods to be paid by the consumer, more is the difficulty for him to pay earn income.

2) Similarly, there is a strong positive correlation(0.98) between AMT_ANNUITY & AMT_CREDIT as well. Here, higher is the loan annuity to be paid by the customer, more difficult, it is for the customer to avail credit from the bank

Also, the boxplot is comparatively short, suggesting that the values are evenly concentrated around the median.

3) Also, there is a strong correlation(0.75) between AMT_ANNUITY & AMT_INCOME_TOTAL, wherein less is the income of the client, more difficulty for the client in making the loan annuity payments to the bank.

4) There seems to be negative correlation(-0.18) between AMT_ANNUITY & CNT_CHILDREN. if the client has more no. of children, it is difficult for him to make pay regular loan payments to the bank.

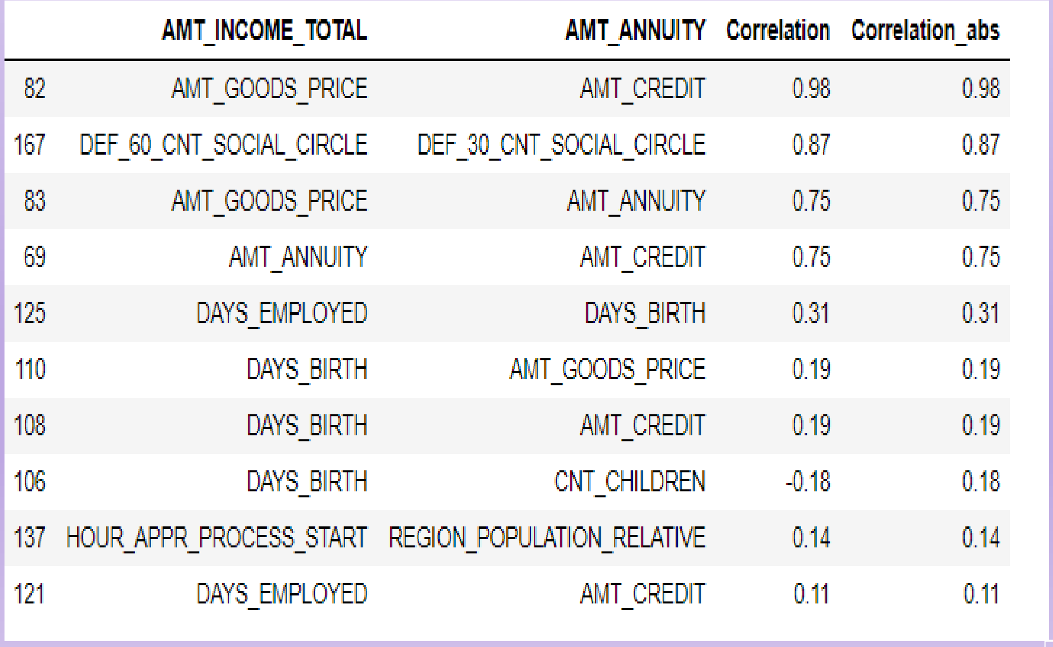
5) Additionally, there is a weak correlation(0.3) between AMT_INCOME_TOTAL & DAYS_EMPLOYED, wherein the employment of the consumer is not greatly impacting his loan application.’

6) Weak correlation(0.14) implies customers residing in densely populated regions have less likely to receive loan facility

```
# For Target=1 (client with payment difficulties)
corr1 = target1.corr()
corr1 = corr1.where(np.triu(np.ones(corr1.shape), k=1).astype(np.bool))
corr1

corrdf1 = corr1.unstack().reset_index()
corrdf1.columns = ['AMT_INCOME_TOTAL', 'AMT_ANNUITY',
'Correlation']
corrdf1.head()

corrdf1.dropna(subset = ['Correlation'], inplace = True)
corrdf1['Correlation'] = round(corrdf1['Correlation'], 2)
```



	AMT_INCOME_TOTAL	AMT_ANNUITY	Correlation	Correlation_abs
82	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
167	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
83	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
69	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
125	DAYS_EMPLOYED	DAYS_BIRTH	0.31	0.31
110	DAYS_BIRTH	AMT_GOODS_PRICE	0.19	0.19
108	DAYS_BIRTH	AMT_CREDIT	0.19	0.19
106	DAYS_BIRTH	CNT_CHILDREN	-0.18	0.18
137	HOUR_APPR_PROCESS_START	REGION_POPULATION_RELATIVE	0.14	0.14
121	DAYS_EMPLOYED	AMT_CREDIT	0.11	0.11

Univariate Analysis for Numerical Variables with Target = 0

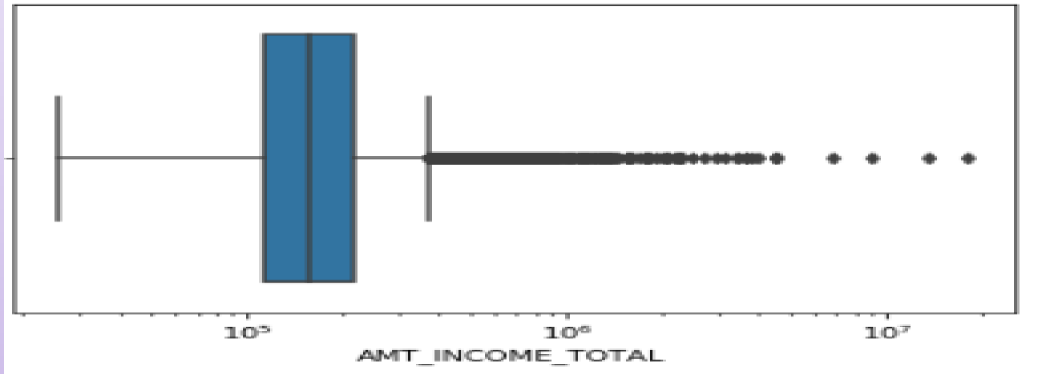
AMT_INCOME_TOTAL

From the graph, some outliers in the column "AMT_INCOME_TOTAL" are observed.

This signifies there are customers with very high salary and the others, or the larger crowd are evenly distributed around the mean.

This conclusion is that customers are capable to repay the loan back to the bank within the defined timelines.

```
# Box plotting for AMT_INCOME_TOTAL
sns.boxplot(x=target0['AMT_INCOME_TOTAL'])
plt.xscale('log')
plt.show()
```



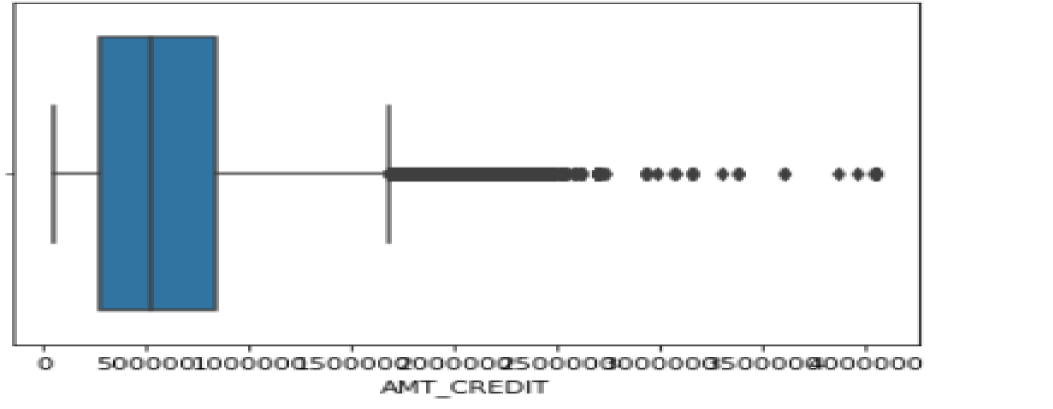
AMT_CREDIT

From the boxplot, there seems to be a huge difference between 3rd Quartile and the max value, leading to the conclusion of more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

It can be said that there are less percentage of borrowers, as observed in lower quartile. The longer upper whisker in the figure means that credit is varied amongst the upper quartiles. Additionally, values are concentrated more above the median value in lower quartiles as well. It concludes that more credit can be availed to the customers, with less credit loss to the bank .

A larger boxplot implies that credit amount is spread over quartiles

```
# Box plotting for AMT_CREDIT
sns.boxplot(x=target0['AMT_CREDIT'])
plt.show()
```



Univariate Analysis for Numerical Variables with Target = 0

AMT_ANNUITY

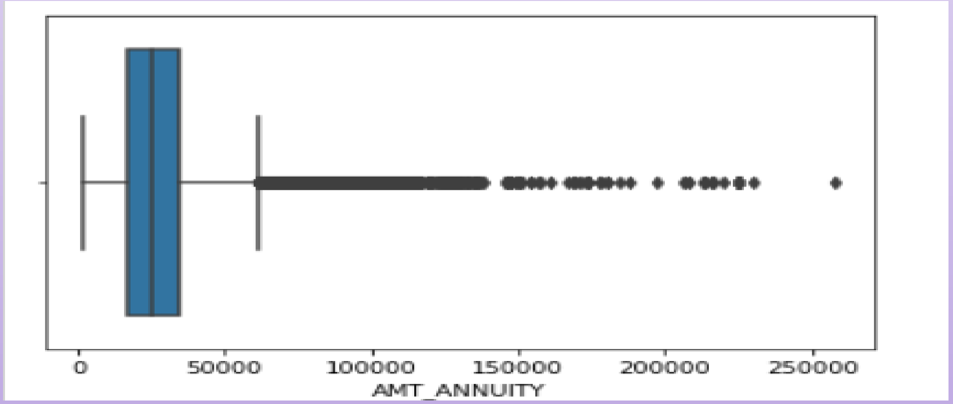
From the boxplot, there seems to more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

It can be said that there is less percentage of loan borrowers in lower quartiles. The values are concentrated more above the median value in lower quartiles as well.

Also, there is a greater probability of loan annuity payments on a regular basis to the bank.

# Boxplot for AMT_ANNUITY

```
sns.boxplot(x=df['AMT_ANNUITY'])
```



Univariate Analysis for Numerical Variables with Target = 1

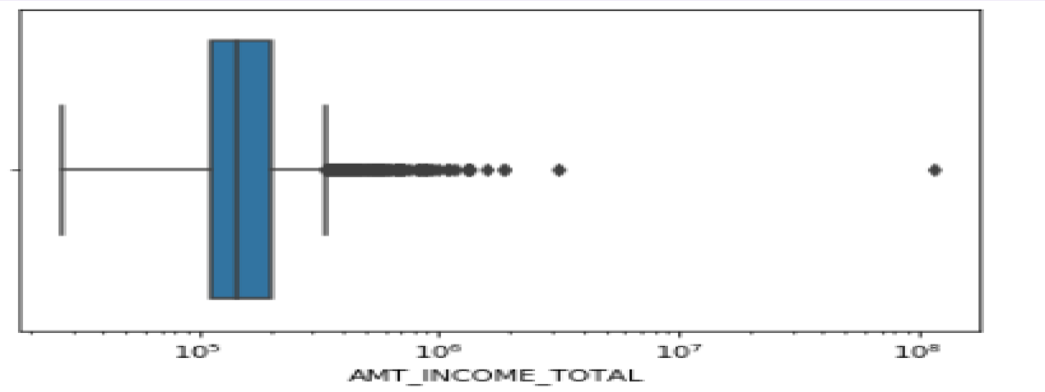
AMT_INCOME_TOTAL

From the boxplot, the It can be said that there are many customers with higher salary in higher quartile. However, the longer lower whisker in the figure means that income is varied amongst the lower quartiles.

Additionally, values are concentrated more above the median value in higher quartiles as well

There seems to be a huge difference between 3rd Quartile and the max value, leading to the conclusion of more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

```
# Box plotting for AMT_INCOME_TOTAL
sns.boxplot(x=target0['AMT_INCOME_TOTAL'])
plt.xscale('log')
plt.show()
```



AMT_CREDIT

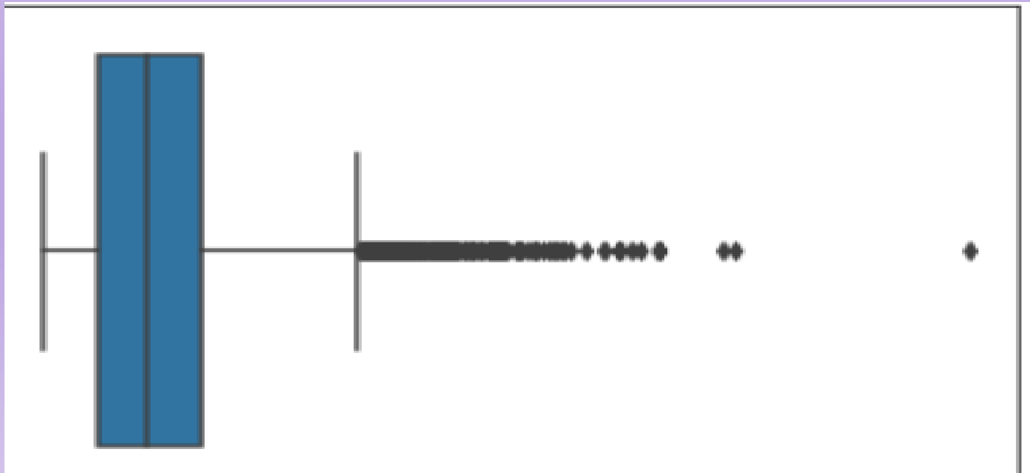
From the boxplot, it is observed can be said that many customers have availed credit as observed in lower quartile. However, the longer upper whisker in the figure means that credit is varied amongst the upper quartiles. The bank may have to consider credit loss if customers are unable to pay the loan.

Additionally, values are concentrated more above the median value in higher quartiles as well

There seems to be a huge difference between 3rd Quartile and the max value, leading to the conclusion of more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

.

```
# Box plotting for AMT_CREDIT
sns.boxplot(x=target0['AMT_CREDIT'])
plt.show()
```



Univariate Analysis for Numerical Variables with Target = 1

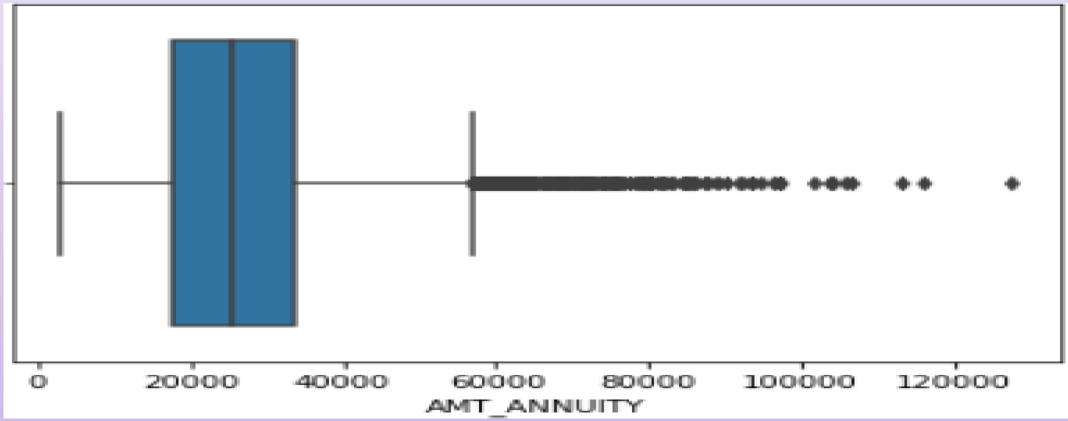
AMT_ANNUITY

From the boxplot, there seems to be a huge difference between 3rd Quartile and the max value, leading to the conclusion of more values concentrated in the 4th quartile. Also, values greater than 1.5(IQR) would lead to the conclusion of outliers.

Additionally, values are concentrated evenly around the median value as well. A larger boxplot implies that loan annuity is spread over quartiles

It is observed that percentage of customers having loan annuity is skewed more in lower quartiles, wherein there may be possibility of credit loss to the bank, if customers default on loan payments. However, the longer upper whisker in the figure means that annuity is varied amongst the upper quartiles.

```
# Box plotting for AMT_INCOME_TOTAL
sns.boxplot(x=target0['AMT_ANNUITY'])
plt.xscale('log')
plt.show()
```



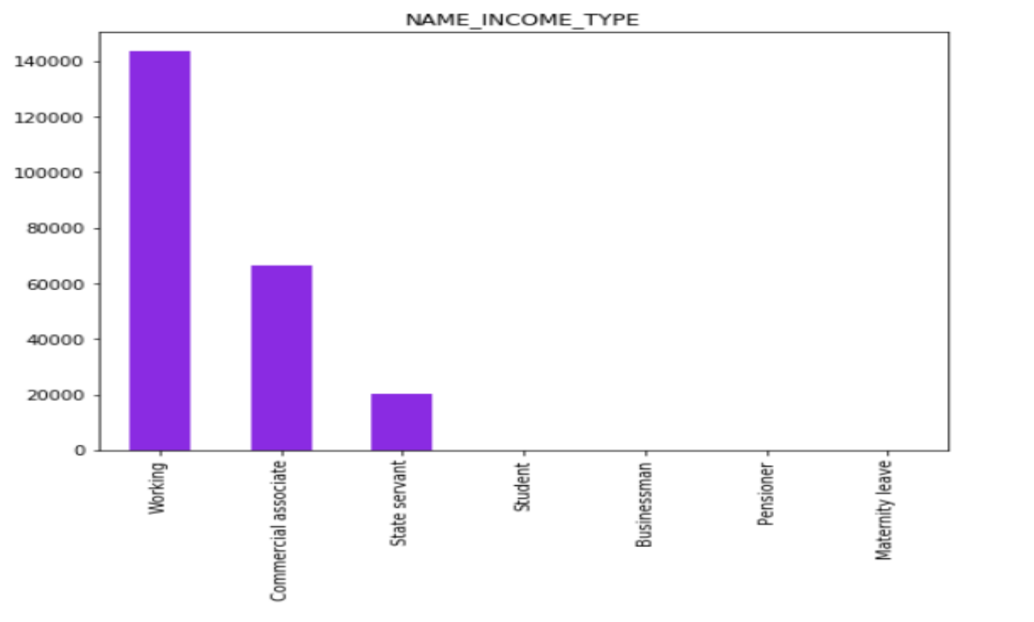
Univariate Analysis for Categorical Variables with Target = 0

NAME_INCOME_TYPE

The number of credits is higher for income type 'Working', 'Commercial Associate', 'State Servant’.

About 62% of the credit is availed to the income type ' Working' and 28% is availed to 'Commercial Associate’

Less no. of credits for income type 'Student’ ,'Businessman’, 'Pensioner'



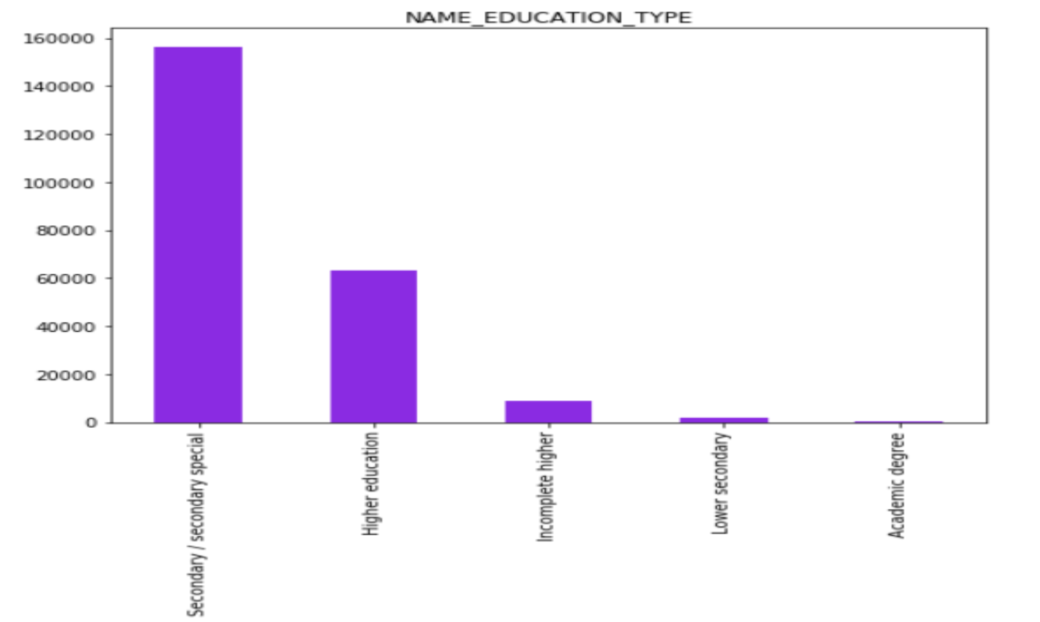
NAME_EDUCATION_TYPE

More educated customer namely 'Secondary / secondary special', 'Higher Education' have incomes.

Credits are mostly available for educated customer with 'Secondary / secondary special' or 'Higher Education' as compared to customer with lower secondary or incomplete higher education

About 95% of customers have higher secondary or more degree

.



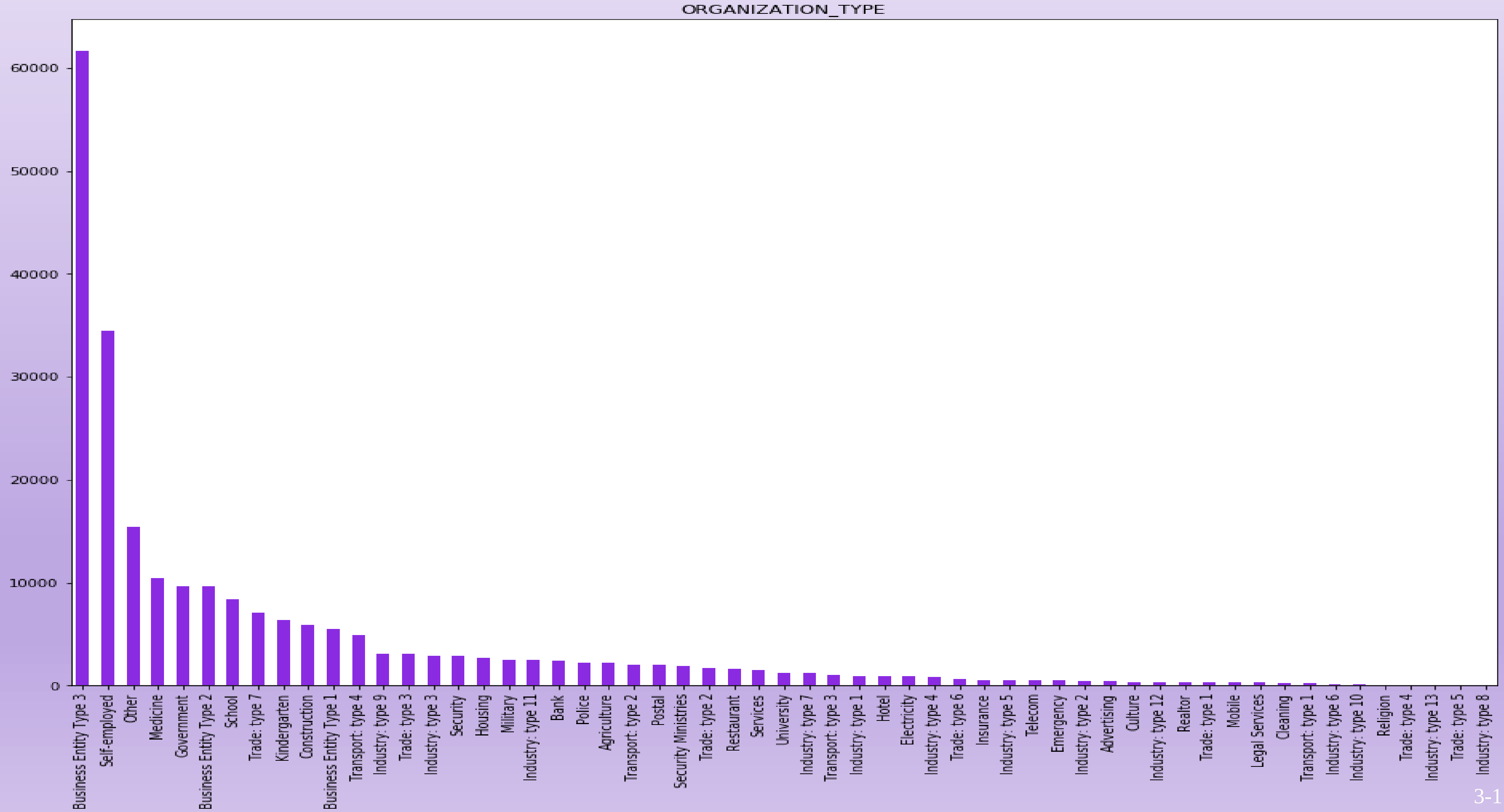
Univariate Analysis for Categorical Variables with Target = 0

ORGANIZATION_TYPE

The number of credits is higher for income type 'Working', 'Commercial Associate', 'State Servant’.

About 62% of the credit is availed to the income type ' Working' and 28% is availed to 'Commercial Associate’

Less no. of credits for income type 'Student’ ,'Businessman’, 'Pensioner'



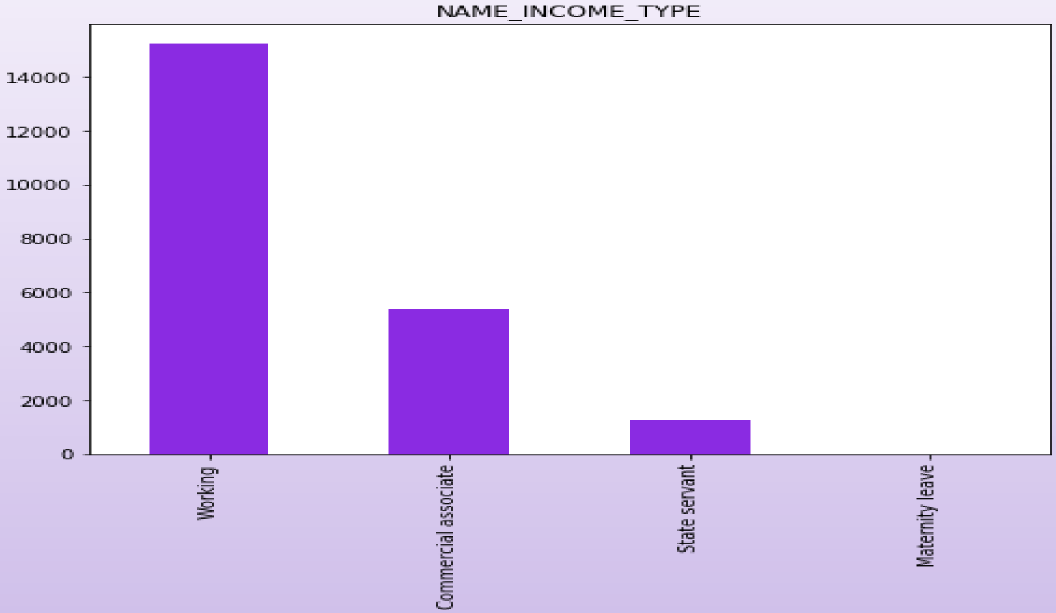
Univariate Analysis for Categorical Variables with Target = 1

NAME_INCOME_TYPE

The number of credits is higher for income type 'Working', 'Commercial Associate','State Servant'.

About 62% of the credit is availed to the income type ' Working' and 28% is availed to 'Commercial Associate'

Less no. of credits for income type 'Student','Businessman','Pensioner'



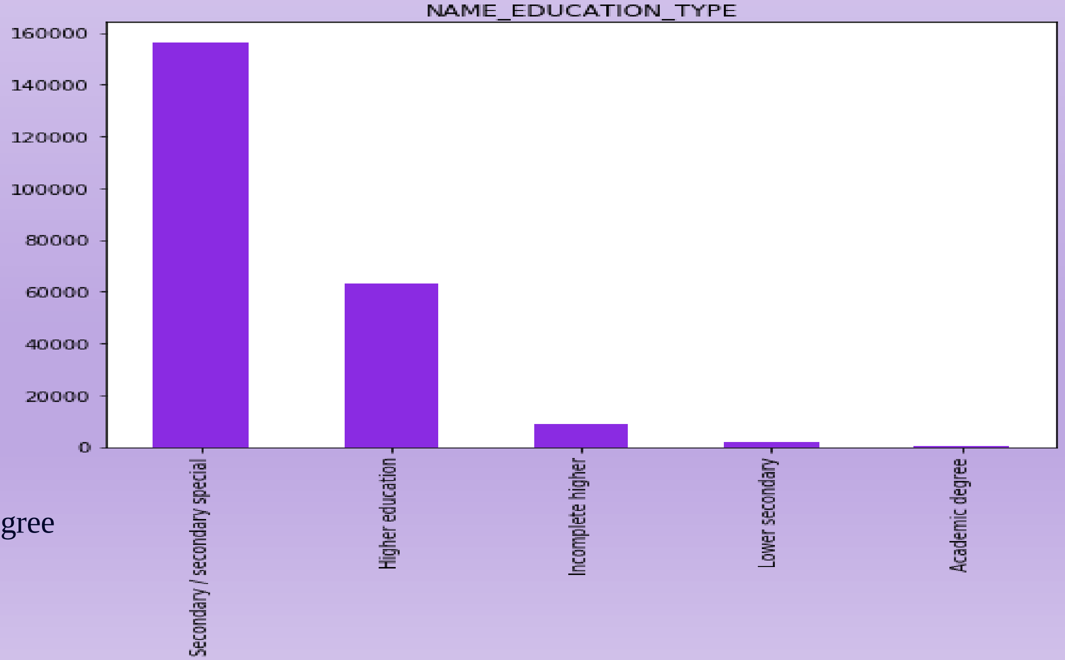
NAME_EDUCATION_TYPE

.

More educated customer namely 'Secondary / secondary special','Higher Education' have incomes.

Credits are mostly available for educated customer with 'Secondary / secondary special' or 'Higher Education' as compared to customer with lower secondary or incomplete higher education

About 95% of customers have higher secondary or more degree



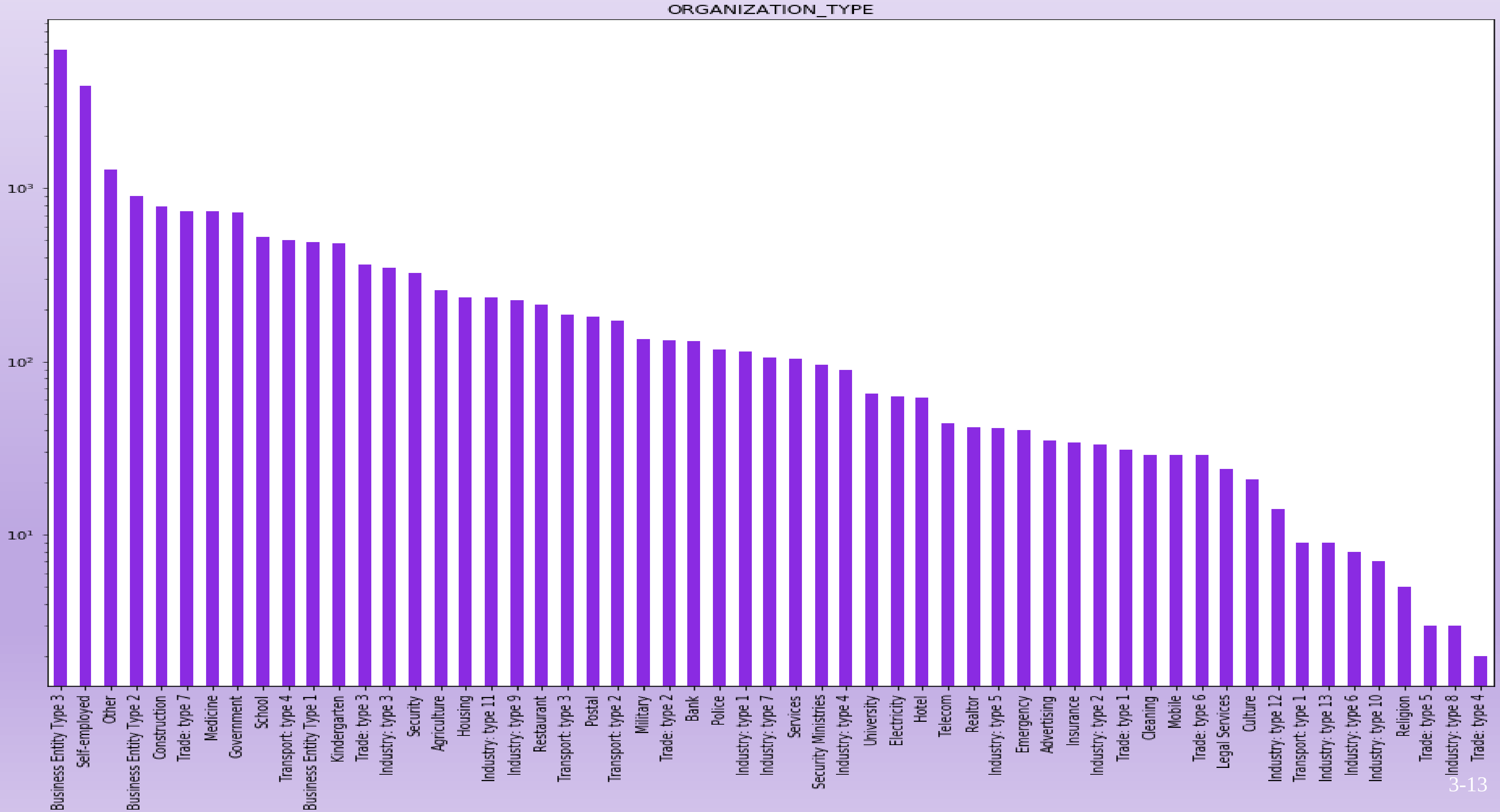
Univariate Analysis for Categorical Variables with Target = 1

ORGANIZATION_TYPE

Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.

Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

Same as type 0 in distribution of organization type.



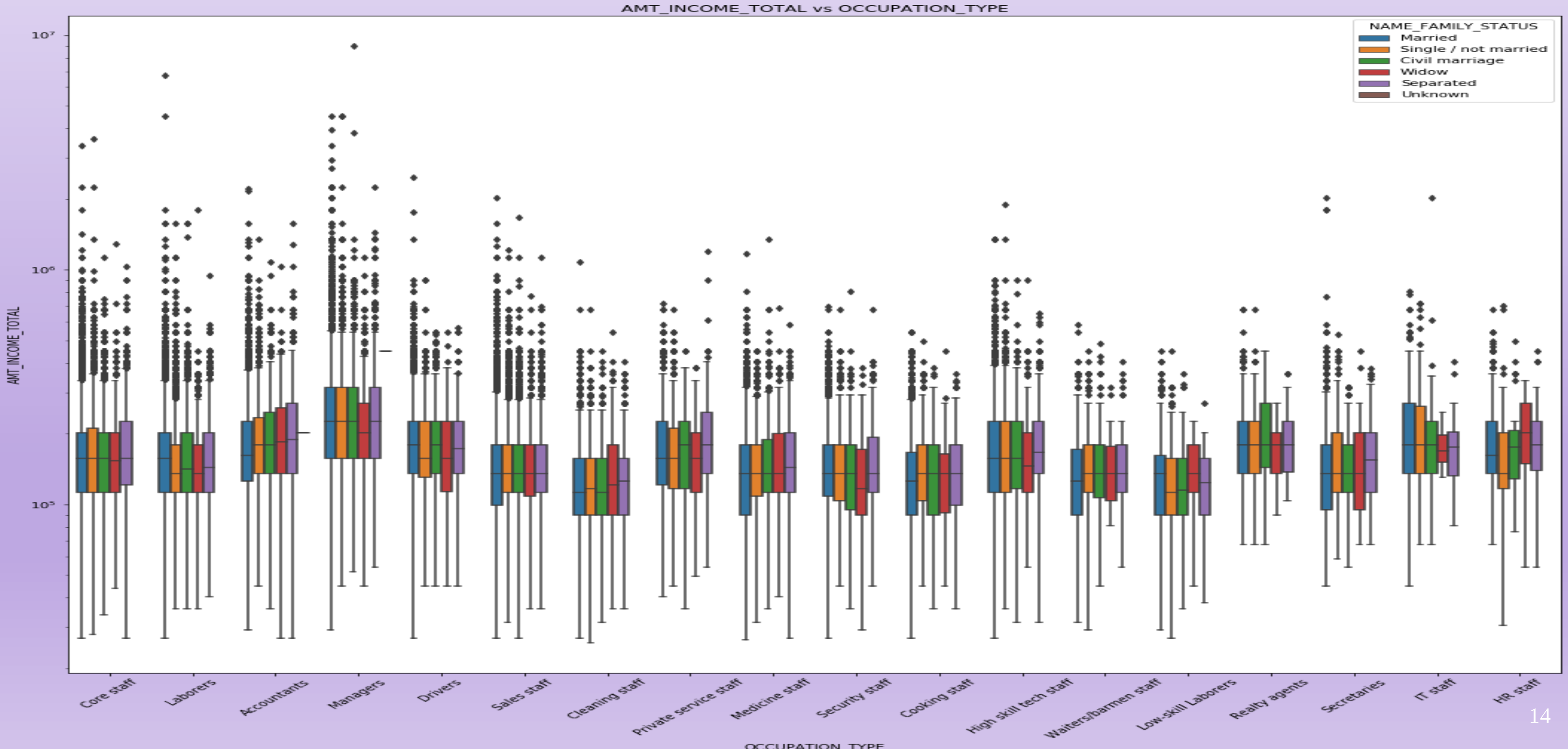
Bivariate Analysis For Numerical For Variables With TARGET = 0 & TARGET = 1

TARGET = 0 AMT_INCOME_TOTAL vs OCCUPATION_TYPE

Inference drawn from the above boxplot states that Managers earn the most followed by Realty Agents, IT staff, High skill Tech staff, Accountants and Core Staff.

For Managers, the income amount is mostly equal with family status. Also, there are more outliers in Managers, High skill tech staff, Accountants, core staff and laborers implying varied income range within mentioned occupation type.

The average income is equally distributed for Managers, Core staff, while single or unmarried men have higher average income and married men the lowest average income



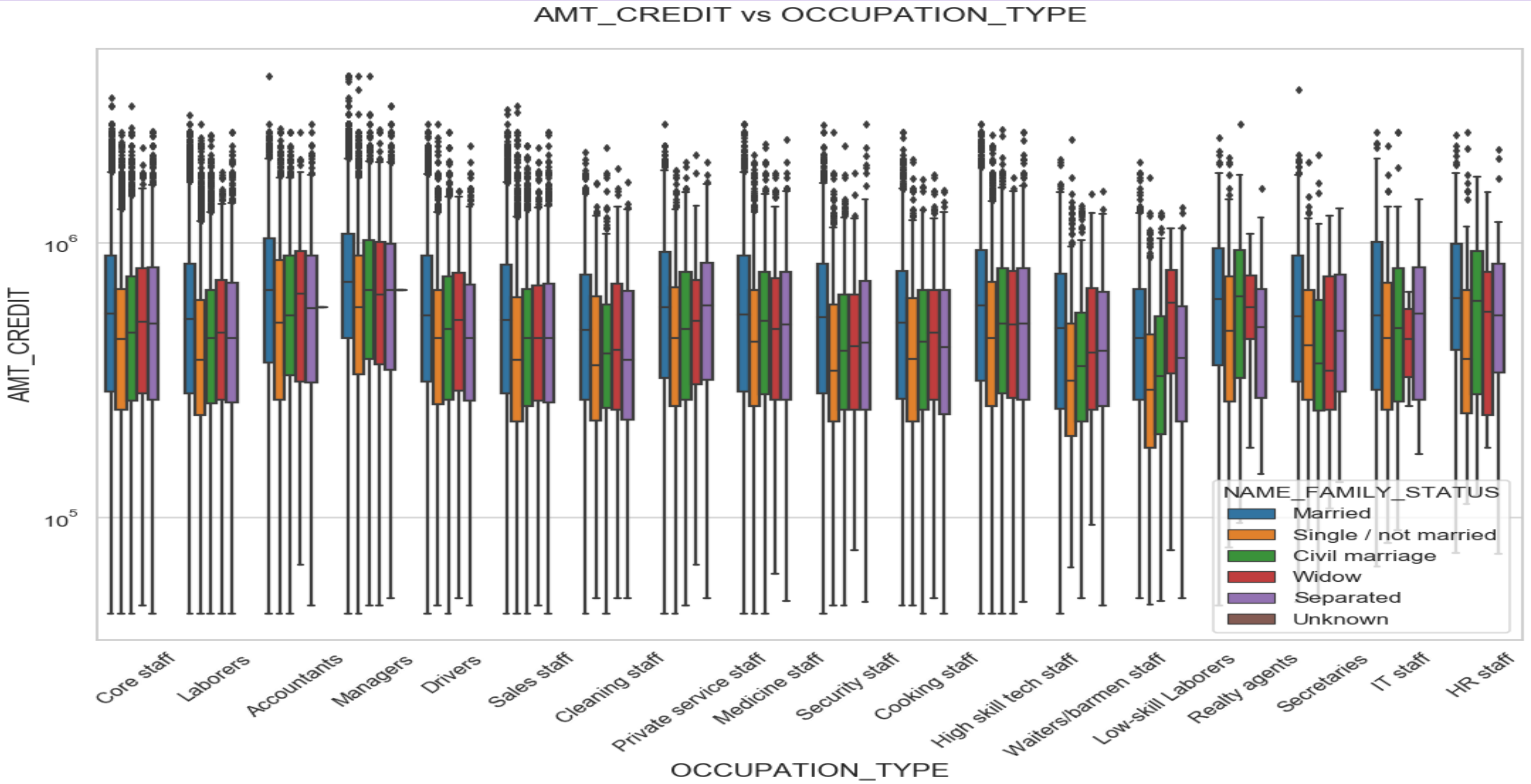
Bivariate Analysis For Numerical For Variables With TARGET = 0 & TARGET = 1

TARGET = 0 AMT_CREDIT vs OCCUPATION_TYPE

The above boxplot shows that the single or unmarried men have less credit, followed by civil marriage and separated, while married men are given more credits.

Manages, Accountants, Private Service staff, Realty agents and IT staff earn the most and are availed more credit within the occupation types while waiters/barmen and low skill laborers have less credits

There are outliers in outstanding credits for Managers, Accountants and Sales staff.



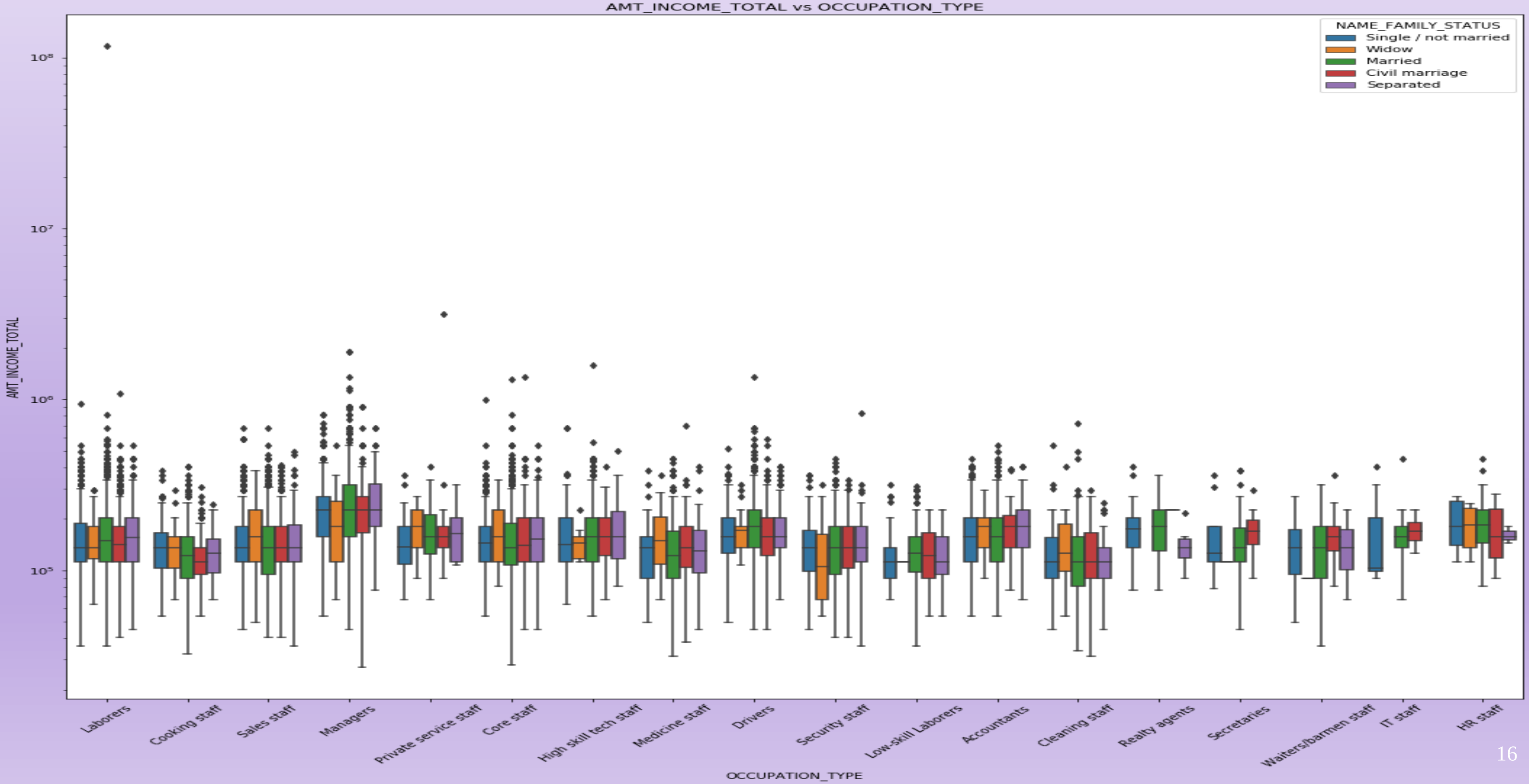
Bivariate Analysis For Numerical For Variables With TARGET = 0 & TARGET = 1 Cont…

TARGET = 1 AMT_INCOME_TOTAL vs OCCUPATION_TYPE

The above boxplot shows that Managers are the ones who earn the most, followed by HR staff, laborers and other categories.

Family types are almost equally distributed amongst the occupation types except for the ones in the IT industry, where there are no one in Separated and none are widows.

There are outliers in income for all occupation types but one standing out in Laborers is fascinating.

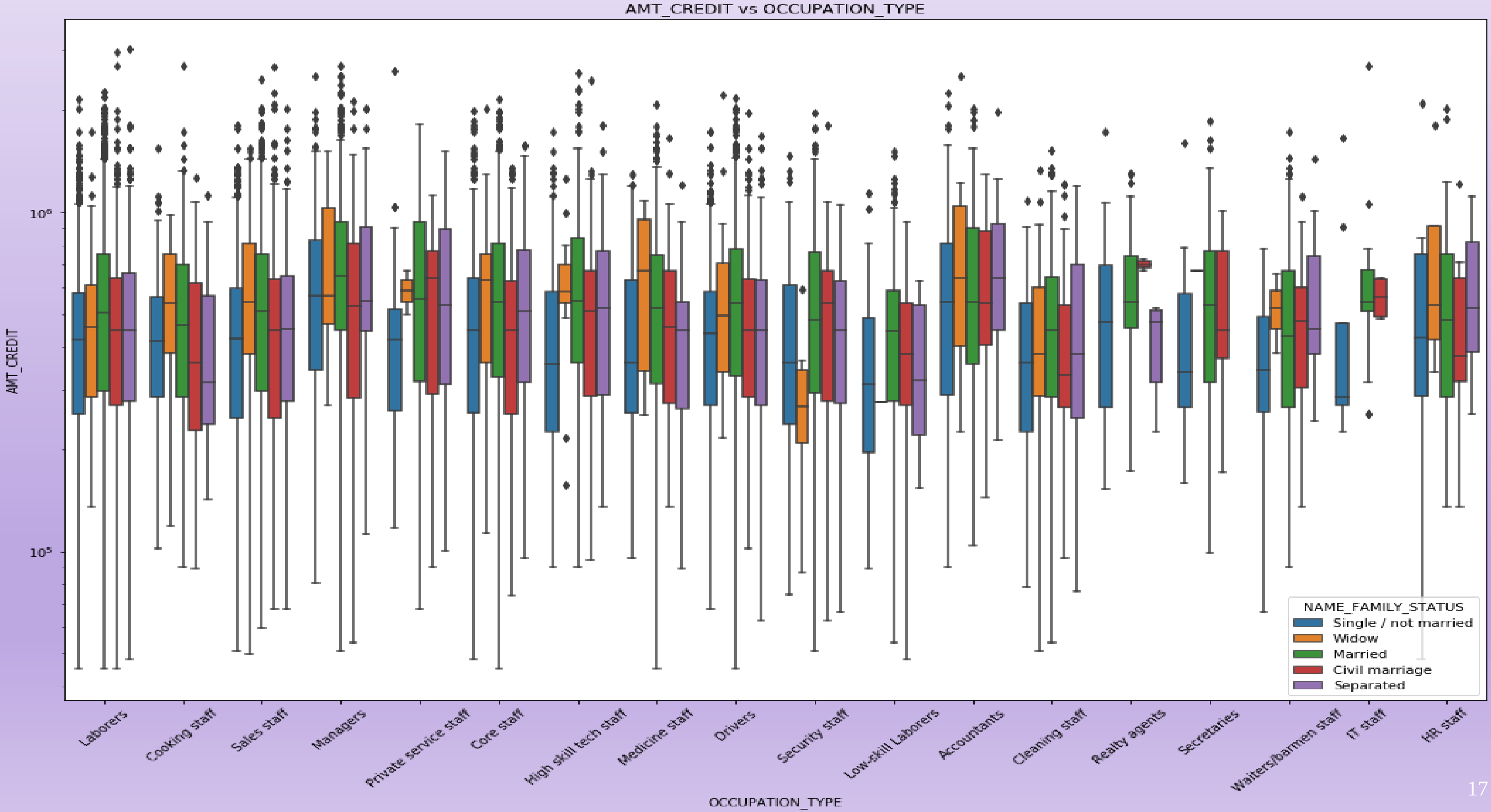


Bivariate Analysis For Numerical For Variables With TARGET = 0 & TARGET = 1 Cont…

TARGET = 1 AMT_CREDIT vs OCCUPATION_TYPE

The above boxplot shows that even though Managers are ones earning the most are in debt, Accounts are the ones who have more debts than managers in the segment Target1.

There are outliers in every occupation with the most in laborers and the least in the High skill tech staffs.



BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES (NAME_EDUCATION_TYPE, NAME_CONTRACT_TYPE) WITH TARGET = 0

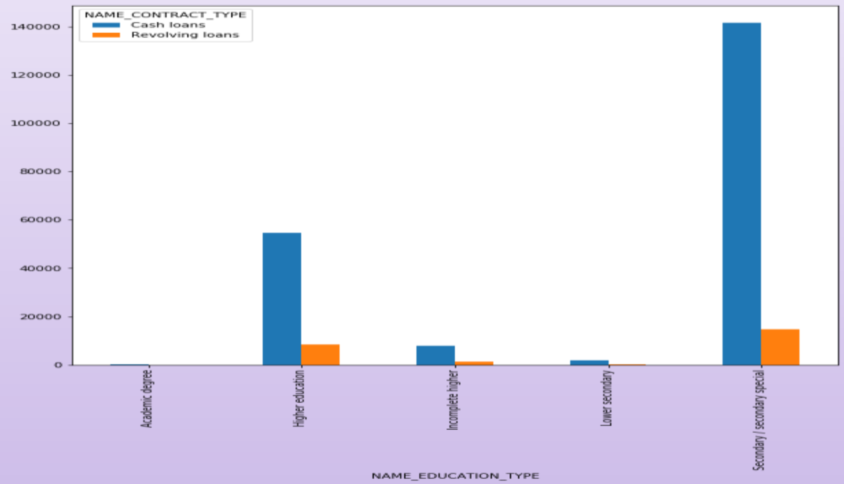
Customers with higher education or secondary/secondary special education have preferred loans(cash loans or revolving loans) as compared with lower educated customers

Customers with higher education or secondary/secondary special education have preferred to avail cash loans as compared to revolving loans

Customers with academic degree have taken less loans(cash or revolving loans)

TARGET = 0

NAME_EDUCATION_TYPE & NAME_CONTRACT_TYPE



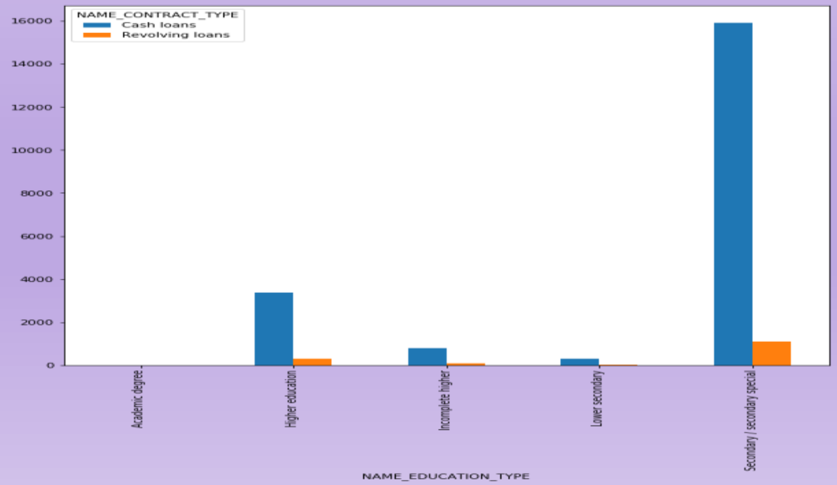
Customers with higher education or secondary/secondary special education have preferred loans(cash loans or revolving loans) as compared with lower educated customers and will face difficulty in repayment of the loans

Customers with higher education or secondary/secondary special education have preferred to avail cash loans as compared to revolving loans imputing that

Customers with academic degree have taken less cash loans, ensuring less credit loss to the banks.

TARGET = 0

NAME_EDUCATION_TYPE & NAME_CONTRACT_TYPE



BIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES (NAME_EDUCATION_TYPE, NAME_CONTRACT_TYPE) WITH TARGET = 0

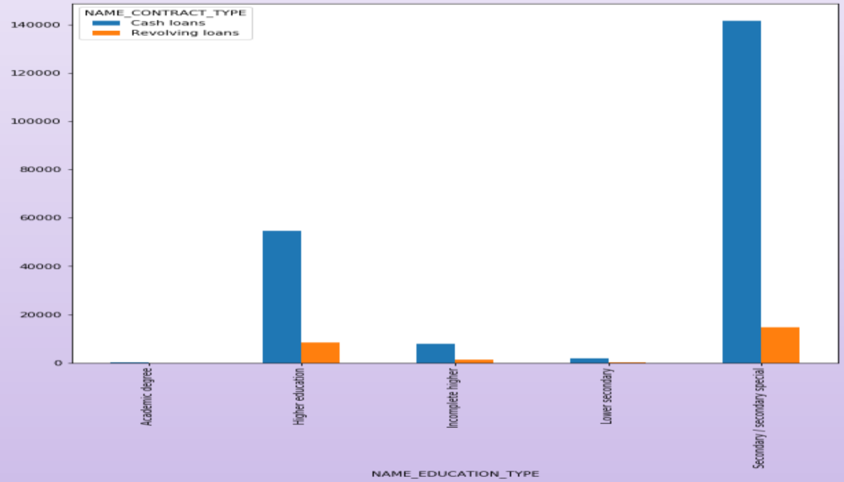
Customers with higher education or secondary/secondary special education have preferred loans(cash loans or revolving loans) as compared with lower educated customers

Customers with higher education or secondary/secondary special education have preferred to avail cash loans as compared to revolving loans

Customers with academic degree have taken less loans(cash or revolving loans)

TARGET = 0

NAME_EDUCATION_TYPE & NAME_CONTRACT_TYPE



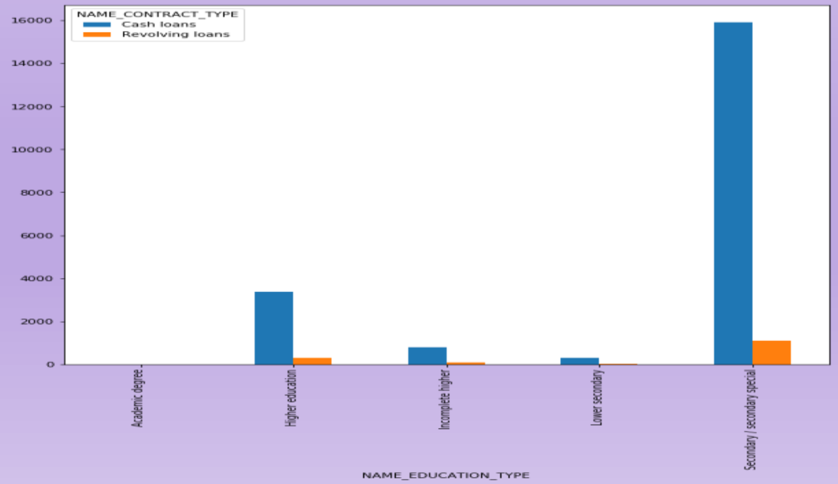
Customers with higher education or secondary/secondary special education have preferred loans(cash loans or revolving loans) as compared with lower educated customers and will face difficulty in repayment of the loans

Customers with higher education or secondary/secondary special education have preferred to avail cash loans as compared to revolving loans imputing that

Customers with academic degree have taken less cash loans, ensuring less credit loss to the banks.

TARGET = 1

NAME_EDUCATION_TYPE & NAME_CONTRACT_TYPE

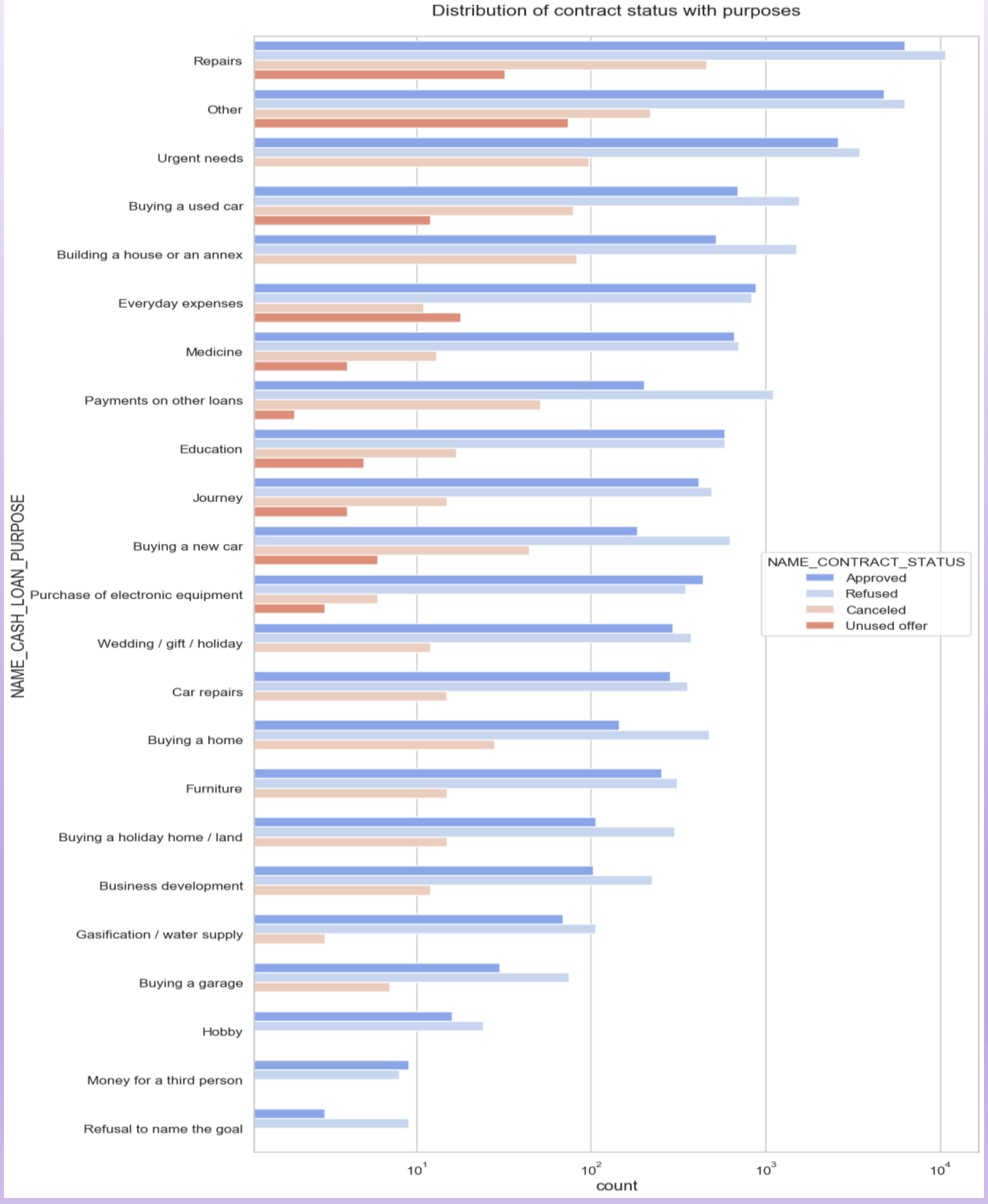


Performing Univariate & Bivariate Analysis for Merged Data

UNIVARIATE ANALYSIS

NAME_CONTRACT_STATUS

Distribution of contract status with purposes



Most number of Approved and Rejections of loans are from 'repairs'.

Both Approved and Rejections in Education are equal

Both Approved and Rejections for Everyday Expenses and Medicine are almost equal.

Refused loans are the smallest in Money for a third party

Conclusions - 1

NAME_CONTRACT_TYPE VS NAME_CONTRACT_STATUS

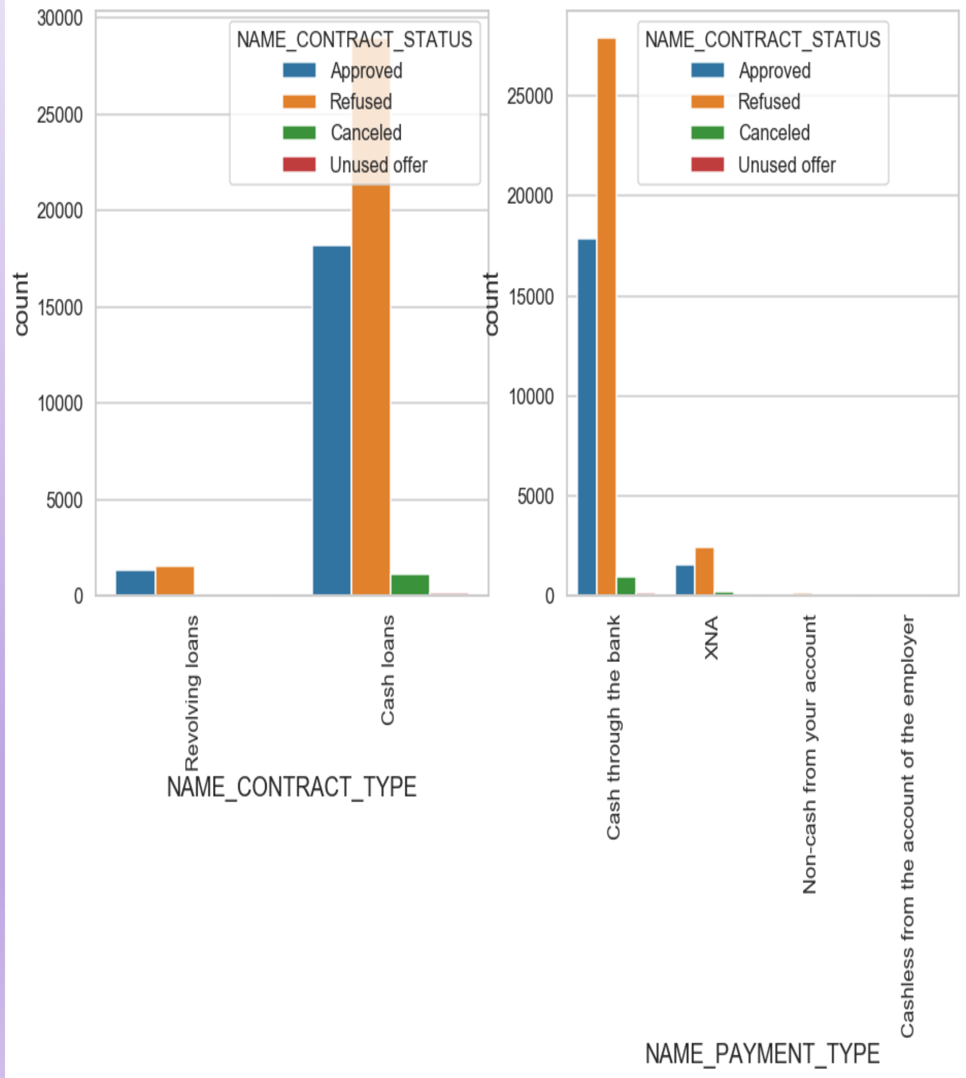
1. Cash loans are significantly higher as compared to Revolving loans.
2. In both Cash loans and Revolving loans there are more rejections than approvals.
3. For Cash loans 59.33% applications are refused while only 38.02% applications are approved.
4. For Revolving loans 59.32% applications are refused while only 40.67% applications are approved.

NAME_PAYMENT_TYPE VS NAME_CONTRACT_STATUS

1. Cash through bank is the most preferred payment method chosen by the customers in all the previous applications.
2. There are more rejections of loan applications as compared to Approvals.
3. For applications with payment method as Cash through bank the rejection percentage is 59.30% while 36.66% applications are approved.

BIVARIATE ANALYSIS

NAME_CONTRACT_TYPE NAME_PAYMENT_TYPE



Conclusions - 2

NAME_TYPE_SUITE VS NAME_CONTRACT_STATUS

1. About 83.216% of Customer applying for loans are Unaccompanied followed by 12.02% accompanied by family.

2. 60.761% of applications of customers who are Unaccompanied are rejected while 37.39% are approved.

3. 56.625% of applications of customers with family are rejected while 40.45% are approved.

NAME_CLIENT_TYPE VS NAME_CONTRACT_STATUS

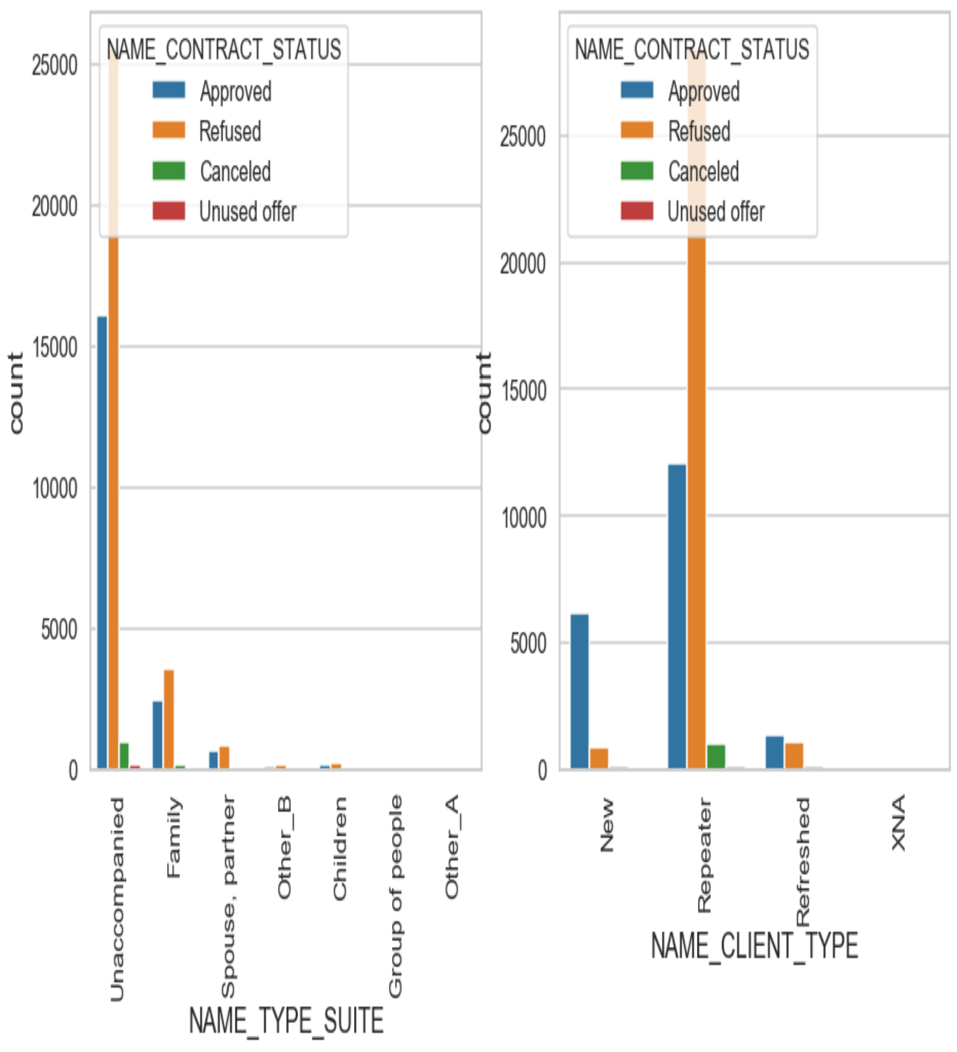
1. It is noticed that 81.013% of customers are Repeat customers, while 13.94% are new customers.

2. The rejection percentage for Repeat customers is 67.21 and approval is 28.80

BIVARIATE ANALYSIS

NAME_TYPE_SUITE

NAME_CLIENT_TYPE



Conclusions - 3

CHANNEL_TYPE VS NAME_CONTRACT_STATUS

1. About 70.79% of customers are applying for loan through "Credit & Cash Offices" while 11.80% are through from AP+ Cash loans and 7.44% are from Countrywide

2. 60.43% of customers applying for loan through Cash and Credit offices are rejected while 38.46% are approved.

3. 65.90% of customers applying for loan through AP+ Cash loans are rejected while 41.19% are approved.

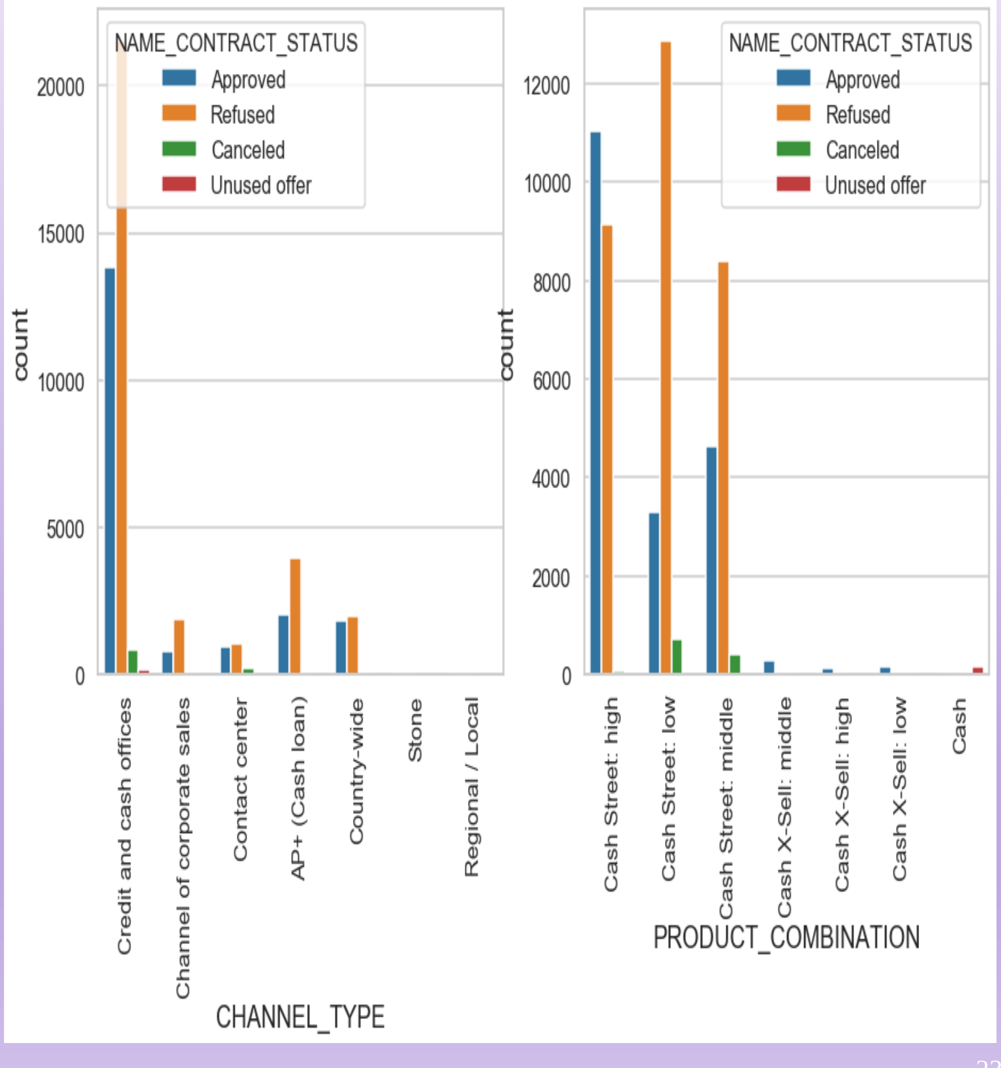
PRODUCT_COMBINATION VS NAME_CONTRACT_STATUS

1. The product Combinations of the customer who have applied for loans is 39.33% for Cash Street: High, while 32.79% for Cash Street: Low and 26.16% for Cash Street: Middle

2. There are more rejections observed on both Cash Street: Low & Cash Street: Middle than the approvals, while for Cash Street: High there are more approvals than rejections

BIVARIATE ANALYSIS

CHANNEL_TYPE PRODUCT_COMBINATION



Final Conclusions

1. Customers applying for cash loans who are unaccompanied and preferring payment type as cash through bank have higher percentage of Rejections.

2. Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

3. Customers applying for loan through AP+ Cash loans and who are accompanied by family members should be given preference on approvals.

4. Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.