EDA CREDIT ASSIGNMENT

BY SATYA IYER

Area’s Of Study

1. Problem Statement
2. Proposed Solution
3. Reading Data and Cleaning Data.
4. Data Visualization using Plots.
   a) Univariate Analysis
      i. Numeric Variable Outlier Analysis .
      ii. Categorical Variable Outlier Analysis.
   b) Bivariate Analysis
      i. TARGET vs all Variables
         a) With Respect to Defaulter.
         b) With Respect to Non Defaulter
      ii. CONTRACT STATUS vs all Variables
         a) With Respect to Approved
         b) With Respect to Cancelled
         c) With Respect to Refused
   c) HEATMAPS.(Multivariate Analysis)

5. Checking Data Imbalance

6. Conclusion

PROBLEM STATEMENT

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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **DEFAULTER :** The client with payment difficulties
- **NON DEFAULTER :** All other cases when the payment is paid on time.

For this we Analyze ***Application Dataset Model.***

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer**: Loan has been cancelled by the client but at different stages of the process.

For this we Analyze ***Previous Application Dataset Model.***

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

For this we Analyze ***Total Previous Application Dataset Model.***

PROPOSED SOLUTION

In this assignment, we are applying the techniques of the Exploratory Analysis, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used by bank to minimize the risk of losing money .

We are also using various Plots for graphical representation of data ,and on the basis of that we can analyze what are variables on which bank can take risk or approve for loan.

2. READING AND CLEANING DATA

1. Reading CSV Dataset Using Panda Libraries

a). Application_data.csv

```
inp0 = pd.read_csv('application_data.csv')
inp0.tail(5)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_
307506	456251	0	Cash loans	M	N	N	0	157500.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	
307508	456253	0	Cash loans	F	N	Y	0	153000.0	
307509	456254	1	Cash loans	F	N	Y	0	171000.0	
307510	456255	0	Cash loans	F	N	N	0	157500.0	

5 rows × 122 columns

b). Previous Application_data.csv

```
#Reading PREVIOUS APPLICATION DATA
inp1 = pd.read_csv('previous_application.csv')
inp1.tail(5)
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
1670209	2300464	352015	Consumer loans	14704.290	267295.5	311400.0	0.0	267295.5
1670210	2357031	334635	Consumer loans	6622.020	87750.0	64291.5	29250.0	87750.0
1670211	2659632	249544	Consumer loans	11520.855	105237.0	102523.5	10525.5	105237.0
1670212	2785582	400317	Cash loans	18821.520	180000.0	191880.0	NaN	180000.0
1670213	2418762	261212	Cash loans	16431.300	360000.0	360000.0	NaN	360000.0

5 rows × 37 columns

c). Total of Previous Application_data.csv

```
inp_total = pd.merge(left=inp0 , right=inp1, how='inner', on='SK_ID_CURR')
```

```
inp_total.shape
```

```
inp_total.tail(5)
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	A
1061184	456255	0	Cash loans	FEMALE	No	No	0	157500.0	
1061185	456255	0	Cash loans	FEMALE	No	No	0	157500.0	
1061186	456255	0	Cash loans	FEMALE	No	No	0	157500.0	
1061187	456255	0	Cash loans	FEMALE	No	No	0	157500.0	
1061188	456255	0	Cash loans	FEMALE	No	No	0	157500.0	

5 rows × 46 columns

Important Point: (Variable Name Created)

1. inp0 = application_data
2. Inp1=previous_application_data
3. Inp_total = joining dataset

4. Cleaning Of Data:

My Approach of handling Null, Missing, XAP, XAN values.

Fetching records for null values more than 40% which can be dropped.

1.Fetching records for null values less than or equal to 40% which can be filled up by fillna method and substitute suitable values and further cleaned .If it is a numeric variable then I am substituting **MEDIAN** value and if it is categorical then substituting **MODE()[0]** value.

2.Replace all Abbreviated value to Full name like(**Y – YES,N– No** etc.)

3.Also Correcting the Values like(**canceled – Cancelled** etc.)

4.Dropping all unnecessary fields which are not required for our analysis.

5.In previous application data we have XAN,XAP values which are not mentioned values. For that I fallowed a approach ,of first converting all XAN and XAP to **np.NaN** value using **Replace().** Then in Future study we can either drop those records are using substitution we can substitute some values and handle the data model.

6.Finally Shape of our data in total.

```
inp_total.shape
```

(1061189, 46)

4. Data Visualization using Plots.

Before Visualization the data we have to group the fields according to its datatypes, for that we creating two different lists one contains Numerical Variable (dtypes :Int, float) and another contains Categorical Variables (dtypes: category, Object).On the bases of these both lists we further analyze the data model into **Univariate**(Analyze Single Variable) ,**Bivariate** (Analyze two variable with respect to each other) , **Multivate**(Analyze more than two variables at a time.) Analysis.

I. Application_data.csv :

```
### Now creating list of numerical variables for outlier detection and finding insights

list_of_int_val =[]
for i in inp0:
    if (inp0[i].dtypes == 'int64') or (inp0[i].dtypes == 'float64'):
        list_of_int_val.append(i)
print(list_of_int_val)
```

```
['SK_ID_CURR', 'TARGET', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'HOUR_APPR_PROCESS_START']
```

```
# created and fetched a list of categorical data feild from our dataset for easy Analysing.
list_of_categorical_data = []

for i in inp0:
    if(inp0[i].dtypes == "object"):
        list_of_categorical_data.append(i)
print(list_of_categorical_data)
```

```
['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE']
```

II. Previous_ Application_data.csv :

```
### Now creating list of numerical variables for outlier detection and finding insights
list_of_int_val1 =[]
for i in inp1:
    if (inp1[i].dtypes == 'int64') or (inp1[i].dtypes == 'float64'):
        list_of_int_val1.append(i)
print(list_of_int_val1)
```

```
['SK_ID_PREV', 'SK_ID_CURR', 'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'DAYS_DECISION', 'CNT_PAYMENT']
```

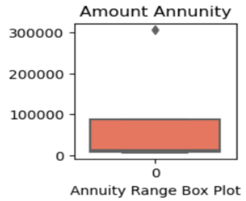
```
# list of Categorical Variable Datatype
list_of_cat =[]
for i in inp1:
    if (inp1[i].dtypes == 'object') :
        list_of_cat.append(i)
print(list_of_cat)
```

```
['NAME_CONTRACT_TYPE_PREV', 'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS', 'NAME_PAYMENT_TYPE', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE', 'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE', 'CHANNEL_TYPE', 'NAME_SELLER_INDUSTRY', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION']
```

4. (i). Analyze Numerical Variables for Outlier

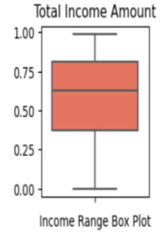
I. Application_data.csv :

1. .inp0 ['AMT_ANNUITY'] :



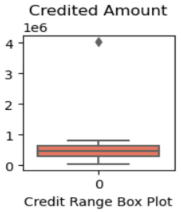
Here also data is in continous fashion ,not much outlier present.

2.inp0 ['AMT_INCOME_TOTAL'] :



Here we can observe that on Y axis we have divided the values in quantile ranging from 0.0 to 0.99 and plotted againt it.We can clearly see that max value of Income ranges from 40% to 80%.There is also an outlier lies on max 0.99%.

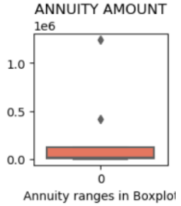
3.inp0 ['AMT_CREDIT'] :



Here we can see that mean,min,max having values in continous fashion .Hence no outlier present in it.

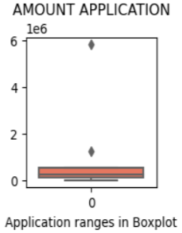
II. Previous_Application_data.csv :

1. .inp1 ['AMT_ANNUITY'] :



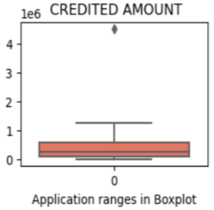
Data is in Continous fashion .No as such outlier present in Loan Annuity in Previous application.

2. .inp1 ['AMT_APPLICATION'] :



Data is in Continous fashion .No as such outlier present For how much credit did client ask on the previous application.

3 .inp1 ['AMT_CREDIT'] :

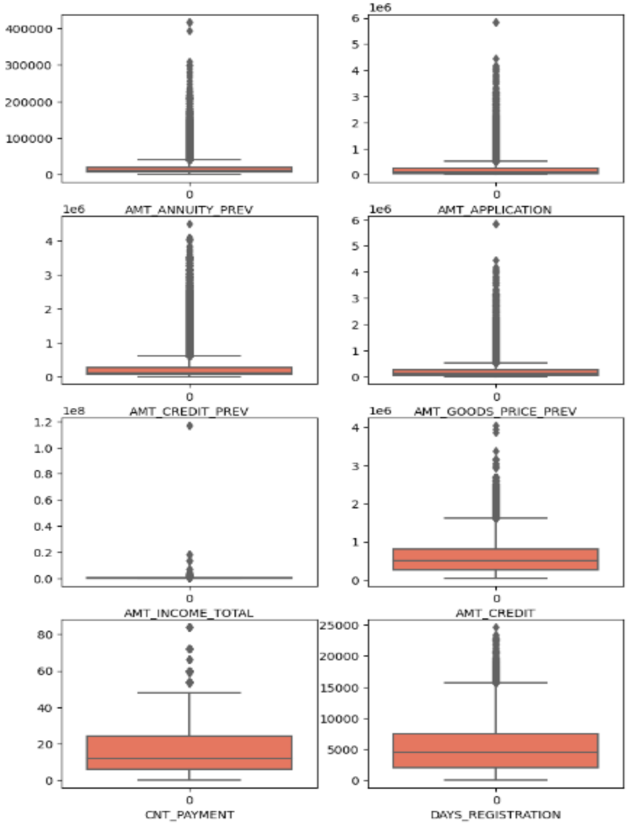


Data is in Continous fashion .No as such outlier present For how much credit did client ask on the previous application.

III. Total_Previous_Application_data.csv :

```
list_to_study1 = ["AMT_ANNUITY_PREV","AMT_APPLICATION","AMT_CREDIT_PREV","AMT_GOODS_PRICE_PREV","AMT_INCOME_TOTAL","AMT_CREDIT",
                  "CNT_PAYMENT","DAYS_REGISTRATION"]
```

```
fig, axes = plt.subplots(4, 2,figsize=(8,13))
for item,ax in zip(list_to_study1, axes.flatten()):
    sns.boxplot(data=inp_total[item],ax=ax,palette = "Reds")
    ax.set_xlabel(item)
plt.show()
```

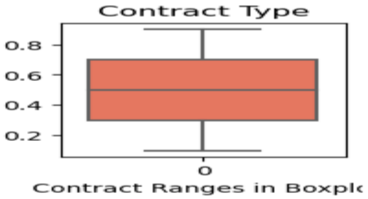


Here We can analyse the **AMT_CREDIT ,AMT_CREDIT_PREV, AMT_GOODS_PRICE_PREV, CNT_PAYMENT, DAYS_REGISTRATION, AMT_INCOME_TOTAL, AMT_APPLICATION** by joining both the dataset together.As such no outlier is present.Data is in continous fashion as we already checked both dataset individually.

4. (ii). Analyze Categorical Variables for Outlier

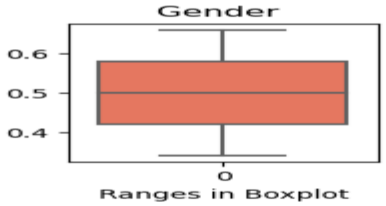
I. Application_data.csv :

1. .inp0 [‘CONTRACT_TYPE’] :



Data is in Continous fashion No outlier Present.

2.inp0 [‘CODE_GENDER’] :



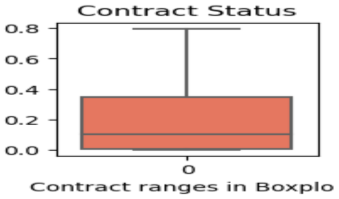
Data is in Continous fashion No outlier Present.

3.inp0 [‘NAME_INCOME_TYPE’] :

Data is in Continous fashion Very rare outlier Present.

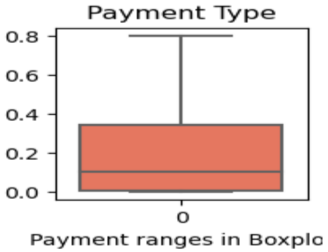
II. Previous_Application_data.csv :

1. .inp1 [‘CONTRACT_STATUS’] :



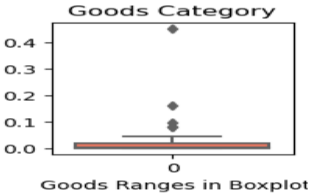
Data is in Continous fashion .No as such outlier present.

2. .inp1 [‘NAME_PAYMENT_TYPE’] :



Data is in Continous fashion .No as such outlier present.

3 .inp1 [‘NAME_GOODS_CATEGORY] :

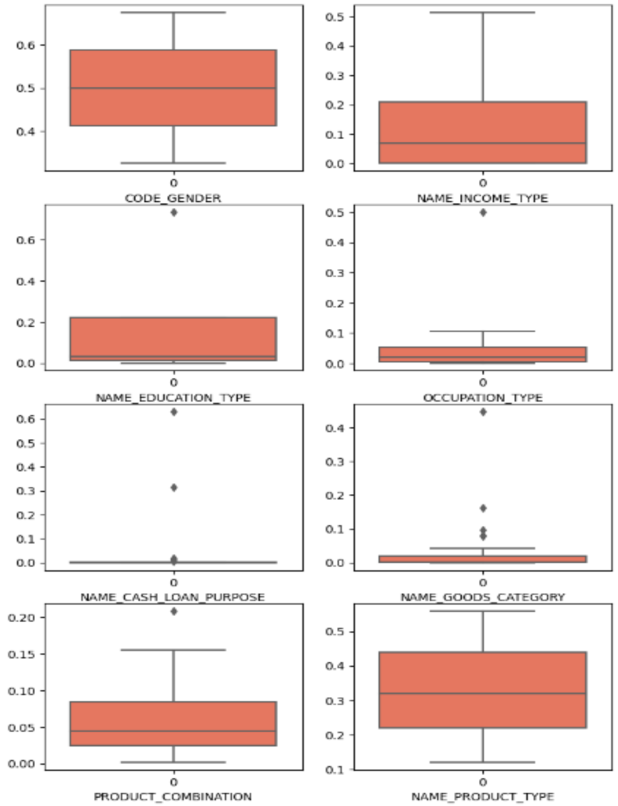


Data is in Continous fashion .But some outlier present.

III. Total_Previous_Application_data.csv :

```
list_to_study2 = ["CODE_GENDER","NAME_INCOME_TYPE","NAME_EDUCATION_TYPE","OCCUPATION_TYPE","NAME_CASH_LOAN_PURPOSE","NAME_GOODS_(
                 "PRODUCT_COMBINATION","NAME_PRODUCT_TYPE"]
```

```
fig, axes = plt.subplots(4, 2,figsize=(8,13))
for item,ax in zip(list_to_study2, axes.flatten()):
    sns.boxplot(data=inp_total[item].value_counts(normalize=True),ax=ax,palette = "Reds")
    ax.set_xlabel(item)
plt.show()
```

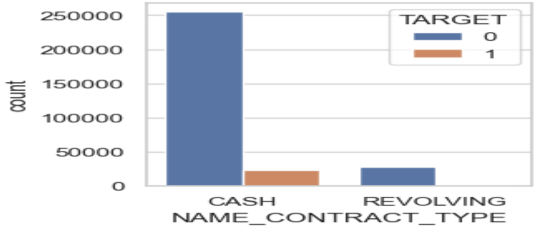


Here We can analyse the **CODE_GENDER ,NAME_INCOME_TYPE, NAME_EDUCATION_TYPE, OCCUPATION_TYPE, NAME_CASH_LOAN_PURPOSE, NAME_GOODS_CATEGORY, PRODUCT_COMBINATION,NAME_PRODUCT_TYPE** by joining both the dataset together.As such no outlier is present.Data is in continous fashion as we already checked both dataset individually

4. (b). BIVARIATE ANALYSIS

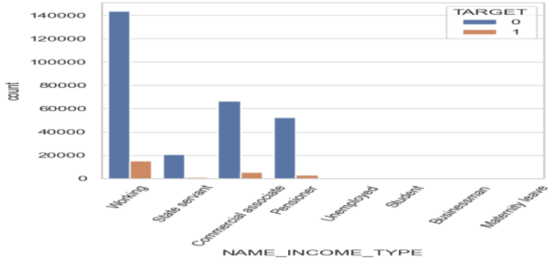
I. Application_data.csv :

1. .inp0 ['NAME_CONTRACT_TYPE'] VS TARGET:



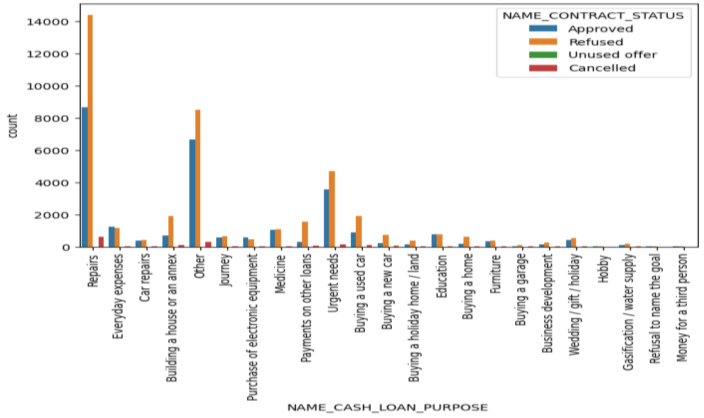
2.inp0 ['CODE_GENDER'] VS TARGET :

3.inp0 ['NAME_INCOME_TYPE'] VS TARGET:

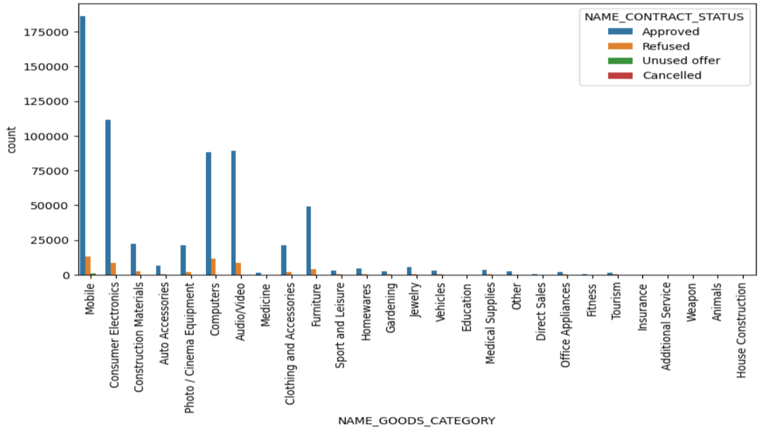


II. Previous_Application_data.csv :

1. .inp1 ['NAME_CASH_LOAN_PURPOSE'] VS CONTRACT_TYPE:

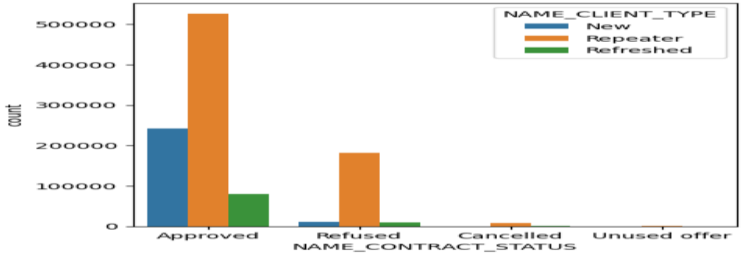


2 .inp1 ['NAME_GOODS_CATEGORY'] VS CONTRACT_TYPE :

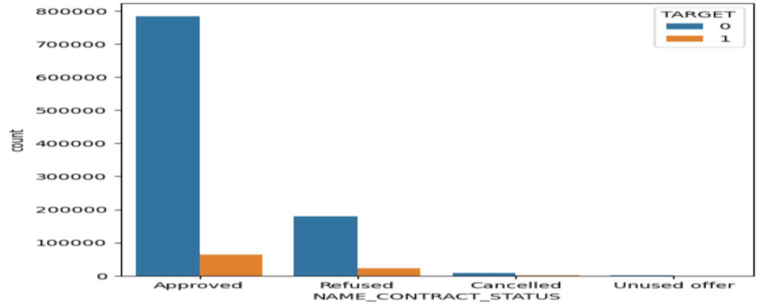


III. Total_Previous_Application_data.csv :

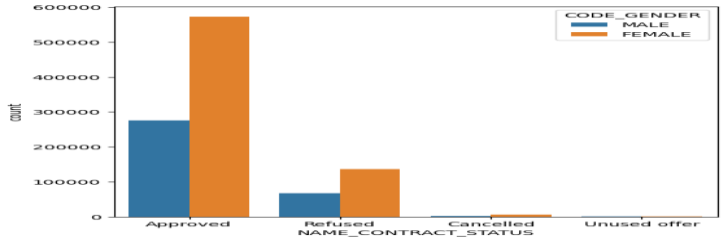
1. .inp_total ['NEW_CLIENT_TYPE'] VS CONTRACT_STATUS:



2. .inp_total ['NEW_CONTRACT_STATUS'] VS TARGET:



3. .inp_total ['NAME_CONTRACT_STATUS] VS CODE_GENDER:

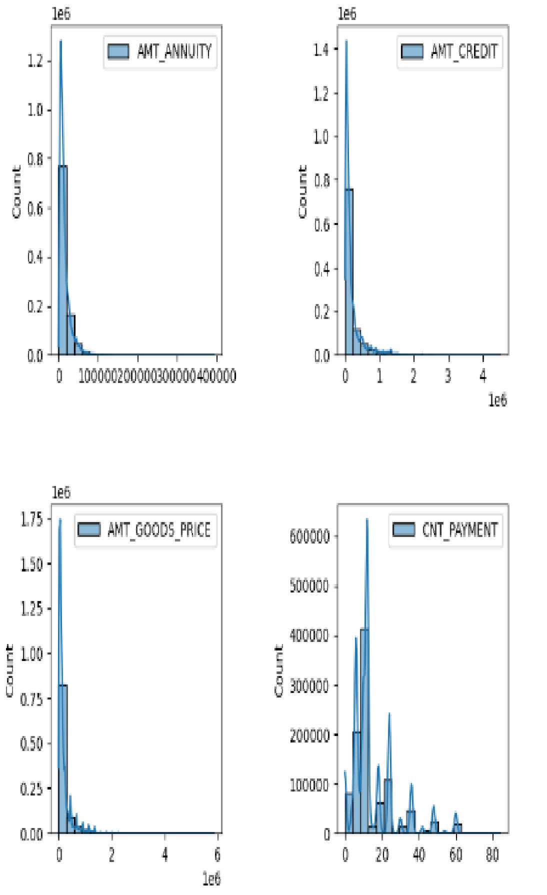


4. (b). BIVARIATE ANALYSIS (contd..)

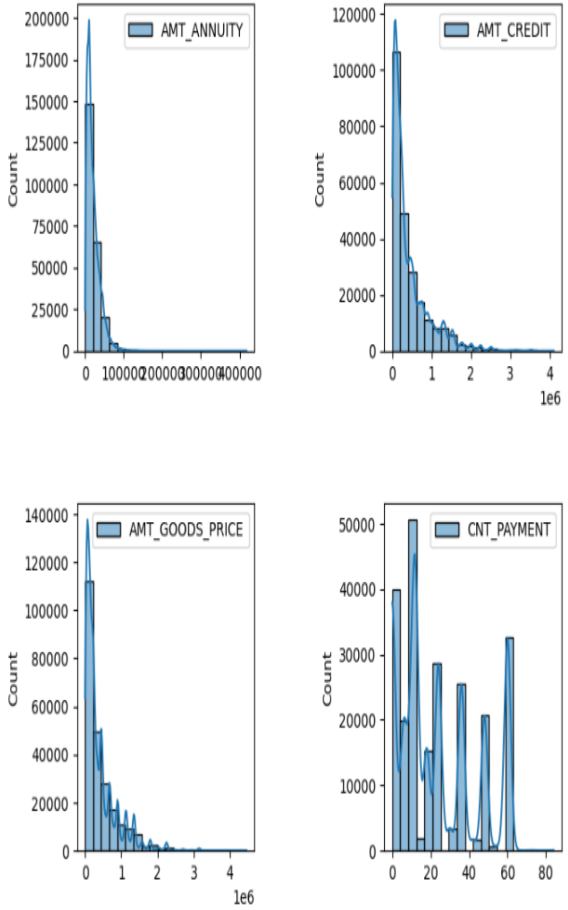
DISTRIBUTION PLOT WITH RESPECT TO CONTRACT STATUS

```
# List of imp numeric feilds only
imp_num_list1 = ['AMT_ANNUITY','AMT_CREDIT','AMT_GOODS_PRICE','CNT_PAYMENT']
```

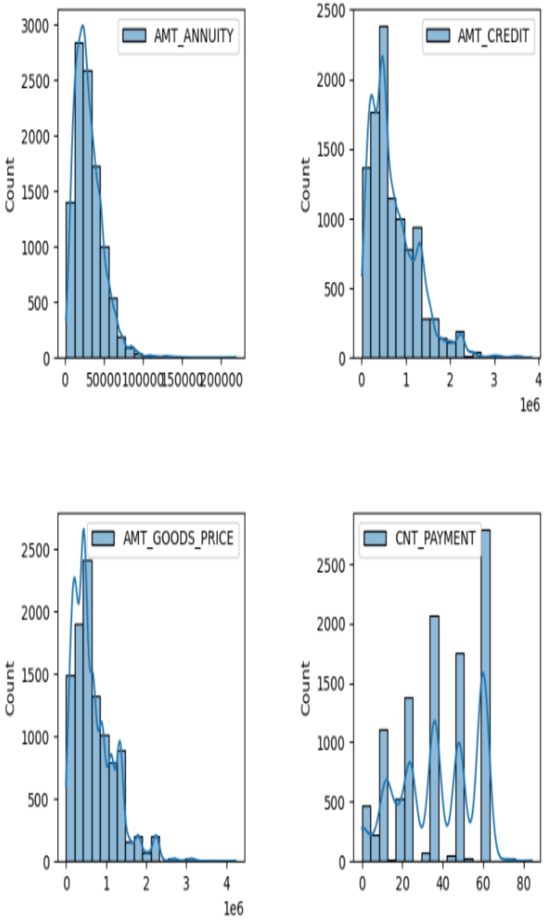
1. With Respect to Approved



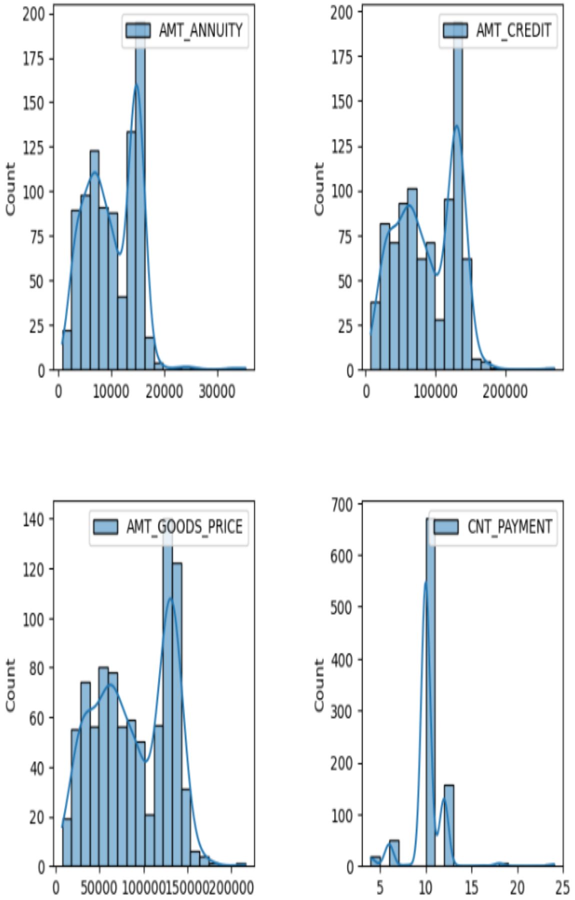
2. With Respect to Refused



3. With Respect to Cancelled



4. With Respect to Unused Offer

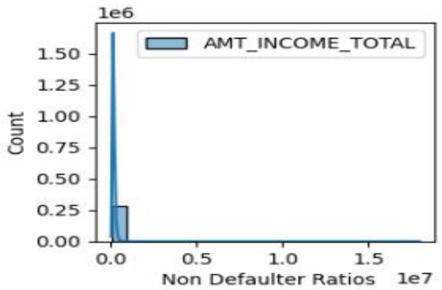


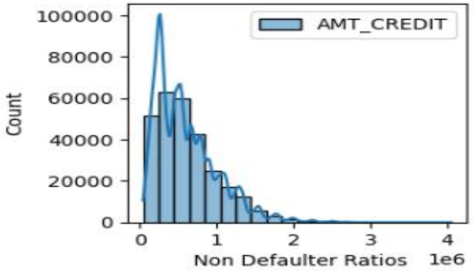
4. (b). BIVARIATE ANALYSIS (contd..)

1. DISTRIBUTION PLOT WITH RESPECT TO TARGET

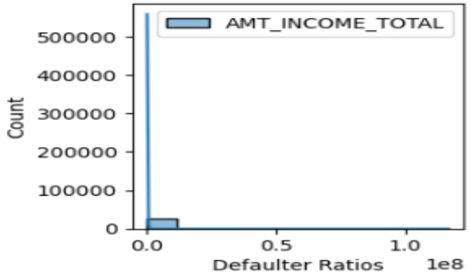
```python
# List of imp numeric feilds only
imp_num_list = ["AMT_INCOME_TOTAL","AMT_CREDIT"]
```

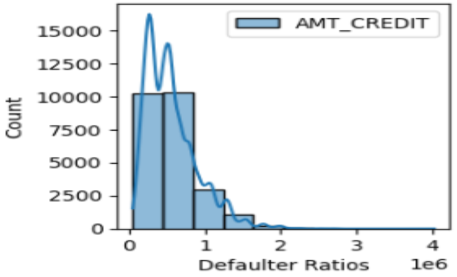
1. With Respect to DEFAULTER





2. With Respect to NON DEFAULTER

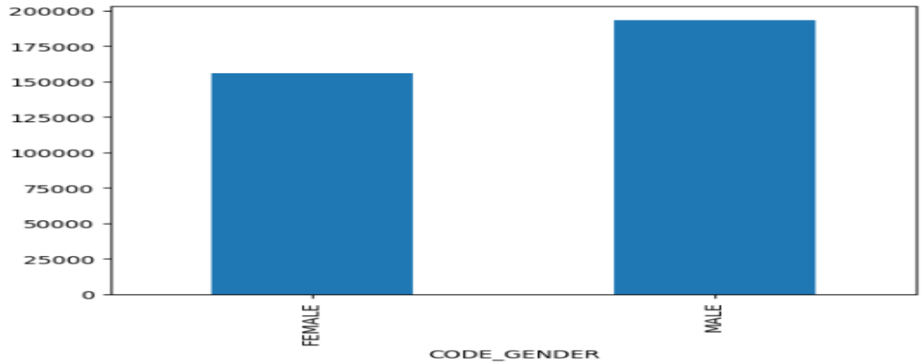




Hist Plot Shows the shape of data of Income of the client,Credit amount of the loan with respect to Target Variable. For Defaulter ar
Ratios.Clearly we can see approx 85% and more are not defaulters where as approx 15% were defaulters. This ratio is quite enougl
Non Defaulters are less in Number as compared to defaulter.

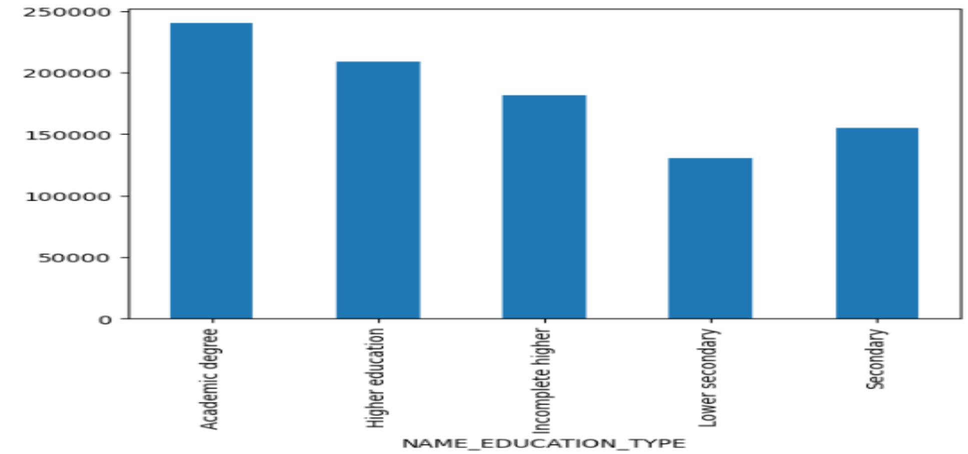
2. DISTRIBUTION PLOT WITH RESPECT TO AMT_INCOME_TOTAL

1. GENDER With AMT_INCOME_TOTAL



Here we can see that MALE is earning more income than FEMALE.

2. EDUCATION With AMT_INCOME_TOTAL

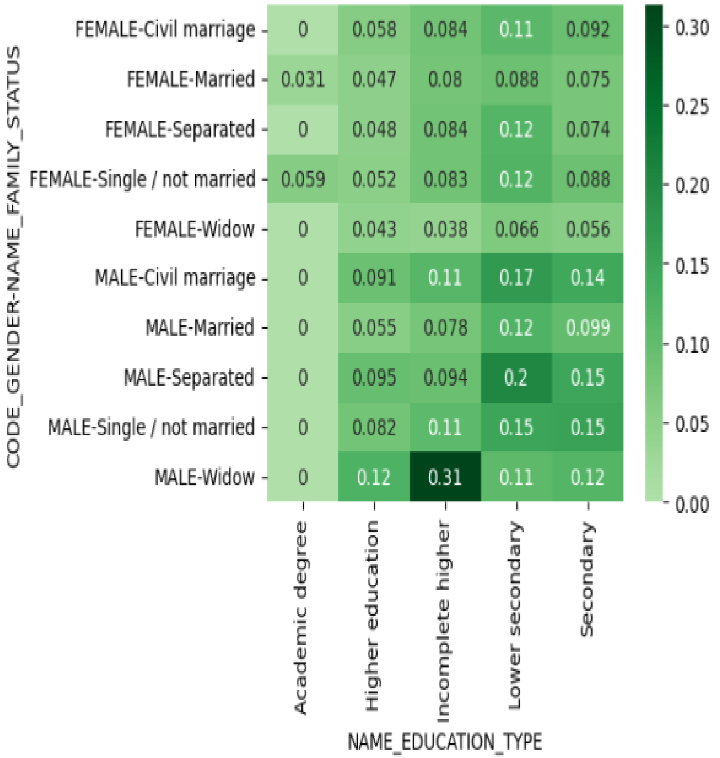


Acadamic Holder is earning More income than others.

4. (c). MULTIVARITE ANALYSIS (HEATMAPS)

Heatmap 1 : GENDER , FAMILY STATUS, EDUCATION TYPE with respect to TARGET

```
#Creating HEATMAP having gender,family status,education type with respect to Target.
plt.figure(figsize=(5,4))
sns.heatmap(pivot_table_1,cmap="Greens",annot=True,center=0.08072881945686496)
plt.show()
```

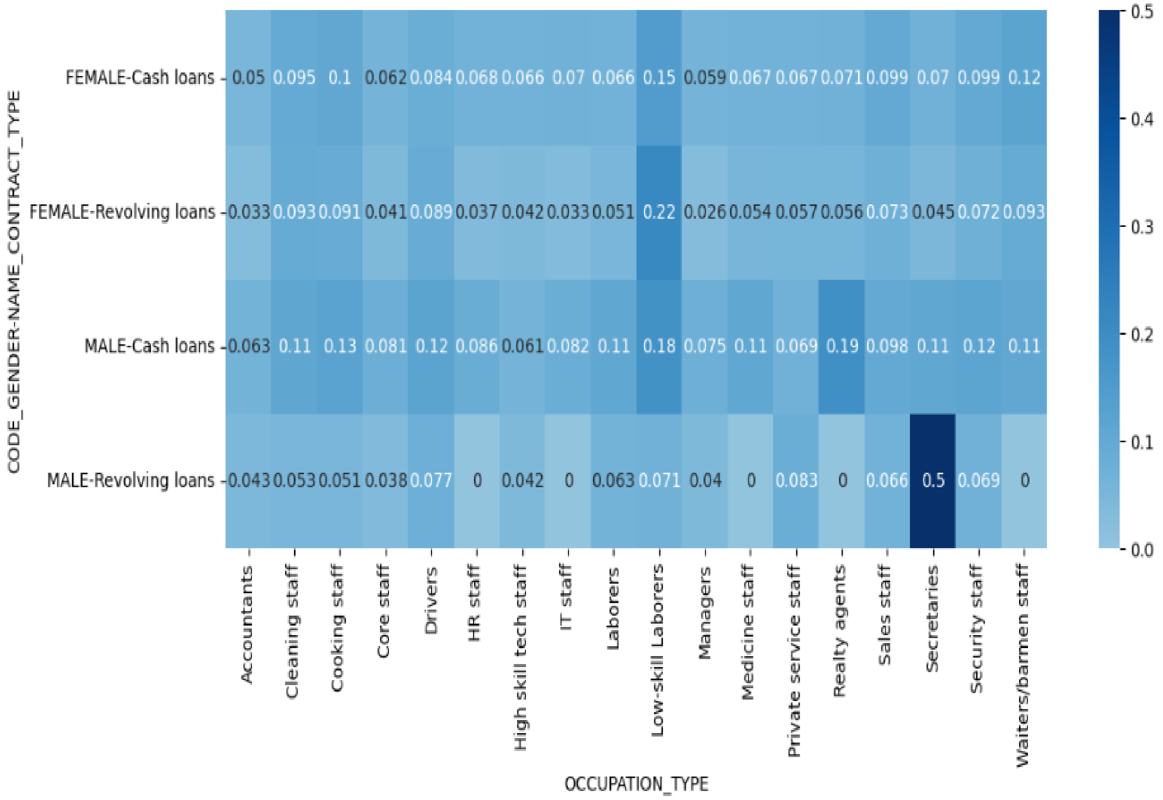


INSIGHTS:

* We can observe that MALE Widow having Incomplete Higher education are more prone to defaulter than others.Whereas Acadamic Degree holder MALE or FEMALE with any family status are very less likely to be defaulter.

Heatmap 2 : GENDER , FAMILY CONTRACT , 0CCUPATION TYPE with respect to TARGET

```
#Creating HEATMAP with gender,family contract type, occupation type with respect to Target.
plt.figure(figsize=(12,5))
sns.heatmap(pivot_table_2,cmap="Blues",annot=True,center=0.08072881945686496)
plt.show()
```

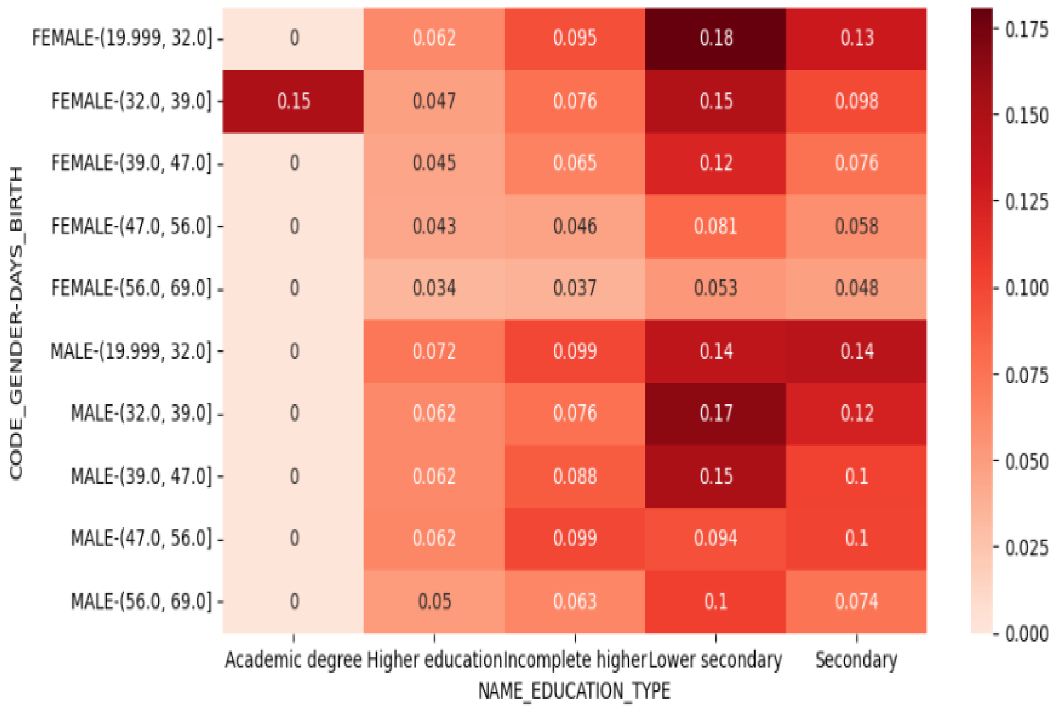


MALE Secretaries having a revolving payment type loans are more likely to be defaulter.

4. (c). MULTIVARITE ANALYSIS (HEATMAPS)(contd..)

Heatmap 3 : GENDER , AGE, EDUCATION TYPE with respect to TARGET

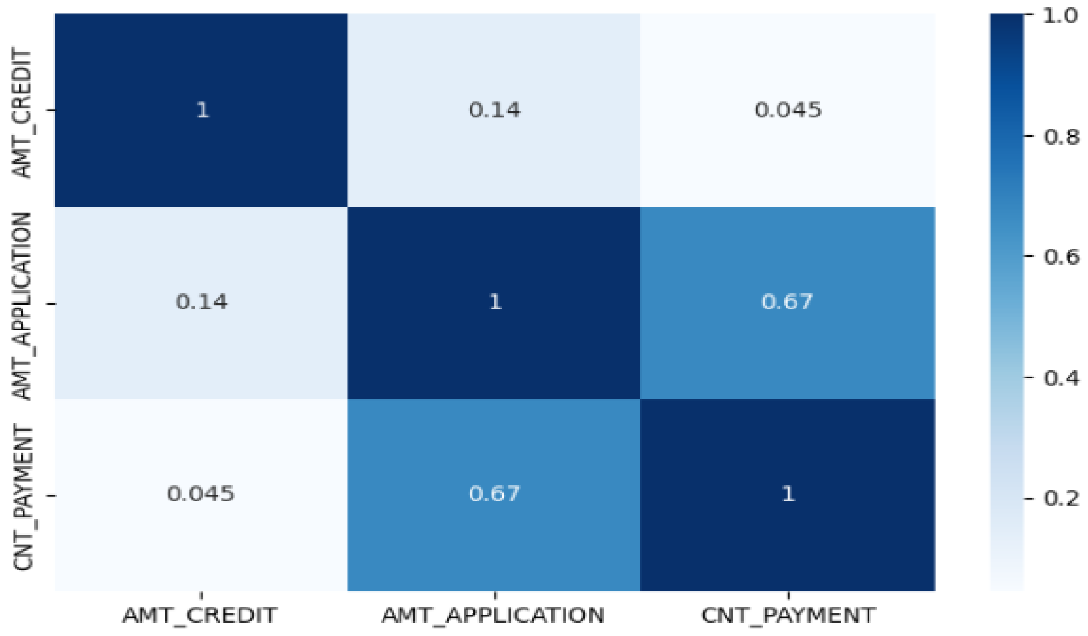
```
#Creating HEATMAP with gender,age,Education type with respect to Target.
plt.figure(figsize=(10,5))
sns.heatmap(pivot_table_3,cmap="Reds",annot=True,center=0.08072881945686496)
plt.show()
```



FEMALES between 19 to 39 yrs ,whose education is Lower Secondary and Academic Degree are more likely to be Defaulter. And MALE between 32 to 39 yrs,whose education is Lower Secondary are more likely to be Defaulter.

Heatmap 4 : With respect to AMT_CREDIT , AMT_APPLICATION, AMT_ANNUITY, PAYMENT

```
#Heatmap with respect to Amt_credit,amt_application,amt_annuity,cnt_payment
plt.figure(figsize=(8,5))
sns.heatmap(result1.corr(),cmap="Blues",annot=True)
plt.show()
```



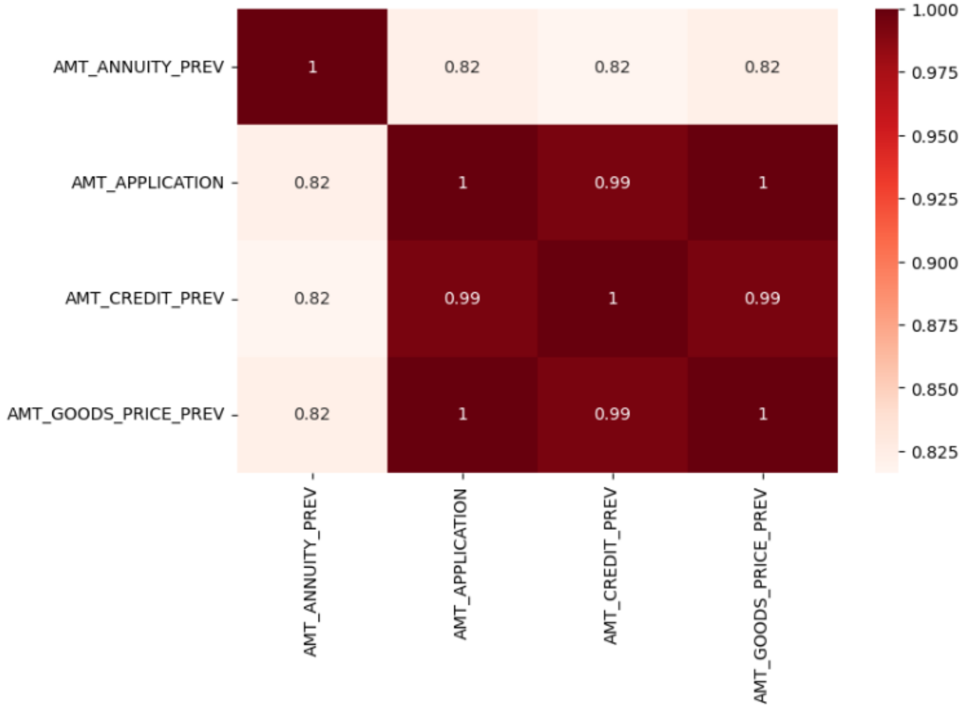
INSIGHTS:

1. AMT_CREDIT & AMT_APPLICATION : Highly correlated, nearly 99%.
2. AMT_CREDIT & AMT_ANNUITY : Highly Correlated, nearly 75%.
3. AMT_ANNUITY & AMT_APPLICATION : Highly correlated, nearly 13%.

4. (c). MULTIVARITE ANALYSIS (HEATMAPS)(contd..)

Heatmap 5 : 'AMT_ANNUITY_PREV', 'AMT_APPLICATION', 'AMT_CREDIT_PREV', 'AMT_GOODS_PRICE_PREV',"NAME_CONTRACT_STATUS"

```
plt.figure(figsize=(8,5))
sns.heatmap(result2.corr(),cmap="Reds",annot=True)
plt.show()
```

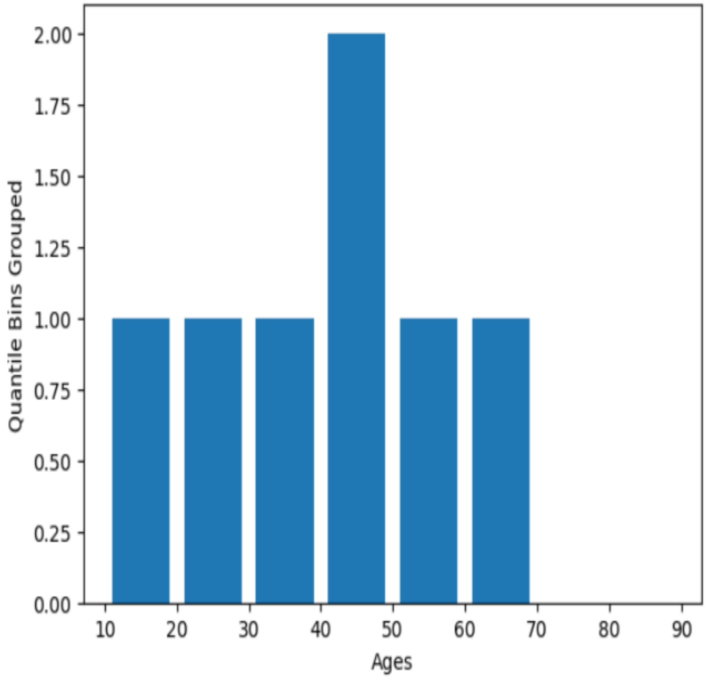


INSIGHTS:

1. AMT_CREDIT_PREV & AMT_APPLICATION : Highly correlated, nearly **99%.**
2. AMT_CREDIT_PREV & AMT_ANNUITY_PREV : Highly Correlated, nearly **82%.**
3. AMT_ANNUITY_PREV & AMT_APPLICATION : Highly correlated, nearly **82%.**
4. AMT_GOODS_PRICE_PREV & AMT_APPLICATION : Highly correlated, nearly **99%.**

EXTRA : Univarite Analysis Of Age using bins

```
# sns.barplot(inp0['DAYS_BIRTH'].describe(),palette = "Reds")
bins = [10,20,30,40,50,60,70,80,90]
plt.hist(inp0['DAYS_BIRTH'].describe(), bins, histtype='bar', rwidth=0.8)
plt.xlabel('Ages')
plt.ylabel('Quantile Bins Grouped')
plt.show()
```



Here we can see the number of people Ages in between 20 to 70 apply for loan.Where as maxmimum loan where applied by people between 40 to 50 Yrs in Age.

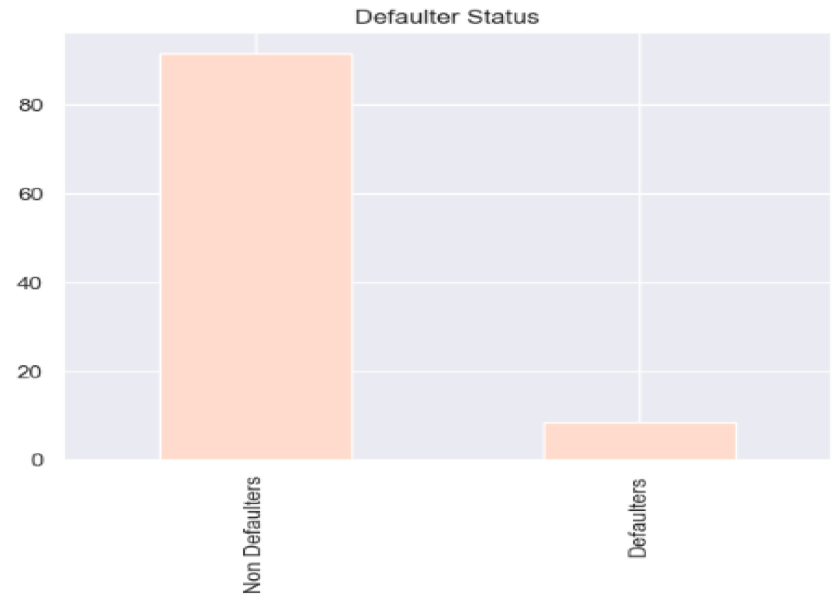
5. Checking Data Imbalance

1. Checking Data Imbalance with TARGET Variable:

```
inp_total['TARGET'].value_counts()
```

```
Non Defaulters     972465
Defaulters          88724
Name: TARGET, dtype: int64
```

```
sns.set_theme(style="darkgrid", palette="Reds")
(inp_total['TARGET'].value_counts(normalize = True) * 100.0).plot.bar()
plt.title("Defaulter Status")
plt.show()
```



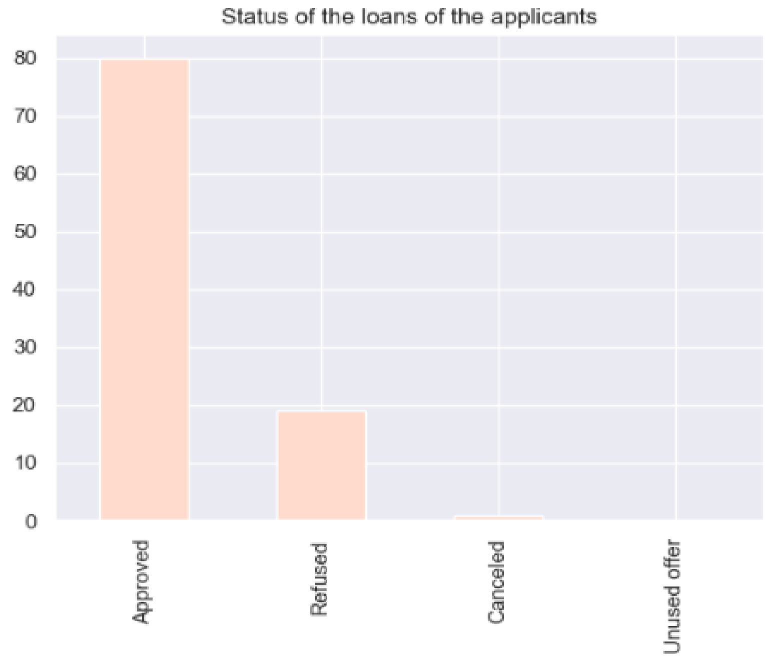
INSIGHTS:

We can see the data imbalance in TARGET Coloumn as 90% domain specific who are not defaulters.

1. About **90%** clients belongs to Others are not Defaulters
2. About **10%** Clients are Defaulters.

2. Checking Data Imbalance with Name_Contract_Type Variable:

```
sns.set_theme(style="darkgrid", palette="Reds")
(inp_total['NAME_CONTRACT_STATUS'].value_counts(normalize = True) * 100.0).plot.bar()
plt.title("Status of the loans of the applicants")
plt.show()
```



INSIGHTS:

We can see the data imbalance here is with Approved as majority of cases are APPROVED.So it is also dominent.

1. About **80%** loans were approved
2. About **19%** loans were refused
3. About **2%** loans were cancelled
4. About **1%** loans were unused offer

6. CONCLUSION

I. After Analyzing the 'application_data.csv' we got fallowing INSIGHTS:

1. Number of Defaulter is more in CASH Payment whereas bit less in Revolving Payment Strategy.
2. Percentage of FEMALES are more for applying loan than MALES.
3. People who owns a car or House are more likely to apply for loan as compared to those who don't own a car.
4. People who are Unaccompanied are more likely to apply for loan as compared to others. And Defaulter cases are also same
5. Working people are more likely to apply for loan whereas Commercial Associate and Pensioners also apply for loan. But defaulters are more in WORKING type people than others.
6. Secondary Education People are more likely to apply for loan and defaulter cases are also more as compared to others
7. Married people are more likely to apply for loan than others and defaulter rate is also more than other categories.
8. People who lives in Houses/Apartments are more likely to apply for loans and become defaulter too as compared to others.
9. People who work as Labores, Core Staff, Managers, Drivers, Sales Staff are more likely to apply for loan.
10. Here we observe that Business Entity Type-3 and Self Employed Organization Working people are more likely to apply for loan than other categories and there Defaulter rate is also high.
11. MALE is earning more income than FEMALE.
12. Businessman is having High Income than others.
13. Academic Holder is earning More income than others.
14. Managers and Accounts have a maximum income. Remaining all are having a average amount of income.
15. Legal Services, Retailer, Military, Security Ministers, Advertising organization's income is more as compared to other organization category. But on an average remaining all have a average amount of income.

6. CONCLUSION(contd..)

II. After Analyzing the 'previous_application_data.csv' we got fallowing INSIGHTS:

1. People who opt for CASH contract type are more likely to be Approved not having chances of Cancellations. Revolving having 70% chances getting approved and 30% of refused and 5% chances of getting cancelled. Consumer have 50-50% of chances of getting Approved or Cancelled.
2. Loan regarding Urgent needs, Repairs, Other are more likely to be approved and Refused than other type of loans.
3. CASH through Bank Payment type are more likely to be approved and very less refused for loan than others.
4. Majorly reasons were not provided for rejection as we can see XAP. But HC ,LIMIT,SCO reasons are likely to be refused.
5. The clients who are Repeater for applying loan are more likely to get approved for loan than NEW Client.
6. Loans for Goods like Mobile, Electronics, Computers, Audio/Video, Furniture are more likely to be approved.
7. Loans for X-sell are more likely to be approved than walk-in and Refused is also work in same fashion.
8. Country-wide ,contact center, Credit and cash offices Stores like channel are more likely to be approved for loan
9. Seller belongs to Consumer Electronics Industry are more likely to be approved for loan than others. Whereas Seller from Connectivity industry also get approved for loan. Refuses for loan also there in both the industries.
10. POS Household with interest, POS mobile with interest, POS industry with interest are more likely to be approved for loan than others.