Case Study to Identify the Loan Defaulters

Data Source : Up Grad University
Prepared By :
Ishu Srivastava
Santosh Kumar

What We Did

This Case Study is for analyze the data provided in 2 csv files containing. To analyze the data we

1. Preparation of Data Source and Data understanding
2. Data Cleaning
3. Removal of Not required columns
4. Suggestion for null Value treatments
5. Identification and treatment of outliers
6. Creation of bins for continuous variables
7. Univariate Analysis of the Categorized columns
8. Analyze the correlation in Continuous variables

Preparation of Data Source and Data understanding

- Read the csv files into Pandas DataFrames
- Check the number of columns and their data types
- Analyze the columns to identify the required columns for further analysis
- Combine the 2 Data Frames for Final analysis

- Initially there were 307511 rows and 122 columns in application_data.csv

- Initially there were 1670214 rows and 37columns in previous_application.csv
- After merging this dataframe with first dataframe, we got 1413701 rows and 158 columns
- Out of which

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1413701 entries, 0 to 1413700
Columns: 158 entries, SK_ID_CURR to NFLAG_INSURED_ON_APPROVAL
dtypes: float64(80), int64(46), object(32)
memory usage: 1.7+ GB
```

Data Cleaning

- Indentify the % of null values in each columns
- removed the column and drop the columns where null values are greater than approx 50%

	Column	Null Values	Total Values	Percent of NULL Values
20	OWN_CAR_AGE	202929	307511	65.990810
27	OCCUPATION_TYPE	96391	307511	31.345545
10	AMT_GOODS_PRICE	278	307511	0.090403
9	AMT_ANNUITY	12	307511	0.003902
28	CNT_FAM_MEMBERS	2	307511	0.000650
29	REGION_RATING_CLIENT	0	307511	0.000000
22	FLAG_EMP_PHONE	0	307511	0.000000
23	FLAG_WORK_PHONE	0	307511	0.000000
24	FLAG_CONT_MOBILE	0	307511	0.000000

	Column	Null Values	Total Values	Percent of NULL Values
24	RATE_INTEREST_PRIMARY	1408910	1413701	99.661102
25	RATE_INTEREST_PRIVILEGED	1408910	1413701	99.661102
21	AMT_DOWN_PAYMENT	749540	1413701	53.019698
23	RATE_DOWN_PAYMENT	749540	1413701	53.019698
30	NAME_TYPE_SUITE_y	694672	1413701	49.138538
22	AMT_GOODS_PRICE_y	319525	1413701	22.602021
18	AMT_ANNUITY_y	307218	1413701	21.731469
10	NAME_TYPE_SUITE_x	3526	1413701	0.249416
9	AMT_GOODS_PRICE_x	1208	1413701	0.085449

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 37 columns):
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1413701 entries, 0 to 1413700
Data columns (total 31 columns):
```

Removal of Not required columns

- In first data Frame we identified 38 columns for our analysis
- In final combined Data Frame we chosen 34 columns for analysis

	Column	Null Values	Total Values	Percent of NULL Values
20	OWN_CAR_AGE	202929	307511	65.990810
27	OCCUPATION_TYPE	96391	307511	31.345545
10	AMT_GOODS_PRICE	278	307511	0.090403
9	AMT_ANNUITY	12	307511	0.003902
28	CNT_FAM_MEMBERS	2	307511	0.000650
29	REGION_RATING_CLIENT	0	307511	0.000000
22	FLAG_EMP_PHONE	0	307511	0.000000
23	FLAG_WORK_PHONE	0	307511	0.000000
24	FLAG_CONT_MOBILE	0	307511	0.000000

	Column	Null Values	Total Values	Percent of NULL Values
24	RATE_INTEREST_PRIMARY	1408910	1413701	99.661102
25	RATE_INTEREST_PRIVILEGED	1408910	1413701	99.661102
21	AMT_DOWN_PAYMENT	749540	1413701	53.019698
23	RATE_DOWN_PAYMENT	749540	1413701	53.019698
30	NAME_TYPE_SUITE_y	694672	1413701	49.138538
22	AMT_GOODS_PRICE_y	319525	1413701	22.602021
18	AMT_ANNUITY_y	307218	1413701	21.731469
10	NAME_TYPE_SUITE_x	3526	1413701	0.249416
9	AMT_GOODS_PRICE_x	1208	1413701	0.085449

Suggestion for null Value treatments

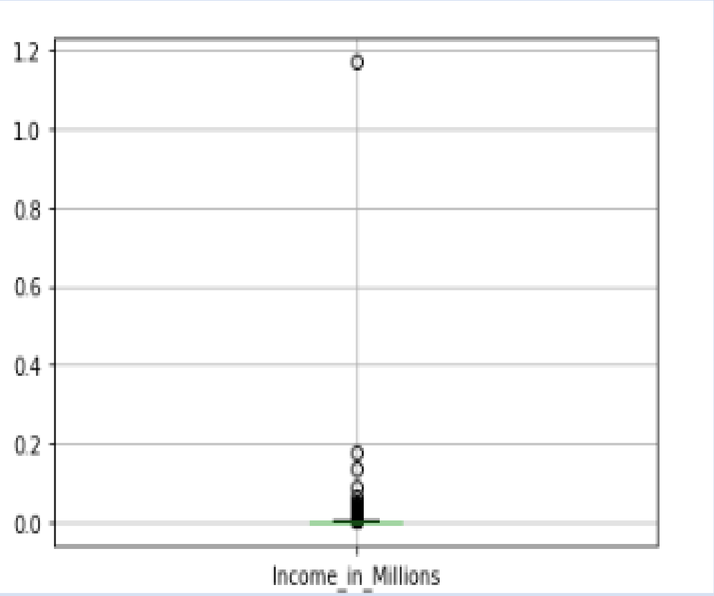
- We have analyzed below categorical columns and suggested the imputation of null values

1. **OCCUPATION_TYPE:** As we can see that approx 75% of the applicants are female where the occupation is not specified, so probably we can impute the nullvaues of OCCUPATION_TYPE column to "Home Maker" or "Unemployed“
2. **AMT_ANNUITY :** Out of 307511 records 21297 records are seems to outliers, which is approx 7% of total records. So we can impute the null values with median value of the column, which is 24903.
3. **AMT_GOODS_PRICE:** AMT_GOODS_PRICE is the price of the goods for which the loan is given. Since we are talking about the loans which have already been disbursed, so the missing value for AMT_GOODS_PRICE implies that this loan is not taken for purchase of any goods, hence we can replace the null values comfirtably with 0
4. **CNT_FAM_MEMBERS:** we can see that there are just 2 such applicants where the CNT_FAM_MEMBERS value is missing. Out of such huge number of records only 2 missing values indicates that that this field was taken very seriously during taking the survey, and hence the chances are very less that these fields are left blank accidently. So we can say that here NULL has been recorded instead of 0

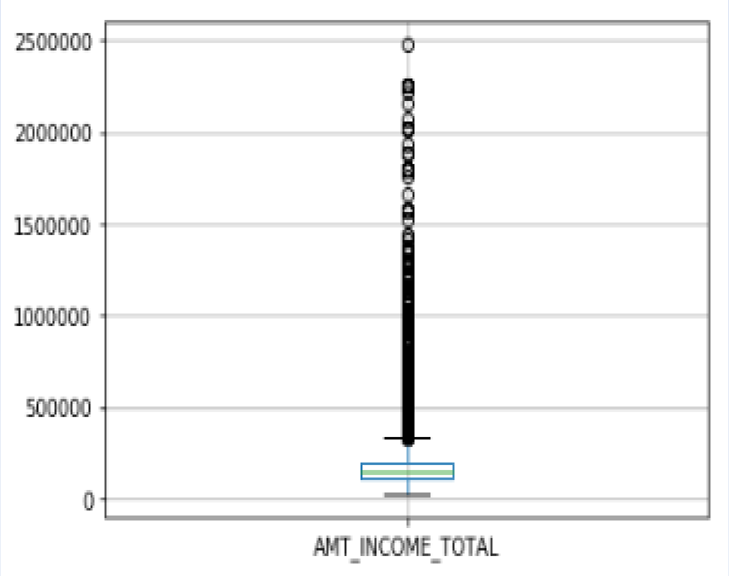
Identification and treatment of outliers

- Plotted the box plot for the continuous variables and treated them accordingly
- Below are few glimpses

- **Before Treatment**

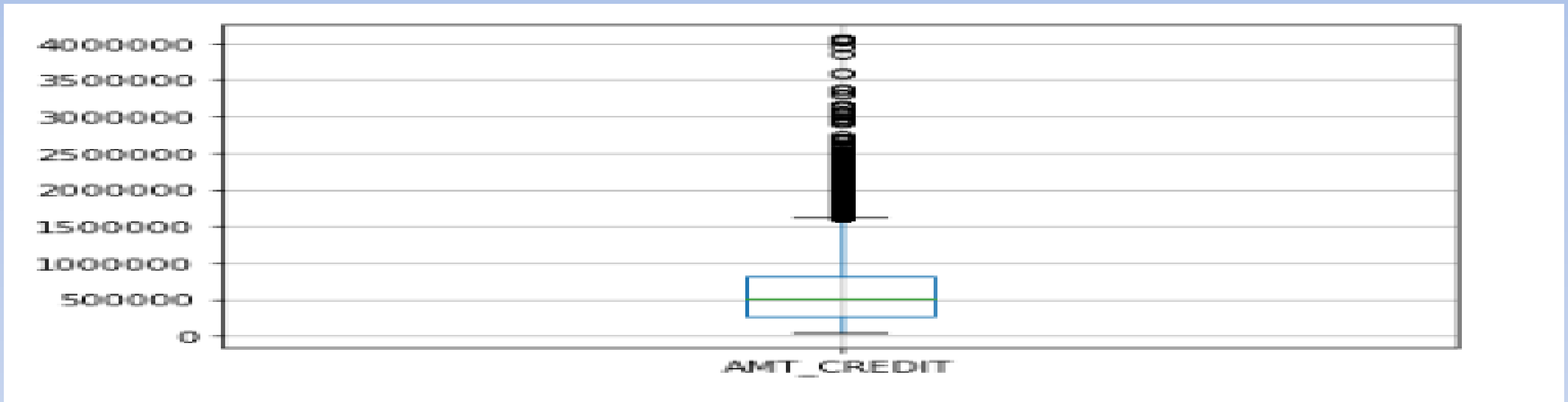


- **AfterTreatment**



Identification and treatment of outliers

Before Treatment



After Treatment

We can see here are also outliers having loan more than 15 Lakhs and the mean is just 5Lakhs,so now lets see how many applicants are there having loan more than 15 Lakh rupees ¶

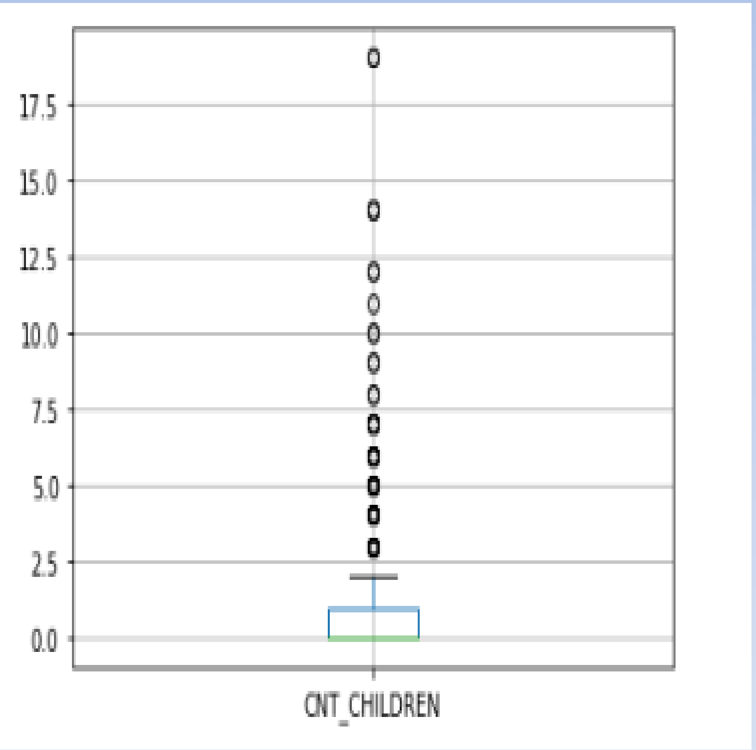
```
]: temp=curr_app.loc[curr_app['AMT_CREDIT']>1500000]
   len(temp.index)
```

```
]: 10753
```

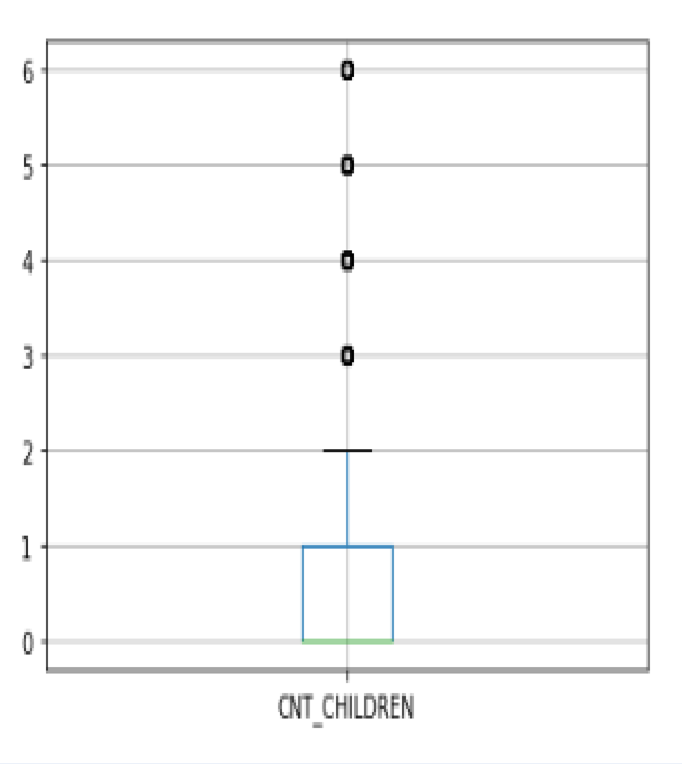
As we can see here as well the entries having loan more than 15 Lakhs are around 10k, Which is approx one third ot the total data, so we cannot drop this as well. So here as well the better way would be create differen bins for Loan amount and catogerize the rows accordingly

Identification and treatment of outliers

- **Before Treatment**



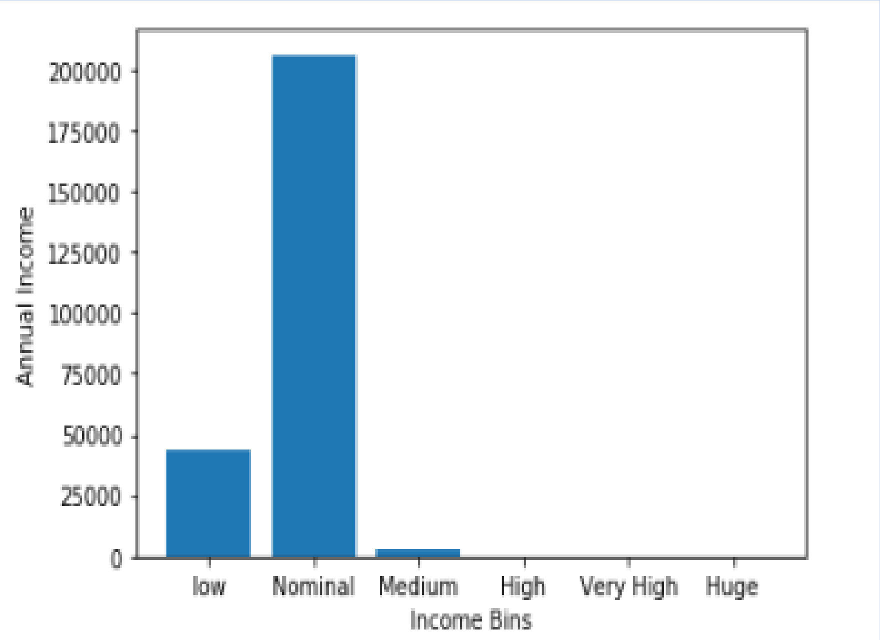
- **AfterTreatment**



Creation of bins for continuous variables

- Created bins for continuous variables
- Analyze those and Provided the insight
- Few glimpses

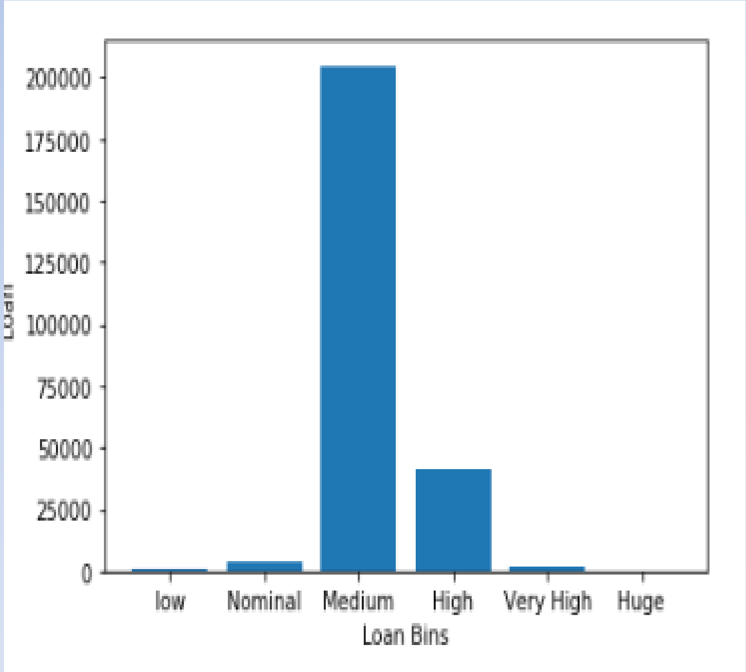
Annual Income Bins



	Income_Cat	AMT_INCOME_TOTAL
0	low	43671
1	Nominal	205923
2	Medium	2303
3	High	160
4	Very High	38
5	Huge	23

Creation of bins for continuous variables

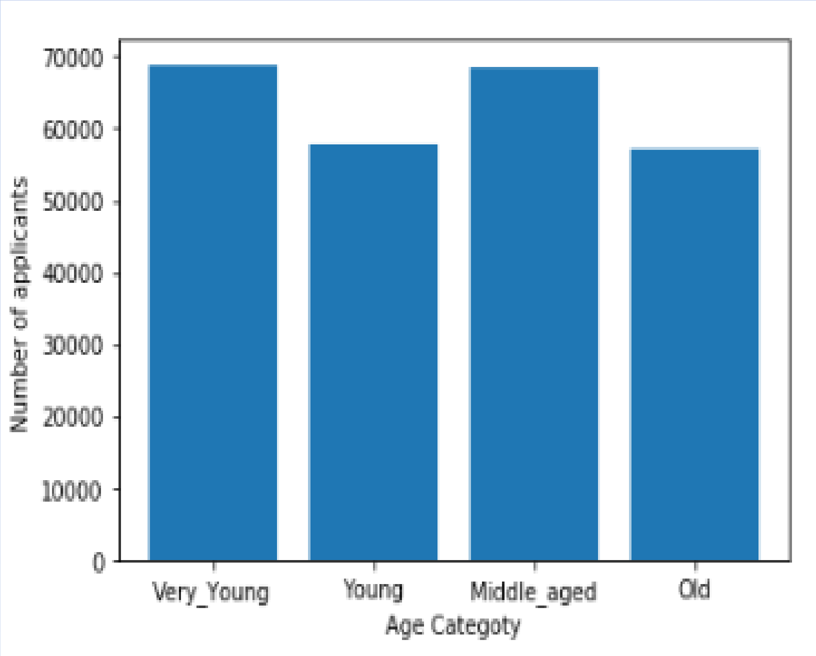
LOAN AMMOUNT BINS



	Loan_cat	AMT_CREDIT
0	low	383.0
1	Nominal	3885.0
2	Medium	204715.0
3	High	41341.0
4	Very High	1784.0
5	Huge	10.0

Creation of bins for continuous variables

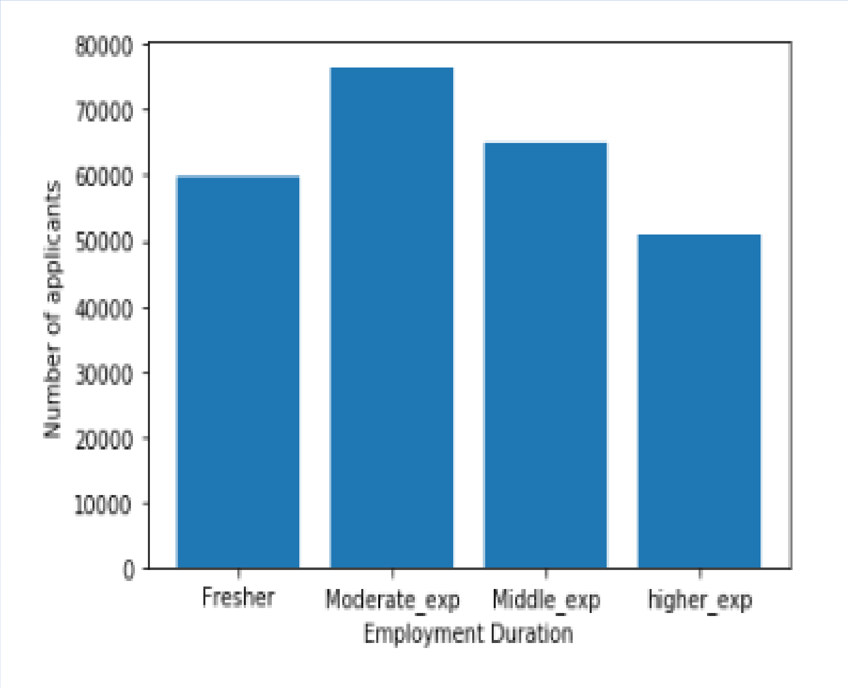
Age Group BINS



	Age_cat	Age_in_Years
0	Very_Young	68862
1	Young	57788
2	Middle_aged	68369
3	Old	57098

Creation of bins for continuous variables

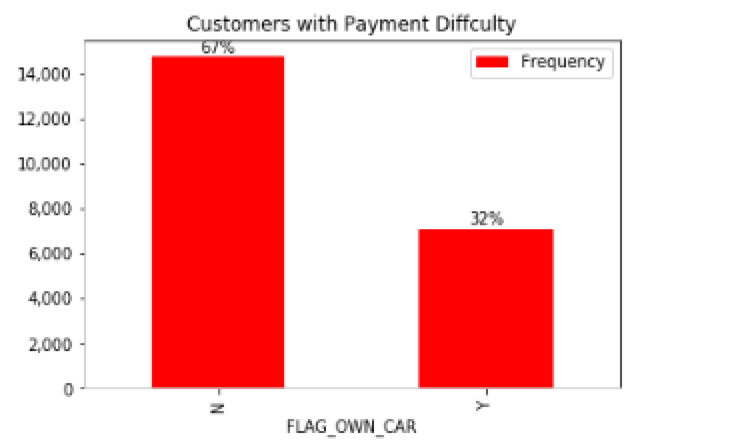
Employment Duration Bins

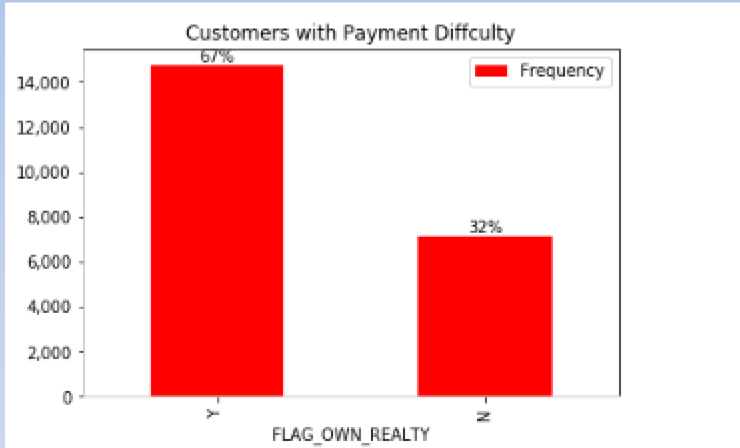


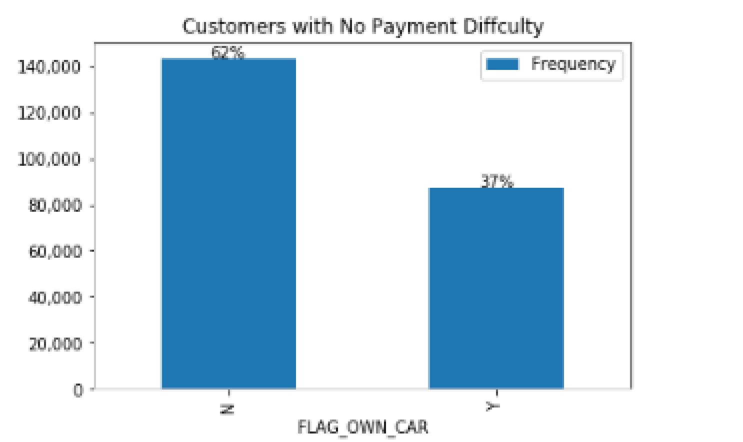
	Exp_cat	Employment_in_years
0	Fresher	59820.0
1	Moderate_exp	76480.0
2	Middle_exp	64867.0
3	higher_exp	50949.0

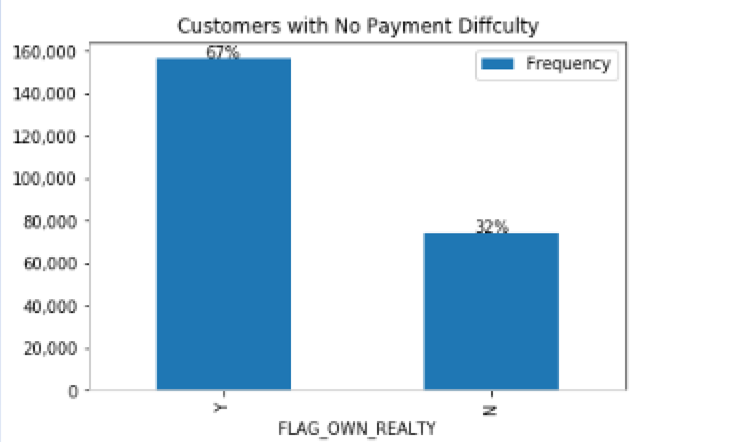
Univariate Analysis of the Categorized columns

- Performed the Univariate analysis of the categorized columns for both the dataframe for both the categotries (defaulters and non-defaulters)
- Provided the insight Few glimpses

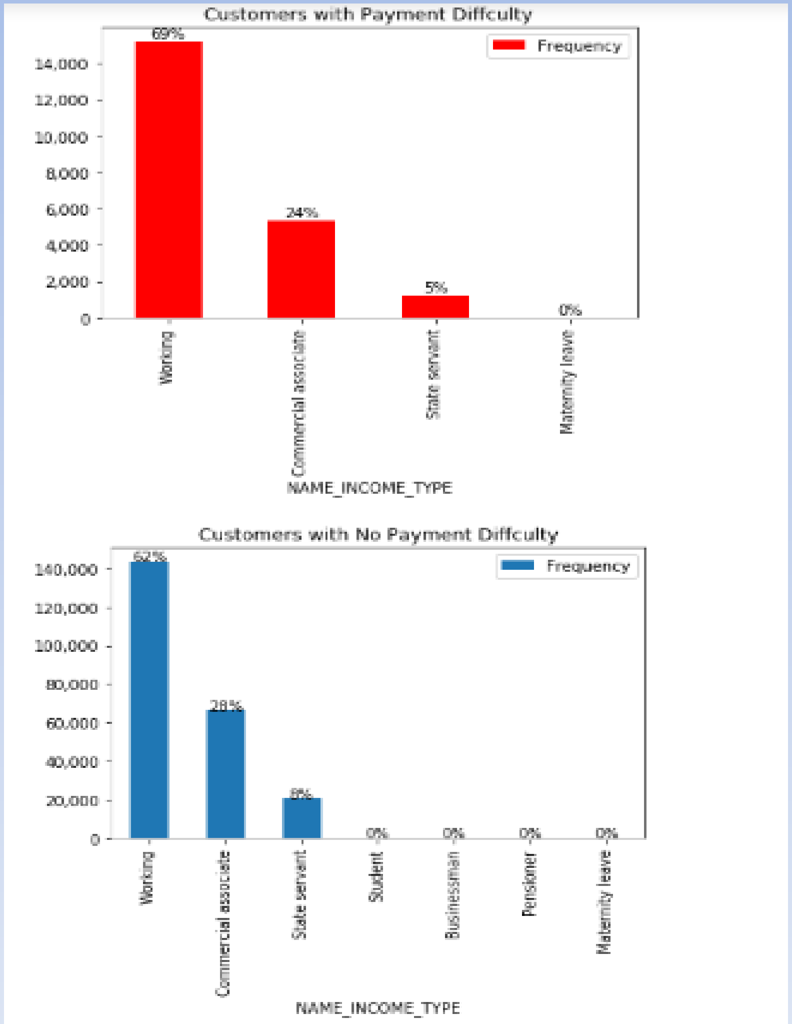


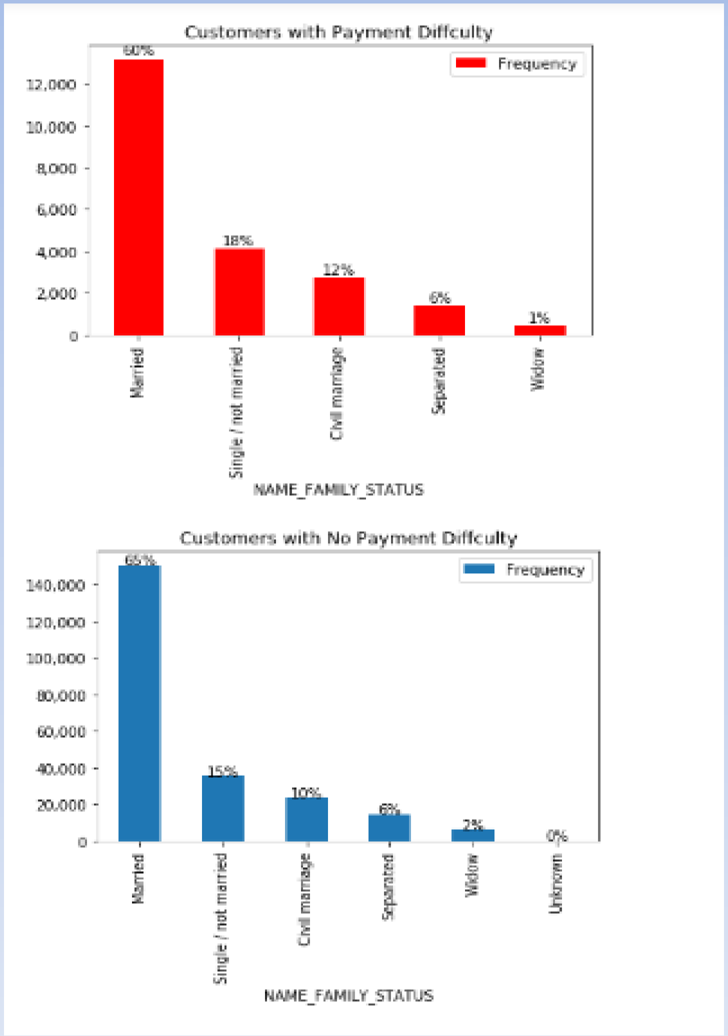






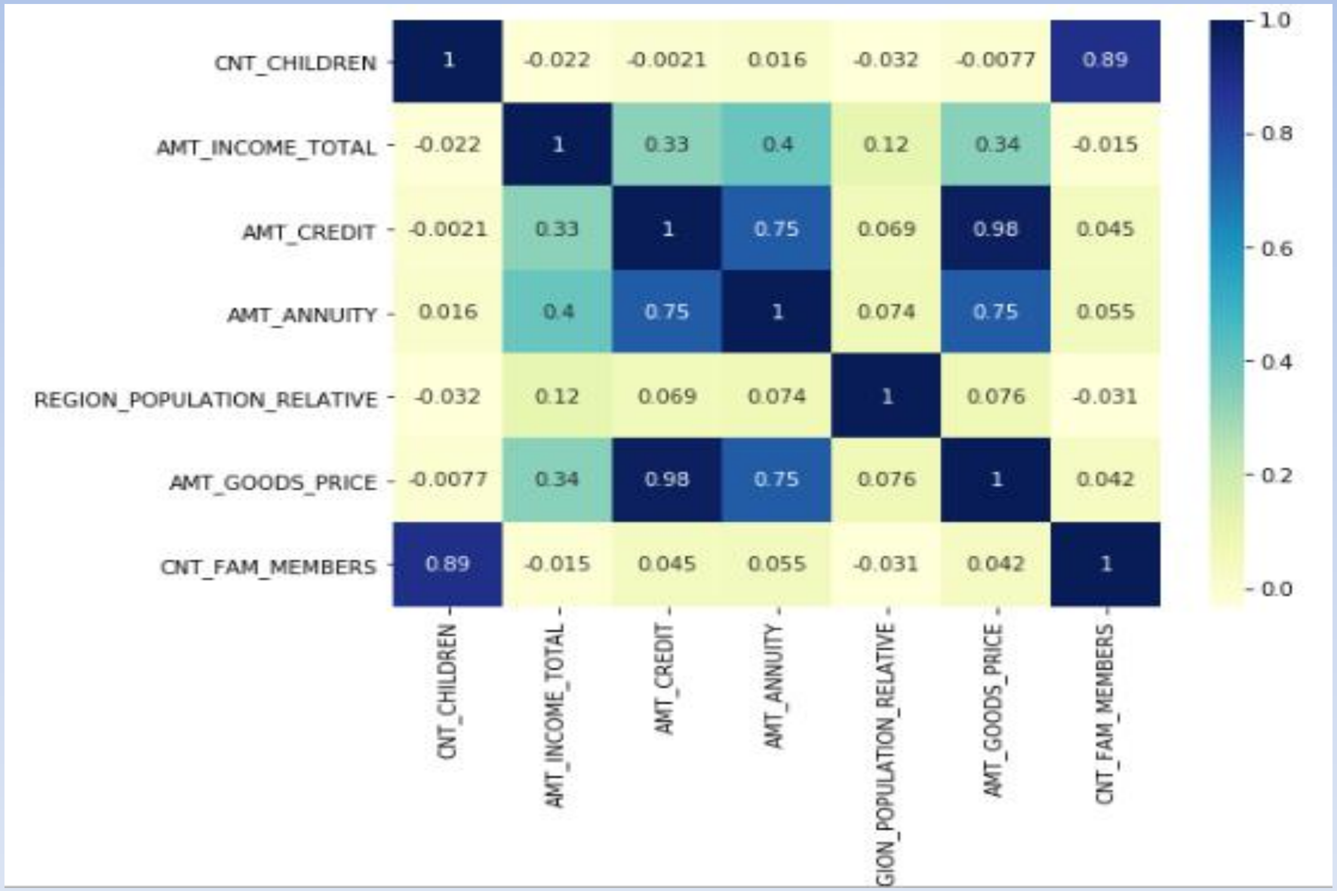
Univariate Analysis of the Categorized columns





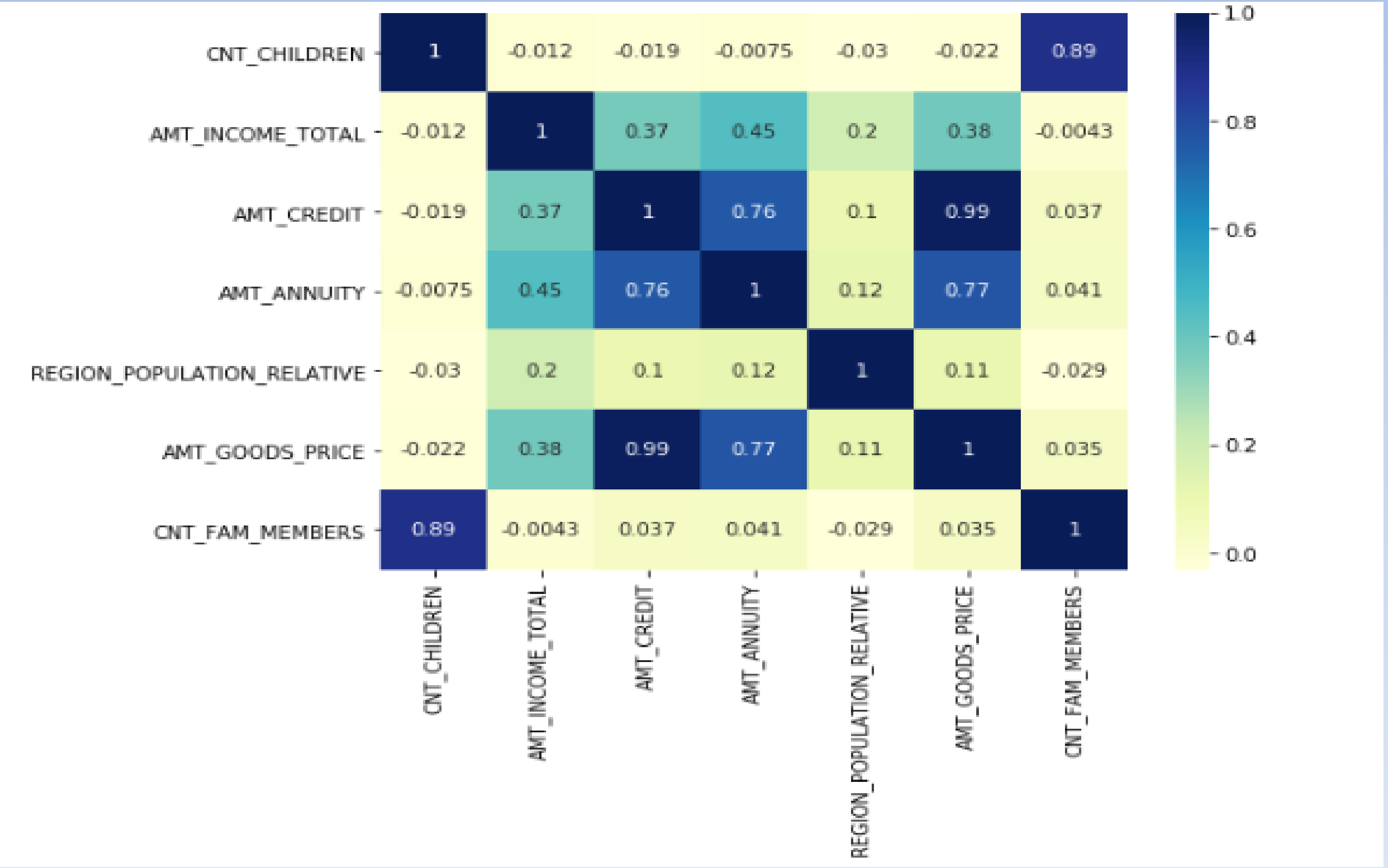
Analyze the correlation in Continuous variables

Defaulters



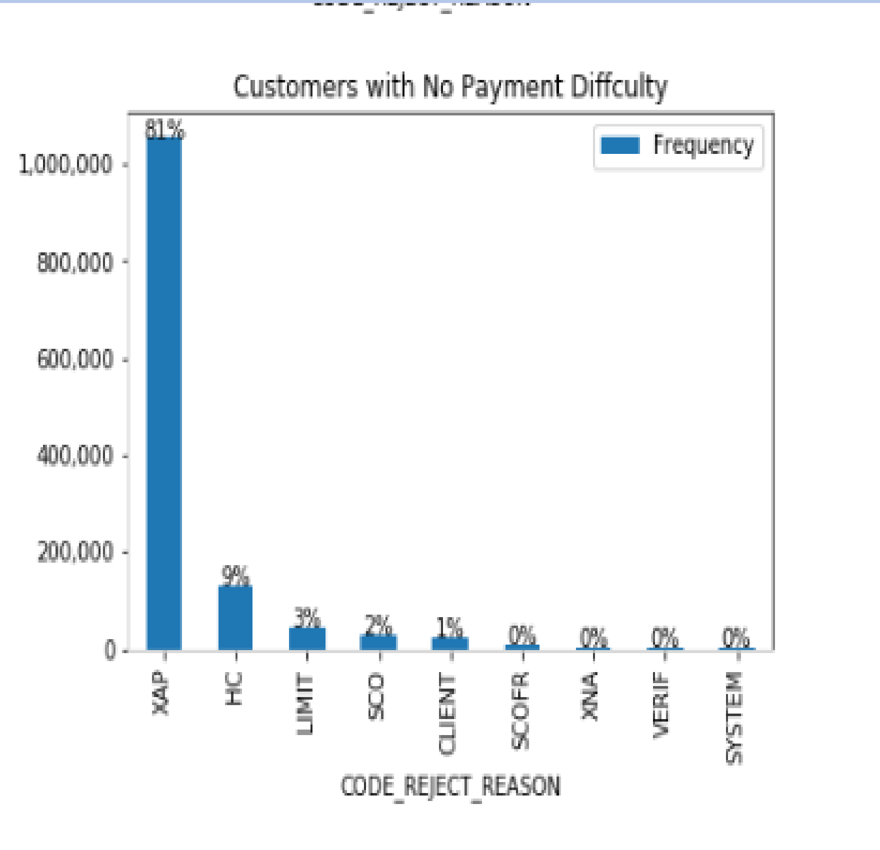
Analyze the correlation in Continuous variables

Non-Defaulters



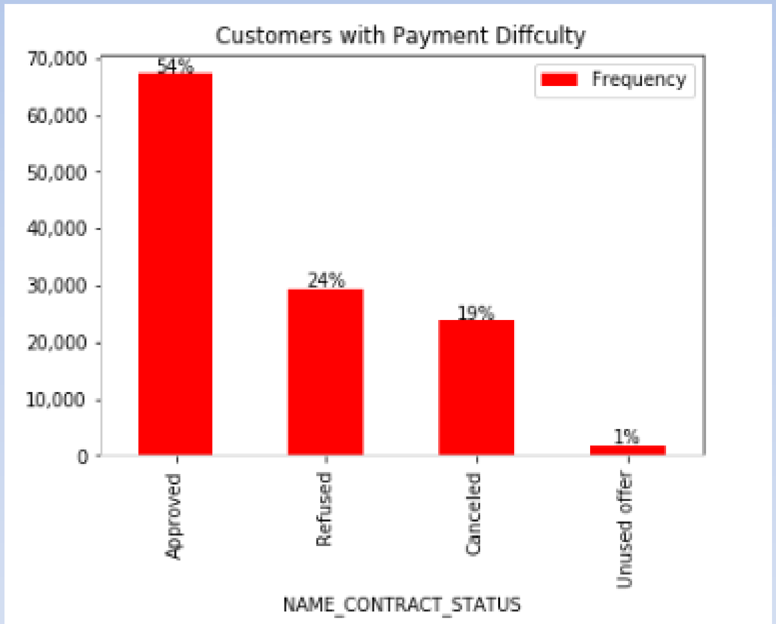
Similar Analysis for the Merged final DataFrame

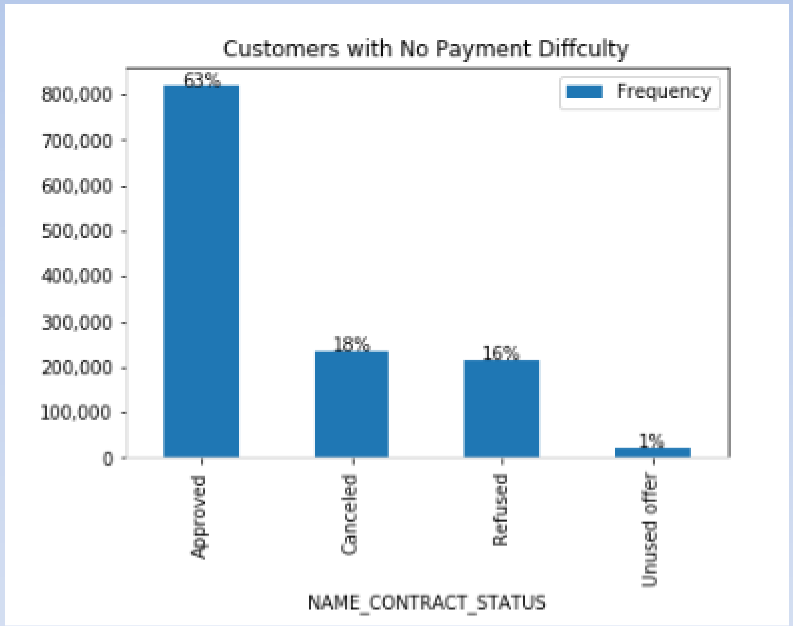
Univariate Analysis



Similar Analysis for the Merged final DataFrame

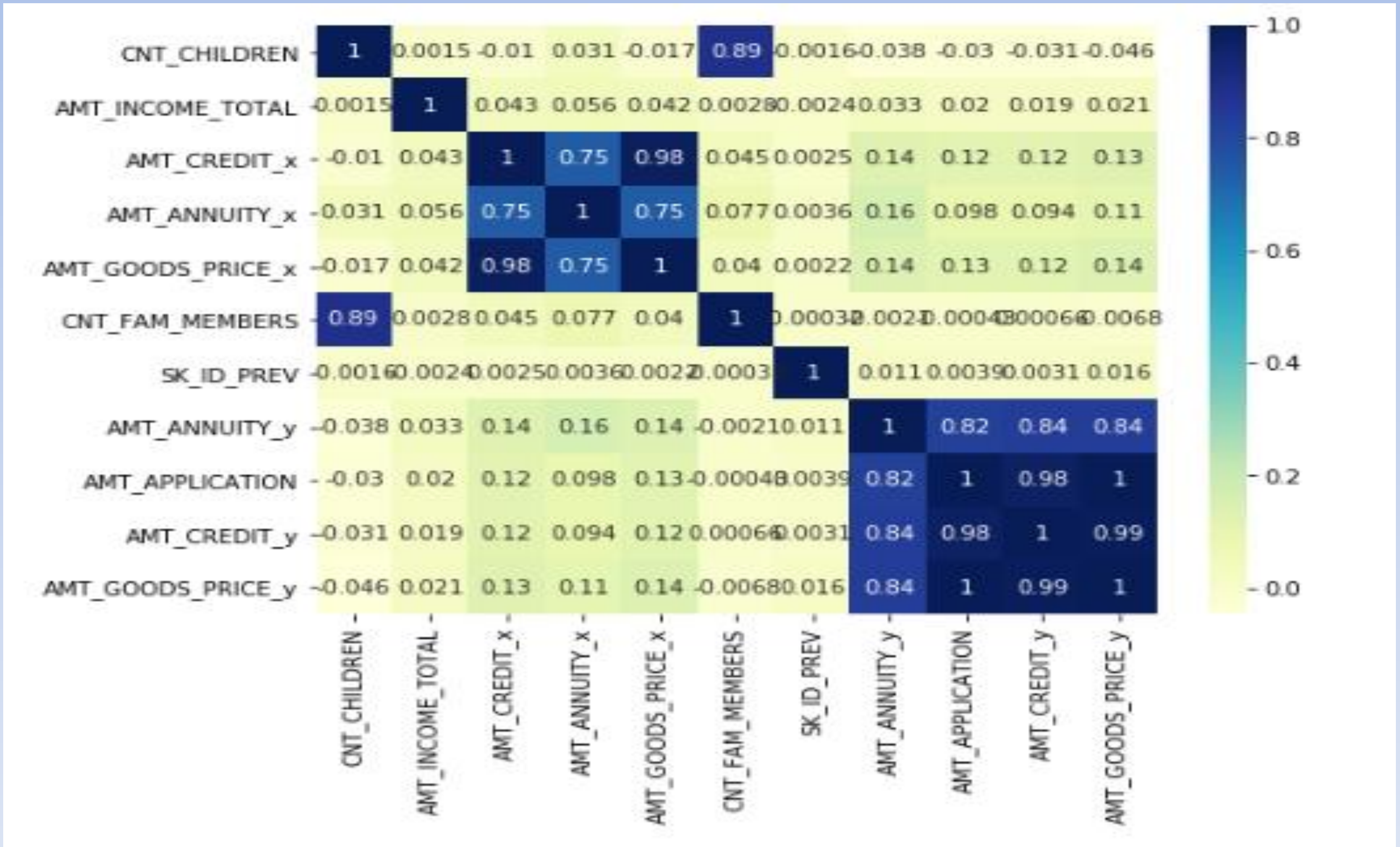
Univariate Analysis





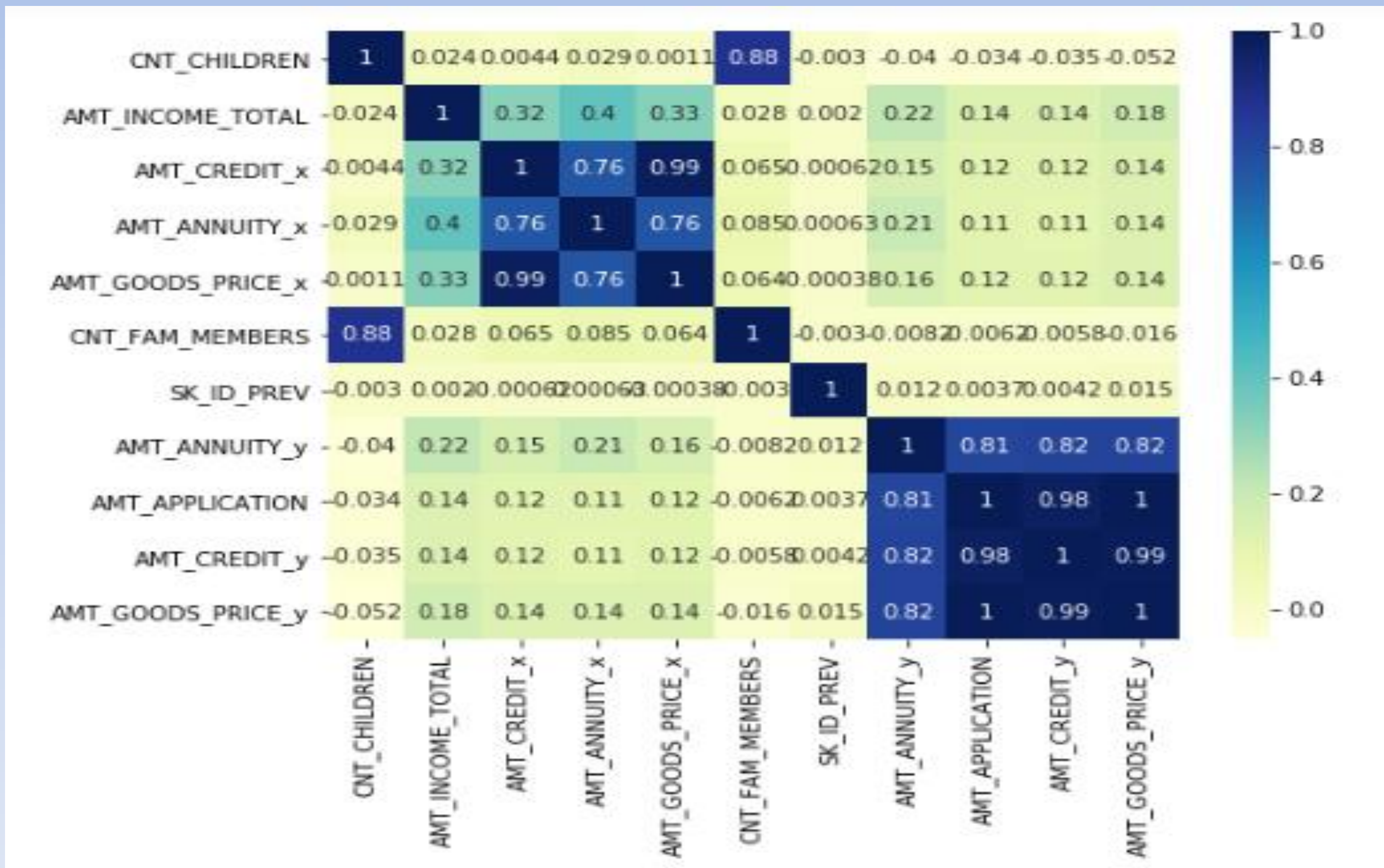
Analyze the correlation in Continuous variables

Defaulters



Analyze the correlation in Continuous variables

Non-Defaulters



Conclusion

- **Based on the data analysis as shown, we can conclude that there is high probability of Default when the applicant –**
    - Doesn’t own a car
    - Own realty
    - Cash Type Loan is applied
    - Education of Applicant is ‘Secondary / Secondary Special’

	CODE_GENDER	Default_Count	Percentage
0	F	11921	54.6
1	M	9914	45.4

	FLAG_OWN_CAR	Default_Count	Percentage
0	False	14753	67.57
1	True	7082	32.43

	FLAG_OWN_REALTY	Default_Count	Percentage
0	False	7110	32.56
1	True	14725	67.44

	NAME_CONTRACT_TYPE	Default_Count	Percentage
0	Cash loans	20371	93.3
1	Revolving loans	1464	6.7

	NAME_EDUCATION_TYPE	Default_Count	Percentage
0	Academic degree	3	0.01
1	Higher education	3669	16.80
2	Incomplete higher	848	3.88
3	Lower secondary	315	1.44
4	Secondary / secondary special	17000	77.86