PROJECT - 6

BANK LOAN CASE STUDY

NAME: BHAVYA SRI DUGGINA

PROJECT DESCRIPTION

This project focuses on leveraging Excel ,python skills and Statistical skills to conduct Bank loan case study

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments.

This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

This EDA project aims to help your finance company make informed decisions regarding loan approval. By understanding the patterns and risk factors associated with loan defaults, you can optimize the loan approval process, reduce financial losses, and ensure that deserving applicants are not rejected. Regular monitoring and adaptation of strategies will be essential to maintain a healthy loan portfolio.

APPROACH

1.IMPORTING THE DATASET INTO EXCEL AND JUPYTER NOTEBOOK

2.DATA CLEANING AND QUALITY CHECK

3.EXPLORE THE DATASET AND EXTRACT THE INSIGHTS

4.GENERATE EFFICIENT REPORT

TECH STACK USED

Tech-stack used in this project are Microsoft Excel 2013,Jupyter Notebook and Microsoft PowerPoint

- **Microsoft Excel 2013:**

Purpose: Microsoft Excel 2013 is a pivotal tool for this bank loan case study project. It is utilized for various data-related tasks, including data cleaning, manipulation, and exploratory data analysis (EDA).

- **Jupyter Notebook**

Purpose: Certainly! Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. In this project I have used Jupyter for Data cleaning and identifying the outliers.

- **Microsoft PowerPoint 2013:**

Purpose: Microsoft PowerPoint 2013 plays a crucial role in this project by enabling the creation of a compelling and informative presentation. It allows us to present the project's objectives, methodologies, findings, and recommendations in a structured and visually engaging manner.

INSIGHTS

A. Identify Missing Data and Deal with it Appropriately: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

- **Find the number of missing values in each column and calculate the missing percentage then treat them**

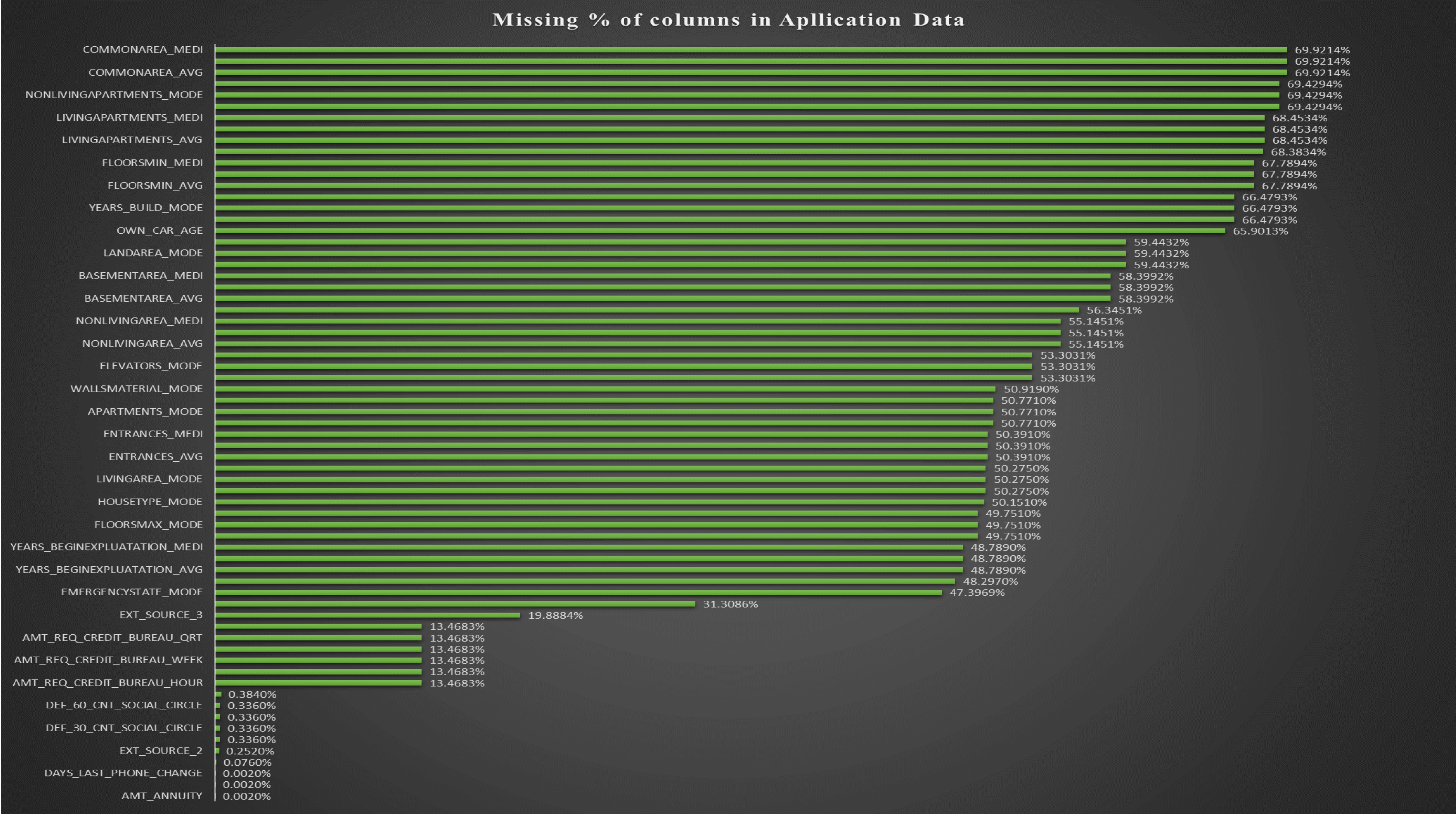
- **By the missing value percentage delete the columns with more than 50% null values**

- **Find the column which are insignificant**

1.flag_mobil column has all ones in the column only have one 0 so we can delete the column

- **Delete the columns which are irrevalant and also near to 50% missing values such as 49.75% mark columns**

- **For the remaining columns which has missing values impute the Mean/Mode/Median values with the null values.For these imputation I have used Jupyter notebook instead of Excel as it is more convinient to impute**



Missing % of columns in Apllication Data
COMMONAREA_MEDI
COMMONAREA_AVG
NONLIVINGAPARTMENTS_MODE
LIVINGAPARTMENTS_MEDI
LIVINGAPARTMENTS_AVG
FLOORSMIN_MEDI
FLOORSMIN_AVG
YEARS_BUILD_MODE
OWN_CAR_AGE
LANDAREA_MODE
BASEMENTAREA_MEDI
BASEMENTAREA_AVG
NONLIVINGAREA_MEDI
NONLIVINGAREA_AVG
ELEVATORS_MODE
WALLSMATERIAL_MODE
APARTMENTS_MODE
ENTRANCES_MEDI
ENTRANCES_AVG
LIVINGAREA_MODE
HOUSETYPE_MODE
FLOORSMAX_MODE
YEARS_BEGINEXPLUATATION_MEDI
YEARS_BEGINEXPLUATATION_AVG
EMERGENCYSTATE_MODE
EXT_SOURCE_3
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_HOUR
DEF_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
EXT_SOURCE_2
DAYS_LAST_PHONE_CHANGE
AMT_ANNUITY
69.9214%
69.9214%
69.9214%
69.4294%
69.4294%
69.4294%
68.4534%
68.4534%
68.4534%
68.3834%
67.7894%
67.7894%
67.7894%
66.4793%
66.4793%
66.4793%
65.9013%
59.4432%
59.4432%
59.4432%
58.3992%
58.3992%
58.3992%
56.3451%
55.1451%
55.1451%
55.1451%
53.3031%
53.3031%
53.3031%
50.9190%
50.7710%
50.7710%
50.7710%
50.3910%
50.3910%
50.3910%
50.2750%
50.2750%
50.2750%
50.1510%
49.7510%
49.7510%
49.7510%
48.7890%
48.7890%
48.7890%
48.2970%
47.3969%
31.3086%
19.8884%
13.4683%
13.4683%
13.4683%
13.4683%
13.4683%
13.4683%
0.3840%
0.3360%
0.3360%
0.3360%
0.3360%
0.2520%
0.0760%
0.0020%
0.0020%
0.0020%

DATASET 1-APPLICATION DATASET

COLUMNS WITH MORE THAN 50% MISSING VALUES DELETE THEM

- HOUSETYPE_MODE
- WALLSMATERIAL_MODE
- BASEMENTAREA_MEDI
- FLOORSMIN_MEDI
- LIVINGAREA_AVG
- ELEVATORS_AVG 0
- LANDAREA_AVG
- LIVINGAPARTMENTS_AVG
- LIVINGAREA_MODE
- ELEVATORS_MODE

- LANDAREA_MODE
- LIVINGAPARTMENTS_MODE
- LIVINGAREA_MEDI
- ELEVATORS_MEDI
- LANDAREA_MEDI
- LIVINGAPARTMENTS_MEDI
- ENTRANCES_AVG
- NONLIVINGAREA_AVG
- OWN_CAR_AGE
- APARTMENTS_AVG
- EXT_SOURCE_1
- YEARS_BUILD_MEDI

- FONDKAPREMONT_MODE
- ENTRANCES_MODE
- NONLIVINGAREA_MODE
- YEARS_BUILD_AVG
- NONLIVINGAPARTMENTS_AVG
- ENTRANCES_MEDI
- NONLIVINGAREA_MEDI
- YEARS_BUILD_MODE
- NONLIVINGAPARTMENTS_MODE

- YEARS_BUILD_MEDI
- NONLIVINGAPARTMENTS_MEDI
- APARTMENTS_MODE
- BASEMENTAREA_AVG

- FLOORSMIN_AVG
- COMMONAREA_AVG
- APARTMENTS_MEDI
- BASEMENTAREA_MODE
- FLOORSMIN_MODE
- COMMONAREA_MODE
- COMMONAREA_MEDI

COLUMNS WHICH ARE NOT NEACESSARY FOR THE ANALYSIS AND NEAR TO 50 % NULL VALUES DELETE THEM

- FLOORSMAX_AVG
- FLOORSMAX_MODE
- FLOORSMAX_MEDI
- EXT_SOURCE_2
- YEARS_BEGINEXPLUATATION_AVG
- YEARS_BEGINEXPLUATATION_MODE
- YEARS_BEGINEXPLUATATION_MEDI
- TOTALAREA_MODE
- EXT_SOURCE_3
- EMERGENCYSTATE_MODE
- FLAG_MOBIL

COLUMNS NEED TO IMPUTATE NULL VALUES WHITH MEAN/MEDAIN / MODE

- OCCUPATION_TYPE
- OBS_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_HOUR
- DEF_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_DAY
- OBS_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_WEEK
- DEF_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_GOODS_PRICE
- AMT_REQ_CREDIT_BUREAU_YEAR
- NAME_TYPE_SUITE

COLUMNS WITH 0.002% NULL VALUES WE CAN DIRECTLY DELETE THE NULL VALUES AS THE % IS VERY INSIGNIFICANT

- CNT_FAM_MEMBERS
- AMT_ANNUITY
- DAYS_LAST_PHONE_CHANGE

Outliers are present in below columns
So impute the Null values with median

- OBS_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_HOUR
- DEF_30_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_DAY
- OBS_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_WEEK
- DEF_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_GOODS_PRICE
- AMT_REQ_CREDIT_BUREAU_YEAR

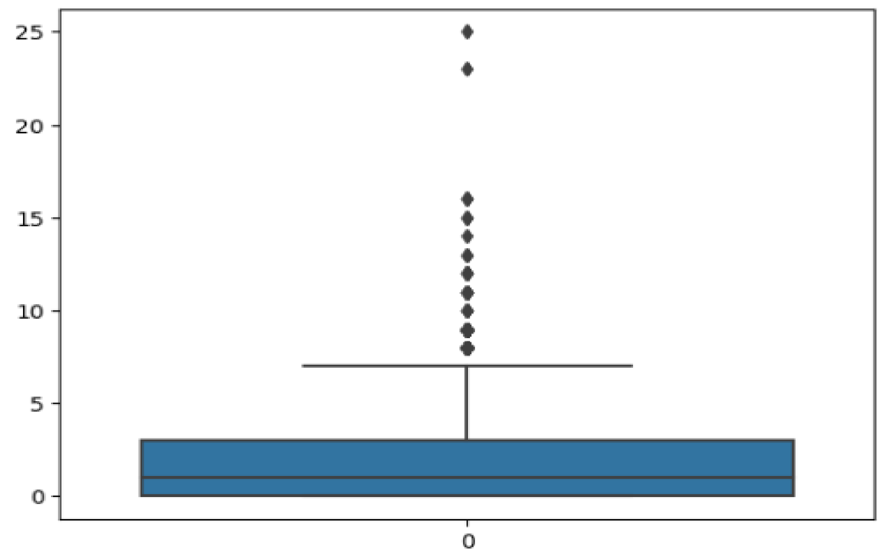
Repeat the same process for the remaining columns detailed process is present in the Jupyter Notebook kindly check it

Step 1: Find the locations of null values in a Specified column using below function
data.loc[data.AMT_REQ_CREDIT_BUREAU_YEAR.isnull()]
Step 2: Check if there exists outliers using box plot
Step 3: fill the null values using fillna() function with median values
Step 4:Check again after filling null values they exist or not

EX: AMT_REQ_CREDIT_BUREAU_YEAR

```
In [30]: sns.boxplot(data.AMT_REQ_CREDIT_BUREAU_YEAR)
```
executed in 142ms, finished 21:14:26 2023-09-15

```
Out[30]: &lt;Axes: &gt;
```

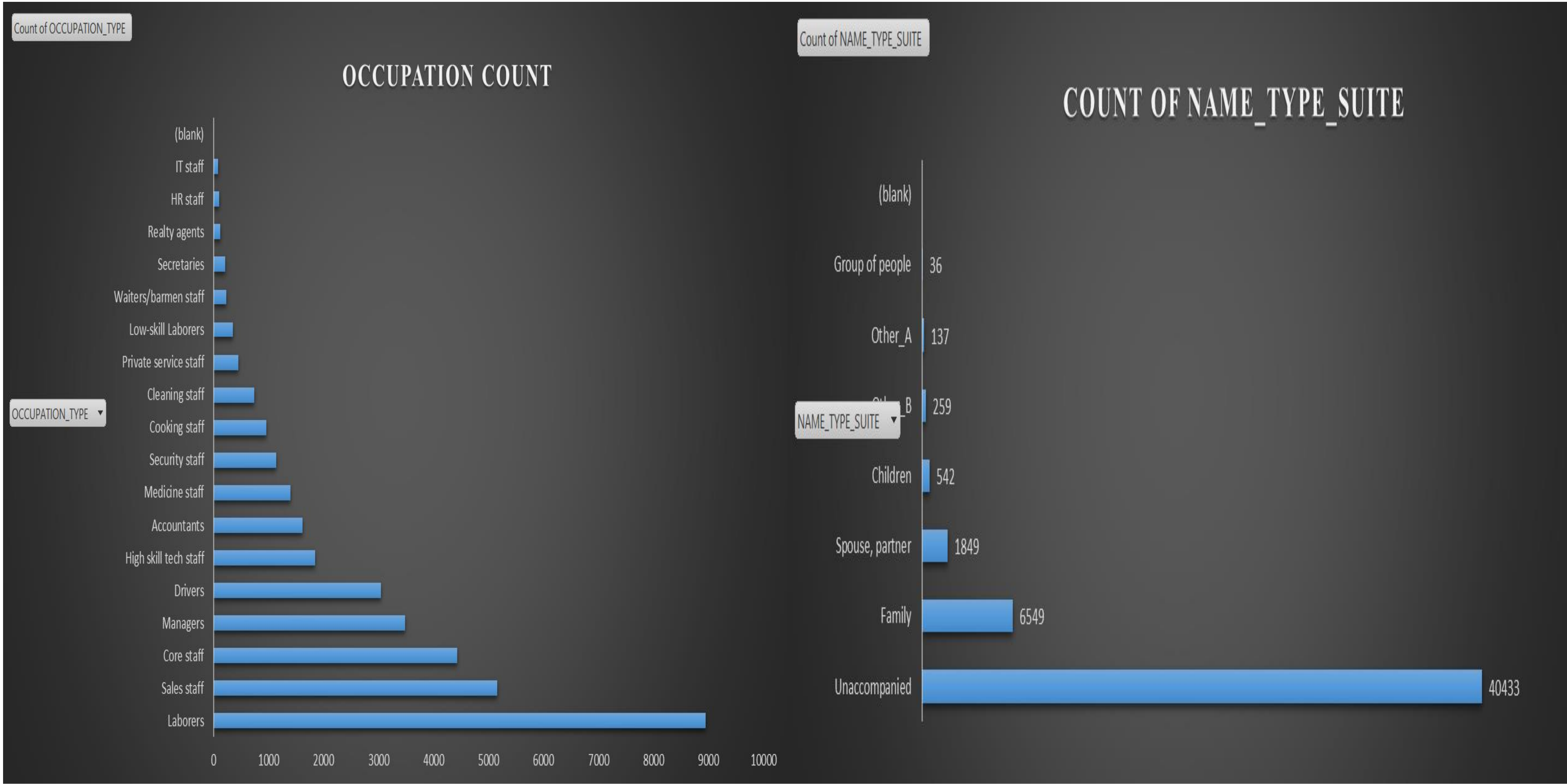



```
In [31]: data.AMT_REQ_CREDIT_BUREAU_YEAR.fillna(data.AMT_REQ_CREDIT_BUREAU_YEAR.median(),inplace=True)
```
executed in 8ms, finished 21:14:26 2023-09-15

```
In [32]: data.AMT_REQ_CREDIT_BUREAU_YEAR.isnull().any()
```
executed in 31ms, finished 21:14:26 2023-09-15

```
Out[32]: False
```

MODE VISUALIZATIONS OF OCCUPATION_TYPE AND NAME_TYPE_SUITE



Fill the Categorical columns null values with Mode

Step 1: Find the locations of null values in a Specified column using below function
data.loc[data.OCCUPATION_TYPE.isnull()]
Step 2: Check the mode of the column using bar or pie plot
Step 3: fill the null values using fillna() function with median values
Step 4:Check again after filling null values they exist or not

EX: OCCUPATION_TYPE

OCCUPATION_TYPE
NAME_TYPE_SUITE

Repeat the same process for the next column detailed process is present in the Jupyter Notebook kindly check it

```
In [53]: data.OCCUPATION_TYPE.mode()
executed in 16ms, finished 21:14:27 2023-09-15
Out[53]: 0    Laborers
Name: OCCUPATION_TYPE, dtype: object

In [54]: data.loc[data.OCCUPATION_TYPE.isnull()].head()
executed in 53ms, finished 21:14:27 2023-09-15
Out[54]:
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TO
8	100011	0	Cash loans	F	N	Y	0	1125
11	100015	0	Cash loans	F	N	Y	0	384
23	100027	0	Cash loans	F	N	Y	0	832
28	100033	0	Cash loans	M	Y	Y	0	2700
30	100035	0	Cash loans	F	N	Y	0	2925

```
In [55]: data.OCCUPATION_TYPE.fillna(data.OCCUPATION_TYPE.mode()[0],inplace=True)
executed in 25ms, finished 21:14:27 2023-09-15

In [56]: data.OCCUPATION_TYPE.isnull().any()
executed in 16ms, finished 21:14:27 2023-09-15
Out[56]: False
```

DATASET 2-PREVIOUS APPLICATION DATASET

COLUMNS WITH MORE THAN 50% MISSING VALUES DELETE THEM

- NAME_TYPE_SUITE
- RATE_INTEREST_PRIMARY
- RATE_INTEREST_PRIVILEGED
- AMT_DOWN_PAYMENT
- RATE_DOWN_PAYMENT

COLUMNS WHICH ARE NOT NECESSARY FOR THE ANALYSIS DELETE THEM

- WEEKDAY_APPR_PROCESS_START
- HOUR_APPR_PROCESS_START
- FLAG_LAST_APPL_PER_CONTRACT
- NFLAG_LAST_APPL_IN_DAY

COLUMN WITH 0.016% NULL VALUES WE CAN DIRECTLY DELETE THE NULL VALUES AS THE % IS VERY INSIGNIFICANT

- PRODUCT_COMBINATION

Outliers are present in below columns
So impute the Null values with median

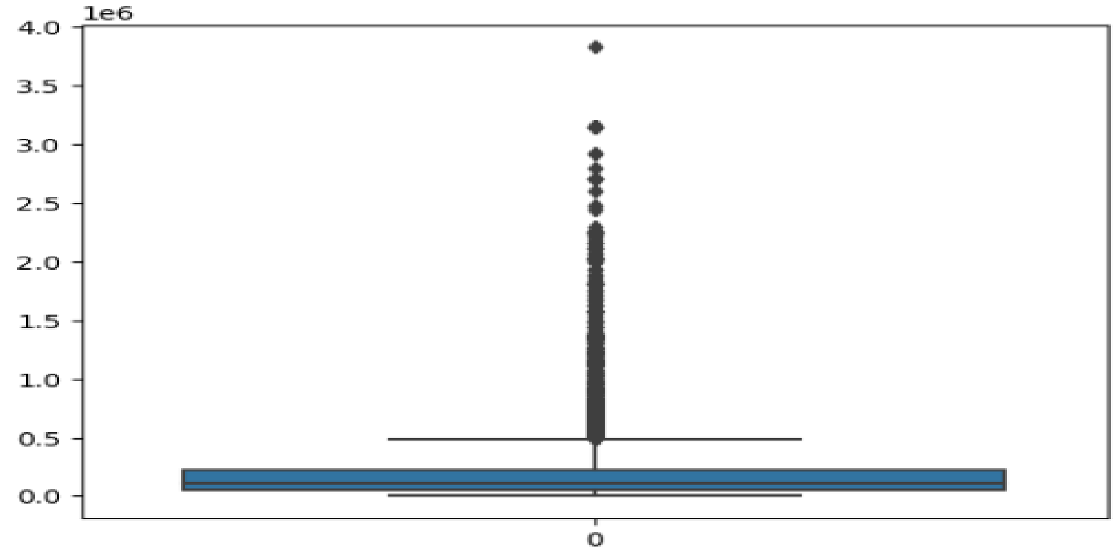
- AMT_ANNUITY
- AMT_GOODS_PRICE
- CNT_PAYMENT

Repeat the same process for the remaining columns detailed process is present in the Jupyter Notebook kindly check it

Step 1: Find the locations of null values in a Specified column using below function
data.loc[data.AMT_GOODS_PRICE.isnull()]
Step 2: Check if there exists outliers using box plot
Step 3: fill the null values using fillna() function with median values
Step 4:Check again after filling null values they exist or not

EX: AMT_GOODS_PRICE

```
In [67]: sns.boxplot(df.AMT_GOODS_PRICE)
executed in 236ms, finished 21:14:29 2023-09-15
Out[67]: <Axes: >
```




```
In [68]: df.AMT_GOODS_PRICE.fillna(df.AMT_GOODS_PRICE.median(),inplace=True)
executed in 12ms, finished 21:14:29 2023-09-15
In [69]: df.AMT_GOODS_PRICE.isnull().any()
executed in 18ms, finished 21:14:29 2023-09-15
Out[69]: False
```

Fill the Categorical columns null values with Mode

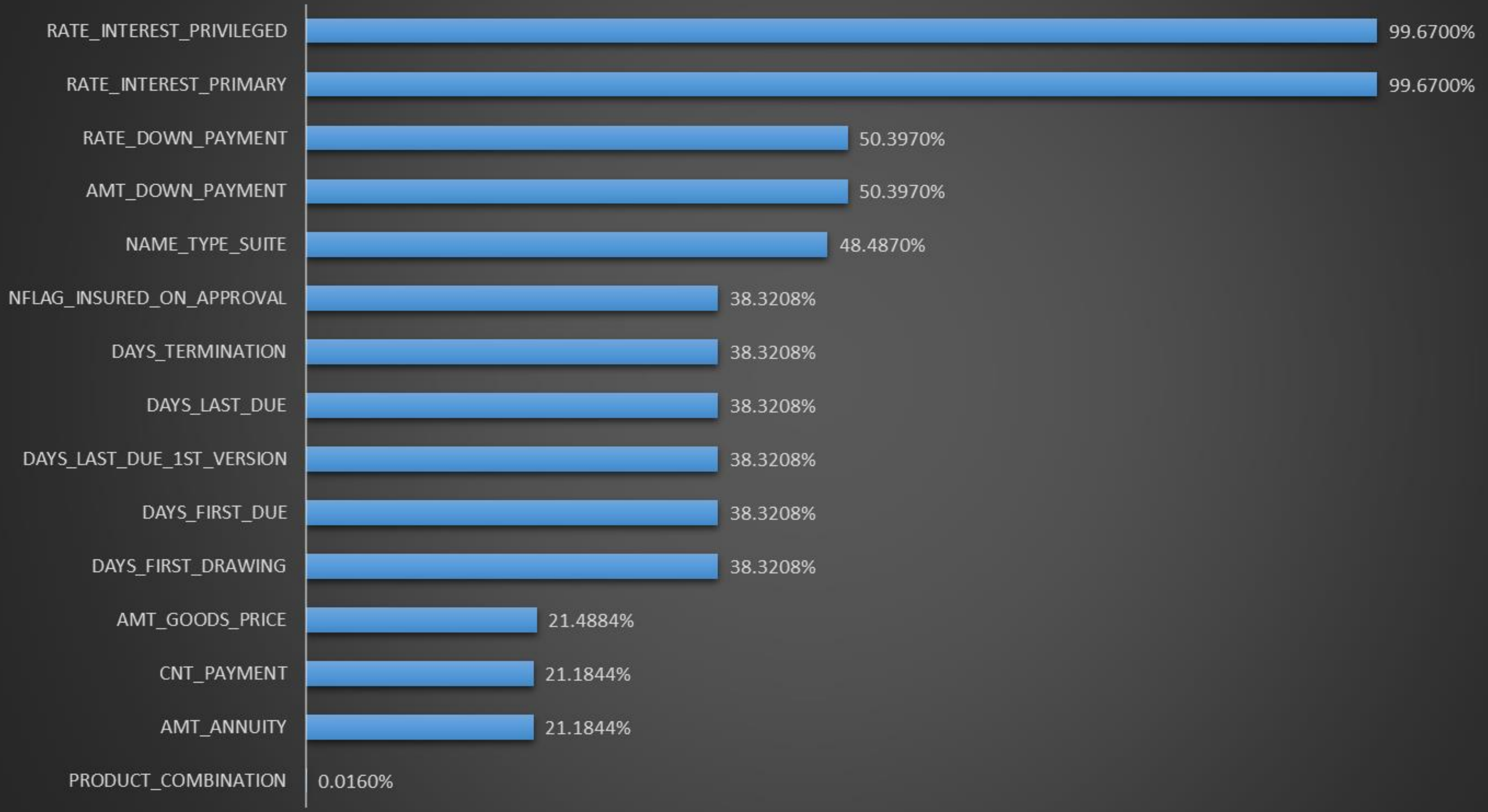
Step 1: Find the locations of null values in a Specified column using below function
data.loc[NFLAG_INSURED_ON_APPROVAL.isnull()]
Step 2: Check the mode of the column using bar or pie plot
Step 3: fill the null values using fillna() function with median values
Step 4:Check again after filling null values they exist or not

EX: NFLAG_INSURED_ON_APPROVAL

NFLAG_INSURED_ON_APPROVAL

```
In [75]: df.NFLAG_INSURED_ON_APPROVAL.mode()
executed in 13ms, finished 21:14:29 2023-09-15

Out[75]: 0    0.0
Name: NFLAG_INSURED_ON_APPROVAL, dtype: float64

In [76]: df.NFLAG_INSURED_ON_APPROVAL.value_counts()
executed in 15ms, finished 21:14:29 2023-09-15

Out[76]: 0.0    20898
1.0     9941
Name: NFLAG_INSURED_ON_APPROVAL, dtype: int64

In [77]: df.NFLAG_INSURED_ON_APPROVAL.fillna(df.NFLAG_INSURED_ON_APPROVAL.mode()[0],inplace=True)
executed in 17ms, finished 21:14:29 2023-09-15

In [78]: df.NFLAG_INSURED_ON_APPROVAL.isnull().any()
executed in 16ms, finished 21:14:29 2023-09-15

Out[78]: False
```

MISSING % OF PREVIOUS APPLICATION DATA



B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

OUTLIERS: An outliers are data points that goes far outside the average value of a group of statistics.

I have used Jupyter Notebook for outlier Detection using Boxplots from Matplotlib library and qantile functions

Function used to detect the outlier

```
def find_outliers_IQR(data):

    q1=data.quantile(0.25)

    q3=data.quantile(0.75)

    IQR=q3-q1

    outliers = data[((data<(q1-1.5*IQR)) | (data>(q3+1.5*IQR)))]

    return outliers
```

OUTLIERS IN APPLICATION DATASET

AMT_ANNUITY

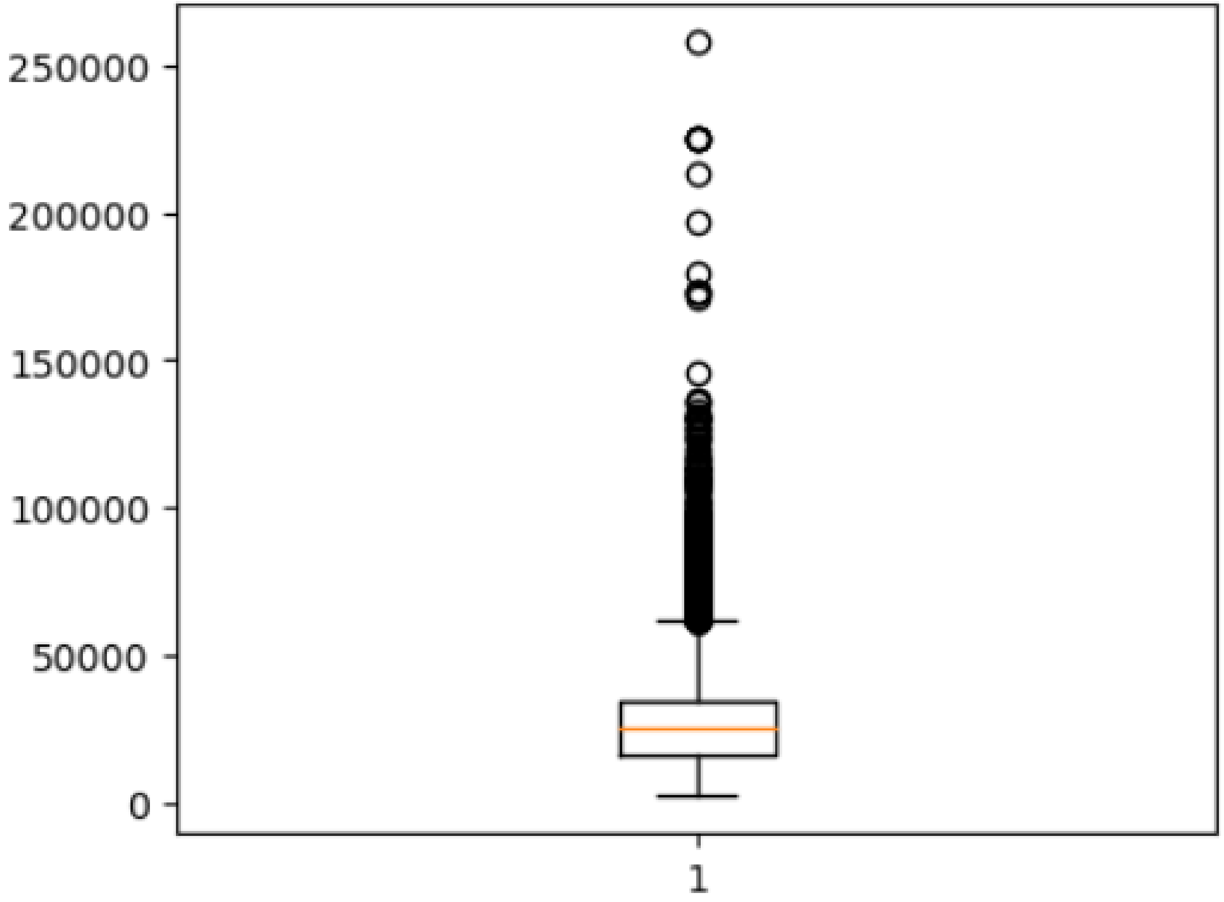
```
outliers = find_outliers_IQR(data["AMT_ANNUITY"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 17ms, finished 21:14:31 2023-09-15

```
number of outliers:1188
max outlier value:258025.5
min outlier value:61875.0
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data.AMT_ANNUITY)
pyplot.show()
```

executed in 115ms, finished 21:14:31 2023-09-15



CNT_CHILDREN

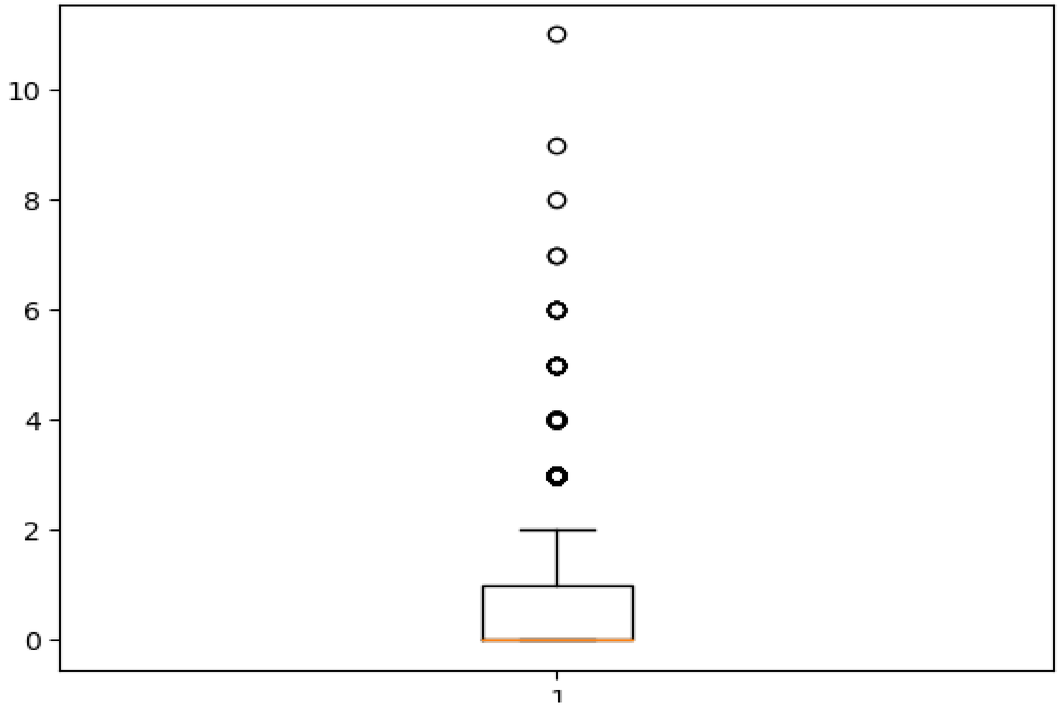
```
outliers = find_outliers_IQR(data["CNT_CHILDREN"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 16ms, finished 21:14:30 2023-09-15

```
number of outliers:723
max outlier value:11
min outlier value:3
```

```
pyplot.boxplot(data.CNT_CHILDREN)
pyplot.show()
```

executed in 90ms, finished 21:14:30 2023-09-15



AMT_INCOME_TOTAL

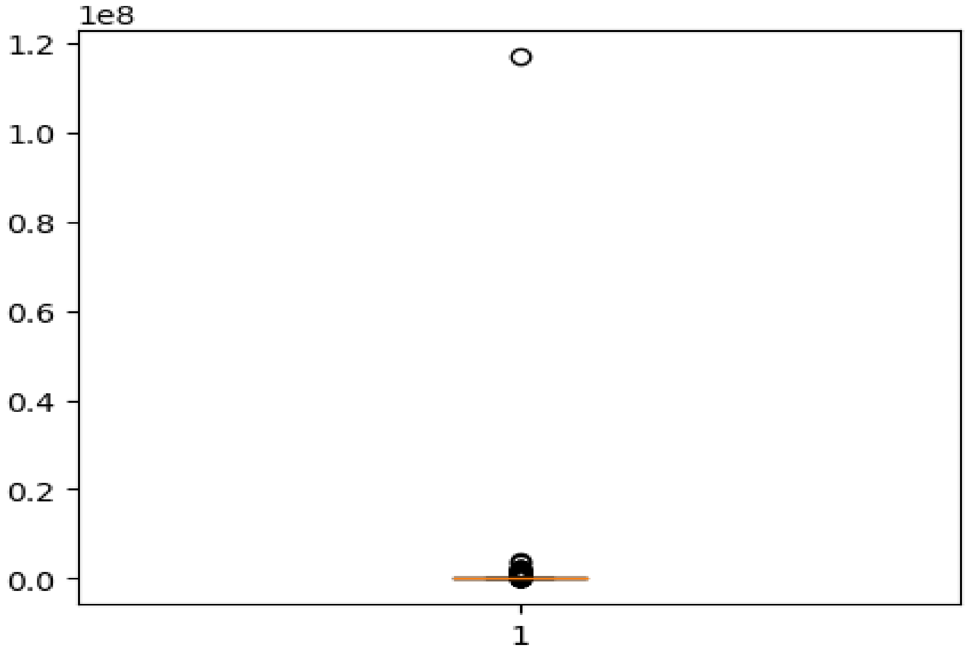
```
outliers = find_outliers_IQR(data["AMT_INCOME_TOTAL"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 16ms, finished 21:14:30 2023-09-15

```
number of outliers:2294
max outlier value:117000000.0
min outlier value:338746.5
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data.AMT_INCOME_TOTAL)
pyplot.show()
```

executed in 106ms, finished 21:14:30 2023-09-15



AMT_GOODS_PRICE

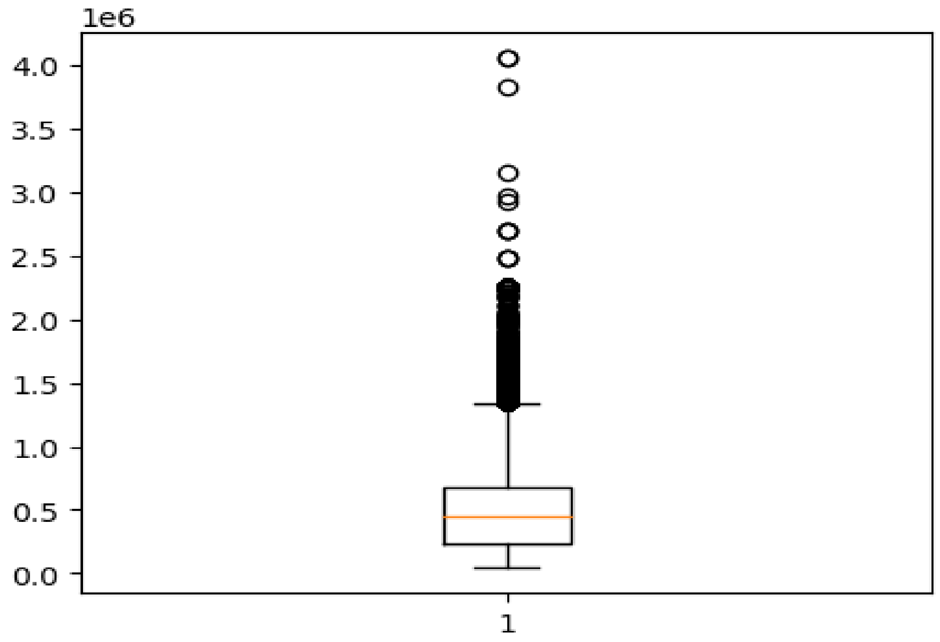
```
outliers = find_outliers_IQR(data["AMT_GOODS_PRICE"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 17ms, finished 21:14:30 2023-09-15

```
number of outliers:2387
max outlier value:4050000.0
min outlier value:1345500.0
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data.AMT_GOODS_PRICE)
pyplot.show()
```

executed in 119ms, finished 21:14:31 2023-09-15



AMT_CREDIT

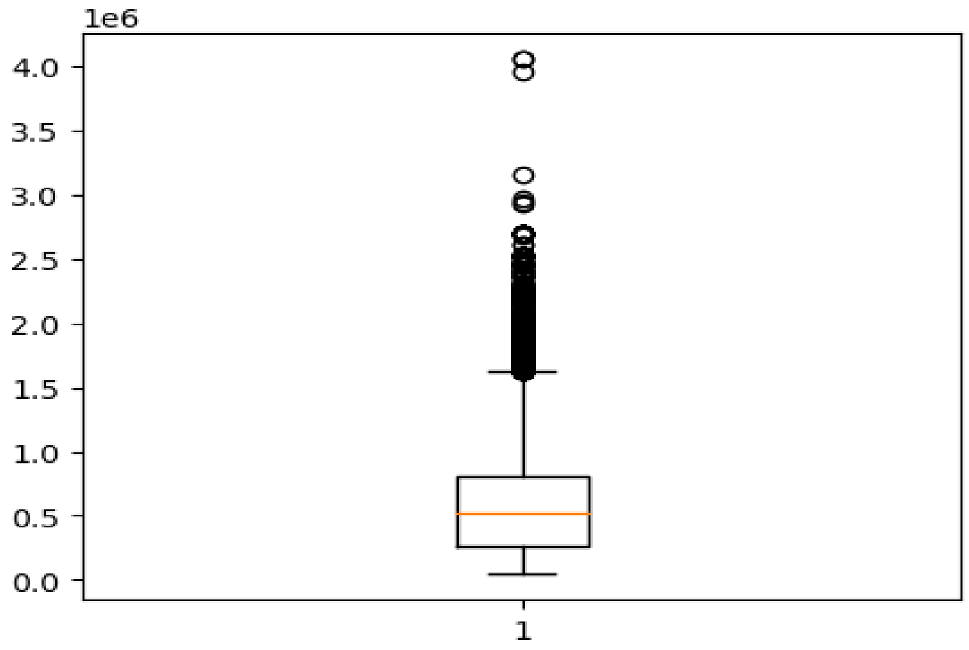
```
outliers = find_outliers_IQR(data["AMT_CREDIT"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 17ms, finished 21:14:31 2023-09-15

```
number of outliers:1063
max outlier value:4050000.0
min outlier value:1620000.0
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data.AMT_CREDIT)
pyplot.show()
```

executed in 113ms, finished 21:14:31 2023-09-15



EMPLOYEEMENT_YEARS

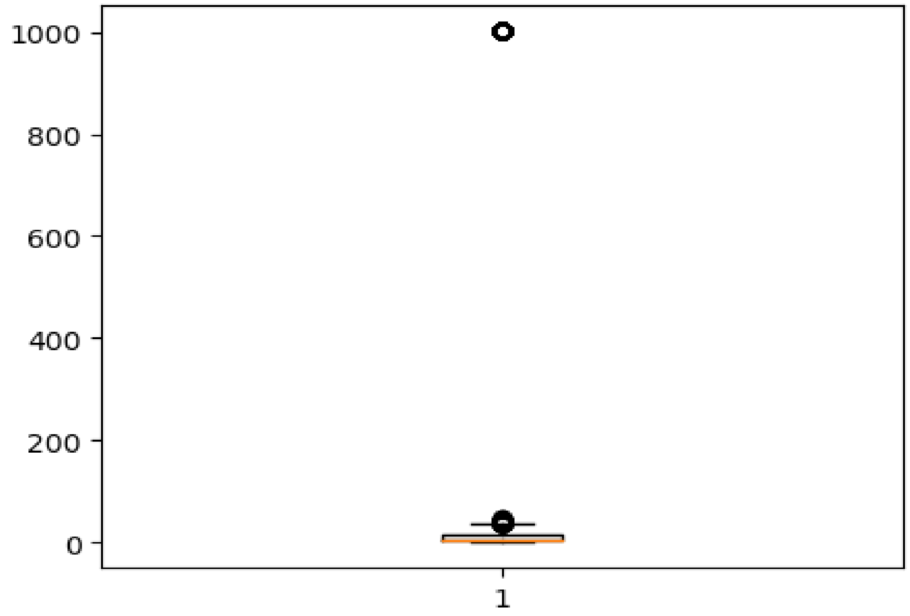
```
outliers = find_outliers_IQR(data["EMPLOYEEMENT YEARS"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 17ms, finished 21:14:31 2023-09-15

```
number of outliers:9076
max outlier value:1001
min outlier value:36
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data["EMPLOYEEMENT YEARS"])
pyplot.show()
```

executed in 115ms, finished 21:14:31 2023-09-15



DAYS_LAST_PHONE_CHANGE

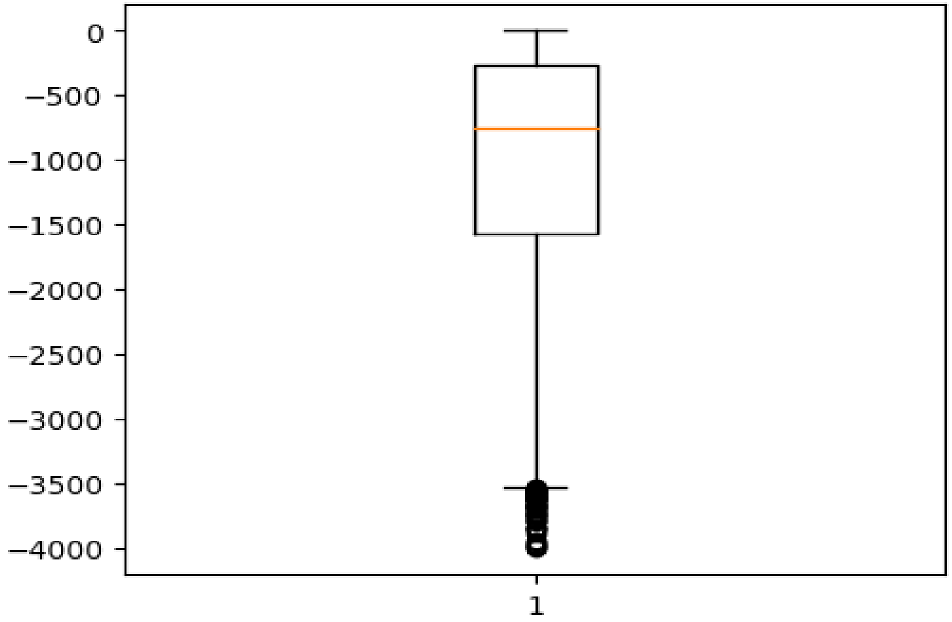
```
outliers = find_outliers_IQR(data["DAYS_LAST_PHONE_CHANGE"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 15ms, finished 21:14:31 2023-09-15

```
number of outliers:63
max outlier value:-3528
min outlier value:-4002
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data.DAYS_LAST_PHONE_CHANGE)
pyplot.show()
```

executed in 106ms, finished 21:14:31 2023-09-15



REGISTRATION_YEARS

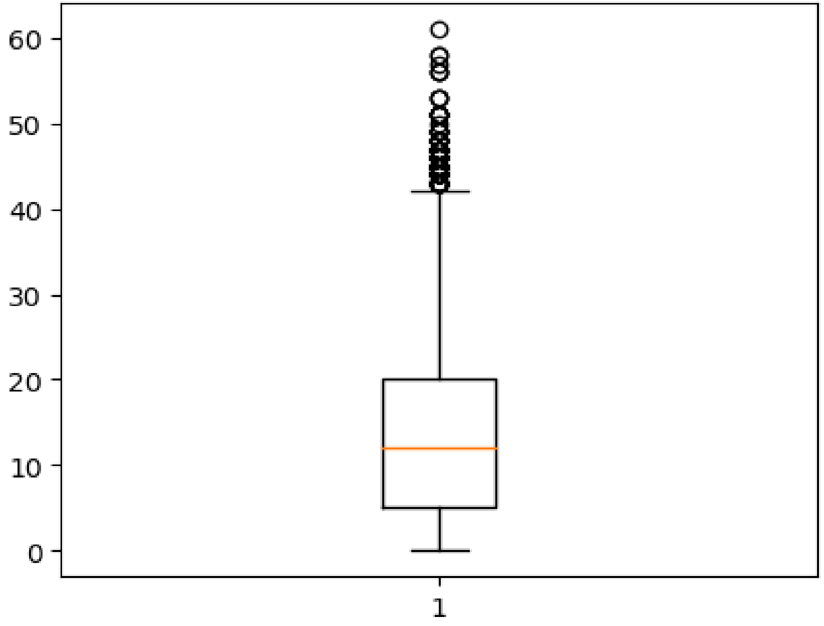
```
outliers = find_outliers_IQR(data["REGISTRATION YEARS"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 16ms, finished 21:14:31 2023-09-15

```
number of outliers:115
max outlier value:61
min outlier value:43
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data["REGISTRATION YEARS"])
pyplot.show()
```

executed in 119ms, finished 21:14:31 2023-09-15



AGE IN YEARS

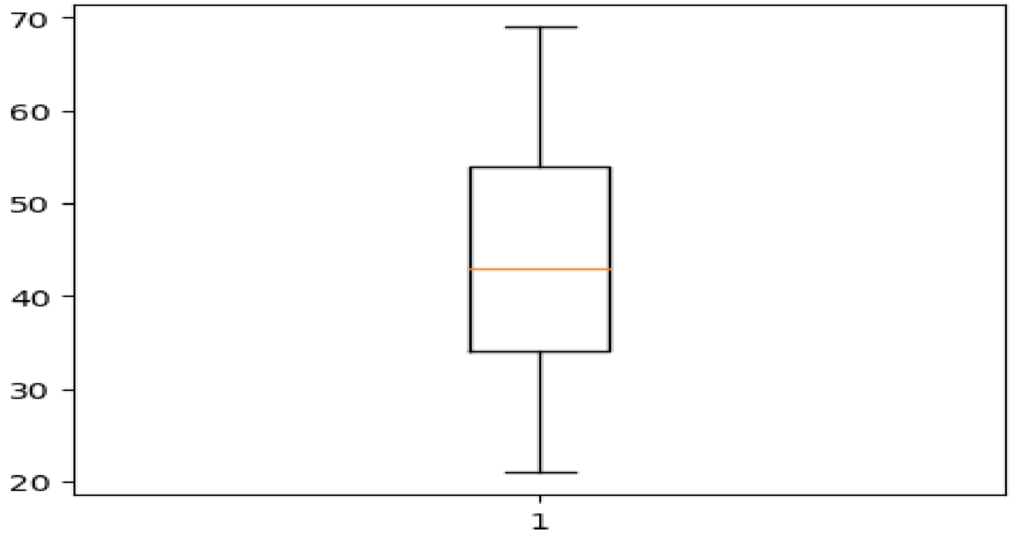
```
outliers = find_outliers_IQR(data["AGE IN YEARS"])

print("number of outliers:"+ str(len(outliers)))

print("max outlier value:"+ str(outliers.max()))

print("min outlier value:"+ str(outliers.min()))
```

executed in 16ms, finished 21:14:31 2023-09-15

```
number of outliers:0
max outlier value:nan
min outlier value:nan
```

```
fig = pyplot.figure(figsize =(5, 4))
pyplot.boxplot(data["AGE IN YEARS"])
pyplot.show()
```

executed in 88ms, finished 21:14:31 2023-09-15

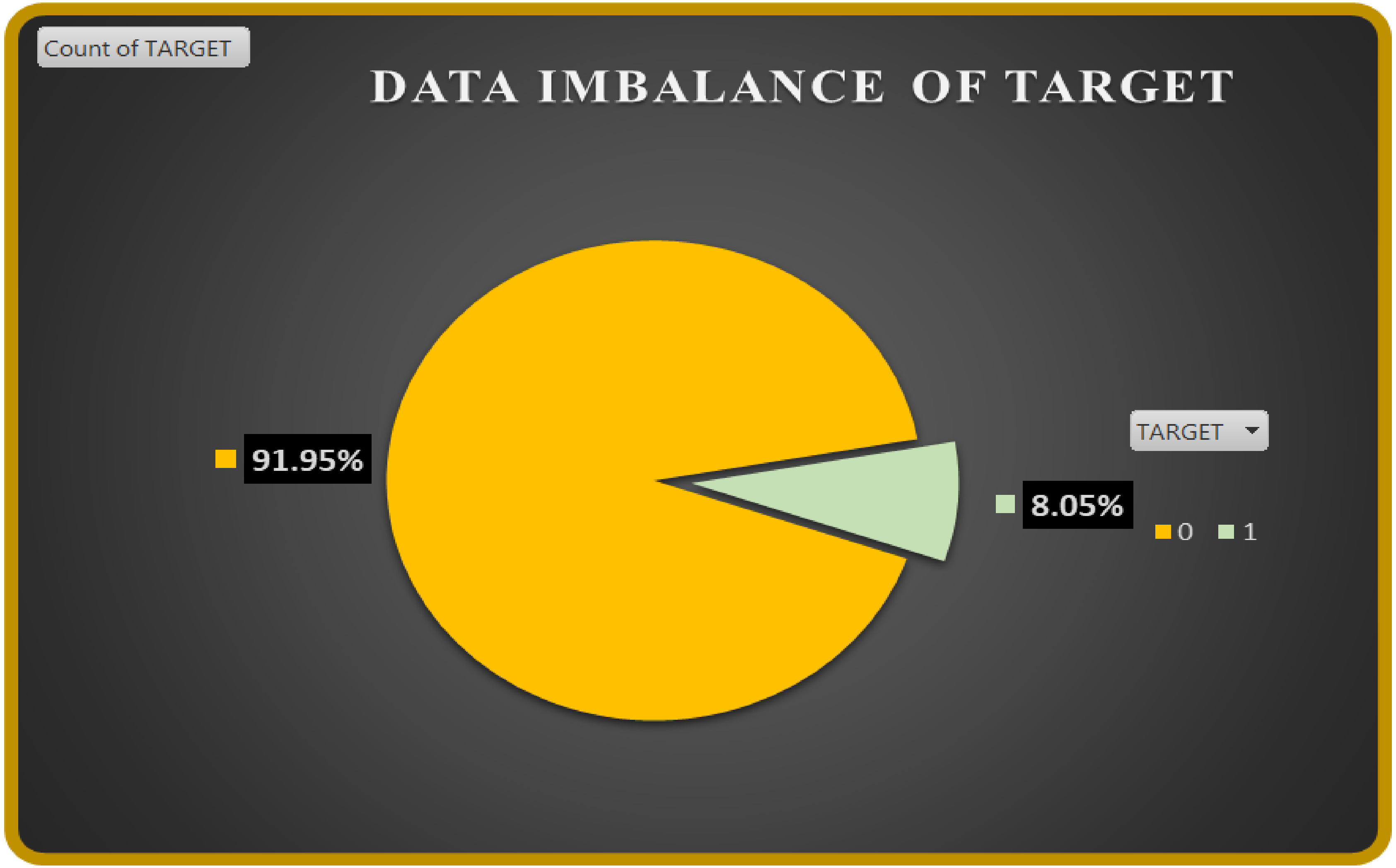


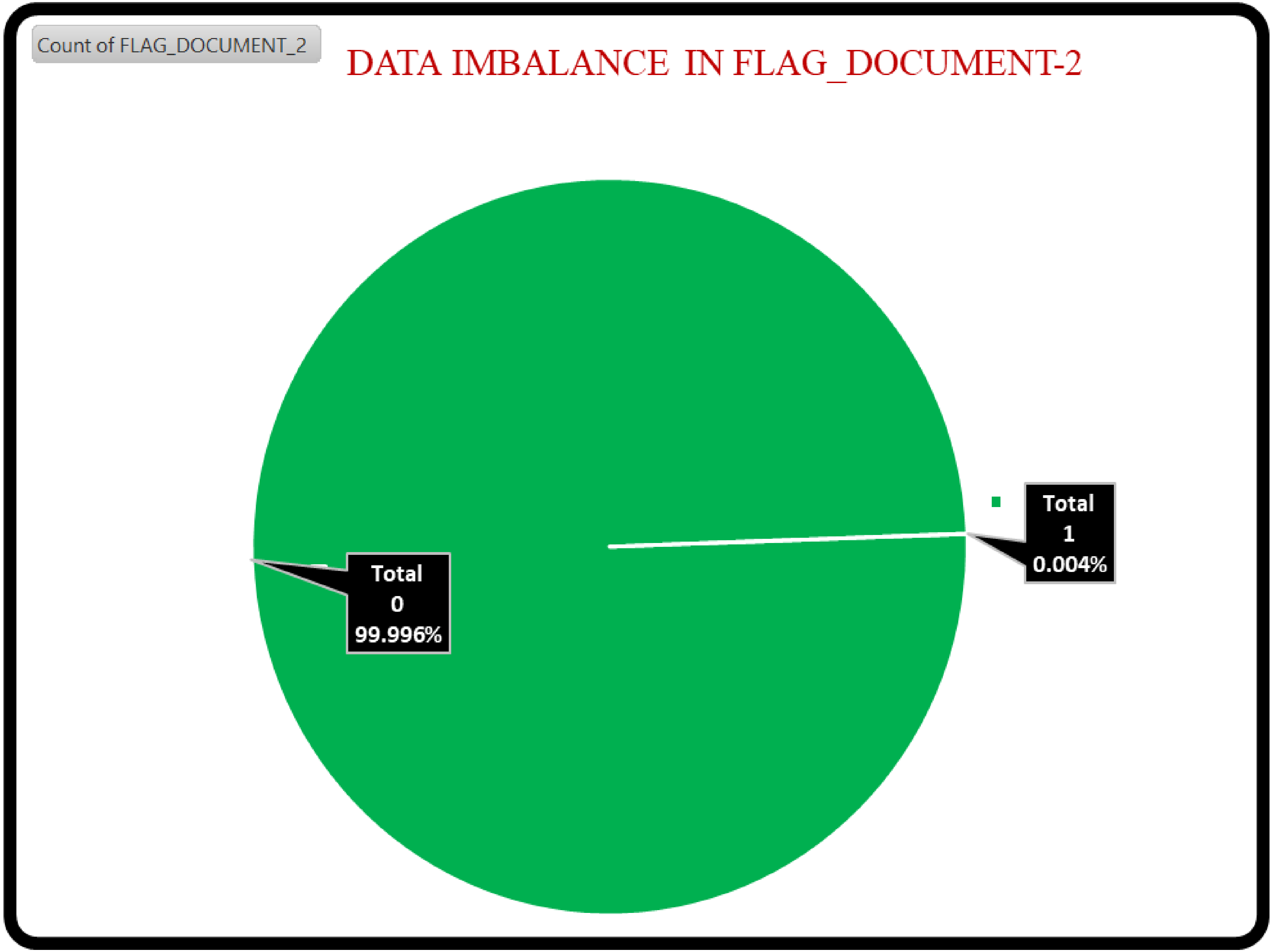
AGE Column has No Outliers

C. Analyze Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

TARGET	Count of TARGET
0	45970
1	4026
Grand Total	**49996**

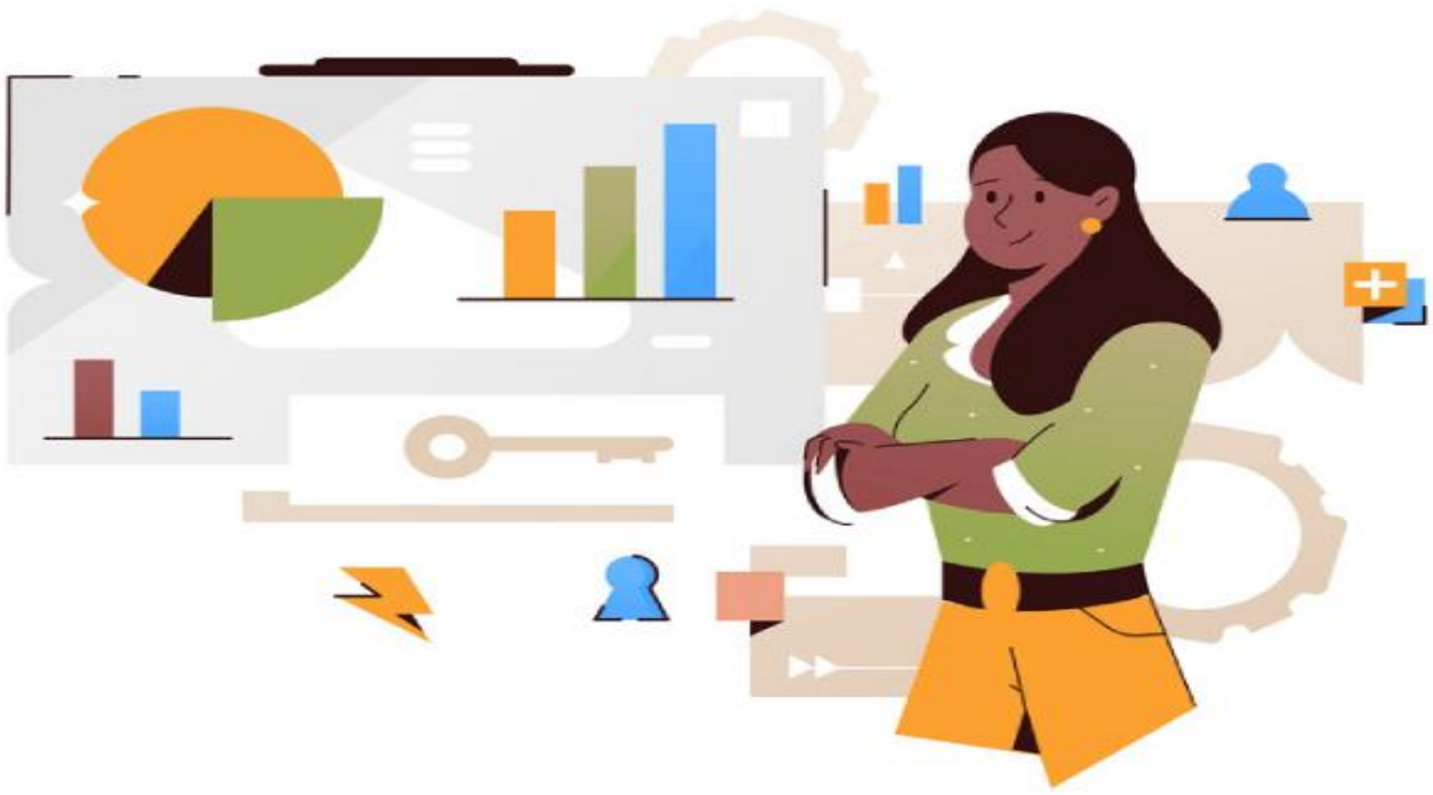




FLAG DOCUMENT -2	Count of FLAG_DOCUMENT_2
0	49994
1	2
Grand Total	**49996**

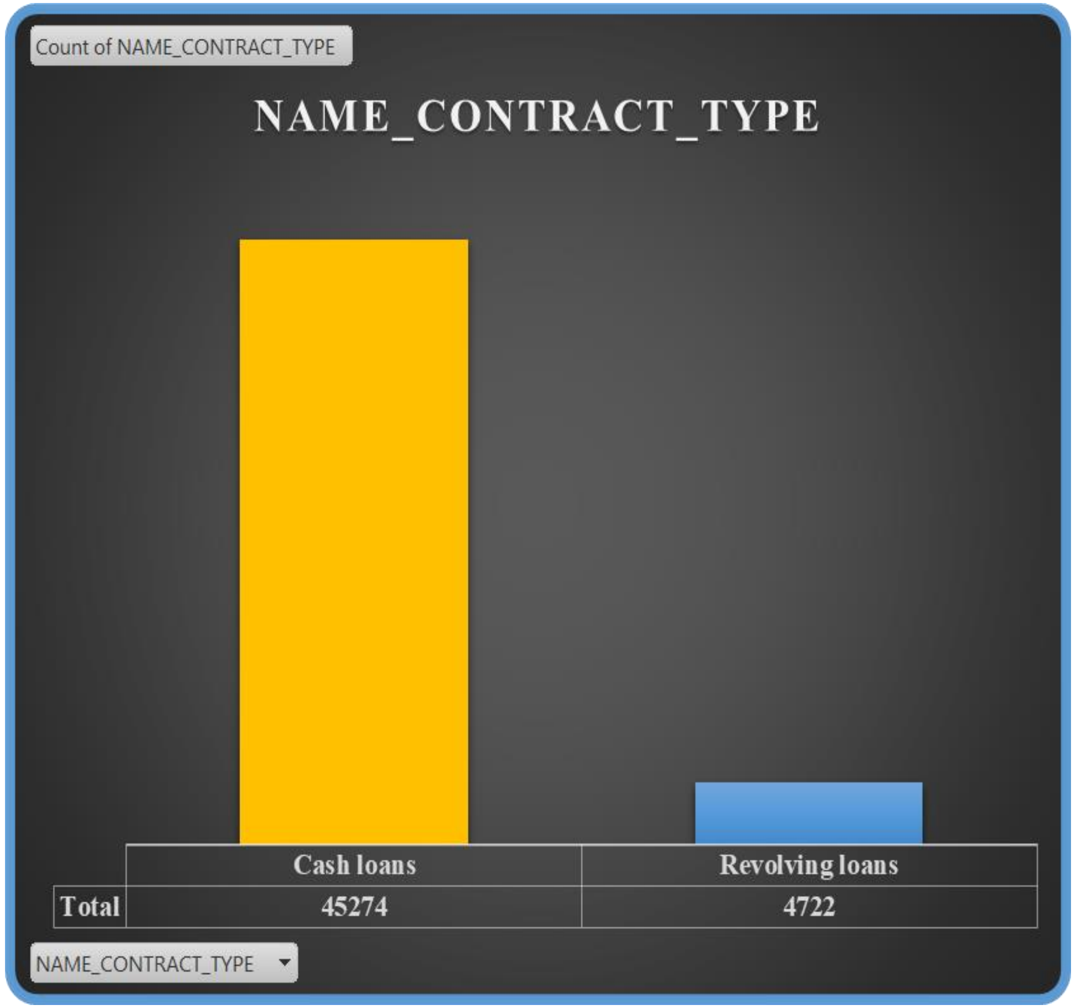
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

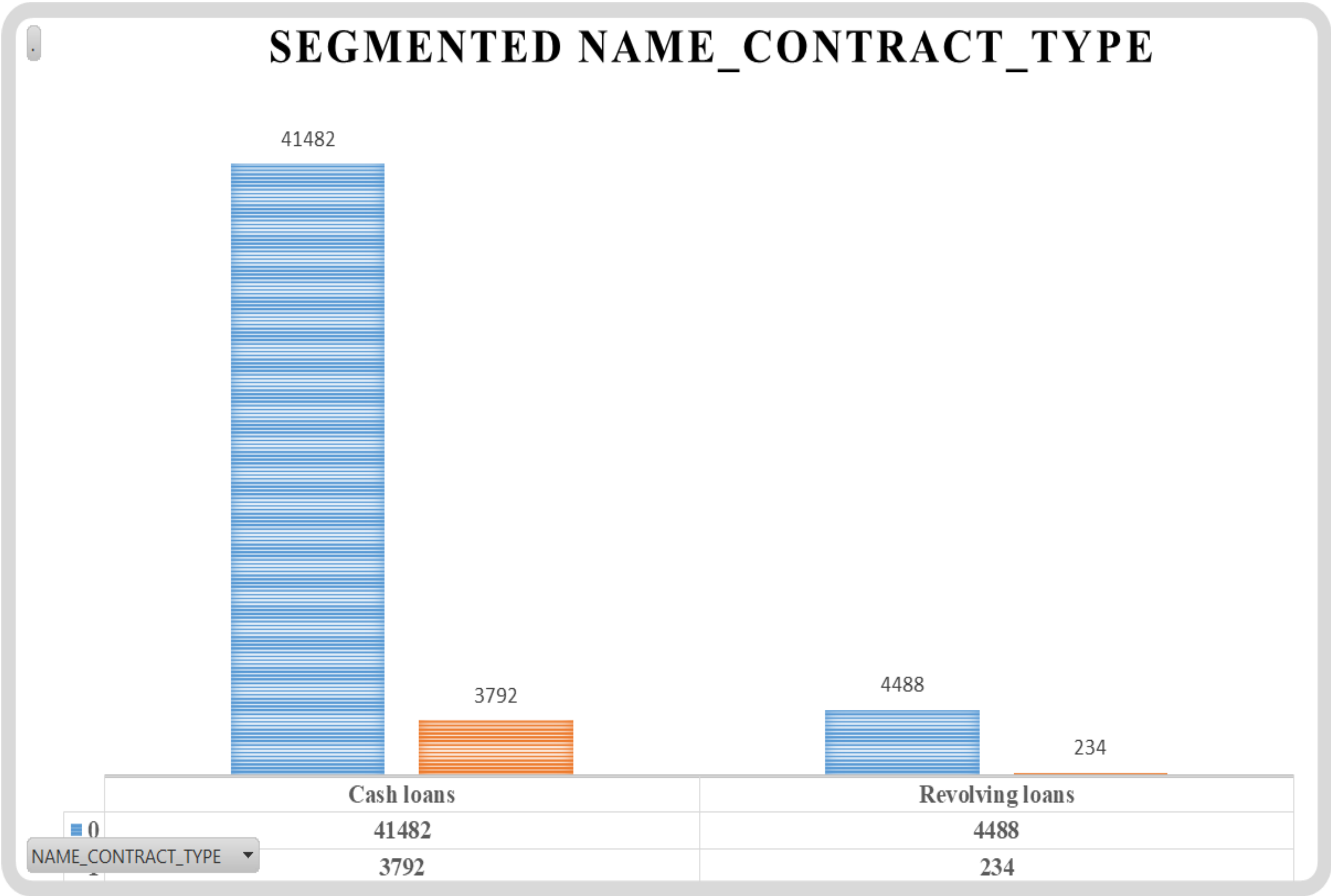
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.



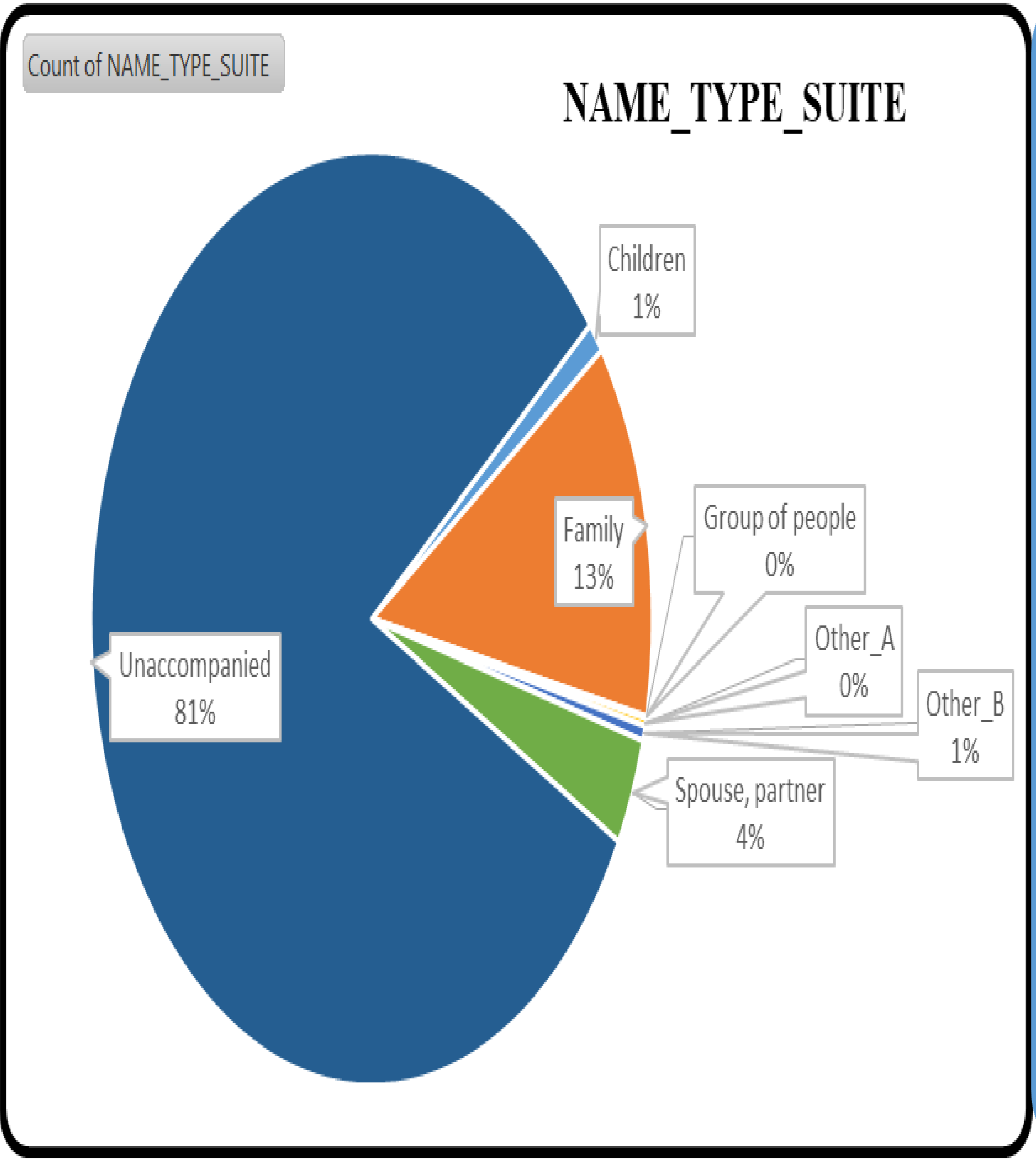
DATASET 1-APPLICATION DATASET VISAULIZATIONS

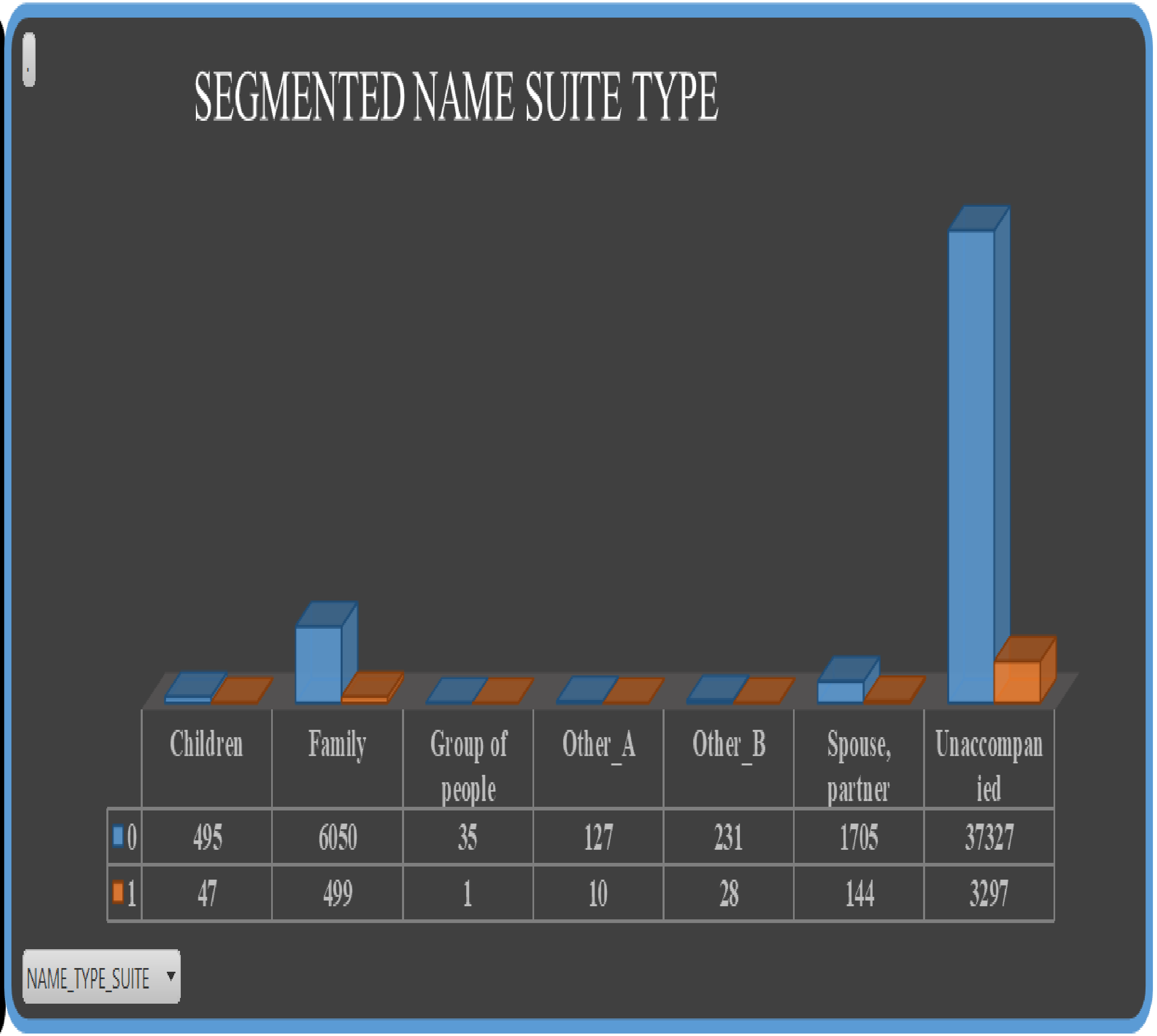
UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS





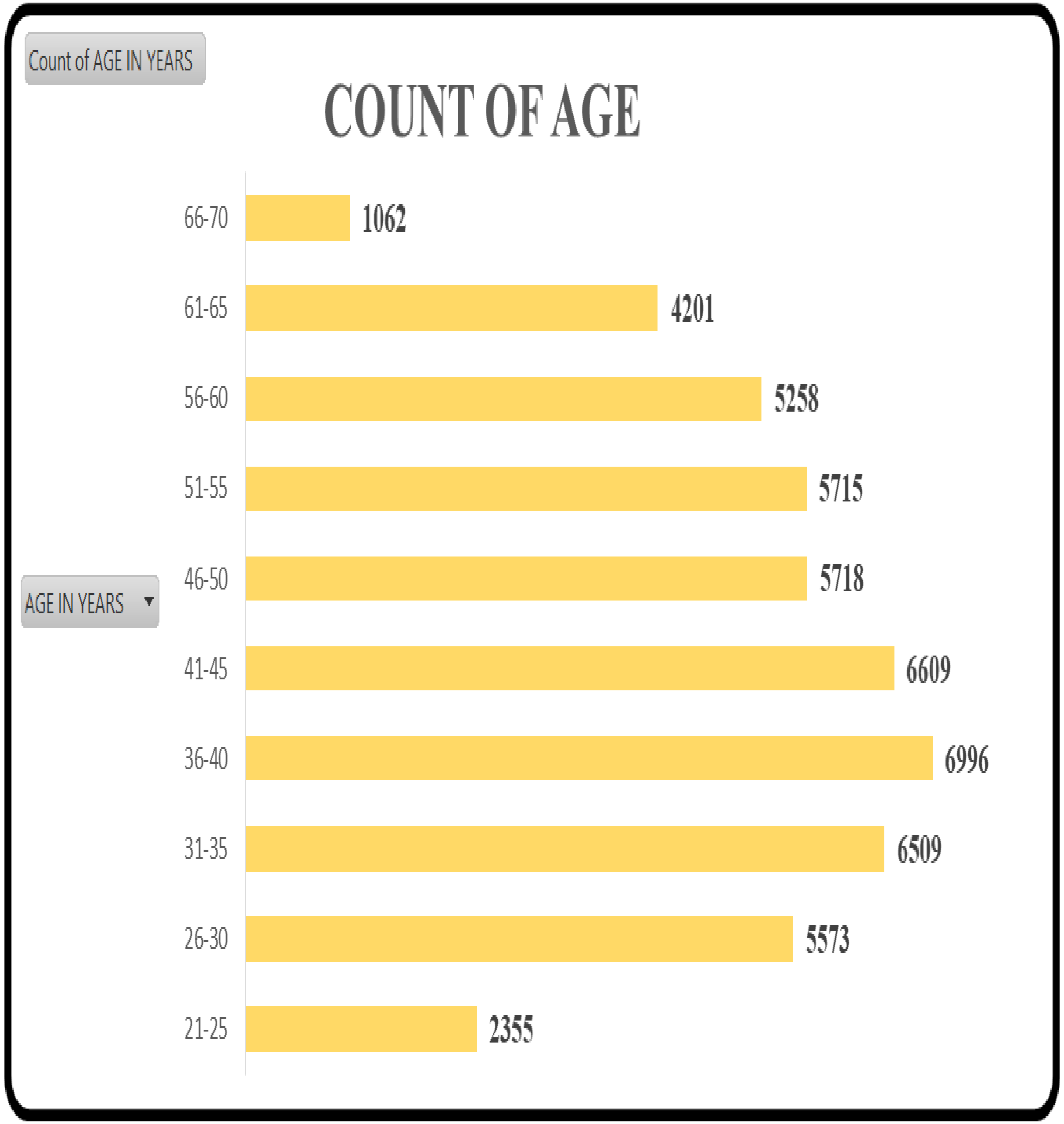
UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS

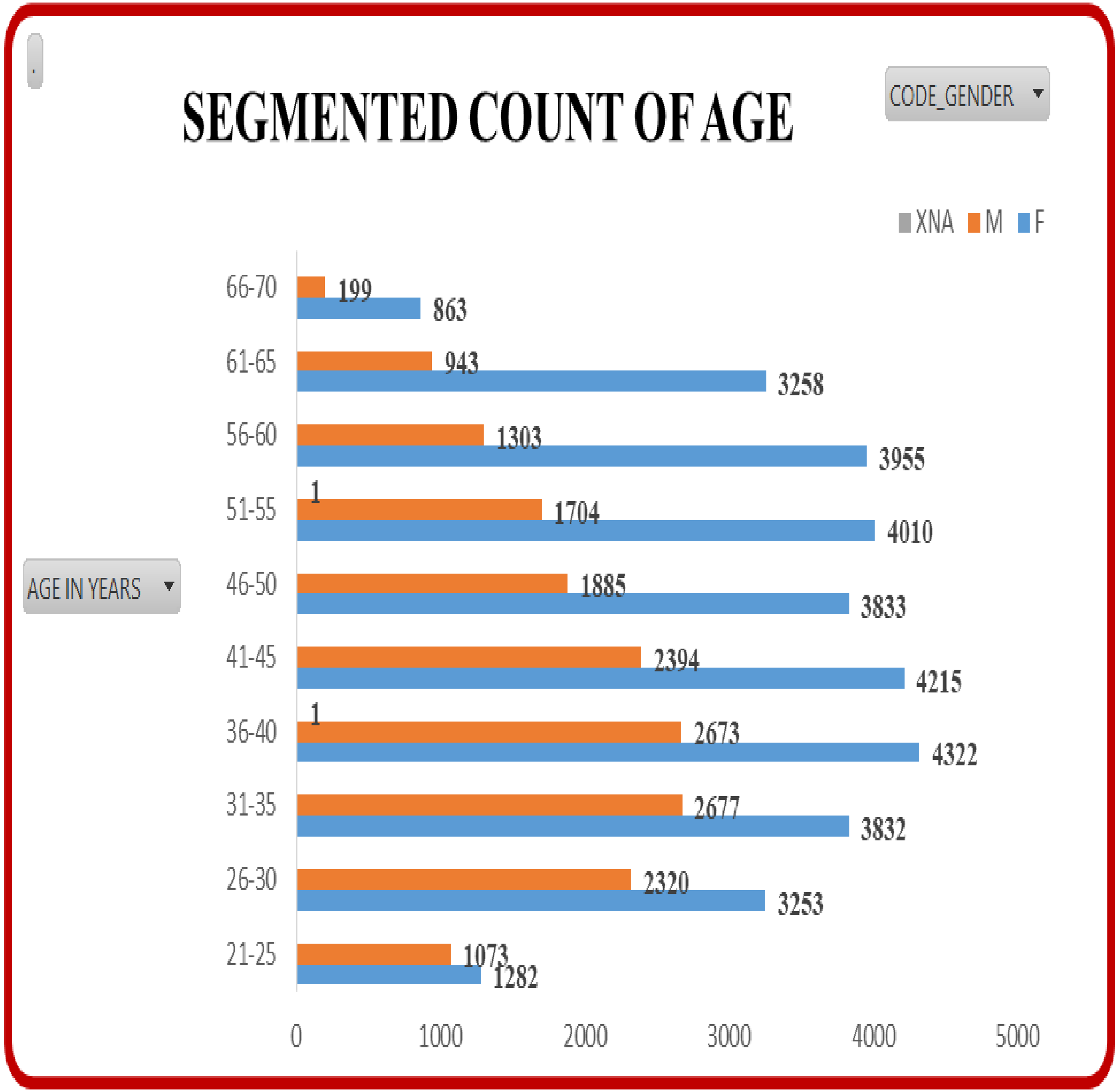




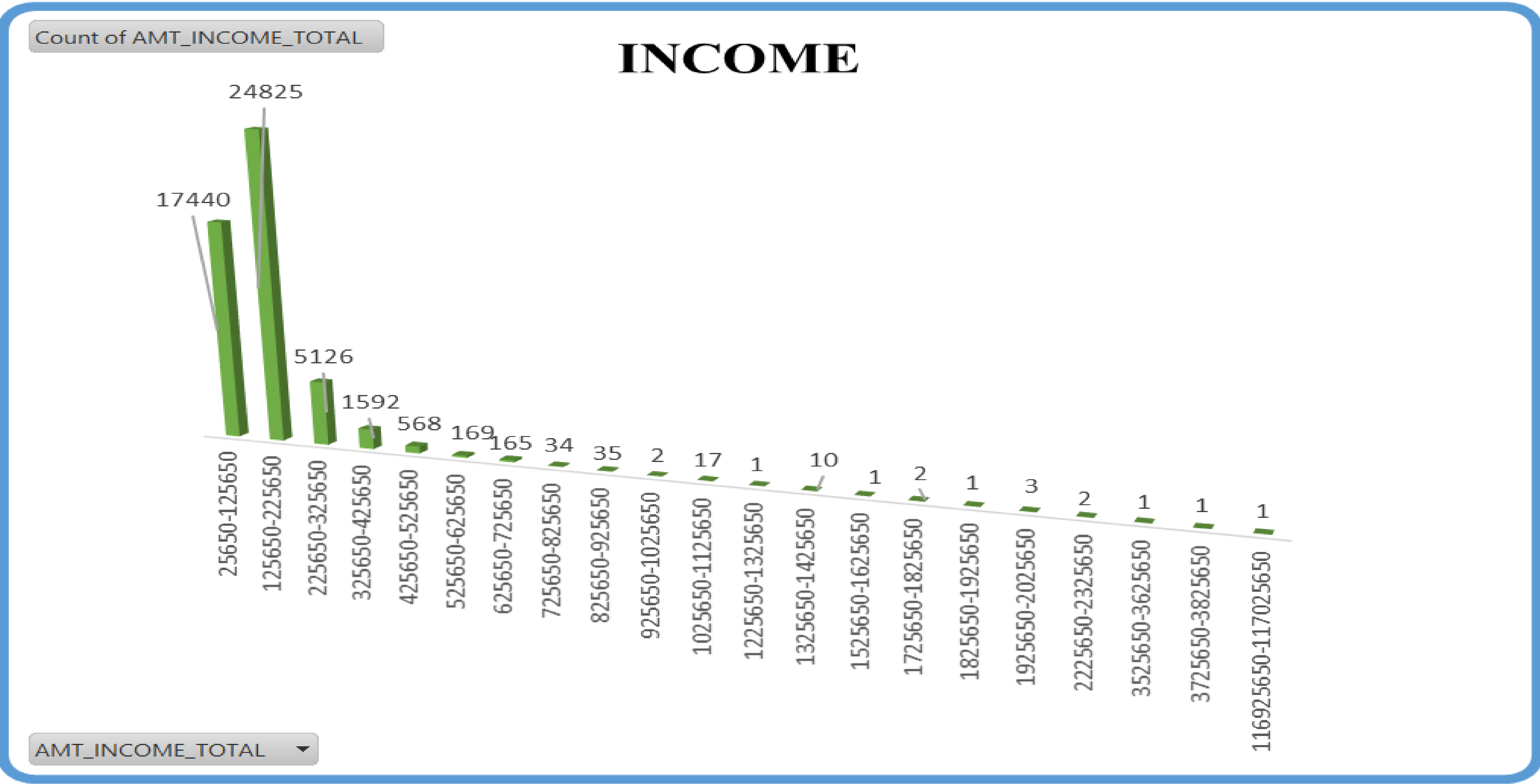
	Children	Family	Group of people	Other_A	Other_B	Spouse, partner	Unaccompanied
0	495	6050	35	127	231	1705	37327
1	47	499	1	10	28	144	3297

UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS

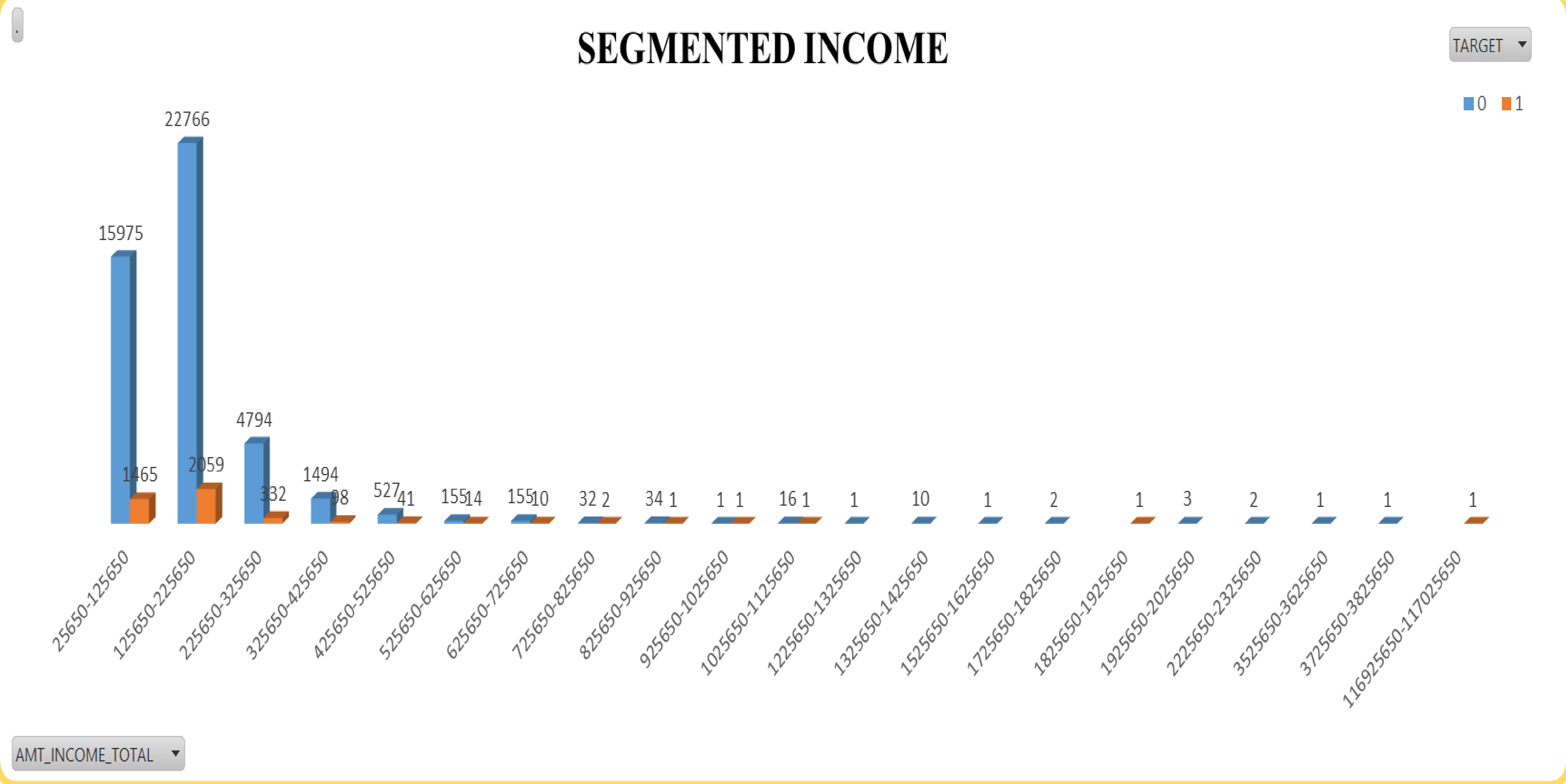




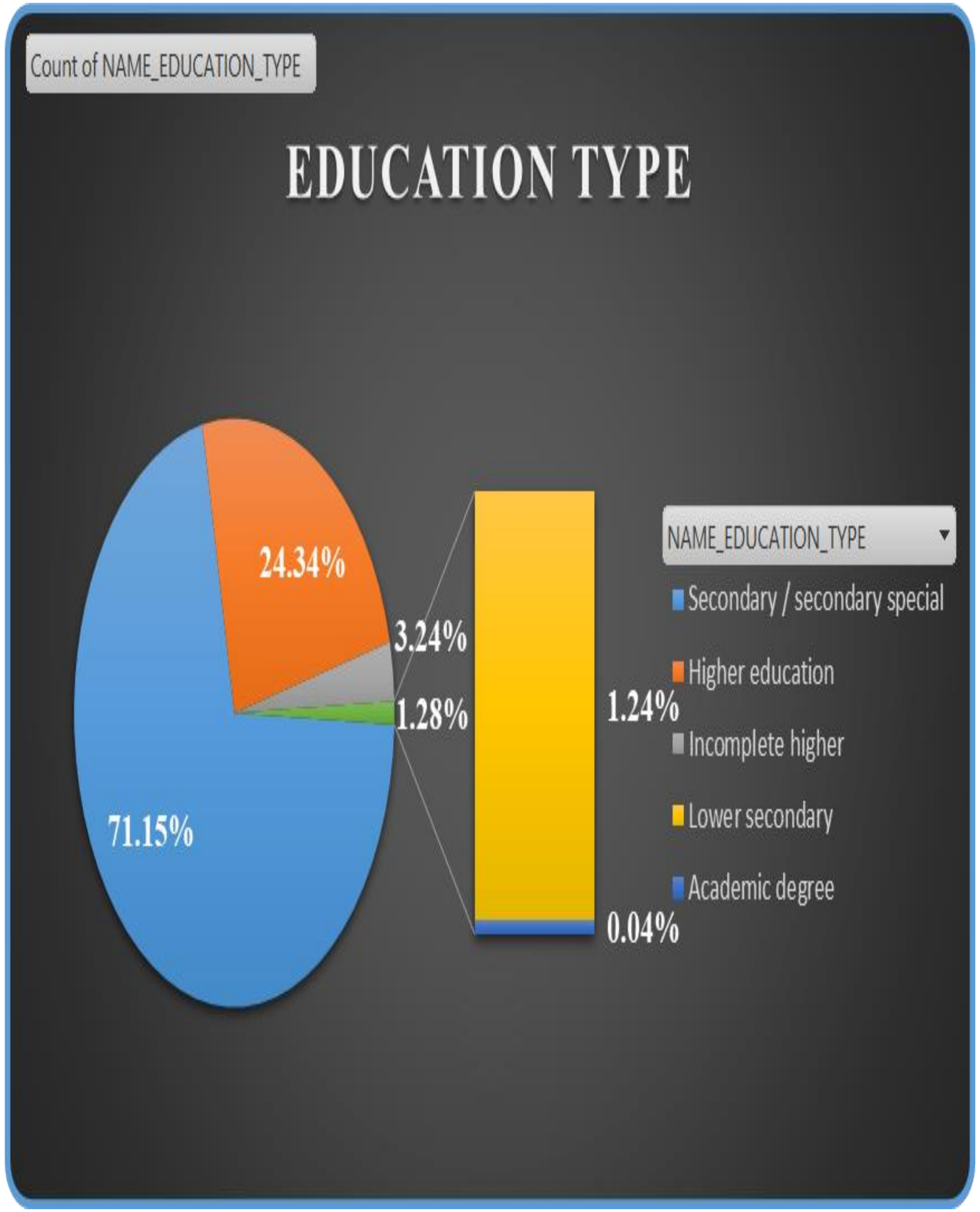
UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS

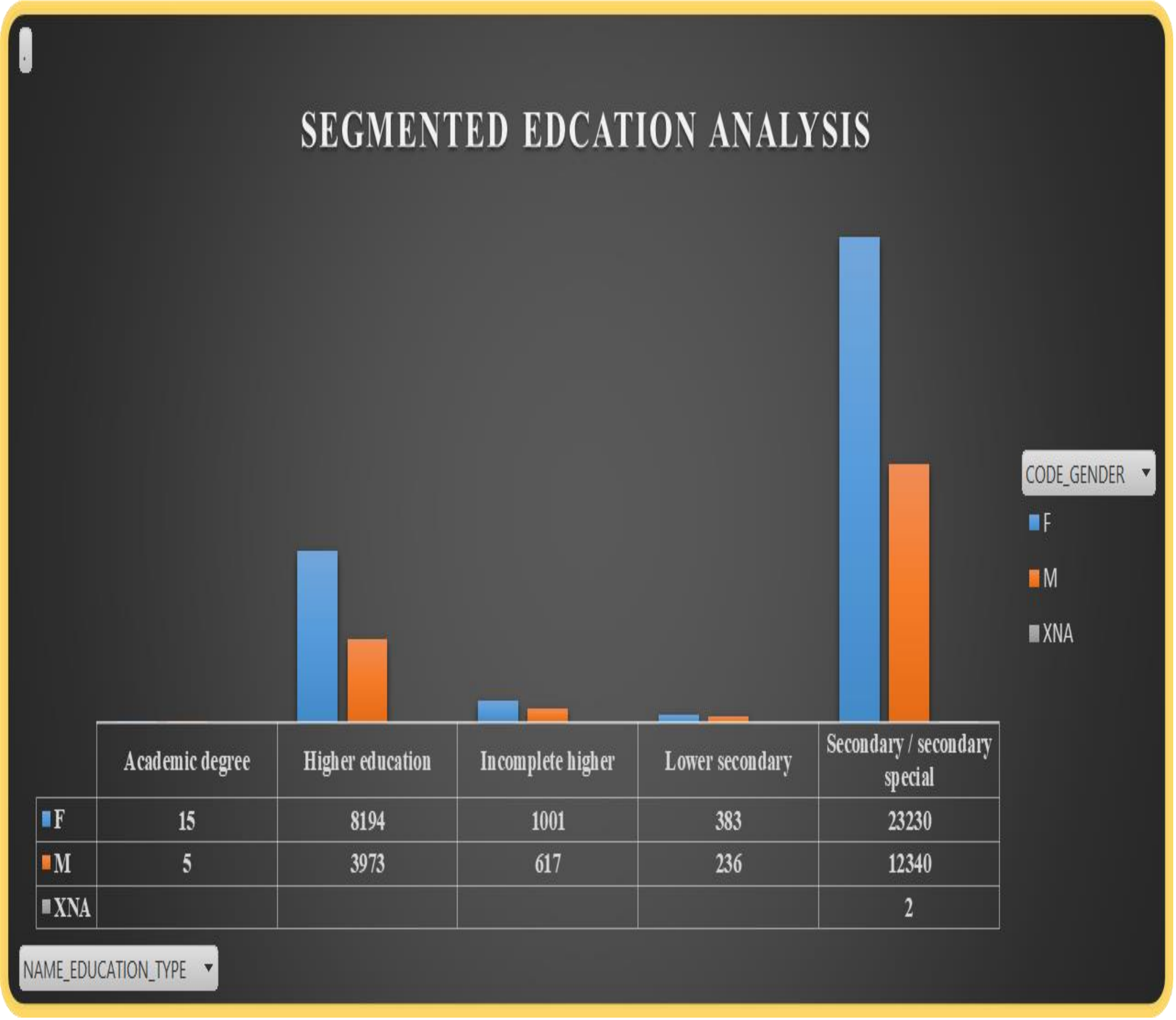


UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS



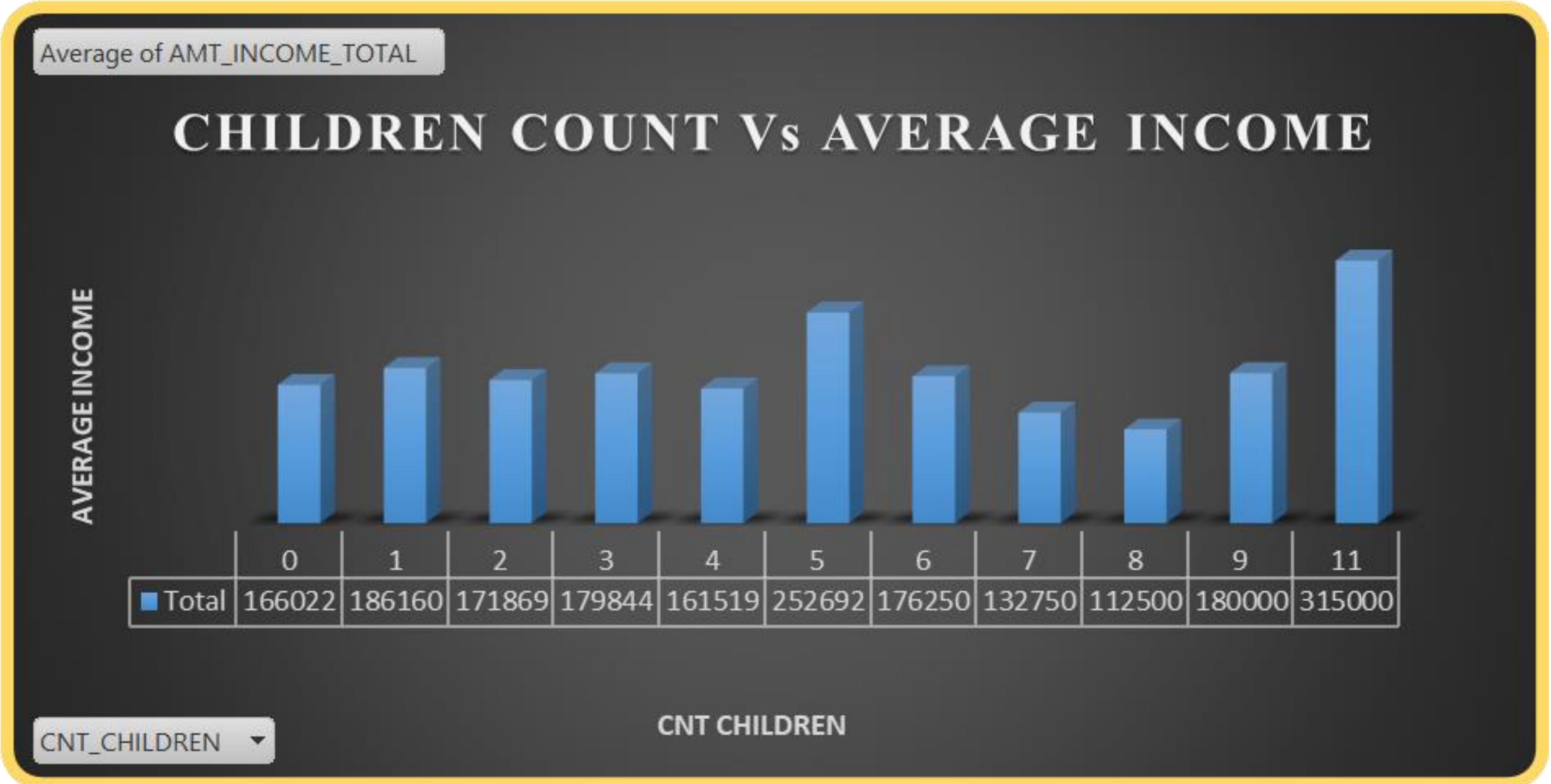
UNIVARIATE /SEGMENTED UNIVARIATE VISUALIZATIONS



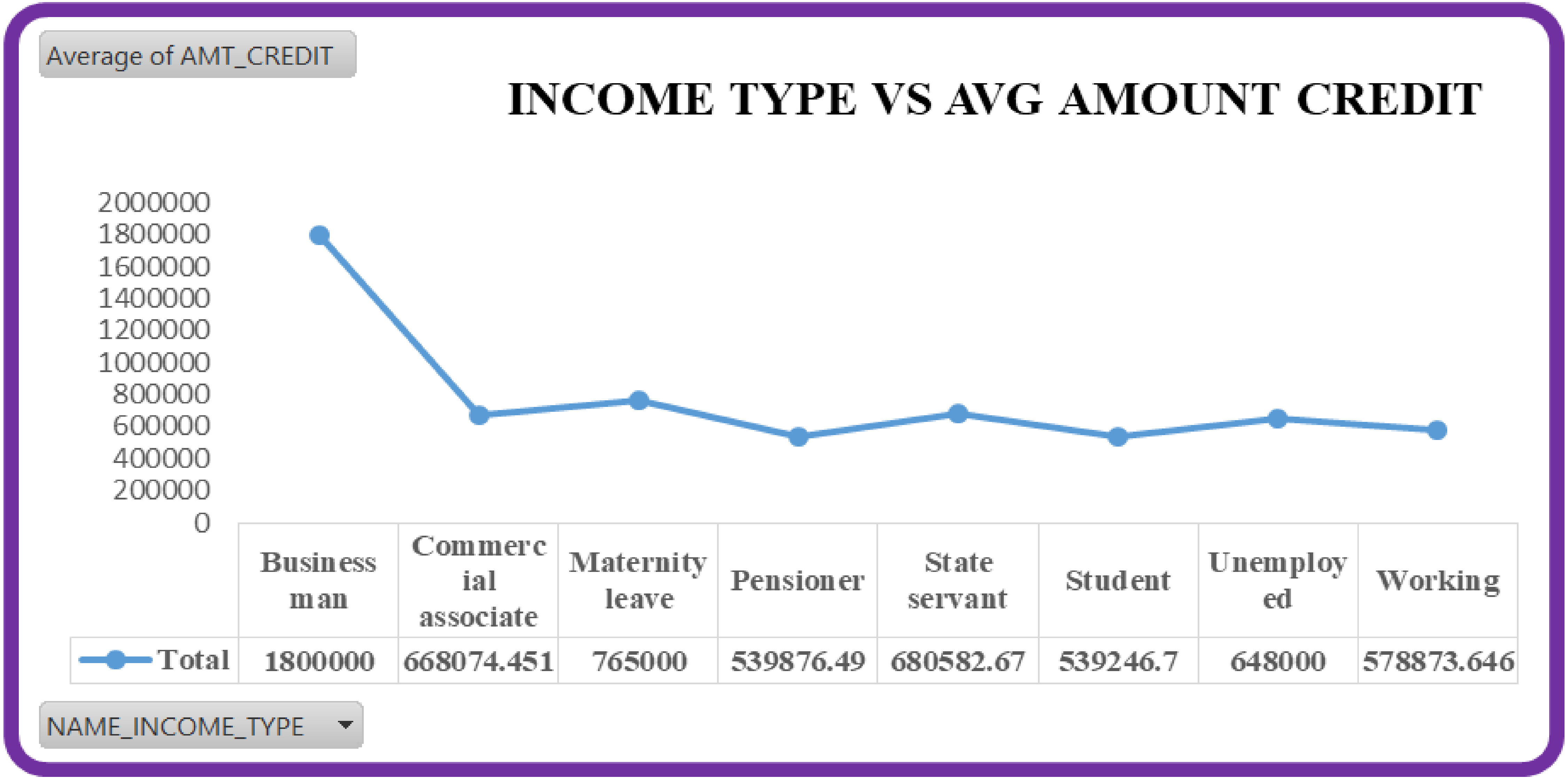


	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
F	15	8194	1001	383	23230
M	5	3973	617	236	12340
XNA					2

BIVARIATE VISUALIZATION ANALYSIS

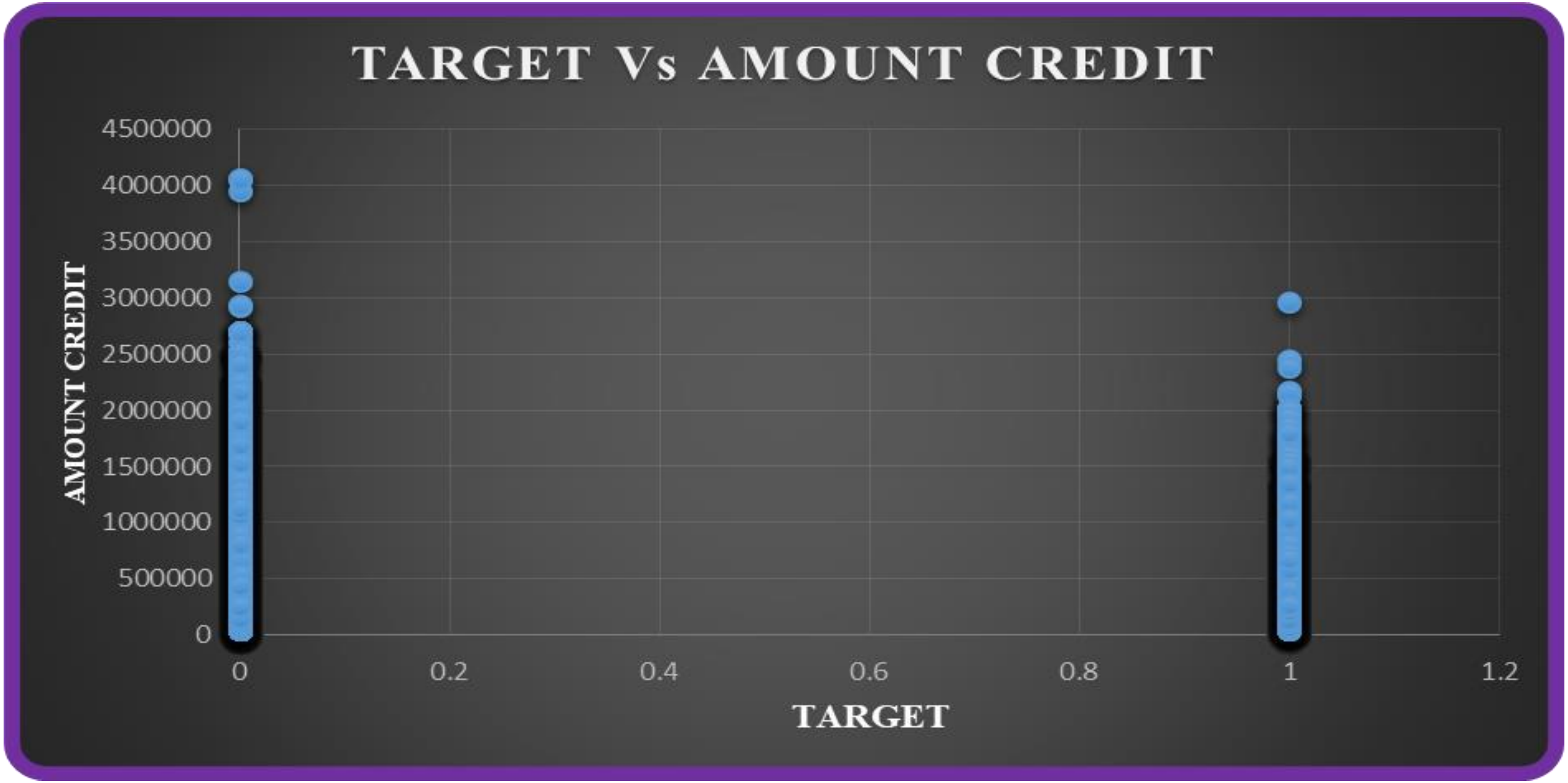


BIVARIATE VISUALIZATION ANALYSIS



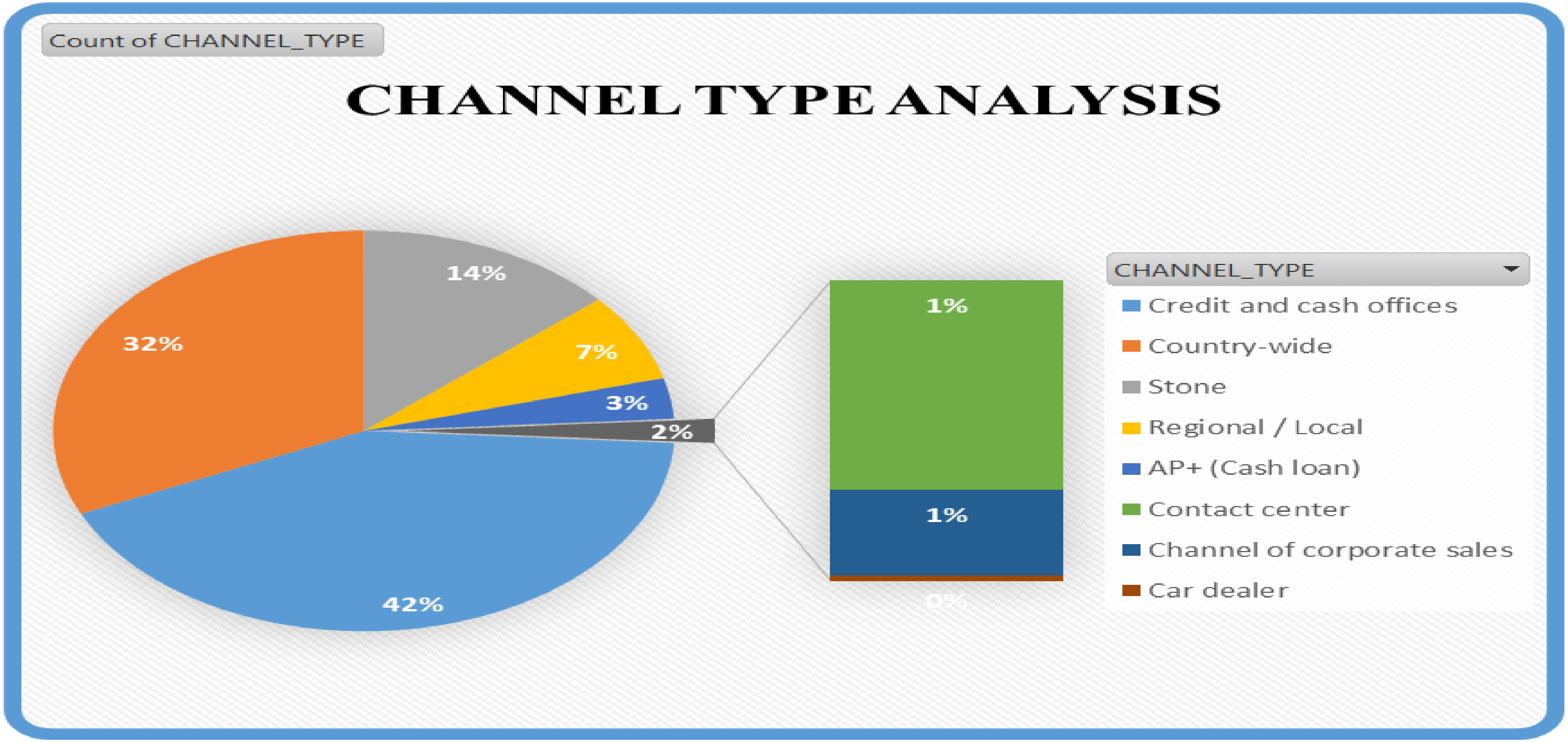
BIVARIATE VISUALIZATION ANALYSIS

BIVARIATE VISUALIZATION ANALYSIS



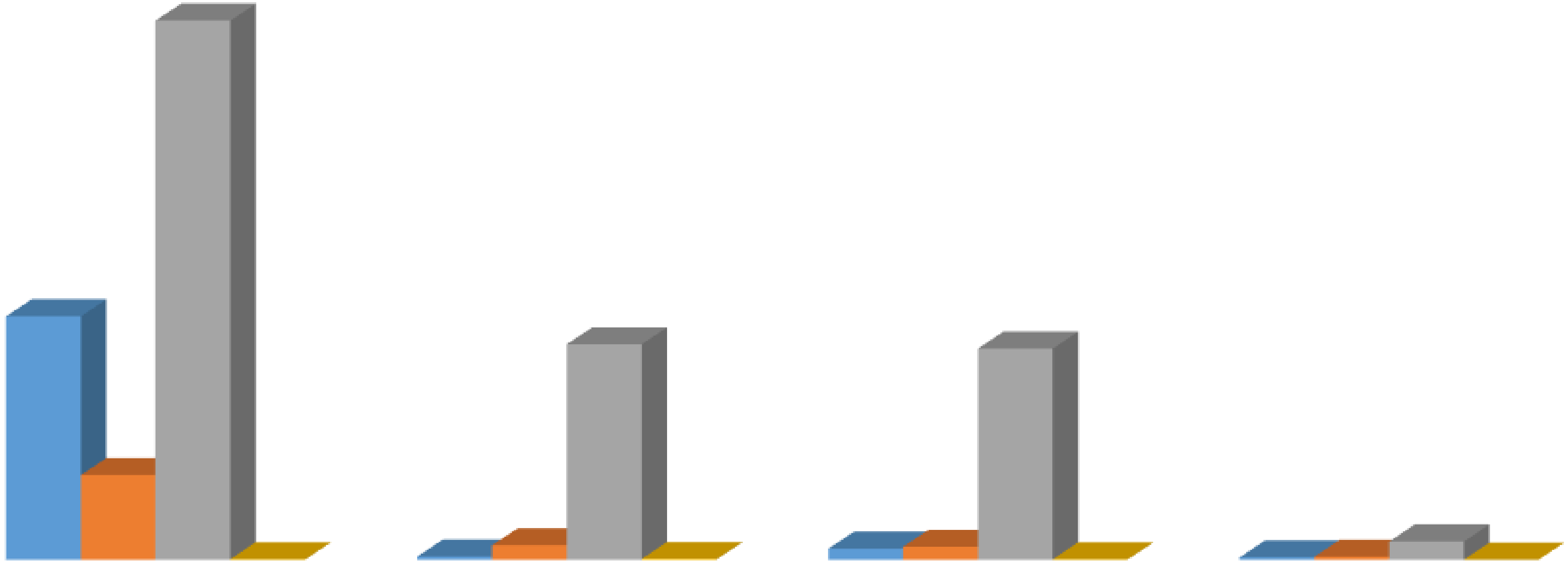
DATASET 2-PREVIOUS APPLICATION DATASET VISAULIZATIONS

UNIVARIATE VISUALIZATION



UNIVARIATE SEGMENTED VISUALIZATION ANALYSIS

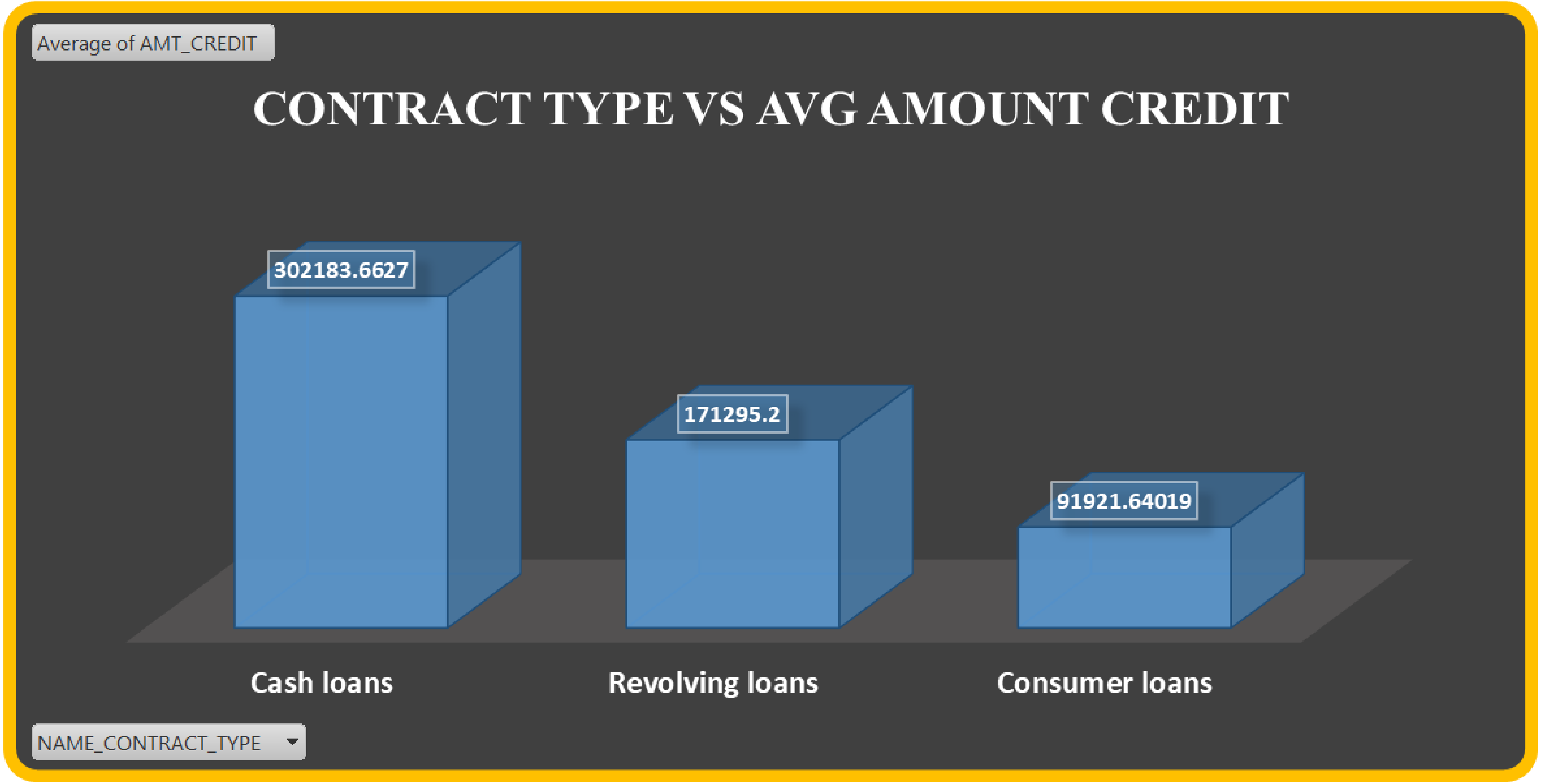
CONTRACT STATUS ANALYSIS



	Approved	Canceled	Refused	Unused offer
New	8944	114	407	82
Refreshed	3110	529	480	108
Repeater	19814	7917	7761	668
XNA	17	27	12	1

NAME_CONTRACT_STATUS

BIVARIATE VISUALIZATION ANALYSIS



E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: Segment the dataset based on different scenarios and identify the top correlations for each segmented data using Excel functions.

CORRELATION OF CLIENTS WHO MADE THE PAYMENT ON TIME CORRECTLY (TARGET-0)

TOP 10 CORRELATIONS	CORRELATION VALUE
OBS_30_CNT_SOCIAL_CIRCLE-OBS_60_CNT_SOCIAL_CIRCLE	0.998357533
AMT_GOODS_PRICE-AMT_CREDIT	0.987001704
REGION_RATING_CLIENT_W_CITY-REGION_RATING_CLIENT	0.950468197
CNT_FAM_MEMBERS-CNT_CHILDREN	0.879243419
LIVE_REGION_NOT_WORK_REGION-REG_REGION_NOT_WORK_REGION	0.861312965
DEF_30_CNT_SOCIAL_CIRCLE-DEF_60_CNT_SOCIAL_CIRCLE	0.850995019
REG_CITY_NOT_WORK_CITY-LIVE_CITY_NOT_WORK_CITY	0.825341967
AMT_GOODS_PRICE-AMT_ANNUITY	0.775843488
AMT_ANNUITY-AMT_CREDIT	0.77077712
EMPLOYEEMENT YEARS-AGE IN YEARS	0.623250115

HEATMAP FOR CLIENTS WHO MADE PAYMENT ON TIME

CORRELATIONS	IT_CHILDR	INCOME_	MT_CRED	MT_ANNUIT	_GOODS_	OPULATION	E IN YEA	YEEMENT	ISTRATION YEARS	_ID_PUBL	_FAM_MEME	_RATING_	ATING_CLIEN	R_PROC	N_NOT_LIV	_NOT_WO	_NOT_W	_NOT_L	_NOT_WO	_NOT_W	CNT_SOCIA	NT_SOCIA	NT_SOCI	NT_SOCIA	AST_PHONE_	CREDIT_BUR
CNT_CHILDREN	1																									
AMT_INCOME_TOTAL	0.036355	1																								
AMT_CREDIT	0.005693	0.37799	1																							
AMT_ANNUITY	0.026387	0.45115	0.77078	1																						
AMT_GOODS_PRICE	0.001502	0.38462	0.987	0.775843488	1																					
REGION_POPULATION	-0.02492	0.18198	0.09553	0.117284582	0.098957	1																				
AGE IN YEARS	-0.33576	-0.0736	0.05121	-0.00970746	0.048916	0.0303618	1																			
EMPLOYEEMENT YEAR	-0.24555	-0.1617	-0.0747	-0.11128659	-0.07246	-0.006781	0.62325	1																		
REGISTRATION YEARS	-0.18276	-0.0689	-0.0079	-0.03442075	-0.01105	0.05835	0.3347	0.208569	1																	
YEARS_ID_PUBLISH	0.032534	-0.0321	0.00794	-0.00966557	0.009078	0.0022903	0.27025	0.273674	0.103729285	1																
CNT_FAM_MEMBERS	0.879243	0.04162	0.06487	0.07789142	0.062883	-0.022999	-0.2843	-0.23479	-0.171083848	0.02514257	1															
REGION_RATING_CLI	0.021286	-0.205	-0.1026	-0.12992087	-0.10484	-0.53934	-0.0091	0.040939	-0.082481658	0.00751263	0.02220199	1														
REGION_RATING_CLI	0.017872	-0.2201	-0.1116	-0.14319754	-0.11313	-0.536865	-0.0072	0.043227	-0.074580818	0.01221065	0.02121299	0.950468	1													
HOUR_APPR_PROCES	-0.00525	0.08539	0.05653	0.053558403	0.065284	0.1676257	-0.0963	-0.09298	0.002303547	-0.0379359	-0.0101054	-0.28282	-0.26175855	1												
REG_REGION_NOT_LI	-0.01039	0.07896	0.02781	0.046176354	0.030313	-0.003187	-0.0605	-0.03795	-0.027693998	-0.0333239	-0.0131606	-0.04282	-0.03858905	0.05119	1											
REG_REGION_NOT_W	0.013847	0.15683	0.05609	0.082476225	0.05748	0.0631966	-0.0957	-0.10988	-0.034555228	-0.0476478	0.00837438	-0.14529	-0.13802473	0.07351	0.449659	1										
LIVE_REGION_NOT_V	0.021747	0.14748	0.05442	0.074841217	0.054607	0.087486	-0.0697	-0.09761	-0.023253175	-0.0333988	0.01713632	-0.14979	-0.14348683	0.05968	0.0804851	0.861313	1									
REG_CITY_NOT_LIVE_	0.020091	0.00994	-0.0214	-0.00527542	-0.02049	-0.046095	-0.1834	-0.09584	-0.067864309	-0.075838	0.01324471	0.035001	0.044869801	0.0197	0.3351148	0.151984	0.02164	1								
REG_CITY_NOT_WOR	0.070985	0.01504	-0.014	0.001608122	-0.01461	-0.038243	-0.2361	-0.25784	-0.091429126	-0.1019364	0.07524461	0.006077	0.026044065	0.0269	0.1426066	0.236685	0.18374	0.4415	1							
LIVE_CITY_NOT_WOR	0.067902	0.01953	0.00396	0.011180524	0.002762	-0.011263	-0.1491	-0.21998	-0.061001824	-0.063114	0.08011429	-0.01931	-0.00352094	0.0151	0.0034963	0.192075	0.23359	0.0292	0.825342	1						
OBS_30_CNT_SOCIAL	0.016179	-0.0331	0.00086	-0.01000069	0.000495	-0.019072	-0.0124	0.005572	-0.01103469	0.01127728	0.02429315	0.035606	0.033430237	-0.008	-0.01512	-0.02527	-0.0203	-0.005	-0.00608	-0.0053	1					
DEF_30_CNT_SOCIAL	-0.00283	-0.032	-0.0135	-0.01974468	-0.01522	0.0089004	-0.0008	0.016653	-0.003129137	-0.0018816	-0.0028244	0.007424	0.005694397	-0.0023	-0.008273	-0.00889	-0.0069	0.0055	0.001007	-0.0022	0.3061583	1				
OBS_60_CNT_SOCIAL	0.016334	-0.0331	0.00117	-0.00968453	0.000718	-0.018015	-0.0124	0.005442	-0.011356368	0.01158201	0.0245776	0.035333	0.033013071	-0.008	-0.015144	-0.02546	-0.0205	-0.006	-0.00606	-0.0052	0.9983575	0.308565	1			
DEF_60_CNT_SOCIAL	-0.00334	-0.0325	-0.0186	-0.02300948	-0.01974	0.0032491	-0.0023	0.01612	-0.006128729	-0.0021104	-0.0045961	0.011422	0.009417385	-0.0061	-0.009386	-0.0137	-0.012	0.0055	0.003317	-0.0002	0.2291725	0.850995	0.23128	1		
DAYS_LAST_PHONE_	-0.0048	-0.0495	-0.0712	-0.06444853	-0.07423	-0.044133	-0.0724	0.029178	-0.04777323	-0.0845802	-0.0250066	0.023518	0.02318117	-0.0146	0.0324053	0.035896	0.0256	0.0502	0.04174	0.01823	-0.014342	0.002504	-0.0151	0.002288	1	
AMT_REQ_CREDIT_BU	0.002614	0.00813	3.5E-05	0.010141172	0.000809	-0.003133	-0.0015	-0.0044	0.003954082	-0.0023185	0.00368478	0.008066	0.007027933	-0.0074	-0.002459	1.12E-05	0.00248	0.0005	0.004277	0.00401	0.0023638	-0.0044	0.00258	-0.0032	-0.00127988	1
AMT_REQ_CREDIT_BU	0.001197	0.00948	0.01349	0.009157148	0.013639	-0.00034	-0.002	0.001518	0.003447755	-0.0031122	0.00064732	0.00219	0.001338327	0.01034	-0.005756	0.000758	0.00291	8E-05	-0.00023	-0.0012	0.0009729	0.003686	0.00087	0.002777	-0.0004527	0.2307631
AMT_REQ_CREDIT_BU	0.004322	0.0095	0.00537	0.018910543	0.005807	0.0026421	0.00228	-0.00624	-0.0004583	0.00473105	0.00611381	-0.00081	-0.00444873	-0.0067	-0.001768	0.003335	0.00545	-0.001	0.002181	0.00243	-0.004288	-0.00504	-0.0049	-0.00573	-0.00599161	0.01212501
AMT_REQ_CREDIT_BU	-0.01162	0.07488	0.06397	0.037986896	0.065703	0.070733	0.00232	-0.03224	0.010873947	0.01403248	-0.0045104	-0.0642	-0.06187879	0.02885	-0.008608	0.004263	0.00996	-0.014	-0.01239	-0.0046	0.0081697	0.007682	0.00813	0.003967	-0.04733187	0.0095461
AMT_REQ_CREDIT_BU	-0.00472	0.0158	0.0268	0.010067047	0.027519	-0.009716	0.02162	0.014687	-0.003230511	0.02431873	-0.0042414	0.011873	0.010766519	-0.0005	-0.000265	-0.00873	-0.0123	-2E-05	-0.00391	-0.0052	0.0088515	0.005354	0.00868	0.00831	-0.01288266	0.00351898
AMT_REQ_CREDIT_BU	-0.03575	0.03135	-0.0316	-0.0041718	-0.03443	0.0046453	0.07022	0.044358	0.022615593	0.04446743	-0.0229288	0.007002	0.004908534	-0.025	-0.019529	-0.0275	-0.0225	-0.007	-0.01195	-0.0129	0.0341607	0.014498	0.03457	0.015198	-0.11761008	0.00409328

For the full proper Heatmap please view the Excel File

CORRELATION OF CLIENTS WHO MADE LATE PAYMENT (TARGET-1)

TOP 10 CORRELATIONS	CORRELATION VALU
OBS_30_CNT_SOCIAL_CIRCLE-OBS_60_CNT_SOCIAL_CIRCLE	0.998065853
AMT_GOODS_PRICE-AMT_CREDIT	0.982267963
REGION_RATING_CLIENT_W_CITY-REGION_RATING_CLIENT	0.950768899
CNT_FAM_MEMBERS-CNT_CHILDREN	0.892521875
DEF_30_CNT_SOCIAL_CIRCLE-DEF_60_CNT_SOCIAL_CIRCLE	0.89051161
LIVE_REGION_NOT_WORK_REGION-REG_REGION_NOT_WORK_REGION	0.806743886
REG_CITY_NOT_WORK_CITY-LIVE_CITY_NOT_WORK_CITY	0.783754676
AMT_ANNUITY-AMT_CREDIT	0.749665201
AMT_GOODS_PRICE-AMT_ANNUITY	0.74950403
EMPLOYEEMENT YEARS-AGE IN YEARS	0.587858433

HEATMAP FOR CLIENTS WHO MADE LATE PAYMENT

CORRELATIONS	T_CHILDR	_INCOME_T	MT_CRED	IT_ANNUI	_GOODS_	PULATIC	E IN YEA	YEEMENT	TRATION	'S_ID_PUE	FAM_MEN	_RATING	TING_CLI	PR_PROCE	N_NOT_LIN	_NOT_WC	_NOT_WT	Y_NOT_LI	_NOT_W	NOT_W	CNT_SOCIA	CNT_SOCIA	CNT_
CNT_CHILDREN	1																						
AMT_INCOME_	0.0101	1																					
AMT_CREDIT	0.0076	0.015271	1																				
AMT_ANNUITY	0.0292	0.018005	0.7497	1																			
AMT_GOODS_	-0.0011	0.01327	0.9823	0.7495	1																		
REGION_POPU	-0.0204	-0.00618	0.0678	0.07312	0.07664	1																	
AGE IN YEARS	-0.2496	-0.008444	0.1424	0.00886	0.14086	0.017	1																
EMPLOYEEMEI	-0.1898	-0.011735	0.0188	-0.0781	0.02318	0.008	0.588	1															
REGISTRATION	-0.1518	0.010368	0.0425	-0.0217	0.04298	0.046	0.288	0.1929	1														
YEARS_ID_PUE	0.0435	0.009176	0.0445	0.0215	0.05024	0.006	0.248	0.23133	0.09149	1													
CNT_FAM_MEI	0.8925	0.013122	0.0612	0.07584	0.05514	-0.017	-0.2	-0.1834	-0.1519	0.0443	1												
REGION_RATII	0.0555	-0.012847	-0.045	-0.0616	-0.0513	-0.43	-0.04	-0.0092	-0.1164	-0.0282	0.05728	1											
REGION_RATII	0.0548	-0.012666	-0.053	-0.0794	-0.0567	-0.432	-0.04	-0.0041	-0.1086	-0.0169	0.05799	0.9508	1										
HOUR_APPR_F	-0.0069	0.014482	0.0454	0.04489	0.05746	0.156	-0.06	-0.0516	0.05826	-0.0035	-0.0239	-0.279	-0.2531	1									
REG_REGION_	-0.0157	0.000595	0.0065	0.03176	0.00708	-0.003	-0.04	-0.0364	-0.0162	-0.0252	-0.0039	-0.031	-0.0295	0.04942	1								
REG_REGION_	-0.0057	0.001666	0.0235	0.06569	0.02502	0.019	-0.08	-0.0869	-0.0166	-0.042	-0.0086	-0.103	-0.0992	0.07615	0.5255	1							
LIVE_REGION_	-0.0004	0.002228	0.0346	0.07424	0.03542	0.06	-0.05	-0.0737	-0.0135	-0.0296	-0.0106	-0.123	-0.1188	0.06606	0.10053	0.806744	1						
REG_CITY_NOT	0.0017	-0.005992	-0.052	-0.0177	-0.0527	-0.035	-0.15	-0.0909	-0.0558	-0.0646	0.00908	0.0472	0.05479	0.00552	0.33817	0.18375	0.0261	1					
REG_CITY_NOT	0.0489	-0.010357	-0.039	0.00218	-0.044	-0.043	-0.23	-0.2499	-0.1013	-0.0842	0.04938	0.017	0.04131	0.0032	0.14759	0.228676	0.1578	0.4673	1				
LIVE_CITY_NO	0.0582	-0.008036	-0.007	0.01356	-0.0131	-0.025	-0.14	-0.2025	-0.0704	-0.0392	0.0563	-0.006	0.0134	-0.0118	-0.0037	0.169078	0.2179	-0.01502	0.7838	1			
OBS_30_CNT_	0.0179	-0.011281	0.0335	0.01382	0.03272	-0.009	0.011	0.00471	0.00488	0.0274	0.03999	0.0264	0.02142	-0.0197	-0.032	-0.03211	-0.0208	-0.04989	-0.0421	-0.024	1		
DEF_30_CNT_	-0.0136	-0.007979	-0.025	-0.0345	-0.0191	0.028	0.021	0.02977	-0.0019	0.027	-0.0065	0.0162	0.01439	0.01767	0.00849	0.001517	-0.0061	0.00342	-0.0156	-0.028	0.365074	1	
OBS_60_CNT_	0.0151	-0.011211	0.0344	0.0141	0.03388	-0.007	0.013	0.00541	0.00504	0.0262	0.03754	0.0255	0.02073	-0.0195	-0.032	-0.03155	-0.02	-0.05043	-0.0416	-0.023	0.998066	0.36806	
DEF_60_CNT_	-0.0185	-0.006727	-0.029	-0.0405	-0.0206	0.027	0.026	0.02379	0.00582	0.028	-0.0089	-8E-04	-0.0002	0.01752	0.00582	0.004932	9E-05	0.00258	-0.0137	-0.025	0.297951	0.890512	0.3
DAYS_LAST_PH	0.0113	0.012457	-0.125	-0.1005	-0.1288	-0.067	-0.12	-0.0194	-0.0788	-0.1378	-0.0057	0.0262	0.02231	-0.0352	0.01769	0.020813	0.0112	0.06899	0.074	0.042	-0.02192	0.004158	
AMT_REQ_CRE	-0.0003	-0.001104	0.0178	0.0374	0.01526	0.009	-0.02	-0.0036	-0.0056	-0.0146	0.00486	-0.009	-0.0113	-0.0331	-0.011	0.022701	0.0319	-0.00109	0.0183	0.014	-0.01409	0.002728	-0.
AMT_REQ_CRE	-0.0306	-0.001447	-0.009	-0.0187	-0.0063	-0.004	0.023	0.04939	0.00195	0.0078	-0.0331	0.0206	0.01995	0.00141	0.0042	0.011146	0.007	-0.01913	-0.0053	8E-04	-0.01703	0.012236	-0.

For the full proper Heatmap please view the Excel File

RESULT

- Throughout this project, the role of Lead Data Analyst has been instrumental in driving data-driven decision-making within the organization , resembling the Bank loan analysis. Through meticulous analysis of diverse aspects of dataset and handling null values and outliers this project has yielded actionable insights that helps the bank loan process.

- In summary, our EDA has provided valuable insights into the challenges posed by customers with insufficient credit history. By adopting a more holistic approach to assessing creditworthiness, leveraging advanced analytics, and continuously improving our lending practices, we aim to strike a balance between mitigating default risks and providing financial support to deserving applicants.

THANK YOU

BHAVYA SRI DUGGINA

Excel file link

Please download the Excel file and view in MS Excel for better visualizations also the file is large to preview

ipynb Notebook link

Please download the ipynb notebook and view in suitable source to view it in correct Format