BANK LOAN CASE STUDY

By Shivam Choudhary

"AGENDA

- PROJECT DESCRIPTION
- DATASET DESCRIPTION
- APPROACH
- TECH STACK USED
- INSIGHTS
- RESULTS

PROJECT DESCRIPTION

When a customer applies for a loan, company faces two risks:

- *If the applicant can repay the loan but is not approved, the company loses business.*
- *If the applicant cannot repay the loan and is approved, the company faces a financial loss.*

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.

DATASET DESCRIPTION

01 - APPLICATION DATA

Contains all the information and details of application submitted by client;

02 - PREVIOUS APPLICATION

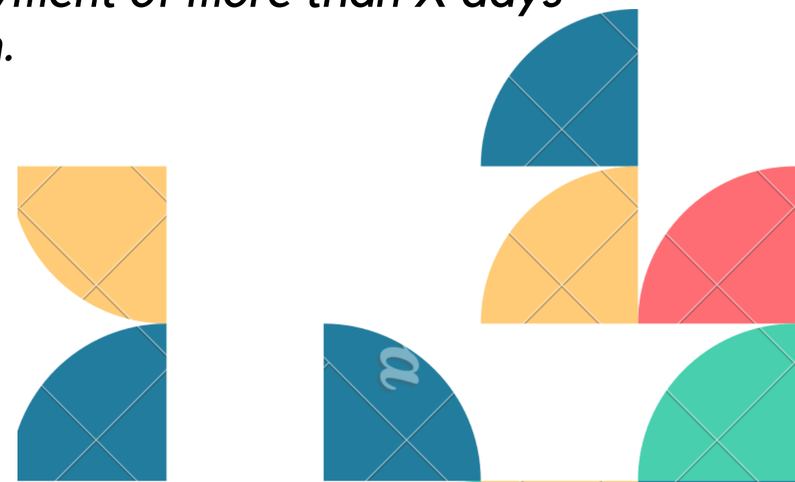
Contains information of the previous application submitted by client

03- COLUMN DESCRIPTION

Contains the breif description about the columns used in both dataset

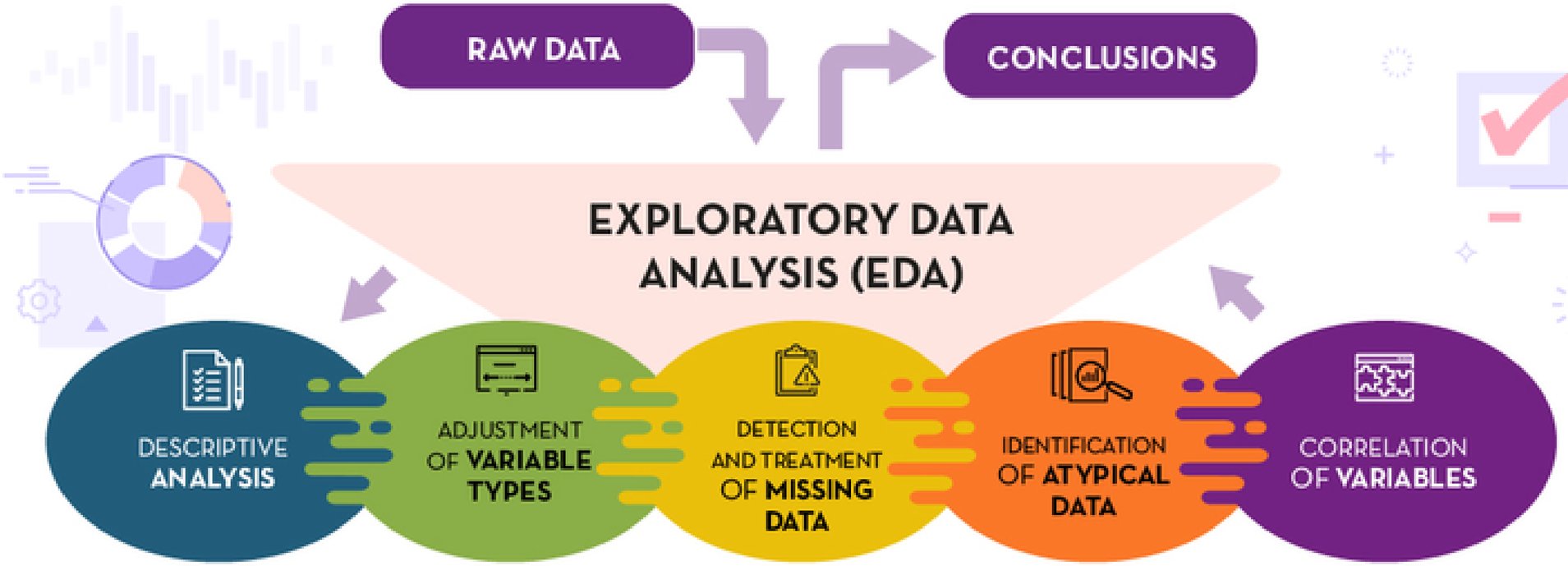
The dataset contains information about loan applications. It includes two types of scenarios:

- *Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.*
- *All other cases: These are cases where the payment was made on time.*



APPROACH

Exploratory Data Analysis (EDA) is an approach that is used to analyze the data and discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations.



TECH STACK USED

Microsoft Excel is one of the most popular applications for data analysis.It is a powerful tool for analyzing and visualizing data. It can help you understand how past events affect your future.Equipped with built-in pivot tables, they are without a doubt the most sought-after analytic tool available. It is an all-in-one data management software that allows you to easily import, explore, clean, analyze, and visualize your data

APPLICATION DATASET CLEANING

Click here for GDrive link for Dataset

APPLICATION DATASET-NULL VALUES

Firstly the percentage of null values needs to be analyzed and those columns that have more than 50% of the null data have to be dropped

Columns In RED Having Missing values greater than 50 % Needs to Be Removed

S.NO	NAME	% OF NULL VALUES	ROUND
1	OWN_CAR_AGE	65.9	66
2	EXT_SOURCE_1	67.6611	68
3	APARTMENTS_AVG	49.92	50
4	BASEMENTAREA_AVG	66.3179	66
5	YEARS_BEGINEXPLUATATION_AVG	57.6395	58
6	YEARS_BUILD_AVG	68.03	68
7	COMMONAREA_AVG	53.3823	53
8	ELEVATORS_AVG	66.807	67
9	ENTRANCES_AVG	54.4	54
10	FLOORSMAX_AVG	50.5783	51
11	FLOORSMIN_AVG	60.0381	60
12	LANDAREA_AVG	50.5076	51
13	LIVINGAPARTMENTS_AVG	66.8149	67
14	LIVINGAREA_AVG	66.674	67
15	NONLIVINGAPARTMENTS_AVG	55.9129	56
16	NONLIVINGAREA_AVG	51.1161	51
17	APARTMENTS_MODE	49.921	50
18	BASEMENTAREA_MODE	66.3179	66
19	YEARS_BEGINEXPLUATATION_MODE	57.6395	58
20	YEARS_BUILD_MODE	68.0302	68
21	COMMONAREA_MODE	53.3823	53
22	ELEVATORS_MODE	66.8071	67
23	ENTRANCES_MODE	54.401	54
24	FLOORSMAX_MODE	49.75	50
25	FLOORSMIN_MODE	67.788	68
26	LANDAREA_MODE	59.442	59
27	LIVINGAPARTMENTS_MODE	68.452	68
28	LIVINGAREA_MODE	50.274	50
29	NONLIVINGAPARTMENTS_MODE	69.428	69
30	NONLIVINGAREA_MODE	55.144	55
31	APARTMENTS_MEDI	50.77	51
32	BASEMENTAREA_MEDI	58.398	58
33	YEARS_BUILD_MEDI	66.478	66
34	COMMONAREA_MEDI	69.92	70
35	ELEVATORS_MEDI	53.302	53
36	ENTRANCES_MEDI	50.39	50
37	FLOORSMAX_MEDI	49.75	50
38	FLOORSMIN_MEDI	67.788	68
39	LANDAREA_MEDI	59.442	59
40	LIVINGAPARTMENTS_MEDI	68.452	68
41	LIVINGAREA_MEDI	50.274	50
42	NONLIVINGAPARTMENTS_MEDI	69.428	69
43	NONLIVINGAREA_MEDI	55.144	55
44	FONDKAPREMONT_MODE	68.382	68
45	HOUSETYPE_MODE	50.15	50
46	WALLSMATERIAL_MODE	50.918	51

APPLICATION DATASET-IRRELEVANT COLUMN

Secondly those columns which are irrelevant and is not required during analysis process needs to be dropped

COLUMNS IN YELLOW NEED TO BE DROPPED DOWN AS THEY ARE IRRELEVANT FOR ANALYSIS

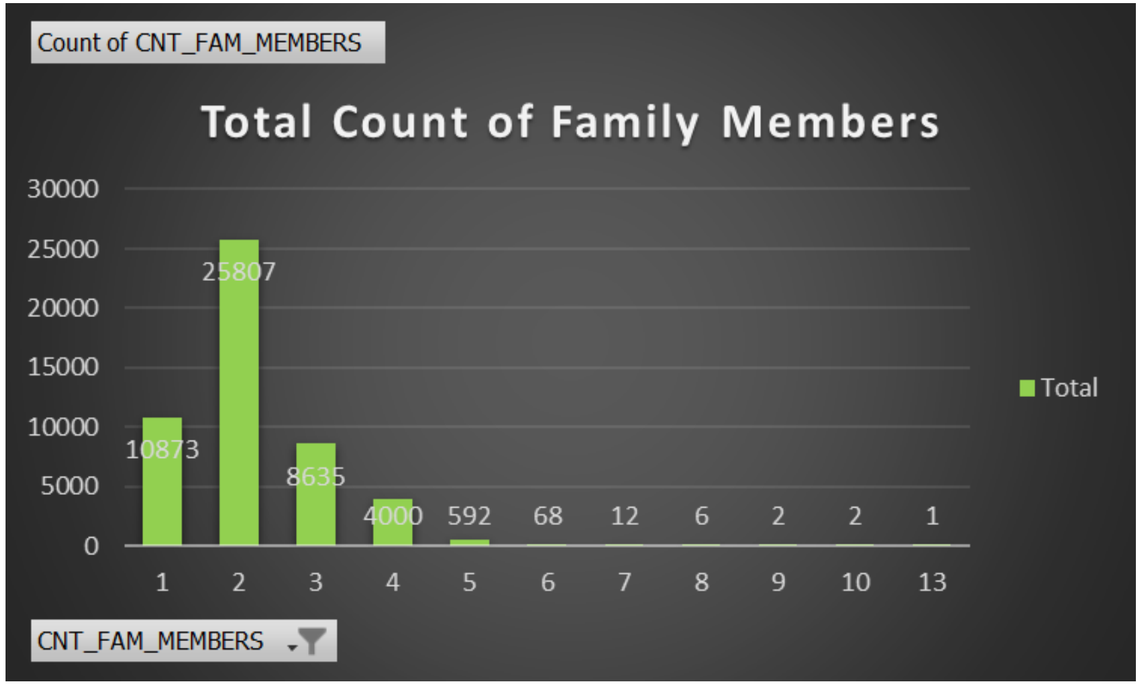
S. NO	COLUMN NAME
1	FLAG_MOBIL
2	FLAG_EMP_PHONE
3	FLAG_WORK_PHONE
4	FLAG_CONT_MOBILE
5	FLAG_PHONE
6	FLAG_EMAIL
7	REGION_RATING_CLIENT
8	REGION_RATING_CLIENT_W_CITY
9	EXT_SOURCE_3
10	YEARS_BEGINEXPLUATATION_MEDI
11	TOTALAREA_MODE
12	EMERGENCYSTATE_MODE
13	DAYS_LAST_PHONE_CHANGE
14	FLAG_DOCUMENT_2
15	FLAG_DOCUMENT_3
16	FLAG_DOCUMENT_4
17	FLAG_DOCUMENT_5
18	FLAG_DOCUMENT_6
19	FLAG_DOCUMENT_7
20	FLAG_DOCUMENT_8
21	FLAG_DOCUMENT_9
22	FLAG_DOCUMENT_10
23	FLAG_DOCUMENT_11
24	FLAG_DOCUMENT_12
25	FLAG_DOCUMENT_13
26	FLAG_DOCUMENT_14
27	FLAG_DOCUMENT_15
28	FLAG_DOCUMENT_16
29	FLAG_DOCUMENT_17
30	FLAG_DOCUMENT_18
31	FLAG_DOCUMENT_19
32	FLAG_DOCUMENT_20
33	FLAG_DOCUMENT_21

APPLICATION DATASET-MISSING VALUES

Columns with less than 50% of the null data have to be replaced with mean or median or the highest occurring categorical variable

REPLACING BLANKS IN 'CNT_FAM_MEMBERS' WITH VALUE '2' SINCE IT IS HIGHEST OCCURING VALUE

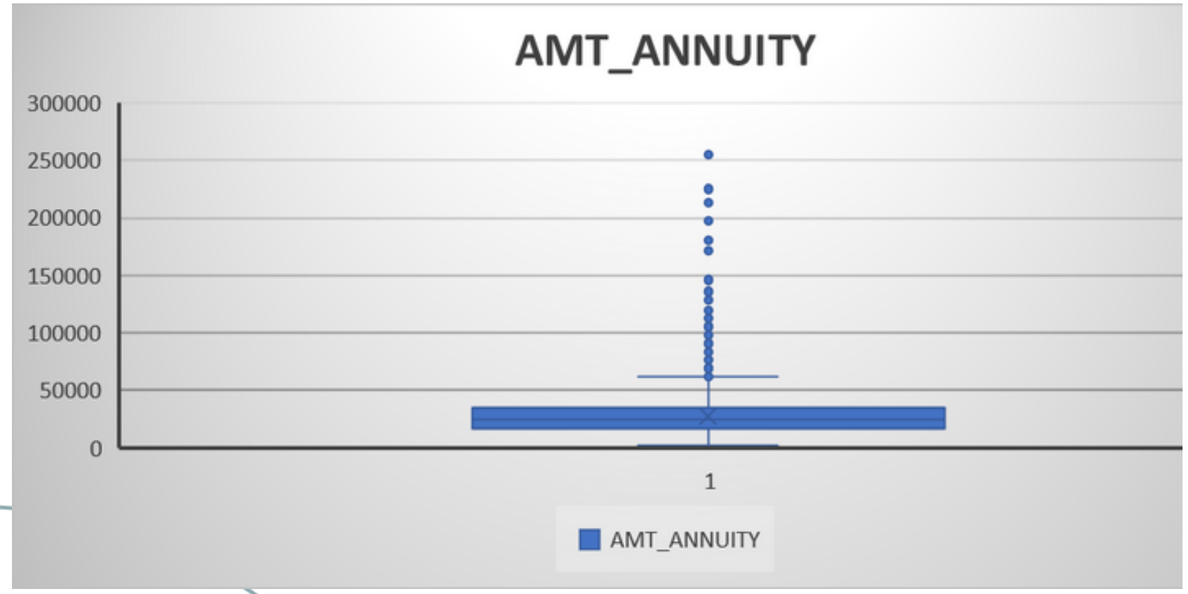
NO. OF FAMILY_MEMBERS	Count of CNT_FAM_MEMBERS
1	10873
2	25807
3	8635
4	4000
5	592
6	68
7	12
8	6
9	2
10	2
13	1
Grand Total	**49998**



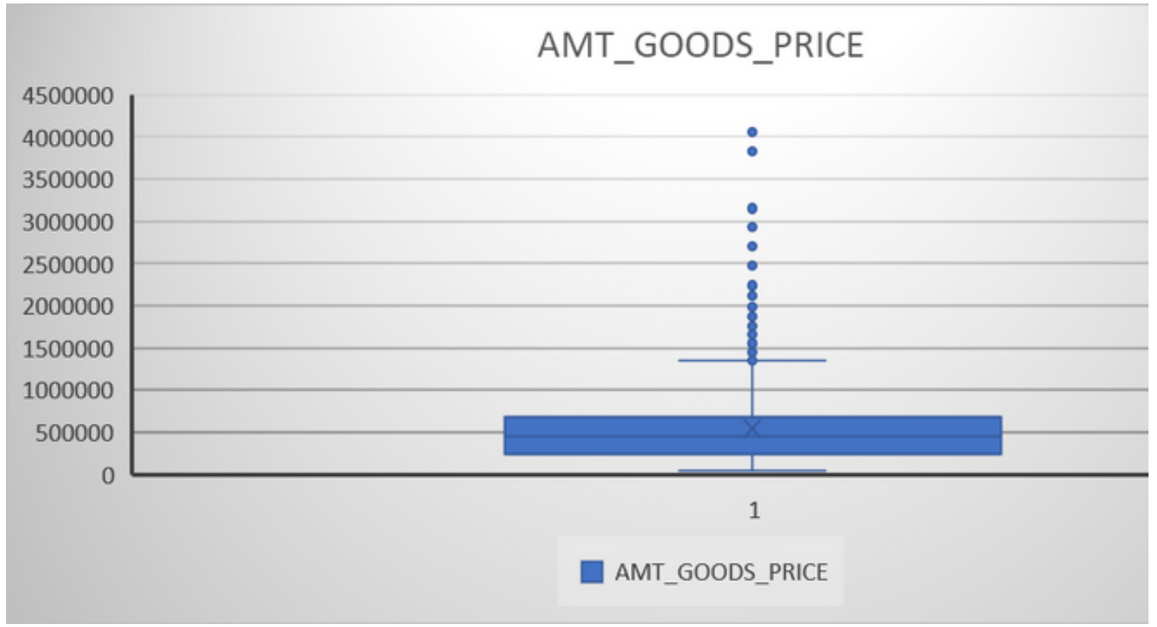
Click here for Task 1 dataset link

APPLICATION DATASET-MISSING VALUES

REPLACING BLANKS IN AMT_ANNUITY WITH *MEDIAN* SINCE THE COLUMN CONTAINS *OUTLIERS* HENCE WE CANNOT USE *MEAN*	
MEDIAN	24939



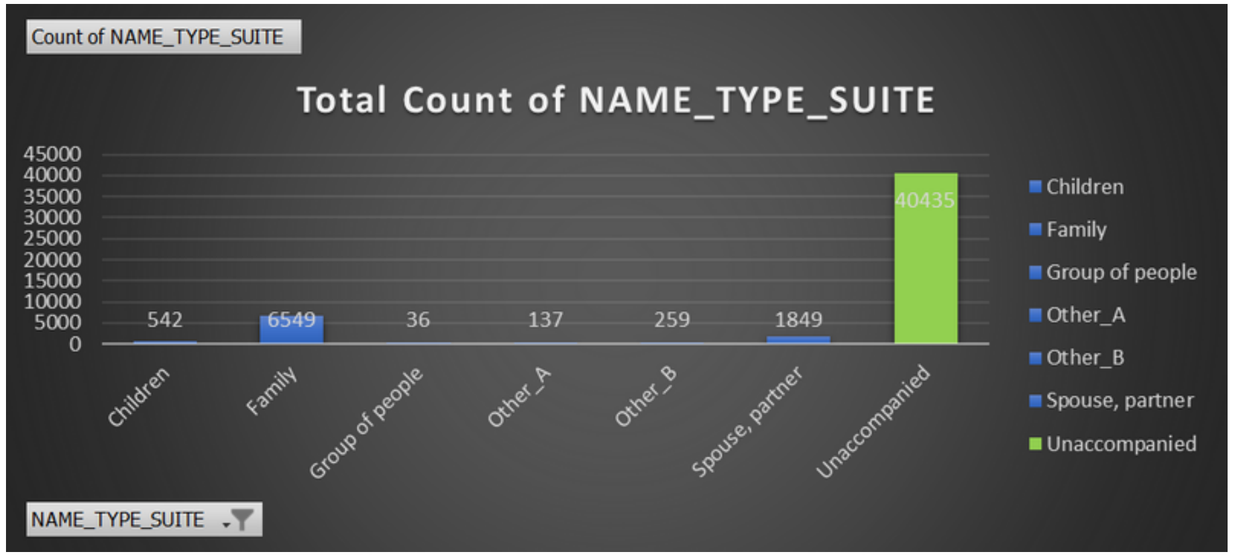
REPLACING BLANKS IN AMT_GOODS_PRICE WITH *MEDIAN* SINCE THE COLUMN CONTAINS *OUTLIERS* HENCE WE CANNOT USE *MEAN*	
MEDIAN	450000



APPLICATION DATASET-MISSING VALUES

REPLACING BLANKS IN COLUMN 'NAME_TYPE_SUITE' WITH HIGHEST OCCURING VALUE 'UNACCOMPANIED'

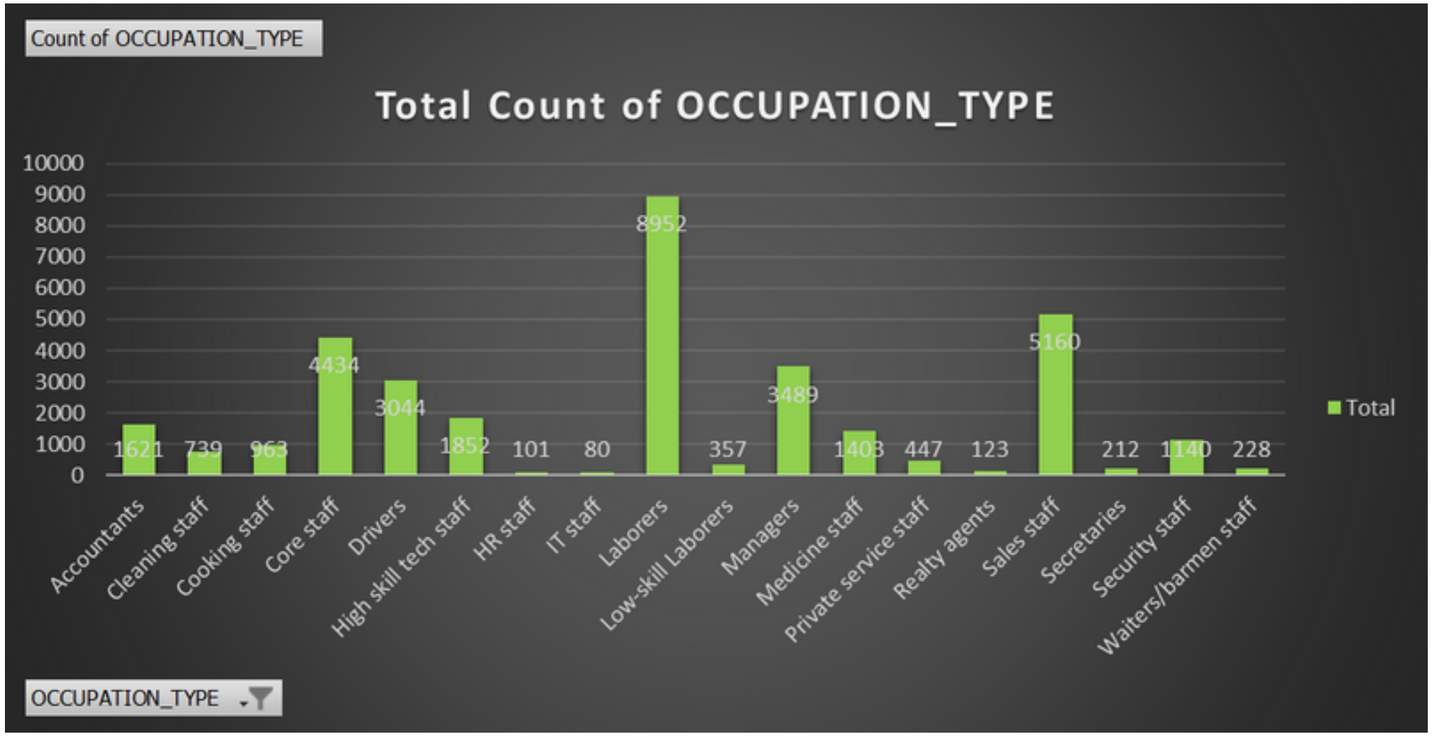
NAME_TYPE_SUITE	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
Grand Total	**49807**



APPLICATION DATASET-MISSING VALUES

REPLACING BLANKS IN COLUMN 'OCCUPATION_TYPE' WITH HIGHEST OCCURING VALUE 'LABORERS'

Occupation	Count of OCCUPATION_TYPE
Accountants	1621
Cleaning staff	739
Cooking staff	963
Core staff	4434
Drivers	3044
High skill tech staff	1852
HR staff	101
IT staff	80
Laborers	8952
Low-skill Laborers	357
Managers	3489
Medicine staff	1403
Private service staff	447
Realty agents	123
Sales staff	5160
Secretaries	212
Security staff	1140
Waiters/barmen staff	228
Grand Total	**34345**

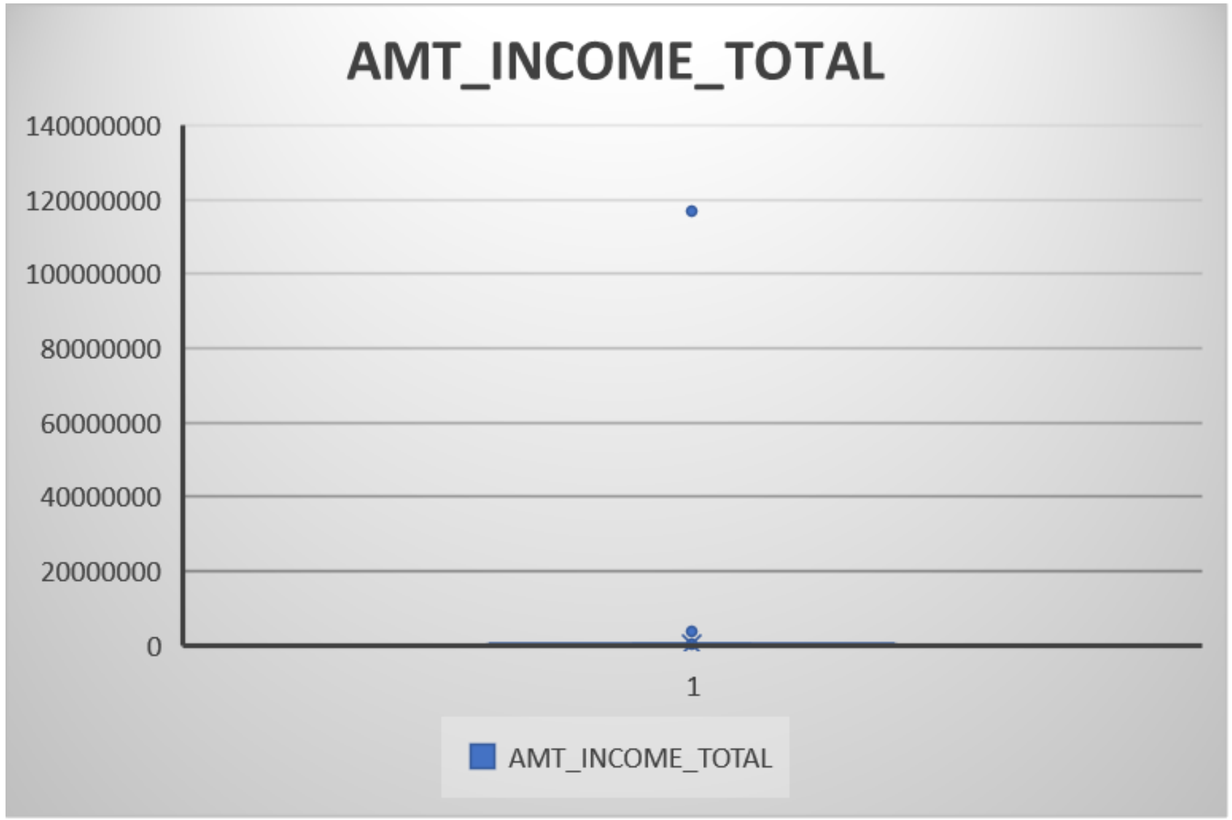


APPLICATION DATASET-OUTLIERS

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

AMT_INCOME_TOTAL	
QUARTILE 1	112500
QUARTILE 3	202500
IQR	90000
LOWER BOUND	-22500
UPPER BOUND	337500

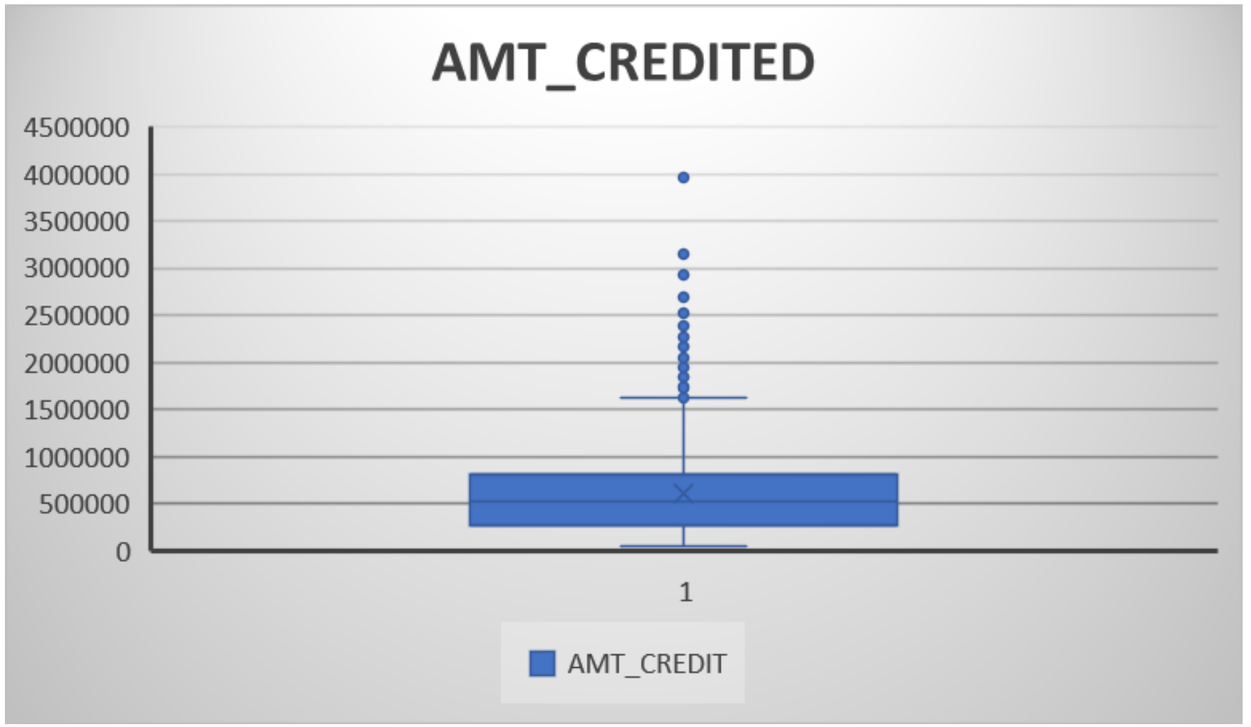
HERE WE CAN SEE THAT THERE LIES AN OUTLIER WHOSE VALUE IN MUCH GREATER THAN UPPER BOUND CALCULATED BUT SINCE INCOME MAY VARY FROM PERSON TO PERSON HENCE WE WILL NOT REMOVE THE OUTLIER.



Click here for Task 2 Dataset Link

APPLICATION DATASET-OUTLIERS

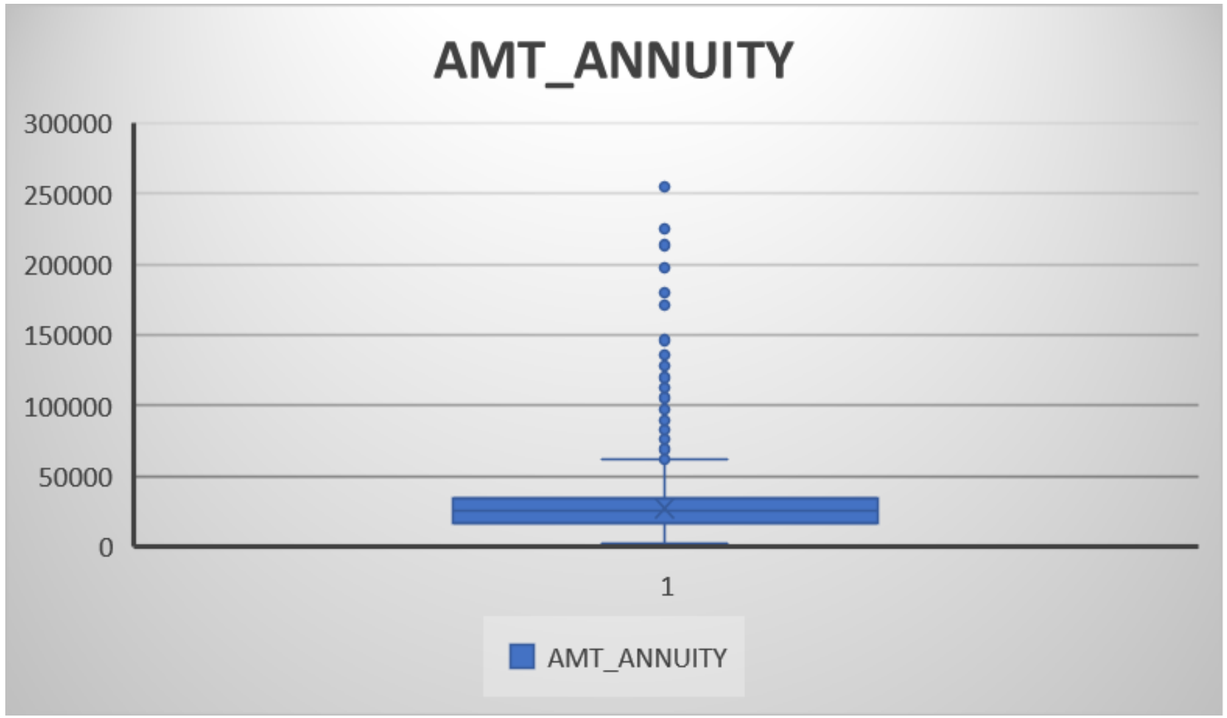
AMT_CREDITED	
QUARTILE 1	270000
QUARTILE 3	808650
IQR	538650
LOWER BOUND	-537975
UPPER BOUND	1616625



HERE WE CAN SEE THAT THERE LIES OUTLIERS
WHOSE VALUE IN MUCH GREATER THAN UPPER BOUND
CALCULATED BUT SINCE CREDIT VARIES FROM
PERSON TO PERSON HENCE WE WILL NOT REMOVE THE OUTLIER.

APPLICATION DATASET-OUTLIERS

AMT_ANNUITY	
QUARTILE 1	16456.5
QUARTILE 3	34596
IQR	18139.5
LOWER BOUND	-10752.8
UPPER BOUND	61805.25



FIRST OUTLIER IN AMT_ANNUITY WHICH IS GREATER THAN 250000 IS REPLACED WITH MEDIAN VALUE.

APPLICATION DATASET-OUTLIERS

AMT_GOODS_PRICE	
QUARTILE 1	238500
QUARTILE 3	679500
IQR	441000
LOWER BOUND	-423000
UPPER BOUND	1341000

HERE WE CAN SEE THAT THERE LIES OUTLIERS WHOSE VALUE IN MUCH GREATER THAN UPPER BOUND CALCULATED BUT SINCE GOODS PRICE VARIES FROM PERSON TO PERSON HENCE WE WILL NOT REMOVE THE OUTLIER.

APPLICATION DATASET-OUTLIERS

DAYS_EMPLOYED	
QUARTILE 1	933
QUARTILE 3	5725
IQR	4792
LOWER BOUND	-6255
UPPER BOUND	12913
MEDIAN	2217

HERE WE CAN SEE THAT THERE EXIST ONLY ONE OUTLIER '365243'.WE CAN REPLACE IT BY MEDIAN OF THE COLUMN

APPLICATION DATASET-ADDITIONAL CLEANING

DAYS_BIRTH	YEARS_BIRTH	DAYS_EMPLOYED	YEARS_EMPLOYED
9461	25.9	637	1.7
16765	45.9	1188	3.3
19046	52.2	225	0.6
19005	52.1	3039	8.3
19932	54.6	3038	8.3
16941	46.4	1588	4.4
13778	37.7	3130	8.6
18850	51.6	449	1.2
20099	55.1	2217	6.1
14469	39.6	2019	5.5
10197	27.9	679	1.9
20417	55.9	2217	6.1
13439	36.8	2717	7.4
14086	38.6	3028	8.3
14583	40	203	0.6
8728	23.9	1157	3.2
12931	35.4	1317	3.6
9776	26.8	191	0.5
17718	48.5	7804	21.4
11348	31.1	2038	5.6
18252	50	4286	11.7
14815	40.6	1652	4.5

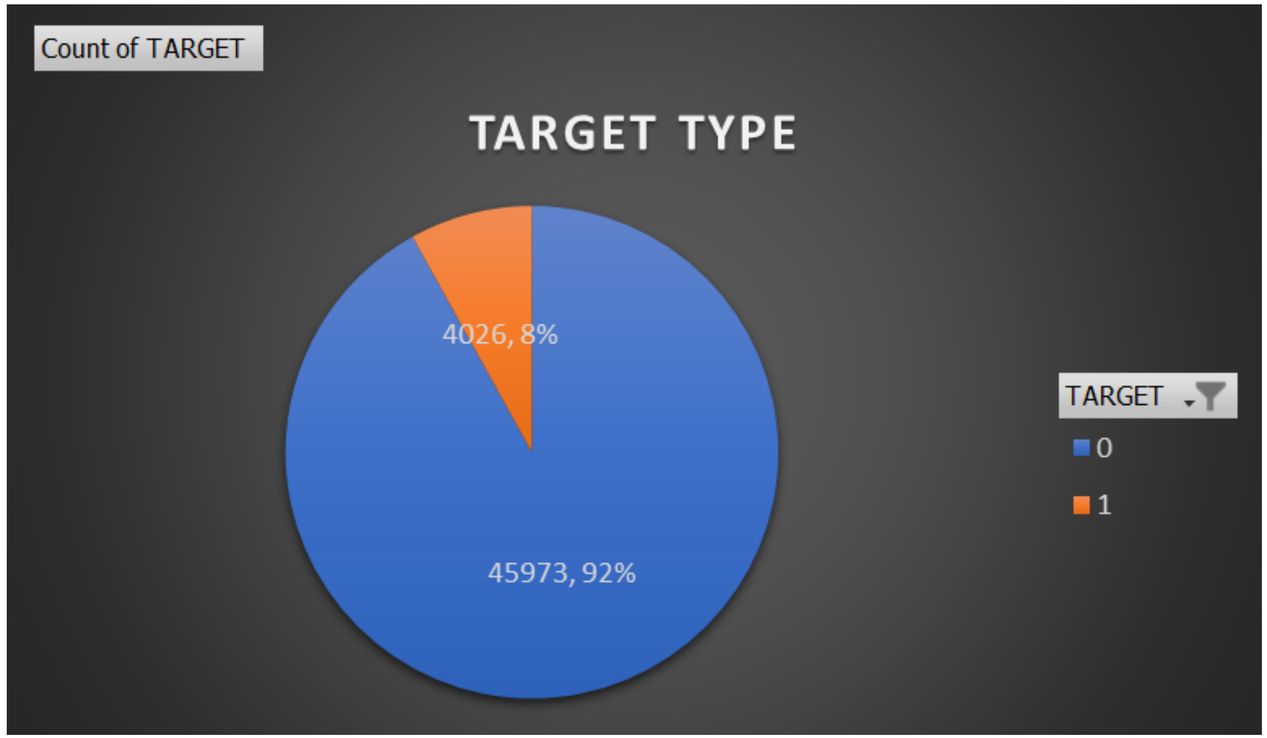
- TWO NEW COLUMNS YEARS_BIRTH AND YEARS_EMPLOYED WERE ADDED IN THE TABLE
- 34 DUPLICATES WERE FOUND AND REMOVED
- IRRELEVANT DATA BELOW ROWS 50000 WERE FOUND AND REMOVED
- REPLACED NEGATIVE DAYS VALUE WITH POSTIVE FOR BETTER CLARITY DURING ANALYSIS

APPLICATION DATASET-DATA IMBALANCE

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

TARGET COLUMN	
TARGET	**Count of TARGET**
0	45973
1	4026
Grand Total	**49999**

From the Pie Chart We can Analyse that
92% of the total client had no payment difficulties
while 8% of the total client had Payment difficulties



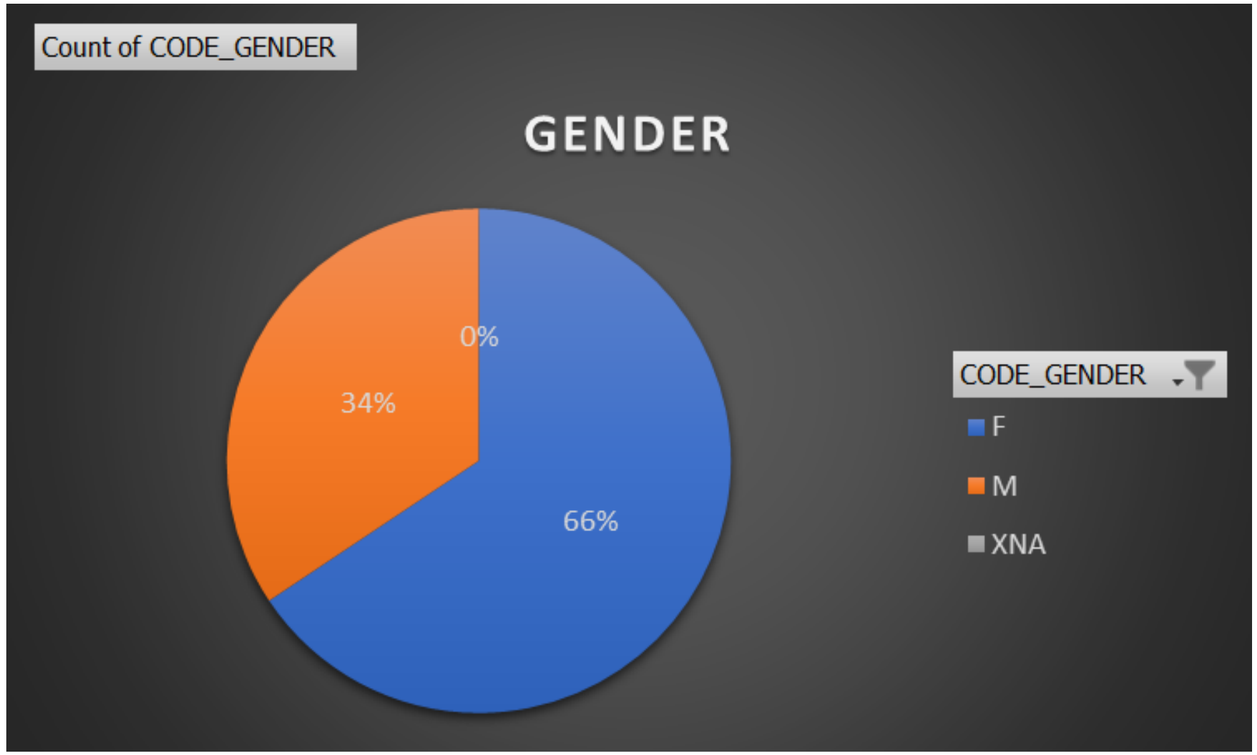
Click here for Task 3 Dataset Link

APPLICATION DATASET-DATA IMBALANCE

CODE_GENDER COLUMN

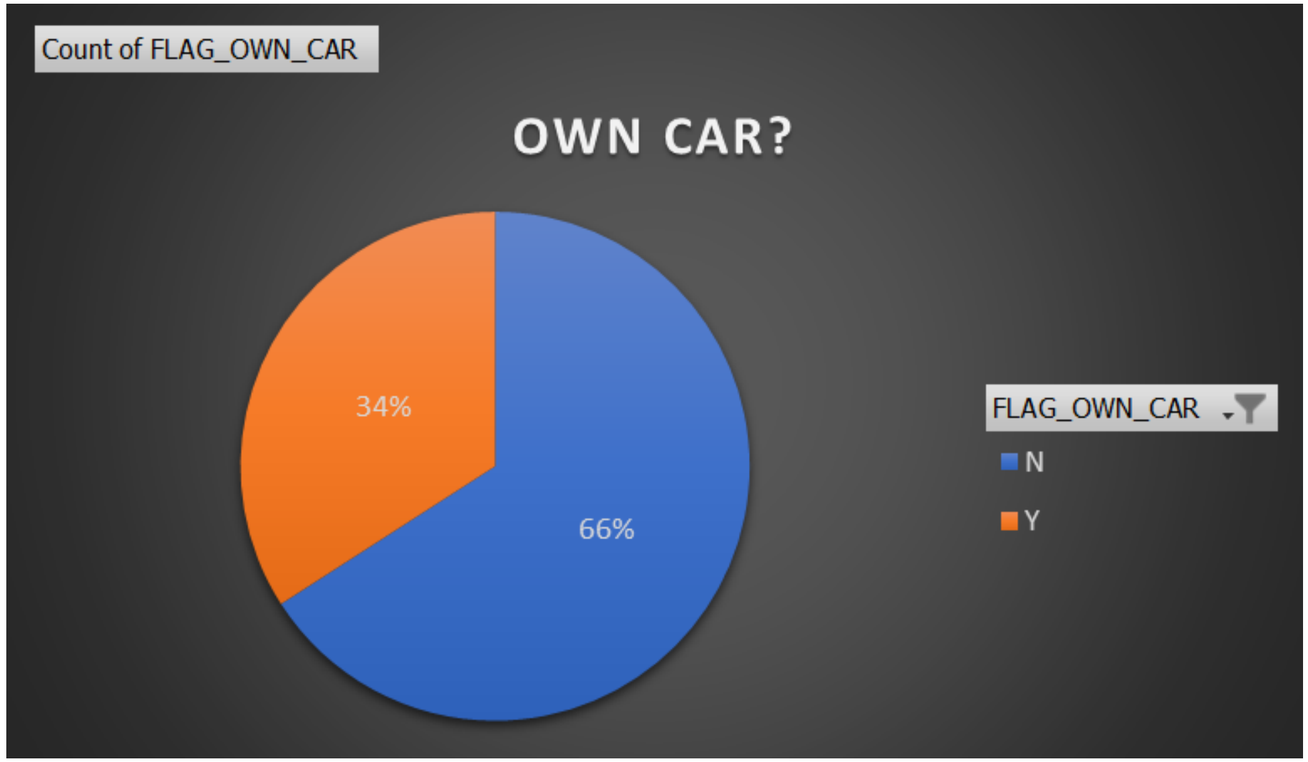
GENDER TYPE	Count of CODE_GENDER
F	32823
M	17174
XNA	2
Grand Total	49999

From the Pie Chart we Can infer that 66% of the total clients are Female while 34% of the total clients are Male while 2 of the clients have a gender as XNA which can be ignored.



APPLICATION DATASET-DATA IMBALANCE

FLAG_OWN_CAR COLUMN	
OWN CAR	Count of FLAG_OWN_CAR
N	32949
Y	17050
Grand Total	49999

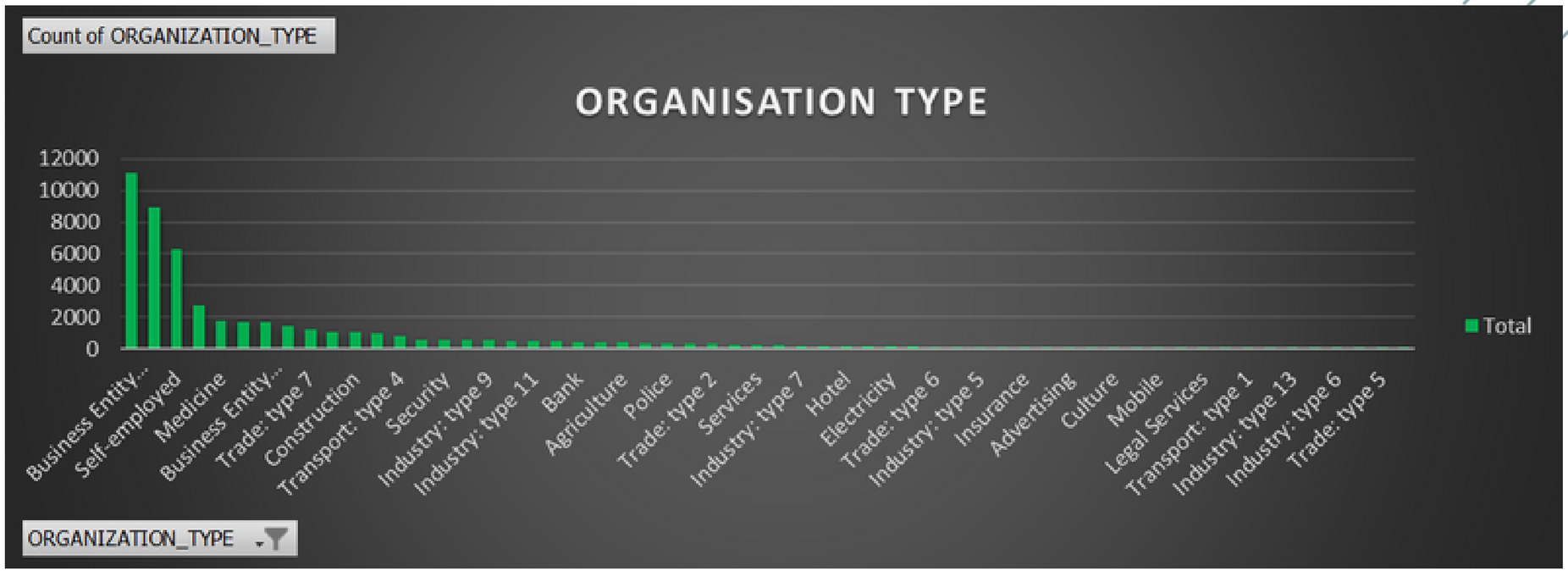


From the Pie Chart we can infer that 66% of the clients don't own a car while 34% of the cleint do own a car

APPLICATION DATASET-DATA IMBALANCE

ORGANISATION_TYPE

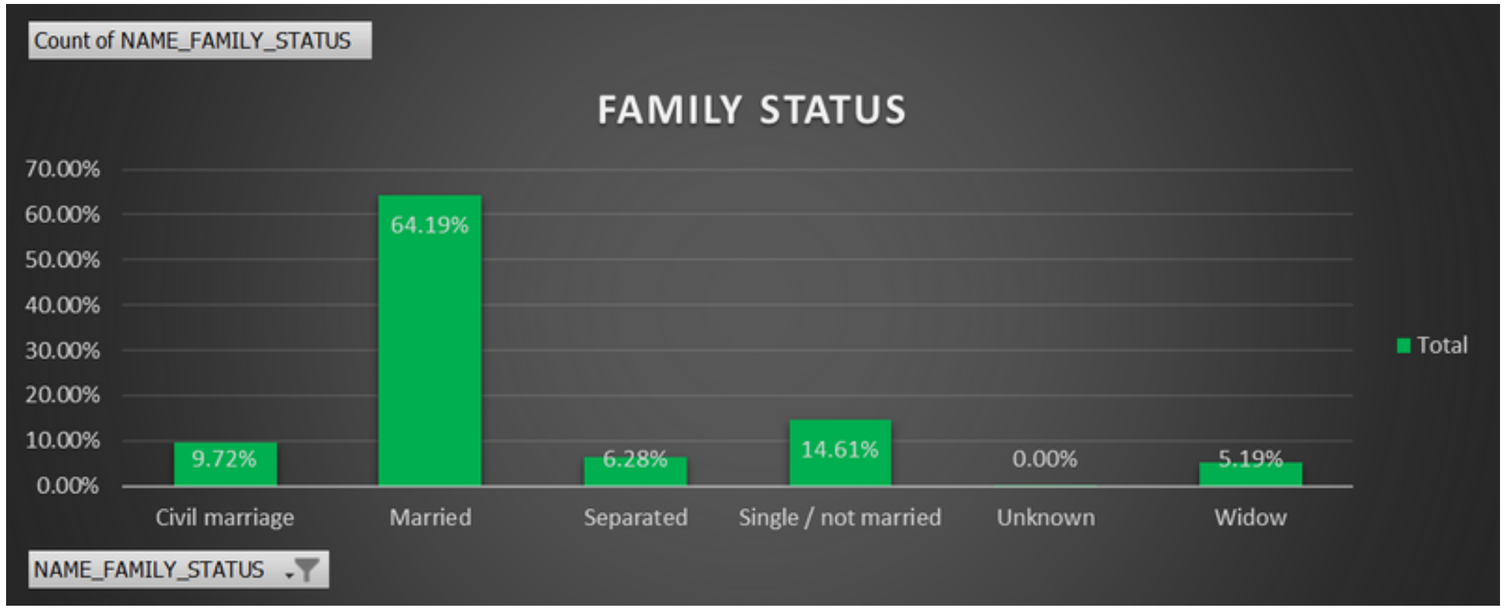
Row Labels	Count of ORGANIZATION_TYPE
Business Entity Typ	11101
XNA	8924
Self-employed	6240
Other	2717
Medicine	1817
Government	1716
Business Entity Typ	1704
School	1450
Trade: type 7	1210
Kindergarten	1090
Construction	1066
Business Entity Typ	953
Transport: type 4	837
Trade: type 3	550
Security	550
Industry: type 3	542
Industry: type 9	537
Housing	489
Industry: type 11	489
Military	458
Bank	435
Transport: type 2	392
Agriculture	392
Postal	370
Police	366
Transport: type 3	191
Hotel	182
Industry: type 1	159
Electricity	147
Industry: type 4	140
Trade: type 6	108
Telecom	106
Industry: type 5	103
Emergency	93
Insurance	89
Industry: type 2	78
Advertising	68
Trade: type 1	66
Culture	64
Realtor	61
Mobile	56
Industry: type 12	53
Legal Services	44
Cleaning	40
Transport: type 1	28
Industry: type 10	21
Industry: type 13	15
Religion	14
Industry: type 6	12
Industry: type 8	8
Trade: type 5	8
Trade: type 4	8
Grand Total	49999



From the Bar Graph we can infer that most of the clients belongs to Business Entity Type 3,XNA and Self-employed

APPLICATION DATASET-DATA IMBALANCE

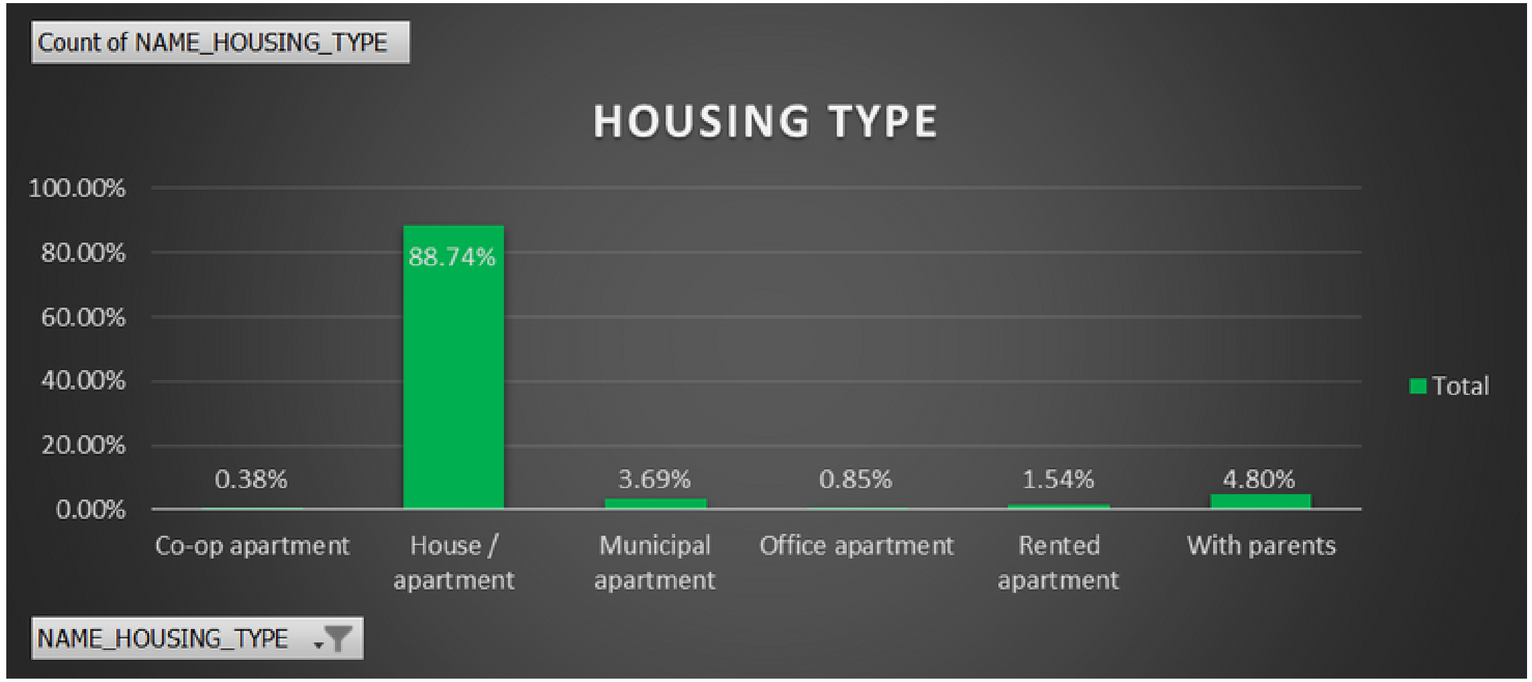
NAME FAMILY STATUS	
FAMILY STATUS	**Count of NAME_FAMILY_STATUS**
Civil marriage	9.72%
Married	64.19%
Separated	6.28%
Single / not marrie	14.61%
Unknown	0.00%
Widow	5.19%
Grand Total	**100.00%**



From the Bar Graph we can infer that 64.19% of the total clients belongs to Married Status

APPLICATION DATASET-DATA IMBALANCE

NAME HOUSING TYPE	
HOUSING TYPE	**Count of NAME_HOUSING_TYPE**
Co-op apartment	0.38%
House / apartment	88.74%
Municipal apartme	3.69%
Office apartment	0.85%
Rented apartment	1.54%
With parents	4.80%
Grand Total	**100.00%**



From the Graph we can infer that 88.74% of Housing_Type of the clients is House/Apartment

APPLICATION DATASET ANALYSIS

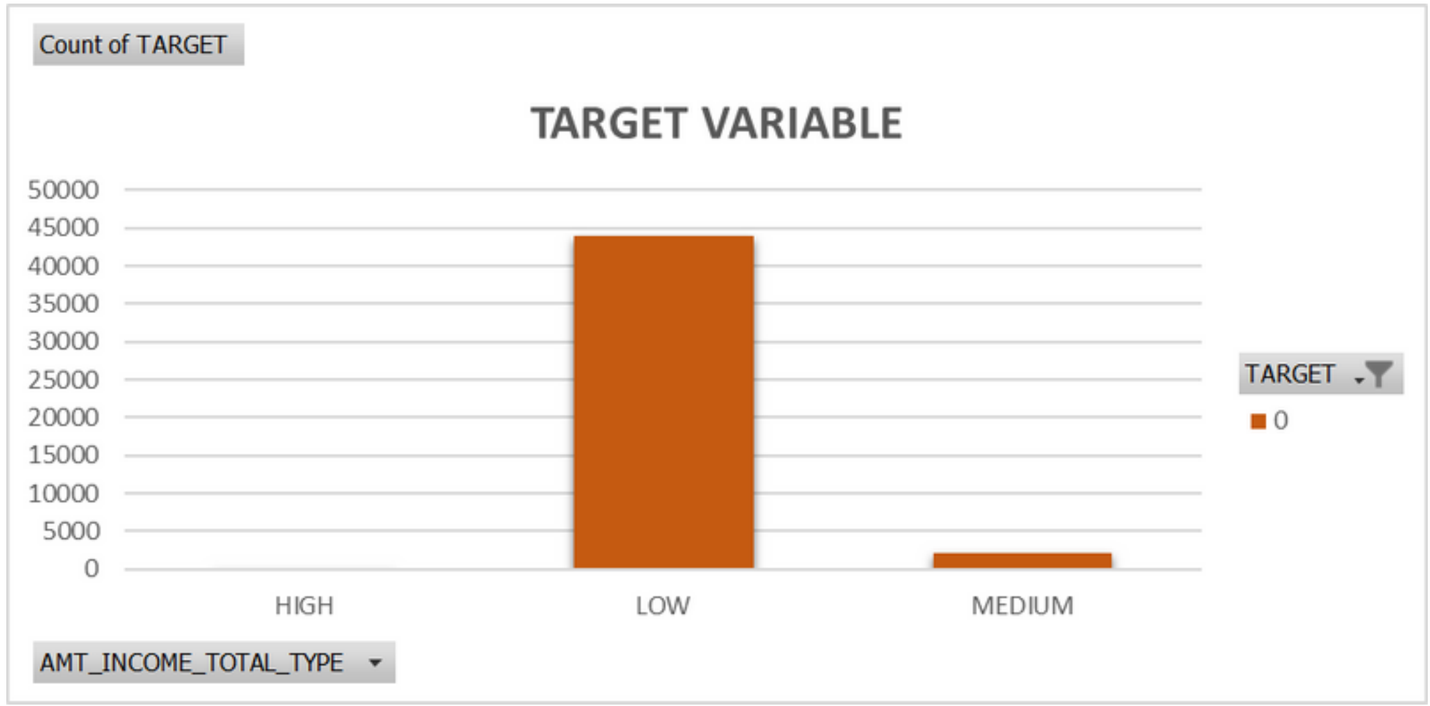
Click here for Dataset Link

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

AMT_INCOME_TOTAL(TARGET:0)

Count of TARGET	Column Labels
INCOME TYPE	0
HIGH	37
LOW	43859
MEDIUM	2077
Grand Total	45973

HIGH	>800000
MEDIUM	>=300000
LOW	<300000

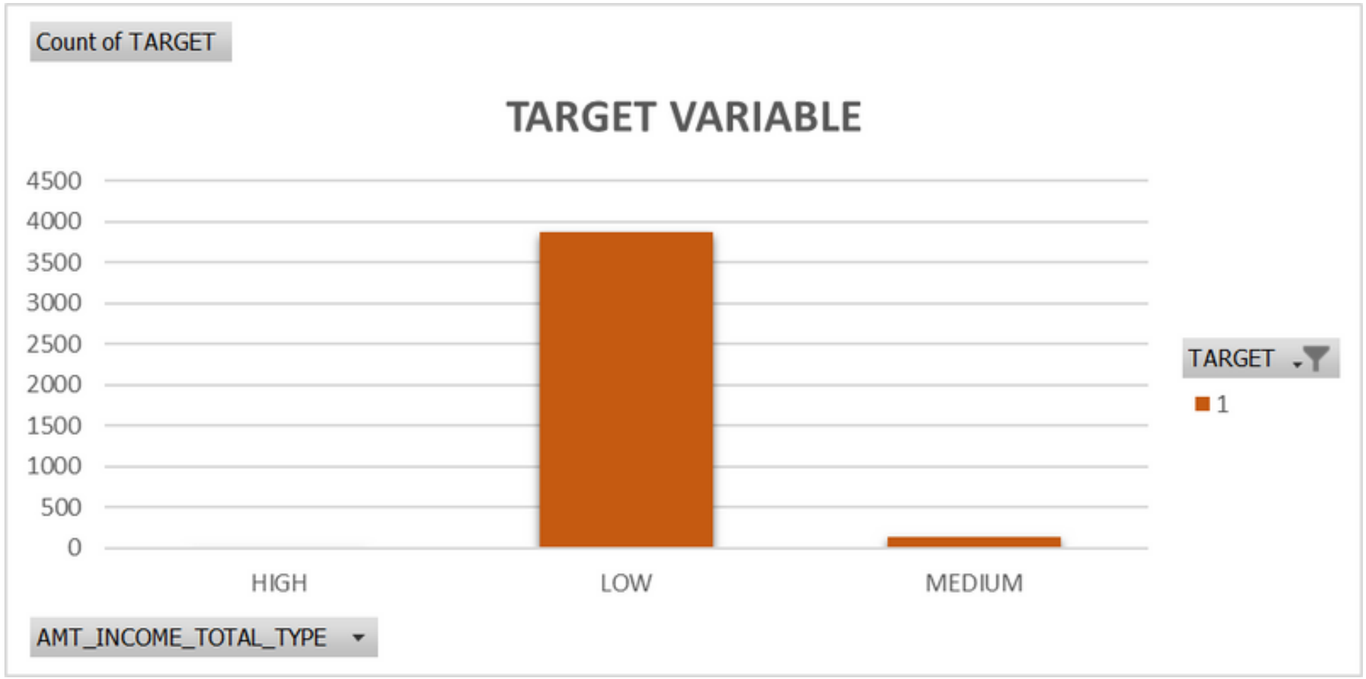


FROM THE ABOVE BAR GRAPH WE CAN INFER THAT CLIENTS HAVING TOTAL INCOME RANGE AS LOW HAVE THE HIGHEST COUNT WHEN IT COMES TO CLIENTS HAVING NO PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

AMT_INCOME_TOTAL(TARGET:1)	
Count of TARGET	**Column Labels**
Row Labels	**1**
HIGH	3
LOW	3880
MEDIUM	143
Grand Total	**4026**

HIGH	>800000
MEDIUM	>=300000
LOW	<300000

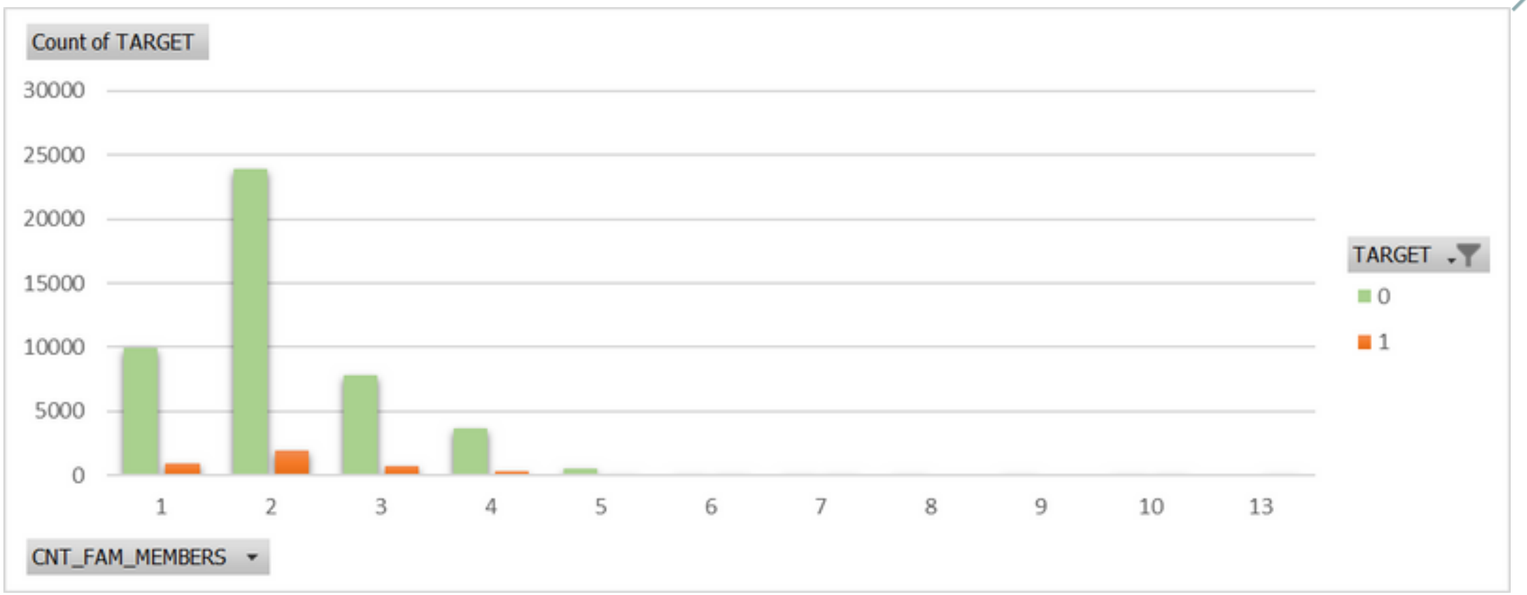


FROM THE ABOVE BAR GRAPH WE CAN INFER THAT CLIENTS HAVING TOTAL INCOME RANGE AS LOW HAVE THE HIGHEST COUNT WHEN IT COMES TO CLIENTS HAVING PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

CNT_FAM_MEMBERS (TARGET VARIABLE)

Count of TARGET	Column Labels	
Row Labels	0	1
1	9951	922
2	23902	1906
3	7858	777
4	3651	349
5	538	54
6	55	13
7	9	3
8	6	
9	2	
10	1	1
13		1
Grand Total	45973	4026

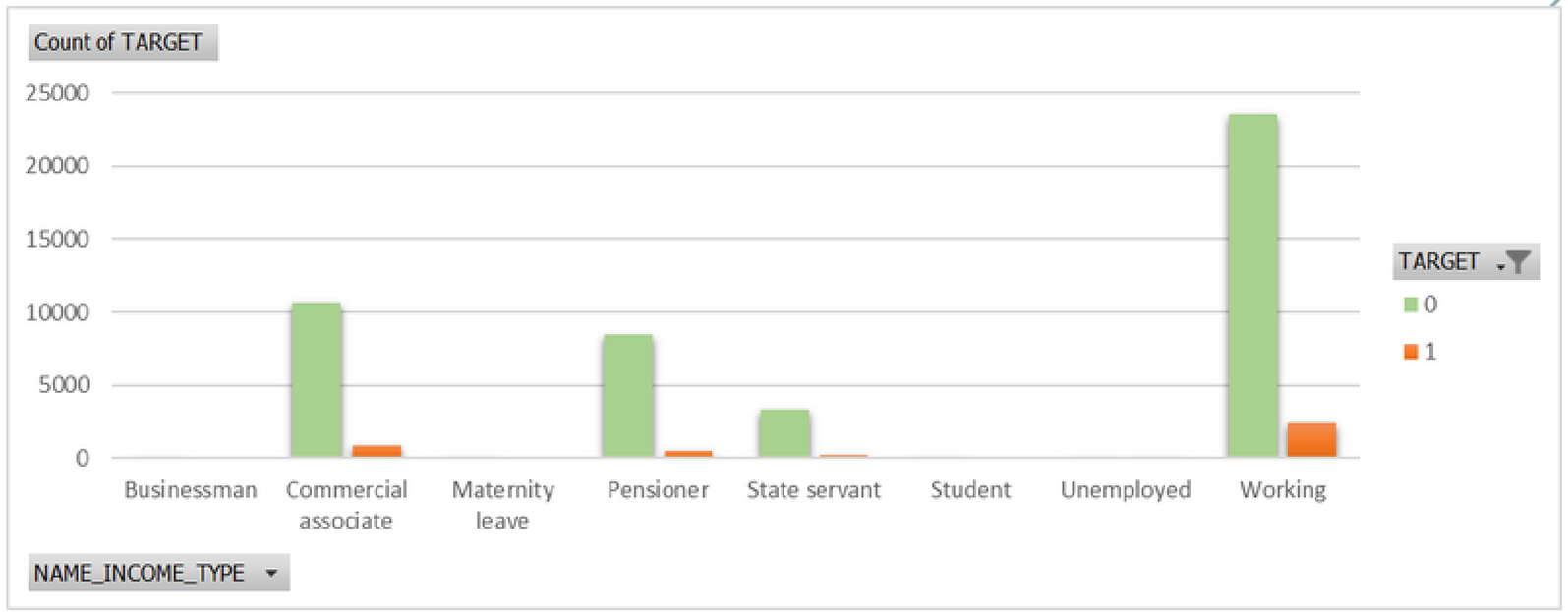


FROM THE ABOVE BAR GRAPH WE CAN INFER THAT CLIENTS HAVING TOTAL COUNT OF FAMILY MEMBERS AS '2' HAVE HIGHEST COUNT WHEN IT COMES TO CLIENT HAVING NO PAYMENT ISSUES AND CLIENTS WITH PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

NAME__INCOME_TYPE(TARGET VARIABLE)

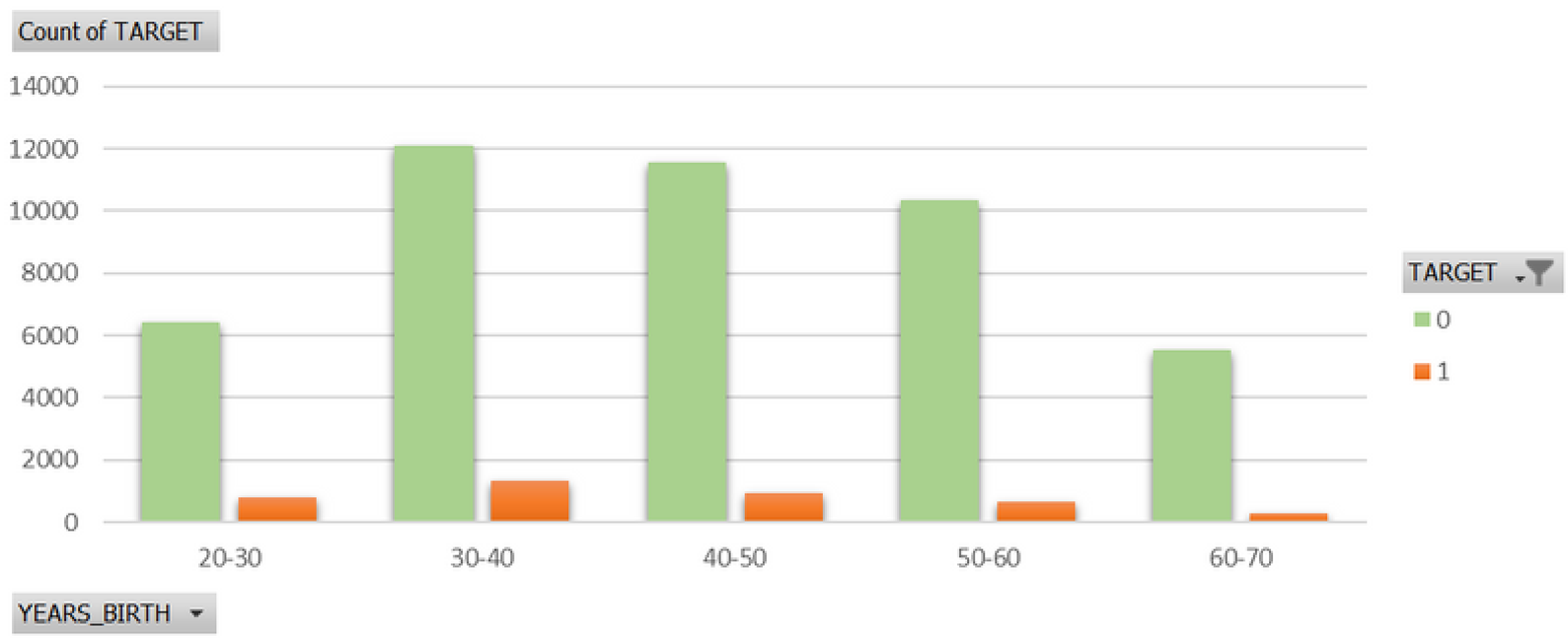
Count of TARGET	Column Labels	
Row Labels	0	1
Businessman	2	
Commercial associate	10679	864
Maternity leave	1	
Pensioner	8419	501
State servant	3314	198
Student	5	
Unemployed	4	2
Working	23549	2461
Grand Total	**45973**	**4026**



FROM THE ABOVE BAR CHART WE CAN INFER THAT CLIENTS HAVING NAME_INCOME_TYPE AS 'WORKING' HAVE THE HIGHEST COUNT FOR CLIENT WITH NO PAYMENT ISSUES AND CLIENT WITH PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

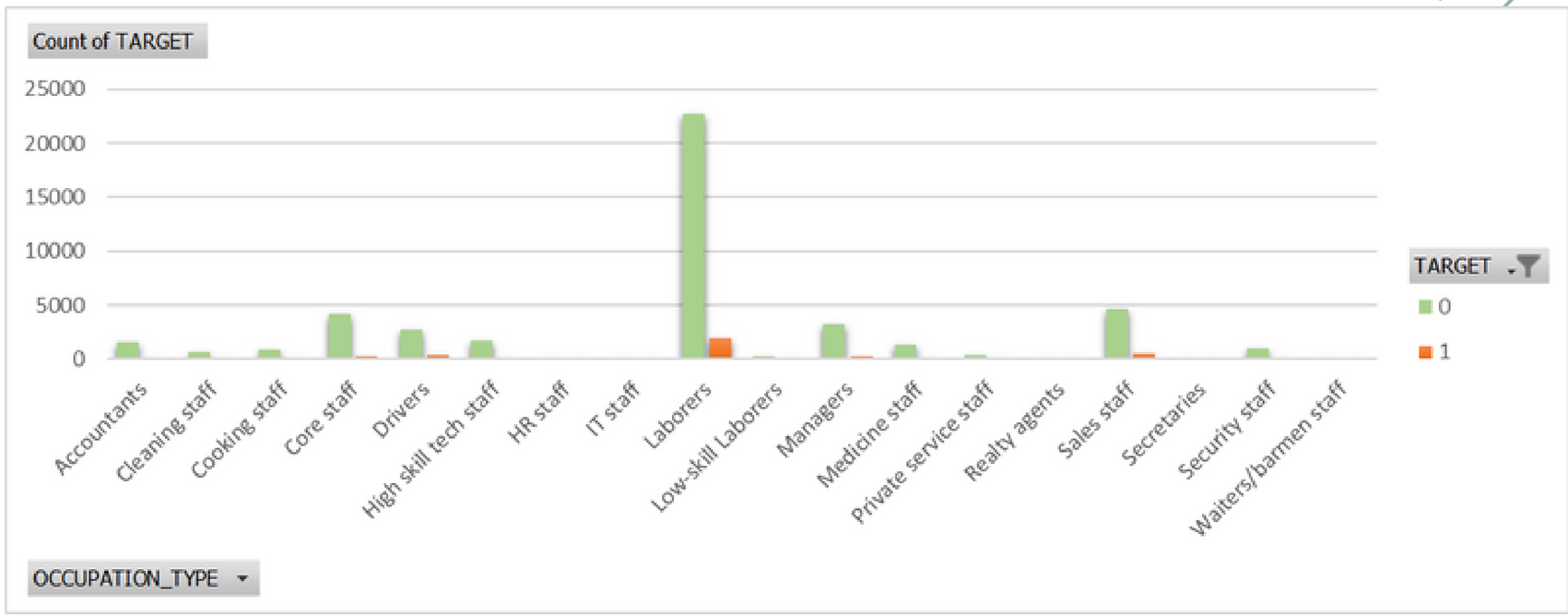
AGE (YEARS_BIRTH)	TARGET VARIABLE	
Count of TARGET	Column Labels	
Row Labels	0	1
20-30	6408	809
30-40	12118	1312
40-50	11565	944
50-60	10357	671
60-70	5525	290
Grand Total	45973	4026



FROM THE BAR GRAPH WE CAN INFER THAT CLIENTS IN THE AGE GROUP 30-40
HAVE THE HIGHEST COUNT FOR CLIENTS WITH NO PAYMENT ISSUES AND CLIENTS WITH PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

OCCUPATION_TYPE	(TARGET VARIABLE)	
Count of TARGET	**Column Labels**	
Row Labels	**0**	**1**
Accountants	1540	81
Cleaning staff	671	68
Cooking staff	862	101
Core staff	4184	250
Drivers	2706	338
High skill tech staff	1734	118
HR staff	92	9
IT staff	76	4
Laborers	22660	1946
Low-skill Laborers	296	61
Managers	3246	243
Medicine staff	1297	106
Private service staff	410	37
Realty agents	110	13
Sales staff	4668	492
Secretaries	203	9
Security staff	1015	125
Waiters/barmen staf	203	25



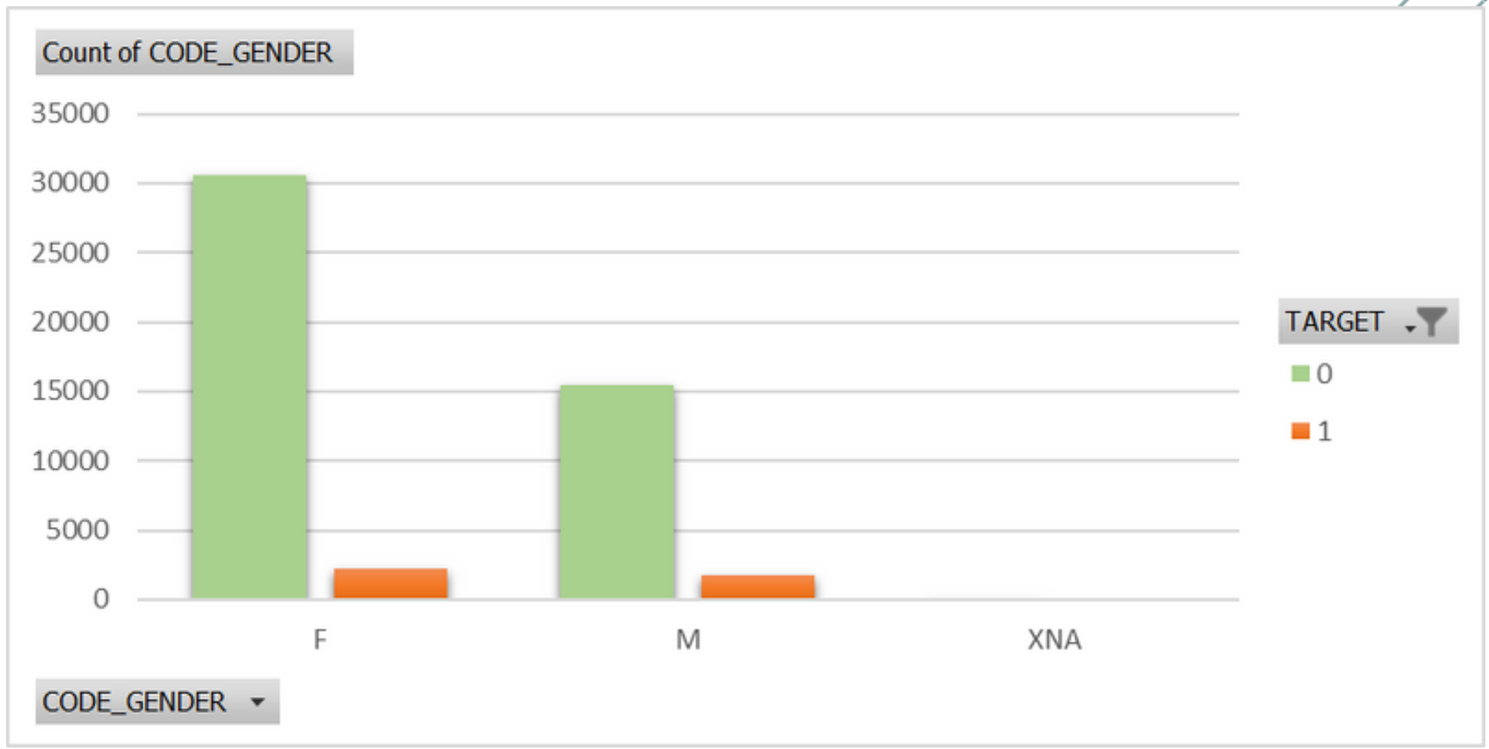
FROM THE BAR GRAPH WE CAN INFER THAT CLIENTS WITH THE OCCUPATION TYPE AS LABORERS' HAVE THE HIGHEST COUNT FOR CLIENTS WITH NO PAYMENT ISSUES AND CLIENTS WITH NO PAYMENT ISSUES

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

CODE_GENDER(TARGET VARIABLE)

Count of CODE_GENI	Column Labels		
Row Labels	0	1	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	45973	4026	49999

% of FEMALE DEFAULTERS	6.9%
% of MALE DFAULTERS	10.26%

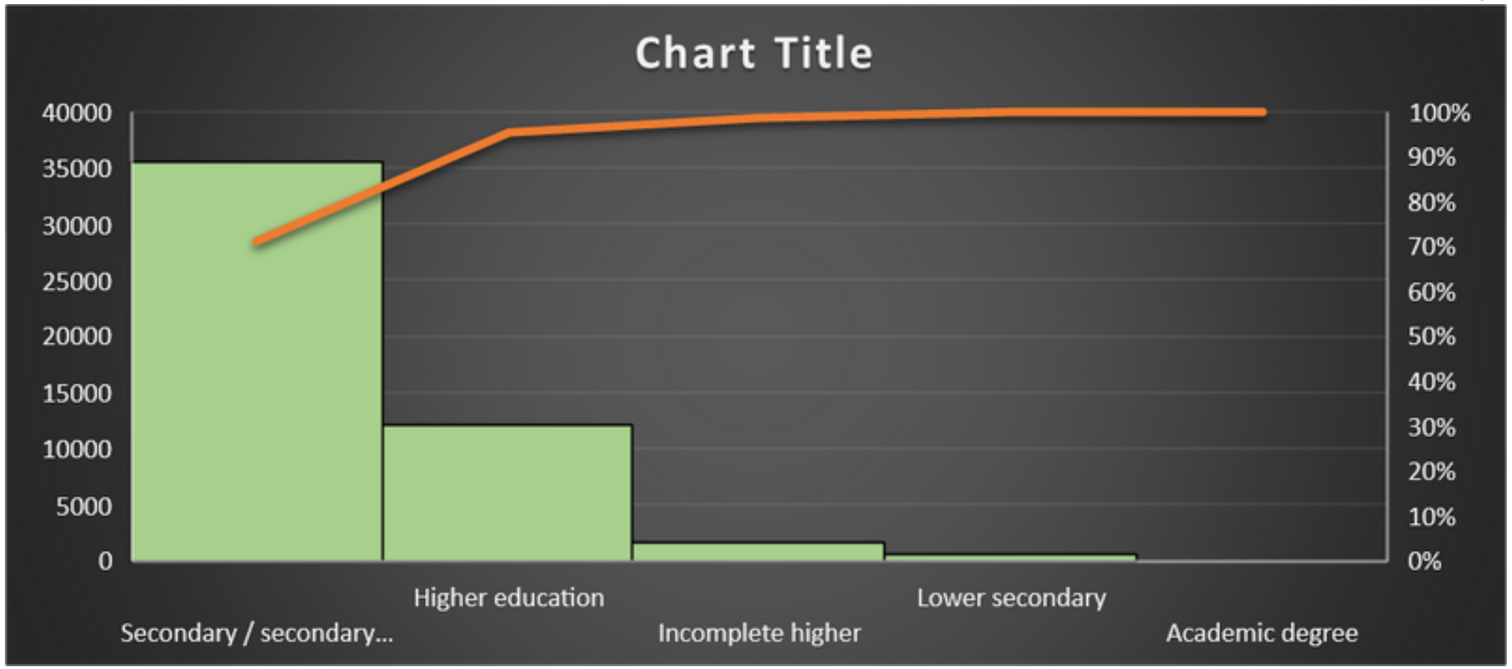


FROM THE ABOVE GRAPH WE CAN INFER THAT GENDER MALE HAS MAXIMUM PERCANTAGE OF DEFAULTERS

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

NAME_EDUCATION_TYPE

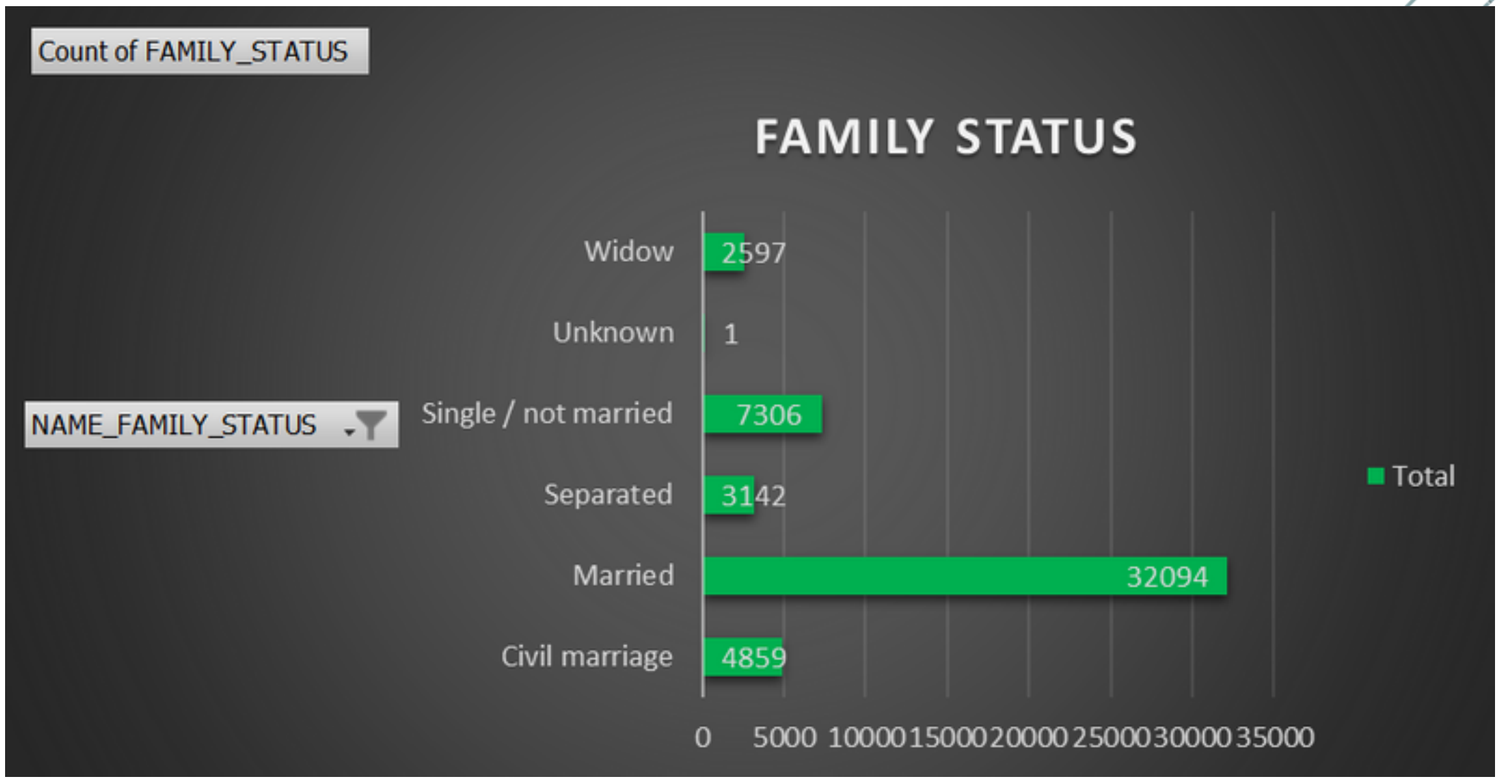
Education Type	Count of Applicants
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary spec	35572



FROM THE ABOVE HISTOGRAM WE CAN INFER THAT MAXIMUM NUMBER OF APPLIACNTS STUDIED SECONDARY

UNIVARIATE/SEGMENTED UNIVARIATE ANALYSIS

NAME FAMILY STATUS	
Row Labels	**Count of FAMILY_STATUS**
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597
Grand Total	**49999**



FROM THE GRAPH WE CAN INFER THAT MAXIMUM COUNT OF APPLICANTS ARE THOSE WHOSE FAMILY STATUS IS 'MARRIED'

BIVARIATE ANALYSIS

AMT_INCOME_TOTAL_TYPE VS CODE_GENDER

Count of TARGET	Column Labels	
INCOME TYPE	1	Grand Total
HIGH	3	3
F	2	2
M	1	1
LOW	3880	3880
F	2209	2209
M	1671	1671
MEDIUM	143	143
F	53	53
M	90	90
Grand Total	4026	4026

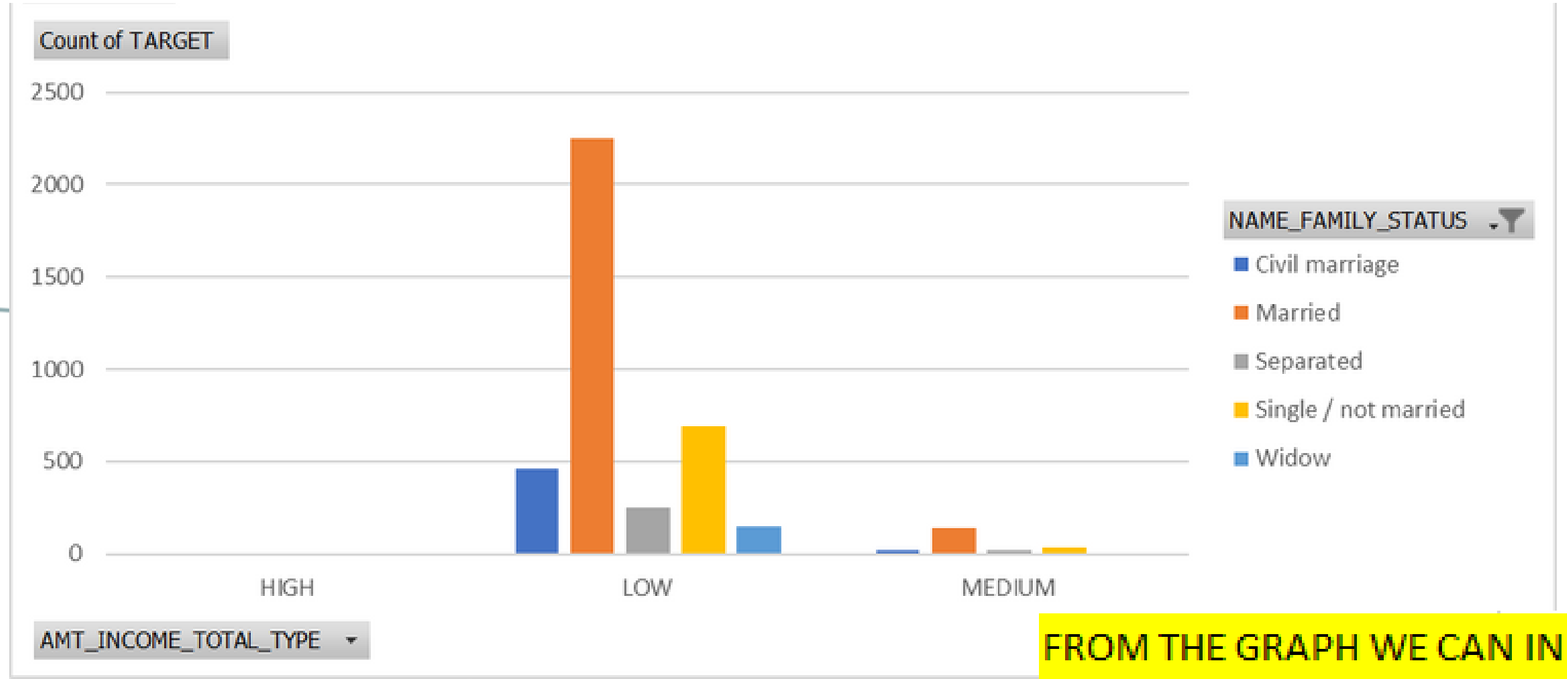
WE CAN INFER FROM THE GRAPH THAT FEMALES BELONGING TO LOW INCOME GROUP HOLD THE HIGHEST COUNT FOR CLIENTS WITH PAYMENT ISSUES

BIVARIATE ANALYSIS

TOTAL_INCOME_TYPE VS FAMILY_STATUS

TARGET 1

Count of TARGET	Column Labels					
INCOME TYPE	Civil marriage	Married	Separated	Single / not married	Widow	Grand Total
HIGH	1	5		1		7
LOW	458	2250	254	691	144	3797
MEDIUM	23	140	18	37	4	222
Grand Total	**482**	**2395**	**272**	**729**	**148**	**4026**

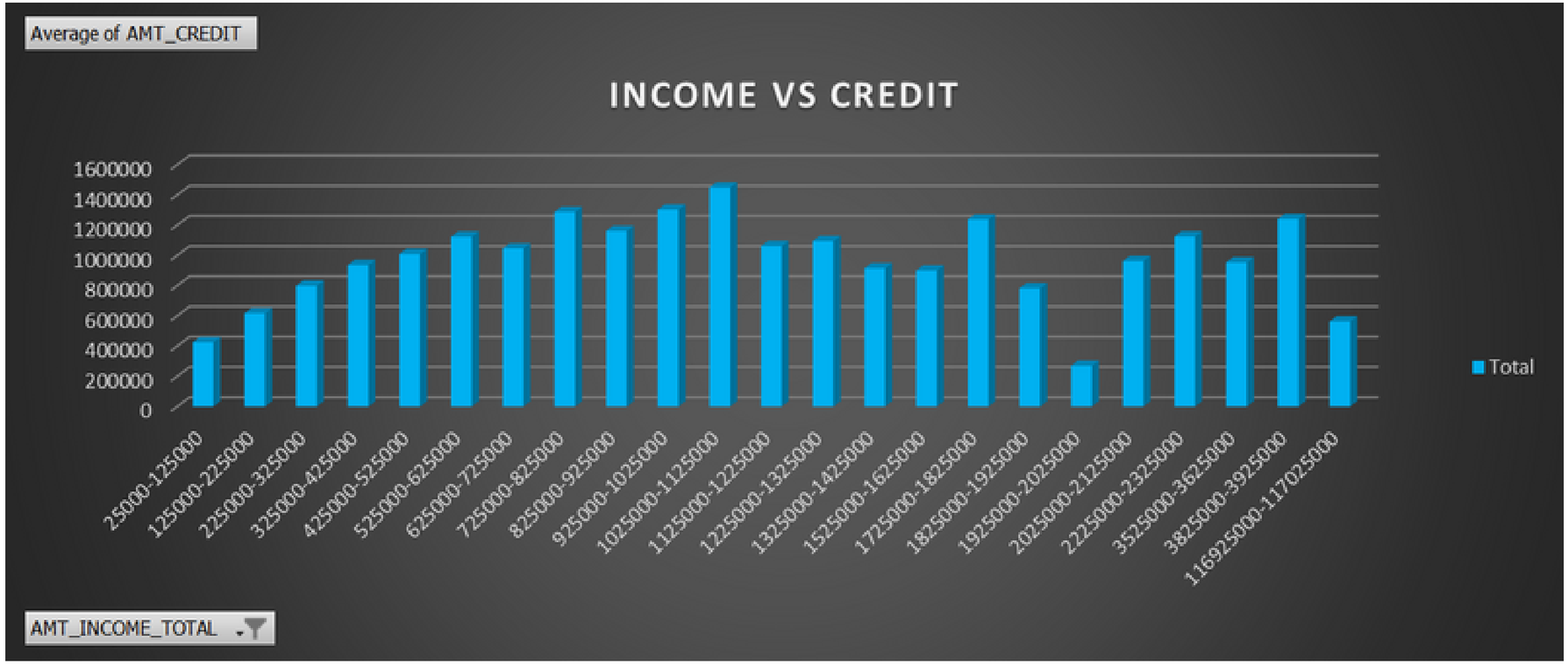


FROM THE GRAPH WE CAN INFER THAT CLIENTS WITH AMT_INCOME_TYPE AS 'LOW' AND FAMILY_STATUS AS 'MARRIED' HAS THE MAXIMUM COUNT FOR CLIENTS WITH PAYMENT ISSUES

BIVARIATE ANALYSIS

AMT_INCOME_ VS AMT_CREDITED

RANGE OF INCOME	Average of AMT_CREDIT
25000-125000	425228.7416
125000-225000	616243.2104
225000-325000	799074.0725
325000-425000	935945.9604
425000-525000	1009091.246
525000-625000	1123616.396
625000-725000	1046201.618
725000-825000	1287182.647
825000-925000	1161345.214
925000-1025000	1303200
1025000-1125000	1450125
1125000-1225000	1062698.885
1225000-1325000	1095111
1325000-1425000	914911.2
1525000-1625000	900000
1725000-1825000	1237500
1825000-1925000	781920
1925000-2025000	269550
2025000-2125000	961827.75
2225000-2325000	1125000
3525000-3625000	953460
3825000-3925000	1241023.5
116925000-11702500	562491
Grand Total	599700.5815



FROM THE GRAPH WE CAN INFER THE AVERAGE CREDIT PROVIDED TO CLIENTS WITHIN THE RANGE OF TOTAL_INCOME

CORRELATON

TASK: SEGMENT THE DATASET BASED ON DIFFERENT SCENARIOS (E.G., CLIENTS WITH PAYMENT DIFFICULTIES AND ALL OTHER CASES) AND IDENTIFY THE TOP CORRELATIONS FOR EACH SEGMENTED DATA USING EXCEL FUNCTIONS.

CORRELATION FOR TARGET VARIABLE 1

	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_INCOME_TOTAL	AMT_CREDIT	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	HOUR_APPR_PROCESS_START
CNT_CHILDREN	1								
CNT_FAM_MEMBERS	0.892521875	1							
AMT_INCOME_TOTAL	0.010110177	0.013121678	1						
AMT_CREDIT	0.007601905	0.06124869	0.015271444	1					
DAYS_BIRTH	-0.2496732	-0.199141397	-0.009033662	0.14250603	1				
DAYS_EMPLOYED	-0.030744275	-0.007464561	-0.008069341	0.09814427	0.28463645	1			
DAYS_REGISTRATION	-0.152113117	-0.151786548	0.009561152	0.0428444	0.28843784	0.146145336	1		
DAYS_ID_PUBLISH	0.042360717	0.044037815	0.009122006	0.0437719	0.24789657	0.109866604	0.09029149	1	
HOUR_APPR_PROCESS_START	-0.006884357	-0.023902962	0.014482013	0.04539638	-0.05789169	0.010604913	0.057808905	-0.005517259	1

CORRELATON

TASK: SEGMENT THE DATASET BASED ON DIFFERENT SCENARIOS (E.G., CLIENTS WITH PAYMENT DIFFICULTIES AND ALL OTHER CASES) AND IDENTIFY THE TOP CORRELATIONS FOR EACH SEGMENTED DATA USING EXCEL FUNCTIONS.

CORRELATION FOR TARGET VARIABLE 0

	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	S_ID_PUBLISH	PR_PROCESS_START
CNT_CHILDREN	1										
CNT_FAM_MEMBERS	0.879238049	1									
AMT_INCOME_TOTAL	0.036319722	0.041599302	1								
AMT_CREDIT	0.005705458	0.064876937	0.377965752	1							
AMT_ANNUITY	0.025692691	0.077398184	0.450224195	0.769917623	1						
AMT_GOODS_PRICE	0.001518097	0.062891858	0.384575912	0.986999774	0.774681662	1					
DAYS_BIRTH	-0.335876269	-0.284379407	-0.073769425	0.051084182	-0.00986325	0.048773297	1				
DAYS_EMPLOYED	-0.052746271	-0.023824409	0.040251296	0.089300681	0.061241873	0.090054399	0.241341736	1			
DAYS_REGISTRATION	-0.183072478	-0.171482728	-0.06893375	-0.00805376	-0.03436184	-0.011260199	0.335028046	0.137554392	1		
DAYS_ID_PUBLISH	0.032537221	0.025054258	-0.032286356	0.008290189	-0.00984231	0.00938552	0.270073313	0.063355271	0.103548902	1	
HOUR_APPR_PROCESS_START	-0.005272551	-0.010117379	0.08543156	0.056524809	0.053504951	0.065271608	-0.09638927	-0.017039741	0.002396446	-0.03797	1

PREVIOUS APPLICATION DATASET CLEANING

Click Here for Dataset Link

PREVIOUS APPLICATION DATASET-NULL VALUES

Firstly the percentage of null values needs to be analyzed and those columns that have more than 50% of the null data have to be dropped

Columns in RED having NULL values Grater than 50% needs to be REMOVED

SNO	COLUMN NAME	NUMBER OF BLANK CELLS	% OF BLANK VALUES
1	AMT_DOWN_PAYMENT	25198	50.4
2	RATE_DOWN_PAYMENT	25198	50.4
3	RATE_INTEREST_PRIMARY	49834	99.67
4	RATE_INTEREST_PRIVILEGED	49834	99.67

PREVIOUS APPLICATION DATASET-IRRELEVANT COLUMN

Secondly those columns which are irrelevant and is not required during analysis process needs to be dropped

Columns in Orange needs to be dropped since they are irrelevant for Analysis

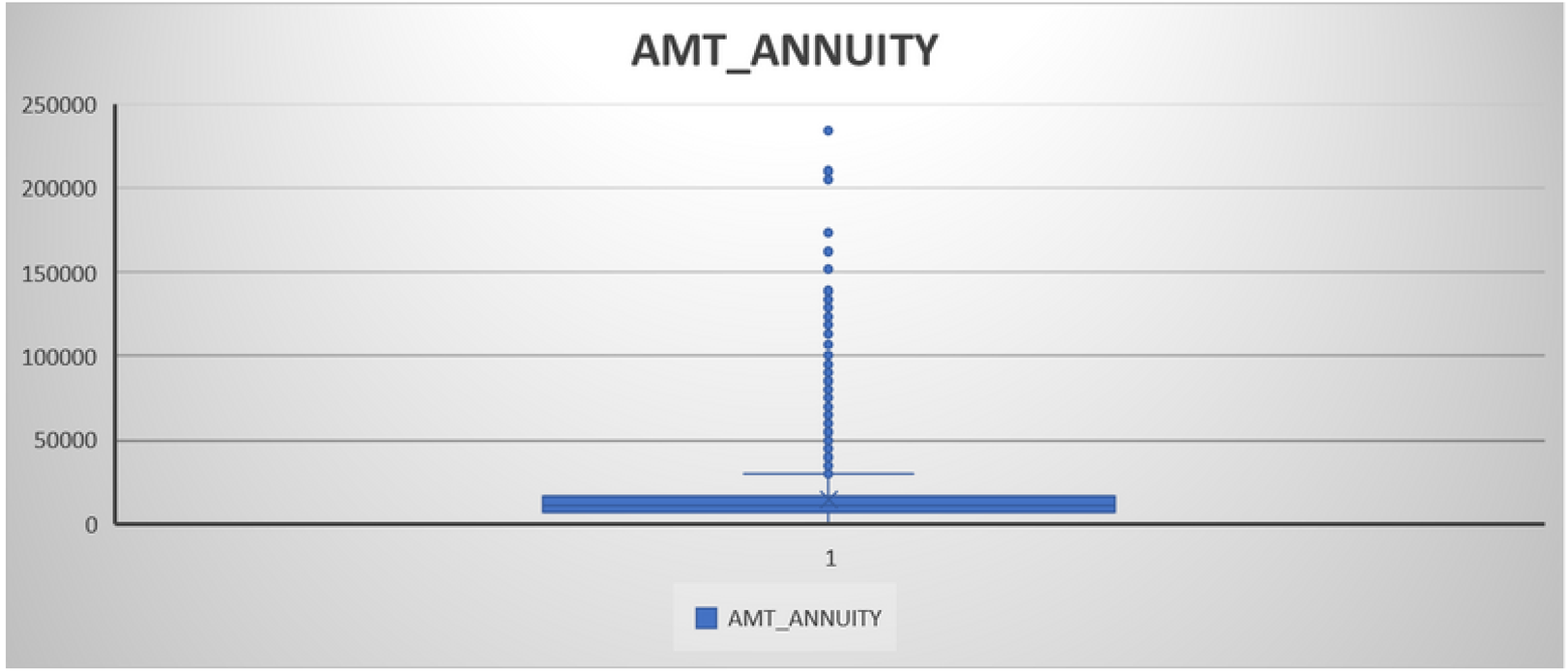
SNO	COLUMN NAME
1	SK_ID_CURR
2	WEEKDAY_APPR_PROCESS_START
3	HOUR_APPR_PROCESS_START
4	FLAG_LAST_APPL_PER_CONTRACT
5	NAME_GOODS_CATEGORY
6	NAME_PRODUCT_TYPE
7	DAYS_FIRST_DRAWING
8	DAYS_FIRST_DUE
9	DAYS_LAST_DUE_1ST_VERSION
10	DAYS_LAST_DUE
11	DAYS_TERMINATION
12	NFLAG_INSURED_ON_APPROVAL

PREVIOUS APPLICATION DATASET ANALYSIS

Click here for Dataset Link

PREVIOUS APPLICATION DATASET-ANALYSIS

AMT_ANNUITY



MEDIAN	10879.92

SINCE THE COLUMN CONTAINS OUTLIER
HENCE WE WILL USE MEDIAN TO FILL
MISSING VALUES

PREVIOUS APPLICATION DATASET-ANALYSIS

AMT_GOODS_PRICE

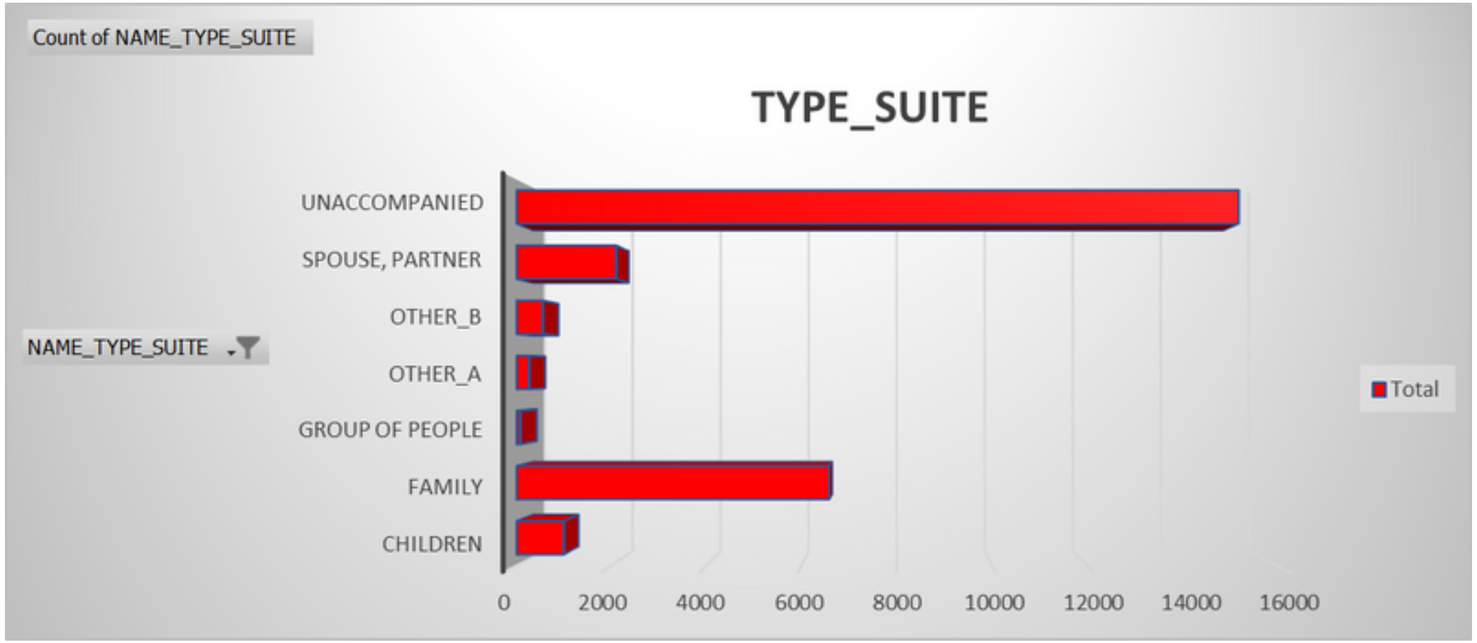
MEDIAN	104017.5

SINCE THE DATA CONSIST OUTLIERS
HENCE WE WILL USE MEDIAN TO FILL
THE BLANKS IN COLUMN

PREVIOUS APPLICATION DATASET-ANALYSIS

NAME_TYPE_SUITE

NAME_TYPE_SUITE	Count of NAME_TYPE_SUITE
Children	993
Family	6581
Group of people	76
Other_A	262
Other_B	551
Spouse, partner	2098
Unaccompanied	15195
Grand Total	**25756**

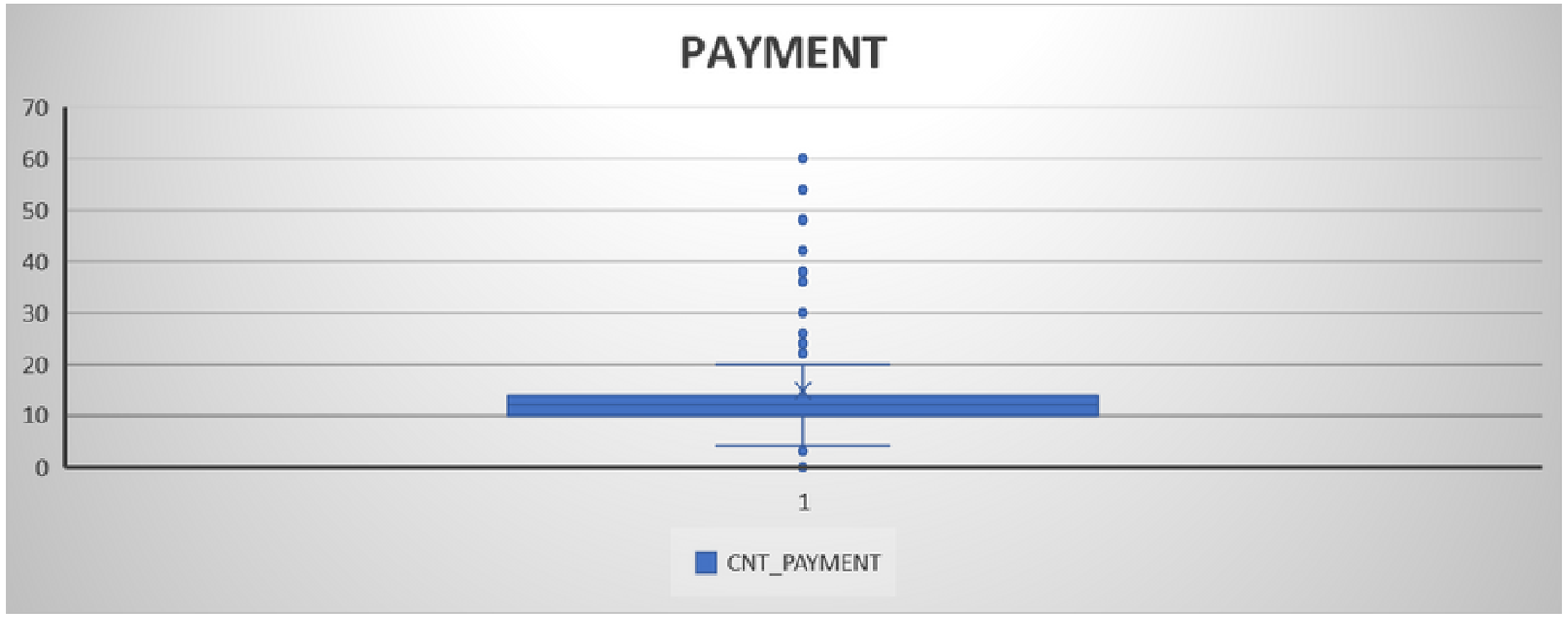


FROM THE GRAPH WE CAN INFER THAT MAXIMUM CLIENTS BELONGS TO UNACCOMPANIED TYPE_SUITE

REPLACING BLANKS WITH 'UNACCOMPANIED'

PREVIOUS APPLICATION DATASET-ANALYSIS

CNT_PAYMENT



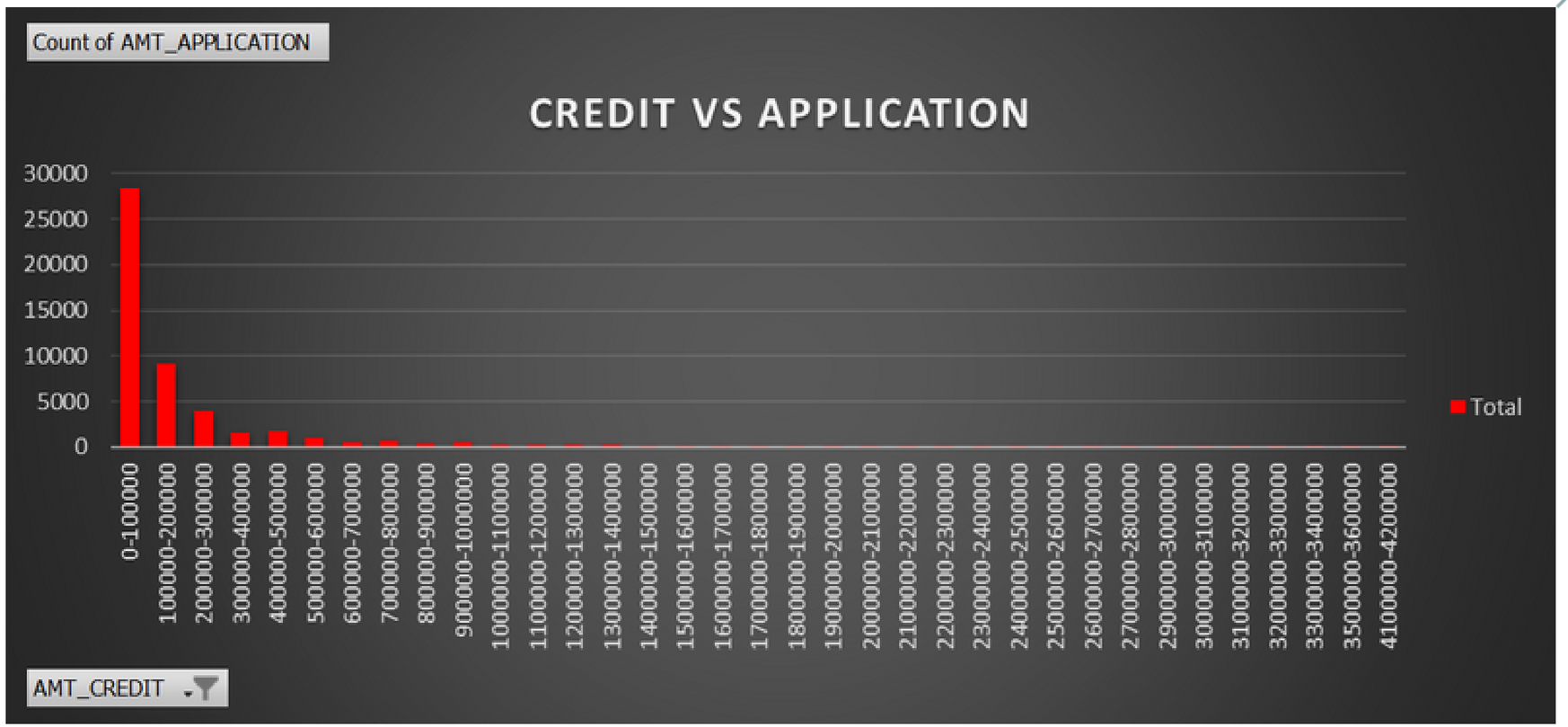
MEDIAN	12

SINCE THE DATA CONSIST OUTLIERS HENCE
WE WILL USE MEDIAN TO DETERMINE MISSING
VALUES

PREVIOUS APPLICATION DATASET-ANALYSIS

AMT_CREDIT VS AMT_APPLICATION

CREDIT RANGE	Count of AMT_APPLICATION
0-100000	28396
100000-200000	9156
200000-300000	4020
300000-400000	1656
400000-500000	1717
500000-600000	1073
600000-700000	615
700000-800000	732
800000-900000	372
900000-1000000	497
1000000-1100000	273
1100000-1200000	288
1200000-1300000	318
1300000-1400000	300
1400000-1500000	102
1500000-1600000	161
1600000-1700000	29
1700000-1800000	48
1800000-1900000	37
1900000-2000000	23
2000000-2100000	46
2100000-2200000	15
2200000-2300000	60
2300000-2400000	9
2400000-2500000	7
2500000-2600000	30
2600000-2700000	1
2700000-2800000	3
2900000-3000000	2
3000000-3100000	2
3100000-3200000	5
3200000-3300000	1
3300000-3400000	2
3500000-3600000	2
4100000-4200000	1
Grand Total	**49999**



FROM THE GRAPH WE CAN INFER THAT MOST OF THE PREVIOUS APPLICATIONS WERE SUBMMITTED FOR CREDIT IN THE RANGE 0-100K

PREVIOUS APPLICATION DATASET-ANALYSIS

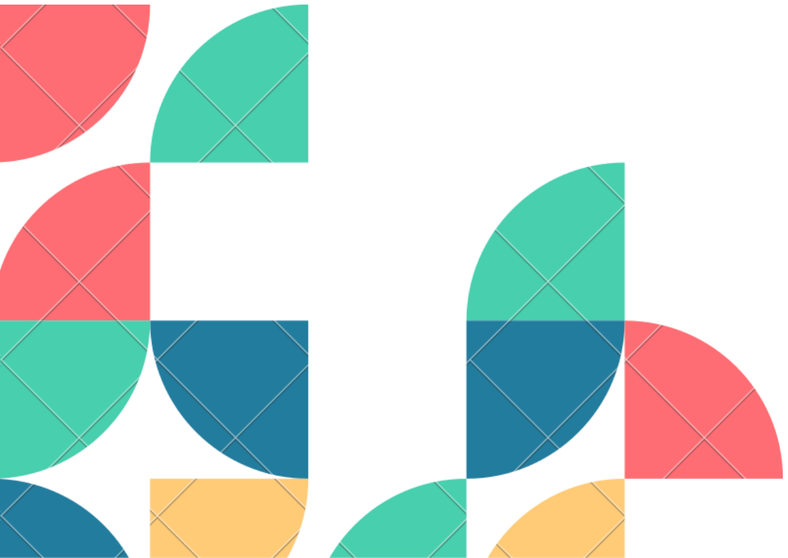
NAME_CONTRACT_STATUS BASED ON LOAN_PURPOSE

PURPOSE_OF_LOAN	Approved	Canceled	Refused	Unused offer
Building a house or an annex	29	2	51	
Business development	1		8	
Buying a garage	2		2	
Buying a holiday home / land	6		16	
Buying a home	6	3	20	
Buying a new car	5		26	
Buying a used car	25	2	56	
Car repairs	16		12	
Education	21		28	
Everyday expenses	24	1	26	
Furniture	11	1	18	
Gasification / water supply	1		10	
Hobby	2		1	
Journey	18	1	21	
Medicine	31		43	
Other	176	3	249	4
Payments on other loans	7	3	53	
Purchase of electronic equipment	13	1	13	1
Refusal to name the goal			1	
Repairs	232	16	431	2
Urgent needs	107	3	150	
Wedding / gift / holiday	10	1	13	
XAP	22986	1396	3919	842
XNA	8156	7162	3493	10

RESULT

CONCLUSION

- The Bank Generally lends more loan to Female Customers as compared to Males as the % of Female Defaulters is less compared to Males. still Bank can look for more Male Cllients
- Clients who fall in the Age Group 31-40 have the highest count for paying off their loans on time.So Bank can Target these agr groups
- Clients having LOW credit amount range tend to pay off their loans on time than compared to HIGH and MEDIUM credit rang
- The percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92 %
- Clients having Education status like Secondary/ Higher Secondary or more tend to pay loan on time so bank can prefer lending loans to clients having such Education Status



- Clients living with their Parents tend to pay off their loans quickly as compared to other housing type. So Bank can lend loan to clients having housing type → Living with Parent
- The Bank should be more cautious when lending money to clients with Repairs purpose because they have high count of Defaulters along with High count of Defaulters

THANK YOU