Name : ADVAIT GURUNATH CHAVAN
Email : advaitchavan135@gmail.com

Task 6: Bank Loan Case Study (Final Project - 2),

Tech Stack Used: Microsoft Excel

Analysis done on the following points:

To identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected.

Analysis is being done into two parts or say two dataset wiz:

1. Application data
2. Previous application data

The cleaned and analyzed data in the form of excel sheets have been uploaded to Google Drive also the excel sheets are large files due to vastness of data, so they won't be visible on google excel sheets online they need to be downloaded and seen offline using Microsoft Excel 2019

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
And those columns with less than 50% of the null data have to be replaced with mean or median or the highest occurring categorical variables

ALL THE COLUMN NAME WHICH ARE HIGHLIGHTED IN BLUE NEED TO BE DROPPED DOWN AS THEY HAVE NULL VALUES GREATER THAN OR EQUAL TO 50%

Column name	Total number of null values	Percentage of null value in that column	ROUND PER
OWN_CAR_AGE	202930	65.99113528	66
EXT_SOURCE_1	173379	56.38139774	56
APARTMENTS_AVG	156061	50.74972928	51
BASEMENTAREA_AVG	179943	58.51595553	59
YEARS_BUILD_AVG	204488	66.49778382	66
COMMON_AREA_AVG	214865	69.87229725	70
ELEVATORS_AVG	163891	53.29597966	53
ENTRANCES_AVG	154828	49.70488861	50
FLOORSMAX_AVG	153021	49.76114676	50
FLOORSMIN_AVG	208642	67.84862981	68
LANDAREA_AVG	182590	59.37673774	59
LIVINGAPARTMENTS_AVG	210199	68.35495316	68
LIVINGAREA_AVG	154350	50.19332642	50
NONLIVINGAPARTMENTS_AVG	213514	69.43296337	69
NONLIVINGAREA_AVG	169682	55.17916432	55
APARTMENTS_MODE	156061	50.74972928	51
BASEMENTAREA_MODE	179943	58.51595553	59
YEARS_BUILD_MODE	204488	66.49778382	66
COMMON_AREA_MODE	214865	69.87229725	70
ELEVATORS_MODE	163891	53.29597966	53
ENTRANCES_MODE	154828	50.34876801	50
FLOORSMAX_MODE	153020	49.76082156	50
FLOORSMIN_MODE	208642	67.84862981	68
LANDAREA_MODE	182590	59.37673774	59
LIVINGAPARTMENTS_MODE	210199	68.35495316	68
LIVINGAREA_MODE	154350	50.19332642	50
NONLIVINGAPARTMENTS_MODE	213514	69.43296337	69
NONLIVINGAREA_MODE	169682	55.17916432	55
APARTMENTS_MEDIAN	156061	50.74972928	51
BASEMENTAREA_MEDIAN	179943	58.51595553	59
YEARS_BUILD_MEDIAN	204488	66.49778382	66
COMMON_AREA_MEDIAN	214865	69.87229725	70
ELEVATORS_MEDIAN	163891	53.29597966	53
ENTRANCES_MEDIAN	154828	50.34876801	50
FLOORSMAX_MEDIAN	153020	49.76082156	50
FLOORSMIN_MEDIAN	208642	67.84862981	68
LANDAREA_MEDIAN	182590	59.37673774	59
LIVINGAPARTMENTS_MEDIAN	210199	68.35495316	68
LIVINGAREA_MEDIAN	154350	50.19332642	50
NONLIVINGAPARTMENTS_MEDIAN	213514	69.43296337	69
NONLIVINGAREA_MEDIAN	169682	55.17916432	55
FONDKAPREMONT_MODE	210295	68.38617155	68
HOUSETYPE_MODE	154297	50.17609126	50
WALLSMATERIAL_MODE	156341	50.84078293	51

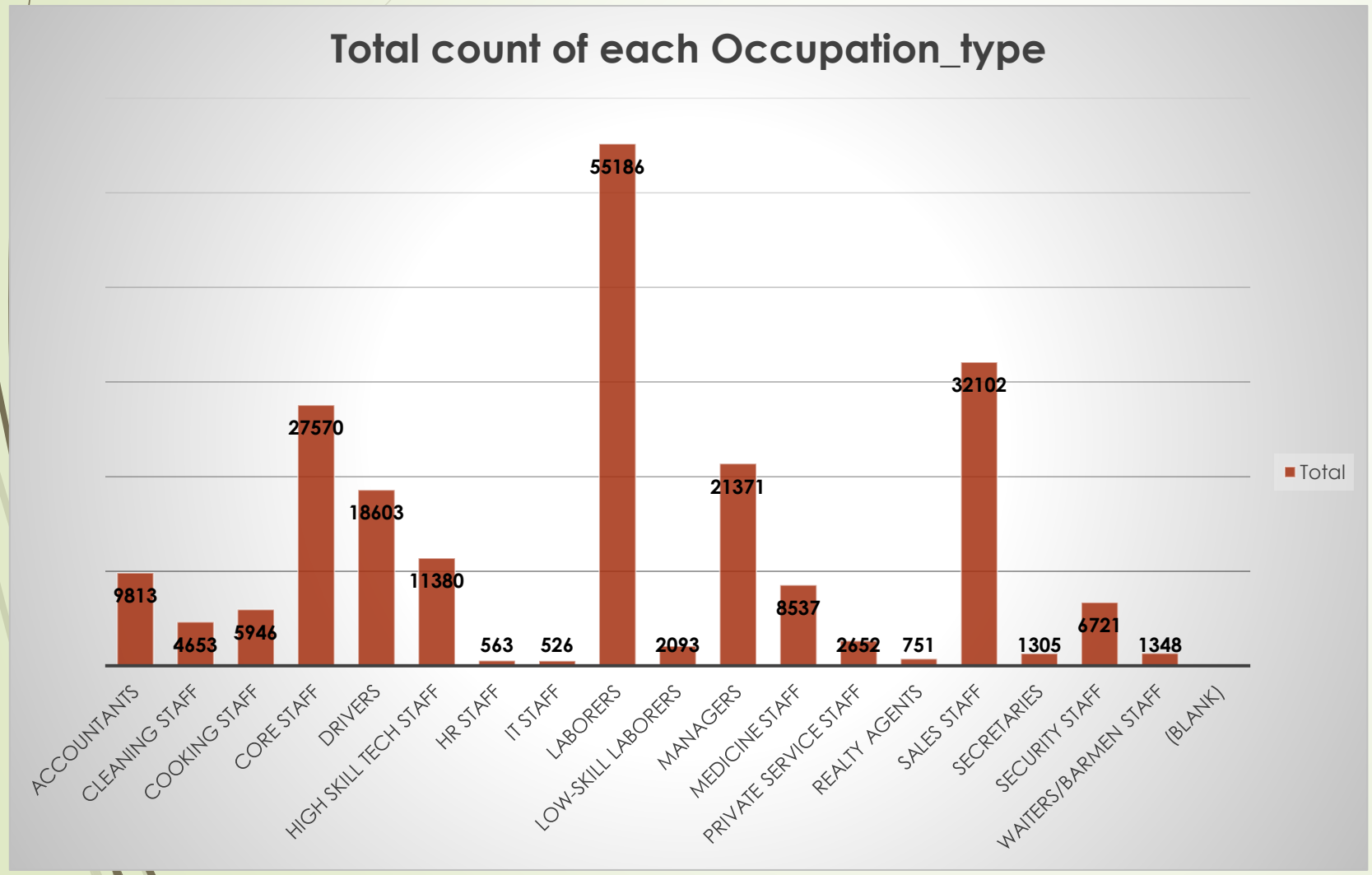
Application Dataset – NULL values

ALL THE COLUMN NAME WHICH ARE HIGHLIGHTED IN GREEN NEED TO BE DROPPED DOWN AS THEY ARE IRRELEVANT COLUMNS FOR DOING OUR ANALYSIS

Column name	Total number of null values	Percentage of null value in that column	ROUND PER
FLAG_MOBIL	1	0.000325192	0
FLAG_EMPLOY_PHONE	55387	18.01138821	18
FLAG_WORK_PHONE	0	0	0
FLAG_CONT_MOBILE	0	0	0
FLAG_PHONE	0	0	0
FLAG_EMAIL	0	0	0
CNT_FAMILY_MEMBERS	2	0.000650383	0
REGION_RATING_CLENT	0	0	0
REGION_RATING_CLENT_W_CITY	0	0	0
EXT_SOURCE_3	60965	19.82530706	20
YEAR_BEGINEXPLUATATION_AVG	150008	48.78134441	49
YEAR_BEGINEXPLUATATION_MODE	150007	48.78101922	49
YEAR_BEGINEXPLUATATION_MEDIAN	150007	48.78101922	49
TOTAL_AREA_MODE	148431	48.26851722	48
EMERGENCYSTATE_MODE	145755	47.39830445	47
DAYS_LAST_PHONE_CHANGE	1	0.000325192	0
FLAG DOC 2	0	0	0
FLAG DOC 3	0	0	0
FLAG DOC 4	0	0	0
FLAG DOC 5	0	0	0
FLAG DOC 6	0	0	0
FLAG DOC 7	0	0	0
FLAG DOC 8	0	0	0
FLAG DOC 9	0	0	0
FLAG DOC 10	0	0	0
FLAG DOC 11	0	0	0
FLAG DOC 12	0	0	0
FLAG DOC 13	0	0	0
FLAG DOC 14	0	0	0
FLAG DOC 15	0	0	0
FLAG DOC 16	0	0	0
FLAG DOC 17	0	0	0
FLAG DOC 18	0	0	0
FLAG DOC 19	0	0	0
FLAG DOC 20	0	0	0
FLAG DOC 21	0	0	0

Application Dataset – NULL values

Replacing Blanks in Occupation_Type column of the Application Dataset with the highest occurring categorical variable

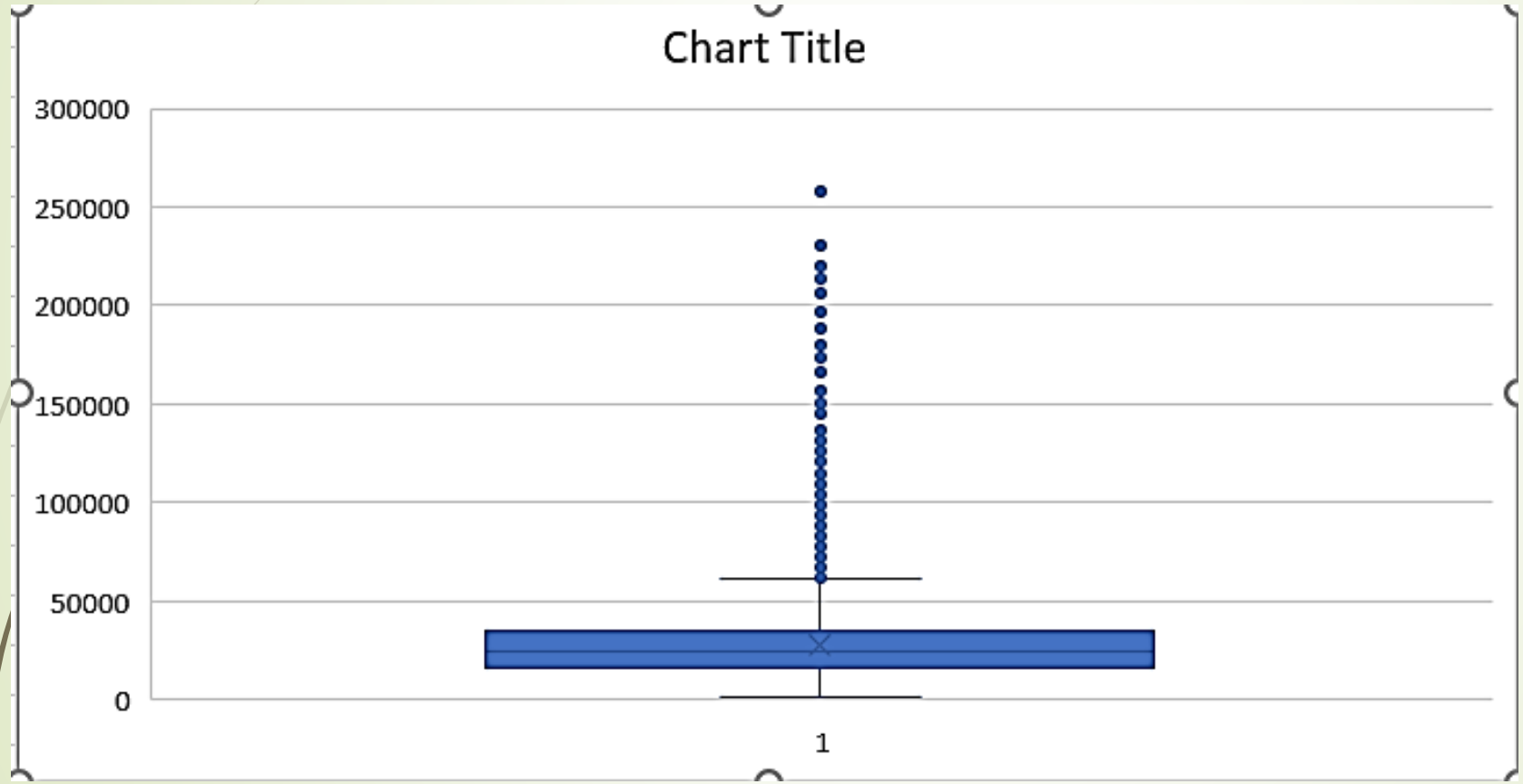


Row Labels	Count of OCCUPATION_TYPE
Accountants	9813
Cleaning staff	4653
Cooking staff	5946
Core staff	27570
Drivers	18603
High skill tech staff	11380
HR staff	563
IT staff	526
Laborers	55186
Low-skill Laborers	2093
Managers	21371
Medicine staff	8537
Private service staff	2652
Realty agents	751
Sales staff	32102
Secretaries	1305
Security staff	6721
Waiters/barmen staff	1348
(blank)	
Grand Total	211120

Highest occurring categorical variable is **'Laborers'**

Application Dataset – NULL values

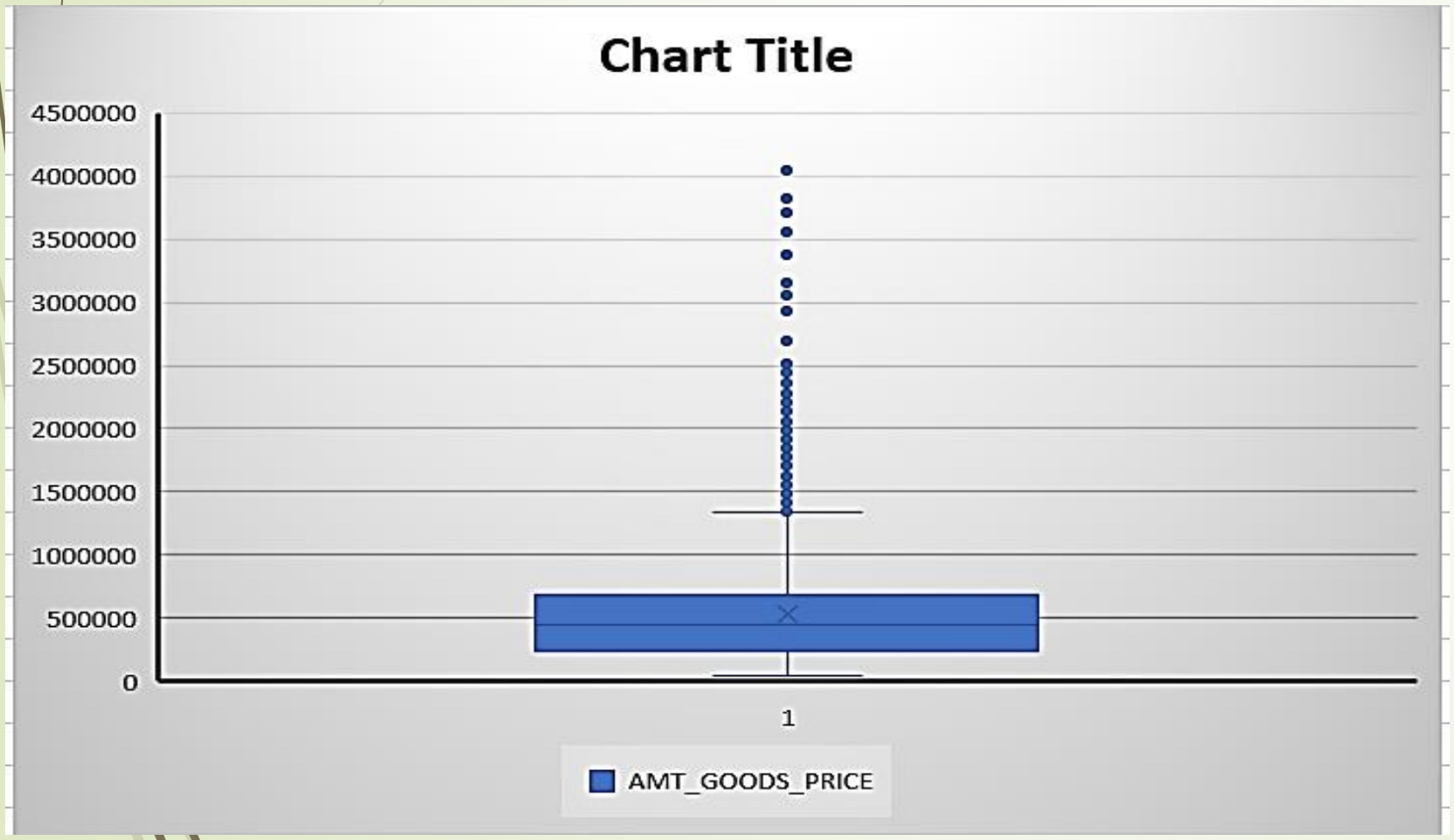
Replacing Blanks in AMT_ANNUTIY column of the Application Dataset with the median of the AMT_ANNUITY as there exists outliers in the AMT_ANNUITY column



Median of AMT_ANNUITY
24903
Replacing Blanks with Median

Application Dataset – NULL values

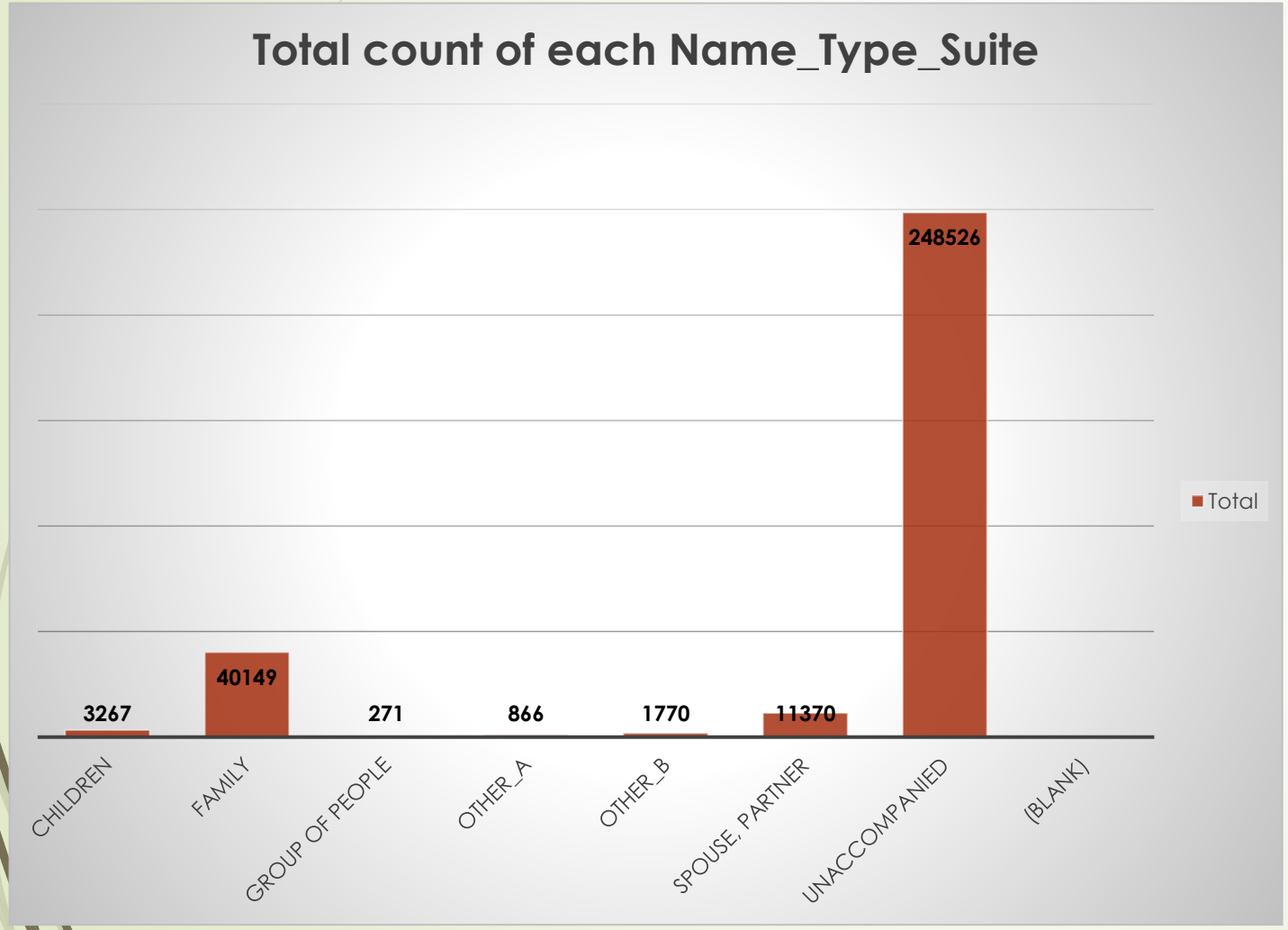
Replacing Blanks in AMT_GOODS_PRICE column of the Application Dataset with the median of the AMT_GOODS_PRICE as there exists outliers in the AMT_GOODS_PRICE column



Median of AMT_GOODS_PRICE
450000
Replacing Blanks with Median

Application Dataset – NULL values

Replacing Blanks in Name_Type_Suite column of the Application Dataset with the highest occurring categorical variable



Row Labels	Count of NAME_TYPE_SUITE
Children	3267
Family	40149
Group of people	271
Other_A	866
Other_B	1770
Spouse, partner	11370
Unaccompanied	248526
(blank)	
Grand Total	306219

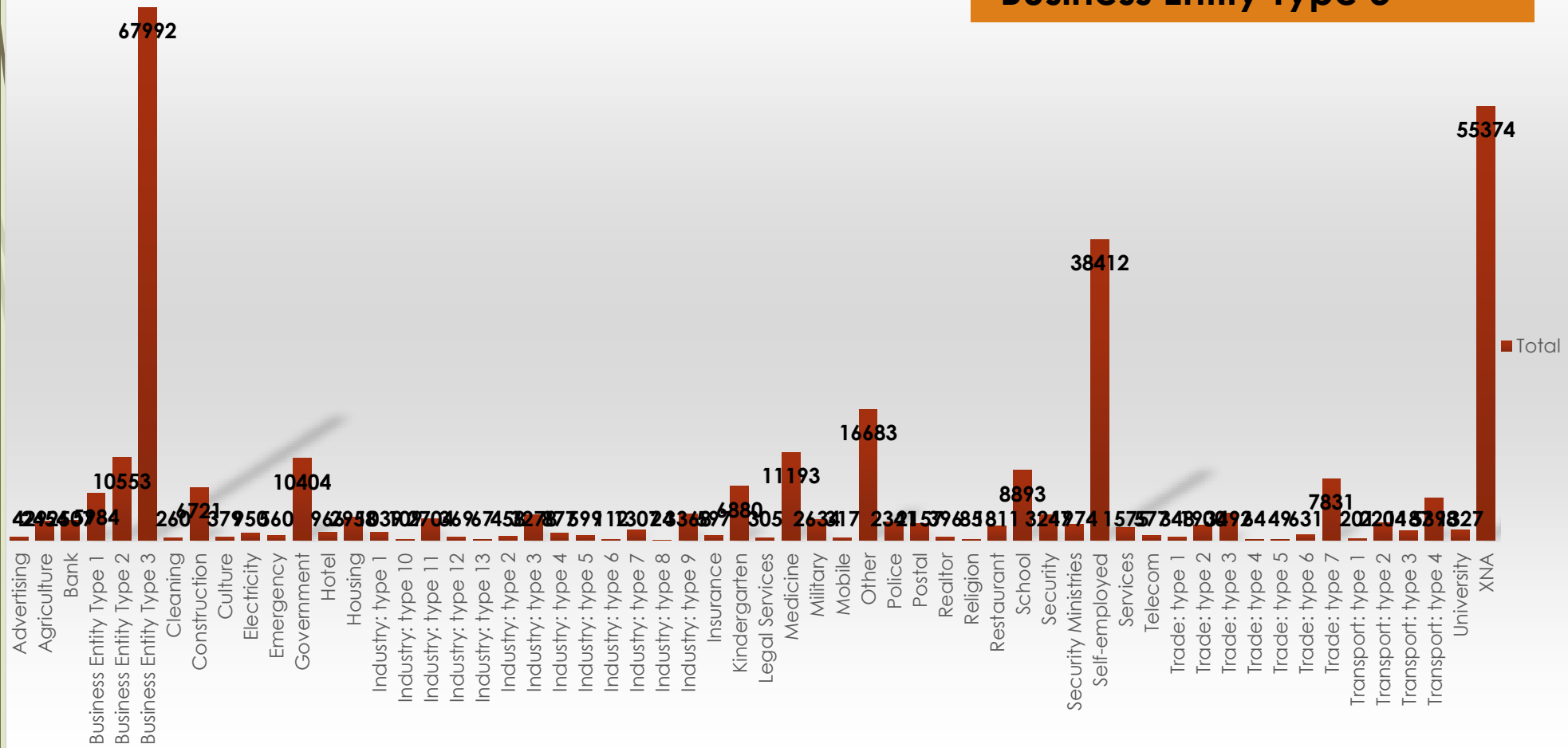
Highest occurring categorical variable is **'Unaccompanied'**

Application Dataset – NULL values

Replacing Blanks in Organization_type column of the Application Dataset with the highest occurring categorical variable

Total count of each Organization_type

Highest occurring categorical variable is **'Business Entity Type 3'**



Row Labels	Count of ORGANIZATION_TYPE
Advertising	429
Agriculture	2454
Bank	2507
Business Entity Type 1	5984
Business Entity Type 2	10553
Business Entity Type 3	67992
Cleaning	260
Construction	6721
Culture	379
Electricity	950
Emergency	560
Government	10404
Hotel	966
Housing	2958
Industry: type 1	1039
Industry: type 10	109
Industry: type 11	2704
Industry: type 12	369
Industry: type 13	67
Industry: type 2	458
Industry: type 3	3278
Industry: type 4	877
Industry: type 5	599
Industry: type 6	112
Industry: type 7	1307
Industry: type 8	24
Industry: type 9	3368
Insurance	597
Kindergarten	6880
Legal Services	305
Medicine	11193
Military	2634
Mobile	317
Other	16683
Police	2341
Postal	2157
Realtor	396
Religion	85
Restaurant	1811
School	8893
Security	3247
Security Ministries	1974
Self-employed	38412
Services	1575
Telecom	577
Trade: type 1	348
Trade: type 2	1900
Trade: type 3	3492
Trade: type 4	64
Trade: type 5	49
Trade: type 6	631
Trade: type 7	7831
Transport: type 1	201
Transport: type 2	2204
Transport: type 3	1187
Transport: type 4	5398
University	1327
XNA	55374
Grand Total	307511

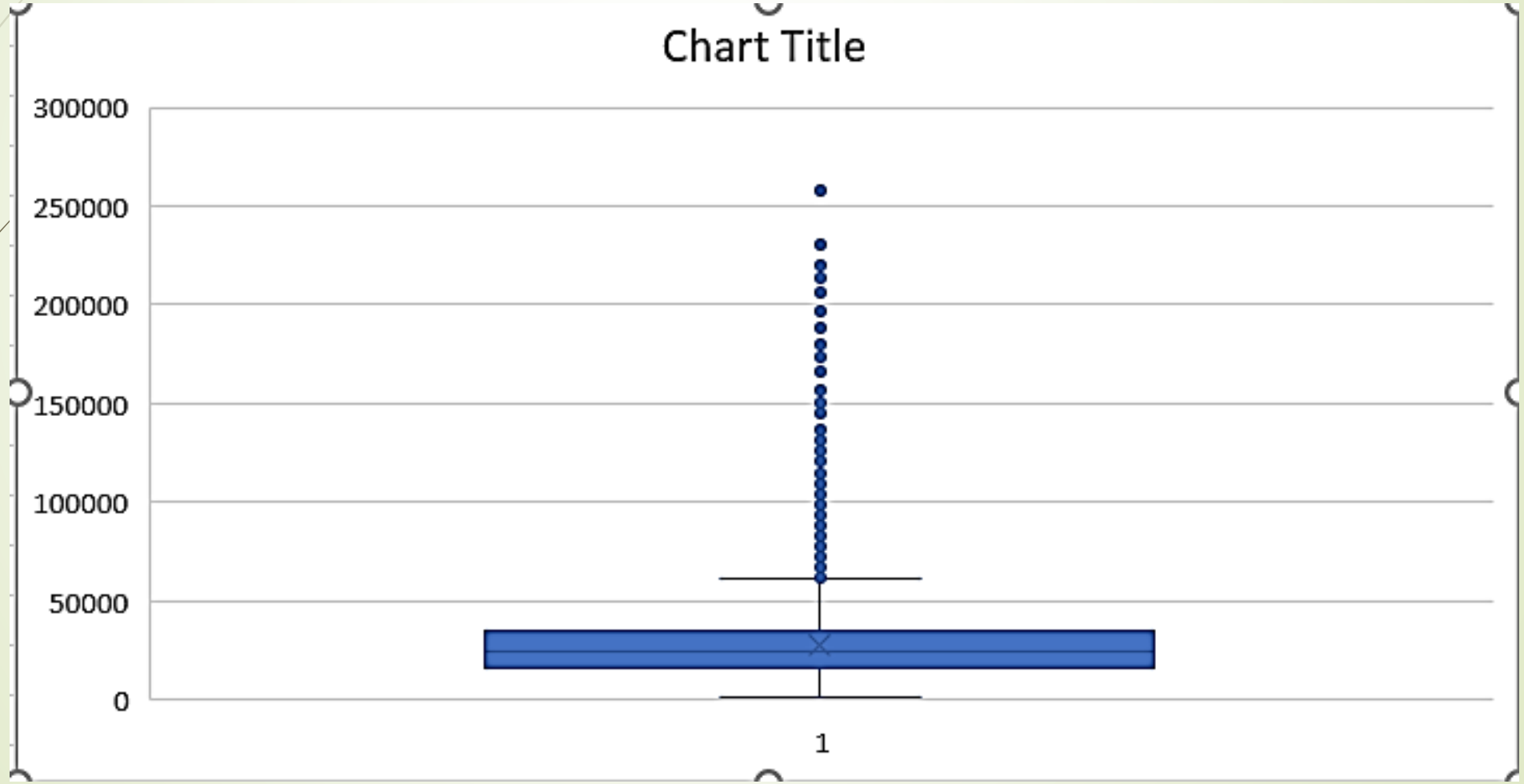
Application Dataset – NULL values

Google Drive Link for Excel sheet of Analysis of Null values done:-

application_data.xlsx - Google Drive

Application Dataset – Outliers

First outlier is in AMT_ANNUITY
which is greater than 250000
this outlier is replaced with 24903
median of AMT_ANNUITY

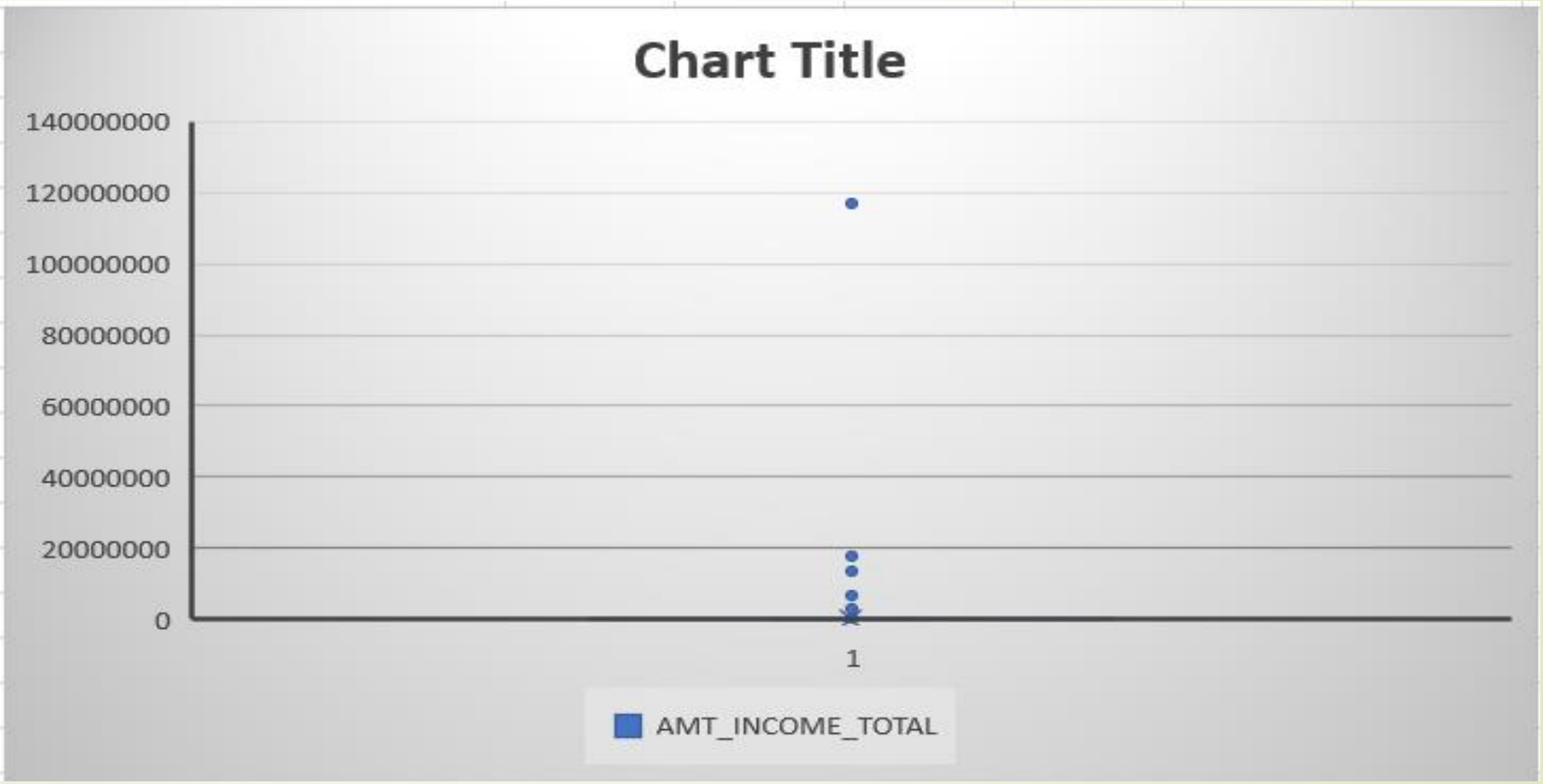


Application Dataset – Outliers

Here we can observe that there is huge difference between
the 25%, 50% and 75% quartile and this is due to presence of outliers
But since the amount of total income varies from person to person
we will not remove the outliers

	Quartiles at AMT_INCOME_TOTAL
MIN	25650
25%	112500
50%	147150
75%	202500
MAX	117000000

outliers at extreme points i.e. max 1.700x10^8



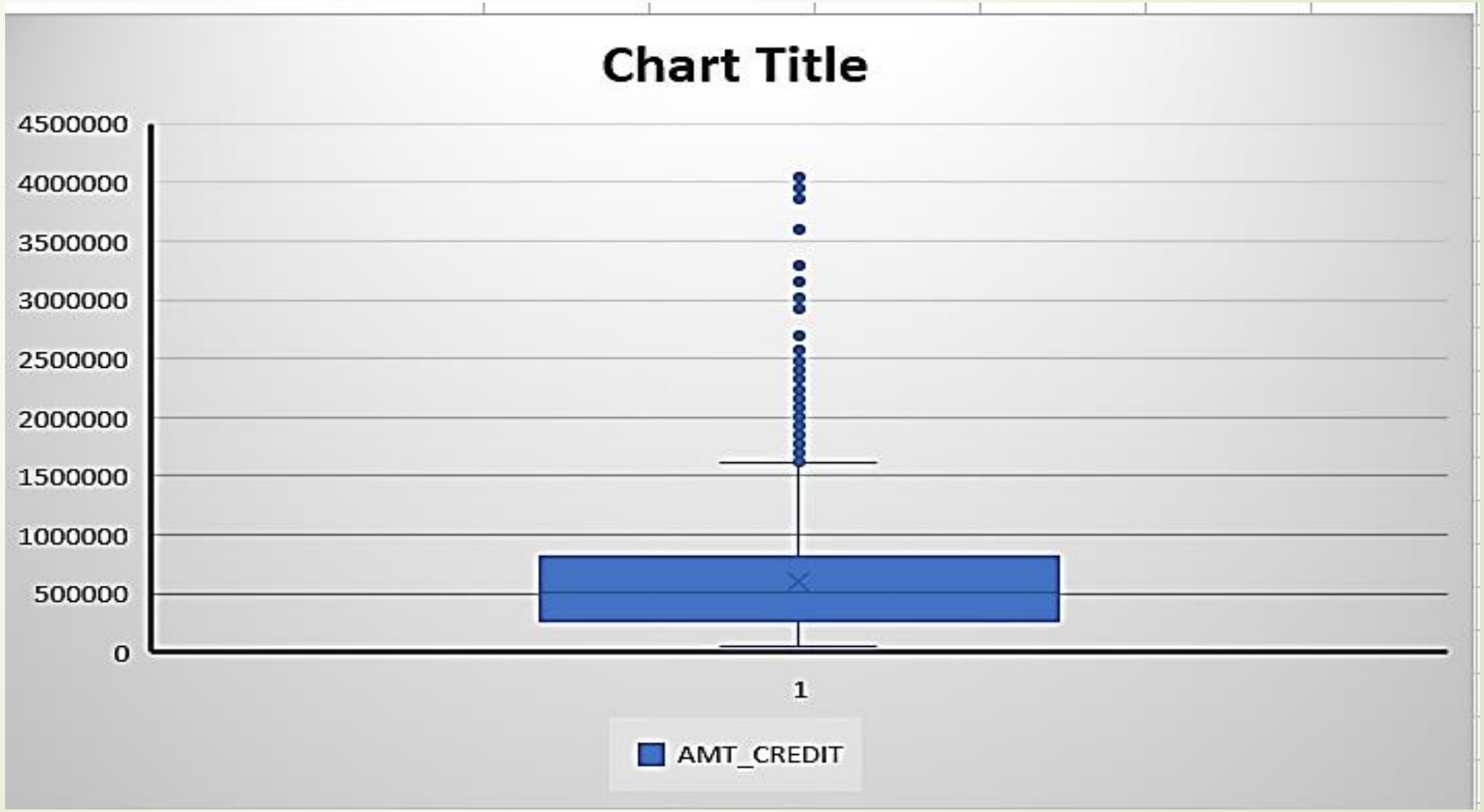
Application Dataset – Outliers

From the chart it is clear that outliers lie in the 98% and near max side of the box plot

Also there is a significant difference between the 75% quartile and the max value and this is due the presence of the outliers

But since the amount of credit varies from person to person we will not remove the outliers

	AMT_CREDIT
	Quartiles at AMT_CREDIT
MIN	45000
25%	270000
50%	513531
75%	808650
MAX	4050000

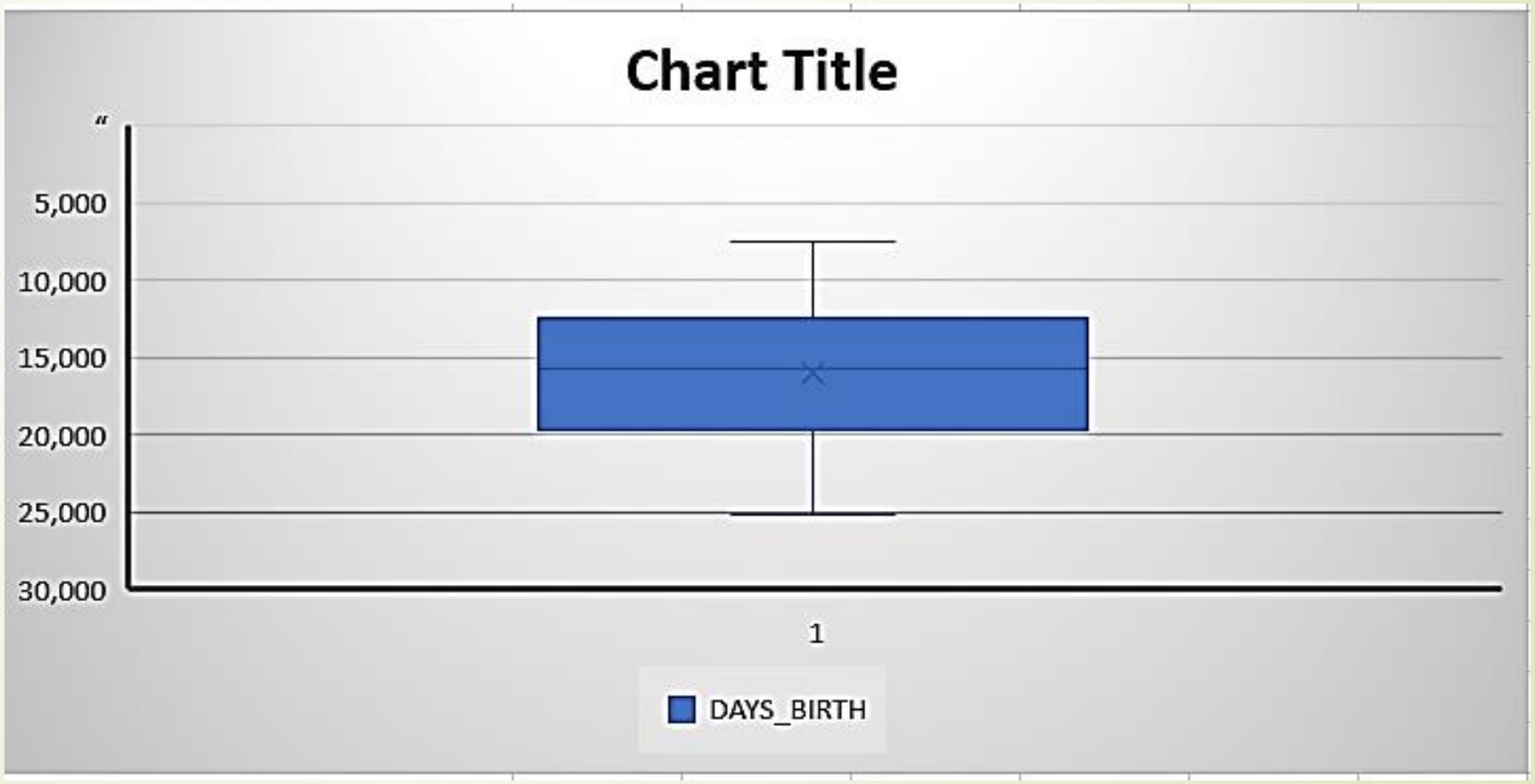


Application Dataset – Outliers

As seen from the boxplot it is clear that there are no outliers

The data of DAYS_BIRTH is well distributed

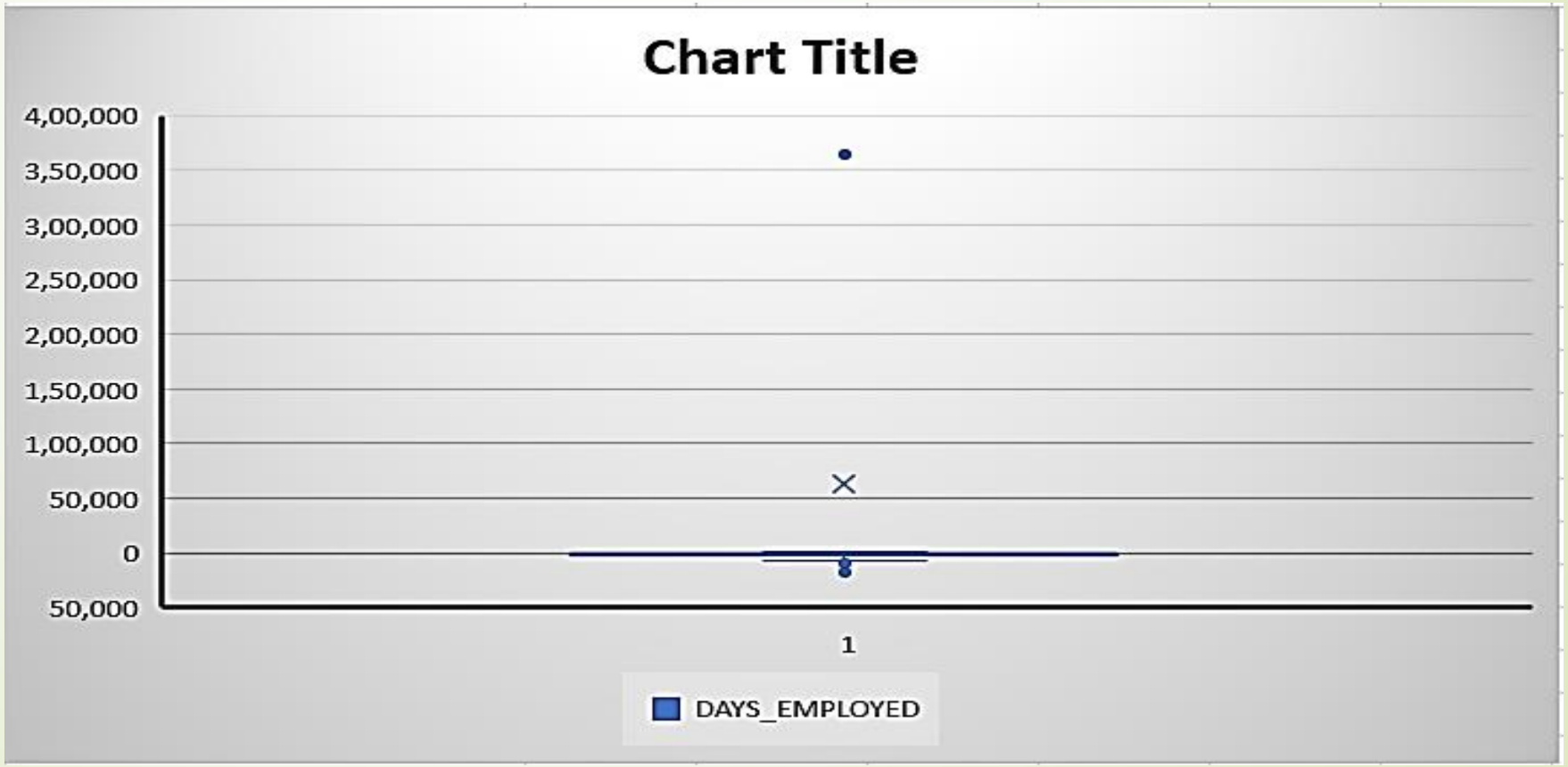
	DAYS_OF_BIRTH
	Quartiles at DAYS_BIRTH
MAX	25,229.00
75%	19,682.00
50%	15,750.00
25%	12,413.00
MIN	7,489.00



Application Dataset – Outliers

There exists only 1 outlier i.e. + or - 365243
Replace with median
1213.00

	DAYS_EMPLOYED
	Quartiles at DAYS_EMPLOYED
MAX	17912.00
75%	2760.00
50%	1213.00
25%	289.00
MIN	365243.00



Application Dataset – Outliers

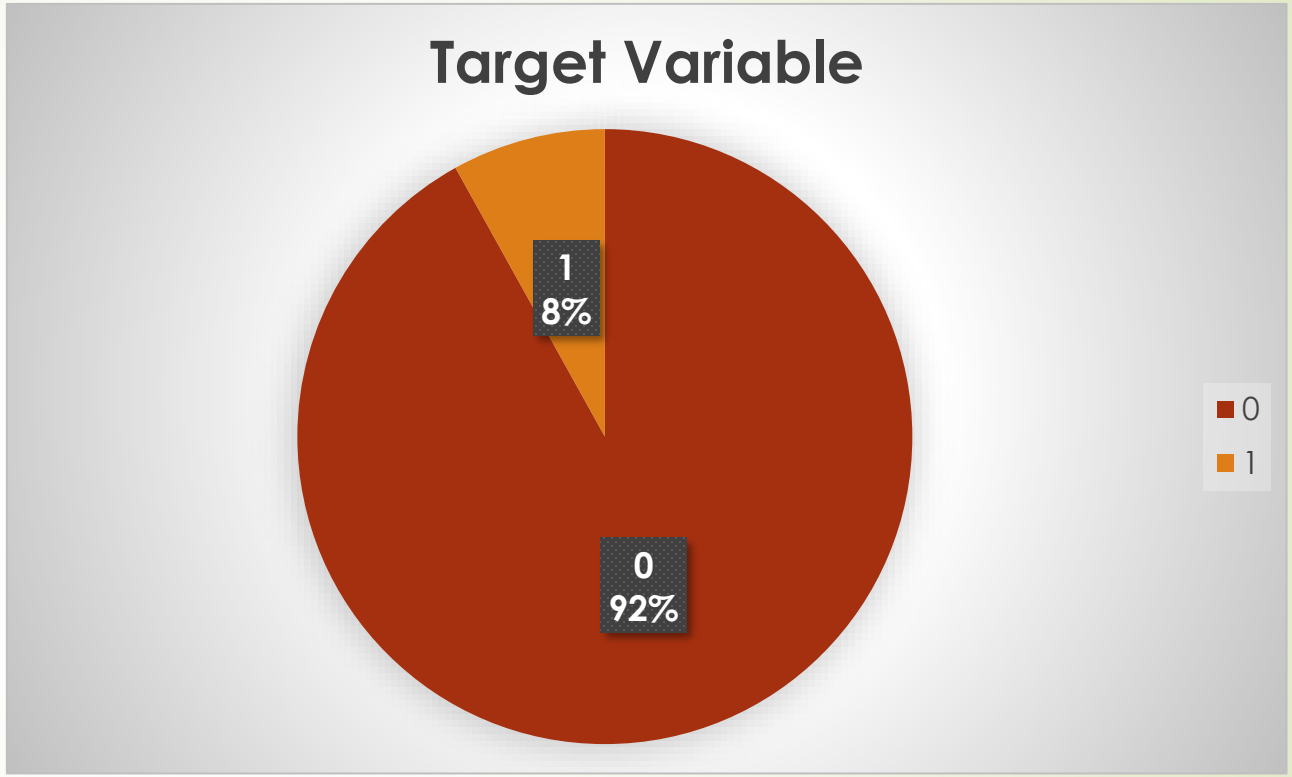
Google Drive Link for Excel sheet of Analysis of Outliers and cleaned Data done:-

application_data_cleaned.xlsx - Google Drive

Application Dataset – Analysis

TARGET VARIABLE

Row Labels	Count of TARGET
0	282686
1	24825
Grand Total	307511



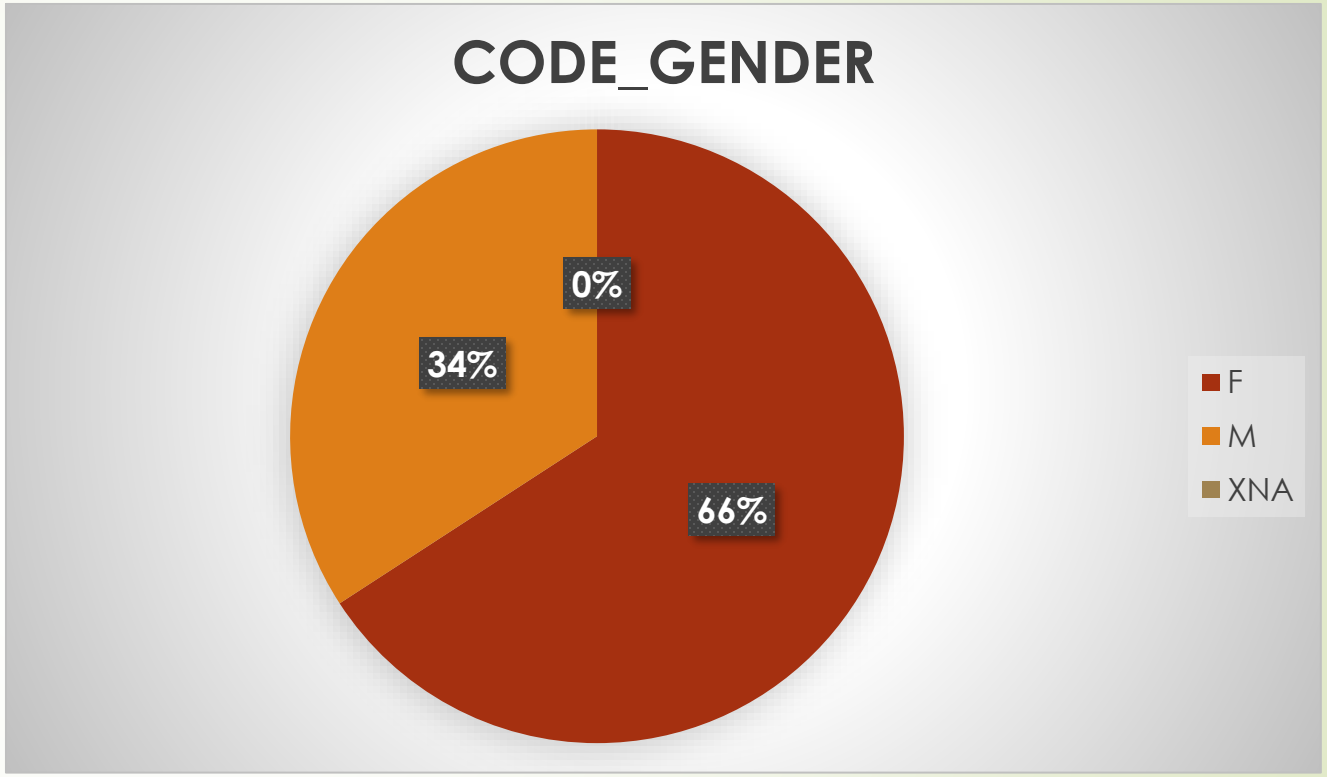
The Target Variable Pie chart shows that almost 92% of the total clients had no problem during payment while 8% of the clients had some or the other problem

0 → No payment issues
1 → Had some payment issues

Application Dataset – Analysis

GENDER VARIABLE

Row Labels	Count of CODE_GENDER
F	202448
M	105059
XNA	4
Grand Total	307511

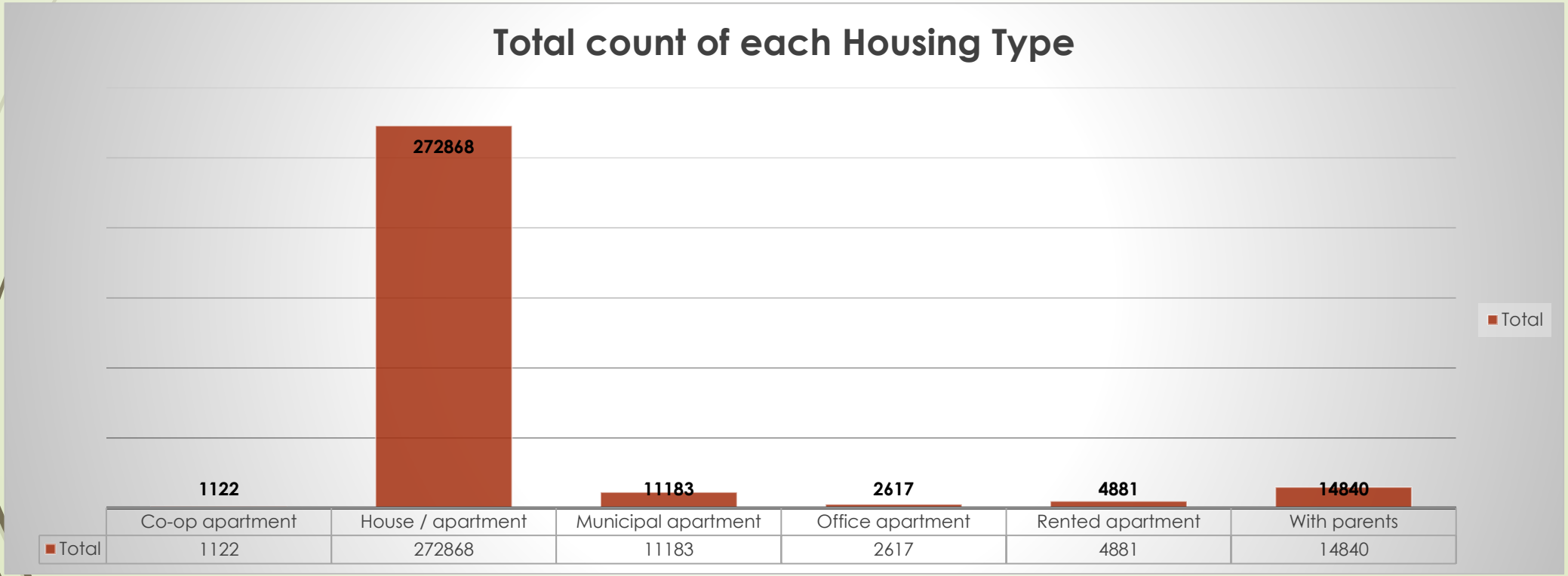


From the GENDER_VARIABLE pie chart
we can infer that almost 66% of
the clients are female and 34% of the
clients are Male
The 4 of the appicants have gender as XNA
which can be ignored

Application Dataset – Analysis

NAME_HOUSING_TYPE

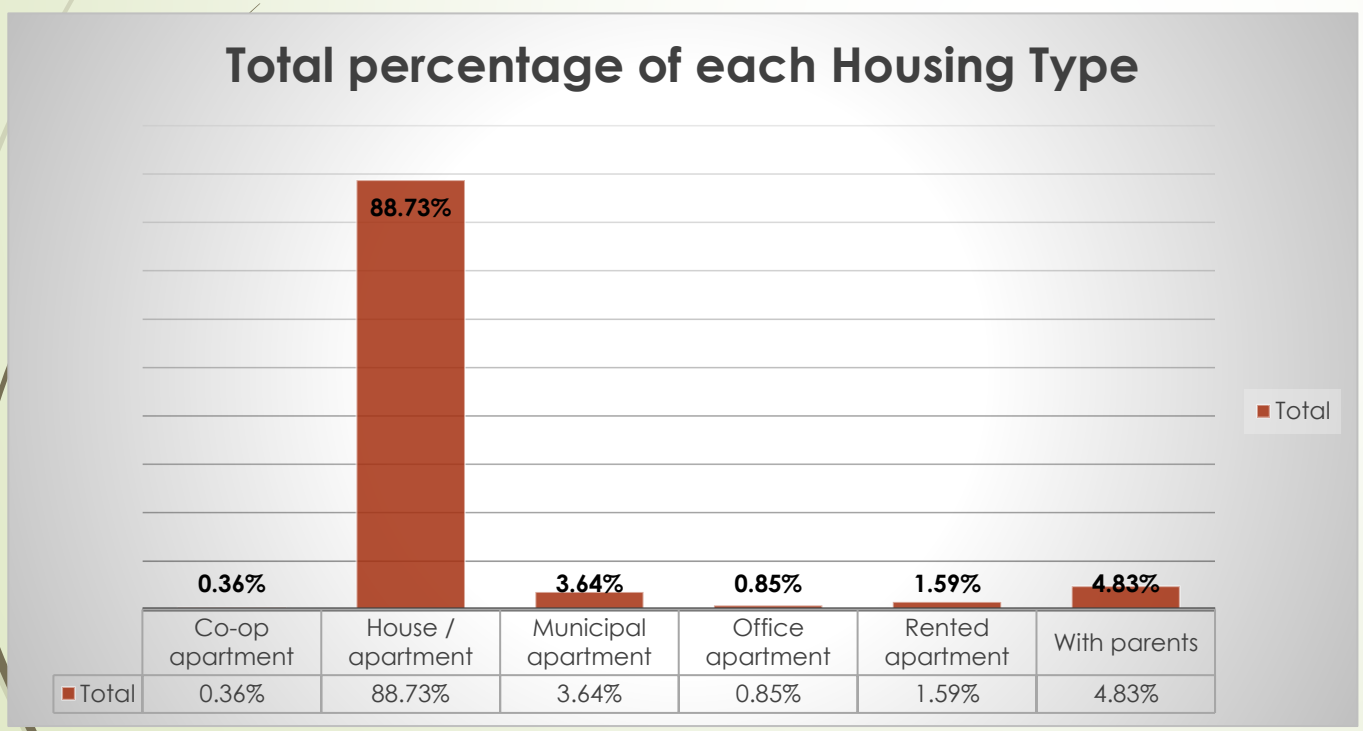
Row Labels	Count of NAME_HOUSING_TYPE
Co-op apartment	1122
House / apartment	272868
Municipal apartment	11183
Office apartment	2617
Rented apartment	4881
With parents	14840
Grand Total	307511



Application Dataset – Analysis

NAME_HOUSING_TYPE

Row Labels	Percentage of NAME_HOUSING_TYPE
Co-op apartment	0.36%
House / apartment	88.73%
Municipal apartment	3.64%
Office apartment	0.85%
Rented apartment	1.59%
With parents	4.83%
Grand Total	100.00%



From the bar graphs of count and percentage

The bank can target those groups who do not have their

own apartment i.e. the bank may consider the people

living in Co-op apartment, Municipal Apartment, Rented

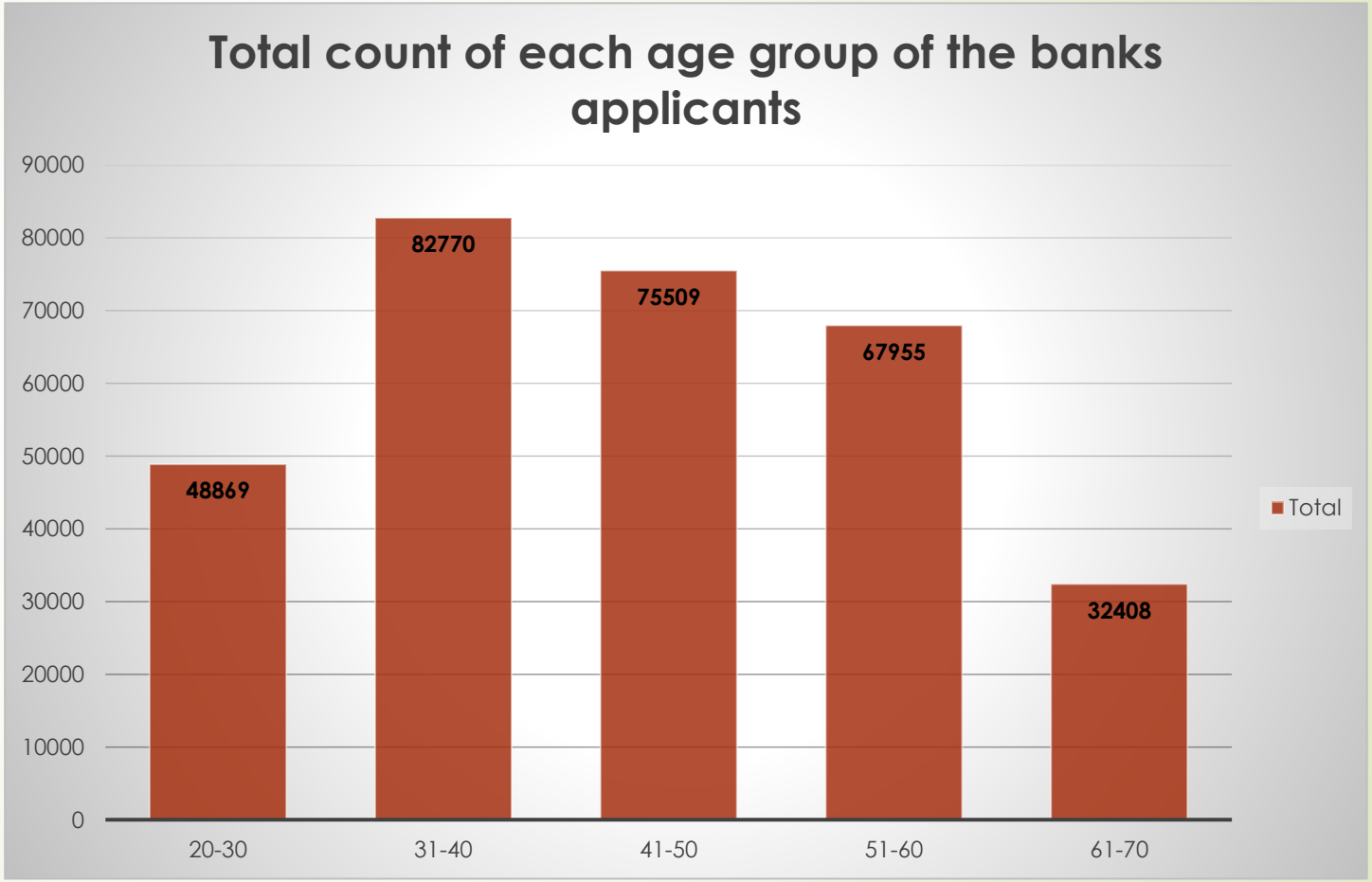
Apartment and people living with their parents

Application Dataset – Analysis

Univariate Analysis

AGE GROUP

Row Labels	Count of YEARS_BIRTH_RANGE
20-30	48869
31-40	82770
41-50	75509
51-60	67955
61-70	32408
Grand Total	307511



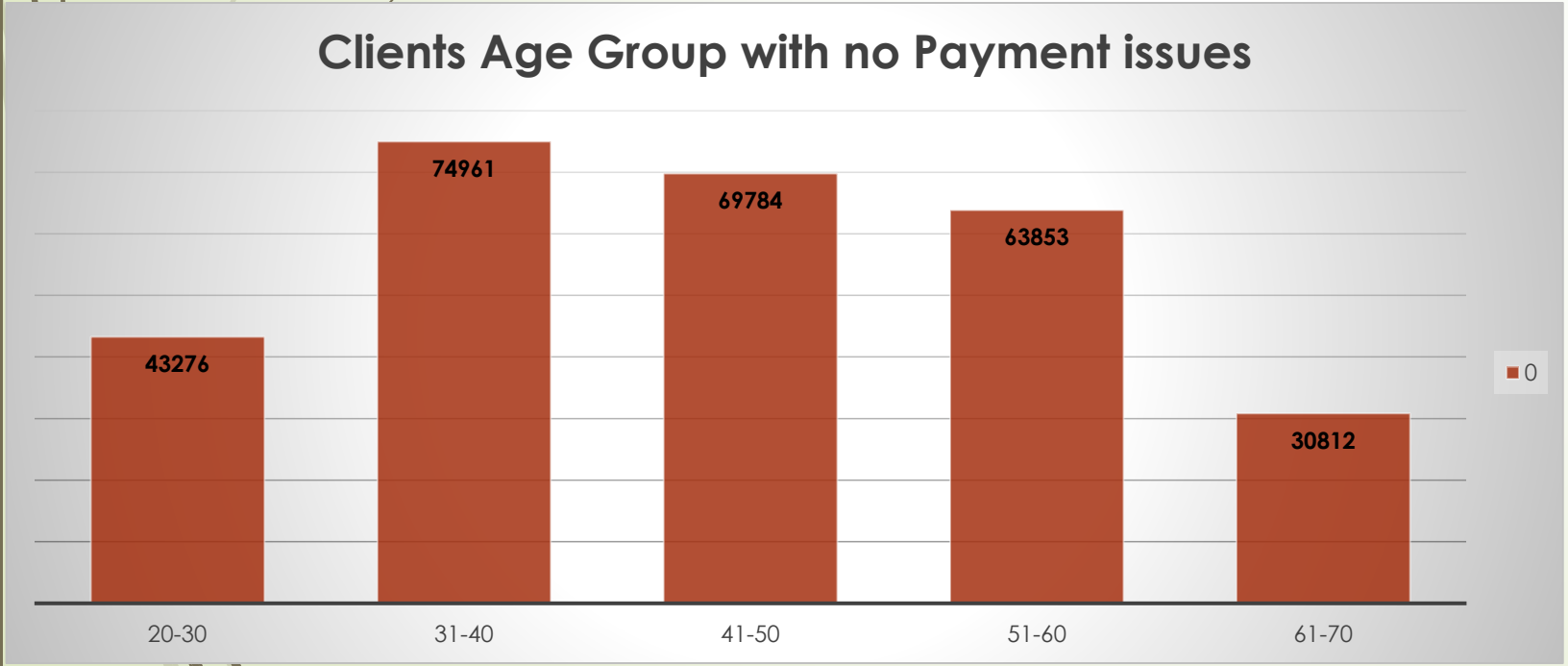
From the adjacent bar plot we can infer that most of the applicants belong to the Age Group '31-40'

Application Dataset – Analysis

Univariate Analysis

AGE GROUP

Count of TARGET	Column Labels	
Row Labels	0	Grand Total
20-30	43276	43276
31-40	74961	74961
41-50	69784	69784
51-60	63853	63853
61-70	30812	30812
Grand Total	282686	282686



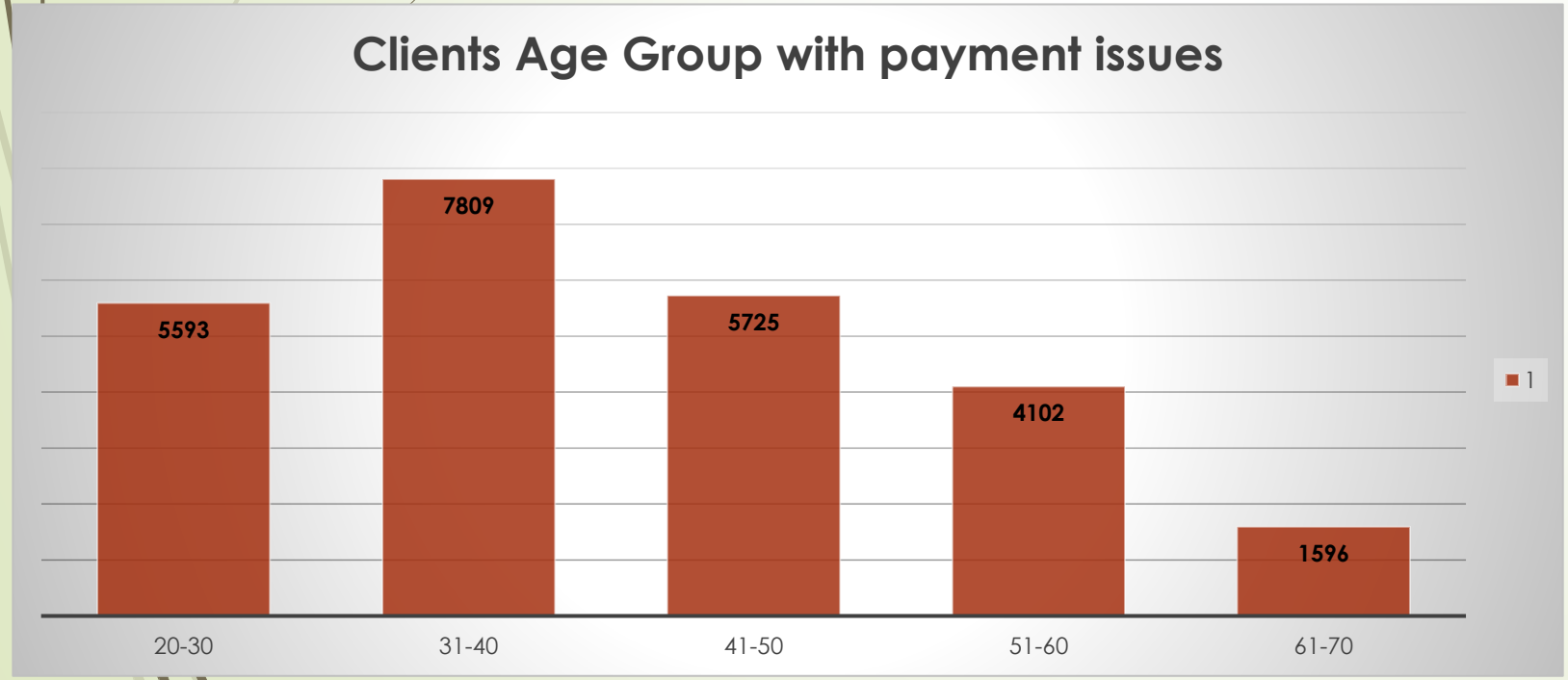
From the adjacent Bar plot we can infer that clients/applicants in the Age Group '31-40' are having the highest number when it comes to doing/returning Payment to Banks

Application Dataset – Analysis

Univariate Analysis

AGE GROUP

Count of TARGET	Column Labels	
Row Labels	1	Grand Total
20-30	5593	5593
31-40	7809	7809
41-50	5725	5725
51-60	4102	4102
61-70	1596	1596
Grand Total	24825	24825



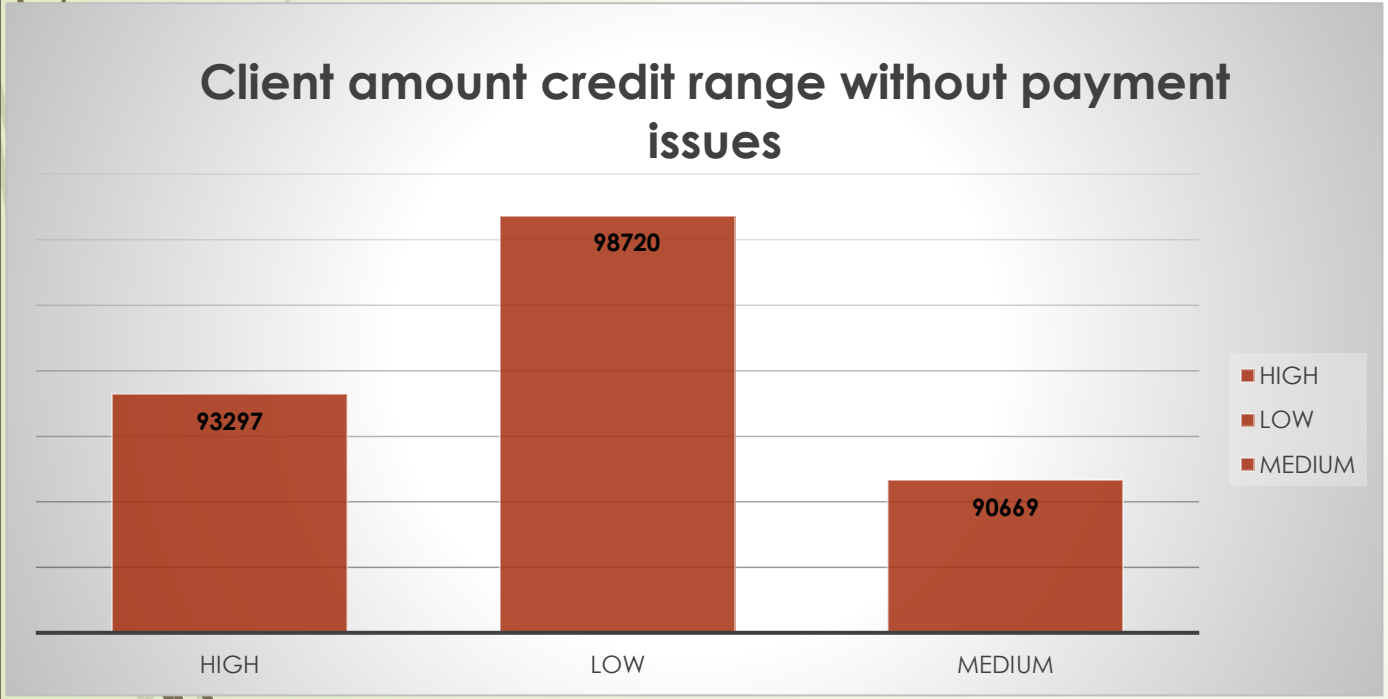
From the adjacent Bar plot we can infer that clients/applicants in the Age Group '31-40' are having the highest number of payment issues when it comes to doing/returning Payment to Banks

Application Dataset – Analysis

Univariate Analysis

Client amount credit range

Count of TARGET	Column Labels	
Row Labels	0	Grand Total
HIGH	93297	93297
LOW	98720	98720
MEDIUM	90669	90669
Grand Total	282686	282686



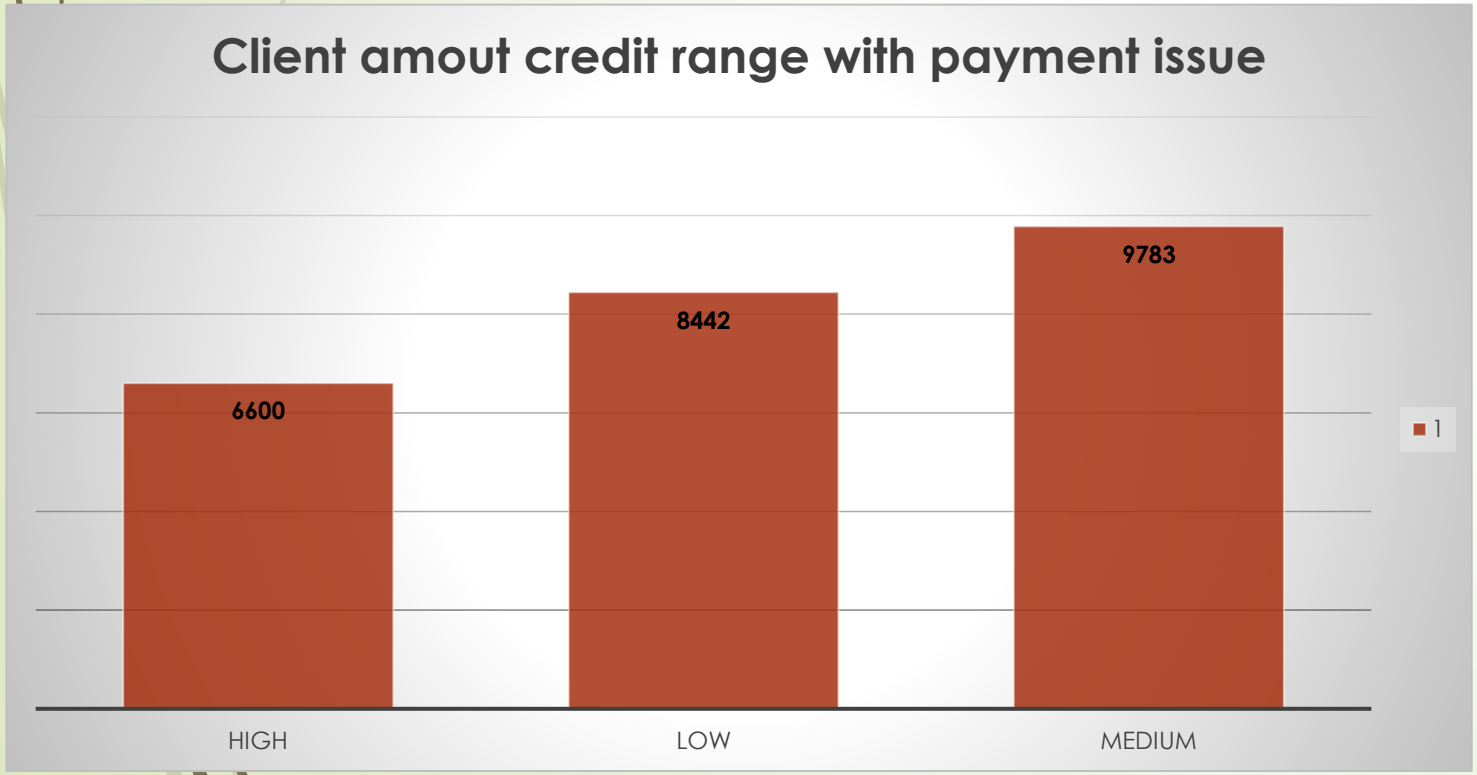
From the adjacent Bar plot we can infer that clients belonging to 'Low' income range have the highest count when it comes to clients with no payment issues

Application Dataset – Analysis

Univariate Analysis

Client amount credit range

Count of TARGET	Column Labels	
Row Labels	1	Grand Total
HIGH	6600	6600
LOW	8442	8442
MEDIUM	9783	9783
Grand Total	24825	24825



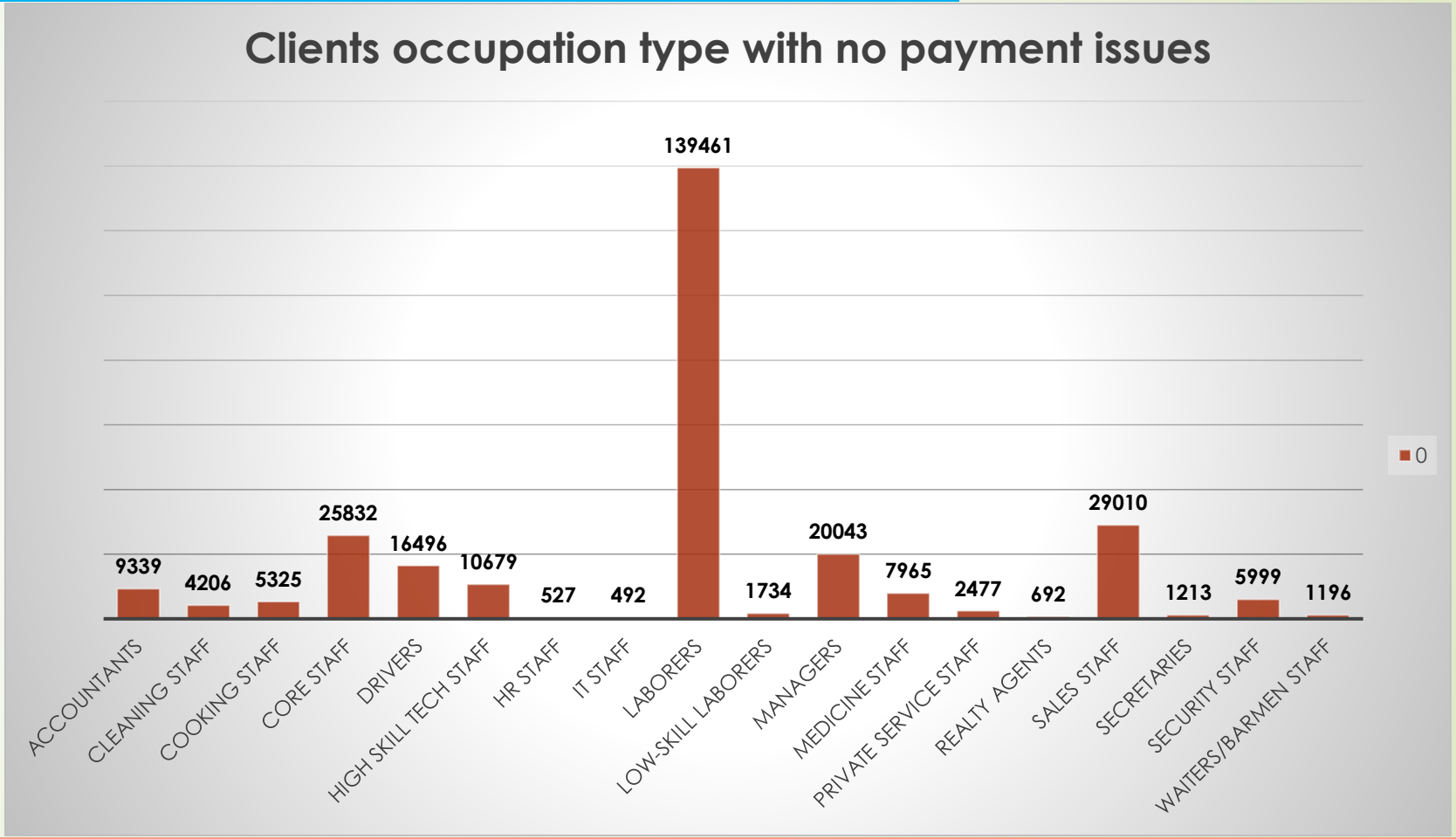
From the adjacent Bar plot we can infer that clients belonging to 'Medium' income range have the highest count when it comes to clients with payment issues

Application Dataset – Analysis

Univariate Analysis

OCCUPATION_TYPE

Count of TARGET	Column Labels	
Row Labels	0	Grand Total
Accountants	9339	9339
Cleaning staff	4206	4206
Cooking staff	5325	5325
Core staff	25832	25832
Drivers	16496	16496
High skill tech staff	10679	10679
HR staff	527	527
IT staff	492	492
Laborers	139461	139461
Low-skill Laborers	1734	1734
Managers	20043	20043
Medicine staff	7965	7965
Private service staff	2477	2477
Realty agents	692	692
Sales staff	29010	29010
Secretaries	1213	1213
Security staff	5999	5999
Waiters/barmen staff	1196	1196
Grand Total	282686	282686



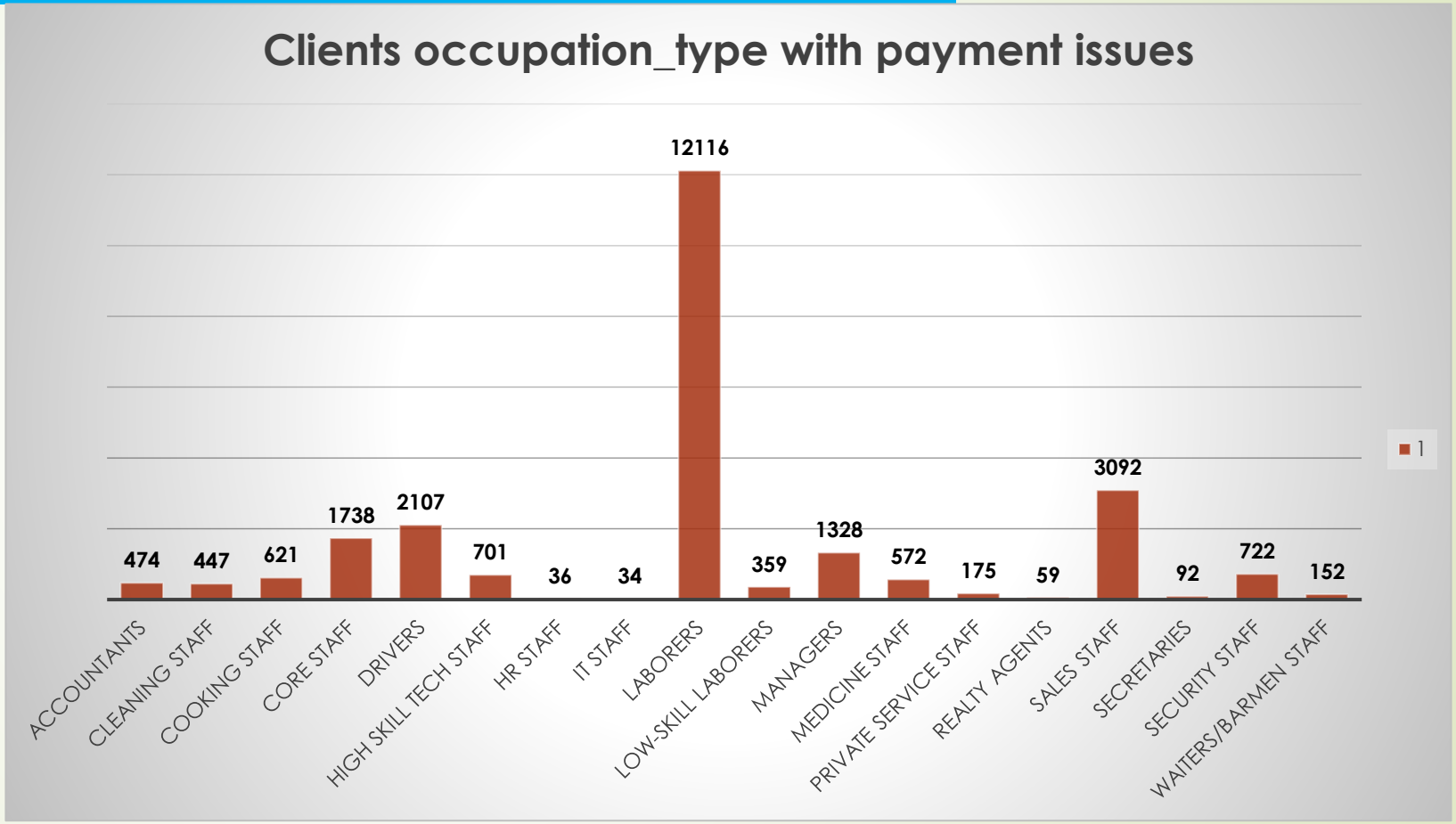
From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with no payment issues

Application Dataset – Analysis

Univariate Analysis

OCCUPATION_TYPE

Count of TARGET	Column Labels	
Row Labels	1	Grand Total
Accountants	474	474
Cleaning staff	447	447
Cooking staff	621	621
Core staff	1738	1738
Drivers	2107	2107
High skill tech staff	701	701
HR staff	36	36
IT staff	34	34
Laborers	12116	12116
Low-skill Laborers	359	359
Managers	1328	1328
Medicine staff	572	572
Private service staff	175	175
Realty agents	59	59
Sales staff	3092	3092
Secretaries	92	92
Security staff	722	722
Waiters/barmen staff	152	152
Grand Total	24825	24825



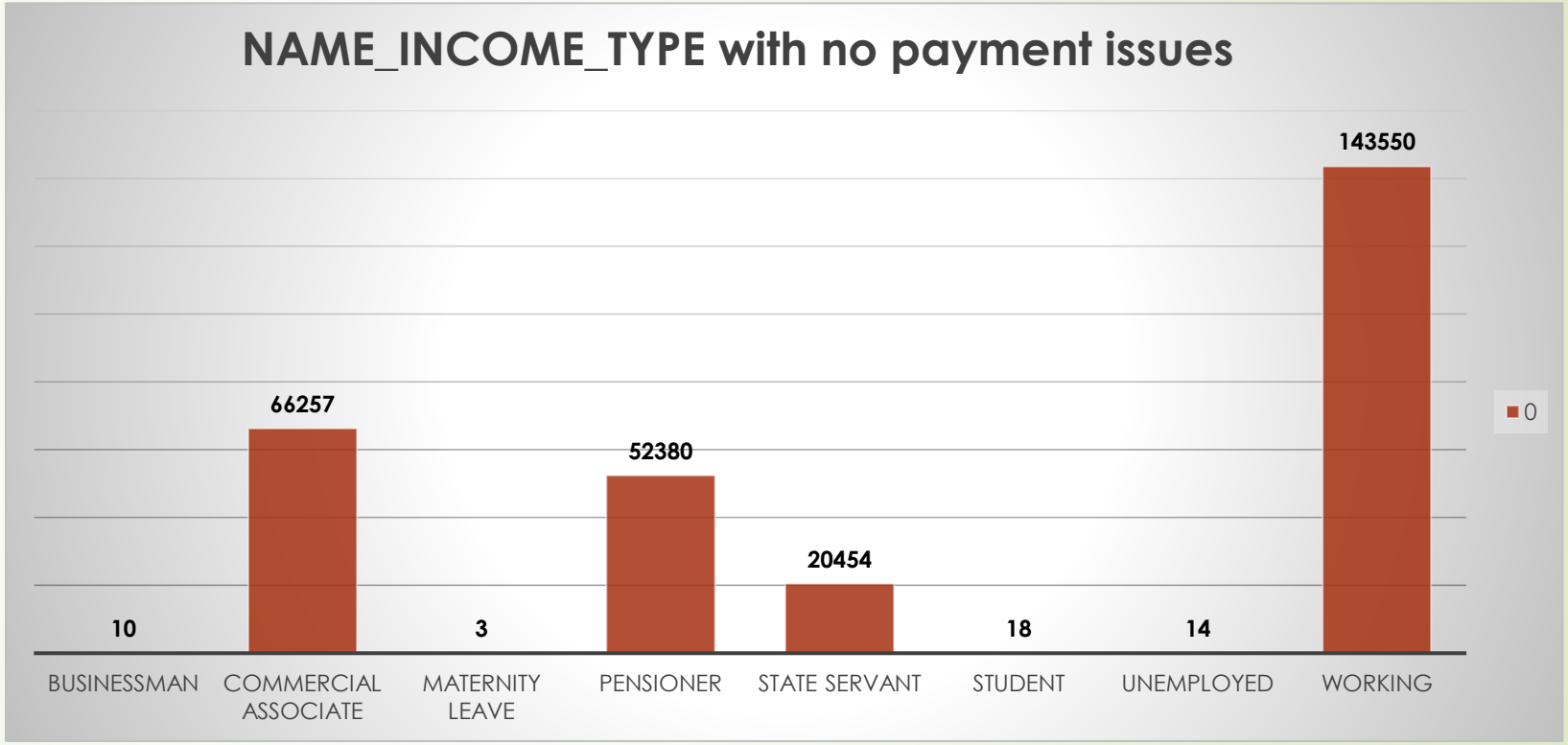
From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with payment issues

Application Dataset – Analysis

Univariate Analysis

NAME_INCOME_TYPE

Count of TARGET	Column Labels	
Row Labels	0	Grand Total
Businessman	10	10
Commercial associate	66257	66257
Maternity leave	3	3
Pensioner	52380	52380
State servant	20454	20454
Student	18	18
Unemployed	14	14
Working	143550	143550
Grand Total	282686	282686



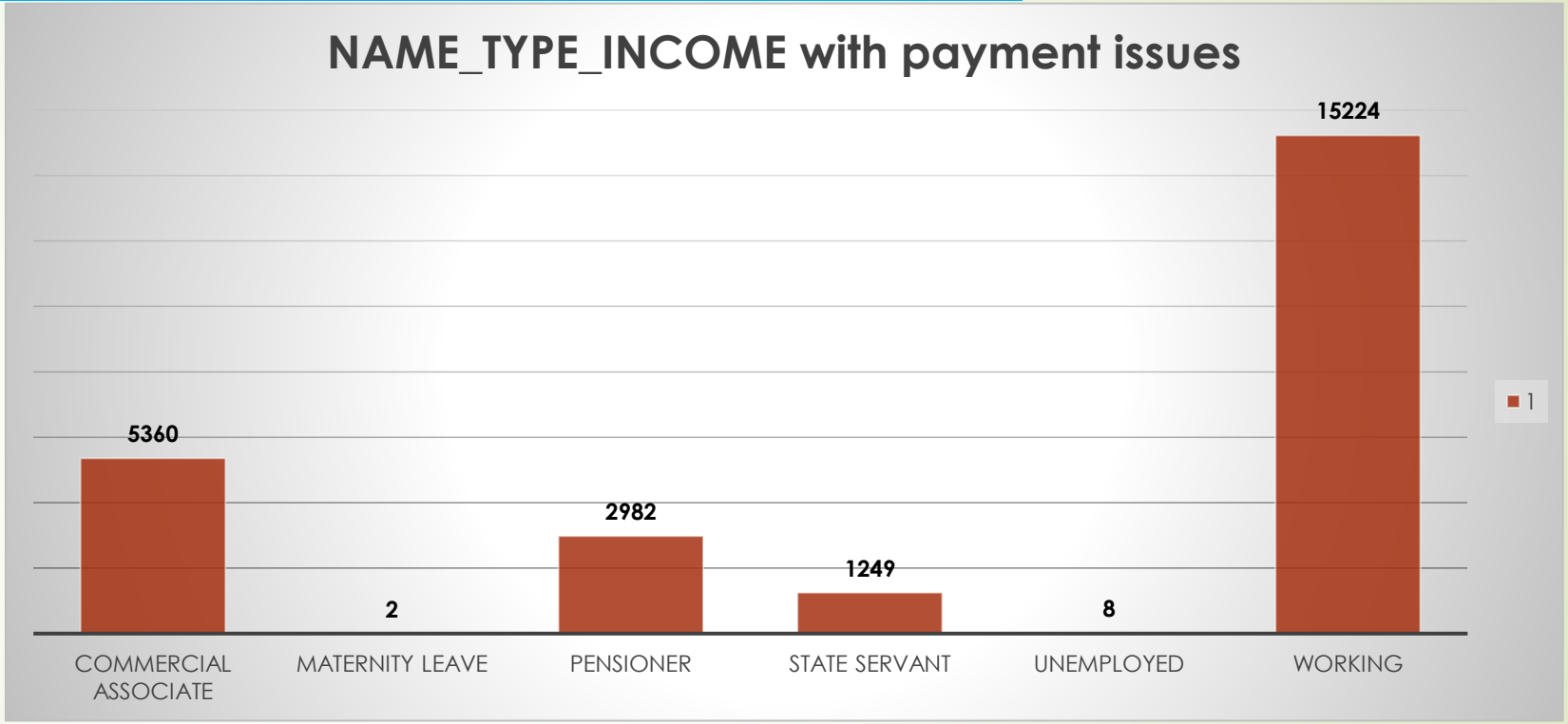
From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with no payment issues

Application Dataset – Analysis

Univariate Analysis

NAME_INCOME_TYPE

Count of TARGET	Column Labels	
Row Labels	1	Grand Total
Commercial associate	5360	5360
Maternity leave	2	2
Pensioner	2982	2982
State servant	1249	1249
Unemployed	8	8
Working	15224	15224
Grand Total	24825	24825



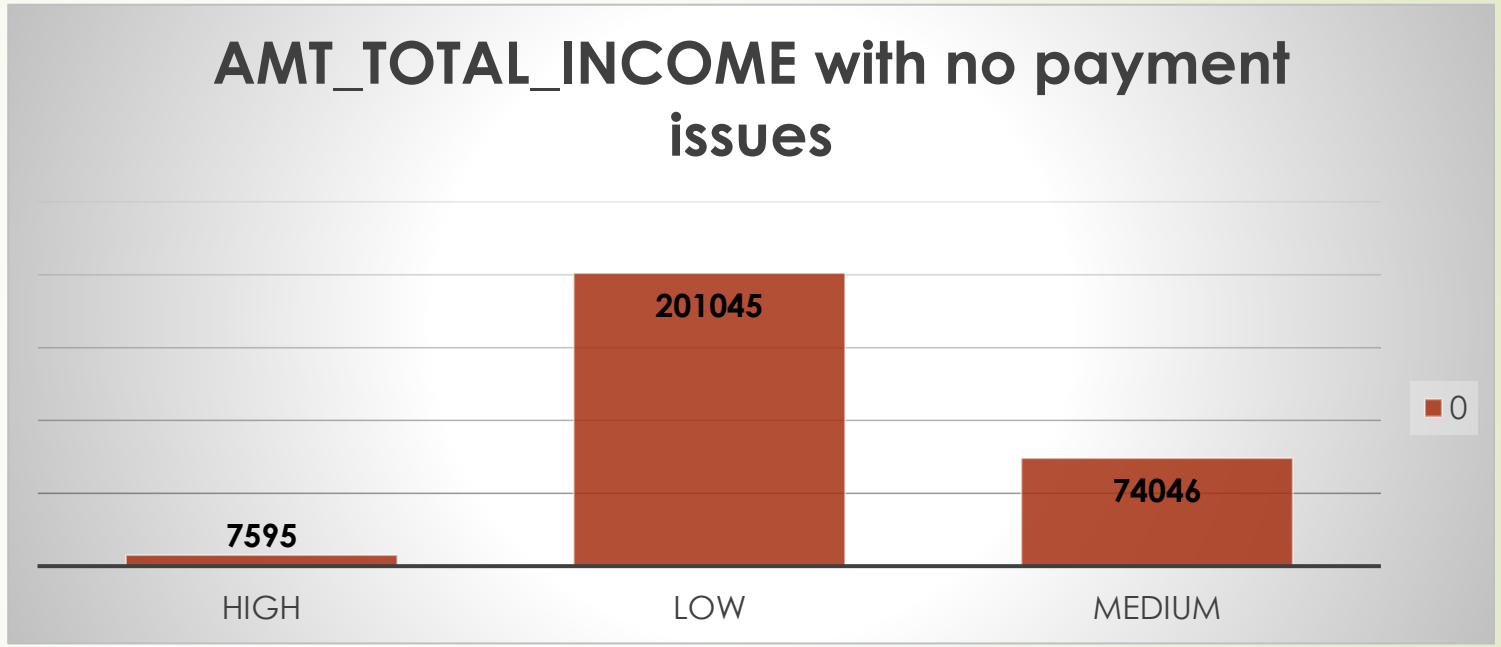
From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with payment issues

Application Dataset – Analysis

Univariate Analysis

AMT_TOTAL INCOME

Count of TARGET	Column Labels	
Row Labels	0	Grand Total
HIGH	7595	7595
LOW	201045	201045
MEDIUM	74046	74046
Grand Total	282686	282686



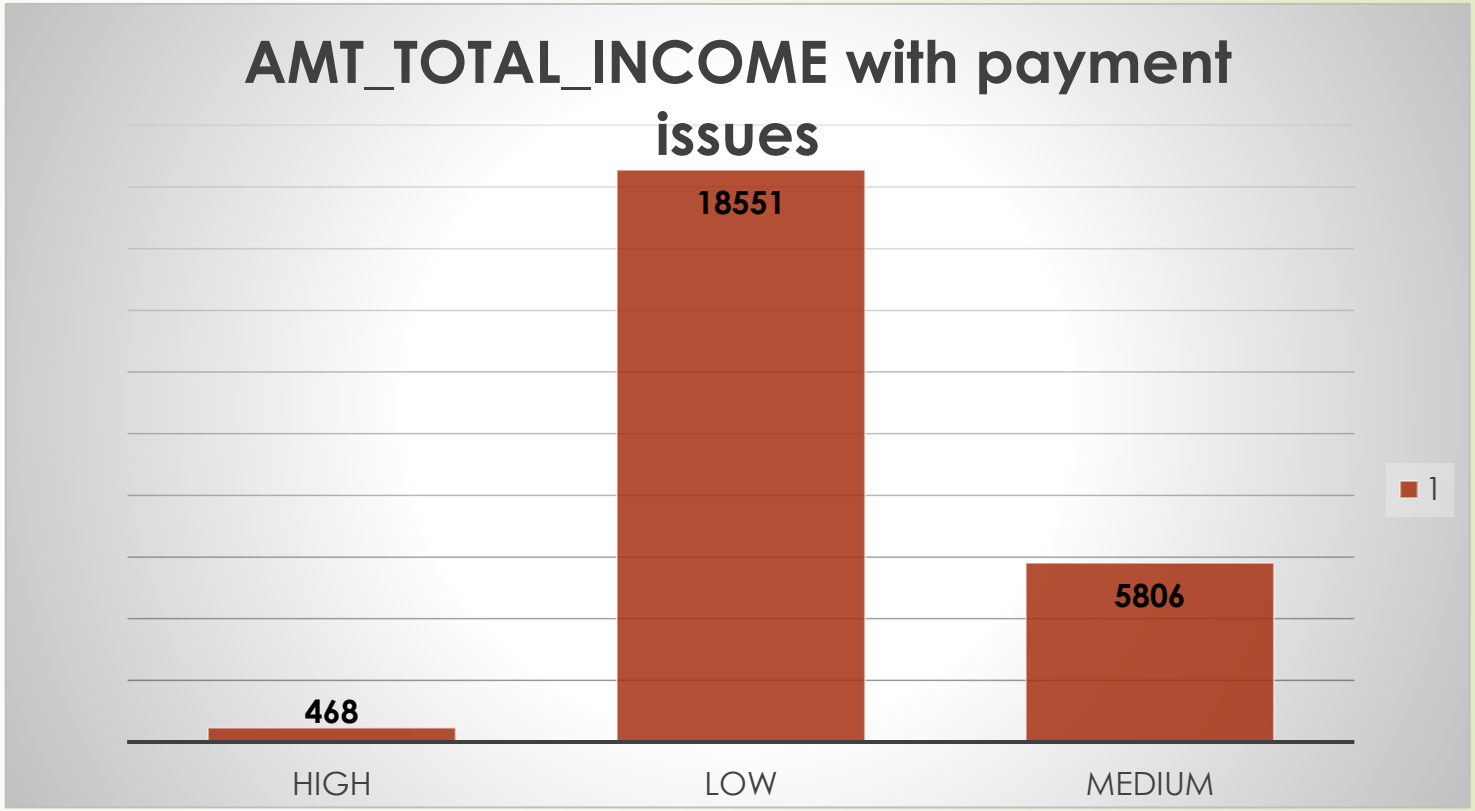
From the above Bar plot we can infer that client having the total income range as 'LOW' have the highest count when it comes to clients having no payment issues

Application Dataset – Analysis

Univariate Analysis

AMT_TOTAL INCOME

Count of TARGET	Column Labels	
Row Labels	1	Grand Total
HIGH	468	468
LOW	18551	18551
MEDIUM	5806	5806
Grand Total	24825	24825



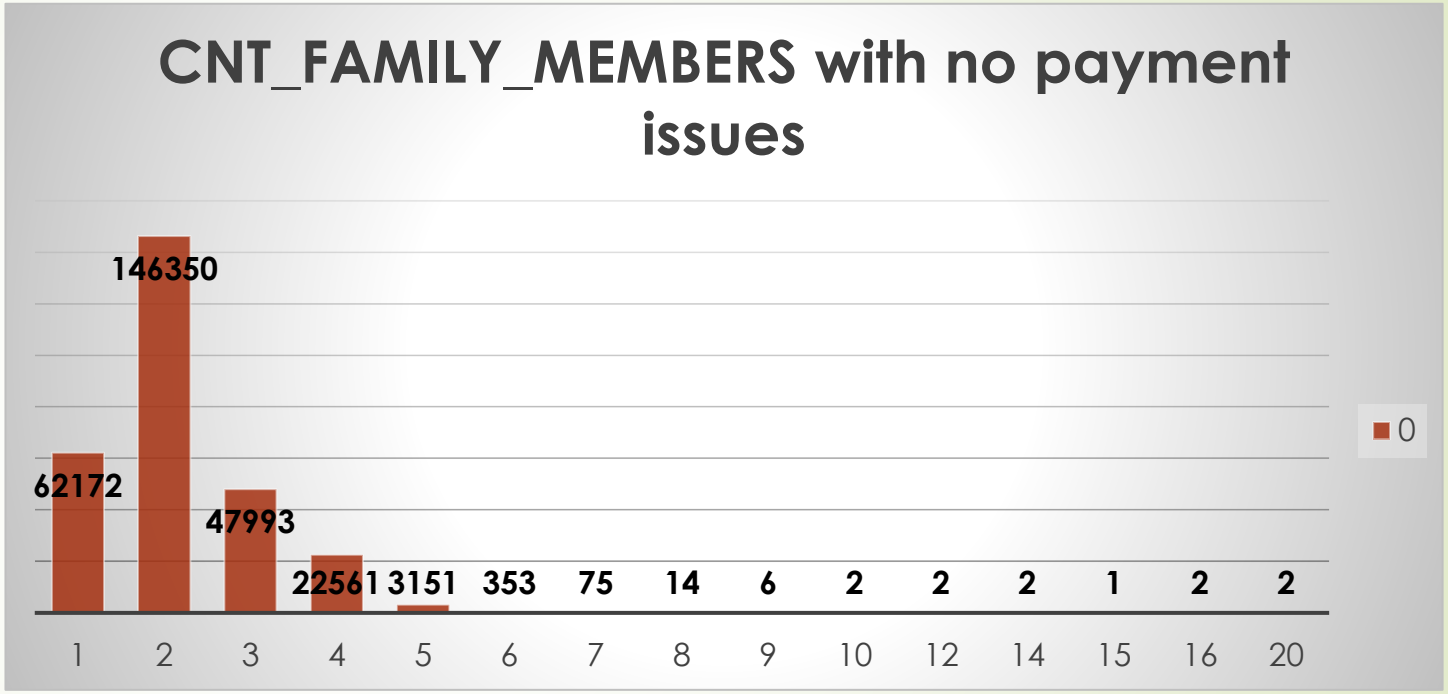
From the above Bar plot we can infer that client having the total income range as 'LOW' have the highest count when it comes to clients having payment issues

Application Dataset – Analysis

Univariate Analysis

CNT_FAMILY_MEMBERS

Count of CNT_FAM_MEMBERS	Column Labels	
Row Labels	0	Grand Total
1	62172	62172
2	146350	146350
3	47993	47993
4	22561	22561
5	3151	3151
6	353	353
7	75	75
8	14	14
9	6	6
10	2	2
12	2	2
14	2	2
15	1	1
16	2	2
20	2	2
Grand Total	282686	282686



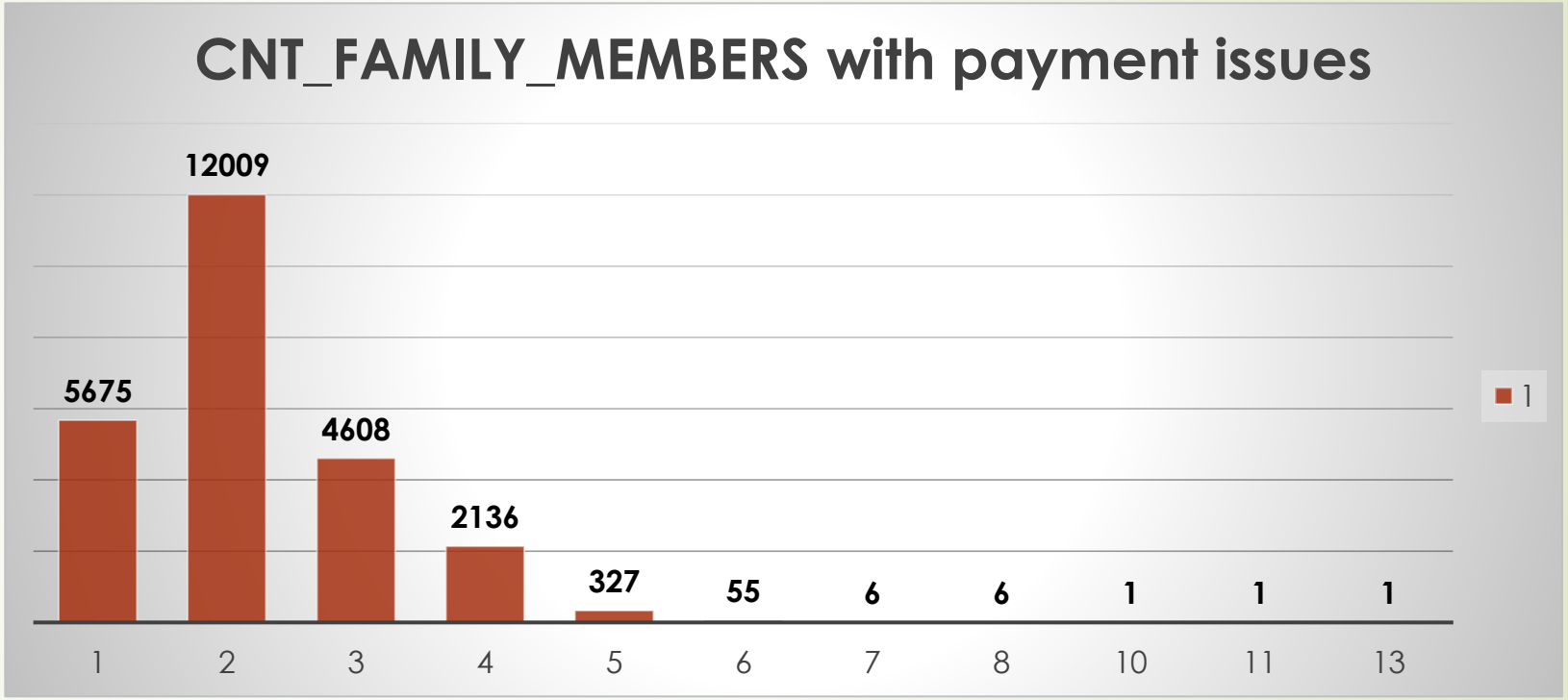
From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having no payment issues

Application Dataset – Analysis

Univariate Analysis

CNT_FAMILY_MEMBERS

Count of CNT_FAM_MEMBERS	Column Labels	
Row Labels	1	Grand Total
1	5675	5675
2	12009	12009
3	4608	4608
4	2136	2136
5	327	327
6	55	55
7	6	6
8	6	6
10	1	1
11	1	1
13	1	1
Grand Total	24825	24825



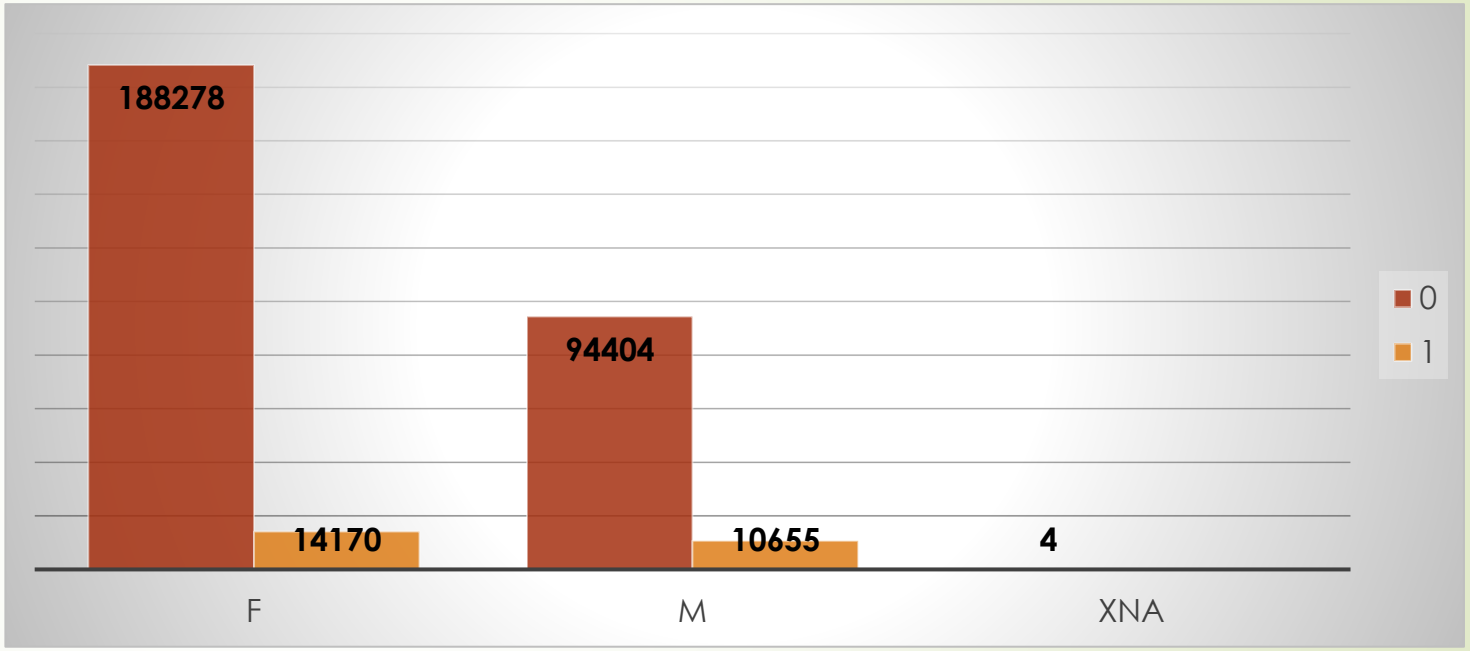
From the above Bar plot we can infer that clients having total count of family members as 2 have the highest count when it comes to clients having payment issues

Application Dataset – Analysis

Univariate Analysis for TARGET variable

CODE_GENDER

Count of CODE_GENDER	Column Labels		
Row Labels	0	1	Grand Total
F	188278	14170	202448
M	94404	10655	105059
XNA	4		4
Grand Total	282686	24825	307511



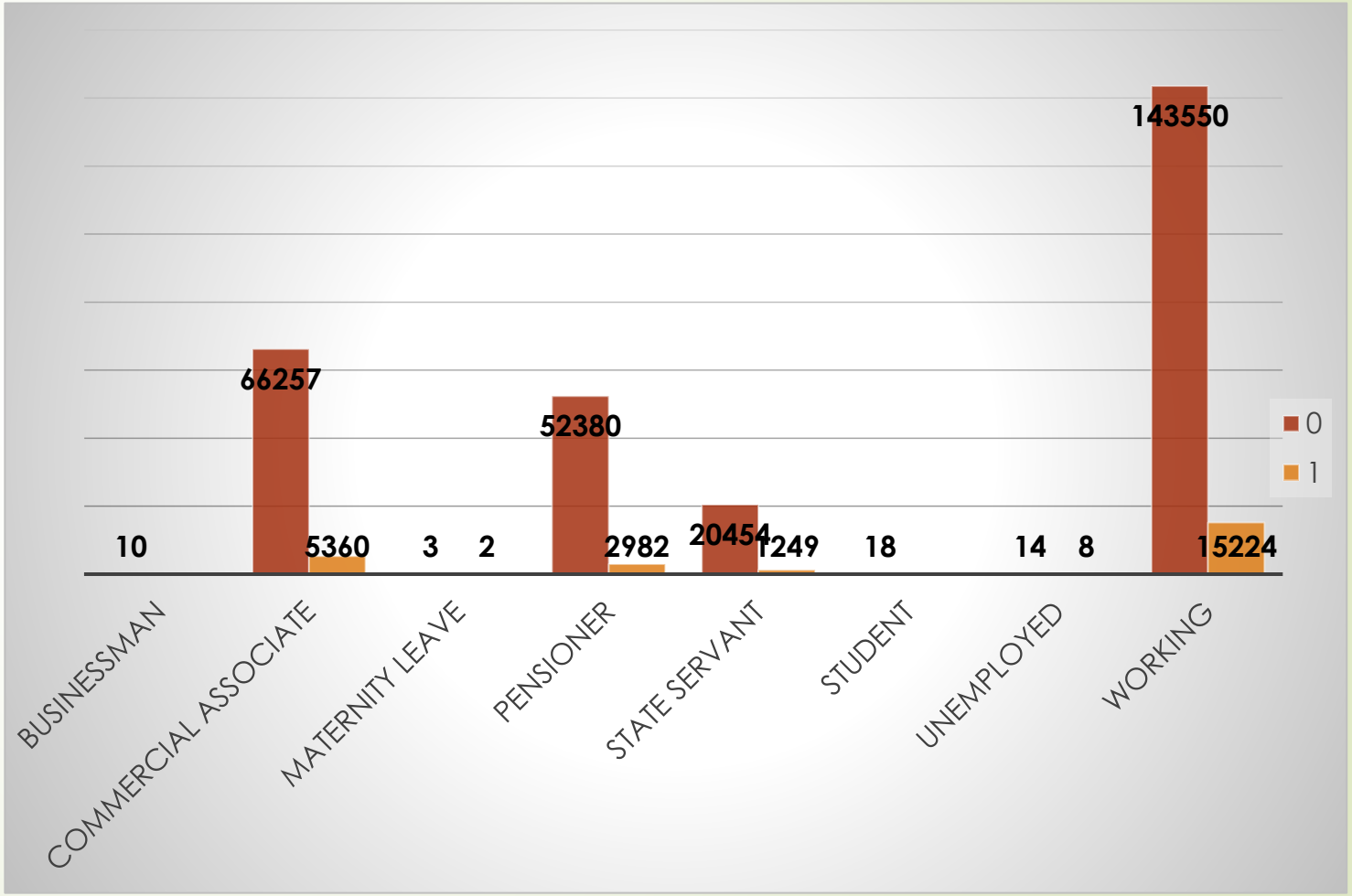
From the above Bar Plot we can infer that Clients with CODE_GENDER = 'F' have the highest number of non-defaulters i.e. 188278-14170 = 174108

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_INCOME_TYPE

Count of NAME_INCOME_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Businessman	10		10
Commercial associate	66257	5360	71617
Maternity leave	3	2	5
Pensioner	52380	2982	55362
State servant	20454	1249	21703
Student	18		18
Unemployed	14	8	22
Working	143550	15224	158774
Grand Total	282686	24825	307511



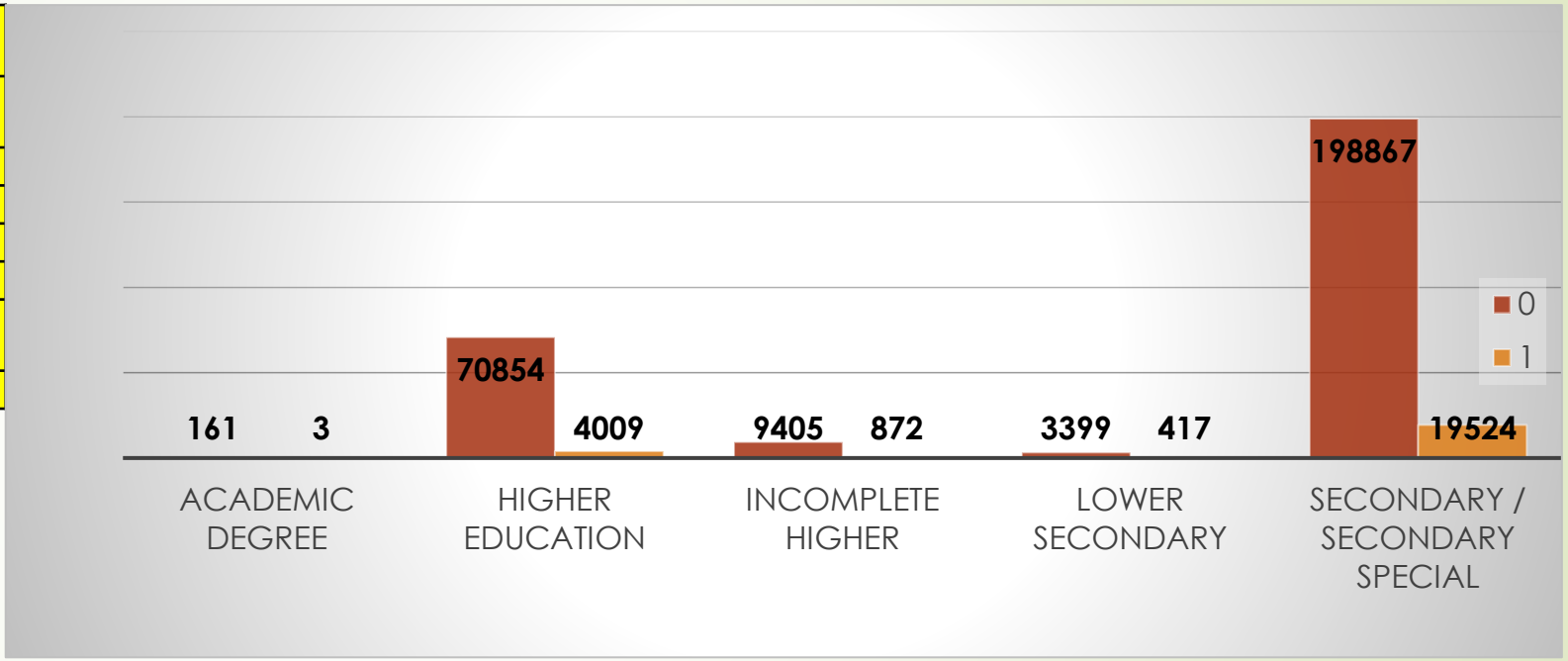
From the adjacent Bar Plot we can infer that clients having NAME_INCOME_TYPE = 'WORKING' having the highest count of Non-defaulters i.e.
143550-15224 = 128326

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_EDUCATION_TYPE

Count of NAME_EDUCATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Academic degree	161	3	164
Higher education	70854	4009	74863
Incomplete higher	9405	872	10277
Lower secondary	3399	417	3816
Secondary / secondary special	198867	19524	218391
Grand Total	282686	24825	307511



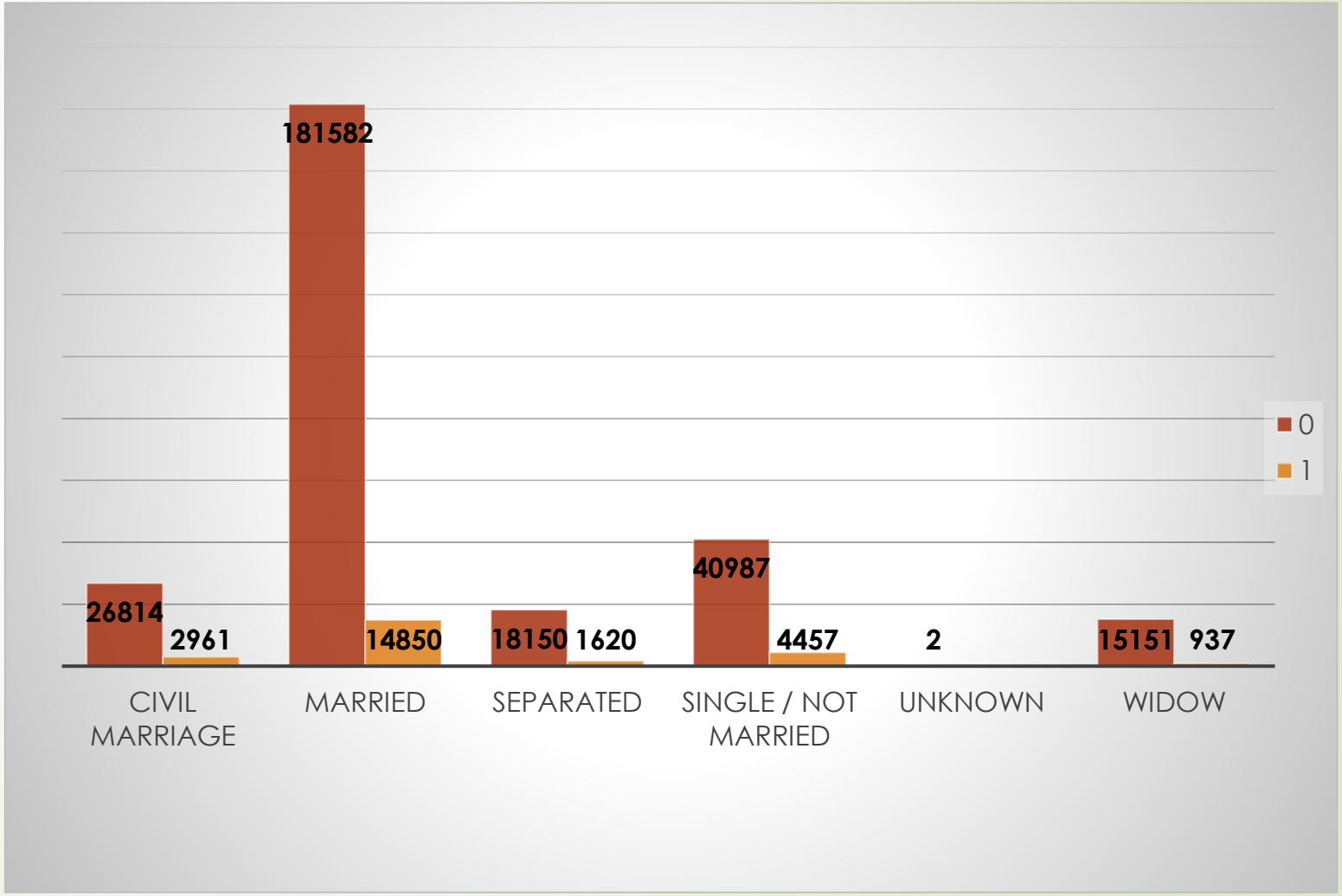
From the above Bar Plot we can infer that clients having NAME_EDUCATION_TYPE = 'SECONDARY/SECONDARY SPECIAL' have the highest count for Non-defaulters i.e.
198867-19524 = 179343

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_FAMILY_STATUS

Count of NAME_FAMILY_STATUS	Column Labels		
Row Labels	0	1	Grand Total
Civil marriage	26814	2961	29775
Married	181582	14850	196432
Separated	18150	1620	19770
Single / not married	40987	4457	45444
Unknown	2		2
Widow	15151	937	16088
Grand Total	282686	24825	307511



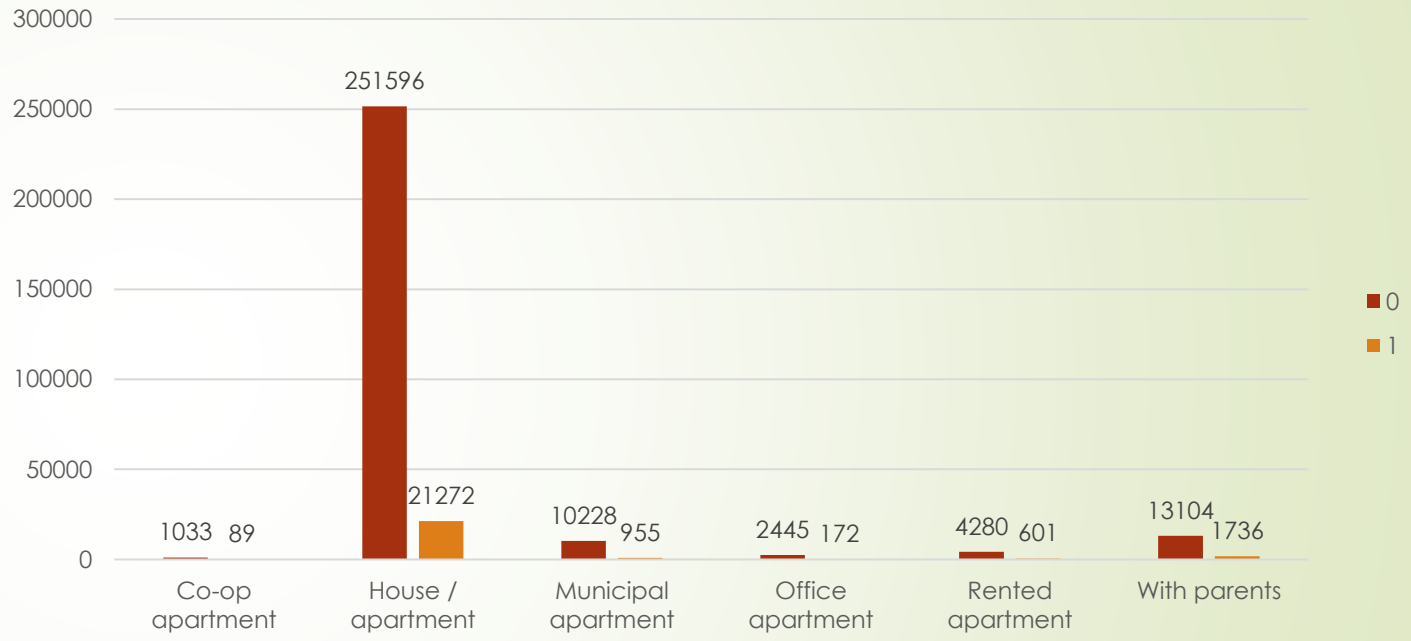
From the adjacent Bar Plot we can infer that clients having NAME_FAMILY_STATUS = 'MARRIED' have the highest count of Non-defaulters i.e.
181582 – 14850 = 166732

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_HOUSING_TYPE

Count of NAME_HOUSING_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Co-op apartment	1033	89	1122
House / apartment	251596	21272	272868
Municipal apartment	10228	955	11183
Office apartment	2445	172	2617
Rented apartment	4280	601	4881
With parents	13104	1736	14840
Grand Total	282686	24825	307511



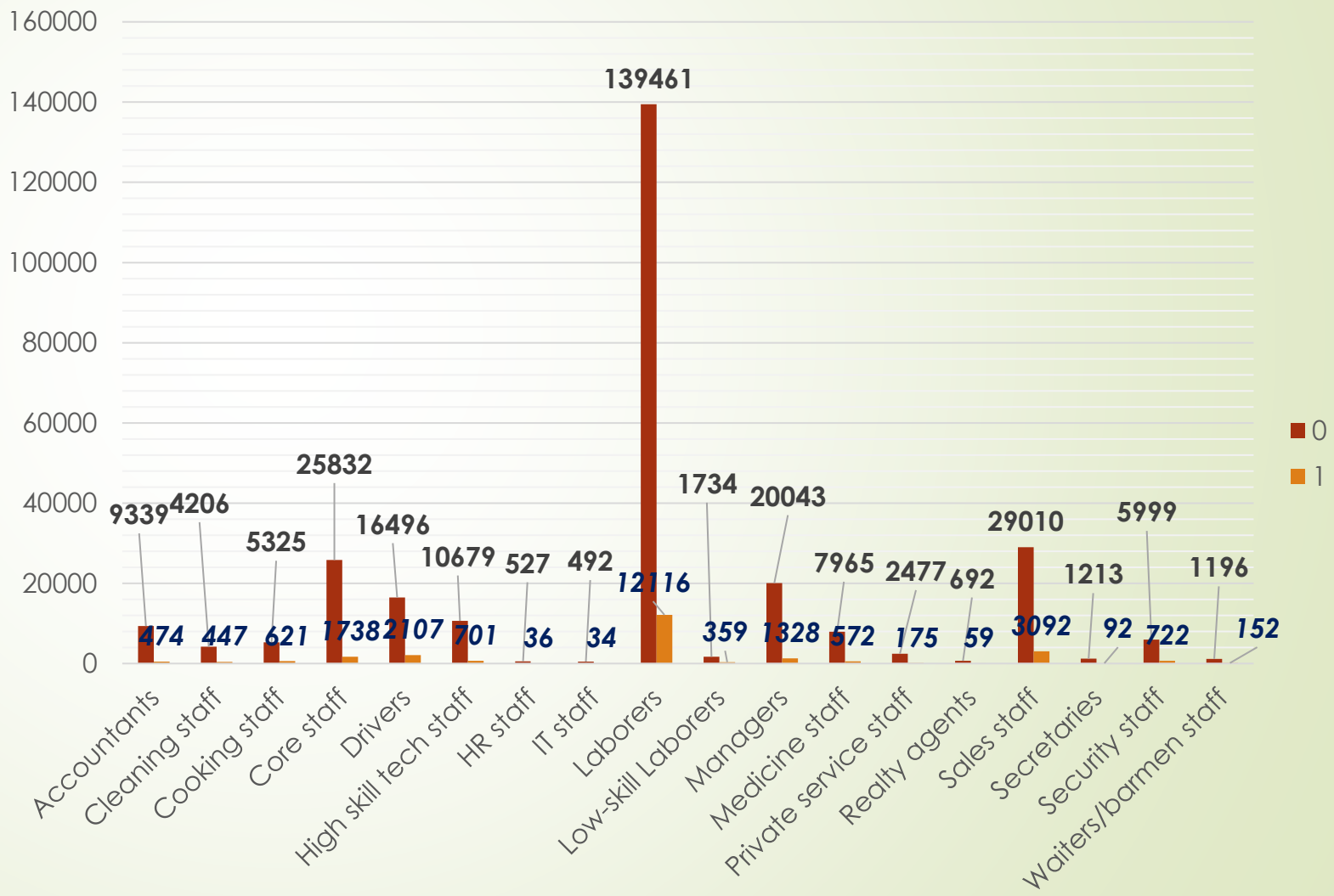
From the above Bar Plot we can infer that clients having NAME_HOUSING_TYPE = 'House/Apartment' have the highest count of Non-defaulters i.e. 251596-21272 = 230324

Application Dataset – Analysis

Univariate Analysis for TARGET variable

OCCUPATION_TYPE

Count of OCCUPATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Accountants	9339	474	9813
Cleaning staff	4206	447	4653
Cooking staff	5325	621	5946
Core staff	25832	1738	27570
Drivers	16496	2107	18603
High skill tech staff	10679	701	11380
HR staff	527	36	563
IT staff	492	34	526
Laborers	139461	12116	151577
Low-skill Laborers	1734	359	2093
Managers	20043	1328	21371
Medicine staff	7965	572	8537
Private service staff	2477	175	2652
Realty agents	692	59	751
Sales staff	29010	3092	32102
Secretaries	1213	92	1305
Security staff	5999	722	6721
Waiters/barmen staff	1196	152	1348
Grand Total	282686	24825	307511



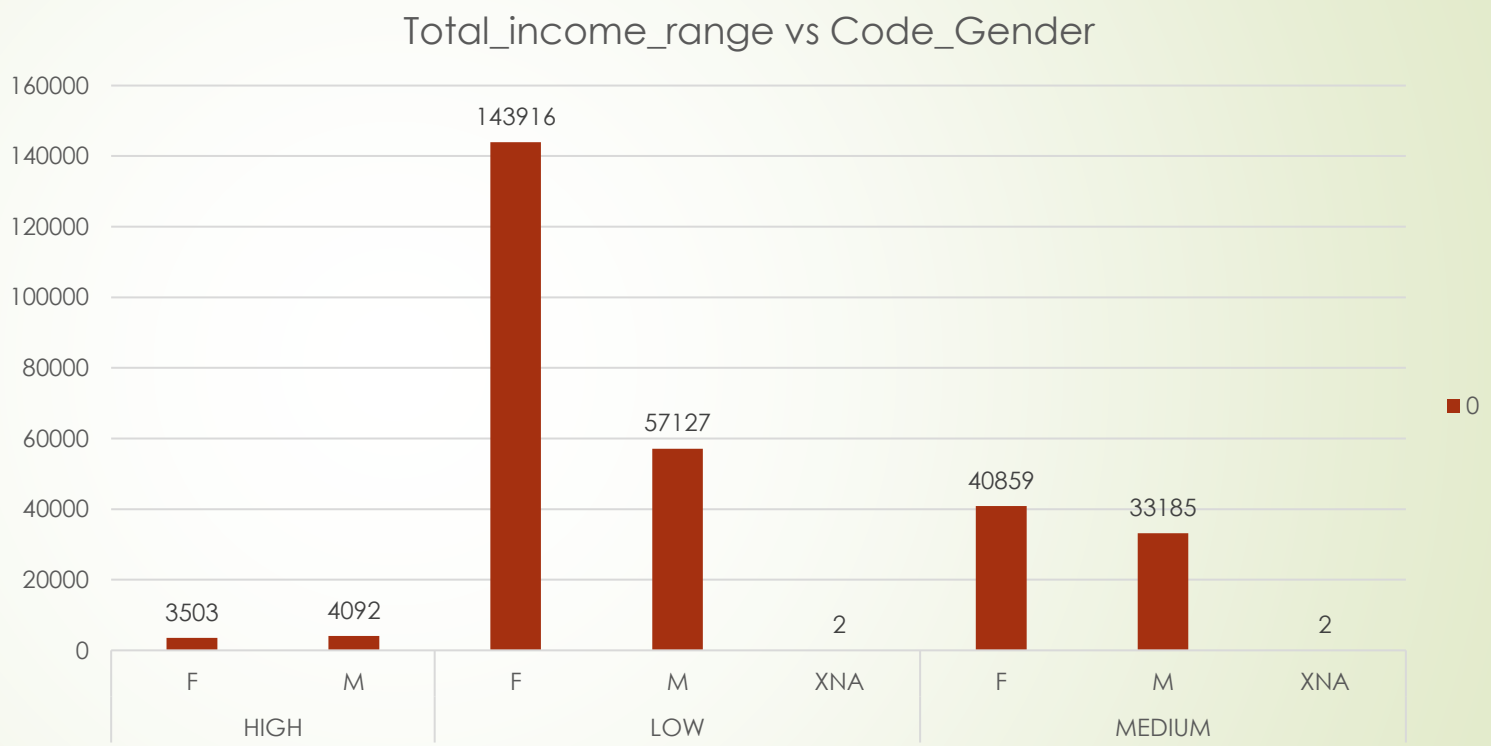
From the adjacent Bar plot we can infer that clients having occupation_type = 'Laborers' have the highest count for Non-defaulters i.e.
139461-12116 = 127345

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Total_income_range vs Code_gender

Count of CODE_GENDER	Column Labels	
Row Labels	0	Grand Total
HIGH	7595	7595
F	3503	3503
M	4092	4092
LOW	201045	201045
F	143916	143916
M	57127	57127
XNA	2	2
MEDIUM	74046	74046
F	40859	40859
M	33185	33185
XNA	2	2
Grand Total	282686	282686



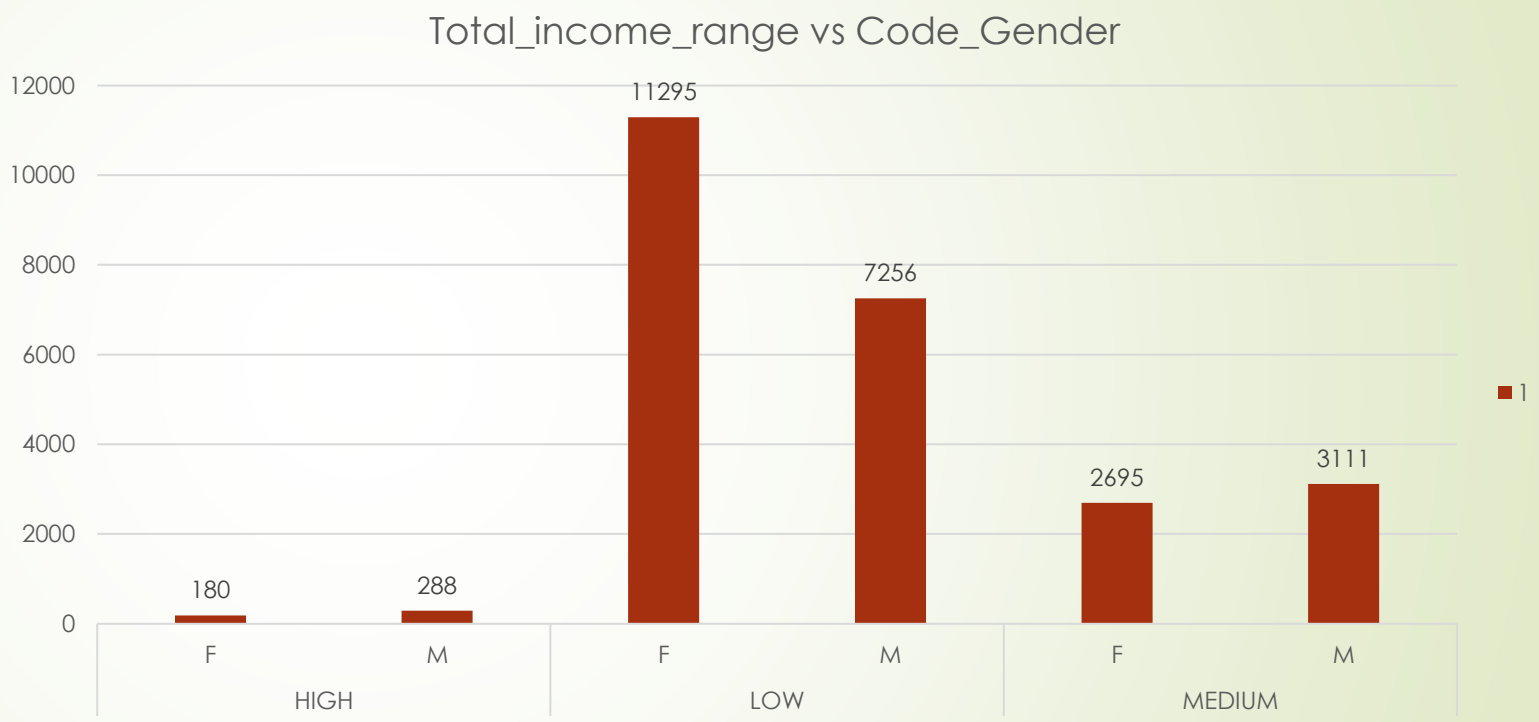
From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with no payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 1: Total_income_range vs Code_gender

Count of CODE_GENDER	Column Labels	
Row Labels	1	Grand Total
HIGH	468	468
F	180	180
M	288	288
LOW	18551	18551
F	11295	11295
M	7256	7256
MEDIUM	5806	5806
F	2695	2695
M	3111	3111
Grand Total	24825	24825



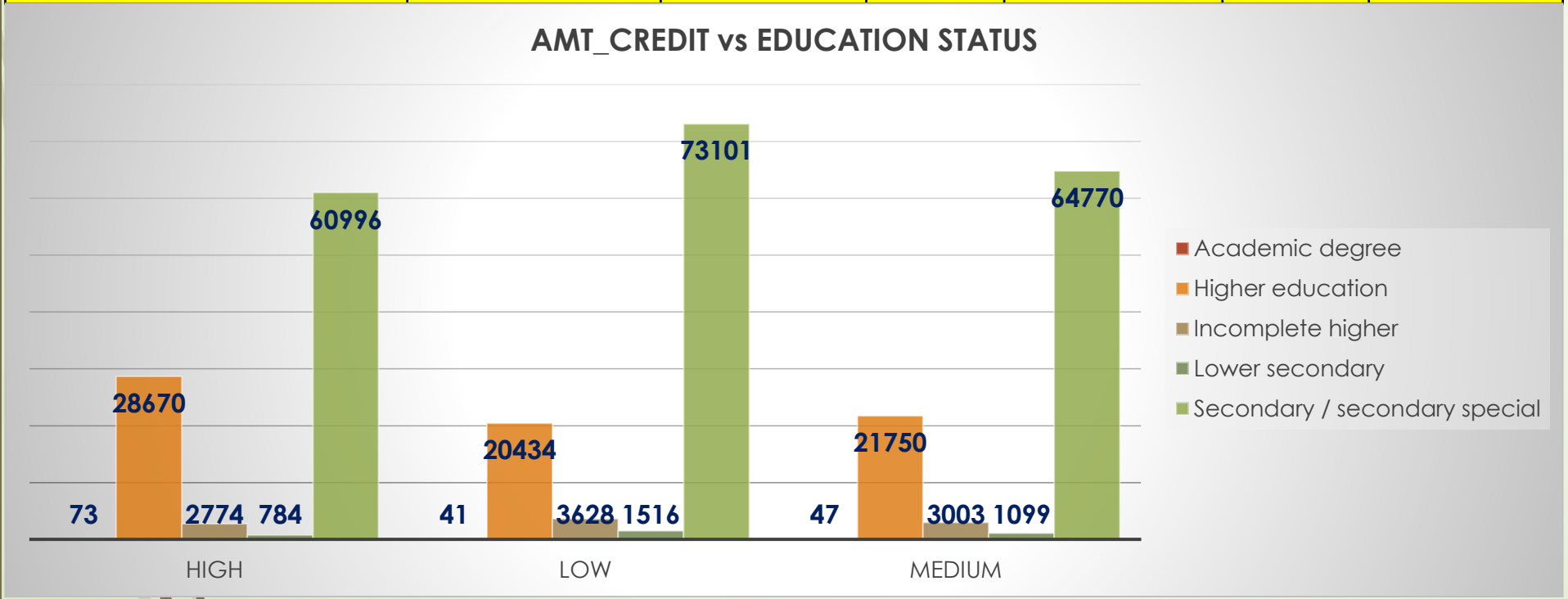
From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Credit Amt vs Education status

TARGET	0					
Count of NAME_EDUCATION_TYPE	Column Labels					
Row Labels	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special	Grand Total
HIGH	73	28670	2774	784	60996	93297
LOW	41	20434	3628	1516	73101	98720
MEDIUM	47	21750	3003	1099	64770	90669
Grand Total	161	70854	9405	3399	198867	282686



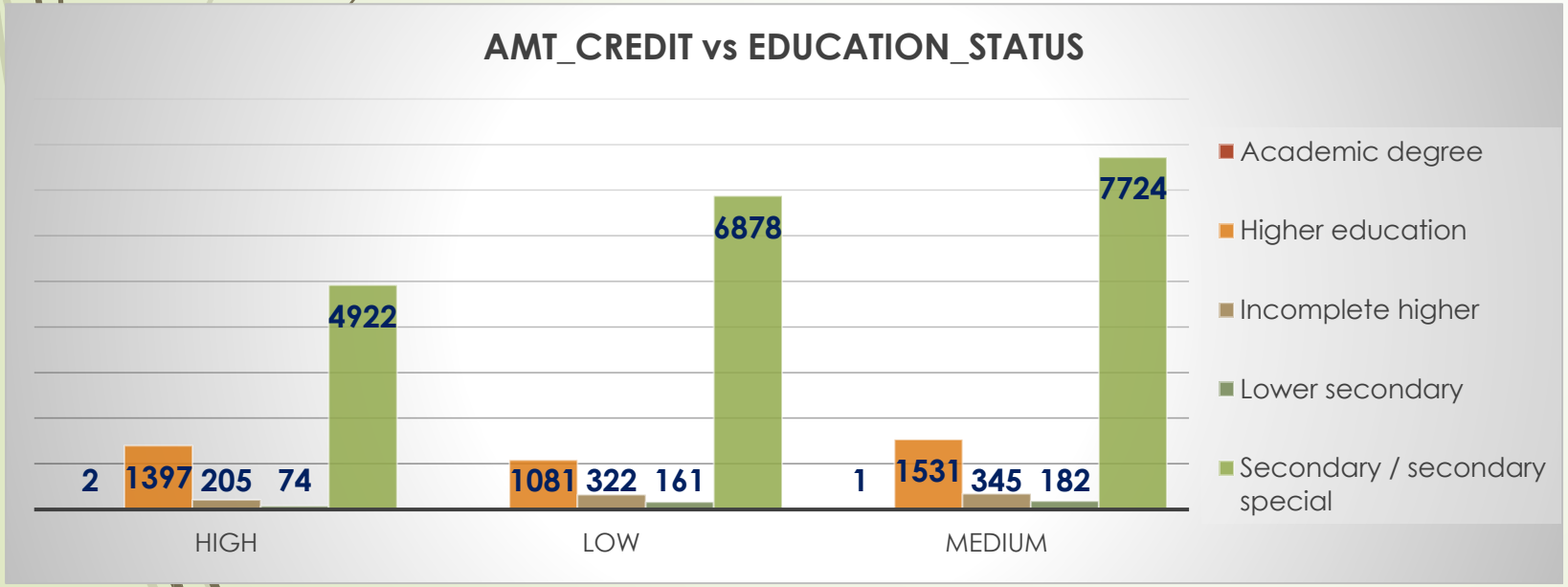
From the adjacent Bar Plot we can infer that clients having credit amt range as 'Low' and education status as 'Secondary/ Secondary Special' have the highest count for clients with no payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 1: Credit Amt vs Education status

TARGET	1					
Count of NAME_EDUCATION_TYPE	Column Labels					
Row Labels	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special	Grand Total
HIGH	2	1397	205	74	4922	6600
LOW		1081	322	161	6878	8442
MEDIUM	1	1531	345	182	7724	9783
Grand Total	3	4009	872	417	19524	24825



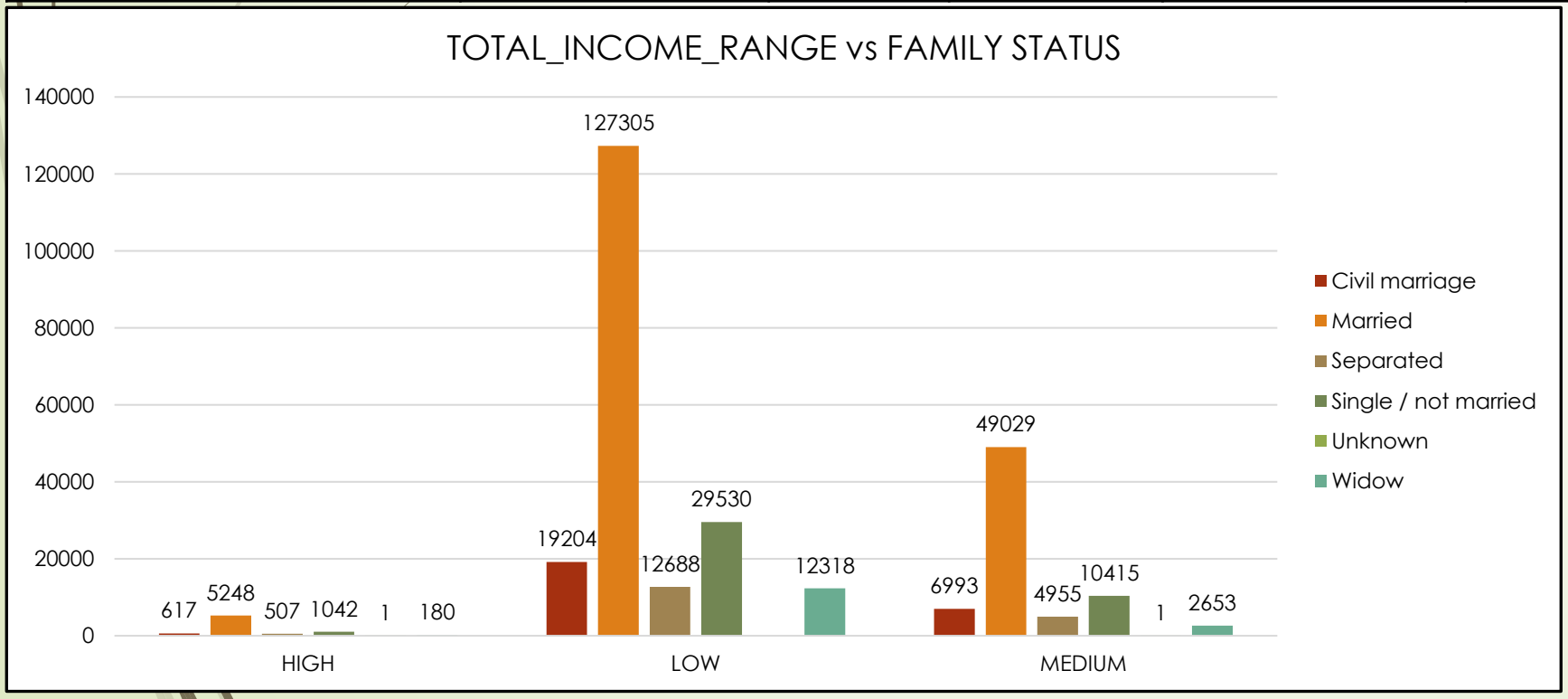
From the adjacent Bar Plot we can infer that clients having credit amt range as 'Medium' and education status as 'Secondary/ Secondary Special' have the highest count for clients with payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Total Income vs Family status

TARGET	0						
Count of NAME_FAMILY_STATUS	Column Labels						
Row Labels	Civil marriage	Married	Separated	Single / not married	Unknown	Widow	Grand Total
HIGH	617	5248	507	1042	1	180	7595
LOW	19204	127305	12688	29530		12318	201045
MEDIUM	6993	49029	4955	10415	1	2653	74046
Grand Total	26814	181582	18150	40987	2	15151	282686



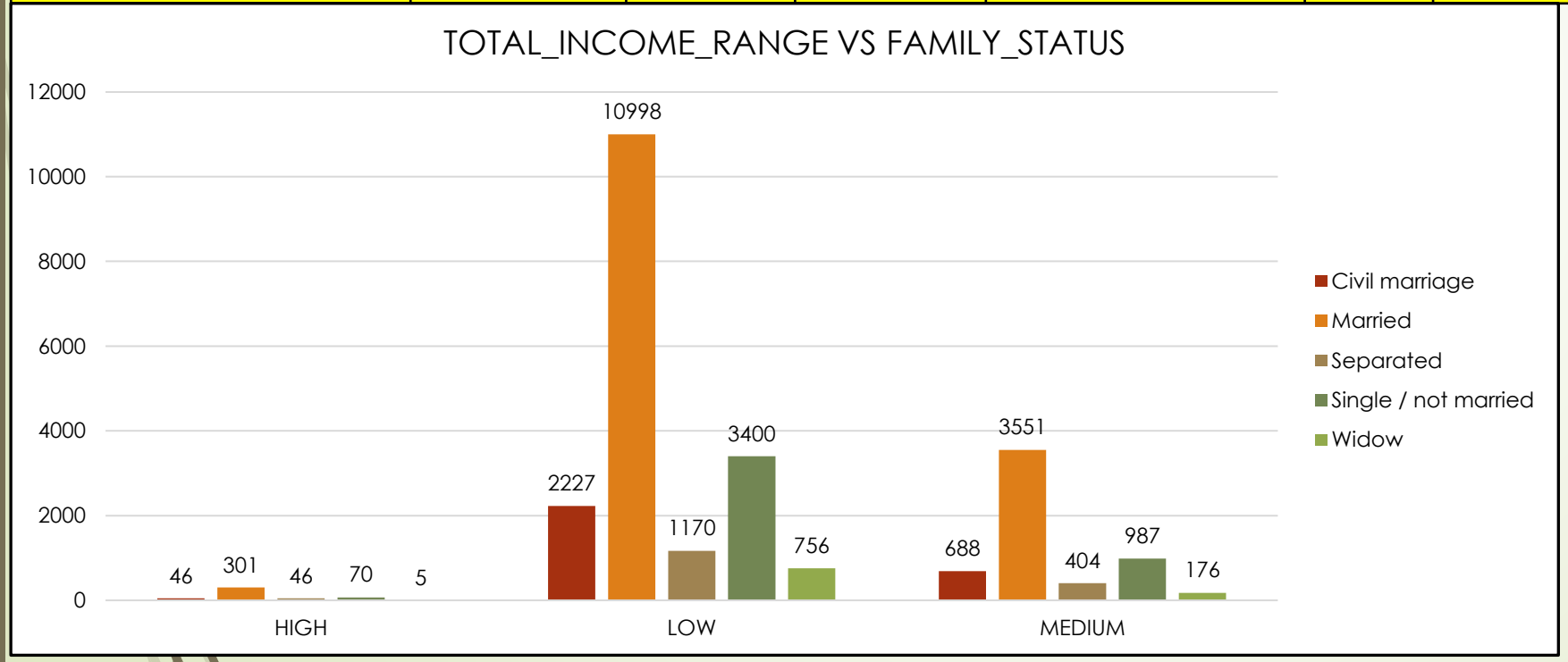
From the adjacent Bar plot we can infer that clients with total_income_range as 'Low' and family_status as 'Married' have the highest count for clients having no payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 1: Total Income vs Family status

TARGET	1					
Count of NAME_FAMILY_STATUS	Column Labels					
Row Labels	Civil marriage	Married	Separated	Single / not married	Widow	Grand Total
HIGH	46	301	46	70	5	468
LOW	2227	10998	1170	3400	756	18551
MEDIUM	688	3551	404	987	176	5806
Grand Total	2961	14850	1620	4457	937	24825



From the adjacent Bar plot we can infer that clients with total_income_range as 'Low' and family_status as 'Married' have the highest count for clients having payment issues

Application Dataset – Analysis

Google Drive Link for Excel sheet of Analysis of Cleaned Data done:-

application_data_cleaned.xlsx - Google Drive

Previous Application Dataset – Dropping, Imputing and analyzing Null values

The following columns of the previous application datasets need to be dropped as they are irrelevant for doing the data analysis

- **HOUR_APPR_PROCESS_START**
- **WEEKDAY_APPR_PROCESS_START_PREV**
- **FLAG_LAST_APPL_PER_CONTRACT**
- **NFLAG_LAST_APPL_IN_DAY**
- **SK_ID_CURR**
- **WEEKDAY_APPR_PROCESS_START**

Removing the rows with the values 'XNA' &'XAP' for the column: NAME_TYPE_SUITE

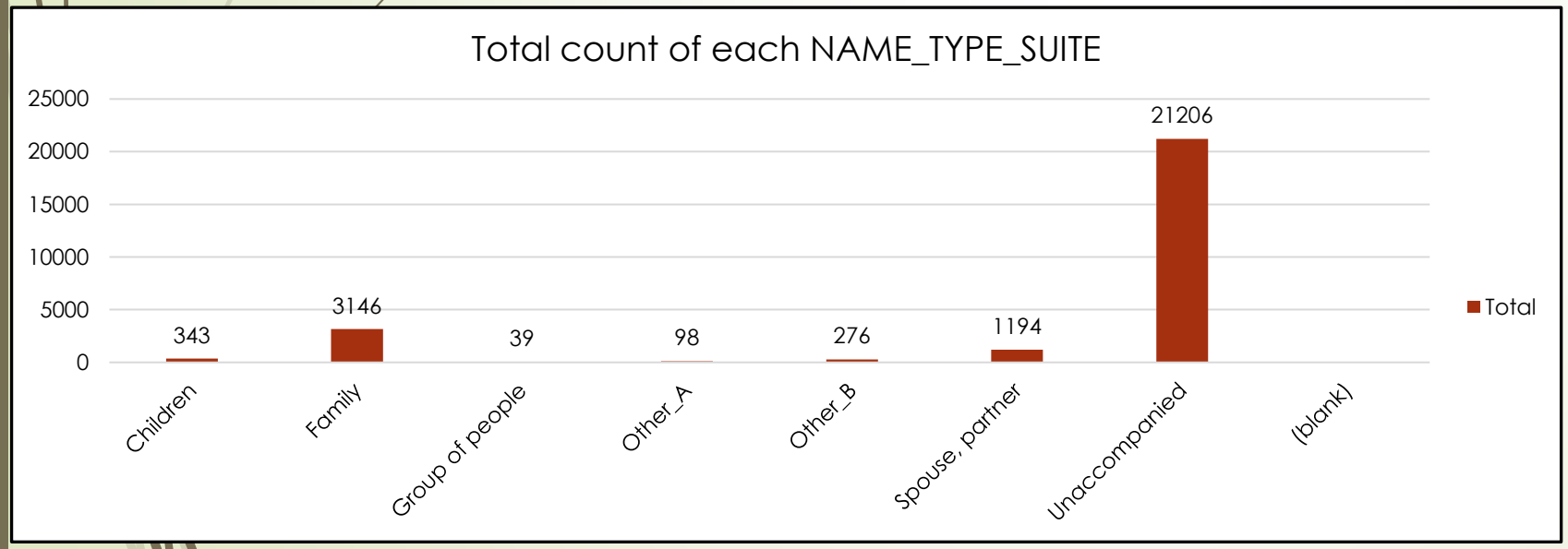
AMT_ANNUITY							
Replace Blanks with 21340							

Median of AMT_ANNUITY
21340

Previous Application Dataset – Dropping, Imputing and analyzing Null values

NAME_TYPE_SUITE

Row Labels	Count of NAME_TYPE_SUITE
Children	343
Family	3146
Group of people	39
Other_A	98
Other_B	276
Spouse, partner	1194
Unaccompanied	21206
(blank)	
Grand Total	26302



Replace Blanks with Unaccompained

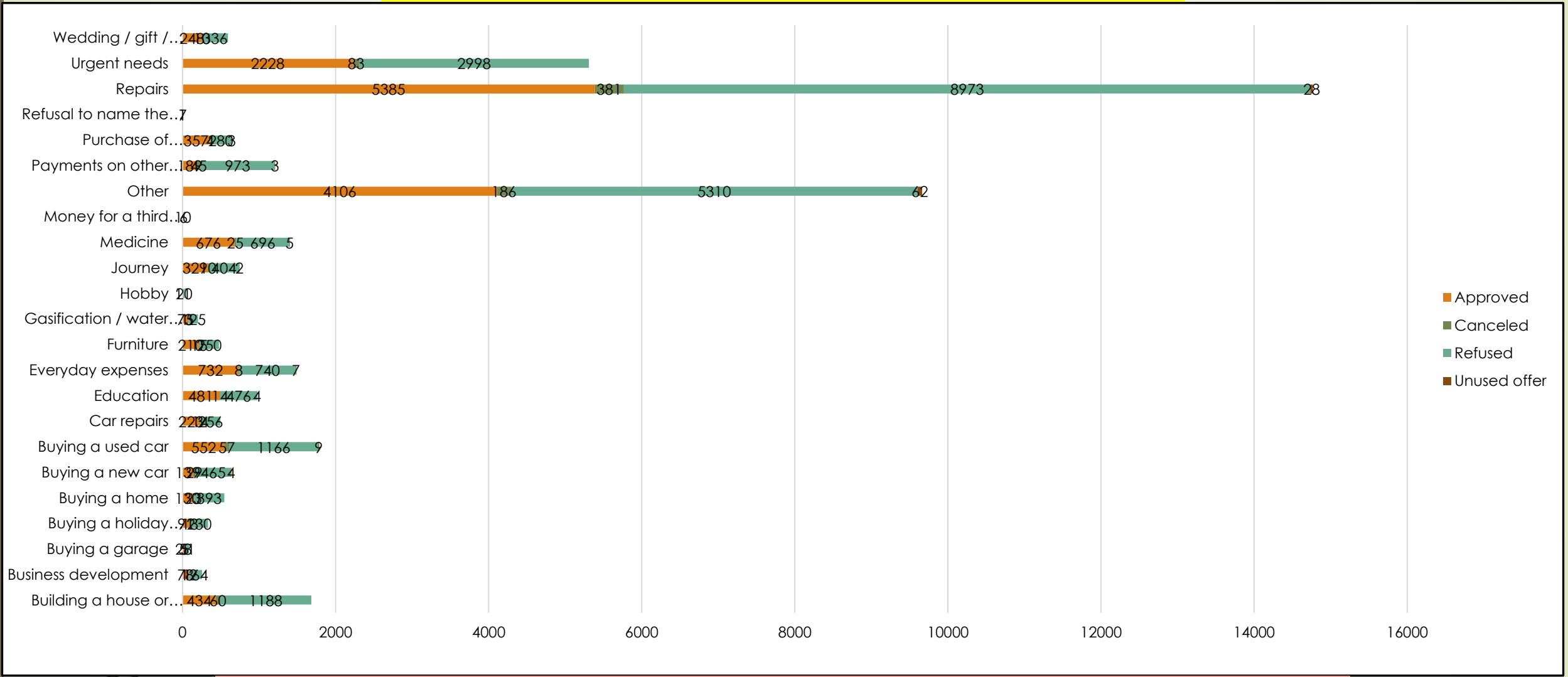
Previous Application Dataset – Analysis of Cleaned Data

Distribution of Name Contract Status

Count of NAME_CONTRACT_STATUS	Column Labels				
Row Labels	Approved	Canceled	Refused	Unused offer	Grand Total
Building a house or an annex	434	60	1188		1682
Business development	78	12	164		254
Buying a garage	28	5	51		84
Buying a holiday home / land	91	13	230		334
Buying a home	130	23	393		546
Buying a new car	139	29	465	4	637
Buying a used car	552	57	1166	9	1784
Car repairs	223	14	256		493
Education	481	14	476	4	975
Everyday expenses	732	8	740	7	1487
Furniture	210	15	250		475
Gasification / water supply	75	3	125		203
Hobby	11		20		31
Journey	329	10	404	2	745
Medicine	676	25	696	5	1402
Money for a third person	10		6		16
Other	4106	186	5310	62	9664
Payments on other loans	189	45	973	3	1210
Purchase of electronic equipment	357	4	280	3	644
Refusal to name the goal	1		7		8
Repairs	5385	381	8973	28	14767
Urgent needs	2228	83	2998		5309
Wedding / gift / holiday	248	10	336		594
Grand Total	16713	997	25507	127	43344

Previous Application Dataset – Analysis of Cleaned Data

Distribution of Name Contract Status



From the above Bar Plot we can infer that Name of Contract status i.e. Repairs work has the highest count of Approved Loans

Previous Application Dataset – Analysis of Cleaned Data

Google Drive Link for Excel sheet of Analysis of Cleaned Data done:-

previous data cleaned.xlsx - Google Sheets

Hence the analysis are being done on both datasets Applications Dataset and Precious Applications Dataset
The following conclusions were drawn from the analysis done

- **The proportion/percentage of the defaulters(target = 1) is around 8% and that of non-defaulters(target = 0) is around 92%**
- **The Bank generally lends more loan to Female clients as compared to Males clients as the count of Female clients in the defaulter's list is less than that of Males. Still Bank can look for more Male clients if their credit amount is satisfied**
- **Also the clients who belong to Working class tend to pay their loans on time followed by the clients who fall under Commercial Associate**
- **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 who fall in the Age Group 31-40 have the highest count for paying off their loans on time followed by the clients who fall in the Age Groups 41-60**
- **Clients having LOW credit amount range tend to pay off their loans on time than compared to HIGH and MEDIUM credit range**

- **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 Parents**

- **Clients taking loan for purchasing New Home i.e. clients taking Home Loans or purchasing New Car i.e. Car Loans and clients who have a income type as State Servant tend to pay their loans on time and hence Bank should prefer clients having such background**

- **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**

Google Drive Folder Link for the Analysed datasets in form of Excel sheets Due to vastness of data the Excel sheets needs to be downloaded and viewed offline:-

trainity_task_6_final_project_2 - Google Drive