Name :Ayithapu Sai Mahathi

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

Columns	count values	XNA	sum	Pecentage of missing value	Count value
0	32950	0	32950	65.90131803	Yes
EXT_SOURCE_1	28172	0	28172	56.3451269	Yes
APARTMENTS_AVG	25385	0	25385	50.77101542	Yes
BASEMENTAREA_AVG	29199	0	29199	58.39916798	Yes
YEARS_BUILD_AVG	33239	0	33239	66.47932959	Yes
COMMONAREA_AVG	34960	0	34960	69.92139843	Yes
ELEVATORS_AVG	26651	0	26651	53.30306606	Yes
ENTRANCES_AVG	25195	0	25195	50.39100782	Yes
FLOORSMIN_AVG	33894	0	33894	67.78935579	Yes
LANDAREA_AVG	29721	0	29721	59.44318886	Yes
LIVINGAPARTMENTS_AVG	34226	0	34226	68.45336907	Yes
LIVINGAREA_AVG	25137	0	25137	50.2750055	Yes
NONLIVINGAPARTMENTS_AVG	34714	0	34714	69.42938859	Yes
NONLIVINGAREA_AVG	27572	0	27572	55.1451029	Yes
APARTMENTS_MODE	25385	0	25385	50.77101542	Yes
BASEMENTAREA_MODE	29199	0	29199	58.39916798	Yes
YEARS_BUILD_MODE	33239	0	33239	66.47932959	Yes
COMMONAREA_MODE	34960	0	34960	69.92139843	Yes
ELEVATORS_MODE	26651	0	26651	53.30306606	Yes
ENTRANCES_MODE	25195	0	25195	50.39100782	Yes
FLOORSMIN_MODE	33894	0	33894	67.78935579	Yes
LANDAREA_MODE	29721	0	29721	59.44318886	Yes
LIVINGAPARTMENTS_MODE	34226	0	34226	68.45336907	Yes
LIVINGAREA_MODE	25137	0	25137	50.2750055	Yes
NONLIVINGAPARTMENTS_MODE	34714	0	34714	69.42938859	Yes
NONLIVINGAREA_MODE	27572	0	27572	55.1451029	Yes
APARTMENTS_MEDI	25385	0	25385	50.77101542	Yes
BASEMENTAREA_MEDI	29199	0	29199	58.39916798	Yes
YEARS_BUILD_MEDI	33239	0	33239	66.47932959	Yes
COMMONAREA_MEDI	34960	0	34960	69.92139843	Yes
ELEVATORS_MEDI	26651	0	26651	53.30306606	Yes
ENTRANCES_MEDI	25195	0	25195	50.39100782	Yes
FLOORSMIN_MEDI	33894	0	33894	67.78935579	Yes
LANDAREA_MEDI	29721	0	29721	59.44318886	Yes
LIVINGAPARTMENTS_MEDI	34226	0	34226	68.45336907	Yes
LIVINGAREA_MEDI	25137	0	25137	50.2750055	Yes
NONLIVINGAPARTMENTS_MEDI	34714	0	34714	69.42938859	Yes
NONLIVINGAREA_MEDI	27572	0	27572	55.1451029	Yes
FONDKAPREMONT_MODE	34191	0	34191	68.38336767	Yes
HOUSETYPE_MODE	25075	0	25075	50.15100302	Yes
WALLSMATERIAL_MODE	25459	0	25459	50.91901838	Yes

Application Dataset – NULL values Dataset

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

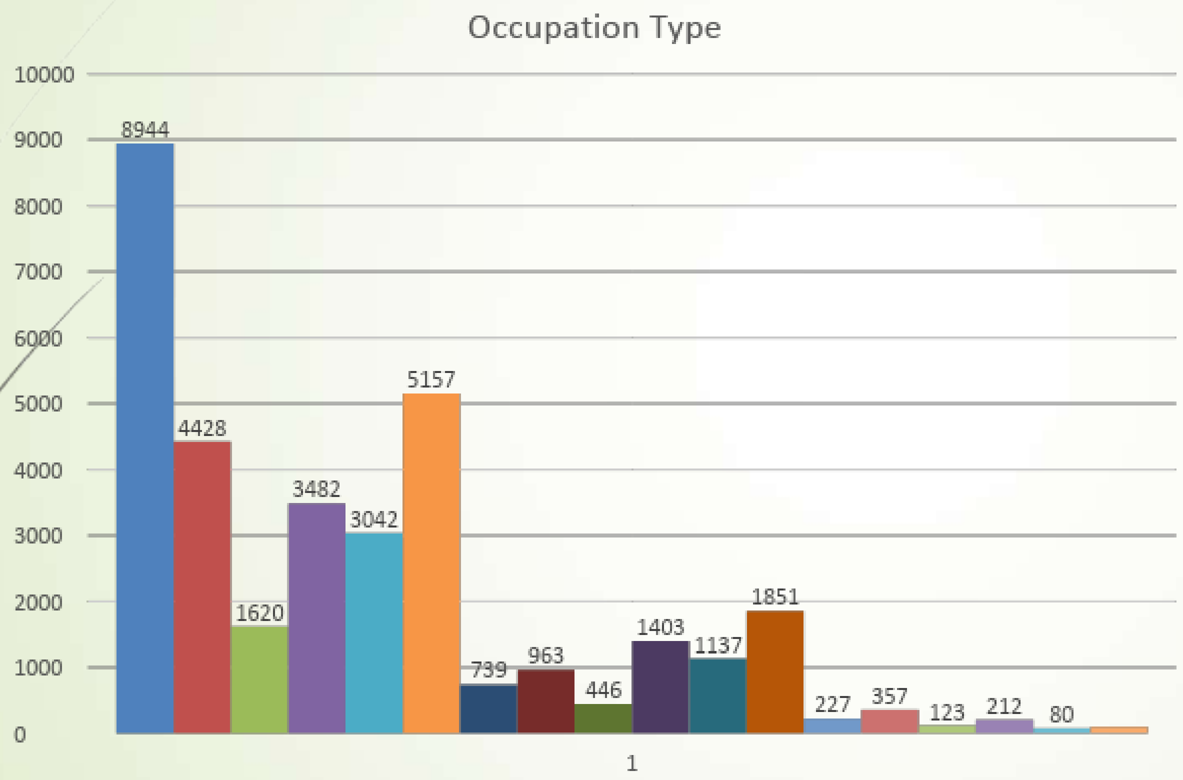
Column name
FLAG_MOBIL
FLAG_EMPLOY_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAMILY_MEMBERS
REGION_RATING_CLENT
REGION_RATING_CLENT_W_CITY
EXT_SOURCE_3
YEAR_BEGINEXPLUATATION_AVG
YEAR_BEGINEXPLUATATION_MODE
YEAR_BEGINEXPLUATATION_MEDIAN
TOTAL_AREA_MODE
EMERGENCYSTATE_MODE
DAYS_LAST_PHONE_CHANGE
FLAG DOC 2
FLAG DOC 3
FLAG DOC 4
FLAG DOC 5
FLAG DOC 6
FLAG DOC 7
FLAG DOC 8
FLAG DOC 9
FLAG DOC 10
FLAG DOC 11
FLAG DOC 12
FLAG DOC 13
FLAG DOC 14
FLAG DOC 15
FLAG DOC 16
FLAG DOC 17
FLAG DOC 18
FLAG DOC 19
FLAG DOC 20
FLAG DOC 21

Application Dataset – NULL values

Dataset

Replacing Blanks in Occupation_Type column of the Application

Dataset with the highest occurring categorical variable



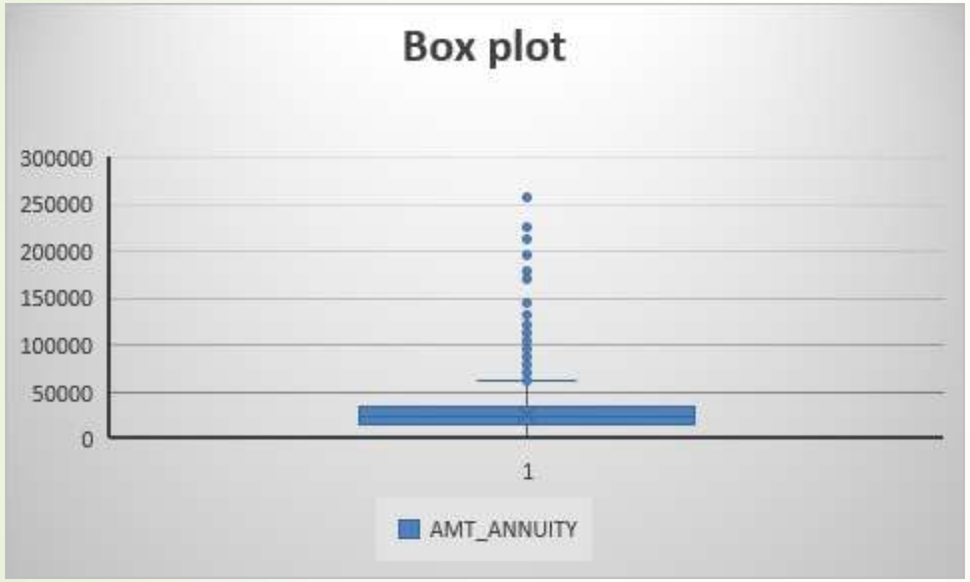
Laborers	8944
Core staff	4428
Accountants	1620
Managers	3482
Drivers	3042
Sales staff	5157
Cleaning staff	739
Cooking staff	963
Private service staff	446
Medicine staff	1403
Security staff	1137
High skill tech staff	1851
Waiters/barmen staff	227
Low-skill Laborers	357
Realty agents	123
Secretaries	212
IT staff	80
HR staff	101
Total	34312

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

Application Dataset – NULL values

Dataset

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

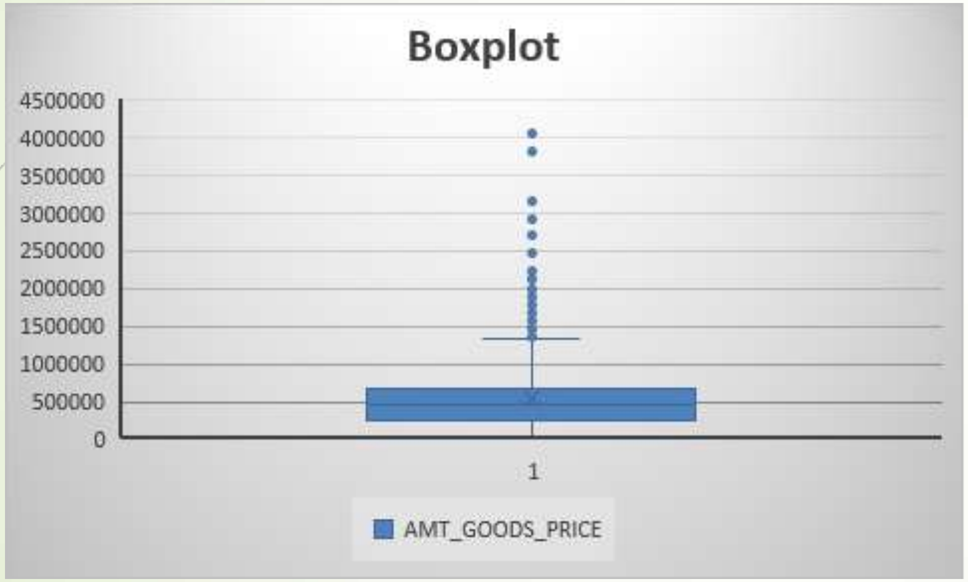


Median	24939

Replacing Blanks with Median

Application Dataset – NULL values
Dataset

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



450000	Median

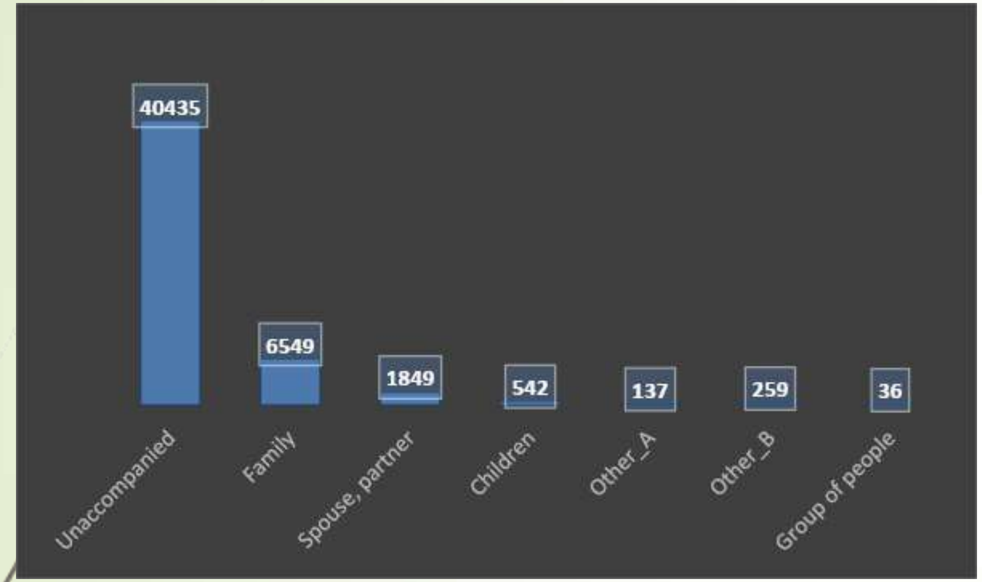
Replaced null values with median

Application Dataset – NULL values

Dataset

Replacing Blanks in Name_Type_Suite column of the Application

Dataset with the highest occurring categorical variable

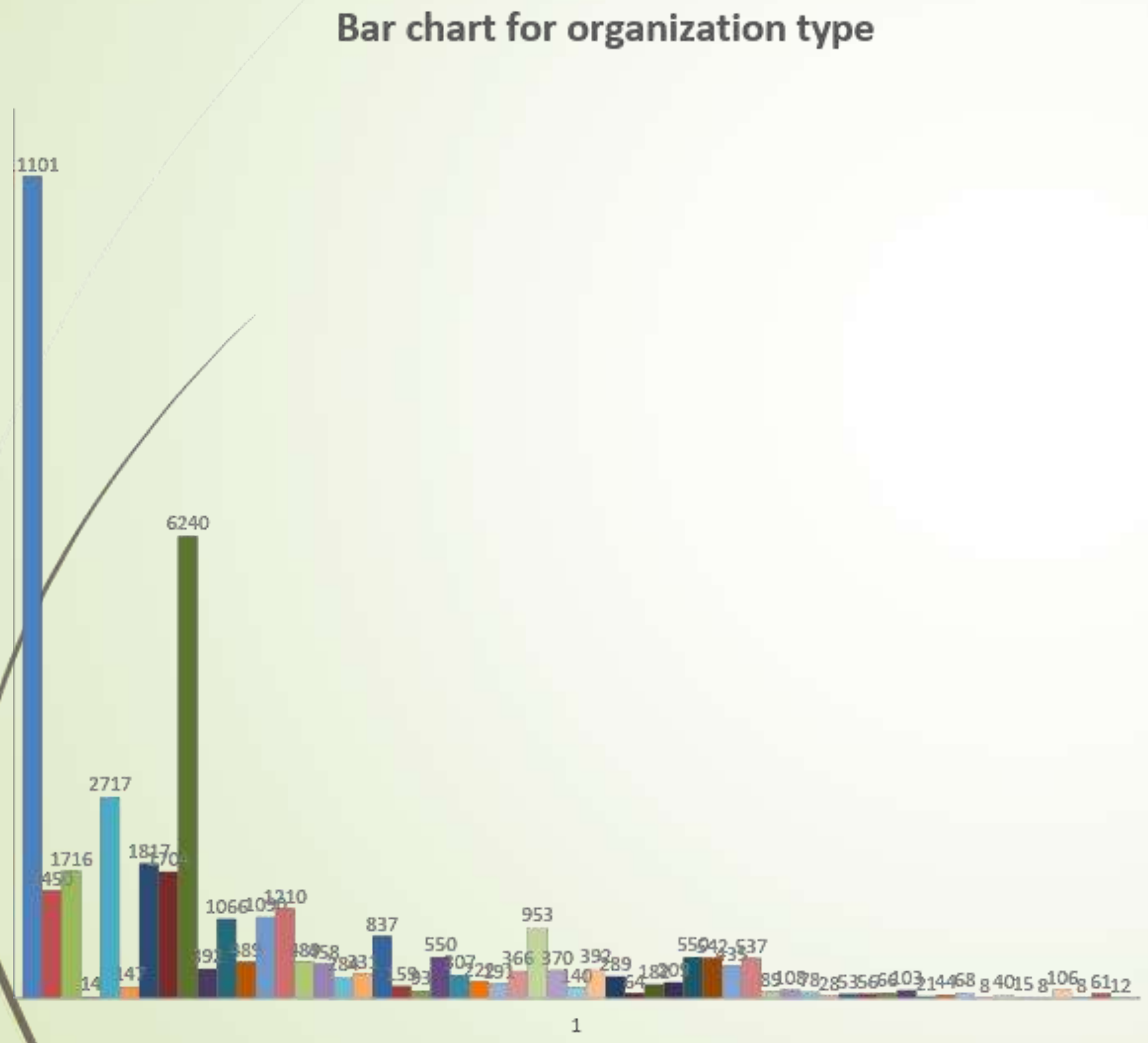


Unaccompanied	40435
Family	6549
Spouse, partner	1849
Children	542
Other_A	137
Other_B	259
Group of people	36

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

Application Dataset – NULL values Dataset

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



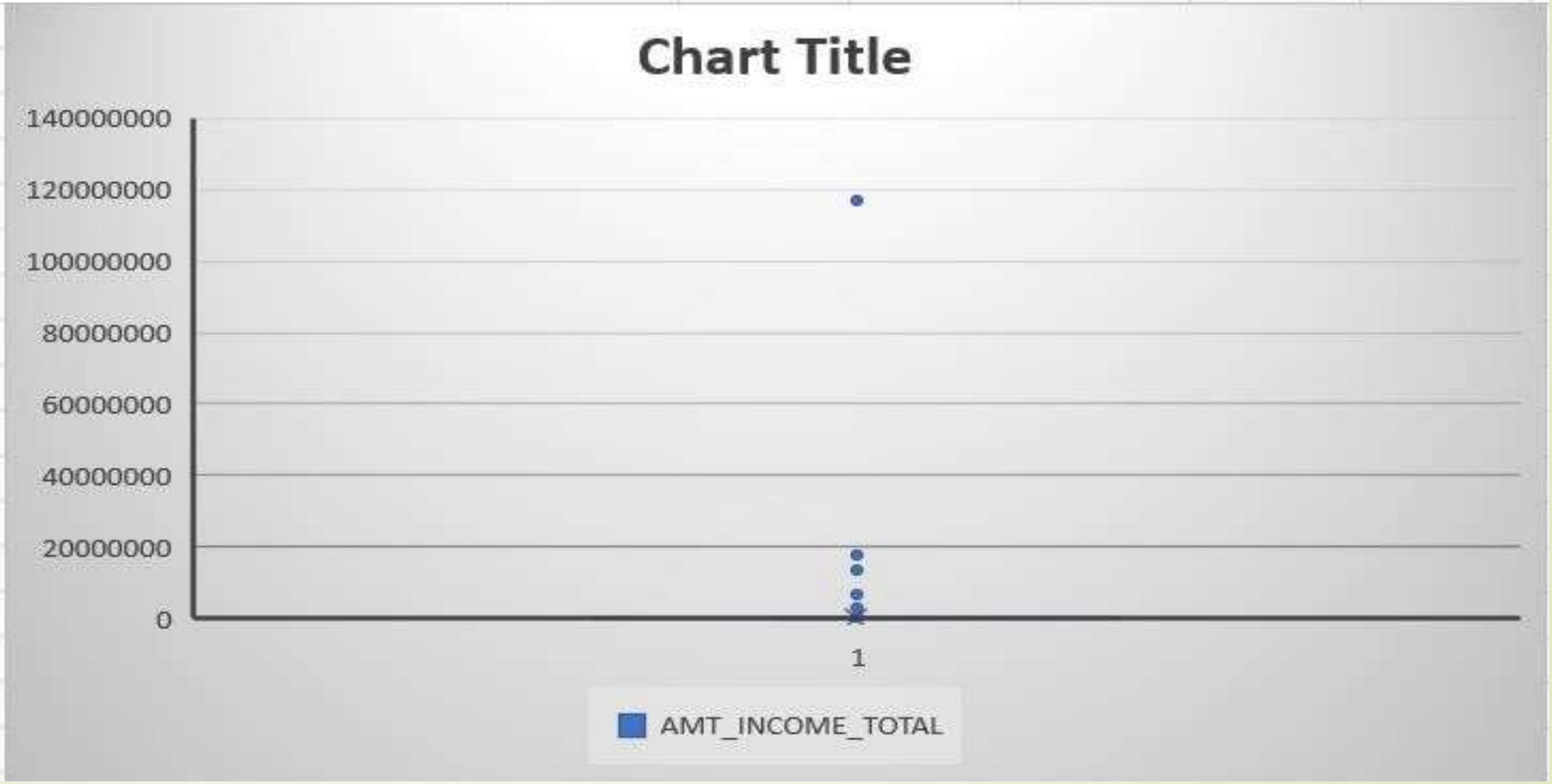
The most commonly occurred organization_type is **Business Entity Type 3**

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

Min	25650
Max	117000000
25%	112500
50%	145800
75%	202500

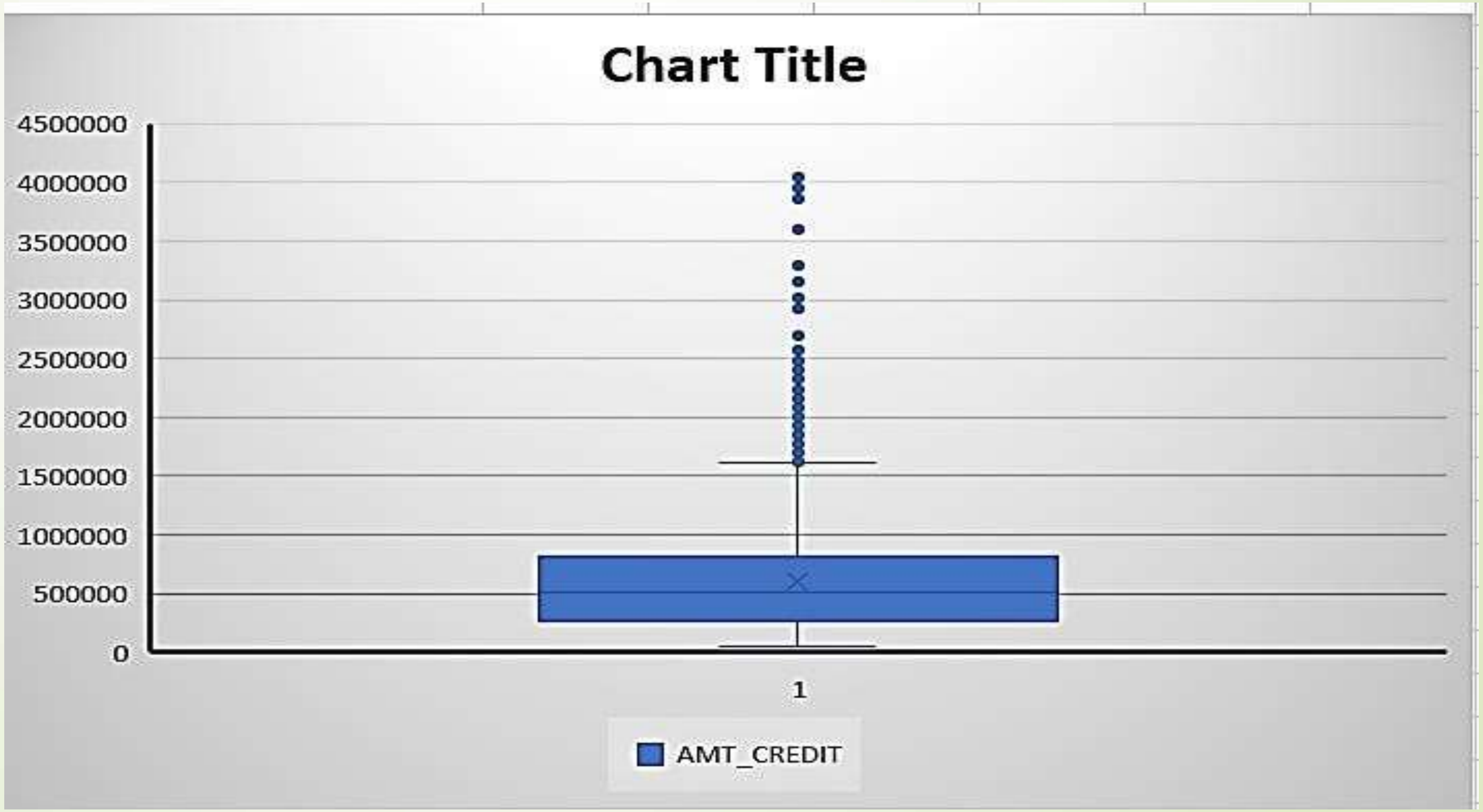
outliers at extreme points i.e. max 1.700x10^8



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

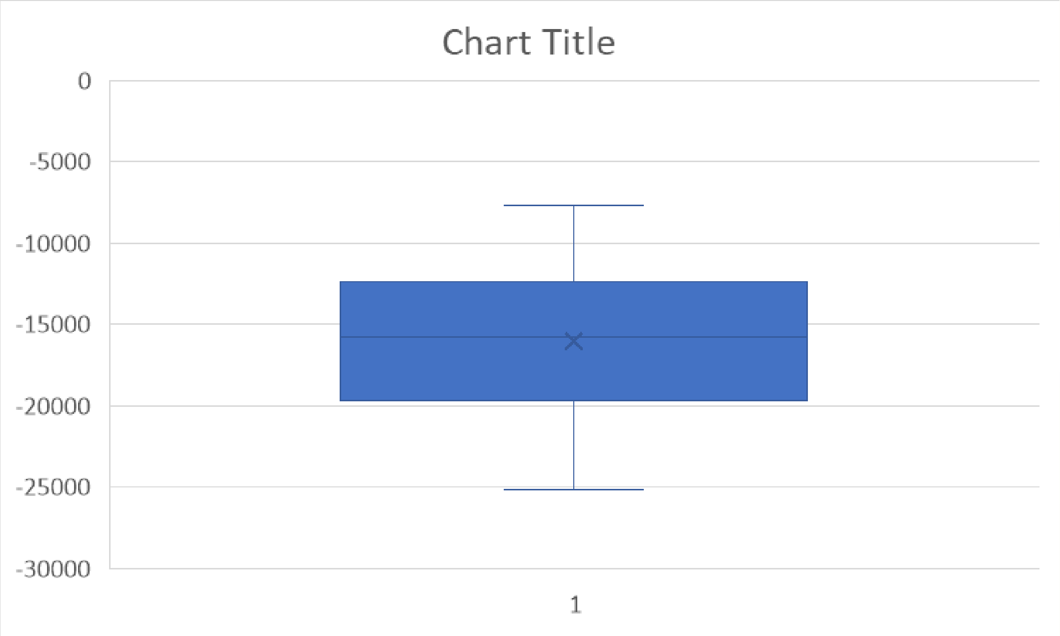
	AMT_CREDIT
	Quartiles at AMT_CREDIT
MIN	45000
25%	270000
50%	513531
75%	808650
MAX	4050000



Application Dataset – Outliers

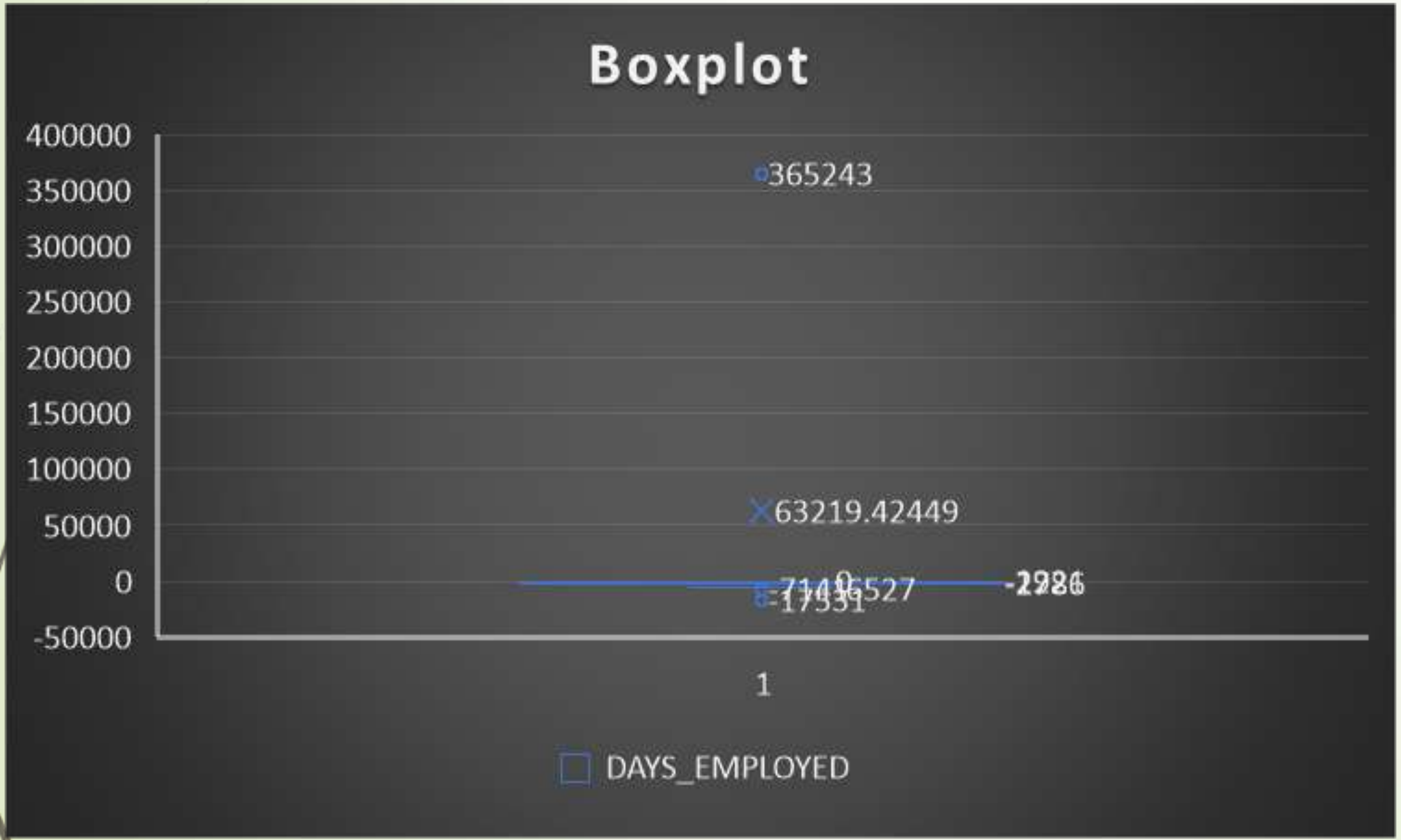
As seen from the boxplot it is clear that there are no outliers

The data of DAYS_BIRTH is well distributed



Application Dataset – Outliers

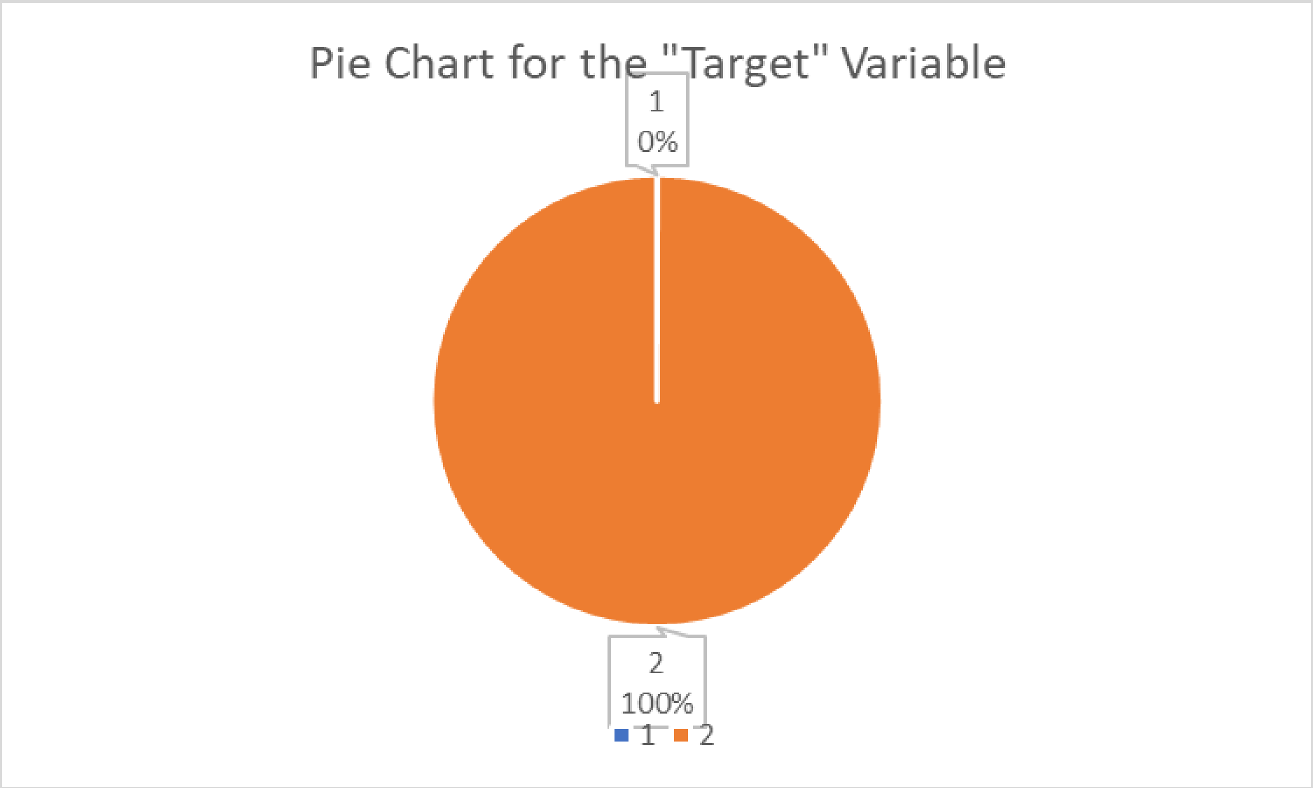
There exists only 1 outlier i.e. + or - 365243
Replace with median
1213.00



Application Dataset – Analysis

TARGET VARIABLE

Row Lables	Count of Target Variable
1	4026
0	45973
Total	49999



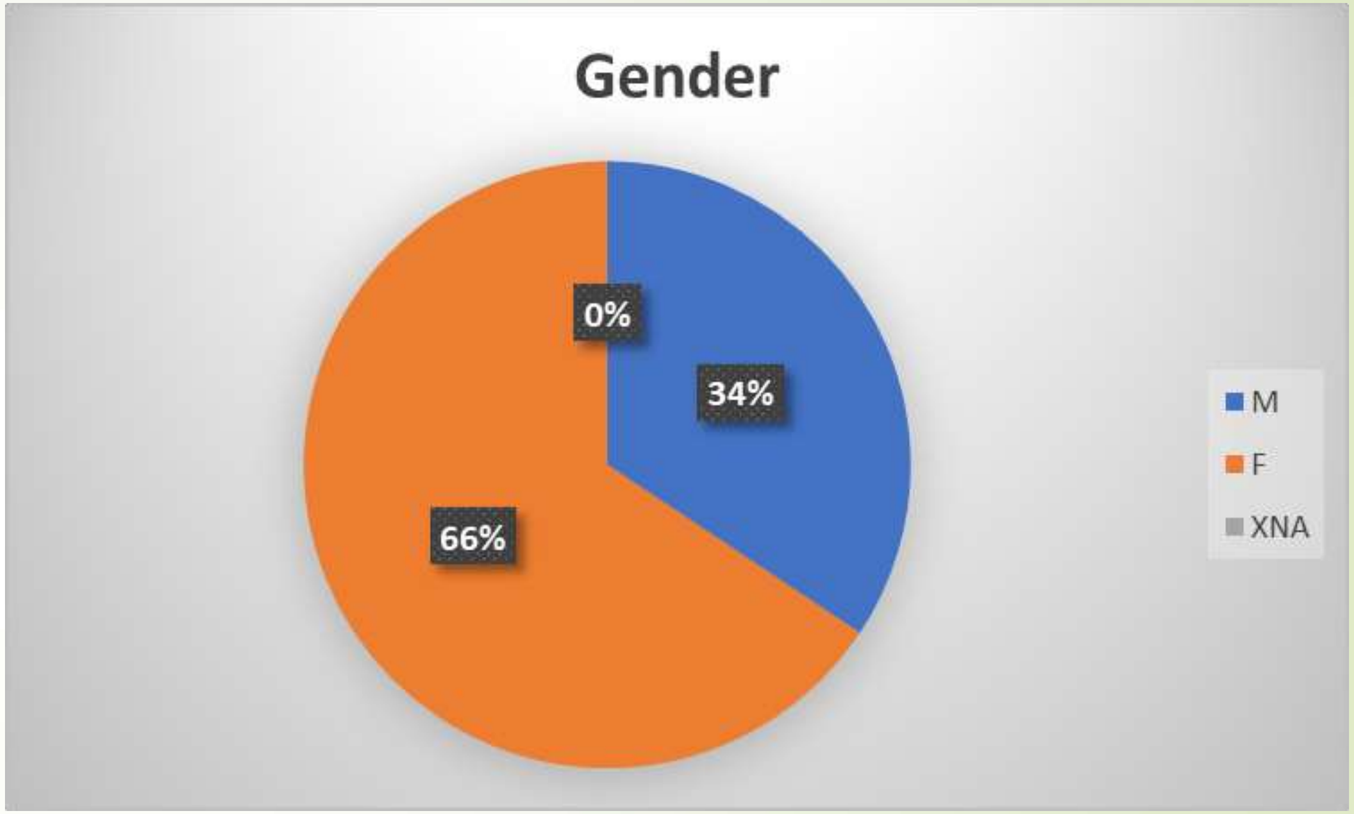
The Target Variable Pie chart shows that almost 100% of the total clients had no problem during payment

1 **No payment issues**
2 **Had some payment issues**

Application Dataset – Analysis

GENDER VARIABLE

Row Labels	Count
M	17174
F	32823
XNA	2
Total	49999



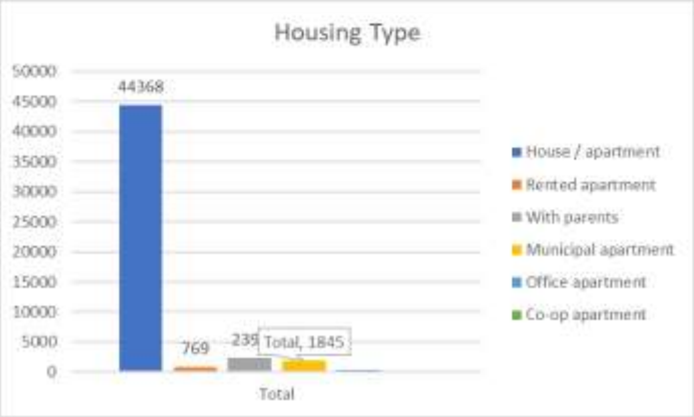
From the GENDER_VARIABLE pie chart we can infer that almost 66% of the clients are female and 34% of the clients are Male

The 2 of the appicants have gender as XNA which can be ignored

Application Dataset – Analysis

NAME_HOUSING_TYPE

NAME_HOUSING_TYPE	Total
House / apartment	44368
Rented apartment	769
With parents	2399
Municipal apartment	1845
Office apartment	427
Co-op apartment	191
Total	49999



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

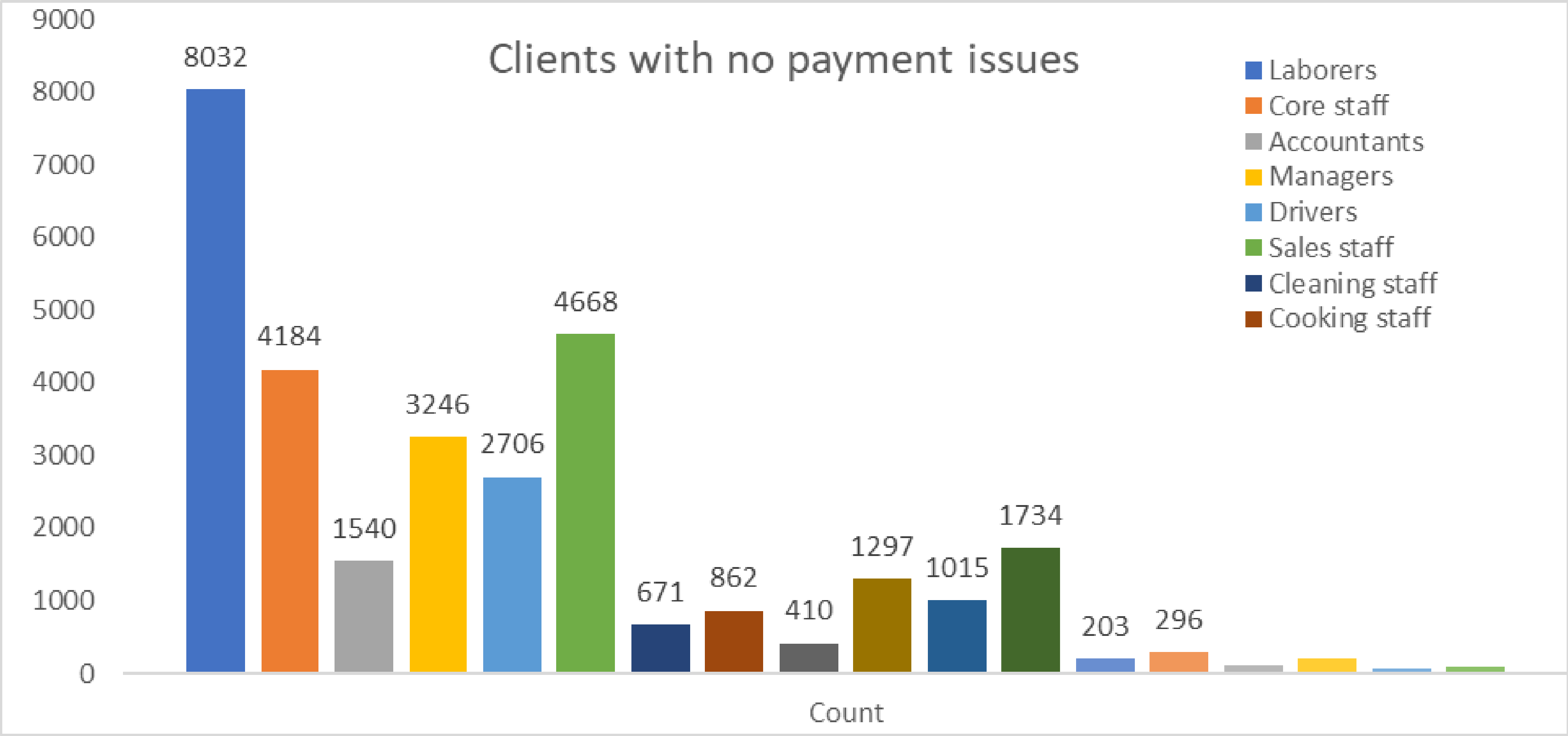
Apartment and people living with their parents

Application Dataset – Analysis

Univariate Analysis

OCCUPATION_TYPE

Occupation_type	Count
Laborers	8032
Core staff	4184
Accountants	1540
Managers	3246
Drivers	2706
Sales staff	4668
Cleaning staff	671
Cooking staff	862
Private service staff	410
Medicine staff	1297
Security staff	1015
High skill tech staff	1734
Waiters/barmen staff	203
Low-skill Laborers	296
Realty agents	110
Secretaries	203
IT staff	76
HR staff	92

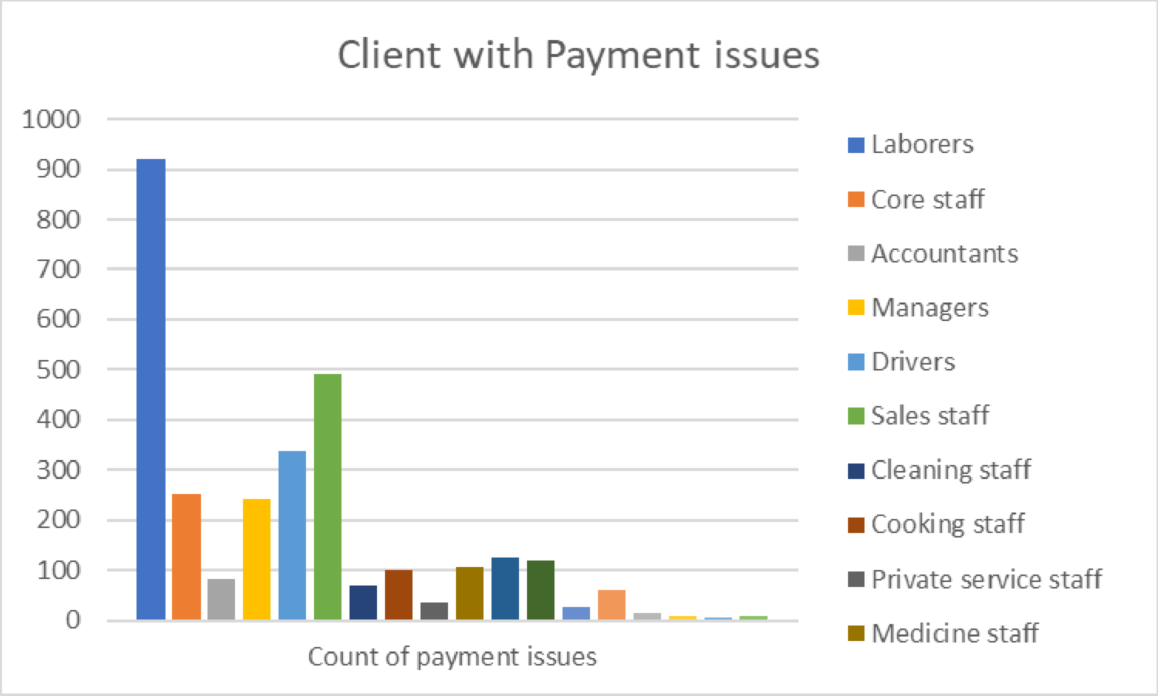


From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with no payment issues

Application Dataset – Analysis

Univariate Analysis

OCCUPATION_TYPE



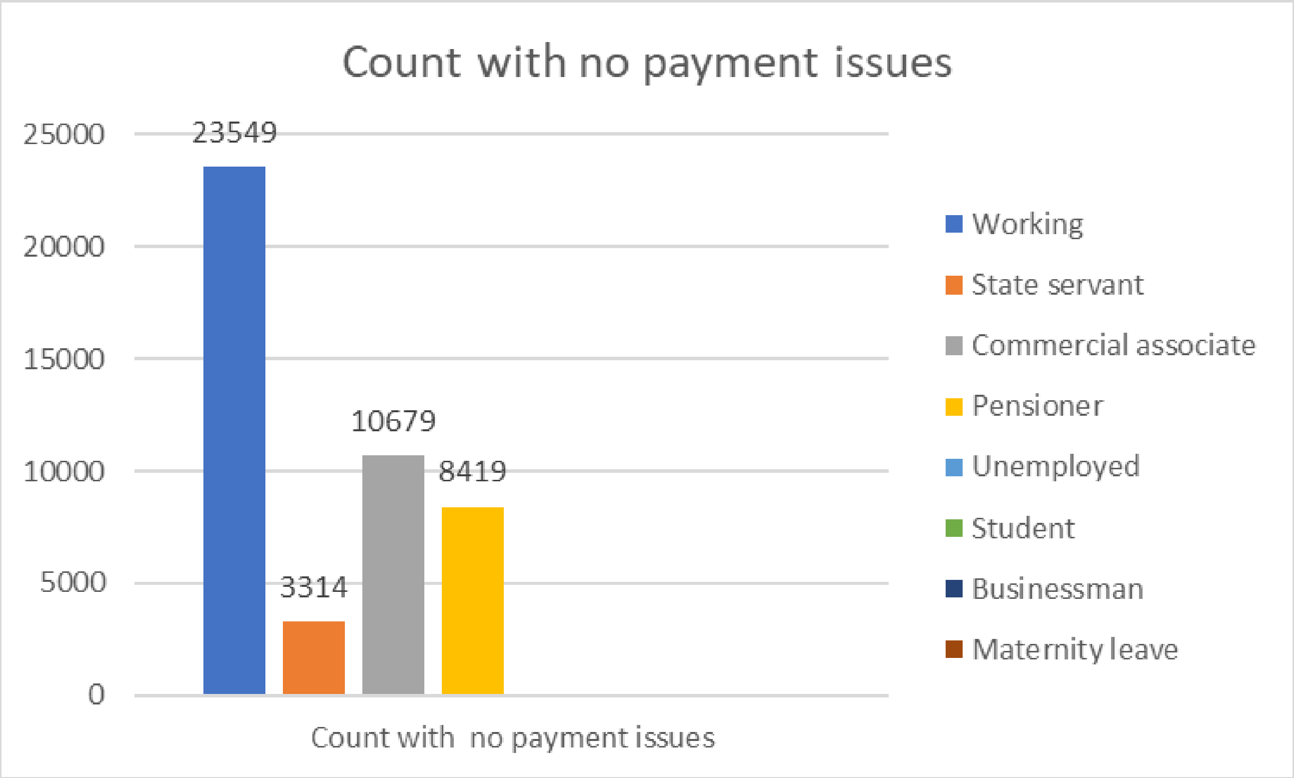
From the above bar plot we can infer that clients with occupation_type 'Laborers' have the highest number of count when it comes to clients with payment issues

Application Dataset – Analysis

Univariate Analysis

NAME_INCOME_TYPE

NAME_INCOME_TYPE	Count with no payment issues
Working	23549
State servant	3314
Commercial associate	10679
Pensioner	8419
Unemployed	4
Student	5
Businessman	2
Maternity leave	1



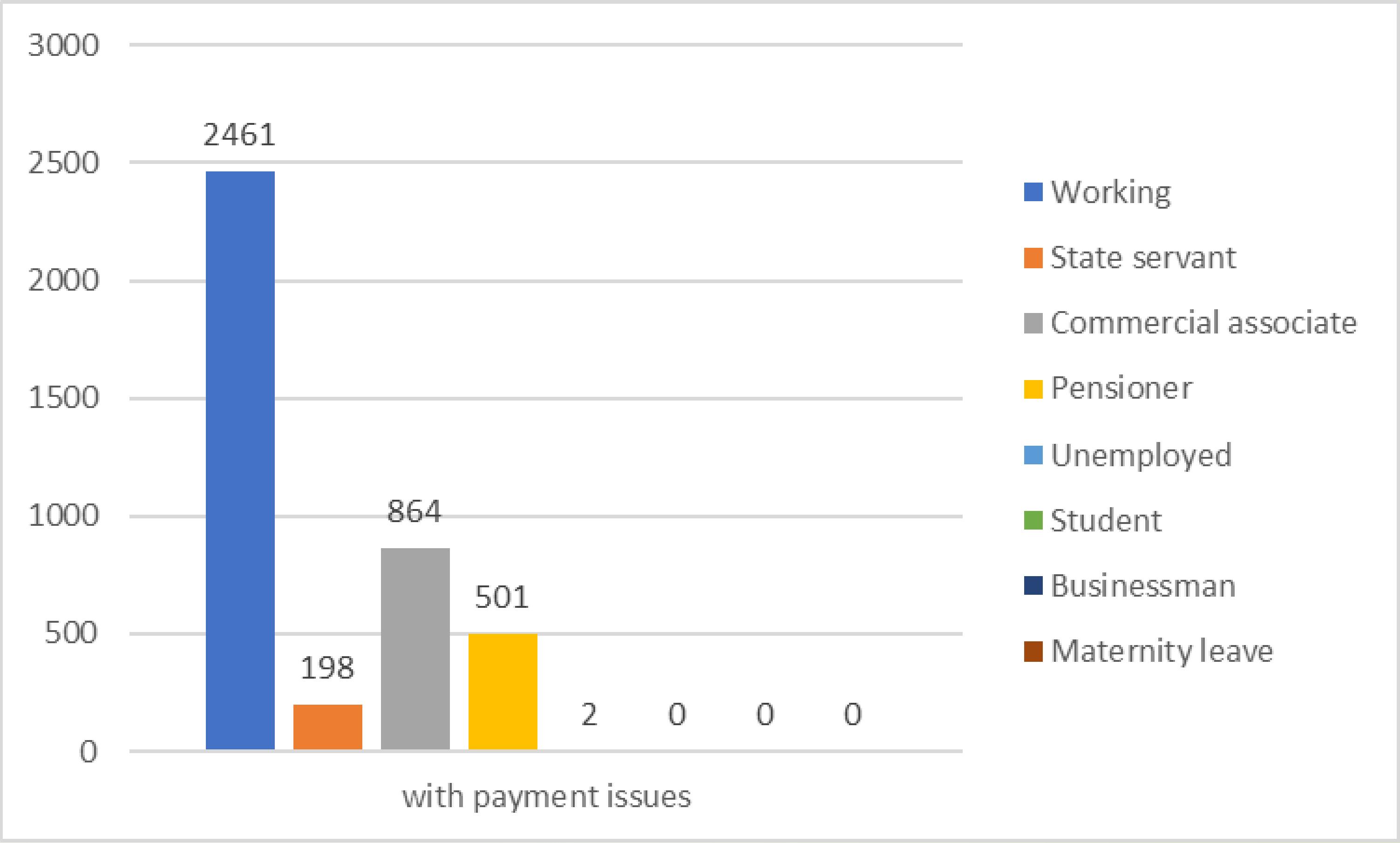
From the above Bar plot we can infer that clients having income_type as 'WORKING' have the highest count when it comes to clients with no payment issues

Application Dataset – Analysis

Univariate Analysis

NAME_INCOME_TYPE

NAME_INCOME_TYPE	Count with no payment issues	with payment issues
Working	23549	2461
State servant	3314	198
Commercial associate	10679	864
Pensioner	8419	501
Unemployed	4	2
Student	5	0
Businessman	2	0
Maternity leave	1	0



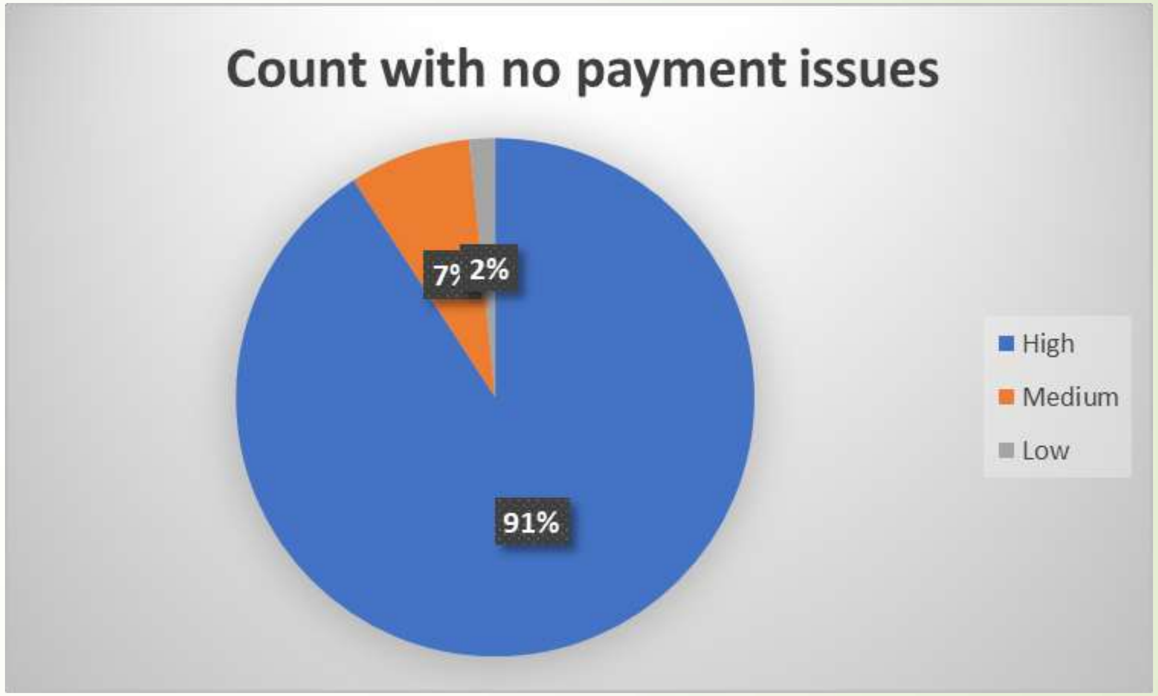
From the above Bar plot we can infer that clients having income_type as 'WORKING' have the

Application Dataset – Analysis

Univariate Analysis

AMT_TOTAL INCOME

amount income total	Count with no payment issues
High	41748
Medium	3484
Low	741



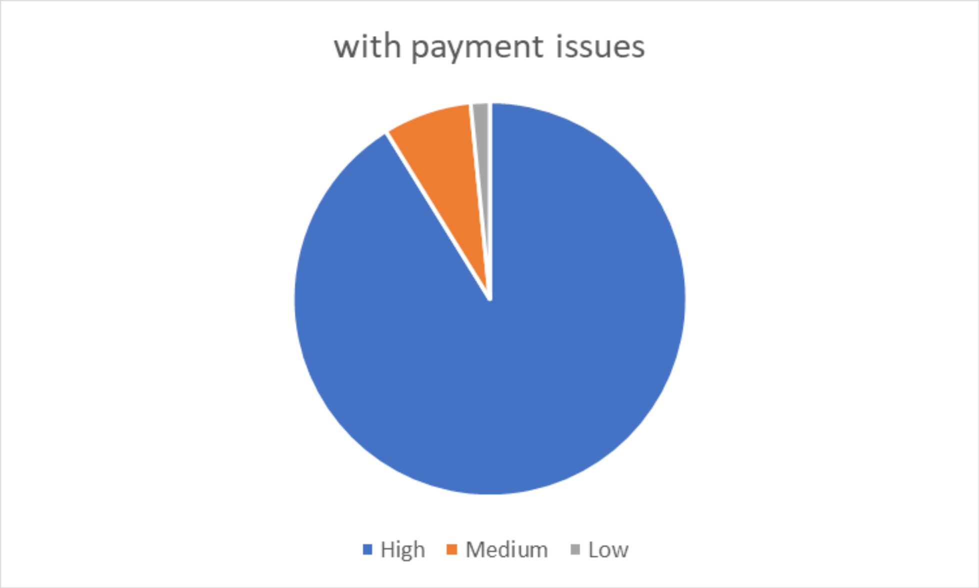
From the above Bar plot we can infer that client having the total income range as 'HIGH' have the highest count when it comes to clients having no payment issues

Application Dataset – Analysis

Univariate Analysis

AMT_TOTAL INCOME

amount income total	Count with no payment issues	with payment issues
High	41748	3670
Medium	3484	293
Low	741	63



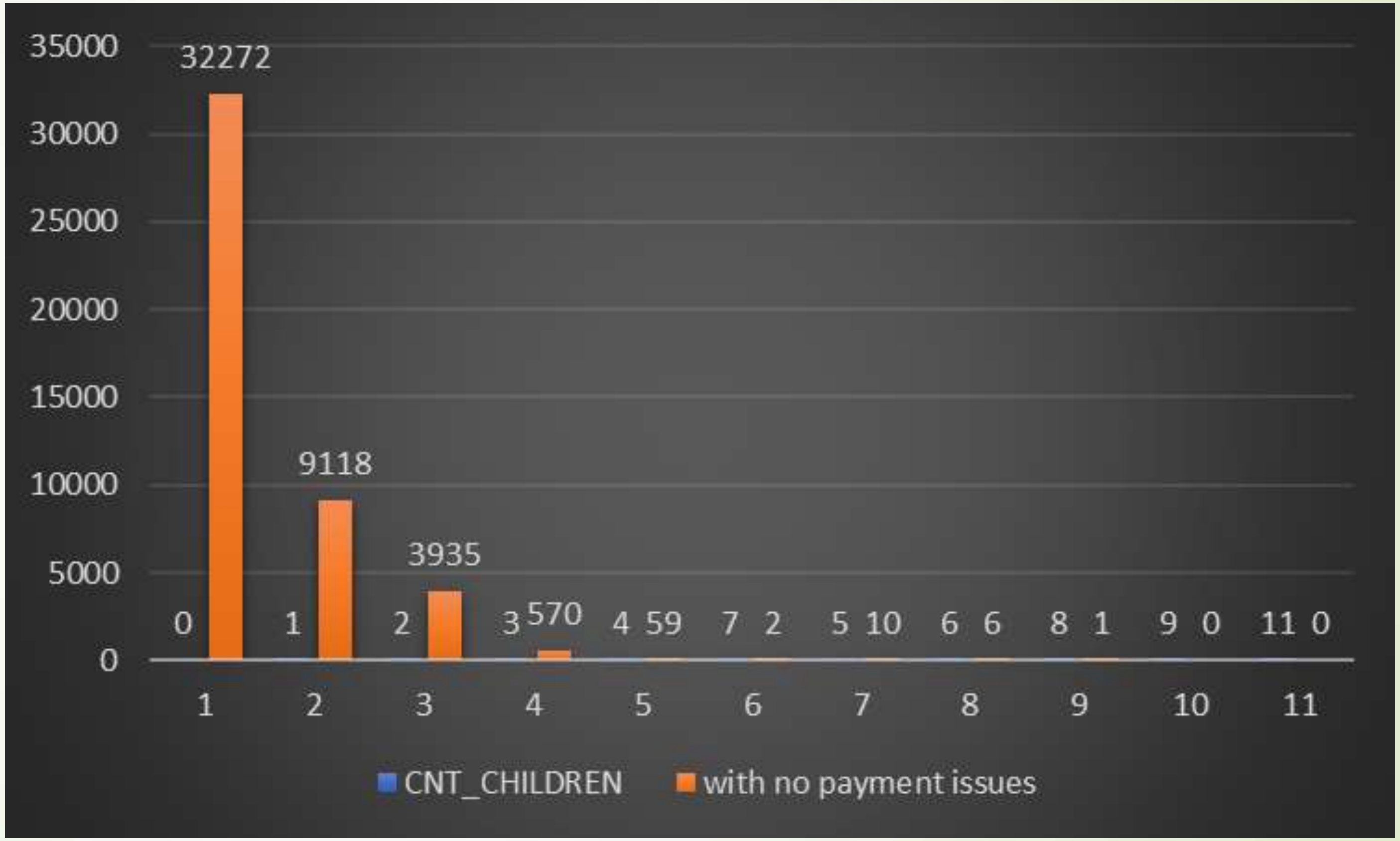
From the above Bar plot we can infer that client having the total income range as 'high"
have the highest count when it comes to clients having payment issues

Application Dataset – Analysis

Univariate Analysis

CNT_FAMILY_MEMBERS

CNT_ CHIL DREN	with no payment issues
0	32272
1	9118
2	3935
3	570
4	59
7	2
5	10
6	6
8	1
9	0
11	0



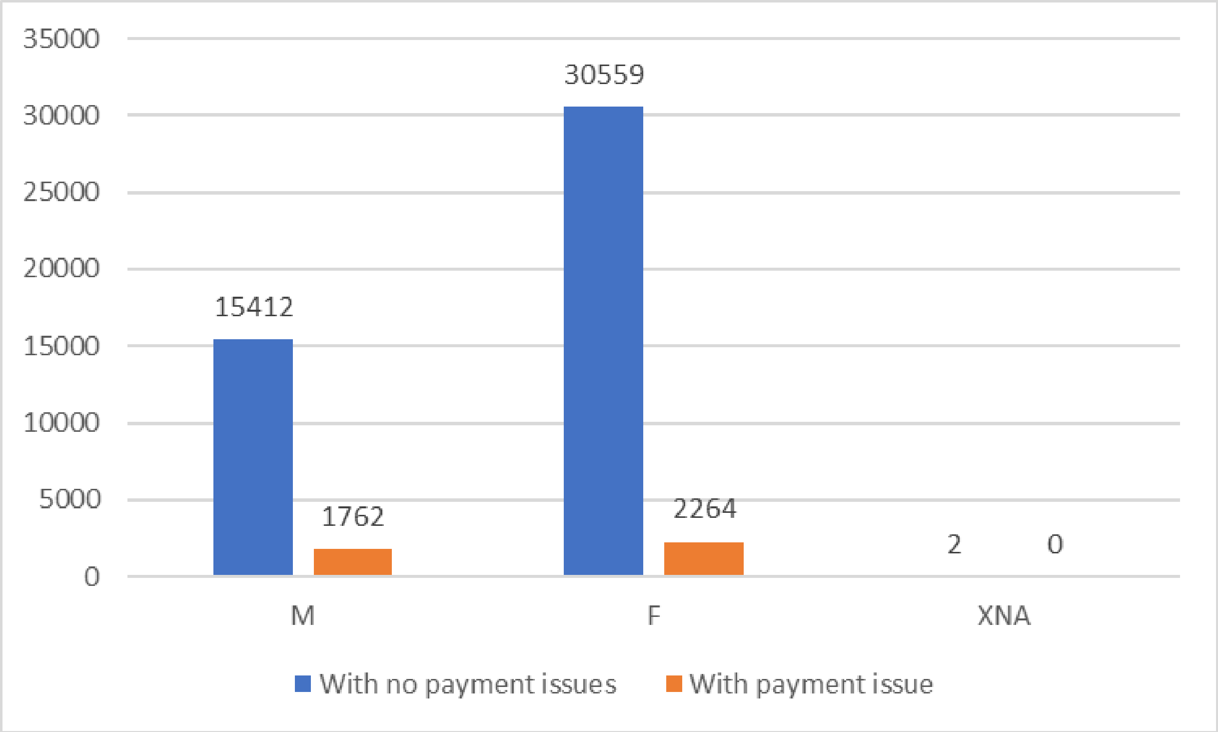
From the above Bar plot we can infer that clients having total count ofchildren as 0 have the highest count when it comes to clients having no payment issues

Application Dataset – Analysis

Univariate Analysis for TARGET variable

CODE_GENDER

CODE_GENDER	With no payment issues	With payment issue
M	15412	1762
F	30559	2264
XNA	2	0



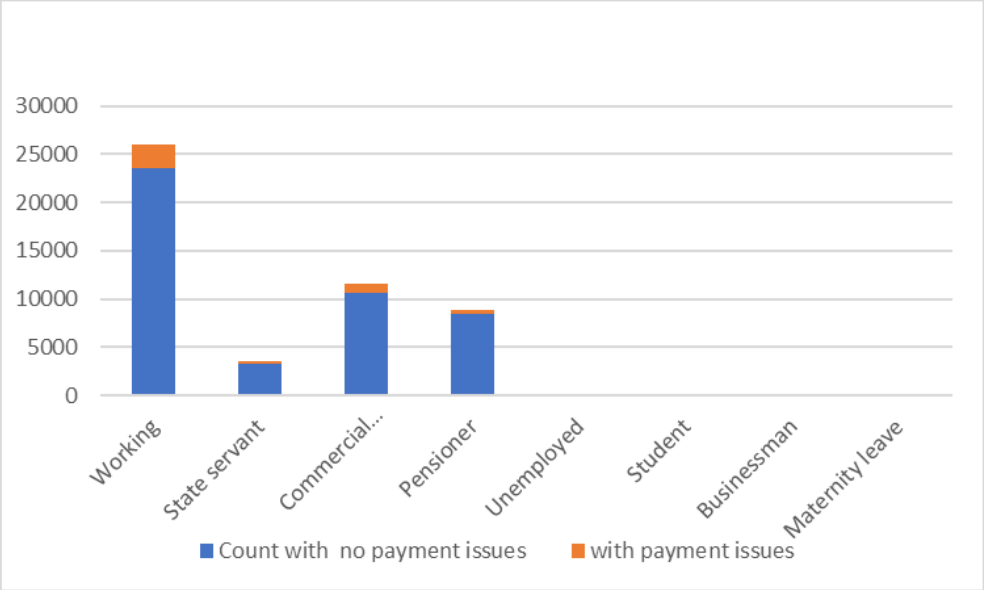
From the above Bar Plot we can infer that Clients with CODE_GENDER = 'F' have the highest number of non-defaulters i.e. 13650

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_INCOME_TYPE

NAME_INCOME_TYPE	Count with no payment issues	with payment issues
Working	23549	2461
State servant	3314	198
Commercial associate	10679	864
Pensioner	8419	501
Unemployed	4	2
Student	5	0
Businessman	2	0
Maternity leave	1	0

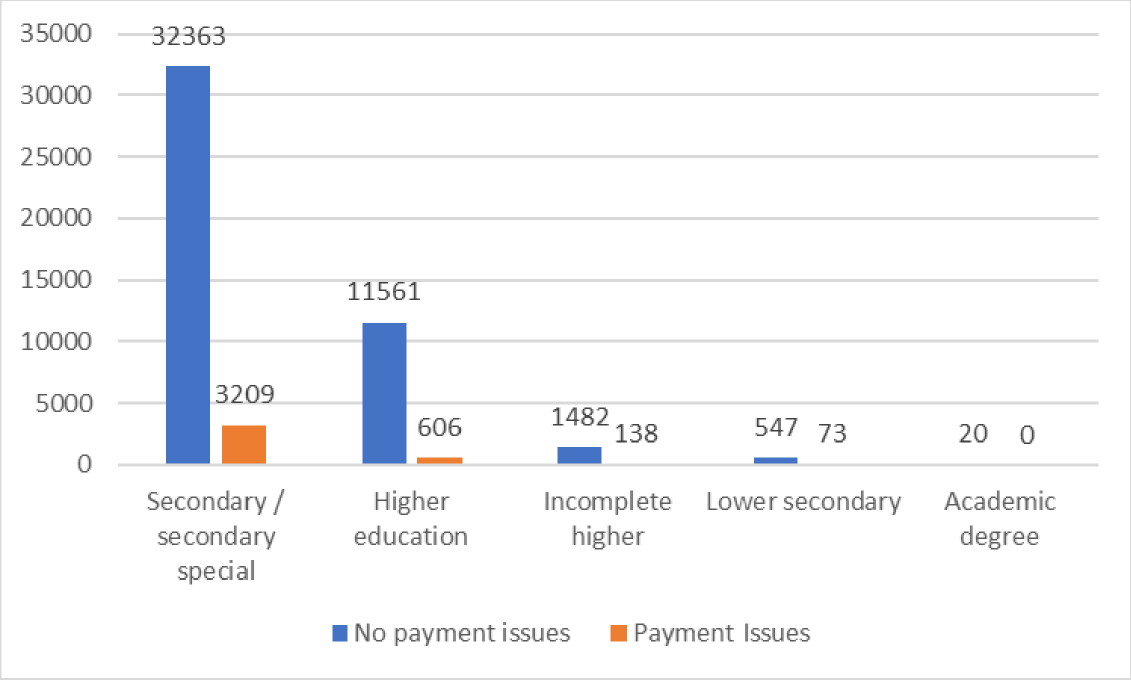


From the adjacent Bar Plot we can infer that clients having NAME_INCOME_TYPE = 'WORKING' having the highest count of Non-defaulters i.e.
23549-2461

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_EDUCATION_TYPE	No payment issues	Payment Issues
Secondary / secondary special	32363	3209
Higher education	11561	606
Incomplete higher	1482	138
Lower secondary	547	73
Academic degree	20	0



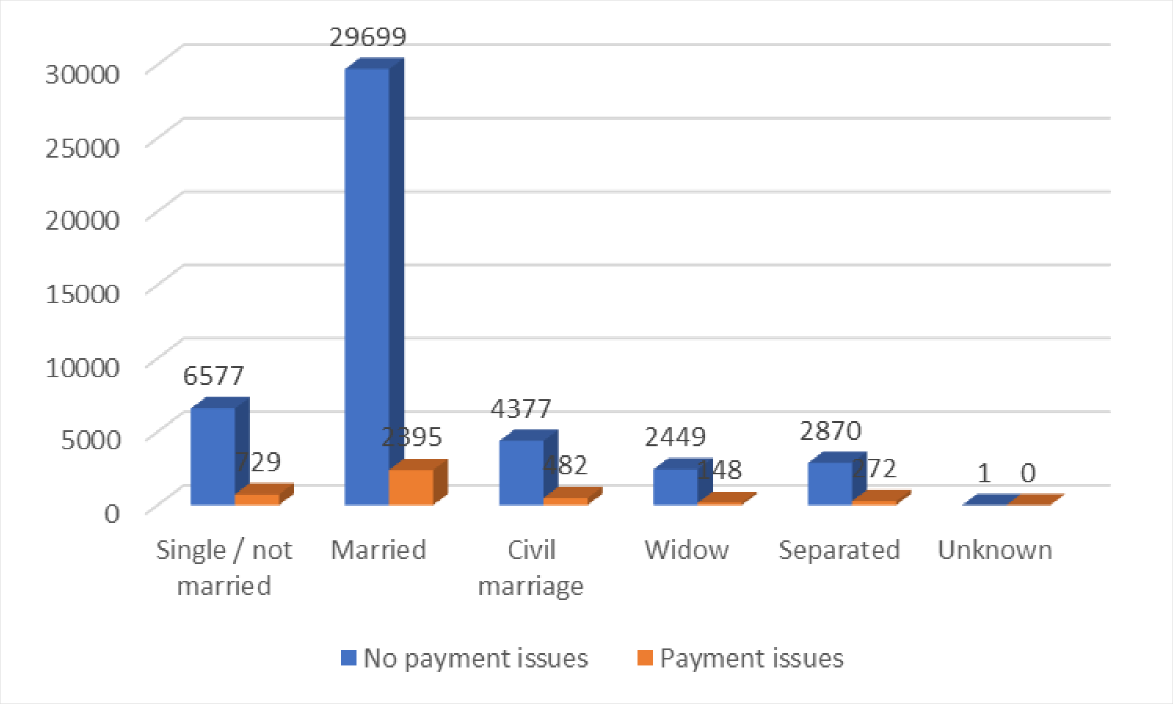
From the above Bar Plot we can infer that clients having NAME_EDUCATION_TYPE = 'SECONDARY/SECONDARY SPECIAL' have the highest count for Non- defaulters i.e. 29154

Application Dataset – Analysis

Univariate Analysis for TARGET variable

NAME_FAMILY_STATUS

NAME_FAMILY_STATUS	No payment issues	Payment issues
Single / not married	6577	729
Married	29699	2395
Civil marriage	4377	482
Widow	2449	148
Separated	2870	272
Unknown	1	0



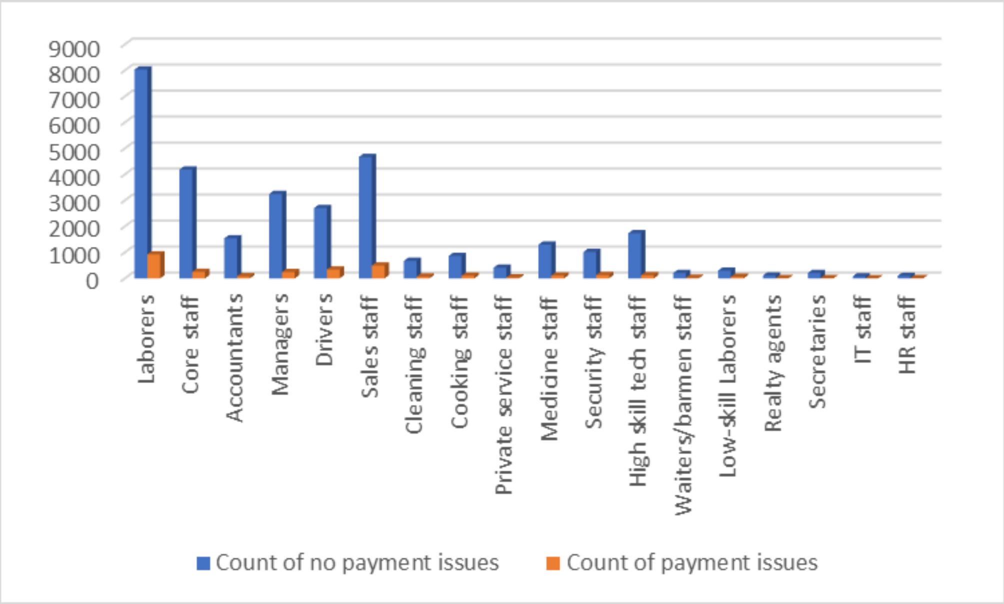
From the adjacent Bar Plot we can infer that clients having NAME_FAMILY_STATUS = 'MARRIED' have the highest count of Non- defaulters i.e. 27304

Application Dataset – Analysis

Univariate Analysis for TARGET variable

OCCUPATION_TYP E

Occupation_type	Count of no payment issues	Count of payment issues
Laborers	8032	920
Core staff	4184	250
Accountants	1540	81
Managers	3246	243
Drivers	2706	338
Sales staff	4668	492
Cleaning staff	671	68
Cooking staff	862	101
Private service staff	410	37
Medicine staff	1297	106
Security staff	1015	125
High skill tech staff	1734	118
Waiters/barmen staff	203	25
Low-skill Laborers	296	61
Realty agents	110	13
Secretaries	203	9
IT staff	76	4
HR staff	92	9



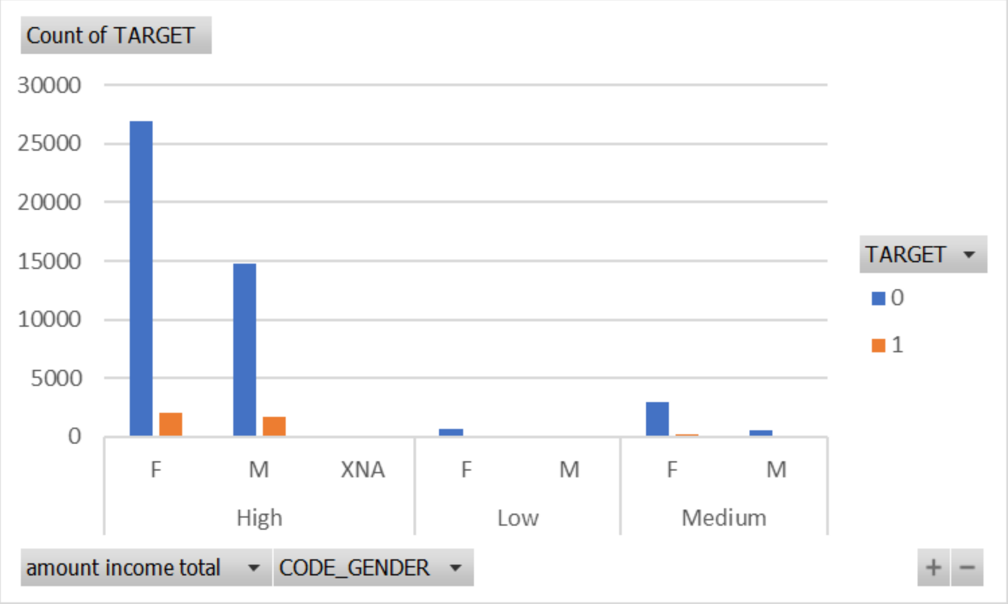
From the adjacent Bar plot we can infer that clients having occupation_type = 'Laborers' have the highest count for Non-defaulters i.e. 7112

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Total_income_range vs Code_gender

Count of TARGET		TARGET		
amount incom	CODE_GENDER	0	1	Grand Total
High	F	26975	1997	28972
	M	14771	1673	16444
	XNA	2		2
Low	F	617	43	660
	M	124	20	144
Medium	F	2967	224	3191
	M	517	69	586
Grand Total		45973	4026	49999



0

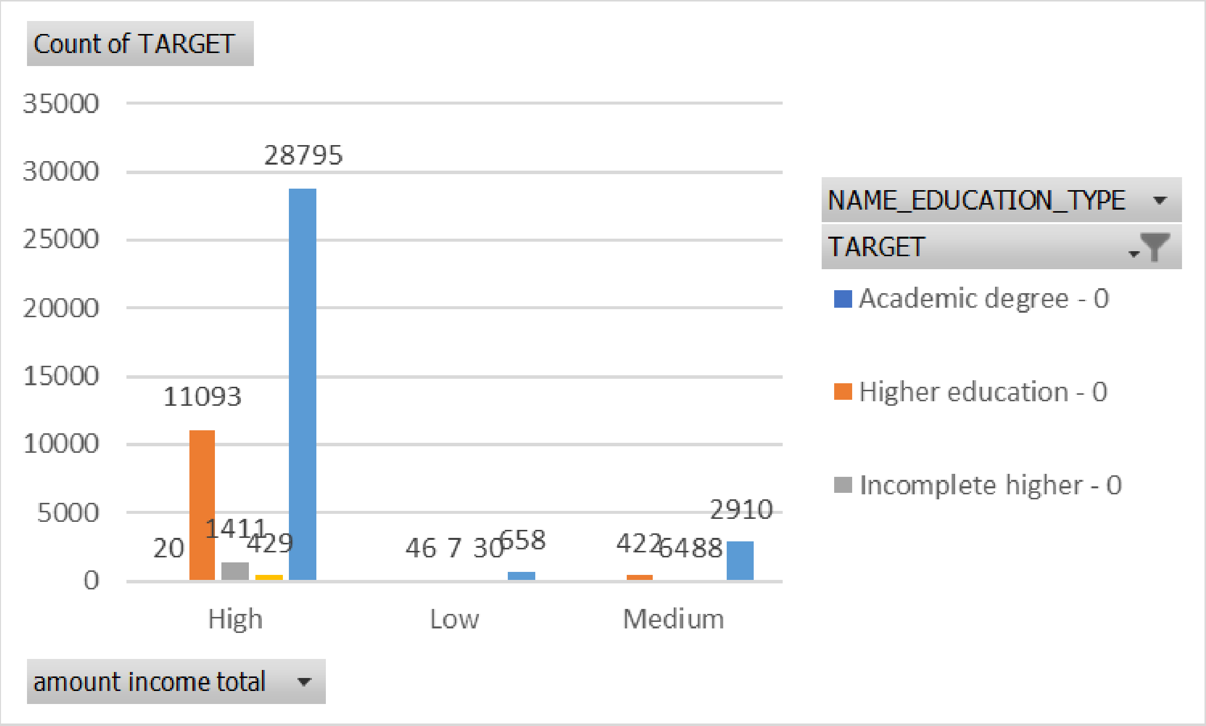
From the above Bar plot we can infer that Females belonging to Low income group are the highest number of clients with no payment issues a

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Credit Amt vs Education status

Count of TARGET	NAME_EDUCATION_TYPE / TARGET					
	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special	Grand Total
amount income total	0	0	0	0	0	
High	20	11093	1411	429	28795	41748
Low		46	7	30	658	741
Medium		422	64	88	2910	3484
Grand Total	**20**	**11561**	**1482**	**547**	**32363**	**45973**



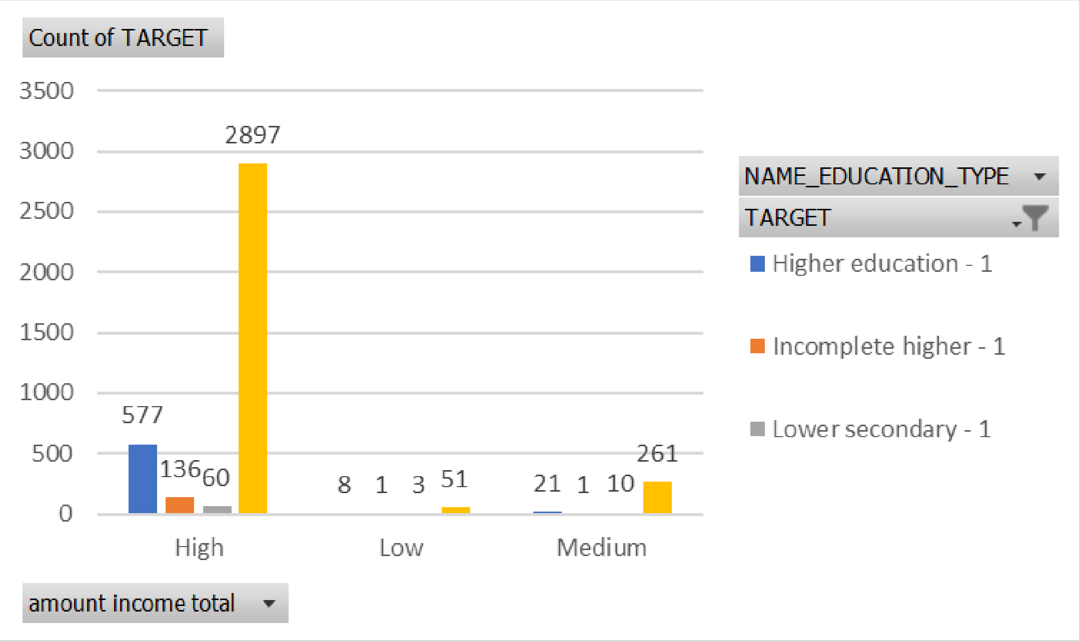
From the adjacent Bar Plot we can infer that clients having credit amt range as 'High' and education status as 'Secondary/ Secondary Special' have the highest count for clients with no payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 1: Credit Amt vs Education status

Count of TARGET	NAME_EDUCATION_TYPE	TARGET			
	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special	Grand Total
amount income total	1	1	1	1	
High	577	136	60	2897	3670
Low	8	1	3	51	63
Medium	21	1	10	261	293
Grand Total	606	138	73	3209	4026



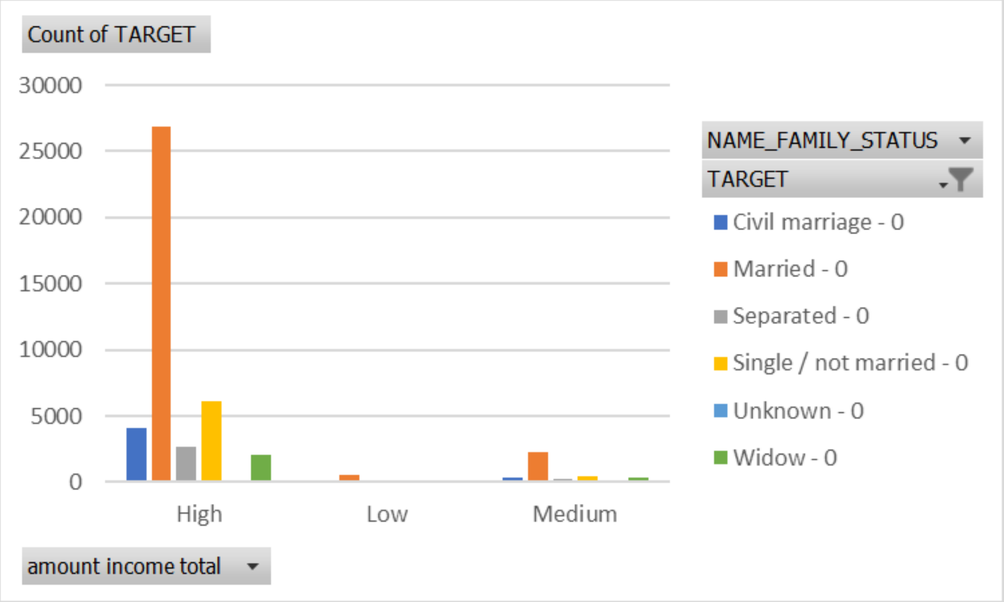
From the adjacent Bar Plot we can infer that clients having credit amt range as 'High' and education status as 'Secondary/ Secondary Special' have the highest count for clients with payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 0: Total Income vs Family status

Count of TARGET	NAME_FAMILY_STATUS / TARGET						
	Civil marriage	Married	Separated	Single / not married	Unknown	Widow	Grand Total
amount income total	0	0	0	0	0	0	
High	4038	26881	2641	6111	1	2076	41748
Low	41	541	29	52		78	741
Medium	298	2277	200	414		295	3484
Grand Total	4377	29699	2870	6577	1	2449	45973



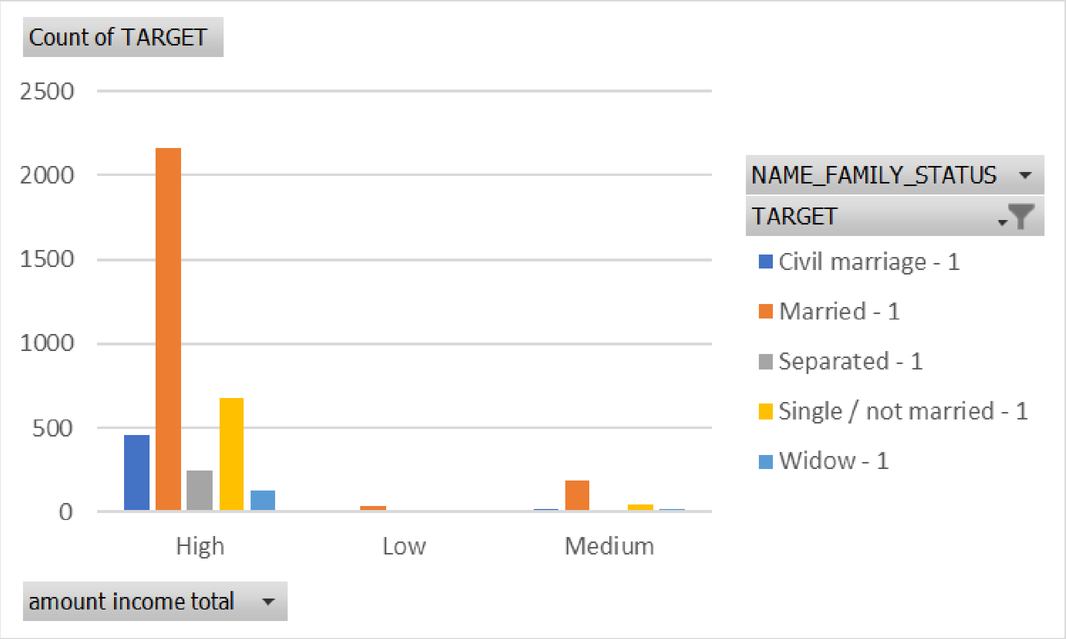
From the adjacent Bar plot we can infer that clients with total_income_range as 'High' and family_status as 'Married' have the highest count for clients having no payment issues

Application Dataset – Analysis

Bivariate Analysis for TARGET variable

Target 1: Total Income vs Family status

Count of TARGET	NAME_FAMILY_STATUS / TARGET					
	Civil marriage	Married	Separated	Single / not married	Widow	Grand Total
amount income total	1	1	1	1	1	
High	455	2164	251	674	126	3670
Low	5	39	6	9	4	63
Medium	22	192	15	46	18	293
Grand Total	482	2395	272	729	148	4026



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

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.

	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION
TARGET	1													
CNT_CHILDREN	0.02636393	1												
AMT_INCOME_TOTAL	0.01089374	0.009588558	1											
AMT_CREDIT	-0.0324283	0.00497156	0.069315897	1										
AMT_ANNUITY	-0.0123991	0.026178823	0.083008508	0.76949891	1									
AMT_GOODS_PRICE	-0.0413065	0.000253361	0.069885575	0.98694373	0.774433947	1								
REGION_POPULATION_RELATIVE	-0.0407992	-0.025555665	0.029841469	0.09511122	0.115111507	0.099190323	1							
DAYS_BIRTH	0.07678768	0.329263754	0.016002774	-0.0593427	0.007712245	-0.057610698	-0.032513748	1						
DAYS_EMPLOYED	-0.0402949	-0.239693041	-0.031615555	-0.0704714	-0.110449038	-0.067845372	-0.004101686	-0.613554	1					
DAYS_REGISTRATION	0.04234268	0.181217183	0.009952379	0.00344857	0.033218936	0.006101039	-0.059322344	0.33363251	-0.204680611	1				
DAYS_ID_PUBLISH	0.04692674	-0.032115773	0.003506646	-0.0122288	0.006716454	-0.013967763	-0.004345136	0.27082514	-0.270382022	0.104298561	1			
HOUR_APPR_PROCESS_START	-0.0320365	-0.006253862	0.01846417	0.05667698	0.053274911	0.066006129	0.167725422	0.09058852	-0.088041494	-0.007923326	0.033747646	1		
REG_REGION_NOT_LIVE_REGION	0.00943872	-0.010655142	0.013772946	0.02577159	0.044803016	0.028022722	-0.003551805	0.05910451	-0.036571015	0.027220331	0.032789191	0.050667774	1	
REG_REGION_NOT_WORK_REGION	-0.0010064	0.012057308	0.027597461	0.0538458	0.081294649	0.055274119	0.060070269	0.09398543	-0.105581514	0.033222575	0.047170119	0.073773517	0.456004449	1
LIVE_REGION_NOT_WORK_REGION	-0.0054979	0.019659478	0.026151596	0.05326216	0.074849042	0.053565657	0.085696541	0.06811994	-0.093697402	0.022333016	0.033126512	0.060378475	0.082027392	0.857141677
REG_CITY_NOT_LIVE_CITY	0.0387731	0.019192202	0.00013564	-0.0249779	-0.006720829	0.02450719	-0.046481668	0.1821089	-0.093212442	0.068076021	0.075613949	0.017045445	0.335486228	0.154537611
REG_CITY_NOT_WORK_CITY	0.04845079	0.070031796	-2.0613E-05	-0.0173774	0.001060535	-0.018650803	-0.040442999	0.23789671	-0.254363741	0.094100615	0.102338904	0.023227801	0.143364173	0.235678978
LIVE_CITY_NOT_WORK_CITY	0.03226132	0.067750525	0.001294886	0.00213578	0.010940322	0.000332962	-0.013596017	0.15055154	-0.217007584	0.063076045	0.062527981	0.011719442	0.003095157	0.190031391
EXT_SOURCE_2	-0.1584243	-0.017641055	0.019517645	0.13812532	0.128928233	0.146936283	0.201241779	-0.0938822	-0.026141544	-0.060998098	-0.047546534	0.157147521	0.01694439	0.028985033
FLOORSMAX_AVG	-0.0407069	-0.011049964	0.023707935	0.10217051	0.131697863	0.107267192	0.312577533	0.0061698	-0.023148459	0.053739494	0.002229333	0.114201316	0.011893385	0.036865906
FLOORSMAX_MODE	-0.0394183	-0.0093124	0.022748888	0.09960313	0.128314923	0.104456903	0.29479127	0.00605383	-0.022800106	0.053076306	0.00200631	0.110908867	0.012619829	0.037209933
FLOORSMAX_MEDI	-0.039939	-0.01039189	0.023619011	0.10159087	0.13024317	0.106550439	0.309406862	0.00693889	-0.022526761	0.05459385	0.002796091	0.112756129	0.011644107	0.037441254
OBS_30_CNT_SOCIAL_CIRCLE	0.0141799	0.016616011	-0.008621505	0.00180568	-0.009325084	0.001503912	-0.018011384	0.01168188	0.004948923	0.0102734	-0.012271066	-0.008882612	-0.016523133	-0.025748377
DEF_30_CNT_SOCIAL_CIRCLE	0.04160309	-0.002964877	-0.007629245	-0.0161974	-0.021818679	-0.017494297	0.00927643	0.00189866	0.016367961	0.004953708	0.001267352	-0.001398603	-0.00597519	-0.007859428
OBS_60_CNT_SOCIAL_CIRCLE	0.01394542	0.016499856	-0.008590555	0.00216318	-0.009011554	0.001793772	-0.016896943	0.01156461	0.004878693	0.010551928	-0.012464099	-0.008884665	-0.016540821	-0.025874283
DEF_60_CNT_SOCIAL_CIRCLE	0.04425977	-0.00395542	-0.007343829	-0.0211519	-0.025276301	-0.021742984	0.003999907	0.00284222	0.01526564	0.006862911	0.001559858	-0.004748971	-0.007143874	-0.011727416
AMT_REQ_CREDIT_BUREAU_HOUR	0.00325823	0.001962836	0.000718114	-0.0004152	0.01231602	0.0001482	-0.002398582	0.00505391	-0.004192452	-0.002899095	0.007149958	-0.010883777	-0.003036368	0.002416539
AMT_REQ_CREDIT_BUREAU_DAY	0.01195613	-0.002127186	0.00098241	0.01084775	0.007005406	0.011115919	-0.000998469	0.00185473	0.005467039	-0.002895391	0.005937983	0.009585075	-0.004736951	0.002080597
AMT_REQ_CREDIT_BUREAU_WEEK	0.00573127	0.000181483	0.000264227	0.00169082	0.02018245	0.002174202	0.00363998	-0.0004191	-0.004249294	-0.000396103	0.00329626	-0.008538022	8.56894E-05	0.002455388
AMT_REQ_CREDIT_BUREAU_MON	-0.0113555	-0.013532921	0.011521369	0.08405784	0.040962609	0.065821111	0.078138453	0.00026739	-0.033607312	-0.01114222	-0.00627554	0.03267141	-0.002991648	0.004592531
AMT_REQ_CREDIT_BUREAU_QRT	-0.0008061	-0.009174798	-0.000441699	0.01647665	0.006668109	0.01742399	-0.007843893	-0.0165092	0.018661699	0.003013469	-0.012459786	-0.003785099	0.001448056	-0.00785144

Application Dataset – Analysis

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

Work Done:https://docs.google.com/spreadsheets/d/14gvQJOW7eMiKNL2F44RyOF2wzLA22f2U/edit?usp=drive_link&ouid=101365768232666009716&rtpof=true&sd=true

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**

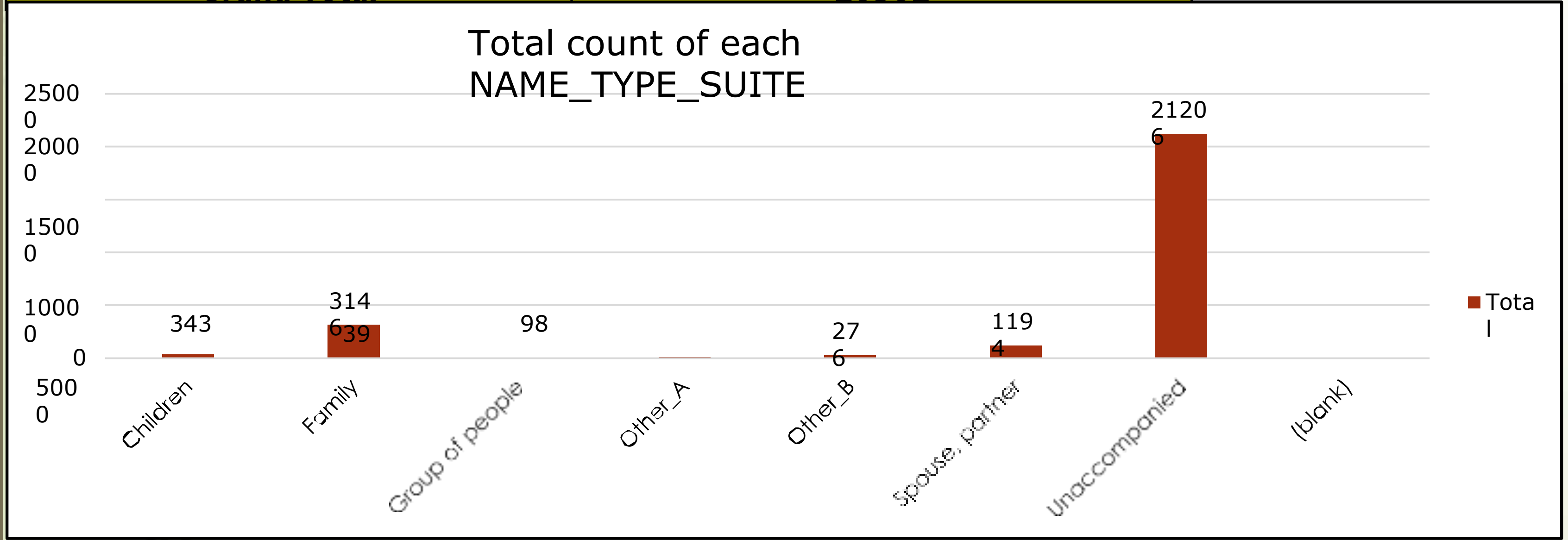
AMT_ANNUITY							
Replace Blanks with 21340							

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

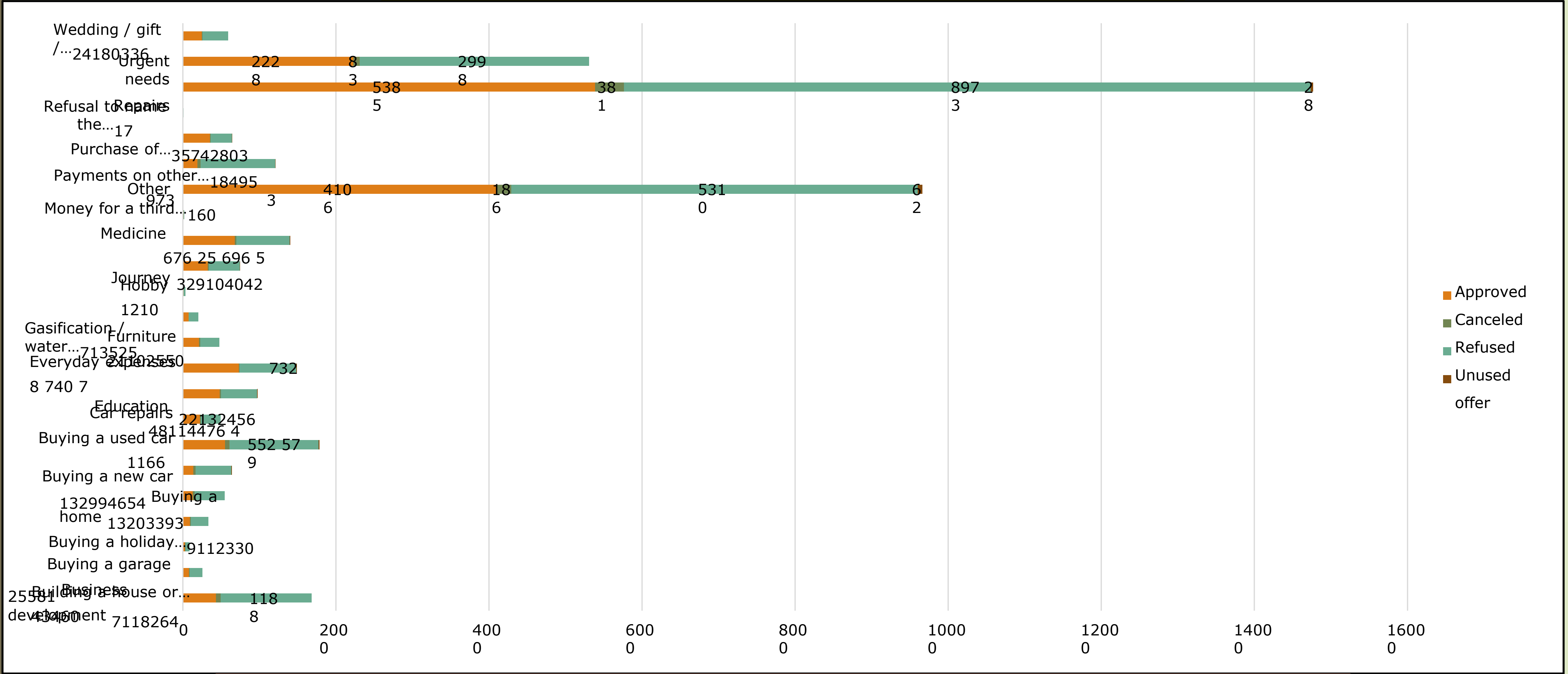
Previous Application Dataset – Analysis of Cleaned Data

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

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

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 0% and that of non-defaulters(target = 0) is around 100%**

- **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 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