Task-6: **Bank Loan Case Study**

EXCEL FILE LINK FOR BANK LOAN PROJECT

- https://drive.google.com/drive/folders/1eCWYStVJzokbkv-pFOrA0p39suJL3IbL?usp=sharing

Excel Tasks:

1. Identify Missing Data and Deal with it Appropriately
2. Identify Outliers in the Dataset
3. Analyse Data Imbalance
4. Perform Univariate, Segmented Univariate, and Bivariate Analysis
5. Identify Top Correlations for Different Scenarios

Project Details:

The Bank Loan Case Study project focuses on using Exploratory Data Analysis (EDA) to analyze data patterns and prevent the rejection of qualified loan applicants. The goal is to utilize Excel, data visualization, and statistical techniques for a thorough data analysis. This project aims to extract valuable insights and identify patterns that can indicate whether a customer may face challenges in repaying their installments.

- **Software Used: Microsoft Excel 365**

DATA HANDLING

- I looked at the data and understood all the columns. I noticed there are 128 columns and 49999 rows. The data has some unnecessary columns, empty values, and blank rows. I've decided to clean up the dataset completely.

1) Identify Missing Data and Deal with it Appropriately

FUNTIONS USED:

You used two formulas, =COUNTBLANK(A2:A50000) and =COUNTBLANK(A2:A50000)/COUNTA(A2:A50000)*100, to count blank values in your data.

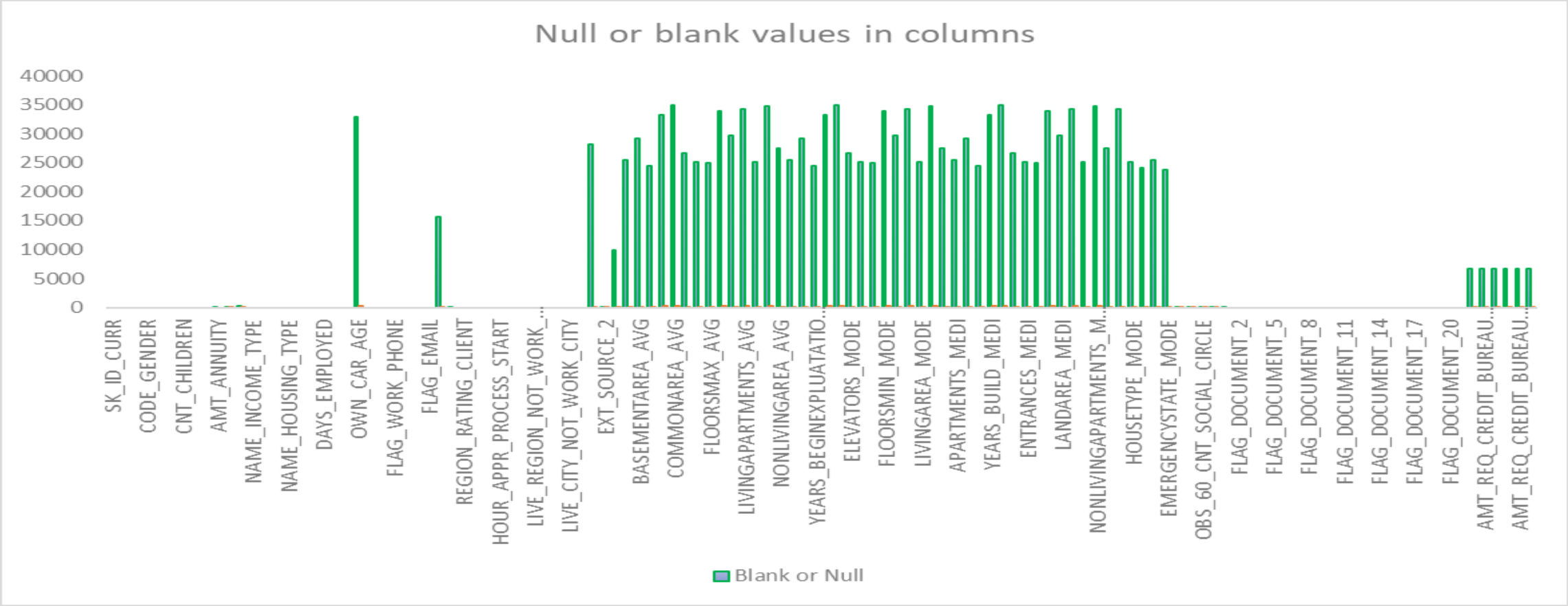
After finding the null values, you deleted columns where the null values were more than 25%. For the columns with less than 25% null values, you replaced them with the median using the =MEDIAN(J2:J50000) formula.

After these steps, you ended up with 72 columns and 49999 rows. This task helped you learn how to handle missing values in a large dataset effectively.

DATA ANALYSIS

1) Identify Missing Data and Deal with it Appropriately

Results: Before Cleaning



DATA ANALYSIS

1) Identify Missing Data and Deal with it Appropriately

Columns	total Null values	<25%	Average
EXT_SOURCE_3	9944	24.83	0.51
AMT_REQ_CREDIT_BUREAU_QRT	6734	15.56	0.26
AMT_REQ_CREDIT_BUREAU_MON	6734	15.56	0.27
AMT_REQ_CREDIT_BUREAU_DAY	6734	15.56	0.01
AMT_REQ_CREDIT_BUREAU_WEEK	6734	15.56	0.03
AMT_REQ_CREDIT_BUREAU_HOUR	6734	15.56	0.01
AMT_REQ_CREDIT_BUREAU_YEAR	6734	15.56	1.88
NAME_TYPE_SUITE	192	0.38	#DIV/0!
OBS_60_CNT_SOCIAL_CIRCLE	168	0.34	1.40
OBS_30_CNT_SOCIAL_CIRCLE	168	0.34	1.42
DEF_30_CNT_SOCIAL_CIRCLE	168	0.34	0.14
DEF_60_CNT_SOCIAL_CIRCLE	168	0.34	0.10
EXT_SOURCE_2	126	0.25	0.51
AMT_GOODS_PRICE	38	0.08	539060.04
DAYS_LAST_PHONE_CHANGE	1	0.00	-964.30
AMT_ANNUITY	1	0.00	27107.38
CNT_FAM_MEMBERS	1	0.00	2.16

I filled in missing values in columns where there were less than 25% null values. For columns, I found the most common text and used that to replace the missing values.

2) Identify Outliers in the Dataset:

Functions Used:

=QUARTILE.EXC(A2:A50000,1) **[QUARTILE-1]**

=QUARTILE.EXC(A2:A50000,3) **[QUARTILE-3]**

=N2-M2 **[IQR]**

=M2-1.5*O2 **[LOWER BOUND]**

=N2+1.5*O2 **[UPPER BOUND]**

I have Calculated Quartile-1, Quartile-2, Inter Quartile Range (IQR), Lower Bound, Upper Bound.

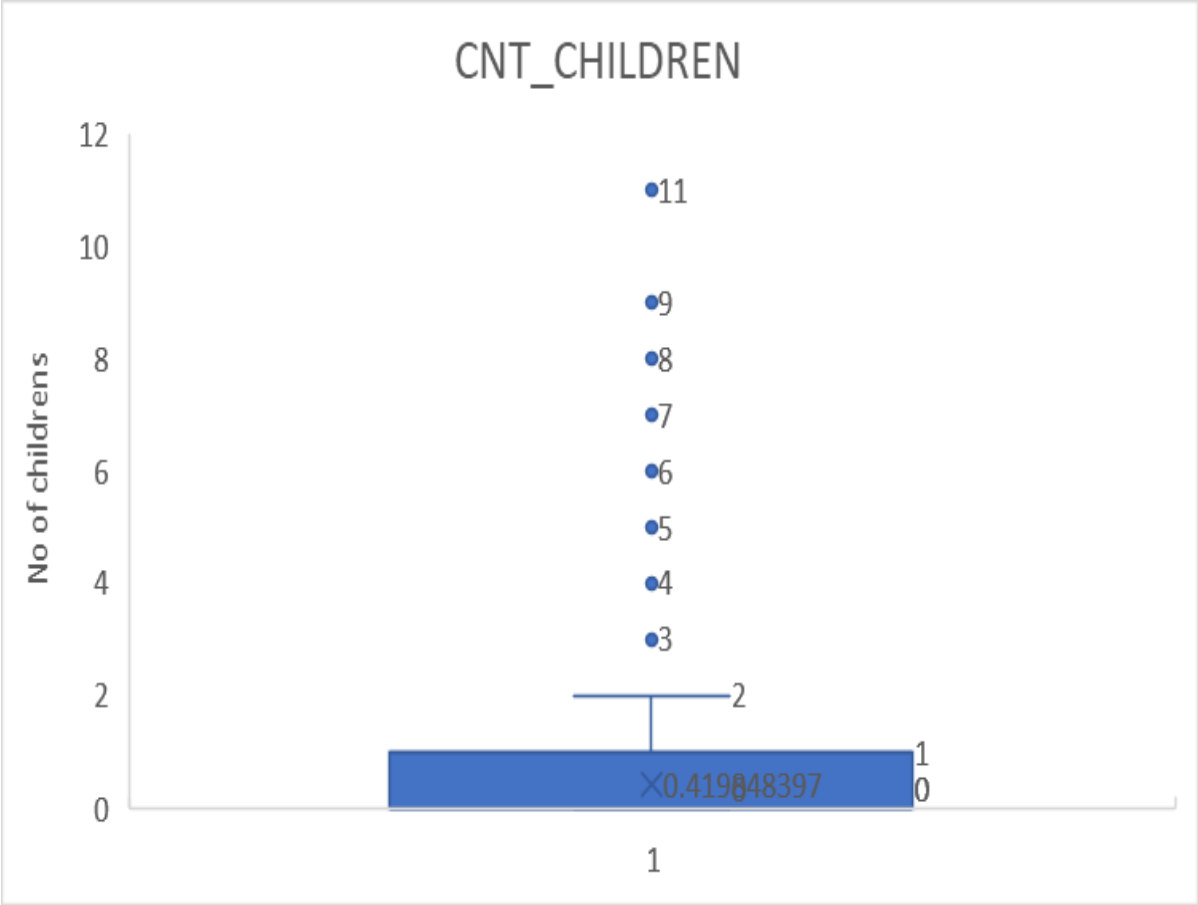
2) Identify Outliers in the Dataset:

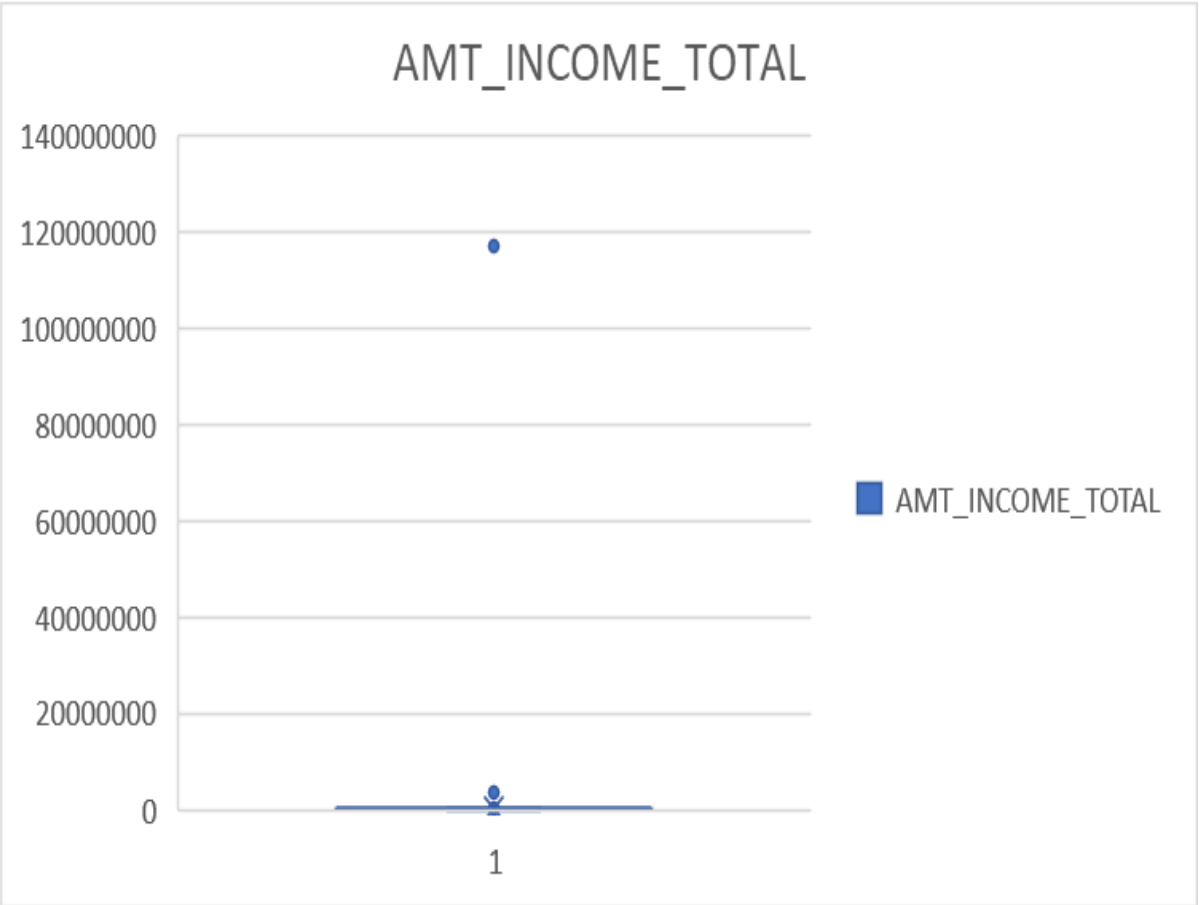
Results:

Column	QUARTILE Q1	QUARTILE Q3	Inter Quartile Range IQR	Lower Bound	Upper Bound
CNT_CHILDREN	0	1	1	-1.5	2.5
AMT_INCOME_TOTAL	112500	202500	90000	-22500	337500
AMT_CREDIT	270000	808650	538650	-537975	1616625
AMT_ANNUITY	16456.5	34596	18139.5	-10752.75	61805.25
AMT_GOODS_PRICE	238500	679500	441000	-423000	1341000
DAYS_BIRTH	-19644	-12378	7266	-30543	-1479
DAYS_EMPLOYED	-2786	-292	2494	-6527	3449
DAYS_REGISTRATION	-7464	-1998	5466	-15663	6201
DAYS_ID_PUBLISH	-4297	-1722	2575	-8159.5	2140.5
DAYS_LAST_PHONE_CHANGE	-1573	-270	1303	-3527.5	1684.5

2) Identify Outliers in the Dataset:

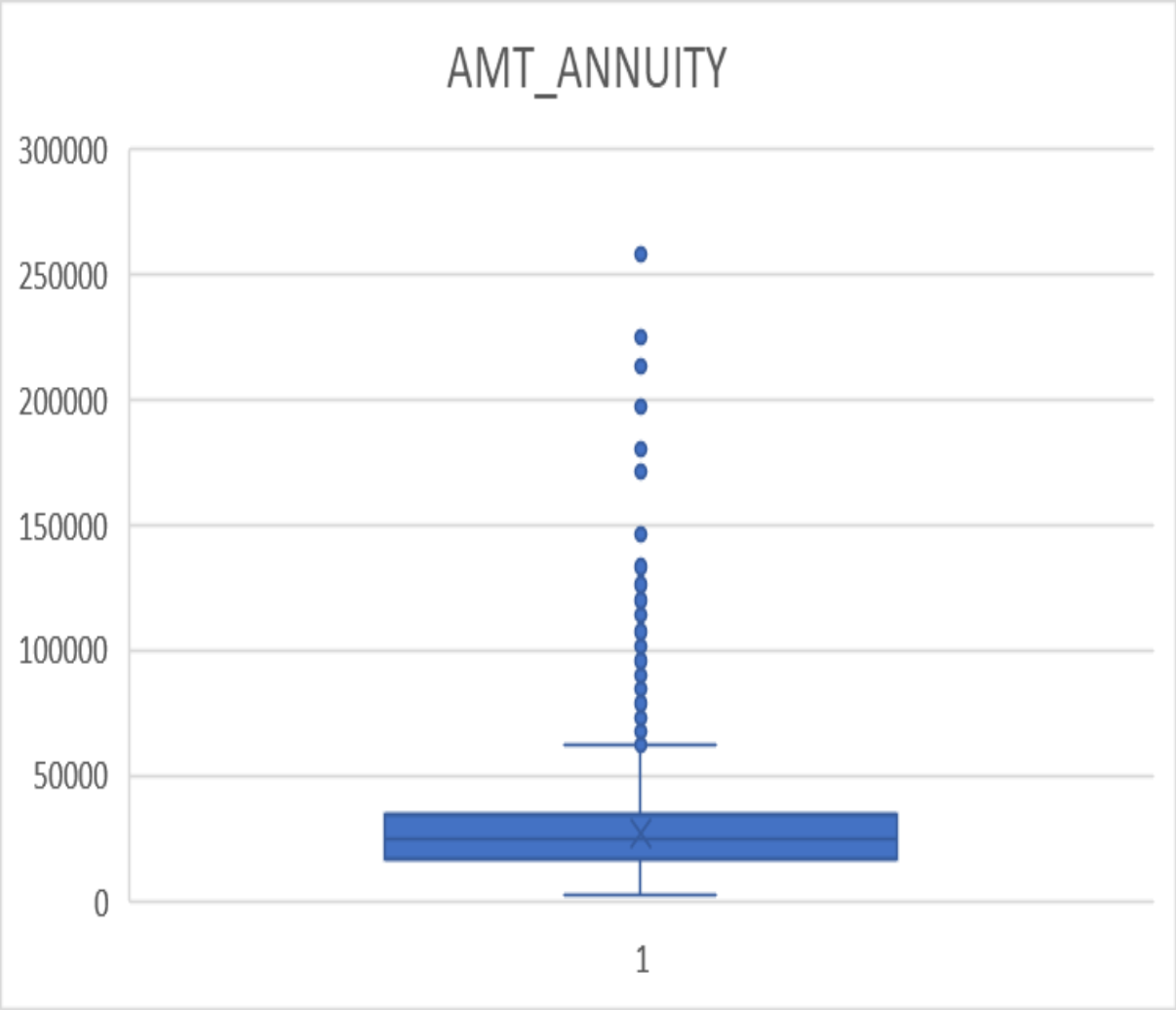
Results:





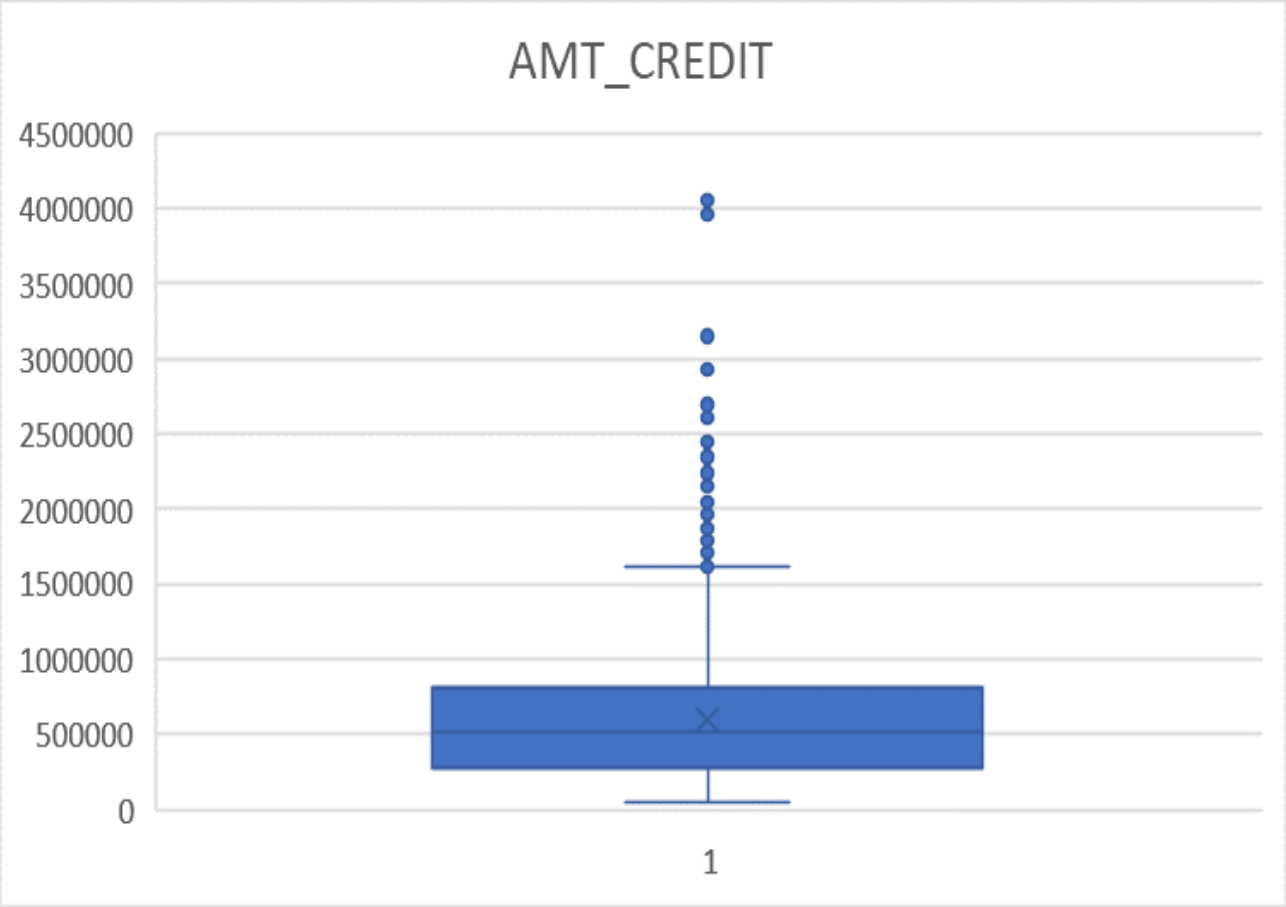
2) Identify Outliers in the Dataset:

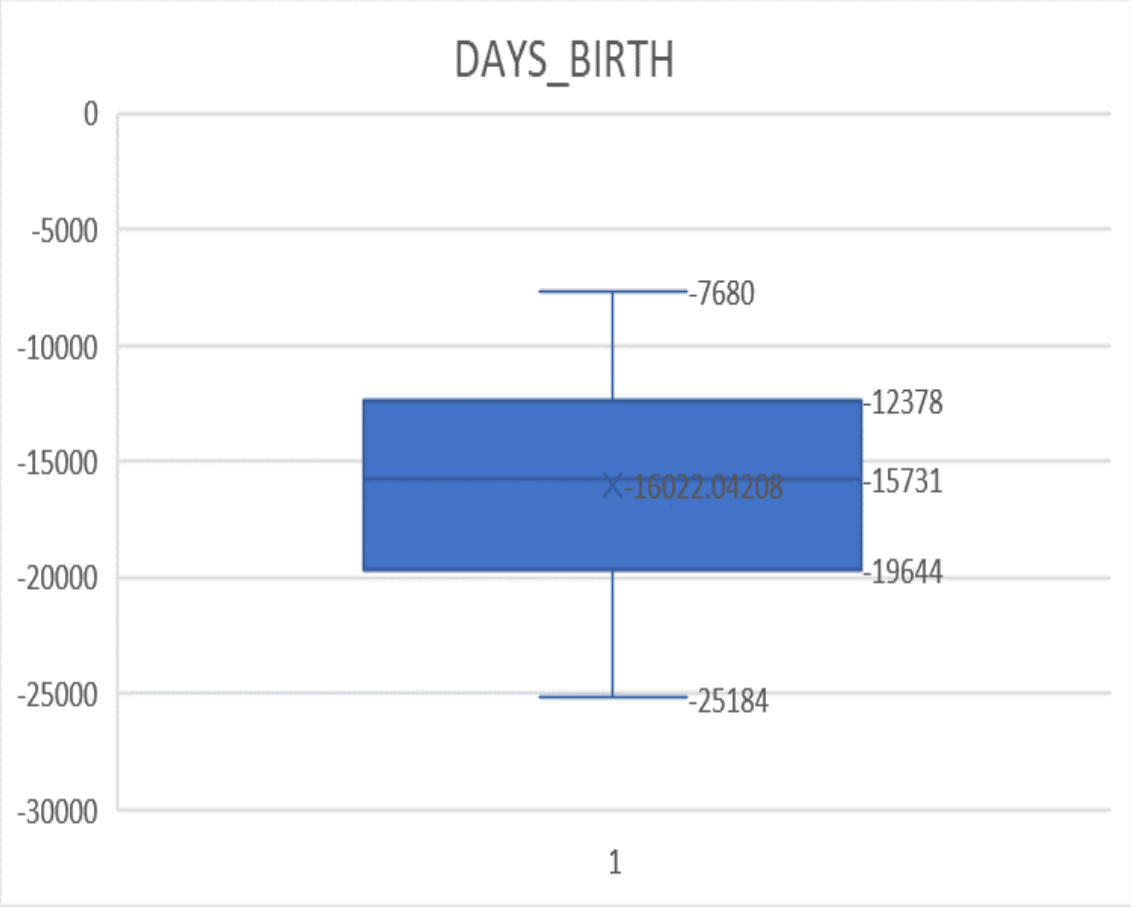
Results:



2) Identify Outliers in the Dataset:

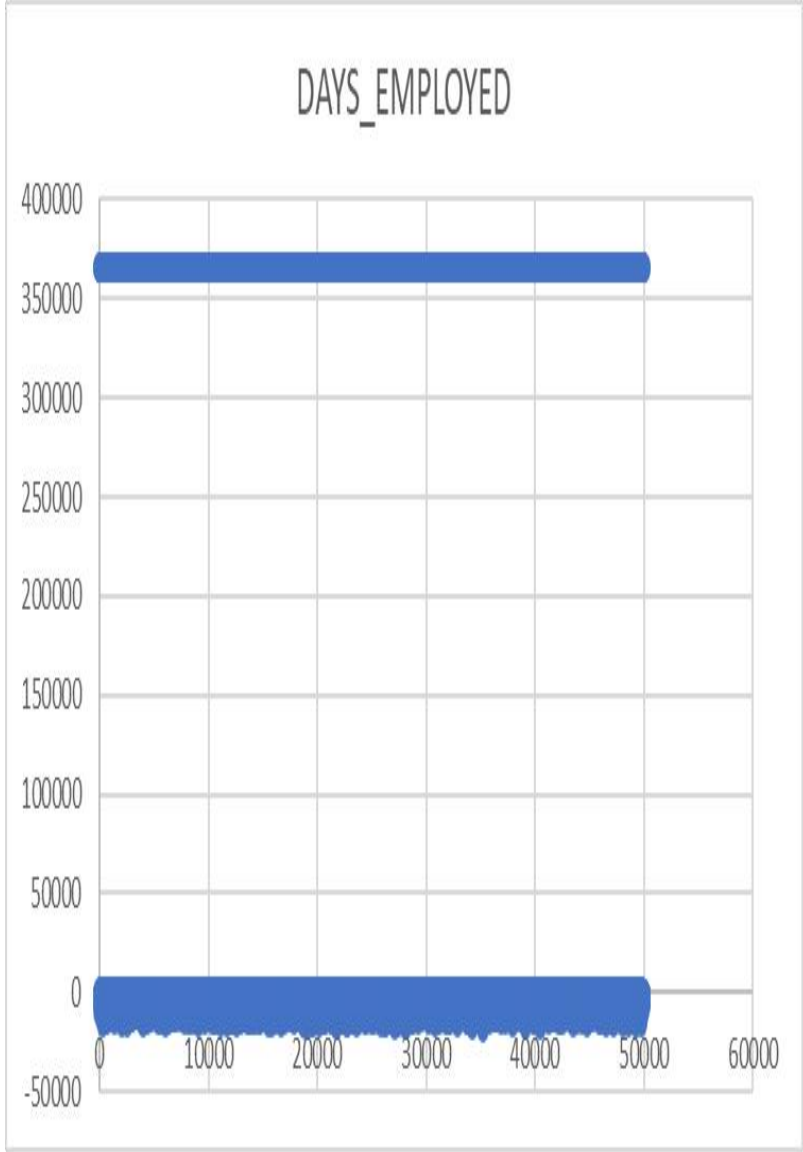
Results:

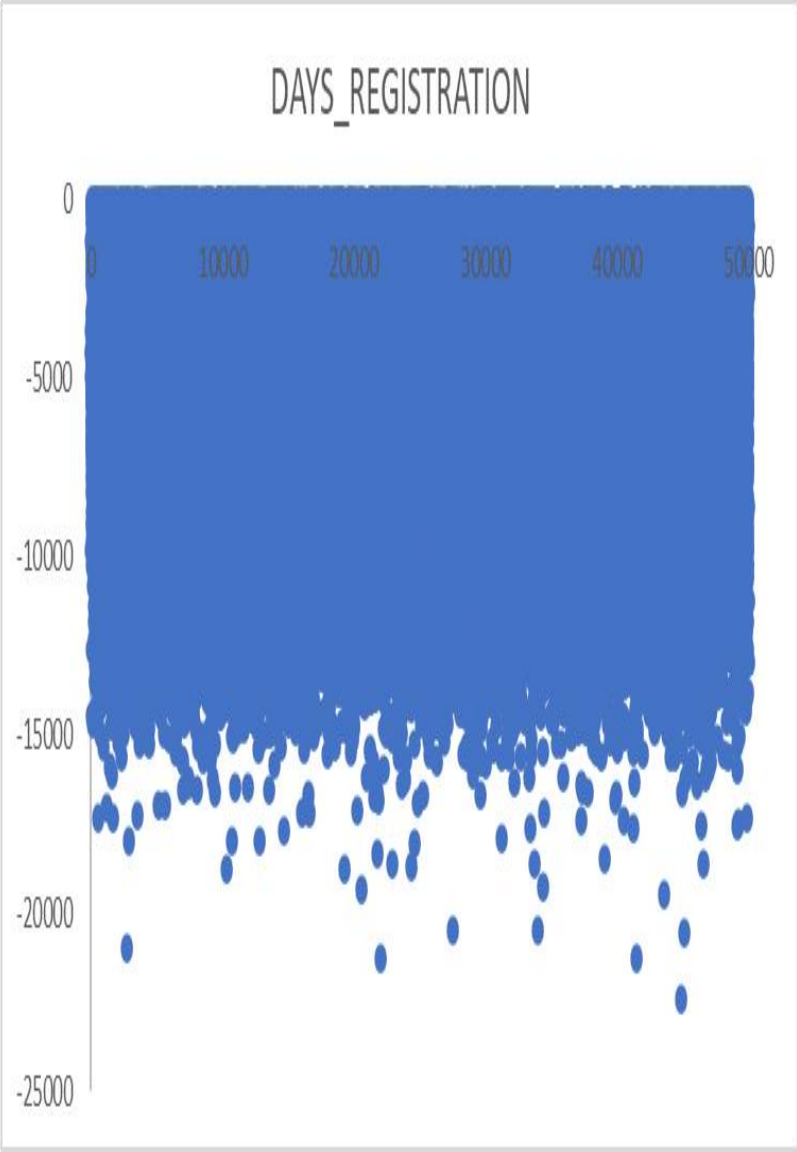


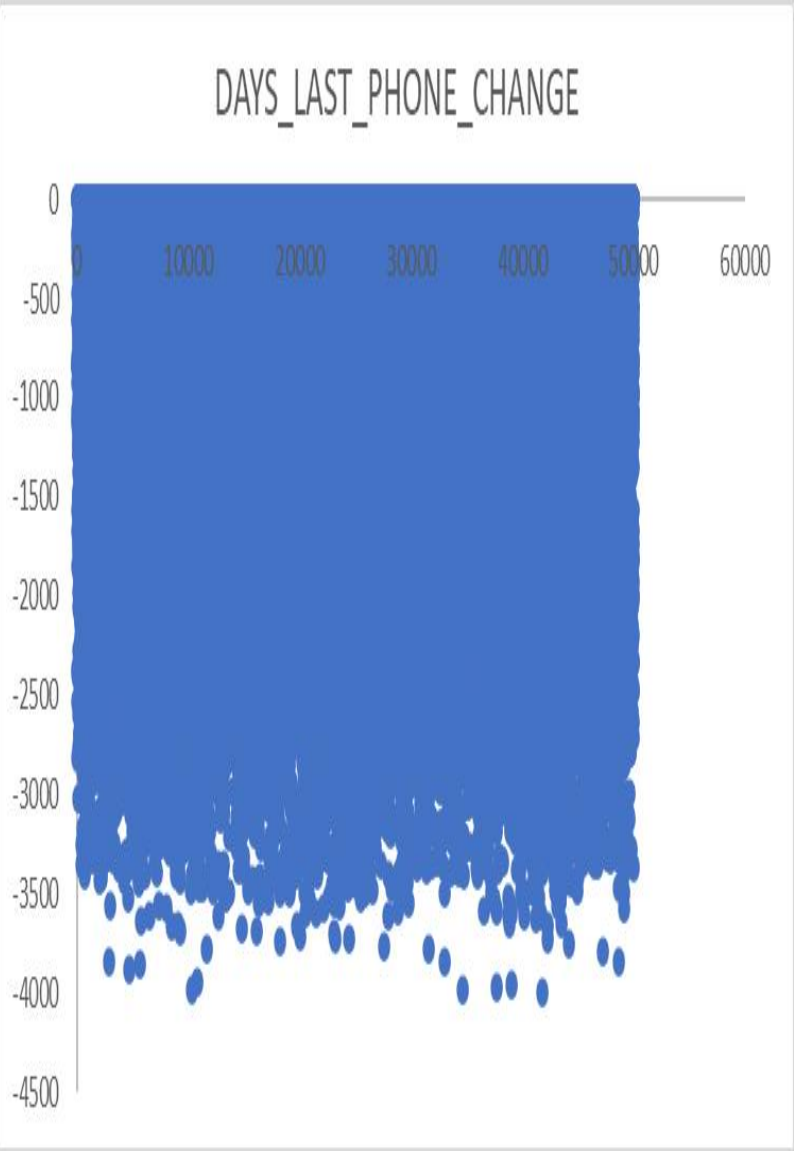


2) Identify Outliers in the Dataset:

Results:







2) Identify Outliers in the Dataset:

Results:

REN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	270000	1293502.5	35698.5	1129500	-16765	-1188	-1186	-291	-828
0	67500	135000	6750	135000	-19046	-225	-4260	-2531	-815
0	135000	312682.5	29686.5	297000	-19005	-3039	-9833	-2437	-617
0	121500	513000	21865.5	513000	-19932	-3038	-4311	-3458	-1106
0	99000	490495.5	27517.5	454500	-16941	-1588	-4970	-477	-2536
1	171000	1560726	41301	1395000	-13778	-3130	-1213	-619	-1562
0	360000	1530000	42075	1530000	-18850	-449	-4597	-2379	-1070
0	112500	1019610	33826.5	913500	-20099	365243	-7427	-3514	0
0	135000	405000	20250	405000	-14469	-2019	-14437	-3992	-1673
1	112500	652500	21177	652500	-10197	-679	-4427	-738	-844
0	38419.155	148365	10678.5	135000	-20417	365243	-5246	-2512	-2396
0	67500	80865	5881.5	67500	-13439	-2717	-311	-3227	-2370
1	225000	918468	28966.5	697500	-14086	-3028	-643	-4911	-4
0	189000	773680.5	32778	679500	-14583	-203	-615	-2056	-188
0	157500	299772	20160	247500	-8728	-1157	-3494	-1368	-925
0	108000	509602.5	26149.5	387000	-12931	-1317	-6392	-3866	-3
1	81000	270000	13500	270000	-9776	-191	-4143	-2427	-2811
0	112500	157500	7875	157500	-17718	-7804	-8751	-1259	-239
1	90000	544491	17563.5	454500	-11348	-2038	-1021	-3964	-1850
0	135000	427500	21375	427500	-18252	-4286	-298	-1800	-296
1	202500	1132573.5	37561.5	927000	-14815	-1652	-2299	-2299	0
1	450000	497520	32521.5	450000	-11146	-4306	-114	-2518	-468
0	83250	239850	23850	225000	-24827	365243	-9012	-3684	-795
2	135000	247500	12703.5	247500	-11286	-746	-108	-3729	-4
0	90000	225000	11074.5	225000	-19334	-3494	-2419	-2893	0
0	112500	979992	27076.5	702000	-18724	-2628	-6573	-1827	-161
1	112500	327024	23827.5	270000	-15948	-1234	-5782	-3153	-2
0	270000	790830	57676.5	675000	-9994	-1796	-4668	-2661	-849
0	90000	180000	9000	180000	-10341	-1010	-4799	-3015	-599
0	292500	665892	24592.5	477000	-15280	-2668	-5266	-3787	-1634
0	112500	512064	25033.5	360000	-11144	-1104	-7846	-2904	-397
0	90000	199008	20893.5	180000	-12974	-4404	-7123	-4464	-2766
1	360000	733315.5	39069	679500	-11694	-2060	-3557	-3557	-697
0	135000	1125000	32895	1125000	-15997	-4585	-5735	-4067	-3019

highlighted the columns using conditional formatting of upper bound and lower bound

3) Analyse Data Imbalance:

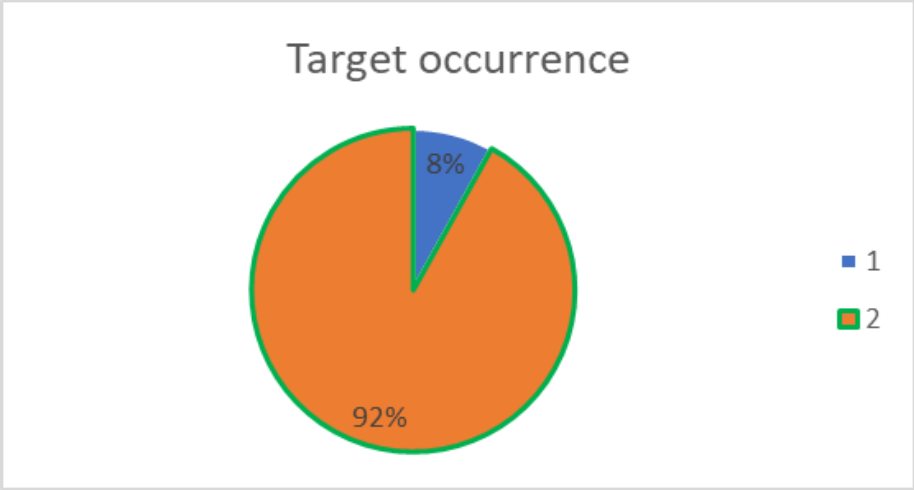
Functions used:

use the **UNIQUE** function to find the unique values in column B from cells 2 to 50000. Additionally,
the **COUNTIF** function to calculate the occurrences of either 1 or 0 in that column.

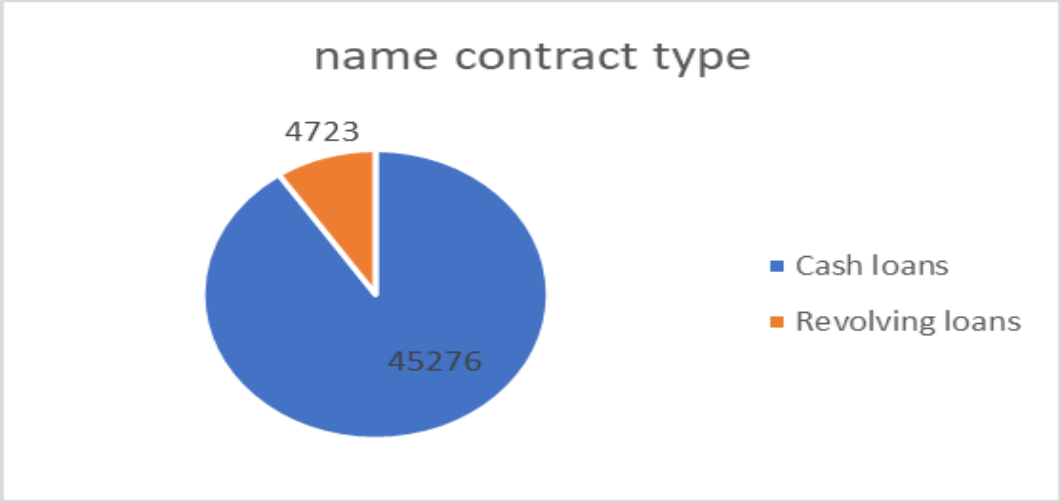
for the imbalance ratio, o divide the count of occurrences for one scenario by the count for the other.

3) Analyse Data Imbalance:

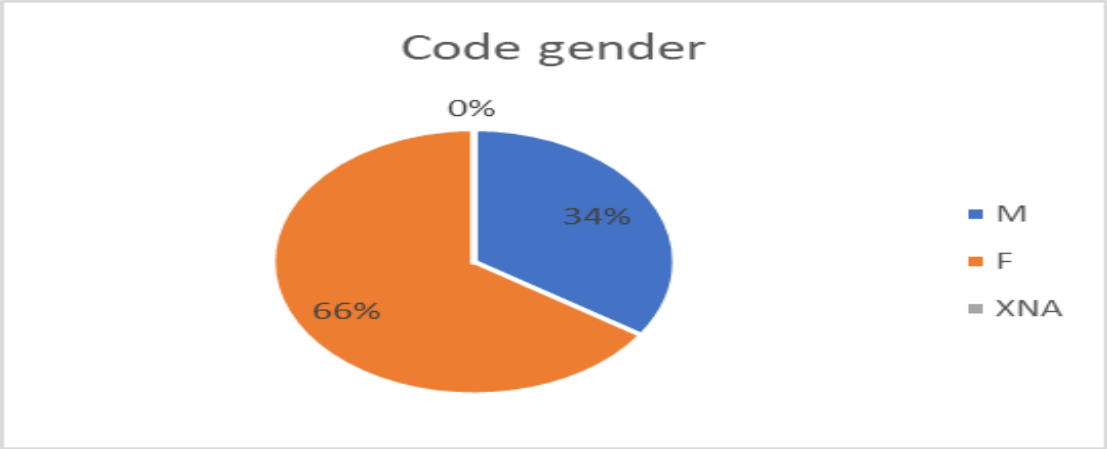
TARGET	occurrence	imbalance ratio
1	4026	8.757314076
0	45973	



NAME_CONTRACT_TYPE	occurrence	imbalance ratio
Cash loans	45276	10.43157523
Revolving loans	4723	

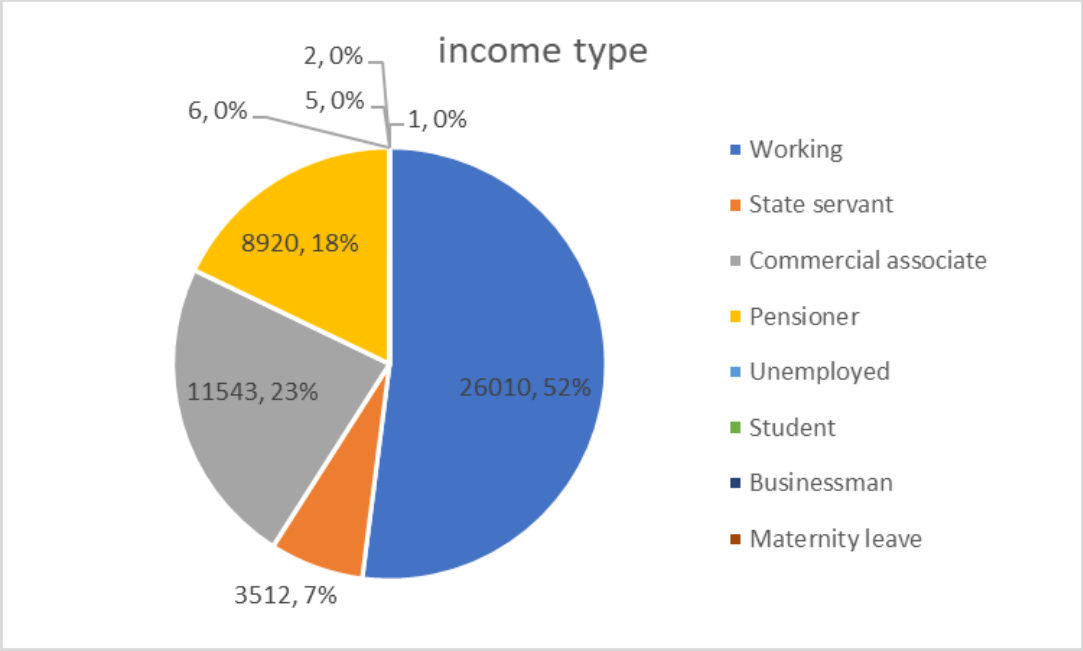


CODE_GENDER	occurrence	imbalance ratio
M	17174	52.32306614
F	32823	
XNA	2	

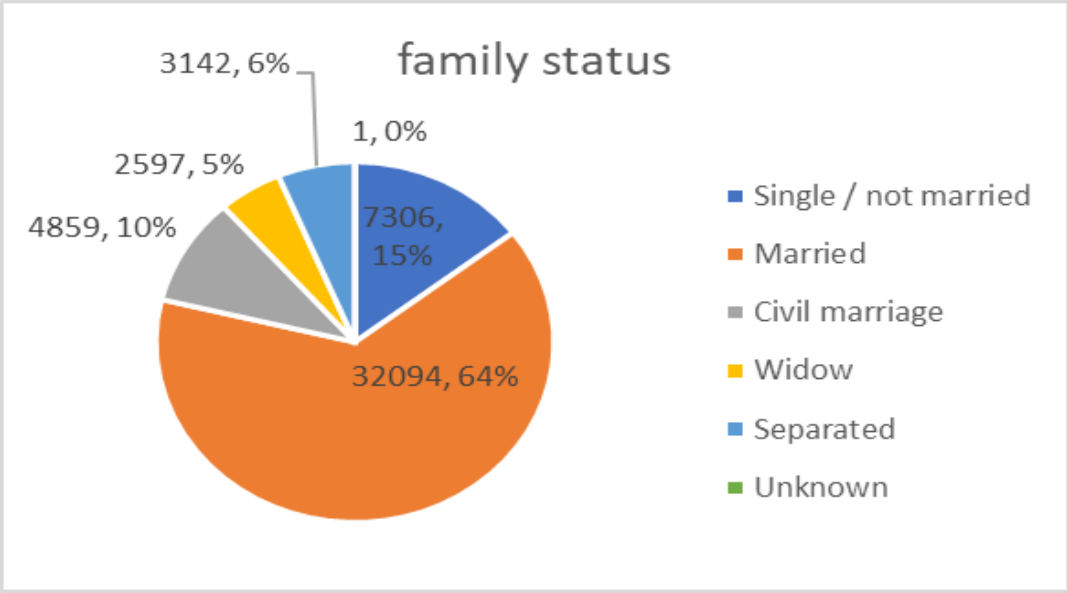


3) Analyse Data Imbalance:

NAME_INCOME_TYPE	occurrence
Working	26010
State servant	3512
Commercial associate	11543
Pensioner	8920
Unemployed	6
Student	5
Businessman	2
Maternity leave	1



NAME_FAMILY_STATUS	occurrence
Single / not married	7306
Married	32094
Civil marriage	4859
Widow	2597
Separated	3142
Unknown	1



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

- I have created pivot table.
- Calculate mean,mode,median,std dev using excel inbuilt functions

columns	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	DAYS_BIRTH	AGE
mean	170767.5905	599700.5815	27107.33399	16022.04208	45
median	145800	514777.5	24939	15731	44
mode	135000	450000	9000	11039	31
std dev	531813.7768	402405.2266	14562.6564	4361.356655	12

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

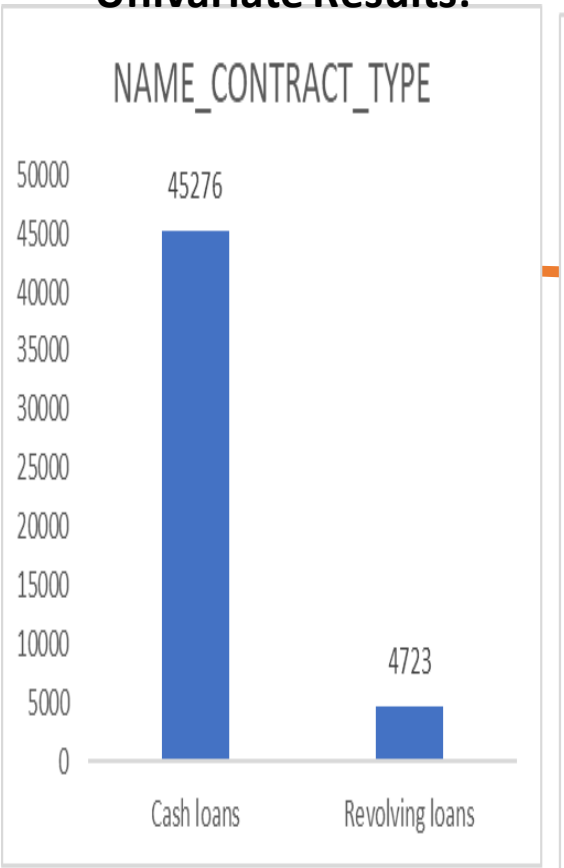
NAME_CONTRACT_TYPE	occurrence
Cash loans	45276
Revolving loans	4723

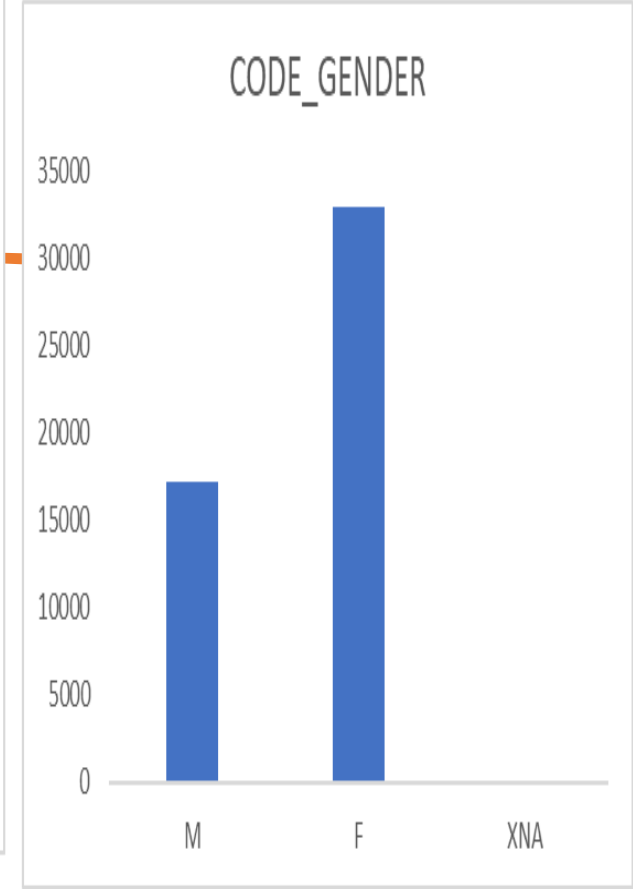
CODE_GENDER	occurrence
M	17174
F	32823
XNA	2

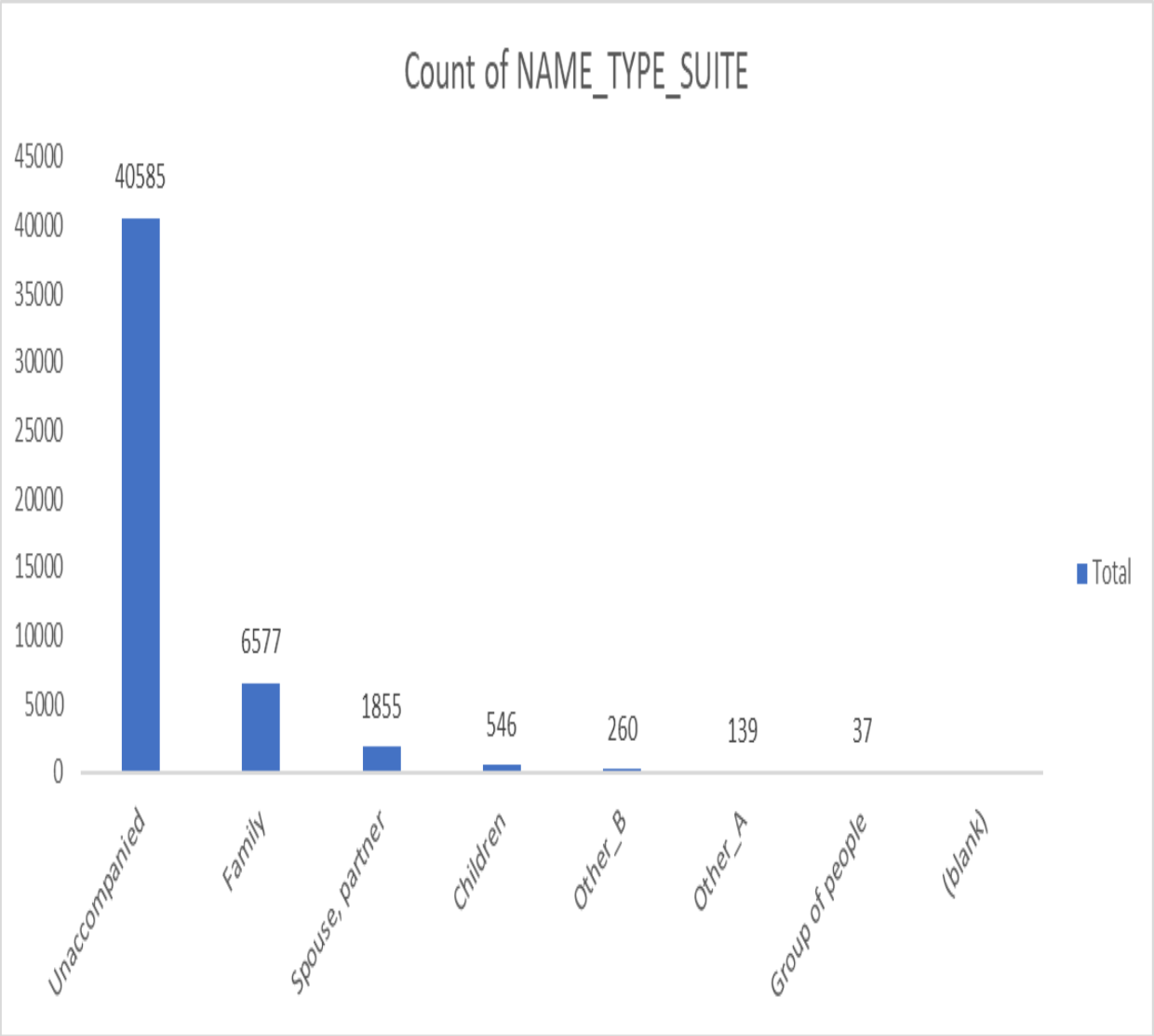
Row Labels	Count of NAME_TYPE_SUITE
Unaccompanied	40585
Family	6577
Spouse, partner	1855
Children	546
Other_B	260
Other_A	139
Group of people	37

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:



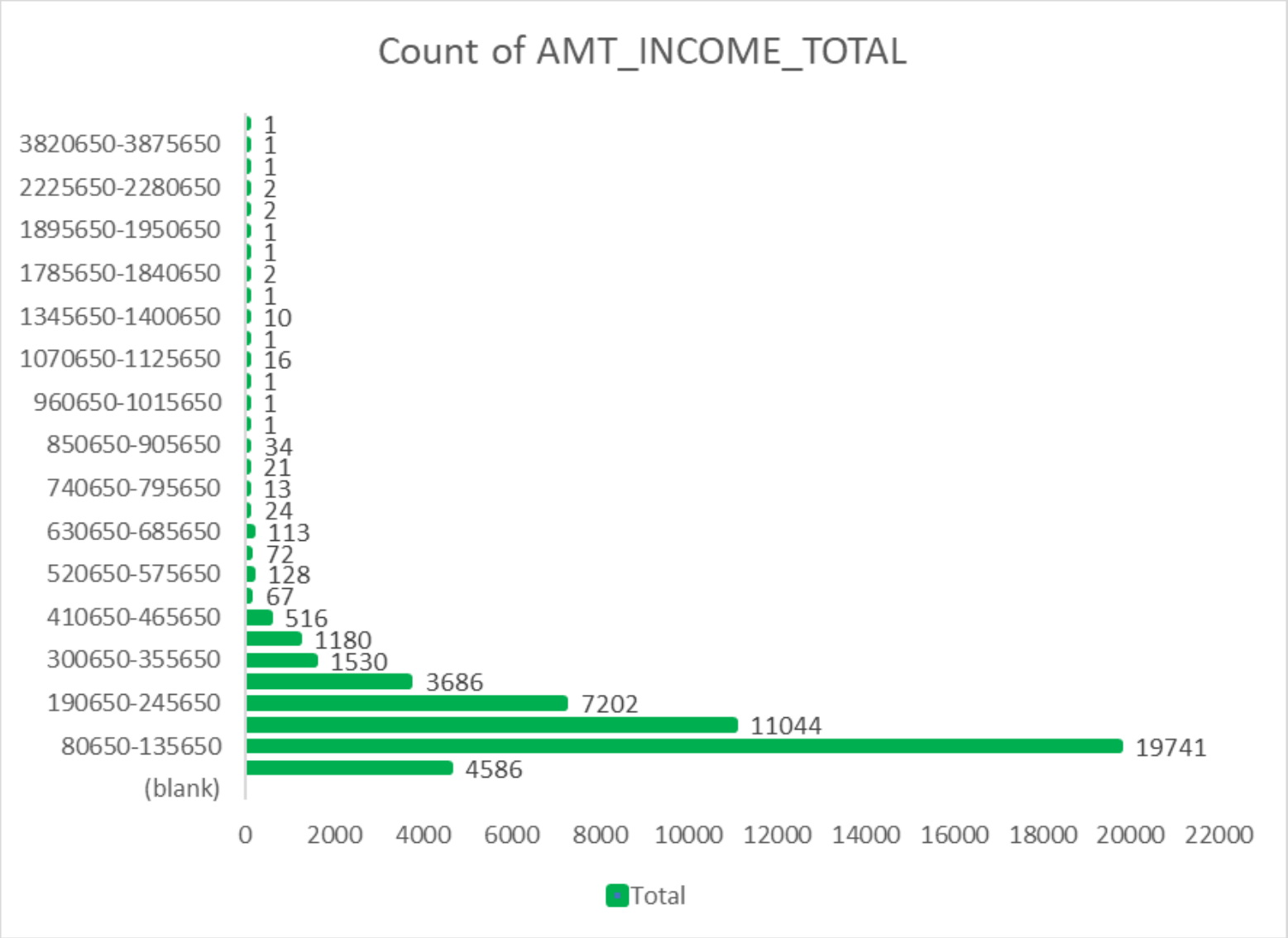




4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count of AMT_INCOME_TOTAL
(blank)	
25650-80650	4586
80650-135650	19741
135650-190650	11044
190650-245650	7202
245650-300650	3686
300650-355650	1530
355650-410650	1180
410650-465650	516
465650-520650	67
520650-575650	128
575650-630650	72
630650-685650	113
685650-740650	24
740650-795650	13

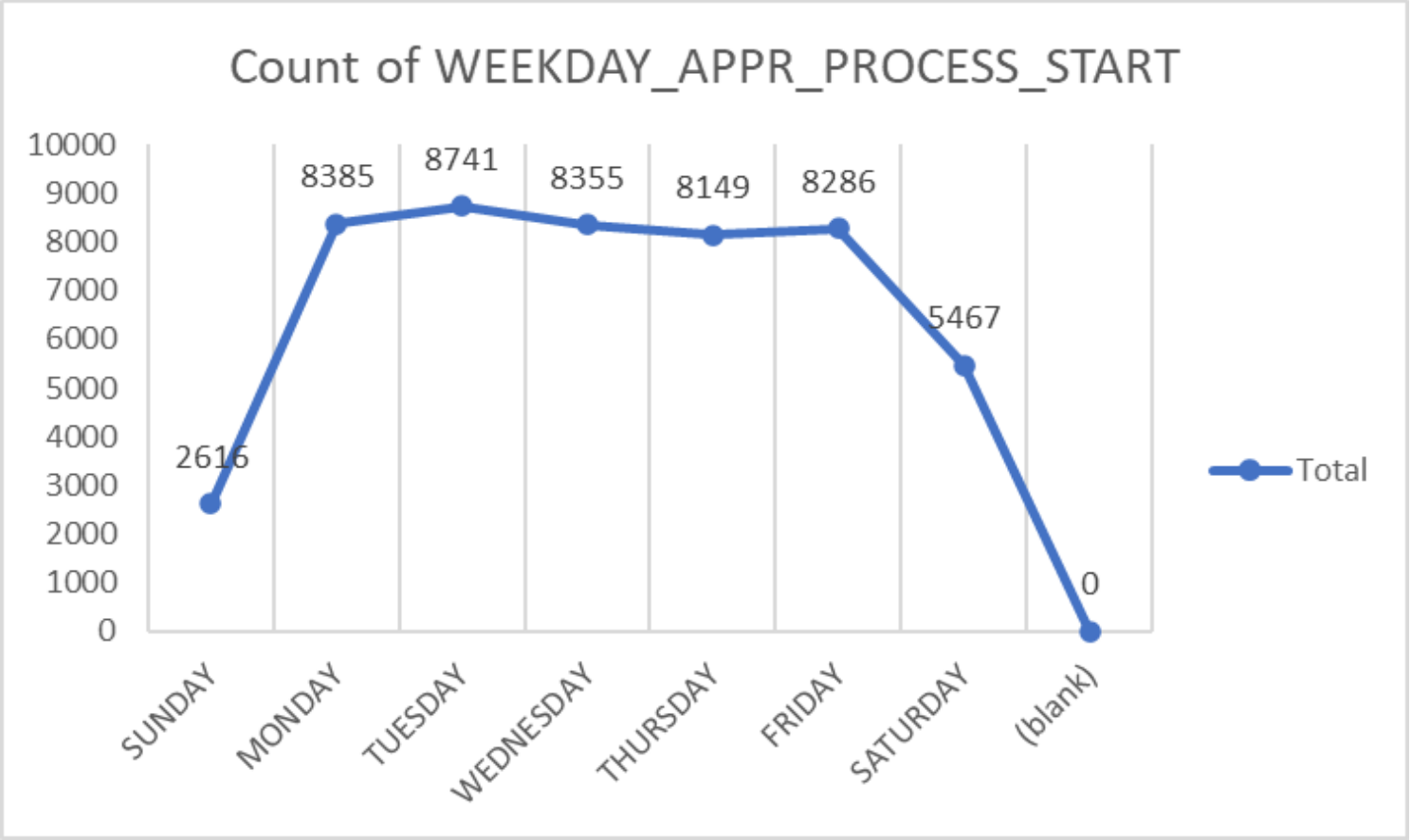


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

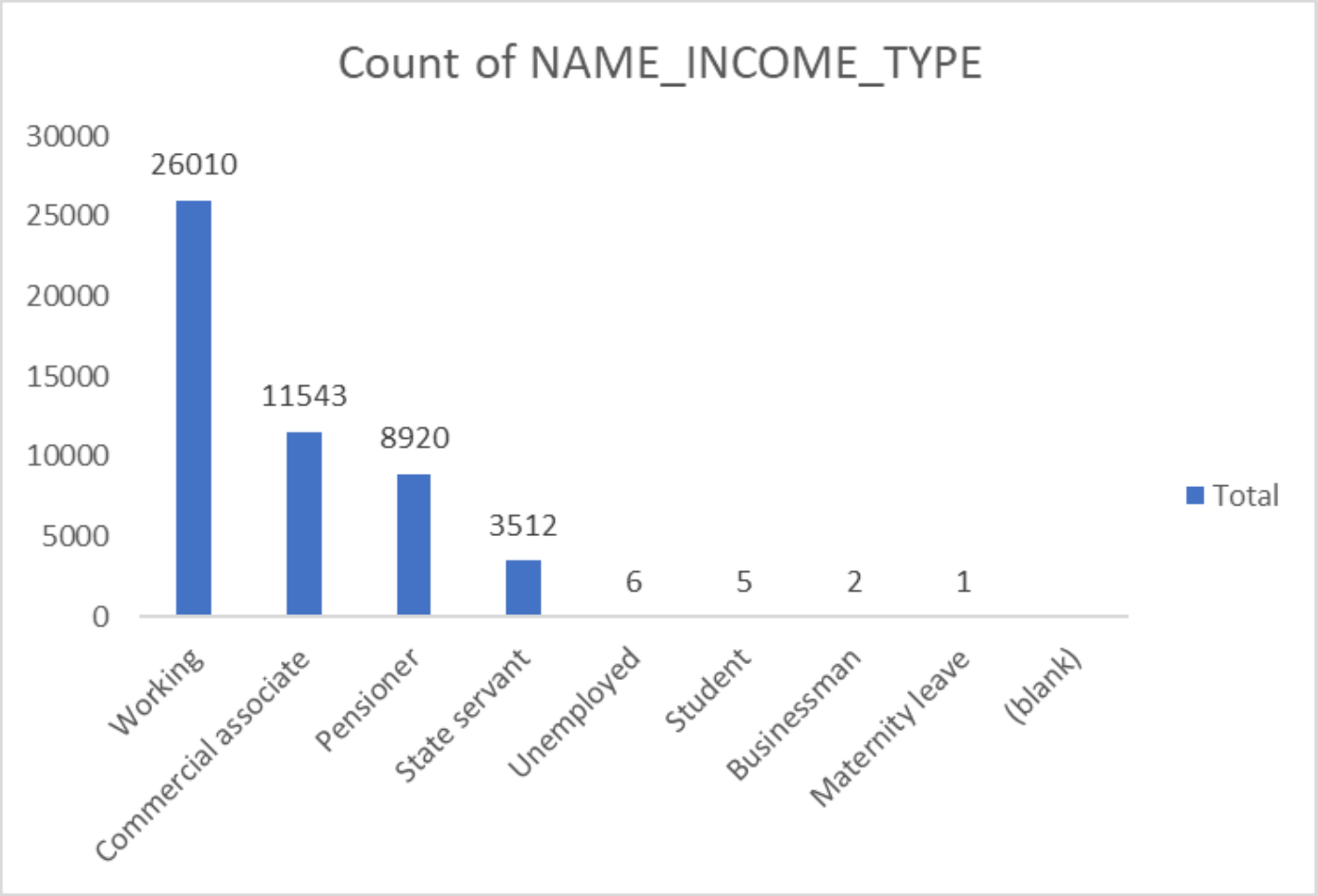
Row Labels	Count of WEEKDAY_APPR_PROCESS_START
SUNDAY	2616
MONDAY	8385
TUESDAY	8741
WEDNESDAY	8355
THURSDAY	8149
FRIDAY	8286
SATURDAY	5467



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

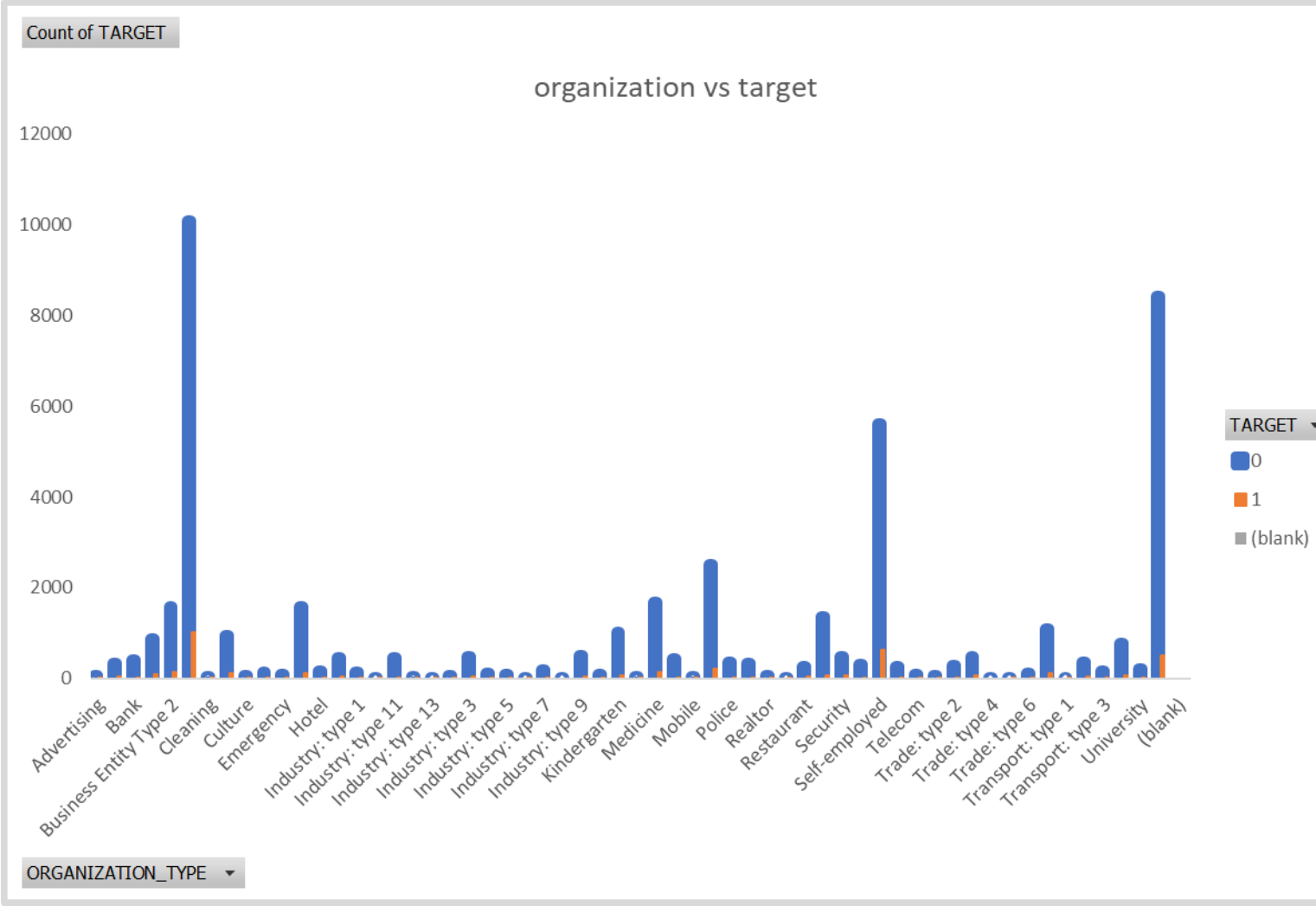
Row Labels	Count of NAME_INCOME_TYPE
Working	26010
Commercial associate	11543
Pensioner	8920
State servant	3512
Unemployed	6
Student	5
Businessman	2
Maternity leave	1



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

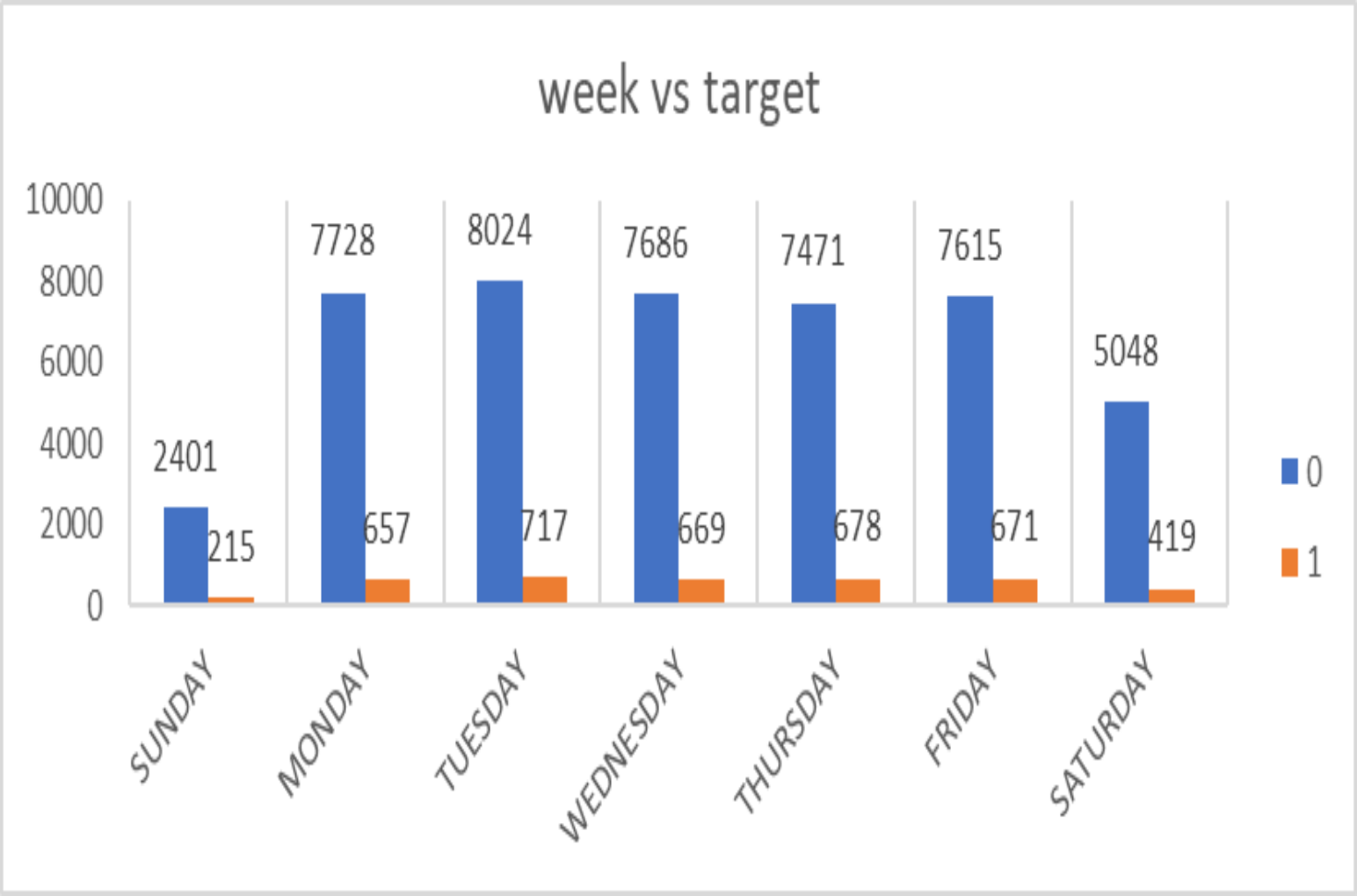
Count of TARGET	Column Labels		
Row Labels	0	1 (blank)	Grand Total
Advertising	61	7	68
Agriculture	341	51	392
Bank	408	27	435
Business Entity Type 1	865	88	953
Business Entity Type 2	1571	133	1704
Business Entity Type 3	10087	1014	11101
Cleaning	37	3	40
Construction	958	108	1066
Culture	62	2	64
Electricity	134	13	147
Emergency	86	7	93
Government	1592	124	1716
Hotel	169	13	182
Housing	447	42	489
Industry: type 1	140	19	159
Industry: type 10	20	1	21
Industry: type 11	461	28	489
Industry: type 12	50	3	53
Industry: type 13	11	4	15
Industry: type 2	68	10	78



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

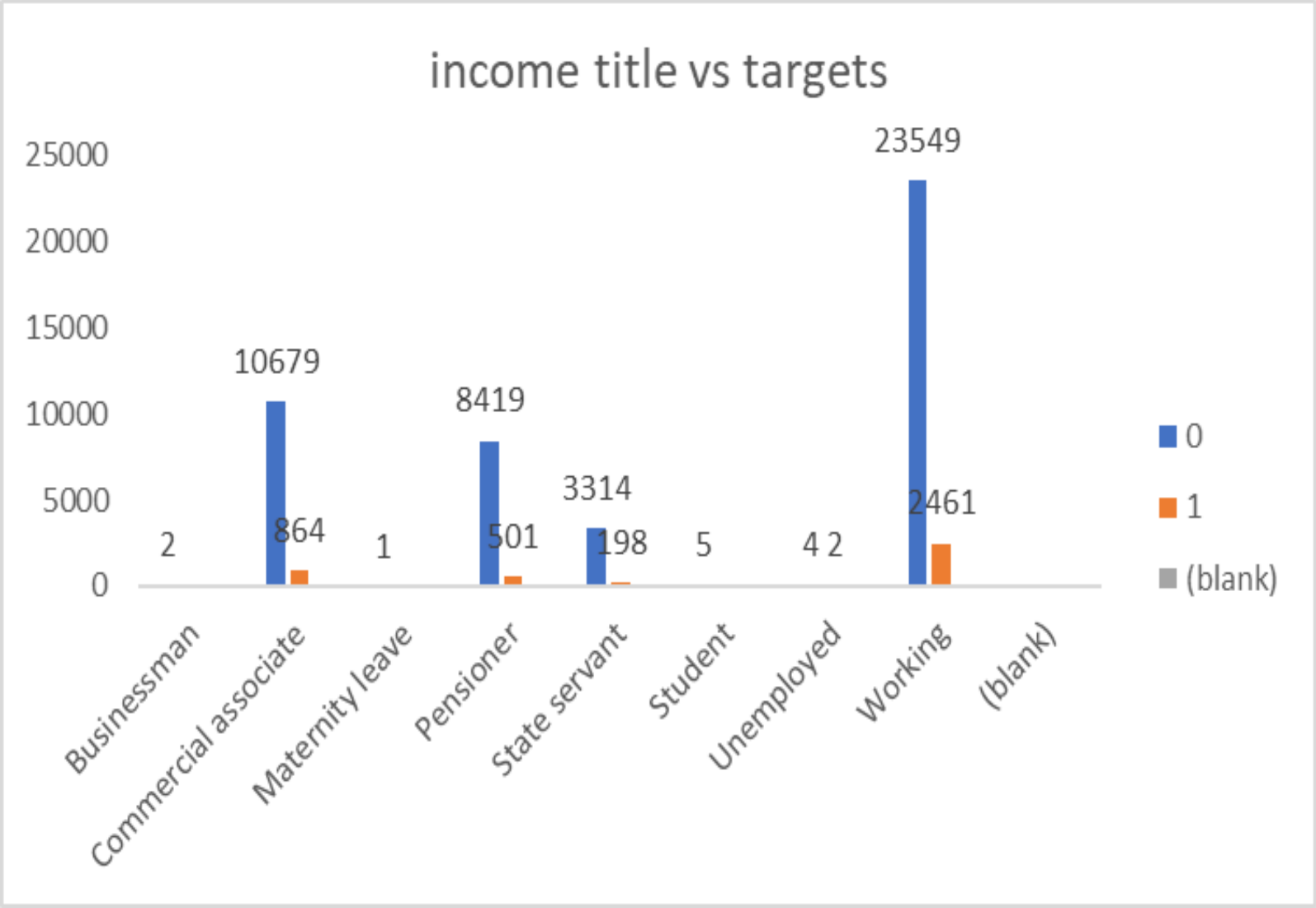
week	Column Labels		
Row Labels	0	1	Grand Total
SUNDAY	2401	215	2616
MONDAY	7728	657	8385
TUESDAY	8024	717	8741
WEDNESDAY	7686	669	8355
THURSDAY	7471	678	8149
FRIDAY	7615	671	8286
SATURDAY	5048	419	5467



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

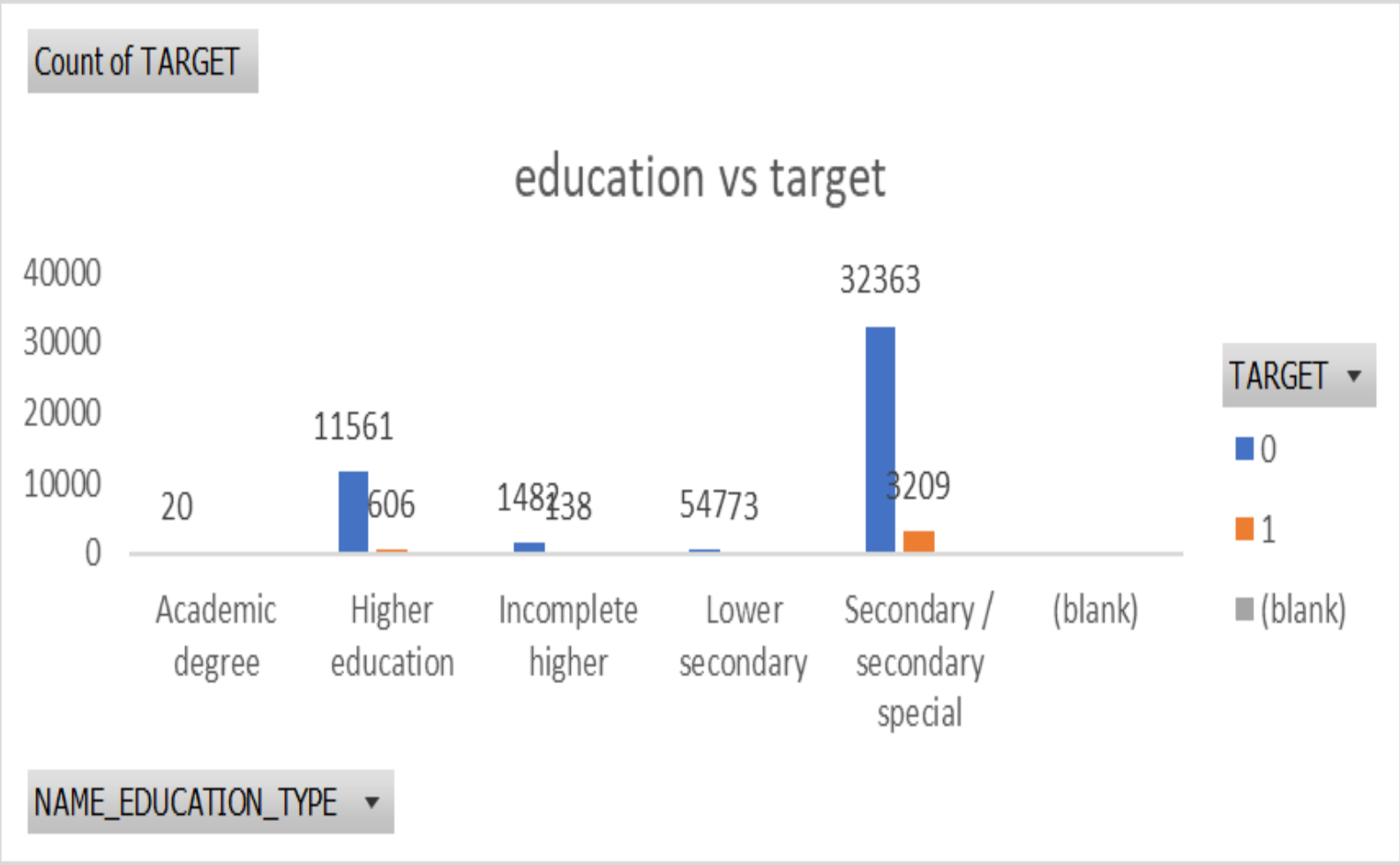
Count of TARGET	Column Labels		
Row Labels	0	1 (blank)	Grand Total
Business man	2		2
Commercial associate	10679	864	11543
Maternity leave	1		1
Pensioner	8419	501	8920
State servant	3314	198	3512
Student	5		5
Unemployed	4	2	6
Working	23549	2461	26010



4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

Count of TARGET	Column Labels	
Row Labels	**0**	**1**
Academic degree	20	
Higher education	11561	606
Incomplete higher	1482	138
Lower secondary	547	73
Secondary / secondary special	32363	3209
(blank)		



DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

K) ORGANIZATION_TYPE VS TARGET

Count of ORGA Column Labels			
Row Labels	0	1	Grand Total
Business Enti	10087	1014	11101
XNA	8421	503	8924
Self-employe	5612	628	6240
Other	2509	208	2717
Medicine	1687	130	1817
Government	1592	124	1716
Business Enti	1571	133	1704
School	1372	78	1450
Trade: type 7	1090	120	1210
Kindergarten	1024	66	1090
Construction	958	108	1066
Business Enti	865	88	953
Transport: typ	770	67	837
Trade: type 3	490	60	550
Security	488	62	550
Industry: type	491	51	542
Industry: type	496	41	537
Housing	447	42	489
Industry: type	461	28	489
Military	432	26	458
Bank	408	27	435
Transport: typ	359	33	392
Agriculture	341	51	392
Postal	343	27	370
Police	348	18	366
Security Minis	315	16	331
Trade: type 2	286	21	307
Restaurant	257	32	289
Services	260	24	284
University	213	9	222
Industry: type	190	19	209
Transport: typ	166	25	191
Hotel	169	13	182
Industry: type	140	19	159
Electricity	134	13	147
Industry: type	125	15	140
Trade: type 6	105	3	108
Telecom	98	8	106
Industry: type	96	7	103
Emergency	86	7	93
Insurance	82	7	89

5) Identify Top Correlations for different scenarios:

Functions used:

I found correlation between target and various columns by using below function:

=CORREL(D2:D50000,C2:C50000)

Then find each corelation where target=0 and target=1

5) Identify Top Correlations for different scenarios:

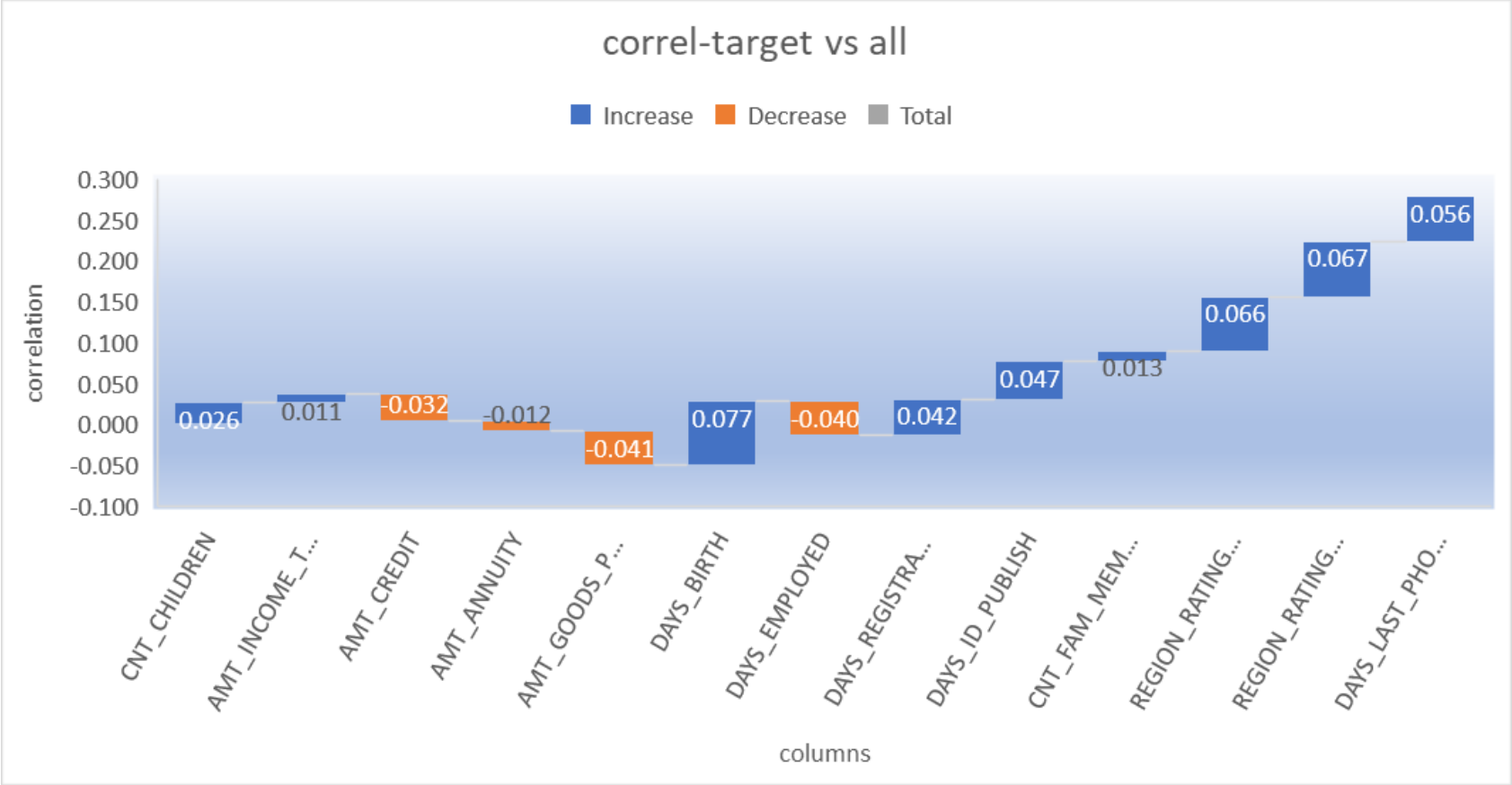
Results: correlation for 0&1

Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.009589	1											
AMT_CREDIT	0.004972	0.069316	1										
AMT_ANNUITY	0.02618	0.083008	0.769499	1									
AMT_GOODS_PRICE	0.000233	0.069892	0.986704	0.774134	1								
DAYS_BIRTH	0.329264	0.016003	-0.05934	0.007708	-0.05767	1							
DAYS_EMPLOYED	-0.23969	-0.03162	-0.07047	-0.11045	-0.06779	-0.61355	1						
DAYS_REGISTRATION	0.181217	0.009952	0.003449	0.033219	0.006084	0.333633	-0.20468	1					
DAYS_ID_PUBLISH	-0.03212	0.003507	-0.01223	0.006717	-0.01403	0.270825	-0.27038	0.104299	1				
CNT_FAM_MEMBERS	0.880453	0.011226	0.063997	0.07738	0.061573	0.277241	-0.22982	0.170109	-0.02607	1			
REGION_RATING_CLIENT	0.025914	-0.03819	-0.10051	-0.1258	-0.10364	0.016779	0.034322	0.087518	-0.00231	0.025985	1		
REGION_RATING_CLIENT_W_CITY	0.022778	-0.04072	-0.10949	-0.13932	-0.11171	0.014552	0.03683	0.079792	-0.00731	0.025165	0.95071	1	
DAYS_LAST_PHONE_CHANGE	-0.00203	-0.0048	-0.07618	-0.06726	-0.07971	0.080196	0.027516	0.052146	0.09138	-0.02271	0.027327	0.026789	1

5) Identify Top Correlations for different scenarios:

Results: correlation for 0&1

columns	target
CNT_CHILDREN	0.026
AMT_INCOME_TOTAL	0.011
AMT_CREDIT	-0.032
AMT_ANNUITY	-0.012
AMT_GOODS_PRICE	-0.041
DAYS_BIRTH	0.077
DAYS_EMPLOYED	-0.040
DAYS_REGISTRATION	0.042
DAYS_ID_PUBLISH	0.047
CNT_FAM_MEMBERS	0.013
REGION_RATING_CLIENT	0.066
REGION_RATING_CLIENT_W_CITY	0.067
DAYS_LAST_PHONE_CHANGE	0.056



5) Identify Top Correlations for different scenarios:

Results: correlation for 1

Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.01011	1											
AMT_CREDIT	0.007602	0.015271	1										
AMT_ANNUITY	0.029173	0.018005	0.749665	1									
AMT_GOODS_PRICE	-0.00108	0.01327	0.982268	0.749504	1								
DAYS_BIRTH	0.249673	0.009034	-0.14251	-0.00875	-0.14101	1							
DAYS_EMPLOYED	-0.18932	-0.01156	0.01604	-0.07956	0.020235	-0.58148	1						
DAYS_REGISTRATION	0.152113	-0.00956	-0.04284	0.021582	-0.04332	0.288438	-0.18872	1					
DAYS_ID_PUBLISH	-0.04236	-0.00912	-0.04377	-0.02132	-0.04972	0.247897	-0.23006	0.090291	1				
CNT_FAM_MEMBERS	0.892522	0.013122	0.061249	0.075838	0.055136	0.199141	-0.18356	0.151787	-0.04404	1			
REGION_RATING_CLIENT	0.055516	-0.01285	-0.04502	-0.06158	-0.0513	0.045027	-0.00915	0.115625	0.025335	0.05728	1		
REGION_RATING_CLIENT_W_CITY	0.054802	-0.01267	-0.05295	-0.07942	-0.05669	0.038087	-0.00414	0.108123	0.014431	0.057988	0.950769	1	
DAYS_LAST_PHONE_CHANGE	0.011339	0.012457	-0.12454	-0.10047	-0.12883	0.124609	-0.01573	0.078605	0.138088	-0.00573	0.026186	0.022309	1

5) Identify Top Correlations for different scenarios:

Results: correlation for 0

Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.03632	1											
AMT_CREDIT	0.005705	0.377966	1										
AMT_ANNUITY	0.026384	0.451135	0.770773	1									
AMT_GOODS_PRICE	0.001518	0.384576	0.987	0.775835	1								
DAYS_BIRTH	0.335876	0.073769	-0.05108	0.009911	-0.04877	1							
DAYS_EMPLOYED	-0.24359	-0.1627	-0.07737	-0.11301	-0.07511	-0.61529	1						
DAYS_REGISTRATION	0.183072	0.068934	0.008054	0.034609	0.01126	0.335028	-0.20437	1					
DAYS_ID_PUBLISH	-0.03254	0.032286	-0.00829	0.009427	-0.00939	0.270073	-0.27222	0.103549	1				
CNT_FAM_MEMBERS	0.879238	0.041599	0.064877	0.077893	0.062892	0.284379	-0.23373	0.171483	-0.02505	1			
REGION_RATING_CLIENT	0.021289	-0.20503	-0.10256	-0.12992	-0.10484	0.009025	0.040506	0.082563	-0.0081	0.022204	1		
REGION_RATING_CLIENT_W_CITY	0.017873	-0.22004	-0.11164	-0.1432	-0.11312	0.007084	0.042899	0.074746	-0.01267	0.021214	0.950468	1	
DAYS_LAST_PHONE_CHANGE	-0.00482	-0.0495	-0.0712	-0.06445	-0.07424	0.07254	0.032952	0.04778	0.085063	-0.02504	0.023515	0.023179	1

Insights for bank loan project

SOME INFORMATION IS MISSING IN OUR DATA. WE'VE REMOVED COLUMNS WITH A LOT OF MISSING INFO AND FILLED IN THE GAPS USING TYPICAL VALUES LIKE THE MIDDLE NUMBER OR THE MOST COMMON ONE.

OUR DATA HAS SOME WEIRD VALUES THAT DON'T FIT THE USUAL PATTERN. WE NEED TO USE SPECIAL METHODS TO HANDLE THESE ODDITIES.

THE DATA IS NOT SPREAD OUT EVENLY ACROSS DIFFERENT CATEGORIES.

PEOPLE WITH LOWER INCOMES, WHO ARE MARRIED, WORKING, AND AROUND 38-39 YEARS OLD, TEND TO APPLY FOR LOANS THE MOST. INTERESTINGLY, THEY ALSO HAVE A HIGHER CHANCE OF NOT REPAYING THOSE LOANS.

DIFFERENT PIECES OF INFORMATION IN OUR DATA ARE CONNECTED, AND THE MOST LINKED ONE IS SOMETHING CALLED "DAYS_BIRTH."

EXCEL FILE LINK FOR BANK LOAN PROJECT

- https://drive.google.com/drive/folders/1eCWYStVJzokbkv-pFOrA0p39suJL3IbL?usp=sharing

Conclusion

I have successfully completed project using Excel, Power point.

I have learned to deal with large datasets which has many missing values and outliers.

Thank You