Task-6: Bank Loan Case Study

Contents:

1. Project Description
2. Tech Stack Used
3. Approach
4. Insights
5. Results and Conclusion

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 Description:

- The Bank Loan Case Study project, My aim is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.
- My task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.
- Through in-depth data analysis using Excel, Data Visualization and Statistics techniques this project seeks to extract valuable insights and to identify patterns that indicate if a customer will have difficulty paying their installments.

- **Software Used: Microsoft Excel 365**

- **NOTE: ALL THE LINKS FOR CLEANED DATASET AND SOLUTIONS DATASET ARE PROVIDED BELOW !!!**

DATA HANDLING

My Approach:

- I have gone through the dataset and understood all the given columns. Then I have observed that there are a total of 128 Columns and 49999 Rows. This dataset consists of unwanted columns, Null values and Blank rows. So, I have decided to Clean this dataset thoroughly.
- Full Results Dataset: https://docs.google.com/spreadsheets/d/1mAHEXTZ_rjQ5VJXV3DHep1SJhlErK55p/edit?usp=sharing&ouid=113249253121491889461&rtpof=true&sd=true

DATA ANALYSIS

1) Identify Missing Data and Deal with it Appropriately

As a data analyst, you come across missing data in the loan application dataset.
It is essential to handle missing data effectively to ensure the accuracy of the analysis.
Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

FUNTIONS I HAVE USED: LINK FOR THE PROJECT

- =COUNTBLANK(A2:A50000)
- =COUNTBLANK(A2:A50000)/COUNTA(A2:A50000)*100
- Firstly, after calculating the Null Values I have deleted the columns which has the null values percentage greater than 25%. Then I have replaced the null values with the median for the columns which has null values less than 25%
- =MEDIAN(J2:J50000)
- By the end, I left with total of 72 Columns and 49999 Rows.
- Thus, In this Task, I learned to handle missing values in a large dataset.

DATA ANALYSIS

1) Identify Missing Data and Deal with it Appropriately

Results: Before Cleaning

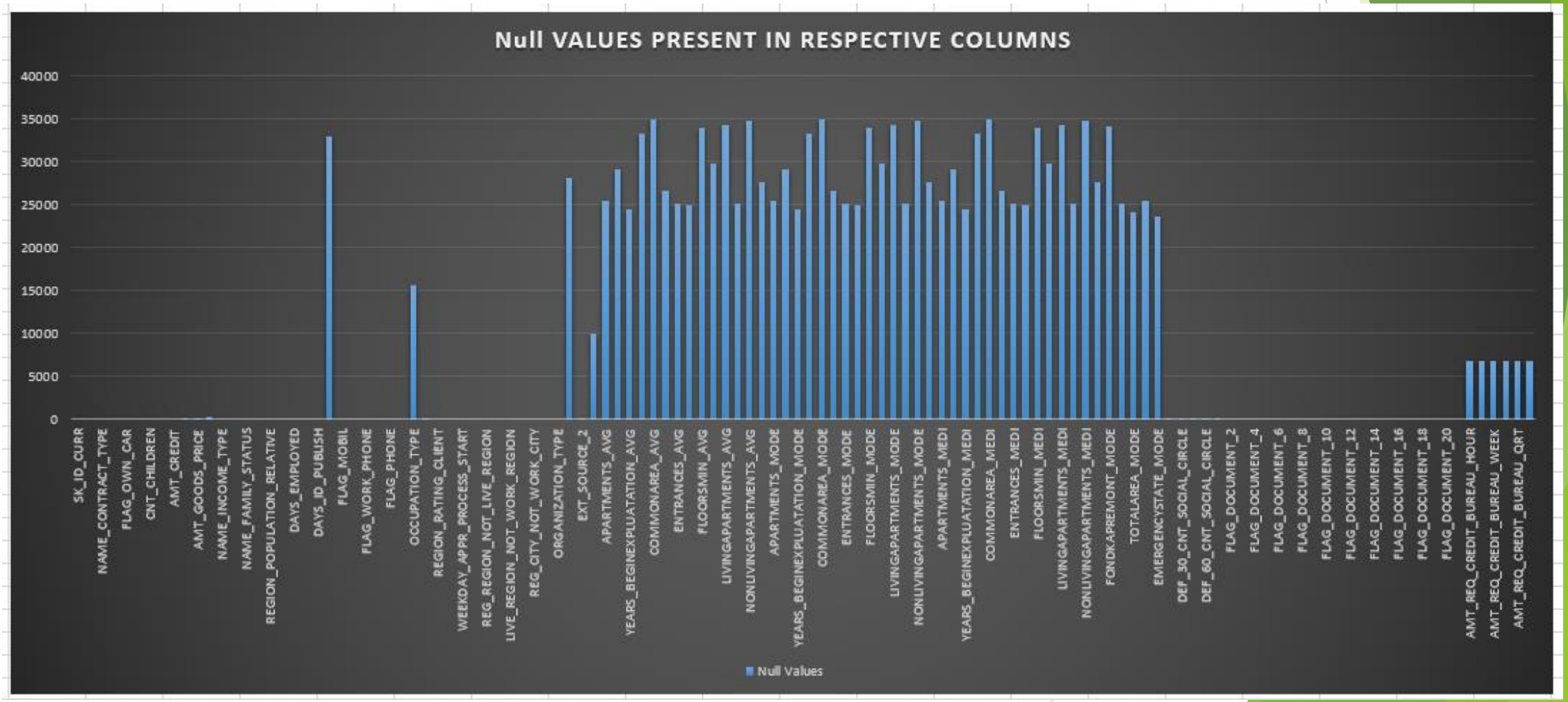
Indicates Null Values | Indicates Null Values Greater Than 25%

Columns	Null Values	Percentage
SK_ID_CURR	0	0
TARGET	0	0
NAME_CONTRACT_TYPE	0	0
CODE_GENDER	0	0
FLAG_OWN_CAR	0	0
FLAG_OWN_REALTY	0	0
CNT_CHILDREN	0	0
AMT_INCOME_TOTAL	0	0
AMT_CREDIT	0	0
AMT_ANNUITY	1	0.00200008
AMT_GOODS_PRICE	38	0.076059326
NAME_TYPE_SUITE	192	0.385487984
NAME_INCOME_TYPE	0	0
NAME_EDUCATION_TYPE	0	0
NAME_FAMILY_STATUS	0	0
NAME_HOUSING_TYPE	0	0
REGION_POPULATION_RELATIVE	0	0
DAYS_BIRTH	0	0
DAYS_EMPLOYED	0	0
DAYS_REGISTRATION	0	0
DAYS_ID_PUBLISH	0	0
OWN_CAR_AGE	32950	193.2664672
FLAG_MOBIL	0	0
FLAG_EMP_PHONE	0	0
FLAG_WORK_PHONE	0	0
FLAG_CONT_MOBILE	0	0
FLAG_PHONE	0	0
FLAG_EMAIL	0	0
OCCUPATION_TYPE	15654	45.57868685
CNT_FAM_MEMBERS	1	0.00200008
REGION_RATING_CLIENT	0	0
REGION_RATING_CLIENT_W_CITY	0	0
WEEKDAY_APPR_PROCESS_START	0	0
HOUR_APPR_PROCESS_START	0	0
REG_REGION_NOT_LIVE_REGION	0	0
REG_REGION_NOT_WORK_REGION	0	0
LIVE_REGION_NOT_WORK_REGION	0	0
REG_CITY_NOT_LIVE_CITY	0	0
REG_CITY_NOT_WORK_CITY	0	0
LIVE_CITY_NOT_WORK_CITY	0	0
ORGANIZATION_TYPE	0	0
EXT_SOURCE_1	28172	129.0695011
EXT_SOURCE_2	126	0.25264171
EXT_SOURCE_3	9944	24.82586444
APARTMENTS_AVG	25385	103.1323637
BASEMENTAREA_AVG	29199	140.3798077
YEARS_BEGINEXPLUATATION_AVG	24394	95.27045499
YEARS_BUILD_AVG	33239	198.323389
COMMONAREA_AVG	34960	232.4622648
ELEVATORS_AVG	26651	114.146822
ENTRANCES_AVG	25195	101.5763587
FLOORSMAX_AVG	24875	99.00891578
FLOORSMIN_AVG	33894	210.45638
LANDAREA_AVG	29721	146.5677088
LIVINGAPARTMENTS_AVG	34226	216.9910607
LIVINGAREA_AVG	25137	101.1061057
NONLIVINGAPARTMENTS_AVG	34714	227.1115473
NONLIVINGAREA_AVG	27572	122.9410978
APARTMENTS_MODE	25385	103.1323637
BASEMENTAREA_MODE	29199	140.3798077
YEARS_BEGINEXPLUATATION_MODE	24394	95.27045499
YEARS_BUILD_MODE	33239	198.323389
COMMONAREA_MODE	34960	232.4622648
ELEVATORS_MODE	26651	114.146822
ENTRANCES_MODE	25195	101.5763587
FLOORSMAX_MODE	24875	99.00891578
FLOORSMIN_MODE	33894	210.45638
LANDAREA_MODE	29721	146.5677088
LIVINGAPARTMENTS_MODE	34226	216.9910607
LIVINGAREA_MODE	25137	101.1061057
NONLIVINGAPARTMENTS_MODE	34714	227.1115473
NONLIVINGAREA_MODE	27572	122.9410978
APARTMENTS_MEDI	25385	103.1323637
BASEMENTAREA_MEDI	29199	140.3798077
YEARS_BEGINEXPLUATATION_MEDI	24394	95.27045499
YEARS_BUILD_MEDI	33239	198.323389
COMMONAREA_MEDI	34960	232.4622648
ELEVATORS_MEDI	26651	114.146822
ENTRANCES_MEDI	25195	101.5763587
FLOORSMAX_MEDI	24875	99.00891578
FLOORSMIN_MEDI	33894	210.45638
LANDAREA_MEDI	29721	146.5677088
LIVINGAPARTMENTS_MEDI	34226	216.9910607
LIVINGAREA_MEDI	25137	101.1061057
NONLIVINGAPARTMENTS_MEDI	34714	227.1115473
NONLIVINGAREA_MEDI	27572	122.9410978
FONDKAPREMONT_MODE	34191	216.2892206
HOUSETYPE_MODE	25075	100.6058418
TOTALAREA_MODE	24148	93.41224711
WALLSMATERIAL_MODE	25459	103.7449063
EMERGENCYSTATE_MODE	23698	90.10303791
OBS_30_CNT_SOCIAL_CIRCLE	168	0.337139532
DEF_30_CNT_SOCIAL_CIRCLE	168	0.337139532
OBS_60_CNT_SOCIAL_CIRCLE	168	0.337139532
DEF_60_CNT_SOCIAL_CIRCLE	168	0.337139532
DAYS_LAST_PHONE_CHANGE	1	0.00200008
FLAG_DOCUMENT_2	0	0
FLAG_DOCUMENT_3	0	0
FLAG_DOCUMENT_4	0	0
FLAG_DOCUMENT_5	0	0
FLAG_DOCUMENT_6	0	0
FLAG_DOCUMENT_7	0	0
FLAG_DOCUMENT_8	0	0
FLAG_DOCUMENT_9	0	0
FLAG_DOCUMENT_10	0	0
FLAG_DOCUMENT_11	0	0
FLAG_DOCUMENT_12	0	0
FLAG_DOCUMENT_13	0	0
FLAG_DOCUMENT_14	0	0
FLAG_DOCUMENT_15	0	0
FLAG_DOCUMENT_16	0	0
FLAG_DOCUMENT_17	0	0
FLAG_DOCUMENT_18	0	0
FLAG_DOCUMENT_19	0	0
FLAG_DOCUMENT_20	0	0
FLAG_DOCUMENT_21	0	0
AMT_REQ_CREDIT_BUREAU_HOUR	6734	15.56454409
AMT_REQ_CREDIT_BUREAU_DAY	6734	15.56454409
AMT_REQ_CREDIT_BUREAU_WEEK	6734	15.56454409
AMT_REQ_CREDIT_BUREAU_MON	6734	15.56454409

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

Results: After Cleaning

COLUMNS	MEDIAN
AMT_ANNUITY	24939
AMT_GOODS_PRICE	450000
CNT_FAM_MEMBERS	2
EXT_SOURCE_2	0.565585366
EXT_SOURCE_3	0.53527625
OBS_30_CNT_SOCIAL_CIRCLE	0
DEF_30_CNT_SOCIAL_CIRCLE	0
OBS_60_CNT_SOCIAL_CIRCLE	0
DEF_60_CNT_SOCIAL_CIRCLE	0
DAYS_LAST_PHONE_CHANGE	-755
AMT_REQ_CREDIT_BUREAU_HOUR	0
AMT_REQ_CREDIT_BUREAU_DAY	0
AMT_REQ_CREDIT_BUREAU_WEEK	0
AMT_REQ_CREDIT_BUREAU_MON	0
AMT_REQ_CREDIT_BUREAU_QRT	0
AMT_REQ_CREDIT_BUREAU_YEAR	1

- I have used these values to replace the null values in the columns which has null values less than 25%.
- For text based columns I used the mode function and replaced the null values with the most repeated text.

DATA ANALYSIS

2) Identify Outliers in the Dataset:

Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Functions I have Used: LINK FOR THE PROJECT

- =QUARTILE.EXC(A2:A50000,1) **[QUARTILE-1]**
- =QUARTILE.EXC(A2:A50000,3) **[QUARTILE-3]**
- =O10-O9 **[IQR]**
- =O9-(1.5*O11) **[LOWER BOUND]**
- =O10+(1.5*O11) **[UPPER BOUND]**
- By using these functions, I have Calculated Quartile-1, Quartile-2, Inter Quartile Range (IQR), Lower Bound, Upper Bound.

DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:

A) CNT_CHILDREN

CALCULATIONS	VALUES
QUARTILE Q1	0
QUARTILE Q3	1
Inter Quartile Range IQR	1
Lower Bound	-1.5
Upper Bound	2.5

B) AMT_INCOME_TOTAL

CALCULATIONS	VALUES
QUARTILE Q1	112500
QUARTILE Q3	202500
Inter Quartile Range IQR	90000
Lower Bound	-22500
Upper Bound	337500

C) AMT_CREDIT

CALCULATIONS	VALUES
QUARTILE Q1	270000
QUARTILE Q3	808650
Inter Quartile Range IQR	538650
Lower Bound	-537975
Upper Bound	1616625

D) AMT_ANNUITY

CALCULATIONS	VALUES
QUARTILE Q1	16456.5
QUARTILE Q3	34596
Inter Quartile Range IQR	18139.5
Lower Bound	-10752.75
Upper Bound	61805.25

E) AMT_GOODS_PRICE

CALCULATIONS	VALUES
QUARTILE Q1	238500
QUARTILE Q3	679500
Inter Quartile Range IQR	441000
Lower Bound	-423000
Upper Bound	1341000

F) DAYS_BIRTH

CALCULATIONS	VALUES
QUARTILE Q1	-19644
QUARTILE Q3	-12378
Inter Quartile Range IQR	7266
Lower Bound	-30543
Upper Bound	-1479

G) DAYS_EMPLOYED

CALCULATIONS	VALUES
QUARTILE Q1	-2786
QUARTILE Q3	-292
Inter Quartile Range IQR	2494
Lower Bound	-6527
Upper Bound	3449

H) DAYS_REGISTRATION

CALCULATIONS	VALUES
QUARTILE Q1	-7464
QUARTILE Q3	-1998
Inter Quartile Range IQR	5466
Lower Bound	-15663
Upper Bound	6201

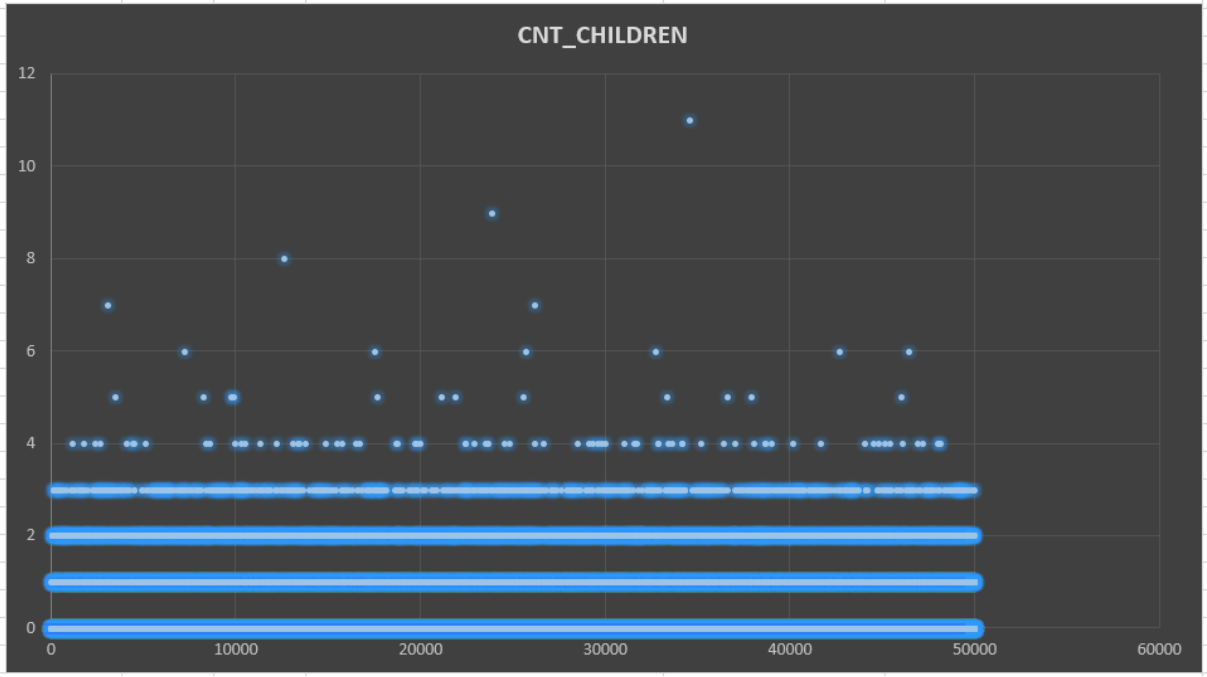
I) DAYS_ID_PUBLISH

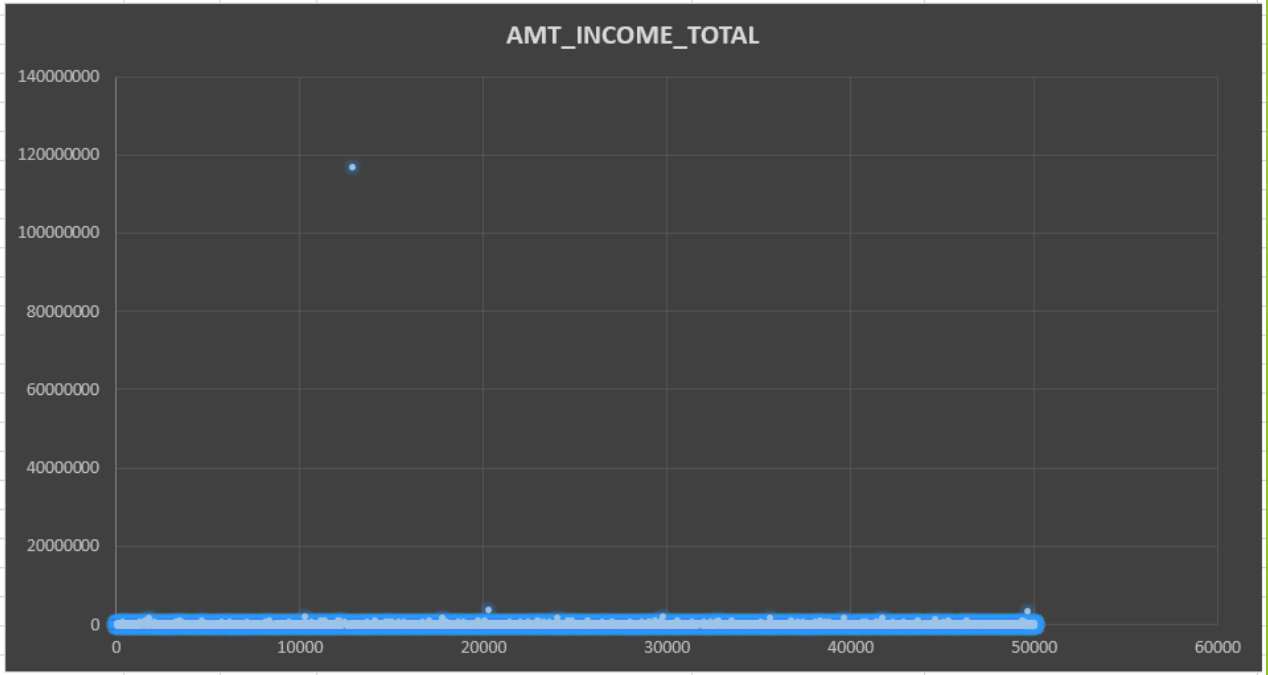
CALCULATIONS	VALUES
QUARTILE Q1	-4297
QUARTILE Q3	-1722
Inter Quartile Range IQR	2575
Lower Bound	-8159.5
Upper Bound	2140.5

DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:

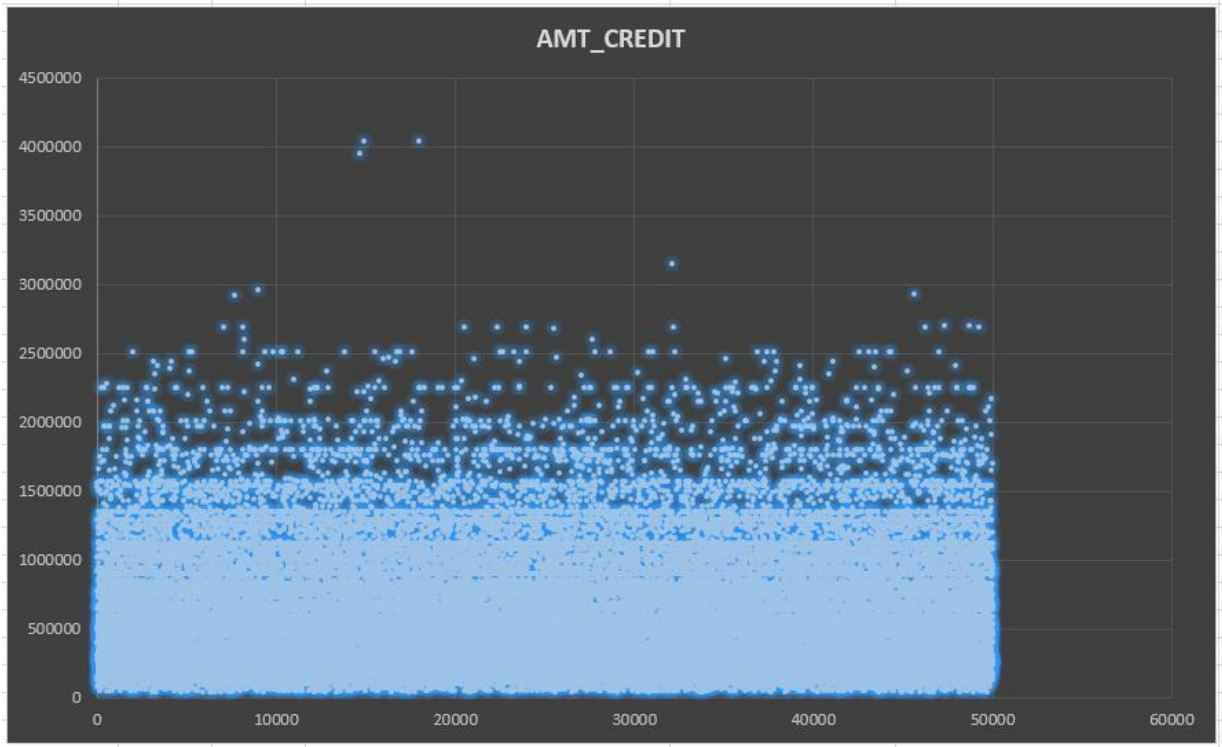


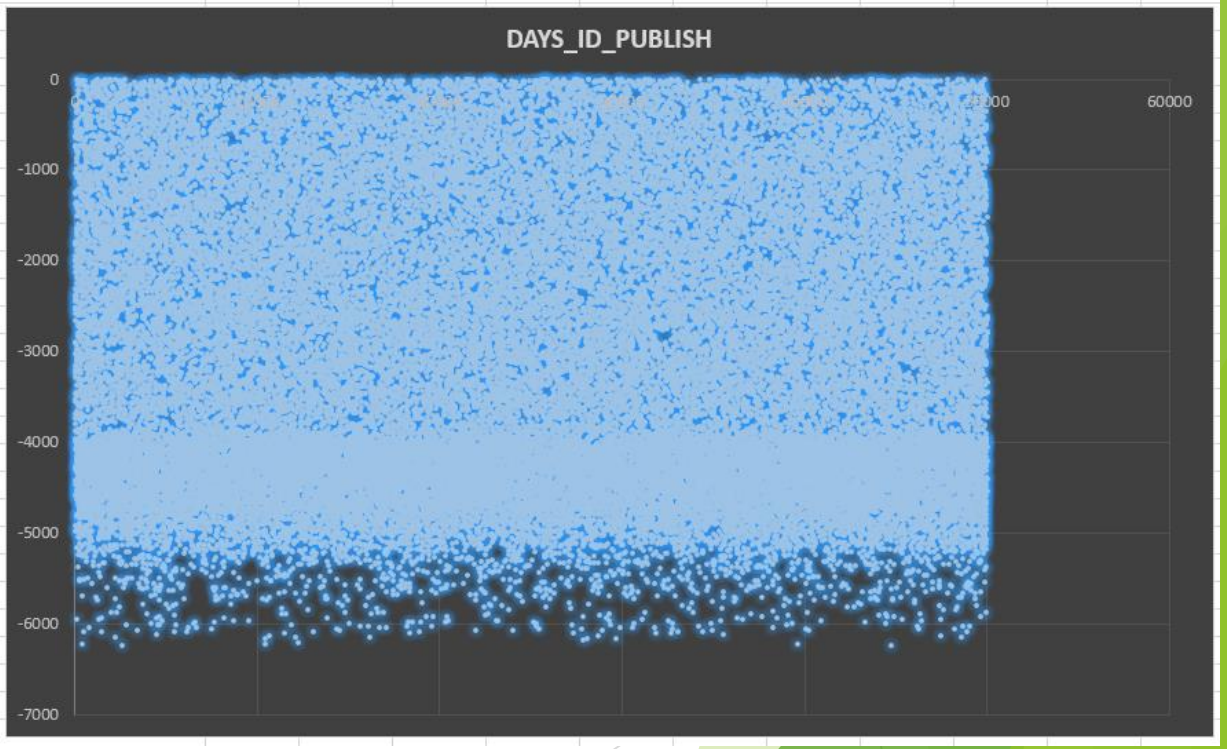


DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:

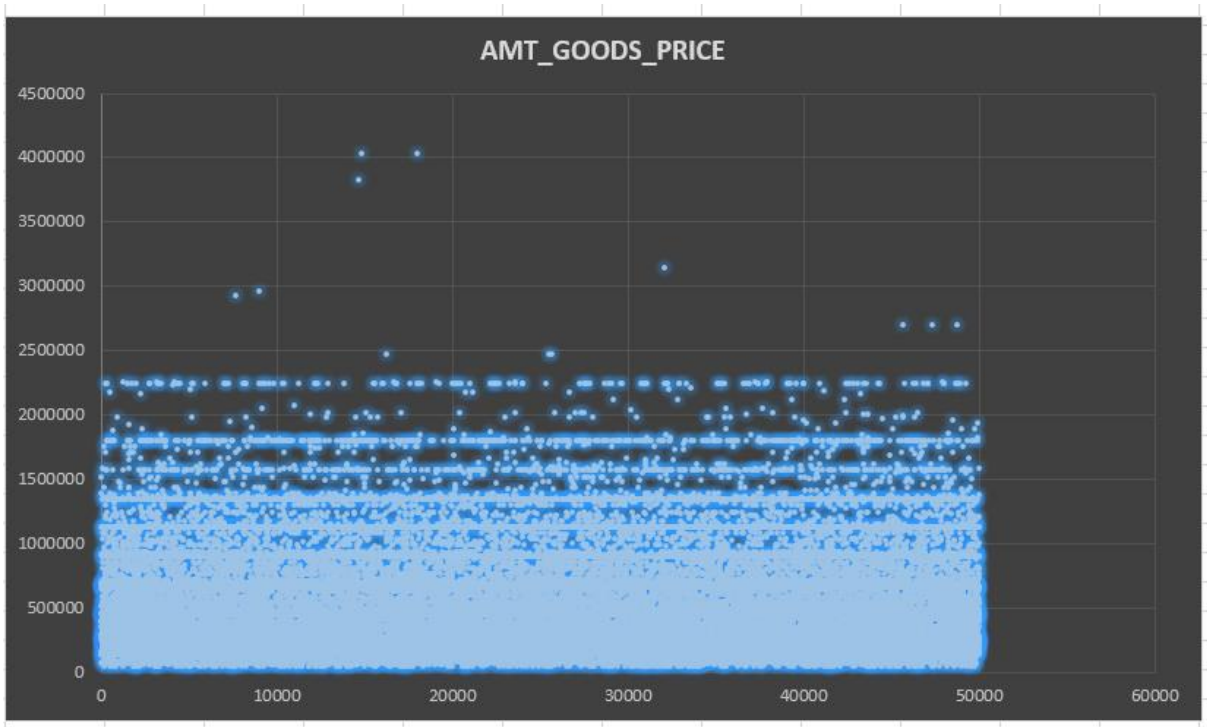




DATA ANALYSIS

2) Identify Outliers in the Dataset:

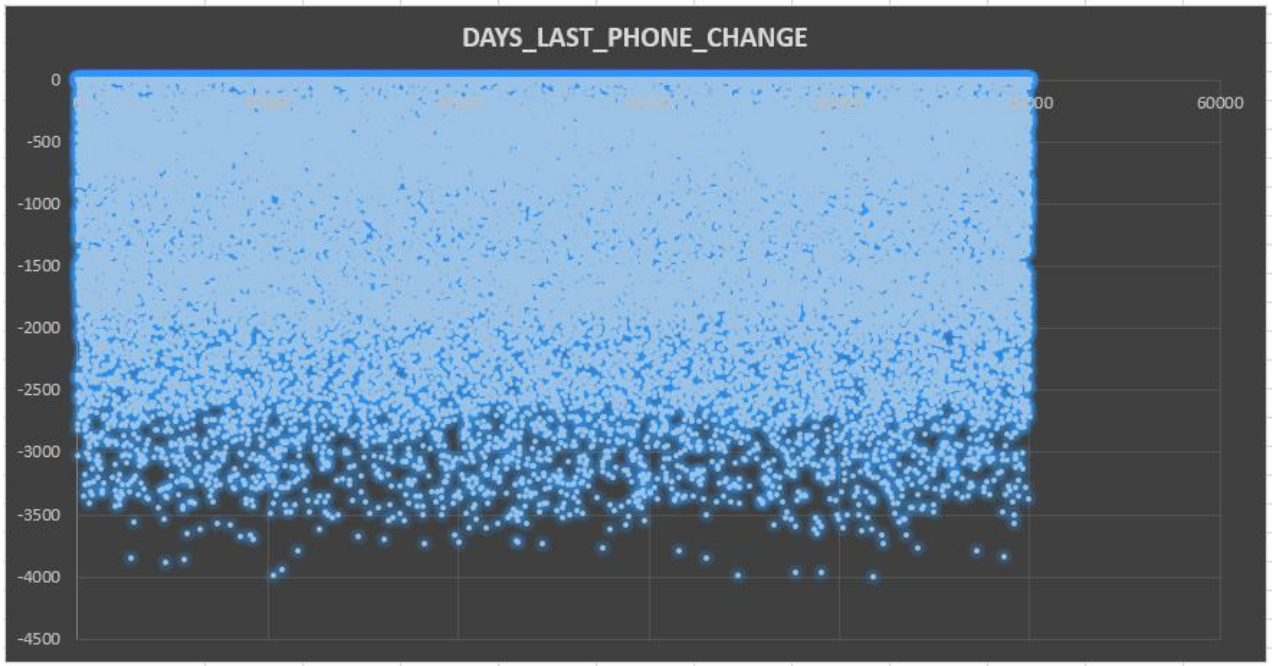
Results:

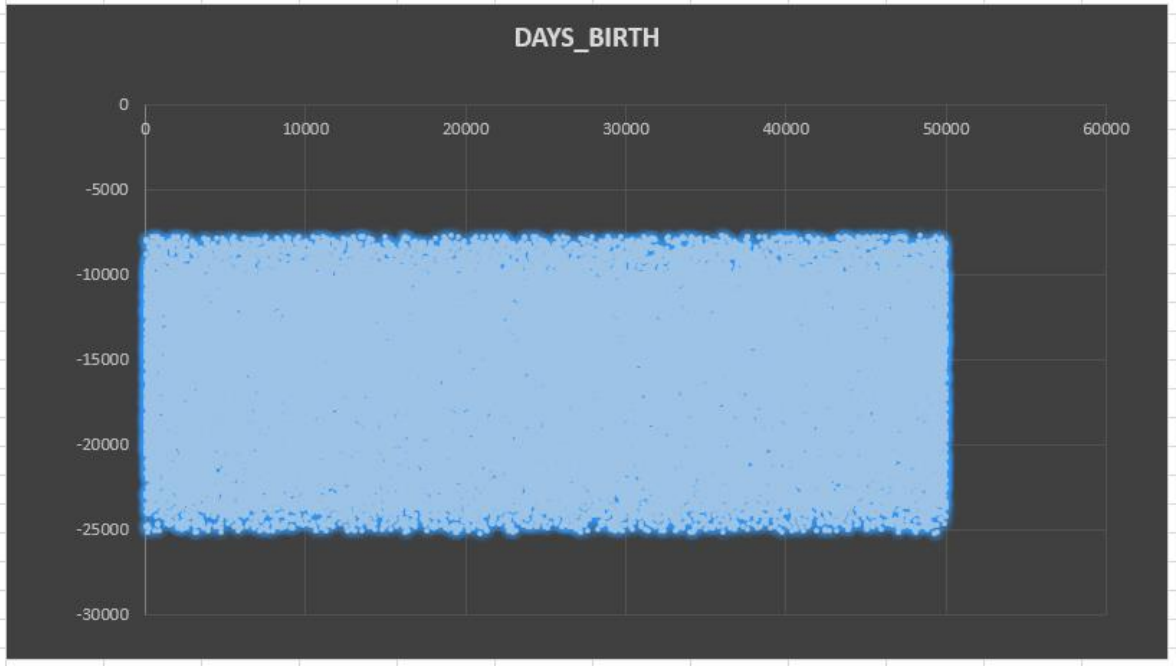


DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:

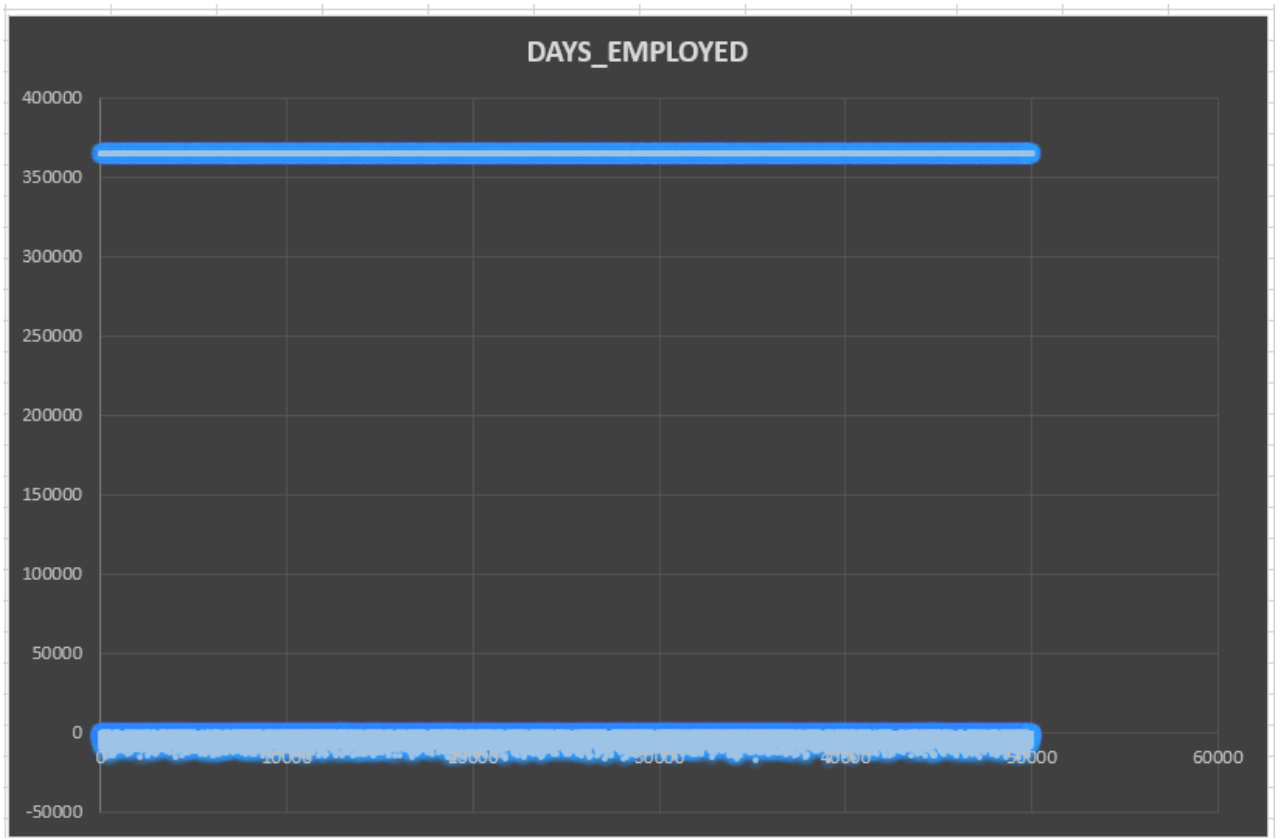




DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:



DATA ANALYSIS

2) Identify Outliers in the Dataset:

Results:

CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANG
0	202500	406597.5	24700.5	351000	-9461	-637	-3648	-2120	-1134
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
0	112500	450000	44509.5	450000	-12158	-1275	-6265	-2009	-1285
2	198000	641173.5	23157	553500	-17199	-768	-63	-735	-2411
0	121500	454500	15151.5	454500	-21077	-1288	-5474	-4270	-1541
0	99000	247275	17338.5	225000	-23920	365243	-9817	-4969	0
0	180000	540000	27000	540000	-16126	-1761	-8236	-4292	-540

- I have also highlighted the columns using conditional formatting of upper bound and lower bound

DATA ANALYSIS

3) Analyse Data Imbalance:

Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.

Functions that I have used: LINK FOR THE PROJECT

- =UNIQUE(B2:B50000)
- =COUNTIF(B2:B50000,1) OR =COUNTIF(B2:B50000,0)
- By using these functions, I have Calculated count/occurrence of a particular scenario in a column.
- I have also calculated the ratio of imbalance between these data.

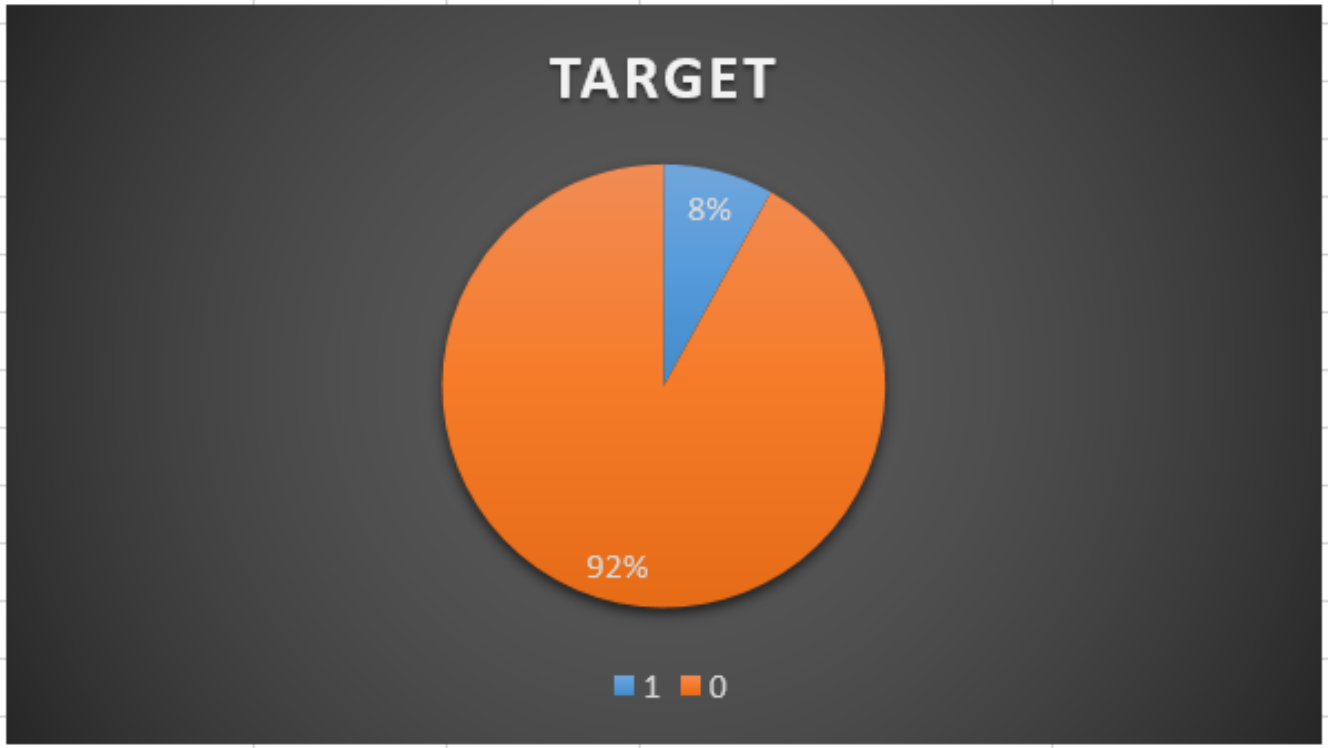
DATA ANALYSIS

3) Analyse Data Imbalance:

Results:

TARGET_COLUMN	OCCURRENCE
1	4026
0	45973

DATA IMBALANCE RATIO IS **8.757314076**

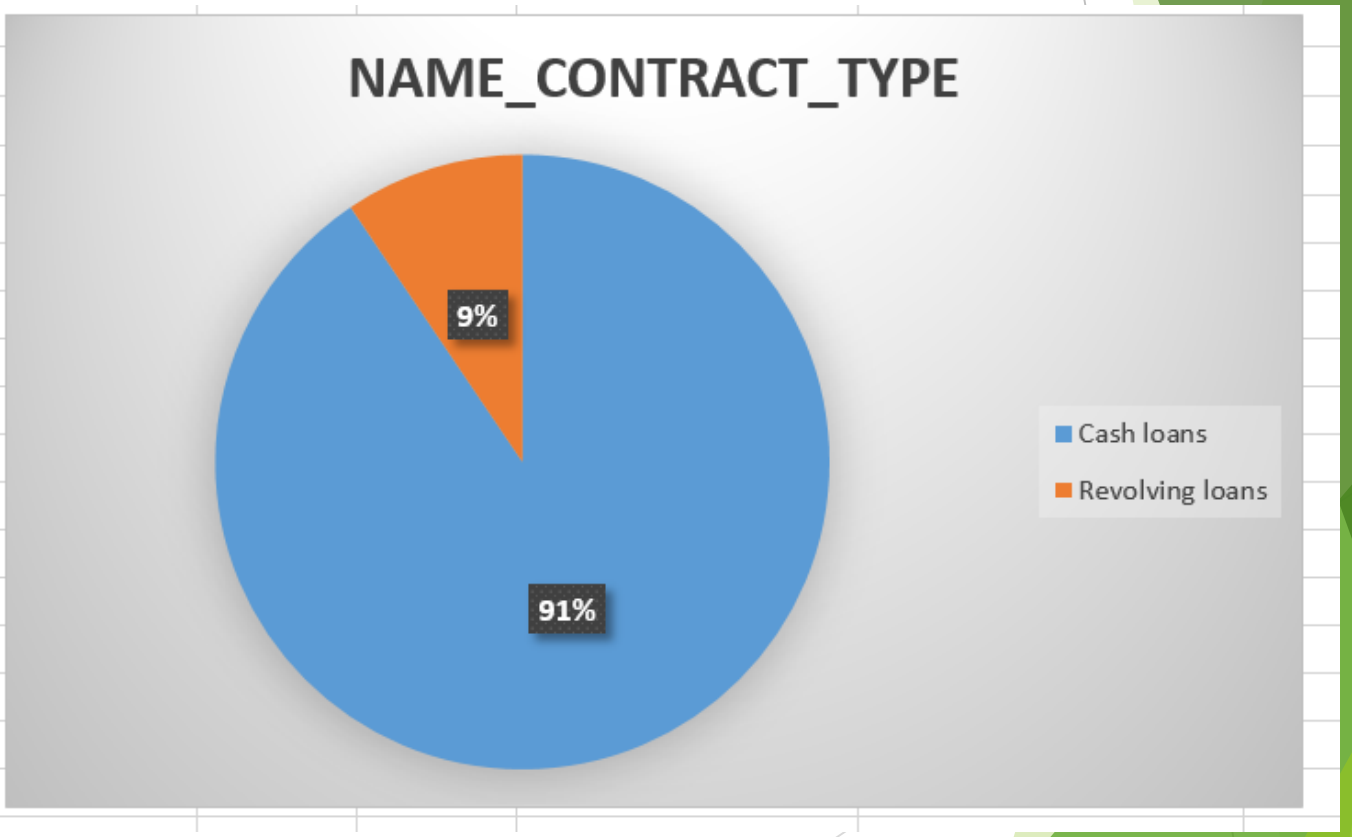


DATA ANALYSIS

3) Analyse Data Imbalance:

Results:

NAME_CONTRACT_TYPE	OCCURRENCE
Cash loans	45276
Revolving loans	4723
DATA IMBALANCE RATIO IS	**10.43157523**

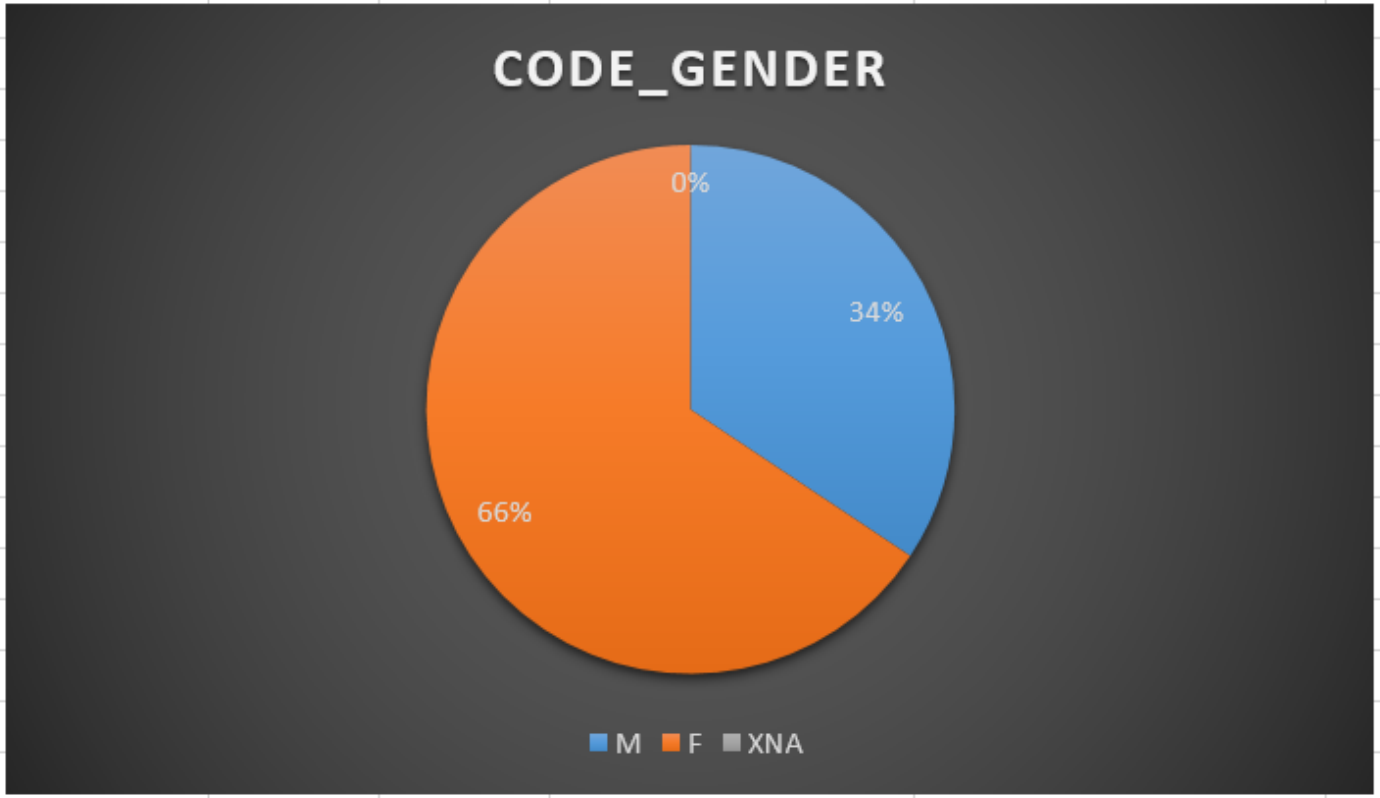


DATA ANALYSIS

3) Analyse Data Imbalance:

Results:

CODE_GENDER	OCCURRENCE
M	17174
F	32823
XNA	2

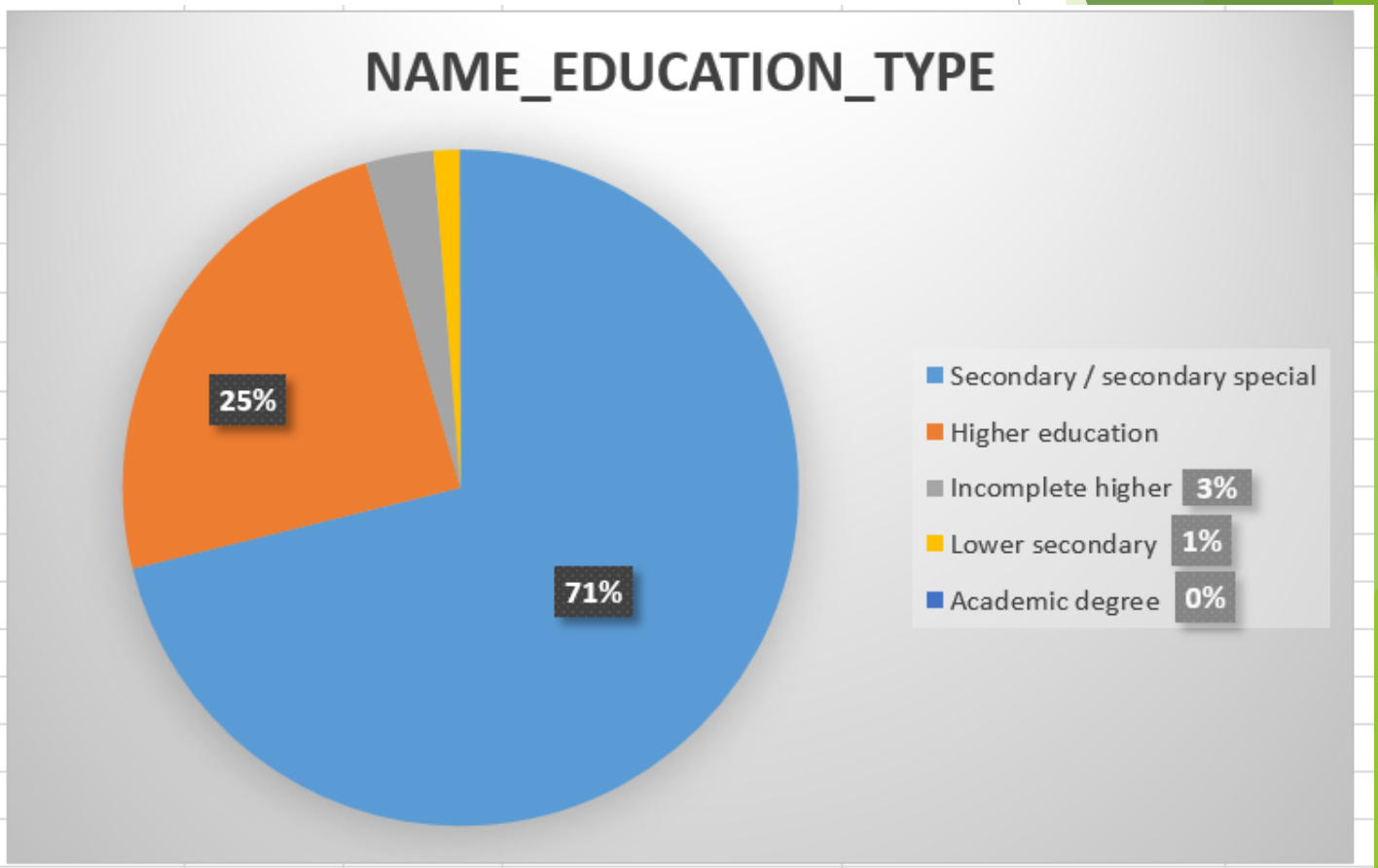


DATA ANALYSIS

3) Analyse Data Imbalance:

Results:

NAME_EDUCATION_TYPE	OCCURRENCE
special	35572
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Academic degree	20

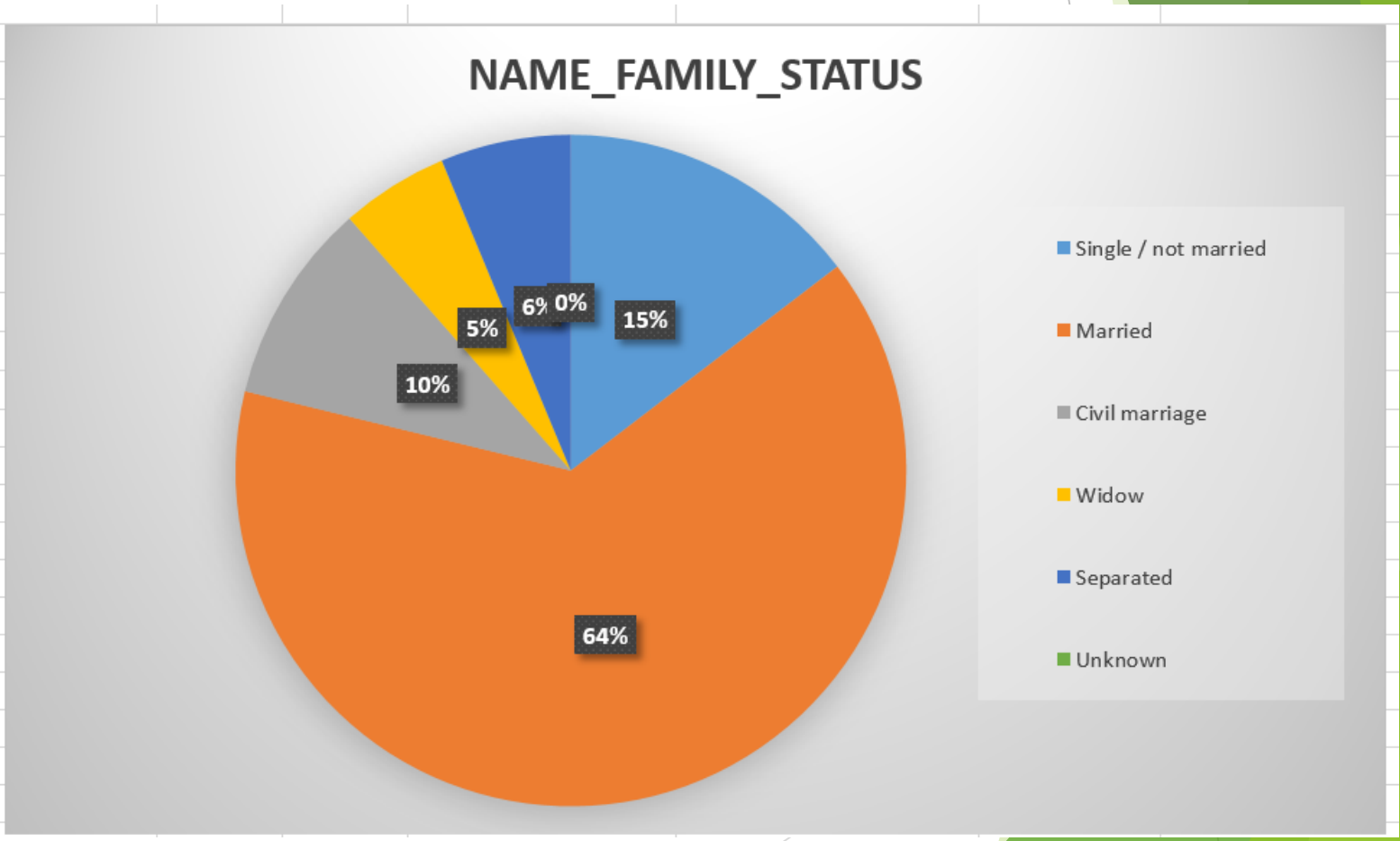


DATA ANALYSIS

3) Analyse Data Imbalance:

Results:

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



DATA ANALYSIS

3) Analyse Data Imbalance:

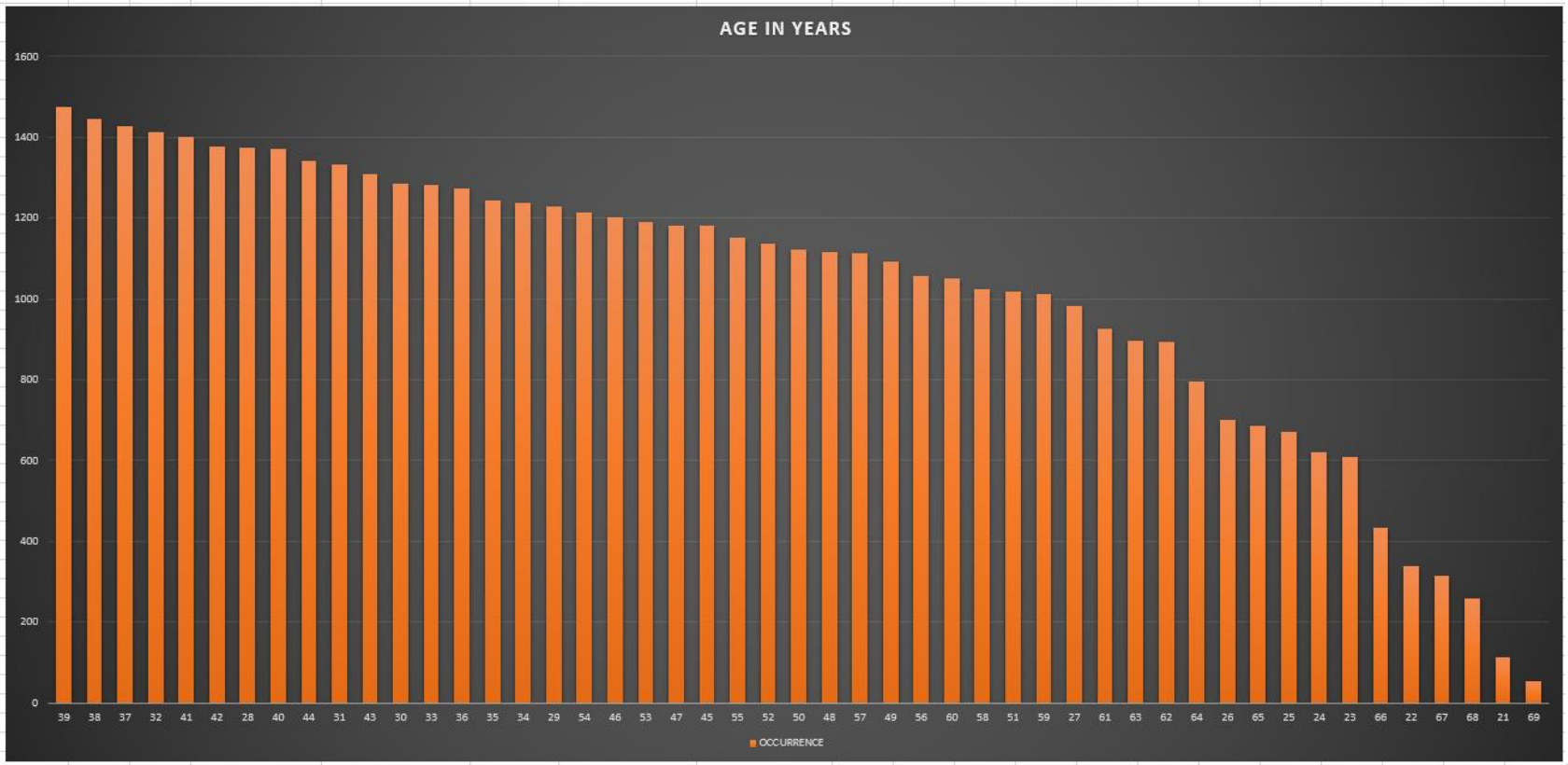
Results:

AGE	OCCURRENCE
39	1475
38	1447
37	1429
32	1414
41	1402
42	1376
28	1374
40	1372
44	1341
31	1332
43	1308
30	1286
33	1282
36	1273
35	1243
34	1239
29	1228
54	1213
46	1202
53	1191
47	1182
45	1182
55	1151
52	1138
50	1123
48	1117
57	1112
49	1092
56	1058
60	1052
58	1024
51	1017
59	1011
27	982
61	925
63	897
62	893
64	797
26	701
65	687
25	671
24	622
23	610
66	435
22	338
67	315
68	258
21	114
69	53

DATA ANALYSIS

3) Analyse Data Imbalance:

Results:



DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Functions that I have used: LINK FOR THE PROJECT

- I have created pivot tables using appropriate columns.
- Then used count function in pivot table to count occurrence them.
- Then plotted using Histogram.

DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Results:

VALUES	AMT_INCOME_TOTAL	AMT_CREDIT
AVERAGE	170767.5905	599700.5815
MEDIAN	145800	514777.5
MODE	135000	450000
STDEV	531819.0951	402415.4339

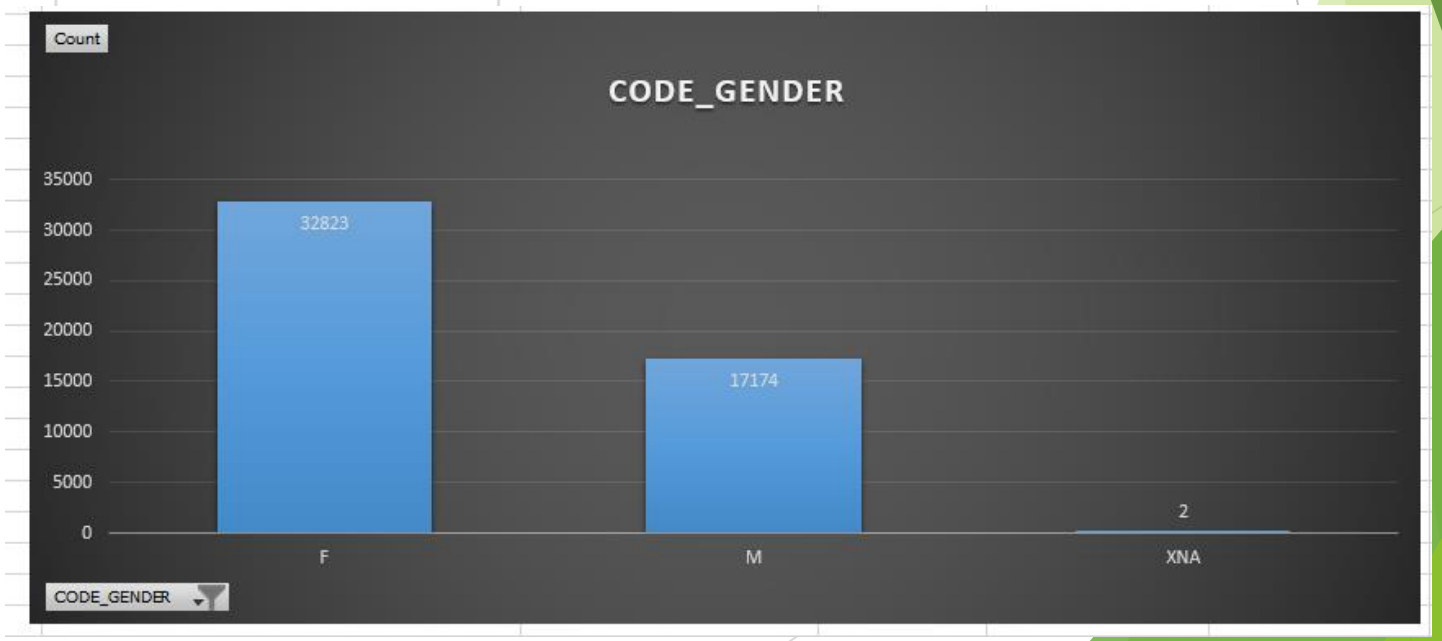
AMT_ANNUITY	DAYS_BIRTH	AGE
27107.33399	16022.04208	44
24939	15731	43
9000	11039	30
14562.80203	4361.40027	12

DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
F	32823
M	17174
XNA	2
Grand Total	**49999**

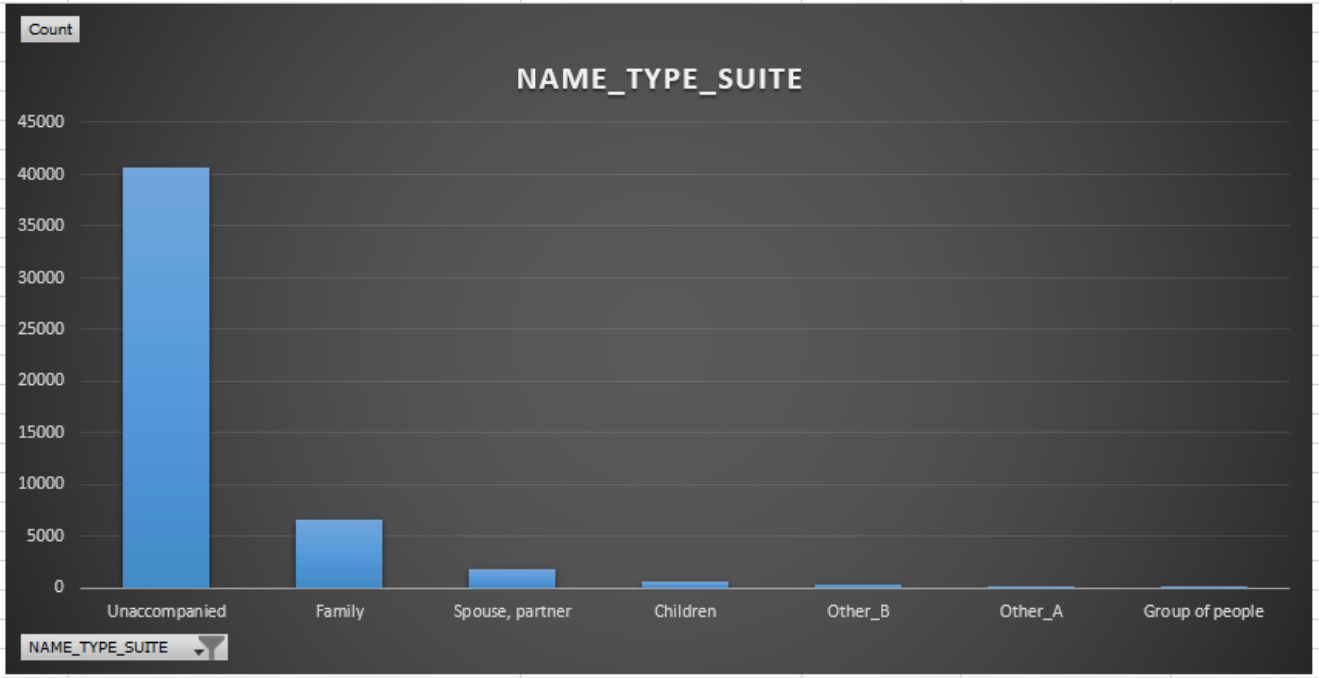


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
Unaccompanied	40585
Family	6577
Spouse, partner	1855
Children	546
Other_B	260
Other_A	139
Group of people	37
Grand Total	**49999**

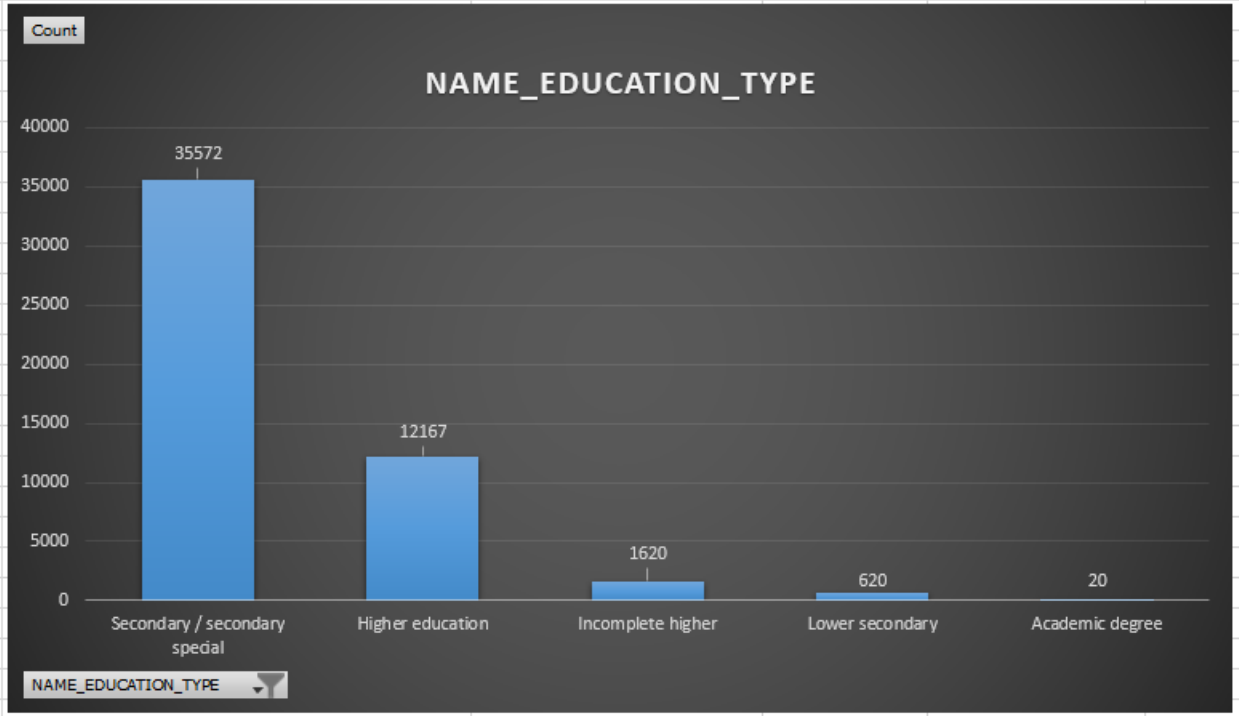


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
Secondary / secondary special	35572
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Academic degree	20
Grand Total	**49999**

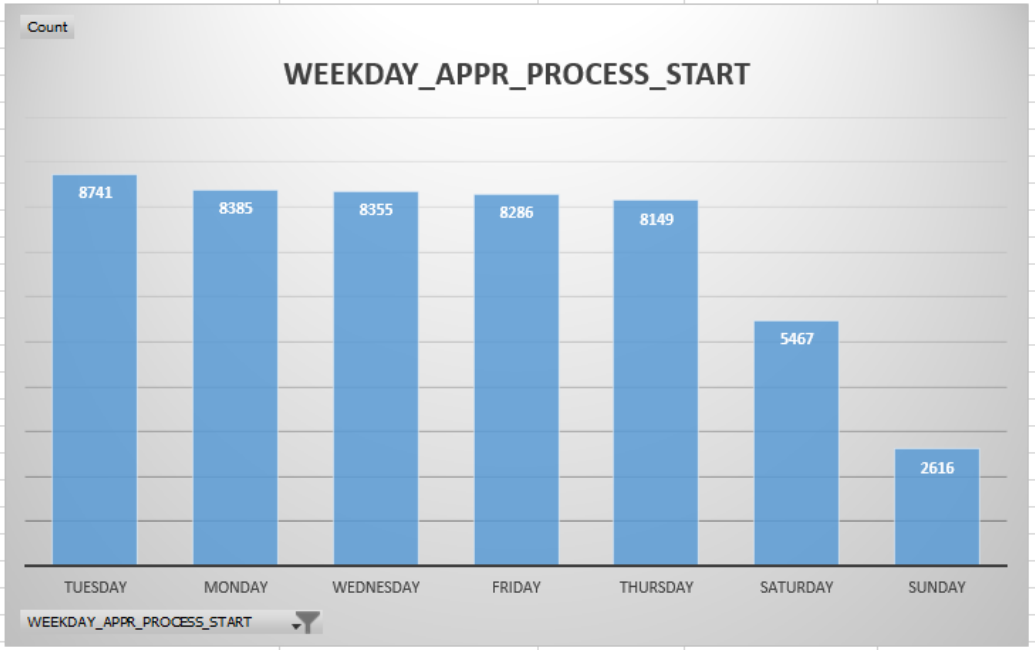


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
TUESDAY	8741
MONDAY	8385
WEDNESDAY	8355
FRIDAY	8286
THURSDAY	8149
SATURDAY	5467
SUNDAY	2616
Grand Total	**49999**

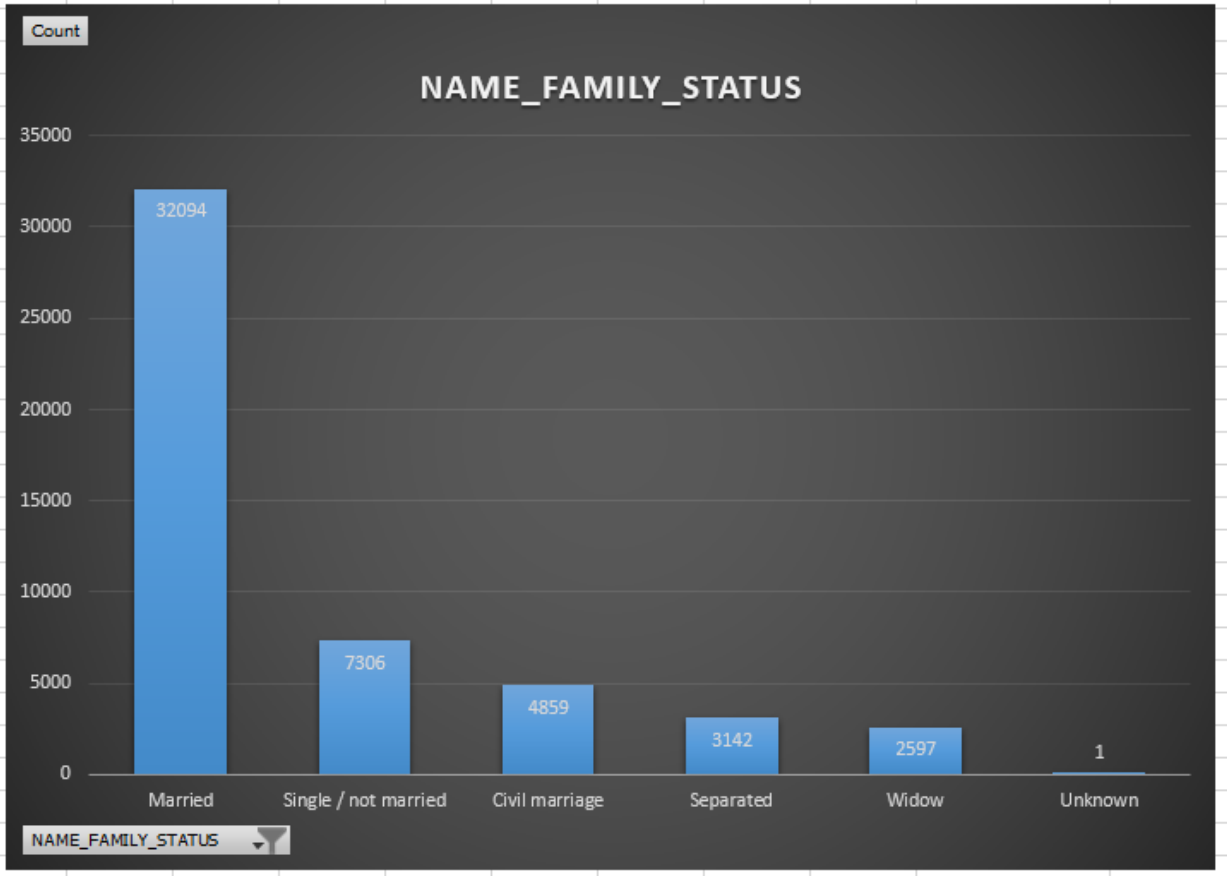


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

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

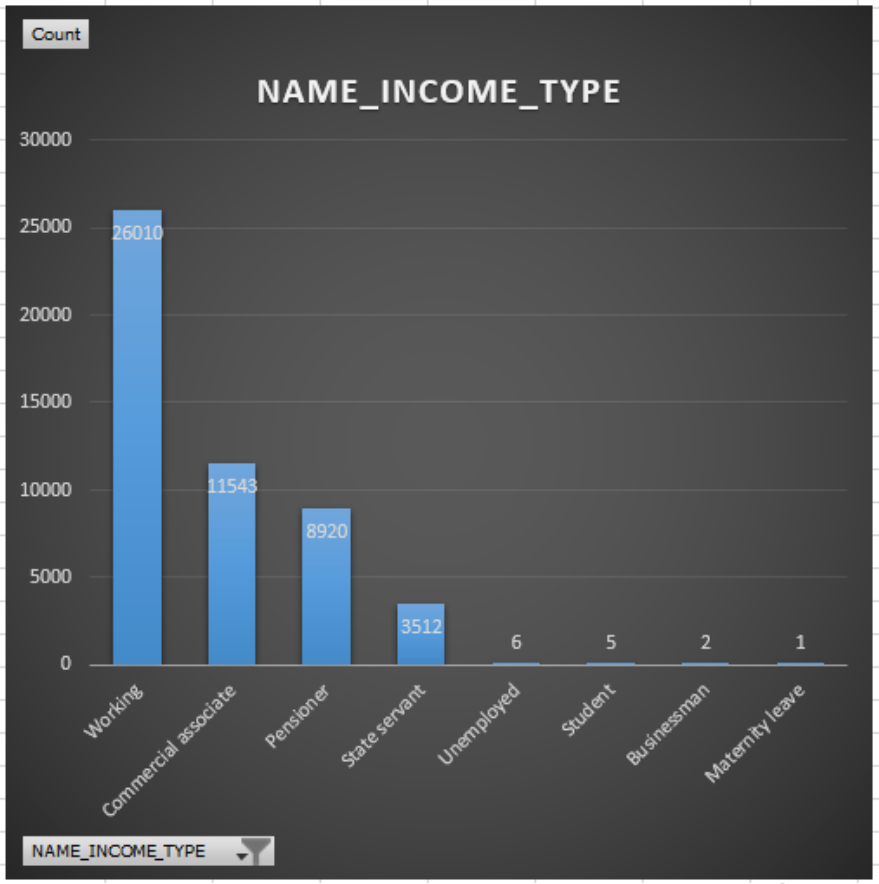


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
Working	26010
Commercial associate	11543
Pensioner	8920
State servant	3512
Unemployed	6
Student	5
Businessman	2
Maternity leave	1
Grand Total	**49999**

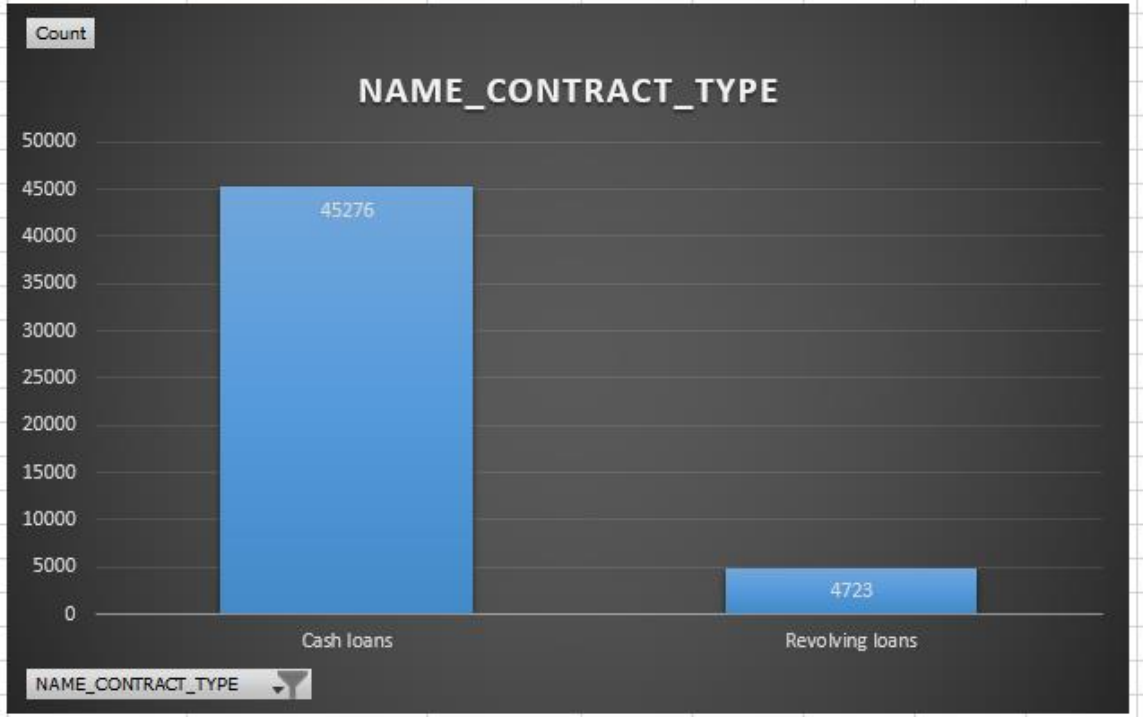


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**

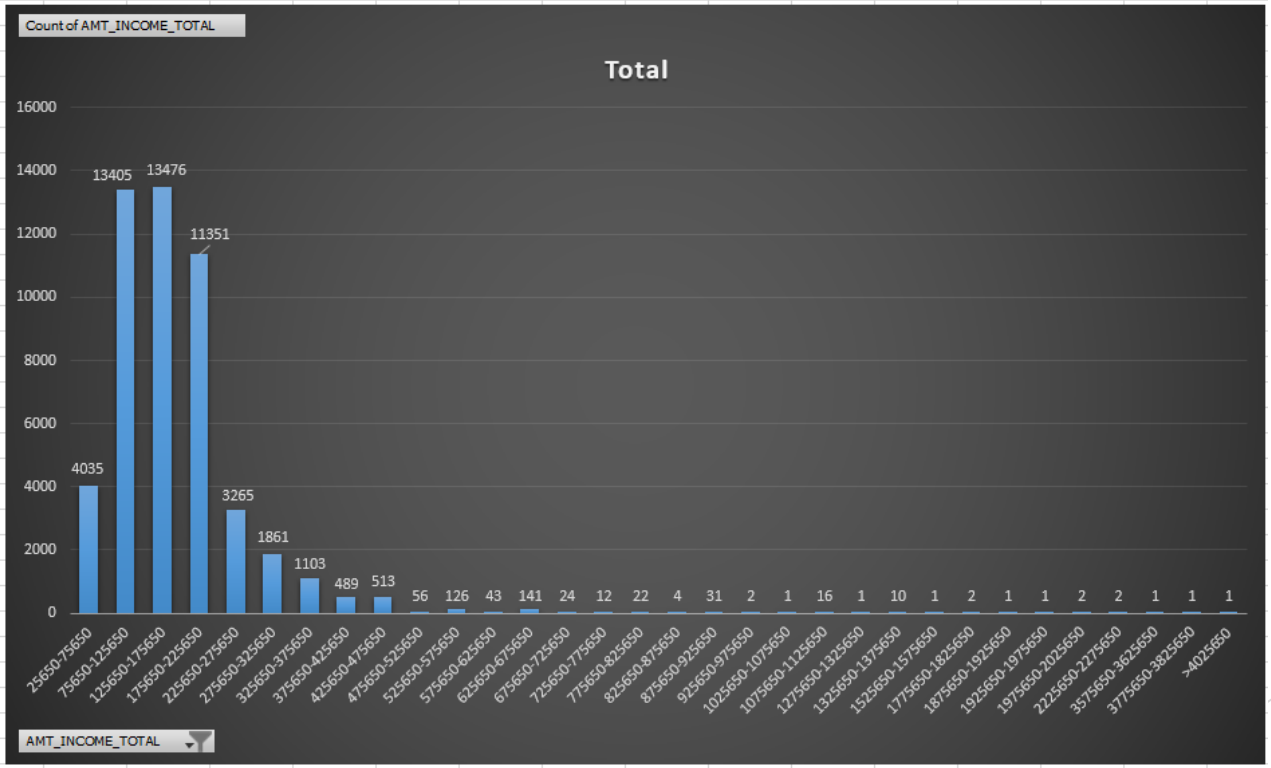


DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:

Row Labels	Count of AMT_INCOME_TOTAL
25650-75650	4035
75650-125650	13405
125650-175650	13476
175650-225650	11351
225650-275650	3265
275650-325650	1861
325650-375650	1103
375650-425650	489
425650-475650	513
475650-525650	56
525650-575650	126
575650-625650	43
625650-675650	141
675650-725650	24
725650-775650	12
775650-825650	22
825650-875650	4
875650-925650	31
925650-975650	2
1025650-1075650	1
1075650-1125650	16
1275650-1325650	1
1325650-1375650	10
1525650-1575650	1
1775650-1825650	2
1875650-1925650	1
1925650-1975650	1
1975650-2025650	2
2225650-2275650	2
3575650-3625650	1
3775650-3825650	1
>4025650	1
Grand Total	49999



DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

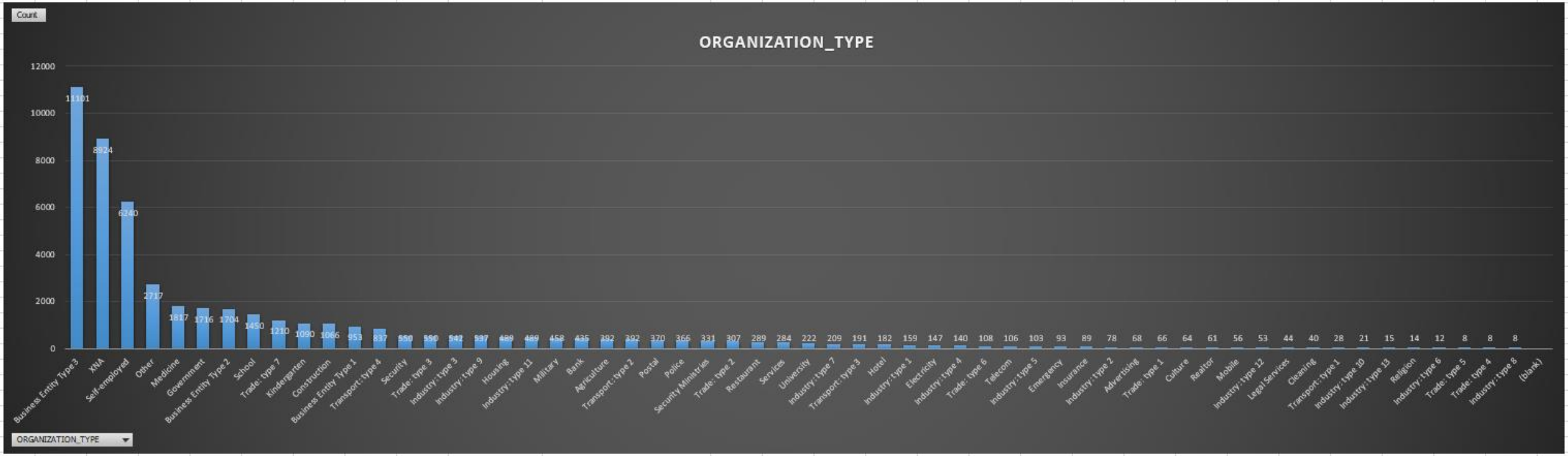
Univariate Results:

Row Labels	Count
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
Security	550
Trade: type 3	550
Industry: type 3	542
Industry: type 9	537
Housing	489
Industry: type 11	489
Military	458
Bank	435
Agriculture	392
Transport: type 2	392
Postal	370
Police	366
Security Ministries	331
Trade: type 2	307
Restaurant	289
Services	284
University	222
Industry: type 7	209
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
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
Trade: type 5	8
Trade: type 4	8
Industry: type 8	8
(blank)	
Grand Total	**49999**

DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Univariate Results:



DATA ANALYSIS

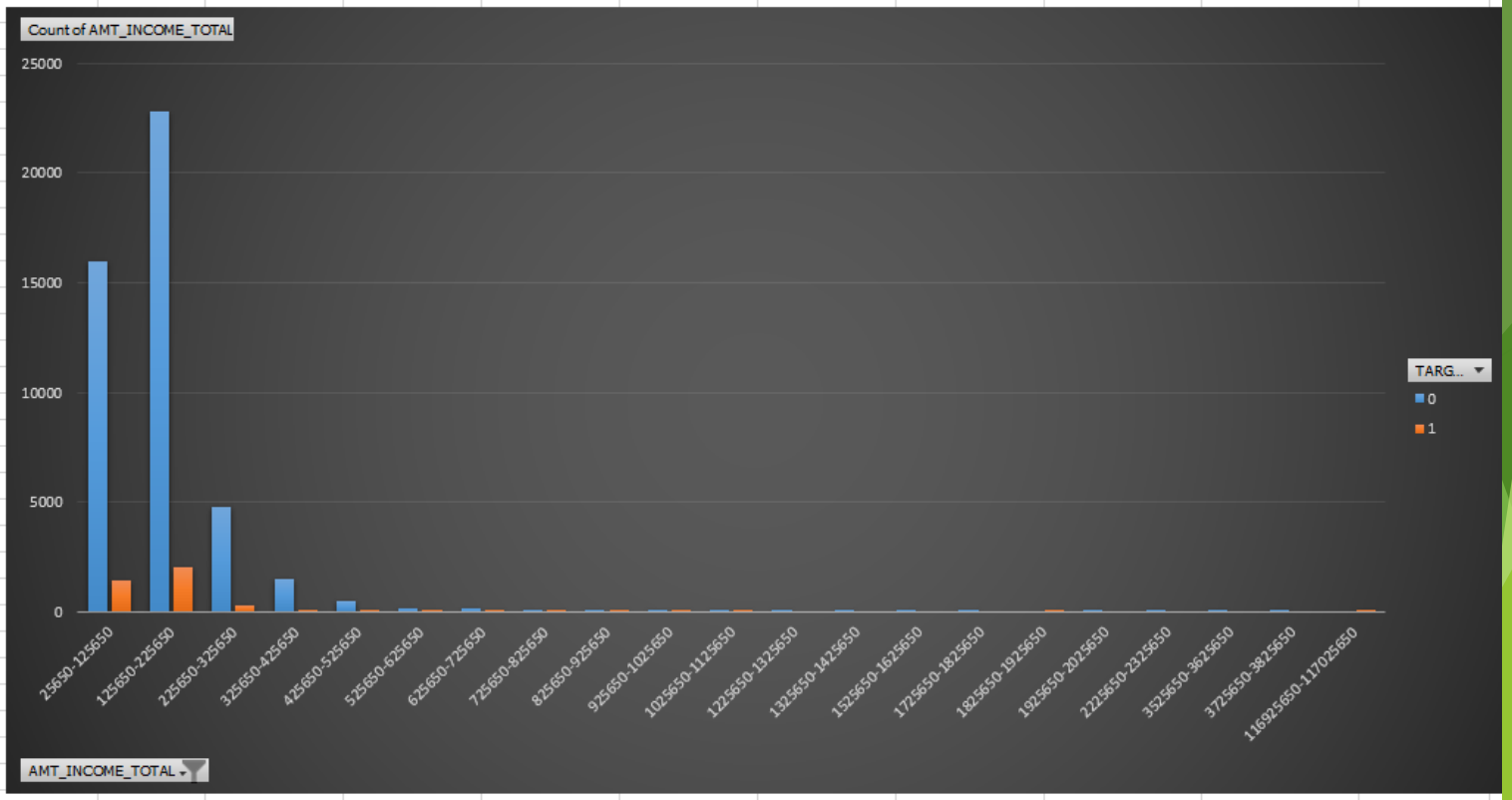
4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

4) Performing Segmented Univariate, and Bivariate Analysis:

A) AMT_INCOME_TOTAL VS TARGET

Count of AMT_INCOME_TOTAL	Column Labels		
Row Labels	0	1	Grand Total
25650-125650	15975	1465	17440
125650-225650	22768	2059	24827
225650-325650	4794	332	5126
325650-425650	1494	98	1592
425650-525650	528	41	569
525650-625650	155	14	169
625650-725650	155	10	165
725650-825650	32	2	34
825650-925650	34	1	35
925650-1025650	1	1	2
1025650-1125650	16	1	17
1225650-1325650	1		1
1325650-1425650	10		10
1525650-1625650	1		1
1725650-1825650	2		2
1825650-1925650		1	1
1925650-2025650	3		3
2225650-2325650	2		2
3525650-3625650	1		1
3725650-3825650	1		1
116925650-117025650		1	1
Grand Total	45973	4026	49999



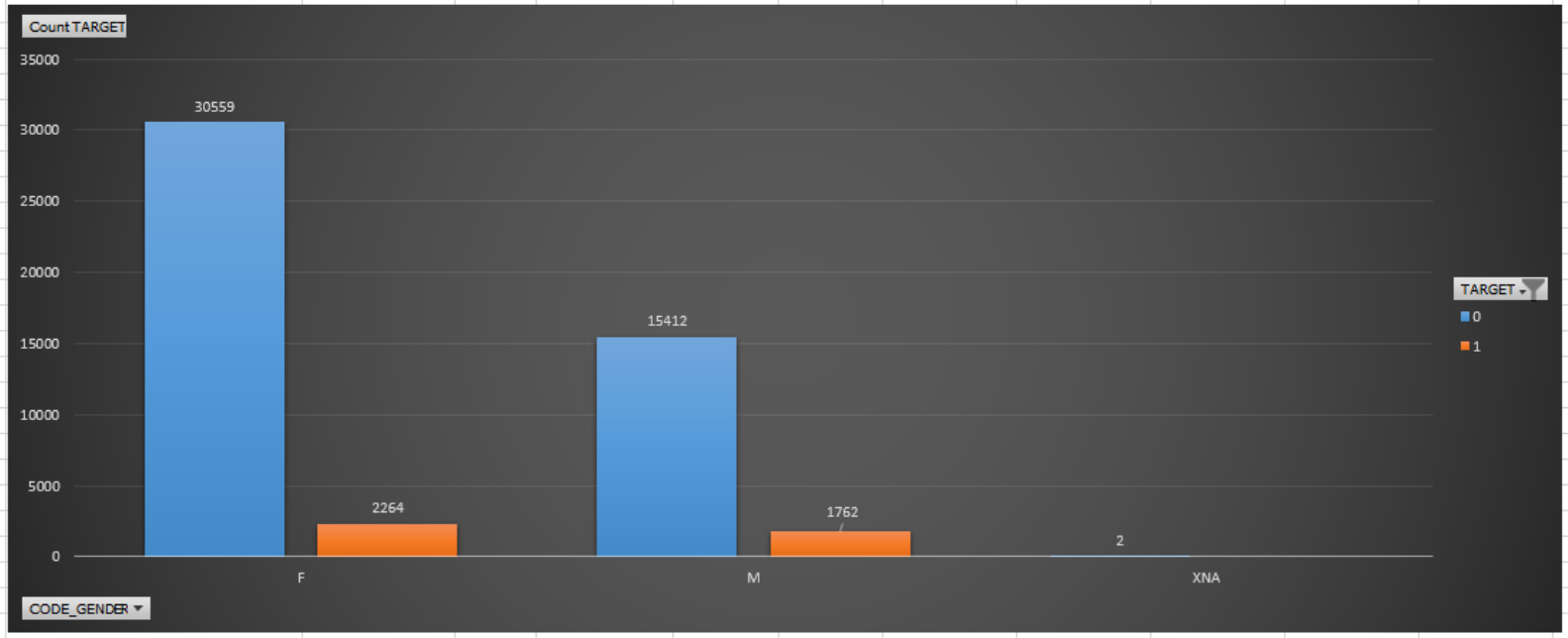
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

B) GENDER VS TARGET

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



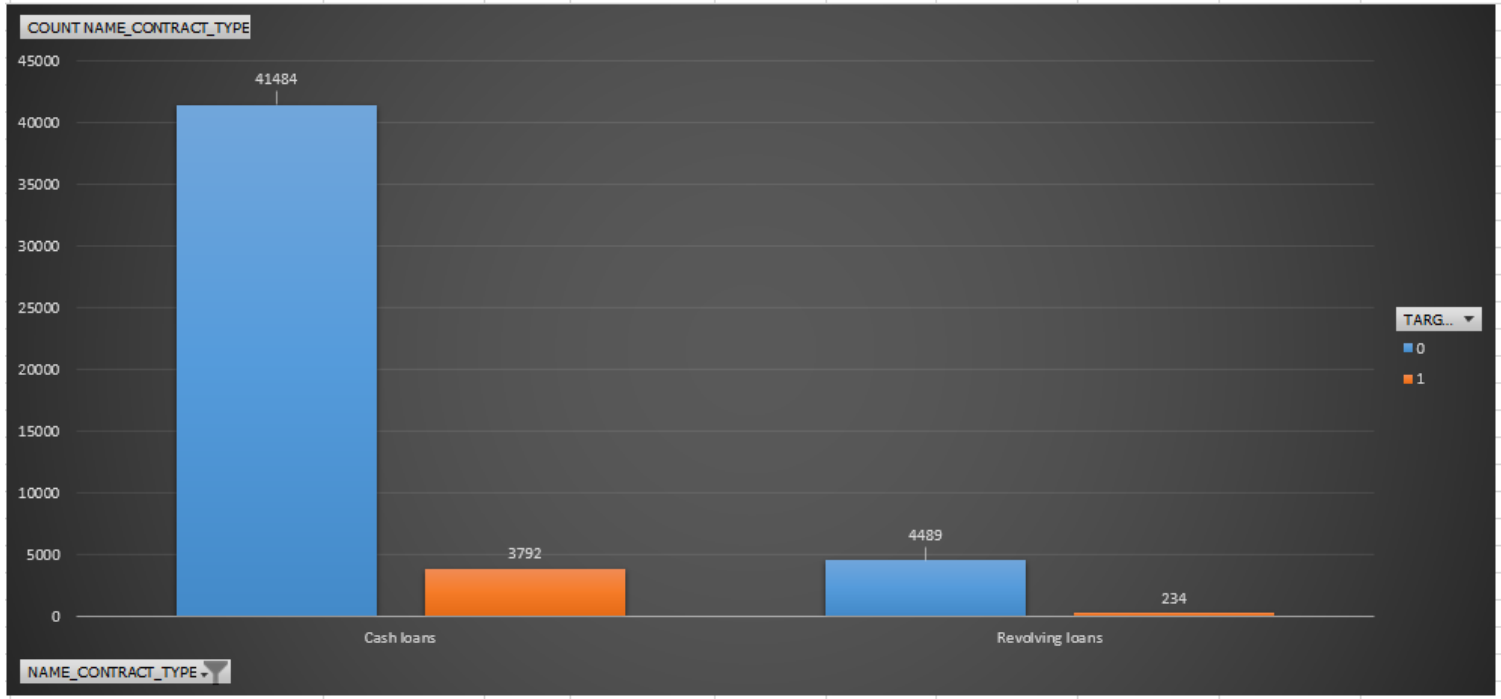
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

C) NAME_CONTRACT_TYPE VS TARGET

COUNT NAME_CONTRACT_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	**45973**	**4026**	**49999**



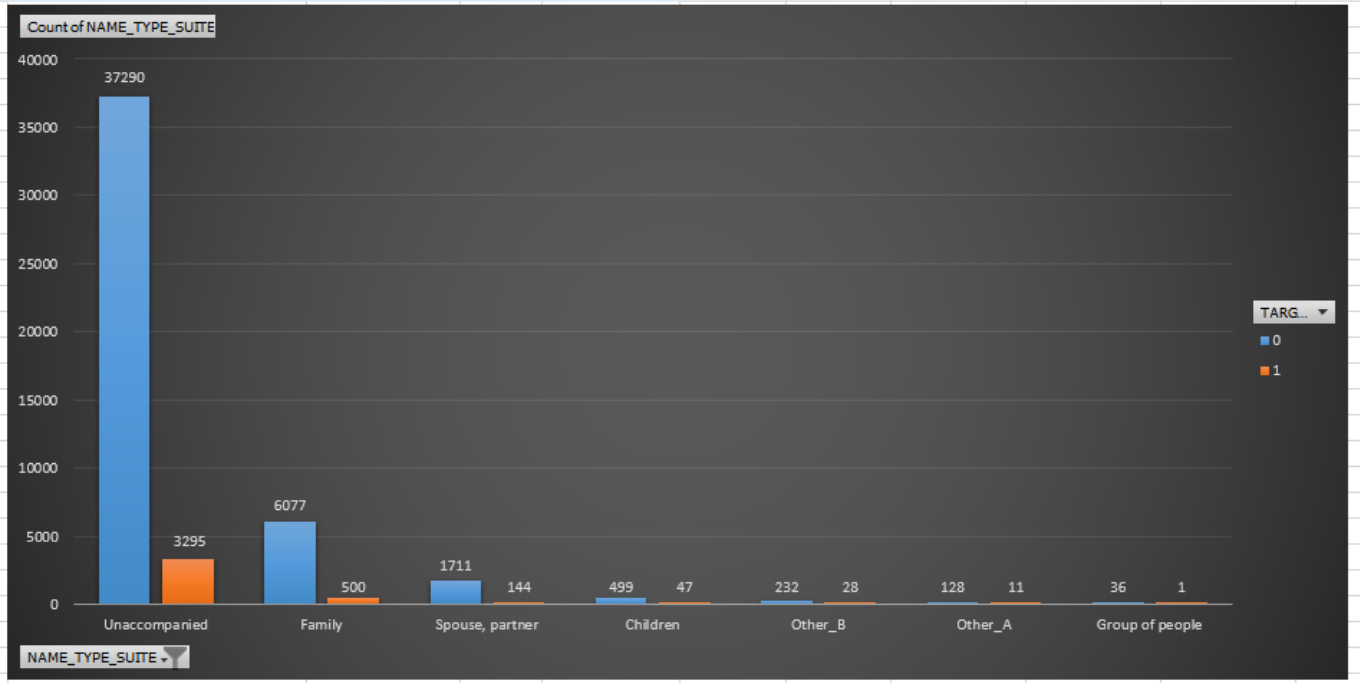
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

D) NAME_TYPE_SUITE VS TARGET

Count of NAME_TYPE_SUITE	Column Labels		
Row Labels	0	1	Grand Total
Unaccompanied	37290	3295	40585
Family	6077	500	6577
Spouse, partner	1711	144	1855
Children	499	47	546
Other_B	232	28	260
Other_A	128	11	139
Group of people	36	1	37
Grand Total	**45973**	**4026**	**49999**



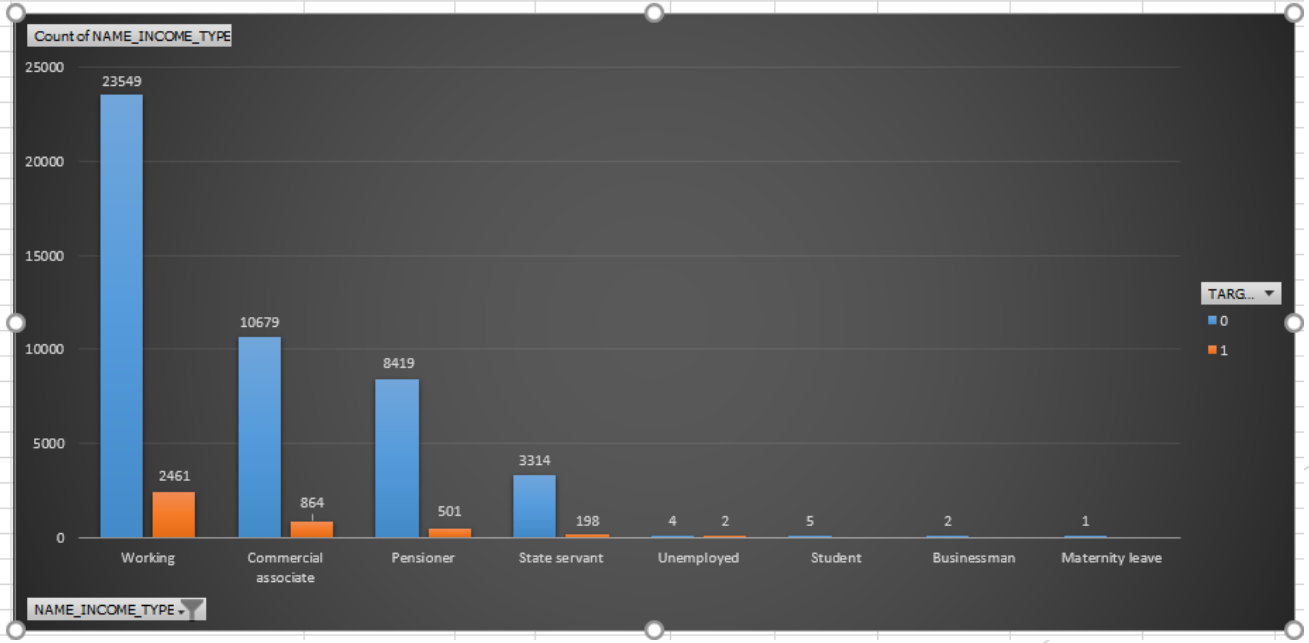
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

E) NAME_INCOME_TYPE VS TARGET

Count of NAME_INCOME_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Working	23549	2461	26010
Commercial associate	10679	864	11543
Pensioner	8419	501	8920
State servant	3314	198	3512
Unemployed	4	2	6
Student	5		5
Businessman	2		2
Maternity leave	1		1
Grand Total	**45973**	**4026**	**49999**



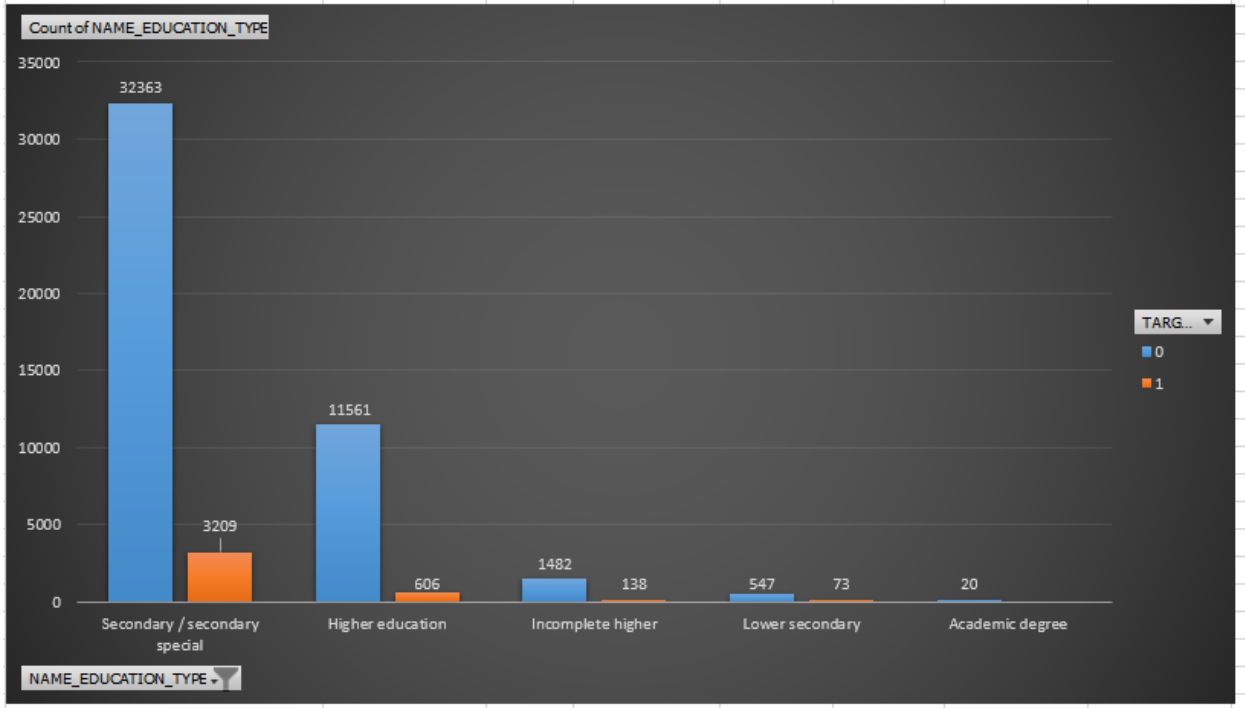
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

F) NAME_EDUCATION_TYPE VS TARGET

Count of NAME_EDUCATION_TYPE	Column Labels		
Row Labels	0	1	Grand Total
Secondary / secondary special	32363	3209	35572
Higher education	11561	606	12167
Incomplete higher	1482	138	1620
Lower secondary	547	73	620
Academic degree	20		20
Grand Total	**45973**	**4026**	**49999**



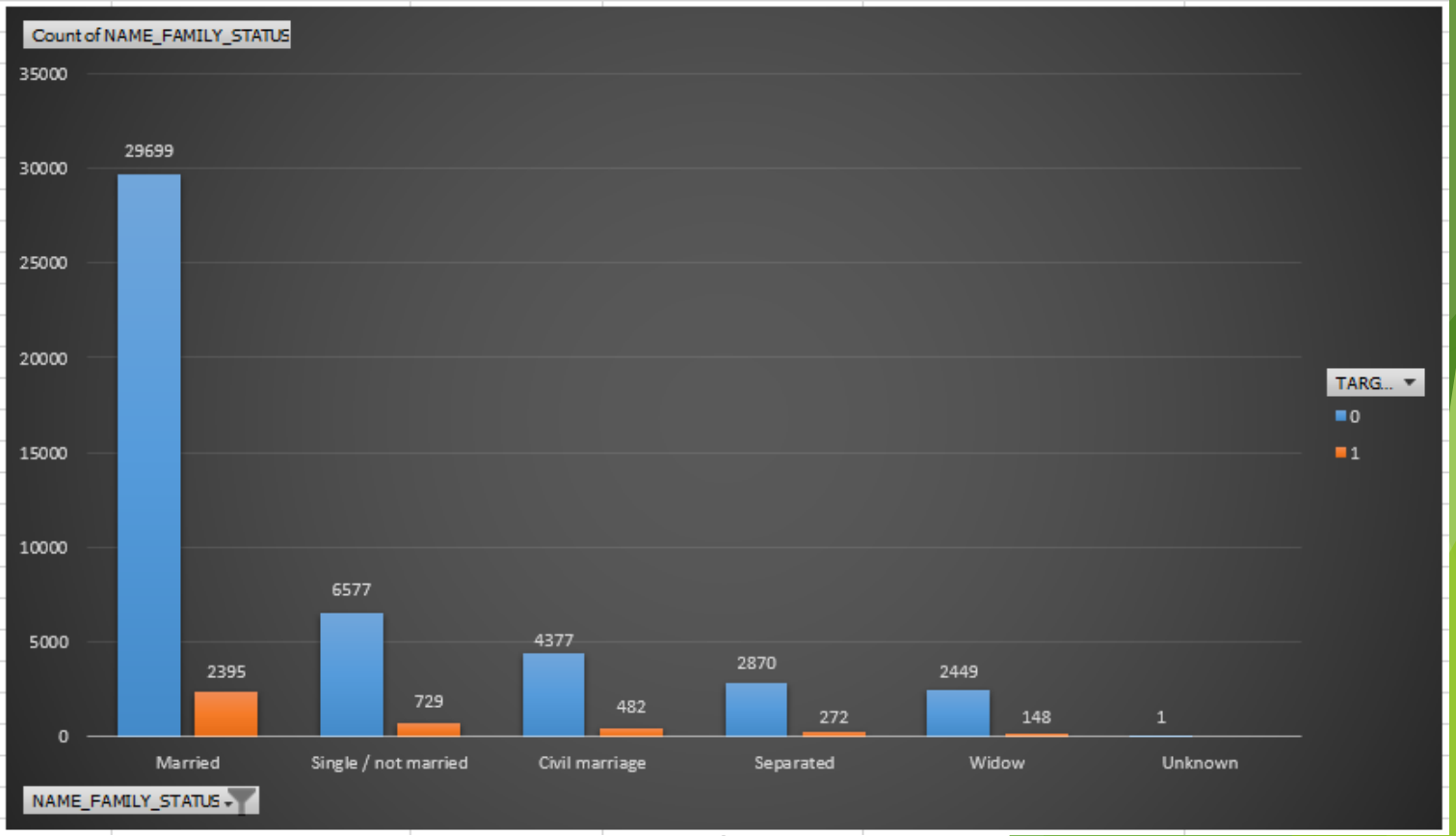
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

I) NAME_FAMILY_STATUS VS TARGET

Count of NAMI	Column Labels		
Row Labels	0	1	Grand Total
Married	29699	2395	32094
Single / not mai	6577	729	7306
Civil marriage	4377	482	4859
Separated	2870	272	3142
Widow	2449	148	2597
Unknown	1		1
Grand Total	**45973**	**4026**	**49999**



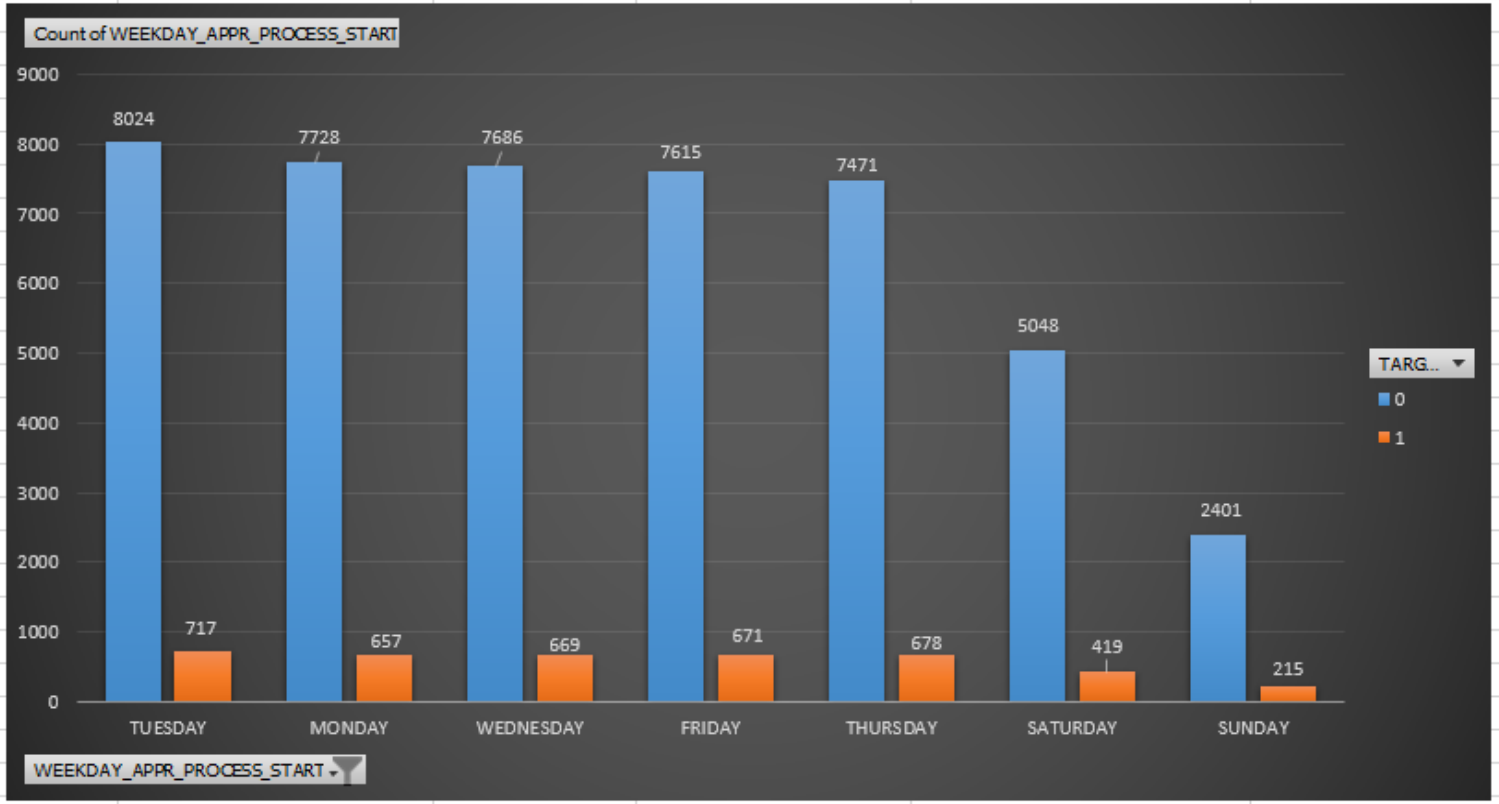
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

J) WEEKDAY_APPR_PROCESS_START VS TARGET

Count of WEEK	Column Labels		
Row Labels	0	1	Grand Total
TUESDAY	8024	717	8741
MONDAY	7728	657	8385
WEDNESDAY	7686	669	8355
FRIDAY	7615	671	8286
THURSDAY	7471	678	8149
SATURDAY	5048	419	5467
SUNDAY	2401	215	2616
Grand Total	**45973**	**4026**	**49999**



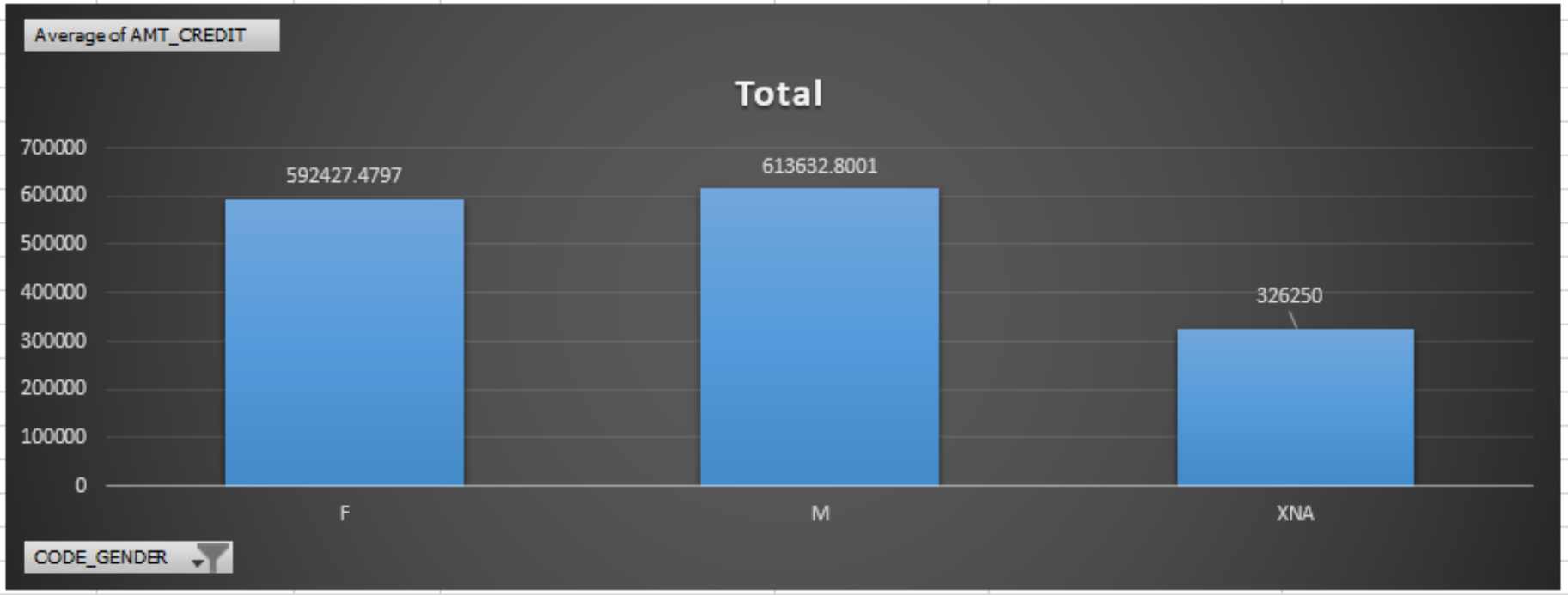
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

L) AMT_CREDIT VS GENDER

Row Labels	Average of AMT_CREDIT
F	592427.4797
M	613632.8001
XNA	326250
Grand Total	**599700.5815**



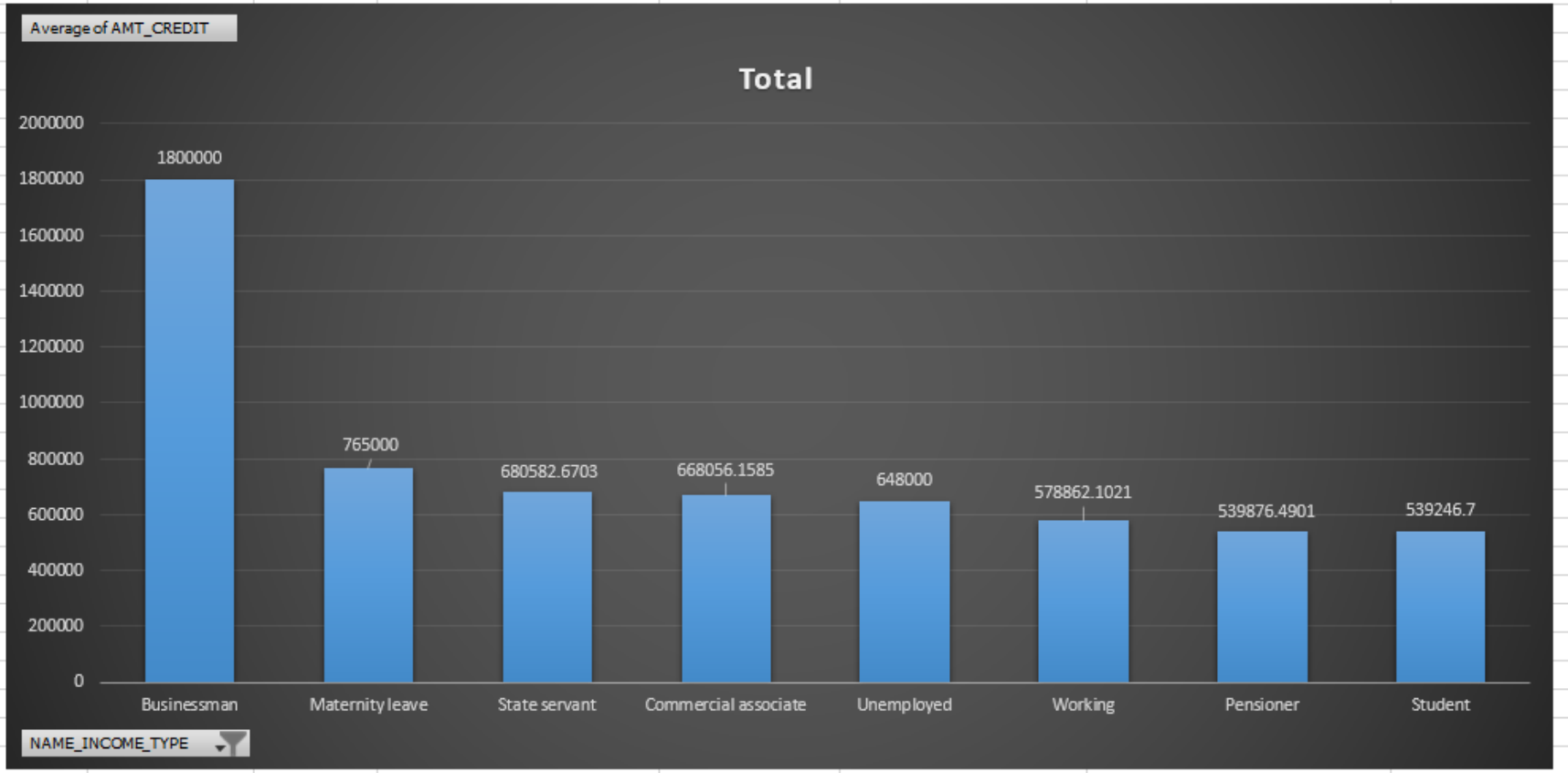
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

M) AGE VS AMT_CREDIT

Row Labels	Average of AMT_CREDIT
Businessman	1800000
Maternity leave	765000
State servant	680582.6703
Commercial ass	668056.1585
Unemployed	648000
Working	578862.1021
Pensioner	539876.4901
Student	539246.7
Grand Total	**599700.5815**



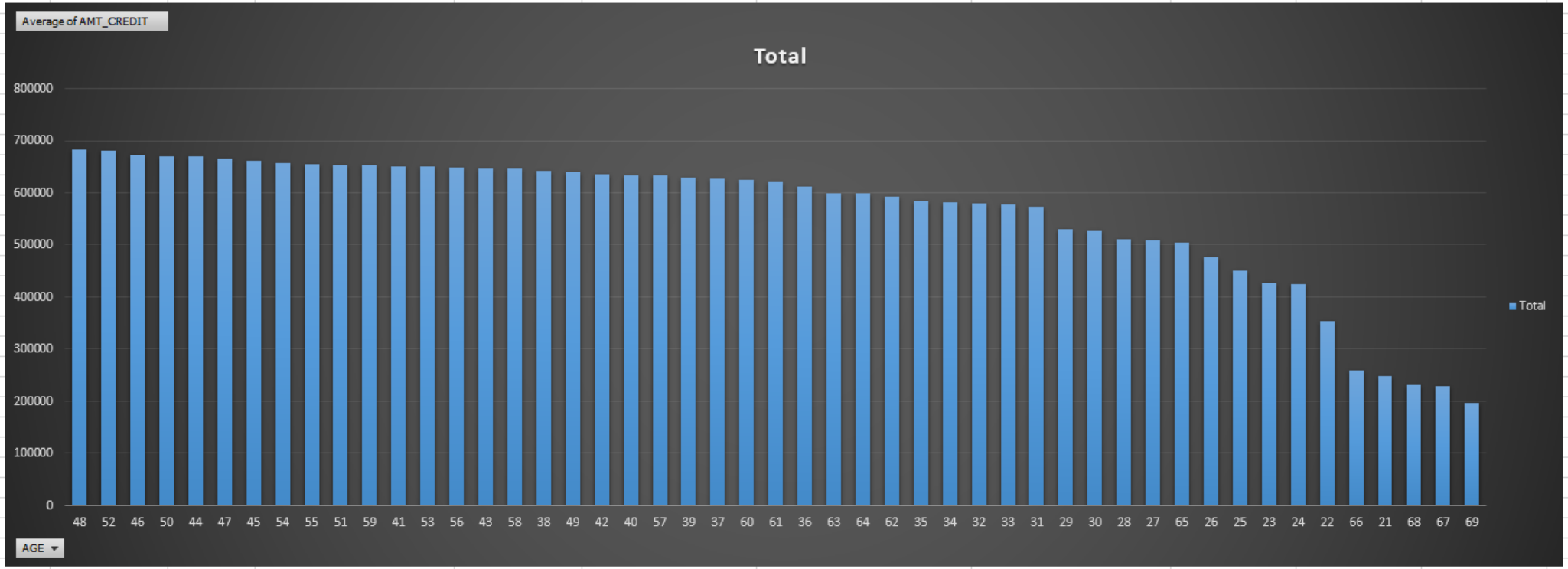
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

O) AGE VS AM_CREDIT

Row Labels	Average of AMT_CREDIT
48	683783.4843
52	679852.404
46	672465.6292
50	670858.1687
44	669239.3087
47	664579.5457
45	660589.3782
54	657256.9897
55	654599.0412
51	652070.4602
59	652052.7596
41	651424.4583
53	650445.9521
56	648321.7989
43	645926.9071
58	645249.9771
38	641919.7256
49	639386.1923
42	635866.6984
40	634010.9005
57	633476.0881
39	629235.6071
37	626716.0392
60	624555.5133
61	620272.9897
36	611890.7946
63	599386.8211
64	599173.6934



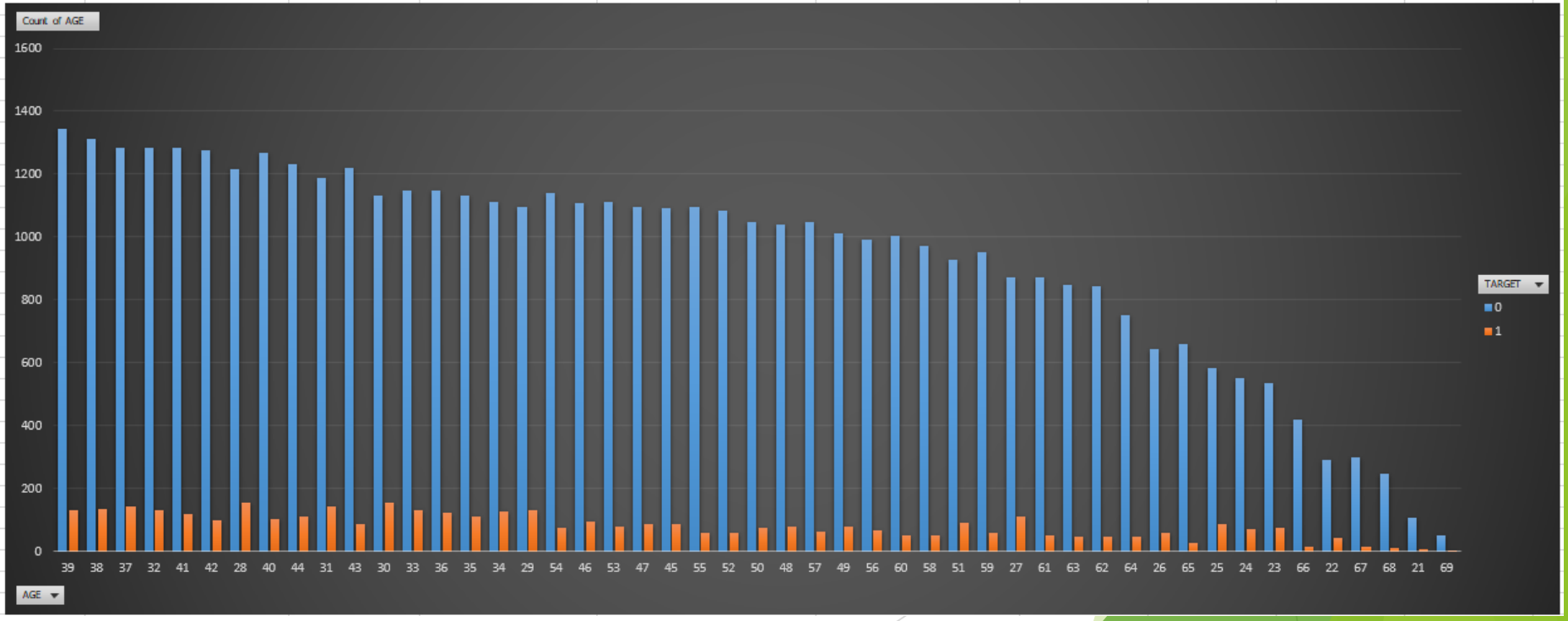
DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

N) AGE VS TARGET

Count of AGE Row Labels	Column Labels 0	1	Grand Total
39	1345	130	1475
38	1311	136	1447
37	1284	145	1429
32	1282	132	1414
41	1282	120	1402
42	1277	99	1376
28	1217	157	1374
40	1269	103	1372
44	1231	110	1341
31	1188	144	1332
43	1219	89	1308
30	1132	154	1286
33	1149	133	1282
36	1148	125	1273
35	1133	110	1243
34	1113	126	1239
29	1097	132	1229
54	1138	75	1213
46	1107	97	1204
53	1112	79	1191
47	1096	86	1182
45	1093	89	1182
55	1095	58	1153
52	1082	59	1141
50	1047	76	1123
48	1038	79	1117
57	1047	65	1112



DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:

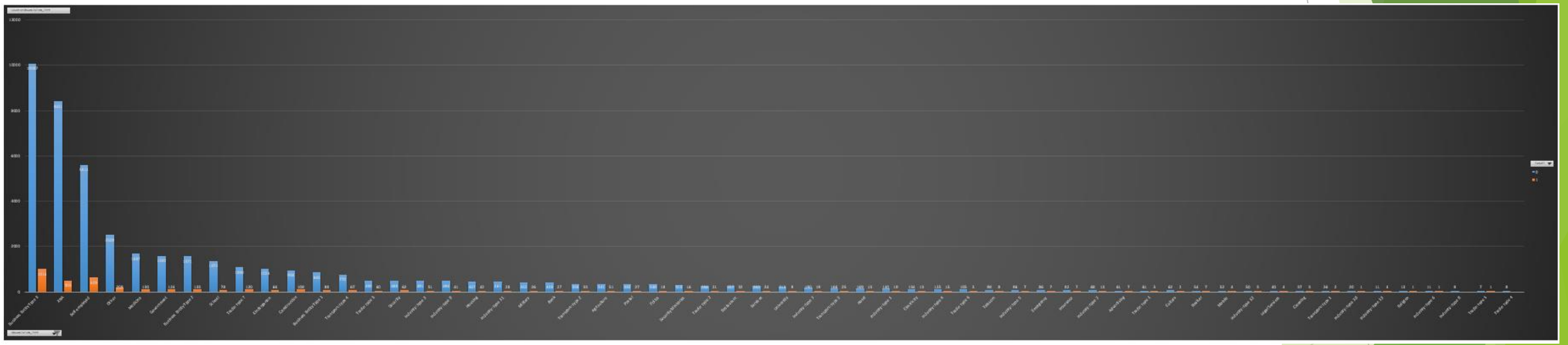
K) ORGANIZATION_TYPE VS TARGET

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

DATA ANALYSIS

4) Perform Univariate, Segmented Univariate and Bivariate Analysis:

Segment Univariate & Bivariate Results:



DATA ANALYSIS

5) Identify Top Correlations for different scenarios:

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Functions that I have used:

- First, I found correlation between target and various columns by using following function:
- =CORREL(D2:D50000,C2:C50000)

https://docs.google.com/spreadsheets/d/1zyX-hMR5zvB5rdqSWwr8bC69_EDzqUQW/edit?usp=sharing&ouid=113249253121491889461&rtpof=true&sd=true

DATA ANALYSIS

5) Identify Top Correlations for different scenarios:

Results:

Column1	CNT_CHILDREN	AMT_INCOME_TOT	AMT_CRED	AMT_ANNUITY	AMT_GOODS_PR	DAYS_BIRT	DAYS_EMPL	DAYS_REGISTR	DAYS_ID_PUB	CNT_FAM_MEMBE	REGION_RATING_CLIE	REGION_RATING_CL	DAYS_LAST_PHONE_CHA
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.009588558	1											
AMT_CREDIT	0.00497156	0.069315897	1										
AMT_ANNUITY	0.026180456	0.083008438	0.769498787	1									
AMT_GOODS_PRICE	0.000232954	0.069831714	0.986704386	0.774134042	1								
DAYS_BIRTH	0.329263754	0.016002774	-0.059342658	0.007708471	-0.057671487	1							
DAYS_EMPLOYED	-0.239693041	-0.031615555	-0.070471393	-0.110447382	-0.06779213	-0.613553972	1						
DAYS_REGISTRATION	0.181217183	0.009952379	0.003448569	0.033218903	0.006083568	0.333632509	-0.204680611	1					
DAYS_ID_PUBLISH	-0.032115773	0.003506646	-0.012228765	0.006716927	-0.014028517	0.270825141	-0.270382022	0.104298561	1				
CNT_FAM_MEMBERS	0.880453292	0.011225511	0.063997155	0.07737959	0.061572677	0.277241347	-0.229816716	0.170108881	-0.026074278	1			
REGION_RATING_CLIENT	0.025913889	-0.038188511	-0.100507425	-0.12580231	-0.103635237	0.016779196	0.034321673	0.087517643	-0.002307011	0.025985394	1		
REGION_RATING_CLIENT_W_CI	0.022777663	-0.040719164	-0.109486833	-0.139321549	-0.111707308	0.014551531	0.036829676	0.079791927	-0.007312572	0.025165113	0.950710179	1	
DAYS_LAST_PHONE_CHANGE	-0.002026164	-0.004804921	-0.076182343	-0.067259706	-0.079714657	0.08019577	0.027515683	0.052146356	0.091380071	-0.02270589	0.027326713	0.026788601	1

NOTE THAT THESE ARE GENERAL CORRELATION SCENARIOS !!!

COULMNS	CORRELATION WITH RESPECT TO TARGET
DAYS_BIRTH	0.076787685
REGION_RATING_CLIENT_W_CITY	0.067079294
REGION_RATING_CLIENT	0.066130148
DAYS_LAST_PHONE_CHANGE	0.056135157
DAYS_ID_PUBLISH	0.046926745
DAYS_REGISTRATION	0.042342679
CNT_CHILDREN	0.026363931
CNT_FAM_MEMBERS	0.01299346
AMT_INCOME_TOTAL	0.010893745
AMT_ANNUITY	-0.0123982
AMT_CREDIT	-0.032428347
DAYS_EMPLOYED	-0.040294905
AMT_GOODS_PRICE	-0.04127611

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
AMT_GOODS_PRICE	AMT_CREDIT	0.986704386
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950710179
CNT_FAM_MEMBERS	CNT_CHILDREN	0.880453292
AMT_GOODS_PRICE	AMT_ANNUITY	0.774134042
AMT_ANNUITY	AMT_CREDIT	0.769498787
DAYS_REGISTRATION	DAYS_BIRTH	0.333632509
DAYS_BIRTH	CNT_CHILDREN	0.329263754
CNT_FAM_MEMBERS	DAYS_BIRTH	0.277241347
DAYS_ID_PUBLISH	DAYS_BIRTH	0.270825141
DAYS_REGISTRATION	CNT_CHILDREN	0.181217183

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
REGION_RATING_CLIEN	AMT_CREDIT	-0.109486833
DAYS_EMPLOYED	AMT_ANNUITY	-0.110447328
REGION_RATING_CLIEN	AMT_GOODS_PRICE	-0.111707308
REGION_RATING_CLIEN	AMT_ANNUITY	-0.12580231
REGION_RATING_CLIEN	AMT_ANNUITY	-0.139321549
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.204680611
CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.229816716
DAYS_EMPLOYED	CNT_CHILDREN	-0.239693041
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.270382022
DAYS_EMPLOYED	DAYS_BIRTH	-0.613553972

DATA ANALYSIS

5) Identify Top Correlations for different scenarios:

Results:

5B) Identify Top Correlations for Different Scenarios:

Column1	CNT_CHILDREN	AMT_INCOME_TO	AMT_CREI	AMT_ANNU	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPL	DAYS_REGISTF	DAYS_ID_PUE	CNT_FAM_MEME	REGION_RATING_CLI	REGION_RATING_C	DAYS_LAST_PHONE_CH
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.036319722	1											
AMT_CREDIT	0.005705458	0.377965752	1										
AMT_ANNUITY	0.02638396	0.451135167	0.770772818	1									
AMT_GOODS_PRICE	0.001518097	0.384575912	0.986999774	0.775835204	1								
DAYS_BIRTH	0.335876269	0.073769425	-0.051084182	0.009911417	-0.048773297	1							
DAYS_EMPLOYED	-0.243591518	-0.162702675	-0.077367219	-0.113005288	-0.075106232	-0.615289978	1						
DAYS_REGISTRATION	0.183072478	0.06893375	0.008053758	0.03460901	0.011260199	0.335028046	-0.204370881	1					
DAYS_ID_PUBLISH	-0.032537221	0.032286356	-0.008290189	0.00942697	-0.00938552	0.270073313	-0.27222439	0.103548902	1				
CNT_FAM_MEMBERS	0.879238049	0.041599302	0.064876937	0.077892626	0.062891858	0.284379407	-0.23373337	0.171482728	-0.025054258	1			
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.102556478	-0.129920896	-0.104841672	0.00902485	0.040505636	0.082562812	-0.008097427	0.022204476	1		
REGION_RATING_CLIENT_W_	0.017873365	-0.220044862	-0.111639948	-0.143197363	-0.113122992	0.00708431	0.042898876	0.074745932	-0.012667326	0.021214058	0.950468157	1	
DAYS_LAST_PHONE_CHANG	-0.004822698	-0.049497956	-0.071203379	-0.064450488	-0.074242871	0.072539576	0.032951867	0.047780168	0.085063175	-0.025039741	0.023514586	0.023179397	1

NOTE THAT THESE ARE GENERAL CORRELATION WHEN TARGET IS 0 SCENARIO !!!

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
AMT_GOODS_PRICE	AMT_CREDIT	0.986999774
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950468157
CNT_FAM_MEMBERS	CNT_CHILDREN	0.879238049
AMT_GOODS_PRICE	AMT_ANNUITY	0.775835204
AMT_ANNUITY	AMT_CREDIT	0.770772818
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451135167
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.384575912
AMT_CREDIT	AMT_INCOME_TOTAL	0.377965752
DAYS_BIRTH	CNT_CHILDREN	0.335876269
DAYS_REGISTRATION	DAYS_BIRTH	0.335028046

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
REGION_RATING_CLIENT	AMT_ANNUITY	-0.129920896
REGION_RATING_CLIENT_W_CITY	AMT_ANNUITY	-0.143197363
DAYS_EMPLOYED	AMT_INCOME_TOTAL	-0.162702675
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.204370881
REGION_RATING_CLIENT	AMT_INCOME_TOTAL	-0.205031899
REGION_RATING_CLIENT_W_CITY	AMT_INCOME_TOTAL	-0.220044862
CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.23373337
DAYS_EMPLOYED	CNT_CHILDREN	-0.243591518
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.27222439
DAYS_EMPLOYED	DAYS_BIRTH	-0.615289978

DATA ANALYSIS

5) Identify Top Correlations for different scenarios:

Results:

Column1	CNT_CHILDREN	AMT_INCOME_T	AMT_CRE	AMT_ANNU	AMT_GOODS_PRICE	DAYS_BIRTH	YED	ATION	ISH	ERS	REGION_RATING_CLI	REGION_RATING_C	ANGE
CNT_CHILDREN	1												
AMT_INCOME_TOTAL	0.010110177	1											
AMT_CREDIT	0.007601905	0.015271444	1										
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1									
AMT_GOODS_PRICE	-0.001079665	0.013269502	0.982267963	0.74950403	1								
DAYS_BIRTH	0.2496732	0.009033662	-0.142506035	-0.008751713	-0.141005898	1							
DAYS_EMPLOYED	-0.189324184	-0.011555963	0.016039571	-0.079556008	0.020235348	-0.581479041	1						
DAYS_REGISTRATION	0.152113117	-0.009561152	-0.042844404	0.021581654	-0.043320226	0.288437837	-0.188718437	1					
DAYS_ID_PUBLISH	-0.042360717	-0.009122006	-0.043771901	-0.02132109	-0.049723232	0.247896571	-0.230063668	0.09029149	1				
CNT_FAM_MEMBERS	0.892521875	0.013121678	0.06124869	0.075838463	0.055135807	0.199141397	-0.183560113	0.151786548	-0.044037815	1			
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.061578289	-0.051296281	0.045027112	-0.009145883	0.115625377	0.025335227	0.057279521	1		
_CITY	0.054802235	-0.01266585	-0.052954314	-0.079418668	-0.056693474	0.038087333	-0.004137686	0.108123203	0.014431344	0.057987728	0.950768899	1	
DAYS_LAST_PHONE_CHANG	0.011339334	0.012457111	-0.124539343	-0.100470941	-0.128832447	0.124609491	-0.015732544	0.078604652	0.138087781	-0.005731154	0.026186488	0.022309455	

NOTE THAT THESE ARE GENERAL CORRELATION WHEN TARGET IS 1 SCENARIO !!!

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
AMT_GOODS_PRICE	AMT_CREDIT	0.982267963
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950768899
CNT_FAM_MEMBERS	CNT_CHILDREN	0.892521875
AMT_GOODS_PRICE	AMT_ANNUITY	0.74950403
AMT_ANNUITY	AMT_CREDIT	0.749665201
DAYS_REGISTRATION	DAYS_BIRTH	0.288437837
DAYS_BIRTH	CNT_CHILDREN	0.2496732
DAYS_ID_PUBLISH	DAYS_BIRTH	0.247896571
CNT_FAM_MEMBERS	DAYS_BIRTH	0.199141397
DAYS_REGISTRATION	CNT_CHILDREN	0.152113117

RELATIONSHIP1	RELATIONSHIP2	CORRELATIONS
DAYS_LAST_PHONE_CHANGE	AMT_ANNUITY	-0.100470941
DAYS_LAST_PHONE_CHANGE	AMT_CREDIT	-0.124539343
DAYS_LAST_PHONE_CHANGE	AMT_GOODS_PRICE	-0.128832447
DAYS_BIRTH	AMT_GOODS_PRICE	-0.141005898
DAYS_BIRTH	AMT_CREDIT	-0.142506035
CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.183560113
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188718437
DAYS_EMPLOYED	CNT_CHILDREN	-0.189324184
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.230063668
DAYS_EMPLOYED	DAYS_BIRTH	-0.581479041

Insights

- There are many missing values in the dataset. The columns having null values above 25% are deleted and the missing values are replaced using median and mode.
- There are many outliers in the dataset. We can use appropriate methods to deal with outliers.
- There is a data imbalance in most of the columns.
- People who has low income, Married, Working and has age 38-39 years have taken the loan mostly and also they are most likely to default the loan.
- There are many correlations between the columns and the highest correlated column is DAYS_BIRTH.

Conclusion

- Finally, I have successfully completed this project using Excel, Power point. I have learned to deal with large datasets which has many missing values and outliers.

Thank You