Bank Loan Case Study

Final Project – 2

By: Mohak Bidaye

Project Description

- This project focuses on analyzing customer loan data to uncover key factors contributing to loan defaults. By leveraging Exploratory Data Analysis (EDA), the aim is to identify patterns and insights that can enhance the company's decision-making process for loan approvals. The objective is to minimize the risk of defaults while ensuring qualified applicants are not unfairly rejected, ultimately balancing business growth and financial stability.
- When a customer applies for a loan, the company encounters two significant risks:

1. Missed Opportunity: If a creditworthy applicant is denied, the company loses potential business.
2. Financial Loss: If a non-creditworthy applicant is approved, the company incurs a loss.

- The loan application process leads to one of four outcomes:

a. Approved: The loan application is approved by the company.

b. Cancelled: The customer withdraws their application during the approval process.

c. Refused: The company rejects the loan application.

d. Unused Offer: The loan is approved, but the customer does not utilize it.

- By analyzing the loan data, this project aims to uncover actionable insights, reduce risk, and optimize the approval process to benefit both the company and its customers.

Approach

- Importing and understanding the data set provided.
- Cleaning data and identifying missing values and dealing with them appropriately.
- Performing analysis and showing insights.
- Visualizing the analysis and insights through graphs and charts.
- Concluding the project with actionable insights which can benefit the bank.

Links

- Excel sheet link:

https://docs.google.com/spreadsheets/d/1alUTJ4Fvwyg6QX5IRclD5pf1rqaKtf3L/edit?usp=drive_link&ouid=109524556463170667809&rtpof=true&sd=true

- Google drive link to download the excel sheet if above link does not work.
- The excel sheet is saved under the name 'Project6(BankLoanProject).'

https://drive.google.com/drive/folders/1hVUPMva915K5R8FuzBxPfIppPWzz6L0a?usp=sharing

Tech Stack Used

- Microsoft Excel 2019
- Microsoft PowerPoint 2019

Task A: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

- Utilized COUNTA() to determine the number of records for each column.
- Calculated the percentage of missing values for every column to assess the extent of null values.
- Formulas used:

I. To find number of records: =COUNTA(B4:B50002)

II. To calculate the percentage of missing values: =1-B2/B2

- Columns with 50% or more missing values (highlighted in red) were removed from the dataset.
- Additional columns that were irrelevant to the analysis (highlighted in yellow) were also deleted.

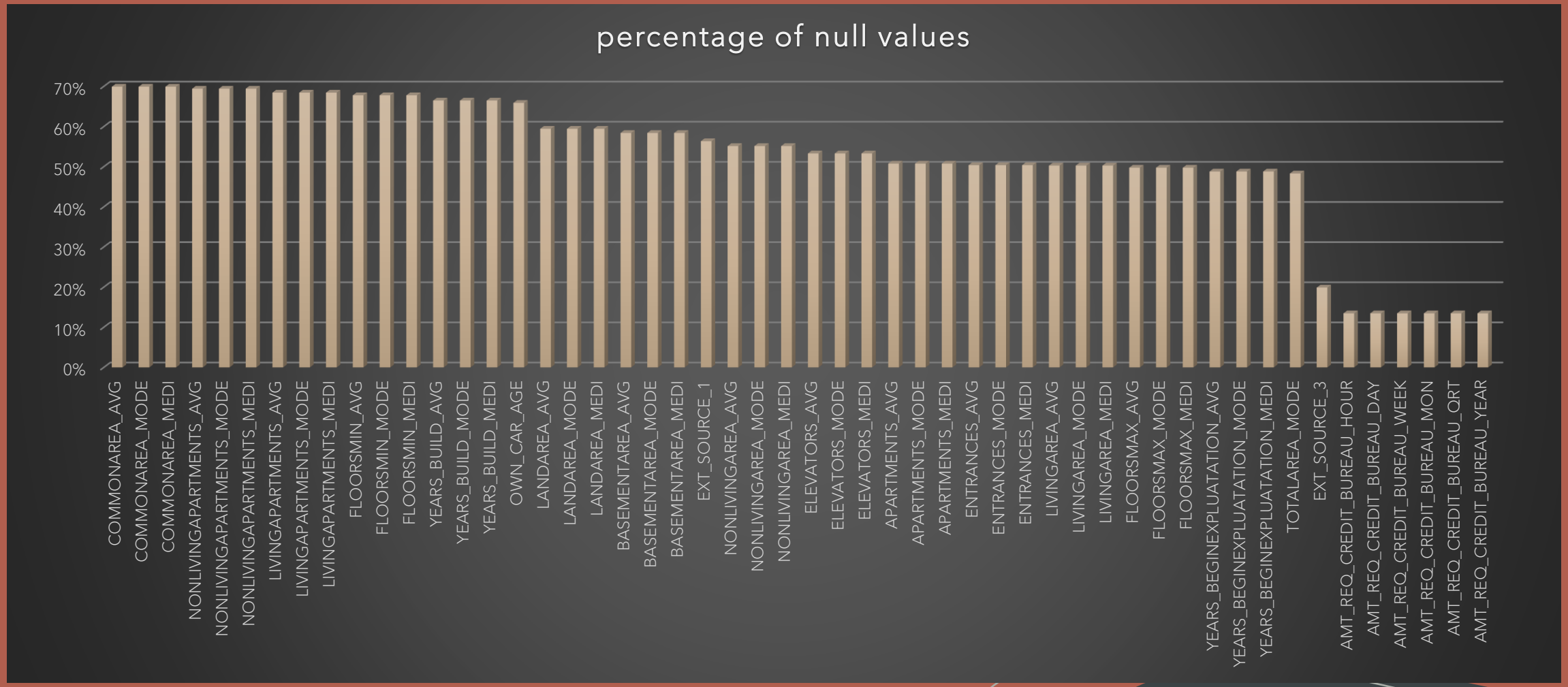
Deleted columns with more than 50% missing values

OWN_CAR_AGE
EXT_SOURCE_1
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE
LIVINGAPARTMENTS_MODE
LIVINGAREA_MODE
NONLIVINGAPARTMENTS_MODE
NONLIVINGAREA_MODE
APARTMENTS_MEDI
BASEMENTAREA_MEDI
YEARS_BUILD_MEDI
COMMONAREA_MEDI
ELEVATORS_MEDI
ENTRANCES_MEDI
FLOORSMAX_MEDI
FLOORSMIN_MEDI
LANDAREA_MEDI
LIVINGAPARTMENTS_MEDI
LIVINGAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
NONLIVINGAREA_MEDI

Deleted columns that were irrelevant to the analysis

FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EXT_SOURCE_2
EXT_SOURCE_3
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EMERGENCYSTATE_MODE
DAYS_LAST_PHONE_CHANGE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_3
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
FLAG_DOCUMENT_7
FLAG_DOCUMENT_8
FLAG_DOCUMENT_9
FLAG_DOCUMENT_10
FLAG_DOCUMENT_11
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21

- Percentage of null values representation:



- Data imputation was done on the missing values.
- Numerical Variables: Missing values were filled using the mean or median, depending on the distribution.
- Categorical Variables: Missing values were imputed using the mode.

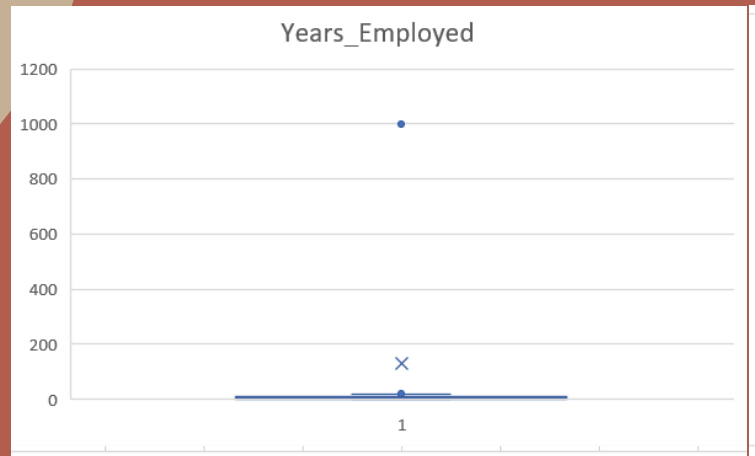
- The columns DAYS_BIRTH and DAYS_EMPLOYED were changed to Age and Year_Employed using ROUND() and ABS()
- Formulas used:

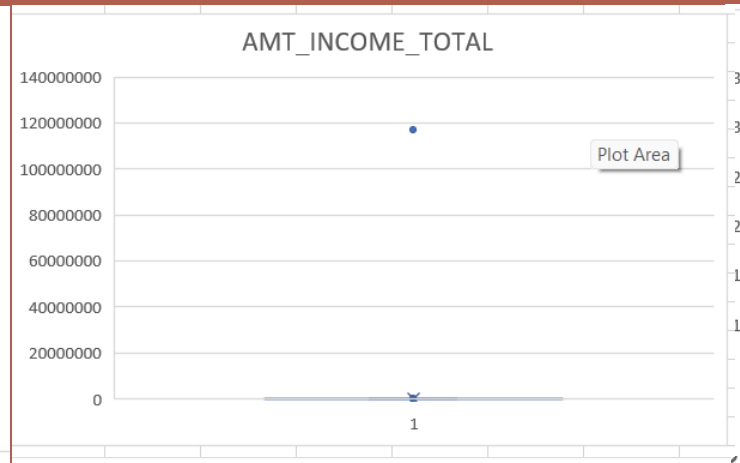
I. =ROUND(ABS(R9/365),0)

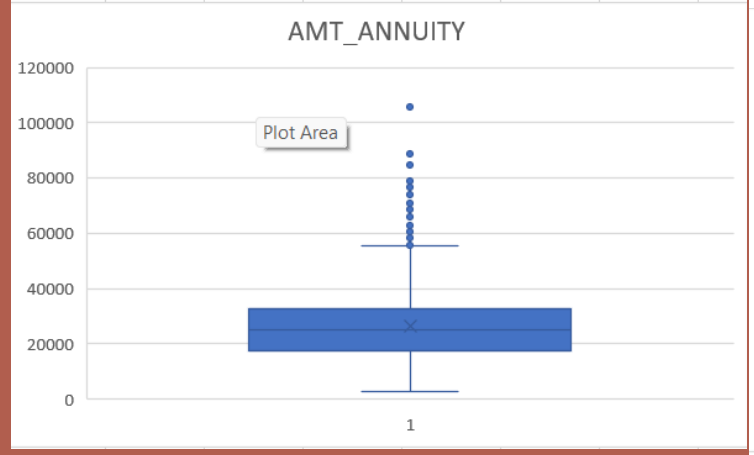
II. =ROUND(ABS(T9/365),0)

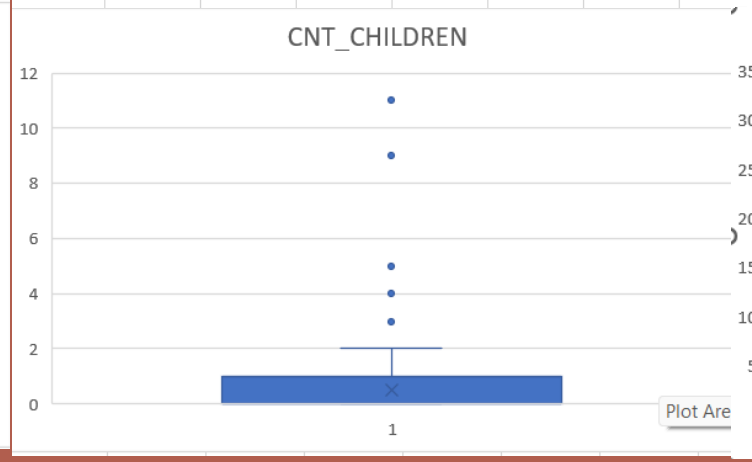
DAYS_BIRTH	Age	DAYS_EMPLOYED	Years_Employed
-18255	50	-386	1
-21914	60	365243	1001
-15566	43	-128	0
-12298	34	-3015	8
-22285	61	365243	1001
-11290	31	-1016	3
-12687	35	-1557	4
-10249	28	-1600	4
-13918	38	-2542	7
-18670	51	-1581	4
-11775	32	-482	1

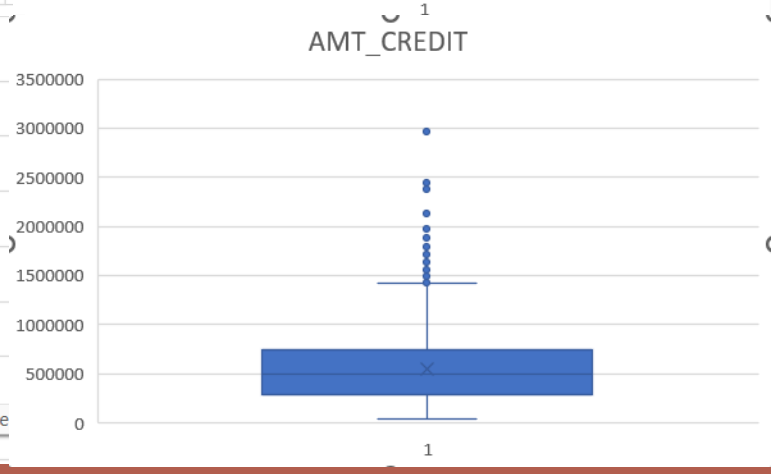
Task B: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables







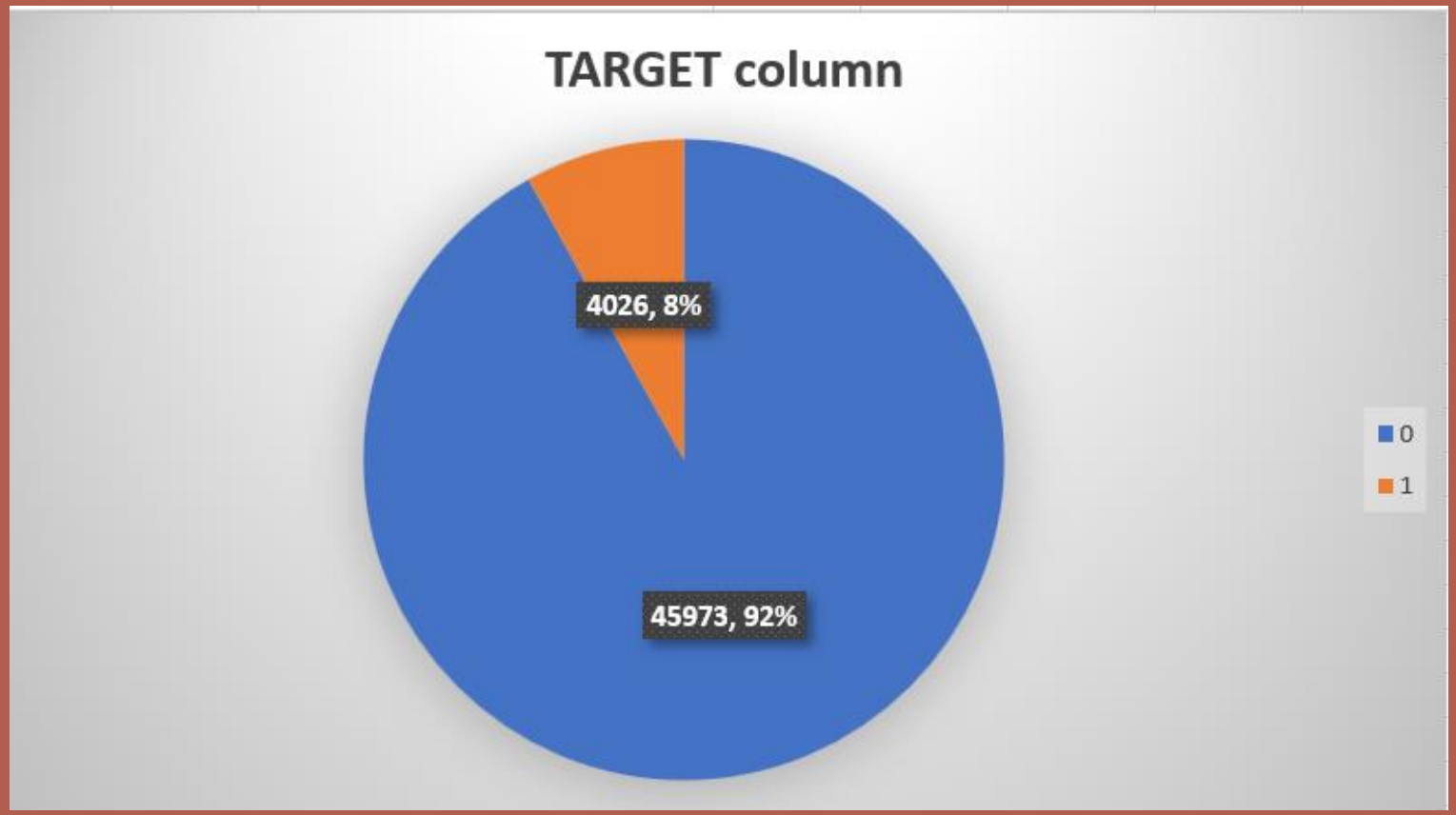




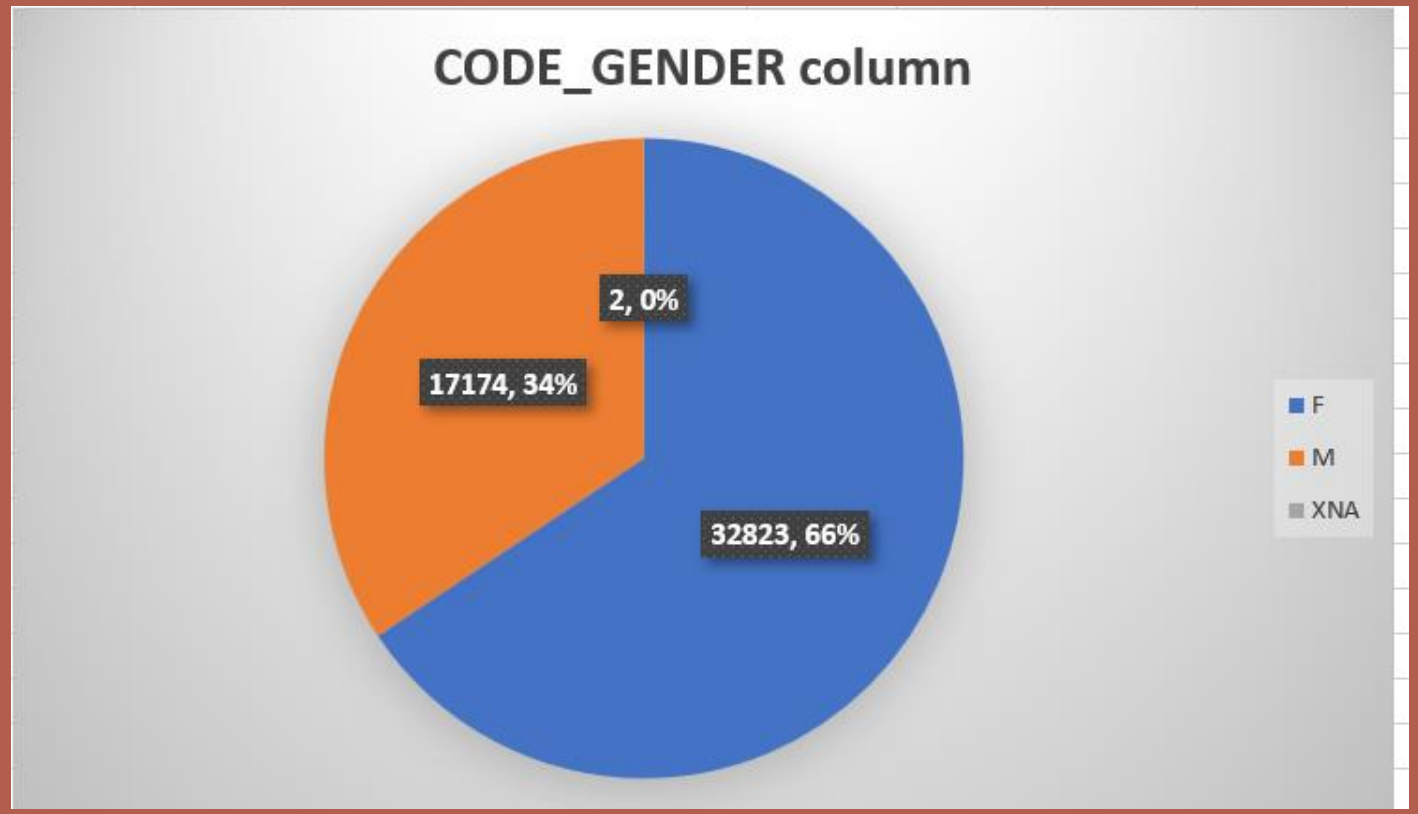
Outlier insights

- AMT_INCOME_TOTAL is a column where one of the outlier is 117000000 which is an extremely high salary to earn.
- Years_Employed has an outlier where it's showing that there are people being employed for 1001 years and that's not possible.
- We also find outliers in CNT_CHILDREN where there is a count of 8+ children.
- We find a large number of outliers in AMT_GOODS_PRICE, AMT_APPLICATION, AMT_CREDIT AND AMT_ANNUITY.

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



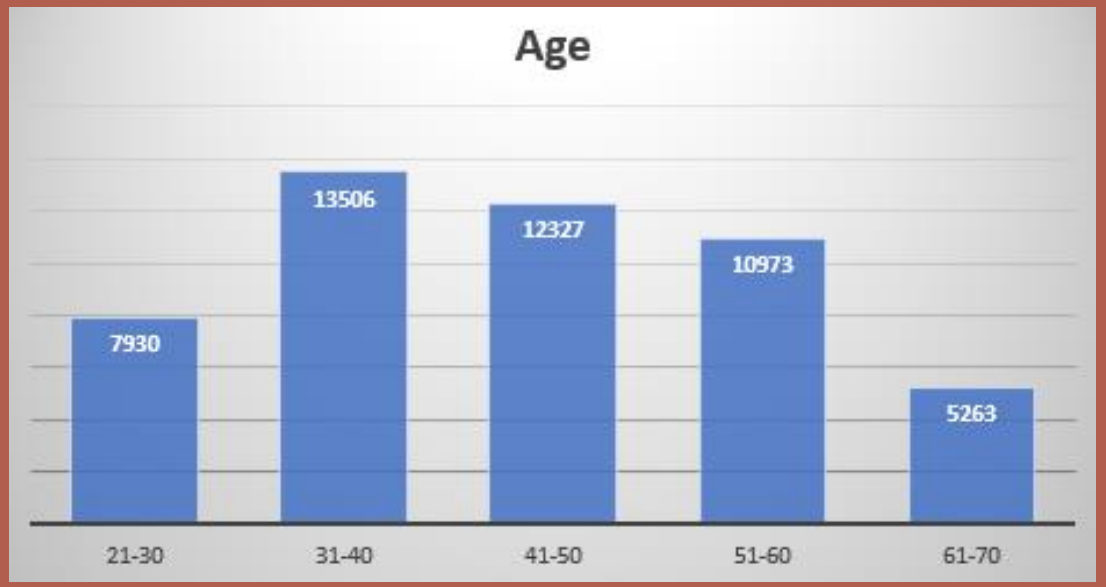
- Here we can see that 92%(blue i.e. 0) clients are loan re-payers and 8%(orange i.e. 1) are defaulters
- This shows that there is an imbalance in the dataset.



Here we see that there is an imbalance in the CODE_GENDER column where 66% of the clients are females and 34% are males.

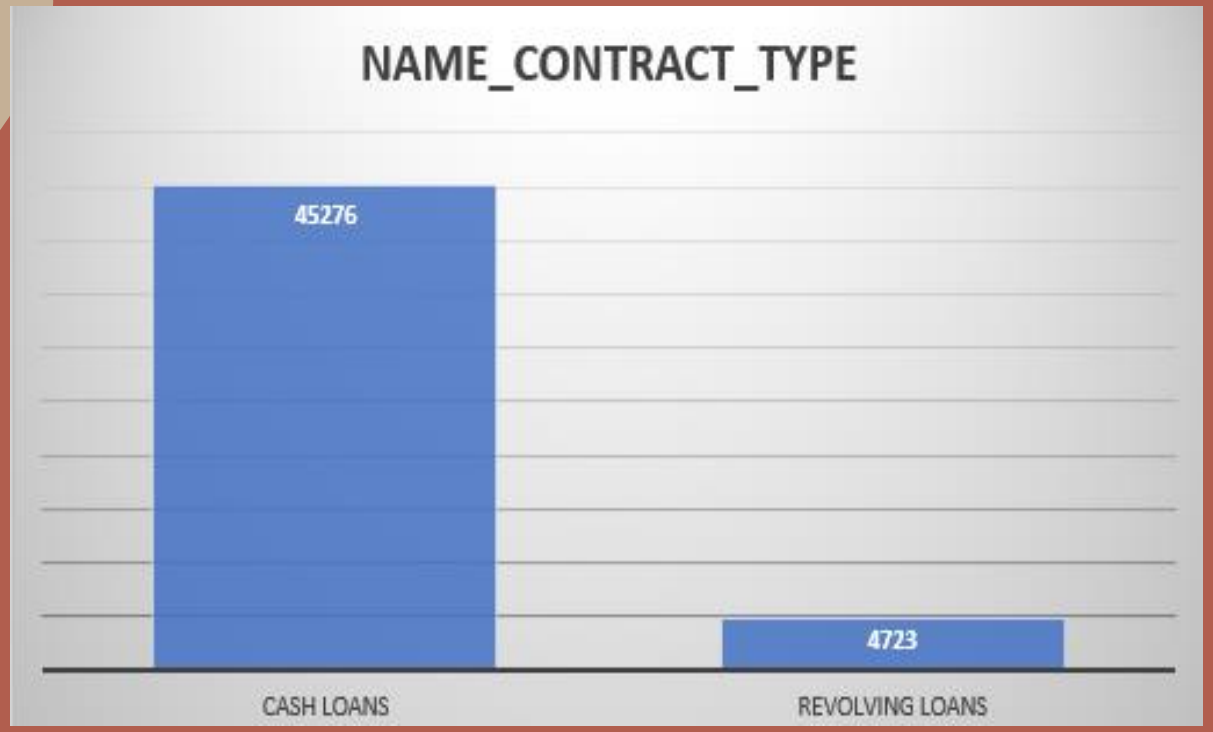
Task D: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

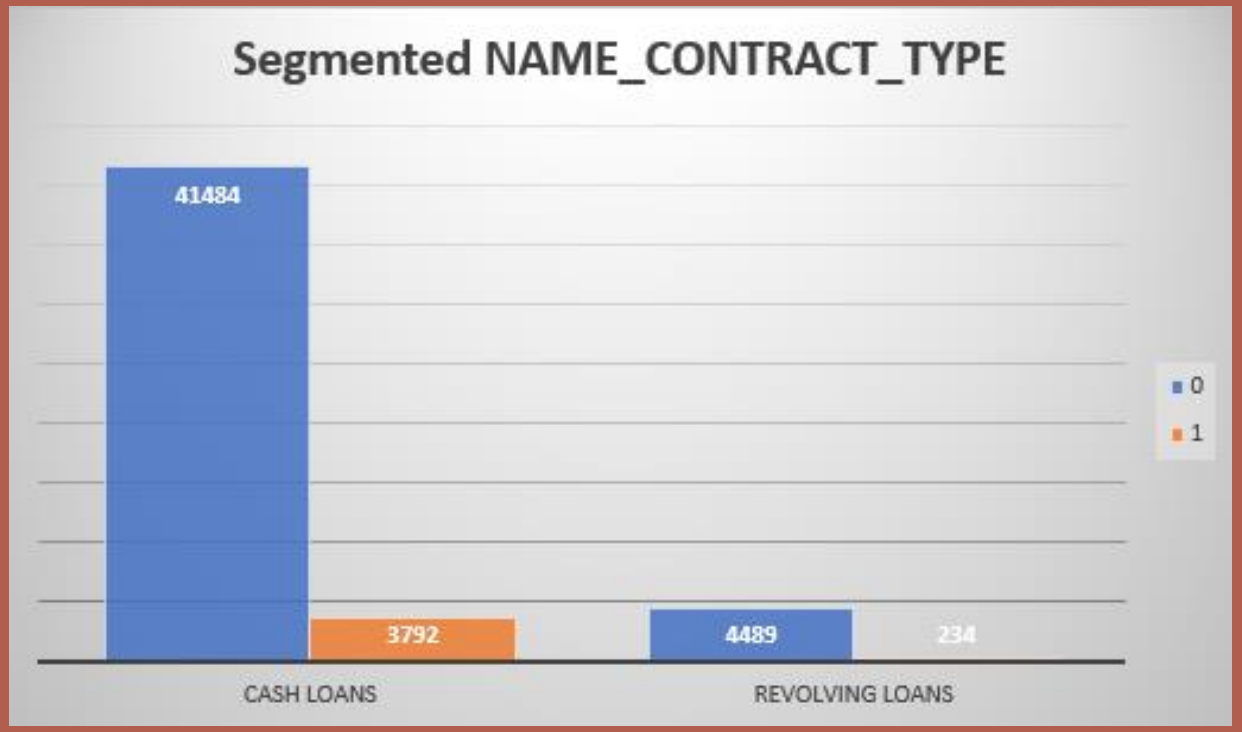
Univariate Analysis



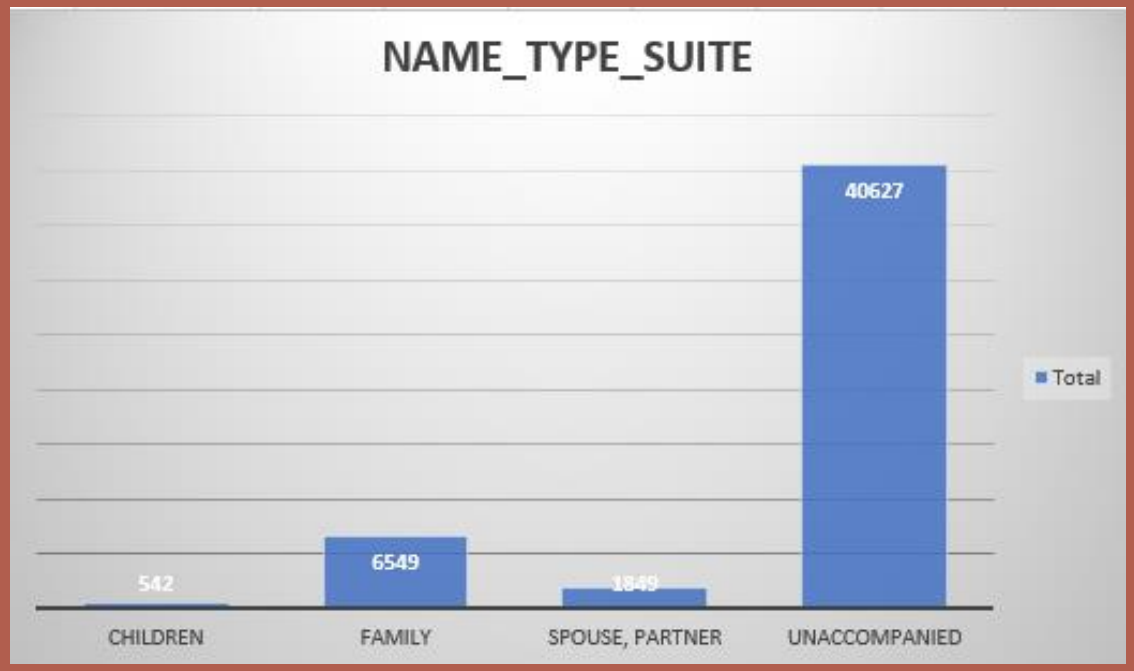
Most of the clients are in the age group of 31-40.

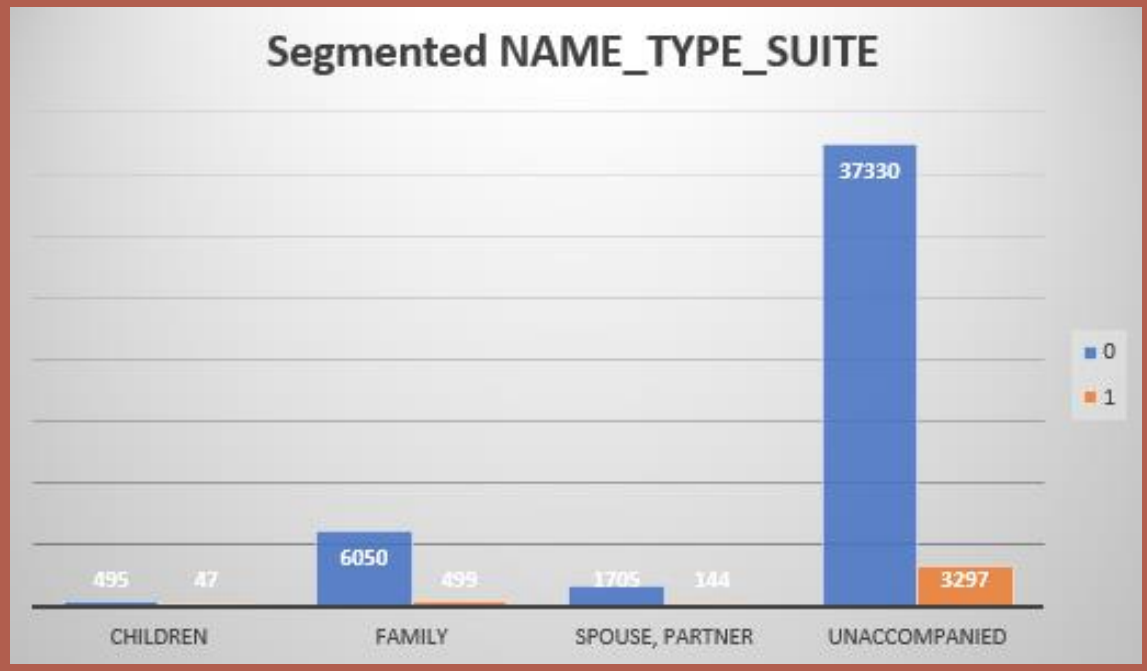
We can see that, as age increases the chances of defaulter decreases



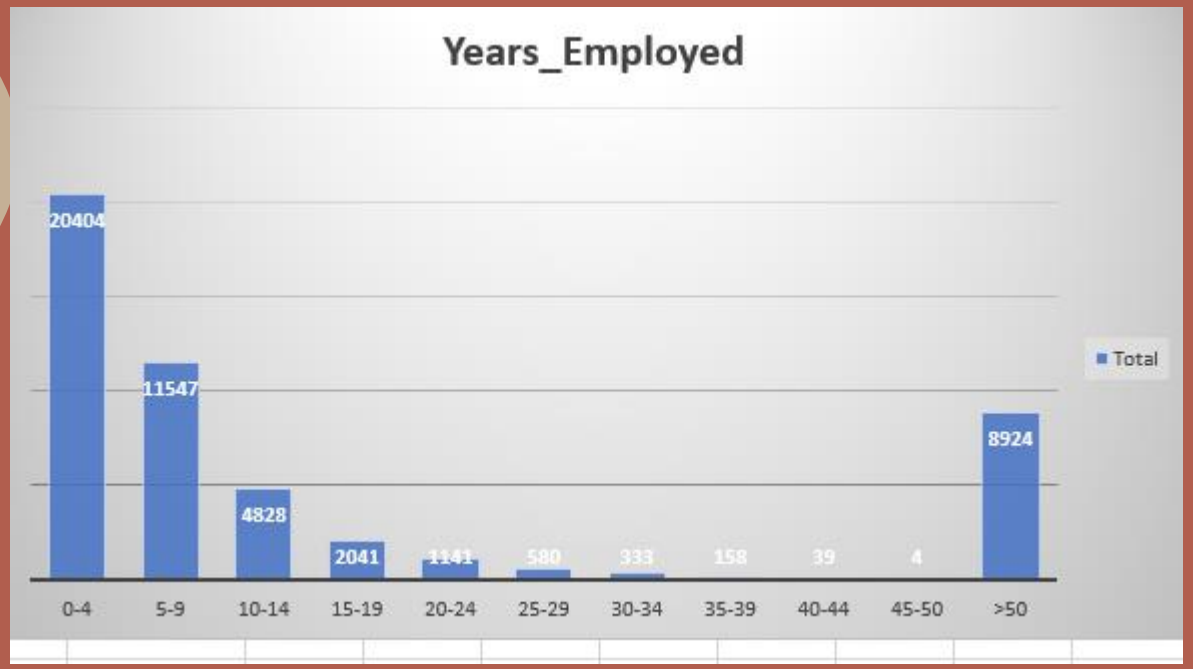


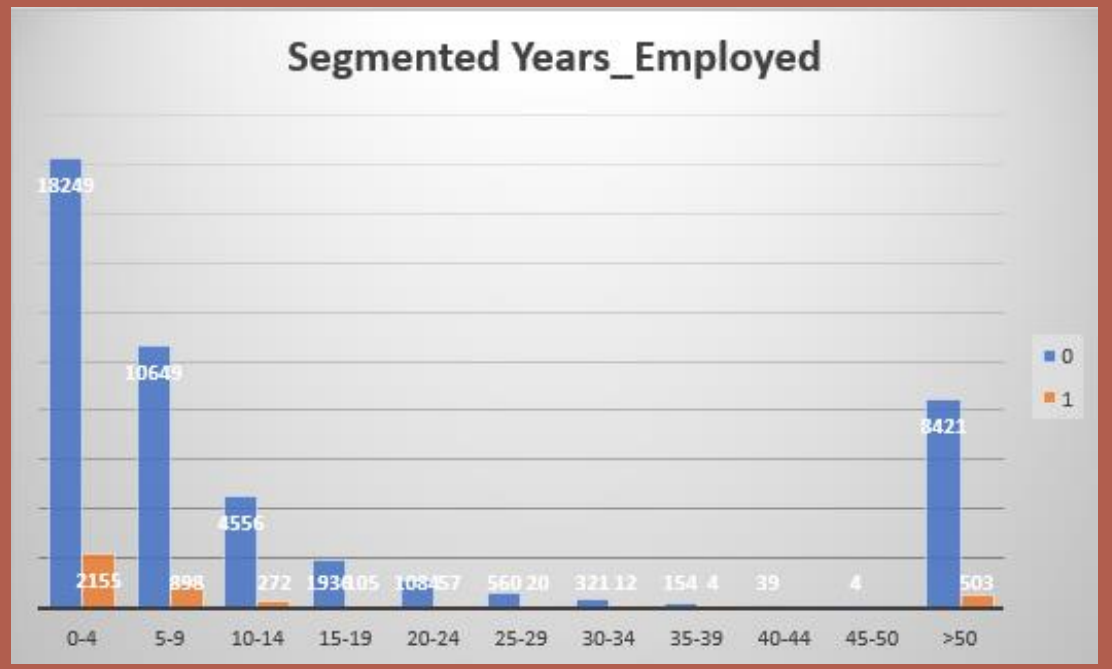
The number of cash loans is larger than the number of revolving loans





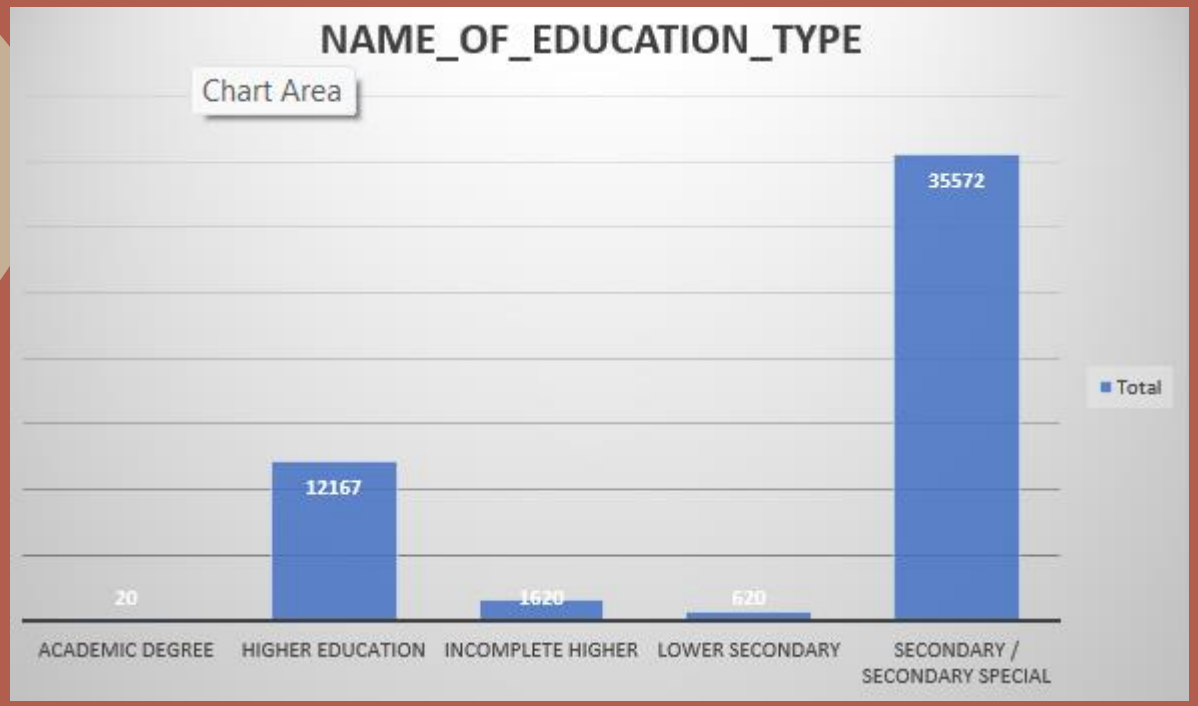
Most of the clients were Unaccompanied while taking loans, this was followed by the Family

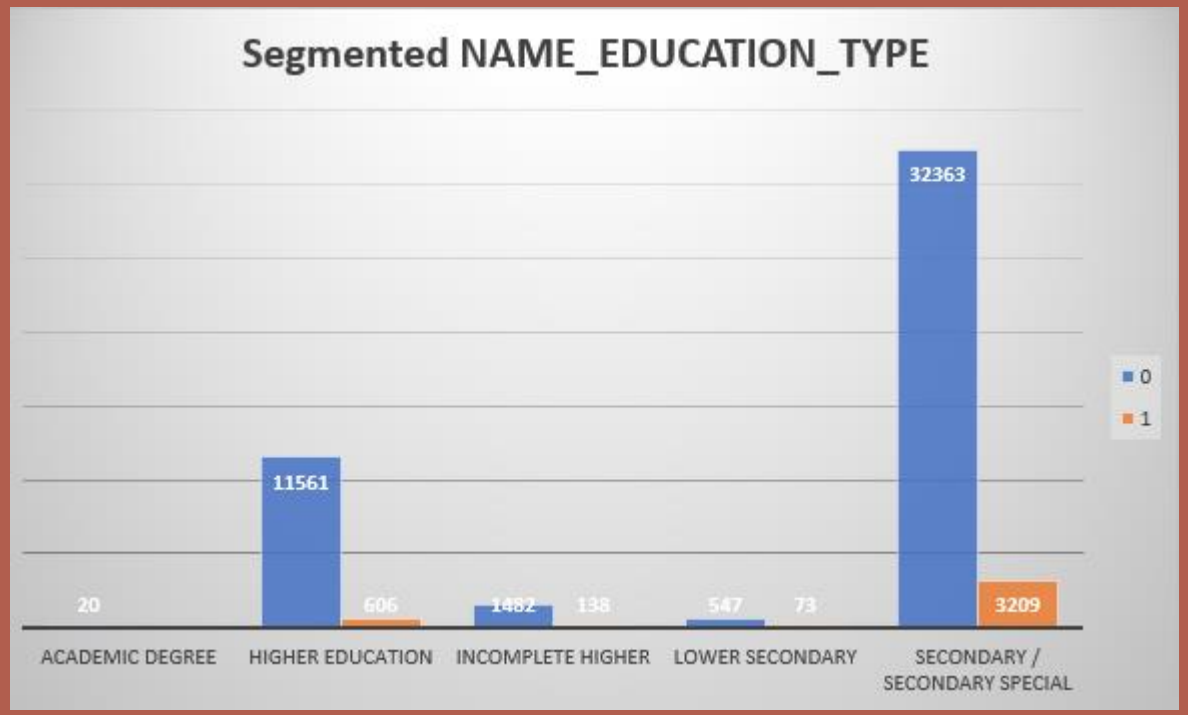




Most of the clients have 0-9 years of experience.

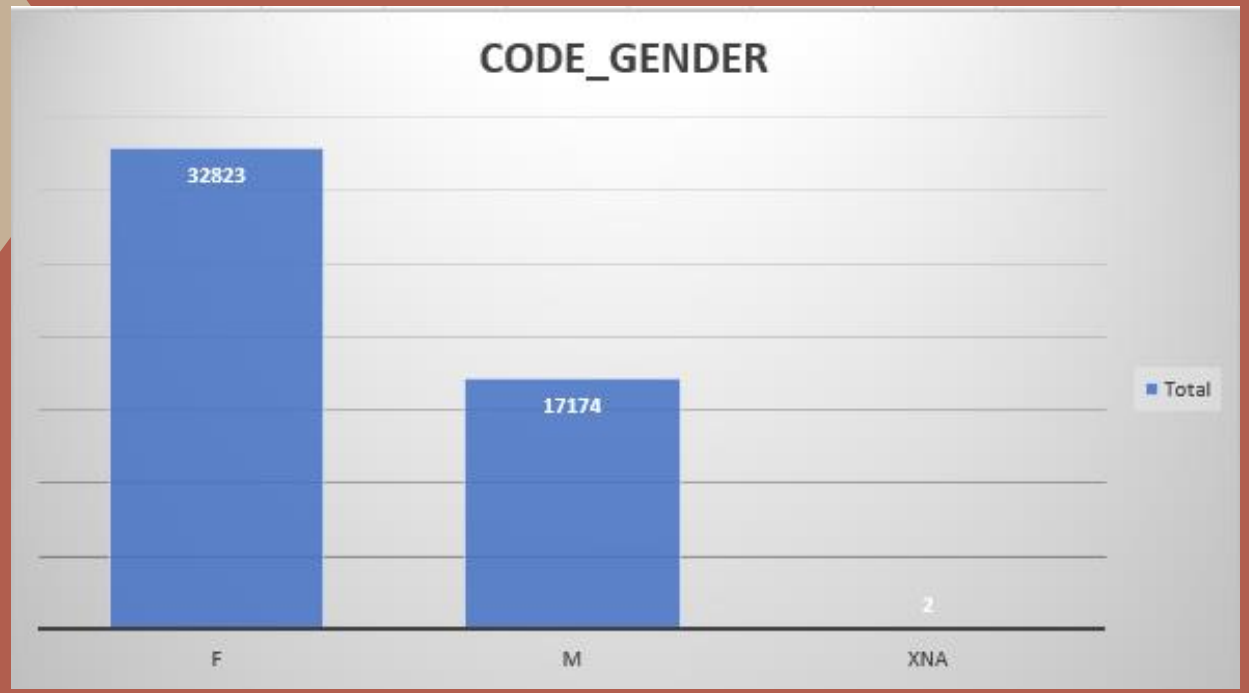
As experience increases, the chances of defaulting decreases.

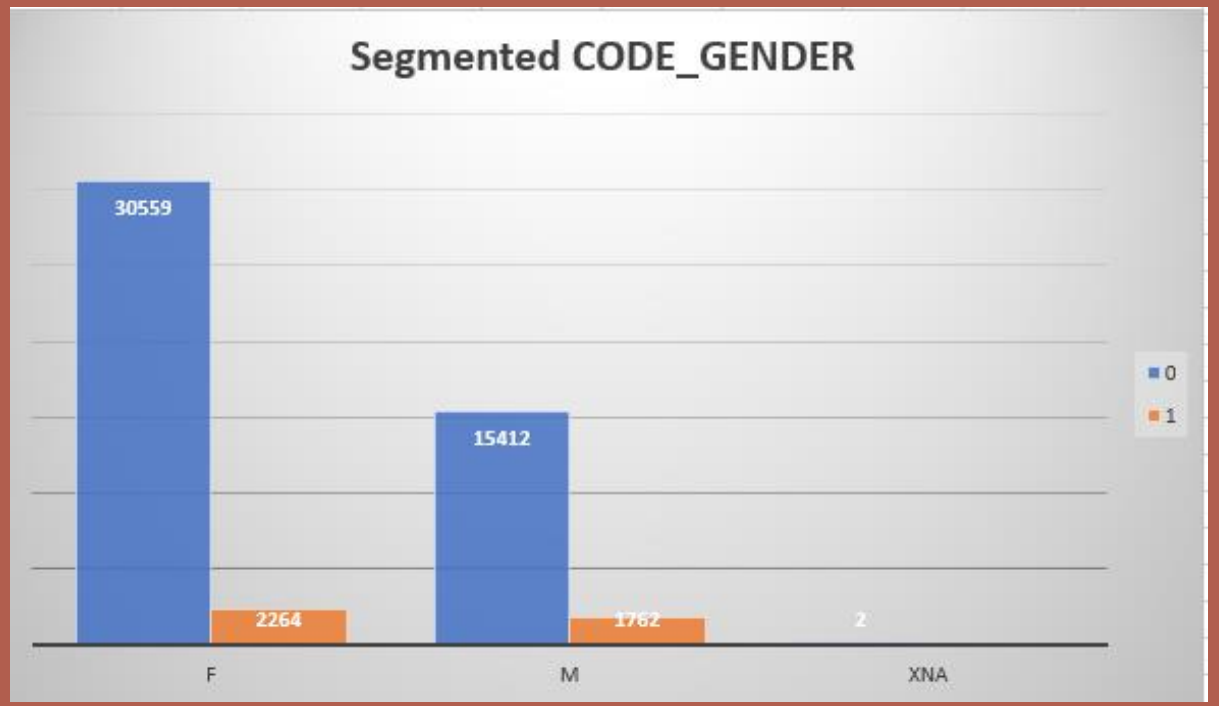




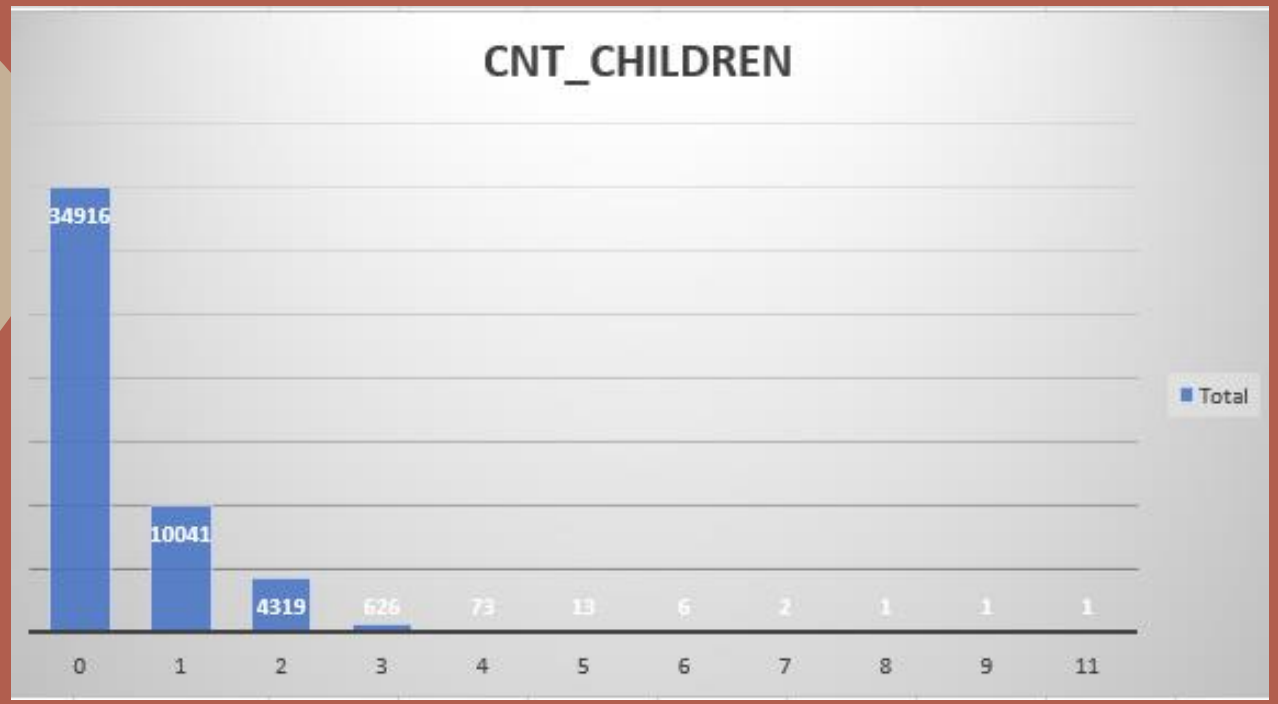
Clients with Secondary Special education have taken the highest number of loans.

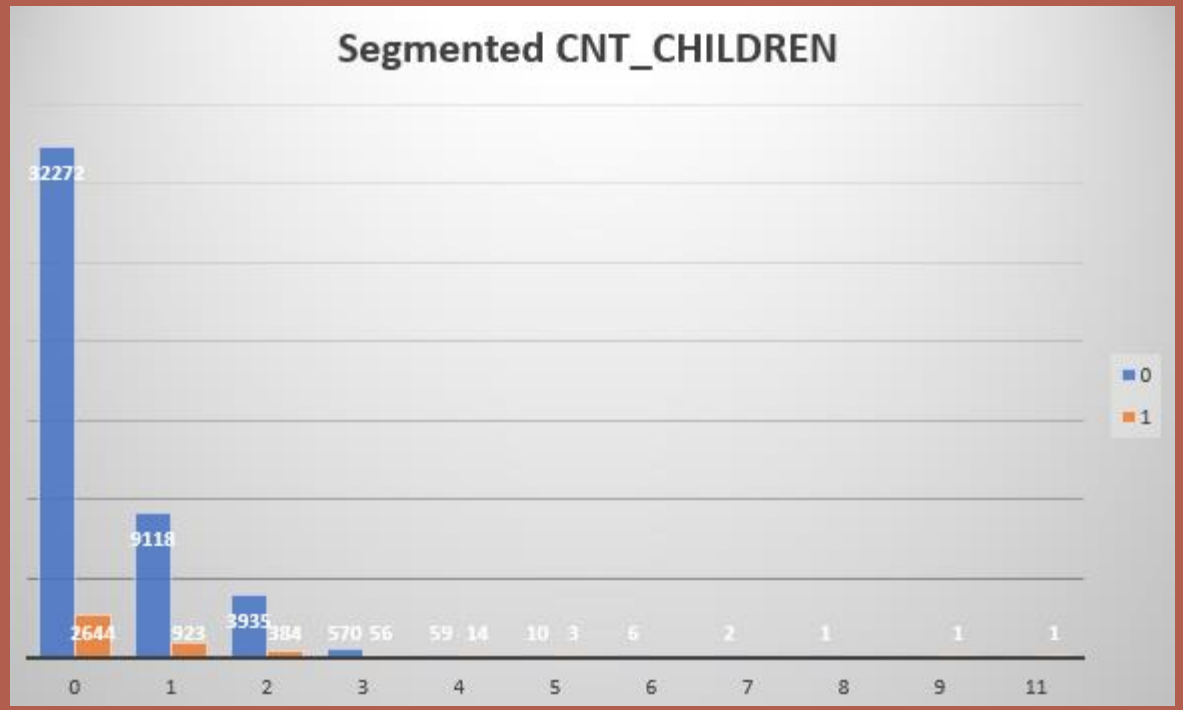
Highest default: Secondary Special
Lowest default: Academic Degree





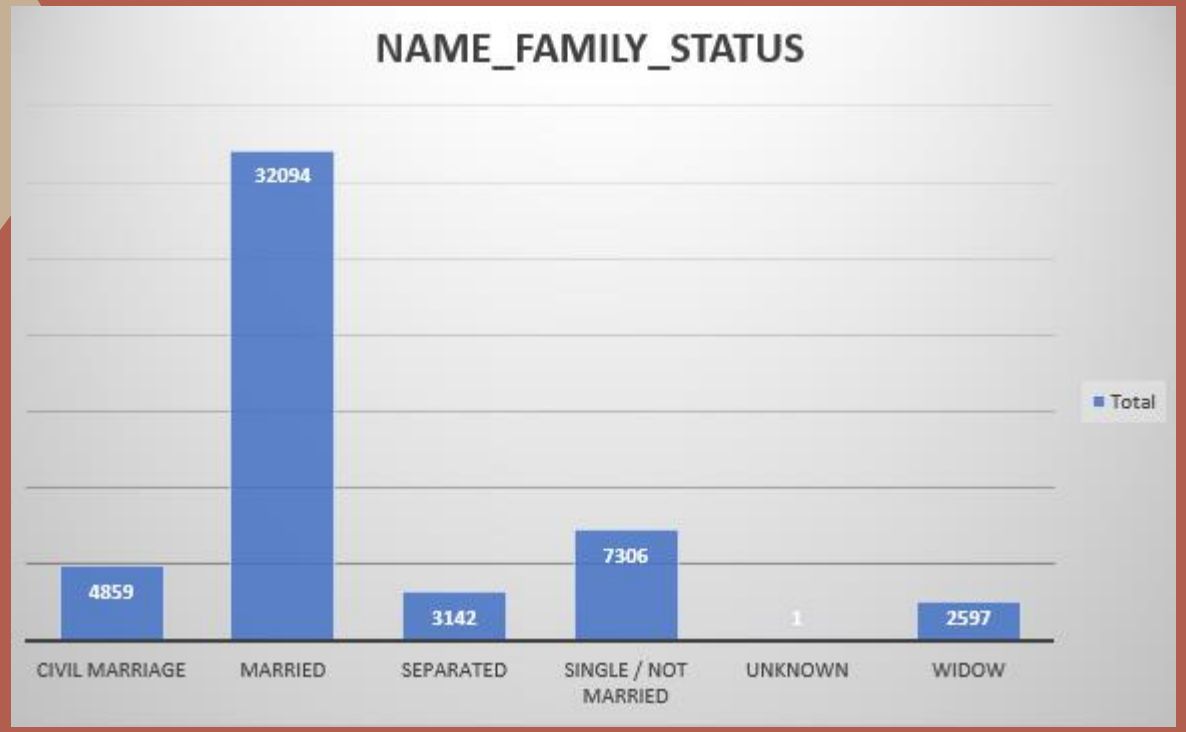
There are less male defaulters in comparison to female defaulters
Females have more defaulters but their ratio of defaulters and non-defaulters are better than males

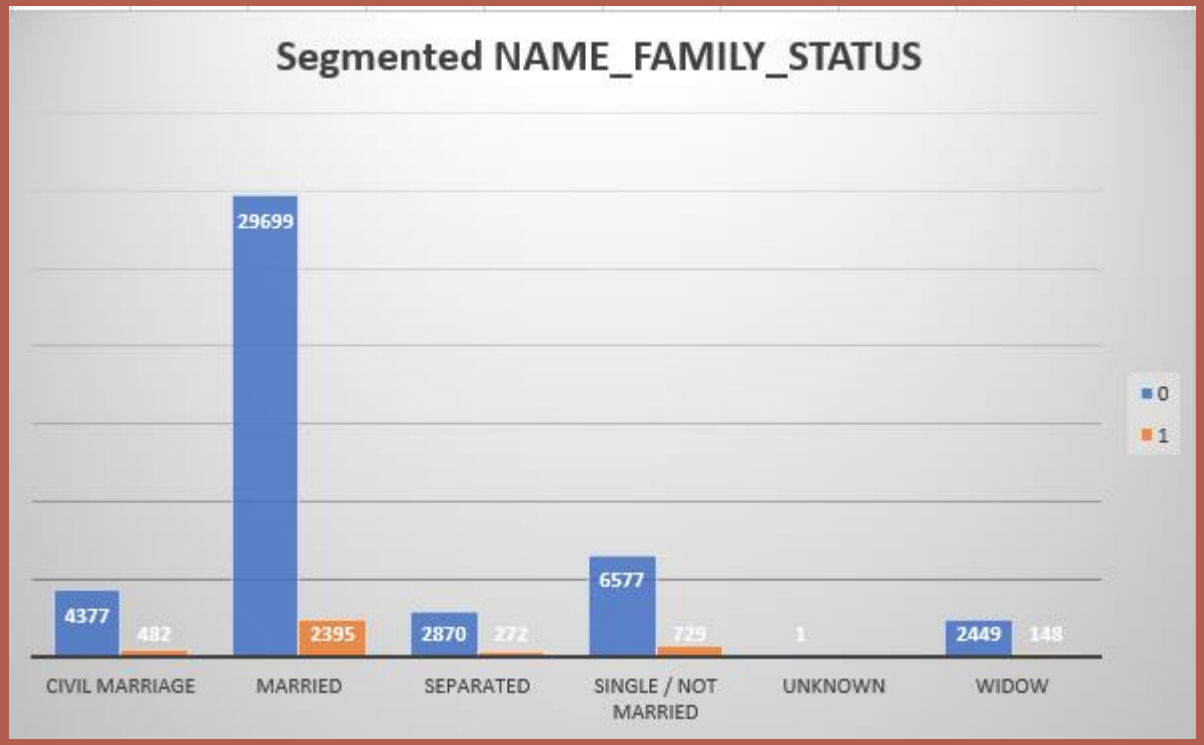




The clients without a child have taken the highest number of loans

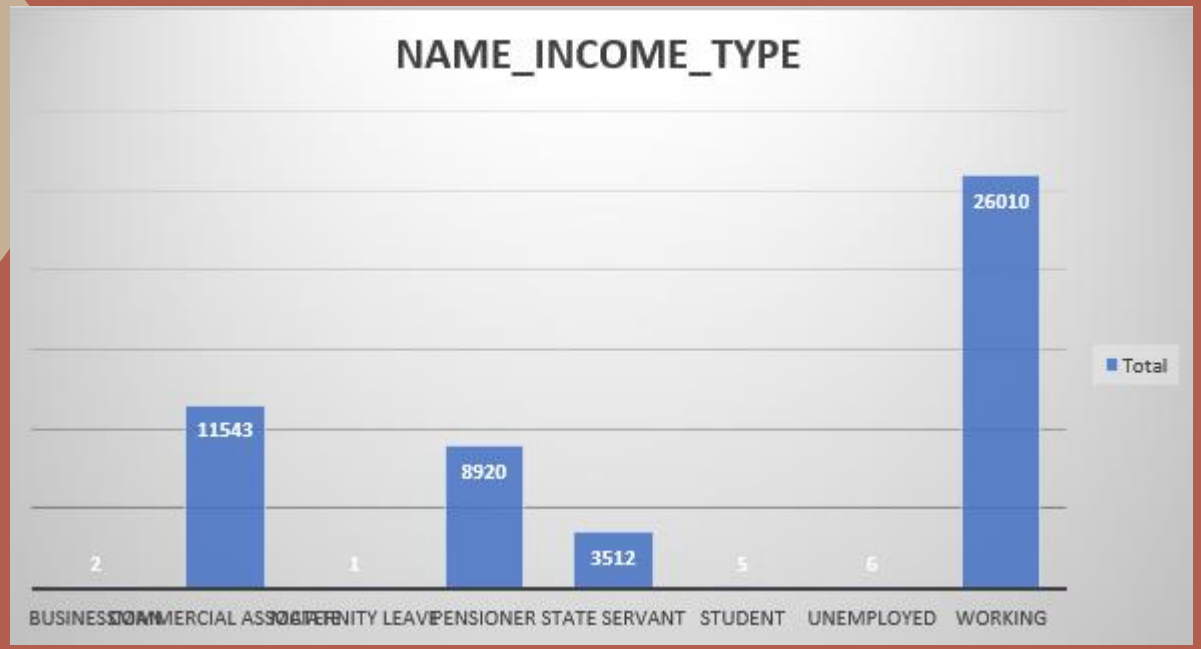
With the increase in number of children, the number of clients who take loans decreases

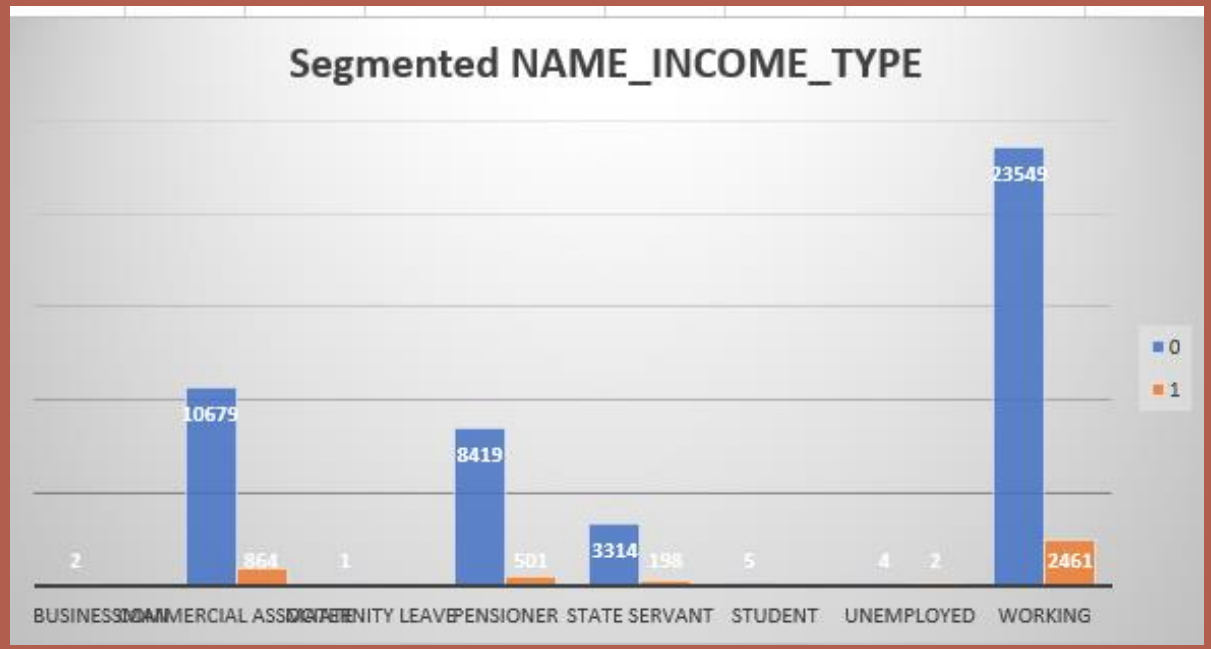




Married clients have taken the highest number of loans whereas clients who are widows have taken the least after ignoring unknown.

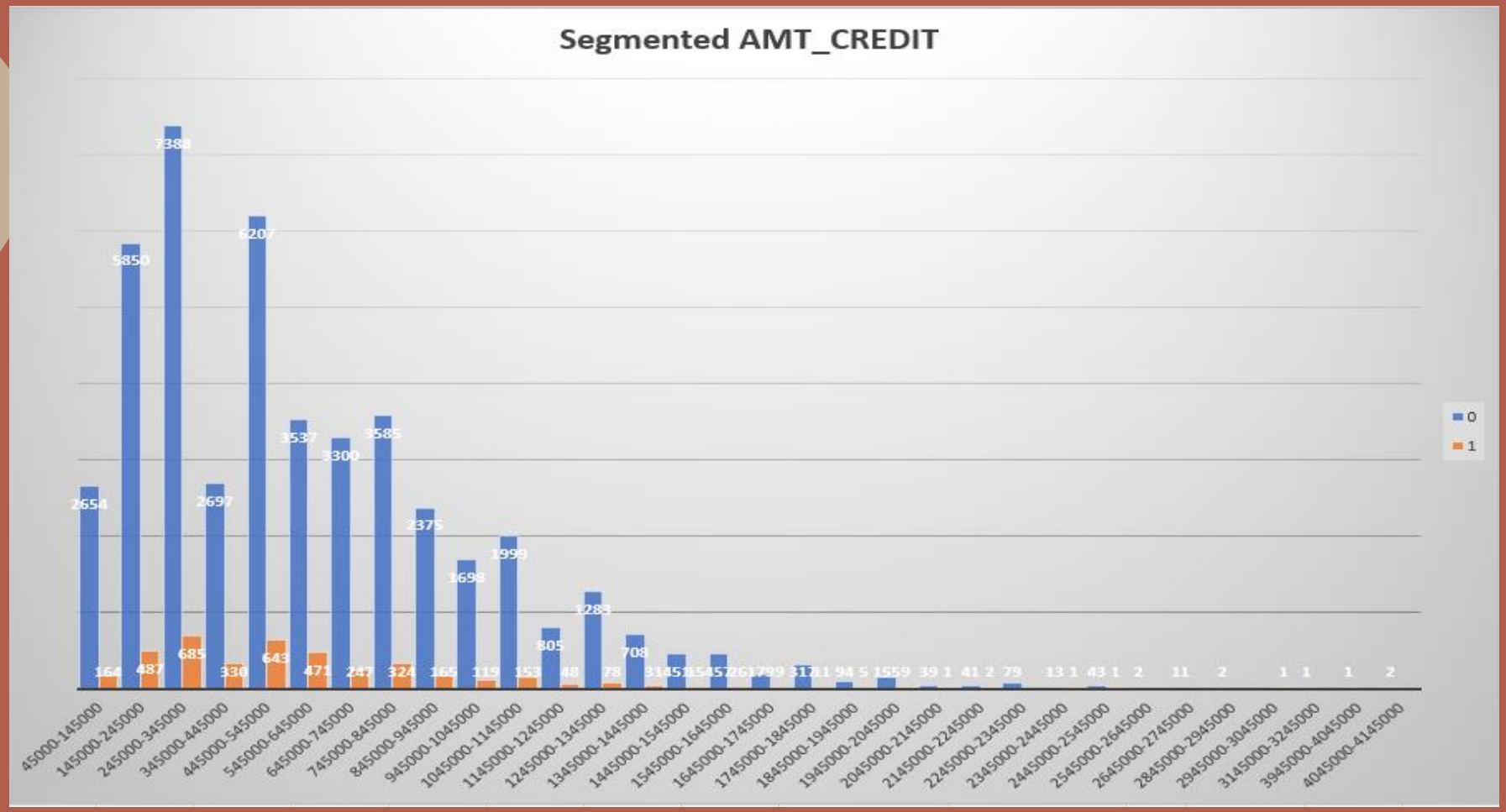
Widows are the least defaulters while Married clients are the highest defaulters.



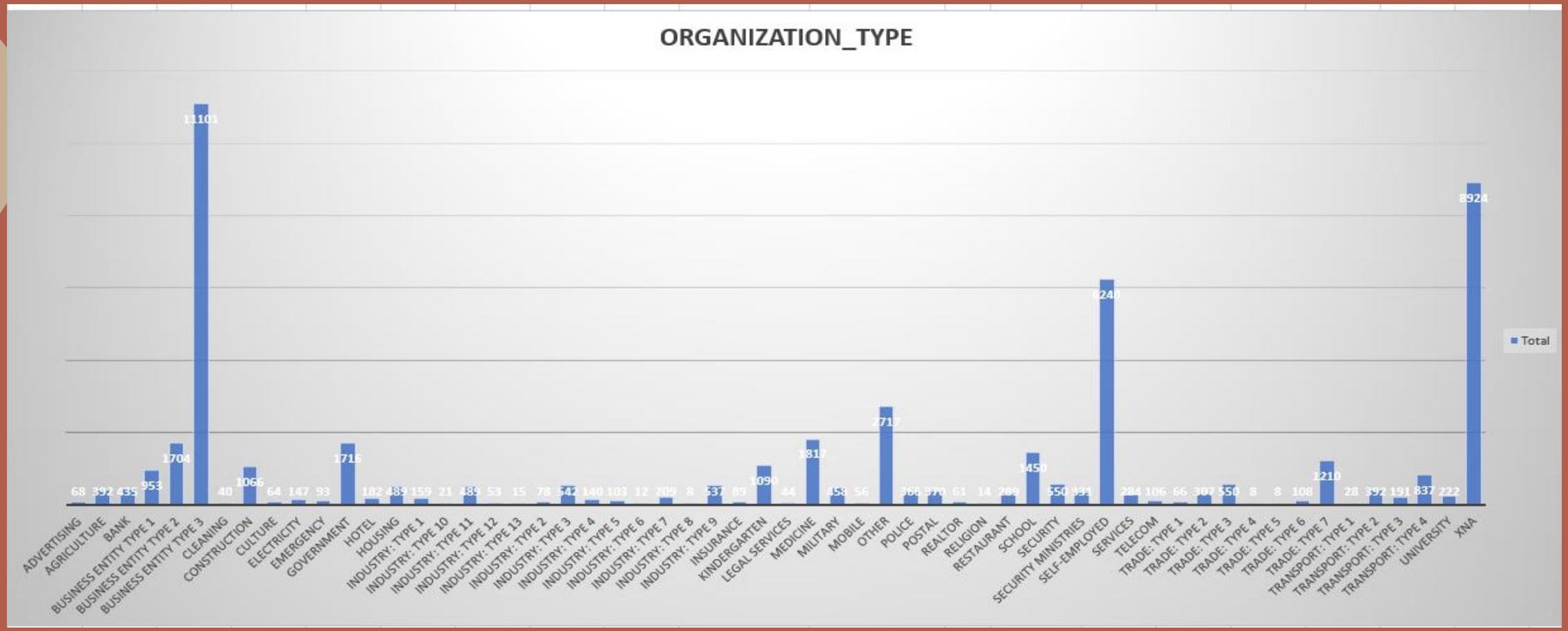


The groups whose income type is working are targeted by banks.

Clients that are businessman or students or at maternity leave are least defaulters.
Clients that are working are highest defaulters.

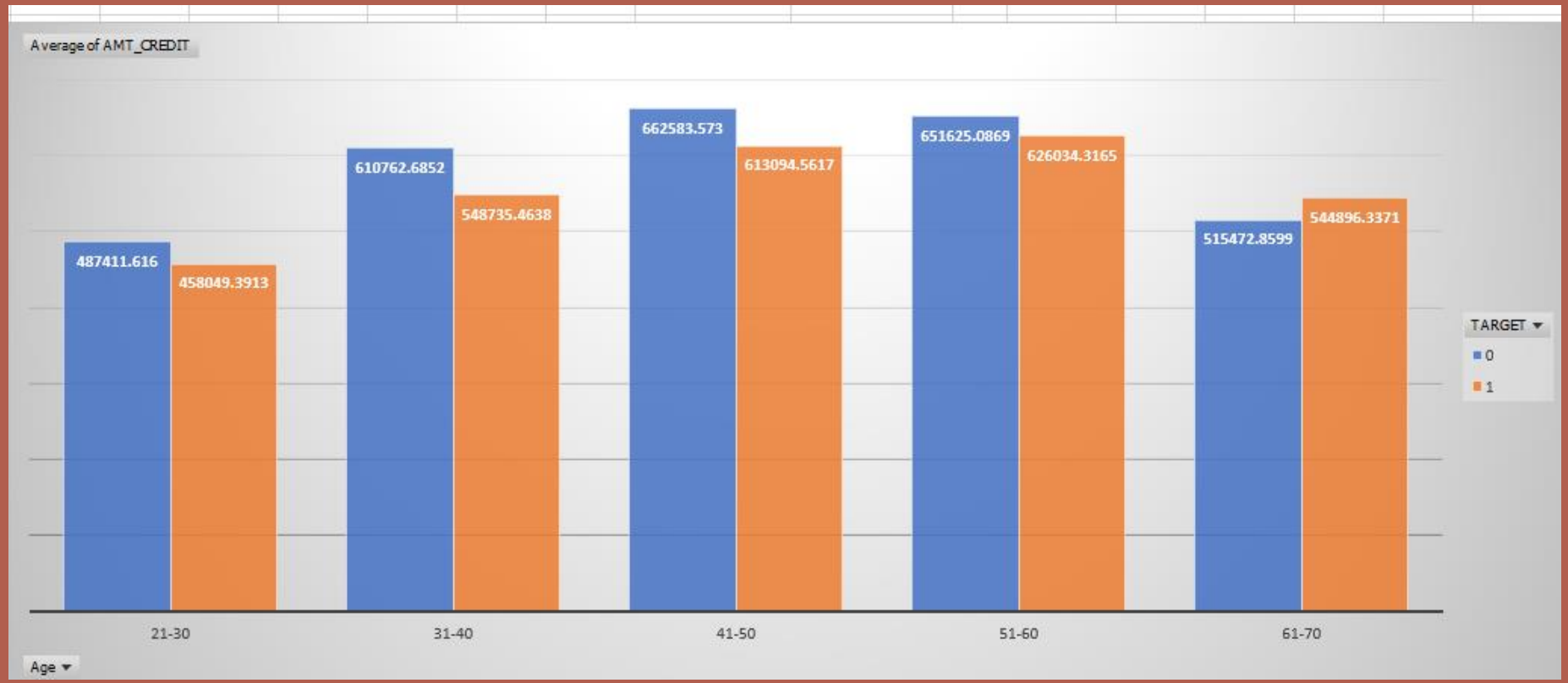


Majority of the clients took the loan between 2L – 3L.

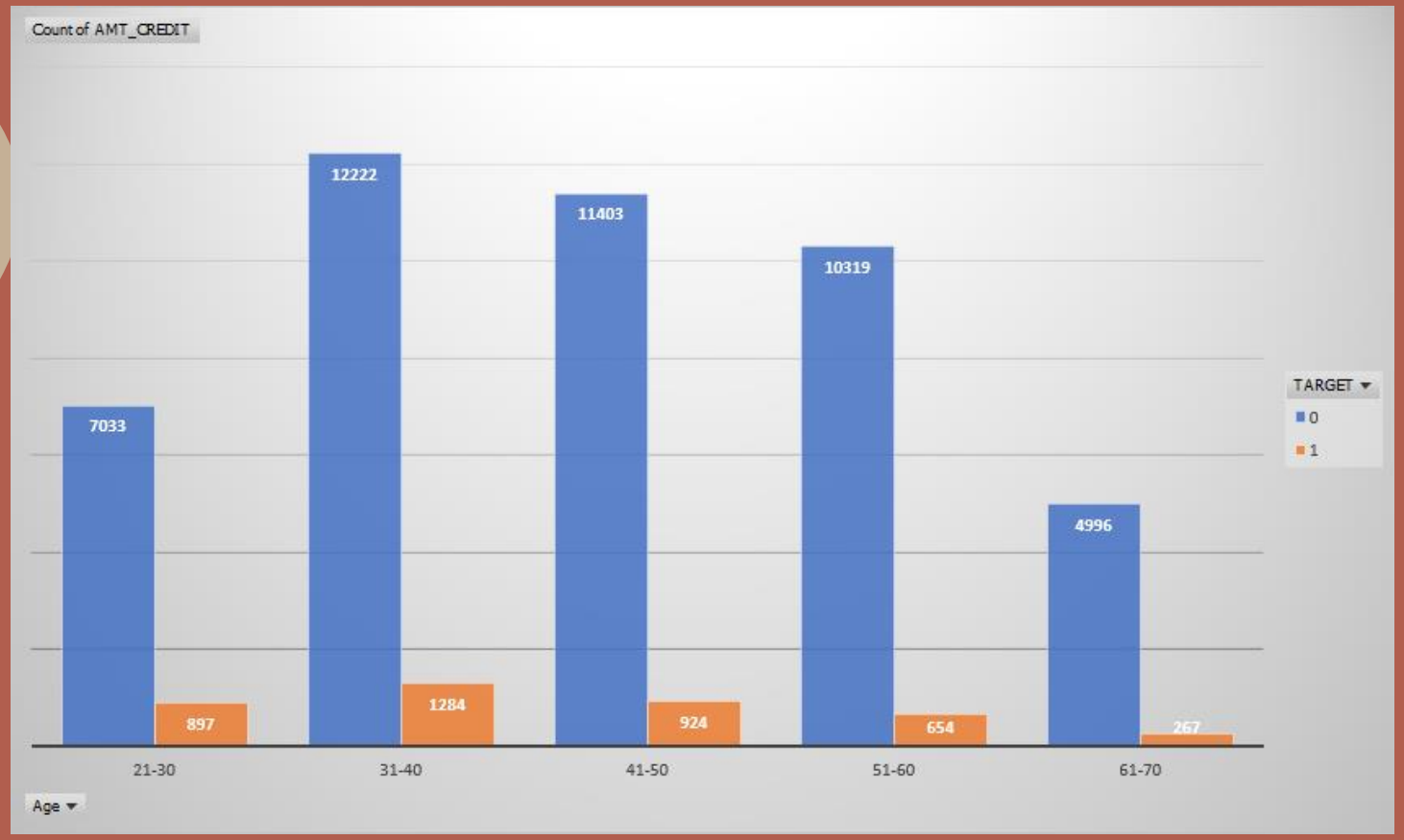


We can see that the clients working in Business Entity Type of organization took the highest number of loans

Bivariate Analysis



Age group 50-60 are the defaulter with the highest amount of loan. The age group 41-50 took the highest amount of loans



- Young Adults (21-30): This group has the highest proportion of payment difficulties mainly due to limited financial experience, unstable income, and higher debt burdens like student loans.
- Middle Age (31-50): Despite having the largest number of credit recipients, these groups show lower proportions of payment difficulties, indicating better financial stability.
- Older Adults (51-70): Older age groups have the lowest payment difficulties, suggesting greater financial security and reliable repayment behavior

Task E: 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.

- The dataset was segmented into 2 groups.
- Group 1: Clients with payment difficulties (Target: 1).
- Group 2: Clients without payment difficulties (Target: 0).
- These groups were formed using Filter feature.
- Next, the CORREL() function was used.

Some of the top correlations for Target 1 are:

AMT_GOODS_PRICE and AMT_CREDIT: 0.982381964
AMT_ANNUITY and AMT_CREDIT: 0.749665201
AMT_ANNUITY and AMT_GOODS_PRICE: 0.749904666

TARGET 1	CNT_CHILDRI	AMT_INCOME_TO1	AMT_CREC	AMT_ANNUIT	AMT_GOODS_PF	REGION_POPULATION_REL	DAYS_BIR	DAYS_EMPLO	DAYS_REGISTRA	DAYS_ID_PUBL	REG_REGION_NOT_LIVE_RE	OBS_30_CNT_SOCIAL_C	DEF_60_CNT_SOCIAL_C	AMT_REQ_CREDIT_BUREAU_	AMT_REQ_CREDIT_BUREAU	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	0.010110177	0.007602	0.02917298	-0.000680096	-0.020359154	0.249673	-0.18932418	0.152113117	-0.042360717	-0.015713279	0.01793193	-0.018505702	0.008160996	-0.011520595	-0.03080113
AMT_INCOME_TOTAL	0.01011018	1	0.015271	0.01800459	0.013298258	-0.006180303	0.009034	-0.01155596	-0.009561152	-0.009122006	0.000594885	-0.011280916	-0.006726958	-0.000864018	-0.003749228	-0.005100984
AMT_CREDIT	0.00760191	0.015271444	1	0.7496652	0.982381964	0.067775624	-0.14251	0.016039571	-0.042844404	-0.043771901	0.006456715	0.033466173	-0.029007236	0.083408196	-0.013361311	-0.016453973
AMT_ANNUITY	0.02917298	0.018004594	0.749665	1	0.749904666	0.073123998	-0.00875	-0.07955601	0.021581654	-0.02132109	0.031753358	0.013819016	-0.040471029	0.071295225	-0.001630664	0.001563273
AMT_GOODS_PRICE	-0.0006801	0.013298258	0.982382	0.74990467	1	0.077209215	-0.14111	0.020465	-0.042814863	-0.049144805	0.00716585	0.032853986	-0.020370507	0.079054172	-0.020117508	-0.023336591
REGION_POPULATION_RELA	-0.0203592	-0.006180303	0.067776	0.073124	0.077209215	1	-0.01647	0.007742909	-0.046130288	-0.005118563	-0.003105241	-0.008875436	0.027142318	0.075395596	0.015310168	0.024023928
DAYS_BIRTH	0.2496732	0.009033662	-0.14251	-0.0087517	-0.14111346	-0.016468731	1	-0.58147904	0.288437837	0.247896571	0.039614727	-0.011150233	-0.025756651	-0.007277397	-0.008783235	-0.090127316
DAYS_EMPLOYED	-0.1893242	-0.011555963	0.01604	-0.079556	0.020465	0.007742909	-0.58148	1	-0.188718437	-0.230063668	-0.035302931	0.003521851	0.023894099	-0.033065614	0.017875877	0.017692456
DAYS_REGISTRATION	0.15211312	-0.009561152	-0.04284	0.02158165	-0.042814863	-0.046130288	0.288438	-0.18871844	1	0.09029149	0.015849157	-0.005793296	-0.006412628	-0.001526001	-0.006290417	-0.025094194
DAYS_ID_PUBLISH	-0.0423607	-0.009122006	-0.04377	-0.0213211	-0.049144805	-0.005118563	0.247897	-0.23006367	0.09029149	1	0.024146053	-0.027313737	-0.027896348	-0.037917309	-0.032671471	-0.08164306
REG_REGION_NOT_LIVE_REC	-0.0157133	0.000594885	0.006457	0.03175936	0.00716585	-0.003105241	0.039615	-0.03530233	0.015849157	0.024146053	1	-0.031976397	0.005817862	0.051549217	-0.010452446	-0.033986108
OBS_30_CNT_SOCIAL_CIRCL	0.01793193	-0.011280916	0.033466	0.01381902	0.032853986	-0.008875436	-0.01115	0.003521851	-0.005793296	-0.027313737	-0.031976397	1	0.29795102	0.016077794	0.034835809	0.050517528
DEF_60_CNT_SOCIAL_CIRCL	-0.0185057	-0.006726958	-0.02901	-0.040471	-0.020370507	0.027142318	-0.02576	0.023894099	-0.006412628	-0.027896348	0.005817862	0.29795102	1	0.013034794	0.025347772	0.020626159
AMT_REQ_CREDIT_BUREAU_	0.008161	-0.000864018	0.083408	0.07129522	0.079054172	0.075395596	-0.00728	-0.03306561	-0.001526001	-0.037917309	0.051549217	0.016077794	0.013034794	1	0.019946401	0.038789503
AMT_REQ_CREDIT_BUREAU_	-0.0115206	-0.003749228	-0.01936	-0.0016307	-0.020117508	0.015310168	-0.00878	0.017875877	-0.006290417	-0.032671471	-0.010452446	0.034835809	0.025347772	0.019946401	1	0.103631744
AMT_REQ_CREDIT_BUREAU_	-0.0308011	-0.005100984	-0.01646	0.00156927	-0.023336591	0.024023928	-0.09013	0.017692456	-0.025094194	-0.08164306	-0.033986108	0.050517528	0.020626159	0.038789503	0.103631744	1

Some of the top correlations of Target 0 are:

AMT_GOODS_PRICE and AMT_CREDIT: 0.986879648
AMT_ANNUITY and AMT_CREDIT: 0.770772818
AMT_ANNUITY and AMT_GOODS_PRICE: 0.775888783

TARGET 0	CNT_CHIL	AMT_INCOME_T	AMT_CR	AMT_ANN	AMT_GOODS_	REGION_POPULATION_R	DAYS_B	DAYS_EMPI	DAYS_REGIST	DAYS_ID_PU	REG_REGION_NOT_LIVE	OBS_30_CNT_SOCIAL_CIR	DEF_60_CNT_SOCIAL_CIR	AMT_REQ_CREDIT_BUREA	AMT_REQ_CREDIT_BURE	AMT_REQ_CREDIT_BUREA	YEAR
CNT_CHILDREN	1	0.036319722	0.005705	0.026384	0.001383048	-0.024912809	0.335876	-0.24359152	0.183072478	-0.03253722	-0.010383386	0.016180239	-0.003330304	-0.011607819	-0.00473083	-0.035734888	
AMT_ANNUITY	0.02638396	0.451135167	0.770773	1	0.775888783	0.11727925	0.009911	-0.11300529	0.03460901	0.00942697	0.046175655	-0.009992103	-0.023010616	0.037985476	0.010059213	-0.004173747	
AMT_CREDIT	0.00570546	0.377965752	1	0.7707728	0.986879648	0.095539444	-0.05108	-0.07736722	0.008053758	-0.00829019	0.027812773	0.000876364	-0.018567338	0.063975989	0.026793294	-0.031568333	
AMT_GOODS_PRICE	0.00138305	0.384486402	0.98688	0.7758888	1	0.099047191	-0.04915	-0.07490119	0.011142315	-0.00975003	0.030399518	0.000529356	-0.019629719	0.065813069	0.027543198	-0.03416296	
AMT_INCOME_TOTAL	0.03631972	1	0.377966	0.4511352	0.384486402	0.181941261	0.073769	-0.16270268	0.06893375	0.032286356	0.078942904	-0.033045993	-0.032535174	0.074854679	0.015777535	0.031323516	
AMT_REQ_CREDIT_BUREAU_	-0.0116078	0.074854679	0.063976	0.0379855	0.065813069	0.070736631	-0.00245	-0.03295459	-0.010724839	-0.01323263	-0.00860612	0.008170201	0.003971647	1	0.011888446	0.019311173	
AMT_REQ_CREDIT_BUREAU_	-0.0047308	0.015777535	0.026793	0.0100592	0.027549198	-0.009694599	-0.02152	0.014577401	0.003127351	-0.02458808	-0.00026605	0.008845246	0.008306707	0.011888446	1	0.121744813	
AMT_REQ_CREDIT_BUREAU_	-0.0357349	0.031323516	-0.03157	-0.004174	-0.03416296	0.004652396	-0.07027	0.044183816	-0.02236176	-0.04469288	-0.019525847	0.034161046	0.015204988	0.019311173	0.121744813	1	
DAYS_BIRTH	0.33587627	0.073769425	-0.05108	0.0099114	-0.049148204	-0.030435419	1	-0.61528998	0.335028046	0.270073313	0.060427036	0.012287026	0.002207122	-0.002452976	-0.021522968	-0.070267716	
DAYS_EMPLOYED	-0.2435915	-0.162702675	-0.07737	-0.113005	-0.074901185	-0.006610653	-0.61529	1	-0.204370881	-0.27222439	-0.03641311	0.005650192	0.016516022	-0.032954589	0.014577401	0.044183816	
DAYS_ID_PUBLISH	-0.0325372	0.032286356	-0.00829	0.009427	-0.00975003	-0.002236288	0.270073	-0.27222439	0.103548902	1	0.033228477	-0.011854044	0.002642424	-0.013232625	-0.024588081	-0.044692876	
DAYS_REGISTRATION	0.18307248	0.06893375	0.008054	0.034609	0.011142315	-0.058501361	0.335028	-0.20437088	1	0.103548902	0.027899954	0.010977833	0.006282428	-0.010724839	0.003127351	-0.02236176	
DEF_60_CNT_SOCIAL_CIRCLI	-0.0033303	-0.032535174	-0.01857	-0.023011	-0.019629719	0.003253593	0.002207	0.016516022	0.006282428	0.002642424	-0.009383908	0.229170262	1	0.003971647	0.008306707	0.015204988	
OBS_30_CNT_SOCIAL_CIRCL	0.0161803	-0.033045993	0.000876	-0.009992	0.000529356	-0.01906908	0.012287	0.005650192	0.010977833	-0.01185404	-0.015119959	1	0.229170262	0.008170201	0.008845246	0.034161046	
REG_REGION_NOT_LIVE_REC	-0.0103834	0.078942904	0.027813	0.0461757	0.030399518	-0.003185217	0.060427	-0.03641311	0.027899954	0.033228477	1	-0.015119959	-0.009383908	-0.00860612	-0.00026605	-0.019525847	
REGION_POPULATION_RELA	-0.0249128	0.181941261	0.095539	0.1172792	0.099047191	1	-0.03044	-0.00661065	-0.058501361	-0.00223629	-0.003185217	-0.01906908	0.003253593	0.070736631	-0.009694599	0.004652396	

Results

- Most clients are loan re-payers, and cash loans are the most common type of loan.
- The bank lends more to females, but males are less likely to default on loans. Females have more defaulters but their ratio of defaulters and non-defaulters are better than males
- Older and more experienced clients have a lower default risk and are more profitable for the bank.
- Clients with higher education levels default less compared to those with lower education qualifications.
- Unemployed clients have the highest default risk and often take larger loans, requiring caution in lending.
- Clients with more children are less likely to take loans.
- As age increases, the loan amounts taken are higher, but default rates are significantly lower, making older clients safer borrowers.
- Avoiding excessively high loans above norms can reduce default rates and improve overall lending outcomes.

Conclusion

- This project was a great learning experience, especially in handling large datasets and working with multiple data sources. It helped me improve my skills in dealing with missing values, outliers, and using Excel more efficiently. It also gave me a good idea of how to work with bank datasets and analyze big data, which has made me feel more confident about handling similar challenges in the future.