Bank Loan Case Study

Project Description:

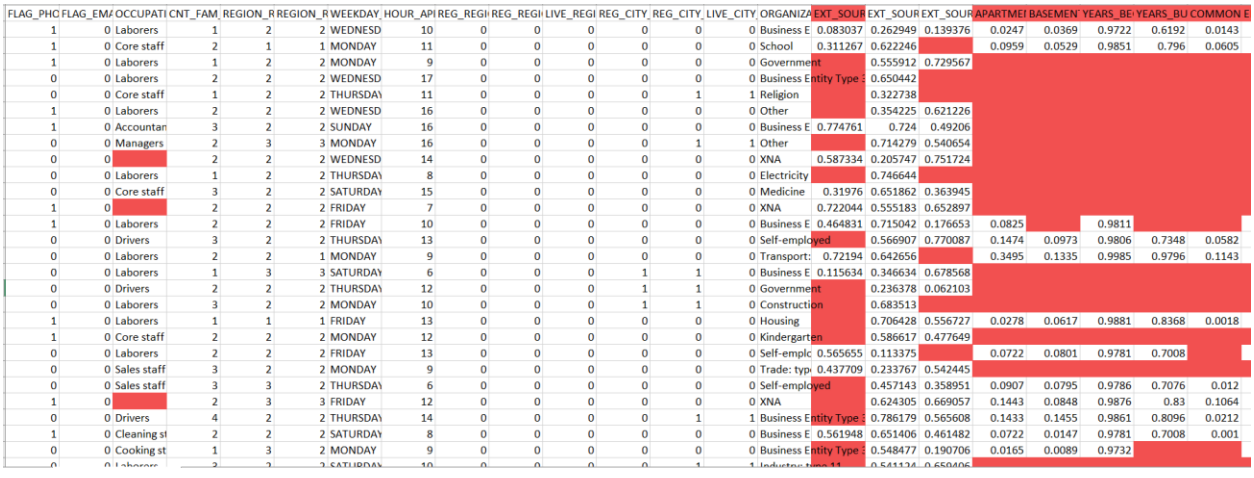
The project involves conducting Exploratory Data Analysis (EDA) on a loan application dataset for a finance company that focuses on lending to urban customers. The intention is to find trends that suggest if a client would struggle to meet their installment payments, assisting the business in making well-informed decisions on loan approval. The dataset includes data on loan applications together with payment histories, customer attributes, and loan outcomes. Customers having payment issues and all other situations are the two categories of scenarios taken into consideration.

LINK TO THE EXCEL FILE:

EXCEL FILE

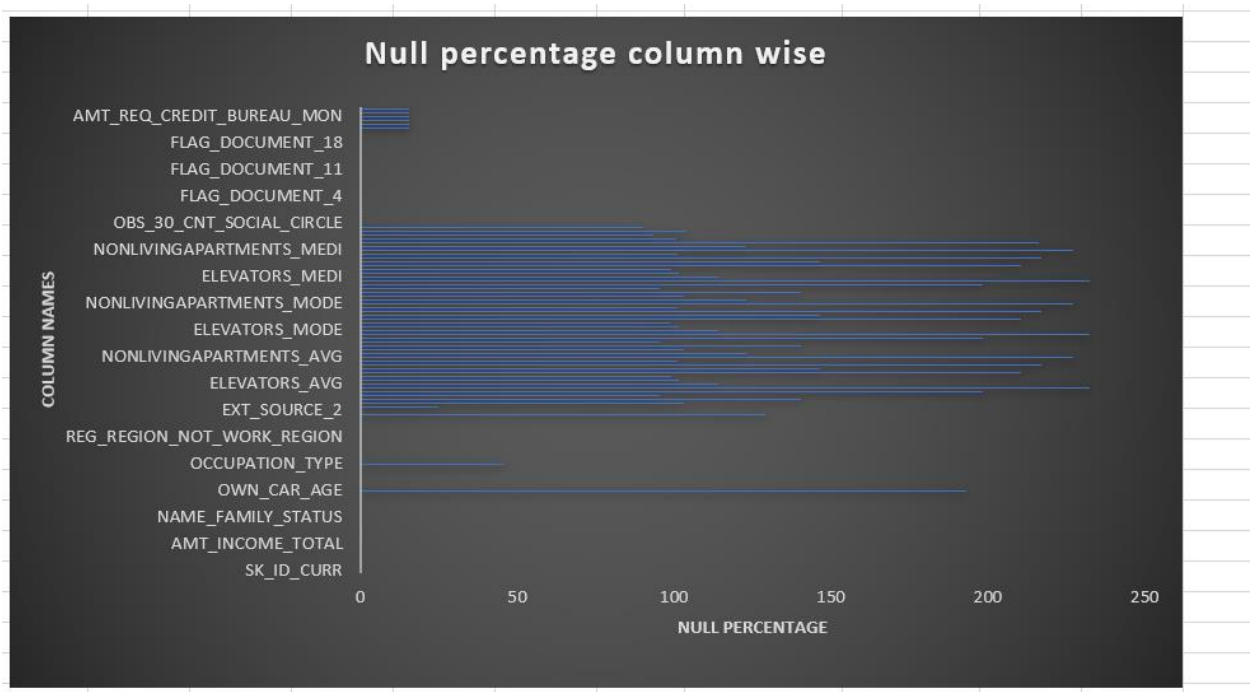
Approach:

Identifying Missing Data:



FLAG_PHO	FLAG_EM/	OCCUPATI	CNT_FAM	REGION_R	REGION_R	WEEKDAY	HOUR_API	REG_REGI	REG_REGI	LIVE_REGI	REG_CITY	REG_CITY	LIVE_CITY	ORGANIZA	EXT_SOUR	EXT_SOUR	EXT_SOUR	APARTME	BASEMEN	YEARS_BE	YEARS_BU	COMMON E
1	0	Laborers	1	2	2	WEDNESD	10	0	0	0	0	0	0	Business E	0.083037	0.262949	0.139376	0.0247	0.0369	0.9722	0.6192	0.0143
1	0	Core staff	2	1	1	MONDAY	11	0	0	0	0	0	0	School	0.311267	0.622246		0.0959	0.0529	0.9851	0.796	0.0605
1	0	Laborers	1	2	2	MONDAY	9	0	0	0	0	0	0	Government		0.555912	0.729567					
0	0	Laborers	2	2	2	WEDNESD	17	0	0	0	0	0	0	Business Entity Type 3	0.650442							
0	0	Core staff	1	2	2	THURSDAY	11	0	0	0	0	1	1	Religion		0.322738						
1	0	Laborers	2	2	2	WEDNESD	16	0	0	0	0	0	0	Other		0.354225	0.621226					
1	0	Accountan	3	2	2	SUNDAY	16	0	0	0	0	0	0	Business E	0.774761	0.724	0.49206					
0	0	Managers	2	3	3	MONDAY	16	0	0	0	0	1	1	Other		0.714279	0.540654					
0	0		2	2	2	WEDNESD	14	0	0	0	0	0	0	XNA	0.587334	0.205747	0.751724					
0	0	Laborers	1	2	2	THURSDAY	8	0	0	0	0	0	0	Electricity		0.746644						
0	0	Core staff	3	2	2	SATURDAY	15	0	0	0	0	0	0	Medicine	0.31976	0.651862	0.363945					
1	0		2	2	2	FRIDAY	7	0	0	0	0	0	0	XNA	0.722044	0.555183	0.652897					
1	0	Laborers	2	2	2	FRIDAY	10	0	0	0	0	0	0	Business E	0.464831	0.715042	0.176653	0.0825		0.9811		
0	0	Drivers	3	2	2	THURSDAY	13	0	0	0	0	0	0	Self-employed	0.566907	0.770087		0.1474	0.0973	0.9806	0.7348	0.0582
0	0	Laborers	2	2	1	MONDAY	9	0	0	0	0	0	0	Transport:	0.72194	0.642656		0.3495	0.1335	0.9985	0.9796	0.1143
0	0	Laborers	1	3	3	SATURDAY	6	0	0	0	0	1	1	Business E	0.115634	0.346634	0.678568					
0	0	Drivers	2	2	2	THURSDAY	12	0	0	0	0	1	1	Government		0.236378	0.062103					
0	0	Laborers	3	2	2	MONDAY	10	0	0	0	0	1	1	Construction		0.683513						
1	0	Laborers	1	1	1	FRIDAY	13	0	0	0	0	0	0	Housing		0.706428	0.556727	0.0278	0.0617	0.9881	0.8368	0.0018
1	0	Core staff	2	2	2	MONDAY	12	0	0	0	0	0	0	Kindergarten		0.586617	0.477649					
0	0	Laborers	2	2	2	FRIDAY	13	0	0	0	0	0	0	Self-emplo	0.565655	0.113375		0.0722	0.0801	0.9781	0.7008	
0	0	Sales staff	3	2	2	MONDAY	9	0	0	0	0	0	0	Trade: type	0.437709	0.233767	0.542445					
0	0	Sales staff	3	3	2	THURSDAY	6	0	0	0	0	0	0	Self-employed		0.457143	0.358951	0.0907	0.0795	0.9786	0.7076	0.012
1	0		2	3	3	FRIDAY	12	0	0	0	0	0	0	XNA		0.624305	0.669057	0.1443	0.0848	0.9876	0.83	0.1064
0	0	Drivers	4	2	2	THURSDAY	14	0	0	0	0	0	1	1	Business Entity Type 3	0.786179	0.565608	0.1433	0.1455	0.9861	0.8096	0.0212
1	0	Cleaning st	2	2	2	SATURDAY	8	0	0	0	0	0	0	Business E	0.561948	0.651406	0.461482	0.0722	0.0147	0.9781	0.7008	0.001
0	0	Cooking st	1	3	2	MONDAY	9	0	0	0	0	0	0	Business Entity Type 3	0.548477	0.190706		0.0165	0.0089	0.9732		

Calculated the null percentage of each column



The columns with more than 50% of the null values are deleted

The other columns that have a null percentage of less than 50% are filled with mean or frequency. Utilized Excel functions such as COUNT, ISBLANK, and IF to identify missing data.

Columns	Average
AMT_ANNUITY	27107.37736
AMT_GOODS_PF	539060.0361
CNT_FAM_MEM	2.158946358
EXT_SOURCE_2	0.513823595
EXT_SOURCE_3	0.511881408
OBS_30_CNT_SC	1.420782244
DEF_30_CNT_SC	0.141819349
OBS_60_CNT_SC	1.403664386
DEF_60_CNT_SC	0.098332363
AMT_REQ_CREC	0.007095805
AMT_REQ_CREC	0.007511846
AMT_REQ_CREC	0.032381833
AMT_REQ_CREC	0.270287761
AMT_REQ_CREC	0.260973073
AMT_REQ_CREC	1.881035479

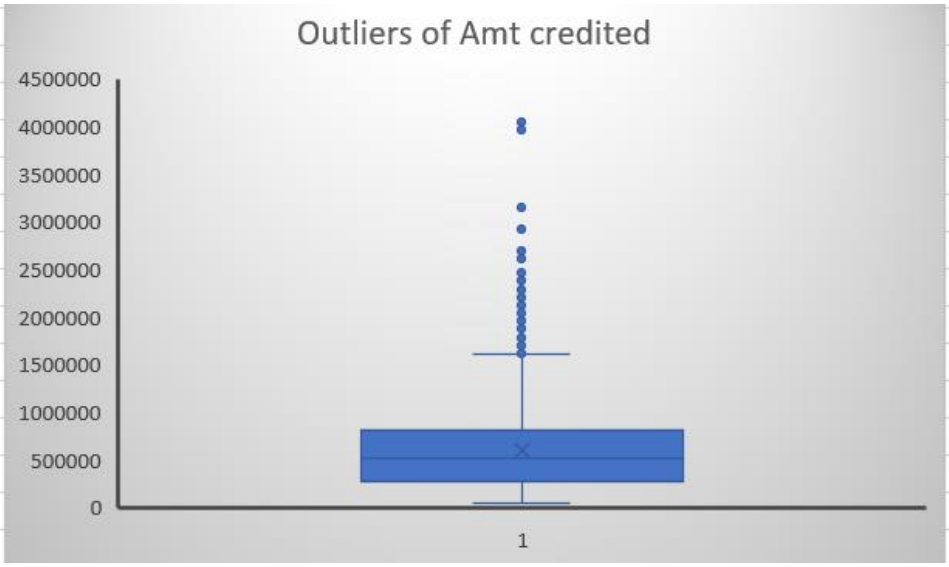
NAME_TYI	FREQUENCY
Unaccomp	40435
Family	6549
Spouse, pa	1849
Children	542
Other_A	137
0	0
Other_B	259
Group of p	36

OCCUPATI	FREQUENCY
Laborers	8952
Core staff	4434
Accountan	1621
Managers	3489
0	0
Drivers	3044
Sales staff	5160
Cleaning st	739
Cooking st	963
Private ser	447
Medicine s	1403
Security st	1140
High skill te	1852
Waiters/ba	228
Low-skill L	357
Realty age	123
Secretaries	212
IT staff	80
HR staff	101

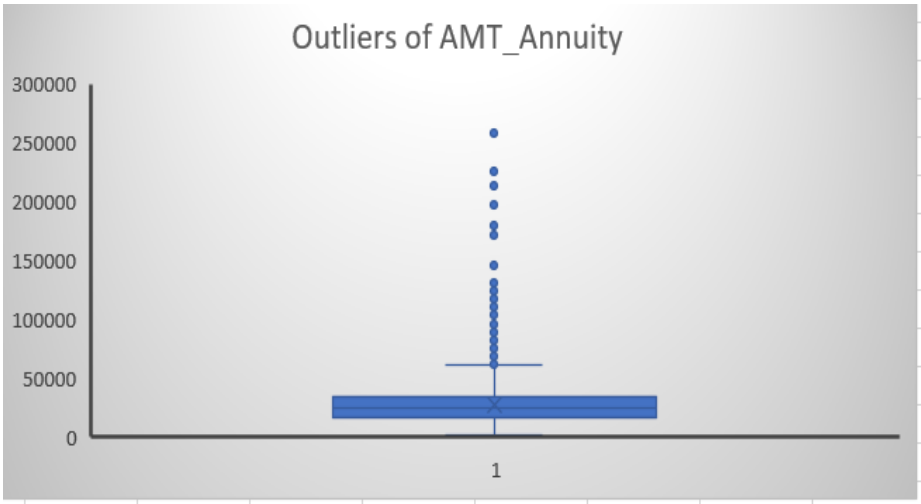
After cleaning, the dataset will look like

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
100002	1	Cash loans	M	N	Y	0	202500	406597.5	Working	Secondary / secondary specia	Single / not married
100003	0	Cash loans	F	N	N	0	270000	1293502.5	State servant	Higher education	Married
100004	0	Revolving loans	M	Y	Y	0	67500	135000	Working	Secondary / secondary specia	Single / not married
100006	0	Cash loans	F	N	Y	0	135000	312682.5	Working	Secondary / secondary specia	Civil marriage
100007	0	Cash loans	M	N	Y	0	121500	513000	Working	Secondary / secondary specia	Single / not married
100008	0	Cash loans	M	N	Y	0	99000	490495.5	State servant	Secondary / secondary specia	Married
100009	0	Cash loans	F	Y	Y	1	171000	1560726	Commercial associate	Higher education	Married
100010	0	Cash loans	M	Y	Y	0	360000	1530000	State servant	Higher education	Married
100011	0	Cash loans	F	N	Y	0	112500	1019610	Pensioner	Secondary / secondary specia	Married
100012	0	Revolving loans	M	N	Y	0	135000	405000	Working	Secondary / secondary specia	Single / not married
100014	0	Cash loans	F	N	Y	1	112500	652500	Working	Higher education	Married
100015	0	Cash loans	F	N	Y	0	38419.155	148365	Pensioner	Secondary / secondary specia	Married
100016	0	Cash loans	F	N	Y	0	67500	80865	Working	Secondary / secondary specia	Married
100017	0	Cash loans	M	Y	N	1	225000	918468	Working	Secondary / secondary specia	Married
100018	0	Cash loans	F	N	Y	0	189000	773680.5	Working	Secondary / secondary specia	Married
100019	0	Cash loans	M	Y	Y	0	157500	299772	Working	Secondary / secondary specia	Single / not married
100020	0	Cash loans	M	N	N	0	108000	509602.5	Working	Secondary / secondary specia	Married
100021	0	Revolving loans	F	N	Y	1	81000	270000	Working	Secondary / secondary specia	Married
100022	0	Revolving loans	F	N	Y	0	112500	157500	Working	Secondary / secondary specia	Widow
100023	0	Cash loans	F	N	Y	1	90000	544491	State servant	Higher education	Single / not married
100024	0	Revolving loans	M	Y	Y	0	135000	427500	Working	Secondary / secondary specia	Married
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	Commercial associate	Secondary / secondary specia	Married
100026	0	Cash loans	F	N	N	1	450000	497520	Working	Secondary / secondary specia	Married
100027	0	Cash loans	F	N	Y	0	83250	239850	Pensioner	Secondary / secondary specia	Married
100029	0	Cash loans	M	Y	N	2	135000	247500	Working	Secondary / secondary specia	Married
100030	0	Cash loans	F	N	Y	0	90000	225000	Working	Secondary / secondary specia	Married
100031	1	Cash loans	F	N	Y	0	112500	979992	Working	Secondary / secondary specia	Widow
100032	0	Cash loans	M	N	Y	1	112500	327024	Working	Secondary / secondary specia	Married
100033	0	Cash loans	M	Y	Y	0	270000	790830	State servant	Higher education	Single / not married

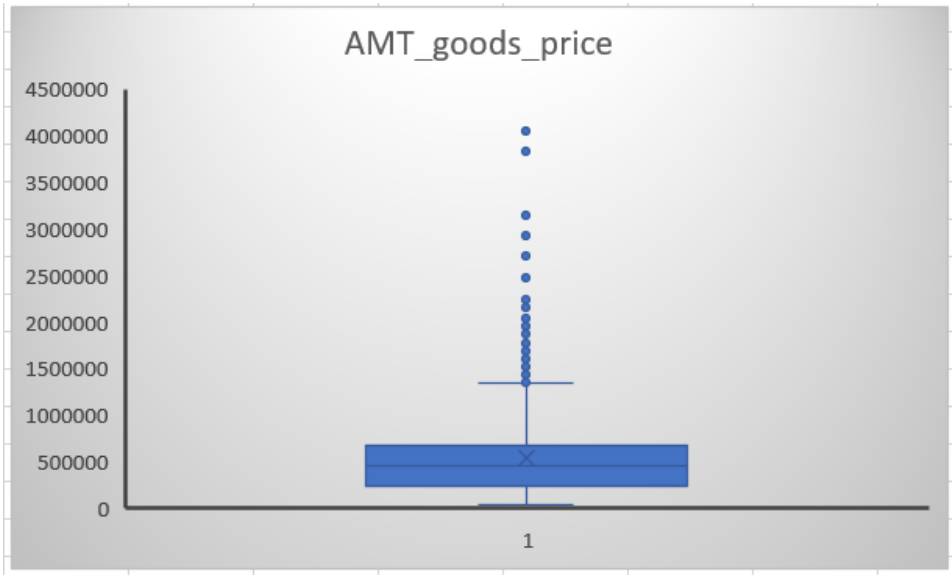
Detecting Outliers: Used Excel statistical functions like QUARTILE, and IQR to detect outliers in numerical variables. Validated outliers against business rules to decide on further investigation.



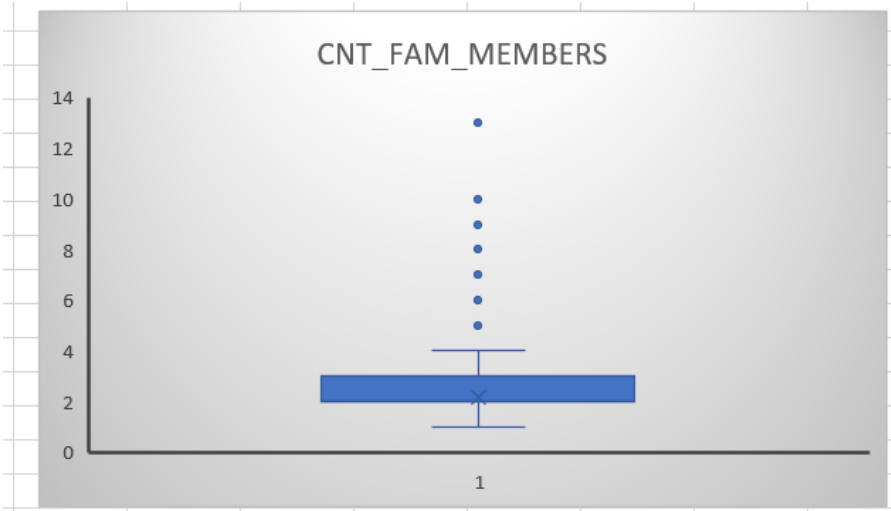
MIN	45000
Q1	270000
MEDIAN	514777.5
Q3	808650
MAX	4050000
IQR	538650



MIN	2052
Q1	16456.5
MEDIAN	24939
Q3	34596
MAX	258025.5
IQR	18139.5

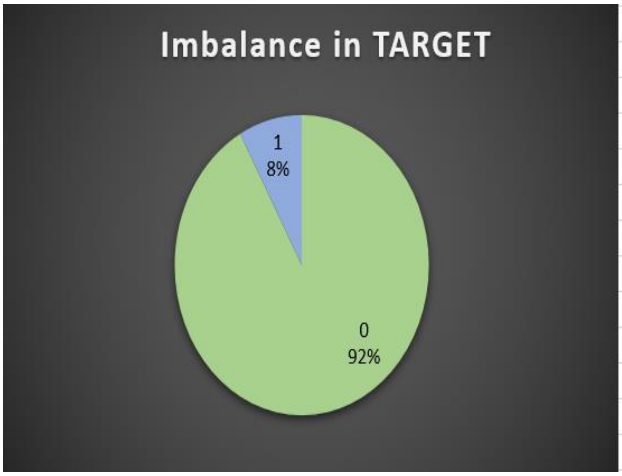


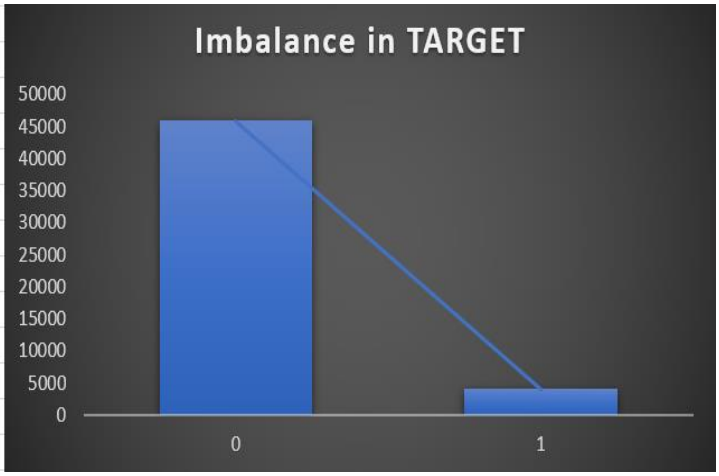
MIN	45000
Q1	238500
MEDIAN	450000
Q3	679500
MAX	4050000
IQR	441000

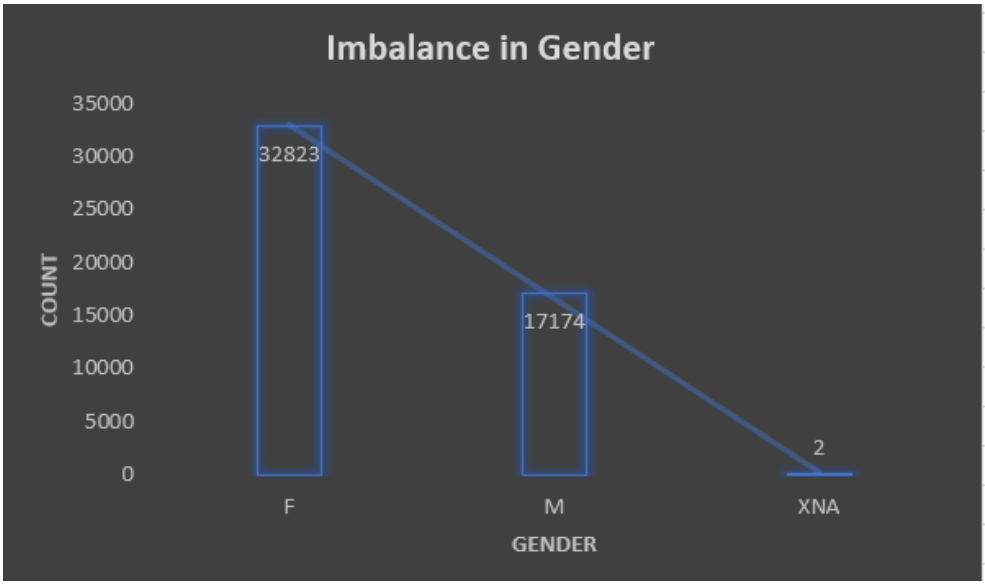


MIN	2052
Q1	16456.5
MEDIAN	24939
Q3	34596
MAX	258025.5
IQR	18139.5

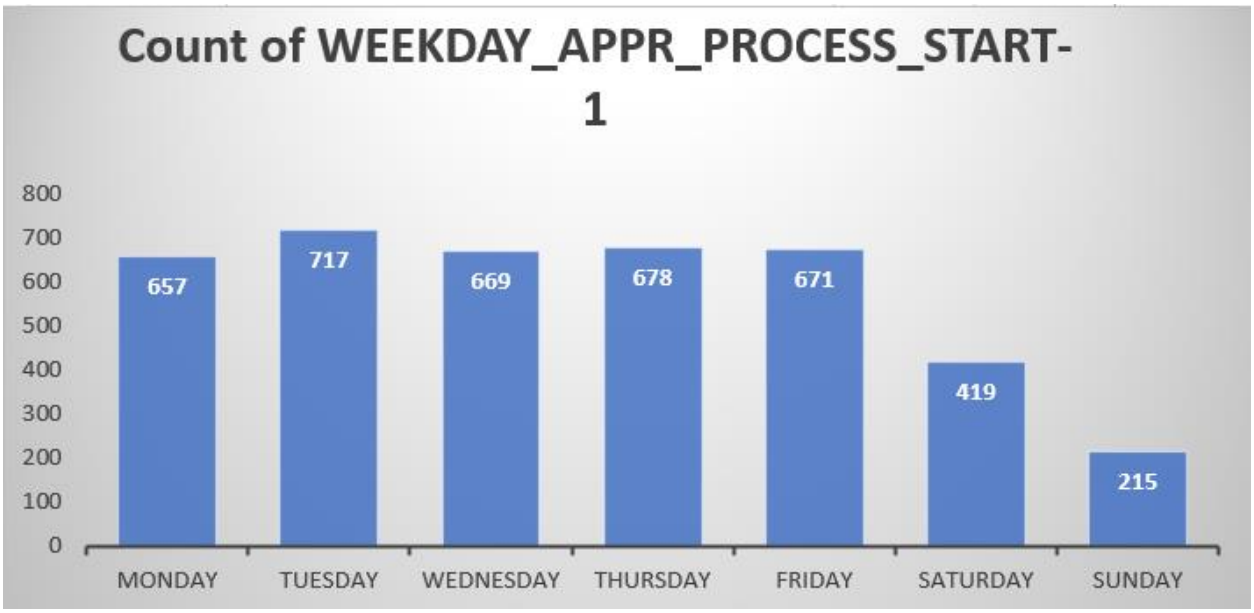
Analyzing Data Imbalance: Calculated the ratio of data imbalance using the Pivot table. Visualized the distribution of the target variable using pie charts and bar charts.

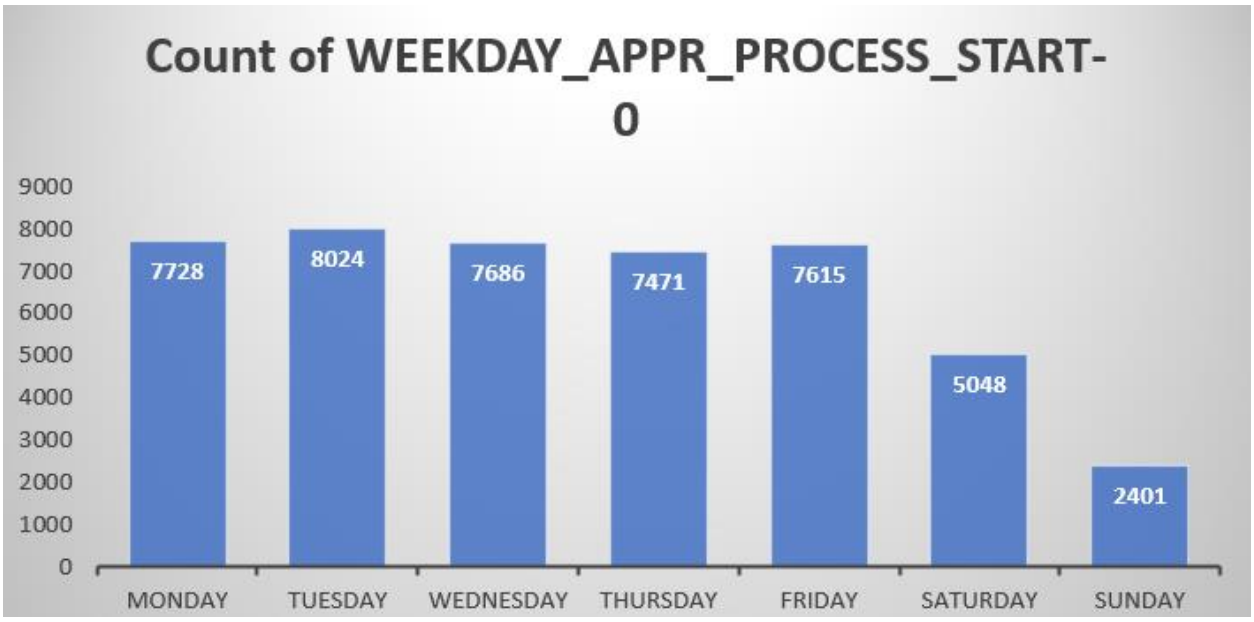




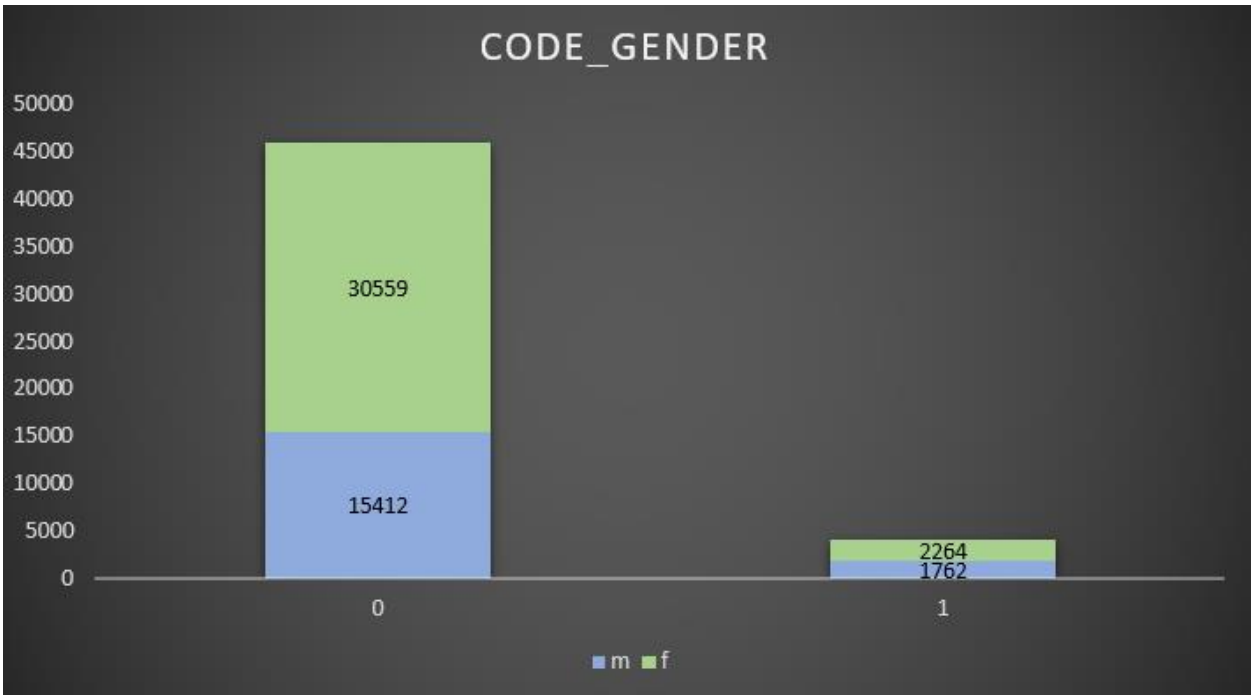


Univariate, Segmented Univariate, and Bivariate Analysis: Conducted descriptive analysis using Excel functions like COUNT, AVERAGE, and pivot tables. Compare variable distributions across scenarios using filters and sorting. Explored relationships between variables and the target variable using scatter plots and heat maps.

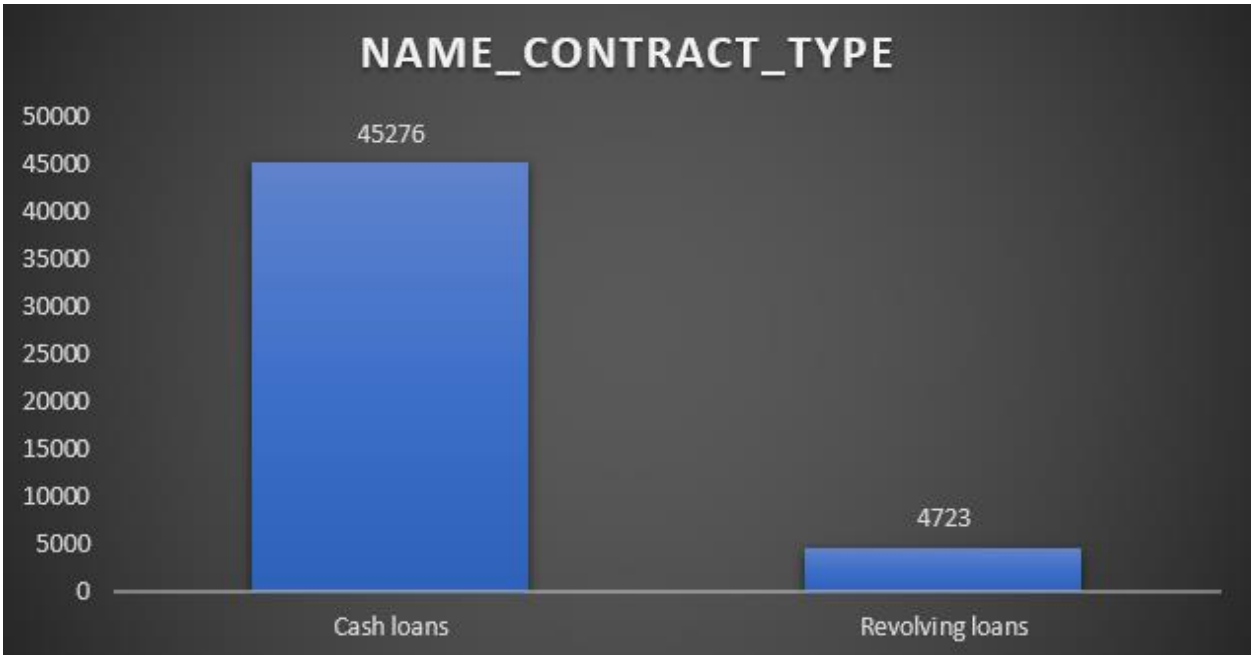


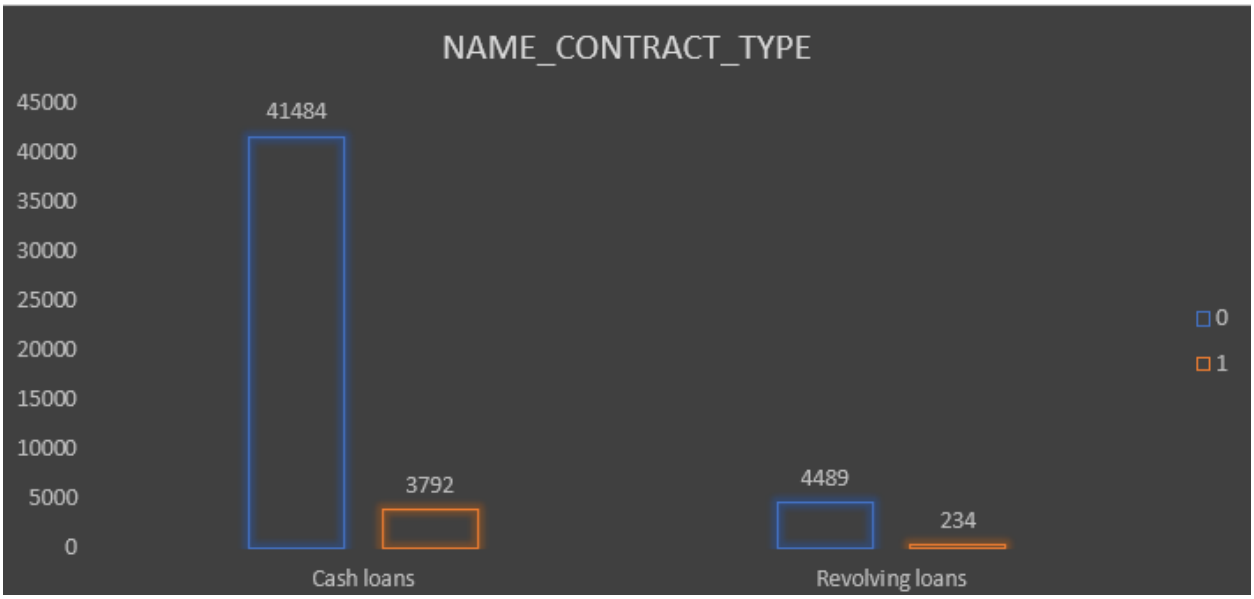


Univariate and Segmented Univariate Analysis:

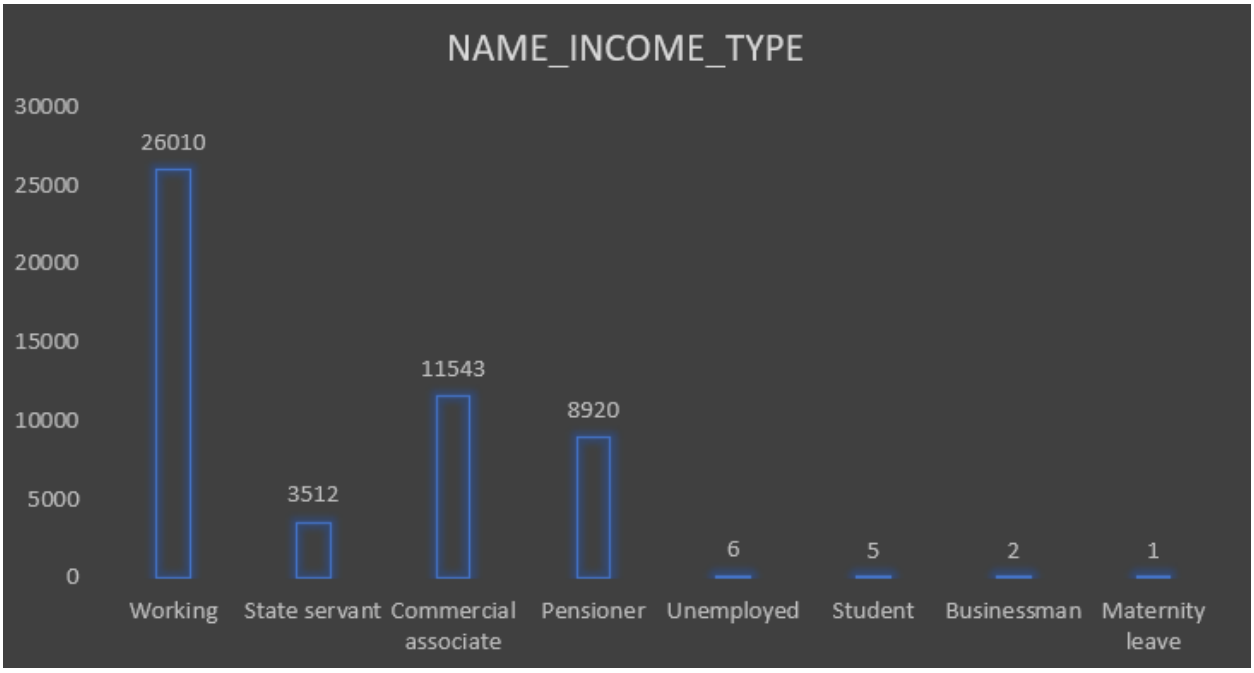


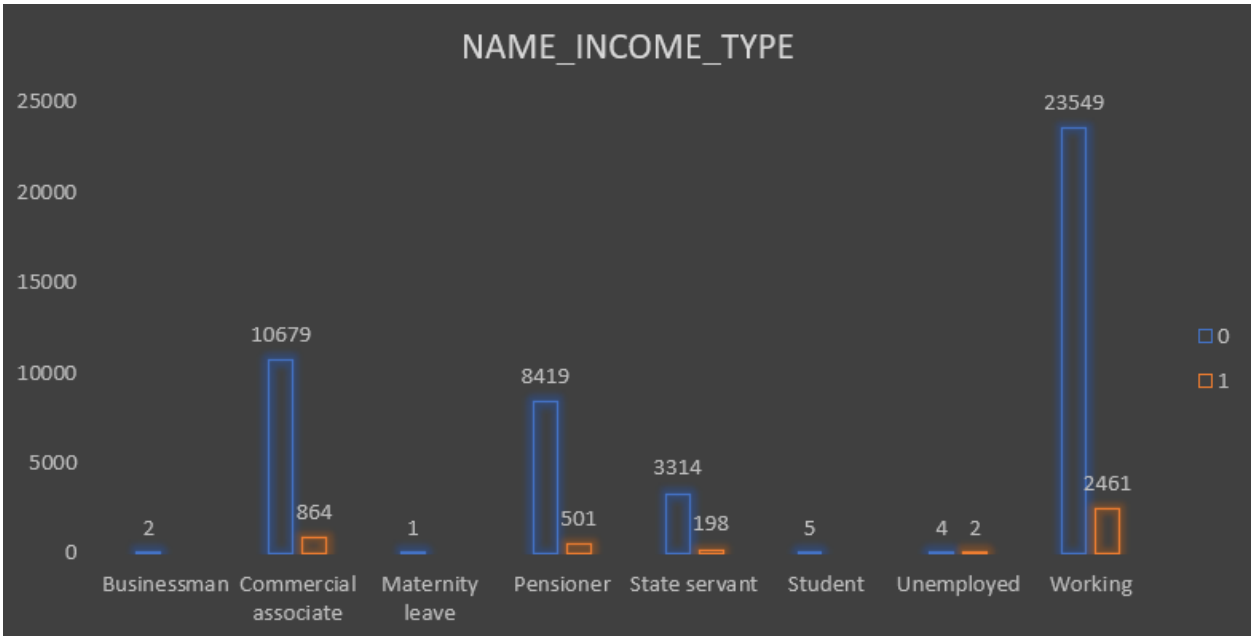
93.10% of females got their loans approved, while the percentage of males that got loans approved is 89.74%.





The total percentage of Cash loans that are rejected is 8.38% and the Total percentage of Revolving loans that are rejected is 4.96%.


NAME_INCOME_TYPE
30000
25000
20000
15000
10000
5000
0
26010
3512
11543
8920
6
5
2
1
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave


NAME_INCOME_TYPE
25000
20000
15000
10000
5000
0
2
10679
864
1
8419
501
3314
198
5
4
2
23549
2461
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
0
1

DISTRIBUTION OF TOTAL INCOME BASED ON GENDER

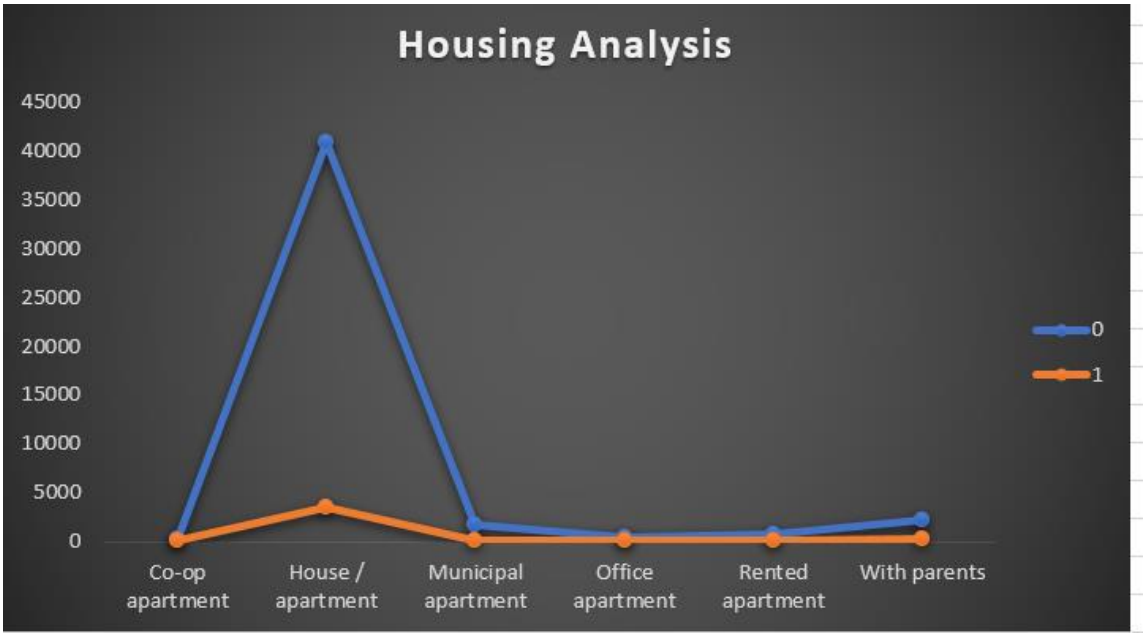
Count of CODE_GENDER	Column Labels		
AMT_INCOME_TOTAL	F	M	XNA
25650-125650	13607	3833	
125650-225650	15291	9534	2
225650-325650	2677	2449	
325650-425650	803	789	
425650-525650	252	317	
525650-625650	76	93	
625650-725650	76	89	
725650-825650	13	21	
825650-925650	11	24	
925650-1025650		2	
1025650-1125650	5	12	
1225650-1325650	1		
1325650-1425650	5	5	
1525650-1625650		1	
1725650-1825650	1	1	
1825650-1925650		1	
1925650-2025650	3		
2225650-2325650		2	
3525650-3625650	1		
3725650-3825650		1	
116925650-117025650	1		
Grand Total	**32823**	**17174**	**2**

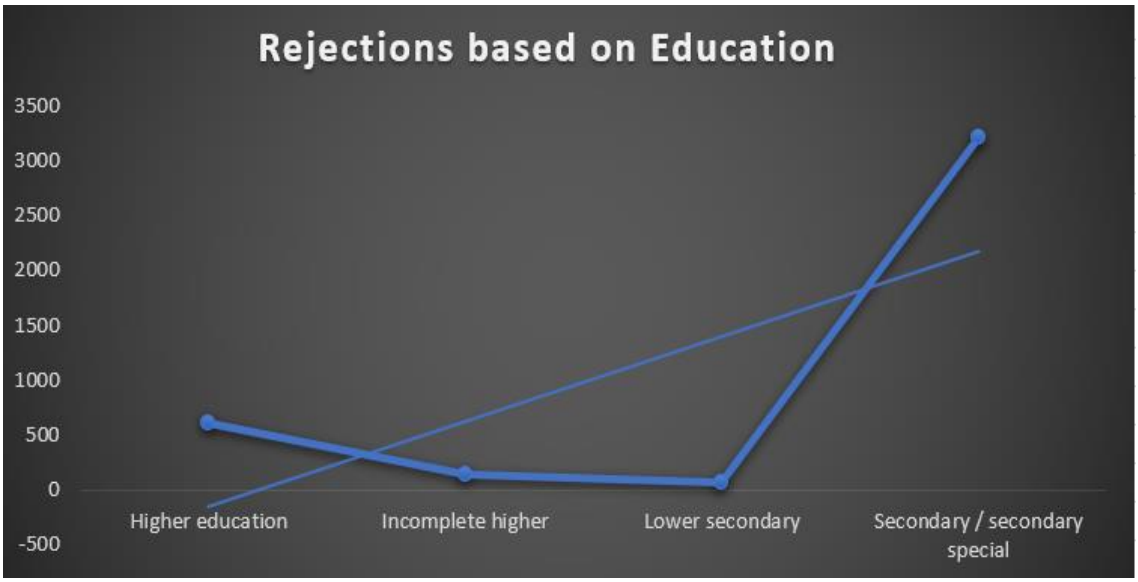
DISTRIBUTION OF TOTAL CREDIT BASED ON GENDER

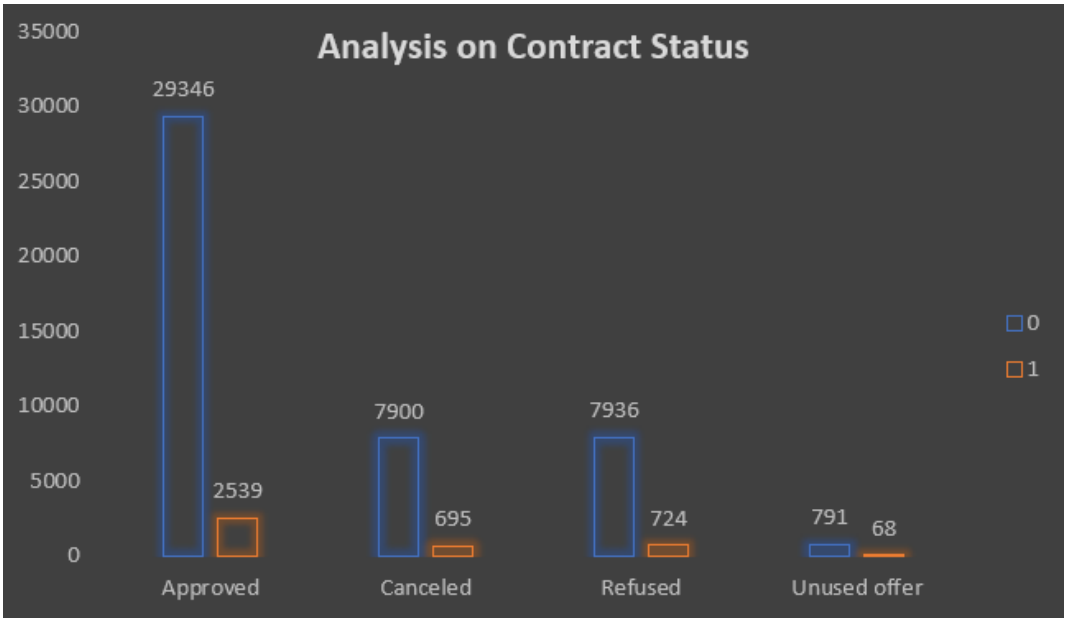
Count of CODE_GENDER	Column Label		
AMT_CREDIT	F	M	XNA
45000-145000	1982	836	
145000-245000	4214	2123	
245000-345000	5322	2750	1
345000-445000	1996	1030	1
445000-545000	4390	2460	
545000-645000	2617	1391	
645000-745000	2347	1200	
745000-845000	2661	1248	
845000-945000	1641	899	
945000-1045000	1178	639	
1045000-1145000	1394	758	
1145000-1245000	551	302	
1245000-1345000	882	479	
1345000-1445000	456	283	
1445000-1545000	285	181	
1545000-1645000	293	190	
1645000-1745000	107	81	
1745000-1845000	202	126	
1845000-1945000	66	33	
1945000-2045000	108	56	
2045000-2145000	28	12	
2145000-2245000	18	25	
2245000-2345000	40	39	

Bins | Sheet4 | Correlation-1 | Correlation-2 | Correlation-3 | Target 0 | Target 1 | Dashboard

At 33.3%, unemployment has the greatest rate of rejections experienced by an individual.







Identifying Top Correlations: Segmented the dataset based on different scenarios and calculated correlation coefficients using Excel functions like CORREL. Visualized correlations using correlation matrices or heatmaps.

TARGET-0

	AMT_INCOME_TOT	AMT_CREDIT	REGION_POPULATI	AMT_ANNUITY	Years_Birth	Years_Employed	Years_Registered	CNT_CHILDREN
AMT_INCOME_TOT	1	0.069315897	0.0298415	0.083009	-0.01582	-0.03151	-0.00995	0.009589
AMT_CREDIT	0.069315897	1	0.0951112	0.769498	0.059487	-0.06774	-0.00345	0.004972
REGION_POPULATI	0.029841469	0.095111221	1	0.115111	0.032471	-0.00416	0.059323	-0.02556
AMT_ANNUITY	0.083008508	0.769497849	0.1151113	1	-0.00751	-0.10871	-0.03322	0.026179
Years_Birth	-0.015823731	0.059486879	0.0324715	-0.00751	1	0.621496	0.333354	-0.32909
Years_Employed	-0.03151033	-0.06773941	-0.0041583	-0.10871	0.621496	1	0.209171	-0.24154
Years_Registered	-0.009952308	-0.003447803	0.0593235	-0.03322	0.333354	0.209171	1	-0.18122
CNT_CHILDREN	0.009588558	0.00497156	-0.0255557	0.026179	-0.32909	-0.24154	-0.18122	1
	AMT_INCOME	AMT_CREDIT	REGION_POI	AMT_ANN	Years_Birt	Years_Emp	Years_Reg	CNT_CHILDREN

TARGET-1

	AMT_INCOME_	AMT_CREDIT	REGION_POPU	AMT_ANNUITY	Years_Birt	Years_Employe	Years_Register	CNT_CHILDREN
AMT_INCOME_	1	0.069315897	0.029841469	0.083008508	-0.01582	-0.03151	-0.00995	0.009589
AMT_CREDIT	0.069315897	1	0.095111221	0.769497849	0.059487	-0.06774	-0.00345	0.004972
REGION_POPU	0.029841469	0.095111221	1	0.115111317	0.032471	-0.00416	0.059322	-0.02556
AMT_ANNUITY	0.083008508	0.769497849	0.115111317	1	-0.00751	-0.10871	-0.03322	0.026179
Years_Birth	-0.015823731	0.059486879	0.032471459	-0.007514338	1	0.621496	0.333354	-0.32909
Years_Employe	-0.03151033	-0.06773941	-0.004158337	-0.108709407	0.621496	1	0.209172	-0.24154
Years_Register	-0.009952379	-0.003448569	0.059322344	-0.033218936	0.333354	0.209172	1	-0.18122
CNT_CHILDREN	0.009588558	0.00497156	-0.025555665	0.026178735	-0.32909	-0.24154	-0.18122	1
	AMT_INCOME_	AMT_CREDIT	REGION_POPU	AMT_ANNUITY	Years_Birt	Years_Emp	Years_Reg	CNT_CHILDREN

Tech-Stack Used:

Microsoft Excel 2021 is used for this project and Excel was chosen for its familiarity, ease of use, and powerful data analysis capabilities, making it suitable for conducting EDA on the loan application dataset.

Insights:

- Identified missing data and outliers, ensuring data quality and reliability for analysis.

- Detected data imbalance, highlighting the need to address class imbalances in the dataset.
- Uncovered key factors influencing loan default through univariate, segmented univariate, and bivariate analysis.
- Discovered strong indicators of loan default through correlation analysis, providing valuable insights for risk assessment.

Result

Several important findings were uncovered by the loan dataset's exploratory data analysis (EDA). First of all, we verified the quality and dependability of our analysis by locating data imbalances, outliers, and missing information. Second, we discovered important elements impacting loan default, such as customer traits and loan features, using a variety of analytical methodologies, including univariate, segmented univariate, and bivariate analysis. Furthermore, by determining the strongest connections between the target variable and the variables, we gathered important information about potential risk indicators for loan default. These insights provide the finance organization the ability to make knowledgeable decisions regarding loan approval, reducing financial risk and enhancing business results.

Conclusion:

In conclusion, this project provided valuable hands-on experience in conducting EDA and analyzing real-world data to derive actionable insights. It underlined how crucial data analytics are to reducing risks and arriving at wise financial decisions. All things considered, the project demonstrated how to use data analysis methods in a real-world setting and emphasized the need to use data to achieve business results.