trainity

Bank Loan Case Study

Made by- Anurag John Phillips

ROADMAP

PROJECT DESCRIPTION

APPROACH

TECH-STACK USED

INSIGHTS

RESULTS

Project Description

In this project, we're diving into the world of "loan defaults" using Exploratory Data Analysis (EDA).

Our main goal is to give loans to good applicants and not turn them away.

A finance company has two big risks:

- They might say "no" to someone who actually deserves the loan, and that means losing business.
- They could give loan to someone who might not be able to pay it back, and that's a financial risk.

The data we're working with has two different situations:

- People who are having a hard time making their payments. This means they're paying late more often, especially in the beginning.
- People who are making their payments on time.

The objective of this project is to use EDA to comprehend how customer characteristics and loan-related factors impact the probability of loan default.

Approach

Understanding the Dataset

application_data.csv

1

2

Cleaning Dataset

Removing Duplicates and Columns, Rows which are not needed

Imputing Data

Using Median/ Mode

3

4

Analysing Dataset

Visualizating Dataset

5

6

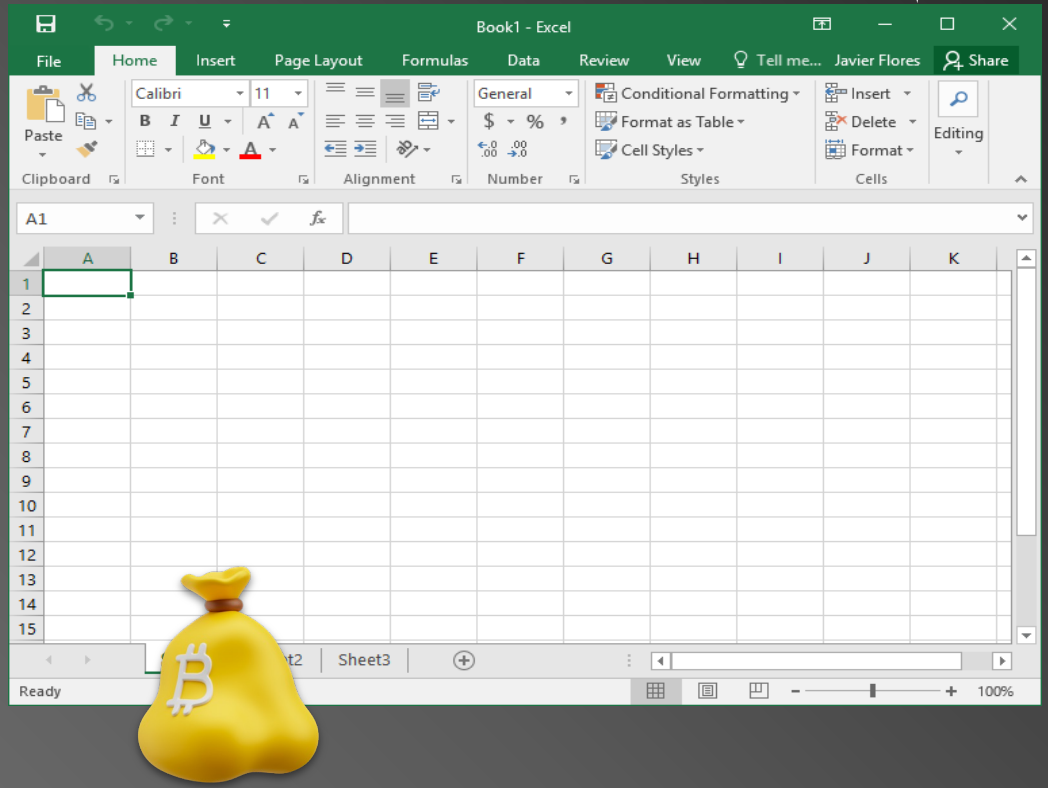
Gathering Insights

Working Files

https://drive.google.com/drive/folders/1r6bM200g8DAag94qym5__Dhdz8aVdh0_?usp=drive_link

Tech-Stack Used

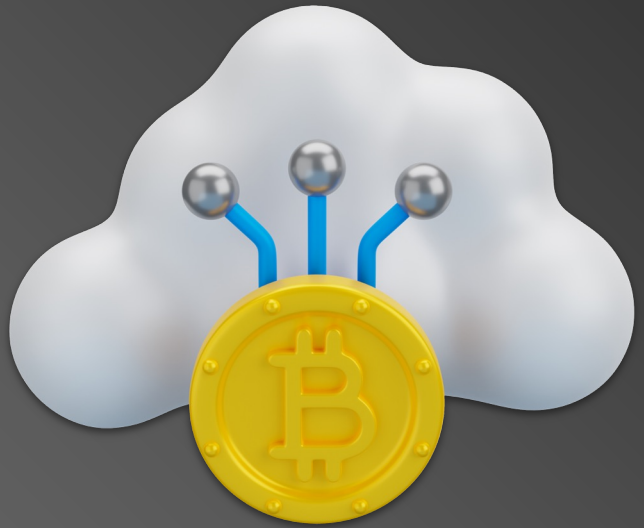
Microsoft Excel for Mac Version 16.74



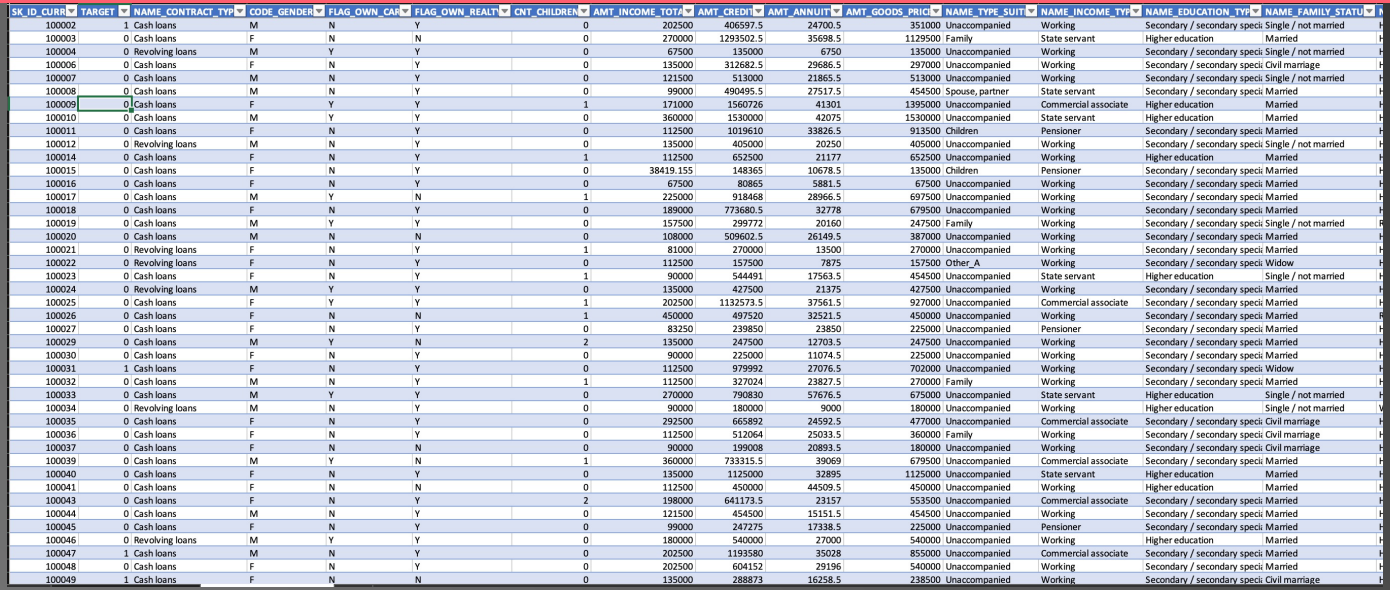
01

Identify Missing Data and Deal with it appropriately

Identify the missing data in the dataset and decide on an appropriate method to deal with it.



Original Dataset



SK_ID_CURR	TARGET	NAME_CONTRACT_TYP	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALT	CNT_CHILDREN	AMT_INCOME_TOTA	AMT_CREDI	AMT_ANNUIT	AMT_GOODS_PRIC	NAME_TYPE_SUIT	NAME_INCOME_TYP	NAME_EDUCATION_TYP	NAME_FAMILY_STATU	H
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary speci	Single / not married	H
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married	H
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary speci	Single / not married	H
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary speci	Civil marriage	H
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary speci	Single / not married	H
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary speci	Married	H
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married	H
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married	H
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary speci	Married	H
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary speci	Single / not married	H
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married	H
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary speci	Married	H
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary speci	Single / not married	R
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary speci	Married	H
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary speci	Married	H
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary speci	Widow	H
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married	H
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Secondary / secondary speci	Married	H
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary / secondary speci	Married	R
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary / secondary speci	Married	H
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working	Secondary / secondary speci	Married	H
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working	Secondary / secondary speci	Widow	H
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working	Secondary / secondary speci	Married	H
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State servant	Higher education	Single / not married	H
100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working	Higher education	Single / not married	V
100035	0	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied	Commercial associate	Secondary / secondary speci	Civil marriage	H
100036	0	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family	Working	Secondary / secondary speci	Civil marriage	H
100037	0	Cash loans	F	N	N	0	90000	199008	20893.5	180000	Unaccompanied	Working	Secondary / secondary speci	Civil marriage	H
100039	0	Cash loans	M	Y	N	1	360000	733315.5	39069	679500	Unaccompanied	Commercial associate	Secondary / secondary speci	Married	H
100040	0	Cash loans	F	N	Y	0	135000	1125000	32895	1125000	Unaccompanied	State servant	Higher education	Married	H
100041	0	Cash loans	F	N	N	0	112500	450000	44509.5	450000	Unaccompanied	Working	Higher education	Married	H
100043	0	Cash loans	F	N	Y	2	198000	641173.5	23157	553500	Unaccompanied	Commercial associate	Secondary / secondary speci	Married	H
100044	0	Cash loans	M	N	Y	0	121500	454500	15151.5	454500	Unaccompanied	Working	Secondary / secondary speci	Married	H
100045	0	Cash loans	F	N	Y	0	99000	247275	17338.5	225000	Unaccompanied	Pensioner	Secondary / secondary speci	Married	H
100046	0	Revolving loans	M	Y	Y	0	180000	540000	27000	540000	Unaccompanied	Working	Higher education	Married	H
100047	1	Cash loans	M	N	Y	0	202500	1193580	35028	855000	Unaccompanied	Commercial associate	Secondary / secondary speci	Married	H
100048	0	Cash loans	F	N	Y	0	202500	604152	29196	540000	Unaccompanied	Working	Secondary / secondary speci	Married	H
100049	1	Cash loans	F	N	N	0	135000	288873	16258.5	238500	Unaccompanied	Working	Secondary / secondary speci	Civil marriage	H

Before Cleaning- 122 Columns, some columns have 50,003 rows while others have 307478 Rows

For my analysis, to make the data uniform across all the columns I took first 50,003 rows and deleted the other rows, then made it into a table and then removed Duplicates which in under **Table** Tab

Calculated Percentage of Missing Values in each Column

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE
0	0	0	0	0	0	0	0	0	1	38	192	0
49999	49999	49999	49999	49999	49999	49999	49999	49999	49998	49961	49807	49999
0	0	0	0	0	0	0	0	0	0.00200008	0.076055326	0.385487984	0
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State servant
100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working
100035	0	Cash loans	F	N	Y	0	202500	665892	24592.5	427000	Unaccompanied	Commercial associate

ORGANIZATION_TYPE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AVG	ENTRANCES_AVG
0	28172	126	9944	25385	29199	24394	33239	34960	26651	25195
99 49999	21827	49873	40055	24614	20800	25605	16760	15039	23348	24804
0 0	129.0695011	0.25264171	24.82586444	103.1323637	140.3798077	95.27045499	198.323389	232.4622648	114.146822	101.5763587
0 Business Entity Type 3	0.083036967	0.262948593	0.13937578	0.0247	0.0369	0.9722	0.6192	0.0143	0	0.069
0 School	0.311267311	0.622245775		0.0959	0.0529	0.9851	0.796	0.0605	0.08	0.0345
0 Government		0.555912083	0.729566691							
0 Business Entity Type 3		0.65044169								
1 Religion		0.322738287								
0 Other		0.354224732	0.621226338							
0 Business Entity Type 3	0.774761413	0.723999852	0.492060094							
1 Other		0.714279286	0.54065445							
0 XNA	0.587334047	0.205747288	0.751723715							
0 Electricity		0.746643629								
0 Medicine	0.319760172	0.651862333	0.363945239							
0 XNA	0.72204445	0.555183162	0.652896552							
0 Business Entity Type 2	0.464831117	0.715041819	0.176652579	0.0825		0.9811			0	0.2069
0 Self-employed		0.566906613	0.77008707	0.1474	0.0973	0.9806	0.7348	0.0582	0.16	0.1379
0 Transport: type 2	0.721939769	0.642656205		0.3495	0.1335	0.9985	0.9796	0.1143	0.4	0.1724
0 Business Entity Type 2	0.115634337	0.346633981	0.678567689							
0 Government		0.23637784	0.062103038							
0 Construction		0.683513346								
0 Housing		0.706428403	0.556727426	0.0278	0.0617	0.9881	0.8368	0.0018	0	0.1034
0 Kindergarten		0.58661714	0.477649155							
0 Self-employed	0.565654882	0.113374513		0.0722	0.0801	0.9781	0.7008		0	0.1379
0 Trade: type 7	0.43770902	0.233766958	0.542445144							
0 Self-employed		0.557142972	0.358951229	0.0907	0.0795	0.9786	0.7076	0.012	0	0.2069
0 XNA		0.624304737	0.669056695	0.1443	0.0848	0.9876	0.83	0.1064	0.14	0.1207
1 Business Entity Type 3		0.786179309	0.565607981	0.1433	0.1455	0.9861	0.8096	0.0212	0	0.3103
0 Business Entity Type 3	0.561948409	0.651405637	0.461482391	0.0722	0.0147	0.9781	0.7008	0.001	0	0.1379
0 Business Entity Type 3		0.54847716	0.190705948	0.0165	0.0089	0.9732			0	0.069
1 Industry: type 11		0.541123702	0.659405532							
0 Military	0.600395905	0.68501099	0.524496446							
0 Business Entity Type 3	0.297913509	0.502779038		0.1505	0.0838	0.9831	0.7688	0.0188	0.16	0.1379
0 Business Entity Type 3		0.479987342	0.410102573	0.0124		0.9697			0	0.069

Yellow Row- Count of number of missing values

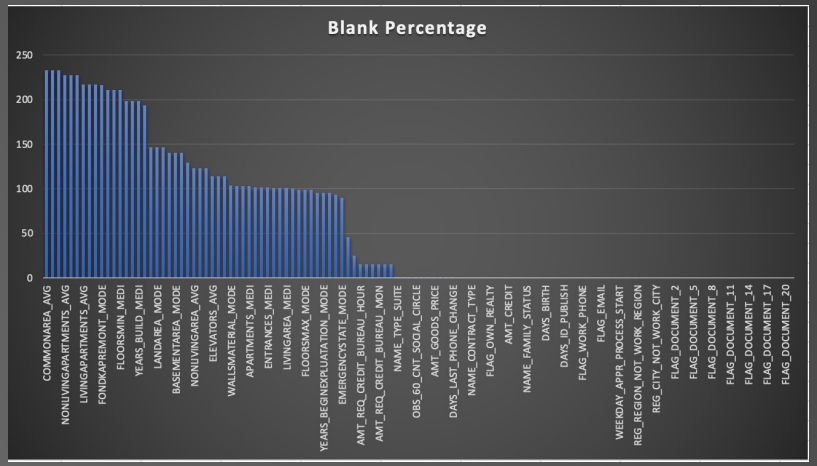
=COUNTBLANK(A5:A50003)

Red Row- Count the number of cells that have value in it

=COUNTA(A5:A50003)

Green Row- Calculated the percentage of missing values

=COUNTBLANK(A:A) / COUNTA(A:A) * 100



Before Imputation

	H	I	J	K	L	M	N	O	P	Q
1	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE
2	0	0	1	38	192	0	0	0	0	0
3	49999	49999	49998	49961	49807	49999	49999	49999	49999	49999
4	0	0	0.00200008	0.076059326	0.385487984	0	0	0	0	0
5	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary specia	Single / not married	House / apartment	0.018801
6	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married	House / apartment	0.003541
7	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary specia	Single / not married	House / apartment	0.010032
8	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary specia	Civil marriage	House / apartment	0.008019
9	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary specia	Single / not married	House / apartment	0.028663
10	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary specia	Married	House / apartment	0.035792
11	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married	House / apartment	0.035792
12	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married	House / apartment	0.003122
13	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary specia	Married	House / apartment	0.018634
14	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary specia	Single / not married	House / apartment	0.019689
15	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married	House / apartment	0.0228
16	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary specia	Married	House / apartment	0.015221
17	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.031329
18	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.016612
19	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.010006
20	157500	299772	20160	247500	Family	Working	Secondary / secondary specia	Single / not married	Rented apartment	0.020713
21	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.018634
22	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.010966
23	112500	157500	7875	157500	Other_A	Working	Secondary / secondary specia	Widow	House / apartment	0.04622
24	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married	House / apartment	0.015221
25	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.015221
26	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Secondary / secondary specia	Married	House / apartment	0.025164
27	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary / secondary specia	Married	Rented apartment	0.020713
28	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary / secondary specia	Married	House / apartment	0.006296
29	135000	247500	12703.5	247500	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.026392
30	90000	225000	11074.5	225000	Unaccompanied	Working	Secondary / secondary specia	Married	House / apartment	0.028663
31	112500	979992	27076.5	702000	Unaccompanied	Working	Secondary / secondary specia	Widow	House / apartment	0.018029
32	112500	327024	23827.5	270000	Family	Working	Secondary / secondary specia	Married	House / apartment	0.019101
33	270000	790830	57676.5	675000	Unaccompanied	State servant	Higher education	Single / not married	House / apartment	0.04622
34	90000	180000	9000	180000	Unaccompanied	Working	Higher education	Single / not married	With parents	0.030755
35	292500	665892	24592.5	477000	Unaccompanied	Commercial associate	Secondary / secondary specia	Civil marriage	House / apartment	0.025164

Numerical Columns- **Median**

Text Columns - **Mode**

After Imputation

SK_ID_CURR	TARGET	NAME_CO	CODE_GE	FLAG_OV	FLAG_OV	CNT_CHI	AMT_INC	AMT_CRE	AMT_ANI	AMT_GO	NAME_TYPE_SUITE	NAME_IN	NAME_EI	NAME_FA	NAME_H	REGION_	DAYS_BIF	DAYS_EN	DAYS_RE
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999	49999
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary /	Single / not	House / apar	0.018801	-9461	-637	-3648
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servan	Higher educa	Married	House / apar	0.003541	-16765	-1188	-1186
100004	0	Revolving loa	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary /	Single / not	House / apar	0.010032	-19046	-225	-4260
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary /	Civil marriag	House / apar	0.008019	-19005	-3039	-9833
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary /	Single / not	House / apar	0.028663	-19932	-3038	-4311
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servan	Secondary /	Married	House / apar	0.035792	-16941	-1588	-4970
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial	Higher educa	Married	House / apar	0.035792	-13778	-3130	-1213
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servan	Higher educa	Married	House / apar	0.003122	-18850	-449	-4597
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary /	Married	House / apar	0.018634	-20099	365243	-7427
100012	0	Revolving loa	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary /	Single / not	House / apar	0.019689	-14469	-2019	-14437
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher educa	Married	House / apar	0.0228	-10197	-679	-4427
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary /	Married	House / apar	0.015221	-20417	365243	-5246
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary /	Married	House / apar	0.031329	-13439	-2717	-311
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary /	Married	House / apar	0.016612	-14086	-3028	-643
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary /	Married	House / apar	0.010006	-14583	-203	-615
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary /	Single / not	Rented apart	0.020713	-8728	-1157	-3494
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied	Working	Secondary /	Married	House / apar	0.018634	-12931	-1317	-6392
100021	0	Revolving loa	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary /	Married	House / apar	0.010966	-9776	-191	-4143
100022	0	Revolving loa	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary /	Widow	House / apar	0.04622	-17718	-7804	-8751
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servan	Higher educa	Single / not	House / apar	0.015221	-11348	-2038	-1021
100024	0	Revolving loa	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary /	Married	House / apar	0.015221	-18252	-4286	-298
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial	Secondary /	Married	House / apar	0.025164	-14815	-1652	-2299
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary /	Married	Rented apart	0.020713	-11146	-4306	-114
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary /	Married	House / apar	0.006296	-24827	365243	-9012
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working	Secondary /	Married	House / apar	0.026392	-11286	-746	-108
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working	Secondary /	Married	House / apar	0.028663	-19334	-3494	-2419
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working	Secondary /	Widow	House / apar	0.018029	-18724	-2628	-6573
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working	Secondary /	Married	House / apar	0.019101	-15948	-1234	-5782
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State servan	Higher educa	Single / not	House / apar	0.04622	-9994	-1796	-4668
100034	0	Revolving loa	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working	Higher educa	Single / not	With parents	0.030755	-10341	-1010	-4799

Working Datasheet

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUIT	AMT_GOODS	NAME_TYPE_SUITE
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccompanied
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied
100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied
100035	0	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied
100036	0	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family
100037	0	Cash loans	F	N	N	0	90000	199008	20893.5	180000	Unaccompanied
100039	0	Cash loans	M	Y	N	1	360000	733315.5	39069	679500	Unaccompanied

After Cleaning- 73 Columns and 50,000 Rows

This completes the first task, here I have identified the missing data in the dataset and used Median and Mode Imputation to deal with it.

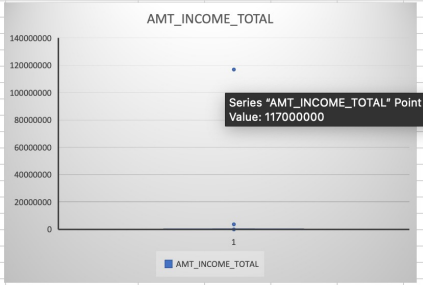
02

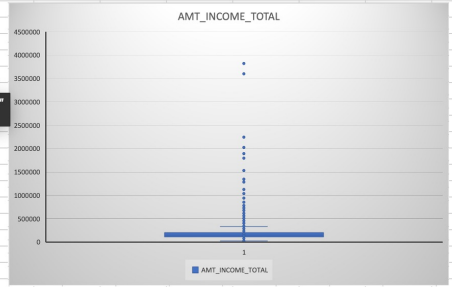
Identify Outliers in the Dataset

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

Outlier - Amt_Income_Total

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	AMT_INCOME_TOTAL														
2	202500		Quartile 1	112500	=QUARTILE(A2:A50000,1)										
3	270000														
4	67500		Quartile 3	202500	=QUARTILE(A2:A50000,3)										
5	135000														
6	121500		Inter Quartile Range IQR	90000	=D4-D2										
7	99000														
8	171000		Lower Threshold	-22500	=D2-(1.5*D6)										
9	360000														
10	112500		Upper Threshold	337500	=D4+(1.5*D6)										
11	135000														
12	112500														
13	38419.155														
14	67500														
15	225000														
16	189000														
17	157500														
18	108000														
19	81000														
20	112500														
21	90000														
22	135000														
23	202500														
24	450000														
25	83250														
26	135000														
27	90000														
28	112500														
29	112500														
30	270000														
31	90000														
32	292500														
33	112500														





QUARTILE 1 (Q1): Q1 is 112,500. This means that 25% of the "AMT_INCOME_TOTAL" values are less than or equal to 112,500.

QUARTILE 3 (Q3): Q3 is 202,500. This means that 75% of the "AMT_INCOME_TOTAL" values are less than or equal to 202,500.

INTERQUARTILE RANGE (IQR): IQR is 90,000, calculated as Q3 - Q1. This means that the middle 50% of the "AMT_INCOME_TOTAL" values fall within the range of 112,500 to 202,500.

LOWER THRESHOLD: Lower Threshold is -22,500. This means that if there are any negative values or extremely low income values in the dataset (less than -22,500), they would be flagged as potential outliers.

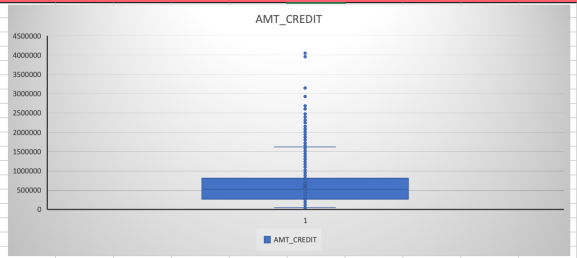
UPPER THRESHOLD: Upper Threshold is 337,500. This means that if there are any extremely high income values in the dataset (greater than 337,500), they would be flagged as potential outliers.

The Lower Threshold and Upper Threshold are calculated based on the interquartile range (IQR) and data points that fall below the Lower Threshold or above the Upper Threshold are considered potential outliers and may require further investigation to determine if they are valid data points or data entry errors.

Outliers

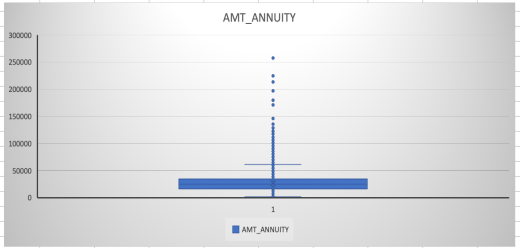
AMT_CREDIT
406597.5
1293502.5
135000
312682.5
513000
490495.5
1560726
1530000
1019610
405000
652500
148365
80865
918468
773680.5
299772
509602.5
270000
157500
544491
427500
1132573.5
497520
239850
247500
225000
979992
327024
790830

Q1	270000
Q3	808650
IQR	538650
Lower Threshold	-537975
Upper Threshold	1616625



AMT_ANNUITY
24700.5
35698.5
6750
29686.5
21865.5
27517.5
41301
42075
33826.5
20250
21177
10678.5
5881.5
28966.5
32778
20160
26149.5
13500
7875
17563.5
21375
37561.5
32521.5
23850
12703.5
11074.5
27076.5

Quartile 1	16456.5
Quartile 3	34596
Inter Quartile Range IQR	18139.5
Lower Threshold	-10752.75
Upper Threshold	61805.25

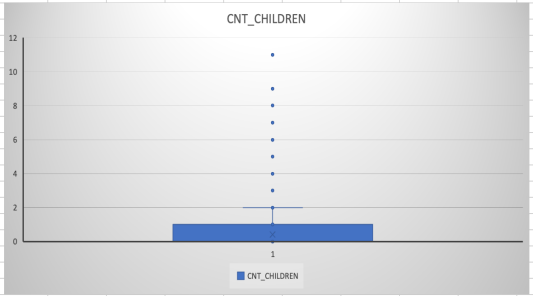


AMT_GOODS_PRICE
351000
1129500
135000
297000
513000
454500
1395000
1530000
913500
405000
652500
135000
67500
697500
679500
247500
387000
270000
157500
454500
427500
927000
450000
225000
247500
225000
702000

Quartile 1	238500
Quartile 3	679500
Inter Quartile Range IQR	441000
Lower Threshold	-423000
Upper Threshold	1341000

CNT_CHILDREN
0
0
0
0
0
1
0
0
0
1
0
0
1
0
0
0
1
0
1
0
1
1
0

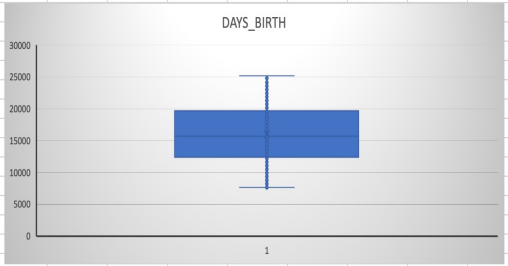
Quartile 1	0
Quartile 3	1
Inter Quartile Range IQR	1
Lower Threshold	-1.5
Upper Threshold	2.5



Outliers

DAYS_BIRTH
9461
16765
19046
19005
19932
16941
13778
18850
20099
14469
10197
20417
13439
14086
14583
8728
12931
9776
17718
11348

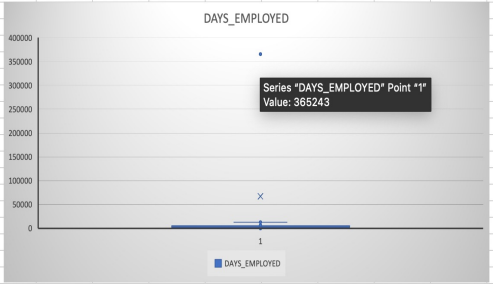
Quartile 1	12378.5
Quartile 3	19644
Inter Quartile Range IQR	7265.5
Lower Threshold	1480.25
Upper Threshold	30542.25



DAYS_EMPLOYED
637
1188
225
3039
3038
1588
3130
449
365243
2019
679
365243
2717
3028
203
1157
1317
191
7804
2038
4286
1652
4306
365243

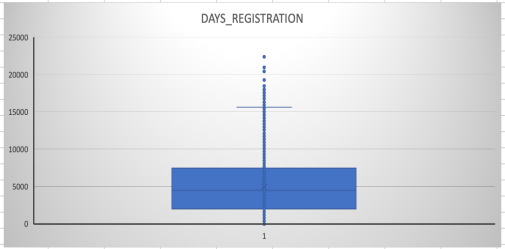
Quartile 1	933
Quartile 3	5718
Inter Quartile Range IQR	4785
Lower Threshold	-6244.5
Upper Threshold	12895.5

365243	365
Employed Years	1000.665753



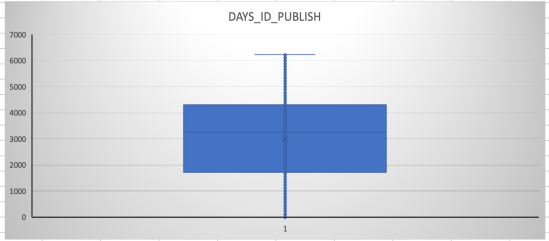
DAYS_REGISTRATION
3648
1186
4260
9833
4311
4970
1213
4597
7427
14437
4427
5246
311
643
615
3494
6392
4143
8751
1021
298
2299
114
9012

Quartile 1	1998
Quartile 3	7463.5
Inter Quartile Range IQR	5465.5
Lower Threshold	-6200.25
Upper Threshold	15661.75



DAYS_ID_PUBLISH
2120
291
2531
2437
3458
477
619
2379
3514
3992
738
2512
3227
4911
2056
1368
3866
2427
1259
3964
1800
2299
2518
3684
3729

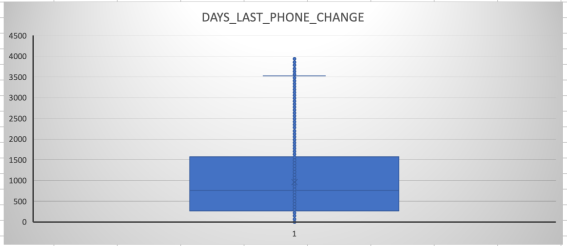
Quartile 1	1722
Quartile 3	4297
Inter Quartile Range IQR	2575
Lower Threshold	-2140.5
Upper Threshold	8159.5



Outliers

DAYS_LAST_PHONE_CHANGE
1134
828
815
617
1106
2536
1562
1070
0
1673
844
2396
2370
4
188
925
3
2811
239
1850
296
0
468
795

Quartile 1	270
Quartile 3	1573
Inter Quartile Range IQR	1303
Lower Threshold	-1684.5
Upper Threshold	3527.5



We observe here that most of the columns have outliers in them except Days_Birth and Days_Id_Publish

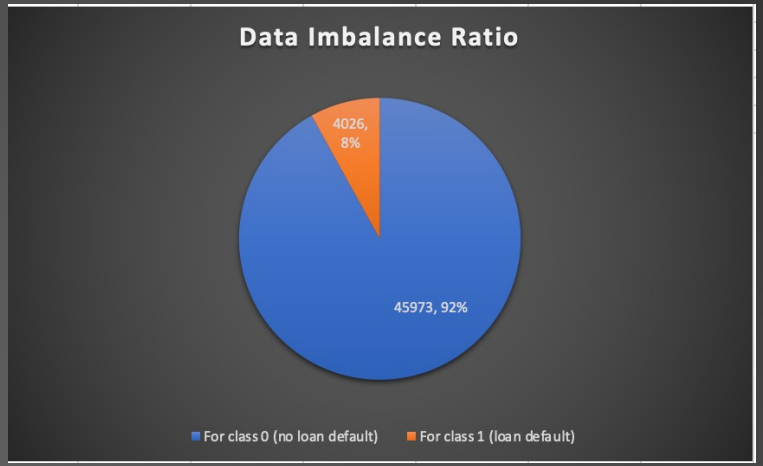
03

Analyse Data Imbalance

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

Data Imbalance Ratio

For class 0 (no loan default)	45973 =COUNTIF(A:A, 0)
For class 1 (loan default)	4026 =COUNTIF(A:A, 1)
Data Imbalance Ratio	**8.75%** =C4/C3



The data imbalance ratio of approximately 8.75% indicates that the 'loan default class' is about 8.75% of the 'no loan default class', highlighting the data imbalance in the target variable.

04

Perform Univariate, Segmented Univariate, and Bivariate Analysis

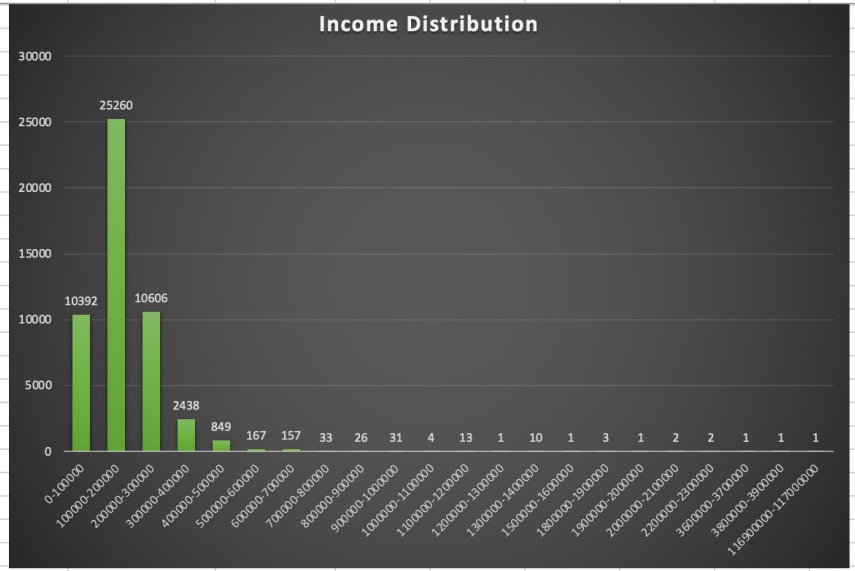
Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features

Univariate Analysis

Univariate analysis is a statistical technique used to analyse and summarize data from a single variable. It focuses on understanding the distribution and characteristics of that particular variable in isolation, without considering any other variables. The primary goal of univariate analysis is to describe and gain insights into the behaviour and patterns of the individual variable.

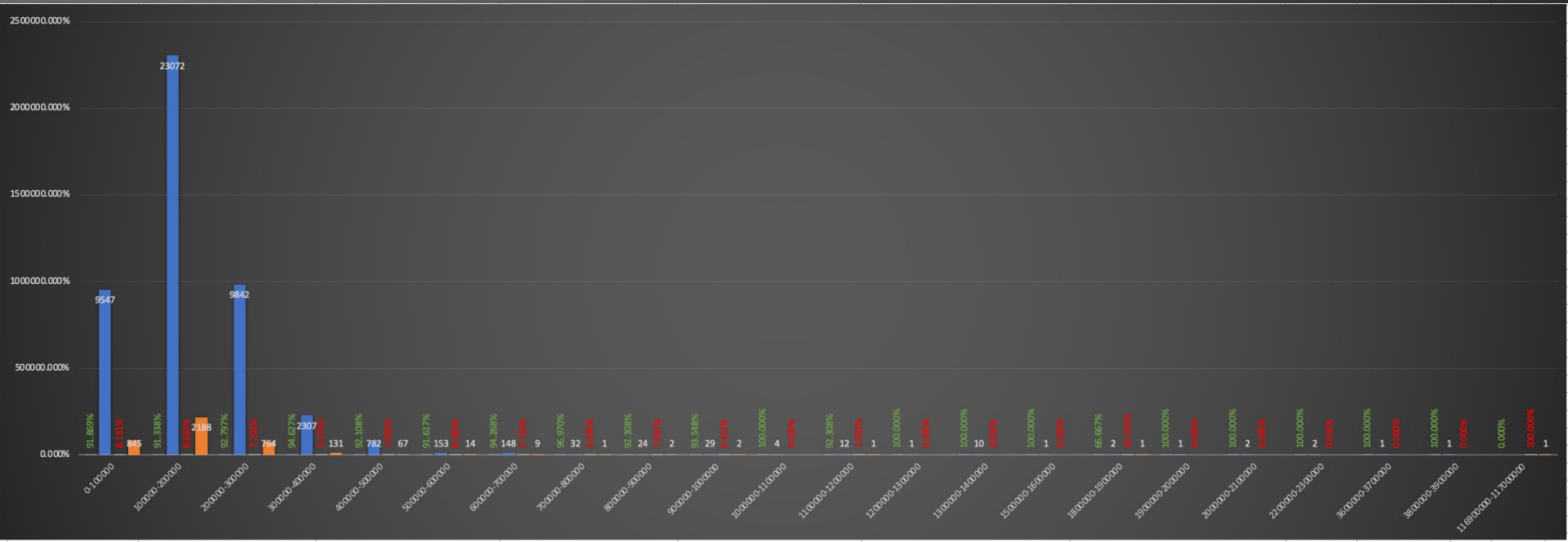
AMT_INCOME_TOTAL
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31500
31531.5
31815
31860
32139
32400
32400
32400
32400
32850
33300

Income Range	Count of AMT_INCOME_TOTAL
0-100000	10392
100000-200000	25260
200000-300000	10606
300000-400000	2438
400000-500000	849
500000-600000	167
600000-700000	157
700000-800000	33
800000-900000	26
900000-1000000	31
1000000-1100000	4
1100000-1200000	13
1200000-1300000	1
1300000-1400000	10
1500000-1600000	1
1800000-1900000	3
1900000-2000000	1
2000000-2100000	2
2200000-2300000	2
3600000-3700000	1
3800000-3900000	1
116900000-117000000	1
Grand Total	**49999**



Segmented Univariate Analysis

Amt_Income_Total - Target



Generally as Income is Increasing the loan default percentage is decreasing.

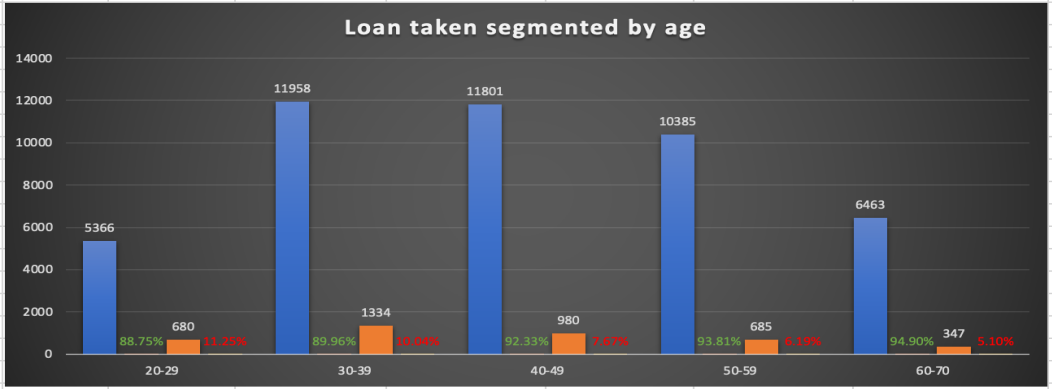
Segmented Univariate Analysis

Age - Target

B2 =ROUNDUP(ABS(C2/365),0)

	A	B	C
1	TARGET	Age	Year_BIRTH
2	1	26	-9461
3	0	46	-16765
4	0	53	-19046
5	0	53	-19005
6	0	55	-19932
7	0	47	-16941
8	0	38	-13778
9	0	52	-18850
10	0	56	-20099
11	0	40	-14469
12	0	28	-10197
13	0	56	-20417
14	0	37	-13439
15	0	39	-14086
16	0	40	-14583
17	0	24	-8728
18	0	36	-12931
19	0	27	-9776
20	0	49	-17718
21	0	32	-11348
22	0	51	-18252
23	0	41	-14815
24	0	31	-11146
25	0	69	-24827
26	0	31	-11286
27	0	53	-19334
28	1	52	-18724
29	0	44	-15948
30	0	28	-9994

Column Labe	0		1		Total Count of TAR(	Total Count of TARGET2
Row Labe	Count of TARGI	Count of TARGET2	Count of TARGET	Count of TARGET2		
20-29	5366	88.75%	680	11.25%	6046	100.00%
30-39	11958	89.96%	1334	10.04%	13292	100.00%
40-49	11801	92.33%	980	7.67%	12781	100.00%
50-59	10385	93.81%	685	6.19%	11070	100.00%
60-70	6463	94.90%	347	5.10%	6810	100.00%
Grand Total	**45973**	**91.95%**	**4026**	**8.05%**	**49999**	**100.00%**



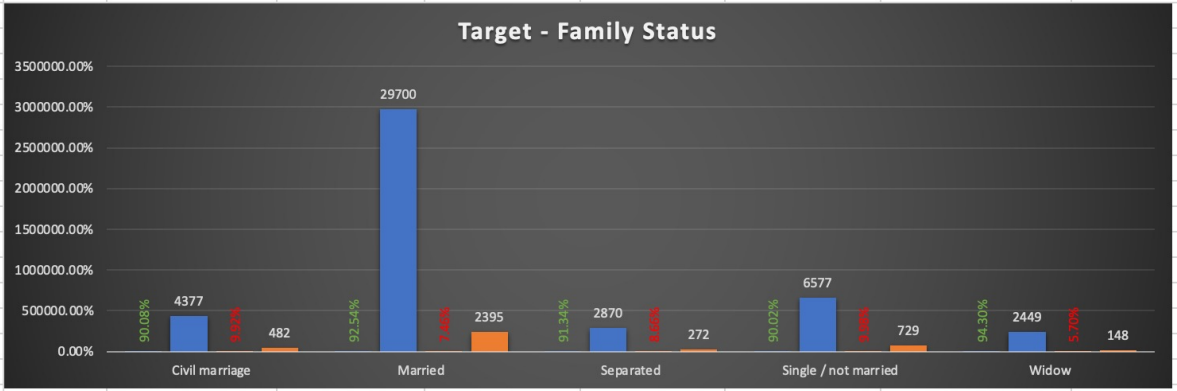
Generally as Age is Increasing the loan default percentage is decreasing.

Segmented Univariate Analysis

Family Status - Target

TARGET	NAME_FAMILY_STATUS
1	Single / not married
0	Married
0	Single / not married
0	Civil marriage
0	Single / not married
0	Married
0	Married
0	Married
0	Married
0	Single / not married
0	Married
0	Married
0	Married
0	Married
0	Married
0	Single / not married
0	Married
0	Married
0	Widow
0	Single / not married
0	Married
0	Married
0	Married
0	Married
0	Married
0	Married
1	Widow

	Column Labels					
	0		1		Total Count of TAR(	Total Count of TARGET2
Row Labe	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
Civil marriag	90.08%	4377	9.92%	482	100.00%	4859
Married	92.54%	29700	7.46%	2395	100.00%	32095
Separated	91.34%	2870	8.66%	272	100.00%	3142
Single / not	90.02%	6577	9.98%	729	100.00%	7306
Widow	94.30%	2449	5.70%	148	100.00%	2597
Grand Total	**91.95%**	**45973**	**8.05%**	**4026**	**100.00%**	**49999**



People who are Married or Widow have lower loan default percentage.

Segmented Univariate Analysis

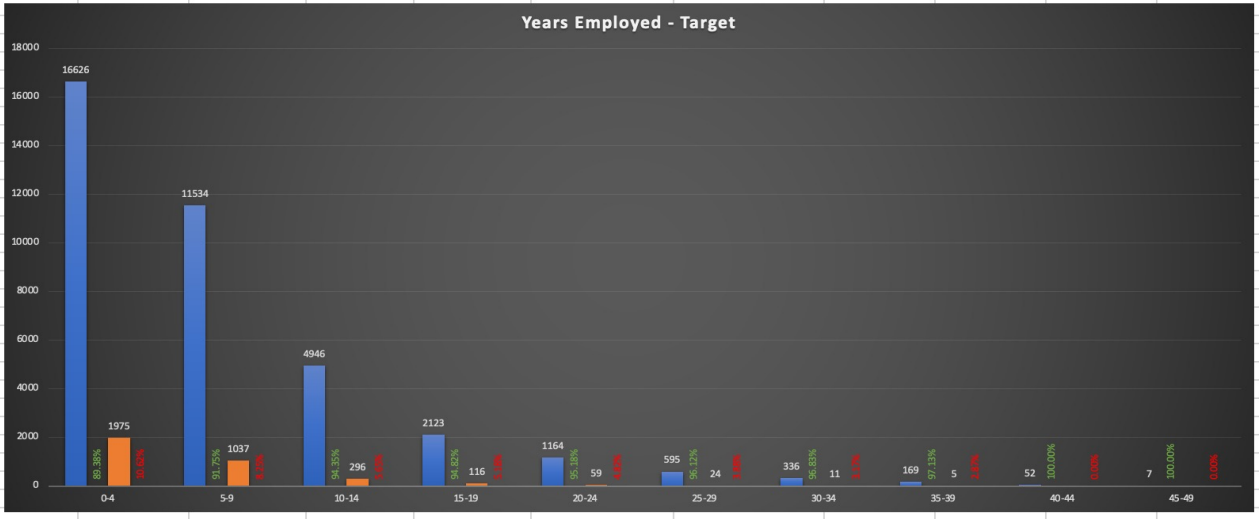
Years Employed - Target

B2 =ROUNDUP(ABS(C2/365),0)

TARGET	Years Employed	DAYS_EMPLOY
1	2	-637
0	4	-1188
0	1	-225
0	9	-3039
0	9	-3038
0	5	-1588
0	9	-3130
0	2	-449
0	1001	365243
0	6	-2019
0	2	-679
0	1001	365243
0	8	-2717
0	9	-3028
0	1	-203
0	4	-1157
0	4	-1317
0	1	-191
0	22	-7804
0	6	-2038
0	12	-4286
0	5	-1652
0	12	-4306
0	1001	365243
0	3	-746
0	10	-3494
1	8	-2628
0	4	-1234
0	5	-1796
0	3	-1010
0	8	-2668
0	4	-1104
0	13	-4404
0	6	-2060
0	13	-4585
0	4	-1275
0	3	-768
0	4	-1288
0	1001	365243
0	5	-1761
1	4	-1262
0	2	-475
1	10	-3597
0	1001	365243

Column Labels	0		1		Total Count of TARGET	Total Count of TARGET2
Row Labels	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
0-4	16626	89.38%	1975	10.62%	18601	100.00%
5-9	11534	91.75%	1037	8.25%	12571	100.00%
10-14	4946	94.35%	296	5.65%	5242	100.00%
15-19	2123	94.82%	116	5.18%	2239	100.00%
20-24	1164	95.18%	59	4.82%	1223	100.00%
25-29	595	96.12%	24	3.88%	619	100.00%
30-34	336	96.83%	11	3.17%	347	100.00%
35-39	169	97.13%	5	2.87%	174	100.00%
40-44	52	100.00%		0.00%	52	100.00%
45-49	7	100.00%		0.00%	7	100.00%
Grand Total	37552	91.42%	3523	8.58%	41075	100.00%

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
0-4	16626	1975	18601
5-9	11534	1037	12571
10-14	4946	296	5242
15-19	2123	116	2239
20-24	1164	59	1223
25-29	595	24	619
30-34	336	11	347
35-39	169	5	174
40-44	52		52
45-49	7		7
1000-1004	8420	503	8923
Grand Total	45972	4026	49998



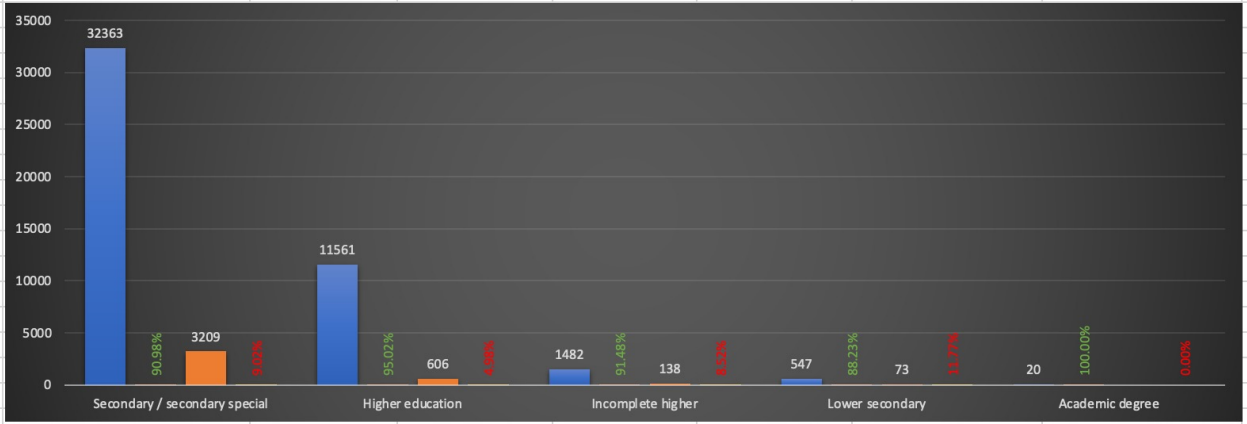
People who are employed for more years have lower loan default percentage.

Segmented Univariate Analysis

Education_Type - Target

TARGET	NAME_EDUCATION_TYPE
1	Secondary / secondary special
0	Higher education
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Higher education
0	Higher education
0	Secondary / secondary special
0	Secondary / secondary special
0	Higher education
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Higher education
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
0	Secondary / secondary special
1	Secondary / secondary special
0	Secondary / secondary special

Column Labels	0		1		Total Count of TARGET	Total Count of TARGET2
Row Labels	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
Secondary / secondary special	32363	90.98%	3209	9.02%	35572	100.00%
Higher education	11561	95.02%	606	4.98%	12167	100.00%
Incomplete higher	1482	91.48%	138	8.52%	1620	100.00%
Lower secondary	547	88.23%	73	11.77%	620	100.00%
Academic degree	20	100.00%		0.00%	20	100.00%
Grand Total	**45973**	**91.95%**	**4026**	**8.05%**	**49999**	**100.00%**



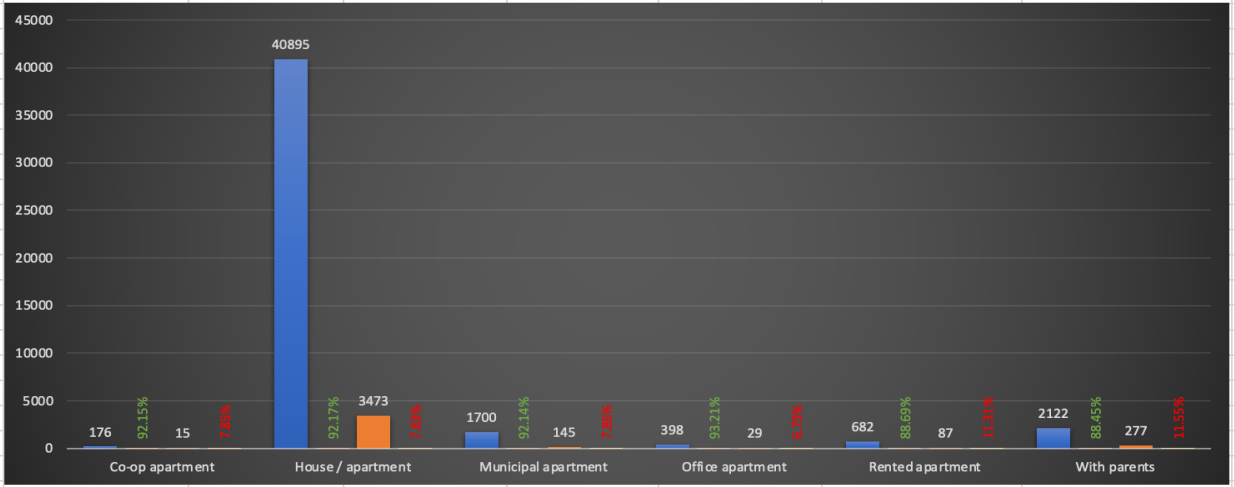
As education level increases, loan default percentage decreases. People who have higher education tend to default less and less on their loan.

Segmented Univariate Analysis

Housing_Type - Target

TARGET	NAME_HOUSING_TYPE
1	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	Rented apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	Rented apartment
0	House / apartment
0	House / apartment
0	House / apartment
1	House / apartment
0	House / apartment
0	House / apartment
0	With parents
0	House / apartment

	Column Labels					
	0		1		Total Count of TARGET	Total Count of TARGET2
Row Labels	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
Co-op apartment	176	92.15%	15	7.85%	191	100.00%
House / apartment	40895	92.17%	3473	7.83%	44368	100.00%
Municipal apartment	1700	92.14%	145	7.86%	1845	100.00%
Office apartment	398	93.21%	29	6.79%	427	100.00%
Rented apartment	682	88.69%	87	11.31%	769	100.00%
With parents	2122	88.45%	277	11.55%	2399	100.00%
Grand Total	**45973**	**91.95%**	**4026**	**8.05%**	**49999**	**100.00%**

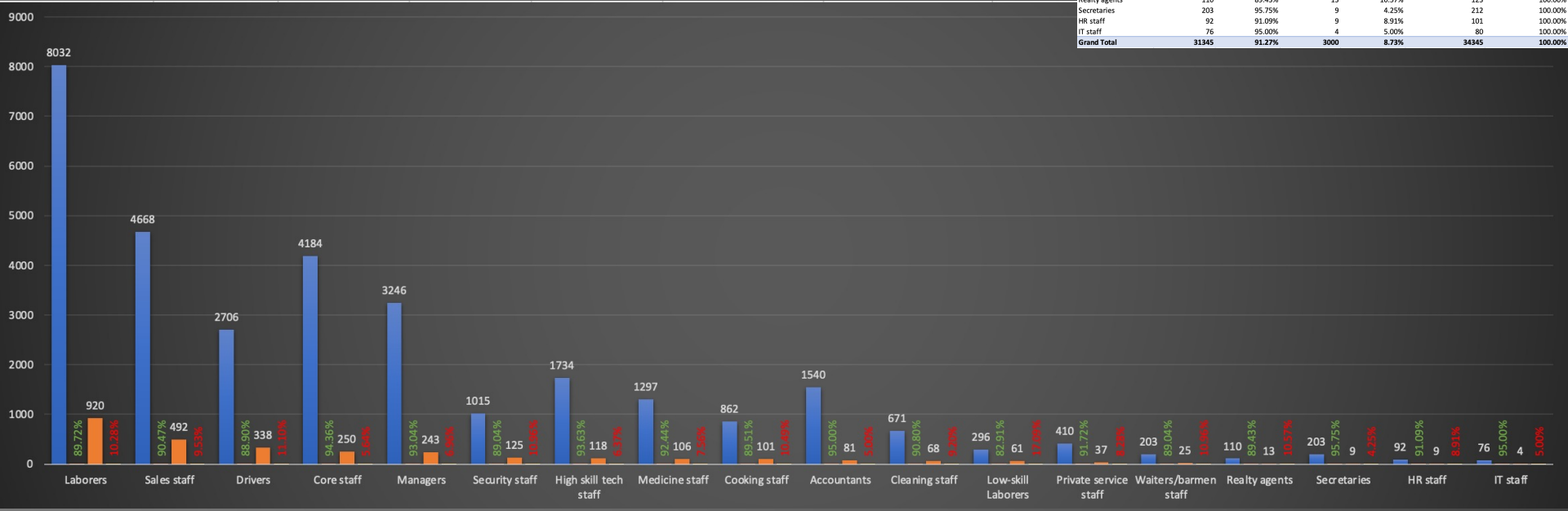


People living in a Rented Apartment or with Parents have higher loan default percentage as compared to people living in apartments or houses.

Segmented Univariate Analysis

Occupation Type- Target

Column Labels	0		1		Total Count of TARGET	Total Count of TARGET2
Row Labels	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
Laborers	8032	89.72%	920	10.28%	8952	100.00%
Sales staff	4668	90.47%	492	9.53%	5160	100.00%
Drivers	2706	88.90%	338	11.10%	3044	100.00%
Core staff	4184	94.36%	250	5.64%	4434	100.00%
Managers	3246	93.04%	243	6.96%	3489	100.00%
Security staff	1015	89.04%	125	10.96%	1140	100.00%
High skill tech staff	1734	93.63%	118	6.37%	1852	100.00%
Medicine staff	1297	92.44%	106	7.56%	1403	100.00%
Cooking staff	862	89.51%	101	10.49%	963	100.00%
Accountants	1540	95.00%	81	5.00%	1621	100.00%
Cleaning staff	671	90.80%	68	9.20%	739	100.00%
Low-skill Laborers	296	82.91%	61	17.09%	357	100.00%
Private service staff	410	91.72%	37	8.28%	447	100.00%
Waiters/barmen staff	203	89.04%	25	10.96%	228	100.00%
Realty agents	110	89.43%	13	10.57%	123	100.00%
Secretaries	203	95.75%	9	4.25%	212	100.00%
HR staff	92	91.09%	9	8.91%	101	100.00%
IT staff	76	95.00%	4	5.00%	80	100.00%
Grand Total	31345	91.27%	3000	8.73%	34345	100.00%



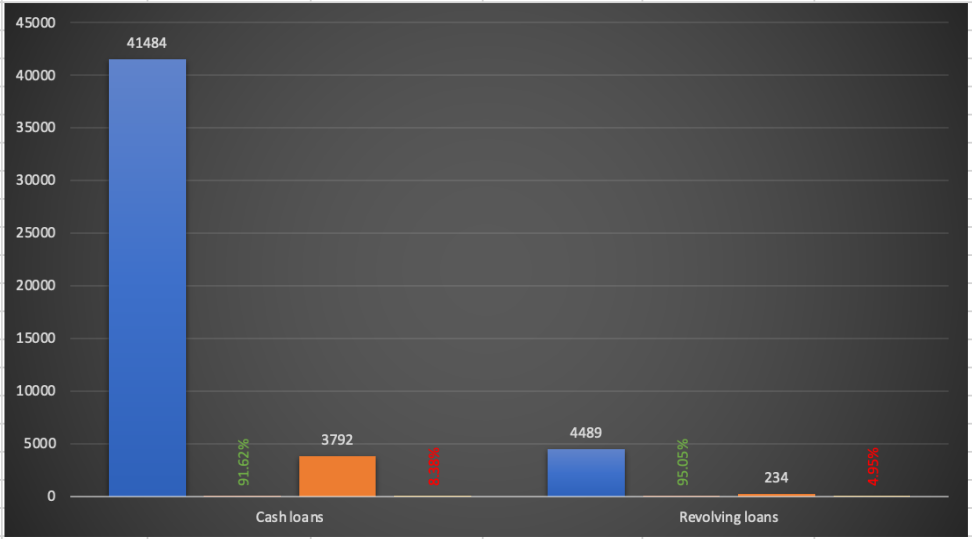
People working as Low Skill Labourers, Drivers, Labourers, Cleaning Staff, Waiters/Barmen Staff etc tend to have higher loan default percentage as compared to people working as IT Staff, Secretaries, Accountants, Core Staff and Managers.

Segmented Univariate Analysis

Contract_Type - Target

TARGET	NAME_CONTRACT_TYPE
1	Cash loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Revolving loans
0	Revolving loans
0	Cash loans
0	Revolving loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
0	Cash loans
1	Cash loans
0	Cash loans
0	Cash loans

Column Labels	0		1		Total Count of TARGET	Total Count of TARGET2
Row Labels	Count of TARGET	Count of TARGET2	Count of TARGET	Count of TARGET2		
Cash loans	41484	91.62%	3792	8.38%	45276	100.00%
Revolving loans	4489	95.05%	234	4.95%	4723	100.00%
Grand Total	**45973**	**91.95%**	**4026**	**8.05%**	**49999**	**100.00%**



Although most people took cash loans but people who took revolving loans have lower loan default percentage.

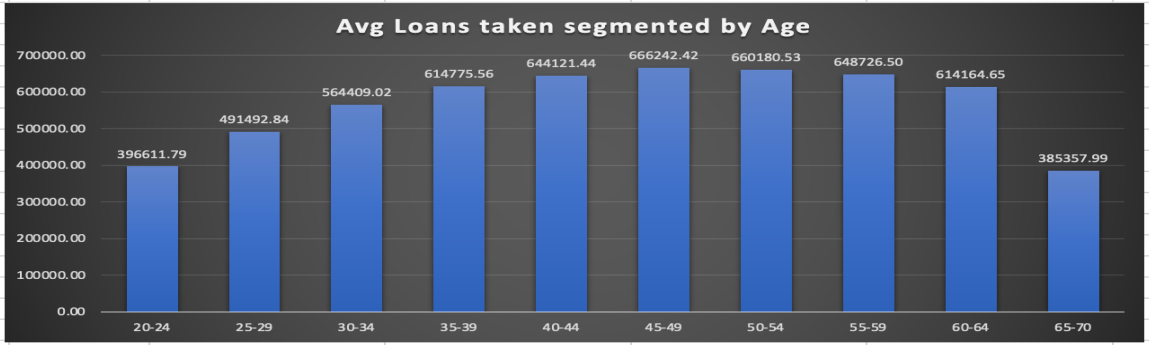
Bivariate Analysis

Age-Amt_Credit

Age	AMT_CREDIT	TARGET
26	406597.5	1
46	1293502.5	0
53	135000	0
53	312682.5	0
55	513000	0
47	490495.5	0
38	1560726	0
52	1530000	0
56	1019610	0
40	405000	0
28	652500	0
56	148365	0
37	80865	0
39	918468	0
40	773680.5	0
24	299772	0
36	509602.5	0
27	270000	0
49	157500	0
32	544491	0
51	427500	0
41	1132573.5	0
31	497520	0
69	239850	0
31	247500	0
53	225000	0
52	979992	1
44	327024	0
28	790830	0
29	180000	0
42	665892	0
31	512064	0
36	199008	0

TARGET	(All)

Row Labels	Average of AMT_CREDIT
20-24	396611.79
25-29	491492.84
30-34	564409.02
35-39	614775.56
40-44	644121.44
45-49	666242.42
50-54	660180.53
55-59	648726.50
60-64	614164.65
65-70	385357.99
Grand Total	599700.5815

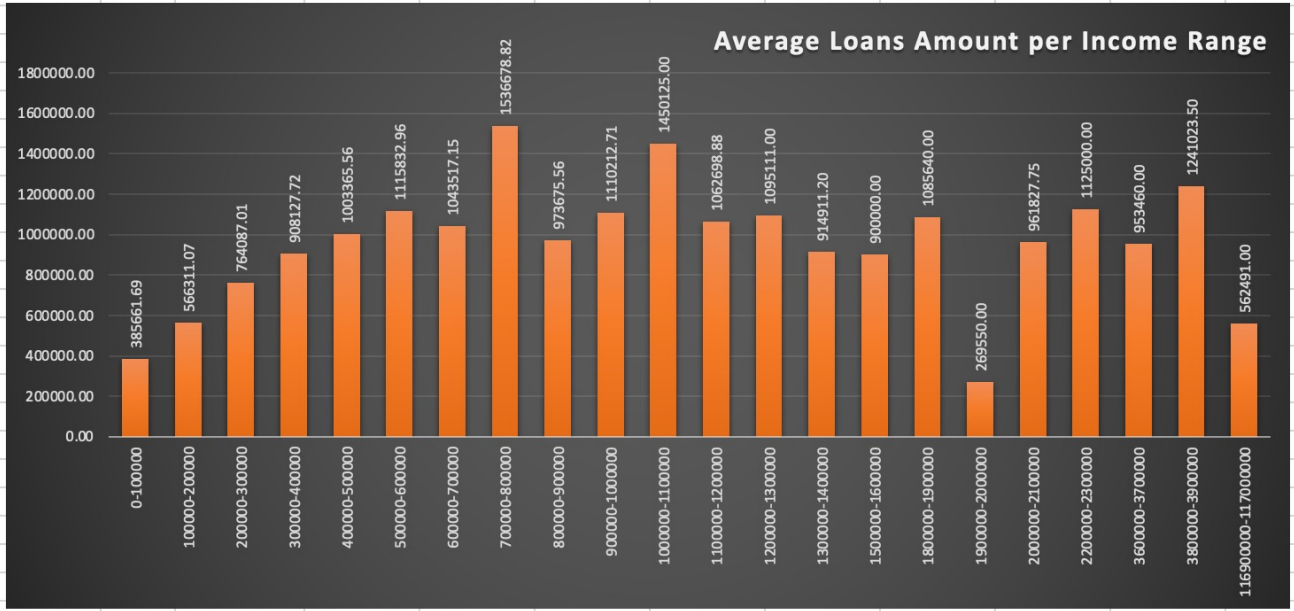


People in the age between 35 to 65 tend to take higher amount loan

Bivariate Analysis

Income - Amt_Credit

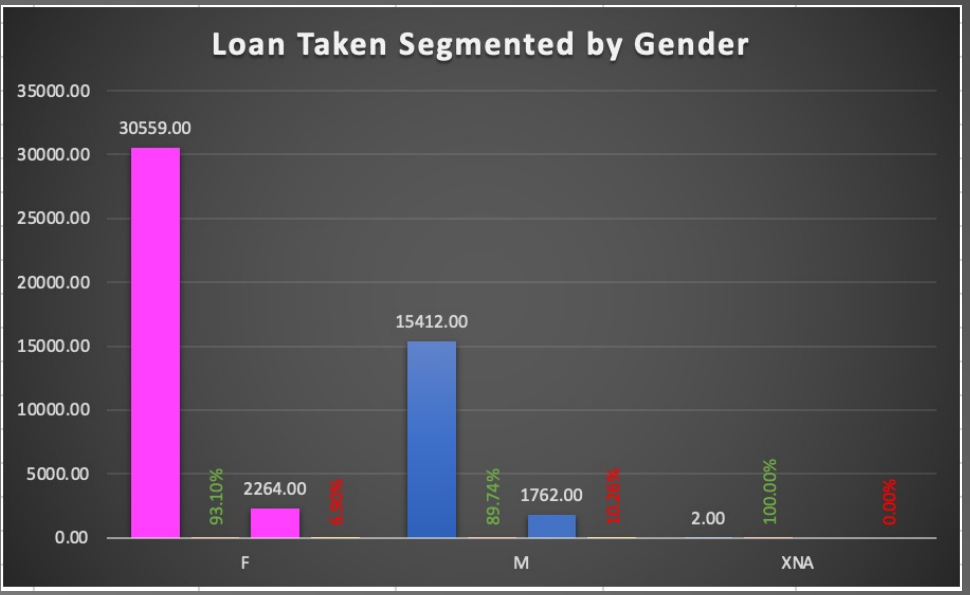
Row Labels	Average of AMT_CREDIT
0-100000	385661.69
100000-200000	566311.07
200000-300000	764087.01
300000-400000	908127.72
400000-500000	1003365.56
500000-600000	1115832.96
600000-700000	1043517.15
700000-800000	1536678.82
800000-900000	973675.56
900000-1000000	1110212.71
1000000-1100000	1450125.00
1100000-1200000	1062698.88
1200000-1300000	1095111.00
1300000-1400000	914911.20
1500000-1600000	900000.00
1800000-1900000	1085640.00
1900000-2000000	269550.00
2000000-2100000	961827.75
2200000-2300000	1125000.00
3600000-3700000	953460.00
3800000-3900000	1241023.50
116900000-117000000	562491.00
Grand Total	**599700.5815**

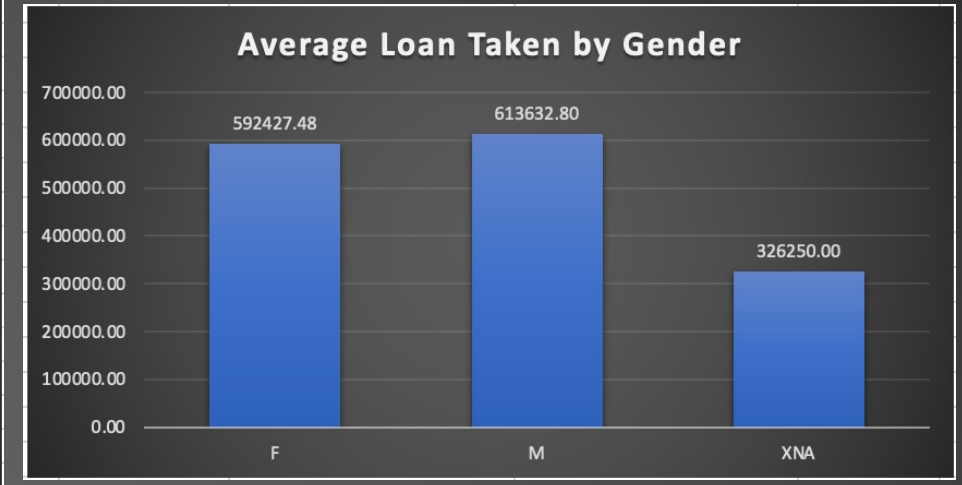


Highest average Loan amount taken is by People whose Income is in 700K-800K.
Here we cant say that as their income increases the amount of loan also increases

Segmented and Bivariate Analysis

Gender-Amt_Credit



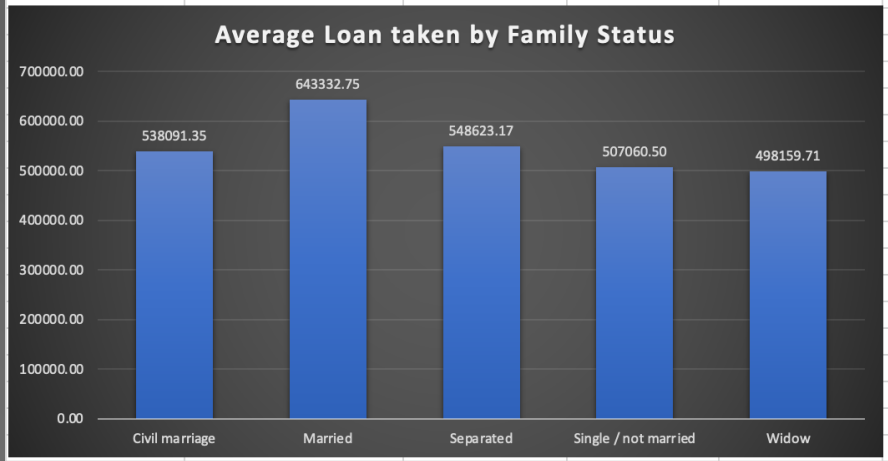


More Females took loan compared to Males and their Loan Default Percentage is Less.
Although fewer Males take loan but their average Loan Amount is more.
This data shows that Females should be preferred over males when giving the loan.

Bivariate Analysis

Family_Status - Amt_Credit

TARGET	(All)
Row Labels	**Average of AMT_CREDIT**
Civil marriage	538091.35
Married	643332.75
Separated	548623.17
Single / not married	507060.50
Widow	498159.71
Grand Total	**599700.58**



Married People tend to take higher amount loan compared to others

05

Identify Top Correlations for Different Scenarios

Segment the dataset based on different scenarios and identify the top correlations for each segmented data using Excel functions.

Correlation

Excel WorkBook

	A	B	C	D	E	F	G	H	I	J	K	L
1	TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_23
2	1	202500	406597.5	24700.5	351000	0.018801	-9461	-637	-3648	-2120	0.262948593	0.13937578
3	0	270000	1293502.5	35698.5	1129500	0.003541	-16765	-1188	-1186	-291	0.622245775	0.53527625
4	0	67500	135000	6750	135000	0.010032	-19046	-225	-4260	-2531	0.555912083	0.729566691
5	0	135000	312682.5	29686.5	297000	0.008019	-19005	-3039	-9833	-2437	0.65044169	0.53527625
6	0	121500	513000	21865.5	513000	0.028663	-19932	-3038	-4311	-3458	0.322738287	0.53527625
7	0	99000	490495.5	27517.5	454500	0.035792	-16941	-1588	-4970	-477	0.354224732	0.621226338
8	0	171000	1560726	41301	1395000	0.035792	-13778	-3130	-1213	-619	0.723999852	0.492060094
9	0	360000	1530000	42075	1530000	0.003122	-18850	-449	-4597	-2379	0.714279286	0.54065445
10	0	112500	1019610	33826.5	913500	0.018634	-20099	365243	-7427	-3514	0.205747288	0.751723715
11	0	135000	405000	20250	405000	0.019689	-14469	-2019	-14437	-3992	0.746643629	0.53527625
12	0	112500	652500	21177	652500	0.0228	-10197	-679	-4427	-738	0.651862333	0.363945239
13	0	38419.155	148365	10678.5	135000	0.015221	-20417	365243	-5246	-2512	0.555183162	0.652896552
14	0	67500	80865	5881.5	67500	0.031329	-13439	-2717	-311	-3227	0.715041819	0.176652579
15	0	225000	918468	28966.5	697500	0.016612	-14086	-3028	-643	-4911	0.566906613	0.77008707
16	0	189000	773680.5	32778	679500	0.010006	-14583	-203	-615	-2056	0.642656205	0.53527625
17	0	157500	299772	20160	247500	0.020713	-8728	-1157	-3494	-1368	0.346633981	0.678567689
18	0	108000	509602.5	26149.5	387000	0.018634	-12931	-1317	-6392	-3866	0.23637784	0.062103038
19	0	81000	270000	13500	270000	0.010966	-9776	-191	-4143	-2427	0.683513346	0.53527625
20	0	112500	157500	7875	157500	0.04622	-17718	-7804	-8751	-1259	0.706428403	0.556727426
21	0	90000	544491	17563.5	454500	0.015221	-11348	-2038	-1021	-3964	0.58661714	0.477649155
22	0	135000	427500	21375	427500	0.015221	-18252	-4286	-298	-1800	0.113374513	0.53527625
23	0	202500	1132573.5	37561.5	927000	0.025164	-14815	-1652	-2299	-2299	0.233766958	0.542445144
24	0	450000	497520	32521.5	450000	0.020713	-11146	-4306	-114	-2518	0.457142972	0.358951229
25	0	83250	239850	23850	225000	0.006296	-24827	365243	-9012	-3684	0.624304737	0.669056695
26	0	135000	247500	12703.5	247500	0.026392	-11286	-746	-108	-3729	0.786179309	0.565607981
27	0	90000	225000	11074.5	225000	0.028663	-19334	-3494	-2419	-2893	0.651405637	0.461482391
28	1	112500	979992	27076.5	702000	0.018029	-18724	-2628	-6573	-1827	0.54847716	0.190705948
29	0	112500	327024	23827.5	270000	0.019101	-15948	-1234	-5782	-3153	0.541123702	0.659405532
30	0	270000	790830	57676.5	675000	0.04622	-9994	-1796	-4668	-2661	0.68501099	0.524496446
31	0	90000	180000	9000	180000	0.030755	-10341	-1010	-4799	-3015	0.502779038	0.53527625
32	0	292500	665892	24592.5	477000	0.025164	-15280	-2668	-5266	-3787	0.479987342	[illegible]0102573
33	0	112500	512064	25033.5	360000	0.008575	-11144	-1104	-7846	-2904	0.627300401	[illegible]3527625
34	0	90000	199008	20893.5	180000	0.010032	-12974	-4404	-7123	-4464	0.559466792	[illegible]137231
35	0	360000	733315.5	39069	679500	0.015221	-11694	-2060	-3557	-3557	0.321744896	[illegible]3556

Corr- Main Sheet | Correlation-Target 0 | Target 0 | Correlation-Target 1 | Target 1 | +

Correlation

Target – 0 (Loan Payees)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_23	DAYS_LAST_PHONE_CHANGE	CNT_CHILDREN
AMT_INCOME_TOTAL	1												
AMT_CREDIT	0.377965752	1											
AMT_ANNUITY	0.451135103	0.770772908	1										
AMT_GOODS_PRICE	0.384675888	0.987235417	0.77613285	1									
REGION_POPULATION_RELATIVE	0.181941261	0.095539444	0.117279106	0.09896058	1								
DAYS_BIRTH	0.073769425	-0.051084182	0.009910977	-0.04869335	-0.030435419	1							
DAYS_EMPLOYED	-0.162702675	-0.077367219	-0.113005115	-0.075172	-0.006610653	-0.61528998	1						
DAYS_REGISTRATION	0.06893375	0.008053758	0.034609001	0.01129847	-0.058501361	0.33502805	-0.204370881	1					
DAYS_ID_PUBLISH	0.032286356	-0.008290189	0.009427021	-0.00930065	-0.002236288	0.27007331	-0.27222439	0.103548902	1				
EXT_SOURCE_2	0.15617334	0.136258463	0.130022952	0.14329839	0.201089847	-0.0803327	-0.034096314	-0.053917105	-0.040900368	1			
EXT_SOURCE_23	-0.073654638	0.028831122	0.018547427	0.03105218	-0.013832665	-0.17888625	0.098594722	-0.099901082	-0.111944396	0.068846786	1		
DAYS_LAST_PHONE_CHANGE	-0.049497956	-0.071203379	-0.064450897	-0.07428594	-0.044127291	0.07253958	0.032951867	0.047780168	0.085063175	-0.184718008	-0.059902129	1	
CNT_CHILDREN	0.036319722	0.005705458	0.026384162	0.00155316	-0.024912809	0.33587627	-0.243591518	0.183072478	-0.032537221	-0.013466269	-0.039280661	-0.004822698	1

Top 10 Correlations for Loan Payees

AMT_GOODS_PRICE	AMT_CREDIT	0.987235417
AMT_GOODS_PRICE	AMT_ANNUITY	0.77613285
AMT_ANNUITY	AMT_CREDIT	0.770772908
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451135103
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.384675888
AMT_CREDIT	AMT_INCOME_TOTAL	0.377965752
CNT_CHILDREN	DAYS_BIRTH	0.335876269
DAYS_REGISTRATION	DAYS_BIRTH	0.335028046
DAYS_ID_PUBLISH	DAYS_BIRTH	0.270073313
EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.201089847

Bottom 10 Correlations for Loan Payees

DAYS_EMPLOYED	DAYS_BIRTH	-0.615289978
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.27222439
CNT_CHILDREN	DAYS_EMPLOYED	-0.243591518
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.204370881
DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	-0.184718008
EXT_SOURCE_23	DAYS_BIRTH	-0.17888625
DAYS_EMPLOYED	AMT_INCOME_TOTAL	-0.162702675
DAYS_EMPLOYED	AMT_ANNUITY	-0.113005115
EXT_SOURCE_23	DAYS_ID_PUBLISH	-0.111944396
EXT_SOURCE_24	DAYS_REGISTRATION	-0.099901082

Correlation

Target – 1 (Loan Defaulters)

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_23	DAYS_LAST_PHONE_CHANGE	CNT_CHILDREN
AMT_INCOME_TOTAL	1												
AMT_CREDIT	0.015271444	1											
AMT_ANNUITY	0.018004594	0.749665201	1										
AMT_GOODS_PRICE	0.013266279	0.982432318	0.749705184	1									
REGION_POPULATION_RELATIVE	-0.006180303	0.067775624	0.073123998	0.076596242	1								
DAYS_BIRTH	0.009033662	-0.142506035	-0.008751713	-0.140996151	-0.016468731	1							
DAYS_EMPLOYED	-0.011555963	0.016039571	-0.079556008	0.020213912	0.007742909	-0.58147904	1						
DAYS_REGISTRATION	-0.009561152	-0.042844404	0.021581654	-0.043371319	-0.046130288	0.28843784	-0.188718437	1					
DAYS_ID_PUBLISH	-0.009122006	-0.043771901	-0.02132109	-0.049784603	-0.005118563	0.24789657	-0.230063668	0.09029149	1				
EXT_SOURCE_2	-0.016228175	0.119184882	0.113693247	0.133319809	0.159220129	-0.11141366	-0.018800184	-0.067763159	-0.03748574	1			
EXT_SOURCE_23	-0.026469291	0.045724291	0.017777631	0.04751535	-0.021648936	-0.13943364	0.083171459	-0.052718825	-0.073389069	0.047324503	1		
DAYS_LAST_PHONE_CHANGE	0.012457111	-0.124539343	-0.100470941	-0.128807068	-0.067105681	0.12460949	-0.015732544	0.078604652	0.138087781	-0.204067856	-0.041135827	1	
CNT_CHILDREN	0.010110177	0.007601905	0.029172977	-0.001116682	-0.020359154	0.2496732	-0.189324184	0.152113117	-0.042360717	-0.01541713	-0.015321808	0.011339334	1

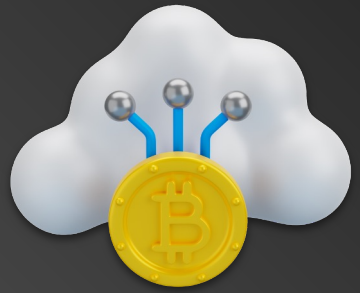
Top 10 Correlations for Loan Defaulters

AMT_GOODS_PRICE	AMT_CREDIT	0.982432318
AMT_GOODS_PRICE	AMT_ANNUITY	0.749705184
AMT_ANNUITY	AMT_CREDIT	0.749665201
DAYS_REGISTRATION	DAYS_BIRTH	0.288437837
CNT_CHILDREN	DAYS_BIRTH	0.2496732
DAYS_ID_PUBLISH	DAYS_BIRTH	0.247896571
EXT_SOURCE_2	*REGION_POPULATION_RELATIVE*	0.159220129
CNT_CHILDREN	DAYS_REGISTRATION	0.152113117
DAYS_LAST_PHONE_CHANGE	DAYS_ID_PUBLISH	0.138087781
EXT_SOURCE_2	*AMT_GOODS_PRICE*	0.133319809

Bottom 10 Correlations for Loan Defaulters

DAYS_EMPLOYED	DAYS_BIRTH	-0.581479041
DAYS_ID_PUBLISH	DAYS_EMPLOYED	-0.230063668
DAYS_LAST_PHONE_CHANGE	EXT_SOURCE_2	-0.204067856
CNT_CHILDREN	DAYS_EMPLOYED	-0.189324184
DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188718437
DAYS_BIRTH	AMT_CREDIT	-0.142506035
DAYS_BIRTH	AMT_GOODS_PRICE	-0.140996151
EXT_SOURCE_23	*DAYS_BIRTH*	-0.139433636
DAYS_LAST_PHONE_CHANGE	AMT_GOODS_PRICE	-0.128807068
DAYS_LAST_PHONE_CHANGE	AMT_CREDIT	-0.124539343

Insights & Results



KEY INSIGHTS FROM OUR ANALYSIS

- **Outliers and Data Distribution:** Most columns display outliers, except for "Days_Birth" and "Days_ID_Publish."
- **Data Imbalance:** The dataset has a substantial data imbalance ratio of 8.75%, emphasizing the need to address class imbalance.
- **Loan Amount and Income:** People with incomes between 100k to 200k tend to apply for loans more frequently. As income increases, loan default percentages decreases.
- **Age and Loan Default:** Loan default likelihood decreases as age increases, implying older applicants have better repayment behaviour.
- **Marital Status:** Married or widowed applicants show lower loan default percentages, suggesting a connection between marital status and loan repayment.
- **Employment Duration:** Longer job tenures associate with lower loan default rates, indicating stable employment positively influences repayments.
- **Education Level:** Higher education correlates with reduced default rates, implying educated individuals prioritize loan repayment.
- **Housing Situation:** Rented apartments or living with parents link to higher default rates, possibly reflecting financial stability differences.
- **Occupation Impact:** Occupations like IT Staff, Secretaries, Accountants, Core Staff, and Managers show lower default percentages than Low Skill Labourers, Drivers, and Cleaning Staff.
- **Loan Type:** Revolving loans have lower defaults, though cash loans are more common.
- **Gender Dynamics:** Despite more female applicants, their loan default rate is lower than males. Males, however, ask for higher average loan amounts.

These insights provide a deep understanding of loan default contributors and inform practical decisions to manage lending risks.

RESULTS

- The main challenge was managing the large dataset. I learned how to handle large datasets.
- The dataset had a lot of missing data and outliers, which I learned how to manage effectively.
- I learned how to use data analysis add-ins in Excel.
- This project helped me in improving my Excel skills and gaining a better understanding of how to navigate complex datasets.

Thank You!