BANK LOAN CASE STUDY.

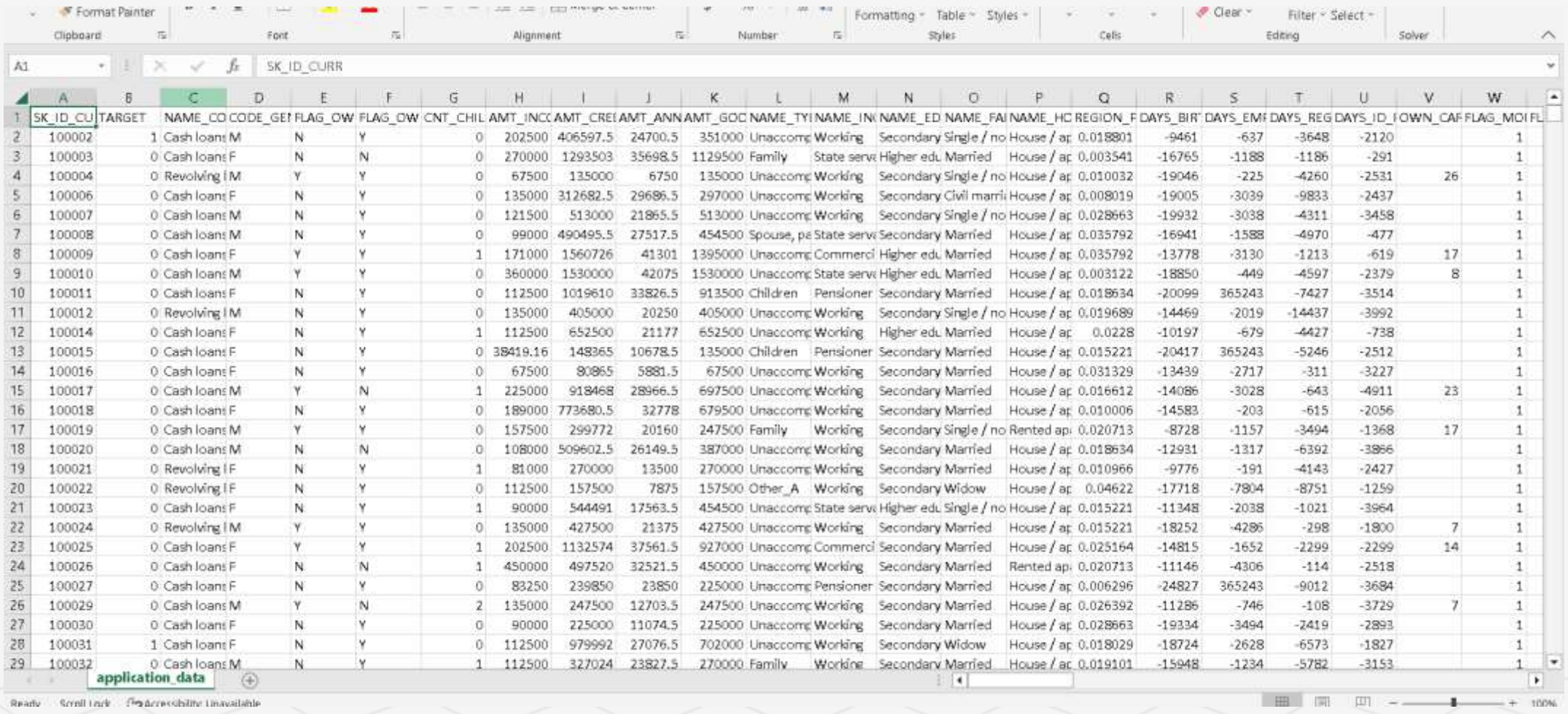
Project Description:- This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Tech Stack Used: Microsoft Excel (Home & Student 2019).

By,

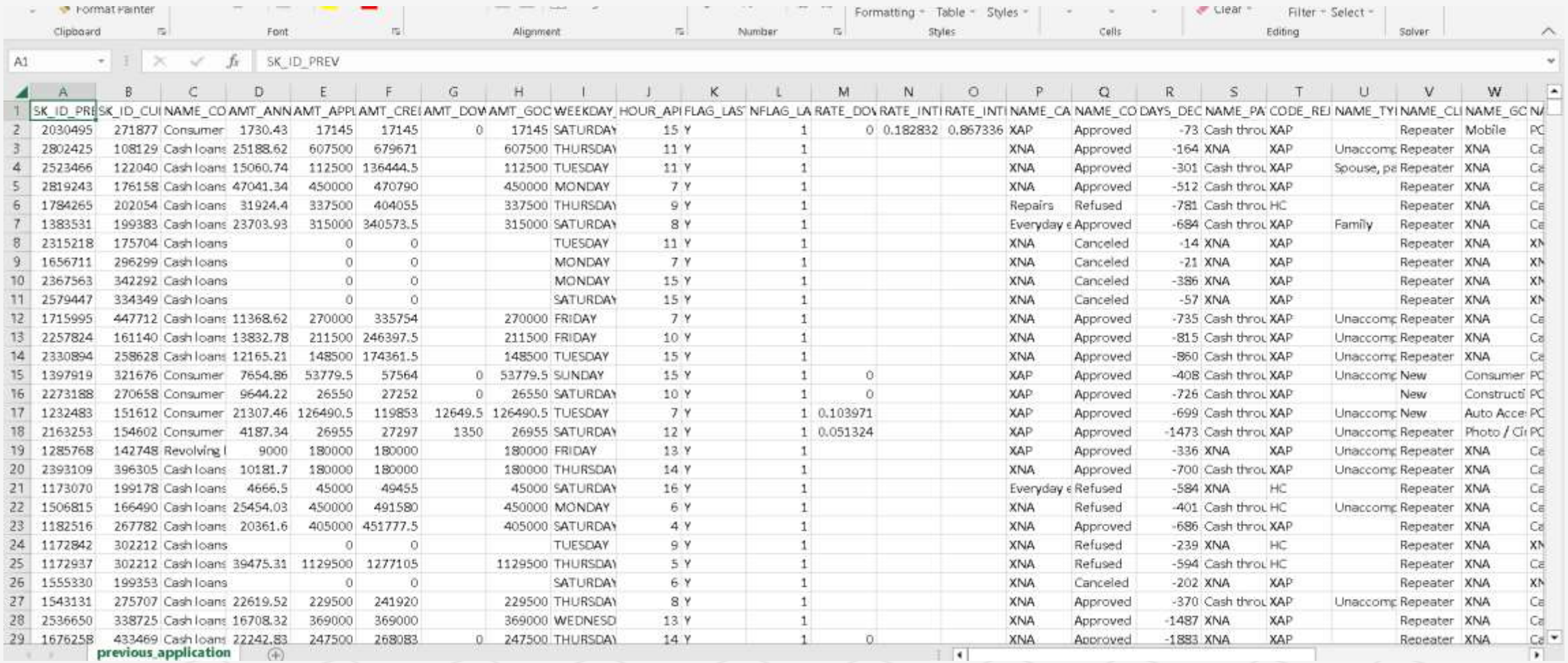
Lohith Kumar.A

Overall Approach of Analysis with Problem Statement: Two csv files for different applicants for accessing loans are given. We will first import it into excel. The goal is to analyze the historical data to understand a borrower's creditworthiness or to assess the risk involved in the granting of a loan. The result of the analysis help banks and financial institutions evaluate their risk and those of their customers. The company will utilize this knowledge for its portfolio and risk assessment.



SK_ID_CU	TARGET	NAME_CO	CODE_GE	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOO	NAME_TY	NAME_IN	NAME_ED	NAME_FA	NAME_HO	REGION_F	DAYS_BIR	DAYS_EM	DAYS_REG	DAYS_ID_	OWN_CAF	FLAG_MOI	FL
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccomp	Working	Secondary	Single / no	House / ap	0.018801	-9461	-637	-3648	-2120		1	
100003	0	Cash loans	F	N	N	0	270000	1293503	35698.5	1129500	Family	State serv	Higher edu	Married	House / ap	0.003541	-16765	-1188	-1186	-291		1	
100004	0	Revolving l	M	Y	Y	0	67500	135000	6750	135000	Unaccomp	Working	Secondary	Single / no	House / ap	0.010032	-19046	-225	-4260	-2531	26	1	
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.008019	-19005	-3039	-9833	-2437		1	
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccomp	Working	Secondary	Single / no	House / ap	0.028663	-19932	-3038	-4311	-3458		1	
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, pa	State serv	Secondary	Married	House / ap	0.035792	-16941	-1588	-4970	-477		1	
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccomp	Commerci	Higher edu	Married	House / ap	0.035792	-13778	-3130	-1213	-619	17	1	
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccomp	State serv	Higher edu	Married	House / ap	0.003122	-18850	-449	-4597	-2379	8	1	
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary	Married	House / ap	0.018634	-20099	365243	-7427	-3514		1	
100012	0	Revolving l	M	N	Y	0	135000	405000	20250	405000	Unaccomp	Working	Secondary	Single / no	House / ap	0.019689	-14469	-2019	-14437	-3992		1	
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccomp	Working	Higher edu	Married	House / ap	0.0228	-10197	-679	-4427	-738		1	
100015	0	Cash loans	F	N	Y	0	38419.16	148365	10678.5	135000	Children	Pensioner	Secondary	Married	House / ap	0.015221	-20417	365243	-5246	-2512		1	
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccomp	Working	Secondary	Married	House / ap	0.031329	-13439	-2717	-311	-3227		1	
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccomp	Working	Secondary	Married	House / ap	0.016612	-14086	-3028	-643	-4911	23	1	
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccomp	Working	Secondary	Married	House / ap	0.010006	-14583	-203	-615	-2056		1	
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary	Single / no	Rented ap	0.020713	-8728	-1157	-3494	-1368	17	1	
100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Unaccomp	Working	Secondary	Married	House / ap	0.018634	-12931	-1317	-6392	-3866		1	
100021	0	Revolving l	F	N	Y	1	81000	270000	13500	270000	Unaccomp	Working	Secondary	Married	House / ap	0.010966	-9776	-191	-4143	-2427		1	
100022	0	Revolving l	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary	Widow	House / ap	0.04622	-17718	-7804	-8751	-1259		1	
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccomp	State serv	Higher edu	Single / no	House / ap	0.015221	-11348	-2038	-1021	-3964		1	
100024	0	Revolving l	M	Y	Y	0	135000	427500	21375	427500	Unaccomp	Working	Secondary	Married	House / ap	0.015221	-18252	-4286	-298	-1800	7	1	
100025	0	Cash loans	F	Y	Y	1	202500	1132574	37561.5	927000	Unaccomp	Commerci	Secondary	Married	House / ap	0.025164	-14815	-1652	-2299	-2299	14	1	
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccomp	Working	Secondary	Married	Rented ap	0.020713	-11146	-4306	-114	-2518		1	
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccomp	Pensioner	Secondary	Married	House / ap	0.006296	-24827	365243	-9012	-3684		1	
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccomp	Working	Secondary	Married	House / ap	0.026392	-11286	-746	-108	-3729	7	1	
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccomp	Working	Secondary	Married	House / ap	0.028663	-19334	-3494	-2419	-2893		1	
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccomp	Working	Secondary	Widow	House / ap	0.018029	-18724	-2628	-6573	-1827		1	
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working	Secondary	Married	House / ap	0.019101	-15948	-1234	-5782	-3153		1	

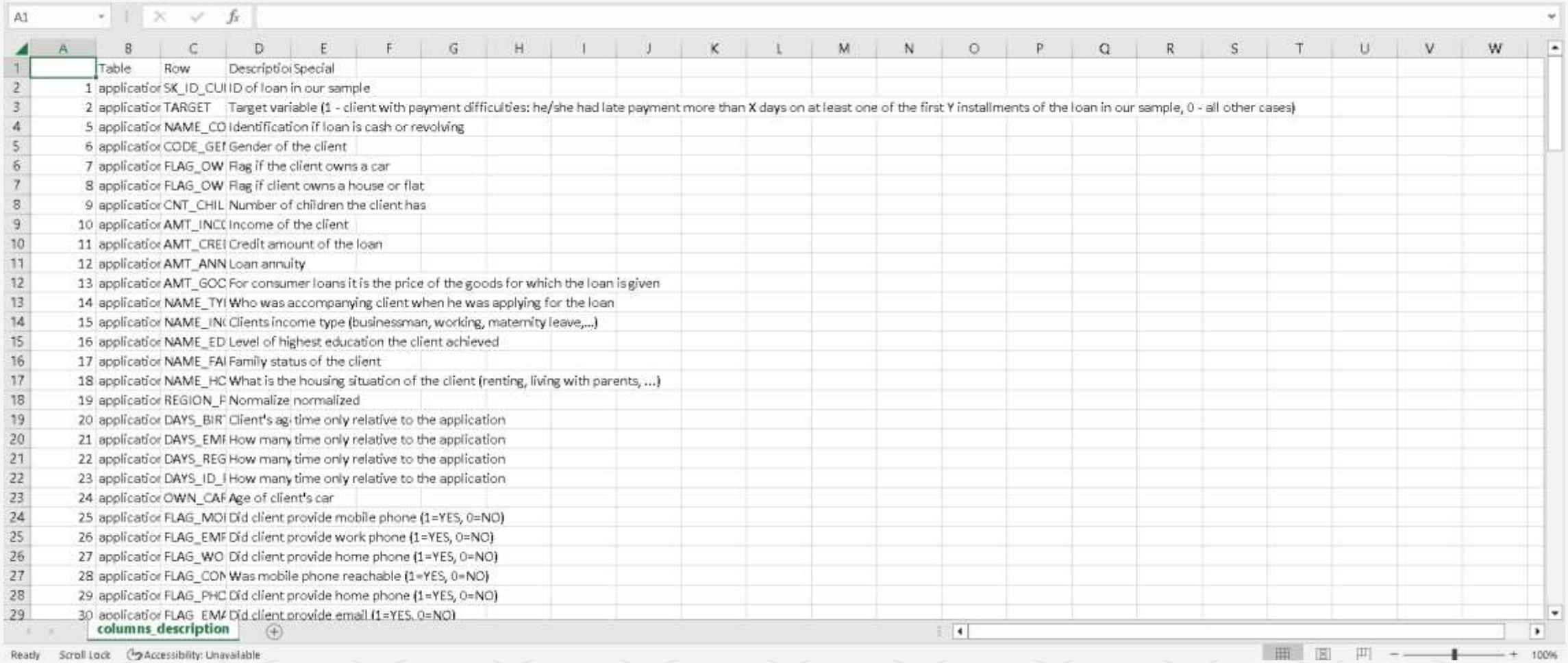
CSV files will be checked for any unnecessary data and unwanted columns/rows, and will be cleaned/removed if necessary.
Will check for outliers, if any, to find if there is skewness in the given columns which would affect the final visualization and insight.



SK_ID_PRE	SK_ID_CUI	NAME_CO	AMT_ANN	AMT_APPL	AMT_CRE	AMT_DOV	AMT_GOC	WEEKDAY_	HOUR_API	FLAG_LAS	NFLAG_LA	RATE_DO	RATE_INTI	RATE_INTI	NAME_CA	NAME_CO	DAYS_DEC	NAME_PA	CODE_REJ	NAME_TYI	NAME_CLI	NAME_GC	NA
2030495	271877	Consumer	1730.43	17145	17145	0	17145	SATURDAY	15	Y	1	0	0.182832	0.867336	XAP	Approved	-73	Cash throu	XAP		Repeater	Mobile	PC
2802425	108129	Cash loans	25188.62	607500	679671		607500	THURSDAY	11	Y	1				XNA	Approved	-164	XNA	XAP	Unaccomp	Repeater	XNA	Ca
2523466	122040	Cash loans	15060.74	112500	136444.5		112500	TUESDAY	11	Y	1				XNA	Approved	-301	Cash throu	XAP	Spouse, pa	Repeater	XNA	Ca
2819243	176158	Cash loans	47041.34	450000	470790		450000	MONDAY	7	Y	1				XNA	Approved	-512	Cash throu	XAP		Repeater	XNA	Ca
1784265	202054	Cash loans	31924.4	337500	404055		337500	THURSDAY	9	Y	1				Repairs	Refused	-781	Cash throu	HC		Repeater	XNA	Ca
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8	Y	1				Everyday e	Approved	-684	Cash throu	XAP	Family	Repeater	XNA	Ca
2315218	175704	Cash loans		0	0			TUESDAY	11	Y	1				XNA	Canceled	-14	XNA	XAP		Repeater	XNA	XN
1656711	296299	Cash loans		0	0			MONDAY	7	Y	1				XNA	Canceled	-21	XNA	XAP		Repeater	XNA	XN
2367563	342292	Cash loans		0	0			MONDAY	15	Y	1				XNA	Canceled	-386	XNA	XAP		Repeater	XNA	XN
2579447	334349	Cash loans		0	0			SATURDAY	15	Y	1				XNA	Canceled	-57	XNA	XAP		Repeater	XNA	XN
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	7	Y	1				XNA	Approved	-735	Cash throu	XAP	Unaccomp	Repeater	XNA	Ca
2257824	161140	Cash loans	13832.78	211500	246397.5		211500	FRIDAY	10	Y	1				XNA	Approved	-815	Cash throu	XAP	Unaccomp	Repeater	XNA	Ca
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15	Y	1				XNA	Approved	-860	Cash throu	XAP	Unaccomp	Repeater	XNA	Ca
1397919	321676	Consumer	7654.86	53779.5	57564	0	53779.5	SUNDAY	15	Y	1	0			XAP	Approved	-408	Cash throu	XAP	Unaccomp	New	Consumer	PC
2273188	270658	Consumer	9644.22	26550	27252	0	26550	SATURDAY	10	Y	1	0			XAP	Approved	-726	Cash throu	XAP		New	Constructi	PC
1232483	151612	Consumer	21307.46	126490.5	119853	12649.5	126490.5	TUESDAY	7	Y	1	0.103971			XAP	Approved	-699	Cash throu	XAP	Unaccomp	New	Auto Acce	PC
2163253	154602	Consumer	4187.34	26955	27297	1350	26955	SATURDAY	12	Y	1	0.051324			XAP	Approved	-1473	Cash throu	XAP	Unaccomp	Repeater	Photo / Ci	PC
1285768	142748	Revolving l	9000	180000	180000		180000	FRIDAY	13	Y	1				XAP	Approved	-336	XNA	XAP	Unaccomp	Repeater	XNA	Ca
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14	Y	1				XNA	Approved	-700	Cash throu	XAP	Unaccomp	Repeater	XNA	Ca
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16	Y	1				Everyday e	Refused	-584	XNA	HC		Repeater	XNA	Ca
1506815	166490	Cash loans	25454.03	450000	491580		450000	MONDAY	6	Y	1				XNA	Refused	-401	Cash throu	HC	Unaccomp	Repeater	XNA	Ca
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4	Y	1				XNA	Approved	-686	Cash throu	XAP		Repeater	XNA	Ca
1172842	302212	Cash loans		0	0			TUESDAY	9	Y	1				XNA	Refused	-239	XNA	HC		Repeater	XNA	XN
1172937	302212	Cash loans	39475.31	1129500	1277105		1129500	THURSDAY	5	Y	1				XNA	Refused	-594	Cash throu	HC		Repeater	XNA	Ca
1555330	199353	Cash loans		0	0			SATURDAY	6	Y	1				XNA	Canceled	-202	XNA	XAP		Repeater	XNA	XN
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8	Y	1				XNA	Approved	-370	Cash throu	XAP	Unaccomp	Repeater	XNA	Ca
2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESD	13	Y	1				XNA	Approved	-1487	XNA	XAP		Repeater	XNA	Ca
1676258	433469	Cash loans	22242.83	247500	268083	0	247500	THURSDAY	14	Y	1	0			XNA	Approved	-1883	XNA	XAP		Repeater	XNA	Ca

previous_application

Imbalance in data will be checked. Different types of analysis will be done to understand the relationships between different variable to find the Driving Factors. Different visualizations will be observed to understand the relationships.



	A	Table	Row	Descriptio	Special
2	1	applicatio	SK_ID_CUI	ID of loan in our sample	
3	2	applicatio	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)	
4	5	applicatio	NAME_CO	Identification if loan is cash or revolving	
5	6	applicatio	CODE_GEI	Gender of the client	
6	7	applicatio	FLAG_OW	Flag if the client owns a car	
7	8	applicatio	FLAG_OW	Flag if client owns a house or flat	
8	9	applicatio	CNT_CHIL	Number of children the client has	
9	10	applicatio	AMT_INCC	Income of the client	
10	11	applicatio	AMT_CRE	Credit amount of the loan	
11	12	applicatio	AMT_ANN	Loan annuity	
12	13	applicatio	AMT_GOC	For consumer loans it is the price of the goods for which the loan is given	
13	14	applicatio	NAME_TYI	Who was accompanying client when he was applying for the loan	
14	15	applicatio	NAME_IN	Clients income type (businessman, working, maternity leave,...)	
15	16	applicatio	NAME_ED	Level of highest education the client achieved	
16	17	applicatio	NAME_FA	Family status of the client	
17	18	applicatio	NAME_HC	What is the housing situation of the client (renting, living with parents, ...)	
18	19	applicatio	REGION_F	Normalize normalized	
19	20	applicatio	DAYS_BIR	Client's ag	time only relative to the application
20	21	applicatio	DAYS_EMI	How many	time only relative to the application
21	22	applicatio	DAYS_REG	How many	time only relative to the application
22	23	applicatio	DAYS_ID_I	How many	time only relative to the application
23	24	applicatio	OWN_CAF	Age of client's car	
24	25	applicatio	FLAG_MOI	Did client provide mobile phone (1=YES, 0=NO)	
25	26	applicatio	FLAG_EMF	Did client provide work phone (1=YES, 0=NO)	
26	27	applicatio	FLAG_WO	Did client provide home phone (1=YES, 0=NO)	
27	28	applicatio	FLAG_CON	Was mobile phone reachable (1=YES, 0=NO)	
28	29	applicatio	FLAG_PHC	Did client provide home phone (1=YES, 0=NO)	
29	30	applicatio	FLAG_EMA	Did client provide email (1=YES, 0=NO)	

columns_description

Ready Scroll Lock Accessibility: Unavailable 100%

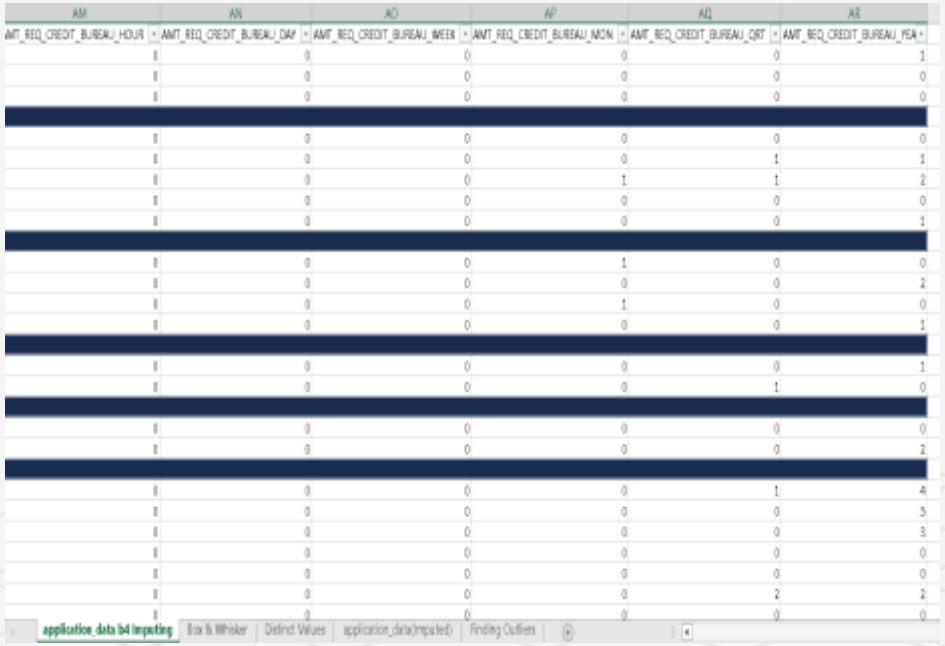
EDA – Data Cleaning (Identifying rows/columns with missing data, Identifying Unnecessary columns and cleaning if required).

Applicant_data.csv(Before Cleaning)

Columns 122, Rows 307512 After Cleaning:-Columns 42 + 4 Columns, Rows 306216

	A	B	C	D	E	F
1		Non Missing Values Count	Missing Value Count	Missing Value Percentage	Dropping Columns	
11	AMT_ANNUITY	307500	12	0.00%	No	
12	AMT_GOODS_PRICE	307234	278	0.09%	no	
13	NAME_TYPE_SUITE	306220	1292	0.42%	no	
23	OWN_CAR_AGE	104583	202929	65.99%	yes	
24	FLAG_MOBIL	307512	0	0.00%	Yes, Not needed	
30	OCCUPATION_TYPE	211121	96391	31.35%	no	
43	EXT_SOURCE_1	134134	173378	56.38%	yes, not needed	
44	EXT_SOURCE_2	306852	660	0.21%	yes, not needed	
45	EXT_SOURCE_3	246547	60965	19.83%	yes, not needed	
46	APARTMENTS_AVG	151451	156061	50.75%	yes	
47	BASEMENTAREA_AVG	127569	179943	58.52%	yes	
48	YEARS_BEGINEXPLUATATION_AVG	157505	150007	48.78%	yes	
49	YEARS_BUILD_AVG	103024	204488	66.50%	yes	
50	COMMONAREA_AVG	92647	214865	69.87%	yes	
51	ELEVATORS_AVG	143621	163891	53.30%	yes	
52	ENTRANCES_AVG	152684	154828	50.35%	yes	

There are date columns with negative values, those needs to be standardized.There are columns having more than 40% null data. These need to be removed. There are more than 50 unwanted columns or columns not desirable for our analysis so I will remove them from the dataset. There are columns with null values less than 40%. They can be treated in 2 ways. I can delete those columns but then I might lose some important information required for my analysis. I can retain it but then I will have to do treatment. If I impute them, I will introduce bias. The decision to delete or retain basically depends on the Understanding of the problem statement, the usefulness of the variable, total size of available data. Here it seems that those columns can be removed So, I have removed them. There are still some columns will very little missing values which will be treated if necessary or left as it is.



AMT_ANNUITY has a smaller number of null values (12). It can be imputed with mean. If it has an outlier which is very large then the null values can be computed with Median.EXIT_SOURCE_2 has 656 null values which is also quite small as compared to total number of rows. Can be imputed with 0.OCCUPATION_TYPE has 96005 null values. Can be imputed by the category which is the most popular (Mode), i.e., ‘Labourers’.

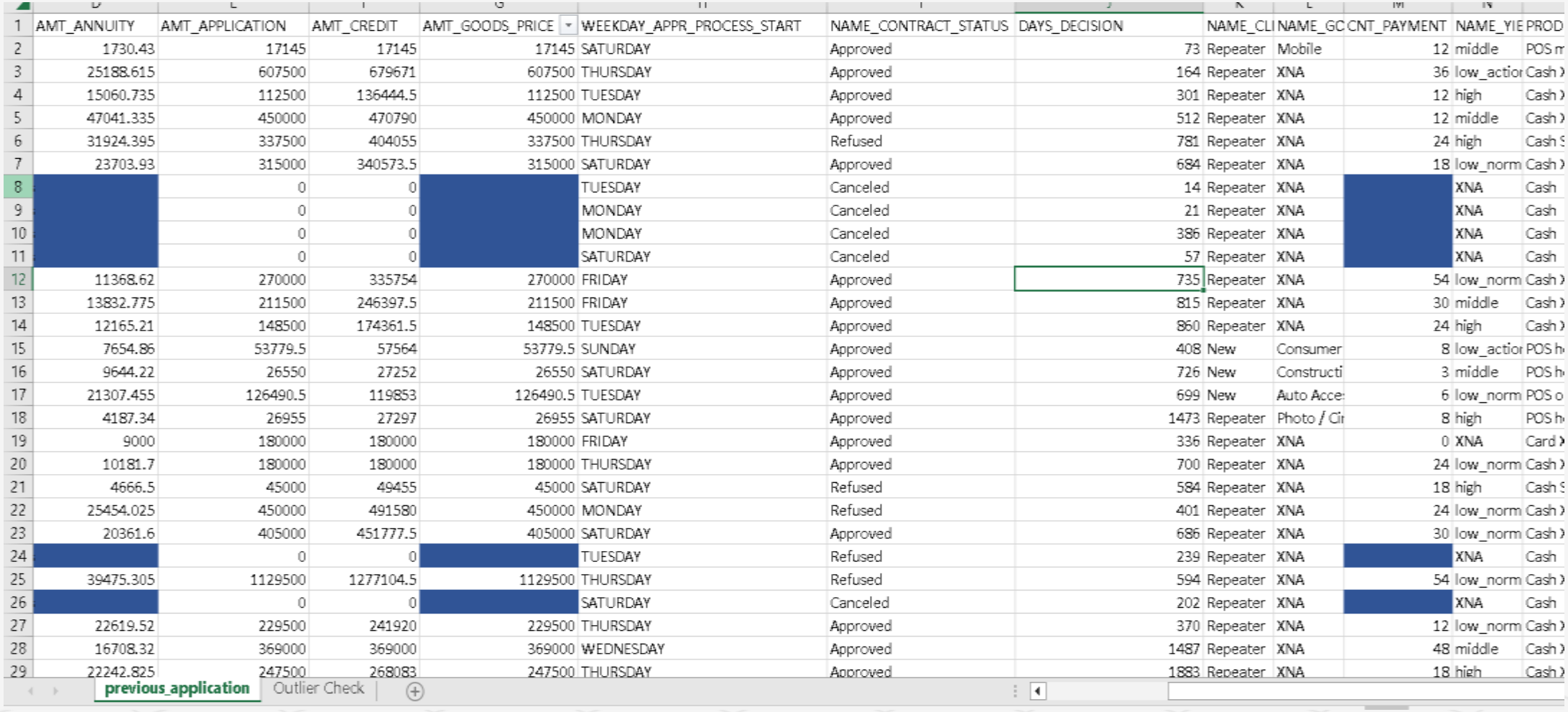
	K	
EDIT	AMT_ANNUITY	AMT_G
315000	15448.5	
270000	16312.5	
755190	36328.5	
450000		
521280	31630.5	
127350	7438.5	
48104.5	42471	
112500	11812.5	
599544	21663	
450000	35554.5	
202500	10125	
675000	35964	
215000	33543	
808650	29709	
27901.5	48825	

W	X	
D_PUBLISH	OCCUPATION_TYPE	WEEKDAY
2120	Laborers	WEDNESD
291	Core staff	MONDAY
2531	Laborers	MONDAY
2437	Laborers	WEDNESD
3458	Core staff	THURSDA
477	Laborers	WEDNESD
619	Accountants	SUNDAY
2379	Managers	MONDAY
3514		WEDNESD
3992	Laborers	THURSDA
738	Core staff	SATURDA
2512		FRIDAY
3227	Laborers	FRIDAY
4911	Drivers	THURSDA
2056	Laborers	MONDAY
1368	Laborers	SATURDA
3866	Drivers	THURSDA
2427	Laborers	MONDAY
1259	Laborers	FRIDAY
3964	Core staff	MONDAY
1800	Laborers	FRIDAY
2299	Sales staff	MONDAY
2518	Sales staff	THURSDA
3684		FRIDAY
3729	Drivers	THURSDA
2893	Cleaning staff	SATURDA
1827	Cooking staff	MONDAY
3153	Laborers	SATURDA

Previous_Application.csv(Before Cleaning)
Columns 37 ,Rows 1048576
There are columns with more than 40% null values and few unnecessary columns that need to be removed.
After cleaning:-Columns-15, Rows- 1048352

	A	B	C	D	E
1		Non Missing Value Counts	Missing Value Counts	Percentage Missing Value	Keep/Remove
5	AMT_ANNUITY	815567	233009	22.22%	
8	AMT_DOWN_PAYMENT	489180	559396	53.35%	R
9	AMT_GOODS_PRICE	807611	240965	22.98%	
14	RATE_DOWN_PAYMENT	489180	559396	53.35%	R
15	RATE_INTEREST_PRIMARY	3722	1044854	99.65%	R
16	RATE_INTEREST_PRIVILEGED	3722	1044854	99.65%	R
22	NAME_TYPE_SUITE	533436	515140	49.13%	R
30	CNT_PAYMENT	815570	233006	22.22%	
32	PRODUCT_COMBINATION	1048352	224	0.02%	
33	DAYS_FIRST_DRAWING	627868	420708	40.12%	R
34	DAYS_FIRST_DUE	627868	420708	40.12%	R
35	DAYS_LAST_DUE_1ST_VERSION	627868	420708	40.12%	R
36	DAYS_LAST_DUE	627868	420708	40.12%	R
37	DAYS_TERMINATION	627868	420708	40.12%	R
38	NFLAG_INSURED_ON_APPROVAL	627868	420708	40.12%	R
39					

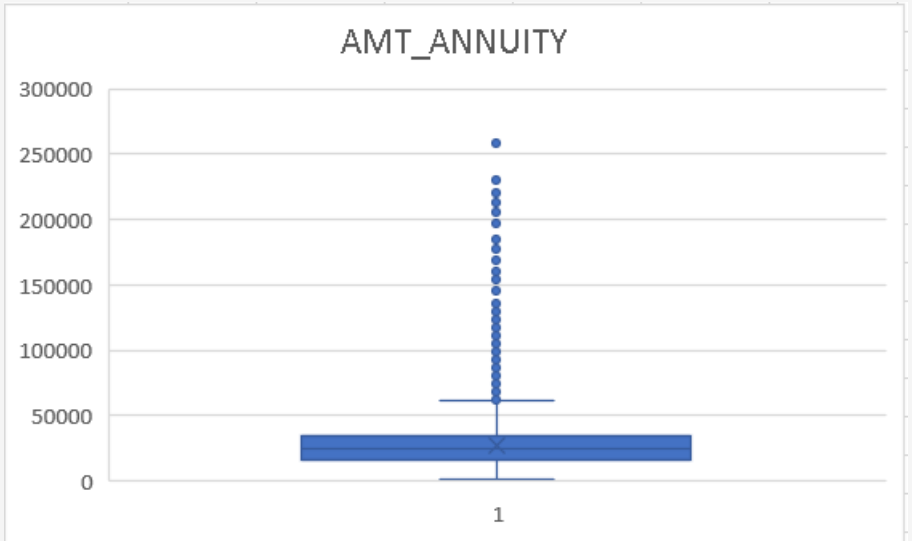
Columns with null values less than 40% are present. They can be treated in 2 ways. I can delete those columns but then I might lose some important information required for my analysis. I can retain it but then I will have to do treatment. If I impute them, I will introduce bias. The decision to delete or retain basically depends on the Understanding of the problem statement, the usefulness of the variable, total size of available data. Here it seems that those columns can be removed So, I have removed them. There are still some columns will very little missing values which will be treated if necessary or left as it is.

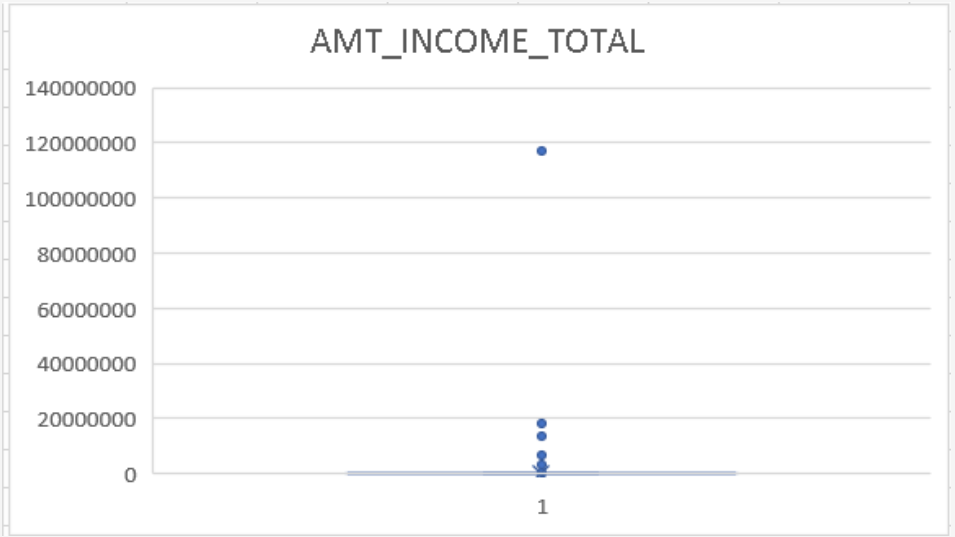


	D	E	F	G	H	I	J	K	L	M	N	
1	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	NAME_CONTRACT_STATUS	DAYS_DECISION	NAME_CLI	NAME_GC	CNT_PAYMENT	NAME_YIE	PROD
2	1730.43	17145	17145	17145	SATURDAY	Approved	73	Repeater	Mobile	12	middle	POS m
3	25188.615	607500	679671	607500	THURSDAY	Approved	164	Repeater	XNA	36	low_actior	Cash X
4	15060.735	112500	136444.5	112500	TUESDAY	Approved	301	Repeater	XNA	12	high	Cash X
5	47041.335	450000	470790	450000	MONDAY	Approved	512	Repeater	XNA	12	middle	Cash X
6	31924.395	337500	404055	337500	THURSDAY	Refused	781	Repeater	XNA	24	high	Cash S
7	23703.93	315000	340573.5	315000	SATURDAY	Approved	684	Repeater	XNA	18	low_norm	Cash X
8		0	0		TUESDAY	Canceled	14	Repeater	XNA		XNA	Cash
9		0	0		MONDAY	Canceled	21	Repeater	XNA		XNA	Cash
10		0	0		MONDAY	Canceled	386	Repeater	XNA		XNA	Cash
11		0	0		SATURDAY	Canceled	57	Repeater	XNA		XNA	Cash
12	11368.62	270000	335754	270000	FRIDAY	Approved	735	Repeater	XNA	54	low_norm	Cash X
13	13832.775	211500	246397.5	211500	FRIDAY	Approved	815	Repeater	XNA	30	middle	Cash X
14	12165.21	148500	174361.5	148500	TUESDAY	Approved	860	Repeater	XNA	24	high	Cash X
15	7654.86	53779.5	57564	53779.5	SUNDAY	Approved	408	New	Consumer	8	low_actior	POS ho
16	9644.22	26550	27252	26550	SATURDAY	Approved	726	New	Constructi	3	middle	POS ho
17	21307.455	126490.5	119853	126490.5	TUESDAY	Approved	699	New	Auto Acce	6	low_norm	POS o
18	4187.34	26955	27297	26955	SATURDAY	Approved	1473	Repeater	Photo / Ci	8	high	POS ho
19	9000	180000	180000	180000	FRIDAY	Approved	336	Repeater	XNA	0	XNA	Card X
20	10181.7	180000	180000	180000	THURSDAY	Approved	700	Repeater	XNA	24	low_norm	Cash X
21	4666.5	45000	49455	45000	SATURDAY	Refused	584	Repeater	XNA	18	high	Cash S
22	25454.025	450000	491580	450000	MONDAY	Refused	401	Repeater	XNA	24	low_norm	Cash X
23	20361.6	405000	451777.5	405000	SATURDAY	Approved	686	Repeater	XNA	30	low_norm	Cash X
24		0	0		TUESDAY	Refused	239	Repeater	XNA		XNA	Cash
25	39475.305	1129500	1277104.5	1129500	THURSDAY	Refused	594	Repeater	XNA	54	low_norm	Cash X
26		0	0		SATURDAY	Canceled	202	Repeater	XNA		XNA	Cash
27	22619.52	229500	241920	229500	THURSDAY	Approved	370	Repeater	XNA	12	low_norm	Cash X
28	16708.32	369000	369000	369000	WEDNESDAY	Approved	1487	Repeater	XNA	48	middle	Cash X
29	22242.825	247500	268083	247500	THURSDAY	Approved	1883	Repeater	XNA	18	high	Cash X

previous_application | Outlier Check

EDA – Identifying Outliers.
Applicatiton_data.csv



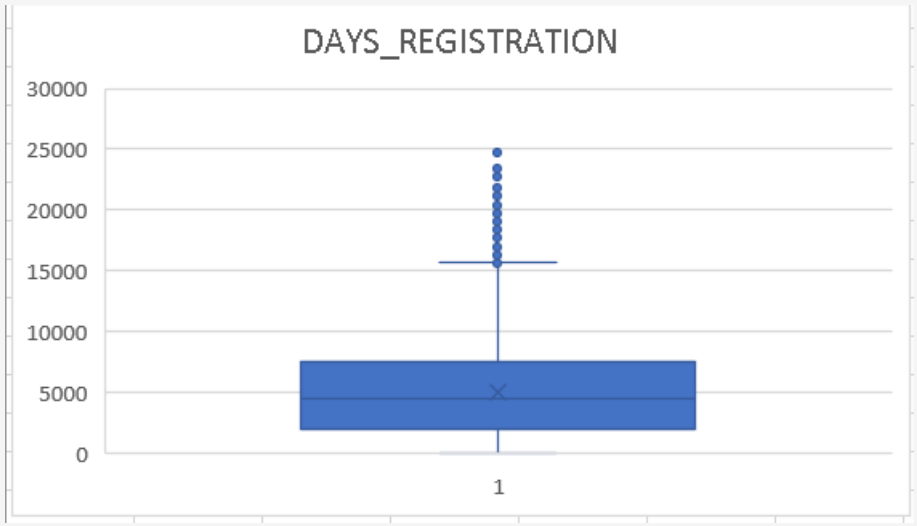


Outliers are those values which are Less Than LOWER BOUND and Also Greater Than UPPER BOUND.
Box and Whisker chart are shown in next page.

	INCOME	CREDIT	ANNUITY	GOODS_PRICE
Count	306215	306215	306203	306215
Mean	168782.9641	598799.7	27122.21047	537947.9476
STD	237517.4738	401959.9	14490.87377	368917.9877
min	25650	45000	1615.5	40500
q1	112500	270000	16551	238500
median	147600	513531	24930	450000
q3	202500	808650	34596	679500
max	117000000	4050000	258025.5	4050000
IQR	**90000**	**538650**	**18045**	**441000**
Upper Bound	**337500**	**1616625**	**61663.5**	**1341000**
Lower Bound	**-22500**	**-537975**	**-10516.5**	**-423000**

	BIRTH	EMPLOYED	REGISTRATION	PUBLISH
Count	306215	306215	306215	306215
Mean	16040.63386	67769.97	4987.987728	2994.331035
STD	4362.848928	139479.9	3522.552007	1509.515617
min	7489	0	0	0
q1	12418	934	2013	1720
median	15756	2221	4507	3255
q3	19685	5714	7481	4299
max	25229	365243	24672	7197
IQR	**7267**	**4780**	**5468**	**2579**
Upper Bound	**30585.5**	**12884**	**15683**	**8167.5**
Lower Bound	**1517.5**	**-6236**	**-6189**	**-2148.5**

In the above data of the next four Numeric columns, it can be seen that they contain outliers as well Except DAYS_BIRTH and DAYS_PUBLISH column. It can also be confirmed by the chart below.



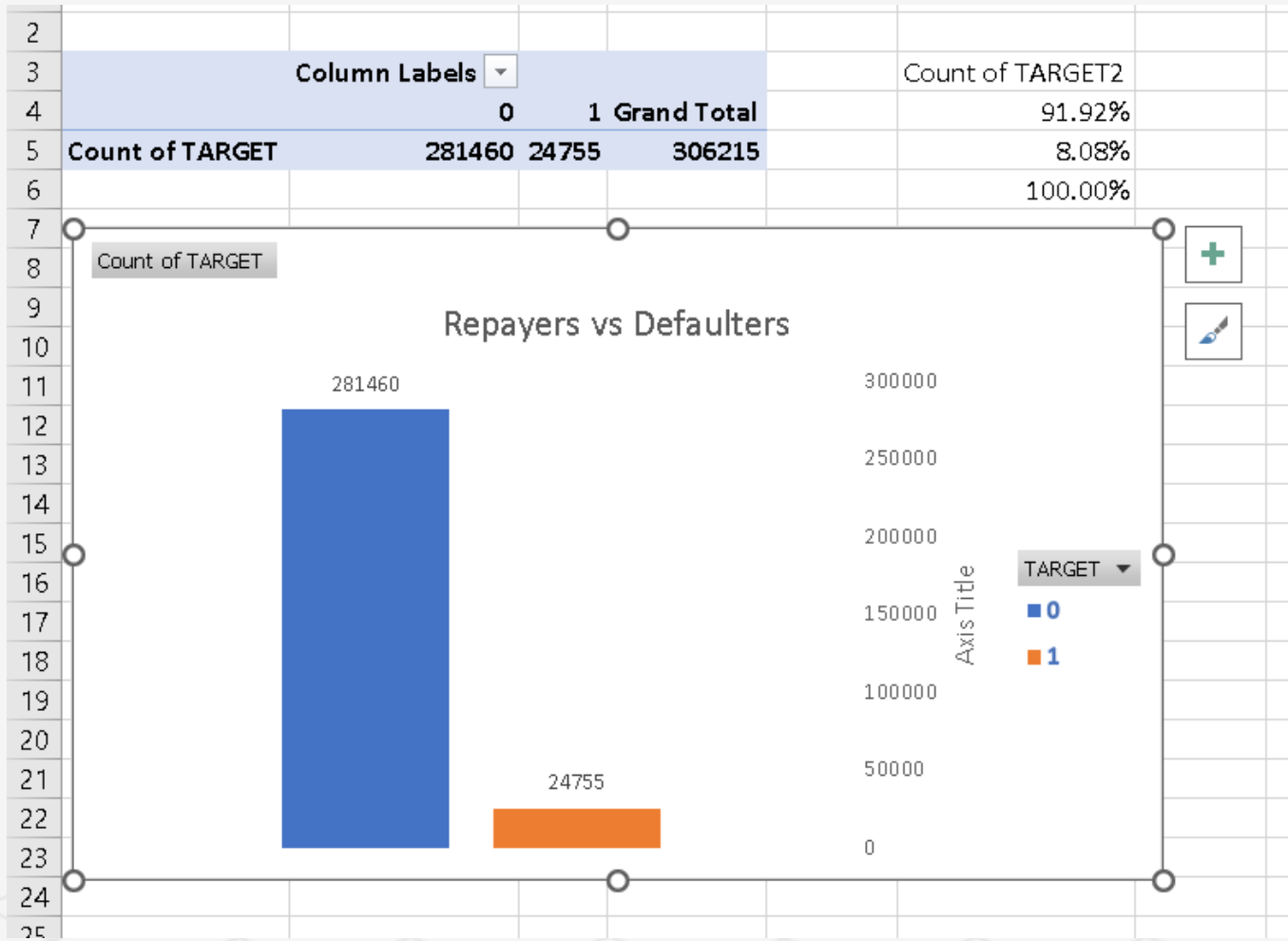
DESCRIPTION

	ANNUITY	APPLICATION	CREDIT	GOODS	DAYS_DECISION	CNT_PAYMENT
Count	815566	1048351	1048351	807610	1048351	815569
Mean	15891.27	174307.0055	195042	226289.2	882.1377382	15.99563863
STD	14745.55	291098.6811	316962	313448.8	779.3166038	14.50810019
min	0	0	0	0	2	0
q1	6301.35	18990	24300	50580	281	6
median	11250	70859.025	80293.5	111511.6	583	12
q3	20523	180000	215404	229500	1303	24
max	418058.1	6905160	6905160	6905160	2922	84
IQR	14221.65	161010	191104	178920	1022	18
UPPER BOUND	41855.48	421515	502059	497880	2836	51
LOWER BOUND	-15031.1	-222525	-262356	-217800	-1252	-21

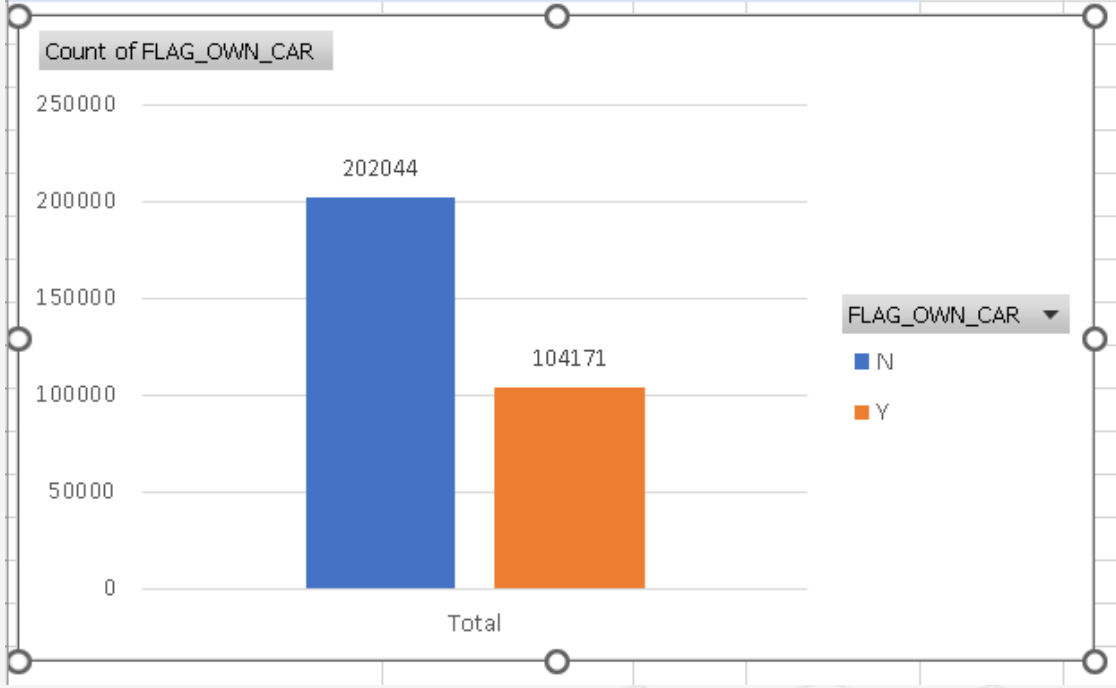
Data Imbalance Check and finding ratio of Imbalance.

Column Labels			
	0	1	Grand Total
Count of TARGET	281460	24755	306215

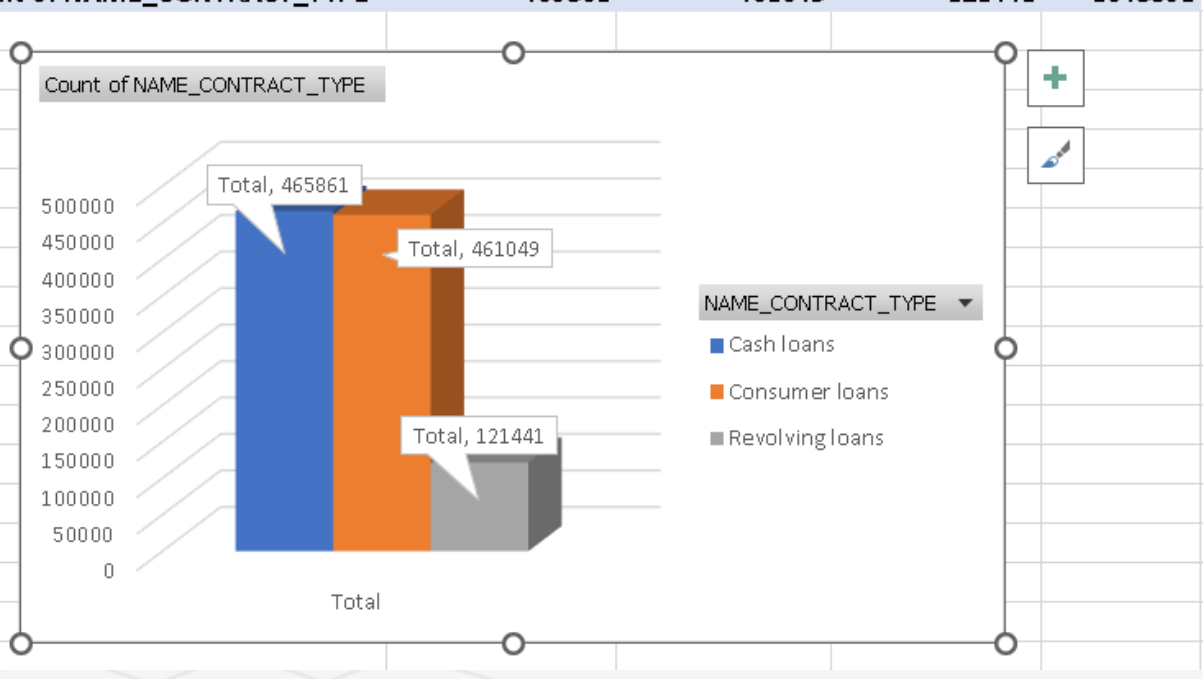
Count of TARGET2
91.92%
8.08%
100.00%



Column Labels			
	N	Y	Grand Total
Count of FLAG_OWN_CAR	202044	104171	306215



Column Labels				
	Cash loans	Consumer loans	Revolving loans	Grand Total
Count of NAME_CONTRACT_TYPE	465861	461049	121441	1048351



DATA IMPUTING BEFORE ANALYSIS
AMT_ANNUITY – Imputing with Mean.
OCCUPATION_TYPE – Imputing with ‘Labourers’

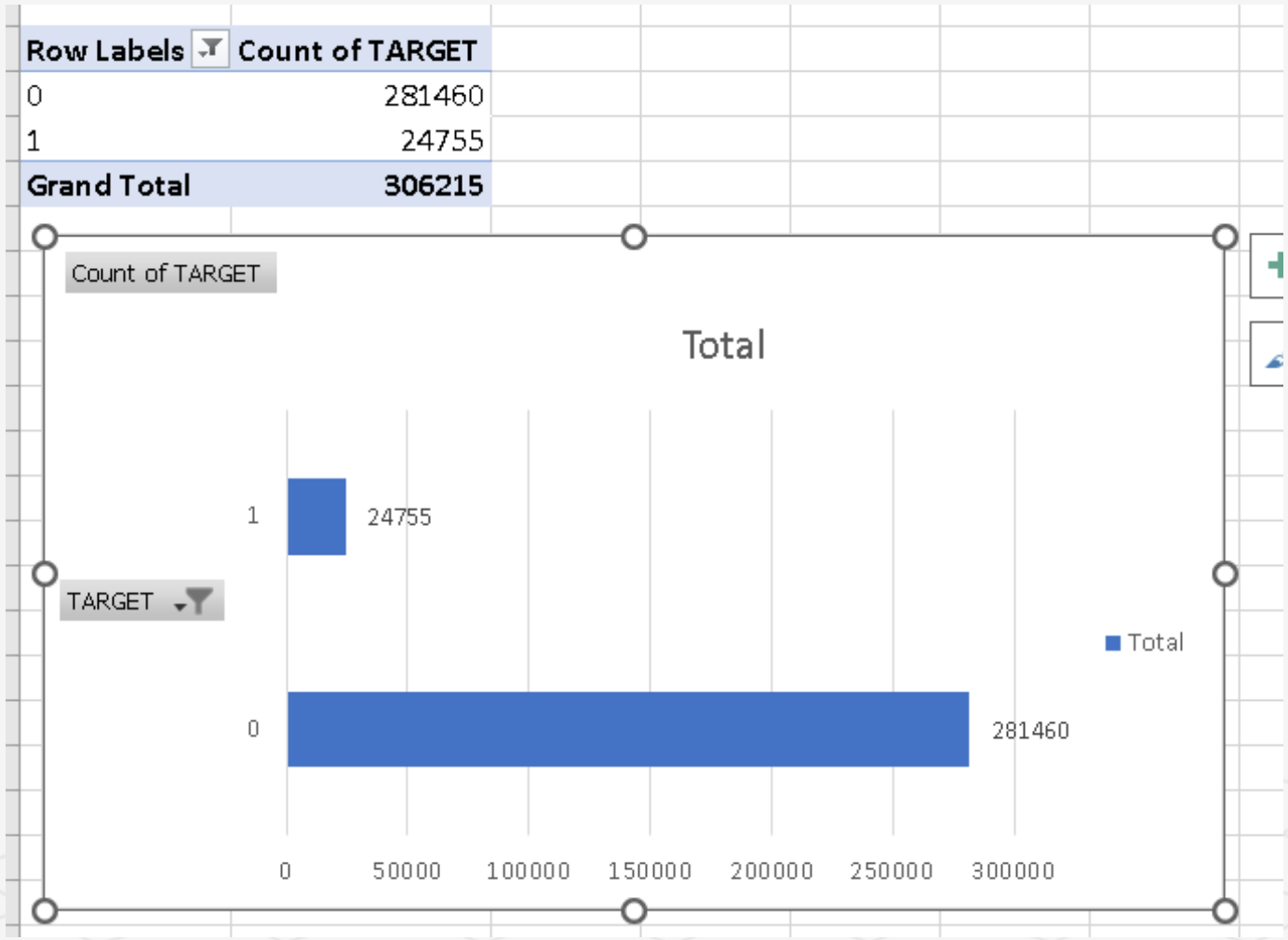
G	H	I	J	K	L
CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SU
0	180000	450000	27122.21047	450000	Unaccompanied
0	94500	450000	27122.21047	450000	Unaccompanied
0	202500	539100	27122.21047	450000	Unaccompanied
0	162000	296280	27122.21047	225000	Unaccompanied
0	202500	360000	27122.21047	360000	Unaccompanied
0	144000	219249	27122.21047	166500	Unaccompanied
0	90000	157500	27122.21047	157500	Unaccompanied
0	202500	929088	27122.21047	720000	Unaccompanied
0	171000	486000	27122.21047	486000	Unaccompanied
0	315000	628069.5	27122.21047	499500	Unaccompanied
0	157500	792000	27122.21047	792000	Family
0	315000	1483231.5	27122.21047	1354500	Unaccompanied

X	Y	Z	AA
	DAYS_ID_PUBLISH_GROUP	OCCUPATION_TYPE	WEEKDAY_APPR_PROCESS_STAR
2120	6	Laborers	WEDNESDAY
291	8	Core staff	MONDAY
2531	6	Laborers	MONDAY
2437	6	Laborers	WEDNESDAY
3458	5	Core staff	THURSDAY
477	8	Laborers	WEDNESDAY
619	8	Accountants	SUNDAY
2379	6	Managers	MONDAY
3514	5	Laborers	WEDNESDAY
3992	5	Laborers	THURSDAY
738	8	Core staff	SATURDAY
2512	6	Laborers	FRIDAY
3227	5	Laborers	FRIDAY
4911	4	Drivers	THURSDAY
2056	6	Laborers	MONDAY
1368	7	Laborers	SATURDAY
3866	5	Drivers	THURSDAY
2427	6	Laborers	MONDAY
1259	7	Laborers	FRIDAY
3964	5	Core staff	MONDAY
1800	7	Laborers	FRIDAY
2299	6	Sales staff	MONDAY

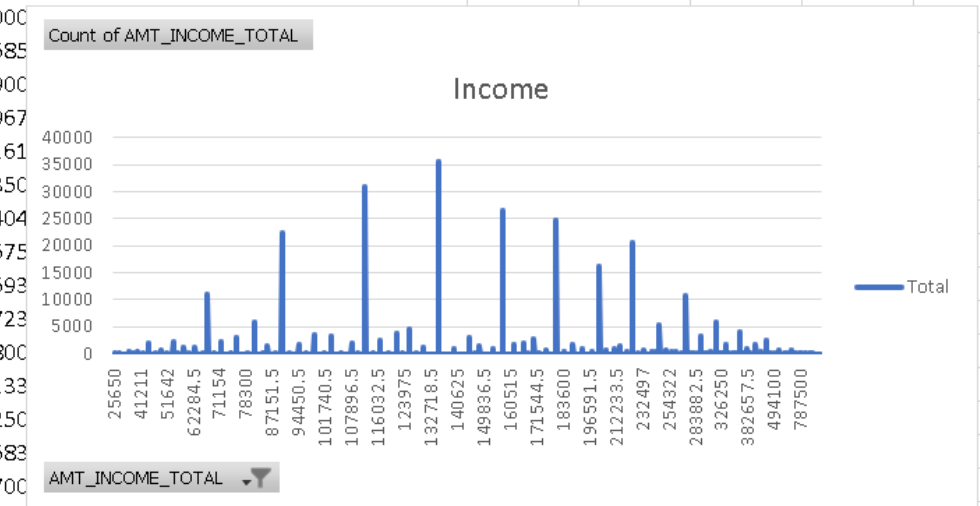
Explain the results of univariate, segmented variate and bivariate analysis.

Univariate Numerical Analysis.

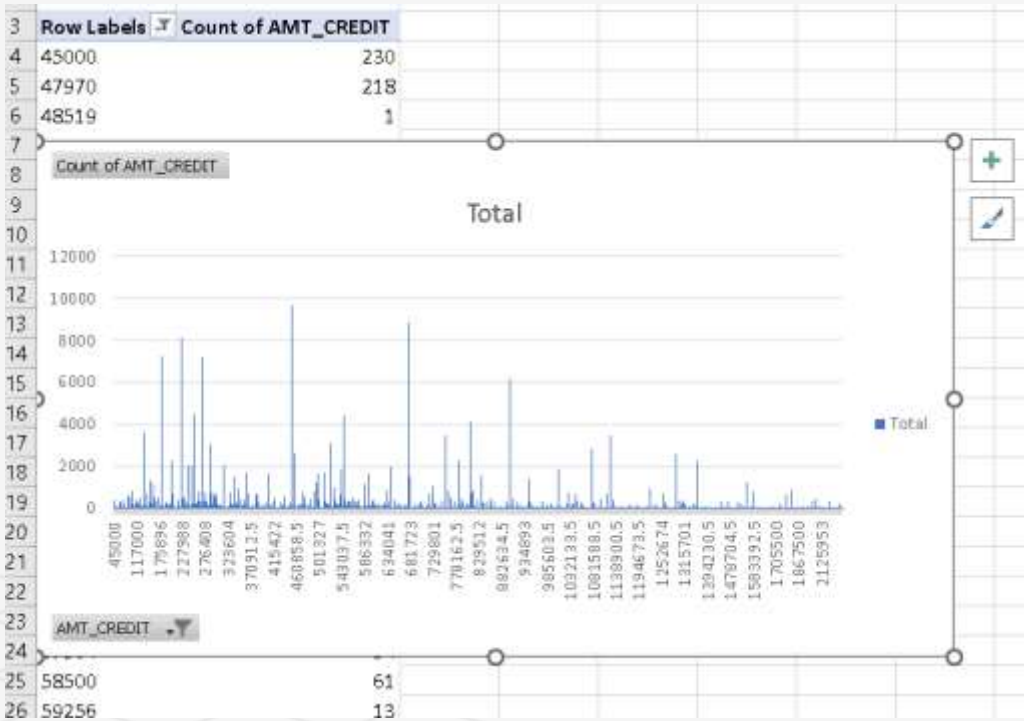
Row Labels	Count of TARGET
0	281460
1	24755
Grand Total	**306215**



Row Labels	Count of AMT_INCOME_TOTAL
25650	2
26100	3
26460	1
26550	2
27000	
27585	
27900	
27967	
28161	
28350	
28404	
28575	
28593	
28723	
28800	
29133	
29250	
29583	
29700	
30150	7

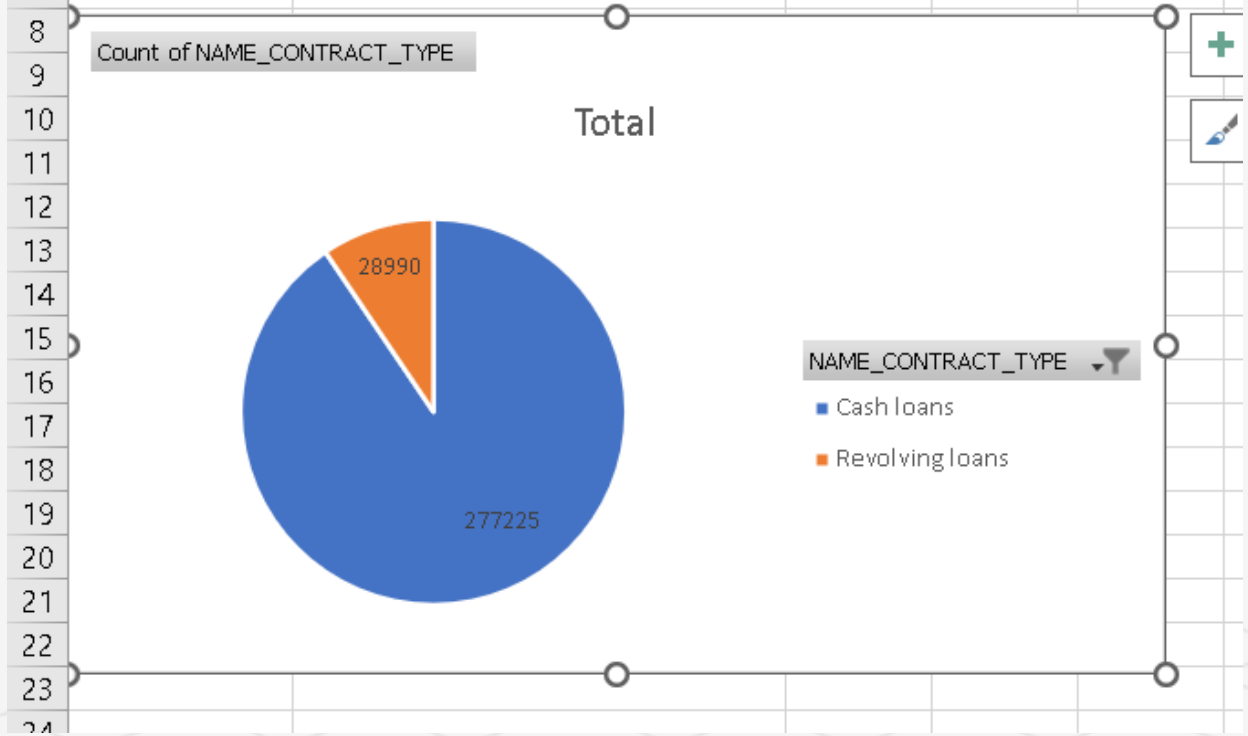


Row Labels	Count of AMT_CREDIT
45000	230
47970	218
48519	1
58500	61
59256	13

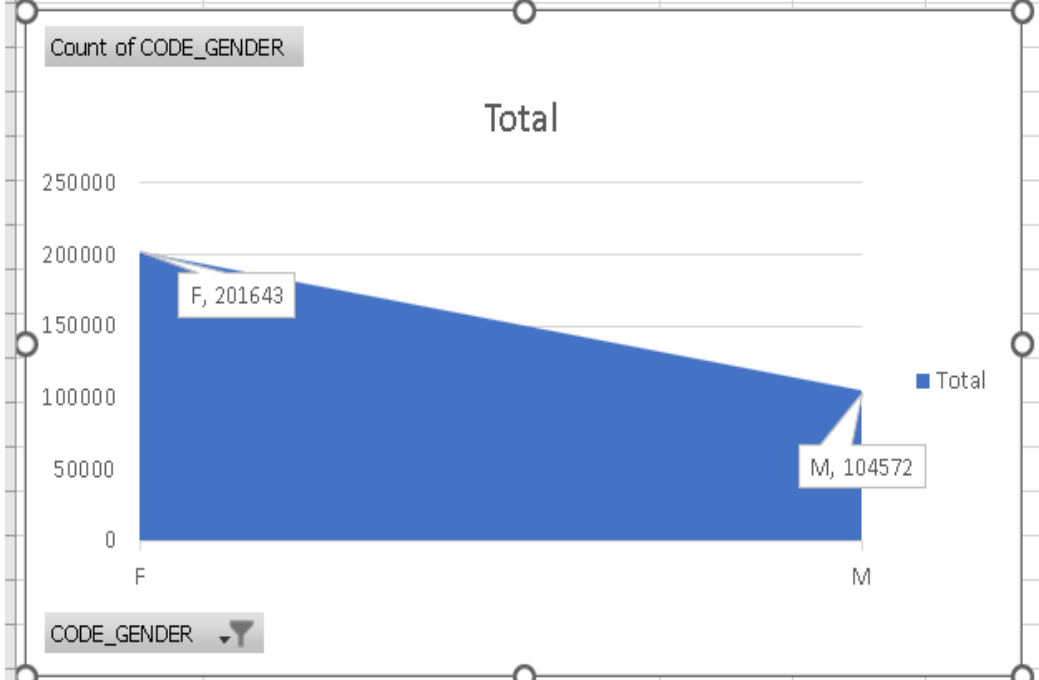


Univariate Categorical Analysis:-

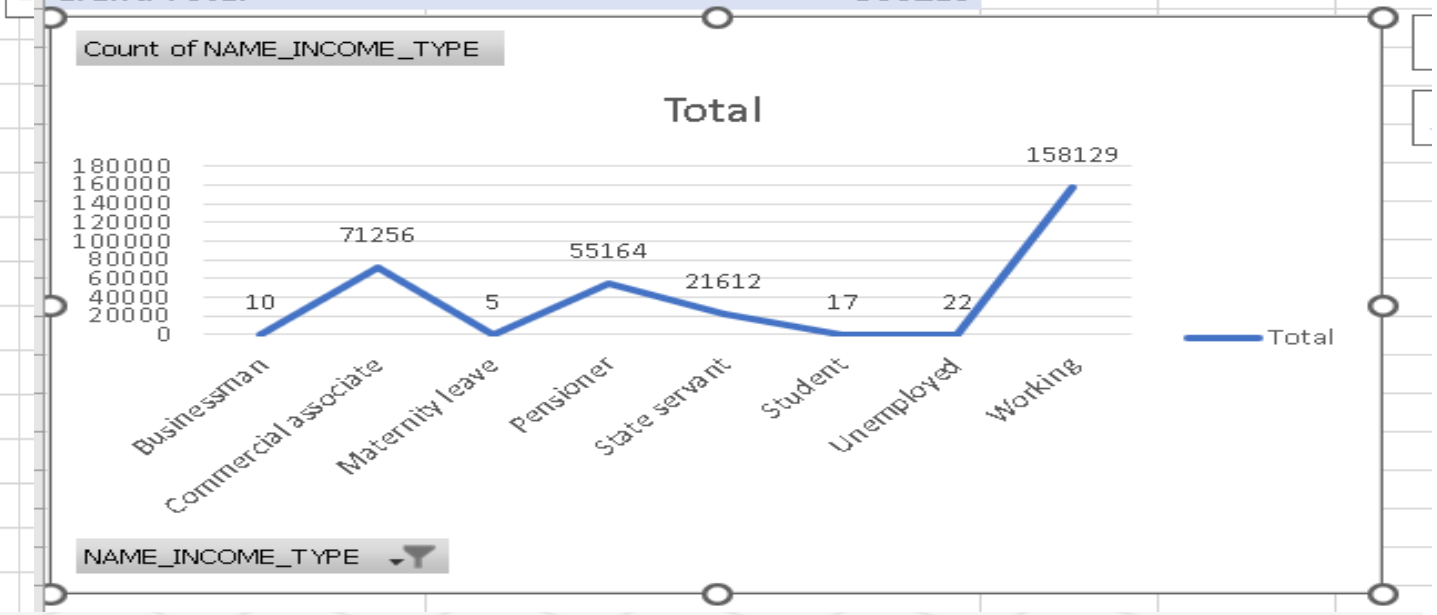
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	277225
Revolving loans	28990
Grand Total	**306215**



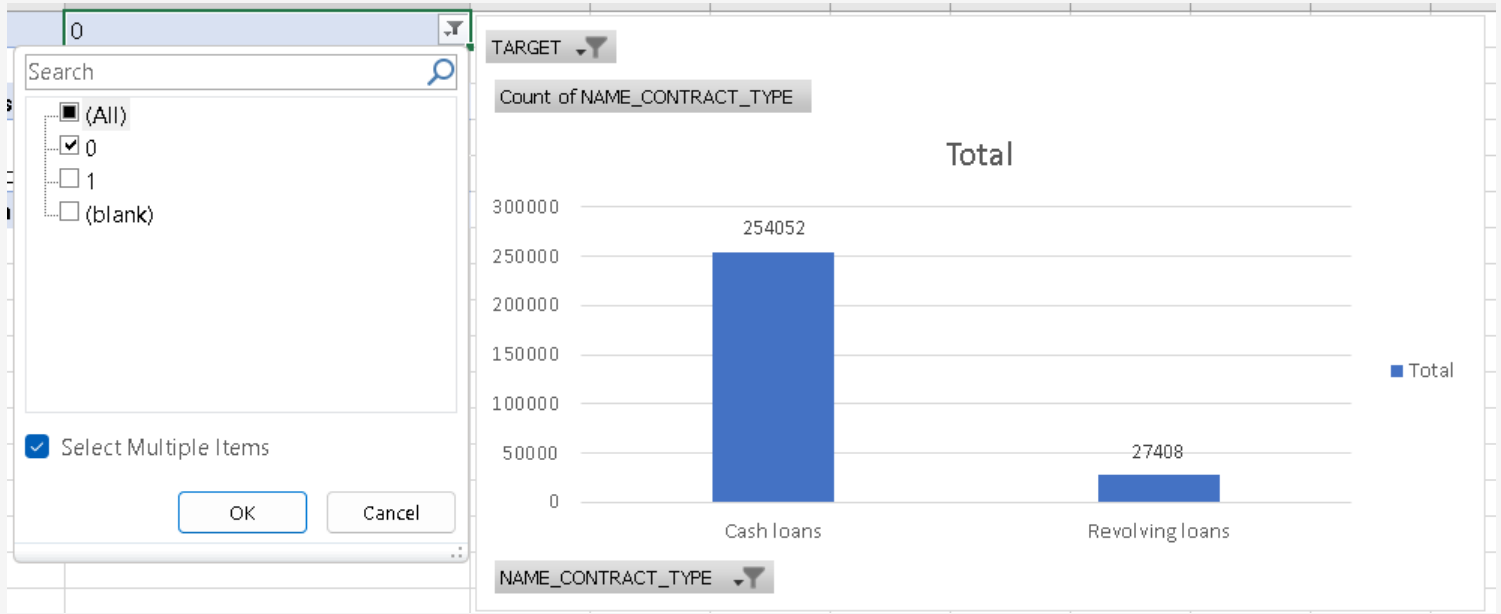
Row Labels	Count of CODE_GENDER
F	201643
M	104572
Grand Total	**306215**

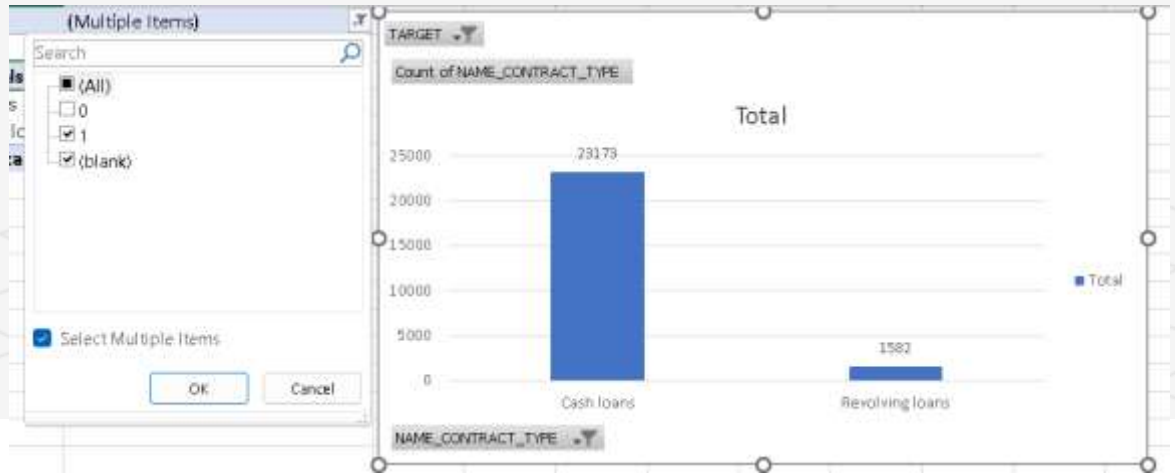


Row Labels	Count of NAME_INCOME_TYPE
Businessman	10
Commercial associate	71256
Maternity leave	5
Pensioner	55164
State servant	21612
Student	17
Unemployed	22
Working	158129
Grand Total	**306215**

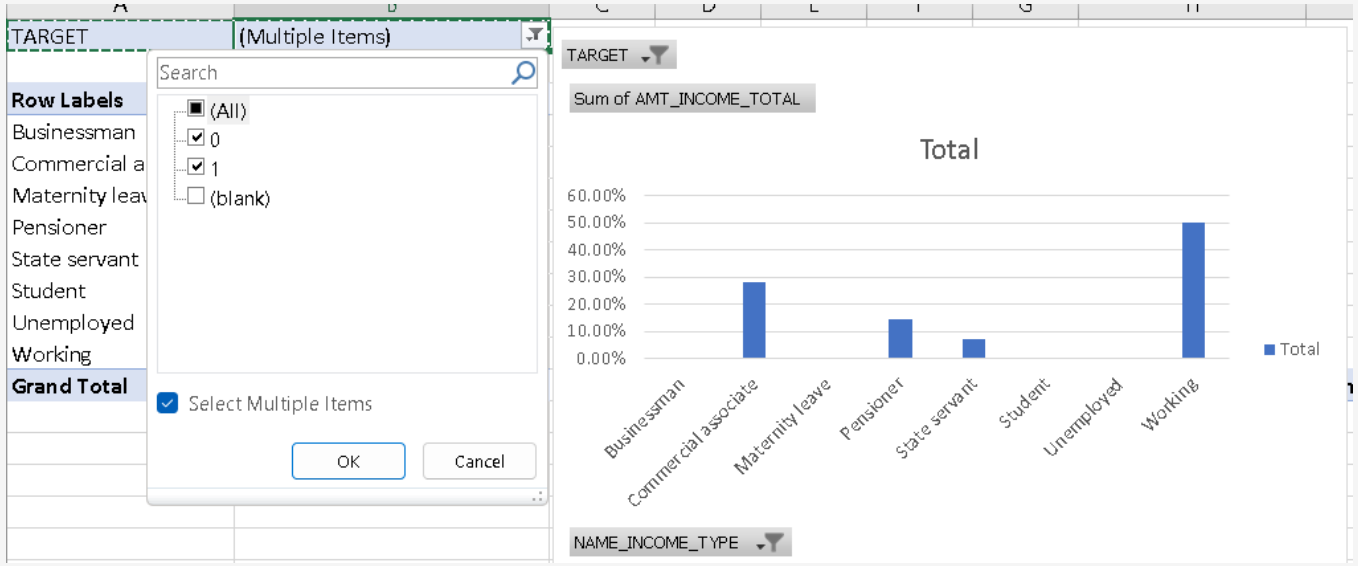


Segmented Univariate Analysis:-





Bivariate Analysis.

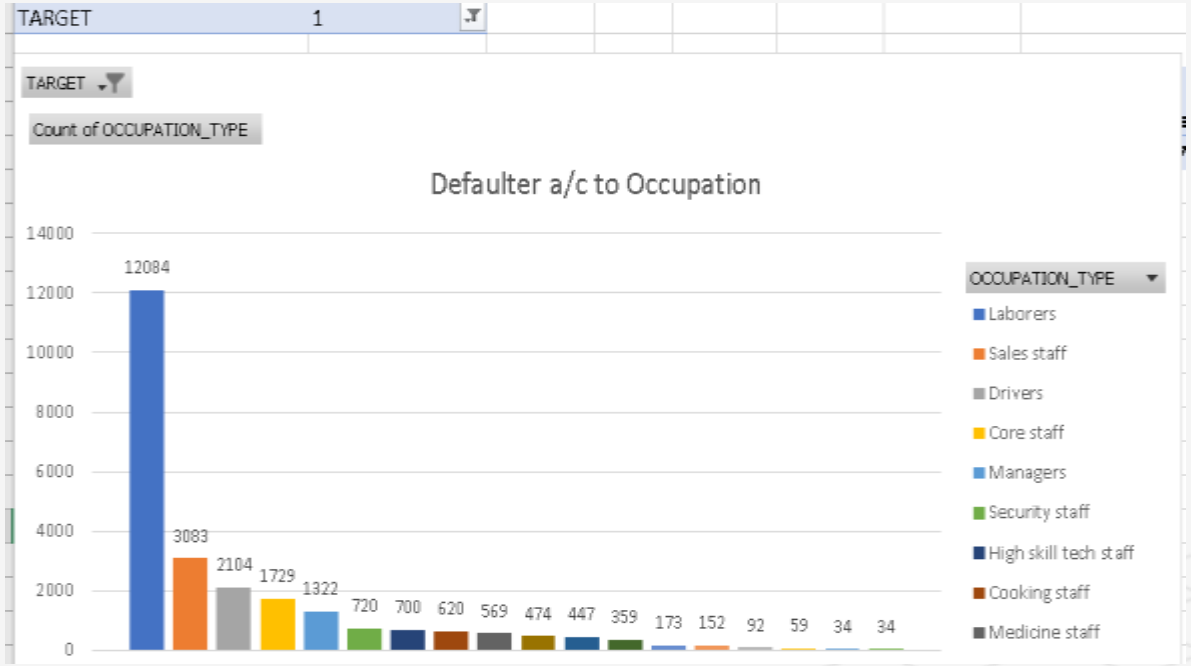


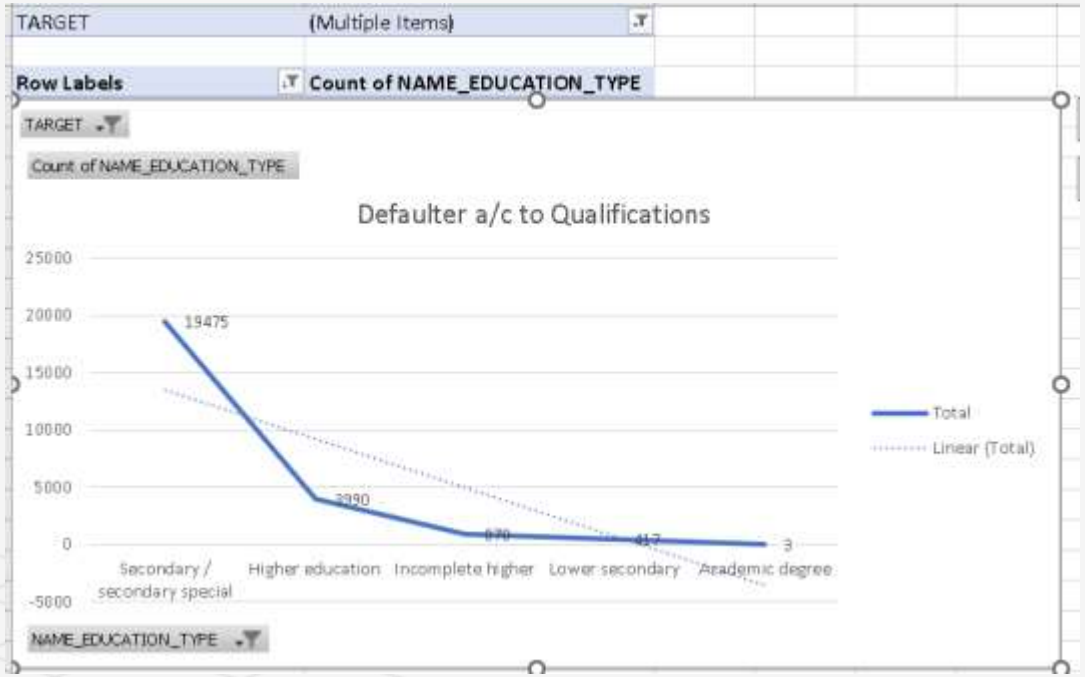
TARGET	0		
NAME_CONTRACT_TYPE	Cash loans		
Sum of AMT_ANNUITY	Column Labels		
Row Labels	F	M	Grand Total
100003	35698.5		35698.5
100006	29686.5		29686.5
100007		21865.5	21865.5
100008		27517.5	27517.5
100009	41301		41301
100010		42075	42075
100011	33826.5		33826.5
100014	21177		21177
100015	10678.5		10678.5
100016	5881.5		5881.5
100017		28966.5	28966.5
100018	32778		32778
100019		20160	20160
100020		26149.5	26149.5
100023	17563.5		17563.5
100025	37561.5		37561.5
100026	32521.5		32521.5
100027	23850		23850
100029		12703.5	12703.5
100030	11074.5		11074.5
100032		23827.5	23827.5
100033		57676.5	57676.5
100035	24592.5		24592.5

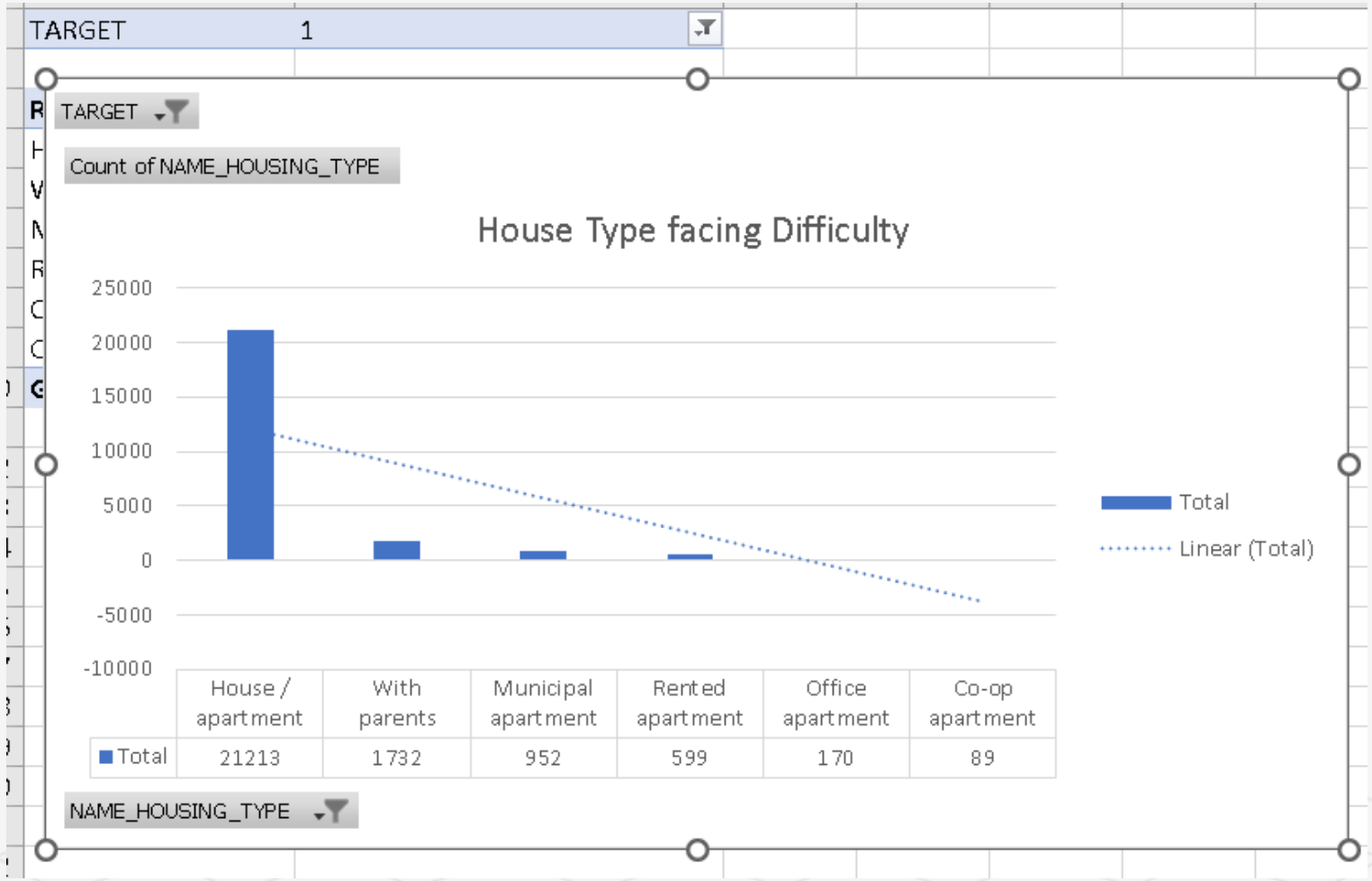
TARGET	(All)		
NAME_CONTRACT_TYPE	Revolving loans		
Sum of AMT_ANNUITY	**Column Labels**		
Row Labels	**F**	**M**	**Grand Total**
100004		6750	6750
100012		20250	20250
100021	13500		13500
100022	7875		7875
100024		21375	21375
100034		9000	9000
100046		27000	27000
100052	9000		9000
100058	6750		6750
100068		12375	12375
100079		13500	13500
100080	22500		22500
100088	6750		6750
100095	6750		6750
100098		13500	13500
100119	9000		9000
100126		9000	9000
100129	6750		6750
100134	9000		9000
100140	33750		33750
100143		13500	13500
100154		9000	9000

TARGET	(All)		
NAME_CONTRACT_TYPE	Cash loans		
Sum of AMT_ANNUITY	**Column Labels**		
Row Labels	**F**	**M**	**Grand Total**
100002		24700.5	24700.5
100003	35698.5		35698.5
100006	29686.5		29686.5
100007		21865.5	21865.5
100008		27517.5	27517.5
100009	41301		41301
100010		42075	42075
100011	33826.5		33826.5
100014	21177		21177
100015	10678.5		10678.5
100016	5881.5		5881.5
100017		28966.5	28966.5
100018	32778		32778
100019		20160	20160
100020		26149.5	26149.5
100023	17563.5		17563.5
100025	37561.5		37561.5
100026	32521.5		32521.5
100027	23850		23850
100029		12703.5	12703.5
100030	11074.5		11074.5
100031	27076.5		27076.5
100032		23827.5	23827.5

Find the top correlations for Client with payment difficulties and all other case.






TARGET 1
TARGET
Count of NAME_HOUSING_TYPE
House Type facing Difficulty
25000
20000
15000
10000
5000
0
-5000
-10000
Total
Linear (Total)
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
Total
21213
1732
952
599
170
89
NAME_HOUSING_TYPE

Insights:-After performing the analysis, we can rectify whether a client will repay the loan or not. Also, the people who are likely to face problem in loan repayment are labourers. Also, people with Secondary /secondary special education might face problem in loan repayment. Moreover, those who are living in house/apartment are facing difficulty in loan repayment (may be because of extra home loan, EMIs and so on). •NAME_EDUCATION_TYPE: Academic degree has less defaults.

NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education • NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot. •REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults. • OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.

• NAME_INCOME_TYPE: Student and Businessmen have no defaults. • REGION_RATING_CLIENT: RATING 1 is safer. • ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%. • DAYS_BIRTH: People above age of 50 have low probability of defaulting • DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate.

• AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default.•NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, buying garage are being repaid mostly. • CNT_CHILDREN: People with zero to two children tend to repay the loans.• CODE_GENDER: Men are at relatively higher default rate • NAME_FAMILY_STATUS: People who have civil marriage or who are single default a lot. •