BANK LOAN ANALYSIS

PROJECT DESCRIPTION – This project is about a case study relayed to a Bank Loan. We have to carry out an EDA (Exploratory Data Analysis). Based on our analysis, we will get the solution for required questions.

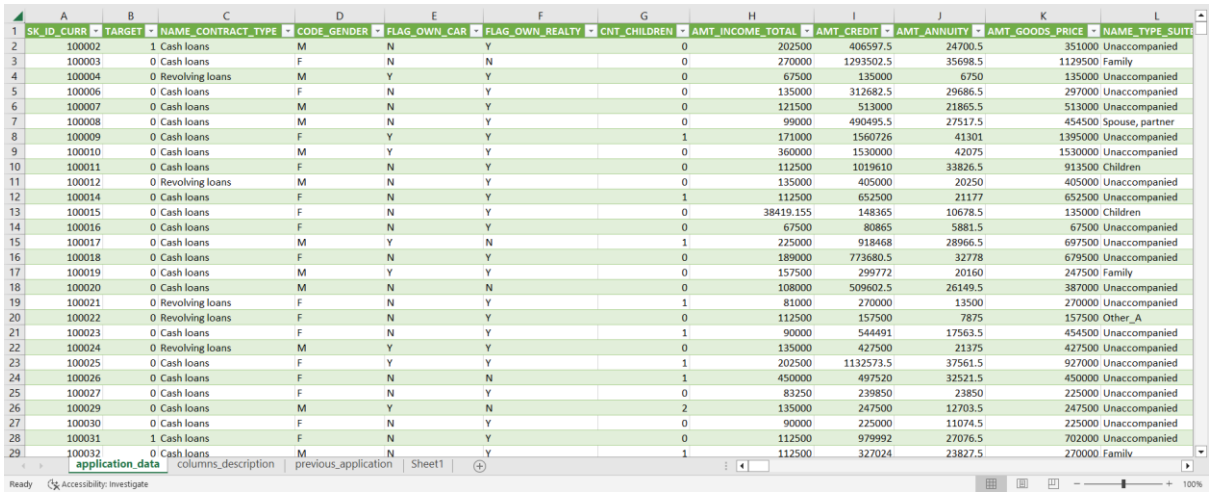
APPROACH – I first analyzed the data. While analyzing, I found out that data had a lot of missing values. So my first task was to get the missing values by performing mean, median and mode functions as required. So, I began by cleaning the data and then finding the outliers so as to make the data standardized.

TECH STACK USED – MS Excel 2019

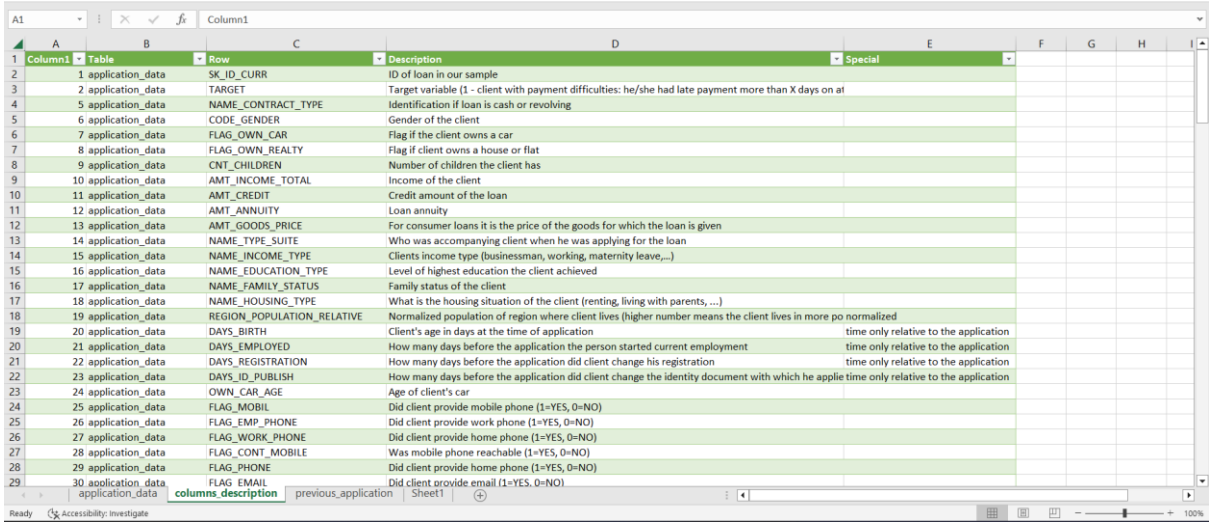
So lets begin with analysis……

1. Present the overall approach of the analysis. Mention the problem statement and the analysis approach briefly

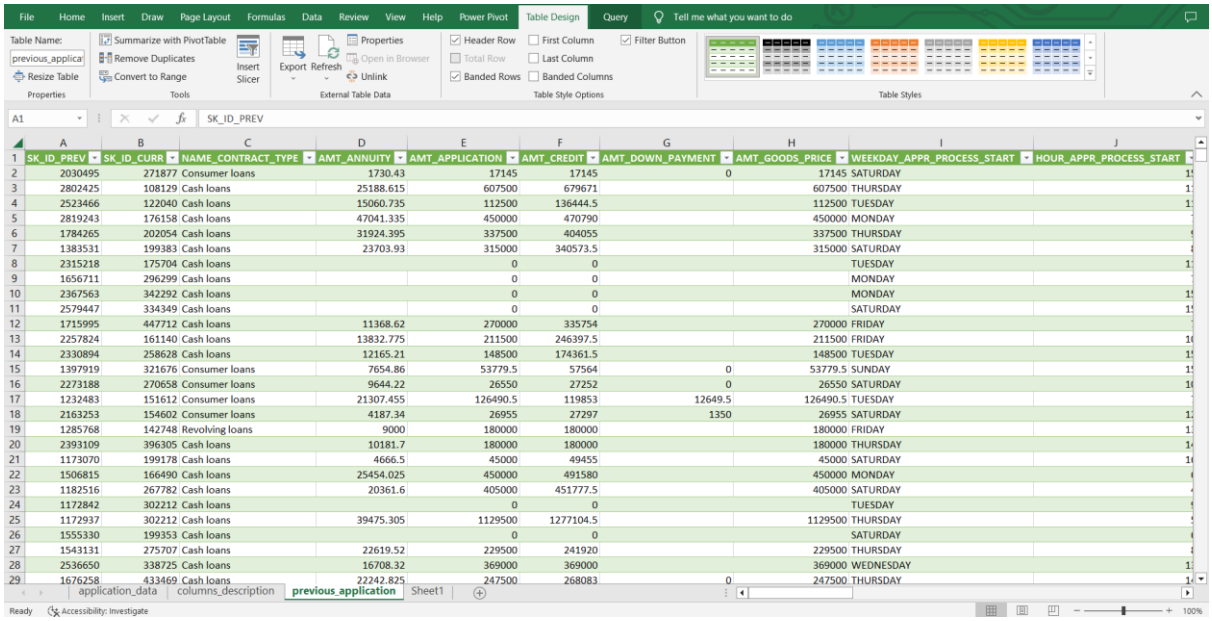
First we imported the data to excel.



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
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

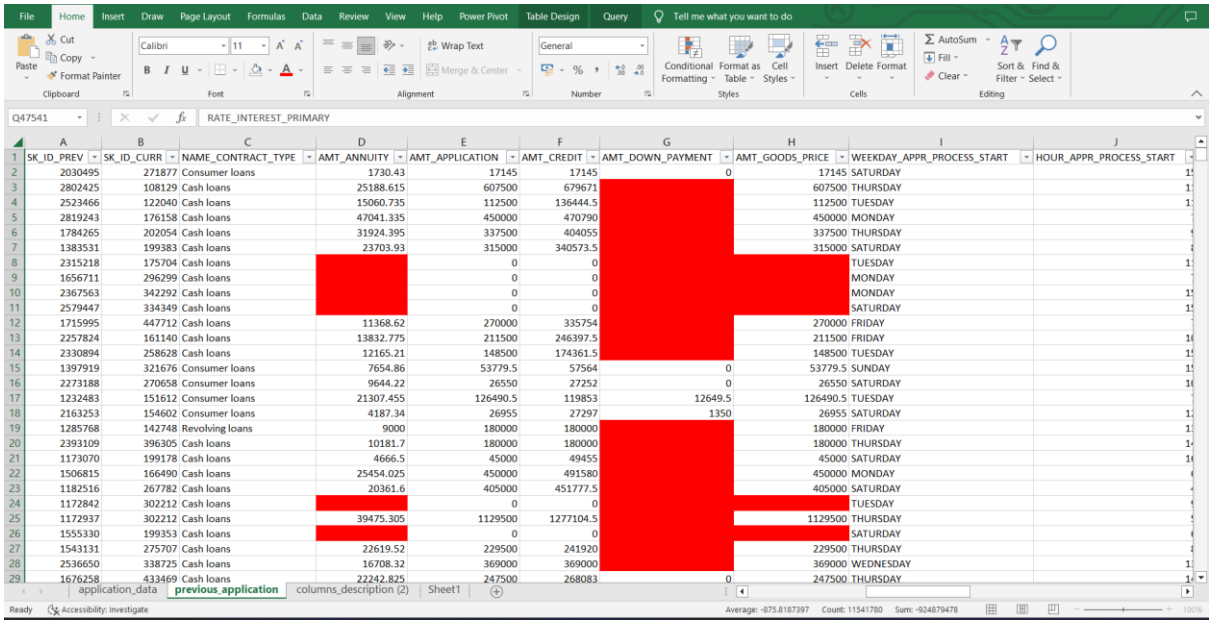


Column1	Table	Row	Description	Special
1	application_data	SK_ID_CURR	ID of loan in our sample	
2	application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at	
5	application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	
6	application_data	CODE_GENDER	Gender of the client	
7	application_data	FLAG_OWN_CAR	Flag if the client owns a car	
8	application_data	FLAG_OWN_REALTY	Flag if client owns a house or flat	
9	application_data	CNT_CHILDREN	Number of children the client has	
10	application_data	AMT_INCOME_TOTAL	Income of the client	
11	application_data	AMT_CREDIT	Credit amount of the loan	
12	application_data	AMT_ANNUITY	Loan annuity	
13	application_data	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given	
14	application_data	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan	
15	application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)	
16	application_data	NAME_EDUCATION_TYPE	Level of highest education the client achieved	
17	application_data	NAME_FAMILY_STATUS	Family status of the client	
18	application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)	
19	application_data	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more po	normalized
20	application_data	DAYS_BIRTH	Client's age in days at the time of application	time only relative to the application
21	application_data	DAYS_EMPLOYED	How many days before the application the person started current employment	time only relative to the application
22	application_data	DAYS_REGISTRATION	How many days before the application did client change his registration	time only relative to the application
23	application_data	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applie	time only relative to the application
24	application_data	OWN_CAR_AGE	Age of client's car	
25	application_data	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)	
26	application_data	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)	
27	application_data	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)	
28	application_data	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)	
29	application_data	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)	
30	application_data	FLAG_EMAIL	Did client provide email (1=YES, 0=NO)	

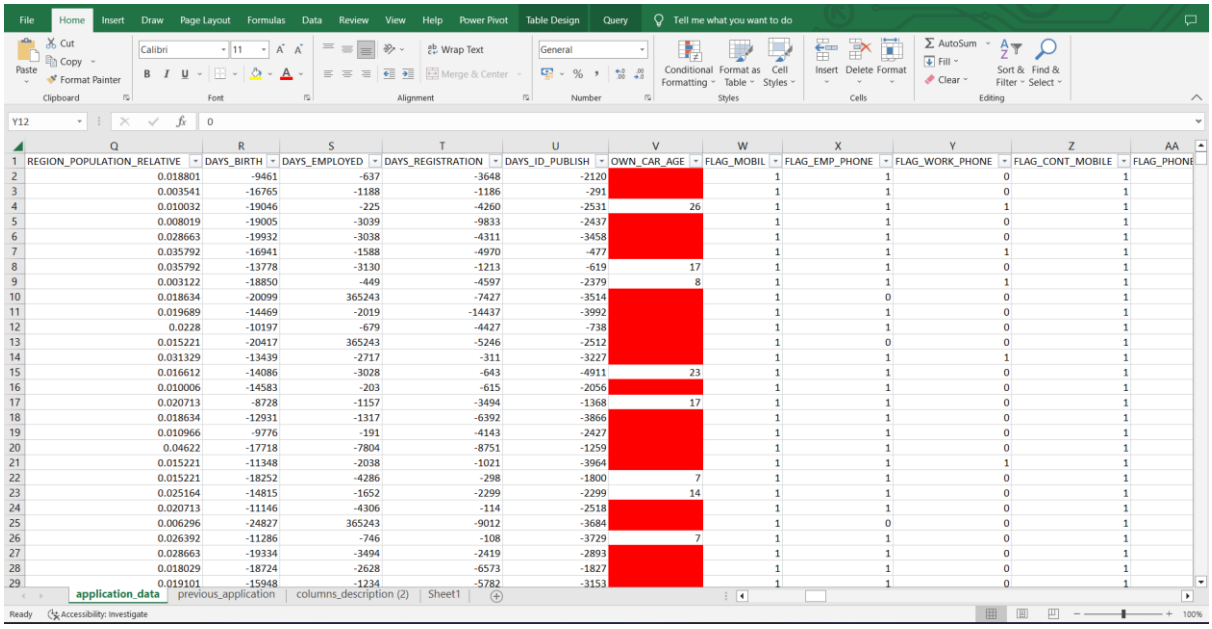


SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	11
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	11
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	9
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8
2315218	175704	Cash loans		0	0			TUESDAY	11
1656711	296299	Cash loans		0	0			MONDAY	
2367563	342292	Cash loans		0	0			MONDAY	15
2579447	334349	Cash loans		0	0			SATURDAY	15
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	16
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	15
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	10
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	11
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	6
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4
1172842	302212	Cash loans		0	0			TUESDAY	9
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	9
1555330	199353	Cash loans		0	0			SATURDAY	6
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8
2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESDAY	11
1676258	433469	Cash loans	22242.825	247500	268083	0	247500	THURSDAY	14

Then, in order to clean data I, highlighted the blank cells first .



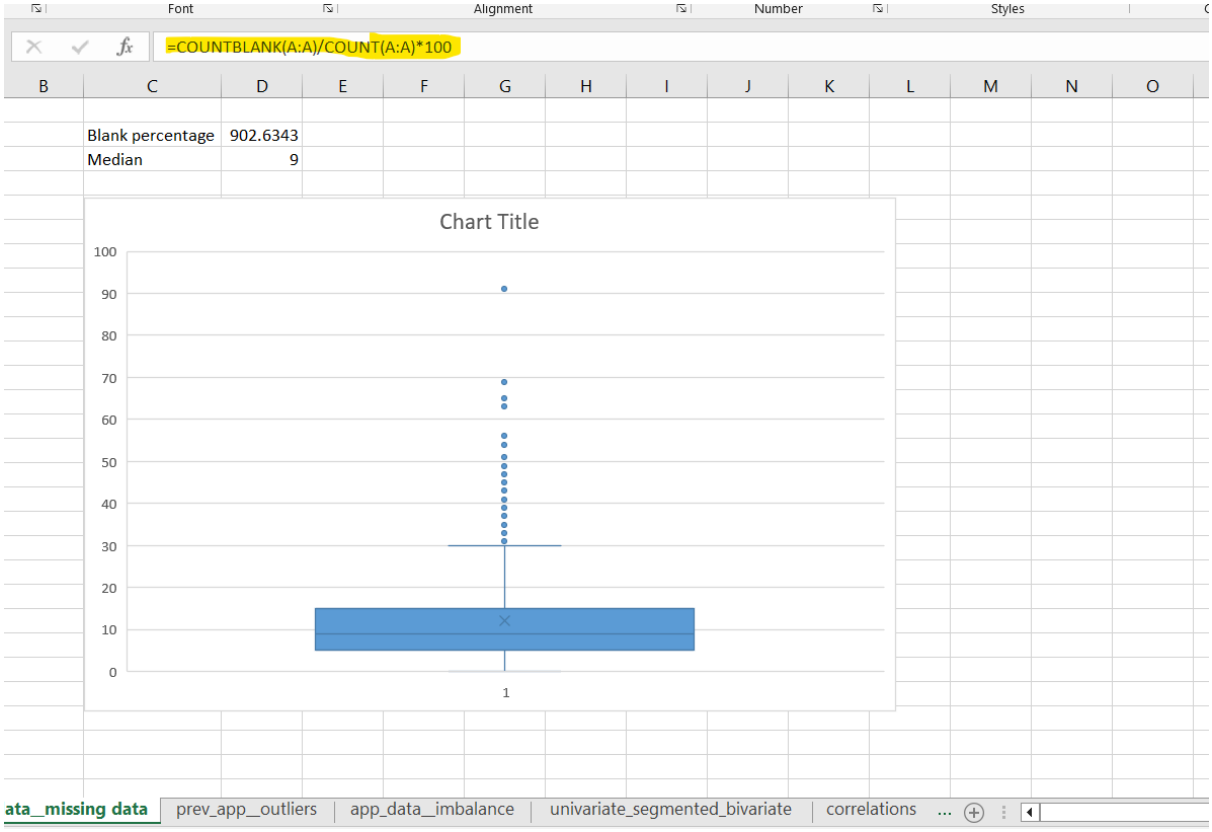
SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15
2802425	108129	Cash loans	25188.615	607500	679671		607500	THURSDAY	11
2523466	122040	Cash loans	15060.735	112500	136444.5		112500	TUESDAY	11
2819243	176158	Cash loans	47041.335	450000	470790		450000	MONDAY	
1784265	202054	Cash loans	31924.395	337500	404055		337500	THURSDAY	9
1383531	199383	Cash loans	23703.93	315000	340573.5		315000	SATURDAY	8
2315218	175704	Cash loans		0	0			TUESDAY	11
1656711	296299	Cash loans		0	0			MONDAY	
2367563	342292	Cash loans		0	0			MONDAY	15
2579447	334349	Cash loans		0	0			SATURDAY	15
1715995	447712	Cash loans	11368.62	270000	335754		270000	FRIDAY	
2257824	161140	Cash loans	13832.775	211500	246397.5		211500	FRIDAY	16
2330894	258628	Cash loans	12165.21	148500	174361.5		148500	TUESDAY	15
1397919	321676	Consumer loans	7654.86	53779.5	57564	0	53779.5	SUNDAY	15
2273188	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	10
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12
1285768	142748	Revolving loans	9000	180000	180000		180000	FRIDAY	11
2393109	396305	Cash loans	10181.7	180000	180000		180000	THURSDAY	14
1173070	199178	Cash loans	4666.5	45000	49455		45000	SATURDAY	16
1506815	166490	Cash loans	25454.025	450000	491580		450000	MONDAY	6
1182516	267782	Cash loans	20361.6	405000	451777.5		405000	SATURDAY	4
1172842	302212	Cash loans		0	0			TUESDAY	9
1172937	302212	Cash loans	39475.305	1129500	1277104.5		1129500	THURSDAY	9
1555330	199353	Cash loans		0	0			SATURDAY	6
1543131	275707	Cash loans	22619.52	229500	241920		229500	THURSDAY	8
2536650	338725	Cash loans	16708.32	369000	369000		369000	WEDNESDAY	11
1676258	433469	Cash loans	22242.825	247500	268083	0	247500	THURSDAY	14



REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	OWN_CAR_AGE	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE
0.018801	-9461	-637	-3648	-2120		1	1	0	1	
0.003541	-16765	-1188	-1186	-291		1	1	0	1	
0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	
0.008019	-19005	-3039	-9833	-2437		1	1	0	1	
0.028663	-19932	-3038	-4311	-3458		1	1	0	1	
0.035792	-16941	-1588	-4970	-477		1	1	1	1	
0.035792	-13778	-3130	-1213	-619	17	1	1	0	1	
0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	
0.018634	-20099	365243	-7427	-3514		1	0	0	1	
0.019689	-14469	-2019	-14437	-3992		1	1	0	1	
0.0228	-10197	-679	-4427	-738		1	1	0	1	
0.015221	-20417	365243	-5246	-2512		1	0	0	1	
0.031329	-13439	-2717	-311	-3227		1	1	1	1	
0.016612	-14086	-3028	-643	-4911	23	1	1	0	1	
0.010006	-14583	-203	-615	-2056		1	1	0	1	
0.020713	-8728	-1157	-3494	-1368	17	1	1	0	1	
0.018634	-12931	-1317	-6392	-3866		1	1	0	1	
0.010966	-9776	-191	-4143	-2427		1	1	0	1	
0.04622	-17718	-7804	-8751	-1259		1	1	0	1	
0.015221	-11348	-2038	-1021	-3964		1	1	1	1	
0.015221	-18252	-4286	-298	-1800	7	1	1	0	1	
0.025164	-14815	-1652	-2299	-2299	14	1	1	0	1	
0.020713	-11146	-4306	-114	-2518		1	1	0	1	
0.006296	-24827	365243	-9012	-3684		1	0	0	1	
0.026392	-11286	-746	-108	-3729	7	1	1	0	1	
0.028663	-19334	-3494	-2419	-2893		1	1	0	1	
0.018029	-18724	-2628	-6573	-1827		1	1	0	1	
0.019101	-15948	-1234	-5782	-3153		1	1	0	1	

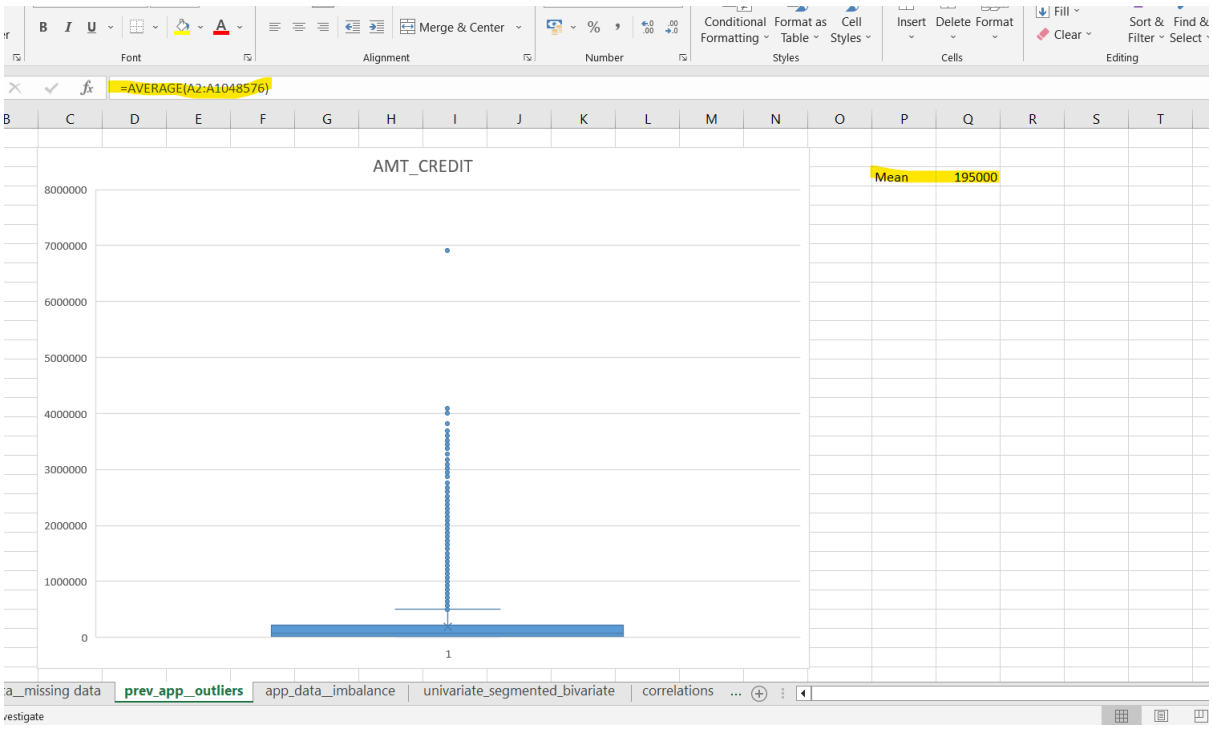
2. **Identify** the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

I found out the blank percentage and median of the column and filled the empty spaces there. (This is just for one table. Actual cleaning and filling of data is shown in excel file attached for other columns).

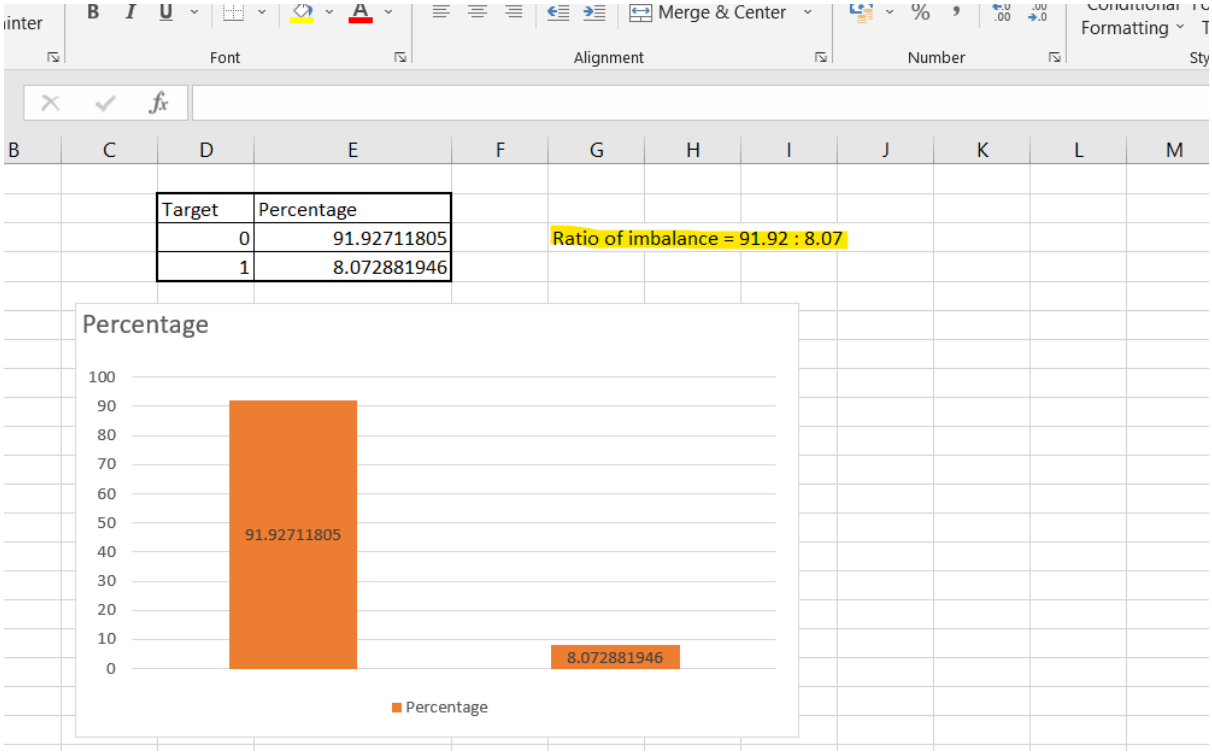


3. Identify if there are **outliers** in the dataset. Also, mention why do you think it is an outlier.

Foe Numerical columns, I found out the outliers and chose the value for the upper whisker as shown below. The credit amount value above 195000 is considered to be an upper whisker.



4. Identify if there is data imbalance in the data. Find the ratio of data imbalance. **The ratio of imbalance for Target Table came out to be 91.92:8.07.**

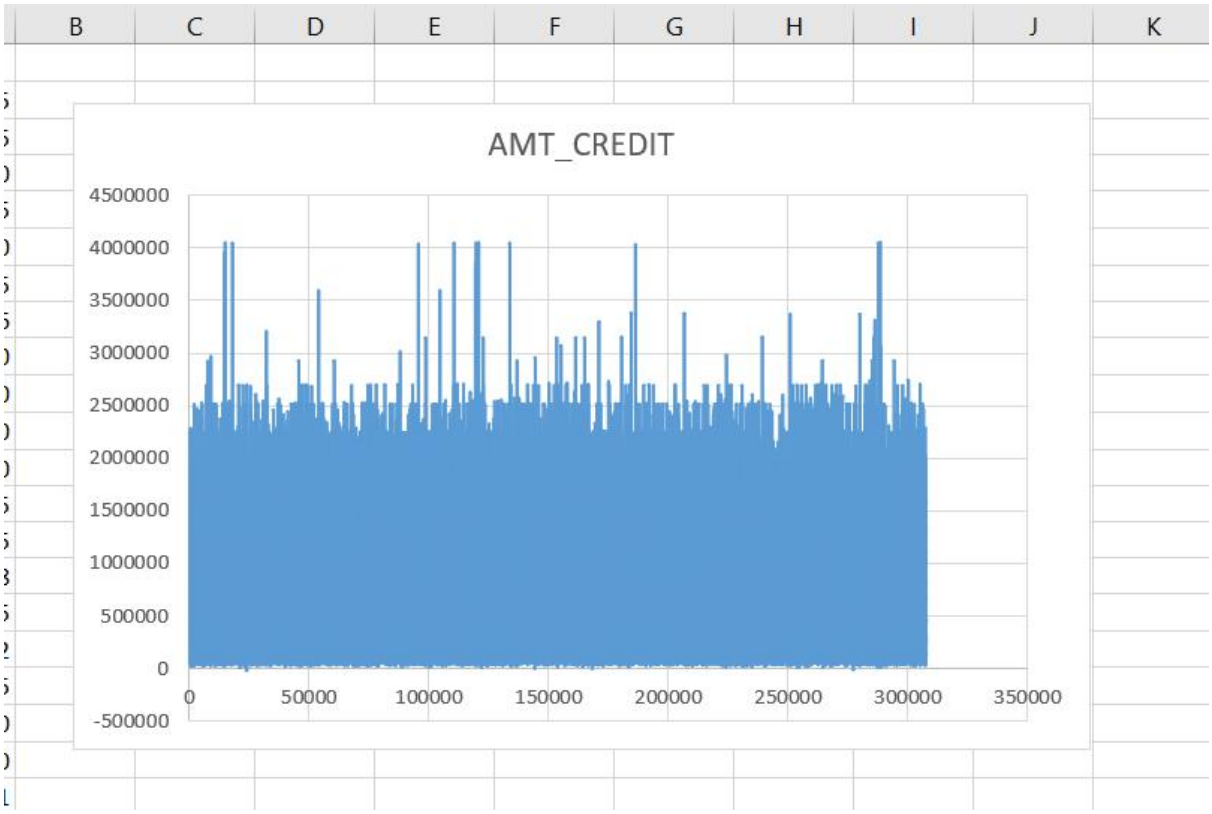


5. Explain the **results of univariate, segmented univariate, bivariate analysis, etc.** in business terms.

The results of **univariate, segmented univariate, bivariate analysis** are as follows –

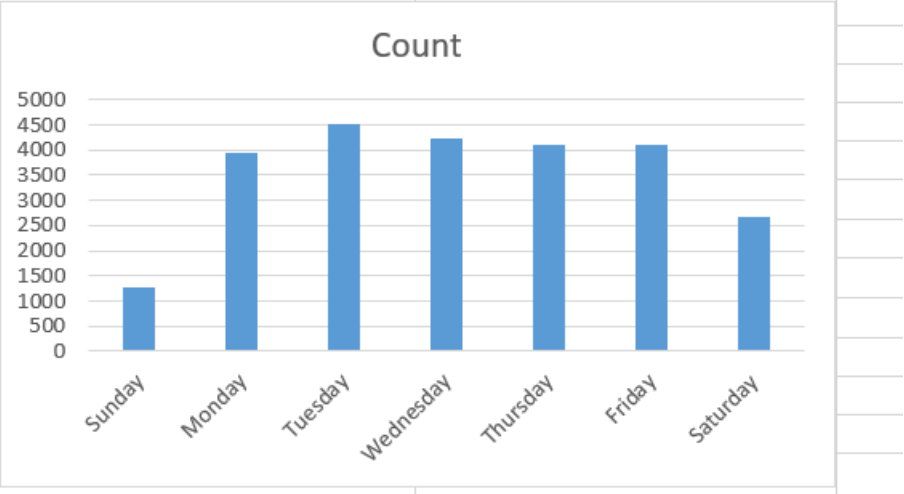
To perform the analysis, I first divided the data into two sets i.e. Target - 0 and Target – 1

AMT_CREDIT

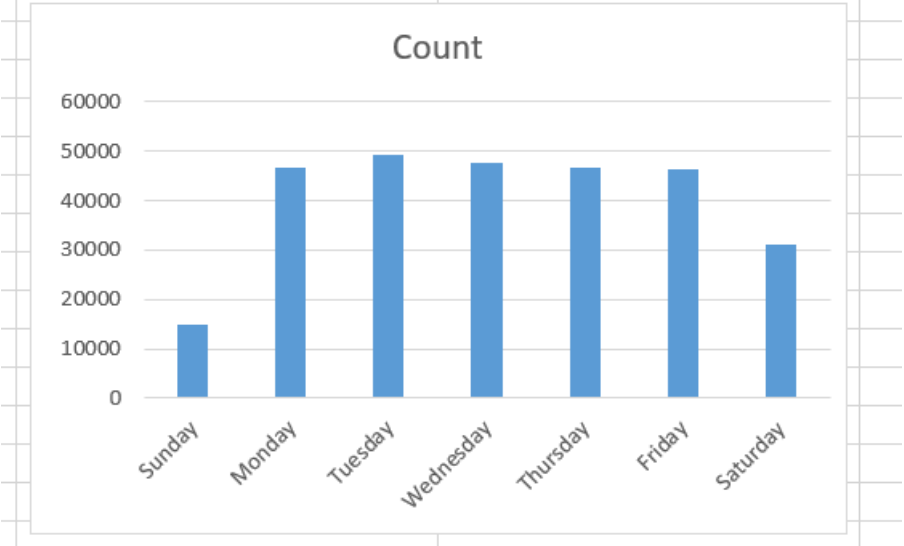


WEEKDAY_APPR_PROCESS_START

Target - 1	
WEEKDAY_APPR_PROCESS_START	**Count**
Sunday	1283
Monday	3934
Tuesday	4501
Wednesday	4238
Thursday	4098
Friday	4101
Saturday	2670



Target - 0	
WEEKDAY_APPR_PROCESS_START	**Count**
Sunday	14898
Monday	46780
Tuesday	49400
Wednesday	47696
Thursday	46493
Friday	46237
Saturday	31182



INSIGHTS – We can conclude that application starting process is less on Saturday and Sunday.

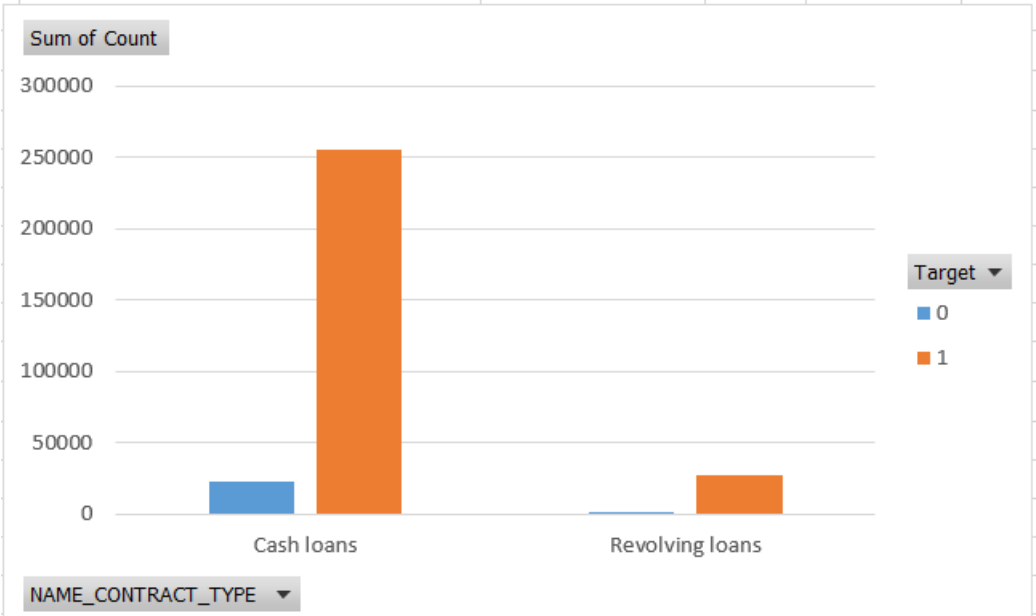
NAME_CONTRACT_TYPE

Target - 1	
NAME_CONTRACT_TYPE	Count
Cash loans	23221
Revolving loans	1604

Target - 0	
NAME_CONTRACT_TYPE	Count
Cash loans	255011
Revolving loans	27675

NAME_CONTRACT_TYPE	Count	Target
Cash loans	255011	1
Revolving loans	27675	1
Cash loans	23221	0
Revolving loans	1604	0

Sum of Count	Column Labels		
Row Labels	0	1	Grand Total
Cash loans	23221	255011	278232
Revolving loans	1604	27675	29279
Grand Total	**24825**	**282686**	**307511**



INSIGHTS – We can conclude that people prefer cash type loans more than other. People take more cash loans.

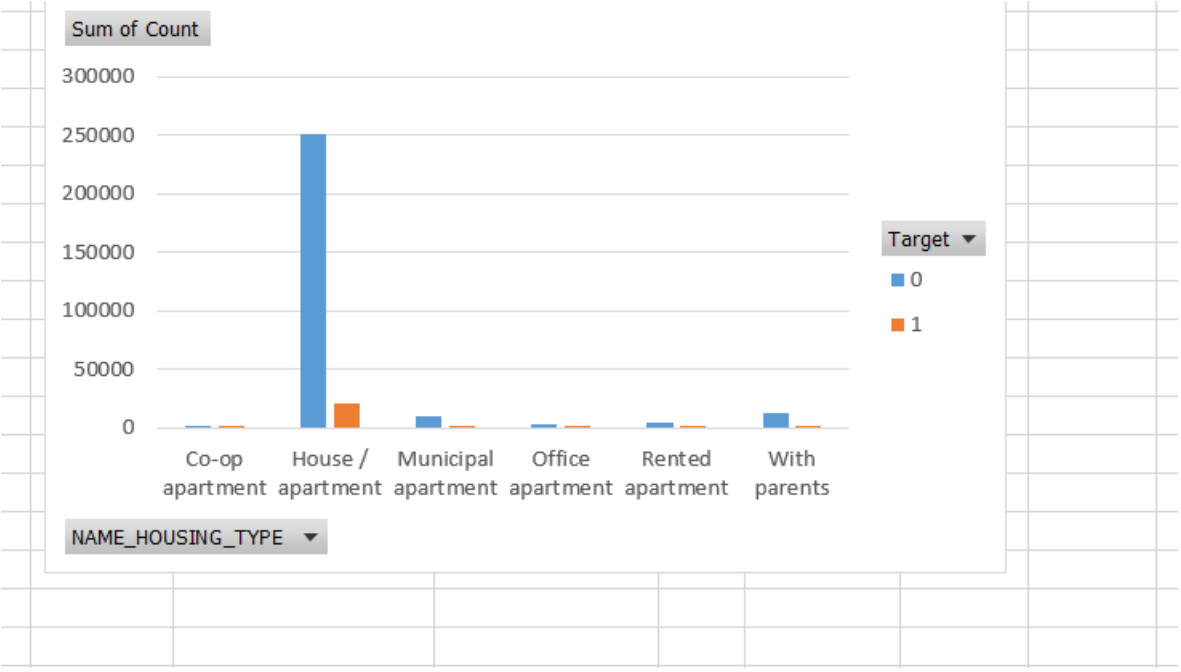
NAME_HOUSING_TYPE

Target - 1	
NAME_HOUSING_TYI	Count
House / apartment	21272
Co-op apartment	89
Municipal apartment	955
Office apartment	172
Rented apartment	601
With parents	1736

Target - 0	
NAME_HOUSING_TYI	Count
House / apartment	251596
Co-op apartment	1033
Municipal apartment	10228
Office apartment	2445
Rented apartment	4280
With parents	13104

NAME_HOUSING_TYI	Count	Target
House / apartment	21272	1
Co-op apartment	89	1
Municipal apartment	955	1
Office apartment	172	1
Rented apartment	601	1
With parents	1736	1
House / apartment	251596	0
Co-op apartment	1033	0
Municipal apartment	10228	0
Office apartment	2445	0
Rented apartment	4280	0
With parents	13104	0

Sum of Count	Column Labels		
Row Labels	0	1	Grand Total
Co-op apartment	1033	89	1122
House / apartment	251596	21272	272868
Municipal apartment	10228	955	11183
Office apartment	2445	172	2617
Rented apartment	4280	601	4881
With parents	13104	1736	14840
Grand Total	**282686**	**24825**	**307511**



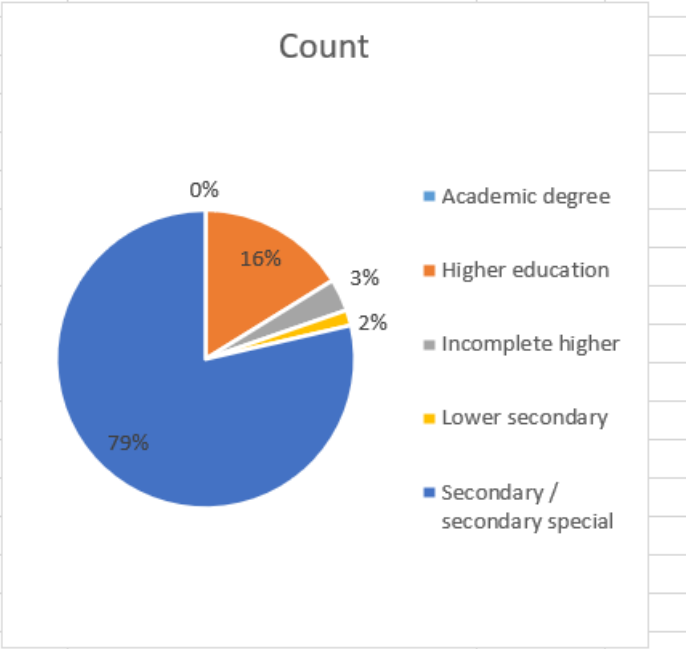
INSIGHTS – We can conclude that people living in houses fall in both the category of default loans and non-default loans.

6. Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable).

To find the correlation, we again divide the data into two sets based on Targets and consider Target – 1 as defaulters.

NAME_EDUCATION_TYPE

NAME_EDUCATION_TYPE	Count
Academic degree	3
Higher education	4009
Incomplete higher	872
Lower secondary	417
Secondary / secondary special	19524



INSIGHTS – We can find that people with education type as Secondary/Secondary Special are more likely to default and people with education type Academic degree default the least.

NAME_CONTRACT_TYPE and AMT_APPLICATION

Row Labels	Sum of AMT_APPLICATION
Cash loans	1.27622E+11
Consumer loans	43263132837
Revolving loans	11849616540
XNA	0
Grand Total	**1.82735E+11**

INSIGHTS – If we sum the total amount for loan in applications, we find that that people mostly take cash loans.

AMT_APPLICATION and AMT_CREDIT

AMT_APPLICATION	AMT_CREDIT			
17145	17145		Correlation Coefficient	
607500	679671		0.975777217	
112500	136444.5			
450000	470790			
337500	404055			
315000	340573.5			
0	0			
0	0			

INSIGHTS – We find that the correlation coefficient is 0.9758 using excel formula =CORREAL.

AMT_INCOME_TOTAL and AMT_ANNUITY

W	X	Y	Z
AMT_INCOME_TOTAL	AMT_ANNUITY		
202500	24700.5		Correlation Coefficient
270000	35698.5		0.191657428
67500	6750		
135000	29686.5		
121500	21865.5		
99000	27517.5		

INSIGHTS – We find that the correlation coefficient is 0.19166 using excel formula =CORREAL.

CONCLUSION – From the above analysis, we can find out what kind of people and can repay loan, what kinds of loan people prefer to take, people taking loans come from what background, what is their source of income, for what type of people, the loan applications are refused and based on which conditions.

RESULTS: -

1. People with academic degree have less defaults.
2. People prefer cash loans more than any other type.
3. People with secondary/secondary special as education type have more chances of defaulting loans.
4. People who have less than 5 years of employment have high default rate.
5. Focused variable for application file – Target.
6. Focused variable for Previous application file – NAME_CONTRACT_STATUS.
7. Important fields to consider for loan repayment are –
8. NAME_EDUCATION_TYPE
9. AMT_INCOME_TOTAL
10. DAYS_EMPLOYED
11. AMT_CREDIT
12. People with lower total income are more likely to default.
13. People with high Credit amount are less likely to default.

NOTE :- All columns shown above are just 1 example in screenshot. For exact results please find excel file attached. Drive link for file -
https://docs.google.com/spreadsheets/d/1gn2C2nSJdSSGcqeoeZQCG 3gRtdYRt/edit?usp=share_link&ouid=112715989555881480949&rtpof=true&sd=true

THANK YOU