Bank Loan Case Study

Table of content

1. Introduction

Bank Loan Case Study, in this project it give an idea of applying exploratory data analysis (EDA) in a real business scenario. Further develop a basic understanding of risk analytics in banking and financial services and understanding how data is used to minimise the risk of losing money while lending to customer. We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. So it would help the loan providing companies to decide for loan approval based on the applicant's profile. Tow tables are given for this type case study first one is `application_data.csv` it contains all the information of the client at the time of application. The data is about whether a client has payment difficulties. Next one is `previous_application.csv` it contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Doing EDA In this projects Bank loan case study, It aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

2. Tech-stack used

In this project I have used:

- Microsoft Office Excel 2007

The purpose for I used Microsoft Office Excel 2007 in this project because it work really well for simple calculations and tracking almost any kind of information. The cells can contain numbers, text, or formula.

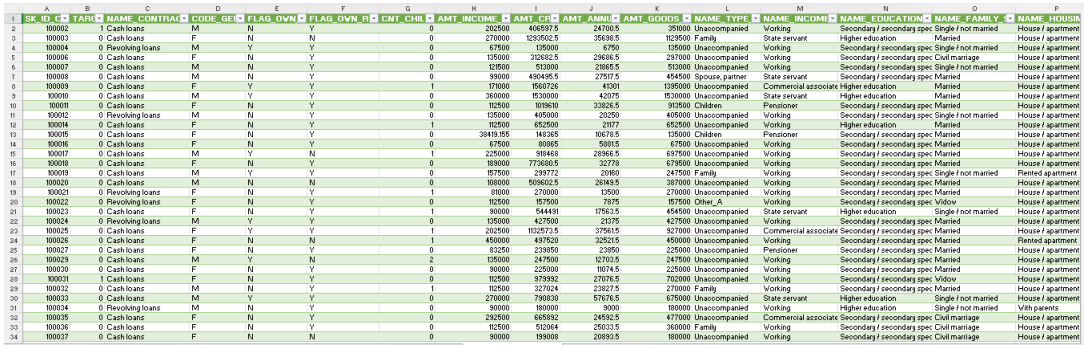
3. Approach

I used Microsoft office excel for implementing this project. Firstly I have imported the given datasets into Microsoft excel. Analysed each column form the table and its attribute, also checked the connection with other columns. I have find null values and duplicate values then I remove null values and duplicates for better results, I removed unwanted columns also. Pivot table helps a lot for doing this project. It helps to connect different columns and helps to perform the task.

4. Exploratory data analysis

1. application_data.

A. Cleaning



SK_ID_C	TARG	NAME_CONTRAC	CODE_GE	FLAG_OWN	FLAG_OWN_R	CNT_CHIL	AMT_INCOME	AMT_CR	AMT_ANNU	AMT_GOODS	NAME_TYPE	NAME_INCOME	NAME_EDUCATION	NAME_FAMILY_S	NAME_HOUSIN
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompanied	Working	Secondary / secondary spec	Single / not married	House / apartment
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher education	Married	House / apartment
100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Unaccompanied	Working	Secondary / secondary spec	Single / not married	House / apartment
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompanied	Working	Secondary / secondary spec	Civil marriage	House / apartment
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompanied	Working	Secondary / secondary spec	Single / not married	House / apartment
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, partner	State servant	Secondary / secondary spec	Married	House / apartment
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompanied	Commercial associate	Higher education	Married	House / apartment
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompanied	State servant	Higher education	Married	House / apartment
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary / secondary spec	Married	House / apartment
100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Unaccompanied	Working	Secondary / secondary spec	Single / not married	House / apartment
100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Unaccompanied	Working	Higher education	Married	House / apartment
100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Children	Pensioner	Secondary / secondary spec	Married	House / apartment
100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Family	Working	Secondary / secondary spec	Single / not married	Rented apartment
100020	0	Cash loans	M	N	N	0	108000	509602.5	26949.5	387000	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Other_A	Working	Secondary / secondary spec	Widow	House / apartment
100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Unaccompanied	State servant	Higher education	Single / not married	House / apartment
100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100025	0	Cash loans	F	Y	Y	1	202500	1132573.5	37561.5	927000	Unaccompanied	Commercial associate	Secondary / secondary spec	Married	House / apartment
100026	0	Cash loans	F	N	N	1	450000	497520	32521.5	450000	Unaccompanied	Working	Secondary / secondary spec	Married	Rented apartment
100027	0	Cash loans	F	N	Y	0	83250	239850	23850	225000	Unaccompanied	Pensioner	Secondary / secondary spec	Married	House / apartment
100029	0	Cash loans	M	Y	N	2	135000	247500	12703.5	247500	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100030	0	Cash loans	F	N	Y	0	90000	225000	11074.5	225000	Unaccompanied	Working	Secondary / secondary spec	Married	House / apartment
100031	1	Cash loans	F	N	Y	0	112500	979992	27076.5	702000	Unaccompanied	Working	Secondary / secondary spec	Widow	House / apartment
100032	0	Cash loans	M	N	Y	1	112500	327024	23827.5	270000	Family	Working	Secondary / secondary spec	Married	House / apartment
100033	0	Cash loans	M	Y	Y	0	270000	790830	57676.5	675000	Unaccompanied	State servant	Higher education	Single / not married	House / apartment
100034	0	Revolving loans	M	N	Y	0	90000	180000	9000	180000	Unaccompanied	Working	Higher education	Single / not married	With parents
100035	0	Cash loans	F	N	Y	0	292500	665892	24592.5	477000	Unaccompanied	Commercial associate	Secondary / secondary spec	Civil marriage	House / apartment
100036	0	Cash loans	F	N	Y	0	112500	512064	25033.5	360000	Family	Working	Secondary / secondary spec	Civil marriage	House / apartment
100037	0	Cash loans	F	N	N	0	90000	193008	20893.5	180000	Unaccompanied	Working	Secondary / secondary spec	Civil marriage	House / apartment

Table 4.1.1

Finding null values:

- for cleaning process we have to identify null values and remove it.
- The table contain 122 columns and 307511 rows. Some of the columns may have contain null values
- I used filter function to find null values in each columns. If the columns contain 50% null values. The columns should deleted
- I used those formulas for this purpose

  =countblank(range)

  =counta(range)

- Delete those columns which contain null values more than 50%. It include those columns like occupation type, EXT-source 1,apartment-avg, basement-avg, fondkapremont_mode, housetype_mode , wallsmaterial_mode.
- Also I replace all the numerical missing values with median, because when the data is skewed, it is good to consider using the median values for replacing the missing values .

- Categorical missing values are low, so when found the mi ssing values I deleted in row vice.

Finding outliers:

- Here box blot is mainly used for finding outliers

AMT_INCOME_TOTAL

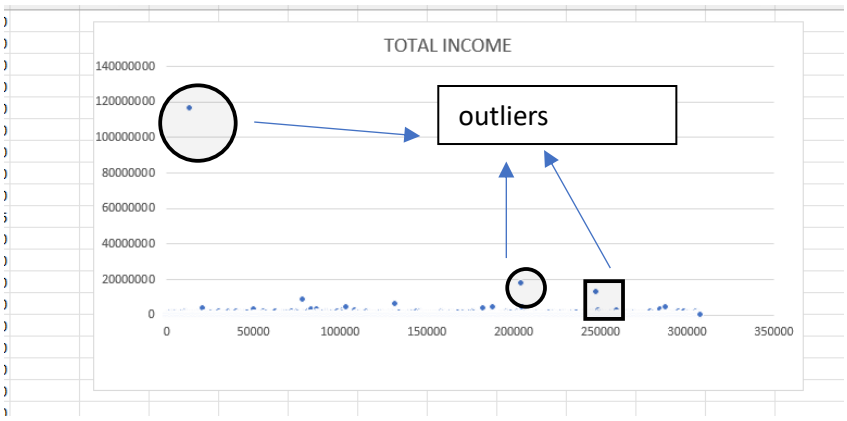


Figure 4.1.1

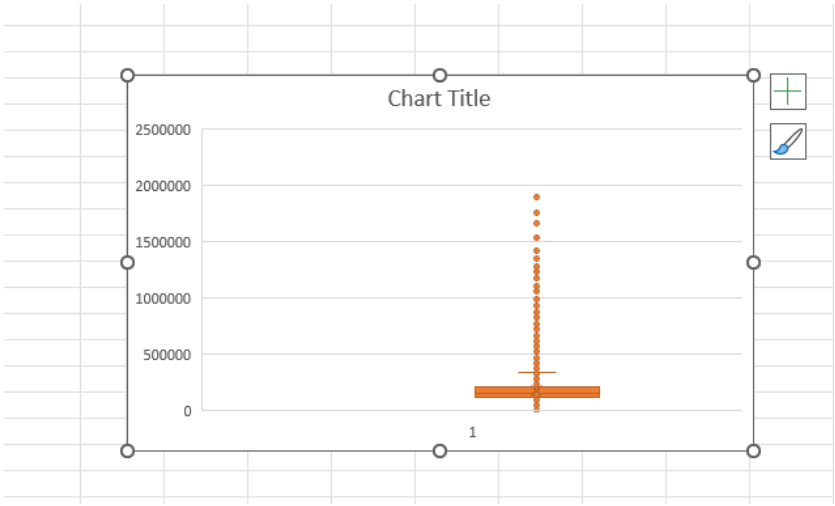


Figure 4.1.2

- From figure 4.1 scatter plot is used. The graph is not pretty clear, but we can notice some outliers there.
- Figure 4.2 shows box plot. We can notice here so many outliers are there
- Values greater than the upper bond is called outliers

CNT_CHILDREN

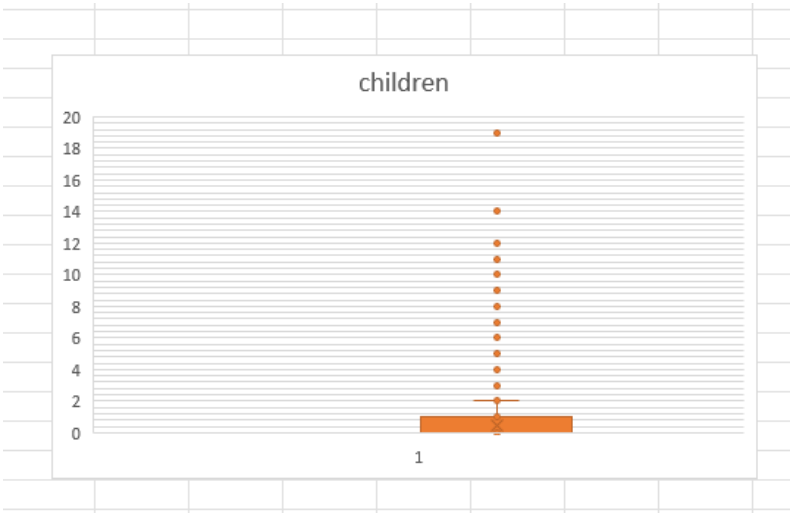


Figure 4.1.3

- From figure 4.3. the box plot shows outliers from columns CNT_CHILDREN, there are few outliers in this column we can see that in this graph.

AMT_CREDIT:

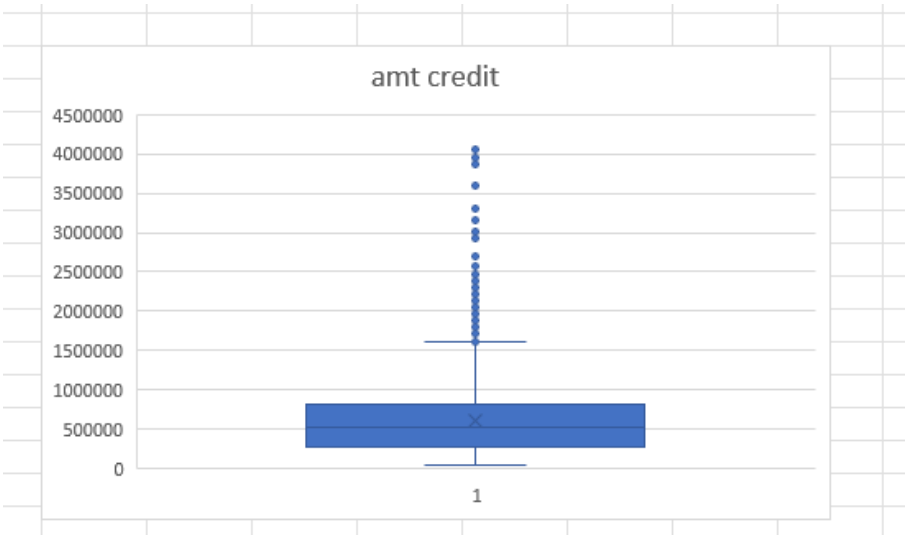


Figure 4.1.4

- From the above figure 4.4. this box plot represent the column amt credit there are outliers in this column also

AMT_ANNUITY:

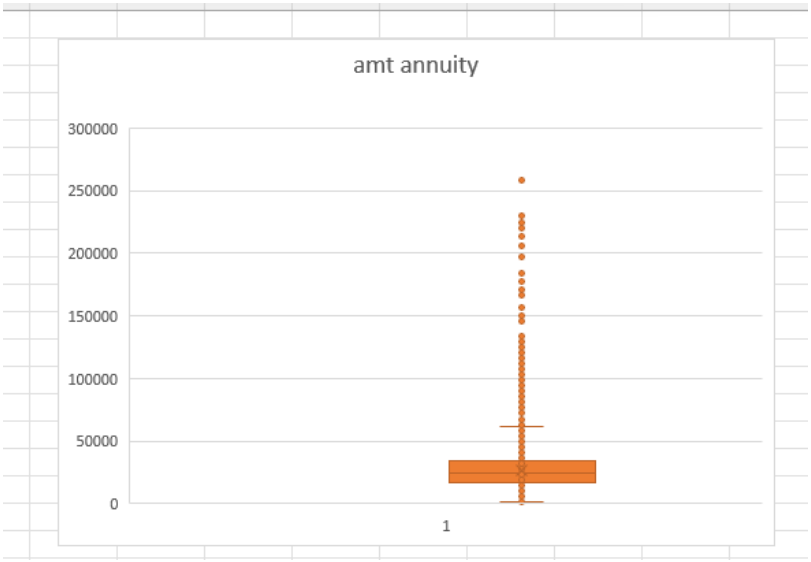


Figure 4.1.5

- from the above figure 4.5 we can infer that there are lot of outliers vested in this graph

AGE:

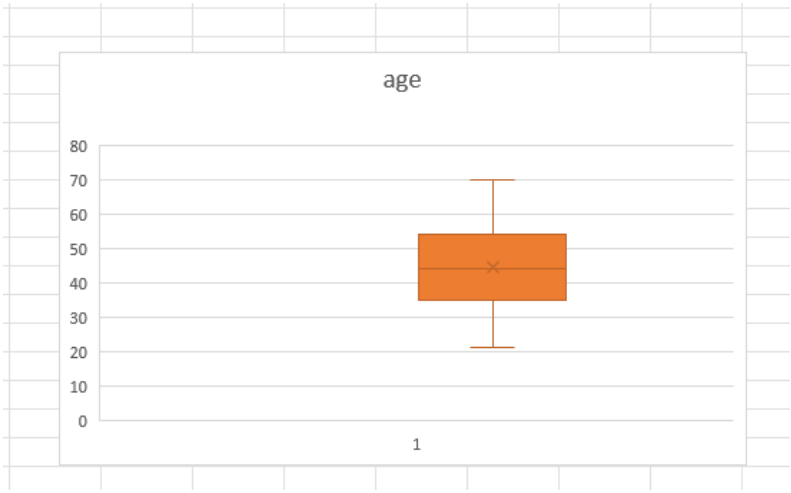


Figure 4.1.6

- from the above graph we can see that there is no outliers spotted from the graph. Its clean box plot plotted over here

DAYS EMPLOYED:

Figure 4.1.7

DAYS REGISTRATION:

Figure 4.1.8

DAYS ID PUBLISHED:

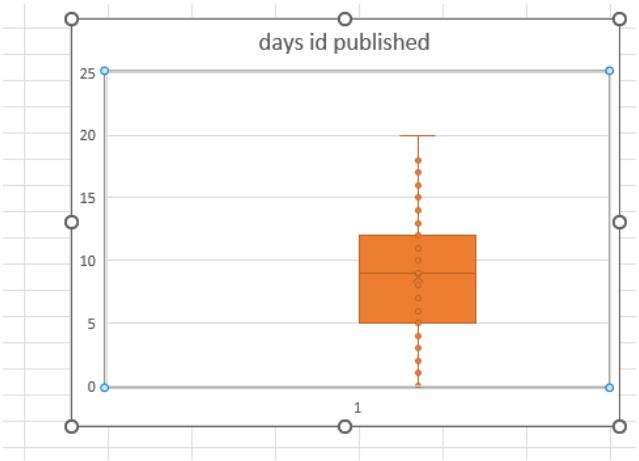


Figure 4.1.9

- From the above figure that shows outliers. In figure 4.9 spotted no outliers

B. Univariate analysis

Count of gender:

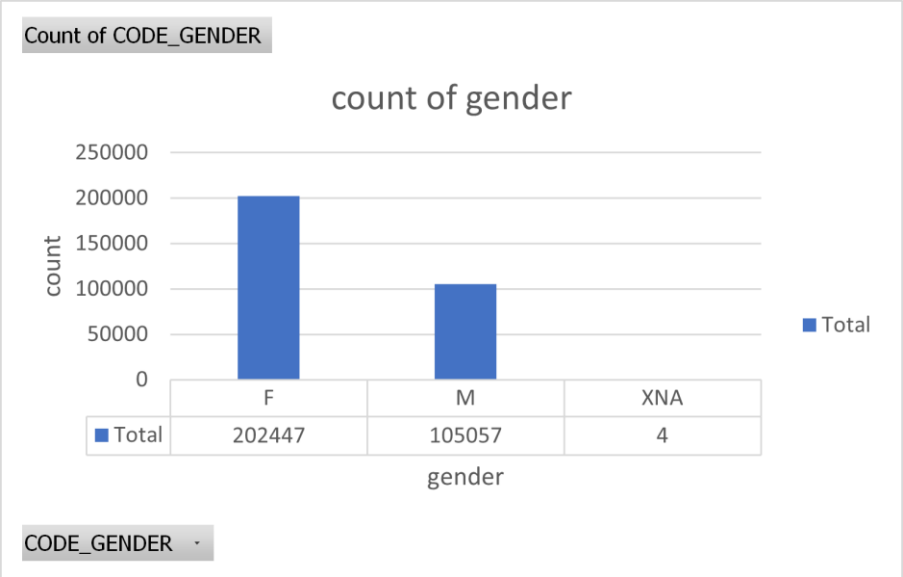


Figure 4.1.10

- We can see from the above figure 4.10 number of female is more than number of male.

Count of annuity income:

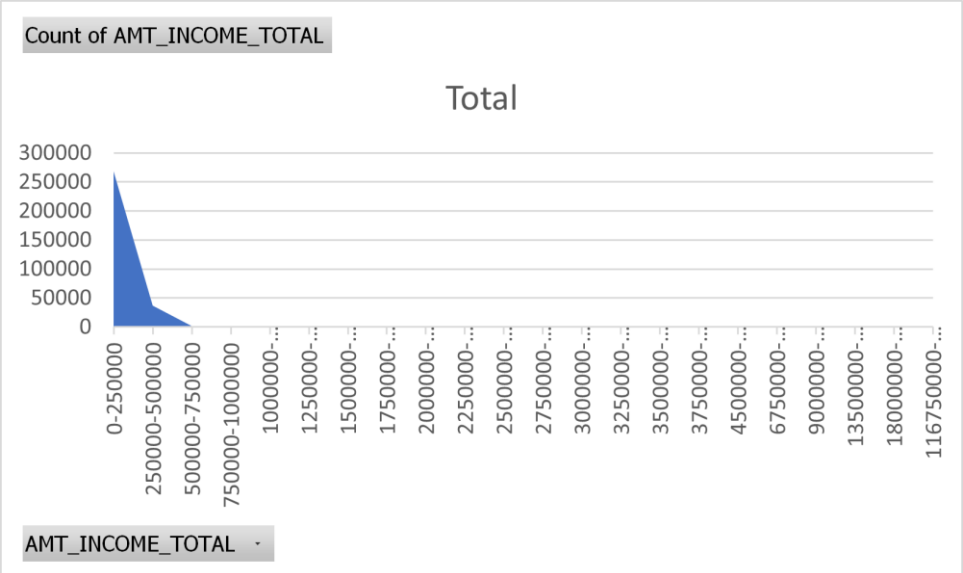


Figure 4.1.11

group	Count of AMT_INCOME_TOTAL
0-250000	267731
250000-500000	37075
500000-750000	2058
750000-1000000	394
1000000-1250000	99
1250000-1500000	67
1500000-1750000	20
1750000-2000000	20
2000000-2250000	11
2250000-2500000	14
2500000-2750000	1
2750000-3000000	1
3000000-3250000	2
3250000-3500000	2
3500000-3750000	2
3750000-4000000	2
4500000-4750000	4
6750000-7000000	1
9000000-9250000	1
13500000-13750000	1
18000000-18250000	1
116750000-117000000	1

Table 4.1.2

- The above figure 4.11 shows range of income total most of the persons total income vested in 0-250000

Count of amount credit:

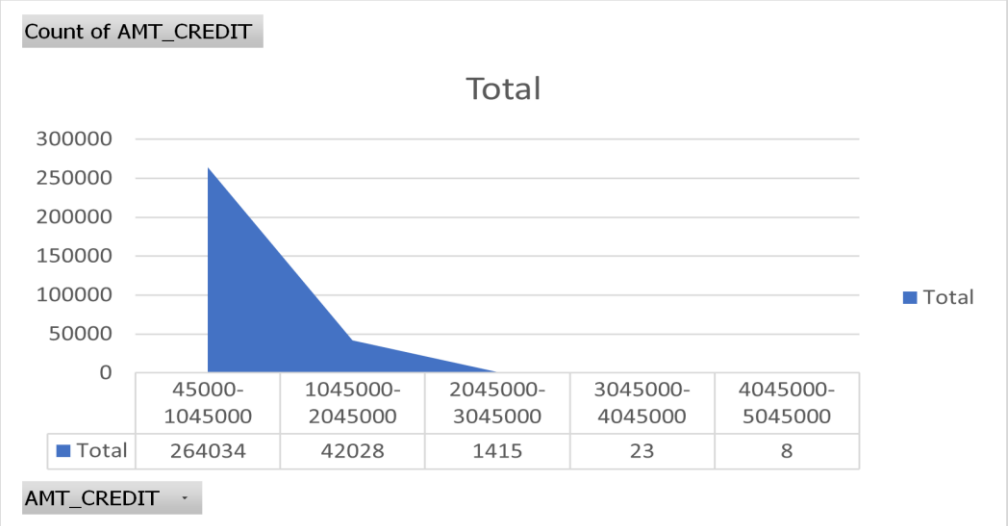


Figure 4.1.12

- The above graph shows the credit amount of loan. We can see here more amount from the range 45000-1045000

Count of annuity amount:

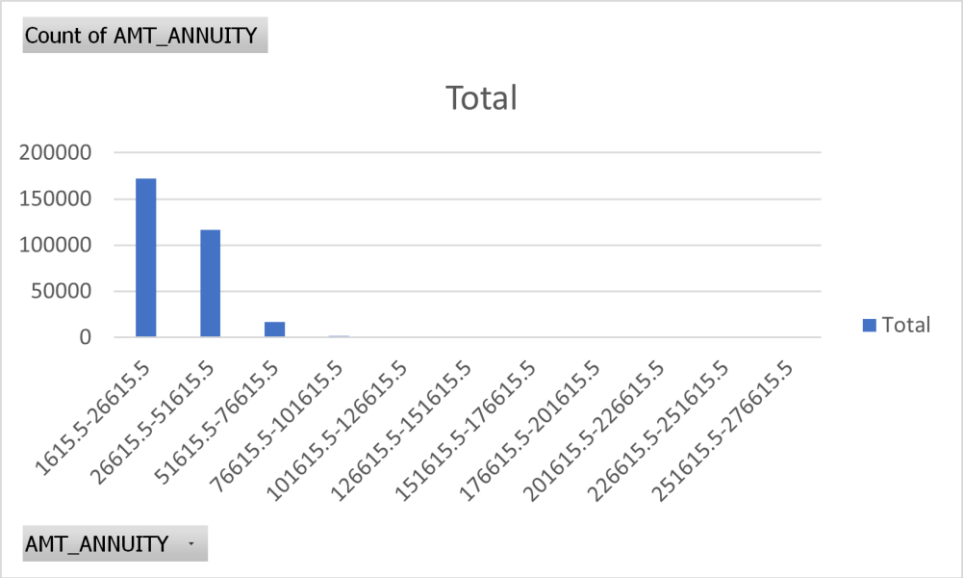


Figure 4.1.13

Row Labels	Count of AMT_ANNUITY
1615.5-26615.5	172042
26615.5-51615.5	116770
51615.5-76615.5	16872
76615.5-101615.5	1348
101615.5-126615.5	327
126615.5-151615.5	86
151615.5-176615.5	19
176615.5-201615.5	8
201615.5-226615.5	34
226615.5-251615.5	1
251615.5-276615.5	1
Grand Total	**307508**

Table 4.1.3

- Figure 4.13 shows loan annuity, more amount vested in 1615-26615

C. Segmented univariate

Count of gender:

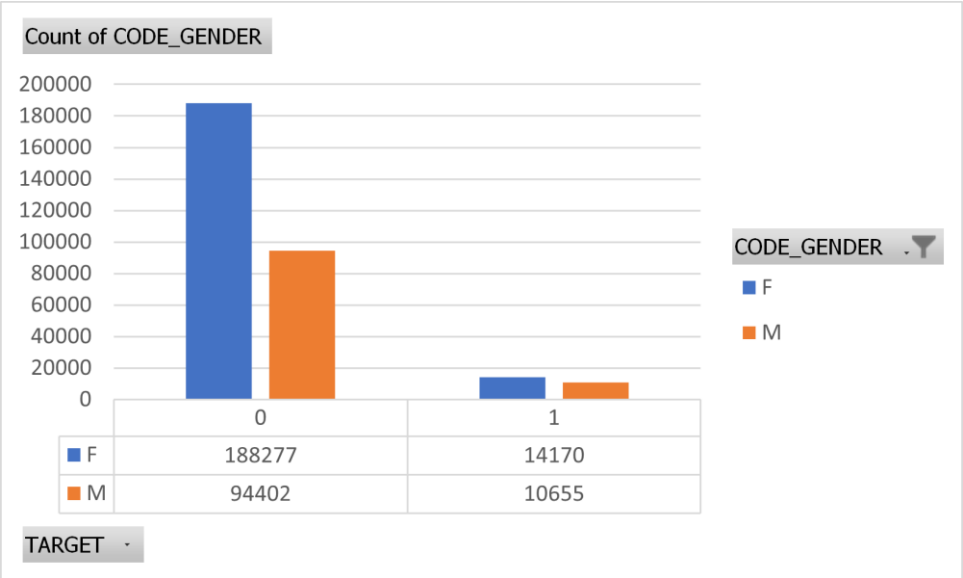


Figure 4.1.14

- From the above graph we can see that both female and male are almost equal in payment difficulties.
- In the other hand number of female client is more than male for not having payment difficulties.

Education type:

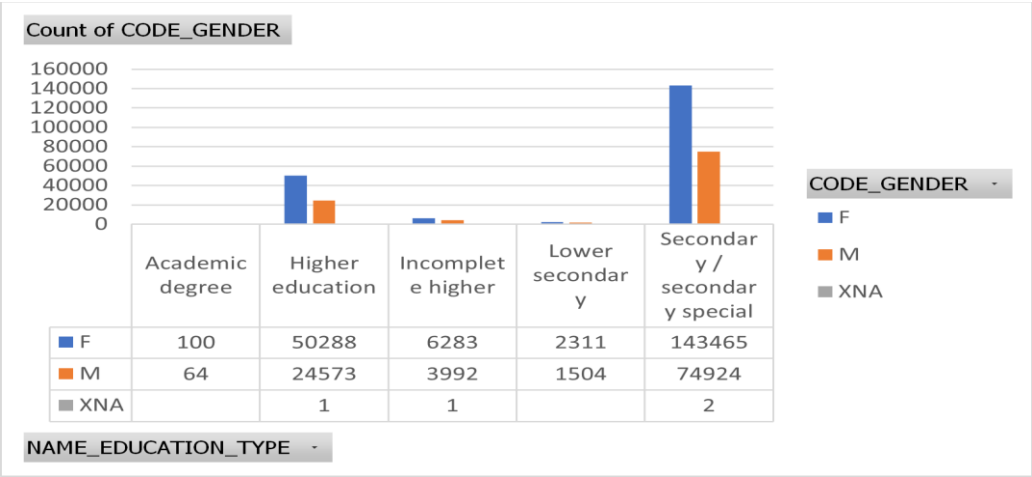


Figure 4.1.15

- Above graph shows education type of the client. We can see from the graph that most of the client have secondary education. On that number of female higher than male client.

Car count:

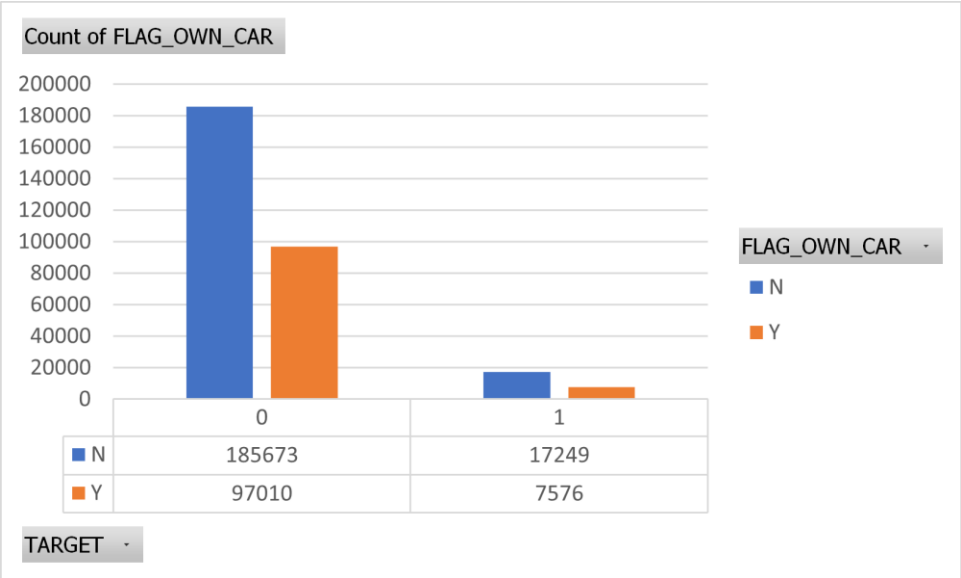


Figure 4.1.16

- From the above graph we can infer that most of the client who not having payment difficulties they don't have a car

Realty count

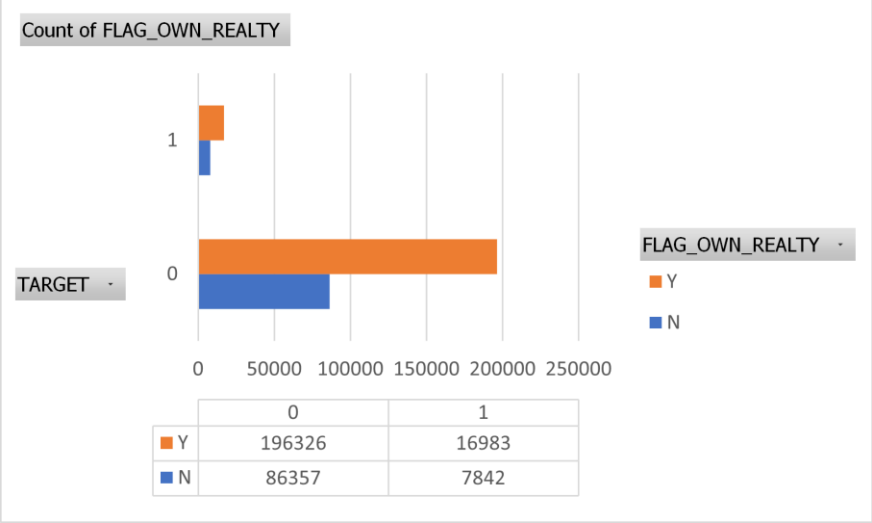


Figure 4.1.17

- We can see from the above graph that most of client own realty. On that most of client has no payment difficulties

Contract type:

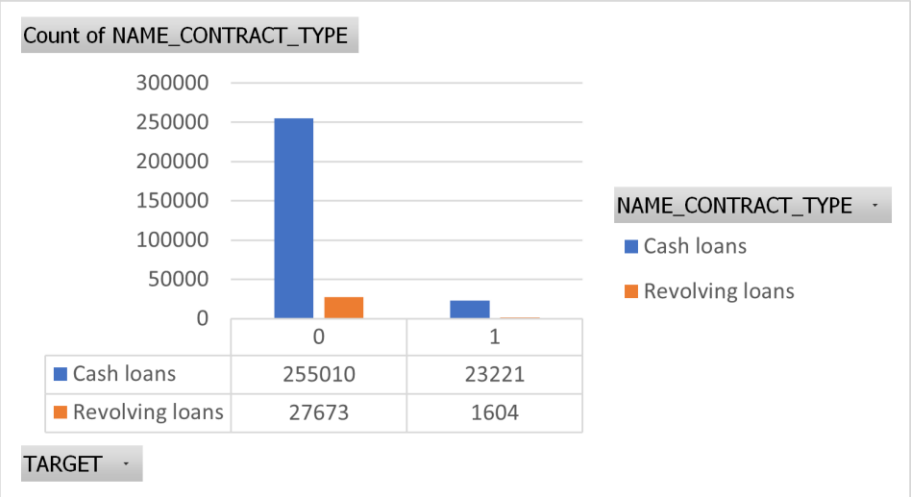


Figure 4.1.18

- From the figure 4.18 that shows most of the client with on payment difficulties uses cash loans. Revolving loans type are very low compared to cash loans.

D. Bivariate analysis

Income type and Annuity:

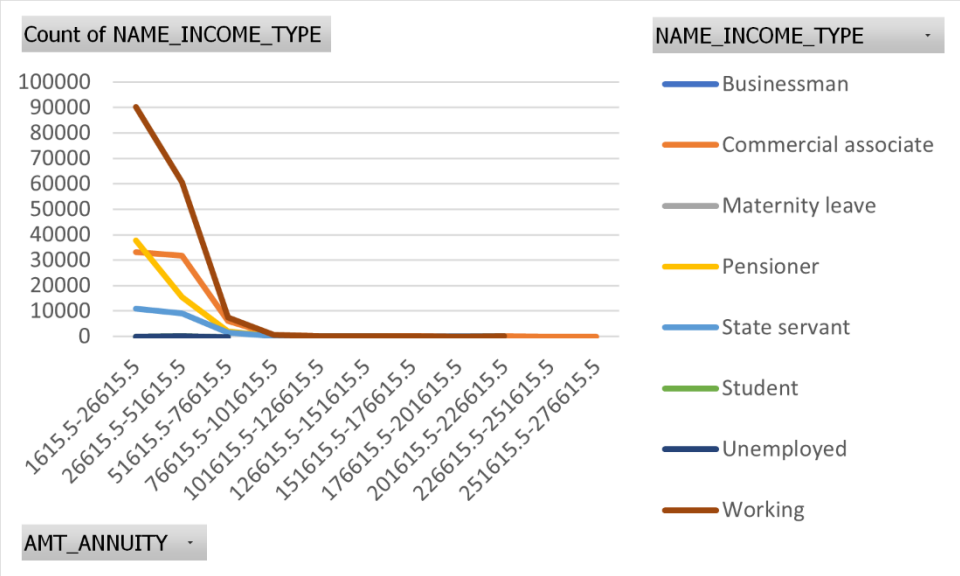


Figure 4.1.19

Count of NAME_INCOME_TYPE	Column Labels								
Row Labels	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working	Grand Total
1615.5-26615.5	2	33082	2	37800	10959	11	12	90174	172042
26615.5-51615.5	3	31702	2	15403	9083	6	9	60562	116770
51615.5-76615.5	2	6035	1	1940	1511	1	1	7381	16872
76615.5-101615.5		578		148	112			510	1348
101615.5-126615.5	2	139		49	25			112	327
126615.5-151615.5		40		17	6			23	86
151615.5-176615.5		9		1	1			8	19
176615.5-201615.5		4		1	2			1	8
201615.5-226615.5	1	24		3	4			2	34
226615.5-251615.5		1							1
251615.5-276615.5		1							1
Grand Total	10	71615	5	55362	21703	18	22	158773	307508

Table 4.1.4

- Above figure 4.19 shows annuity income vs income type. We can infer from the graph that higher annuity income from the income type of work. Just below we have commercial associate.

Income type vs gender:

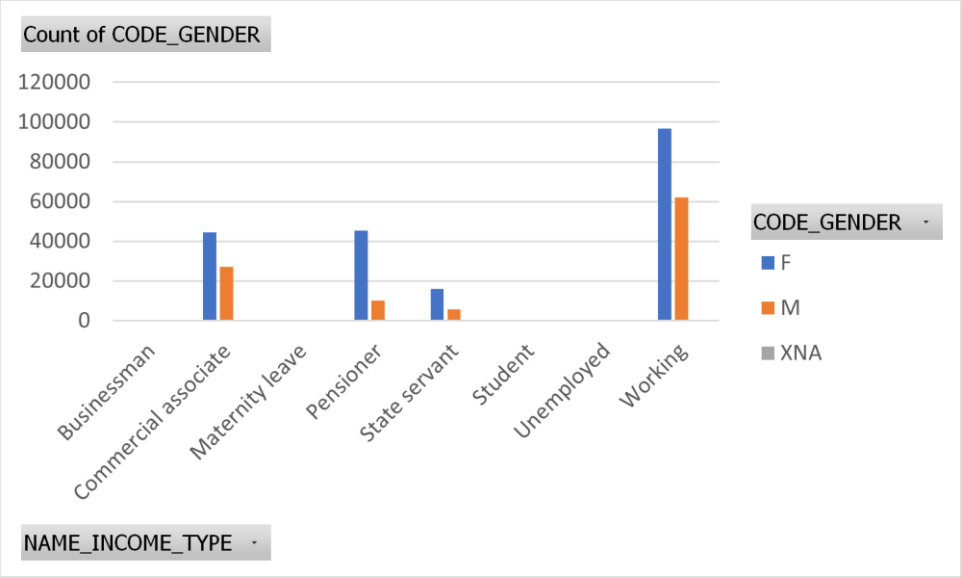


Figure 4.1.20

Count of CODE_GENDER	Column Labels			
Row Labels	F	M	XNA	Grand Total
Businessman	3	7		10
Commercial associate	44518	27096	1	71615
Maternity leave	4	1		5
Pensioner	45261	10101		55362
State servant	15856	5847		21703
Student	7	11		18
Unemployed	17	5		22
Working	96781	61989	3	158773
Grand Total	**202447**	**105057**	**4**	**307508**

Table 4.1.5

- Above column chart shows income type vs gender. Here we can see that number of female is more than male from the income type of working
- Also we can see that female more in commercial associate and pensioner. Because we already seen that female client are client are higher than male.

Total income vs income type:

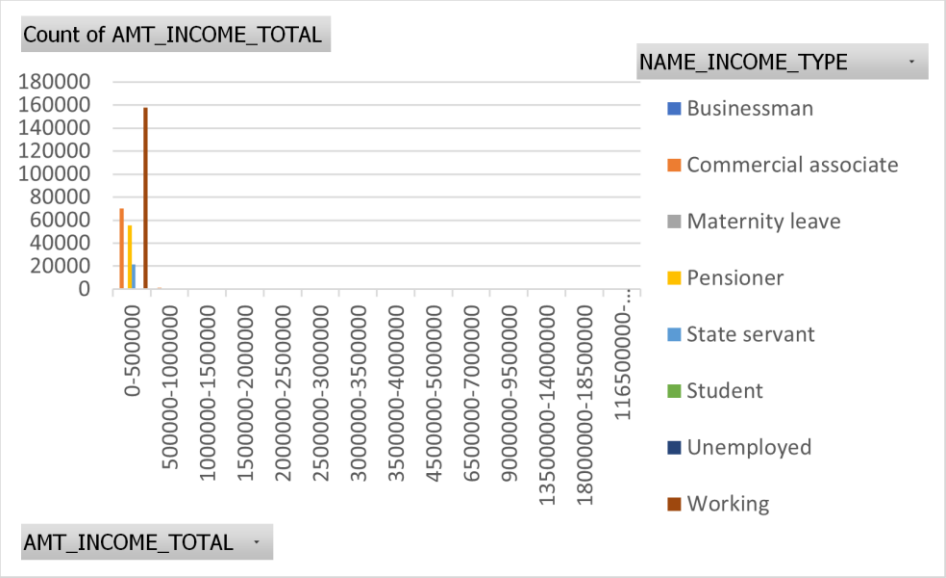


Figure 4.1.21

Count of AMT_INCOME_TOTAL	Column Labels								
group	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working	Grand Total
0-500000	5	70121	5	55203	21496	17	22	157937	304806
500000-1000000	4	1344		149	193	1		761	2452
1000000-1500000		107		6	8			45	166
1500000-2000000		24		2	3			11	40
2000000-2500000	1	12		2	2			8	25
2500000-3000000		1						1	2
3000000-3500000		1			1			2	4
3500000-4000000		3						1	4
4500000-5000000								4	4
6500000-7000000								1	1
9000000-9500000								1	1
13500000-14000000		1							1
18000000-18500000		1							1
116500000-117000000								1	1
Grand Total	10	71615	5	55362	21703	18	22	158773	307508

Table 4.1.6

- The above graph is about total income vs income type. Working group have the higher amount income.

Target and income type:

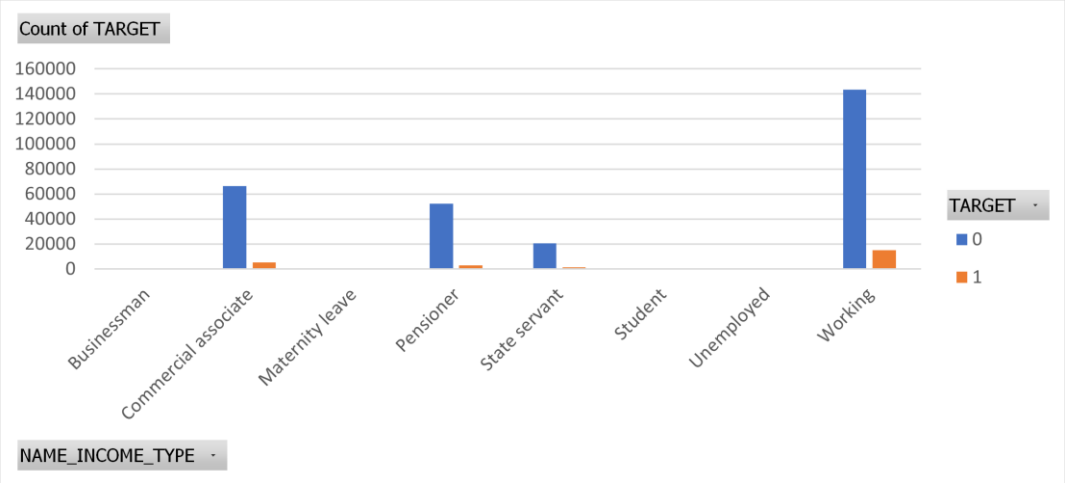


Figure 4.1.22

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Businessman	10		10
Commercial associate	66255	5360	71615
Maternity leave	3	2	5
Pensioner	52380	2982	55362
State servant	20454	1249	21703
Student	18		18
Unemployed	14	8	22
Working	143549	15224	158773
Grand Total	**282683**	**24825**	**307508**

Table 4.1.7

- From the above figure we can get that most of the client with no payment difficulties has income from working.

E. Correlation

		AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
2	AMT_INCOME_TOTAL	1			
3	AMT_CREDIT	0.156869636	1		
4	AMT_ANNUITY	0.19165522	0.770127383	1	
5	AMT_GOODS_PRICE	0.159632806	0.986734628	0.774838069	1

Figure 4.1.23

	TARGET	T_CHILDR	INCOME_T	MT_CREDIT	MT_ANNUI	GOODS_P	PULATION	age	ys employ	s registrat	s id publis	GAPARTM	VINGAREA	TMENTS_I	ENTAREA_I	EXPLUATA	S_BUILD_	LAG_PHON	LAG_EMAI	FAM_MEM	_RATING_	TING_CLIEN
TARGET	1																					
CNT_CHILDREN	0.01919	1																				
AMT_INCOME_TOTAL	-0.00398	0.01288	1																			
AMT_CREDIT	-0.03037	0.00214	0.15687	1																		
AMT_ANNUITY	-0.01281	0.02138	0.19166	0.77013	1																	
AMT_GOODS_PRICE	-0.03962	-0.00186	0.15963	0.98673	0.77484	1																
REGION_POPULATION_RELATIVE	-0.03723	-0.02557	0.0748	0.09974	0.11842	0.10352	1															
age	-0.07824	-0.33091	-0.02725	0.05539	-0.00949	0.05346	0.02964	1														
days employed	-0.04705	-0.24159	-0.06383	-0.06432	-0.10285	-0.06219	-0.00409	0.62389	1													
days registration	-0.04197	-0.1834	-0.02778	-0.00953	-0.03838	-0.01147	0.05382	0.33168	0.21458	1												
days id published	-0.05113	0.0281	-0.00867	0.00685	-0.01108	0.00959	0.00412	0.2715	0.2745	0.1014	1											
NONLIVINGAPARTMENTS_AVG	-0.00591	-0.00173	0.01307	0.01534	0.02032	0.0157	0.0341	0.00127	-0.00341	-0.01192	0.00394	1										
NONLIVINGAREA_AVG	-0.01945	-0.0101	0.03498	0.04167	0.05216	0.04602	0.09144	0.00183	-0.01123	-0.01435	0.00309	0.17513	1									
APARTMENTS_MODE	-0.03852	-0.02879	0.05222	0.06525	0.07902	0.07089	0.17594	0.00885	-0.01429	0.03229	0.01277	0.18024	0.37088	1								
BASEMENTAREA_MODE	-0.03213	-0.0239	0.03796	0.04968	0.05535	0.05314	0.1137	0.01443	-0.0054	0.05099	0.01621	0.12359	0.31293	0.73085	1							
YEARS_BEGINEXPLUATATION_MODE	-0.04094	-0.03761	0.04976	0.05647	0.06689	0.06122	0.10154	0.01896	-0.01005	0.06995	0.01443	0.09924	0.2535	0.58027	0.51283	1						
YEARS_BUILD_MODE	-0.03302	-0.02389	0.04129	0.05271	0.06008	0.05157	0.10546	0.01557	-0.00525	0.03567	0.0125	0.14818	0.19152	0.50259	0.50405	0.68203	1					
FLAG_PHONE	-0.02381	-0.02991	0.00015	0.02621	0.0108	0.04192	0.09094	0.04242	0.01685	0.075	0.03751	0.0067	0.02971	0.06859	0.05919	0.06548	0.04252	1				
FLAG_EMAIL	-0.00176	0.02262	0.03838	0.01663	0.07171	0.01703	0.04001	-0.08818	-0.063	-0.0343	-0.02762	0.00381	0.02258	0.04203	0.03244	0.04167	0.03349	0.01466	1			
CNT_FAM_MEMBERS	0.00931	0.87916	0.01634	0.06316	0.07554	0.06114	-0.02421	-0.2789	-0.23436	-0.17344	0.0209	-0.00524	-0.01627	-0.04289	-0.03518	-0.0643	-0.04041	-0.01542	0.01808	1		
REGION_RATING_CLIENT	0.0589	0.02542	-0.08547	-0.10178	-0.12852	-0.10376	-0.53288	-0.00941	0.03304	-0.08016	0.00495	-0.02366	-0.09014	-0.15666	-0.09214	-0.1359	-0.10059	-0.08383	-0.05206	0.02969	1	
REGION_RATING_CLIENT_W_CITY	0.06089	0.02478	-0.09173	-0.11092	-0.14167	-0.11218	-0.53154	-0.00812	0.03488	-0.07401	0.00755	-0.03054	-0.0991	-0.18609	-0.12177	-0.16592	-0.13479	-0.07913	-0.05078	0.03078	0.95084	1

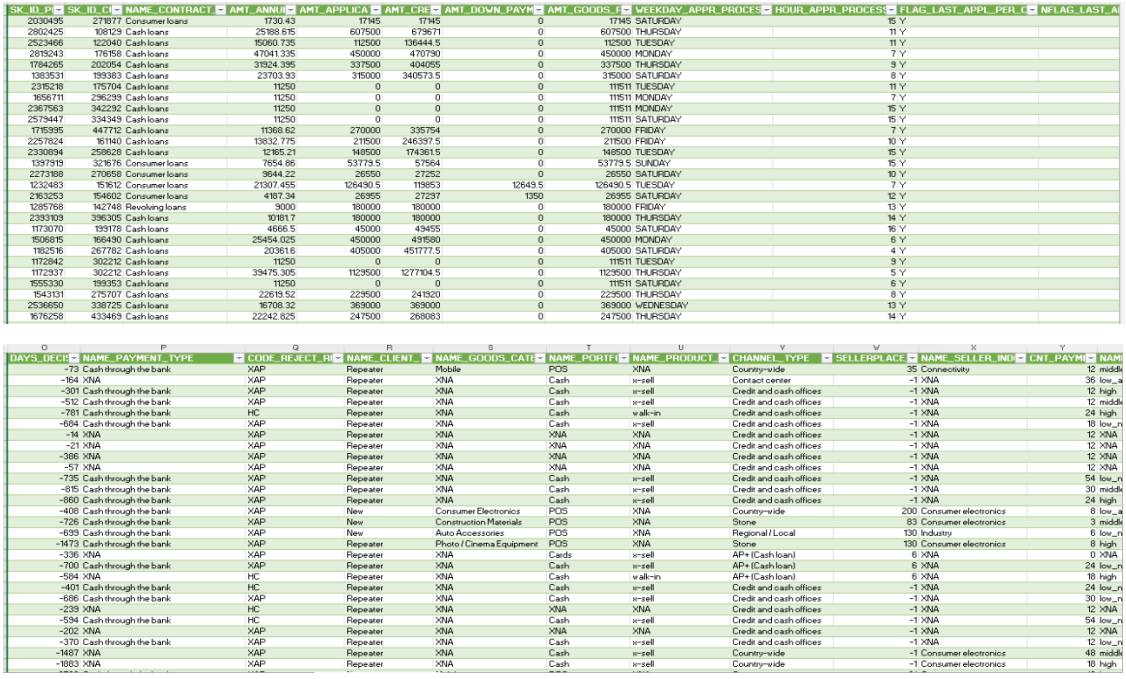
Figure 4.1.24

- Here are some top correlation.
  - AMT_ANNUITY to AMT_CREDIT
  - AMT_GOODS_PRICE to AMT_CREDIT
  - AMT_GOODS_PRICE to AMT_ANNUITY
  - days employed to age
  - BASEMENTAREA_MODE to APARTMENTS_MODE
  - YEARS_BUILD_MODE to YEARS_BEGINEXPLUATATION_MODE
  - CNT_FAM_MEMBERS to CNT_CHILDREN
  - REGION_RATING_CLIENT_W_CITY to REGION_RATING_CLIENT

- some negative correlations are
    - REGION_RATING_CLIENT to REGION_POPULATION_RELATIVE
    - REGION_POPULATION_RELATIVE to REGION_RATING_CLIENT_W_CITY
    - age to CNT_CHILDREN

2. previous_application.

A. Data cleaning



SK_ID_P	SK_ID_C	NAME_CONTRACT	AMT_ANNU	AMT_APPLICA	AMT_CRE	AMT_DOWN_PAYM	AMT_GOODS_P	WEEKDAY_APPR_PROCES	HOUR_APPR_PROCES	FLAG_LAST_APPL_PER_C	NFLAG_LAST_A
2030495	271877	Consumer loans	1730.43	17145	17145	0	17145	SATURDAY	15	Y	
2802425	108129	Cash loans	25188.615	607500	679671	0	607500	THURSDAY	11	Y	
2523466	122040	Cash loans	15060.735	112500	136444.5	0	112500	TUESDAY	11	Y	
2819243	176158	Cash loans	47041.335	450000	470790	0	450000	MONDAY	7	Y	
1784265	202054	Cash loans	31924.395	337500	404055	0	337500	THURSDAY	9	Y	
1383531	199383	Cash loans	23703.93	315000	340573.5	0	315000	SATURDAY	8	Y	
2315218	175704	Cash loans	11250	0	0	0	111511	TUESDAY	11	Y	
1656711	296299	Cash loans	11250	0	0	0	111511	MONDAY	7	Y	
2367563	342292	Cash loans	11250	0	0	0	111511	MONDAY	15	Y	
2579447	334349	Cash loans	11250	0	0	0	111511	SATURDAY	15	Y	
1715995	447712	Cash loans	11368.62	270000	335754	0	270000	FRIDAY	7	Y	
2257824	161140	Cash loans	13832.775	211500	246397.5	0	211500	FRIDAY	10	Y	
2330894	258628	Cash loans	12165.21	148500	174361.5	0	148500	TUESDAY	15	Y	
1397919	321676	Consumer loans	7654.86	53775.5	57564	0	53775.5	SUNDAY	15	Y	
2273198	270658	Consumer loans	9644.22	26550	27252	0	26550	SATURDAY	10	Y	
1232483	151612	Consumer loans	21307.455	126490.5	119853	12649.5	126490.5	TUESDAY	7	Y	
2163253	154602	Consumer loans	4187.34	26955	27297	1350	26955	SATURDAY	12	Y	
1285768	142748	Revolving loans	9000	180000	180000	0	180000	FRIDAY	13	Y	
2393109	396305	Cash loans	10161.7	180000	180000	0	180000	THURSDAY	14	Y	
1173070	139178	Cash loans	4666.5	45000	49455	0	45000	SATURDAY	16	Y	
1506815	166490	Cash loans	25454.025	450000	491580	0	450000	MONDAY	6	Y	
1182516	267782	Cash loans	20361.6	405000	451777.5	0	405000	SATURDAY	4	Y	
1172842	302212	Cash loans	11250	0	0	0	111511	TUESDAY	9	Y	
1172937	302212	Cash loans	39475.305	1129500	1277104.5	0	1129500	THURSDAY	5	Y	
1555330	199353	Cash loans	11250	0	0	0	111511	SATURDAY	6	Y	
1543131	275707	Cash loans	22619.52	229500	241920	0	229500	THURSDAY	8	Y	
2536650	338725	Cash loans	16708.32	369000	369000	0	369000	WEDNESDAY	13	Y	
1676258	433469	Cash loans	22242.825	247500	268083	0	247500	THURSDAY	14	Y	

O	P	Q	R	S	T	U	V	W	X	Y	
DAYS_DECI	NAME_PAYMENT_TYPE	CODE_REJECT_RE	NAME_CLIENT_	NAME_GOODS_CATE	NAME_PORTFO	NAME_PRODUCT_	CHANNEL_TYPE	SELLERPLACE_	NAME_SELLER_IND	CNT_PAYMI	NAM
-73	Cash through the bank	XAP	Repeater	Mobile	POS	XNA	Country-wide	35	Connectivity	12	middl
-164	XNA	XAP	Repeater	XNA	Cash	x-sell	Contact center	-1	XNA	36	low_a
-301	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	12	high
-512	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	12	middl
-781	Cash through the bank	HC	Repeater	XNA	Cash	walk-in	Credit and cash offices	-1	XNA	24	high
-684	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	18	low_n
-14	XNA	XAP	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-21	XNA	XAP	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-386	XNA	XAP	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-57	XNA	XAP	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-735	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	54	low_n
-815	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	30	middl
-860	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	24	high
-408	Cash through the bank	XAP	New	Consumer Electronics	POS	XNA	Country-wide	200	Consumer electronics	6	low_a
-726	Cash through the bank	XAP	New	Construction Materials	POS	XNA	Stone	83	Consumer electronics	3	middl
-699	Cash through the bank	XAP	New	Auto Accessories	POS	XNA	Regional / Local	130	Industry	6	low_n
-1473	Cash through the bank	XAP	Repeater	Photo / Cinema Equipment	POS	XNA	Stone	130	Consumer electronics	8	high
-336	XNA	XAP	Repeater	XNA	Cards	x-sell	AP+ (Cash loan)	6	XNA	0	XNA
-700	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	AP+ (Cash loan)	6	XNA	24	low_n
-584	XNA	HC	Repeater	XNA	Cash	walk-in	AP+ (Cash loan)	6	XNA	18	high
-401	Cash through the bank	HC	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	24	low_n
-686	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	30	low_n
-239	XNA	HC	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-594	Cash through the bank	HC	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	54	low_n
-202	XNA	XAP	Repeater	XNA	XNA	XNA	Credit and cash offices	-1	XNA	12	XNA
-370	Cash through the bank	XAP	Repeater	XNA	Cash	x-sell	Credit and cash offices	-1	XNA	12	low_n
-1487	XNA	XAP	Repeater	XNA	Cash	x-sell	Country-wide	-1	Consumer electronics	48	middl
-1883	XNA	XAP	Repeater	XNA	Cash	x-sell	Country-wide	-1	Consumer electronics	18	high

Table 4.2.1

Finding null values:

- for cleaning process we have to identify null values and remove it.
- The table contain 37 columns and 1670214 rows. Some of the columns may have contain null values
- I used filter function to find null values in each columns. If the columns contain 50% null values. The columns should be deleted
- I used those formulas for this purpose

=countblank(range)

=counta(range)

- Deleting those columns which contain null values more than 50%. It include columns like rate down pay, rate interest, rate interest privilege, name type suit etc.
- Also I replace all the numerical missing values with median, because when the data is skewed, it is good to consider using the median values for replacing the missing values
- Categorical missing values are low, so when found the mi ssing values I deleted in row vice.

Finding outliers:

- Here box blot is mainly used for finding outliers

Amount credit of previous:

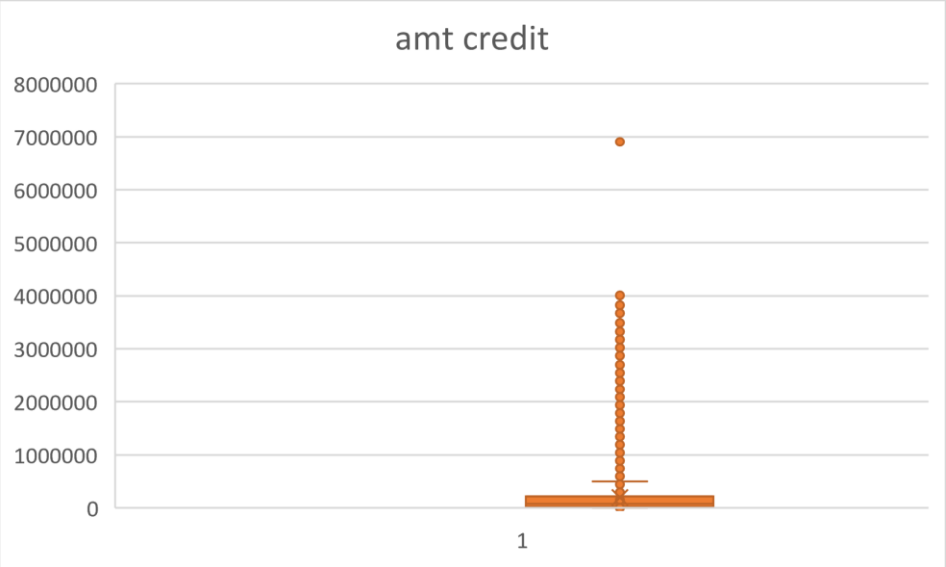


Figure 4.2.1

Amount annuity of previous:

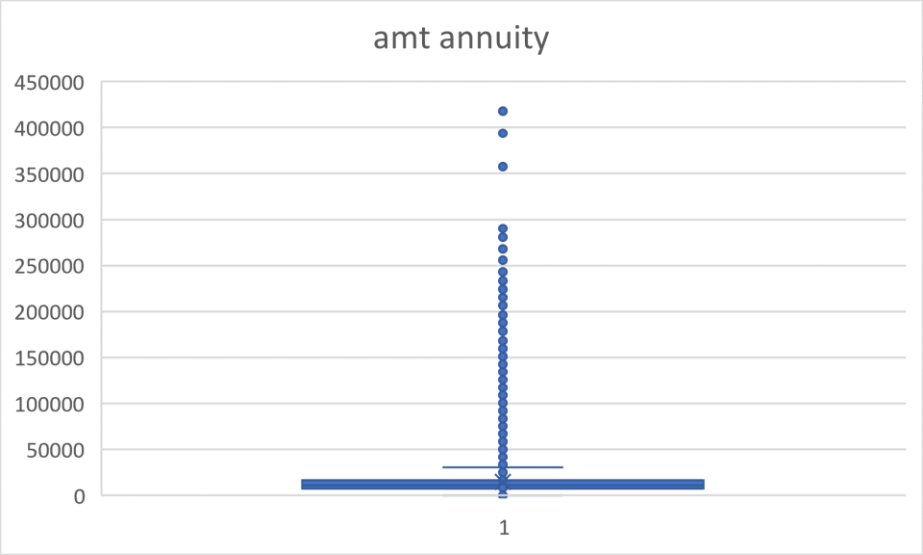


Figure 4.2.2

Amount application of previous:

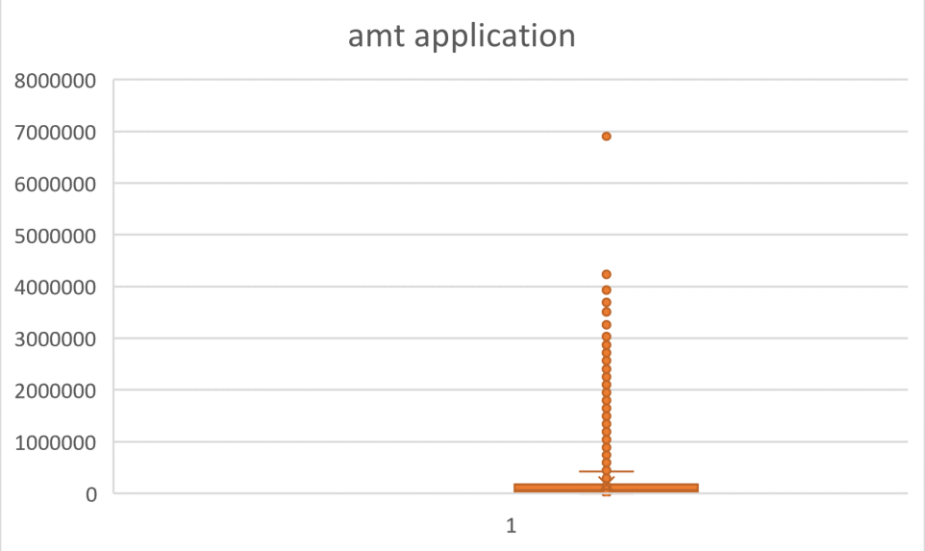


Figure 4.2.3

- We can see from the above box plots there are lots of outliers are in this columns.

B. Univariate analysis

Contract type

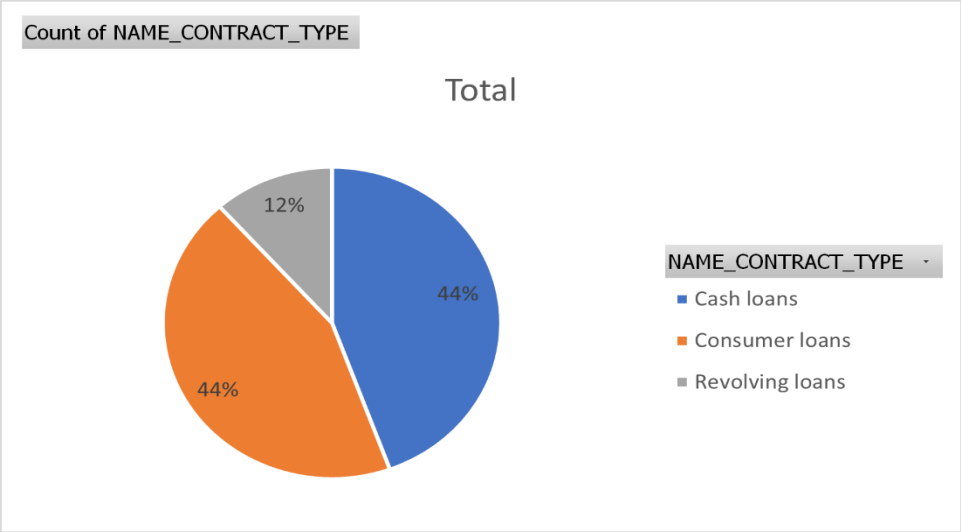


Figure 4.2.4

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

Table 4.2.2

- From the above pie chart we can analysis that 44% are cash loan and consumer loan. Rest 12% are revolving loans.

Amount of annuity:

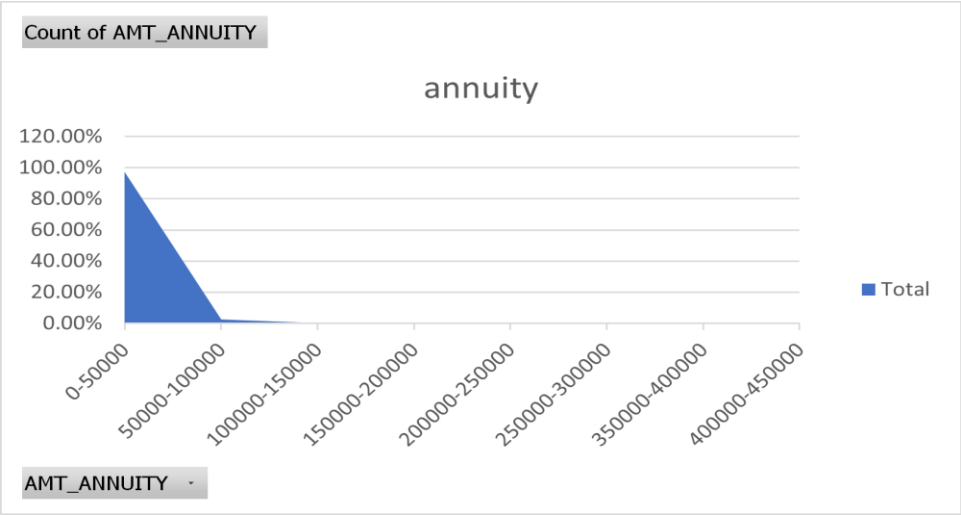


Figure 4.2.5

Row Labels	Count of AMT_ANNUITY
0-50000	97.24%
50000-100000	2.60%
100000-150000	0.14%
150000-200000	0.01%
200000-250000	0.01%
250000-300000	0.00%
350000-400000	0.00%
400000-450000	0.00%
Grand Total	**100.00%**

Table 4.2.3

- above figure 4.2.5 shows that the amount of annuity from the previous application. There are 97% of client have annuity amount around 50000.

Amount of application:

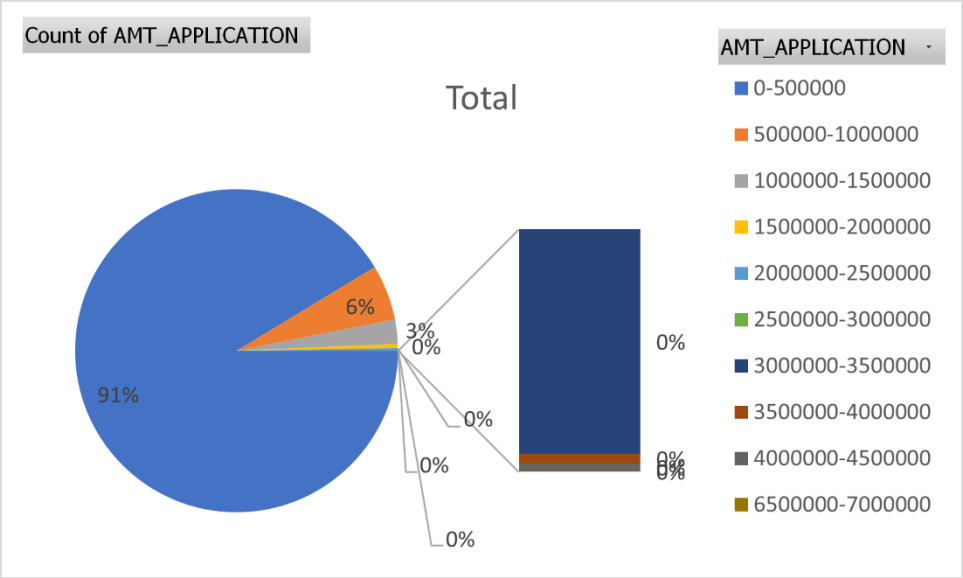


Figure 4.2.6

Row Labels	Count of AMT_APPLICATION
0-500000	957392
500000-1000000	58212
1000000-1500000	25535
1500000-2000000	4314
2000000-2500000	2395
2500000-3000000	211
3000000-3500000	270
3500000-4000000	12
4000000-4500000	9
6500000-7000000	1
Grand Total	**1048351**

Table 4.2.4

- above graph shows the application amount of the previous application. Here we can see that most of the client have application amount from the range 0-500000.

Amount of credit:

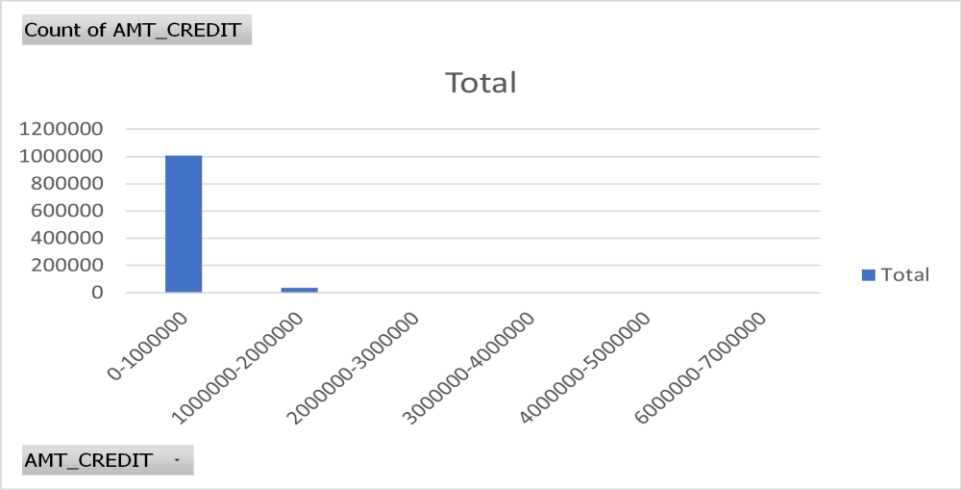


Figure 4.2.7

Row Labels	Count of AMT_CREDIT
0-1000000	1008932
1000000-2000000	35460
2000000-3000000	3586
3000000-4000000	361
4000000-5000000	11
6000000-7000000	1
Grand Total	**1048351**

Table 4.2.5

- most of the client credit amount vested in the range of 0 to 1000000.

Loan purpose:

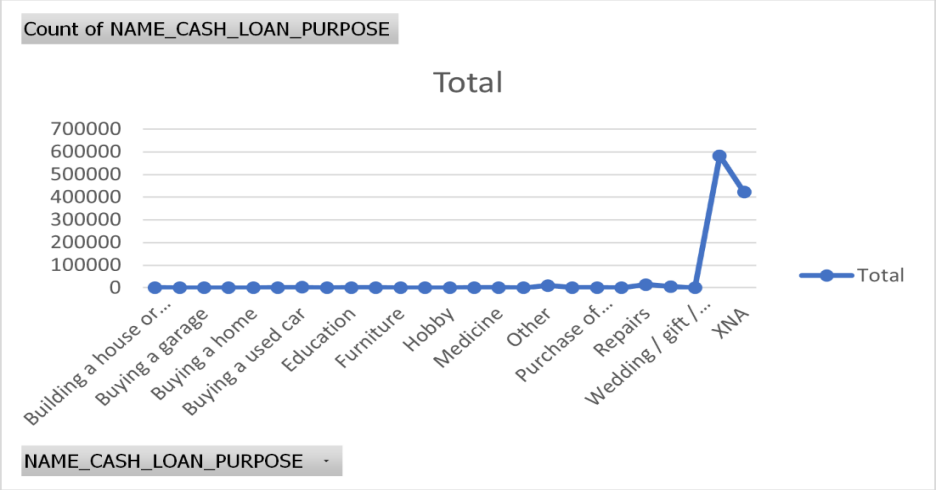


Figure 4.2.8

Row Labels	Count of NAME_CASH_LOAN_PURPOSE
Building a house or an annex	1682
Business development	254
Buying a garage	84
Buying a holiday home / land	334
Buying a home	546
Buying a new car	637
Buying a used car	1784
Car repairs	493
Education	975
Everyday expenses	1487
Furniture	475
Gasification / water supply	203
Hobby	31
Journey	745
Medicine	1402
Money for a third person	16
Other	9664
Payments on other loans	1210
Purchase of electronic equipmen	644
Refusal to name the goal	8
Repairs	14767
Urgent needs	5309
Wedding / gift / holiday	594
XAP	582490
XNA	422517
Grand Total	**1048351**

Table 4.2.6

- above graph shows the purpose of the loan. Known reason for loan taken by the most of the client is for repairs.

Contract status:

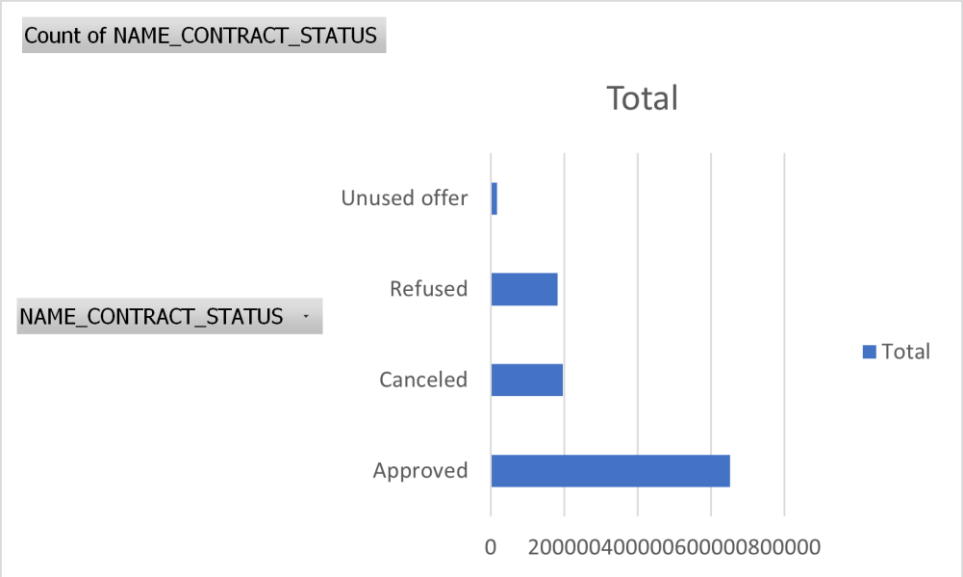


Figure 4.2.9

Row Labels	Count of NAME_CONTRACT_STATUS
Approved	652486
Canceled	197026
Refused	182064
Unused offer	16775
Grand Total	**1048351**

Table 4.2.7

- we can see that most of the applications are approved by the company

C. Segmented univariate

Contract status:

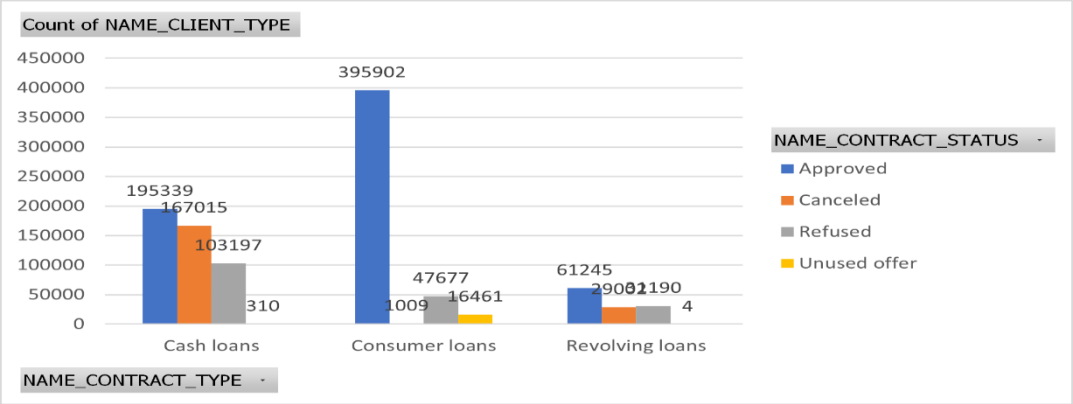


Figure 4.2.10

Count of NAME_CLIENT_TYPE	Column Labels				
Row Labels	Approved	Canceled	Refused	Unused offer	Grand Total
Cash loans	195339	167015	103197	310	465861
Consumer loans	395902	1009	47677	16461	461049
Revolving loans	61245	29002	31190	4	121441
Grand Total	**652486**	**197026**	**182064**	**16775**	**1048351**

Table 4.2.8

- Consumer loan type, which a company approved by the most.

  Most of the refused and cancelled loan type is cash loan.

Goods category and contract type:

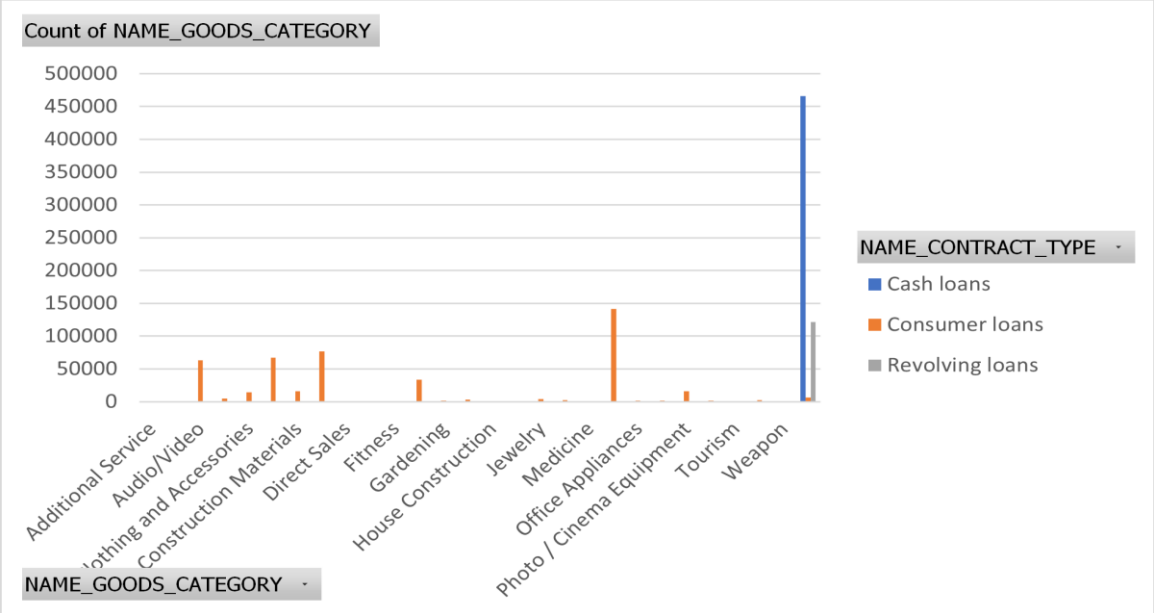


Figure 4.2.11

Row Labels	Cash loans	Consumer loans	Revolving loans	Grand Total
Additional Service		89		89
Animals		1		1
Audio/Video		62859	51	62910
Auto Accessories		4706		4706
Clothing and Accessories		14764		14764
Computers		67067	15	67082
Construction Materials		15754		15754
Consumer Electronics		76879	228	77107
Direct Sales		275		275
Education		71		71
Fitness		129		129
Furniture		33919	12	33931
Gardening		1686		1686
Homewares		3149		3149
House Construction		1		1
Insurance		35		35
Jewelry		4032		4032
Medical Supplies		2461		2461
Medicine		985		985
Office Appliances		1477		1477
Other		1589	1	1590
Photo / Cinema Equipment		15777	2	15779
Sport and Leisure		1906		1906
Tourism		1020		1020
Vehicles		2162		2162
Weapon		46		46
XNA	465861	6487	121118	593466
Grand Total	**465861**	**461049**	**121441**	**1048351**

Figure 4.2.9

- We can see from the above graph that consumer loan are used for different goods.

Product combination and contract type:

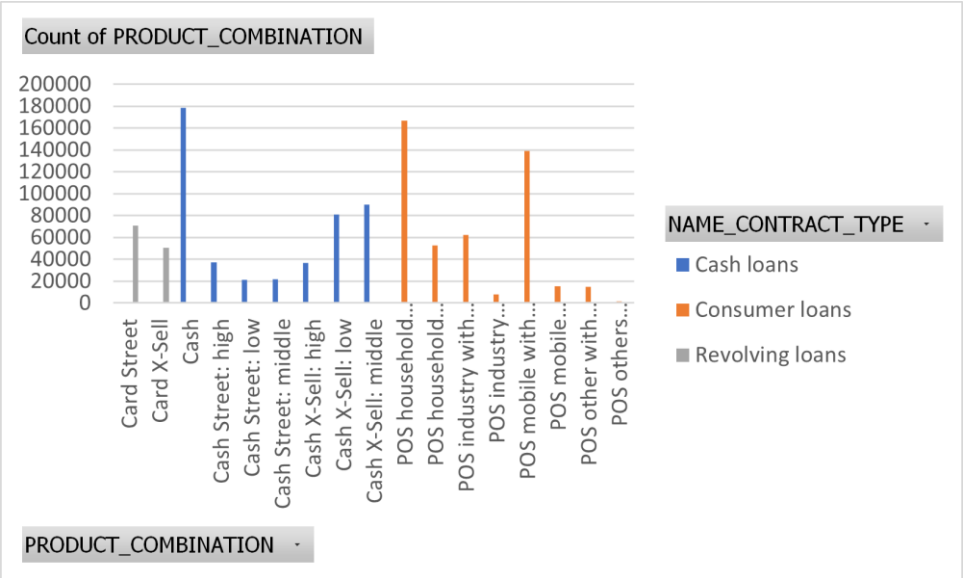


Figure 4.2.12

Row Labels	Cash loans	Consumer loans	Revolving loans	Grand Total
Card Street			70951	70951
Card X-Sell			50490	50490
Cash	178352			178352
Cash Street: high	37235			37235
Cash Street: low	21166			21166
Cash Street: middle	21616			21616
Cash X-Sell: high	36813			36813
Cash X-Sell: low	80873			80873
Cash X-Sell: middle	89806			89806
POS household with interest		166869		166869
POS household without interest		52747		52747
POS industry with interest		62492		62492
POS industry without interest		7856		7856
POS mobile with interest		139176		139176
POS mobile without interest		15181		15181
POS other with interest		15072		15072
POS others without interest		1656		1656
Grand Total	**465861**	**461049**	**121441**	**1048351**

Table 4.2.10

- We can see from the above graph about the product combination and contract type

D.Bivariate analysis

Amount of annuity vs application:

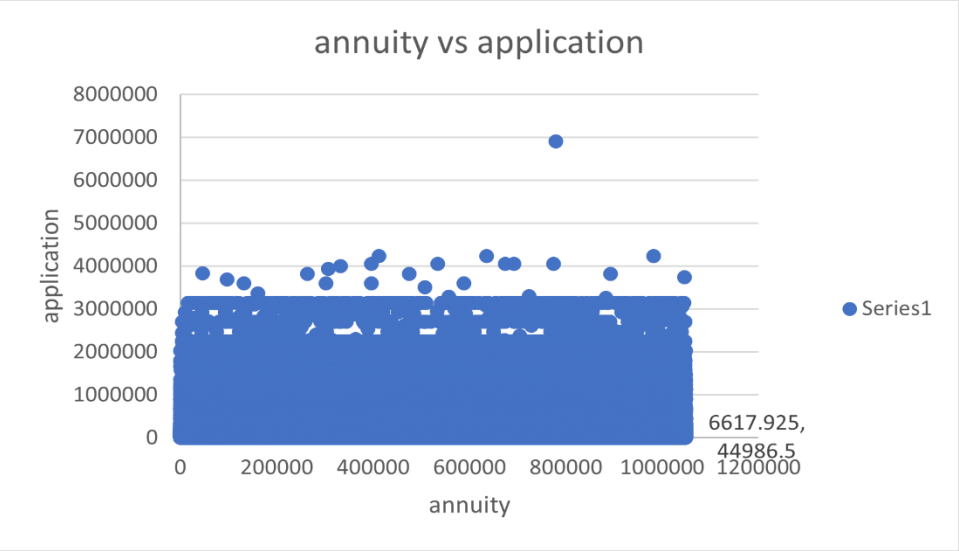


Figure 4.2.13

- Above graph shows the application amount vs annuity amount. We can also see a few outliers plotted over here

Payment type and channel type:

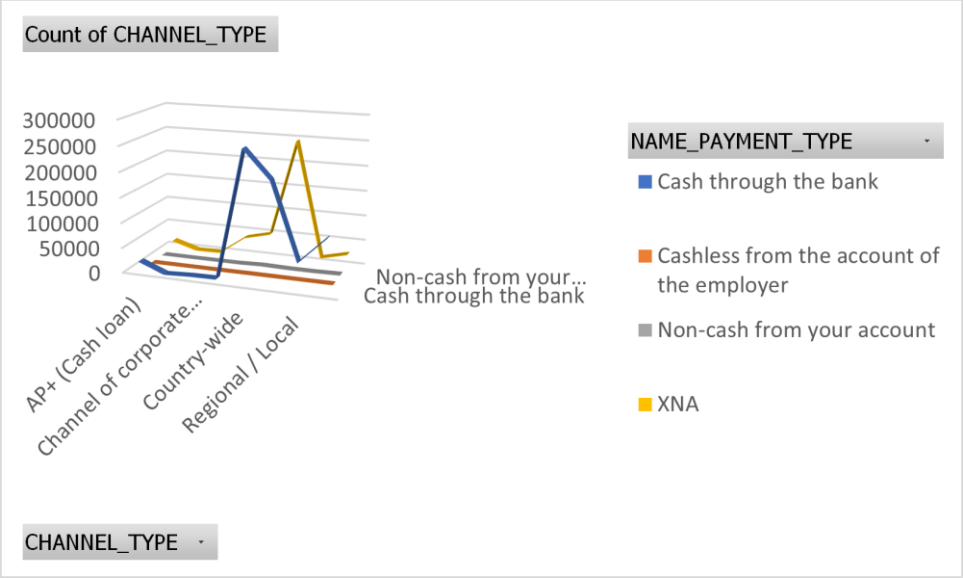


Figure 4.2.14

Count of CHANNEL_TYPE	Column Labels				
Row Labels	Cash through the bank	Cashless from the account of the employer	Non-cash from your account	XNA	Grand Total
AP+ (Cash loan)	20267	33	185	15320	35805
Car dealer	259	1		35	295
Channel of corporate sales	2787		2	1169	3958
Contact center	3507	26	51	38378	41962
Country-wide	259136	259	1794	51280	312469
Credit and cash offices	204210	163	720	245952	451045
Regional / Local	54476	42	707	13396	68621
Stone	105962	153	1774	26307	134196
Grand Total	**650604**	**677**	**5233**	**391837**	**1048351**

Figure 4.2.11

- Payment type that most of the client is preferred is cash through the bank. Channel type is credit and cash offices and country wide.

E. Correlation

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	days first drawing	days first due	days last due	days termination
AMT_ANNUITY	1						
AMT_APPLICATION	0.805166123	1					
AMT_CREDIT	0.81098189	0.975775339	1				
days first drawing	0.047564821	0.047654981	-0.023020849	1			
days first due	-0.052791034	-0.036210838	-0.002812151	-0.01706311	1		
days last due	0.047036601	0.10182193	0.133667943	-0.29223176	0.411940424	1	
days termination	0.035469243	0.086770291	0.126796082	-0.421058639	0.340352758	0.935737498	1

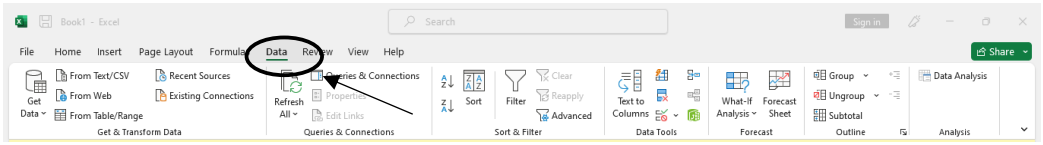
Figure 4.2.15

- Top correlations are
    - AMT_APPLICATION to AMT_ANNUITY
    - AMT_ANNUITY to AMT_CREDIT
    - AMT_CREDIT to AMT_APPLICATION
    - days termination to days last due
- top negative correlations are
    - days termination to days first drawing
    - days last due to days first drawing

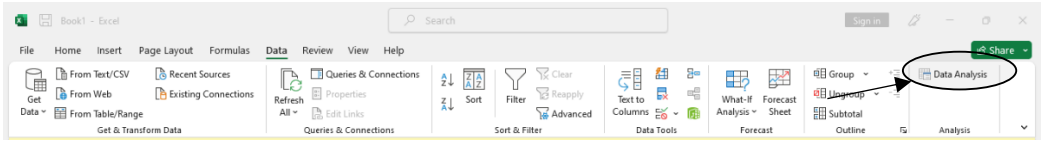
5. Result

- (cleaning) basic things to do in the data analysis process is data cleaning. Here 3 lakh and 10 lakh plus rows are in the tables. I delete some columns which has missing value more than 50%. for find the percentage of missing values I have used some function called =countblank(which count number of missing values) and =counta(which count number of values). Then after finding outliers, for finding outliers I have used box plot.
- (data analysing) for analysing data and provide interpretation I have used pivot table and pivot chart
- (correlation) step for finding correlation

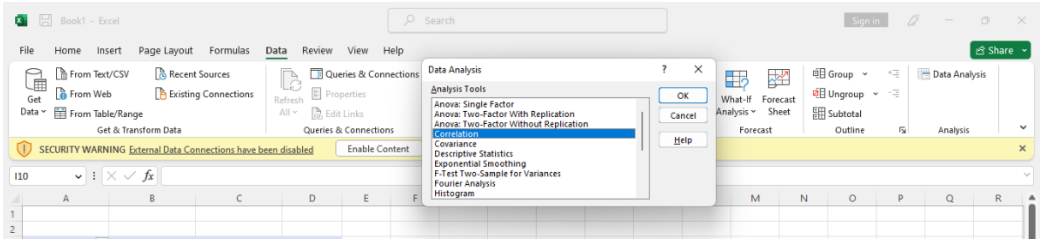
1) click on data tab



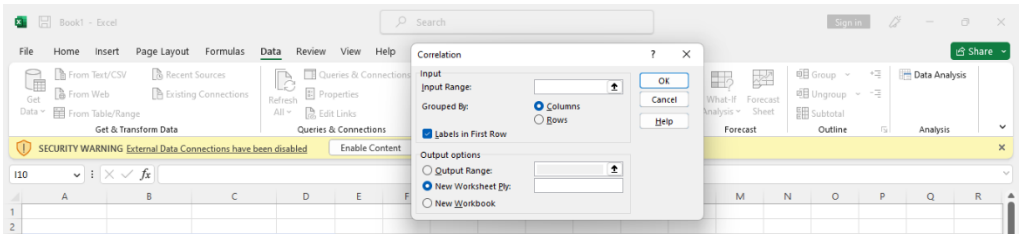
2) from analysis group click on data analysis



3) then a box will open.



4) select correlation and click ok



5) then after select range click ok.

6. conclusion

In this project bank loan case study we analysis the datasets and check if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. In this project I have use Microsoft excel for applying EDA(exploratory data analysis) which increase my excel skills and knowledge. I have mostly use pivot table and pivot chart for analysing the data's. The tables are contain 3 lakh and 10 lakh plus rows so it make difficult to analysis each column. When I use pivot table for analysing the table it make me much easier for analysing each columns and reduce the time also. We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. This will improve our analytical skills and visualization skills.