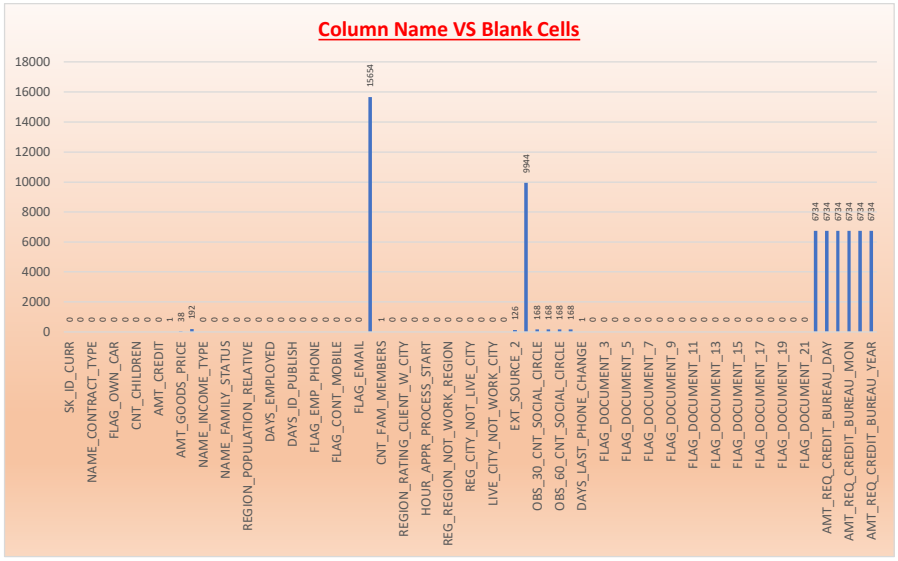
A. Identify Missing Data and Deal with it Appropriately:		Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.			
After Cleaning the Data	Total Rows	49999	After Cleaning the Data	Total Rows	49999
Current Loan Application Data / Column Names	Blank Cells	Percentage of missing values (for each column)	Previous Loan Application Data / Column Names	Blank Cells	Percentage of missing values (for each column)
SK_ID_CURR	0	0	SK_ID_PREV	0	0
TARGET	0	0	SK_ID_CURR	0	0
NAME_CONTRACT_TYPE	0	0	NAME_CONTRACT_TYPE	0	0
CODE_GENDER	0	0	AMT_ANNUITY	10592	21.18
FLAG_OWN_CAR	0	0	AMT_APPLICATION	0	0
FLAG_OWN_REALTY	0	0	AMT_CREDIT	0	0
CNT_CHILDREN	0	0	AMT_GOODS_PRICE	10744	21.49
AMT_INCOME_TOTAL	0	0	WEEKDAY_APPR_PROCESS_START	0	0
AMT_CREDIT	0	0	HOUR_APPR_PROCESS_START	0	0
AMT_ANNUITY	1	0	FLAG_LAST_APPL_PER_CONTRACT	0	0
AMT_GOODS_PRICE	38	0.08	NFLAG_LAST_APPL_IN_DAY	0	0
NAME_TYPE_SUITE	192	0.38	NAME_CASH_LOAN_PURPOSE	0	0
NAME_INCOME_TYPE	0	0	NAME_CONTRACT_STATUS	0	0
NAME_EDUCATION_TYPE	0	0	DAYS_DECISION	0	0
NAME_FAMILY_STATUS	0	0	NAME_PAYMENT_TYPE	0	0
NAME_HOUSING_TYPE	0	0	CODE_REJECT_REASON	0	0
REGION_POPULATION_RELATIVE	0	0	NAME_CLIENT_TYPE	0	0
DAYS_BIRTH	0	0	NAME_GOODS_CATEGORY	0	0
DAYS_EMPLOYED	0	0	NAME_PORTFOLIO	0	0
DAYS_REGISTRATION	0	0	NAME_PRODUCT_TYPE	0	0
DAYS_ID_PUBLISH	0	0	CHANNEL_TYPE	0	0
FLAG_MOBIL	0	0	SELLERPLACE_AREA	0	0
FLAG_EMP_PHONE	0	0	NAME_SELLER_INDUSTRY	0	0
FLAG_WORK_PHONE	0	0	CNT_PAYMENT	10592	21.18
FLAG_CONT_MOBILE	0	0	NAME_YIELD_GROUP	0	0
FLAG_PHONE	0	0	PRODUCT_COMBINATION	8	0.02
FLAG_EMAIL	0	0	DAYS_FIRST_DRAWING	19160	38.32
OCCUPATION_TYPE	15654	31.31	DAYS_FIRST_DUE	19160	38.32
CNT_FAM_MEMBERS	1	0	DAYS_LAST_DUE_1ST_VERSION	19160	38.32
REGION_RATING_CLIENT	0	0	DAYS_LAST_DUE	19160	38.32
REGION_RATING_CLIENT_W_CITY	0	0	DAYS_TERMINATION	19160	38.32
WEEKDAY_APPR_PROCESS_START	0	0	NFLAG_INSURED_ON_APPROVAL	19160	38.32
HOUR_APPR_PROCESS_START	0	0			
REG_REGION_NOT_LIVE_REGION	0	0			
REG_REGION_NOT_WORK_REGION	0	0			
LIVE_REGION_NOT_WORK_REGION	0	0			
REG_CITY_NOT_LIVE_CITY	0	0			
REG_CITY_NOT_WORK_CITY	0	0			
LIVE_CITY_NOT_WORK_CITY	0	0			
ORGANIZATION_TYPE	0	0			
EXT_SOURCE_2	126	0.25			
EXT_SOURCE_3	9944	19.89			
OBS_30_CNT_SOCIAL_CIRCLE	168	0.34			
DEF_30_CNT_SOCIAL_CIRCLE	168	0.34			
OBS_60_CNT_SOCIAL_CIRCLE	168	0.34			
DEF_60_CNT_SOCIAL_CIRCLE	168	0.34			
DAYS_LAST_PHONE_CHANGE	1	0			

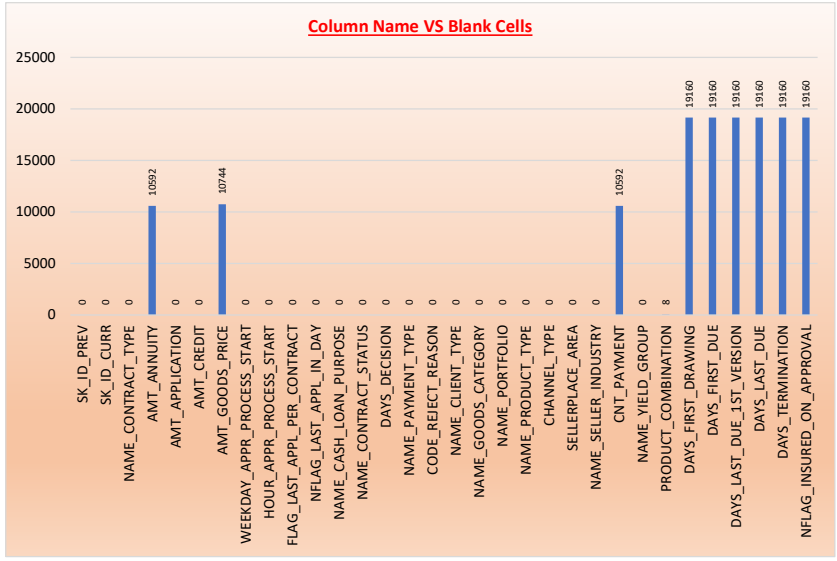
FLAG_DOCUMENT_2	0	0			
FLAG_DOCUMENT_3	0	0			
FLAG_DOCUMENT_4	0	0			
FLAG_DOCUMENT_5	0	0			
FLAG_DOCUMENT_6	0	0			
FLAG_DOCUMENT_7	0	0			
FLAG_DOCUMENT_8	0	0			
FLAG_DOCUMENT_9	0	0			
FLAG_DOCUMENT_10	0	0			
FLAG_DOCUMENT_11	0	0			
FLAG_DOCUMENT_12	0	0			
FLAG_DOCUMENT_13	0	0			
FLAG_DOCUMENT_14	0	0			
FLAG_DOCUMENT_15	0	0			
FLAG_DOCUMENT_16	0	0			
FLAG_DOCUMENT_17	0	0			
FLAG_DOCUMENT_18	0	0			
FLAG_DOCUMENT_19	0	0			
FLAG_DOCUMENT_20	0	0			
FLAG_DOCUMENT_21	0	0			
AMT_REQ_CREDIT_BUREAU_HOUR	6734	13.47			
AMT_REQ_CREDIT_BUREAU_DAY	6734	13.47			
AMT_REQ_CREDIT_BUREAU_WEEK	6734	13.47			
AMT_REQ_CREDIT_BUREAU_MON	6734	13.47			
AMT_REQ_CREDIT_BUREAU_QRT	6734	13.47			
AMT_REQ_CREDIT_BUREAU_YEAR	6734	13.47			

GRAPHICAL REPRESENTATION

Current Loan Application Data



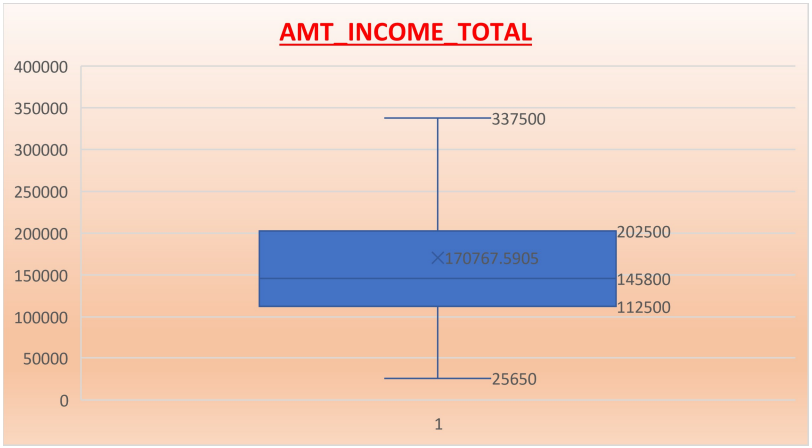
Previous Load Application Data

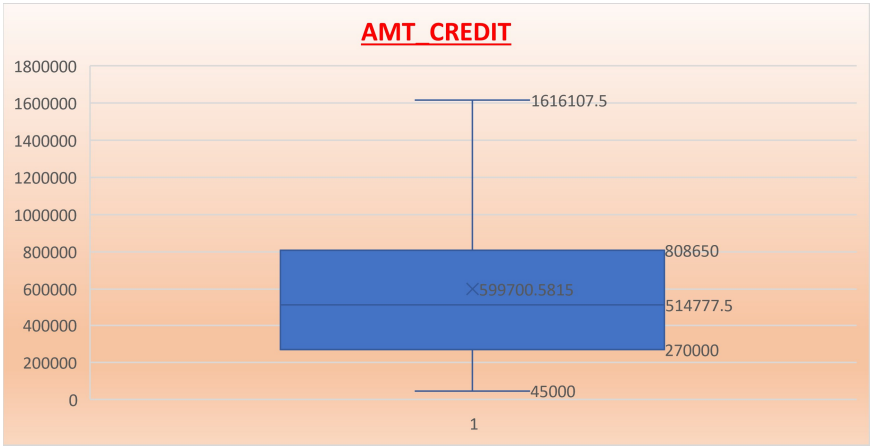


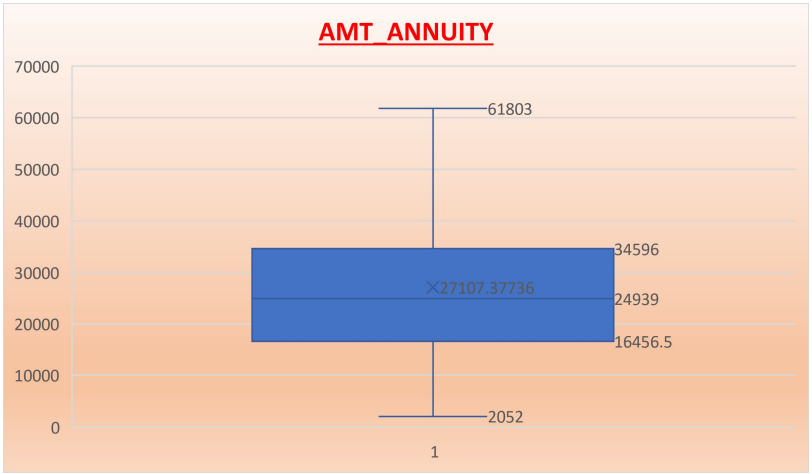
NOTE: **Replaced the null values in the numerical columns with average and median values**

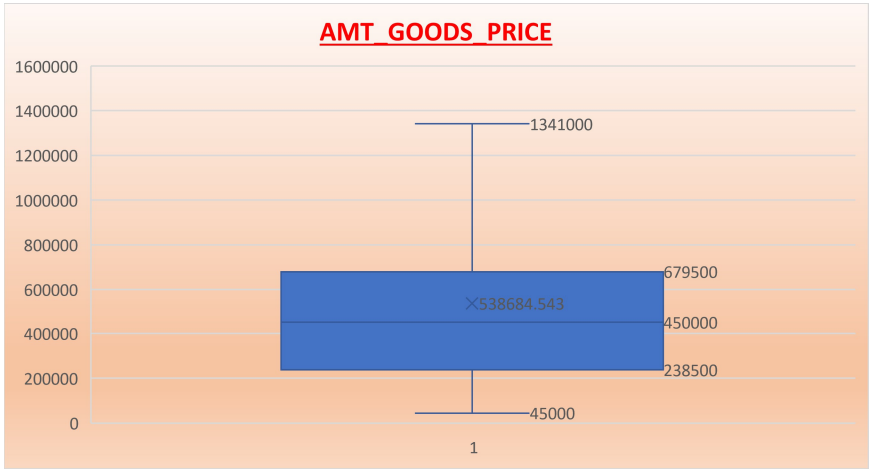
B. Identify Outliers in the Dataset:			Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.	
	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
Mean	170767.5905	599700.5815	27107.37736	538684.543
Median	145800	514777.5	24939	450000
Q1	112500	270000	16456.5	238500
Q2	145800	514777.5	24939	450000
Q3	202500	808650	34596	679500
IQR	90000	538650	18139.5	441000
Min	25650	45000	2052	45000
Max	117000000	4050000	258025.5	4050000
Upper Bound	337500	1616625	61805.25	1341000
Lower Bound	-22500	-537975	-10752.75	-423000

GRAPHICAL REPRESENTATION









C. Analyze Data Imbalance:

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

TARGET	COUNT	PERCENTAGE
1	4026	8.05
0	45973	91.95
TOTAL	49999	100

NAME_CONTRACT_TYPE	COUNT	PERCENTAGE
Cash loans	45276	90.55
Revolving loans	4723	9.45
TOTAL	49999	100

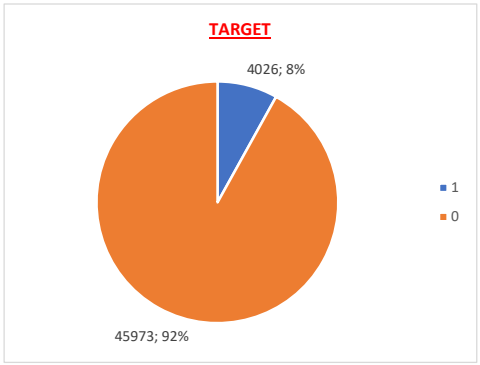
CODE_GENDER	COUNT	PERCENTAGE
M	17174	34.35
F	32823	65.65
XNA	2	0.004
TOTAL	49999	100

FLAG_OWN_CAR	COUNT	PERCENTAGE
N	32949	65.90
Y	17050	34.10
TOTAL	49999	100

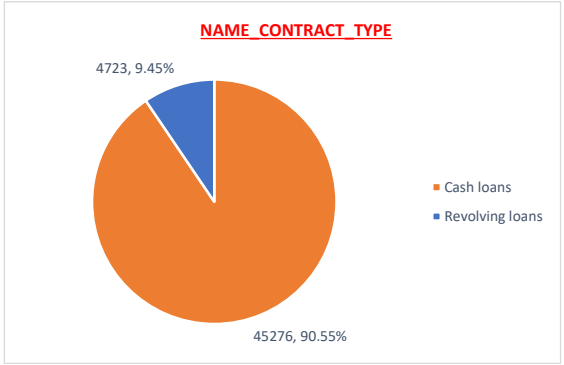
FLAG_OWN_REALTY	COUNT	PERCENTAGE
Y	34691	69.38
N	15308	30.62
TOTAL	49999	100

GRAPHICAL REPRESENTATION

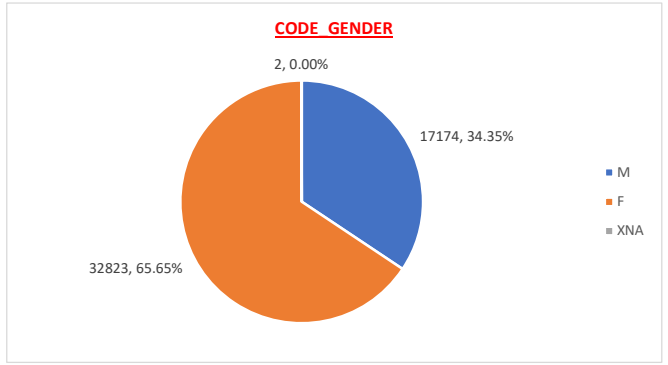
TARGET



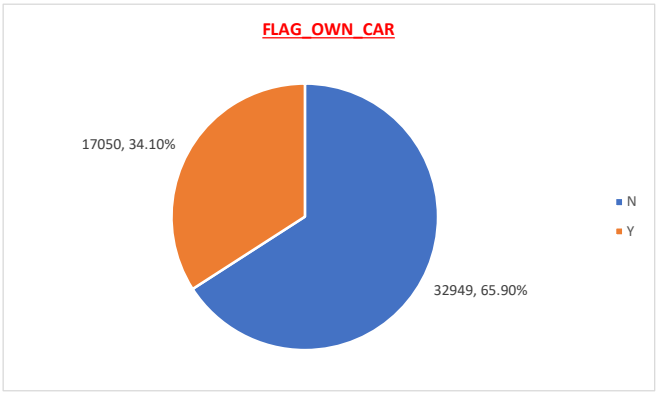
NAME_CONTRACT_TYPE



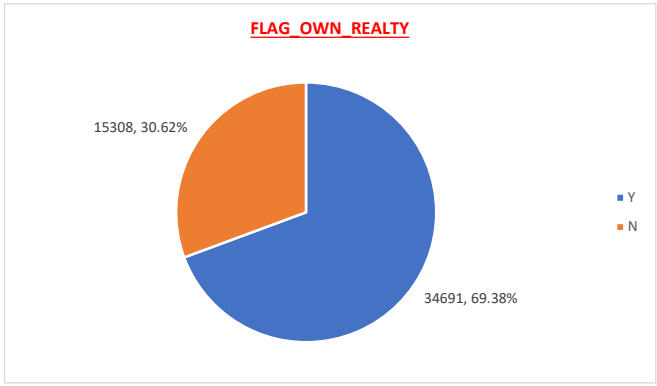
CODE_GENDER



FLAG_OWN_CAR



FLAG_OWN_REALTY

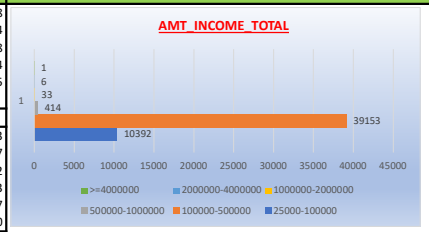


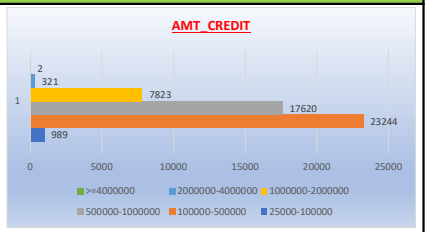
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables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

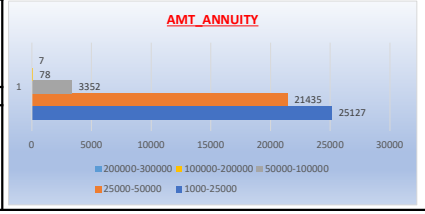
Univariate Analysis

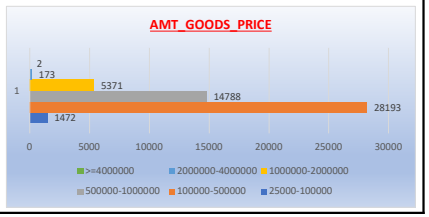
Column Name	Amount	Count	Average	Median	Mode	Min	Max	Variance	Standard Deviation
AMT_INCOME_TOTAL	25000-100000	10392	77672.0987	81000	90000	25650	99796.5	242031611.4	15557.36518
	100000-500000	39153	186093.023	162000	135000	100278	495000	5258402476	72514.84314
	500000-1000000	414	650549.9457	641250	540000	508500	967500	11922076293	109188.2608
	1000000-2000000	33	1292045.455	1125000	1125000	1035000	1935000	60656036932	246284.4634
	2000000-4000000	6	2662500	2250000	2250000	2025000	3825000	6.76688E+11	822610.175
	>=4000000	1	117000000	117000000	#N/A	117000000	117000000	#DIV/0!	#DIV/0!
AMT_CREDIT	25000-100000	989	75090.75379	76410	95940	45000	99576	281377452.4	16774.3093
	100000-500000	23244	294140.2676	270000	450000	100246.5	499500	11624580043	107817.3457
	500000-1000000	17620	703456.8449	675000	675000	500211	999886.5	19351428277	139109.4112
	1000000-2000000	7823	1272288.435	1223010	1125000	1000858.5	1998000	53279362242	230823.2273
	2000000-4000000	321	2233788.799	2220259.5	2250000	2008804.5	3956274	51348949857	226603.0667
	>=4000000	2	4050000	4050000	4050000	4050000	4050000	0	0
AMT_ANNUITY	1000-25000	25127	16324.75479	16515	9000	2052	24997.5	28779890.52	5364.689229
	25000-50000	21435	34216.66498	32760	37800	25002	49990.5	45164678.05	6720.4671
	50000-100000	3352	59880.5455	56607.75	51948	50004	99045	93330126.25	9660.751847
	100000-200000	78	120506.7692	112909.5	109597.5	101250	197230.5	390915367	19771.57978
	200000-300000	7	228045.2143	225000	225000	213291	258025.5	193811836.8	13921.63197
AMT_GOODS_PRICE	25000-100000	1472	75108.375	81000	90000	45000	99000	317944425.5	17830.9962
	100000-500000	28193	306678.394	270000	450000	103500	499500	13743583866	117233.0323
	500000-1000000	14788	734239.9652	679500	675000	502893	999000	15750332084	125500.327
	1000000-2000000	5371	1288697.797	1206000	1125000	1003500	1998000	49193470914	221796.0119
	2000000-4000000	173	2250338.15	2250000	2250000	2002500	3825000	35885357373	189434.3089
	>=4000000	2	4050000	4050000	4050000	4050000	4050000	0	0

GRAPHICAL REPRESENTATION









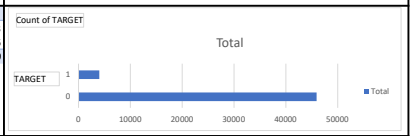
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:	Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables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

Segmented Variate Analysis

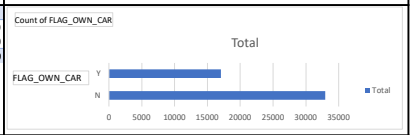
GRAPHICAL REPRESENTATION

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**

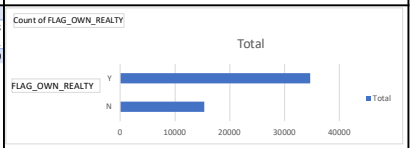
Row Labels	Count of TARGET
0	45973
1	4026
Grand Total	**49999**



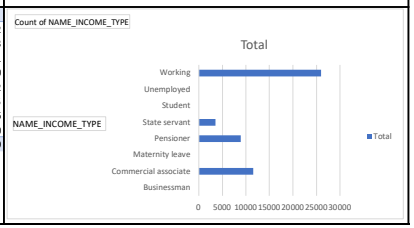
Row Labels	Count of FLAG_OWN_CAR
N	32949
Y	17050
Grand Total	**49999**



Row Labels	Count of FLAG_OWN_REALTY
N	15308
Y	34691
Grand Total	**49999**



Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010
Grand Total	**49999**



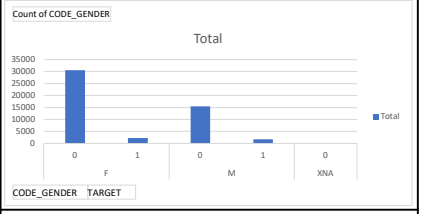
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:	Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables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

Bivariate Analysis

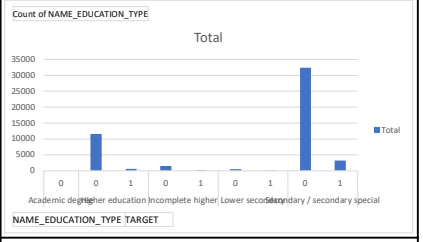
GRAPHICAL REPRESENTATION

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	**45276**
0	41484
1	3792
Revolving loans	**4723**
0	4489
1	234
Grand Total	**49999**

Row Labels	Count of CODE_GENDER
F	**32823**
0	30559
1	2264
M	**17174**
0	15412
1	1762
XNA	**2**
0	2
Grand Total	**49999**



Row Labels	Count of NAME_EDUCATION_TYPE
Academic degree	**20**
0	20
Higher education	**12167**
0	11561
1	606
Incomplete higher	**1620**
0	1482
1	138
Lower secondary	**620**
0	547
1	73
Secondary / secondary special	**35572**
0	32363
1	3209
Grand Total	**49999**



Row Labels	Count of NAME_INCOME_TYPE
Businessman	**2**
0	2
Commercial associate	**11543**
0	10679
1	864
Maternity leave	**1**
0	1
Pensioner	**8920**
0	8419
1	501
State servant	**3512**
0	3314
1	198
Student	**5**
0	5
Unemployed	**6**
0	4
1	2
Working	**26010**
0	23549
1	2461
Grand Total	**49999**

E. Identify Top Correlations for Different Scenarios:	Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions							
Target '0' Correlation								
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
CNT_CHILDREN	1	0.036319722	0.005705458	0.02638212	0.001383048	-0.024912809	0.021288992	0.017873365
RANK	1	2	6	3	7	8	4	5
AMT_INCOME_TOTAL	0.036319722	1	0.377965752	0.451135629	0.384486402	0.181941261	-0.205031899	-0.220044862
RANK	6	1	4	2	3	5	7	8
AMT_CREDIT	0.005705458	0.377965752	1	0.770771802	0.986879648	0.095539444	-0.102556478	-0.111639948
RANK	6	4	1	3	2	5	7	8
AMT_ANNUITY	0.02638212	0.451135629	0.770771802	1	0.775888179	0.117280527	-0.129921191	-0.143197531
RANK	6	4	3	1	2	5	7	8
AMT_GOODS_PRICE	0.001383048	0.384486402	0.986879648	0.775888179	1	0.099047191	-0.104516599	-0.112777867
RANK	6	4	2	3	1	5	7	8
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539444	0.117280527	0.099047191	1	-0.539333113	-0.536859601
RANK	6	2	5	3	4	1	8	7
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.102556478	-0.129921191	-0.104516599	-0.539333113	1	0.950468157
RANK	3	7	4	6	5	8	1	2
REGION_RATING_CLIENT_W_CITY	0.017873365	-0.220044862	-0.111639948	-0.143197531	-0.112777867	-0.536859601	0.950468157	1
RANK	3	7	4	6	5	8	2	1
Target '1' Correlation								
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
CNT_CHILDREN	1	0.010110177	0.007601905	0.029172977	-0.000680095	-0.020359154	0.055515557	0.054802235
RANK	1	5	6	4	7	8	2	3
AMT_INCOME_TOTAL	0.010110177	1	0.015271444	0.018004594	0.013298258	-0.006180303	-0.012846697	-0.01266585
RANK	5	1	3	2	4	6	8	7
AMT_CREDIT	0.007601905	0.015271444	1	0.749665201	0.982381964	0.067775624	-0.045024534	-0.052954314
RANK	6	5	1	3	2	4	7	8
AMT_ANNUITY	0.029172977	0.018004594	0.749665201	1	0.749904665	0.073123998	-0.061578289	-0.079418668
RANK	5	6	3	1	2	4	7	8
AMT_GOODS_PRICE	-0.000680095	0.013298258	0.982381964	0.749904665	1	0.077209215	-0.051709204	-0.05713431
RANK	6	5	2	3	1	4	7	8
REGION_POPULATION_RELATIVE	-0.020359154	-0.006180303	0.067775624	0.073123998	0.077209215	1	-0.430032303	-0.431675881
RANK	6	5	4	3	2	1	7	8
REGION_RATING_CLIENT	0.055515557	-0.012846697	-0.045024534	-0.061578289	-0.051709204	-0.430032303	1	0.950768899
RANK	3	4	5	7	6	8	1	2
REGION_RATING_CLIENT_W_CITY	0.054802235	-0.01266585	-0.052954314	-0.079418668	-0.05713431	-0.431675881	0.950768899	1
RANK	3	4	5	7	6	8	2	1

Project 06 - Bank Loan Case Study	
Data Cleaning	
Current Loan Application Data	Initially 122 columns & 50000 rows
	Deleted columns which contains more than 40% of blank cells
	Finally we have 73 colums & 50000 rows
Previous Loan Application Data	Initially 37 columns & 50000 rows
	Deleted columns which contains more than 40% of blank cells
	Finally we have 32 colums & 50000 rows

FINAL SUMMARY

Sr No	Question	Answer
1	Project Description	Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans. Your task is to use Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.
2	Approach	1. Understood the dataset provided 2. Have done the data cleaning as described in the data cleaning table 3. Used appropriate functions and formulas to get the requried answers for each questions
2(A)	Identify Missing Data and Deal with it Appropriately	COUNTBLANKS - Used to count the blank cells
		Calculated the percentage of blank cells for each column
		Highlighted the percentage above 40% and removed all those columns
		Visualized the final data set representing the count of blank cells for each column
2(B)	Identify Outliers in the Dataset	Calculated the mean, median, range, Q1, Q2, Q3, IQR and lower & upper bounds
		Visualized the box plot for the same
2(C)	Analyze Data Imbalance	Calculated the count of different segments for each variables
		Visualized the same using pie chart
2(D)	Perform Univariate, Segmented Univariate, and Bivariate Analysis	Univariate - Calculated the mean, median, average, min, max, mode , variance, standard deviation for different numerical columns (created with different segements of values)
		Univariate - Visualized the same using bar chart
		Segmented - Used pivot tables to show the count of different segments for each variable
		Segmented - Visualized the same using pivot chart
		Bivariate - Used pivot table to show the count of each segments for different variables according to target variable
		Bivariate - Visualized the same using pivot chart
2(E)	Identify Top Correlations for Different Scenarios	Created a two new data sets containing datas of target variable '0' and '1'
		Created corelation matrix for different varibles and their rank using 'CORREL' & 'RANK' function
3	Tech-Stack Used	Microsoft Office 2019
4	Insights	In this projects I have used different excel functions, formulas, charts to extract the answers for all the questions which has helped me to improve the way of thinking while working on excel and selecting appropriate functions according to the questions.
5	Result	This project has helped me to improve my skills on advanced excel functions, pivot tables and charts