Bank Loan Case Study

Victor Shah

Project Description

As a data analyst at a finance company that specializes in lending various types of loans to urban customers, my primary responsibility is about using Exploratory Data Analysis (EDA) to analyze patterns in the data and ensure that capable applicants are not rejected.

Some of the challenges faced by the company are : some customers who don’t have a sufficient credit history take advantage of this and default on their loans.

Approach

- Downloading the dataset: The first step is downloading the excel file (.csv) into the local device. Make sure the downloaded file is having the extension (.xlsx)
- Understanding the worksheet: The next step is to examine the structure of the table holding the data in the Excel Sheet. (application_data.csv, columns_description.csv, previous_application.csv)
- Identifying the key tables: Identification of the primary key from the dataset of excel files.

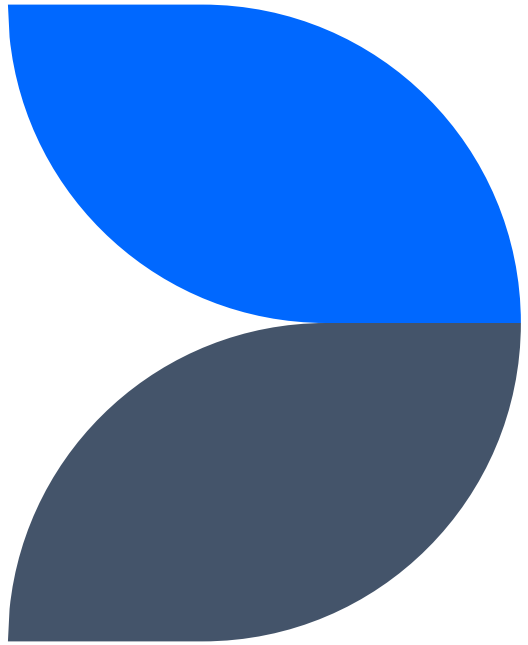
- Data Cleaning: This is the preprocessing step that makes the data suitable for analysis. It includes handling missing values, removing duplicates.

Analyzing the four possible outcomes:

1. Approved: The company has approved the loan application.
2. Cancelled: The customer cancelled the application during the approval process.
3. Refused: The company rejected the loan.
4. Unused Offer: The loan was approved but the customer did not use it.

- Data Visualization: To use EDA to understand how customer attributes and loan attributes influence the likelihood of default.

DATA ANALYTICS TASKS



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.

Missing Data for the dataset application_data.csv

	AMT_ANN	AMT_GOC	NAME_TYPE_SUI	CNT_FAM	EXT_SOUF	EXT_SOUF	OBS_30_C	DEF_30_C	OBS_60_C	DEF_60_C	DAYS_LAS	AMT_REQ	AMT_REQ	AMT_REQ	AMT_REQ	AMT_REQ	AMT_REQ
Count Blank	1	38	192	1	126	9915	168	168	168	168	1	6719	6719	6719	6719	6719	6719
Blank %	0.002%	0.076%	0.384%	0.002%	0.252%	19.830%	0.336%	0.336%	0.336%	0.336%	0.002%	13%	13%	13%	13%	13%	13%
Mean	27112.51	538508.2	-	2.159194	0.513673	0.51192	1.421141	0.141844	1.404017	0.09829	-965.972	0.007079	0.007519	0.032468	0.270609	0.26114	1.885699
Median	24939	450000	-	2	0.565467	0.535276	0	0	0	0	-757	0	0	0	0	0	1
Mode	-	-	Unaccompanied	-	-	-	-	-	-	-	-	-	-	-	-	-	-

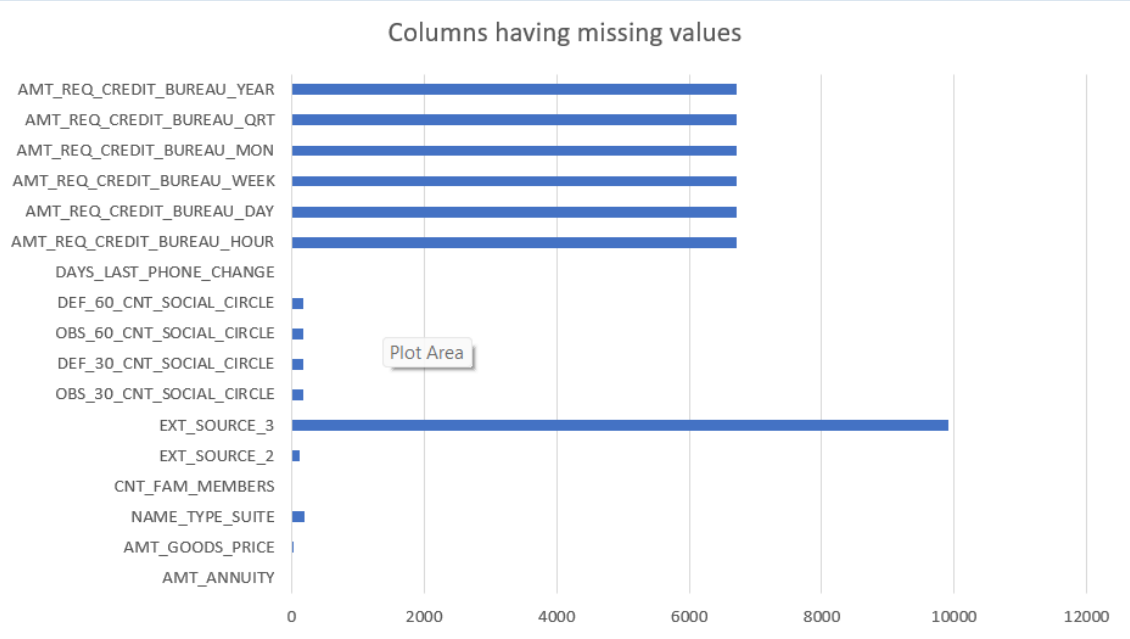
Missing Data for the dataset previous_application.csv

	AMT_ANN	AMT_GOC	CNT_PAYI	PRODUCT_COMBINATION
Count Blank	10592	10744	10592	8
Blank %	21%	21%	21%	0%
Mean	15482.6	215141.4	15.55589	-
Median	10879.92	104017.5	12	-
Mode	-	-	-	POS household without interest

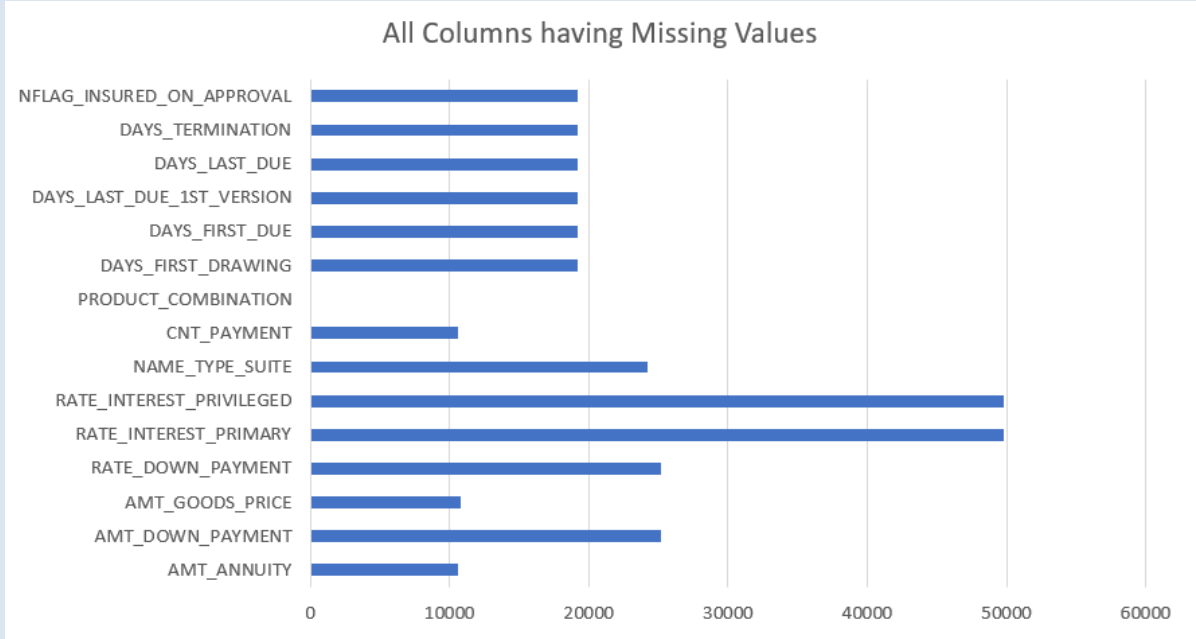
Create a bar chart or column chart to visualize the proportion of missing values for each variable.

Missing Data Visualization for:

dataset application_data.csv



previous_application.csv



B. Identify Outliers in the Dataset:

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

Outlier check representation table for the dataset application_data.csv

	Quartile Range					Inner Quartile Range	
	Min	Q1	Median	Q3	Max	Lower Limit	Upper Limit
AMT_INCOME_TOTAL	25650	112500	145800	202500	3825000	-22500	337500
AMT_CREDIT	45000	270000	514777.5	808650	4050000	-537975	1616625
AMT_ANNUITY	2052	112500	145800	202500	3825000	-22500	337500
AMT_GOODS_PRICE	45000	238500	450000	679500	4050000	-423000	1341000
REGION_POPULATION_R	0.000533	0.010006	0.01885	0.028663	0.072508	-0.0179795	0.0566485
DAYS_EMPLOYED	-17531	-2786	-1221	-292	365243	-6527	3449
DAYS_REGISTRATION	-22392	-7463.75	-4490	-1998	0	-15662.375	6200.625
EXT_SOURCE_3	0.000527265	0.417099668	0.53527625	0.638043528	0.896009549	0.085683878	0.969459318

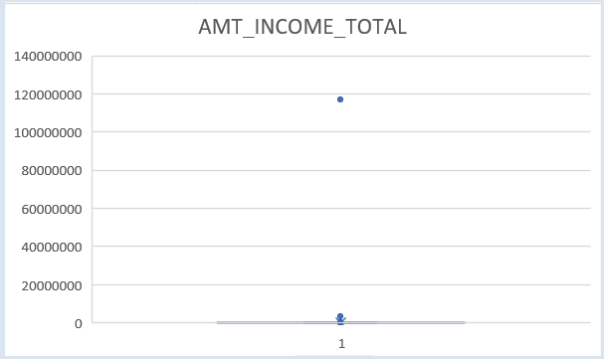
Outlier check representation table for the dataset previous_application.csv

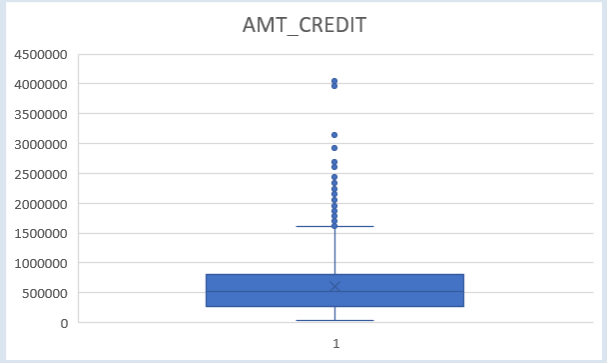
	Quartile Range					Inner Quartile Range	
	Min	Q1	Median	Q3	Max	Lower Limit	Upper Limit
AMT_ANNUITY	0	7189.74	10879.92	16256.16	234478.4	-6409.89	29855.79
AMT_APPLICATION	0	22045.5	71550	180000	3826373	-214886.25	416931.75
AMT_CREDIT	0	26055	78907.5	198105.8	4104351	-232021.13	456181.875
AMT_GOODS_PRICE	0	63663.75	104017.5	180000	3826373	-110840.63	354504.375

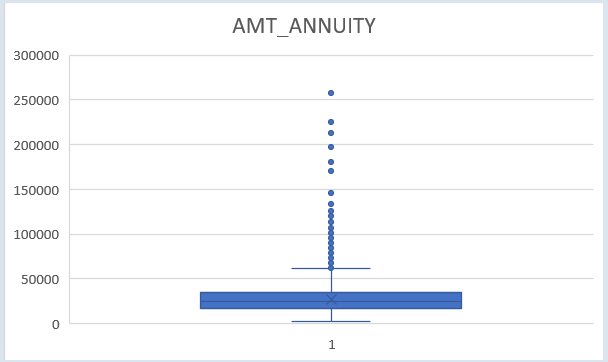
Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

Data Visualization for:

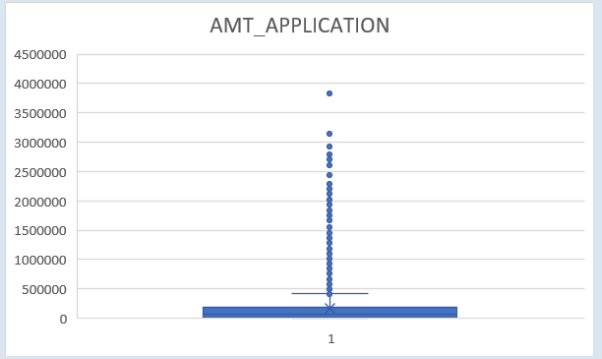
dataset application_data.csv

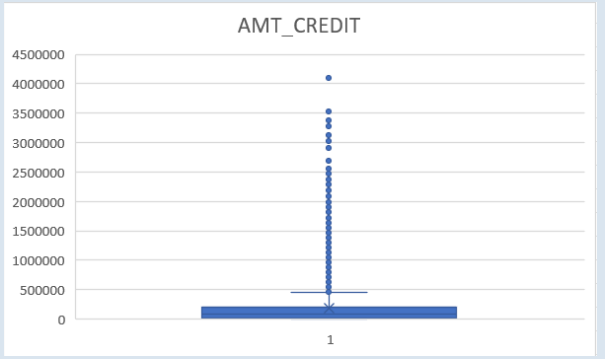


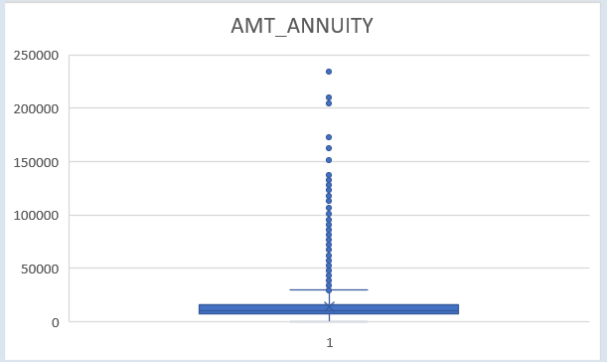




previous_application.csv







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.

Data Imbalance check representation table for the dataset application_data.csv

	1 - client with payment difficulties	0 - all other cases	SUM
TARGET	4025	45973	49998

	Cash loans	Revolving Loans	SUM
NAME_CONTRACT_TYPE	45275	4723	49998

	Male	Female	SUM
CODE_GENDER	17174	32822	49996

	1	0	SUM
FLAG_OWN_CAR	17050	32948	49998
FLAG_OWN_REALTY	34690	15308	49998
FLAG_MOBIL	49997	1	49998
FLAG_EMP_PHONE	41072	8926	49998
FLAG_WORK_PHONE	9963	40035	49998
FLAG_CONT_MOBILE	49897	101	49998
FLAG_PHONE	13886	36112	49998
FLAG_EMAIL	2783	47215	49998
	1	0	SUM
REG_REGION_NOT_LIVE_REGION	750	49248	49998
REG_REGION_NOT_WORK_REGION	2496	47502	49998
LIVE_REGION_NOT_WORK_REGION	1982	48016	49998
REG_CITY_NOT_LIVE_CITY	3998	46000	49998
REG_CITY_NOT_WORK_CITY	11608	38390	49998
LIVE_CITY_NOT_WORK_CITY	8985	41013	49998

Data Imbalance check representation table for the dataset previous_application.csv

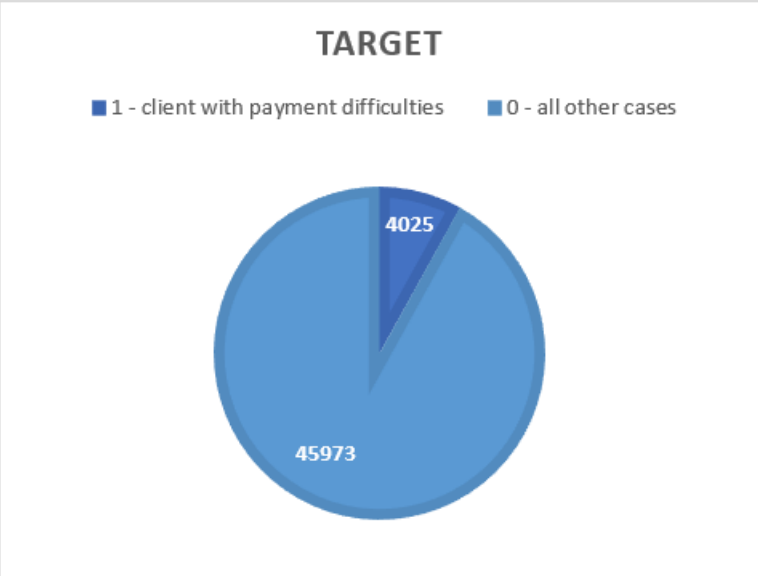
	Consumer loans	Cash loans	Revolving loans	XNA	SUM
NAME_CONTRACT_TYPE	23510	20856	5625	8	49999

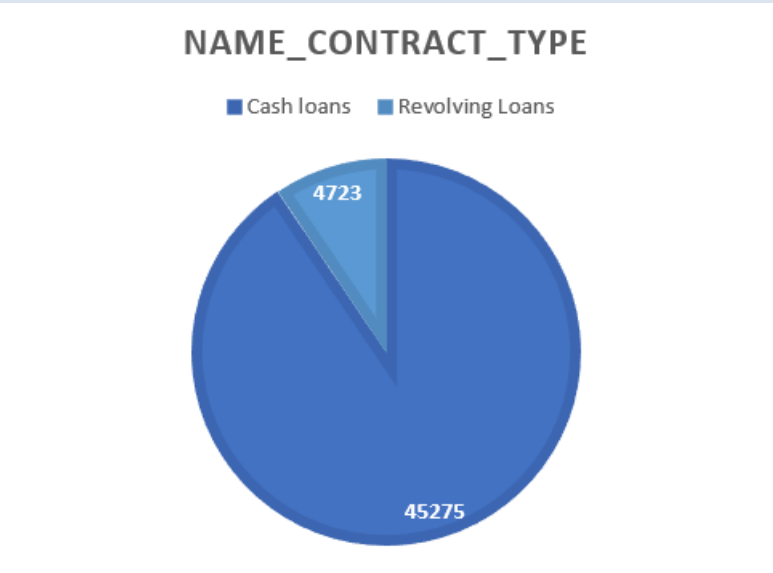
	Approved	Refused	Canceled	Unused offer	SUM
NAME_CONTRACT_STATUS	31885	8660	8595	859	49999

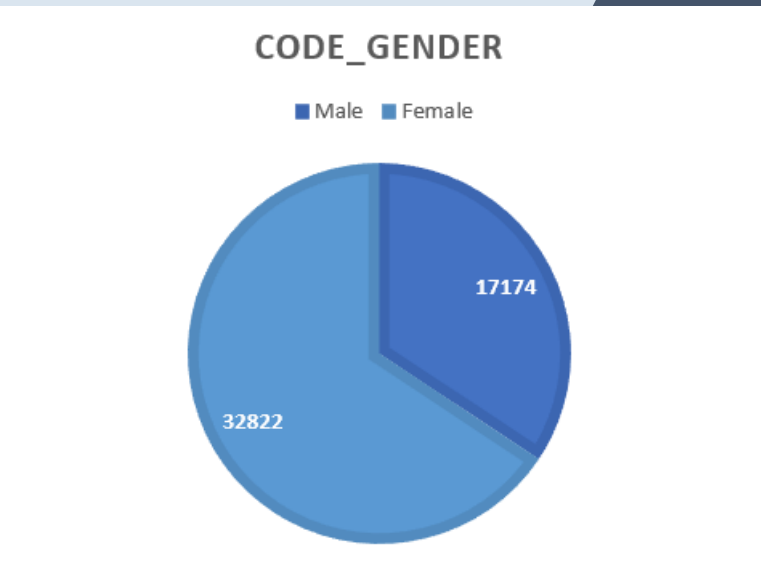
	Repeater	New	Refreshed	XNA	SUM
NAME_CLIENT_TYPE	36167	9548	4227	57	49999

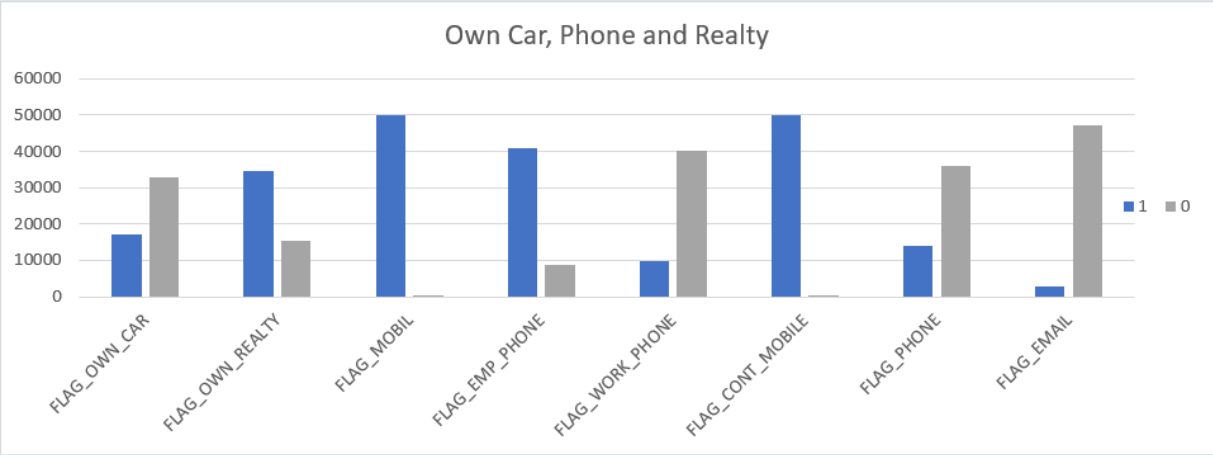
Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

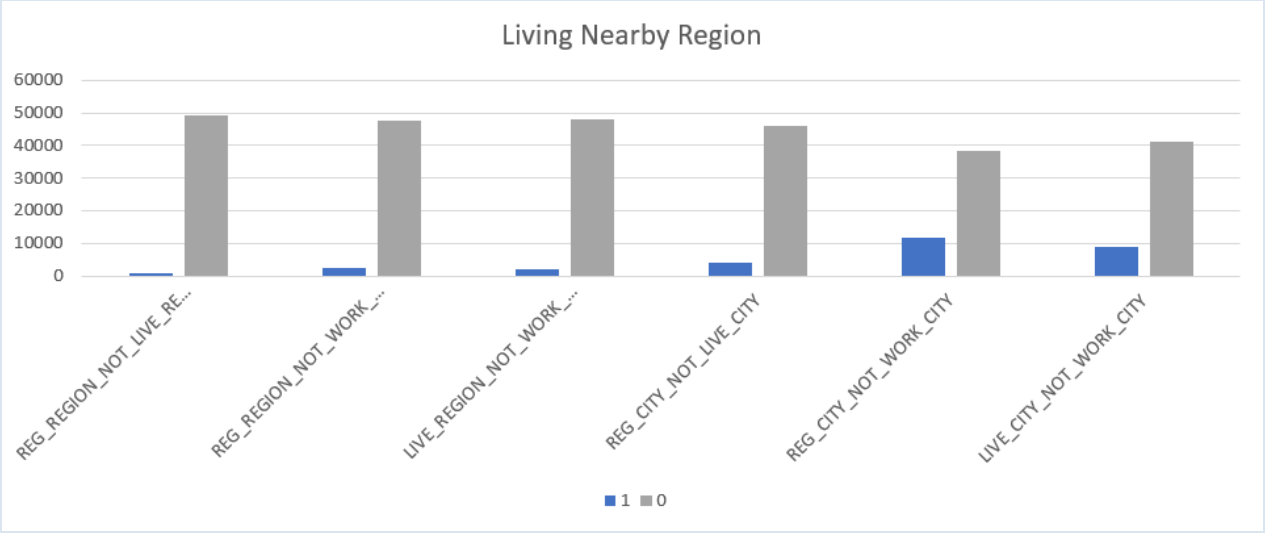
dataset application_data.csv





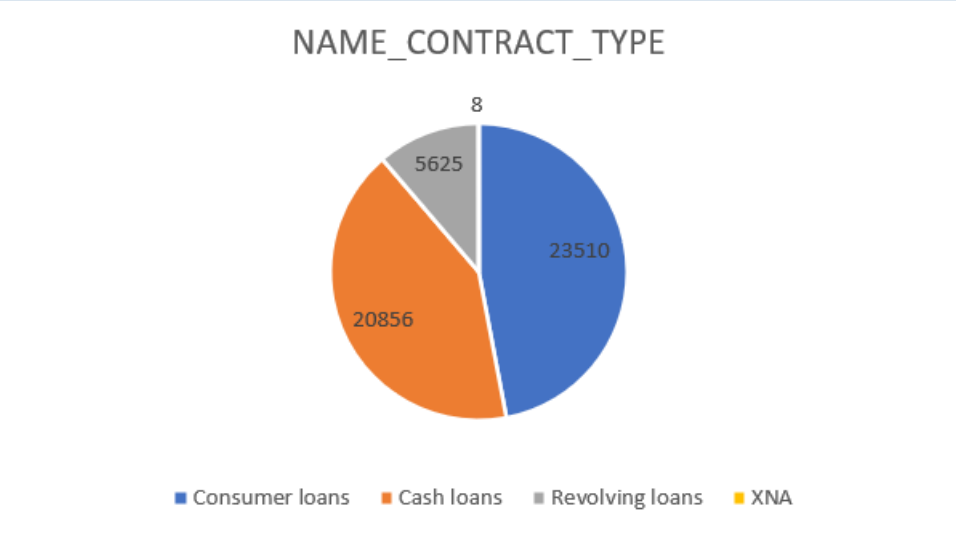


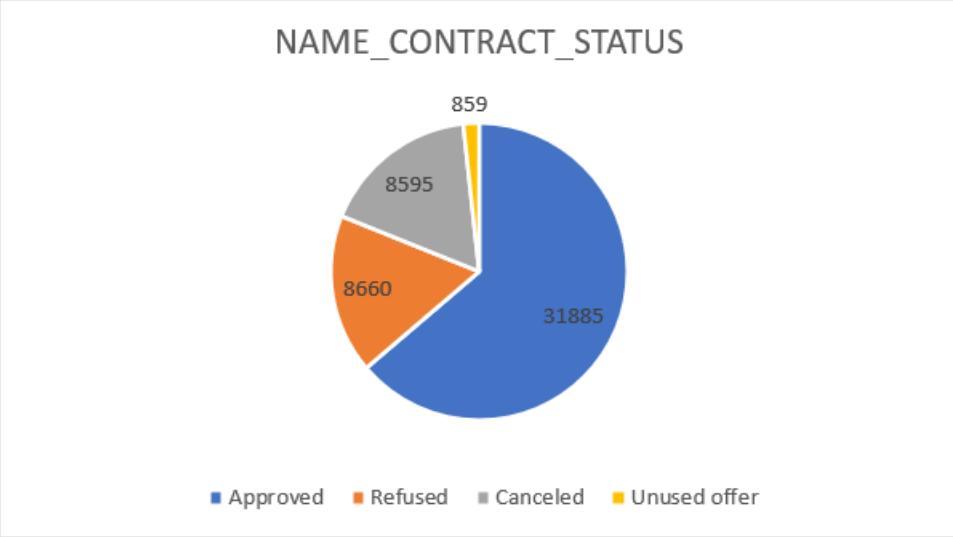


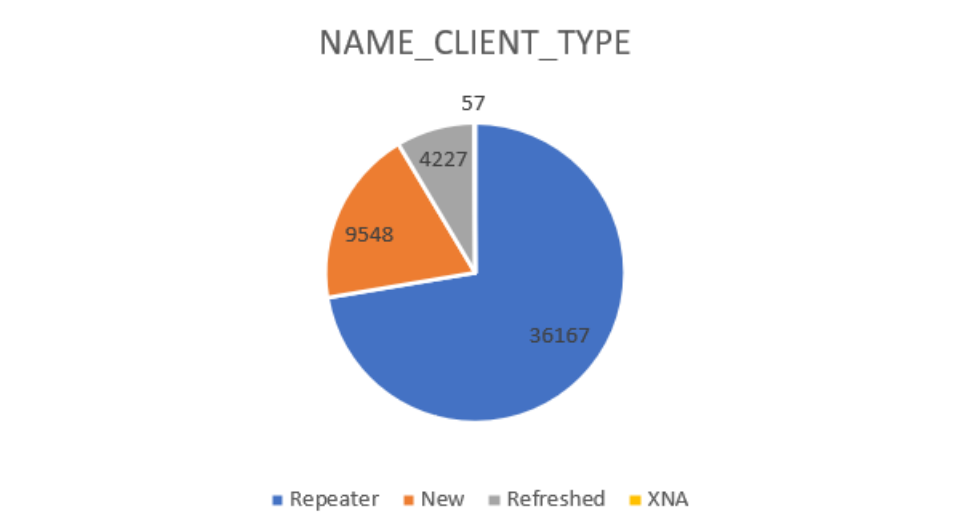


Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

previous_application.csv







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

Performing Univariate Analysis on application_data.csv

	Descriptive Statistics					
	Mean	Median	Mode	Max	Min	Standard Deviation
AMT_INCOME_TOTAL	168430.9124	145800	135000	3825000	25650	99166.40542
AMT_CREDIT	599701.3258	514777.5	450000	4050000	45000	402419.4239
AMT_ANNUITY	27107.35224	24939	9000	258025.5	2052	14562.94709
AMT_GOODS_PRICE	538994.039	450000	450000	4050000	45000	369724.3268

		Frequency	Percentage	Cumulavtive Percentage
NAME_CONTRACT_TYPE	Cash loans	45275	91%	91%
	Revolving loans	4723	9%	100%

		Frequency	Percentage	Cumulavtive Percentage
WEEKDAY_APPR_PROCESS_START	SUNDAY	2616	5%	5%
	MONDAY	8385	17%	22%
	TUESDAY	8740	17%	39%
	WEDNESDAY	8355	17%	56%
	THURSDAY	8149	16%	72%
	FRIDAY	8286	17%	89%
	SATURDAY	5467	11%	100%

		Frequency	Percentage	Cumulavtive Percentage
NAME_FAMILY_STATUS	Civil marriage	4859	10%	10%
	Single / not married	7306	15%	25%
	Married	32093	64%	89%
	Separated	3142	6%	95%
	Widow	2597	5%	100%

Performing Univariate Analysis on previous_application.csv

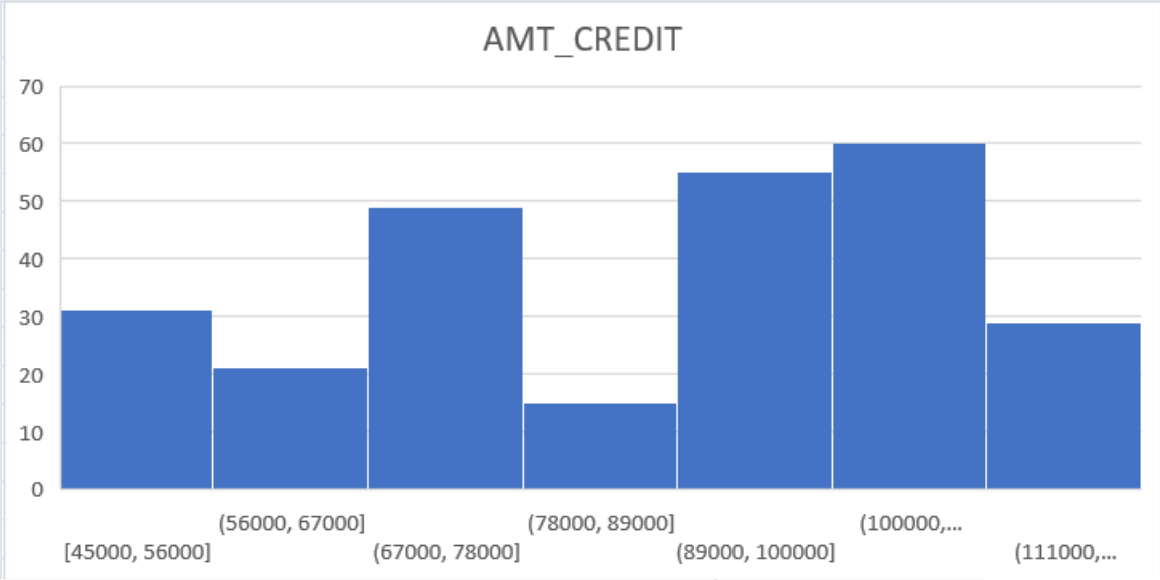
	Descriptive Statistics					
	Mean	Median	Mode	Max	Min	Standard Deviation
AMT_ANNUITY	14507.54628	10879.92	10879.92	234478.395	0	13036.66787
AMT_APPLICATION	168892.4546	71550	0	3826372.5	0	282203.5105
AMT_CREDIT	188542.8855	78907.5	0	4104351	0	308473.6014
AMT_GOODS_PRICE	191262.6324	104017.5	104017.5	3826372.5	0	271892.6356

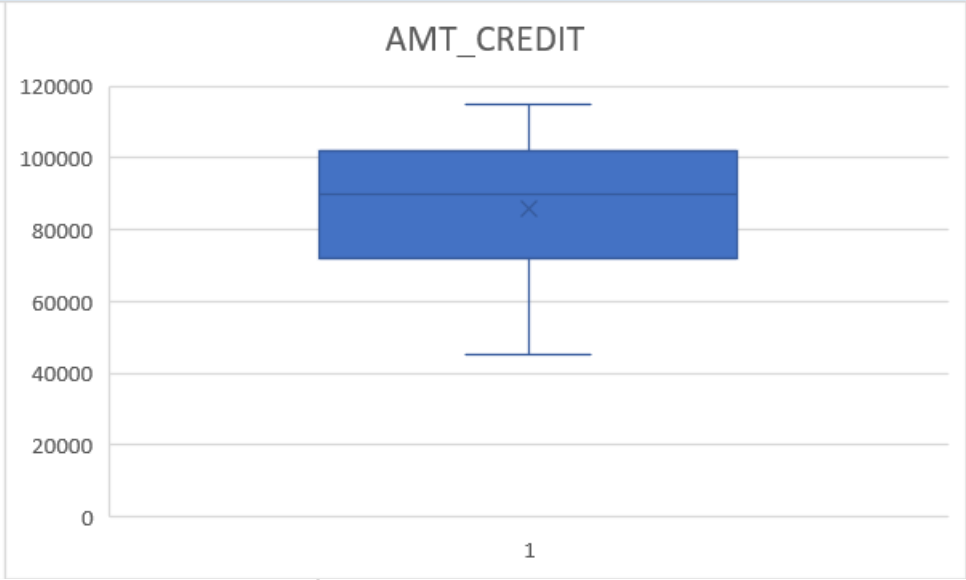
		Frequency	Percentage	Cumulavtive Percentage
NAME_CONTRACT_TYPE	Cash loans	20856	41.71%	42%
	Revolving loans	5625	11.25%	53%
	Consumer loans	23509	47.02%	100%
	XNA	8	0.02%	100%

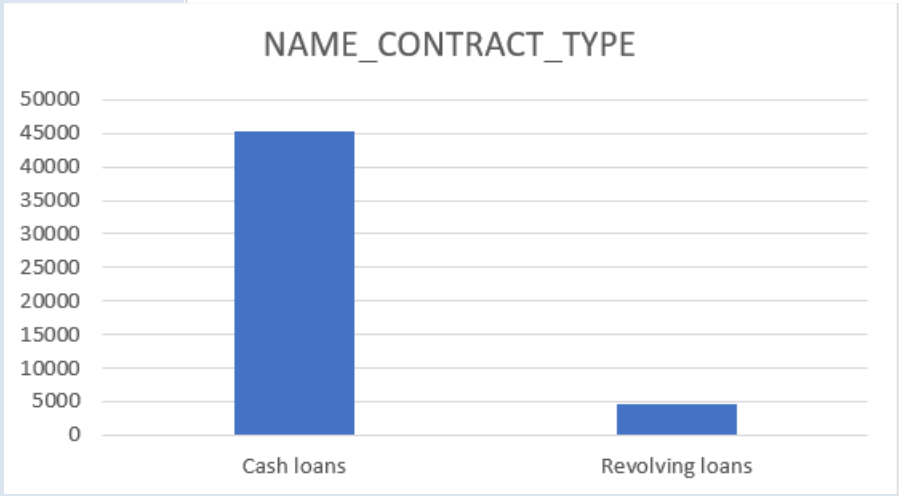
		Frequency	Percentage	Cumulavtive Percentage
WEEKDAY_APPR_PROCES	MONDAY	7419	15%	15%
	TUESDAY	7504	15%	30%
	WEDNESDAY	7649	15%	45%
	THURSDAY	7460	15%	60%
	FRIDAY	7554	15%	75%
	SATURDAY	7380	15%	90%
	SUNDAY	5033	10%	100%

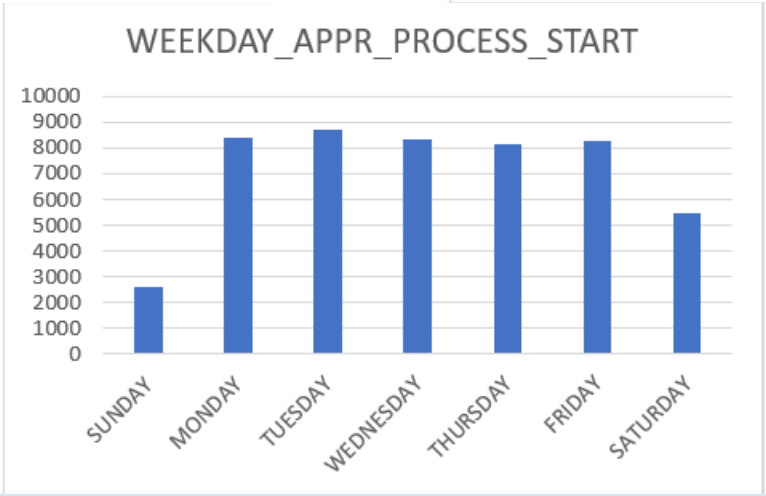
Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

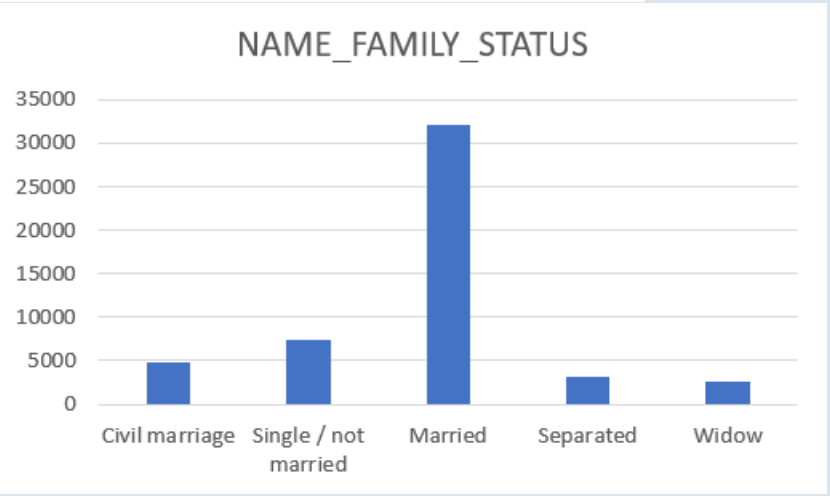
Data Visualization for application_data.csv





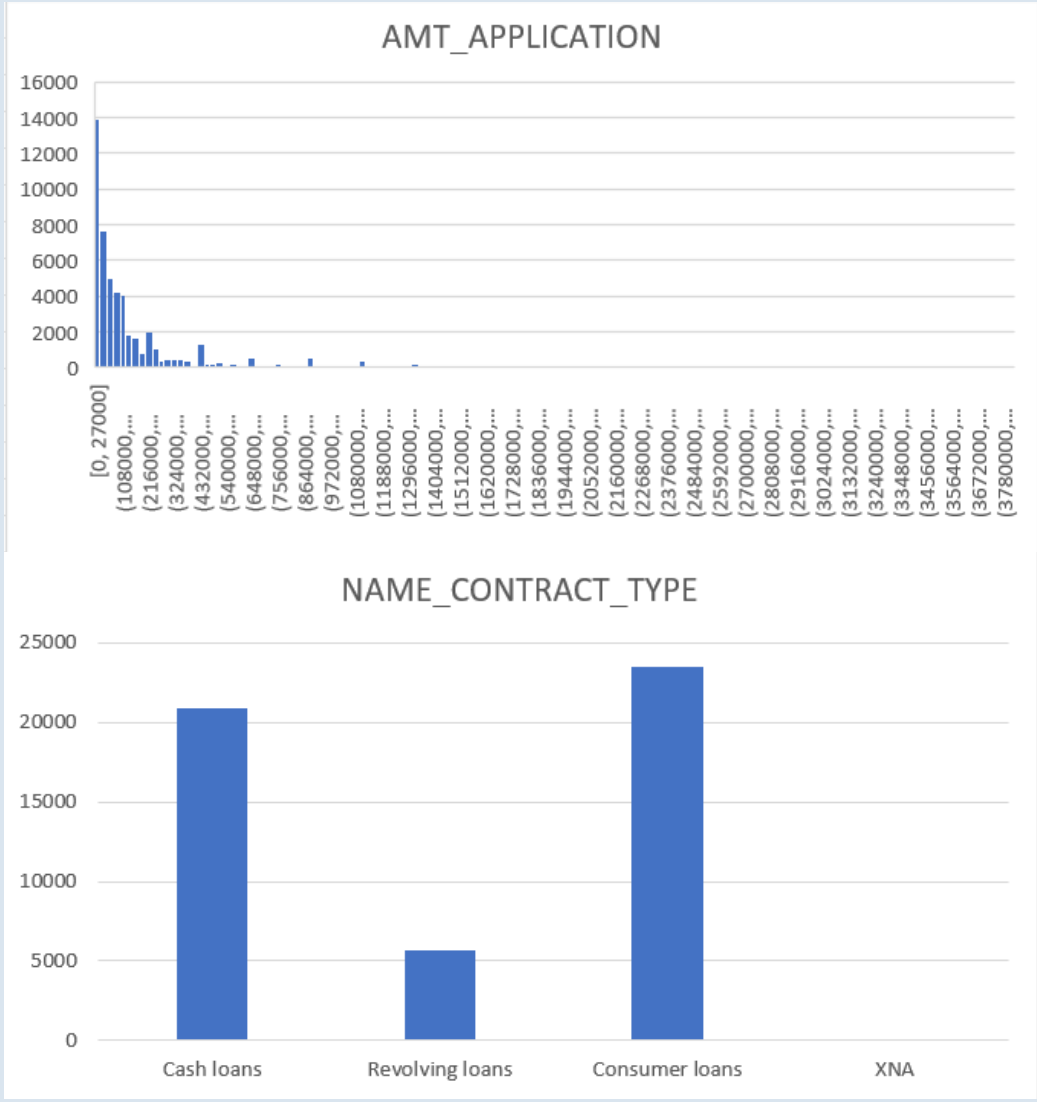


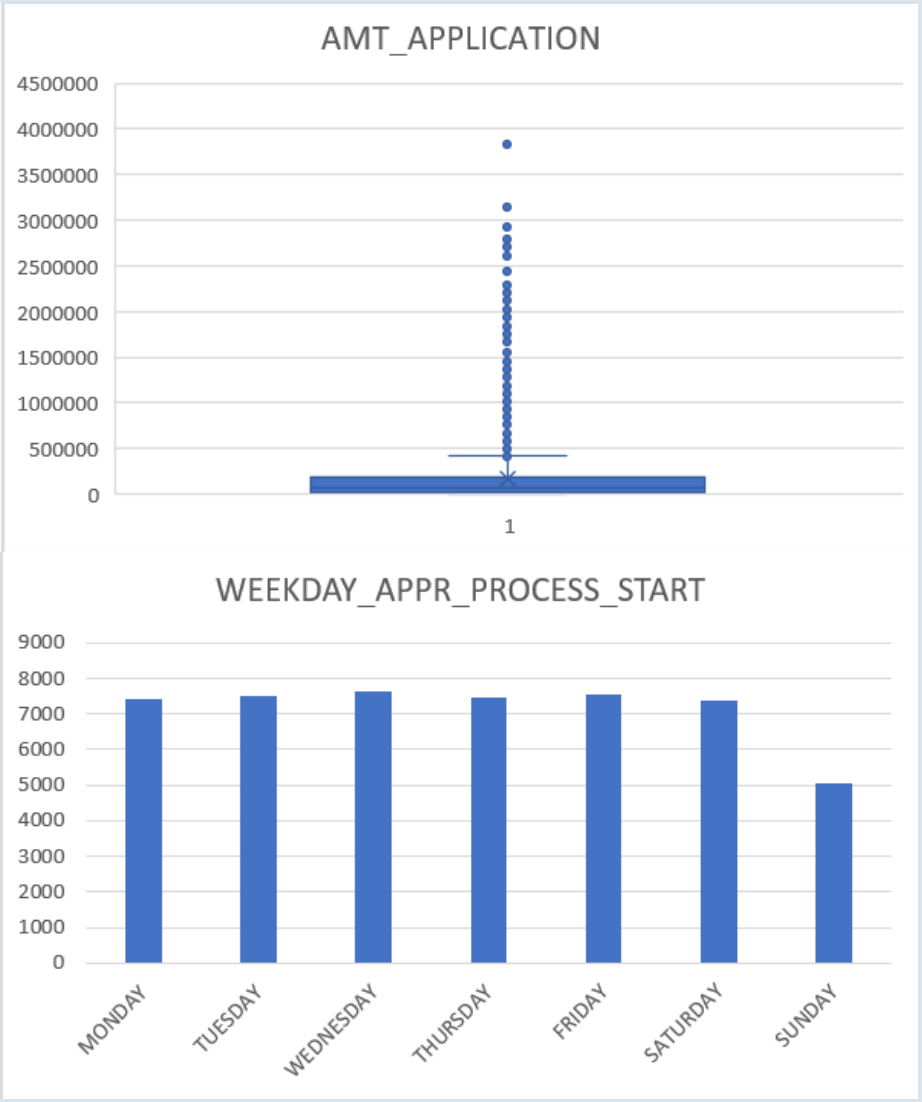




Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Data Visualization for previous_application.csv





SEGMENTED UNIVARIATE ANALYSIS

Performing Segmented Univariate Analysis on application_data.csv

				Descriptive Statistics					
	Range	Lower Limit	Upper Limit	Mean	Median	Mode	Max	Min	Standard Deviation
AMT_CREDIT	45000-115000	45000	115000	86805.09641	90000	101880	114682.5	45000	20558.0743
	115000-185000	115000	185000	157868.9876	157500	180000	184500	115128	20487.47843
	185000-255000	185000	255000	228875.1885	225000	225000	254799	185314.5	19100.9912
	255000-325000	255000	325000	285380.596	284031	270000	324216	255429	18540.94747
	325000-395000	325000	395000	355203.3283	354519	360000	394303.5	325377	19584.35721
	395000-465000	395000	465000	438912.5202	450000	450000	463941	395640	18722.34748
	465000-535000	465000	535000	503985.6369	508495.5	521280	534672	466213.5	18077.59364

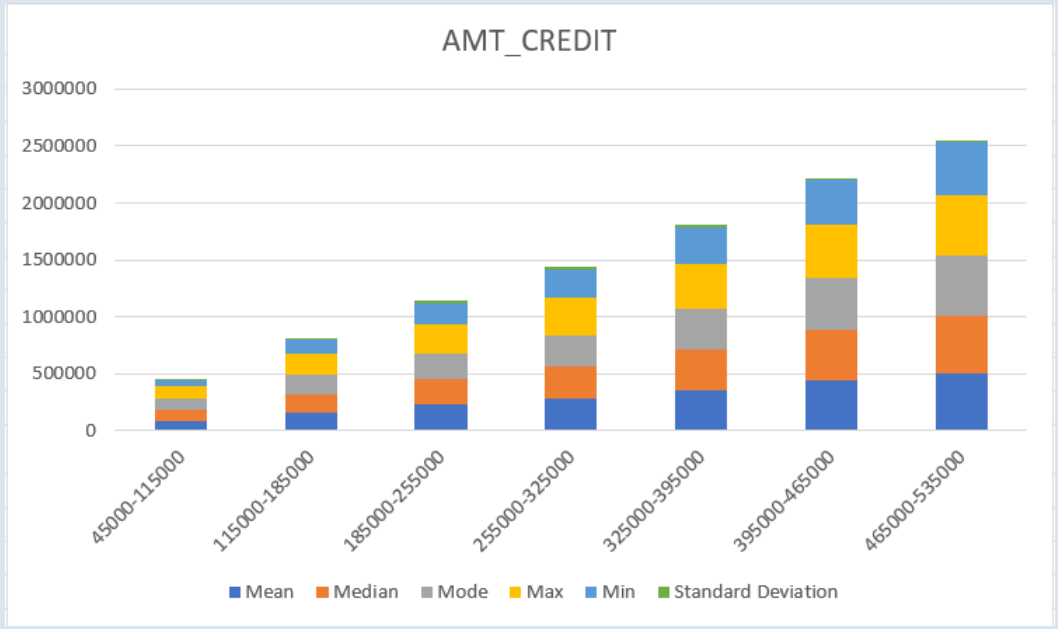
Performing Segmented Univariate Analysis on previous_application.csv

				Descriptive Statistics					
	Range	Lower Limit	Upper Limit	Mean	Median	Mode	Max	Min	Standard Deviation
AMT_APPLICATION	0-27000	0	27000	21342.26408	0	0	27000	0	8891.048344
	27000-54000	27000	54000	40774.36853	42363	45000	54000	27000	7322.29008
	54000-81000	54000	81000	67274.78224	67500	67500	81000	54000	7519.493604
	81000-108000	81000	108000	93283.03132	90000	90000	108000	81000	7148.047743
	108000-135000	108000	135000	124243.5271	124787.925	135000	135000	108000	9460.18311
	135000-162000	135000	162000	149000.5166	139770	135000	162000	135000	9181.600447
	162000-189000	162000	189000	177371.3604	180000	180000	189000	162000	6654.232382

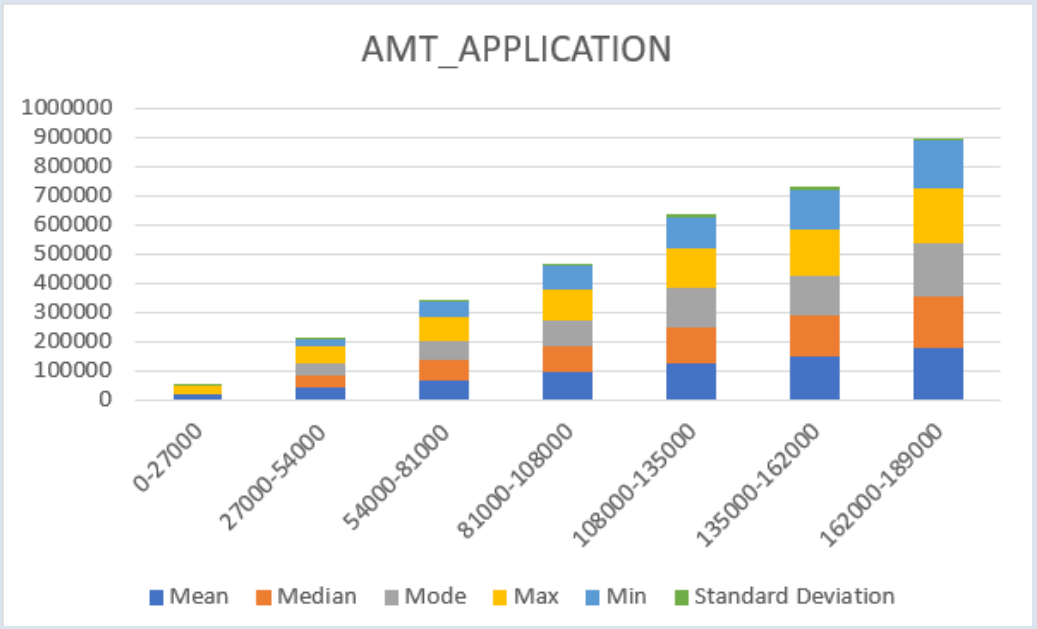
Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Data Visualization for

application_data.csv



previous_application.csv



BIVARIATE ANALYSIS

Performing Bivariate Analysis on application_data.csv

Row Labels	Sum of TARGET	Sum of AMT_INCOME_TOTAL
Cash loans	**3791**	**7636970885**
F	2118	4624132941
M	1673	3012837944
Revolving loans	**234**	**783873373.5**
F	145	467046171
M	89	316827202.5
Grand Total	**4025**	**8420844258**

Row Labels	Sum of TARGET	Sum of AMT_INCOME_TOTAL	Sum of AMT_CREDIT	Sum of AMT_ANNUITY	Sum of AMT_GOODS_PRICE
SUNDAY	**215**	**436869045**	**1518417734**	**69021985.5**	**1376845263**
Cash loans	203	387974295	1422815234	64199110.5	1280365263
Revolving loan	12	48894750	95602500	4822875	96480000
MONDAY	**657**	**1405089416**	**5007227162**	**225352944**	**4482567351**
Cash loans	620	1274935390	4758827162	212932944	4234594851
Revolving loan	37	130154026.5	248400000	12420000	247972500
TUESDAY	**716**	**1463819370**	**5220734027**	**237088341**	**4694774121**
Cash loans	684	1343617652	4989636527	225330966	4462731621
Revolving loan	32	120201718.5	231097500	11757375	232042500
WEDNESDAY	**669**	**1396582918**	**5073303303**	**228845349**	**4563625748**
Cash loans	623	1266795547	4816623303	215898849	4304335748
Revolving loan	46	129787371	256680000	12946500	259290000
THURSDAY	**678**	**1387289277**	**4947565028**	**224026456.5**	**4444137797**
Cash loans	630	1266119375	4712980028	212146456.5	4207775297
Revolving loan	48	121169902.5	234585000	11880000	236362500
FRIDAY	**671**	**1406276773**	**4959807606**	**223360677**	**4463306429**
Cash loans	639	1265199568	4699077606	210324177	4201653929
Revolving loan	32	141077205	260730000	13036500	261652500
SATURDAY	**419**	**925281958.5**	**3256812027**	**147617644.5**	**2923367256**
Cash loans	392	832329058.5	3075124527	138281269.5	2740712256
Revolving loan	27	92952900	181687500	9336375	182655000
Grand Total	**4025**	**8421208758**	**29983866885**	**1355313398**	**26948623964**

HEAT MAP

BIVARIATE ANALYSIS

Performing Bivariate Analysis on previous_application.csv

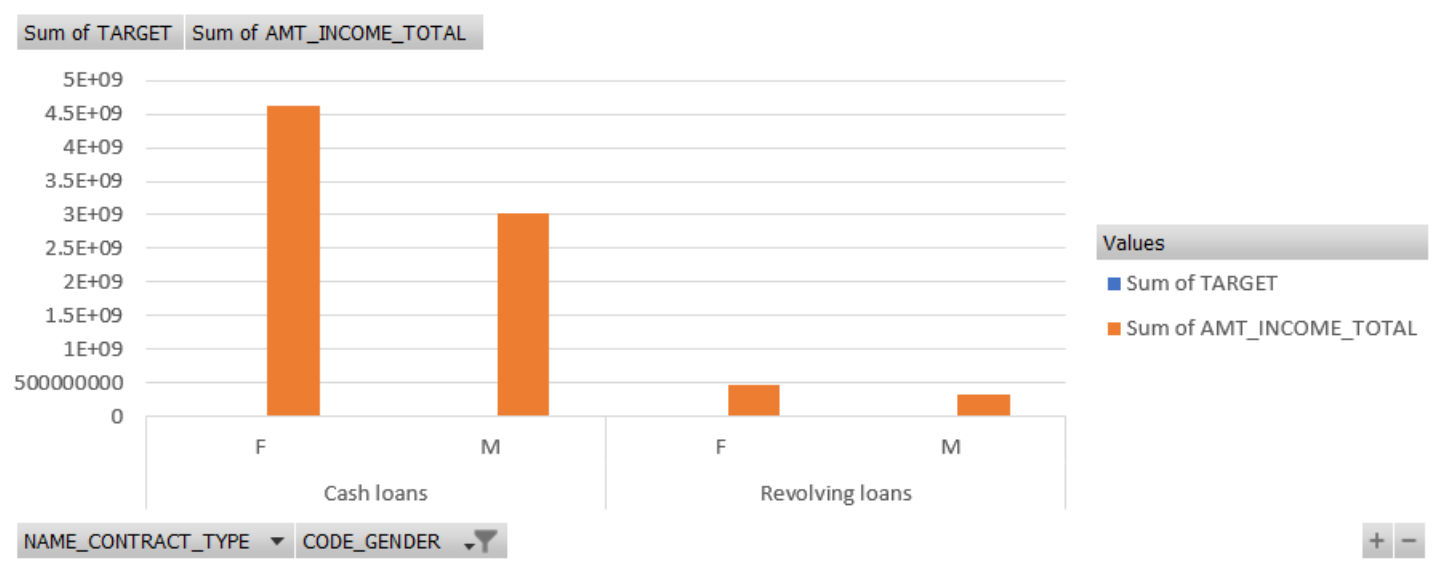
Row Labels	Sum of AMT_ANNUITY	Sum of HOUR_APPR_PROCESS_START	Sum of AMT_APPLICATION	Sum of AMT_CREDIT	Sum of AMT_GOODS_PRICE
Cash loans	**423134726.3**	**248155**	**5676764422**	**6302342470**	**6495382147**
SUNDAY	21173750.69	12658	273460500	302666386.5	322140690
MONDAY	71332304.63	41474	944151174.9	1049916956	1085198905
TUESDAY	71976471.08	41937	993598281.2	1102107542	1124660331
WEDNESDAY	71708973.84	43071	953804714.5	1060698254	1098701092
THURSDAY	68625560.3	40711	930573030.7	1035795280	1060074818
FRIDAY	67672390.1	39577	913836550.7	1013279607	1040009778
SATURDAY	50645275.7	28727	667340169.8	737878443.8	764596532.3
Consumer loans	**239206019.1**	**306778**	**2209887762**	**2161077761**	**2214360515**
SUNDAY	37147425.39	45801	341996032.4	336277127.1	342204067.4
MONDAY	31086549.14	40963	288463879.4	282110982.4	289504054.4
TUESDAY	31275484.74	40545	289250394.1	282720416.8	289978516.6
WEDNESDAY	32457547.26	41556	303579348.5	295389308.5	303995418.5
THURSDAY	31480683.3	41082	293797879.9	285755083.9	294109932.4
FRIDAY	32727498.84	43678	301035820.5	294813404.3	302180013
SATURDAY	43030830.47	53153	391764407.6	384011437.8	392388512.6
Revolving loans	**62935021.8**	**68876**	**557801654**	**963535500**	**852365556.5**
SUNDAY	3897499.32	4614	33387255.9	62977500	51626835.9
MONDAY	10532501.28	11129	96277500	161212500	144530550
TUESDAY	10486176.12	11242	95666395.5	161239500	143750875.5
WEDNESDAY	10226069.28	11150	86442286.5	150651000	135330511.5
THURSDAY	10358591.04	11368	85246562.25	157050000	135517997.3
FRIDAY	9749193.12	10774	92675921.4	151389000	137729516.4
SATURDAY	7684991.64	8599	68105732.4	119016000	103879269.9
XNA	**87039.36**	**95**	**0**	**0**	**832140**
SUNDAY	10879.92	13	0	0	104017.5
MONDAY	10879.92	13	0	0	104017.5
TUESDAY	21759.84	27	0	0	208035
WEDNESDAY	21759.84	28	0	0	208035
SATURDAY	21759.84	14	0	0	208035
Grand Total	**725362806.6**	**623904**	**8444453838**	**9426955730**	**9562940358**

HEAT MAP

Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Data Visualization for

application_data.csv



previous_application.csv

Due to the absence of the target variable the pivot table was not made. The only visualization for the dependability between the variables was shown through the heatmap in the previous slide.

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.

Top 10 Correlated Variables of application_data.csv

Correlation	Correlation Coefficient
DAYS_BIRTH	0.076744119
REGION_RATING_CLIENT_W_CITY	0.067091093
REGION_RATING_CLIENT	0.066144582
DAYS_LAST_PHONE_CHANGE	0.05606376
REG_CITY_NOT_WORK_CITY	0.048493636
DAYS_ID_PUBLISH	0.046961132
FLAG_DOCUMENT_3	0.045012592
DEF_60_CNT_SOCIAL_CIRCLE	0.044419006
DAYS_REGISTRATION	0.042381842
DEF_30_CNT_SOCIAL_CIRCLE	0.041794766

Correlated Variables of previous_application.csv

Due to the absence of the target variable, there was no means to find out the correlation between the variables with the target variable.

Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.Data Visualization for

application_data.csv

	DAYS_BIRTH	REGION_RATING_CLIENT_W_	REGION_RATING_CLIENT	DAYS_LAST_PHONE_CHANGE	REG_CITY_NOT_WORK_CITY
DAYS_BIRTH	1	0.014552576	0.016780889	0.080179098	0.237907474
REGION_RATING_CLIENT_W_CITY	0.014552576	1	0.950710189	0.02679039	0.030502884
REGION_RATING_CLIENT	0.016780889	0.950710189	1	0.027329455	0.010193291
DAYS_LAST_PHONE_CHANGE	0.080179098	0.02679039	0.027329455	1	0.046876914
REG_CITY_NOT_WORK_CITY	0.237907474	0.030502884	0.010193291	0.046876914	1

previous_application.csv

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE
AMT_ANNUITY	1	0.810049781	0.975771049	0.97235717
AMT_APPLICATION	0.810049781	1	0.975771049	0.988711762
AMT_CREDIT	0.975771049	0.975771049	1	0.97235717
AMT_GOODS_PRICE	0.97235717	0.988711762	0.97235717	1

Tech-Stack Used

- Microsoft Excel: It is a spreadsheet program from Microsoft and a component of its Office product for business application. This enables users to format, calculate and organize data in a spreadsheet.
- MS Excel Functions: They are predefined formulas that perform calculations by using specific values, called arguments, in a particular order or structure. Some of the functions are:

1. Text functions: clean(), substitute(), replace(), concatenate(), trim(), etc.
2. Mathematical and Statistical functions: sum(), sumif(), count(), max(), average(), median(), mode(), stdev(), etc.

- Data Visualization in Excel: Bar, Column, Scatter, Heatmap, Stacked Chart, etc.

Insights

- We were able to identify the missing data and performed descriptive statistics with it. This is necessary to ensure the accuracy of the analysis.
- Identifying the Outliers in the dataset. They distort the results and can bring a significant impact to the analysis.
- Analyzing the data imbalance. To check for biases in the dataset this is a necessary step to be undertaken.
- Lastly performing the Univariate, Segmented Univariate, and Bivariate Analysis to gain factors driving for loan default.

Results

- Remembering to adapt excel functions on specific dataset.
- These learned insights helped me understand specific business questions which were addressed by MS Excel
- Learning about Excel Text and Statistical functions. The importance of average(), median(),mode(), text() functions.
- We were able to build different charts for visualization for answering the business questions. Some of the charts used were bar graph, stacked Chart and heatmap.
- Achieving the ability to learn and write MS Excel functions to execute different business questions.
- Solving Company related problems using different visualization charts offered by Excel