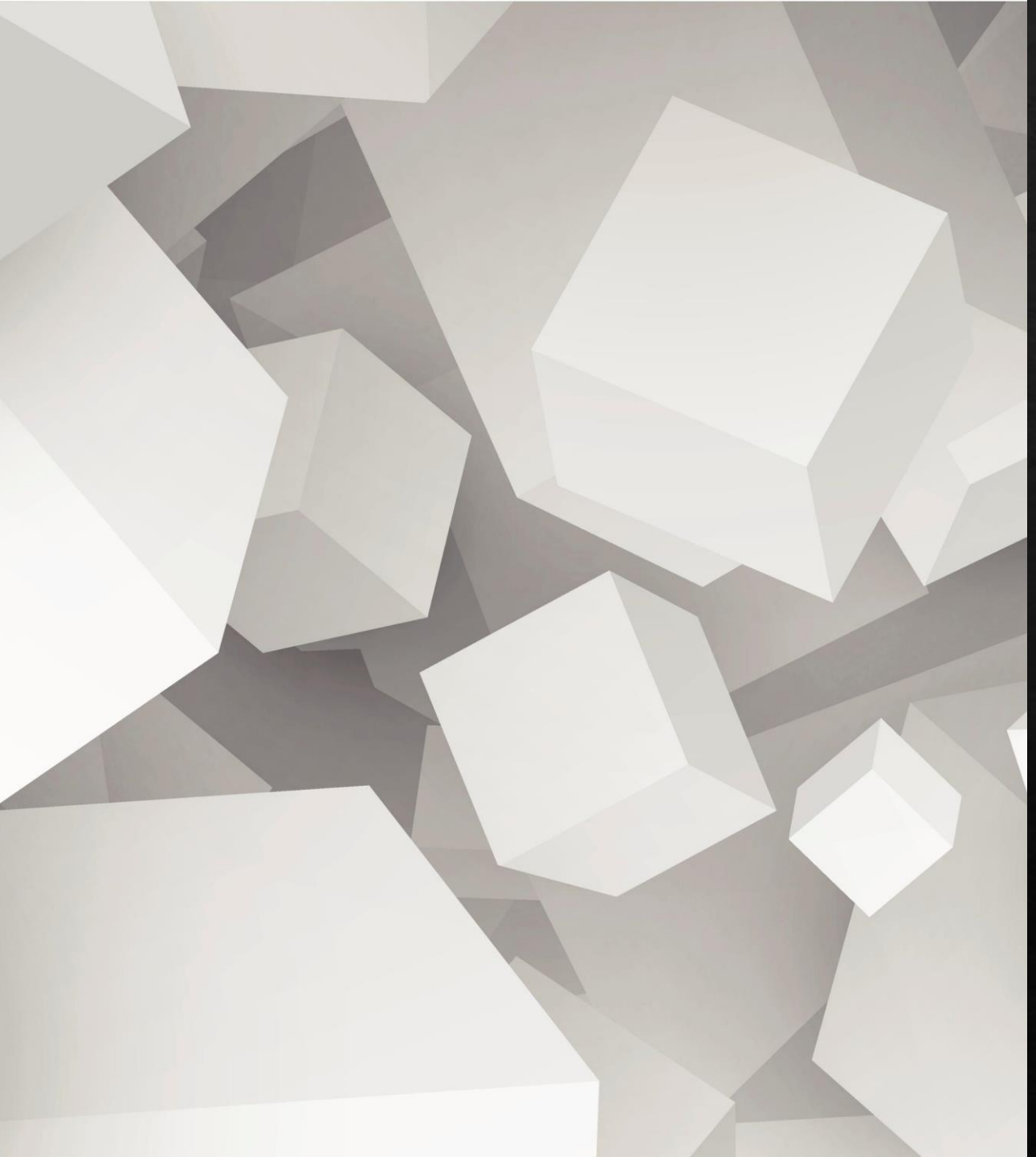


Bank Loan Case Study

- Subrat Sindhu
- Dataset link
- https://docs.google.com/spreadsheets/d/1svTz30f2jRmB99UQXBuBPPkBYWkQLv17/edit?usp=drive_link&ouid=104966616771126772005&rtpof=true&sd=true
- https://docs.google.com/spreadsheets/d/1aS5WynZgoHc4SCYszH6jnmWDppsBAu2s/edit?usp=drive_link&ouid=104966616771126772005&rtpof=true&sd=true
- Video Link
- https://drive.google.com/file/d/1cNOJgeqtexZ9Ufk2zhNKdxSNe4tb-KpQ/view?usp=drive_link

Project Description

- My main duty as a data analyst at a financial company that specializes in providing different kinds of loans to urban clients is to use exploratory data analysis (EDA) to find patterns in the data and make sure that qualified applicants are not turned down.
- Among the difficulties the business faces are: some clients who don't have a long enough credit history take advantage of this and fail to make loan payments.

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: The preprocessing stage that prepares the data for analysis is this one. It involves addressing missing values and getting rid of duplicates.
- Under this project I have removed null values and outliers from the datasets and replaced them with appropriate values.
- Data Visualization: To use EDA to understand how customer attributes and loan attributes influence the likelihood of default.
- For data visualization I have used stacked bar chart, histograms, pie charts and box plot, column charts and heat maps to determine the correlation and visualization of data.

Data Analysis 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

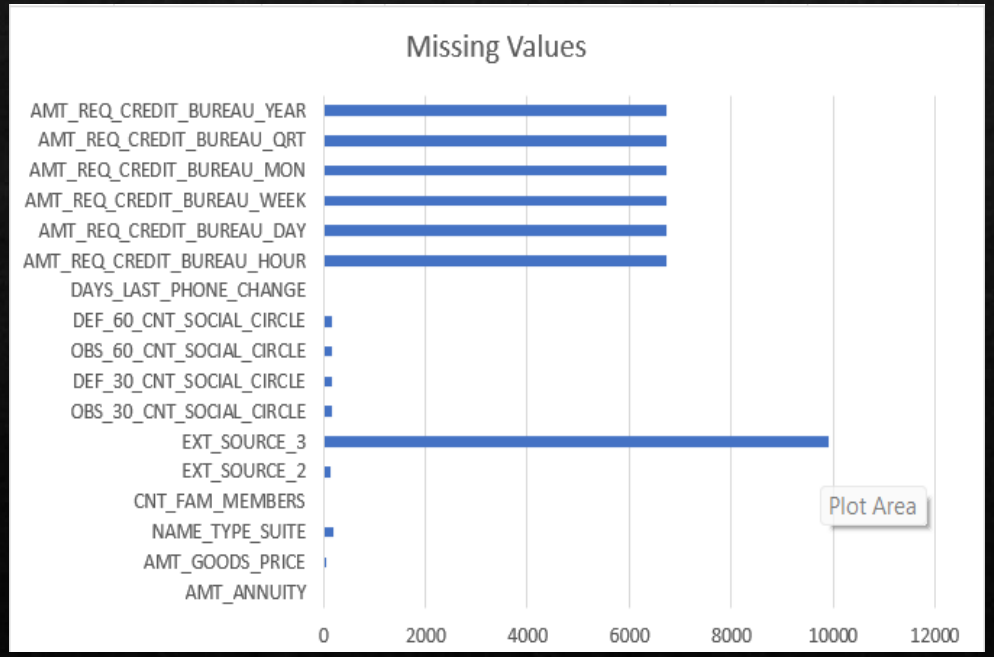
Removing the colums in which blank% is more than 25%																	
	AMT_ANN	AMT_GOODS_	NAME_TYPI	CNT_FAM_MI	EXT_SOURCE_	EXT_SOURCE	OBS_30_CNT_	DEF_30_CNT_	OBS_60_C	DEF_60_C	DAYS_LAS	AMT_REQ_CRE	AMT_REQ_CREDIT	AMT_REQ_CREDIT_	AMT_REQ_CR	AMT_REQ_CREC	AMT_REQ_CRI
Blank Cou	1	38	192	1	126	9915	168	168	168	168	1	6719	6719	6719	6719	6719	6719
Blank%	0.002%	0.076%	0.384%	0.002%	0.252%	20%	0.336%	0.336%	0.336%	0.336%	0.002%	13%	13%	13%	13%	13%	13%
Mean	27112.5	538508.243		2.15919449	0.51367294	0.5119197	1.42114064	0.14184411	1.40402	0.09829	-965.97	0.007078537	0.007519495	0.032468437	0.27060899	0.261139993	1.88569903
Median	24939	450000		2	0.56546716	0.5352763	0	0	0	0	-757	0	0	0	0	0	1
Mode			Unaccompanied														

- Missing Data for the dataset previous_application.csv

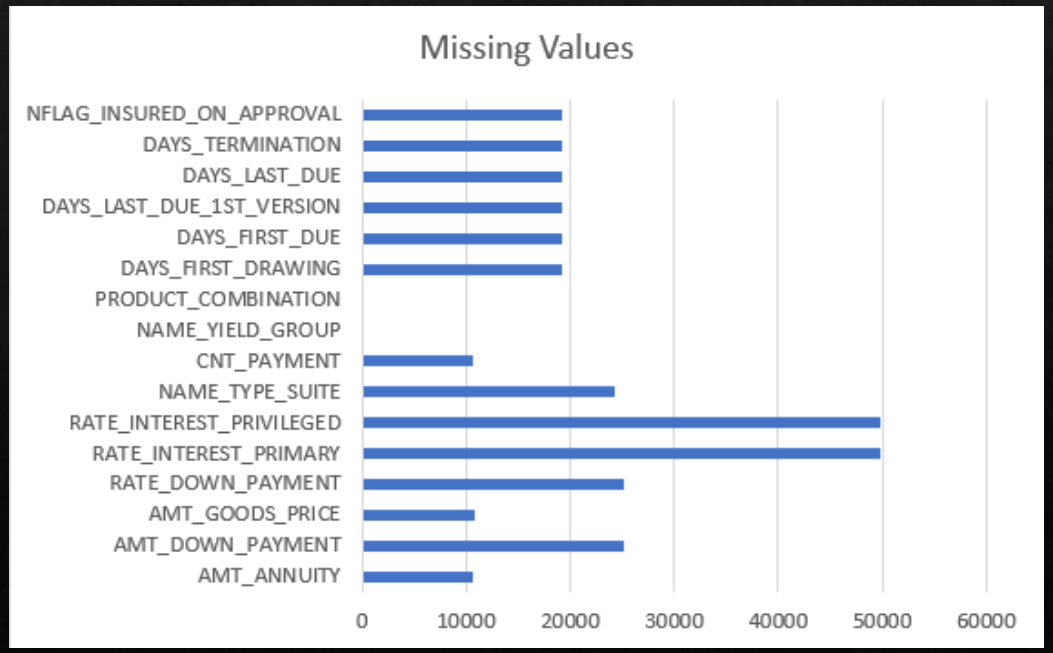
Removing the columns containing blank% more than 25%				
	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_PAYMENT	PRODUCT_COMBINATION
Blank Coun	10591	10743	10591	7
Blank%	21%	21%	21%	0%
Mean	15482.59685	215141.4173	15.55589109	#DIV/0!
Median	10879.92	104017.5	12	#NUM!
Mode				POS household without interest

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

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 table for application_data.csv dataset.

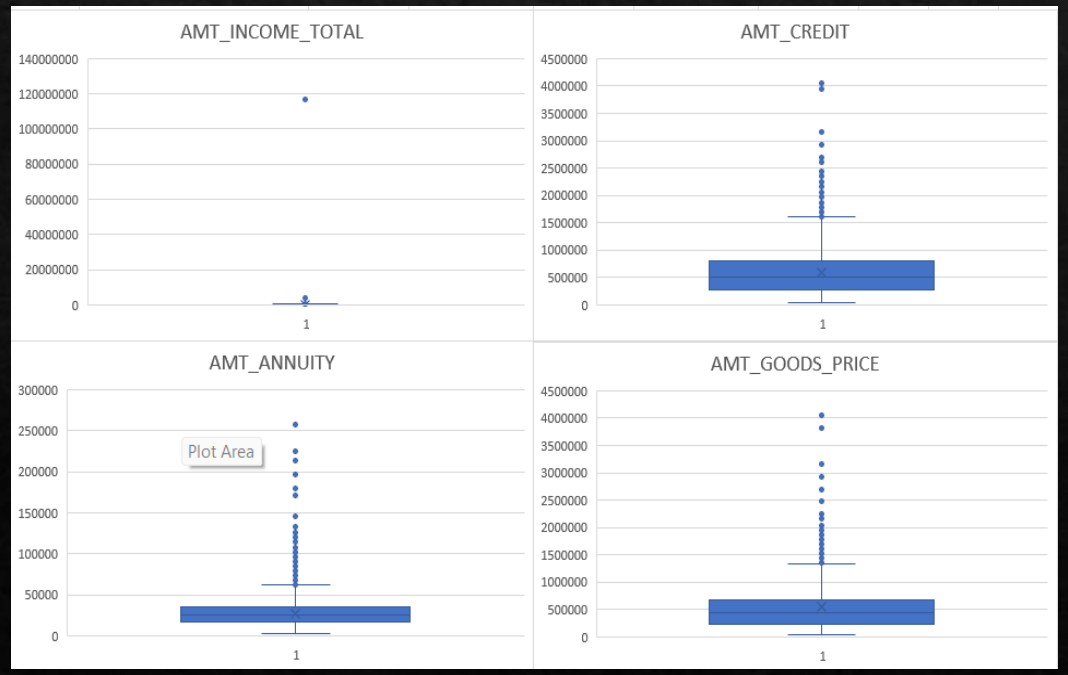
	Quartile Range						IQR	
	Minimum	Q1	Median	Q3	Maximum	IQR	Lower	Upper
AMT_INCOME_TOTAL	25650	112500	145800	202500	3825000	90000	-22500	337500
AMT_CREDIT	45000	270000	514777.5	808650	4050000	538650	-537975	1616625
AMT_ANNUITY	2052	16456.5	24939	34596	258025.5	18139.5	-10752.8	61805.25
AMT_GOODS_PRICE	45000	238500	450000	679500	4050000	441000	-423000	1341000
REGION_POPULATION_RELATIVE	0.000533	0.010006	0.01885	0.028663	0.072508	0.018657	-0.01798	0.056649
DAYS_EMPLOYED	-17531	-2786	-1221	-292	365243	2494	-6527	3449
DAYS_REGISTRATION	-22392	-7463.75	-4490	-1998	0	5465.75	-15662.4	6200.625
EXT_SOURCE_3	0.000527265	0.417099668	0.53527625	0.638043528	0.896009549	0.220944	0.085684	0.969459

- Outlier check table for application_data.csv dataset.

	Quartile						IQR	
	Minimum	Q1	Median	Q3	Maximum	IQR	Lower	Upper
AMT_ANNUITY	0	7189.74	10879.92	16256.16	234478.4	9066.42	-6409.89	29855.79
AMT_APPLICATION	0	22045.5	71550	180000	3826373	157954.5	-214886	416931.8
AMT_CREDIT	0	26055	78907.5	198105.8	4104351	172050.8	-232021	456181.9
AMT_GOODS_PRICE	0	63663.75	104017.5	180000	3826373	116336.3	-110841	354504.4

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

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.

Imbalance check table for application_data.csv dataset.

	0	1	SUM
TARGET	45973	4025	49998
	Cash Loans	Revolving loans	SUM
NAME_CONTRACT_TYPE	45275	4723	49998
	M	F	SUM
CODE_GENDER	17174	32822	49996

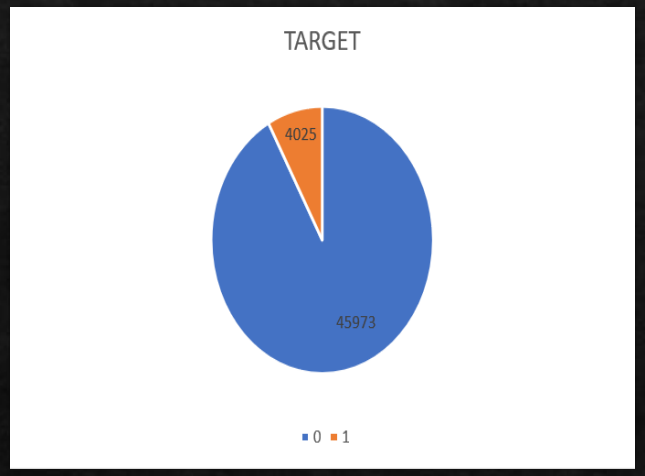
	Y	N	SUM
FLAG_OWN_CAR	17050	32948	49998
FLAG_OWN_REALTY	34690	15308	49998

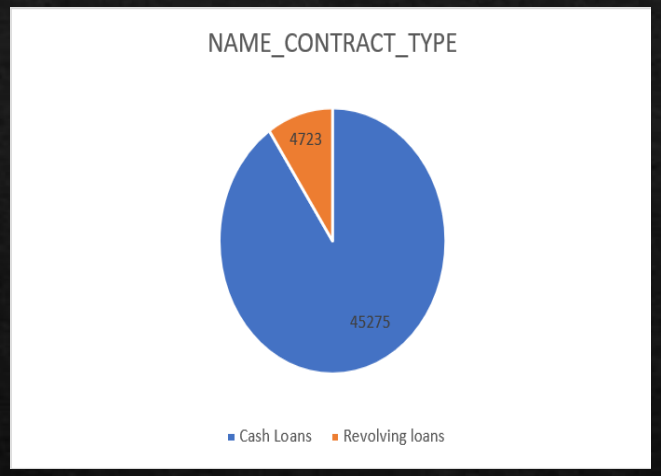
	0	1	SUM
FLAG_MOBIL	1	49997	49998
FLAG_EMP_PHONE	8926	41072	49998
FLAG_WORK_PHONE	40035	9963	49998
FLAG_CONT_MOBILE	101	49897	49998
FLAG_PHONE	36112	13886	49998
FLAG_EMAIL	47215	2783	49998
REG_REGION_NOT_LIVE_REGION	49248	750	49998
REG_REGION_NOT_WORK_REGION	47502	2496	49998
LIVE_REGION_NOT_WORK_REGION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

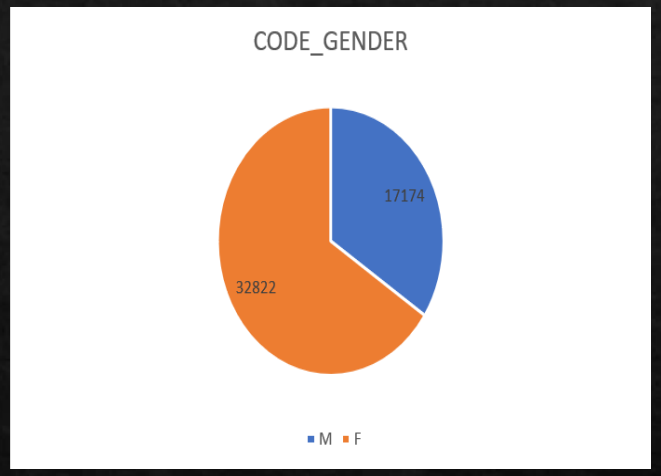
- Imbalance check table for previous_application.csv dataset.

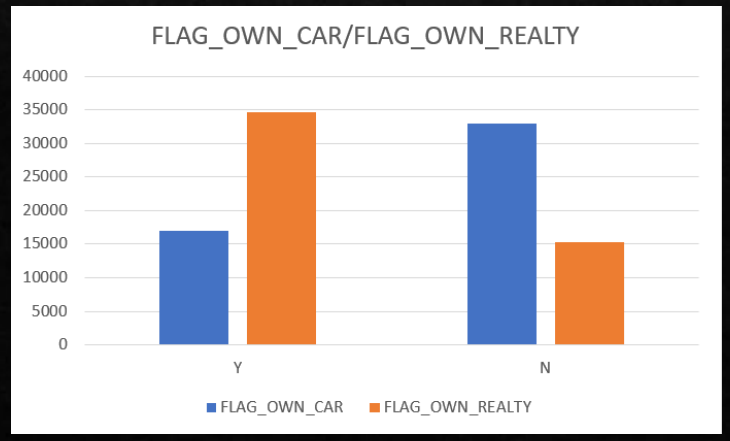
	Cash Loans	Consumer Loans	Revolving Loans	XNA	SUM
NAME_CONTRACT_TYPE	20856	23510	5625	8	49999
	Approved	Canceled	Refused	Unused offer	SUM
NAME_CONTRACT_STATUS	31885	8595	8660	859	49999
	Y	N	SUM		
FLAG_LAST_APPL_PER_CONTRACT	49747	252	49999		
	1	10	SUM		
NFLAG_LAST_APPL_IN_DAY	49824	175	49999		

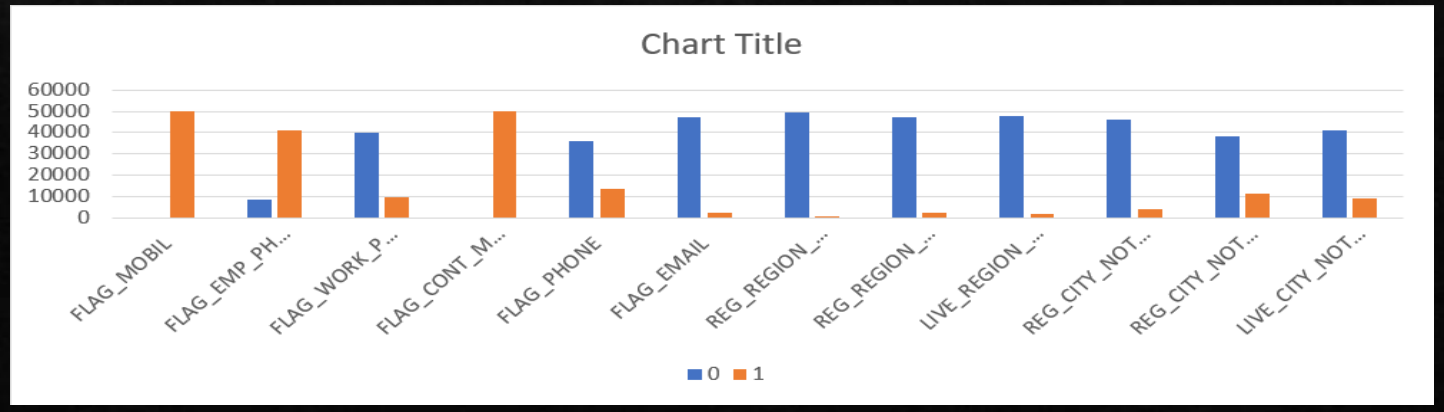
Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.



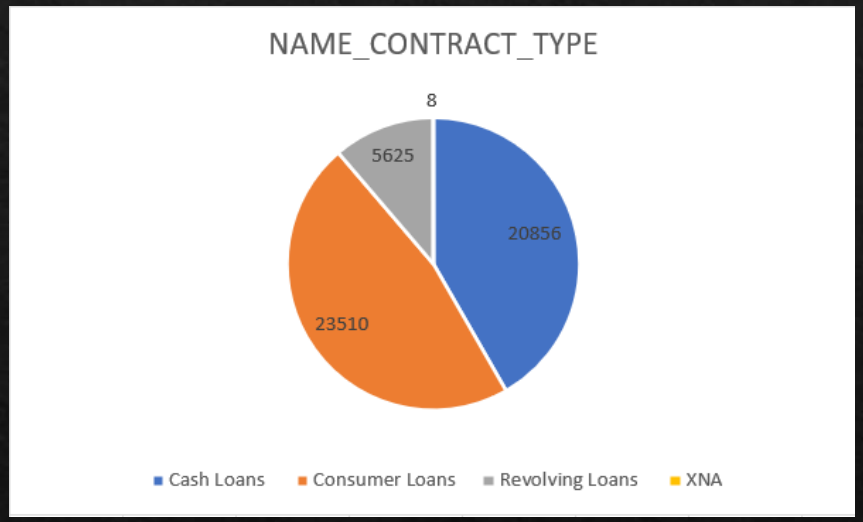


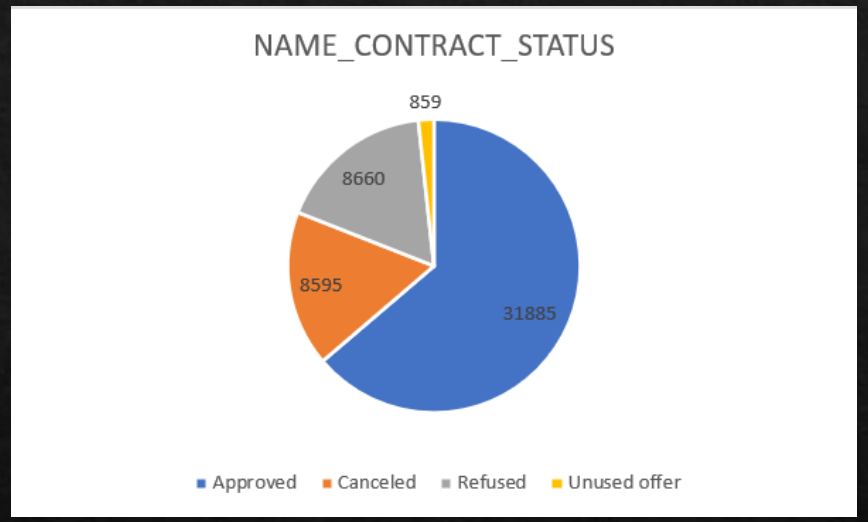


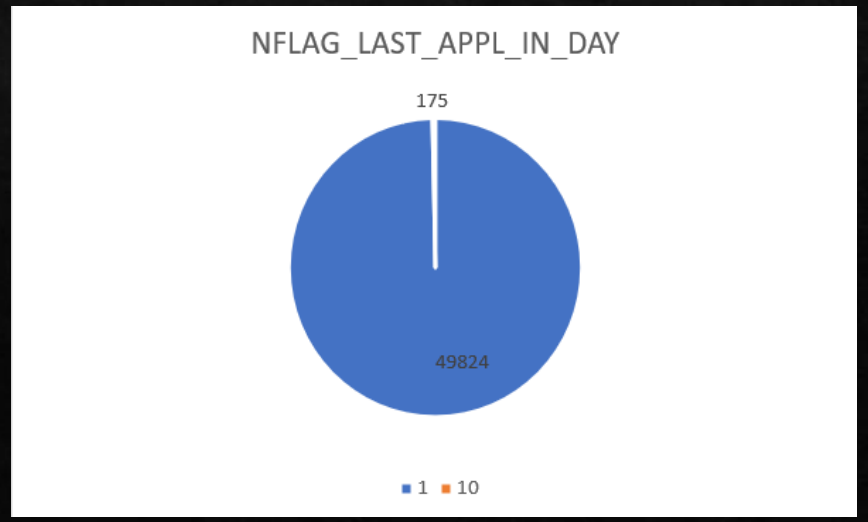




- previous_application.csv







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.

- Univariate Analysis on application_data.csv

Univariate Analysis						
	Mean	Median	Mode	Max	Min	Std. Dev
AMT_INCOME_TOTAL	168430.91	145800	135000	3825000	25650	99166.41
AMT_CREDIT	599701.33	514777.5	450000	4050000	45000	402419.4
AMT_ANNUITY	27107.35	24939	9000	258025.5	2052	14562.95
AMT_GOODS_PRICE	538994.04	450000	450000	4050000	45000	369724.3

NAME_CONTRACT_TYPE	Frequency	Cumulative Frequency	Precentage
Cash Loan	45275	45275	91%
Revolving Loan	4723	49998	9%

NAME_TYPE_SUITE	Frequency	Cumulative Frequency	Precentage
Children	542	542	1%
Family	6549	7091	13%
Group of people	36	7127	0%
Other_A	137	7264	0%
Other_B	259	7523	1%
Spouse,partner	1849	9372	4%
Unacompanied	40626	49998	81%

NAME_FAMILY_STATUS	Frequency	Cumulative Frequency	Precentage
Civil marriage	4859	4859	9.718%
Married	32093	36952	64.189%
Separate	3142	40094	6.284%
Single / not married	7306	47400	14.613%
Unknown	1	47401	0.002%
Widow	2597	49998	5.194%

- Univariate Analysis on previous_application.csv

Univariate Analysis

	Mean	Median	Mode	Max	Min	Std. Dev
AMT_ANNUITY	14507.55	10879.92	10879.92	234478.395	0	13036.66787
AMT_APPLICATION	168892.45	71550	0	3826372.5	0	282203.5105
AMT_CREDIT	188542.89	78907.5	0	4104351	0	308473.6014
AMT_GOODS_PRICE	191262.63	104017.5	104017.5	3826372.5	0	271892.6356

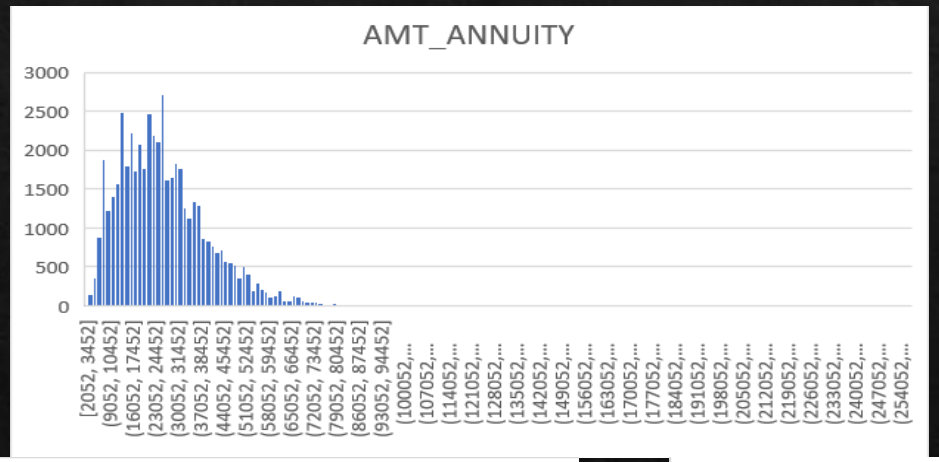
NAME_CONTRACT_TYPE	Frequency	Cumulative Frequency	Precentage
Cash loans	20856	20856	41.713%
Consumer loans	23510	44366	47.021%
Revolving loans	5625	49991	11.250%
XNA	8	49999	0.016%

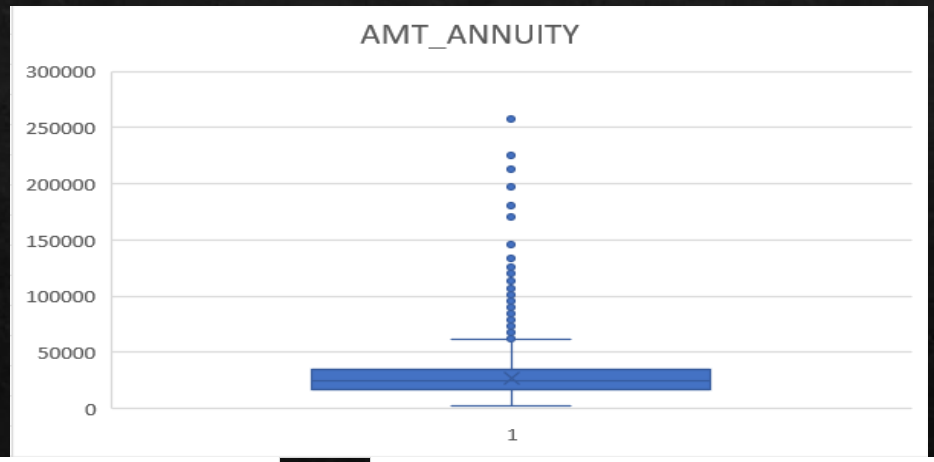
WEEKDAY_APPR_PROCESS_START	Frequency	Cumulative Frequency	Precentage
MONDAY	7419	7419	14.84%
TUESDAY	7504	14923	15.01%
WEDNESDAY	7649	22572	15.30%
THURSDAY	7460	30032	14.92%
FRIDAY	7554	37586	15.11%
SATURDAY	7380	44966	14.76%
SUNDAY	5033	49999	10.07%

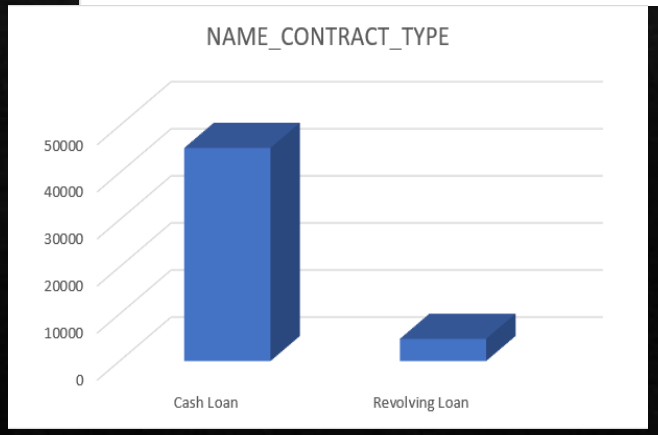
NAME_CONTRACT_STATUS	Frequency	Cumulative Frequency	Precentage
Approved	31885	31885	63.77%
Candeled	8595	40480	17.19%
Refused	8660	49140	17.32%
Unused offer	859	49999	1.72%

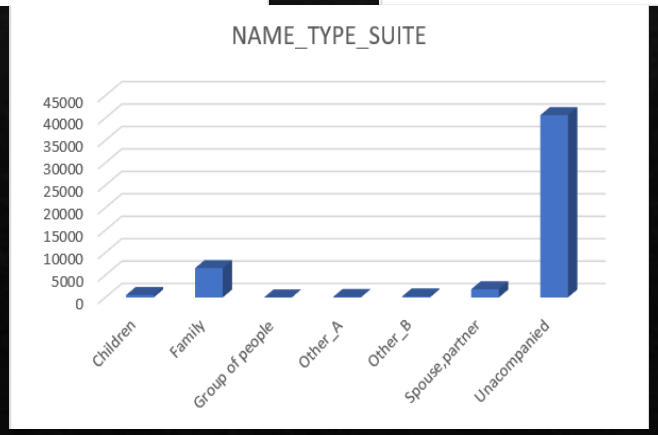
Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios.

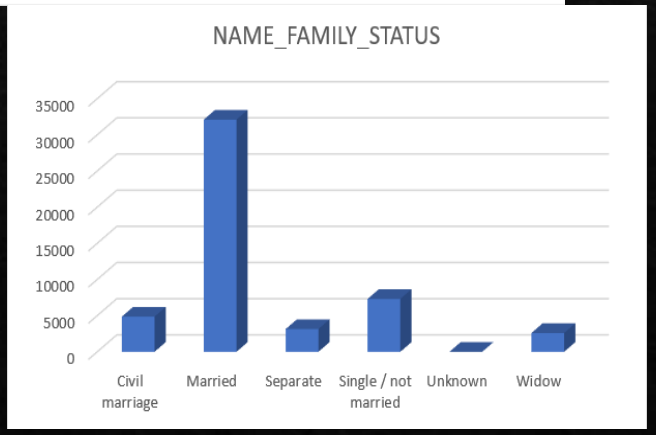
- Data Visualization for application_data.csv



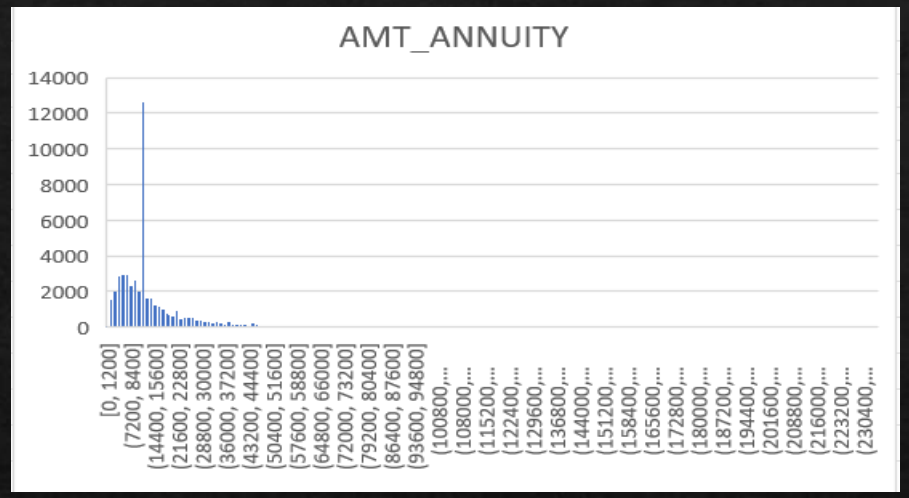


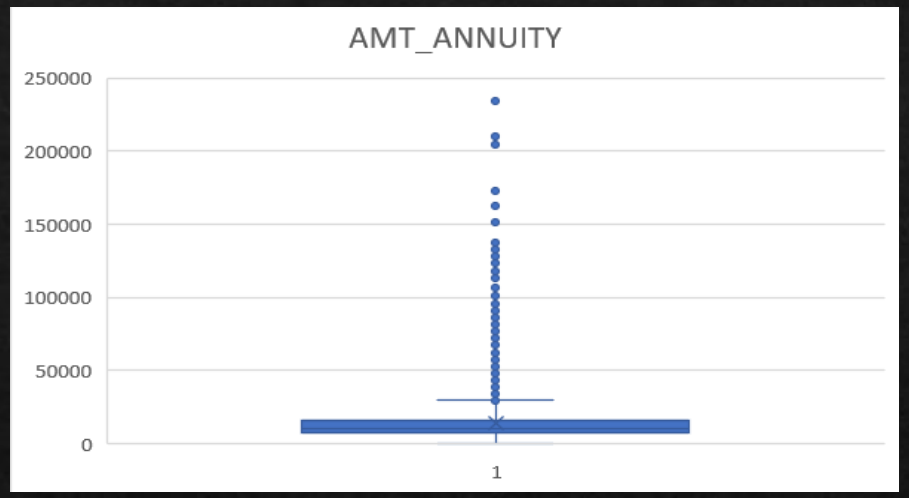


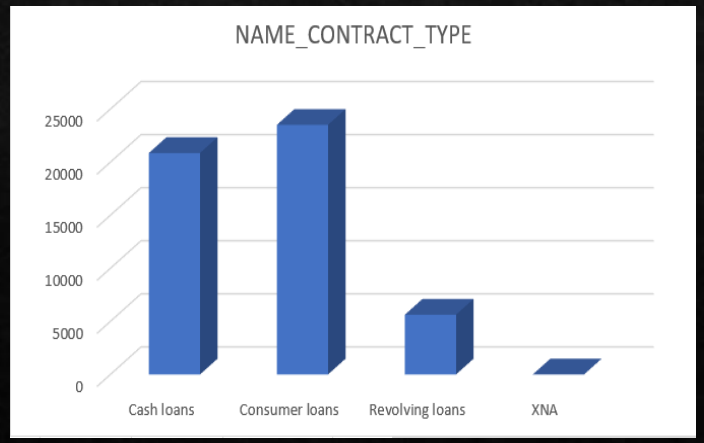


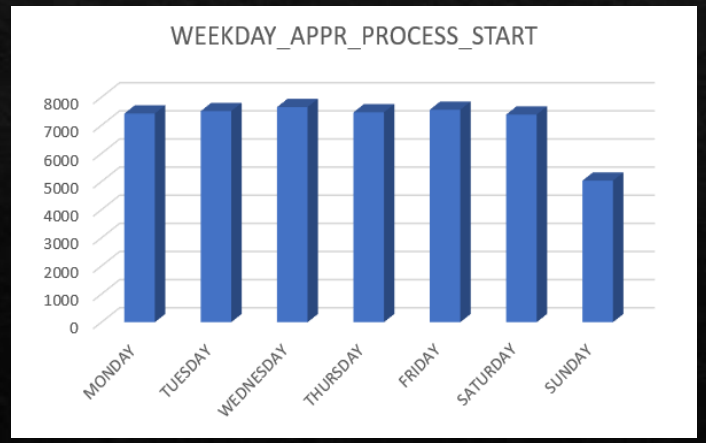


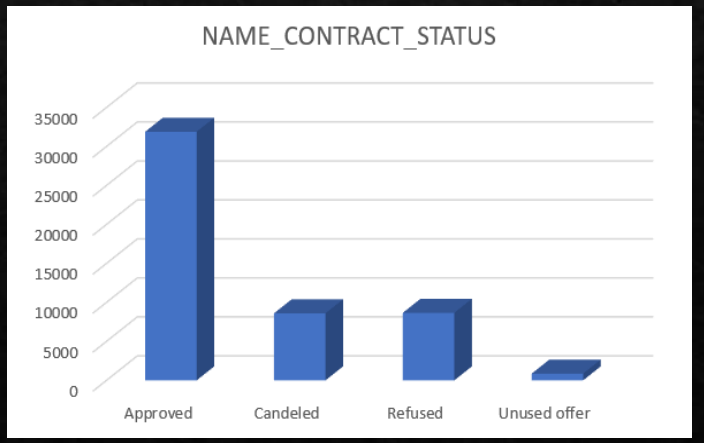
Data Visualization for previous_application.csv











Segmented Analysis

- Segmented Analysis on application_data.csv

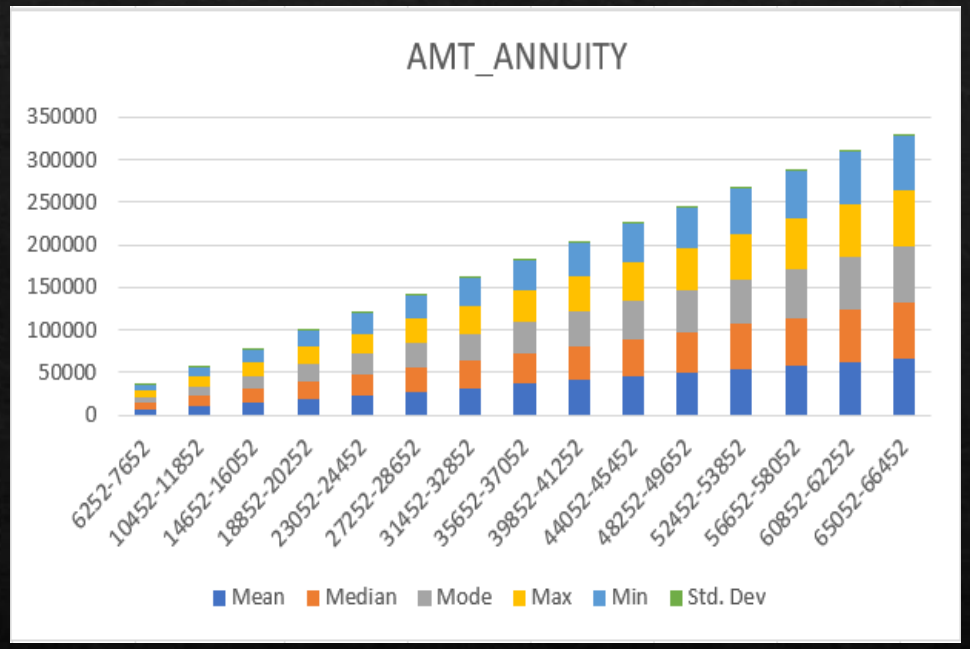
Segmented Univariate Analysis									
AMT_ANNUITY	Range	Lower	Upper	Mean	Median	Mode	Max	Min	Std. Dev
	6252-7652	6252	7652	6887.3237	6750	6750	7645.5	6259.5	311.862
	10452-11852	10452	11852	11165.146	11250	11250	11848.5	10453.5	382.3971
	14652-16052	14652	16052	15395.244	15383.25	15750	16051.5	14652	444.9684
	18852-20252	18852	20252	19630.171	19696.5	20250	20250	18855	439.5438
	23052-24452	23052	24452	23780.658	23773.5	23773.5	24448.5	23053.5	387.2439
	27252-28652	27252	28652	27942.114	27877.5	28408.5	28651.5	27256.5	421.7275
	31452-32852	31452	32852	32057.918	32053.5	31653	32845.5	31455	423.8527
	35652-37052	35652	37052	36315.907	36333	36459	37048.5	35653.5	376.506
	39852-41252	39852	41252	40419.939	40320	40320	41224.5	39852	353.8249
	44052-45452	44052	45452	44853.519	44896.5	45000	45450	44055	377.6419
	48252-49652	48252	49652	48922.29	48838.5	48631.5	49639.5	48258	405.6621
	52452-53852	52452	53852	53240.92	53253	52452	53847	52452	420.4757
	56652-58052	56652	58052	57288.823	57197.25	57685.5	58050	56664	376.8762
	60852-62252	60852	62252	61666.465	61875	61875	62212.5	60867	383.5999
	65052-66452	65052	66452	65775.357	65866.5	65866.5	66402	65052	436.0876

- Segmented Analysis on previous_application.csv

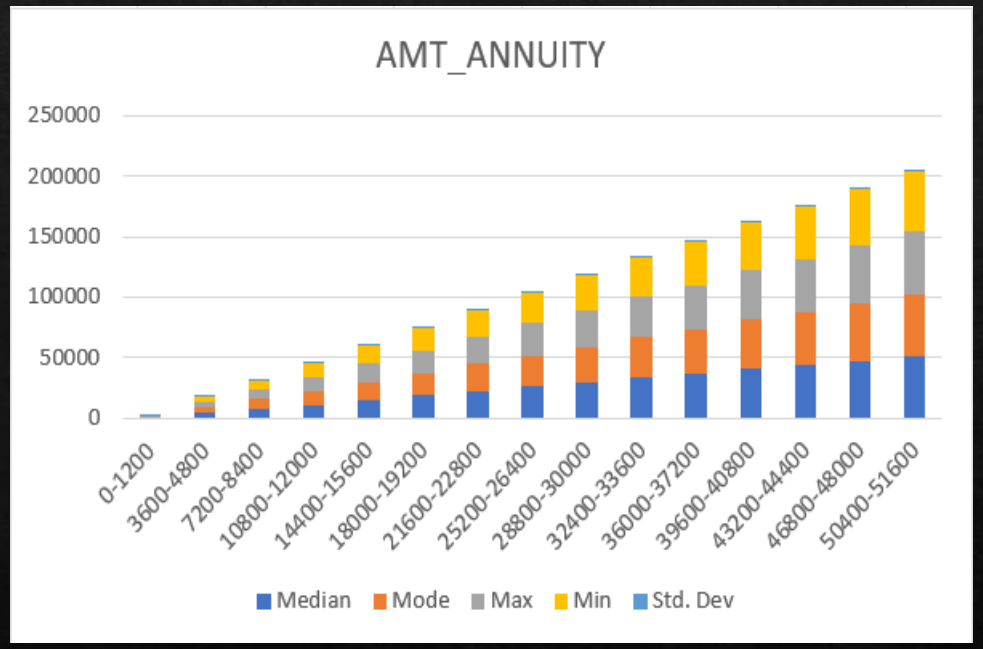
Segmented Univariate Analysis

AMT_ANNUITY	Range	Lower	Upper	Mean	Median	Mode	Max	Min	Std. Dev
	0-1200	0	1200	973.7225	0	0	1190.295	0	425.4972227
	3600-4800	3600	4800	4246.030774	4300.3125	4500	4799.925	3600.315	330.1166462
	7200-8400	7200	8400	7796.388679	7833.6675	7875	8399.07	7200	336.9384402
	10800-12000	10800	12000	10955.94446	10879.92	10879.92	11999.61	10801.26	213.4799693
	14400-15600	14400	15600	14972.49225	14932.935	14625	15599.925	14400.045	340.9628611
	18000-19200	18000	19200	18619.95069	18549.36	18000	19197	18000	390.5412539
	21600-22800	21600	22800	22291.45694	22482	22500	22787.685	21600.27	320.7564248
	25200-26400	25200	26400	25803.6321	25798.5	25932.915	26394.66	25200	339.4300469
	28800-30000	28800	30000	29416.52116	29375.55	29250	29999.7	28810.26	328.6109336
	32400-33600	32400	33600	33000.73437	33011.46	33490.485	33594.975	32400.405	346.0355301
	36000-37200	36000	37200	36580.71621	36517.005	36000	37199.835	36000	367.255557
	39600-40800	39600	40800	40247.0649	40270.095	40783.995	40784.76	39604.5	338.8952746
	43200-44400	43200	44400	43798.34195	43790.22	43551	44396.37	43220.925	369.7663438
	46800-48000	46800	48000	47412.97461	47406.825	47041.335	47999.565	46850.265	334.6258644
	50400-51600	50400	51600	50995.93671	50999.895	51334.695	51571.17	50401.17	350.1102216

- Data Visualization for application_data.csv and previous_application.csv(Segmented Analysis)



application_data.csv



previous_application.csv

Bivariate analysis

- Bivariate analysis for application_data.csv

Row Labels	Sum of TARGET
Businessman	**0**
M	0
Commercial associate	**864**
F	460
M	404
Maternity leave	**0**
M	0
Pensioner	**501**
F	375
M	126
State servant	**198**
F	130
M	68
Student	**0**
F	0
M	0
Unemployed	**2**
F	1
M	1
Working	**2460**
F	1297
M	1163
Grand Total	**4025**

HEAT MAP

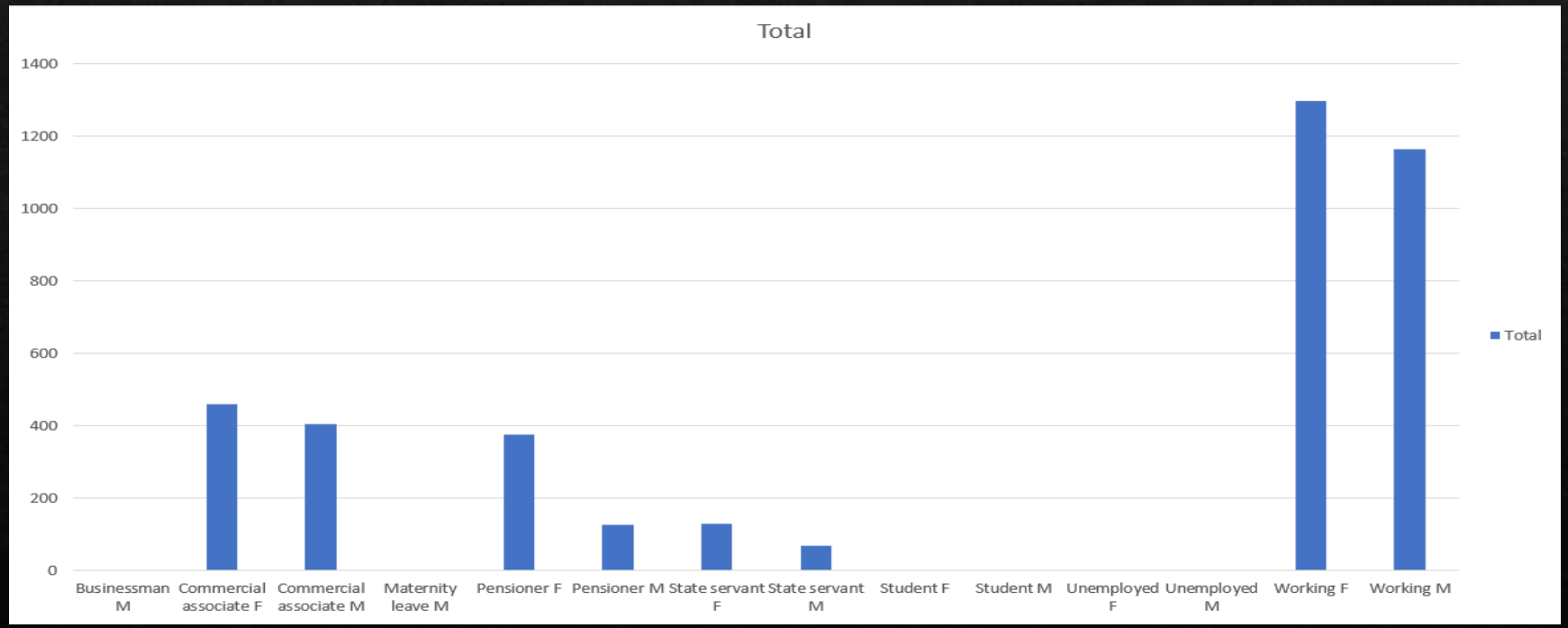
Sum of TARGET	Column Labe											
Row Labels	0	1	2	3	4	5	6	7	8	9	11	Grand Total
Civil marriage	329	109	36	5	3	0						482
Married	1393	629	310	48	10	3	0	0	0		1	2394
Separated	190	64	18	0	0							272
Single / not married	595	111	19	2	1					1		729
Unknown	0											0
Widow	137	9	1	1		0						148
Grand Total	**2644**	**922**	**384**	**56**	**14**	**3**	**0**	**0**	**0**	**1**	**1**	**4025**

- Bivariate analysis for previous_application.csv

Row Labels	Sum of AMT_ANNUITY	Sum of AMT_APPLICATION	Sum of AMT_CREDIT	Sum of AMT_GOODS_PRICE	Sum of HOUR_APPR_PROCESS_START
0	38768.13	488565	488565	592582.5	0
1	80628.48	1798555.5	1859710.5	1798555.5	4
2	513774.18	4825534.5	5319535.5	5969727	60
3	2484593.28	30069531	35054361	35686476	435
4	4430193.75	50726685.6	61004646	59880225.6	1088
5	7698808.98	91949580.32	103831255.8	106616047.8	2390
6	12666710.48	142978667.4	164177344.7	167006709.9	5022
7	21820476.96	265272460.9	298494553.1	304070988.4	9723
8	31563233.51	375469194.8	424854967.9	426333752.3	17152
9	54748753.16	647217917.1	733896236.7	748070252.1	34452
10	78914063.48	922631674.9	1043166832	1054629882	53150
11	85062345.26	1009943029	1135683737	1144021587	63657
12	80571167.82	949394011.8	1057102660	1068426549	66924
13	74287398.15	844992363.1	945764021.2	964573023.1	67613
14	67082125.37	780279706.3	858103934.7	878694278.8	65436
15	63040700.16	717251513.6	799398766.8	813675736.1	66345
16	52123857.89	600304785.8	662615249.6	673325070.8	57168
17	39080741.63	450380686.5	489309463.7	505644961.5	47345
18	27820988.55	323531585.3	356734483.7	353488625.3	34236
19	14495368.59	157405514.6	170315146	169783597.1	20520
20	5340253.275	61036445.43	62447867.15	63209742.93	8600
21	1253580.345	13854461	14151198.15	14374548.5	2163
22	201272.85	2262168	2732143.5	2574220.5	352
23	43002.36	389200.5	449050.5	493218	69
Grand Total	**725362806.6**	**8444453838**	**9426955730**	**9562940358**	**623904**

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



Due to the absence of the target variable in **previous_application.csv**, pivot table was not made.

E. Identify Top Correlations for Different Scenarios: Segment the dataset based on different scenarios and identify the top correlations for each segmented data using Excel functions.

- Top 10 Correlated Variables of application_data.csv dataset.

Variables	Correlations
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

- Due to the absence of target variable, it isn't possible to find correlation between variables and target variable in previous_application.csv.

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.

- application_data.csv

	DAYS_BIRTH	REGION_RATING_CLIENT_W_CITY	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_applicatiov.csv

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE
AMT_ANNUITY	1	0.810049781	0.815868523	0.820433175
AMT_APPLICATION	0.810049781	1	0.975771049	0.988711762
AMT_CREDIT	0.815868523	0.975771049	1	0.97235717
AMT_GOODS_PRICE	0.820433175	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 applications. 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:
- Text functions: clean(), substitute(), replace(), concatenate(), trim(), search(), find(), etc.
- Mathematical and Statistical functions: sum(), sumif(), count(), countif(), round(), avg(), min(), max(), subtotal(),averageif(),median(),mode(), etc.
- Data Visualization in Excel: Bar, Column, Line, Histogram, Pie, Scatter, Boxpot, Heatmap, 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

Thankyou