Project 6

Bank Loan Case Study

Final Project-2

Project description:

In this project, as an data analyst, I am tasked with analyzing loan application data to help a finance company improve its loan approval process. The company faces a significant risk due to some applicants defaulting on their loans, particularly those with limited or no credit history. The main challenge is to identify patterns in customer and loan attributes that can predict whether an applicant is likely to default.

Using Exploratory Data Analysis (EDA), the goal is to uncover insights into the factors influencing loan defaults. This analysis will help the company make informed decisions, such as approving or rejecting loans, reducing loan amounts, or adjusting interest rates for risky applicants. By identifying key attributes that correlate with payment difficulties, the company can minimize financial loss and ensure that capable applicants are not denied loans.

Approach (Concise and Simple):

1. **Data Exploration**: Understand the dataset by reviewing customer and loan attributes.
2. **Data Cleaning**: Handle missing data, correct any errors, and remove outliers that may impact the analysis.
3. **Univariate Analysis**: Analyze individual features to understand their distribution and key statistics.
4. **Bivariate Analysis**: Explore relationships between customer/loan attributes and loan status (approved, defaulted, rejected).
5. **Correlation Analysis**: Identify which variables are most related to loan defaults.
6. **Insights Generation**: Summarize patterns and insights to recommend better loan approval strategies.

This structured approach helps in understanding key factors that influence loan defaults, providing the company with data-driven strategies to minimize financial risks.

Tech-stack used: Microsoft Excel 2019

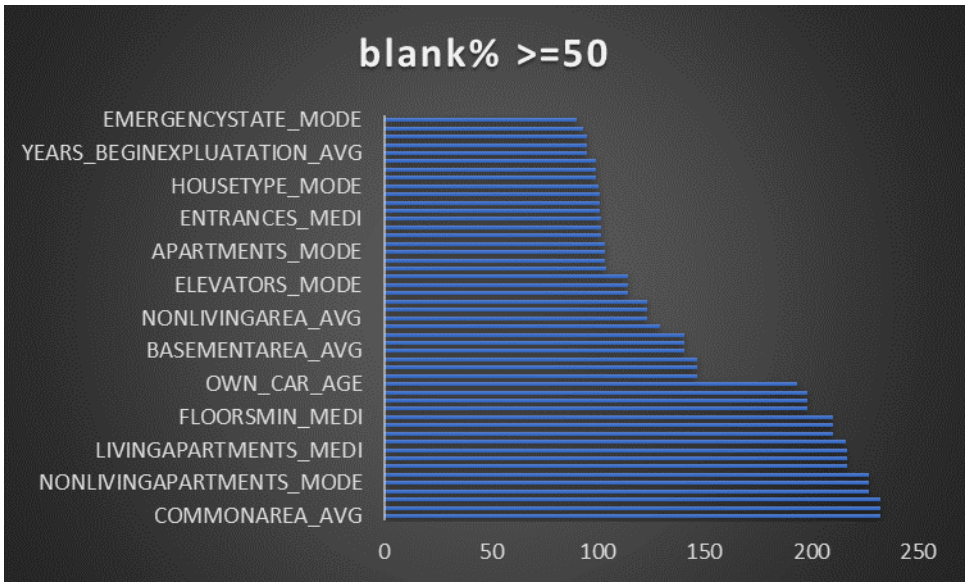
Excel is an excellent tool for handling, analyzing, and visualizing datasets of moderate size. It offers functionalities for data cleaning and statistical analysis. Excel's built-in charts and pivot tables are useful for creating visual representations like bar charts, pie charts, and histograms, which help in understanding trends in the dataset related to rejections, interviews, job types, and vacancies.

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.

Output: PRIMARY FILE

Total columns initially	121	**total column after cleaning**	72
total rows initially	49999	**total rows after cleaning**	49999



Handling of missing data:

1.Categorical data – Using mode

NAME_TYPE_SUITE
OCCUPATION_TYPE

2.Numerical data – Using median for the data with outliers and average for data without the outliers

COLUMN	percentage BLANKS	median	mode
OCCUPATION_TYPE	31.30862617		Unknown
AMT_REQ_CREDIT_BUREAU_HOUR	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_DAY	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_WEEK	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_MON	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_QRT	13.46826937	0	
AMT_REQ_CREDIT_BUREAU_YEAR	13.46826937	1	
NAME_TYPE_SUITE	0.38400768		Unaccompanied
OBS_30_CNT_SOCIAL_CIRCLE	0.33600672	0	
DEF_30_CNT_SOCIAL_CIRCLE	0.33600672	0	
OBS_60_CNT_SOCIAL_CIRCLE	0.33600672	0	
DEF_60_CNT_SOCIAL_CIRCLE	0.33600672	0	
AMT_GOODS_PRICE	0.07600152	450000	
AMT_ANNUITY	0.00200004	24939	
CNT_FAM_MEMBERS	0.00200004	2	
DAYS_LAST_PHONE_CHANGE	0.00200004	0	

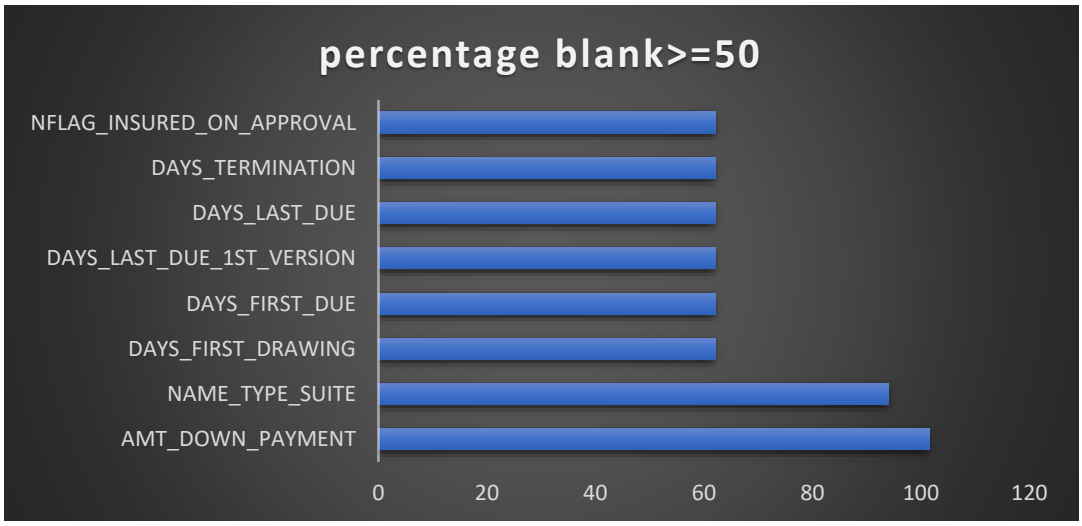
The data above shows there were total 121 column but some column had more than 50% of data missing which can not be imputed so we drop those columns

The missing data was handled by replacing with median where there were outliers present and for the categorical data mode was used to fill the missing values.

After cleaning 72 columns were left.

Output: SECONDARY FILE

The columns with more than 50% blanks



drop columns	
AMT_DOWN_PAYMENT	RATE_INTEREST_PRIMARY
NAME_TYPE_SUITE	RATE_INTERSET_PRIVILAGED
DAYS_FIRST_DRAWING	
DAYS_FIRST_DUE	
DAYS_LAST_DUE_1ST_VERSION	
DAYS_LAST_DUE	
DAYS_TERMINATION	
NFLAG_INSURED_ON_APPROVAL	

The handling of the missing data:

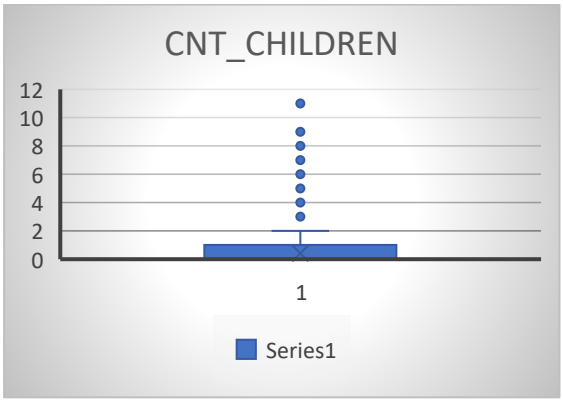
column	percentblank	median	replace
AMT_GOODS_PRICE	27.36976181	104017.5	
AMT_ANNUITY	26.87847337	10879.92	
CNT_PAYMENT	26.87847337		0
PRODUCT_COMBINATION	0.016002881		unknown

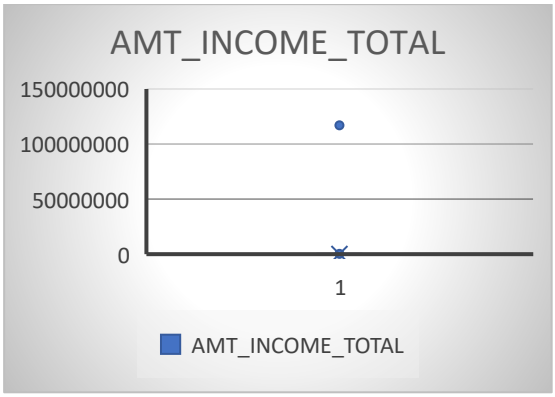
Insights:

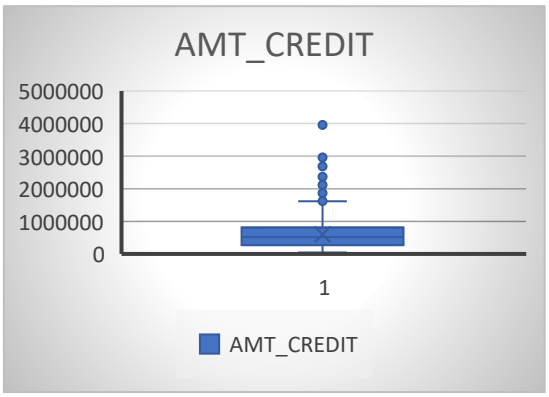
- There were about 36 columns in total initially and after dropping the columns with more than 50% of missing data the total column left was 26.
- The missing data was imputed using median as outliers were present and for categorical data in product_combination we put unknown in blanks as we cannot fil it with mode and in cnt_payment with filled the blanks with 0.

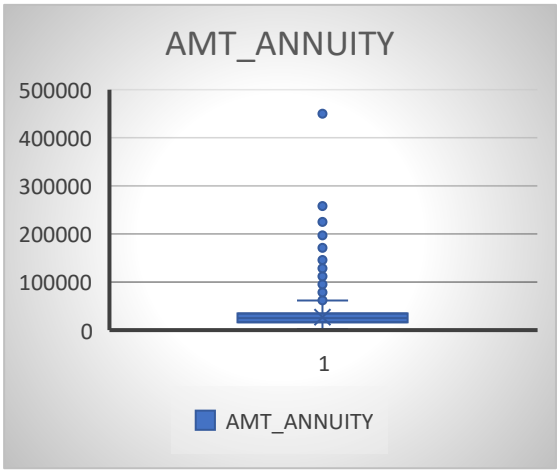
B. Identify Outliers in the Dataset: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Output: PRIMARY FILE









Insights:

- Based on the box plot analysis of the variables CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE, the presence of outliers in each column indicates variability in the dataset.
- For CNT_CHILDREN, outliers may suggest atypical family structures that could influence repayment behavior or it might be an error during data collection.
- In AMT_INCOME_TOTAL, outliers point to extreme income levels which highlight a potential risk factor for default if high-income clients are over-leveraged.
- The outliers in AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE suggest that certain clients are taking on unusually large loans or purchasing high-value goods, which may require closer scrutiny to mitigate risk.

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.

OUTPUT: PRIMARY FILE

TARGET = 1: Clients who had **payment difficulties**, meaning they were **late by more than a certain number of days** on at least one of the early instalments of their loan.

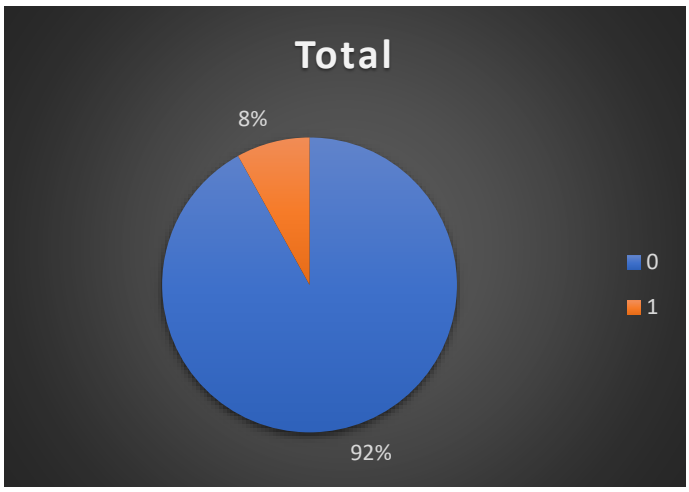
TARGET = 0: Clients who had **no payment difficulties**, meaning they made all their payments on time.

This means

Defaulted clients (TARGET = 1): These clients struggled to meet their loan repayment obligations. They missed or were late in making payments by more than a specified threshold on at least one instalment.

Non-defaulted clients (TARGET = 0): These clients managed their loan repayments well and made their payments on time without any issues.

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



Insights:

- Analysis shows that the number of non-defaulters (clients with TARGET = 0) is greater than the number of defaulters (clients with TARGET = 1). The data imbalance, with more non-defaulters than defaulters, reflects a generally good lending portfolio.

- The higher number of non-defaulters suggests that the lending institution's risk assessment and management strategies are effective in identifying and approving clients who are likely to repay their loans.

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.

OUTPUT: PRIMARY FILE

Univariate analysis/ segmented univariate analysis

1. For numerical data

univariate analysis of AMT_CREDIT	
AMT_CREDIT	
Mean	599700.5815
Standard Error	1799.674528
Median	514777.5
Mode	450000
Standard Deviation	402415.4339
Sample Variance	1.61938E+11
Kurtosis	1.917459058
Skewness	1.223668739
Range	4005000
Minimum	45000
Maximum	4050000
Sum	29984429376
Count	49999

univariate analysis of AMT_INCOME_TOT	
AMT_INCOME_TOTAL	
Mean	170767.5905
Standard Error	2378.391081
Median	145800
Mode	135000
Standard Deviation	531819.0951
Sample Variance	2.82832E+11
Kurtosis	46582.52582
Skewness	212.0777967
Range	116974350
Minimum	25650
Maximum	117000000
Sum	8538208758
Count	49999

Insights:

1.Income Distribution:

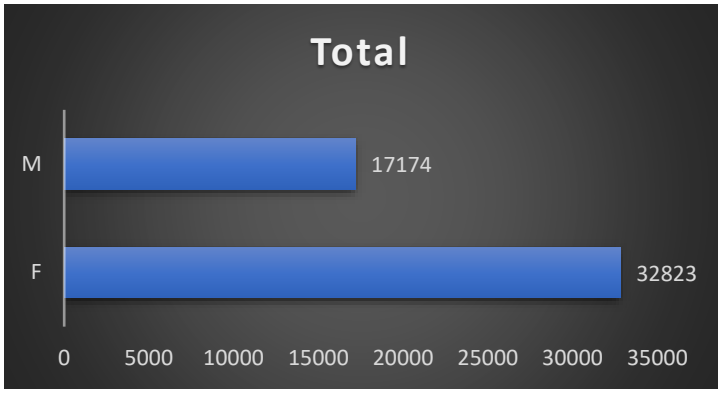
- **Mean and Median Income**: The mean income is higher than the median which indicate a right-skewed distribution where a small number of clients earn much more than the majority. This suggests a concentration of wealth and potential risk if the high earners are over-leveraged.
- **Income Range**: A wide range between the minimum and maximum income values indicates significant variability in client financial backgrounds.

2.Credit Amount Analysis:

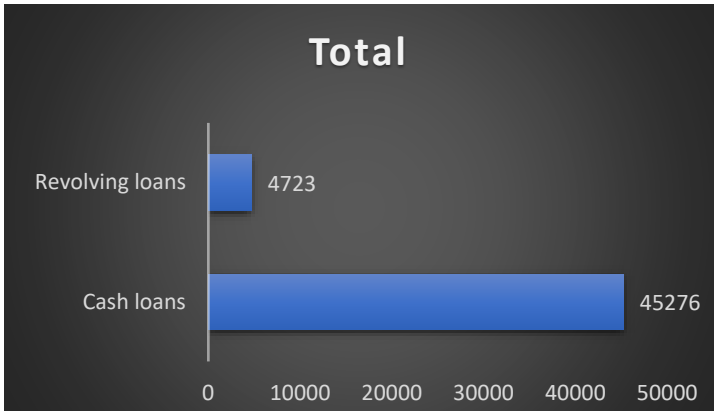
- **Average and Median Loan Amounts**: The average loan amount is considerably higher than the median, this suggests that a few clients are taking out disproportionately large loans. Understanding the reasons behind high loan amounts can help identify potential risks.

2.FOR CATEGORICAL DATA:

Univariate analysis for code_gender	
Row Labels	**Count of CODE_GENDER**
F	32823
M	17174
Grand Total	**49997**



Univariate ananlysis for name_contract	
Row Labels	**Count of NAME_CONTRACT_TYPE**
Cash loans	45276
Revolving loans	4723
Grand Total	**49999**



Insights:

1.Gender Distribution:

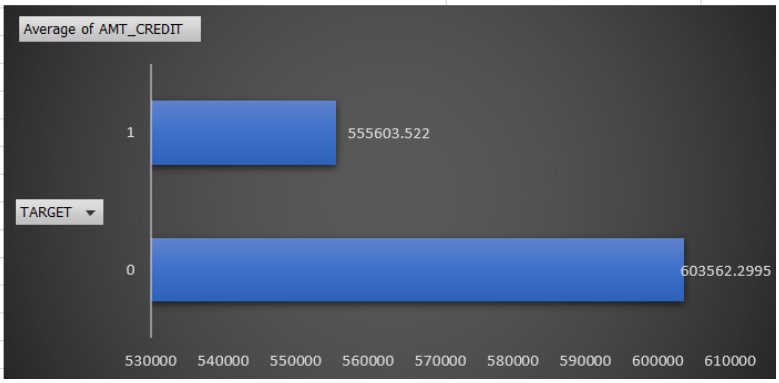
- **Higher Female Clientele**: The predominance of female clients in the dataset suggests that the bank may have successfully targeted or appealed to female borrowers.
- **Implications for Marketing and Product Development**: Understanding the needs of female borrowers could help the institution tailor products or services that meet their specific financial goals. This might include offerings like family loans, education loans or flexible repayment options.

2.Prevalence of Cash Loans:

- **Dominance of Cash Loans**: The higher number of cash loans indicates a strong demand for immediate liquidity among borrowers.
- **Understanding Loan Utilization**: Analyzing the reasons behind taking cash loans can provide valuable insights into client behavior. For example, if cash loans are primarily used for emergencies, lenders might consider offering more tailored financial products that provide better support in urgent situations.

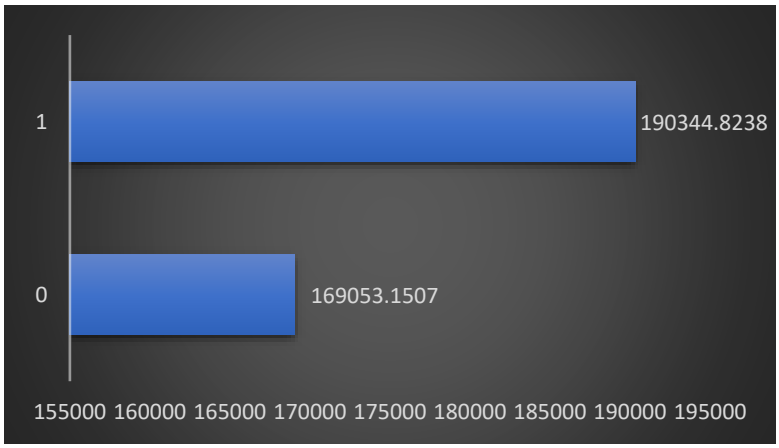
1.

Segmenting loan ammount by loan default status	
Row Labels	**Average of AMT_CREDIT**
0	603562.2995
1	555603.522
Grand Total	**599700.5815**



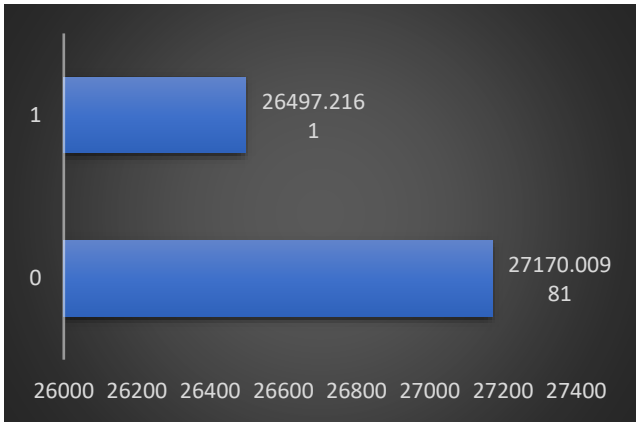
2.

segmenting income by loan default status	
Row Labels	**Average of AMT_INCOME_TOTAL**
0	169053.1507
1	190344.8238
Grand Total	**170767.5905**



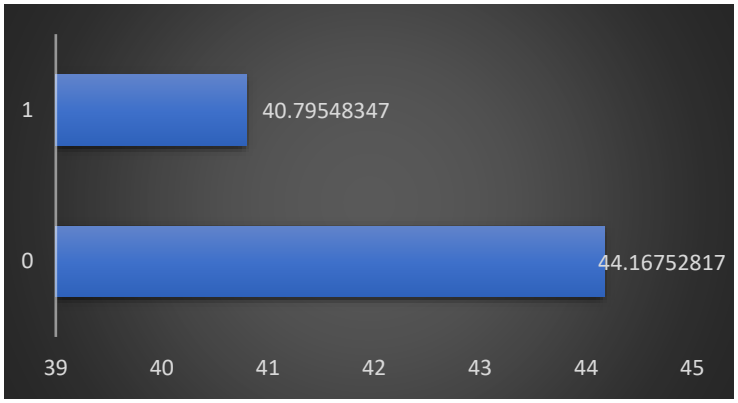
3.

segmenting loan annuity by default status	
Row Labels	**Average of AMT_ANNUITY**
0	27170.00981
1	26497.2161
Grand Total	**27115.83538**



4.

segmenting clients age by default status	
Row Labels	**Average of client's age**
0	44.16752817
1	40.79548347
Grand Total	**43.8960057**



Insights:

1.Non-Defaulters Have Higher Credit, Annuity, and Age:

- **Higher Credit Amounts:** The finding that non-defaulters have greater credit amounts suggests that these clients may be more financially stable and capable of managing higher loans. This could indicate a more favorable risk profile among non-defaulters.
- **Loan Annuity Insights:** A higher average loan annuity among non-defaulters may suggest that they are able to comfortably meet their repayment obligations.
- **Client Age Factor:** The greater age of non-defaulters might imply that older clients tend to have more established financial habits and stability. This could point to the importance of age as a potential risk factor in loan approvals.

2.Defaulters Have Higher Income:

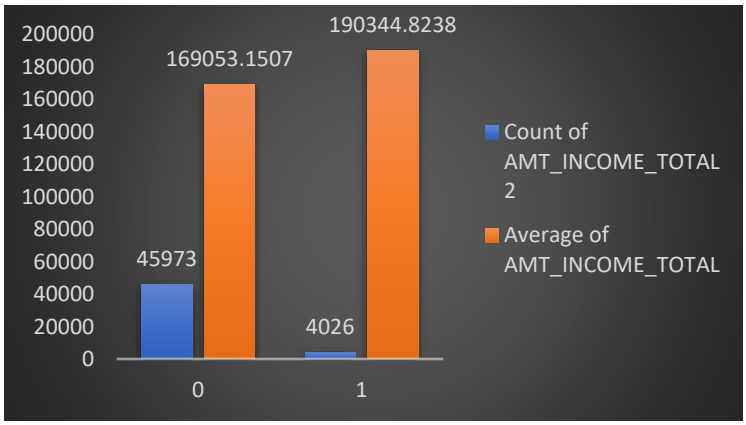
- **Income Disparity in Defaulters**: The clients in the default group have greater income which indicate that higher income alone does not guarantee repayment ability. This might suggest that income is not the only factor influencing loan default risk and that other elements, such as financial management skills or unexpected financial burdens, come into play.

Bivariate analysis:

1.

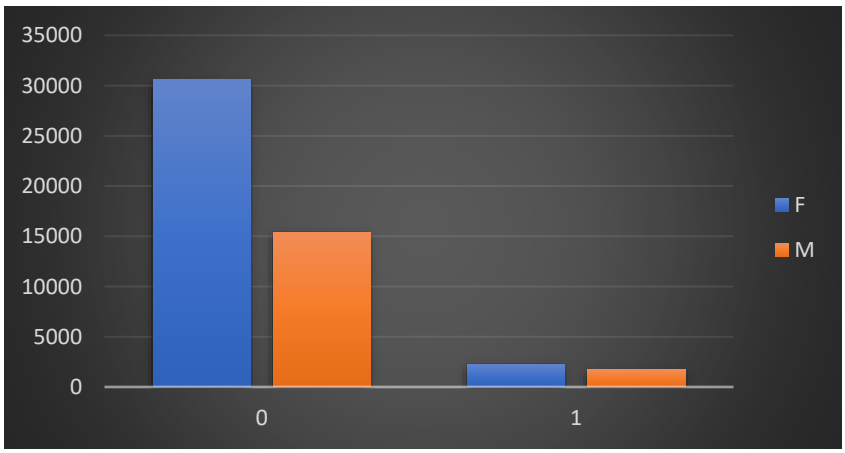
relationship analysis

Row Labels	Count of AMT_INCOME_TOTAL2	Average of AMT_INCOME_TOTAL
0	45973	169053.1507
1	4026	190344.8238
Grand Total	**49999**	**170767.5905**



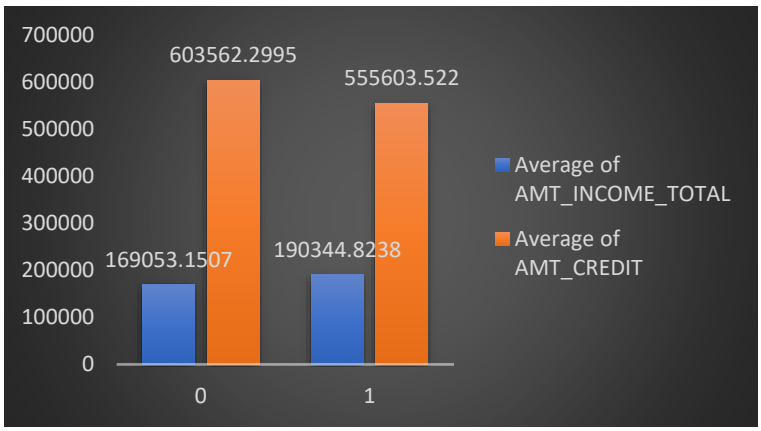
2.

Count of CODE_ Column Labels			
Row Labels	F	M	Grand Total
0	30559	15412	45971
1	2264	1762	4026
Grand Total	**32823**	**17174**	**49997**



3.

Row Labels	Average of AMT_INCOME_TOTAL	Average of AMT_CREDIT
0	169053.1507	603562.2995
1	190344.8238	555603.522
Grand Total	**170767.5905**	**599700.5815**



Insights:

1.Higher Count of Non-Defaulters (TARGET = 0):

The larger count of clients classified as non-defaulters indicates that a significant portion of the client base is effectively managing their loans and obligations. This may suggest that the lending institution's risk assessment processes are working well for the majority of borrowers.

Since there are more non-defaulters, this demographic could be a focus for future marketing and lending strategies, as they represent a stable client base. Understanding their demographics can help in tailoring products that fit their needs.

2. Higher Average Income Among Defaulters (TARGET = 1):

Income Discrepancy: The finding that the average income is higher among defaulters raises important questions about the nature of financial risk. It suggests that having a higher income does not automatically correlate with better repayment behavior.

3. The finding that non-defaulters have a higher average credit amount implies that these clients may be more confident in their repayment capabilities, resulting in them taking larger loans. This could indicate that they have a more stable financial foundation, potentially backed by good financial management practices.

OUTPUT: SECONDARY FILE

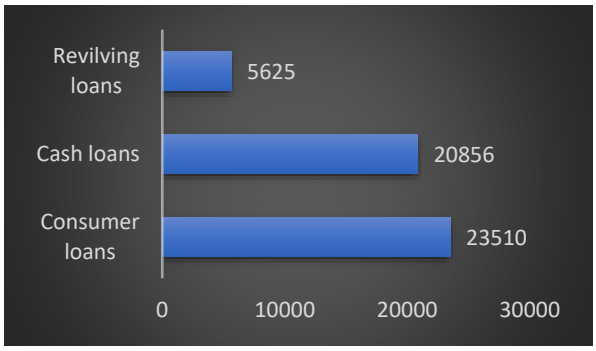
Univariate/segmented univariate analysis:

univariate analysis	
AMT_ANNUITY	
Mean	14507.54628
Standard Error	58.30233413
Median	10879.92
Mode	10879.92
Standard Deviation	13036.66787
Sample Variance	169954709.1
Kurtosis	17.66913934
Skewness	3.164915849
Range	234478.395
Minimum	0
Maximum	234478.395
Sum	725362806.6
Count	49999

AMT_CREDIT	
Mean	188542.8855
Standard Error	1379.549679
Median	78907.5
Mode	0
Standard Deviation	308473.6014
Sample Variance	95155962744
Kurtosis	14.88061385
Skewness	3.344679263
Range	4104351
Minimum	0
Maximum	4104351
Sum	9426955730
Count	49999

NAME_CONTRACT_TYPE

type	total
Consumer loans	23510
Cash loans	20856
Revilving loans	5625



Insights:

1.Annuity Distribution:

- **Average Loan Annuity**: The mean or average annuity amount indicates the typical loan payment clients are making on a periodic basis.
- **Comparison of Mean and Median**: The mean is higher than the median, it suggests that there are a few clients with very high annuities, which are pulling the average up.
- **Loan Annuity Variability**: A high standard deviation or wide range indicates significant variability in loan annuities across the client base, meaning that some clients have very large payments while others have much smaller ones.

2.Credit Amount Analysis:

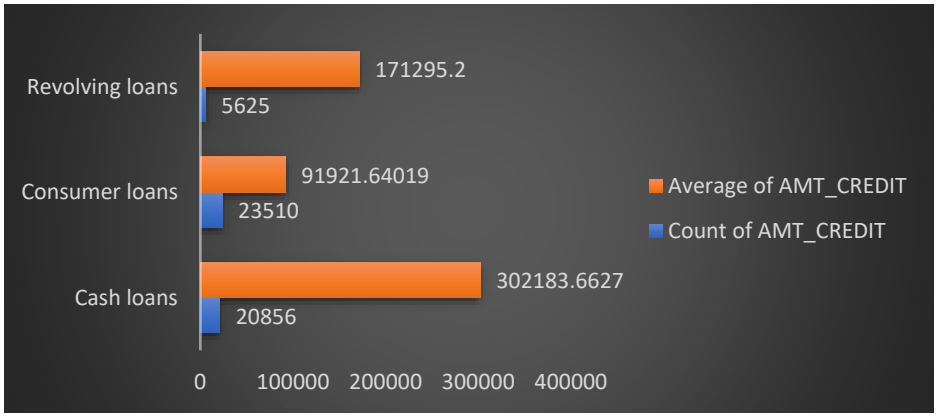
- **Average and Median Loan Amounts**: The average loan amount is considerably higher than the median, this suggests that a few clients are taking out disproportionately large loans. Understanding the reasons behind high loan amounts can help identify potential risks.

3.Contract type analysis:

- **Revolving Loans vs. Cash Loans**: The revolving loans are typically smaller than cash loans is consistent with the nature of revolving credit. Revolving loans, such as credit cards, generally offer smaller more flexible borrowing limits that can be reused once paid off. This reflects a lower average loan size and suggests that clients using revolving loans may be seeking short-term, smaller financing solutions. Cash loans, often used for one-time, larger tend to have higher amounts
- **Prevalence of Consumer Loans**: If consumer loans re the most frequent type, it suggests that the majority of clients are borrowing for everyday expenses or personal consumption. This indicates that the lender's target market is primarily focused on consumers seeking financing for personal or family needs rather than business or investment purposes.

Segmented univariate analysis:

Row Labels	Count of AMT_CREDIT	Average of AMT_CREDIT
Cash loans	20856	302183.6627
Consumer loans	23510	91921.64019
Revolving loans	5625	171295.2
Grand Total	**49991**	**188573.0578**



Insights:

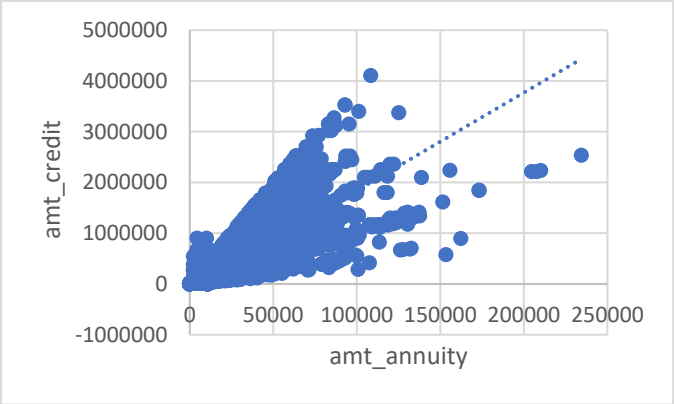
1. Higher Average Credit for Cash Loans:

- Cash loans typically have higher average amounts as they are often used for long-term purchases such as homes, vehicles, or other large expenses.

- Revolving loans, such as credit cards or lines of credit, tend to offer smaller, flexible borrowing limits compared to cash loans. Clients may be using these loans to cover short-term expenses or maintain liquidity, which explains the lower average credit amount.
- The consumer loans have the highest count which indicates strong demand for this type of loan. Consumer loans are often used for a wide range of personal expenses (e.g., home improvement, education, large purchases). Banks can leverage this insight by focusing on expanding and marketing consumer loans, as they clearly cater to a large segment of the borrower population.

Bivariate analysis:

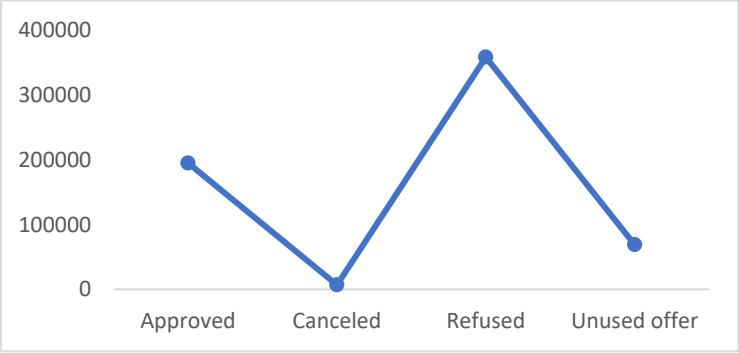
1. AMT_CREDIT vs. AMT_ANNUITY



2.

AMT_CREDIT vs. NAME_CONTRACT_STATUS

Row Labels	Average of AMT_CREDIT
Approved	194728.2553
Canceled	6867.494764
Refused	357962.9779
Unused offer	68754.13737
Grand Total	**188542.8855**



Insights:

1. The linear trendline suggests a positive correlation between the loan amount (AMT_CREDIT) and the loan annuity (AMT_ANNUITY). This means that as the loan amount increases, the annuity payments also tend to increase proportionally. This is logical because larger loans typically require higher regular payments.

There are some outliers they could indicate unusual cases where the annuity is disproportionately high or low relative to the loan amount. These could be due to special loan terms, financial difficulties, or miscalculated risk profiles.

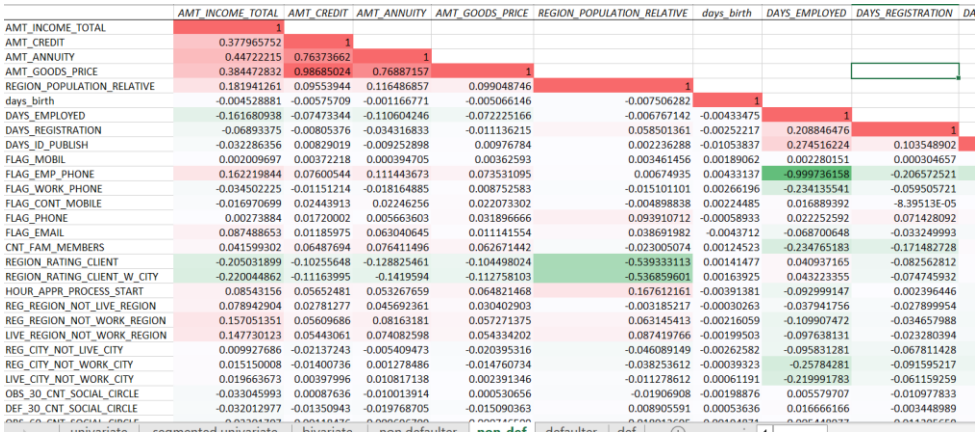
2. Larger loans come with higher risk, both in terms of repayment capability and potential default. Banksmay be refusing these larger loans as part of their risk management strategy.

The cancelled loans have the least average AMT_CREDIT suggests that clients who initially apply for smaller loan amounts may be more likely to cancel their applications. This could indicate that clients seeking smaller loans may have less financial commitment leading them to cancel.

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**.**

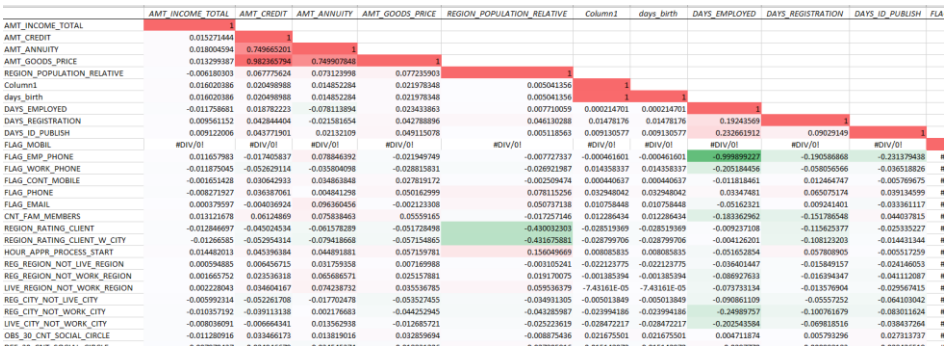
OUTPUT: Primary file

NON-DEFAULTERS:



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	days_birth	DAYS_EMPLOYED	DAYS_REGISTRATION
AMT_INCOME_TOTAL	1							
AMT_CREDIT	0.377965752	1						
AMT_ANNUITY	0.44722215	0.76373662	1					
AMT_GOODS_PRICE	0.384472832	0.98685024	0.76887157	1				
REGION_POPULATION_RELATIVE	0.181941261	0.09553944	0.116486857	0.099048746	1			
days_birth	-0.004528881	-0.00575709	-0.001166771	-0.005066146	-0.007506282	1		
DAYS_EMPLOYED	-0.161680938	-0.07473344	-0.110604246	-0.072225166	-0.006767142	-0.00433475	1	
DAYS_REGISTRATION	-0.06893375	-0.00805376	-0.034316833	-0.011136215	0.058501361	-0.00252217	0.208846476	1
DAYS_ID_PUBLISH	-0.032286356	0.00829019	-0.009252898	0.00976784	0.002236288	-0.01053837	0.274516224	0.103548902
FLAG_MOBIL	0.002009697	0.00372218	0.000394705	0.00362593	0.003461456	0.00189062	0.002280151	0.000304657
FLAG_EMP_PHONE	0.162219844	0.07600544	0.111443673	0.073531095	0.00674935	0.00433137	-0.999736158	-0.206572521
FLAG_WORK_PHONE	-0.034502225	-0.01151214	-0.018164885	0.008752583	-0.015101101	0.00266196	-0.234135541	-0.059505721
FLAG_CONT_MOBILE	-0.016970699	0.02443913	0.02246256	0.022073302	-0.004898838	0.00224485	0.016889392	-8.39513E-05
FLAG_PHONE	0.00273884	0.01720002	0.005663603	0.031896666	0.093910712	-0.00058933	0.022252592	0.071428092
FLAG_EMAIL	0.087488653	0.01185975	0.063040645	0.011141554	0.038691982	-0.0043712	-0.068700648	-0.033249993
CNT_FAM_MEMBERS	0.041599302	0.06487694	0.076411496	0.062671442	-0.023005074	0.00124523	-0.234765183	-0.171482728
REGION_RATING_CLIENT	-0.205031899	-0.10255648	-0.128825461	-0.104498024	-0.539333113	0.00141477	0.040937165	-0.082562812
REGION_RATING_CLIENT_W_CITY	-0.220044862	-0.11163995	-0.1419594	-0.112758103	-0.536859601	0.00163925	0.043223355	-0.074745932
HOUR_APPR_PROCESS_START	0.08543156	0.05652481	0.053267659	0.064821468	0.167612161	-0.00391381	-0.092999147	0.002396446
REG_REGION_NOT_LIVE_REGION	0.078942904	0.02781277	0.045692361	0.030402903	-0.003185217	-0.00030263	-0.037941756	-0.027899954
REG_REGION_NOT_WORK_REGION	0.157051351	0.05609686	0.08163181	0.057271375	0.063145413	-0.00216059	-0.109907472	-0.034657988
LIVE_REGION_NOT_WORK_REGION	0.147730123	0.05443061	0.074082598	0.054334202	0.087419766	-0.00199503	-0.097638131	-0.023280394
REG_CITY_NOT_LIVE_CITY	0.009927686	-0.02137243	-0.005409473	-0.020395316	-0.046089149	-0.00262582	-0.095831281	-0.067811428
REG_CITY_NOT_WORK_CITY	0.015150008	-0.01400736	0.001278486	-0.014760734	-0.038253612	-0.00039323	-0.25784281	-0.091595217
LIVE_CITY_NOT_WORK_CITY	0.019663673	0.00397996	0.010817138	0.002391346	-0.011278612	0.00061191	-0.219991783	-0.061159259
OBS_30_CNT_SOCIAL_CIRCLE	-0.033045993	0.00087636	-0.010013914	0.000530656	-0.01906908	-0.00198876	0.005579707	-0.010977833
DEF_30_CNT_SOCIAL_CIRCLE	-0.032012977	-0.01350943	-0.019768705	-0.015090363	0.008905591	0.00053636	0.016666166	-0.003448989

Defaulter:



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	Column1	days_birth	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	FLA
AMT_INCOME_TOTAL	1										
AMT_CREDIT	0.015271444	1									
AMT_ANNUITY	0.018004594	0.749665201	1								
AMT_GOODS_PRICE	0.013299387	0.982365794	0.749907848	1							
REGION_POPULATION_RELATIVE	-0.006180303	0.067775624	0.073123998	0.077235903	1						
Column1	0.016020386	0.020498988	0.014852284	0.021978348	0.005041356	1					
days_birth	0.016020386	0.020498988	0.014852284	0.021978348	0.005041356	1	1				
DAYS_EMPLOYED	-0.011758681	0.018782223	-0.078113894	0.024333863	0.007710059	0.000214701	0.000214701	1			
DAYS_REGISTRATION	0.009561152	0.042844404	-0.021581654	0.042788896	0.046130288	0.01478176	0.01478176	0.19241569	1		
DAYS_ID_PUBLISH	0.009122006	0.043771901	0.02132109	0.049115078	0.005118563	0.009130577	0.009130577	0.232661912	0.09029149	1	
FLAG_MOBIL	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	
FLAG_EMP_PHONE	0.011657983	-0.017405837	0.078846392	-0.021949749	-0.007727337	-0.000461601	-0.000461601	-0.999899227	-0.190586868	-0.231379438	
FLAG_WORK_PHONE	-0.011875045	-0.052629114	-0.035804098	-0.028815831	-0.026921987	0.014358337	0.014358337	-0.205184456	-0.058056566	-0.036518826	
FLAG_CONT_MOBILE	-0.001651428	0.030642933	0.034863848	0.027819172	-0.002509474	0.000440637	0.000440637	-0.011818461	0.012464747	-0.005769675	
FLAG_PHONE	-0.008271927	0.036387061	0.004841298	0.050162999	0.078115256	0.032948042	0.032948042	0.03347481	0.065075174	0.039134599	
FLAG_EMAIL	0.000179597	-0.004036924	0.096360456	-0.002123308	0.050737138	0.010758448	0.010758448	-0.05162321	0.009241401	-0.033361117	
CNT_FAM_MEMBERS	0.013121678	0.06124869	0.075838463	0.05559165	-0.017257146	0.012286434	0.012286434	-0.183362962	-0.151786548	0.044037815	
REGION_RATING_CLIENT	-0.012846697	-0.045024534	-0.061578289	-0.051728498	-0.430032303	-0.028519369	-0.028519369	-0.009237108	-0.115625377	-0.025335227	
REGION_RATING_CLIENT_W_CITY	-0.01266585	-0.052954314	-0.079418668	-0.057154865	-0.431675881	-0.028799706	-0.028799706	-0.004126201	-0.108123203	-0.014431344	
HOUR_APPR_PROCESS_START	0.014482013	0.045396384	0.044891881	0.057159781	0.156049669	0.008085835	0.008085835	-0.051652854	0.057808905	-0.005517259	
REG_REGION_NOT_LIVE_REGION	0.000594885	0.006456715	0.031759358	0.007169988	-0.003105241	-0.022123775	-0.022123775	-0.036401447	-0.015849157	-0.024146053	
REG_REGION_NOT_WORK_REGION	0.001665752	0.023536318	0.065686571	0.025157881	0.019170075	-0.001385394	-0.001385394	-0.086927633	-0.016394347	-0.041112087	
LIVE_REGION_NOT_WORK_REGION	0.002228043	0.034604167	0.074238732	0.035536785	0.059536379	-7.431616-05	-7.431616-05	-0.073733134	-0.013576904	-0.029567415	
REG_CITY_NOT_LIVE_CITY	-0.005992314	-0.052261708	-0.017702478	-0.035527455	-0.034931305	-0.005013849	-0.005013849	-0.090861109	-0.05557252	-0.064103042	
REG_CITY_NOT_WORK_CITY	-0.010357192	-0.039113138	0.002176683	-0.044252945	-0.043285987	-0.023994186	-0.023994186	-0.24989757	-0.100761679	-0.083011624	
LIVE_CITY_NOT_WORK_CITY	-0.008036091	-0.006664341	0.013562938	-0.012685721	-0.025223619	-0.028472217	-0.028472217	-0.202543584	-0.069818516	-0.038437264	
OBS_30_CNT_SOCIAL_CIRCLE	-0.011280916	0.033466173	0.013819016	0.032859694	-0.008875436	0.021675501	0.021675501	0.004711874	0.005793296	0.027313737	

Insights:

Positive correlation

If certain variables (e.g. AMT_CREDIT, AMT_ANNUITY) show a strong positive correlation with the other variable, it suggests that as these amounts increase, the likelihood of defaulting also increases. The values in red shows the positive correlation.

Negative Correlation

When one variable increases, the other variable tends to decrease. Conversely, when one variable decreases, the other variable tends to increase. The green ones shows the negative correlation between variables.

In a loan default analysis, negative correlations can be significant predictors of risk.

RESULT:

In this project, I conducted a comprehensive analysis of loan applications to understand the factors influencing loan defaults. By performing univariate analysis to explore the distribution of individual variables, such as income, credit amount, and loan types. I conducted bivariate analysis to investigate relationships between variables, particularly focusing on how attributes like average income and credit amount correlate with the likelihood of defaulting. I calculated correlation coefficients to identify strong predictors, discovering that higher income often correlates with a lower likelihood of default.I also assessed data imbalance within the target variable, noting that non-defaulters significantly outnumbered defaulters, which highlighted potential challenges in predictive modeling. Additionally, I examined outliers using box plots, providing insights into how extreme values can influence the analysis.

Overall, this project enhanced my analytical skills and deepened my understanding of credit risk assessment. I learned how to utilize Excel tools for data analysis effectively, interpret statistical relationships, and derive actionable insights to inform lending strategies. This experience has equipped me with valuable knowledge that I can apply in real-world data-driven decision-making scenarios.

Drive link:

This my excel workbook 1 and 2