BANK LOAN ANALYSIS

PROJECT MADE BY

SAMIYA ALAM

Samiyaalam1710@gmail.com

Project description

This project aimed to analyse Bank loan data to identify patterns and key indicators of loan default. The main objective was to gain insights into the factors influencing loan repayment behaviours. When a customer applies for a loan, our company faces two primary risks:

- **Lost Business:** If the applicant can repay the loan but is not approved, the company loses business.
- **Financial Loss:** If the applicant cannot repay the loan and is approved, the company faces a financial loss.

The dataset we analysed contains information about loan applications and includes two types of scenarios:

1. **Customers with Payment Difficulties:** Customers who had a late payment of more than X days on at least one of the first Y instalments of the loan.
2. **All Other Cases:** Cases where the payment was made on time.

When a customer applies for a loan, there are four possible outcomes:

- **Approved:** The company has approved the loan application.
- **Cancelled:** The customer cancelled the application during the approval process.
- **Refused:** The company rejected the loan.
- **Unused Offer:** The loan was approved, but the customer did not use it.

Business Objectives:

The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants. The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

APPROACH

Initially, I started with two datasets: current applications and previous applications. The first step involved thorough data cleaning, which included the removal of blanks and unnecessary or unrelated columns. It was determined that several columns in the provided datasets were not essential for the risk analysis, so these were excluded to streamline the analysis process.

Next, I conducted a comprehensive risk assessment analysis. This was followed by executing the required tasks and generating various visual insights. Utilizing Excel pivot tables etc, Also I created bar charts, box plots, and column charts to effectively present the findings and insights.

The approach included cleaning the dataset, handling missing values, performing exploratory data analysis, and identifying correlations between variables and loan default. Various Excel functions and features were utilized to perform univariate, segmented univariate, and bivariate analysis.

Tech-Stack Used

- **Microsoft Excel:** Used for data cleaning, analysis, and visualization
- **Microsoft Word**: Used for writing the project's information and insights

TASKS PERFORMED

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.

B. Identify Outliers in the Dataset:

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

C. Analyse Data Imbalance:

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

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.

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.

Task A: Identify Missing Data and Deal with it Appropriately

In this task, I analyzed missing data and identified unwanted or unrelated columns and blanks. I utilized Excel's COUNTBLANK functions to calculate the count and percentage of missing values in each column. I started cleaning the dataset in Excel, determined the missing percentage for every column, and removed columns with a missing percentage above 40% as well as those deemed unnecessary.

% OF NULL	0	0	0	0	0	0	0	0	0	0.002	0.07606	0.38549	0	0	0	0	0	0	0	0	0	193.266
COUNT OF NULL	0	0	0	0	0	0	0	0	0	1	38	192	0	0	0	0	0	0	0	0	0	32950
	SK_ID_CURR	TARGET	NAME_CC	CODE_GE	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOC	NAME_TY	NAME_IN	NAME_ED	NAME_FA	NAME_HC	REGION_F	DAYS_BIR	DAYS_EM	DAYS_REG	DAYS_ID_	OWN_CAR
	100002	1	Cash loan	M	N	Y	0	202500	406598	24700.5	351000	Unaccom	Working	Secondary	Single / nc	House / a	0.0188	-9461	-637	-3648	-2120	
	100003	0	Cash loan	F	N	N	0	270000	1293503	35698.5	1129500	Family	State serv	Higher edu	Married	House / a	0.00354	-16765	-1188	-1186	-291	
	100004	0	Revolving	M	Y	Y	0	67500	135000	6750	135000	Unaccom	Working	Secondary	Single / nc	House / a	0.01003	-19046	-225	-4260	-2531	26
	100006	0	Cash loan	F	N	Y	0	135000	312683	29686.5	297000	Unaccom	Working	Secondary	Civil marri	House / a	0.00802	-19005	-3039	-9833	-2437	
	100007	0	Cash loan	M	N	Y	0	121500	513000	21865.5	513000	Unaccom	Working	Secondary	Single / nc	House / a	0.02866	-19932	-3038	-4311	-3458	
	100008	0	Cash loan	M	N	Y	0	99000	490496	27517.5	454500	Spouse, pa	State serv	Secondary	Married	House / a	0.03579	-16941	-1588	-4970	-477	
	100009	0	Cash loan	F	Y	Y	1	171000	1560726	41301	1395000	Unaccom	Commerci	Higher edu	Married	House / a	0.03579	-13778	-3130	-1213	-619	17
	100010	0	Cash loan	M	Y	Y	0	360000	1530000	42075	1530000	Unaccom	State serv	Higher edu	Married	House / a	0.00312	-18850	-449	-4597	-2379	8
	100011	0	Cash loan	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary	Married	House / a	0.01863	-20099	365243	-7427	-3514	
	100012	0	Revolving	M	N	Y	0	135000	405000	20250	405000	Unaccom	Working	Secondary	Single / nc	House / a	0.01969	-14469	-2019	-14437	-3992	

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
28172	126	9944	25385	29199
EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMEN
0.083036967	0.262948593	0.13937578	0.0247	0.0369
0.311267311	0.622245775		0.0959	0.0529
	0.555912083	0.729566691		
	0.65044169			
	0.322738287			
	0.354224732	0.621226338		
0.774761413	0.723999852	0.492060094		
	0.714279286	0.54065445		
0.587334047	0.205747288	0.751723715		
	0.746643629			
0.319760172	0.651862333	0.363945239		
0.72204445	0.555183162	0.652896552		
0.464831117	0.715041819	0.176652579	0.0825	
	0.566906613	0.77008707	0.1474	0.0973
0.721939769	0.642656205		0.3495	0.1335
0.115634337	0.346633981	0.678567689		
	0.23637784	0.062103038		
	0.683513346			
	0.706428403	0.556727426	0.0278	0.0617
	0.58661714	0.477649155		
0.565654882	0.113374513		0.0722	0.0801
0.43770902	0.233766958	0.542445144		
	0.457142972	0.358951229	0.0907	0.0795
	0.624304737	0.669056695	0.1443	0.0848
	0.786179309	0.565607981	0.1433	0.1455
0.561948409	0.651405637	0.461482391	0.0722	0.0147

RED colour cells represent empty cells here

After identifying the remaining missing values, I utilized the median function for imputation for missing values.

0	0	
0	0	
EXT_SOURCE_2	EXT_SOURCE_3	OBS_
0.262948593	0.13937578	
0.622245775	0.53527625	
0.555912083	0.729566691	
0.65044169	0.53527625	
0.322738287	0.53527625	
0.354224732	0.621226338	
0.723999852	0.492060094	
0.714279286	0.54065445	
0.205747288	0.751723715	
0.746643629	0.53527625	
0.651862333	0.363945239	
0.555183162	0.652896552	
0.715041819	0.176652579	
0.566906613	0.77008707	
0.642656205	0.53527625	
0.346633981	0.678567689	
0.23637784	0.062103038	
0.683513346	0.53527625	
0.706428403	0.556727426	
0.58661714	0.477649155	
0.113374513	0.53527625	

I calculated the median value using the Excel function =MEDIAN(range) for each numerical column.

Subsequently, I employed these median values to impute any missing data within their respective columns.

This iterative process ensured that the dataset was eventually devoid of any missing values, achieving a comprehensive and reliable dataset for analysis with 0 null values in numerical columns.

Hence, I successfully handled the missing data.

Task B: Identify Outliers in the Dataset

Outliers are data points that significantly deviate from the majority of values in a dataset, either being much larger or considerably smaller. These outliers can indicate variability in measurements, experimental errors, or novel occurrences.

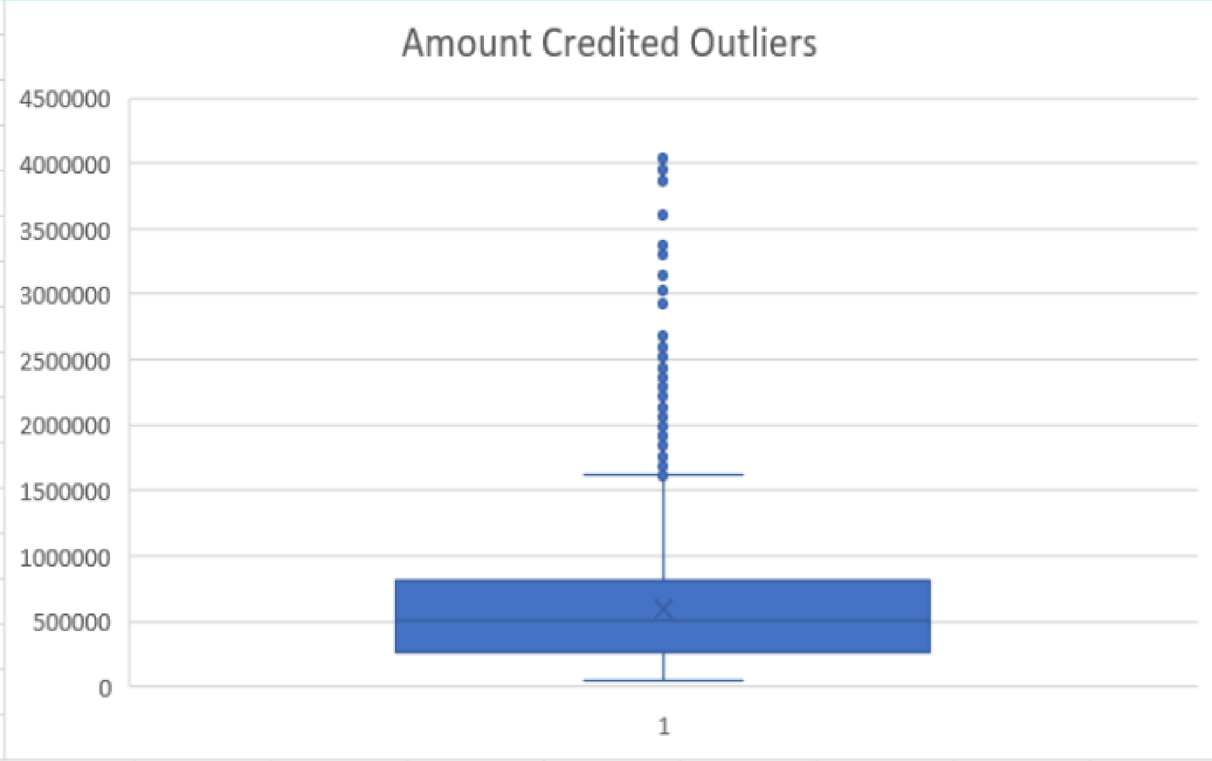
In the datasets, I identified a substantial number of outliers. Using Excel formula, I calculated the quartile ranges Q1 (first quartile) and Q3 (third quartile) and determined the interquartile range (IQR) as IQR=Q3−Q1.From this, I computed the upper and lower bounds using the following formulas:

- **Lower Bound:** Q1−1.5×IQR
- **Upper Bound:** Q3+1.5×IQR

Any data point below the lower bound or above the upper bound is considered an outlier. Applying these bounds with Excel functions, I was able to identify the data points that qualified as outliers based on these conditions using excel formulas and functions.

	SK_ID_CURR	OUTLINERS	TARGET	NAME_CO	CODE_GEI	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INCOME_TOTAL	OUTLIERS	AMT_CREDIT	OUTLIERS
% OF NULL	0		0	0	0	0	0	0	0		0	
COUNT OF NULL	0		0	0	0	0	0	0	0		0	
QUARTILE 1	114570.5		0	#NUM!	#NUM!	#NUM!	#NUM!	0	112500		270000	
QUARTILE 3	143438.5		0	#NUM!	#NUM!	#NUM!	#NUM!	1	202500		808650	
IQR	28868		0	#NUM!	#NUM!	#NUM!	#NUM!	1	90000		538650	
LOWER BOUND	71268.5		0	#NUM!	#NUM!	#NUM!	#NUM!	-1.5	-22500		-537975	
UPPER BOUND	186740.5		0	#NUM!	#NUM!	#NUM!	#NUM!	2.5	337500		1616625	
	100002	Normal	1	Cash loans	M	N	Y	0	202500	Normal	406597.5	Normal
	100003	Normal	0	Cash loans	F	N	N	0	270000	Normal	1293502.5	Normal
	100004	Normal	0	Revolving	M	Y	Y	0	67500	Normal	135000	Normal
	100006	Normal	0	Cash loans	F	N	Y	0	135000	Normal	312682.5	Normal
	100007	Normal	0	Cash loans	M	N	Y	0	121500	Normal	513000	Normal
	100008	Normal	0	Cash loans	M	N	Y	0	99000	Normal	490495.5	Normal
	100009	Normal	0	Cash loans	F	Y	Y	1	171000	Normal	1560726	Normal
	100010	Normal	0	Cash loans	M	Y	Y	0	360000	Outlier	1530000	Normal
	100011	Normal	0	Cash loans	F	N	Y	0	112500	Normal	1019610	Normal
	100012	Normal	0	Revolving	M	N	Y	0	135000	Normal	405000	Normal
	100014	Normal	0	Cash loans	F	N	Y	1	112500	Normal	652500	Normal
	100015	Normal	0	Cash loans	F	N	Y	0	38419.155	Normal	148365	Normal
	100016	Normal	0	Cash loans	F	N	Y	0	67500	Normal	80865	Normal
	100017	Normal	0	Cash loans	M	Y	N	1	225000	Normal	918468	Normal

90000	Normal	199008	Normal
360000	Outlier	733315.5	Normal
135000	Normal	1125000	Normal
112500	Normal	450000	Normal
198000	Normal	641173.5	Normal
121500	Normal	454500	Normal
99000	Normal	247275	Normal
180000	Normal	540000	Normal
202500	Normal	1193580	Normal
202500	Normal	604152	Normal
135000	Normal	288873	Normal
108000	Normal	746280	Normal
202500	Normal	661702.5	Normal
90000	Normal	180000	Normal
202500	Normal	305221.5	Normal
99000	Normal	260640	Normal
130500	Normal	1350000	Normal
360000	Outlier	1506816	Normal
54000	Normal	135000	Normal
540000	Outlier	675000	Normal
76500	Normal	454500	Normal
225000	Normal	314055	Normal
81000	Normal	675000	Normal
180000	Normal	837427.5	Normal
67500	Normal	298728	Normal
81000	Normal	247500	Normal
360000	Outlier	640458	Normal
540000	Outlier	1227901.5	Normal
180000	Normal	1663987.5	Outlier
180000	Normal	1080000	Normal
324000	Normal	1130760	Normal
112500	Normal	95940	Normal



Task C: Analyse Data Imbalance

In this task, I analysed the data distribution to determine if there was any imbalance in the dataset. I inferred that the accuracy of data representation in visual charts is crucial, particularly when counting the applicants grouped as 0 and 1 in the target variable.

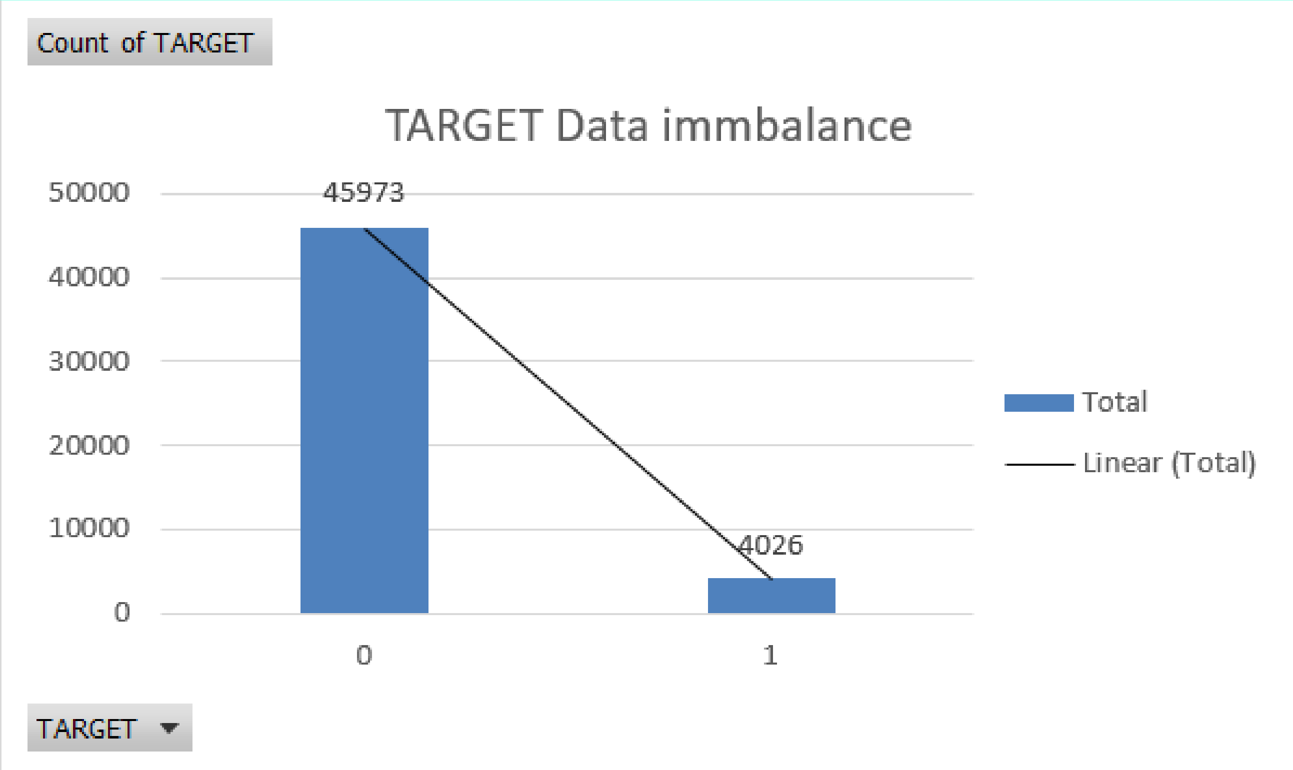
Here,

1 represents clients with payment difficulties (those who had a late payment of more than X days on at least one of the first Y installments of the loan)

0 represents all other cases.

To visually represent this analysis, I created a chart to illustrate any class imbalance present in the dataset.

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



Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis

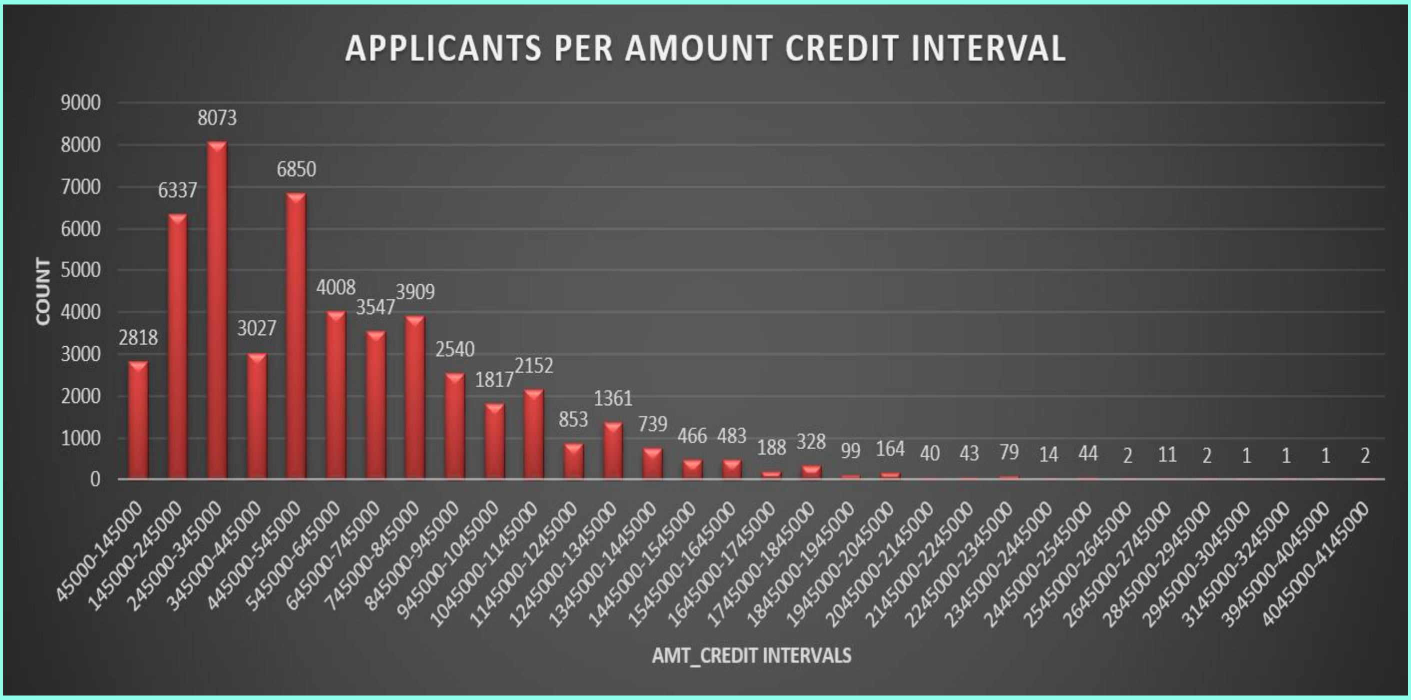
I conducted **univariate analysis** to understand the distribution of individual variables, **segmented univariate analysis** to compare variable distributions across different scenarios, and **bivariate analysis** to explore relationships between variables and the target variable using Excel functions and features.

Univariate analysis

For the univariate analysis, let's consider amt_credit, I segmented the data into intervals and calculated the count of individuals within each interval. Additionally, I determined the average amt_credit and its standard deviation. To achieve this, I utilized Excel functions such as COUNT, AVERAGE, and along with other statistical functions, to perform a comprehensive descriptive analysis.

It can be seen in the table and visualisation below

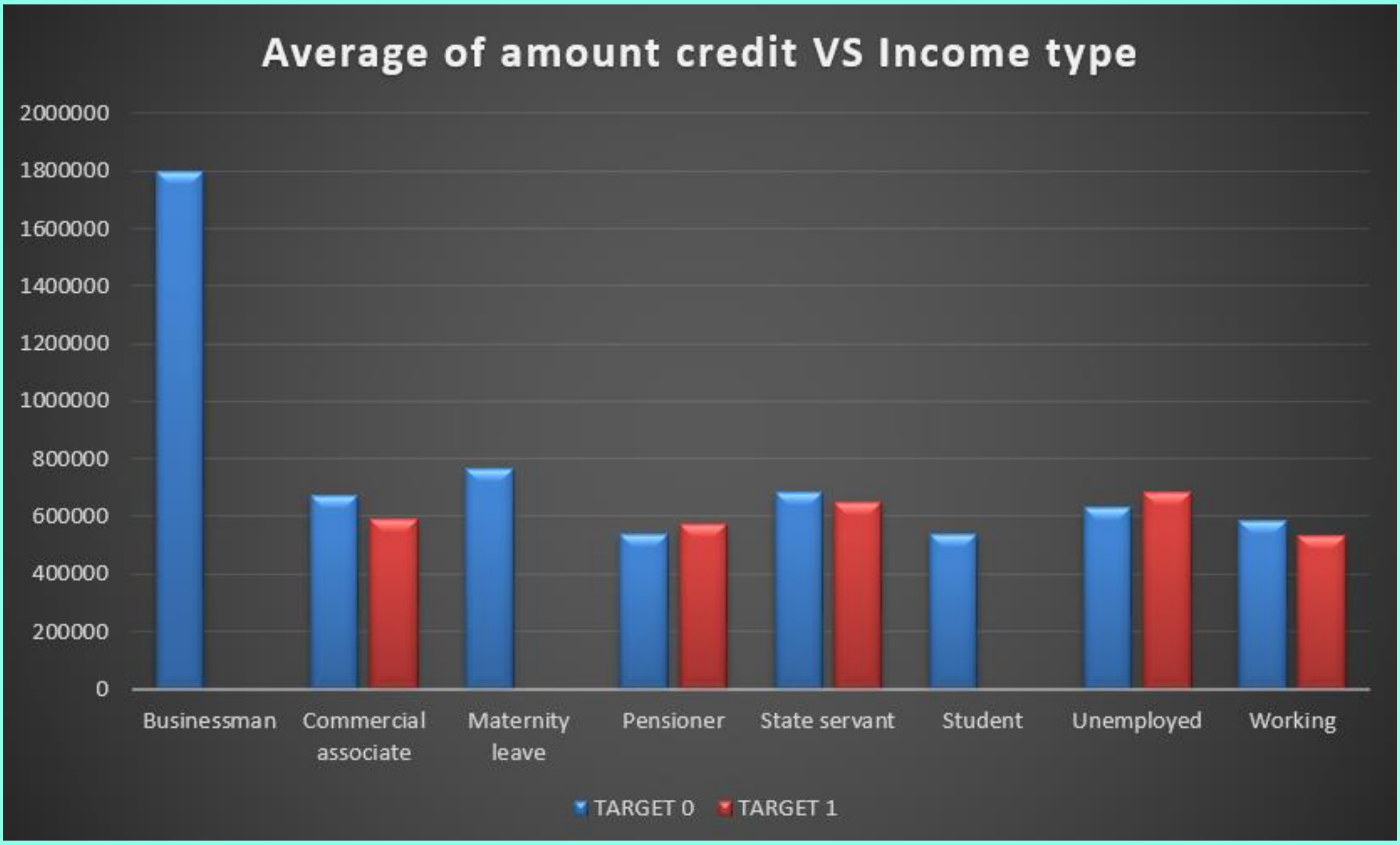
AMT_CREDIT INTERVAL	Count	Average	StdDev of AMT_CREDIT
45000-145000	2818	108054.6499	28289.91212
145000-245000	6337	199038.1032	27959.9842
245000-345000	8073	286191.8027	28121.05194
345000-445000	3027	390796.6858	27959.33647
445000-545000	6850	490149.314	33220.67024
545000-645000	4008	586633.6594	32646.29156
645000-745000	3547	686696.8461	24038.49111
745000-845000	3909	792026.1009	28063.93238
845000-945000	2540	899279.4366	25178.34138
945000-1045000	1817	999939.1849	28270.57535
1045000-1145000	2152	1096985.873	27034.75595
1145000-1245000	853	1200328.022	27836.76354
1245000-1345000	1361	1287608.565	23801.31597
1345000-1445000	739	1370870.689	32454.46452
1445000-1545000	466	1496707.02	26638.55267
1545000-1645000	483	1568933.189	26636.99603
1645000-1745000	188	1701837.527	27553.9759
1745000-1845000	328	1785691.372	22050.64186
1845000-1945000	99	1897560.318	28537.21907
1945000-2045000	164	1994419.262	23939.97124
2045000-2145000	40	2086003.463	20256.32528
2145000-2245000	43	2181237.279	28752.2815
2245000-2345000	79	2255811.152	16700.72213
2345000-2445000	14	2384827.071	24918.98298
2445000-2545000	44	2500671.784	28045.59426
2545000-2645000	2	2606400	0
2645000-2745000	11	2695577.727	3267.637804
2845000-2945000	2	2928330	4709.331163



Segmented univariate analysis

For the segmented analysis, I focused on the NAME_INCOME_TYPE variable and segmented the data based on the target variable (0 and 1). I calculated the average and count of amt_credit across different income types for both segments (clients with payment difficulties and all other cases). Using Excel's filtering and pivot table features, I efficiently analyzed and compared the distributions of amt_credit across the different income types within each segment.

INCOME TYPE	Average of AMT_CREDIT	Count of TARGET
0	**603562.2995**	**45973**
Businessman	1800000	2
Commercial associate	674204.1047	10679
Maternity leave	765000	1
Pensioner	538034.2905	8419
State servant	682281.7971	3314
Student	539246.7	5
Unemployed	630000	4
Working	583777.2373	23549
1	**555603.522**	**4026**
Commercial associate	592067.8281	864
Pensioner	570833.5329	501
State servant	652143.75	198
Unemployed	684000	2
Working	531829.7901	2461
Grand Total	**599700.5815**	**49999**



I performed these univariate, segmented univariate on many different scenarios

Bivariate analysis

For the bivariate analysis, I explored the relationships between pairs of variables to gain deeper insights into the factors influencing loan defaults. I explored relationships between variables and the target variable using Excel functions and features.

For example I found how AMT_INCOME_TOTAL (Income of the client) influences the AMT_CREDIT(Credit amount of the loan)

I also found the correlation between income and credit for the target

Correlation between amt_credit and amt_income_total	0.069316

Income	Average of AMT_CREDIT
25650-125650	425228.7416
125650-225650	632779.0874
225650-325650	839540.732
325650-425650	935945.9604
425650-525650	1009091.246
525650-625650	1123616.396
625650-725650	1046201.618
725650-825650	1287182.647
825650-925650	1161345.214
925650-1025650	1303200
1025650-1125650	1153857.971
1225650-1325650	1095111
1325650-1425650	914911.2
1525650-1625650	900000
1725650-1825650	1237500
1825650-1925650	781920
1925650-2025650	731068.5
2225650-2325650	1125000
3525650-3625650	953460
3725650-3825650	1241023.5
116925650-117025650	562491

SCATTER PLOT:

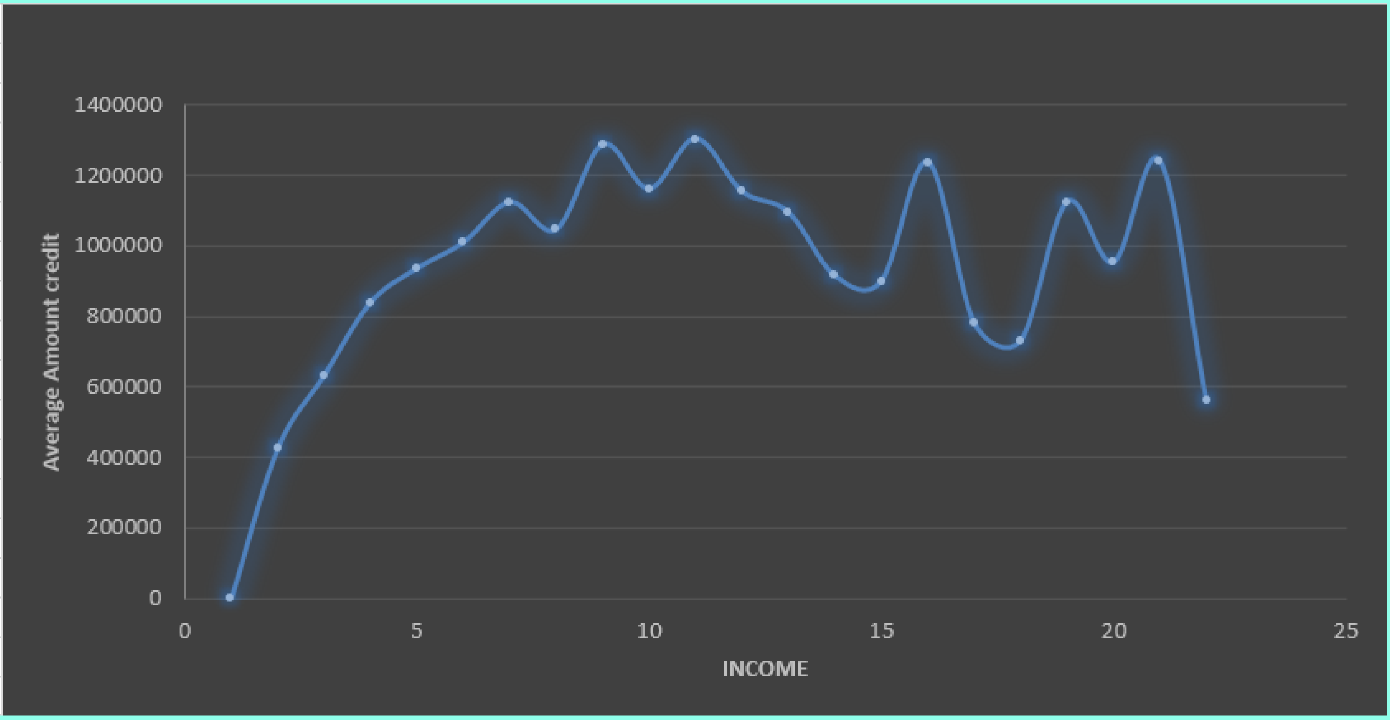
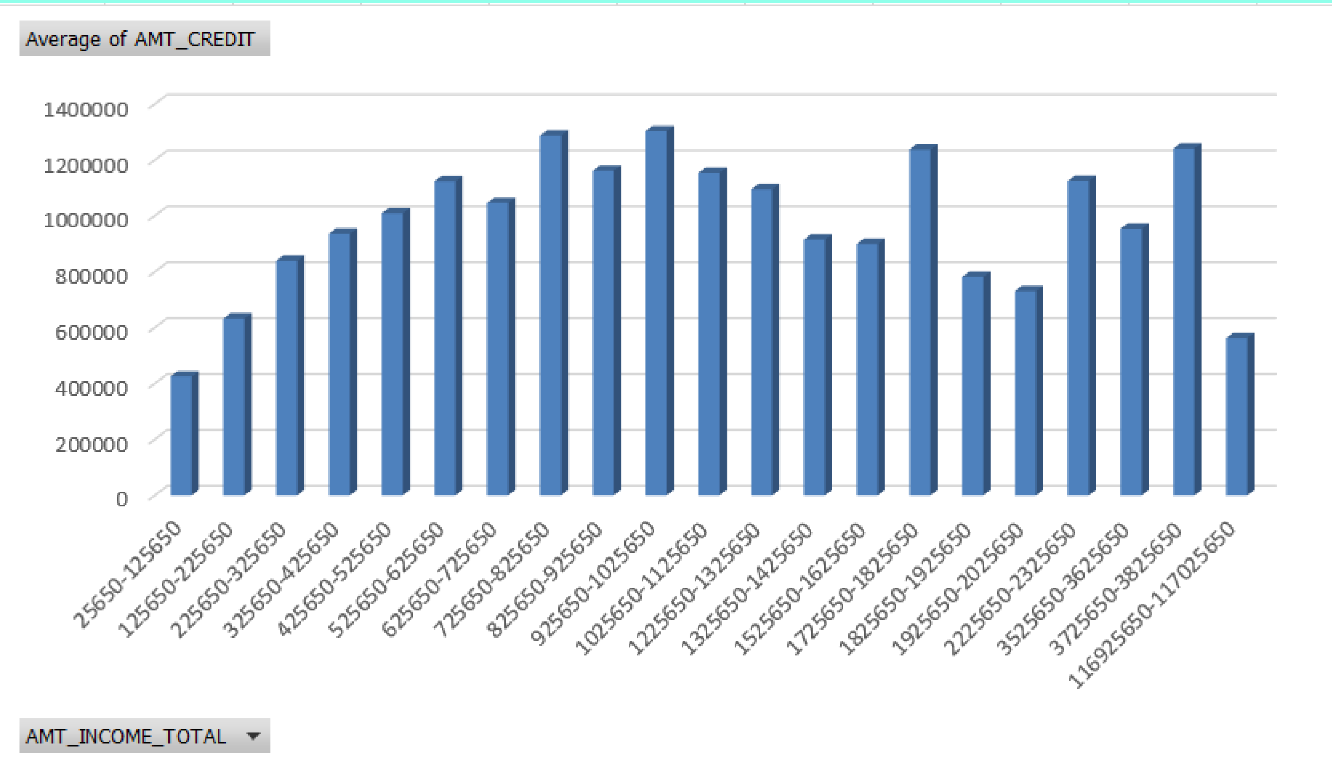


CHART:



Then I conducted the bivariate analysis for different columns and target variable. It is included in the excel sheet.

Task E: Identify Top Correlations for Different Scenarios

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan defaults.

In this task, I segmented the dataset into distinct scenarios,

TARGET 1 as clients with payment difficulties versus

TARGET 0 as all other cases

to identify key correlations using Excel's powerful functions and tools.

1. **Target 0 Analysis**:

I focused on instances where the target variable is 0 and selected pertinent columns from the dataset:

- TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, REGION_POPULATION_RELATIVE, DAYS_BIRTH(yrs), DAYS_EMPLOYED(YRS), DAYS_ID_PUBLISH(YRS), REGION_RATING_CLIENT.

By calculating correlations among these variables, I constructed a heatmap to visualize and interpret the relationships inherent in this scenario.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(YRS)	DAYS_ID_PUBLISH(YRS)	REGION_RATING_CLIENT
CNT_CHILDREN	1							
AMT_INCOME_TOTAL	0.047239208	1						
AMT_CREDIT	0.010694145	0.405722186	1					
REGION_POPULATION_RELATIVE	-0.026180136	0.175147495	0.069697098	1				
DAYS_BIRTH(yrs)	-0.321838399	-0.07298948	0.051561603	0.032925565	1			
DAYS_EMPLOYED(YRS)	-0.249818283	-0.18264393	-0.083301882	6.84645E-05	0.632546882	1		
DAYS_ID_PUBLISH(YRS)	0.044435064	-0.033099227	0.019064097	-0.001192028	0.26245279	0.258827904	1	
REGION_RATING_CLIENT	0.011249256	-0.230803611	-0.103141909	-0.533898789	0.002879069	0.042684004	0.015308781	1

CORRELATION MATRIX HEATMAP FOR TARGET 0

Target 1 Analysis:

- Similarly, I conducted correlation analysis for cases where the target variable is 1 using the same selected columns.

- The resulting heatmap provided insights into the correlations specific to this scenario, highlighting significant relationships between variables like CNT_CHILDREN, AMT_INCOME_TOTAL, and DAYS_BIRTH(YEARS) etc.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(yrs)	DAYS_EMPLOYED(YRS)	DAYS_ID_PUBLISH(YRS)	REGION_RATING_CLIENT
CNT_CHILDREN	1							
AMT_INCOME_TOTAL	-0.067868884	1						
AMT_CREDIT	0.052602922	0.379065503	1					
REGION_POPULATION_RELATIVE	-0.009045767	0.143754478	0.061340672	1				
DAYS_BIRTH(yrs)	-0.234570815	0.038939008	0.165483538	-0.052235285	1			
DAYS_EMPLOYED(YRS)	-0.161840193	-0.107733917	-0.043275678	-0.114078999	0.544596527	1		
DAYS_ID_PUBLISH(YRS)	0.100470237	-0.019024168	0.094938793	0.021928602	0.287903471	0.224197387	1	
REGION_RATING_CLIENT	-0.024645026	-0.14432102	-0.014778391	-0.497637597	0.100112506	0.09049179	0.019338163	1

CORRELATION MATRIX HEATMAP FOR TARGET 1

Through these analyses, I deciphered nuanced correlations within each segmented dataset, offering valuable insights into factors influencing loan scenarios categorized by payment regularity. This approach leverages Excel's functionality to uncover meaningful patterns essential for informed decision-making in financial contexts.

EXCEL SHEET: https://docs.google.com/spreadsheets/d/10L00Wej3YLT2TeuAF8ZqVjzUNXfmubzb/edit?usp=sharing&ouid=118309411958556729568&rtpof=true&sd=true

RESULT

Through this project, I successfully identified and addressed missing data, detected and managed outliers, analysed data imbalance, and conducted comprehensive univariate, segmented univariate, and bivariate analyses. By employing advanced Excel functions and statistical techniques, I gained valuable insights into the key factors influencing loan default. This enhanced my understanding of the critical attributes and their interrelationships within the loan application dataset. The project highlighted the importance of data cleaning, accurate representation, and detailed analysis in assessing loan risk, ultimately contributing to a more robust risk management framework for the Bank Loan Case Study.