BANK LOAN CASE STUDY

PROJECT DESCRIPTION: The case study aims to identify a pattern if a client has difficulty paying their installment. This project also provides an example of how EDA (EXPLORATORY DATA ANALYSIS) might be used in the real-world corporate setting. Some learning about risk analytics in banking. The goal is to detect what are the trends that show whether a client has trouble paying their installments. The process starts from data cleaning to the basic method of data analysis such as univariate, segment univariate and bivariate analysis and correlation between variables

APPROACH: To perform the project a systematic approach was followed. A dataset was downloaded. Microsoft Excel 2021 was selected as the primary tool for data analysis such as cleaning, missing data and other EDAs as needed due to its versatility and robust capabilities in handling tabular data. Specific techniques such as pivot tables, charts, and formulas were employed to analyze the dataset and extract meaningful insights.

TECH-STACK USED: Software: Microsoft Excel 2021
Purpose: Excel was chosen for its extensive data analysis functions, including pivot tables, charts, and statistical functions and visualization of the hiring data

INSIGHTS: various key insights were uncovered through the data analytics process. Meaningful trends and visualization patterns were observed in the data

A. IDENTIFY MISSING DATA AND DEAL WITH IT APPROPRIATELY: As a data analyst, you come across missing data in the loan application dataset. It is essential to handle missing data effectively to ensure the accuracy of the analysis.

Task: Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.

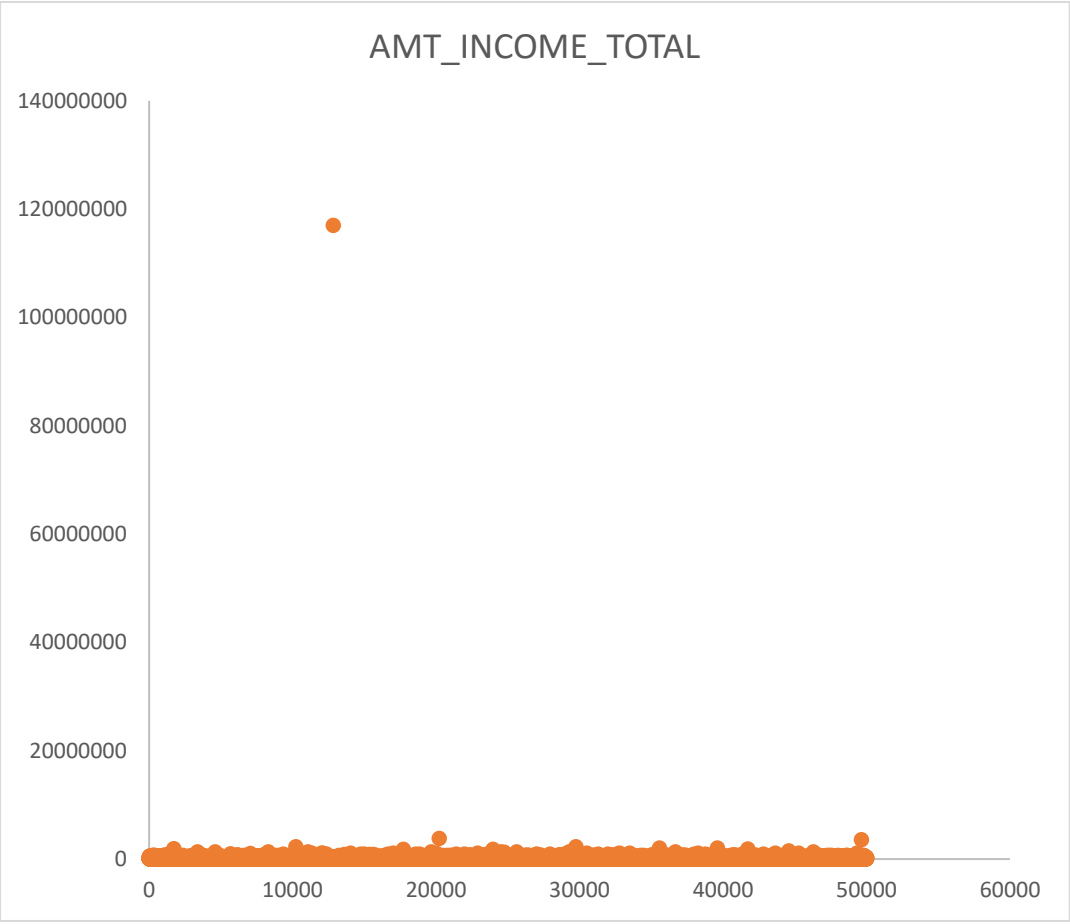
PROPORTION MISSING VALUE	MODIFIED METHOD
< 50%	FOR NUMERICAL VALUES AVERAGE AND MEDIAN WILL BE
	FOR CATEGORICAL VALUES MODE WILL BE PREFERRED
>50	REMOVE THE WHOLE COLUMN COMPLETELY

SL.NO	REPLACING MISSING COLUMNS	FORMULA(AVERAGE)	MEDIAN
1	AVERAGE AMT_ANNUITY	27107.38	24939
2	AVERAGE AMT_GOODS_PRICE	539060.04	450000
3	NAME_TYPE_SUITE	Unaccompanied	MODE
4	OCCUPATION_TYPE	Laborers	MODE
5	AVERAGE CNT_FAM_MEMBERS	2.16	2
6	OBS_30_CNT_SOCIAL_CIRCLE	1.42	0
7	DEF_30_CNT_SOCIAL_CIRCLE	0.14	0
8	OBS_60_CNT_SOCIAL_CIRCLE	1.40	0
9	DEF_60_CNT_SOCIAL_CIRCLE	0.10	0
10	DAYS_LAST_PHONE_CHANGE	-964	-755
11	AMT_REQ_CREDIT_BUREAU_HOUR	0	0
12	AMT_REQ_CREDIT_BUREAU_DAY	0	0
13	AMT_REQ_CREDIT_BUREAU_WEEK	0	0
14	AMT_REQ_CREDIT_BUREAU_MON	0	0
15	AMT_REQ_CREDIT_BUREAU_QRT	0	0
16	AMT_REQ_CREDIT_BUREAU_YEAR	2	2

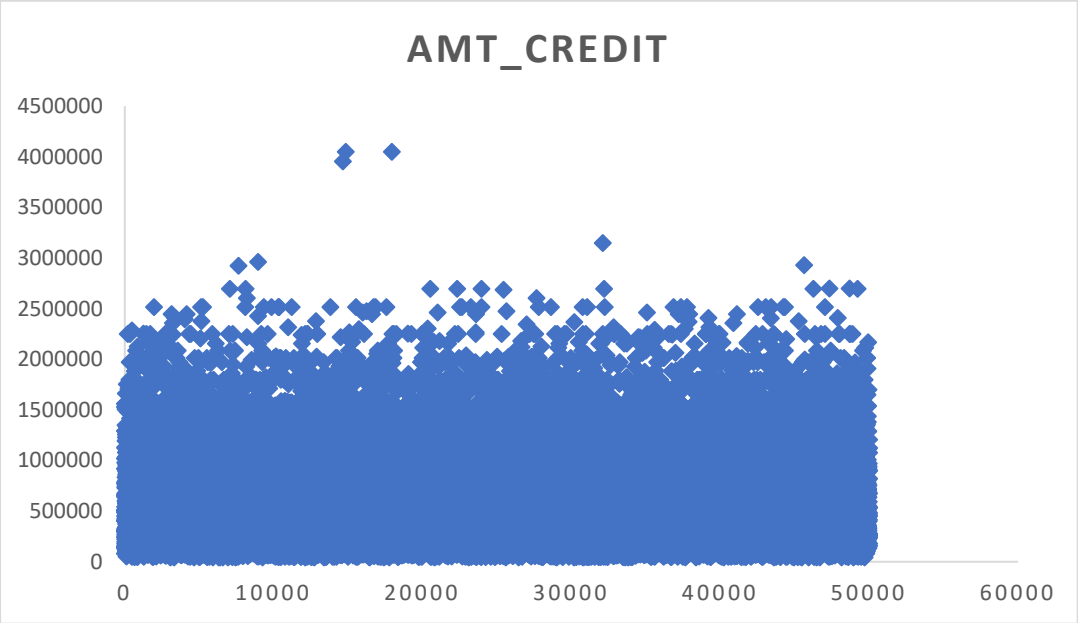
B. Identify Outliers in the Dataset: Outliers can significantly impact the analysis and distort the results. You need to identify outliers in the loan application dataset.

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

Q1	112500
Q3	202500
IQR(Q3-Q1)	90000
MIN	-22500
MAX	225000
MEDIAN	145800



Q1	270000
Q3	808650
IQR(Q3-Q1)	538650
MIN	-537975
MAX	1616625
MEDIAN	514777.5

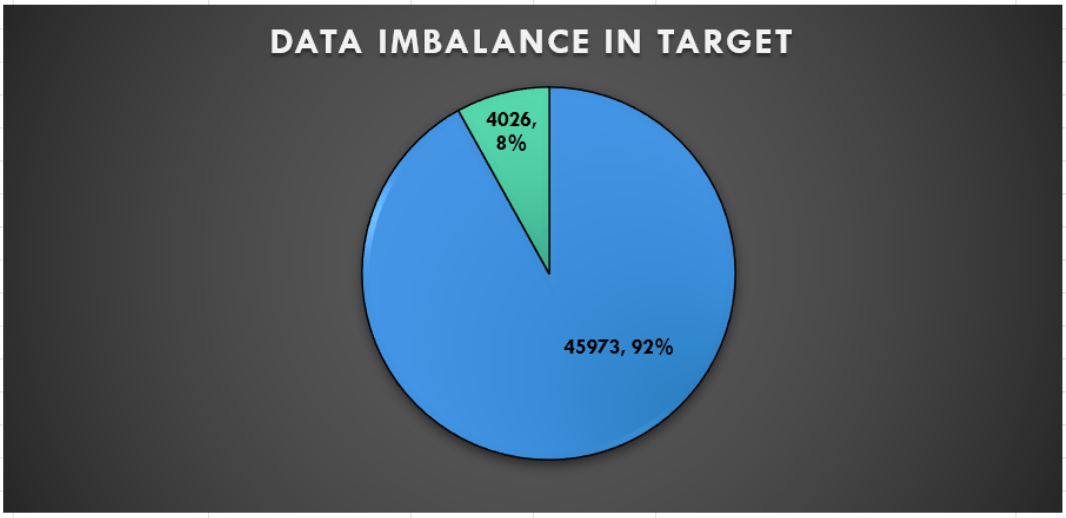


C. Analyse Data Imbalance: Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

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

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

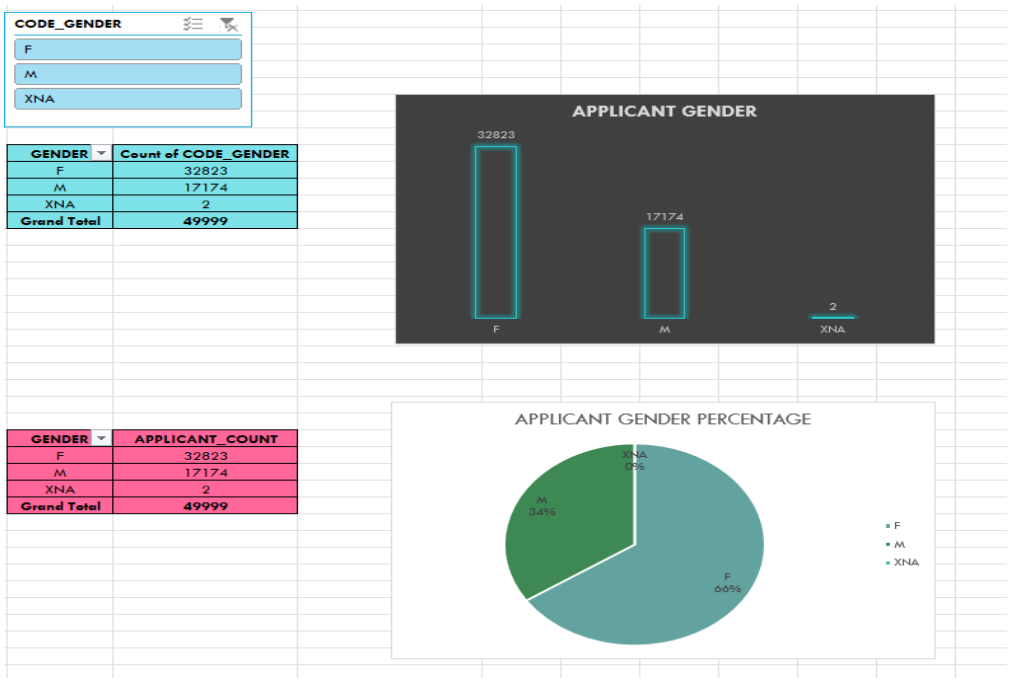
TARGET		
0	NO PAYMENT DIFFICULTY	45973
1	DEFAULT CHANCE	4026
	RATIO OF DATA IMBALANCE	11.42



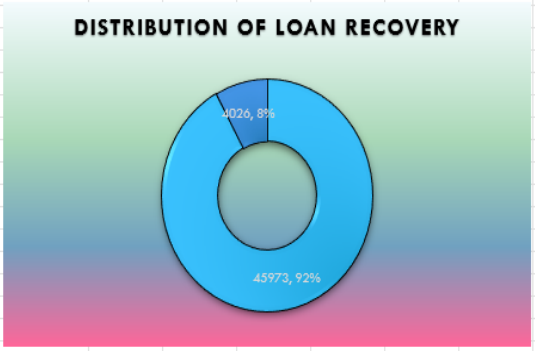
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis: To gain insights into the driving factors of loan default, it is important to conduct various analyses on consumer and loan attributes.

Task: 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:



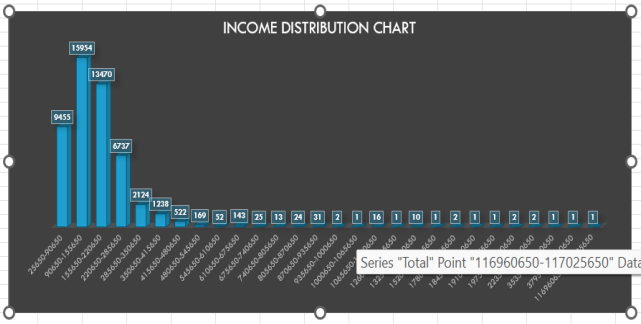
CATEGOR	LOAN PAYMENT
0	45973
1	4026
Grand Total	**49999**

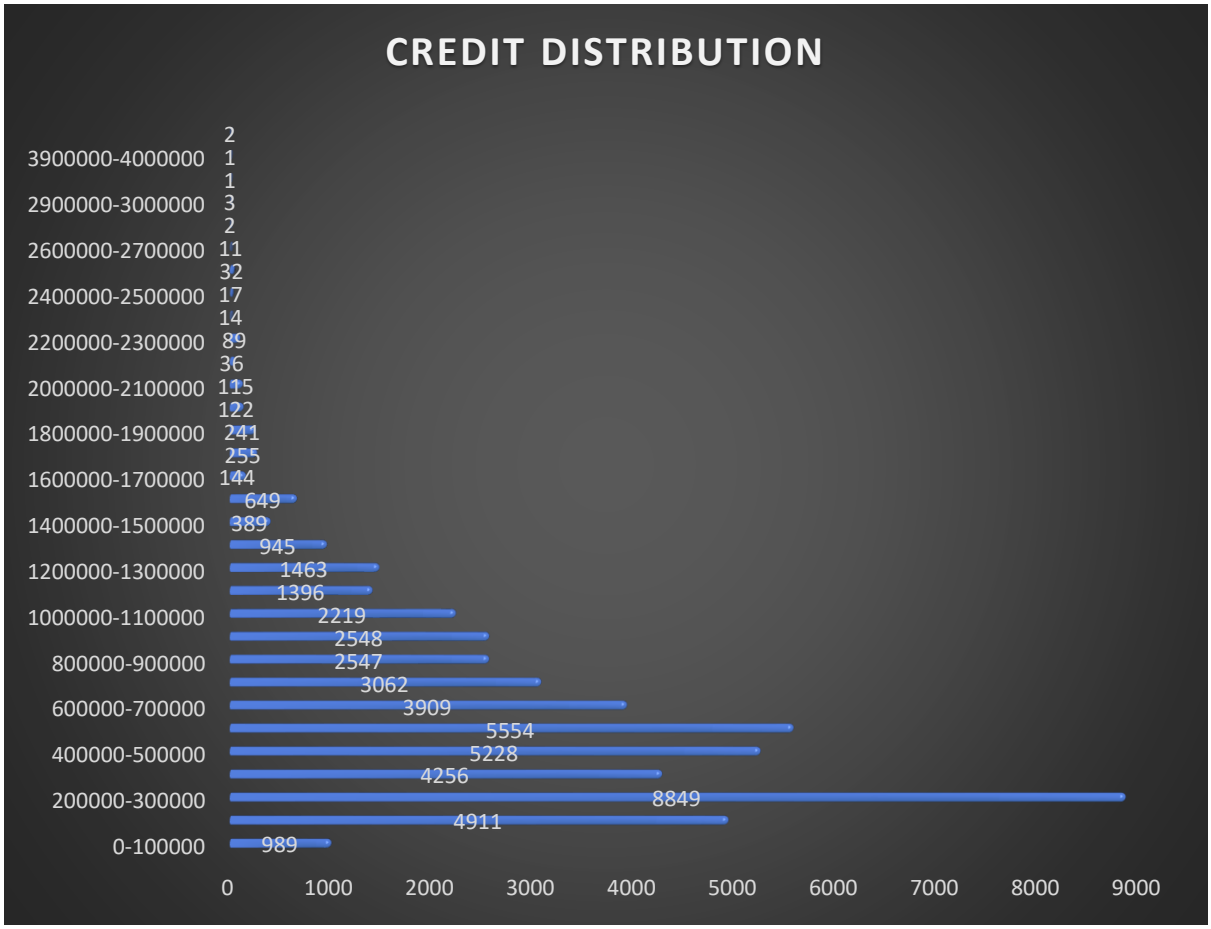


Column1	
Mean	170768
Standard Error	2378
Median	145800
Mode	135000
Standard Deviation	531819.0951
Sample Variance	282831549942
Kurtosis	46582.52582
Skewness	212.0777967
Range	116974350
Minimum	25650
Maximum	117000000
Sum	8538208758
Count	49999

Range	116974350
Bins	1
	116974350
	25650
	117000000

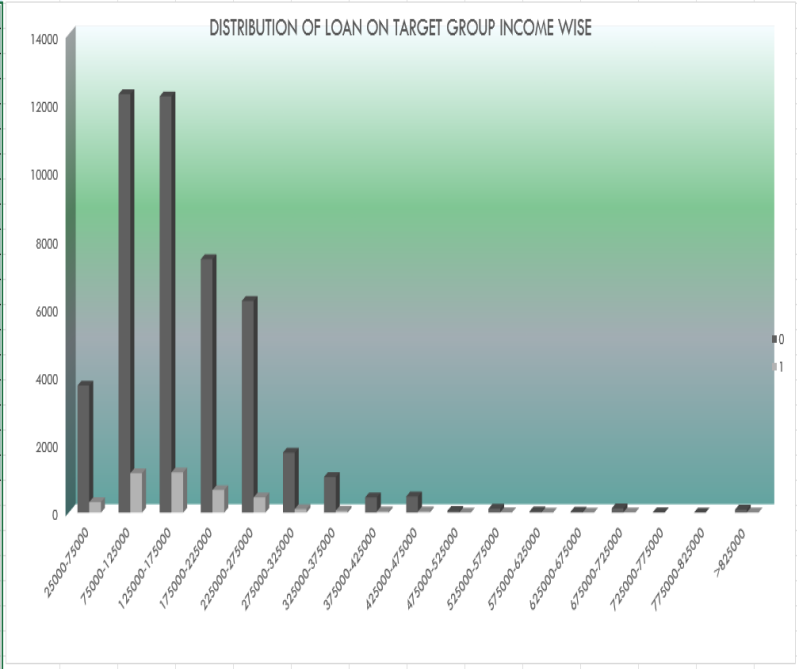
INCOME	Count of AMT_INCOME_TOTAL
25650-90650	9455
90650-155650	15954
155650-220650	13470
220650-285650	6737
285650-350650	2124
350650-415650	1238
415650-480650	522
480650-545650	169
545650-610650	52
610650-675650	143
675650-740650	25
740650-805650	13
805650-870650	24
870650-935650	31
935650-1000650	2
1000650-1065650	1
1065650-1130650	16
1260650-1325650	1
1325650-1390650	10
1520650-1585650	1
1780650-1845650	2
1845650-1910650	1
1910650-1975650	1
1975650-2040650	2
2235650-2300650	2
3535650-3600650	1
3795650-3860650	1
116960650-117025650	1
Grand Total	**49999**





SEGMENT UNIVARIATE ANALYSIS:

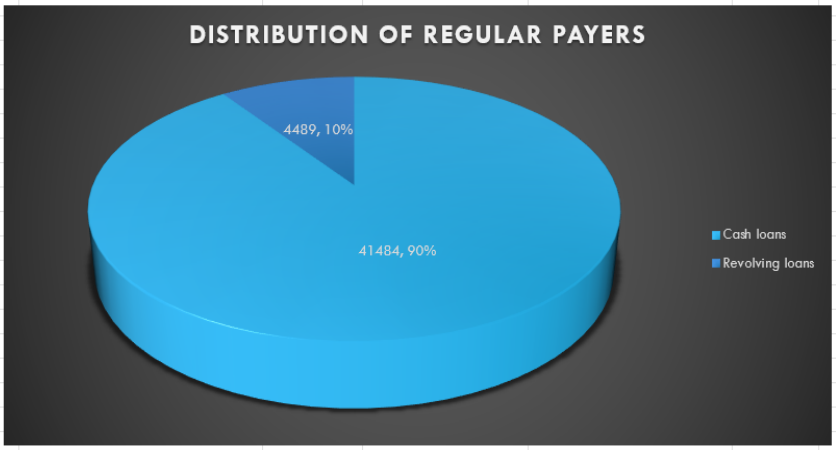
ount of TARGET	TARGET GROUP		
INCOME AMOUNT	0	1	Grand Total
5000-75000	3721	309	4030
5000-125000	12254	1156	13410
25000-175000	12186	1179	13365
75000-225000	7421	661	8082
25000-275000	6198	447	6645
75000-325000	1757	104	1861
25000-375000	1045	58	1103
75000-425000	449	40	489
25000-475000	475	38	513
75000-525000	53	3	56
25000-575000	116	10	126
75000-625000	39	4	43
25000-675000	29	1	30
75000-725000	126	9	135
25000-775000	12		12
75000-825000	2		2
-825000	90	7	97
Grand Total	45973	4026	49999



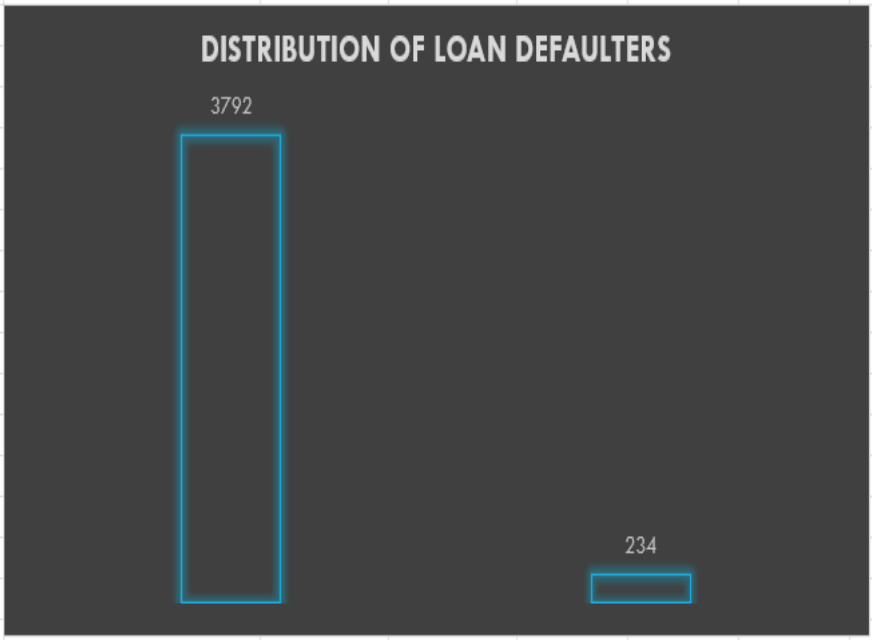
Count of TARGET	LOAN REGULAR PAYERS	
CONTRACY TYPE	0	Grand Total
Cash loans	41484	41484
Revolving loans	4489	4489
Grand Total	**45973**	**45973**

0	LOAN REGU
1	DEFAULTERS

Count of TARGET	LOAN REGULAR PAYERS
CONTRACY TYPE	0
Cash loans	41484
Revolving loans	4489
Grand Total	**45973**



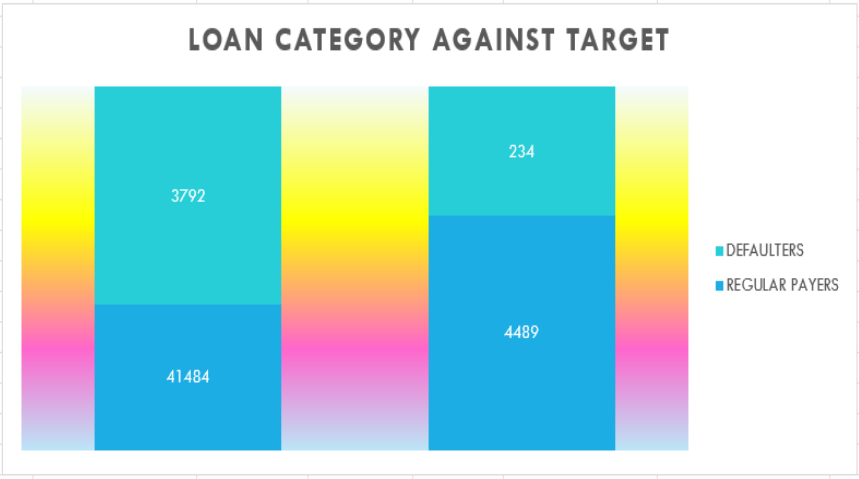
Count of TARGET	LOAN DEFAULTERS	
CONTRACT	1	Grand Total
Cash loans	3792	3792
Revolving loans	234	234
Grand Total	**4026**	**4026**



BIVARIATE ANALYSIS:

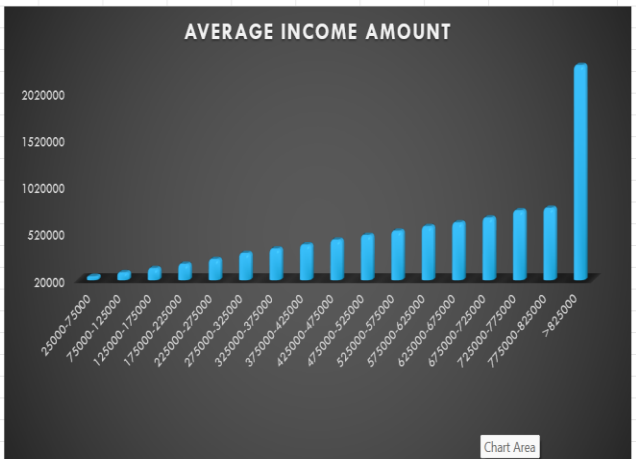
Count of TARGET	TARGET GROUP		
CONTRACT TYPE	REGULAR PAYERS	DEFAULTERS	Grand Total
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	45973	4026	49999

CONTRACT TYPE	REGULAR PAYERS	DEFAULTERS
Cash loans	41484	3792
Revolving loans	4489	234
Grand Total	45973	4026

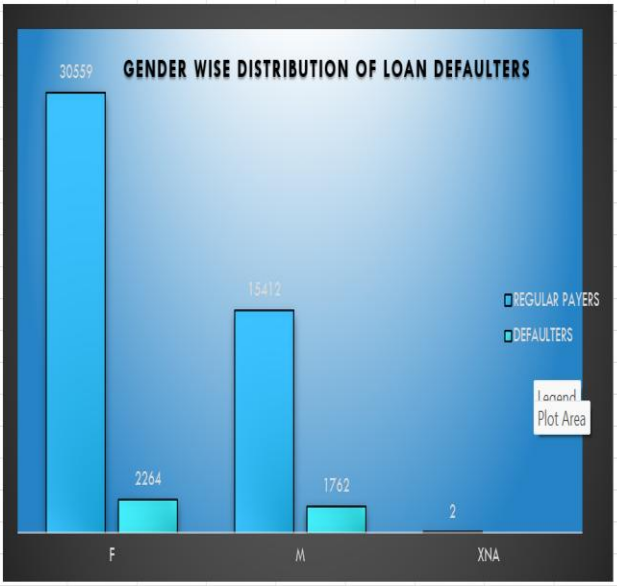


0 REGULAR PAYERS
1 EFAULTERS

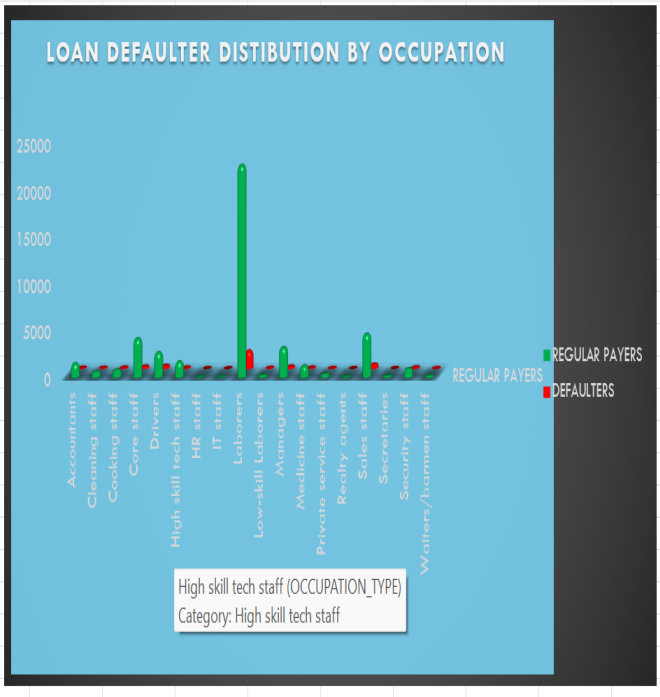
INCOME AMOUNT	Average of AMT_INCOME_TOTAL
25000-75000	60893.22
75000-125000	101124.41
125000-175000	145135.07
175000-225000	190600.10
225000-275000	241976.08
275000-325000	305219.24
325000-375000	352801.16
375000-425000	396784.97
425000-475000	447067.15
475000-525000	496896.43
525000-575000	541946.43
575000-625000	590860.47
625000-675000	630750.00
675000-725000	681650.00
725000-775000	758473.13
775000-825000	785250.00
>825000	2319355.67
Grand Total	170767.59



Count of TARGET	TARGET GROUP		
GENDER	REGULAR PAYERS	DEFAULTERS	Grand Total
F	30559	2264	32823
M	15412	1762	17174
XNA	2		2
Grand Total	45973	4026	49999



Count of TARGET	TARGET GROUP		
OCCUPATION	REGULAR PAYERS	DEFAULTERS	Grand Total
Accountants	1540	81	1621
Cleaning staff	671	68	739
Cooking staff	862	101	963
Core staff	4184	250	4434
Drivers	2706	338	3044
High skill tech staff	1734	118	1852
HR staff	92	9	101
IT staff	76	4	80
Laborers	22660	1946	24606
Low-skill Laborers	296	61	357
Managers	3246	243	3489
Medicine staff	1297	106	1403
Private service staff	410	37	447
Realty agents	110	13	123
Sales staff	4668	492	5160
Secretaries	203	9	212
Security staff	1015	125	1140
Waiters/barmen staff	203	25	228
Grand Total	45973	4026	49999



E. Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

TARGET (0):

CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
1.00	0.04	0.01	0.00	-0.02	-0.34	-0.25	0.03	0.02	-0.04
0.04	1.00	0.38	0.38	0.18	-0.07	-0.16	-0.03	-0.21	0.02
0.01	0.38	1.00	0.99	0.10	0.05	-0.07	0.01	-0.10	-0.04
0.00	0.38	0.99	1.00	0.10	0.05	-0.07	0.01	-0.10	-0.05
-0.02	0.18	0.10	0.10	1.00	0.03	-0.01	0.00	-0.54	0.01
-0.34	-0.07	0.05	0.05	0.03	1.00	0.62	0.27	-0.01	0.06
-0.25	-0.16	-0.07	-0.07	-0.01	0.62	1.00	0.27	0.04	0.05
0.03	-0.03	0.01	0.01	0.00	0.27	0.27	1.00	0.01	0.03
0.02	-0.21	-0.10	-0.10	-0.54	-0.01	0.04	0.01	1.00	0.01
-0.04	0.02	-0.04	-0.05	0.01	0.06	0.05	0.03	0.01	1.00

TARGET (1):

AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	REGION_RATING_CLIENT	AMT_REQ_CREDIT_BUREAU_YEAR
0.01	0.01	0.00	-0.02	-0.25	-0.19	0.04	0.06	-0.04
1.00	0.02	0.01	-0.01	-0.01	-0.01	0.01	-0.01	-0.01
0.02	1.00	0.98	0.07	0.14	0.02	0.04	-0.05	-0.03
0.01	0.98	1.00	0.08	-0.14	0.14	0.02	0.02	-0.05
-0.01	0.07	0.08	1.00	0.02	0.01	0.01	-0.43	0.02
-0.01	0.14	0.14	0.02	1.00	0.59	0.25	-0.05	0.09
-0.01	0.02	0.02	0.01	0.59	1.00	0.23	-0.01	0.02
0.01	0.04	0.05	0.01	0.25	0.23	1.00	-0.03	0.06
-0.01	-0.05	-0.05	-0.43	-0.05	-0.01	-0.03	1.00	0.02
-0.01	-0.03	-0.03	0.02	0.09	0.02	0.06	0.02	1.00

TOP 5 CORRELATION (TAGET 0)- LOAN PAYERS		
VARIABLE 1	VARIABLE 2	CORRELATION
DAYS_EMPLOYED	DAYS_BIRTH	0.62
AMT_CREDIT	AMT_INCOME_TOTAL	0.38
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.27
DAYS_ID_PUBLISH	DAYS_BIRTH	0.27
DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.27

TOP 5 CORRELATION (TARGET 1)- DEFAULTERS		
VARIABLE 1	VARIABLE 2	CORRELATION
AMT_GOODS_PRICE	AMT_CREDIT	0.98
DAYS_BIRTH	DAYS_EMPLOYED	0.59
DAYS_ID_PUBLISH	DAYS_BIRTH	0.25
DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.23
DAYS_BIRTH	AMT_GOODS_PRICE	0.14

RESULTS: The Bank Lona case study was executed starting with data cleaning process as missing data, identifying outliers and other exploratory data analysis then followed by different analysis and then identify the correlation between different variables

Excel link for file: PROJECT-6-BANK LOAN ANALYSIS.xlsx

PPT VIDEO LINK:
https://drive.google.com/file/d/1oyk84T1y7JQfTAi1rjHOdRKqGhRIW3AE/view?usp=sharing