

# **PROJECT 6:**

## **BANK LOAN CASE STUDY:**

### **DESCRIPTION:**

The main aim of this project to apply Exploratory Data Analysis. As a data analyst at finance company that specializes in lending various types of loans to various customers, there are various challenges such as: some customers who don't have a sufficient credit history take advantage of this and default on their loans. EDA plays an important role in analysing patterns in data and ensure that capable applicants are not rejected.

### **APPROACH:**

#### ❖ DATA GATHERING:

I downloaded the data provided and Imported that data in the Excel workbook.





#### ❖ DATA CLEANING:

The next step I followed is data cleaning. There were many blank cells in the file which needed to be removed to get the accurate output. Also, there were cells with not required data which needed to be removed.

#### ❖ INSIGHTS:

The final step is to perform various operations including sorting, Mean, Median, Mode, Variance, etc. to derive outputs so that we can gather insights from the data.

### **Possibilities while applying for a loan:**

-  **Approved:** The company has approved the loan application.
-  **Rejected:** The company rejected the loan application.
-  **Cancelled:** The applicant cancelled the application mid-way
-  **Unused:** The loan was approved but not availed by the customer.

## Risks involved

Applicant can repay the loan but it is not approved.

Applicant can't repay the loan but it is approved.

## TECH - STACK USED:

The Tech-stack used to complete this project is MS-EXCEL, 2013. The purpose of using MS-Excel is that it provides features like PIVOT TABLE, operations like MAX, MIN, SUM, COUNTIF etc. using which we can derive insights easily and data visualization is also possible using excel.

## DATA ANALYTICS TASKS:

**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.
- ❖ By applying COUNTA() calculated the total values in a column.
- ❖ By applying formula for calculating null values (1-cell/target column) ( $=1-A2/(\$B\$2)$ ), calculated null values.
- ❖ Deleted all the columns with null values greater than 30%.

FILEHOMEINSERTPAGE LAYOUTFORMULASDATAREVIEWVIEWPOWERPivot

ClipboardFontAlignmentNumber

Conditional FormattingStylesCell StylesCellsEditing

AW1: =1-AW2/(\$B\$2)

	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	
1	0.00%	56.35%	0.25%	19.89%	50.77%	58.40%	48.79%	66.48%	69.92%	53.30%	
2	49999	21827	49873	40055	24614	20800	25605	16760	15039	23348	
3	ORGANIZATION_TYPE	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATAT	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AVG	ENTR
4	Business Entity Type 3	0.083036967	0.262948593	0.13937578	0.0247	0.0369	0.9722	0.6192	0.0143	0	
5	School	0.311267311	0.622245775		0.0959	0.0529	0.9851	0.796	0.0605	0.08	
6	Government		0.555912083	0.729566691							
7	Business Entity Type 3		0.65044169								
8	Religion		0.322738287								
9	Other		0.354224732	0.621226338							
10	Business Entity Type 3	0.774761413	0.72399852	0.492060094							
11	Other		0.714279286	0.54065445							
12	XNA	0.587334047	0.205747288	0.751723715							
13	Electricity		0.746643629								
14	Medicine	0.319760172	0.651862333	0.363945239							
15	XNA	0.72204445	0.555183162	0.652896552							
16	Business Entity Type 2	0.464831117	0.715041819	0.176652579	0.0825		0.9811			0	
17	Self-employed		0.566906613	0.77008707	0.1474	0.0973	0.9806	0.7348	0.0582	0.16	
18	Transport: type 2	0.721939769	0.642656205		0.3495	0.1335	0.9985	0.9796	0.1143	0.4	
19	Business Entity Type 2	0.115634337	0.346633981	0.678567689							
20	Government		0.23637784	0.062103038							
21	Construction		0.683513346								
22	Housing		0.706428403	0.556727426	0.0278	0.0617	0.9881	0.8368	0.0018	0	
23	Kindergarten		0.58661714	0.477649155							

EDAFinal\_application\_dataamt\_income\_total\_outlierOutlier for Days\_employedoutlier\_cnt ...

READY

Type here to search

GBP/INR -0.33%

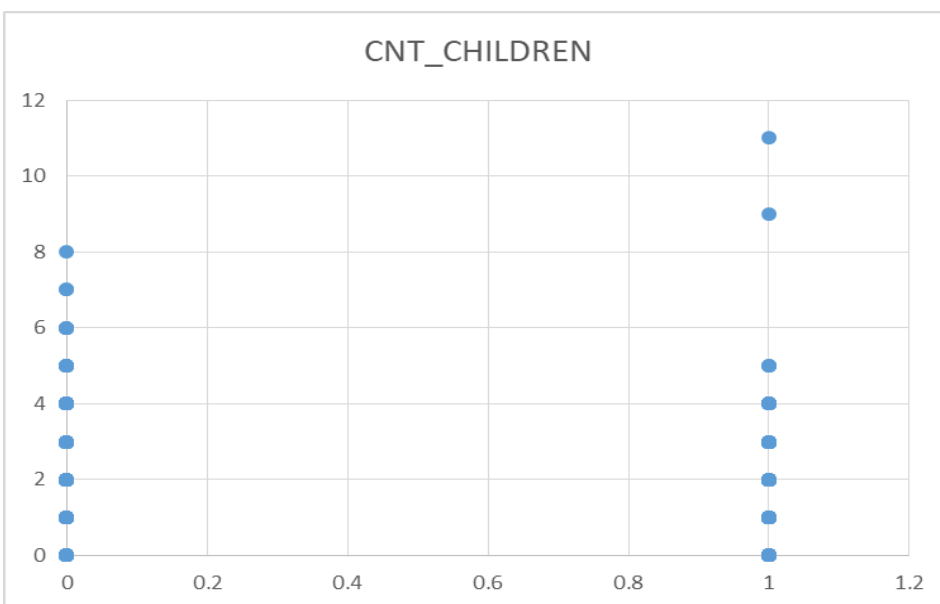
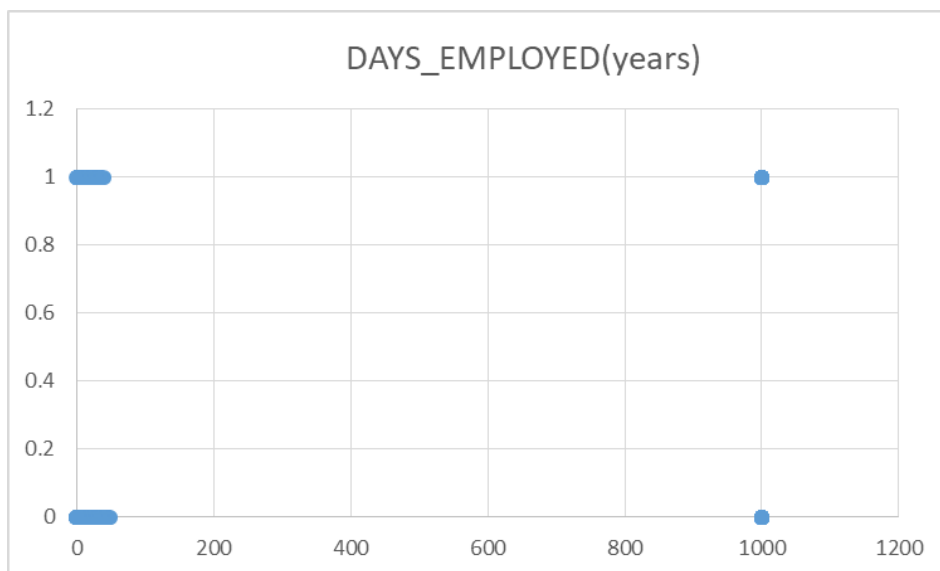
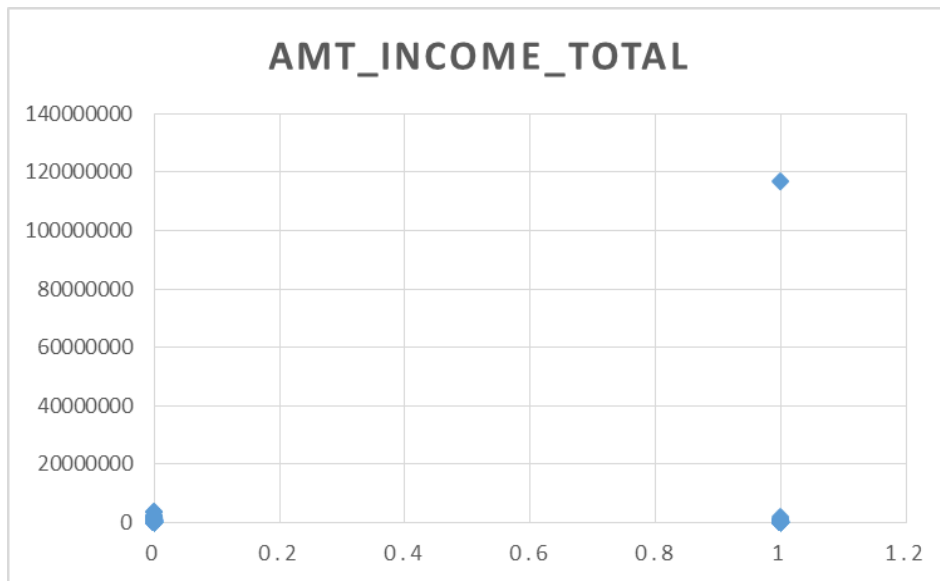
02-09-2024

**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.
  - Use =QUARTILE (Range,1) to find Q1
  - Use =QUARTILE(Range,3) to find Q3
  - Find Interquartile range(IQR) (=Q3-Q1)
  - Find upper limit of threshold range (=Q3+1.5\*IQR)
  - Find lower limit of threshold range (=Q1-1.5\*IQR)

**Insights:**

**The Outliers can be depicted by following graphs:**

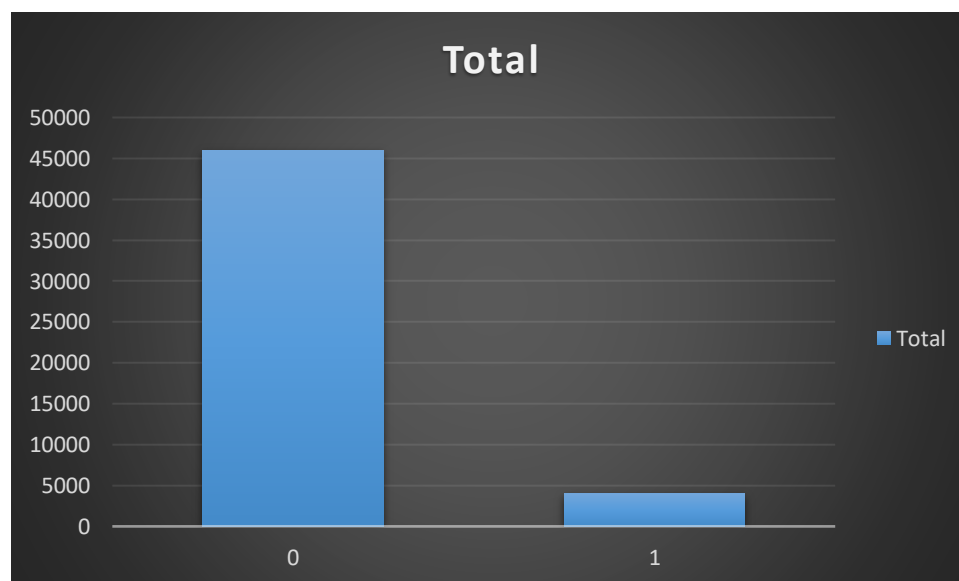


**C. Analyse Data Imbalance:** Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.

- **Task:** Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
  - Used Pivot Table to count number of 0s and 1s in the “Target” column.
  - `=GETPIVOTDATA("TARGET",$A$3,"TARGET",0)/GETPIVOTDATA("TARGET",$A$3,"TARGET",1)` To calculate Ratio of 0s and 1s

### INSIGHTS:

Ratio of 0 and 1	Contribution	
11.41903	0	92%
	1	8%

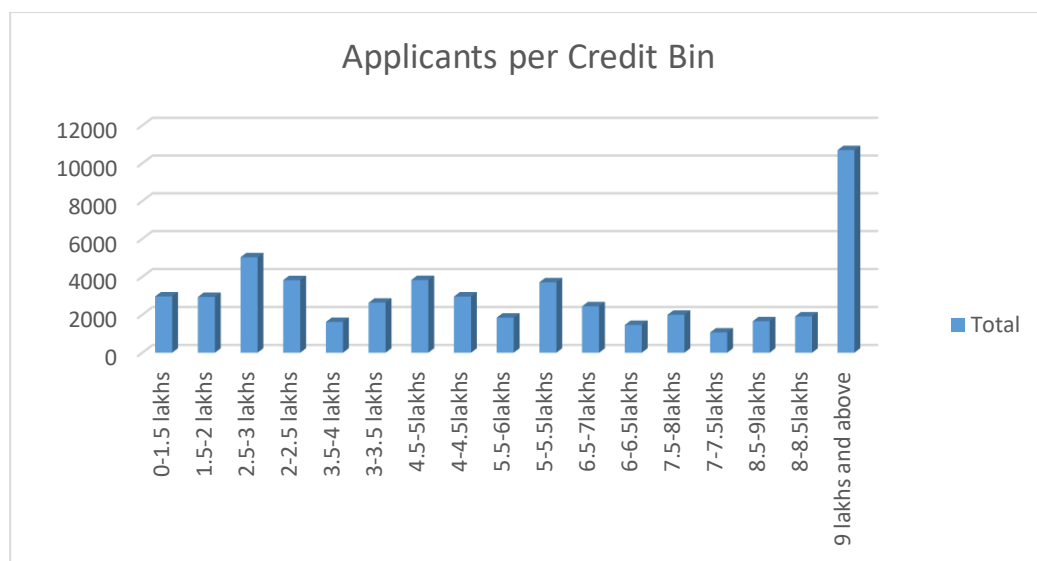


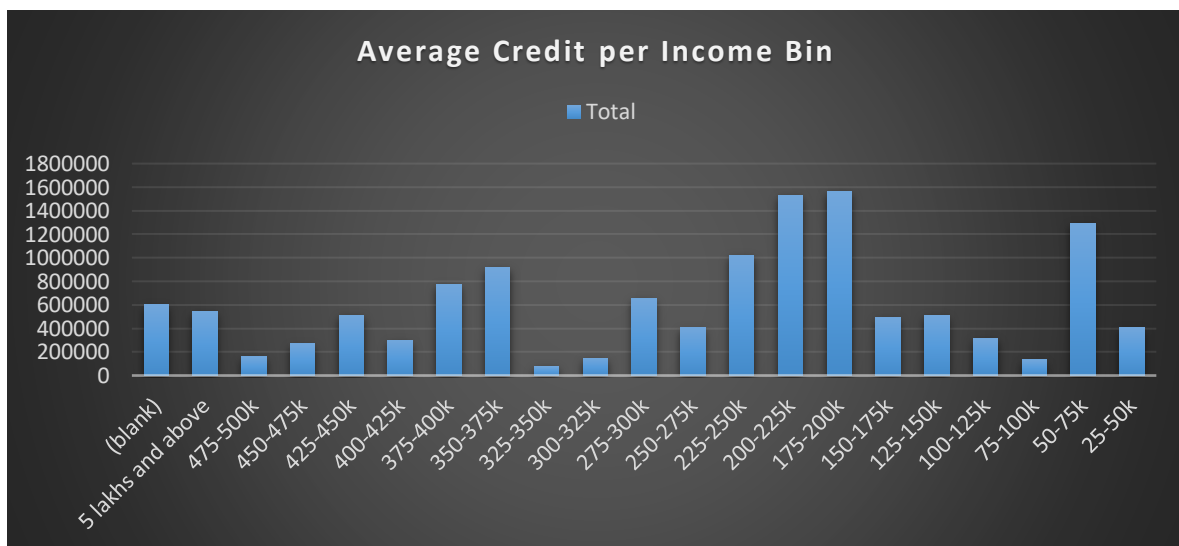
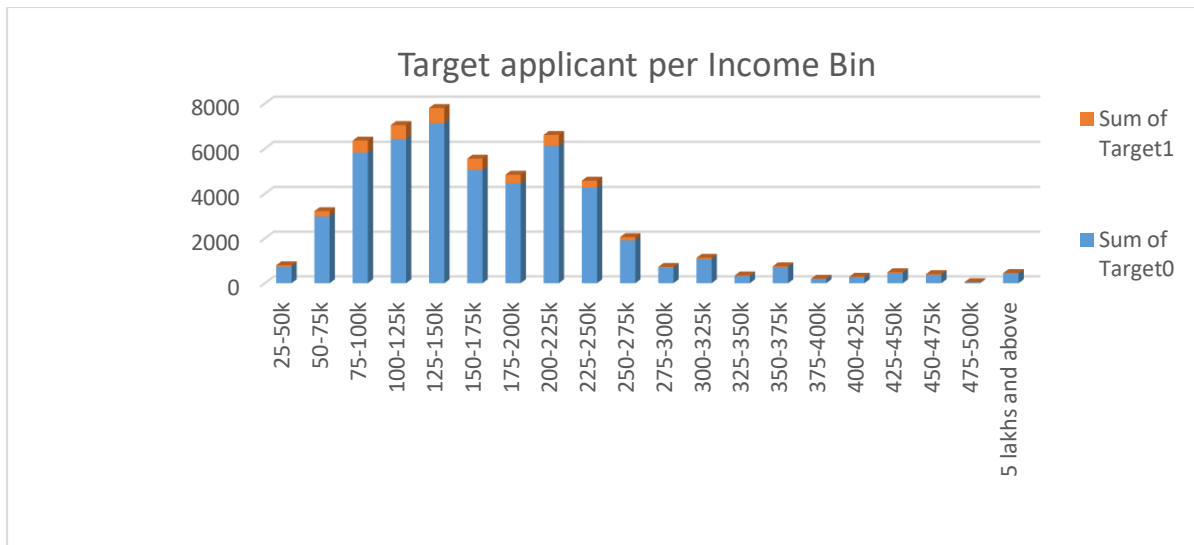
The customers with payment difficulties is far less than customers with no payment difficulties.

**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.
- Calculated the maximum and minimum values of Income\_Total by using =MAX() and MIN() functions
  - Created Income bins for target 0 and 1 for segmented analysis.
  - Calculated the maximum and minimum values of Income\_credits by using =MAX() and MIN() functions
  - Created Credit bins for applicants for univariate analysis.
  - Using pivot table, created Income bins and average of Income credit table for Bivariate analysis.

## INSIGHTS:





**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.
  - Separated the data with 0s( no payment difficulties) and 1s( payment difficulties) by using “sorting” function of MS Excel.
  - Used CORREL() function to find correlation between different entities.
  - Created Heat maps to depict the correlations

**INSIGHTS:**

bank loan case study - Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I
1	CORRELATION For APPLICANTS WHO MADE PAYMENT ON TIME								
2									
3	Cnt_Children	1	0.036319722	0.005705458	-0.024912809	-0.335876269	-0.245521512	0.032537221	0.021288992
4	Amt_Income_Total	0.036319722	1	0.377965752	0.181941261	-0.073769425	-0.161680938	-0.032286356	-0.205031899
5	Amt_credit	0.005705458	0.377965752	1	0.095539444	0.051084182	-0.074733443	0.008290189	-0.102556478
6	Region_population_relative	-0.024912809	0.181941261	0.095539444	1	0.030435419	-0.006767142	0.002236288	-0.539333113
7	Days_birth(years)	-0.335876269	-0.073769425	0.051084182	0.030435419	1	0.623474675	0.270073313	-0.00902485
8	Days_Employed(years)	-0.245521512	-0.161680938	-0.074733443	-0.006767142	0.623474675	1	0.274516224	0.040937165
9	Days_ID_published(years)	0.032537221	-0.032286356	0.008290189	0.002236288	0.270073313	0.274516224	1	0.008097427
10	Region_Rating_client	0.021288992	-0.205031899	-0.102556478	-0.539333113	-0.00902485	0.040937165	0.008097427	1
11		Cnt_Children	Amt_Income_Total	Amt_credit	Region_population_relative	Days_birth(years)	Days_Employed(years)	Days_ID_published(years)	Region_Rating_client
12									
13									

bank loan case study - Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I	J
1	CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES									
2										
3	Cnt_Children	1	0.010110177	0.00760191	-0.020359154	-0.2496732	-0.189773227	0.042360717	0.055515557	
4	Amt_Income_Total	0.010110177	1	0.01527144	-0.006180303	-0.009033662	-0.011758681	0.009122006	-0.012846697	
5	Amt_credit	0.007601905	0.015271444	1	0.067775624	0.142506035	0.018782223	0.043771901	-0.045024534	
6	Region_population_relative	-0.020359154	-0.006180303	0.06777562	1	0.016468731	0.007710059	0.005118563	-0.430032303	
7	Days_birth(years)	-0.2496732	-0.009033662	0.14250603	0.016468731	1	0.588242824	0.247896571	-0.045027112	
8	Days_Employed(years)	-0.189773227	-0.011758681	0.01878222	0.007710059	0.588242824	1	0.232661912	-0.009237108	
9	Days_ID_published(years)	0.042360717	0.009122006	0.0437719	0.005118563	0.247896571	0.232661912	1	-0.025335227	
10	Region_Rating_client	0.055515557	-0.012846697	-0.0450245	-0.430032303	-0.045027112	-0.009237108	-0.025335227	1	
11		Cnt_Children	Amt_Income_Total	Amt_credit	Region_population_relative	Days_birth(years)	Days_Employed(years)	Days_ID_published(years)	Region_Rating_client	
12										
13										

The positive numbers indicate positive correlation between entities; that is with increase in value of one entity, the value of another entity will increase. Whereas, the negative numbers indicate negative correlation between entities; that is with increase in value of one entity, the value of another entity will decrease.

[EXCEL LINK](#)