

TRINITY

# BANK LOAN CASE STUDY

MOGULAGANI PRASHANTH

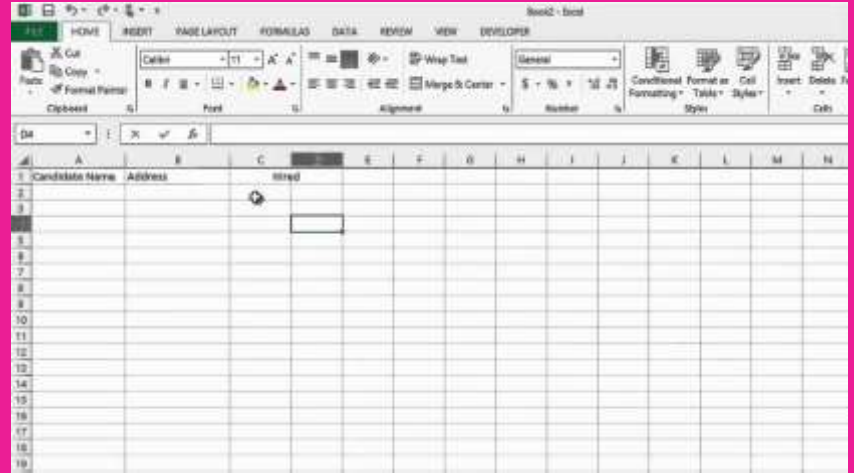
# PROJECT DESCRIPTION

*Consider yourself a data analyst for a financial institution that specialises in providing different types of loans to urban clients. A problem exists for your business because some clients who don't have enough credit history take advantage of this and default on their loans. It is your responsibility to use exploratory data analysis (EDA) to examine data patterns and make sure that qualified candidates are not turned away.*



# APPROACH

1. *We are using MS excel to solve the problems.*
2. *Microsoft Excel is an application developed by Microsoft for Windows, macOS, Android, and iOS*
3. *We use MS excel formulas to analyze the solutions.*



## A. Identifv Missing Data and Deal with it Appropriately

- **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.

### ***FORMULA USED:***

BELOW FORMULA IS FOR REPLACING THE BLANKS WITH THE AVERAGE VALUES OF THE RESPECTIVE COLUMN.

```
=IF(ISBLANK(BZ3), AVERAGE(BZ$3:BZ$50001), BZ3)
```

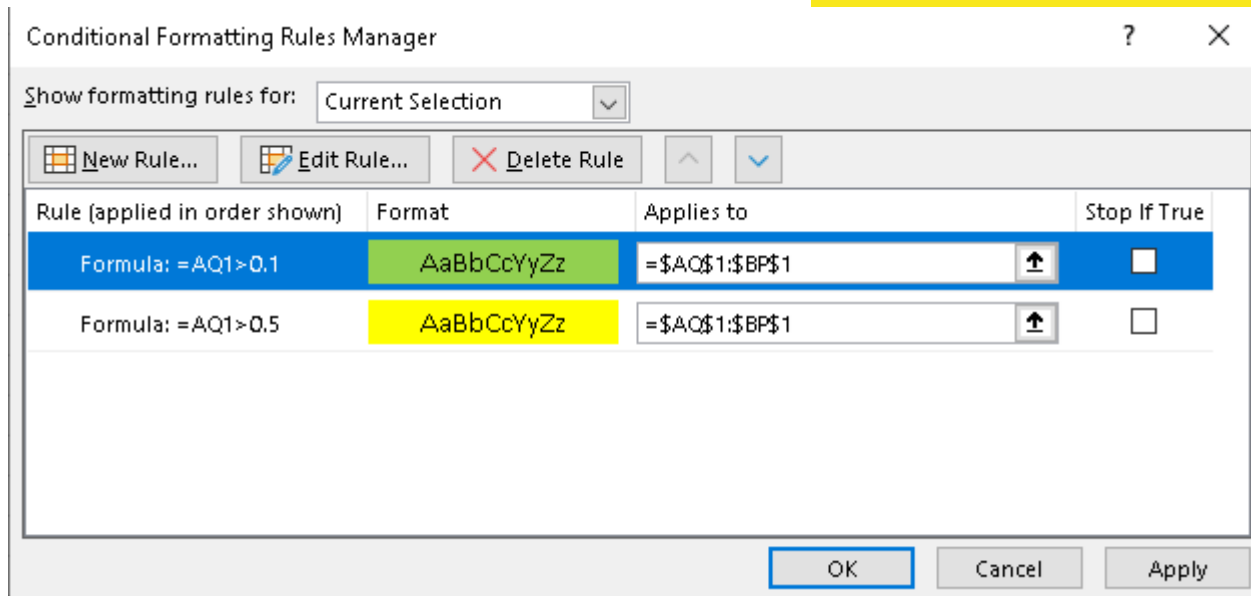
### **FORMULAS USED :**

1. **THE BELOW FORMULA IS USED TO FIND OUT NULL VALUES OR MISSING VALUES FROM THE GIVEN DATA SET AND TO GET PERCENTAGE OF MISSING OR NULL VALUES.**

```
=COUNTBLANK(A3:A50001) / COUNTA(A3:A50001)
```

---

**HERE WE USED CONDITIONAL FORMATTING TO FIND OUT BLANKS MORE THAN 50 PERCENT AND REMOVE THEM.**



## B. IDENTIFY OUTLIERS IN THE DATASET

- **Task:** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- TO FIND OUT THE OUTLIERS FIRST YOU NEED TO FIND OUT THE QUARTILES , INTERQUARTILE RANGE TO FIND OUT THE UPPER BOUND AND LOWER BOUND IN THE GIVEN DATASET.
- VALUES GREATER OR LOWER THAN RESPECTIVE UPPER AND LOWER BOUNDS ARE CALLED AS OUTLIERS.

### FORMULAS USED :

#### FIRST QUARTILE:

```
=QUARTILE.INC(B$7:B$50005,1)
```

#### THIRD QUARTILE:

```
=QUARTILE.INC(B$7:B$50004,3)
```

#### IQR:

```
=B$4-B$5
```

#### UPPER BOUND:

```
=B$4+1.5*B$3
```

#### LOWER BOUND:

```
=B$5-1.5*B$3
```

HERE CONDITIONAL FORMATTING IS USED TO IDENTIFY THE OUTLIERS AS SHOWN TOWARDS RIGHT:

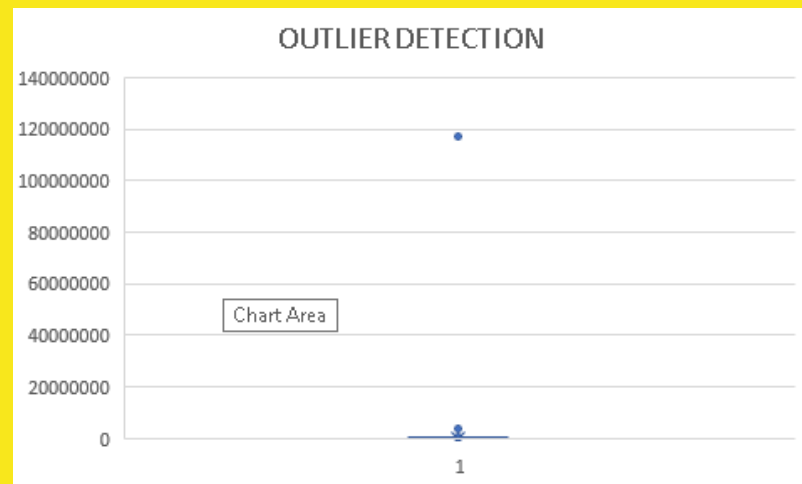
DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE
-637	-3648	-2120	1	1	0
-1398	-1186	-291	1	1	0
-225	-4260	-2531	1	1	0
-3039	-9893	-2487	1	1	0
-3038	-4311	-3458	1	1	0
-1598	-4970	-477	1	1	0
-3130	-1213	-619	1	1	0
-448	-4597	-2379	1	1	0
805141	-7427	-3514	1	1	0
-2019	-14437	-3992	1	1	0
-679	-4427	-798	1	1	0
805141	-5246	-2512	1	1	0
-3717	-311	-9227	1	1	0
-3028	-643	-4511	1	1	0
-203	-415	-2956	1	1	0
-1157	-3494	-1368	1	1	0
-1817	-6392	-3866	1	1	0
-191	-4143	-2427	1	1	0
-7004	-8751	-1259	1	1	0
-2038	-1021	-3964	1	1	0
-4286	-199	-1800	1	1	0

ABOVE IS THE VISUAL IDENTIFICATION OF OUTLIERS COLOURED IN ORANGE IN THE GIVEN DATASET.

Conditional Formatting Rules Manager

Show formatting rules for: Current Selection

Rule (applied in order shown)	Format	Applies to	Stop If True
Cell Value not between ...	AaBbCcYyZz	=\$BM\$7:\$BM\$50005	<input type="checkbox"/>
Cell Value not between ...	AaBbCcYyZz	=\$BO\$7:\$BO\$50005	<input type="checkbox"/>
Cell Value not between ...	AaBbCcYyZz	=\$BQ\$7:\$BQ\$50005	<input type="checkbox"/>
Cell Value not between ...	AaBbCcYyZz	=\$VW\$7:\$VW\$50005	<input type="checkbox"/>
Cell Value not between ...	AaBbCcYyZz	=\$U\$7:\$U\$50005	<input type="checkbox"/>



## C. ANALYZE DATA IMBALANCE:

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

HERE DATA IMBALANCE IS CALCULATED OR ANALYZED FOR TARGET VARIABLE.

BELOW IS THE FORMULA TO FIND THE PERCENTAGE RATIO OF IMBALANCE:

```
=(C$2/SUM(C$2:D$2))*100
```

### FORMULAS USED:

BELOW FORMULA IS TO FIND THE UNIQUES VALUES IN THE TARGET VARIABLE.

```
=UNIQUE(A$2:A$5000)
```

#### COUNT OF MINORITY

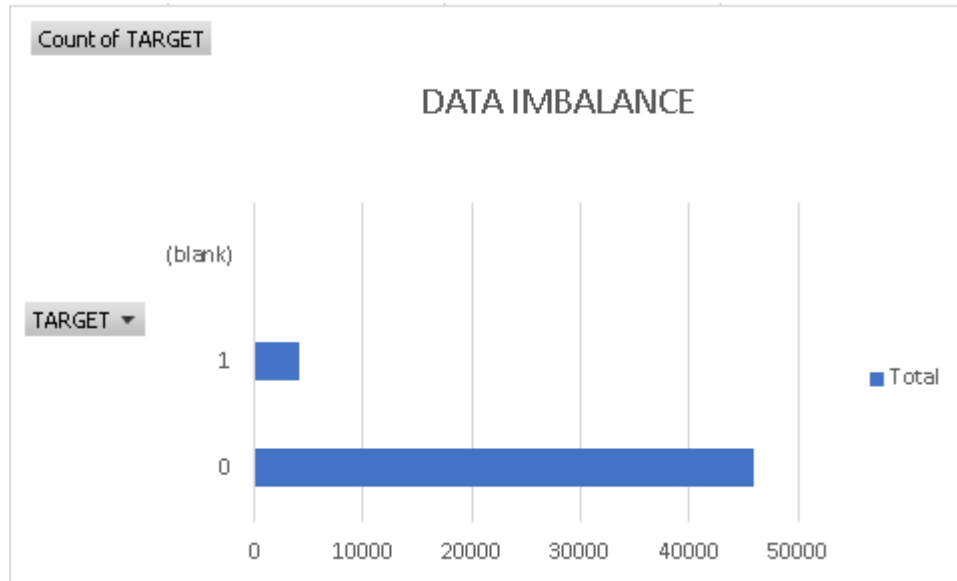
```
=COUNTIF(A$2:A$5000, "1")
```

#### COUNT OF MAJORITY

```
=COUNTIF(A$2:A$5000, "0")
```



**BELOW IS THE BAR CHART THAT TALKS ABOUT THE DATA IMBALANCE IN OUR TARGET VARIABLE:**



## D.PERFORM UNIVARIATE,SEGMENTED UNIVARIATE AND BIVARIATE ANALYSIS:

- **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.

### 1.UNIVARIATE ANALYSIS:

HERE UNIVARIATE ANALYSIS IS PERFORMED ON AMT\_CREDIT VARIABLE FROM THE GIVEN DATASET.

#### DESCRIPTIVE STATISTICS FOR AMT\_CREDIT VARIABLE:

AVERAGE/MEAN	MEDIAN	MODE	MAX	MIN	COUNT
597519.1377	508495.5	450000	2517300	45000	4999

## FORMULAS USED:

=AVERAGE(A\$2:A\$5000)

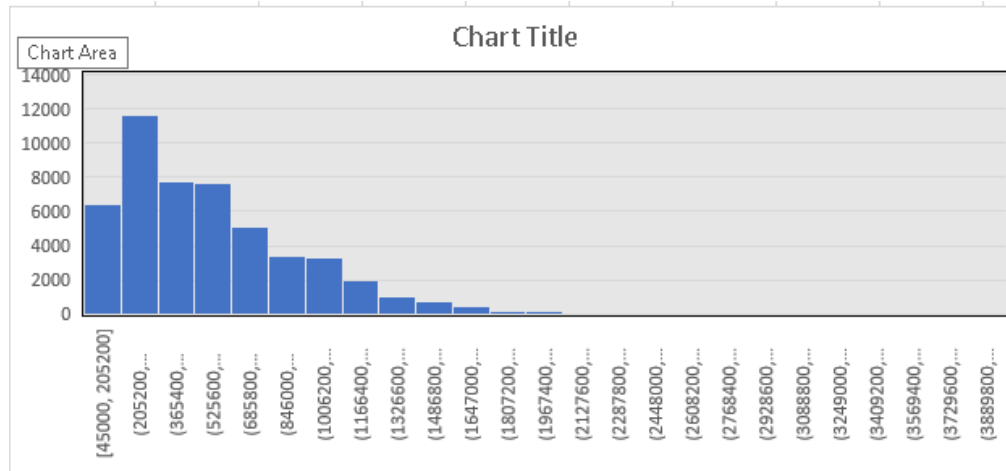
=MEDIAN(A\$2:A\$5000)

{=MODE.MULT(A\$2:A\$5000)}

=MAX(A\$2:A\$5000)

=MIN(A\$2:A\$5000)

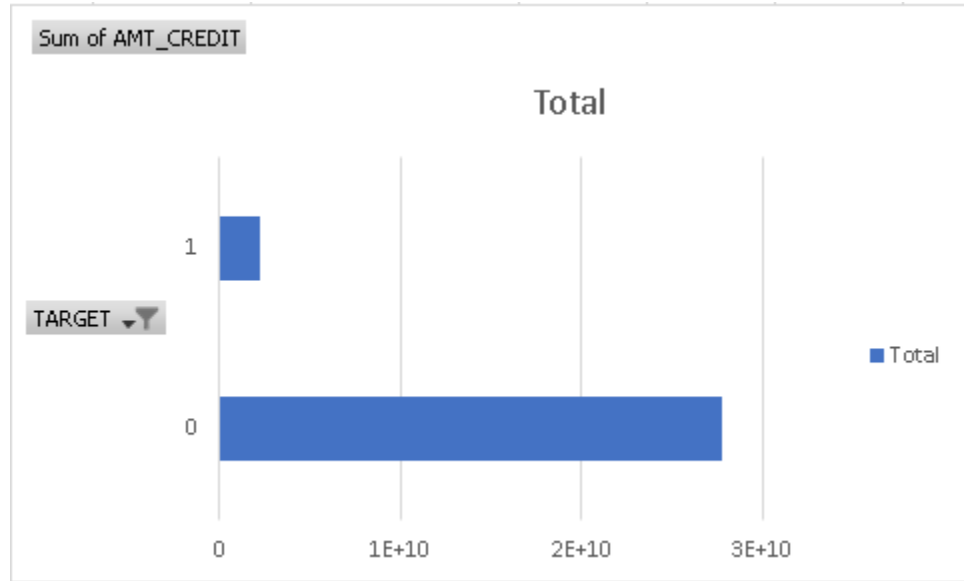
=COUNT(A\$2:A\$5000)



This histogram provides a visual representation of the distribution of AMT\_CREDIT, showing the frequency or count of data points in different bins or intervals of credit amount of the loan.

## 2.SEGMENTED UNIVARIATE ANALYSIS:

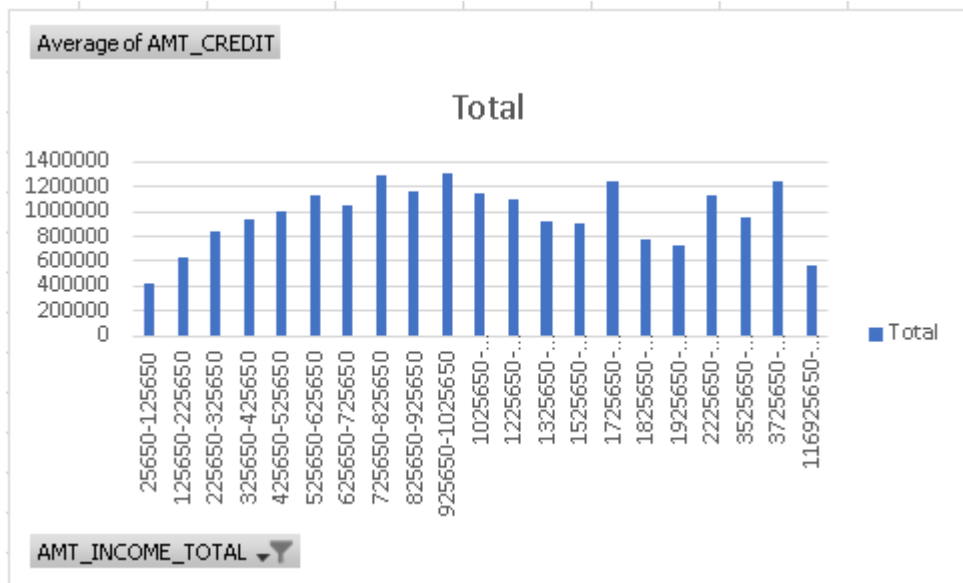
TARGET	Sum of AMT_CREDIT
0	27747569597
1	2236859780
Grand Total	29984429376



HERE THE ABOVE ANALYSIS IS DONE USING PIVOT TABLE AND BAR CHART IS BEING USED FOR VISUALIZATION.

### 3.BIVARIATE ANALYSIS:

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
<b>Grand Total</b>	<b>599700.5815</b>



THE ABOVE ANALYSIS IS DONE BY USING PIVOT TABLE BY PULLING 2 VARIABLES INTO VALUES AND ROWS FIELD IN THE PIVOT TABLE FIELDS AND COLUMN CHART IS USED VISUALIZE THE DATA.

# E.IDENTIFY TOP CORRELATIONS FOR DIFFERENT SCENARIOS:

- **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.

HERE WE ARE SEGMENTING THE TARGET VARIABLE INTO 2 THAT IS : TARGET - 0 , TARGET - 1.

FOR TARGET 1:

FORMULA USED:

```
=CORREL(D$2:D$49988,B$2:B$49988)
```

## DATASET:

TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	CNT_CHILDREN
1	202500	406597.5	0
1	112500	979992	0
1	202500	1193580	0
1	135000	288873	0
1	81000	252000	0
1	315000	953460	0
1	157500	723996	1
1	292500	675000	0
1	157500	245619	0
1	111915	225000	0
1	180000	540000	3
1	202500	436032	1
1	135000	495216	0
1	157500	1710000	0
1	73341	135000	0
1	121500	263686.5	1

## OUTPUT:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
CNT_CHILDREN	1	0.00960733	0.005042552
AMT_INCOME_TOTAL	0.00960733	1	0.069319162
AMT_CREDIT	0.005042552	0.069319162	1

New Formatting Rule ? X




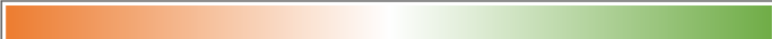
Select a Rule Type:

- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

**Format all cells based on their values:**

Format Style: 3-Color Scale

	Minimum	Midpoint	Maximum
Type:	Lowest Value	Percentile	Highest Value
Value:	(Lowest value)	50	(Highest value)
Color:			
Preview:			

HERE CONDITIONAL FORMATTING IS USED TO DIFFERENTIATE BETWEEN LOWEST , MIDPOINT AND HIGHEST VALUES USING TRI-COLORS



FOR TARGET 0:

DATASET:

TARGET	AMT_INCOME_TOTAL	AMT_CREDIT	CNT_CHILDREN
0	270000	1293502.5	0
0	67500	135000	0
0	135000	312682.5	0
0	121500	513000	0
0	99000	490495.5	0
0	171000	1560726	1
0	360000	1530000	0
0	112500	1019610	0
0	135000	405000	0
0	112500	652500	1
0	38419.155	148365	0
0	67500	80865	0
0	225000	918468	1
0	189000	773680.5	0
0	157500	299772	0
0	108000	509602.5	0
0	81000	270000	1

OUTPUT:

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
CNT_CHILDREN	1	0.00960733	0.005042552
AMT_INCOME_TOTAL	0.00960733	1	0.069319162
AMT_CREDIT	0.005042552	0.069319162	1

New Formatting Rule ? X





Select a Rule Type:

- Format all cells based on their values
- Format only cells that contain
- Format only top or bottom ranked values
- Format only values that are above or below average
- Format only unique or duplicate values
- Use a formula to determine which cells to format

Edit the Rule Description:

**Format all cells based on their values:**

Format Style: 3-Color Scale ▼

	Minimum	Midpoint	Maximum
Type:	Lowest Value ▼	Percentile ▼	Highest Value ▼
Value:	(Lowest value) ↑	50 ↑	(Highest value) ↑
Color:	 ▼	 ▼	 ▼
Preview:			

OK Cancel

HERE CONDITIONAL FORMATTING IS USED TO DIFFERENTIATE BETWEEN LOWEST , MIDPOINT AND HIGHEST VALUES USING TRI-COLORS

# RESULT

*I became acquainted with new EXCEL features, lingo, and methods.*

*By obtaining the appropriate insights from the problem description, practical problems can be resolved. Thanks to the concepts, I was able to comprehend the description of the problem. This project has improved my problem-solving skills and taught me how to apply the theoretical concepts I learned in training to actual-world circumstances.*

## **LINK OF EXCEL FILE :**

[https://docs.google.com/spreadsheets/d/1DUH2ds\\_jyipUgcc12XI0tzXog8IWNEle/edit?usp=sharing&ouid=108547673521600619650&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1DUH2ds_jyipUgcc12XI0tzXog8IWNEle/edit?usp=sharing&ouid=108547673521600619650&rtpof=true&sd=true)

