BANK LOAN

CASE STUDY

ASHIQ PAUL

PROJECT DESCRIPTION

Bank Loan Case Study

This project focused on understanding what drives some loan applicants to default, especially those with limited credit history. Using Excel, I explored the loan dataset to find patterns in customer and loan features. The goal was to help the company make smarter lending decisions, like knowing when to approve, reject, or adjust loan terms, based on real data.

DATASET

Link to the Dataset (106MB)

APPROACH

The Approach taken for this project

Data Cleaning
Cleaned the data by identifying and handling missing values

Outlier Detection
Found and reviewed unusually high or low values

Class Imbalance Check
Looked at how many clients defaulted vs who didn't

Univariate Analysis
Explored each column to understand its distribution

Segmented & Bivariate Analysis
Compared groups and checked how features relate to defaulting.

Correlation Analysis
Found which features are most linked to defaults.

TECH STACK

The Tech-Stacks Used

Microsoft Excel 2025

Used for the cleaning of data, pivot table, calculations and chart visualizations

Microsoft Word / Adobe Acrobat / Canva

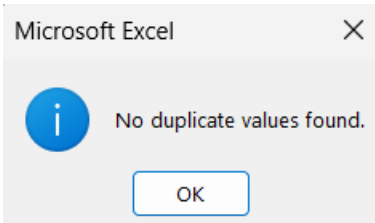
Microsoft Word Used for preparing and structuring the report and converting to PDF. Canva used for the title page.

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.

There are 50,000 rows and 122 columns in the dataset.

First, we check for any duplicate values.



There were no duplicates so we can move on.

I have converted the values with Days into Years, so I used the ABS (Absolute Value) formula and divide the days with 365.

```
=ABS([@[DAYS_BIRTH]]/365)
```

We can also round it with the ROUND formula

```
=ROUND(ABS([@[DAYS_BIRTH]]/365),0)
```

I have highlighted those columns with Light Blue colour.

There are also way too many blanks present in the dataset.

I used the COUNTBLANK formula to find the blanks present in each of the columns

```
=COUNTBLANK(B4:B50002)
```

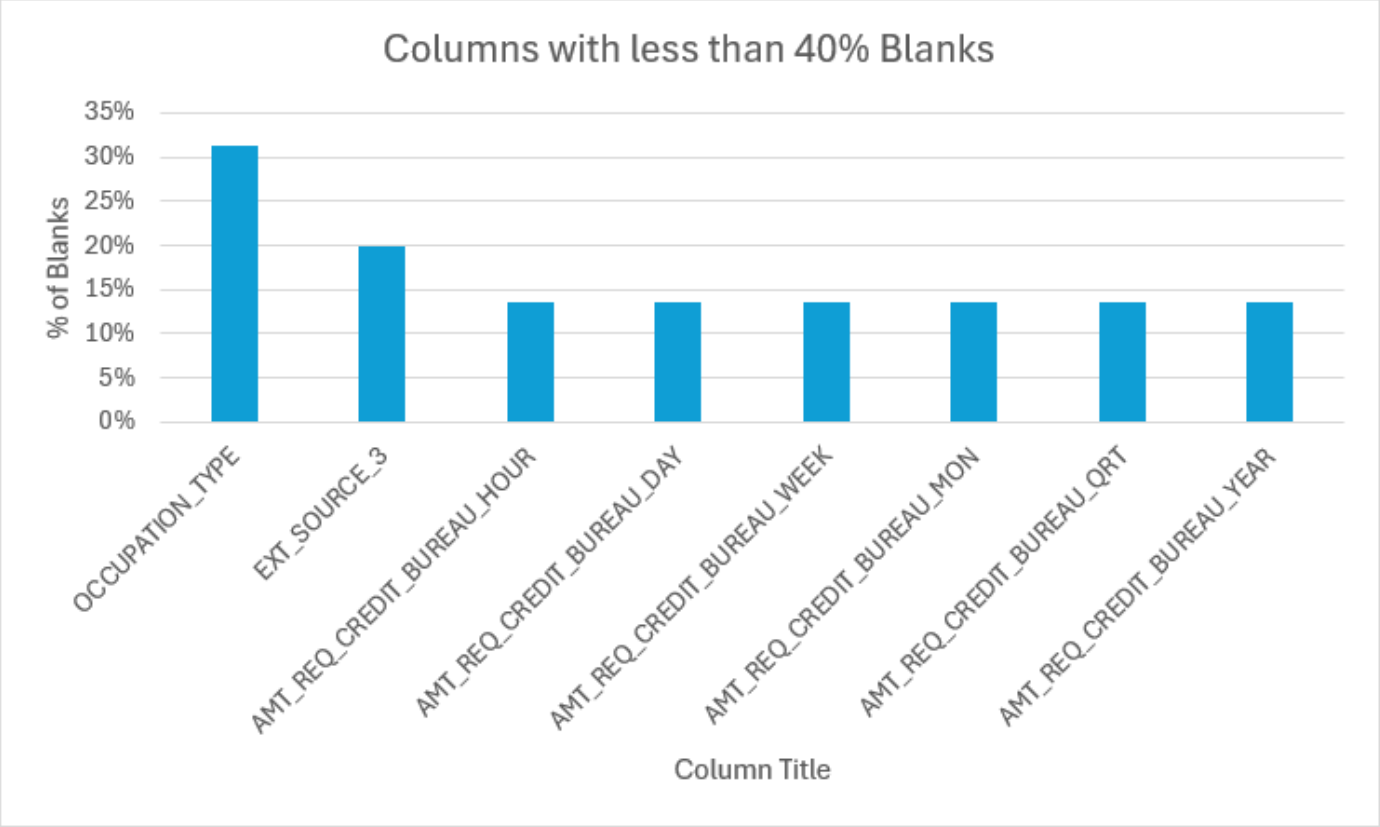
Then found the percentage of blanks present with this formula

```
=(B2/COUNT($B$4:$B$50002))
```

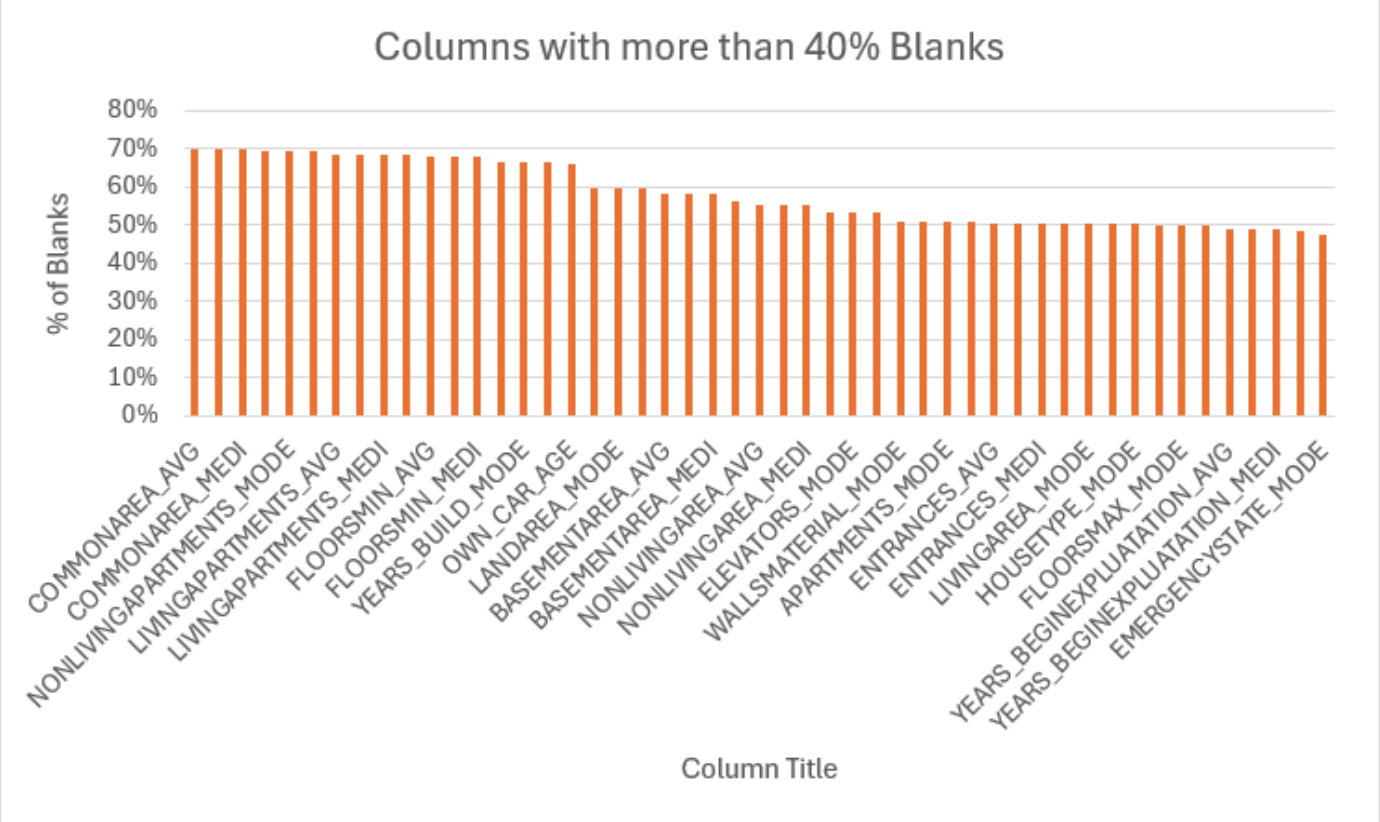
and then formatted as percentage using % in Excel.

The percentage helps us determine the steps we need to take, that is either to add in missing values, leave it as “Unknown” or drop the columns.

SITUATION	ACTION
<40%	The columns having numerical missing values can be found using AVERAGEIFS or MEDIAN The columns with categorical missing values can be found either using MODE or “Unknown”
>40%	The columns with more than 40% missing values can be removed (I have marked them with Red colour)
Unnecessary Columns	There are some columns which are not necessary for the analysis, and we can remove those as well



There are 8 columns with less than 40% Blanks, we can find the missing values using AVERAGEIFS or MEDIAN, MODE and Unknown



There are also 49 columns with more than 40% Blanks, I have marked columns as ‘without blanks’ with Red colour highlight using Conditional Formatting in the ‘EDA’ Sheet, we need to remove those columns

Now we find the missing values

The missing values of the numerical values can be found using the MEDIAN formula

```
=MEDIAN(K6:K50004)
```

MEDIAN			
AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	EXT_SOURCE_2
24939	450000	2	0.565585366
EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE
0.53527625	0	0	0
DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY
0	-755	0	0
AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
0	0	0	1

For the Column NAME_TYPE_SUITE

The missing values of categorical fields can be found using the COUNTIF formula

```
=COUNTIF(M6:M50004, CA6)
```

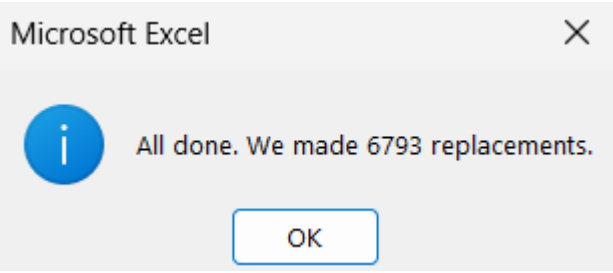
Title	Count
Unaccompanied	40435
Family	6549
Spouse, partner	1849
Children	542
Other_A	137
Other_B	259
Group of people	36
(Blanks)	192
Total	**49999**

From this we can see that the value “Unaccompanied” has more counts, so we will fill the missing fields with that value.

For the Column OCCUPATION_TYPE

There are 15,654 missing values, so we are going to filling them with the value “Unknown”. We can use Find and Replace in Excel to fill all the blanks of this column with the value “Unknown”.

The same can be done for the MEDIAN values as well



All the missing values have been found and dealt with appropriately.

The column header has been marked with Orange fill.

B. Identify Outliers in the Dataset

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

The Quartile and Inter Quartile Range (IQR) can be found for the outliers using the QUARTILE formula, and the deference between the two Quartile will be the IQR

```
=QUARTILE.INC(B2:B50002, 1)
```

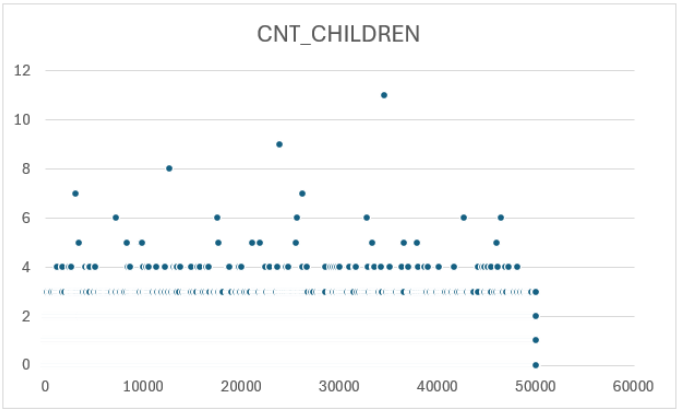
```
=QUARTILE.INC(B3:B50003, 3)
```

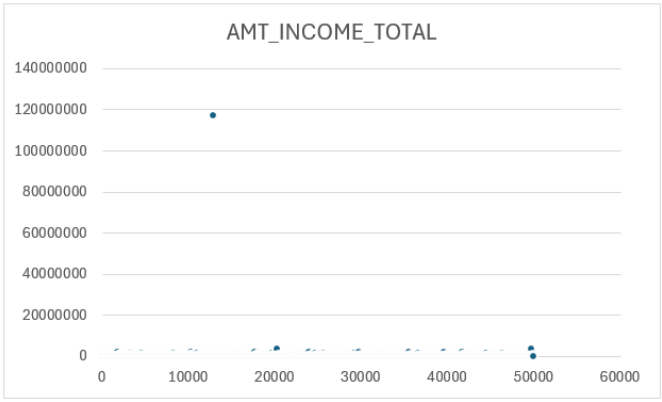
```
=G4-G3
```

```
=G4+(1.5*G5)
```

```
=G3-(1.5*G5)
```

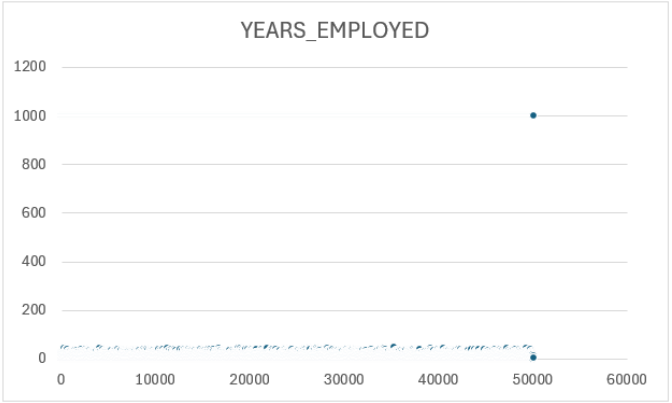
Column Title	CNT_CHILDREN	AMT_INCOME_TOTAL	YEARS_EMPLOYED
Q1	0	112500	3
Q3	1	202500	16
IQR	1	90000	13
Upper Limit	2.5	337500	35.5
Lower Limit	-1.5	-22500	-16.5





In CNT_CHILDREN, it shows that there are people who have more than 8 Children, which is rare and unrealistic in this generation.

In AMT_INCOME_TOTAL, the income 117,000,000 is extremely high.



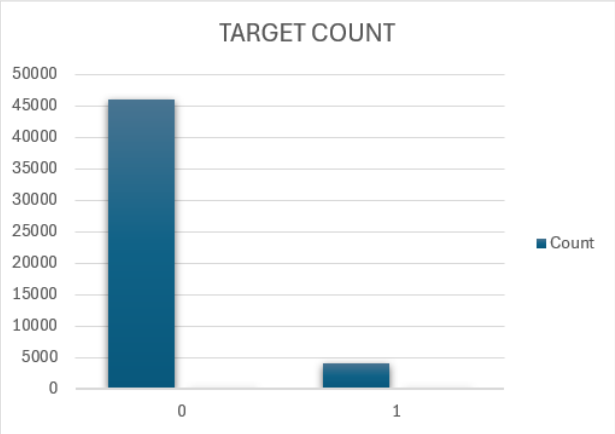
In YEARS_EMPLOYED, it shows that people are employed for over 1,000 years which is impossible.

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.

Using Pivot Table, the COUNT and Percentage of the TARGET Column can be found

Target	Count	% of Count
0	45973	91.95%
1	4026	8.05%
Grand Total	**49999**	**100.00%**



The dataset is imbalanced because of more instances of class 0 (92%) than class 1 (8%).

Hence it will be difficult to predict class 1.

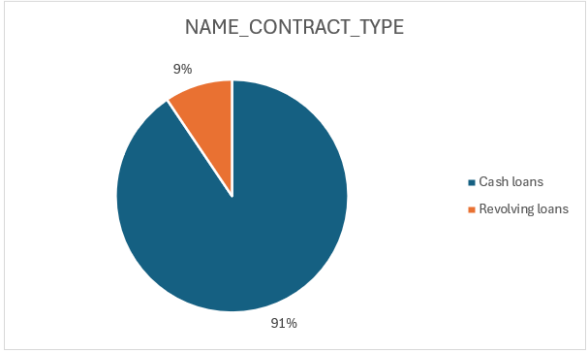
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

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.

Univariate Analysis

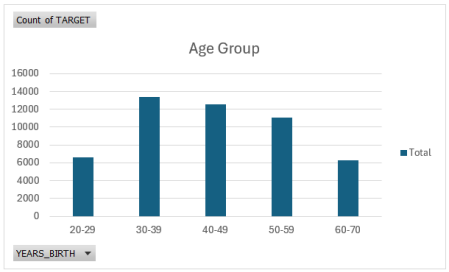
Using Pivot Table, the COUNT and Percentage of the columns can be found

NAME_CONTRACT_TYPE	Count	% of Count
Cash loans	45276	90.55%
Revolving loans	4723	9.45%
Grand Total	**49999**	**100.00%**



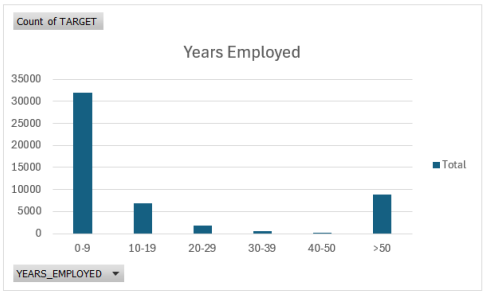
We can see that the most loan contract type is Cash Loans with 91% than Revolving Loans of just 9%.

Age Group	Count of TARGET
20-29	6644
30-39	13420
40-49	12576
50-59	11044
60-70	6315
Grand Total	**49999**



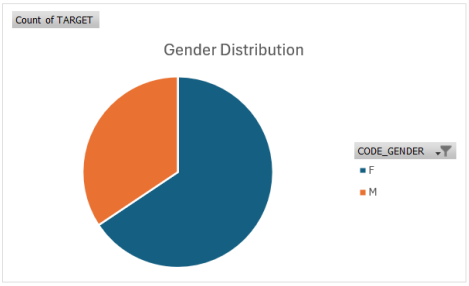
The highest number of applicants are from the age group 30 to 39, followed by 40 to 49

Years Employed	Count of TARGET
0-9	31951
10-19	6869
20-29	1721
30-39	491
40-50	43
>50	8924
Grand Total	**49999**



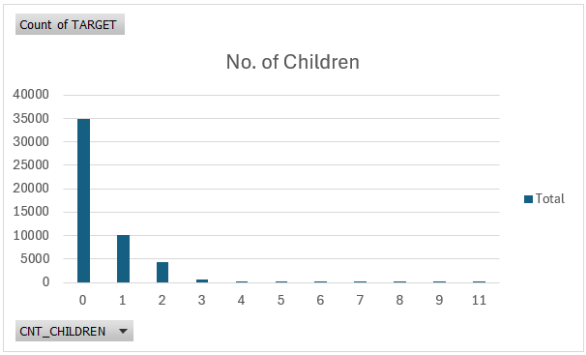
Most of the applicants are employed for 0 to 9 years

Gender	Count of TARGET
F	32823
M	17174
Grand Total	**49997**



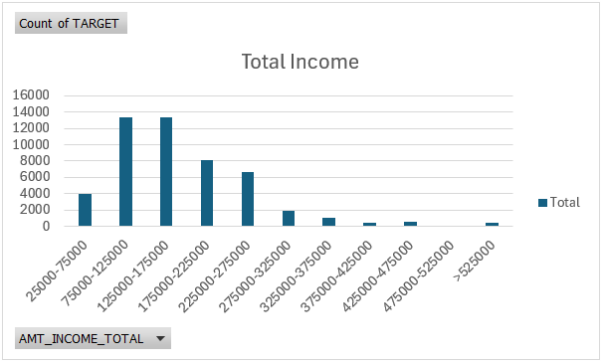
There are more female applicants than male

Children	Count of TARGET
0	34916
1	10041
2	4319
3	626
4	73
5	13
6	6
7	2
8	1
9	1
11	1
Grand Total	**49999**



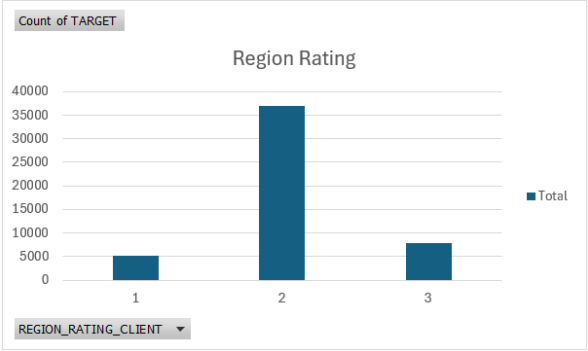
Most of the applicants have no children

Total Income	Count of TARGET
25000-75000	4030
75000-125000	13410
125000-175000	13365
175000-225000	8082
225000-275000	6645
275000-325000	1861
325000-375000	1103
375000-425000	489
425000-475000	513
475000-525000	47
>525000	454
Grand Total	**49999**



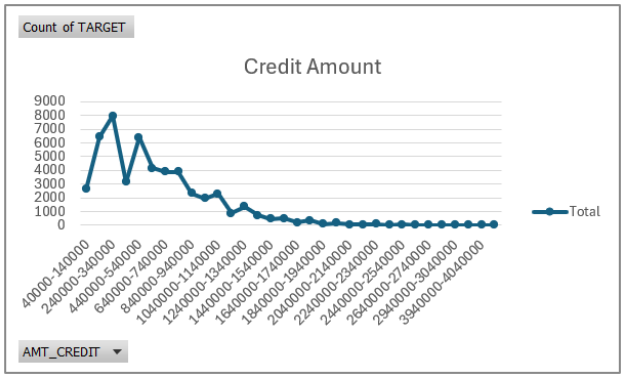
The greatest number of applicants have an annual income between 75,000 and 175,000

Region Rating	Count of TARGET
1	5226
2	36964
3	7809
Grand Total	**49999**

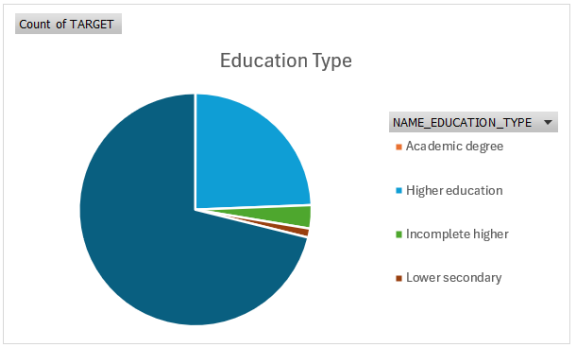


Region 2 has more applicants

Credit Amount	Count of TARGET
40000-140000	2615
140000-240000	6416
240000-340000	7974
340000-440000	3147
440000-540000	6396
540000-640000	4164
640000-740000	3885
740000-840000	3879
840000-940000	2338
940000-1040000	1958
1040000-1140000	2288
1140000-1240000	860
1240000-1340000	1363
1340000-1440000	726
1440000-1540000	468
1540000-1640000	491
1640000-1740000	195
1740000-1840000	331
1840000-1940000	95
1940000-2040000	170
2040000-2140000	40
2140000-2240000	41
2240000-2340000	81
2340000-2440000	14
2440000-2540000	44
2540000-2640000	2
2640000-2740000	11
2840000-2940000	2
2940000-3040000	1
3140000-3240000	1
3940000-4040000	1
4040000-4140000	2
Grand Total	**49999**

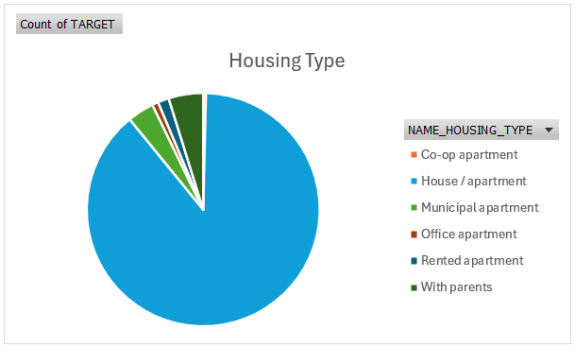


Education Type	Count of TARGET
Academic degree	20
Higher education	12167
Incomplete higher	1620
Lower secondary	620
Secondary / secondary special	35572
Grand Total	**49999**



Most of the applicants have Secondary / Secondary Special education

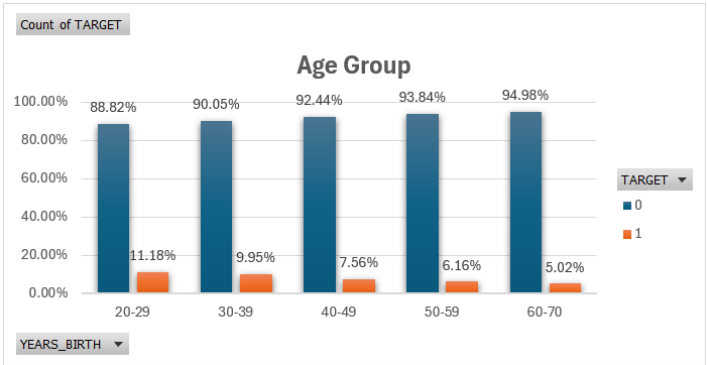
Housing Type	Count of TARGET
Co-op apartment	191
House / apartment	44368
Municipal apartment	1845
Office apartment	427
Rented apartment	769
With parents	2399
Grand Total	**49999**



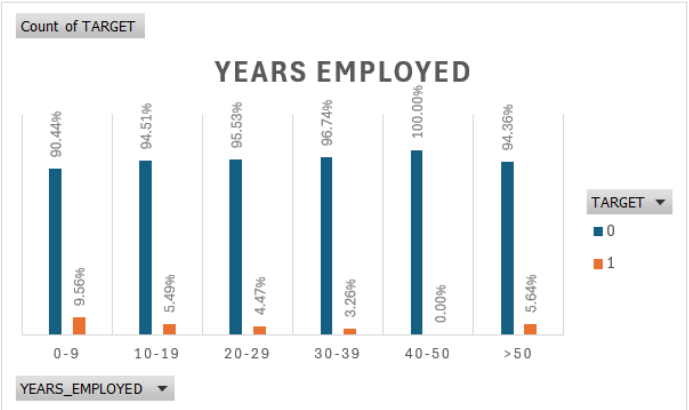
Most of the applicants have own a house or apartment

Segmented Univariate Analysis

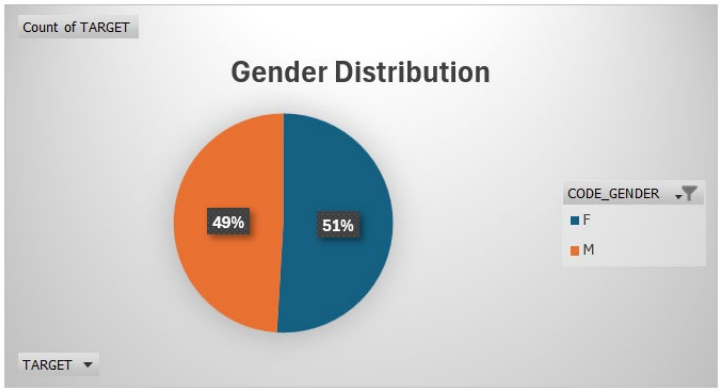
Count of TARGET	Target		
Age Group	0	1	Grand Total
20-29	88.82%	11.18%	100.00%
30-39	90.05%	9.95%	100.00%
40-49	92.44%	7.56%	100.00%
50-59	93.84%	6.16%	100.00%
60-70	94.98%	5.02%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



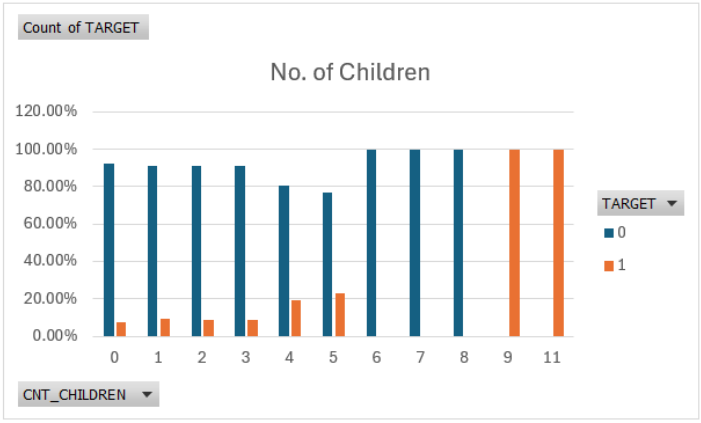
Count of TARGET	Target		
Years Employed	0	1	Grand Total
0-9	90.44%	9.56%	100.00%
10-19	94.51%	5.49%	100.00%
20-29	95.53%	4.47%	100.00%
30-39	96.74%	3.26%	100.00%
40-50	100.00%	0.00%	100.00%
>50	94.36%	5.64%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



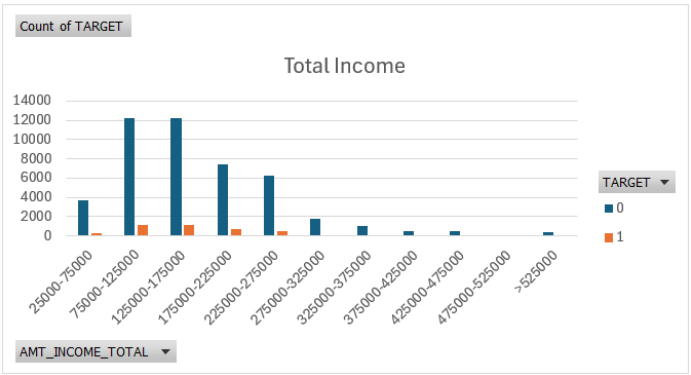
Count of TARGET	Target		
Gender	0	1	Grand Total
F	93.10%	6.90%	100.00%
M	89.74%	10.26%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



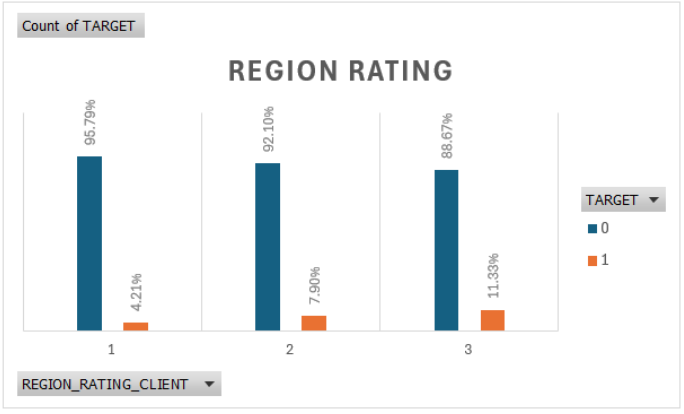
Count of TARGET	Target		
Children	0	1	Grand Total
0	92.43%	7.57%	100.00%
1	90.81%	9.19%	100.00%
2	91.11%	8.89%	100.00%
3	91.05%	8.95%	100.00%
4	80.82%	19.18%	100.00%
5	76.92%	23.08%	100.00%
6	100.00%	0.00%	100.00%
7	100.00%	0.00%	100.00%
8	100.00%	0.00%	100.00%
9	0.00%	100.00%	100.00%
11	0.00%	100.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



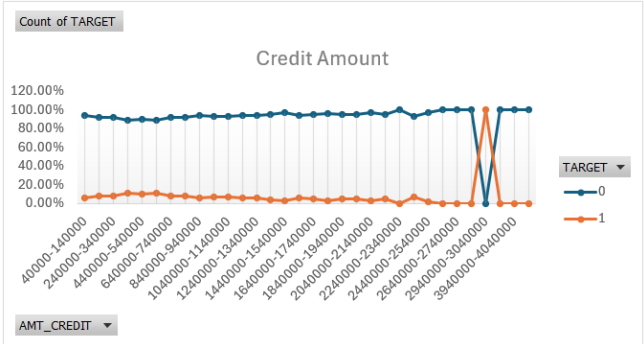
Count of TARGET	Target		
Total Income	0	1	Grand Total
25000-75000	3721	309	4030
75000-125000	12254	1156	13410
125000-175000	12186	1179	13365
175000-225000	7421	661	8082
225000-275000	6198	447	6645
275000-325000	1757	104	1861
325000-375000	1045	58	1103
375000-425000	449	40	489
425000-475000	475	38	513
475000-525000	44	3	47
>525000	423	31	454
Grand Total	**45973**	**4026**	**49999**



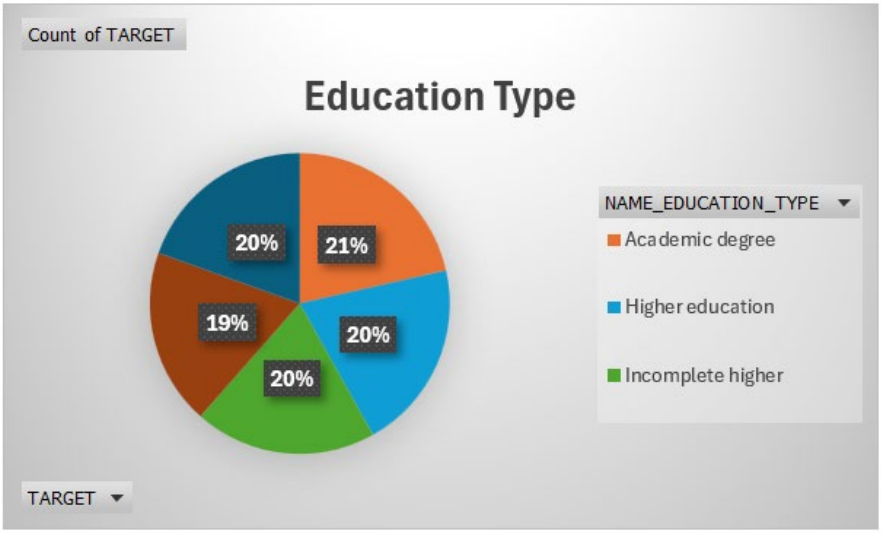
Count of TARGET	Target		
Region Rating	0	1	Grand Total
1	95.79%	4.21%	100.00%
2	92.10%	7.90%	100.00%
3	88.67%	11.33%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



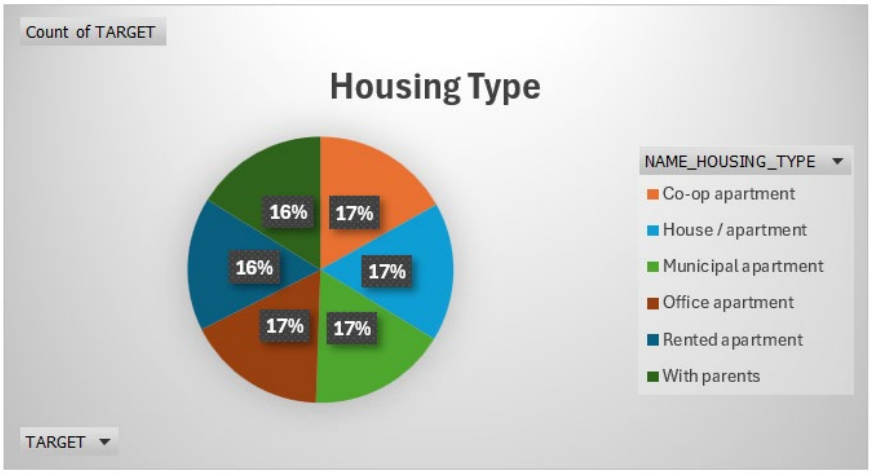
Count of TARGET	Column Labels		
Credit Amount	0	1	Grand Total
40000-140000	94.07%	5.93%	100.00%
140000-240000	92.43%	7.57%	100.00%
240000-340000	91.76%	8.24%	100.00%
340000-440000	88.85%	11.15%	100.00%
440000-540000	90.35%	9.65%	100.00%
540000-640000	89.34%	10.66%	100.00%
640000-740000	92.02%	7.98%	100.00%
740000-840000	91.65%	8.35%	100.00%
840000-940000	93.97%	6.03%	100.00%
940000-1040000	93.05%	6.95%	100.00%
1040000-1140000	92.74%	7.26%	100.00%
1140000-1240000	94.42%	5.58%	100.00%
1240000-1340000	94.35%	5.65%	100.00%
1340000-1440000	95.59%	4.41%	100.00%
1440000-1540000	97.01%	2.99%	100.00%
1540000-1640000	94.30%	5.70%	100.00%
1640000-1740000	95.38%	4.62%	100.00%
1740000-1840000	96.68%	3.32%	100.00%
1840000-1940000	94.74%	5.26%	100.00%
1940000-2040000	94.71%	5.29%	100.00%
2040000-2140000	97.50%	2.50%	100.00%
2140000-2240000	95.12%	4.88%	100.00%
2240000-2340000	100.00%	0.00%	100.00%
2340000-2440000	92.86%	7.14%	100.00%
2440000-2540000	97.73%	2.27%	100.00%
2540000-2640000	100.00%	0.00%	100.00%
2640000-2740000	100.00%	0.00%	100.00%
2840000-2940000	100.00%	0.00%	100.00%
2940000-3040000	0.00%	100.00%	100.00%
3140000-3240000	100.00%	0.00%	100.00%
3940000-4040000	100.00%	0.00%	100.00%
4040000-4140000	100.00%	0.00%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



Count of TARGET	Column Labels		
Education Type	0	1	Grand Total
Academic degree	100.00%	0.00%	100.00%
Higher education	95.02%	4.98%	100.00%
Incomplete higher	91.48%	8.52%	100.00%
Lower secondary	88.23%	11.77%	100.00%
Secondary / secondar	90.98%	9.02%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



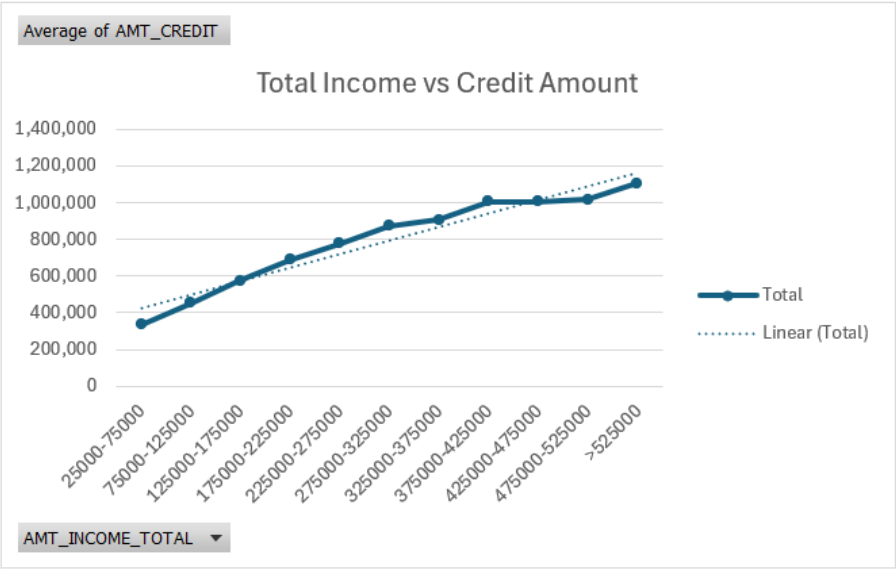
Count of TARGET	Column Labels		
Housing Type	0	1	Grand Total
Co-op apartment	92.15%	7.85%	100.00%
House / apartment	92.17%	7.83%	100.00%
Municipal apartment	92.14%	7.86%	100.00%
Office apartment	93.21%	6.79%	100.00%
Rented apartment	88.69%	11.31%	100.00%
With parents	88.45%	11.55%	100.00%
Grand Total	**91.95%**	**8.05%**	**100.00%**



Bivariate Analysis

For the Bivariate Analysis, we use the AMT_INCOM_TOTAL column in the pivot table along with the average of AMT_CREDIT column and prepare a line graph with trendline to understand the relationship between the Total Income and the Credit Amount.

Total Income	Average of AMT_CREDIT
25000-75000	335,766
75000-125000	452,114
125000-175000	573,427
175000-225000	687,047
225000-275000	778,381
275000-325000	872,960
325000-375000	904,707
375000-425000	1,006,409
425000-475000	1,004,628
475000-525000	1,015,150
>525000	1,105,365
Grand Total	**599,701**



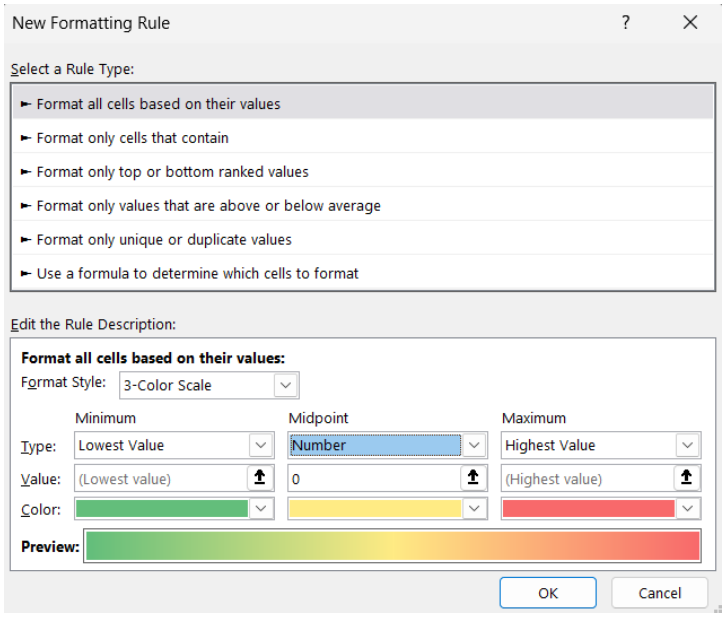
E. Identify Top Correlations for Different Scenarios

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.

I made two sheets, one for Target 0 and one for Target 1 to get the data ready for correlation. I used TRANSPOSE to rearrange the data where needed, then calculated the correlations using the CORREL formula

```
=CORREL(Target_0!B:B, Target_0!B:B)
```

To make the results into a heatmap and easier to read, I added a 3-color scale with conditional formatting: red for high values, green for low, and yellow for values around zero. I also made sure #DIV/0! errors were shaded the same as zero so they wouldn't stand out unnecessarily.



TARGET 0: CORRELATION FOR CLIENTS WHO MADE PAYMENTS ON TIME

COLUMN TITLE	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH
CNT_CHILDREN	1.000	0.036	0.006	0.026	0.002	-0.025	-0.336	-0.246	-0.183	0.033
AMT_INCOME_TOTAL	0.036	1.000	0.378	0.451	0.385	0.182	-0.074	-0.162	-0.069	-0.032
AMT_CREDIT	0.006	0.378	1.000	0.771	0.987	0.096	0.051	-0.075	-0.008	0.008
AMT_ANNUITY	0.026	0.451	0.771	1.000	0.776	0.117	-0.010	-0.111	-0.034	-0.010
AMT_GOODS_PRICE	0.002	0.385	0.987	0.776	1.000	0.099	0.049	-0.072	-0.011	0.009
REGION_POPULATION_RELATIVE	-0.025	0.182	0.096	0.117	0.099	1.000	0.030	-0.007	0.058	0.002
YEARS_BIRTH	-0.336	-0.074	0.051	-0.010	0.049	0.030	1.000	0.623	0.335	0.270
YEARS_EMPLOYED	-0.246	-0.162	-0.075	-0.111	-0.072	-0.007	0.623	1.000	0.209	0.274
YEARS_REGISTRATION	-0.183	-0.069	-0.008	-0.034	-0.011	0.058	0.335	0.209	1.000	0.104
YEARS_ID_PUBLISH	0.033	-0.032	0.008	-0.010	0.009	0.002	0.270	0.274	0.104	1.000

TARGET 1: CORRELATION FOR CLIENTS WHO HAVE PAYMENT DIFFICULTIES

COLUMN TITLE	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH
CNT_CHILDREN	1.000	0.010	0.008	0.029	-0.001	-0.020	-0.250	-0.190	-0.152	0.043
AMT_INCOME_TOTAL	0.010	1.000	0.015	0.018	0.013	-0.006	-0.008	-0.012	0.010	0.009
AMT_CREDIT	0.008	0.015	1.000	0.750	0.982	0.068	0.142	0.019	0.043	0.044
AMT_ANNUITY	0.029	0.018	0.750	1.000	0.750	0.073	0.009	-0.078	-0.022	0.021
AMT_GOODS_PRICE	-0.001	0.013	0.982	0.750	1.000	0.077	0.141	0.023	0.043	0.050
REGION_POPULATION_RELATIVE	-0.020	-0.006	0.068	0.073	0.077	1.000	0.017	0.008	0.046	0.006
YEARS_BIRTH	-0.250	-0.008	0.142	0.009	0.141	0.017	1.000	0.588	0.288	0.248
YEARS_EMPLOYED	-0.190	-0.012	0.019	-0.078	0.023	0.008	0.588	1.000	0.193	0.231
YEARS_REGISTRATION	-0.152	0.010	0.043	-0.022	0.043	0.046	0.288	0.193	1.000	0.091
YEARS_ID_PUBLISH	0.043	0.009	0.044	0.021	0.050	0.006	0.248	0.231	0.091	1.000

RESULT

The analysis showed that factors like short employment length, lower income, and specific loan purposes (like the consumer goods) were strongly linked to defaults.
I also spotted data quality issues like invalid gender values.
These insights can help the company reduce risk by adjusting approvals, setting better loan terms, or flagging high-risk profiles, without losing reliable customers.

THANK YOU
