Final Project-2,Trainity
Bank Loan Case Study.

By: Syed Ali Ashraf.

Project Objective

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.

Tech Stack used:

MS Excel(Advanced Excel Features + VBA Macros) , Statistical Knowledge and AI.

Task A. Identify Missing Data and Deal with it Appropriately .

There were two sheets of valuable information.

Workbook 1: application_data.csv : Provides details about the current loan applications.

Workbook 2: previous_application.csv : Contains information about previous loan applications.

Workbook 1:

Workbook 1 had 122 Columns and 49999 rows.

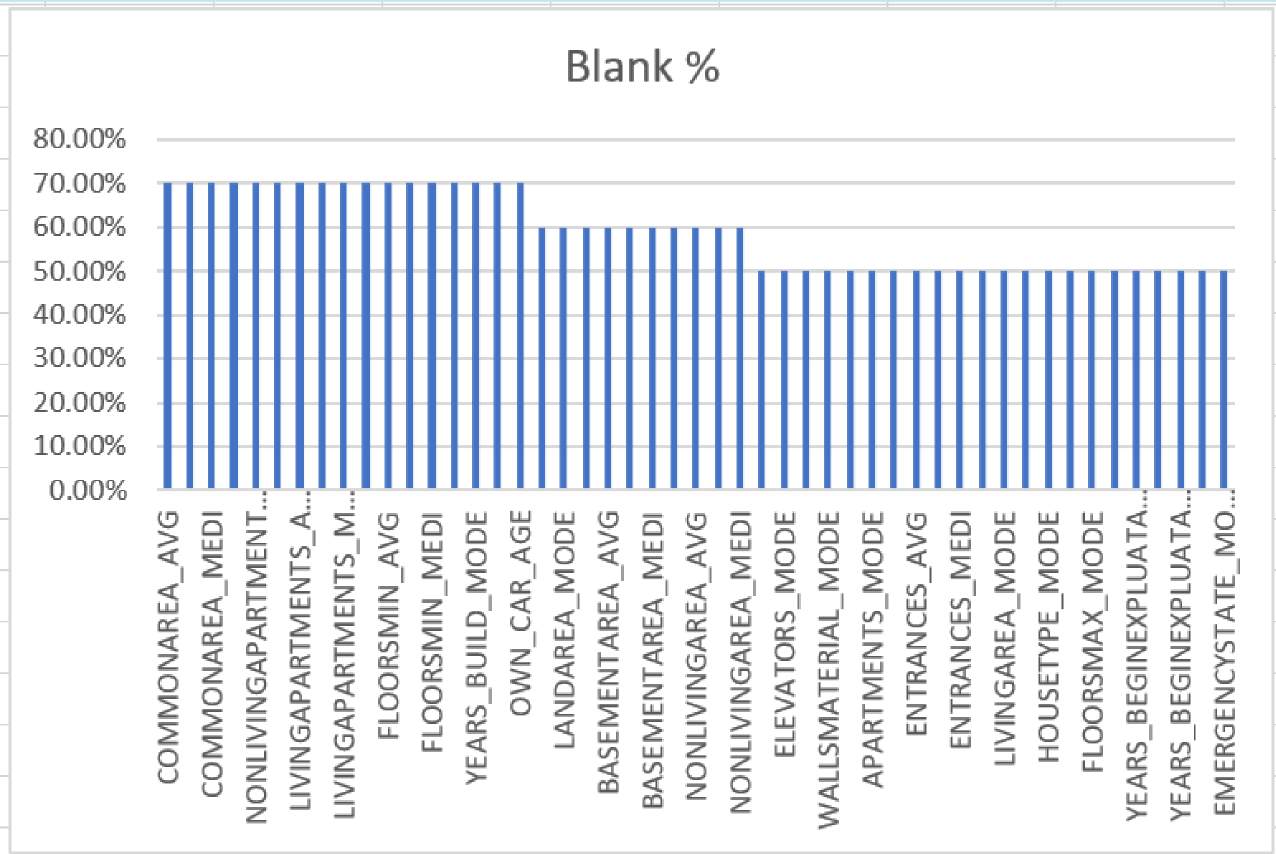
Firstly, Blank cells in each column was calculated using countblank() and then Blank % was calculated for each column.
Calculations: Current_data workbook, sheet: Working_data.

Then Columns with Blank % >=50% were removed using VBA.

Refer: Sheet name: Removed Blank Columns for full data.

Columns	Blanks	Blank %
COMMONAREA_AVG	34960	70.00%
COMMONAREA_MODE	34960	70.00%
COMMONAREA_MEDI	34960	70.00%
NONLIVINGAPARTMENTS_AVG	34714	70.00%
NONLIVINGAPARTMENTS_MODE	34714	70.00%
NONLIVINGAPARTMENTS_MEDI	34714	70.00%
LIVINGAPARTMENTS_AVG	34226	70.00%
LIVINGAPARTMENTS_MODE	34226	70.00%
LIVINGAPARTMENTS_MEDI	34226	70.00%
FONDKAPREMONT_MODE	34191	70.00%
FLOORSMIN_AVG	33894	70.00%
FLOORSMIN_MODE	33894	70.00%
FLOORSMIN_MEDI	33894	70.00%
YEARS_BUILD_AVG	33239	70.00%
YEARS_BUILD_MODE	33239	70.00%
YEARS_BUILD_MEDI	33239	70.00%
OWN_CAR_AGE	32950	70.00%

Column Chart:



As per the column_description many other columns were removed.
Columns : DAYS_BIRTH AND DAYS_EMPLOYED were converted into years (cell/365).

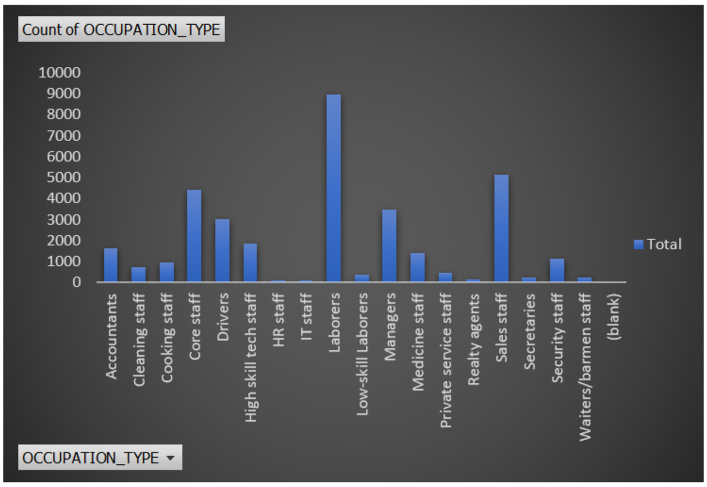
Further for handling missing values in important numerical column like: AMT_CREDIT, AMT_ANNUITY, EXT_SOURCE_2 AND EXT_SOURCE_3

Median was calculated and blank values were replaced with median value in respective columns.

For blank cells in Categorical Columns like: OCCUPATION_TYPE and NAME_TYPE_SUITE, blank cells were replaced by most count values.

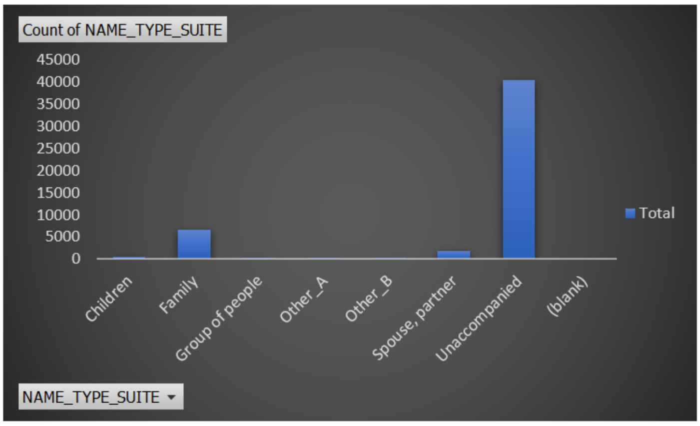
In Occupation_type: replaced by Laborers.

Row Labels	Count of OCCUPATION_TYPE
Accountants	1621
Cleaning staff	739
Cooking staff	963
Core staff	4434
Drivers	3044
High skill tech staff	1852
HR staff	101
IT staff	80
Laborers	8952
Low-skill Laborers	357
Managers	3489
Medicine staff	1403
Private service staff	447
Realty agents	123
Sales staff	5160
Secretaries	212
Security staff	1140
Waiters/barmen staff	228
(blank)	



In Name_type_suite: replaced by Unaccompanied.

Row Labels	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
(blank)	
Grand Total	**49807**



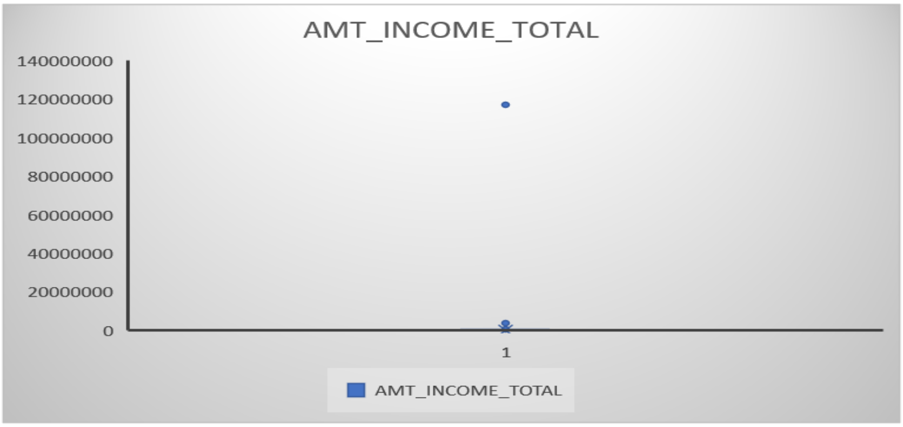
Task B. Identify Outliers in the Dataset:

Outliers are values that can impact the analysis and distort the results. This task is in continuance with Task A.

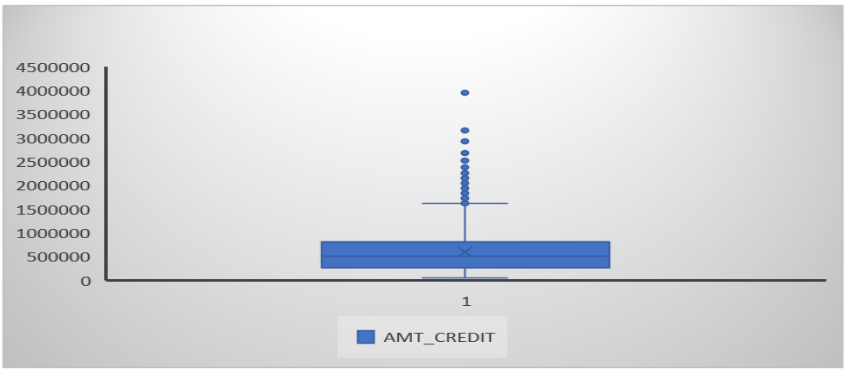
Outliers were calculated for important numerical columns such as

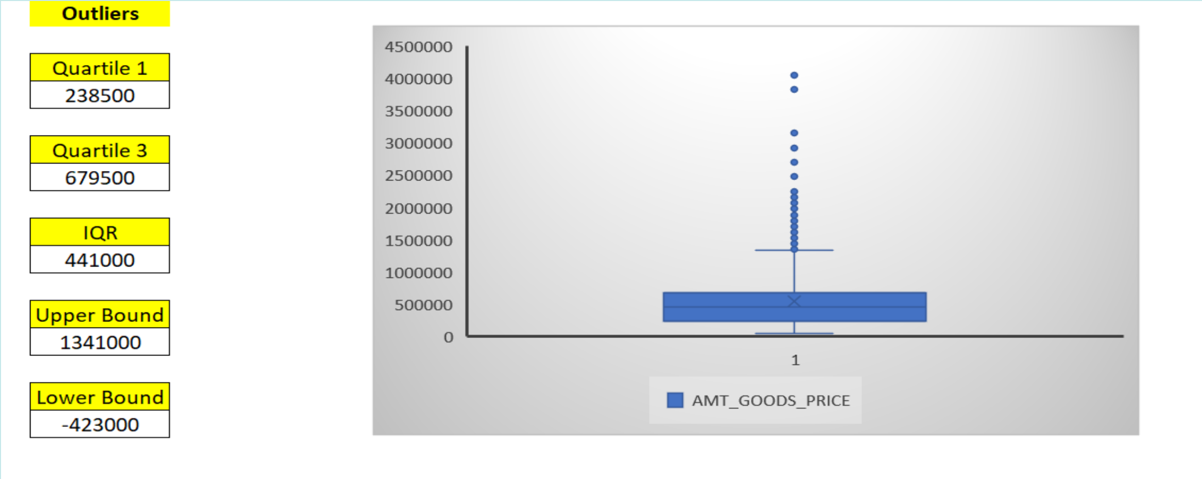
AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE AND YEARS_EMPLOYED as shown below in graphs.

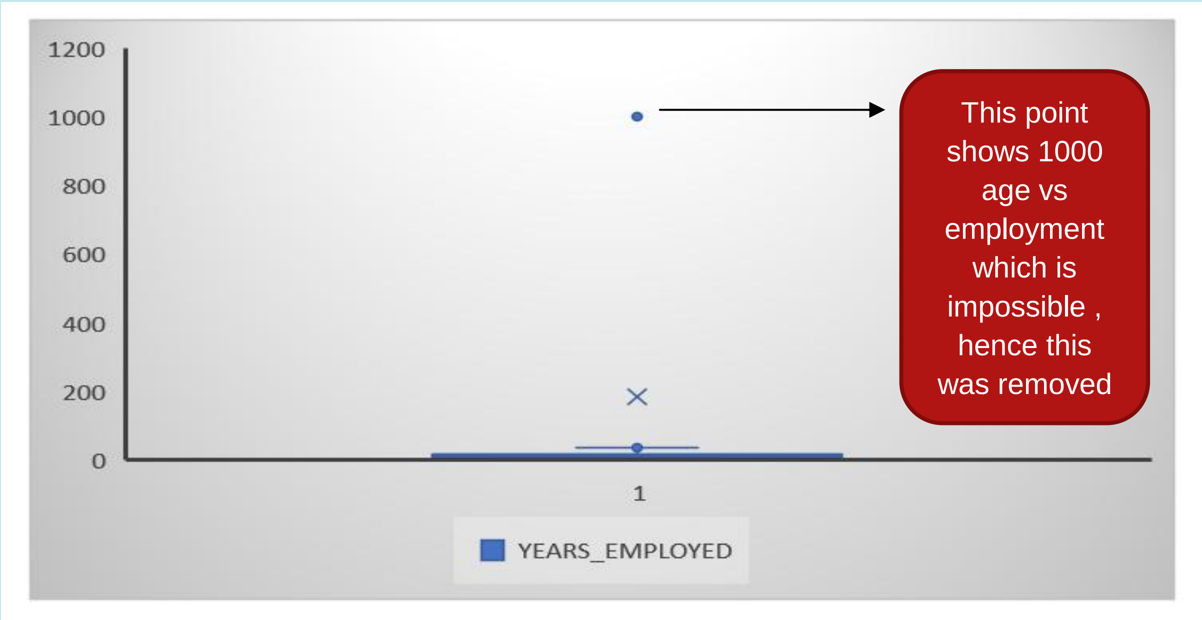
Outliers	
Quartile 1	112500
Quartile 3	202500
IQR	90000
Upper Bound	337500
Lower Bound	-22500



Outliers	
Quartile 1	270000
Quartile 3	808650
IQR	538650
Upper Bound	1616625
Lower Bound	-537975







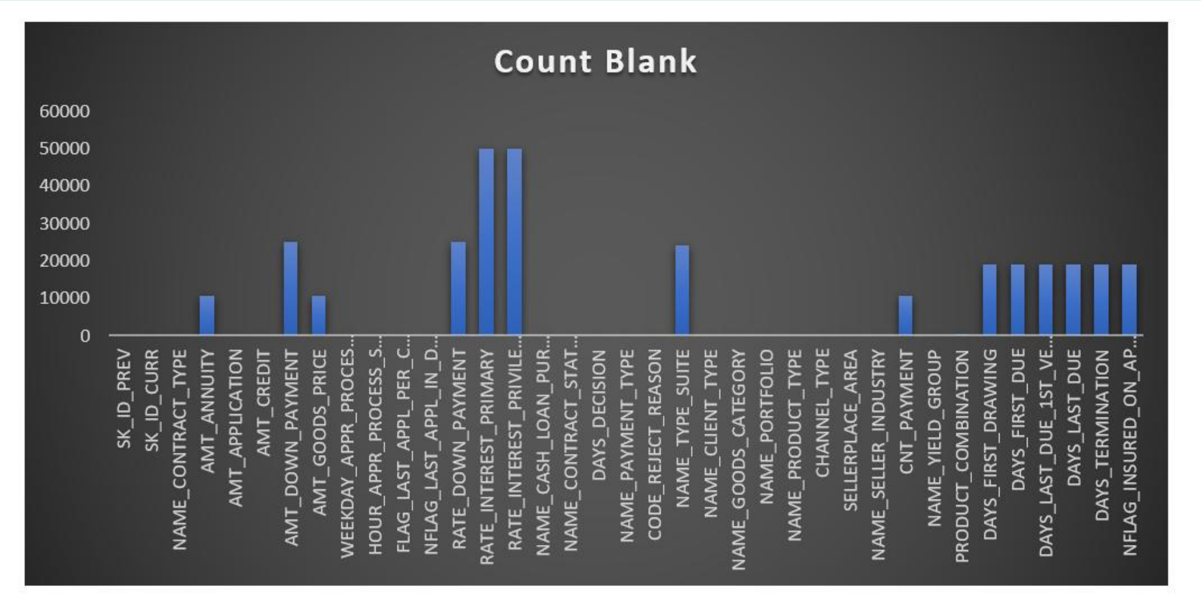
After calculating outliers and removing outlier from years_employed, The Current Application Data was finally cleaned and fit for analysis with 28 columns and 41075 rows.

Workbook 2: previous_application.csv.

This had 37 columns and 49999 rows.

Firstly, blank values for each column was calculated using countblank() along with blank% .

Column Chart:



Columns having >=50% blank columns were removed

Column	Count Blank	Blank %
RATE_INTEREST_PRIMARY	49834	100.00%
RATE_INTEREST_PRIVILEGED	49834	100.00%
AMT_DOWN_PAYMENT	25198	50.00%
RATE_DOWN_PAYMENT	25198	50.00%
NAME_TYPE_SUITE	24243	50.00%

For Numerical columns such as AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT AND AMT_GOODS_PRICE,
Median values were calculated and blank cells were replaced in each column respectively.

previous_application.csv dataset was cleaned and shared as
Final Prev_data in workbook named Previous_Data.xlsx.

Task C. Analyze Data Imbalance.

TARGET: Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

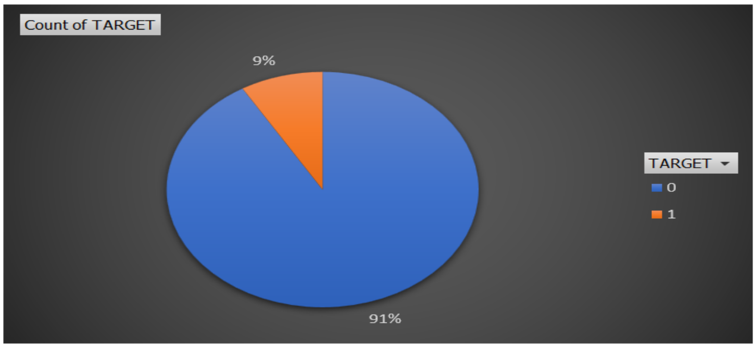
1 = Late Payment
0 = Payment on Time.

Dataset: application_data

Primarily , Pivot Tables and Charts were used .

1)TARGET

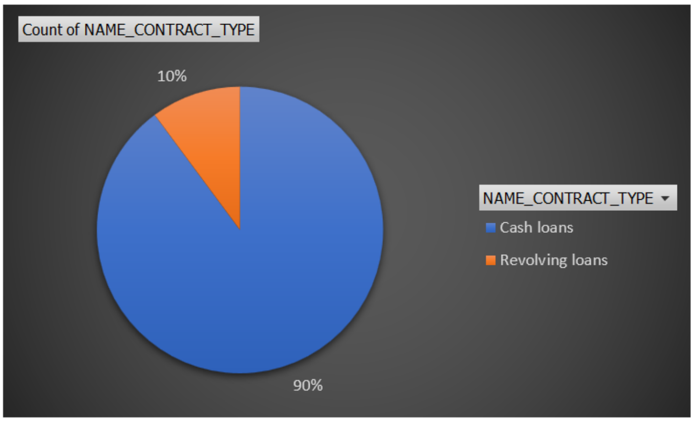
Row Labels	Count of TARGET
0	37552
1	3523
Grand Total	**41075**



91% clients repay on time while near 9% are considered as defaulters.`

2)NAME_CONTRACT_TYPE

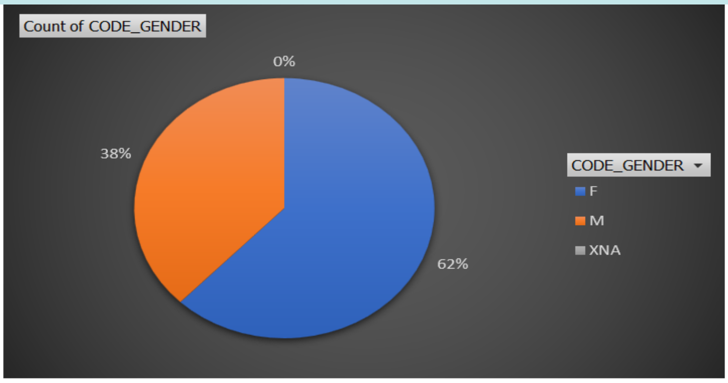
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	36898
Revolving loans	4177
Grand Total	**41075**



90% of clients take cash loan and only 10 % via revolving loans.

3) GENDER

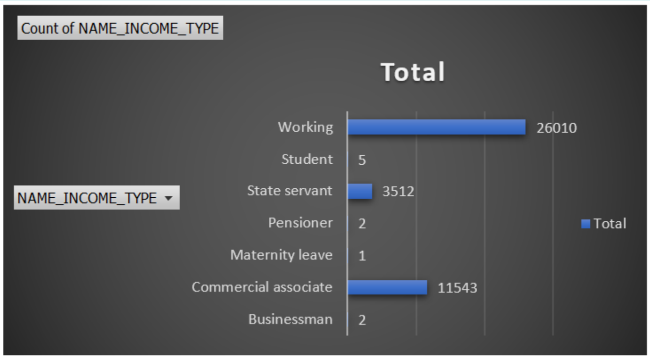
Row Labels	Count of CODE_GENDER
F	25538
M	15535
XNA	2
Grand Total	**41075**



62% of clients taking loan are females .

4)NAME_INCOME_TYPE

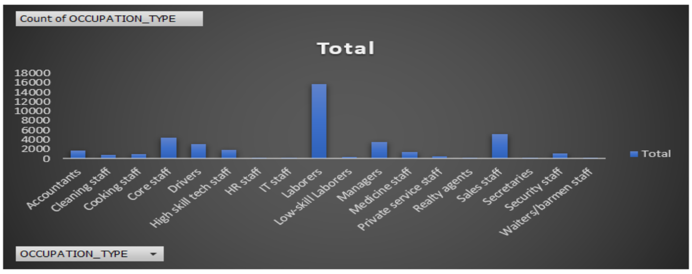
Row Labels	Count of NAME_INCOME_TYPE
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	2
State servant	3512
Student	5
Working	26010
Grand Total	**41075**



Working group is more involved in direct bank loan than others especially with respect to businessman.

5)OCCUPATION_TYPE

Row Labels	Count of OCCUPATION_TYPE
Accountants	1621
Cleaning staff	739
Cooking staff	963
Core staff	4434
Drivers	3044
High skill tech staff	1852
HR staff	101
IT staff	80
Laborers	15682
Low-skill Laborers	357
Managers	3489
Medicine staff	1403
Private service staff	447
Realty agents	123
Sales staff	5160
Secretaries	212
Security staff	1140
Waiters/barmen staff	228
Grand Total	**41075**



Occupation type : Laborers is quite high with respect to other jobs.

Task D. Univariate, Segmented Univariate, and Bivariate Analysis.

Univariate Analysis

1)AMT_INCOME_TOTAL

Average	**178592.5717**
Median	**157500**
STDEV	**585391.2655**
Max	**117000000**
Min	**25650**

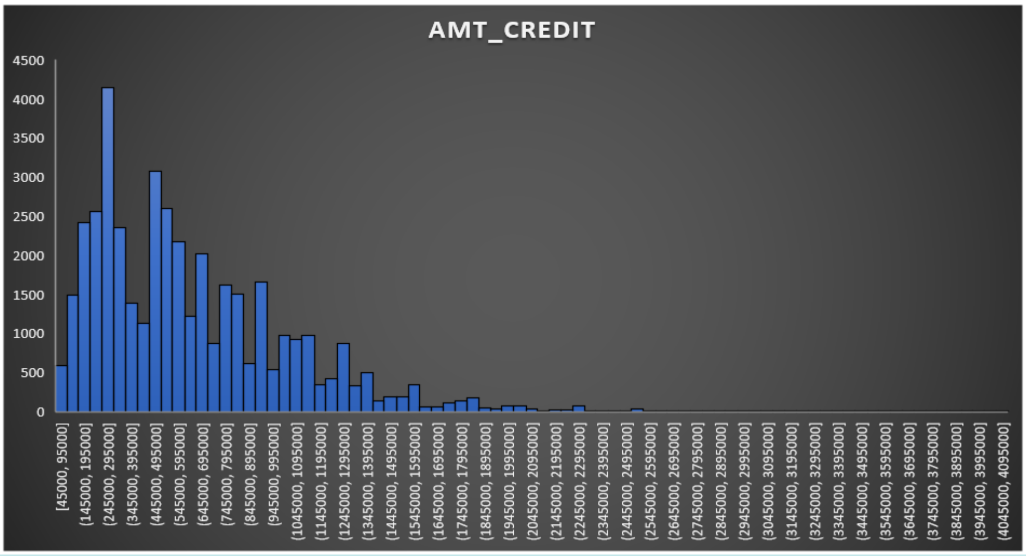
Average (178,592.57) : The mean total income, meaning most incomes cluster around this amount.

Median (157,500) : Since it is lower than the average, it suggests some extremely high incomes are pulling the mean up.

Standard Deviation (585,391.26) : This measures income variability. A high standard deviation indicates significant differences among individuals' incomes, with some earning much more or less than others.

2)AMT_CREDIT

Average	612666
Median	521280
STDEV	406424
Max	4050000
Min	45000



Average (612,665.81) : The mean loan amount granted.

Median (521,280) : The middle value, meaning half of the loans are below this amount and half are above. Since it's lower than the average, larger loans are pulling the mean upward.

Standard Deviation (406,423.81) – It shows spread out the loan amounts are. A high value suggests loans vary significantly in size.

Max (4,050,000) :The highest recorded loan, indicating the bank grants some very large loans.

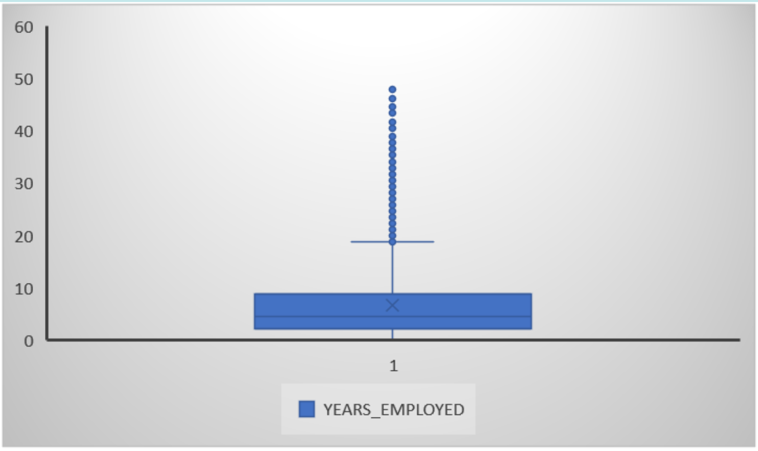
<u>Key Takeaways:</u>

Loan Amounts Vary Widely : The high standard deviation shows significant differences in loan sizes.

Presence of Large Loans : Since the average is higher than the median, there are some large loans skewing the data.

3)YEARS_EMPLOYED

Average	7
Median	5
STDEV	6.446793
Max	48
Min	0



Average (7 years) : On average, loan applicants have been employed for about 7 years.

Median (5 years) : The middle value, meaning half of the applicants have worked **less than 5 years**, while the other half have worked **more than 5 years**.

Standard Deviation (6.45 years) : This indicates how much employment durations vary. A fairly high standard deviation suggests applicants have **diverse work histories**, ranging from very short to long-term employment.

Max (48 years) : The longest tenure recorded, likely representing an applicant near retirement or with a stable career spanning decades.

Min (0 years) : Some applicants have no recorded employment history, possibly indicating **students, unemployed individuals, or self-employed applicants** who don't report traditional employment.

<u>Key Takeaways</u>:

Significant Variation in Employment History : Some applicants have **decades of experience**, while others are just starting or have gaps.

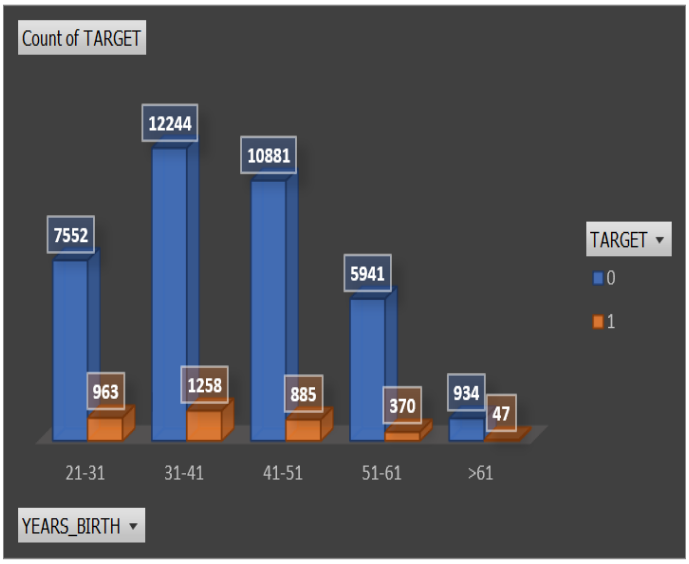
Employment Stability & Loan Eligibility : Banks may assess stability based on tenure, preferring longer employment histories for larger loans.

Potential Risk Factors : Individuals with shorter work histories could pose a **higher risk** for loan repayment, depending on other financial factors.

Segmented Univariate and Bivariate Analysis

1) **YEARS_BIRTH vs Target.**
   (ages were grouped)

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
21-31	7552	963	8515
31-41	12244	1258	13502
41-51	10881	885	11766
51-61	5941	370	6311
>61	934	47	981
Grand Total	37552	3523	41075



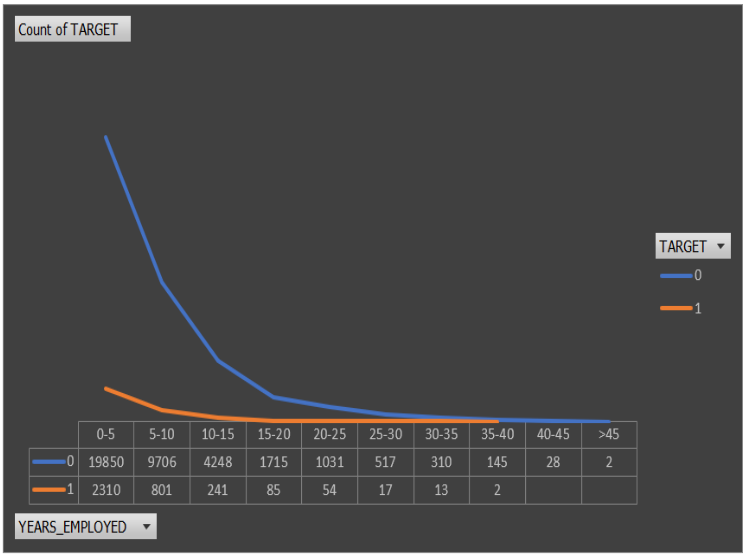
Line Chart explains that

-with increase in age ,the possibility of being a defaulter decreases

- and upward trend lies between 21-41 age group.

2)YEARS_EMPLOYED vs Target

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
0-5	19850	2310	22160
5-10	9706	801	10507
10-15	4248	241	4489
15-20	1715	85	1800
20-25	1031	54	1085
25-30	517	17	534
30-35	310	13	323
35-40	145	2	147
40-45	28		28
>45	2		2
Grand Total	37552	3523	41075

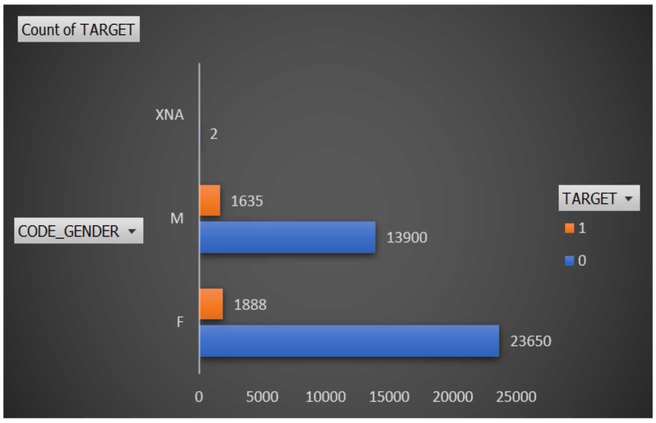


Line Chart explains that

-Lesser the experience of work , more the chances of default.

3)GENDER vs Target

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
F	23650	1888	25538
M	13900	1635	15535
XNA	2		2
Grand Total	37552	3523	41075

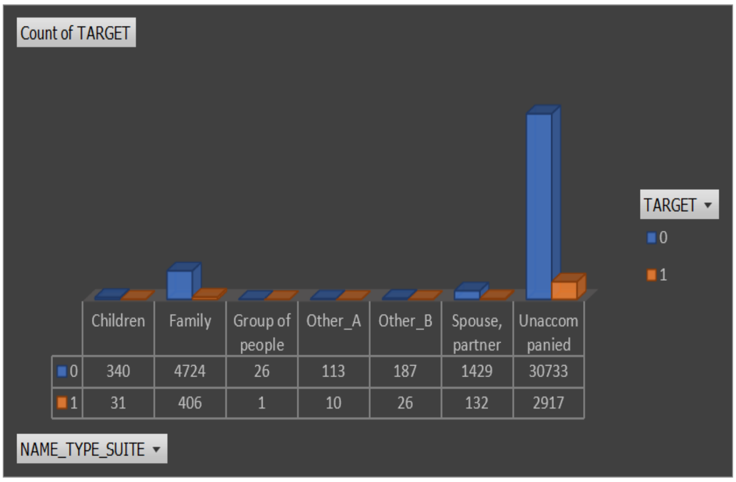


Bar chart explains that-

Females have higher chances on defaulting but the margin is narrow.

4)NAME_TYPE_SUITE vs Target

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Children	340	31	371
Family	4724	406	5130
Group of people	26	1	27
Other_A	113	10	123
Other_B	187	26	213
Spouse, partner	1429	132	1561
Unaccompanied	30733	2917	33650
Grand Total	**37552**	**3523**	**41075**



Column Chart explains-

Majority of clients belong to Unaccompanied group followed by family and Ratio of default counts with Total number of clients is close,

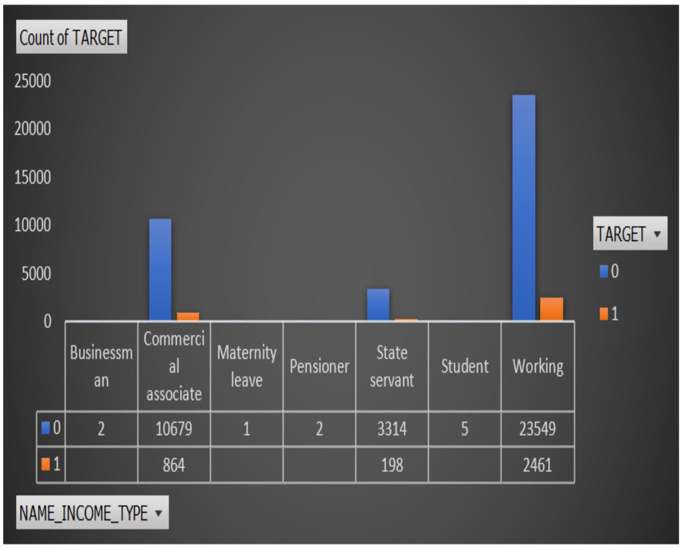
For Family = (Default Count (406)/Total Count(4724))*100 **= 8.59%**

For Unaccompanied = (Default Count (2917)/Total Count(30733))*100 **= 9.49%**

Unaccompanied type suite has more chances of defaulting.

5) NAME_INCOME_TYPE vs Target

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Businessman	2		2
Commercial associate	10679	864	11543
Maternity leave	1		1
Pensioner	2		2
State servant	3314	198	3512
Student	5		5
Working	23549	2461	26010
Grand Total	37552	3523	41075

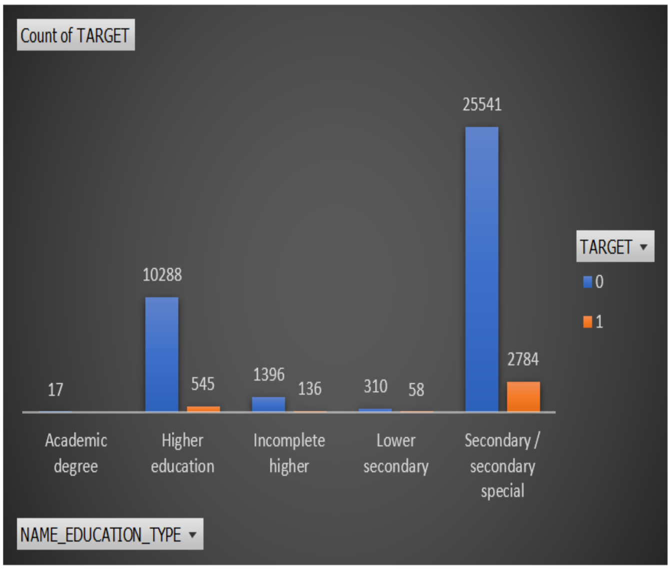


Businessman type have no defaults, thus safe for giving loan

Working people may default but % of default (<1%) is too low.

6) NAME_EDUCATION_TYPE vs Target

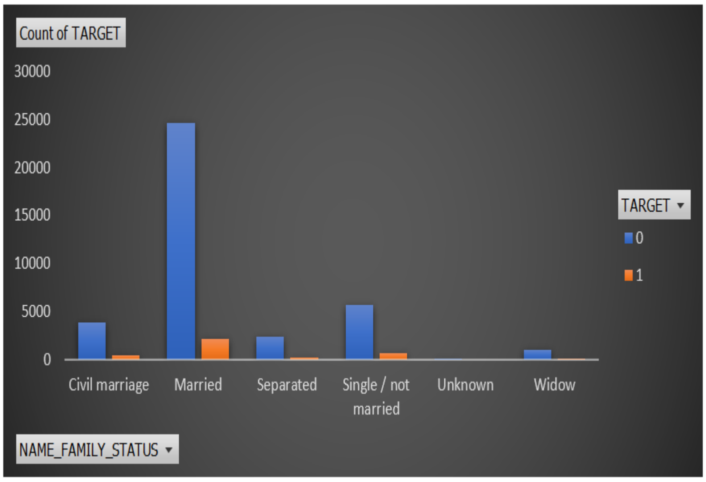
Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Academic degree	17		17
Higher education	10288	545	10833
Incomplete higher	1396	136	1532
Lower secondary	310	58	368
Secondary / secondary special	25541	2784	28325
Grand Total	37552	3523	41075

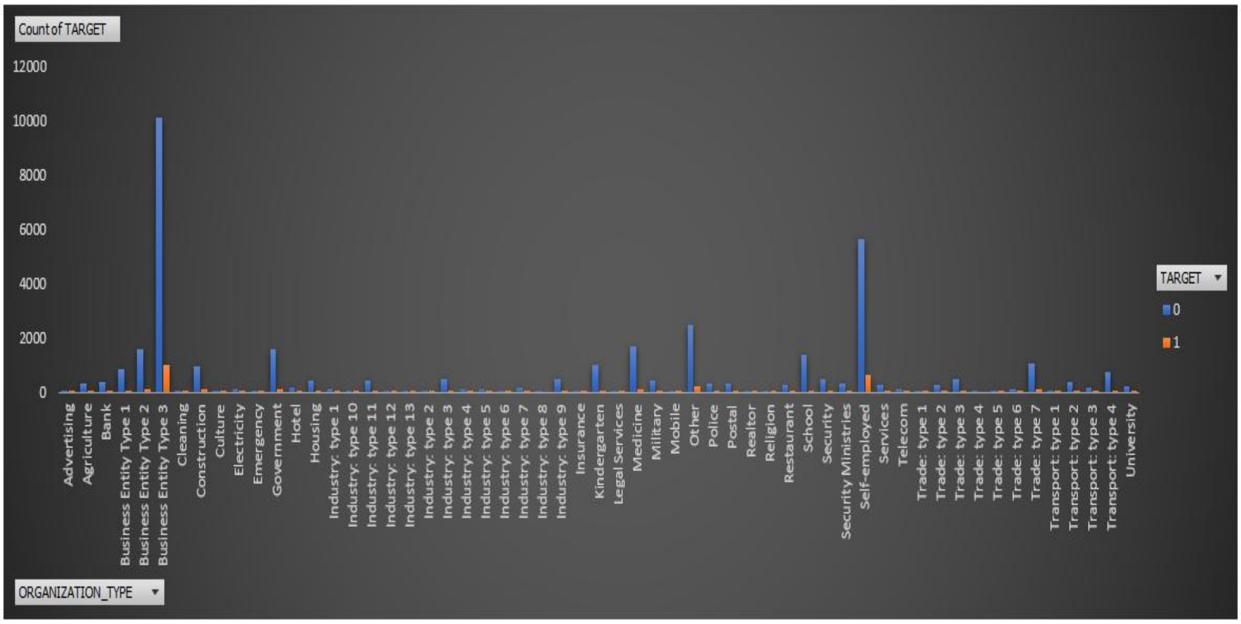


Defaulting % for both higher education and secondary special with respect to credit is <1%.

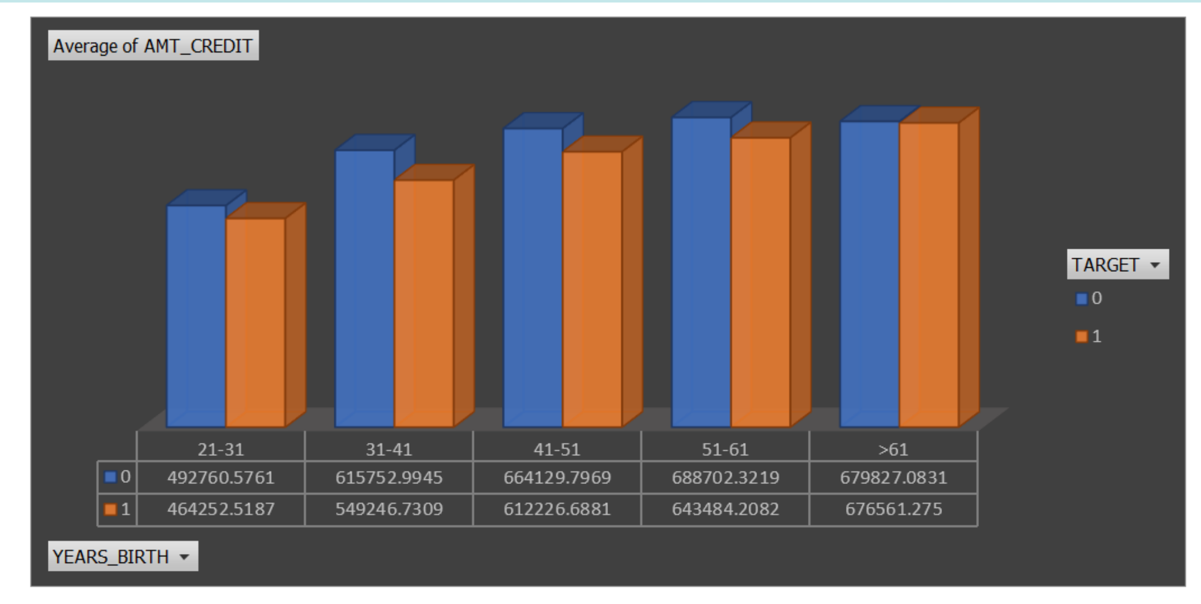
7) Next two column charts show Family_status and Organization distribution wrt Target.

Count of TARGET	Column Labels		
Row Labels	0	1	Grand Total
Civil marriage	3824	453	4277
Married	24658	2105	26763
Separated	2353	225	2578
Single / not married	5680	673	6353
Unknown	1		1
Widow	1036	67	1103
Grand Total	37552	3523	41075

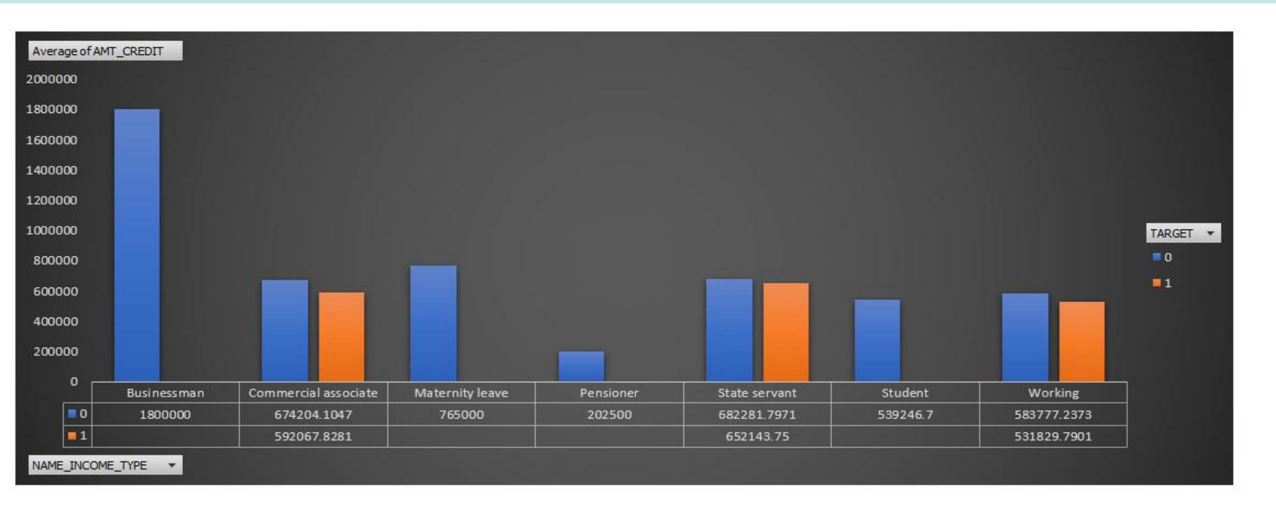




8) AMT_CREDIT vs YEARS_BIRTH wrt Target



9)AMT_CREDIT vs NAME_INCOME_TYPE vs Target



Task E. Identify Top Correlations for Different Scenarios:

Correlation Coefficient calculated for important numerical columns wrt Target and the key takeaways were this –

Variables	Correlation Coefficient with Target
AMT_INCOME_TOTAL	0.010397979
AMT_CREDIT	-0.044691819
YEARS_BIRTH	-0.066971531
YEARS_EMPLOYED	-0.076540348
REGION_RATING_CLIENT	0.073869421

Key Takeaways:

- A value close to **0** (like 0.0104) indicates that **income levels do not significantly impact loan default risk**.
Whether an applicant has a high or low income, their chances of defaulting on the loan are **almost independent** of income.

- The slightly **negative value (-0.0447)** suggests that **as loan amounts increase, the likelihood of default (TARGET = 1) decreases, but only slightly**.

- The slightly **negative correlation (-0.0670)** suggests that **as age increases, default risk decreases— but only slightly**.

- The slightly **negative correlation (-0.0765)** suggests that **longer employment tenure slightly reduces default risk**, but not significantly.

- The **slightly positive correlation (0.0739)** suggests that individuals **from regions with higher ratings may have a slightly increased tendency to default**, but the effect is very small.

CONCLUSION

Loan Default Insights

- **Only 9% of applicants are defaulters**, while **91% repay on time**.
- **Cash loans dominate (90%)**, with revolving loans being far less common (10%).
- **Females represent 62% of applicants**, but have a **slightly higher default rate than males**.
- **Laborers make up a significant portion of loan applicants**, indicating that blue-collar workers frequently seek loans.
- **Business owners are the safest borrowers**, with **zero recorded defaults**.

Key Predictive Patterns

- **Income has little impact on loan defaults** (**correlation = 0.0104**), meaning **high-income borrowers can still default**.
- **Loan amount has a weak negative correlation (-0.0447)** with default risk—larger loans seem **slightly less risky**.
- **Age has a weak negative effect (-0.0670)—older applicants default less frequently**.
- **Longer employment tenure slightly reduces defaults (-0.0765)**, suggesting **job stability improves repayment likelihood**.
- **Region rating has a weak positive impact (0.0739)—higher-rated regions show slightly more defaults**, though the effect is minimal.

Segmented Analysis Findings

- **Younger borrowers (21–41 years) have higher default tendencies**.
- **Shorter work experience is linked to more defaults**.
- **Unaccompanied applicants have a higher default rate (9.49%)**, compared to **families (8.59%)**.
- **Higher education applicants default less** (<1%).**Final Takeaways**

- **Employment history, age, and loan amount** are more influential than income alone.
- **Family applicants default less than unaccompanied borrowers**.
- **Certain occupational groups pose higher risks**, requiring careful assessment for loan approvals.

- **As I have made multiple workbooks while cleaning and analysing , all the workbooks will be in a folder over the link mentioned below:**

Google Drive Link for excel sheets : workbook link

Thank you.