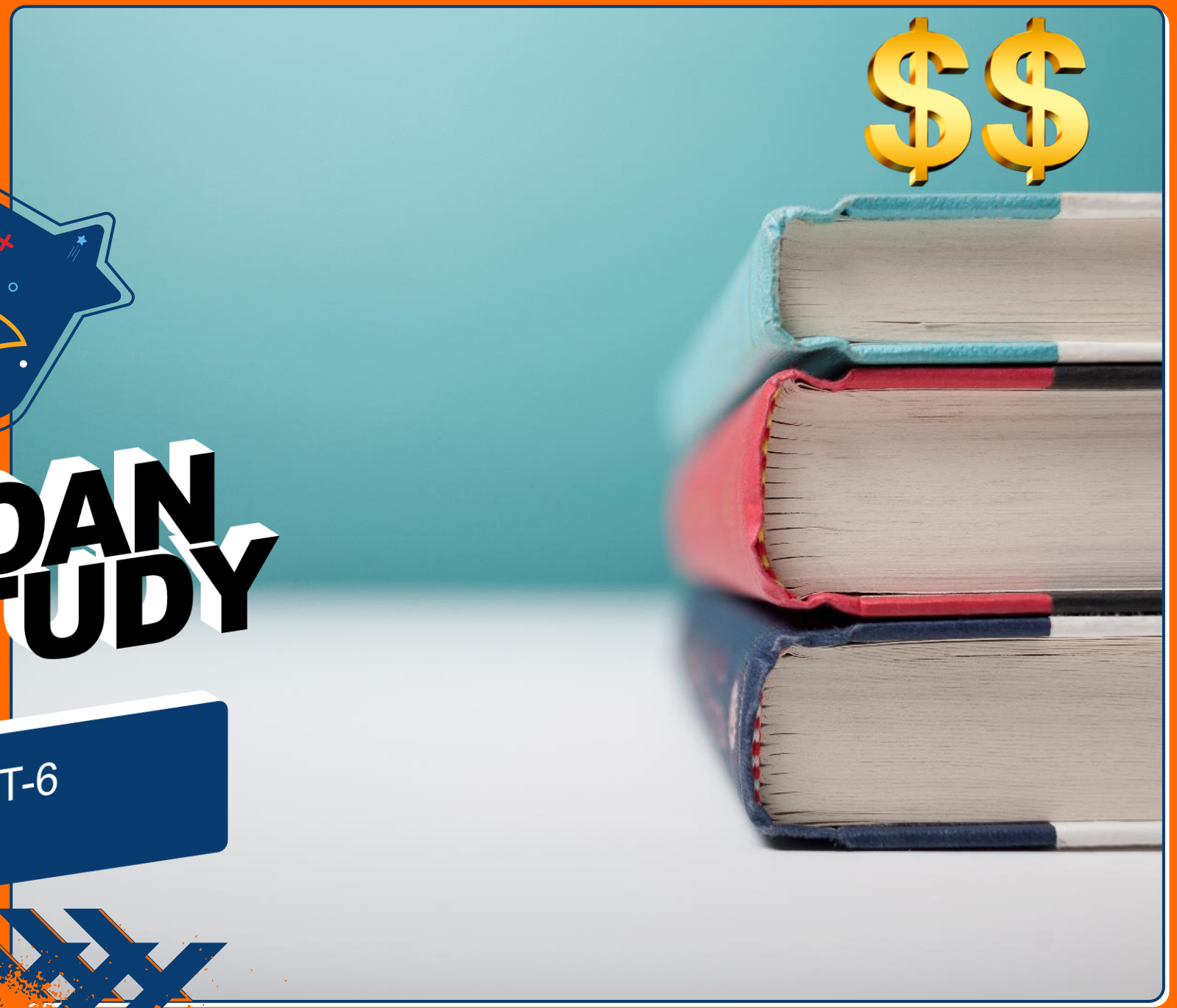


BANK LOAN CASE STUDY

TRAINITY PROJECT-6

CONTENT

- DESCRIPTION
- APPROACH
- TECH-STACK USED
- INSIGHTS
- RESULT

PROJECT DESCRIPTION

In this project, I aim to analyze loan application data to identify patterns that indicate the likelihood of loan default. Our company specializes in providing various types of loans to urban customers, and we face two main risks:

1. Losing business if we reject capable applicants who can repay the loan.
2. Incurring financial losses if we approve applicants who later default on their loans.

To address these challenges, I will use Exploratory Data Analysis (EDA) to gain insights into customer attributes and loan characteristics that influence default rates. The dataset includes information on loan applications, with outcomes such as approved, cancelled, refused, and unused offers. Additionally, it distinguishes between customers who have had payment difficulties and those who have not. The primary goal is to identify key factors that predict loan default. This will enable us to make informed decisions about loan approvals, such as denying risky applications, adjusting loan amounts, or applying higher interest rates for high-risk applicants

APPROACH:

- Import both the dataset into MS-Excel.
- Observed & understood the data to make plan of action for further analysis.
- I cleaned the data and handled missing data carefully by filling mean/median/mode values so that data become consistent while analyzing.
- Further, I used EDA(Exploratory Data analysis) techniques to performed univariate, bivariate and segmented univariate analysis.
- Moreover, I used Excel's charting tools, Graphs to visually represent insights and patterns in clear and concise way to inform decision-making on loan application.

TECH-STACK USED

Microsoft Excel 2016, part of the Microsoft Office 2016 suite, is a powerful spreadsheet application widely used for data organization, analysis, and visualization. It supports large datasets and offers tools for sorting, filtering, and organizing information. Excel 2016 includes a wide range of built-in functions for calculations, advanced data analysis tools like Power Query and Power Pivot, and enhanced charting options such as bar, line, pie, waterfall, and treemap charts.

SOFTWARE: MICROSOFT EXCEL 2016

DATA CLEANING (APPLICATION_DATA)

COLUMN	ROWS	NULL COLUMN
16384	1048576	50

There are 50 columns having null value percentage of more than 30% are identified and deleted after that I started the EDA.

Also removed all the irrelevant rows and columns. here all the irrelevant columns are mentioned.

OWN_CAR_AGE
OCCUPATION_TYPE
EXT_SOURCE_1
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATIO
N_AVG YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AV
G NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BEGINEXPLUATATIO
N_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE

OWN_CAR_AGE
OCCUPATION_TYPE
EXT_SOURCE_1
APARTMENTS_AVG
BASEMENTAREA_AVG
YEARS_BEGINEXPLUATATIO
N_AVG YEARS_BUILD_AVG
COMMONAREA_AVG
ELEVATORS_AVG
ENTRANCES_AVG
FLOORSMAX_AVG
FLOORSMIN_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_AVG
NONLIVINGAPARTMENTS_AV
G NONLIVINGAREA_AVG
APARTMENTS_MODE
BASEMENTAREA_MODE
YEARS_BEGINEXPLUATATIO
N_MODE
YEARS_BUILD_MODE
COMMONAREA_MODE
ELEVATORS_MODE
ENTRANCES_MODE
FLOORSMAX_MODE
FLOORSMIN_MODE
LANDAREA_MODE

DATA CLEANING (PRIEVIOUS_DATA)

- similarly in previous data 11 columns having null value percentage of more than 30% are identified and deleted after that I started the EDA. Also remove all the irrelevant rows and columns

RATE_INTEREST_PRIVILEGED
NAME_TYPE_SUITE
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSIO
N DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROV
AL AMT_DOWN_PAYMENT
RATE_DOWN_PAYMENT
RATE_INTEREST_PRIMARY

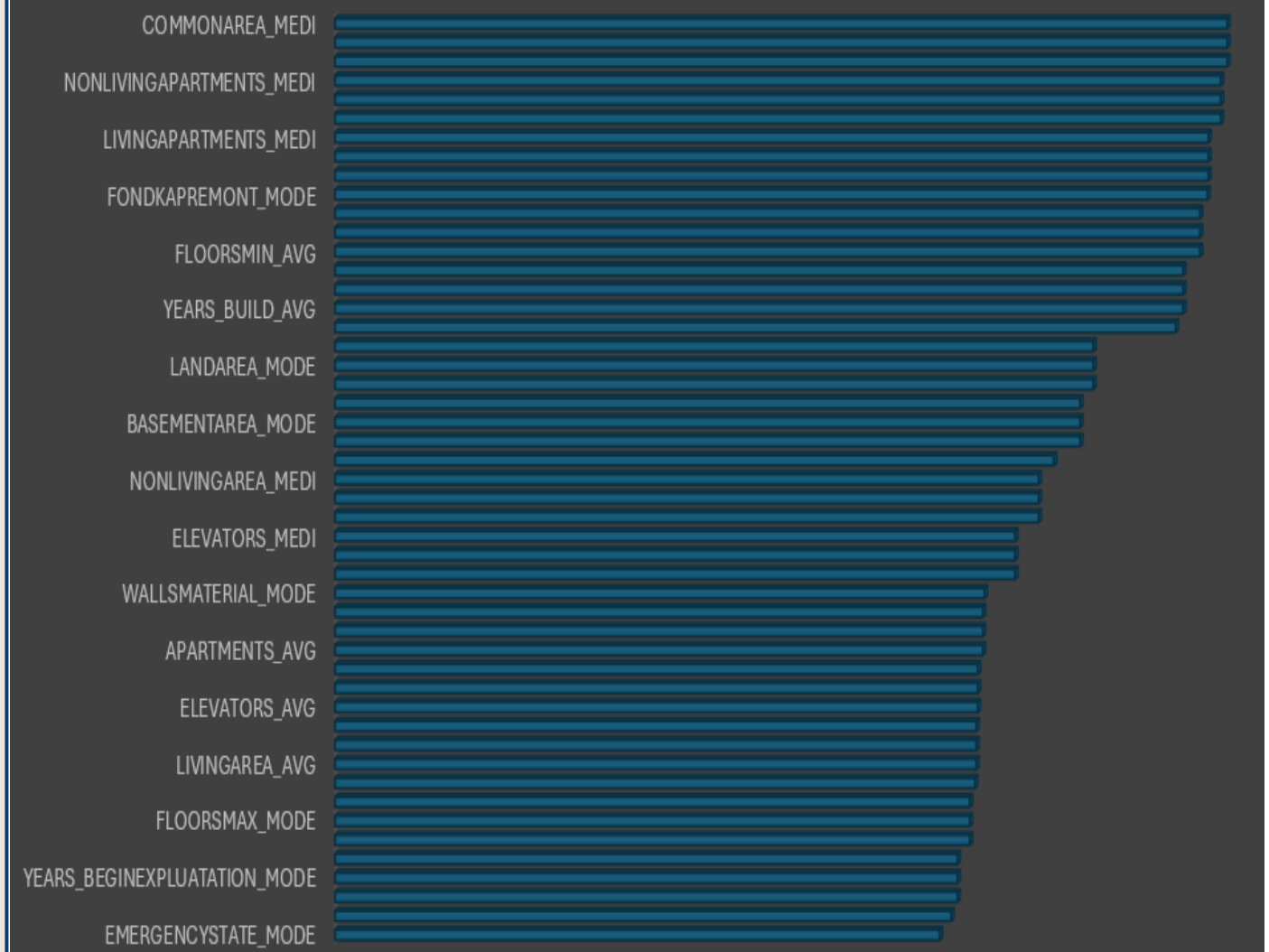
A. IDENTIFY MISSING DATA AND DEAL WITH IT APPROPRIATELY

- Using COUNTA function and arithmetic function I calculated the percentage of null values in each columns.

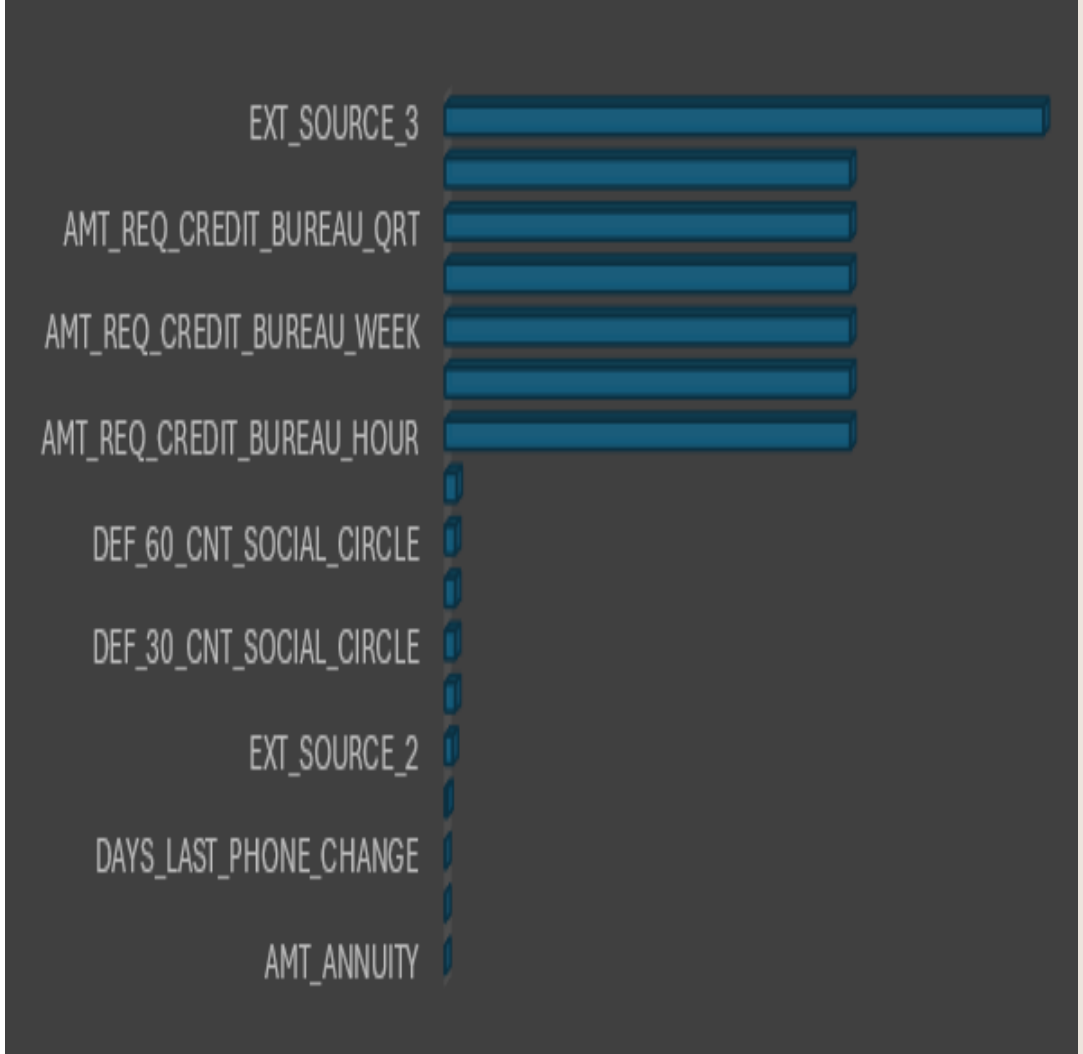
- Then I removed the all columns with more than 30% null values. And filled the null value of the remaining columns with its median or mode depending upon the data type.

- I changed the three columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION from days to years using ABS function, keeping only the variable important for drawing the conclusion.

COLUMNS WITH NULL VALUE > 30%



COLUMNS WITH NULL VALUE<30%



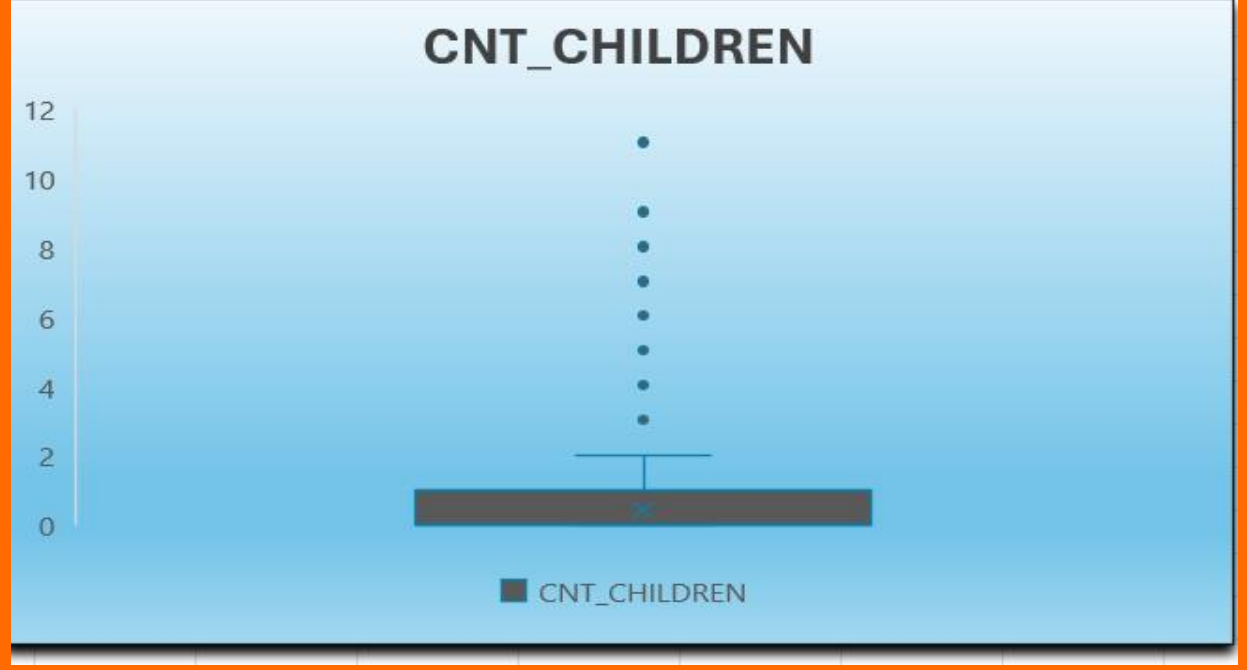
B. IDENTIFY OUTLIERS IN THE DATASET

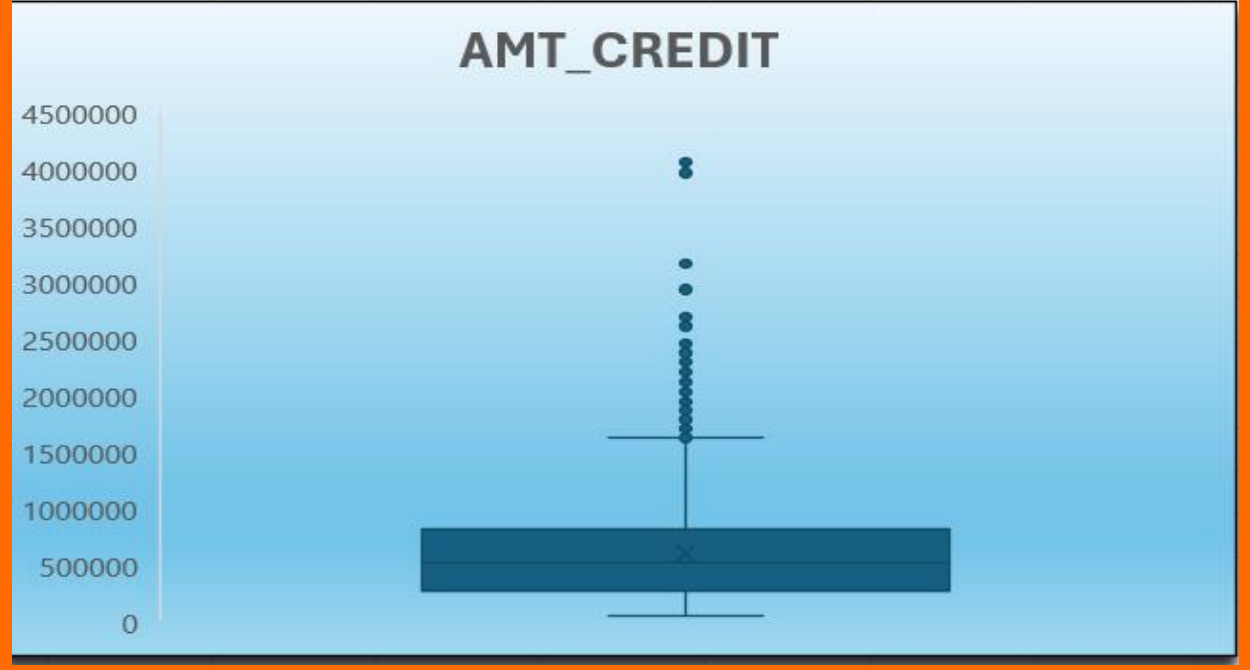
- An outlier is a single data point that goes far outside the average value of a group of statistics. Outliers may be exceptions that stand outside individual samples of populations as well. In a more general context, an outlier is an individual that is markedly different from the norm in some respect.

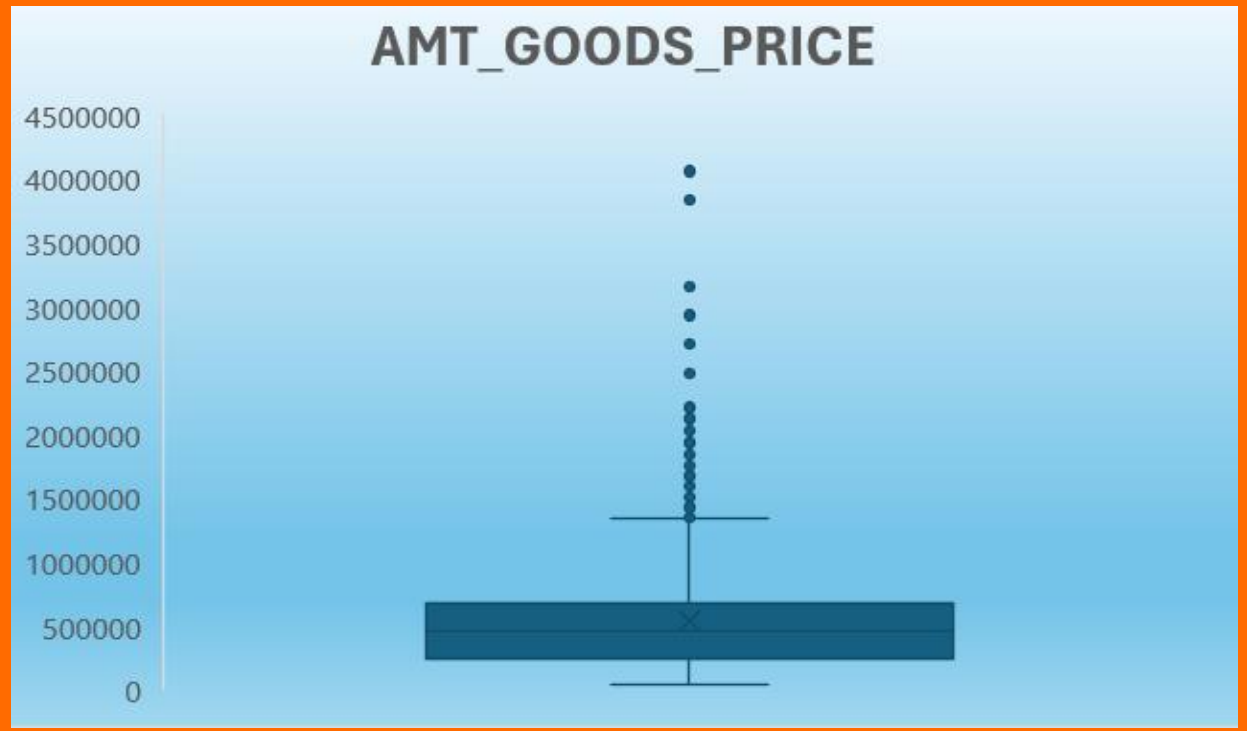
    - I found the presence of a few outliers in DAYS_DECISION which indicated the time taken for the decision was too high. Which is not a good sign for any business.

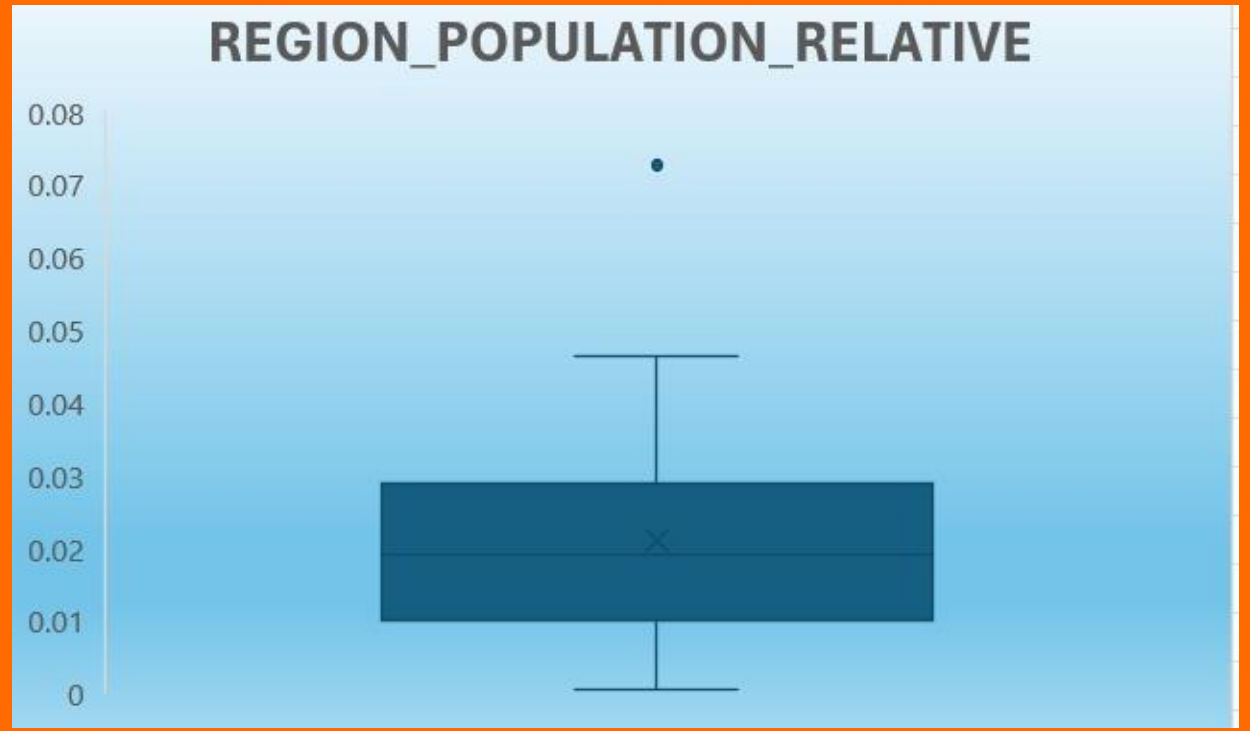
    - Secondly, I found a large number of outliers in AMT_GOODS_PRICE, AMT_APPLICATION, AMT_CREDIT AND AMT_ANNUITY.

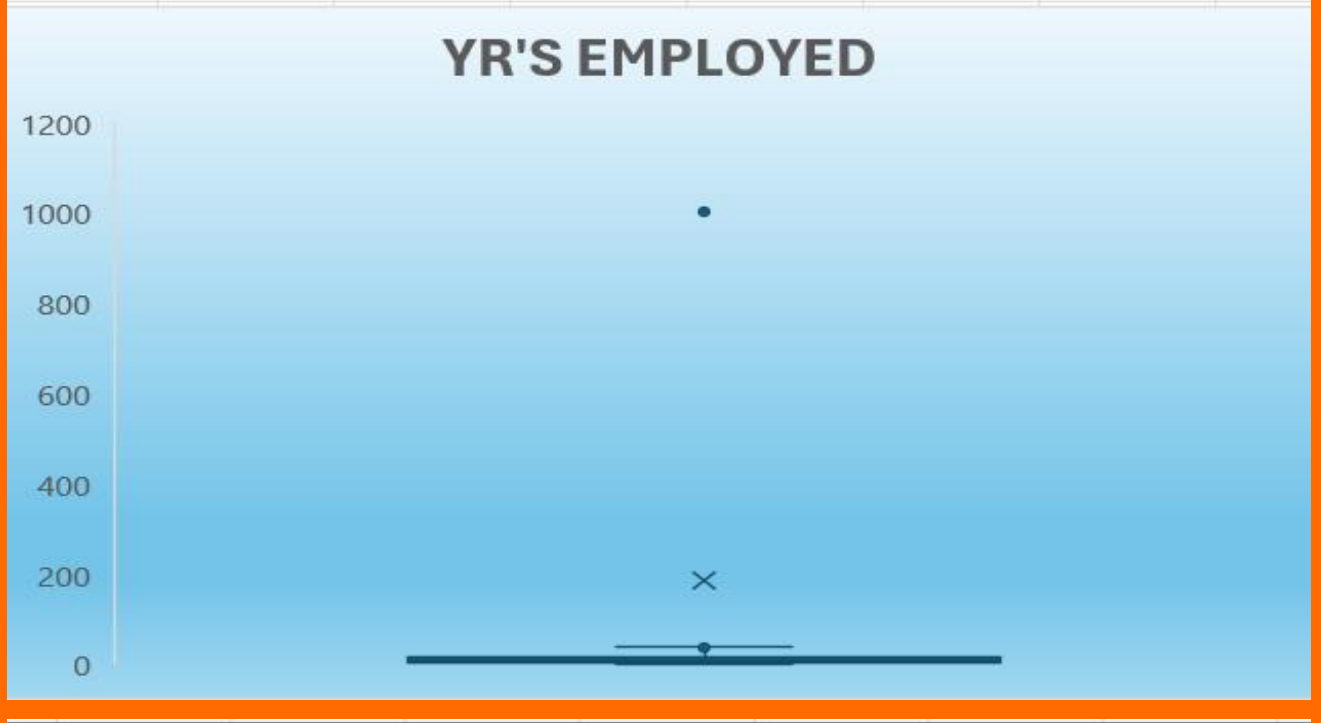
    - I found a few outliers in CNT_PAYMTENT.

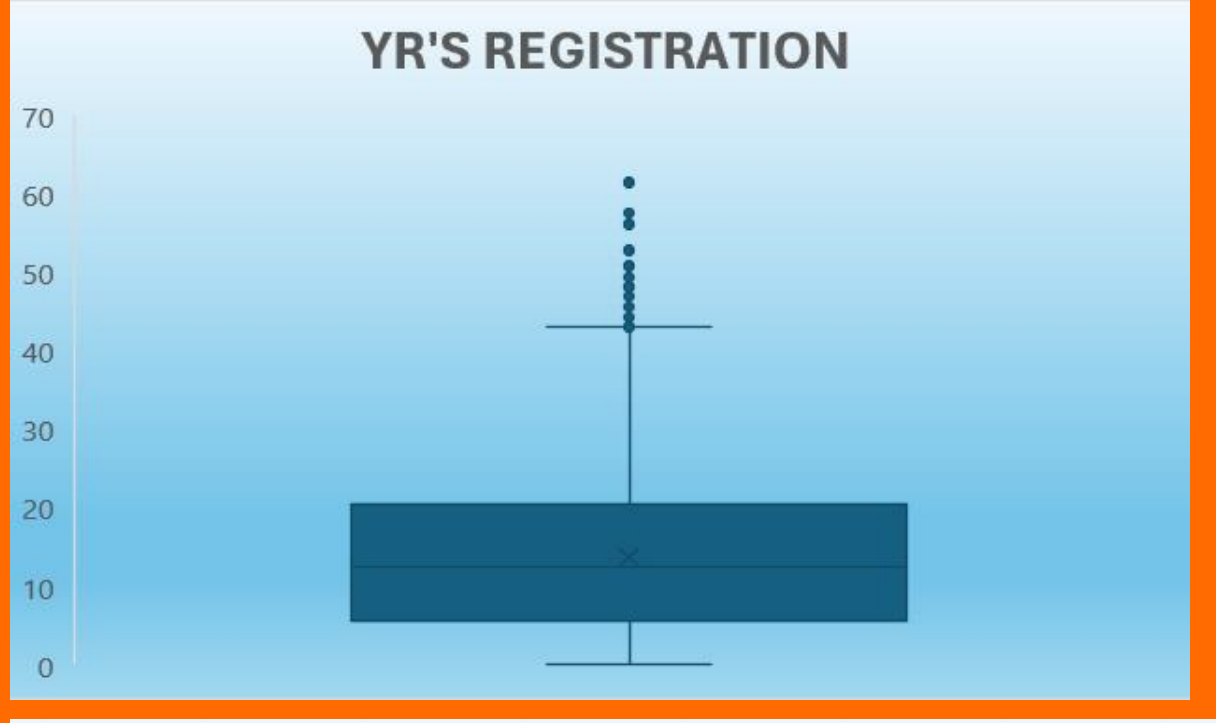

CNT_CHILDREN
12
10
8
6
4
2
0
CNT_CHILDREN

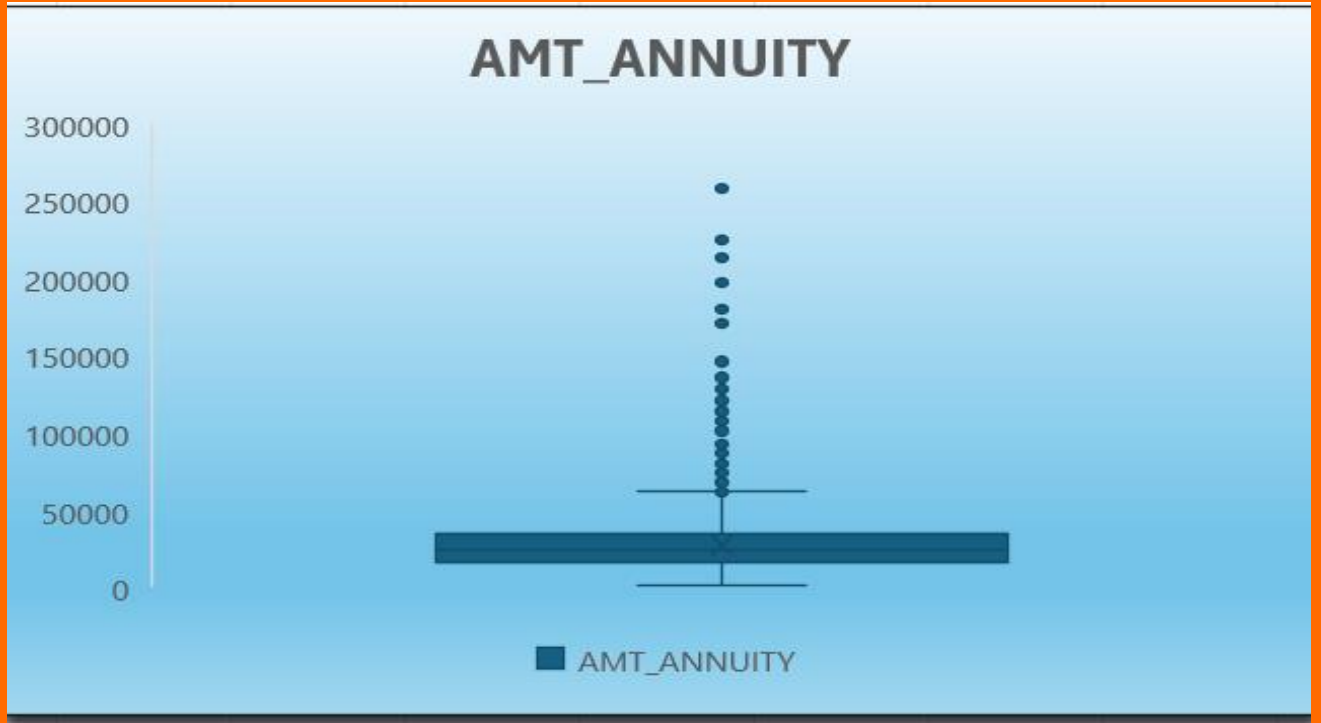

AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0

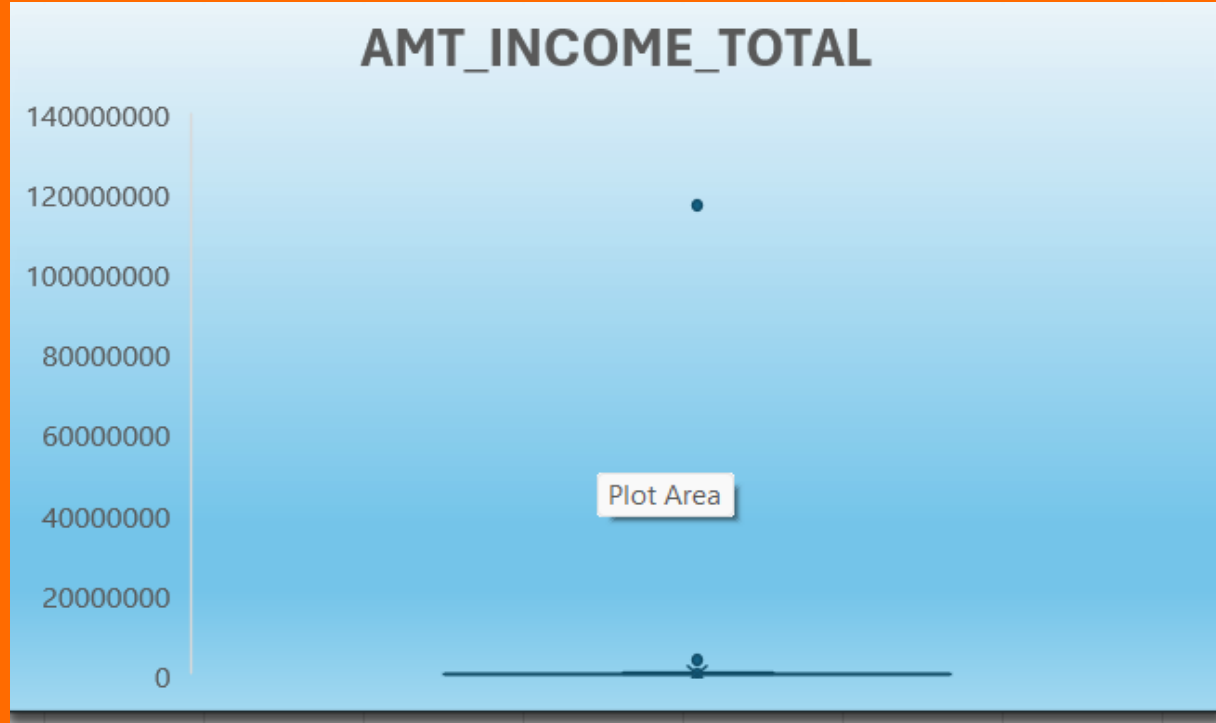

AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0


REGION_POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0


YR'S EMPLOYED
1200
1000
800
600
400
200
0

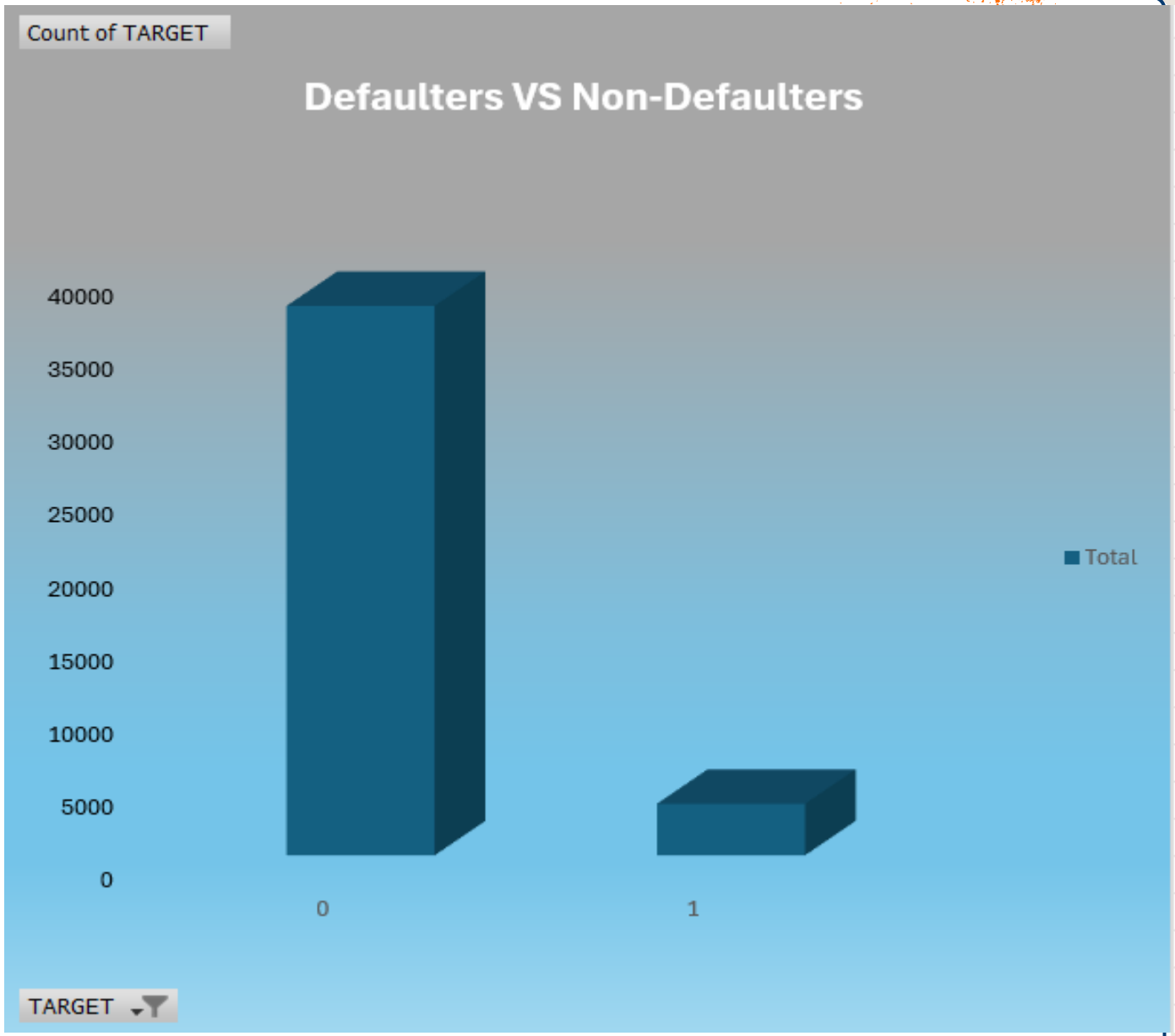

YR'S REGISTRATION
70
60
50
40
30
20
10
0


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
AMT_ANNUITY

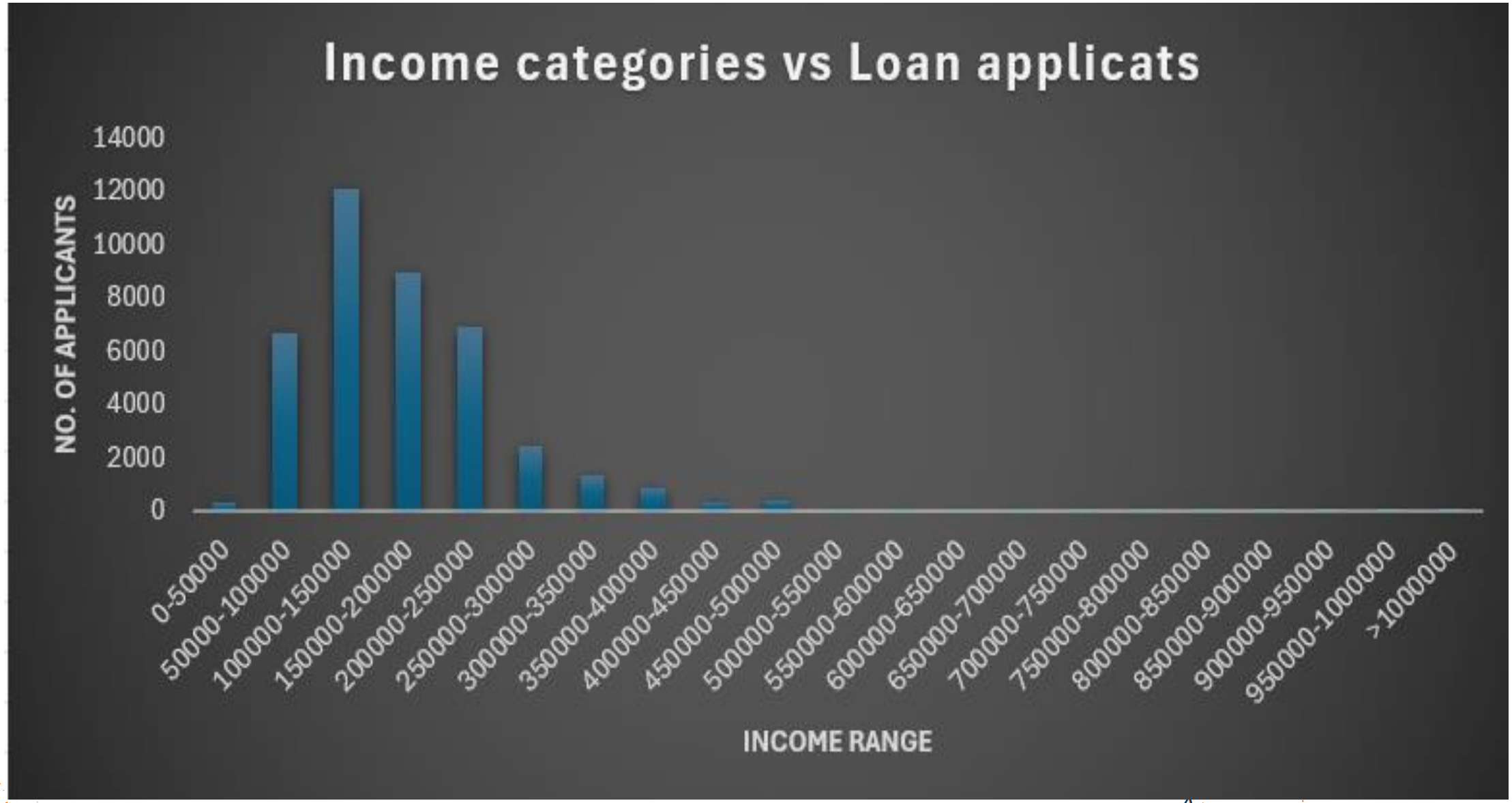

AMT_INCOME_TOTAL
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
Plot Area

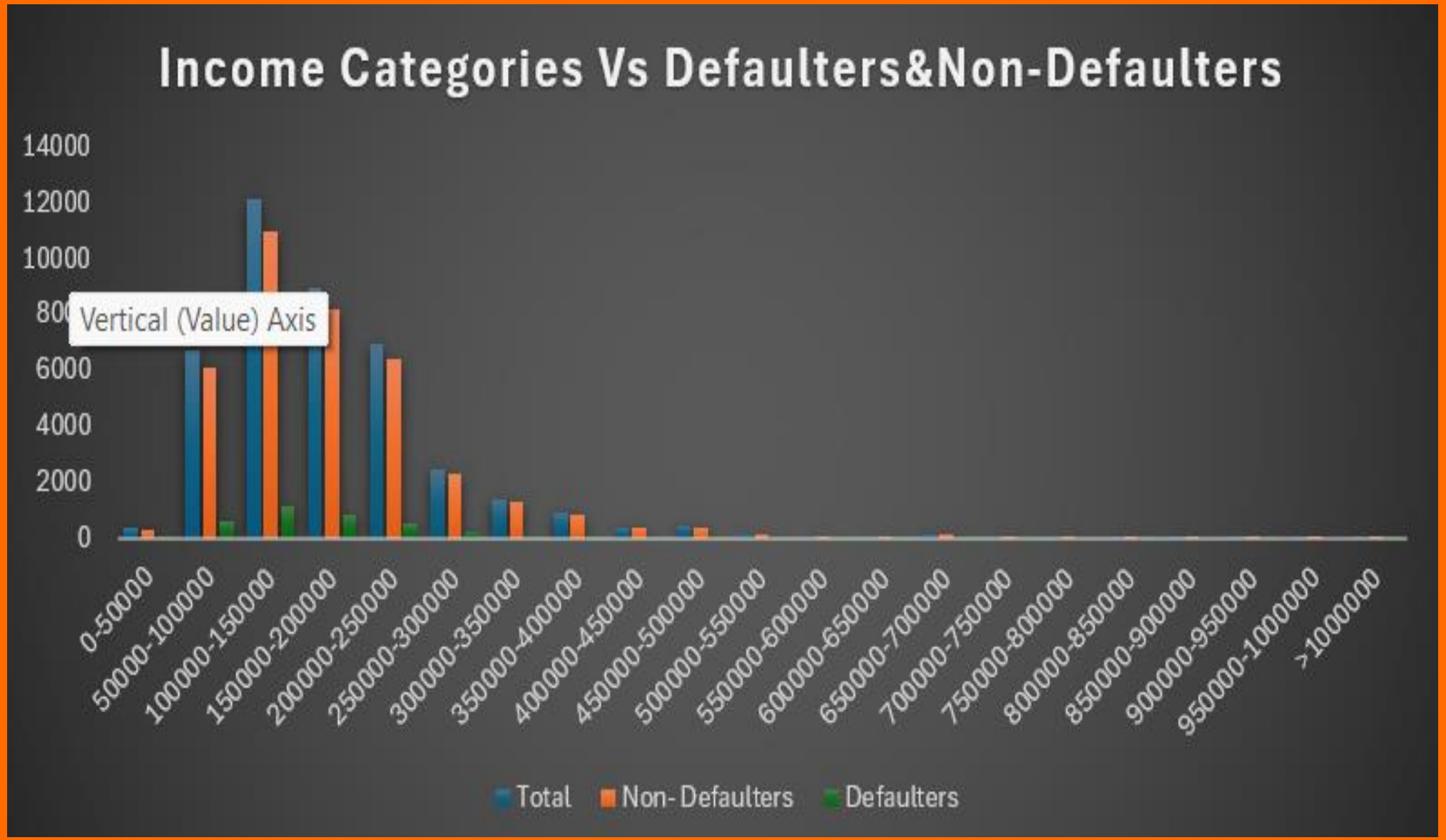
C. ANALYZE DATA IMBALANCE

- Here We see that, we have 91% as loan re-payers,
9% as Defaulters Which gives us a clear indication
that the data is highly imbalanced



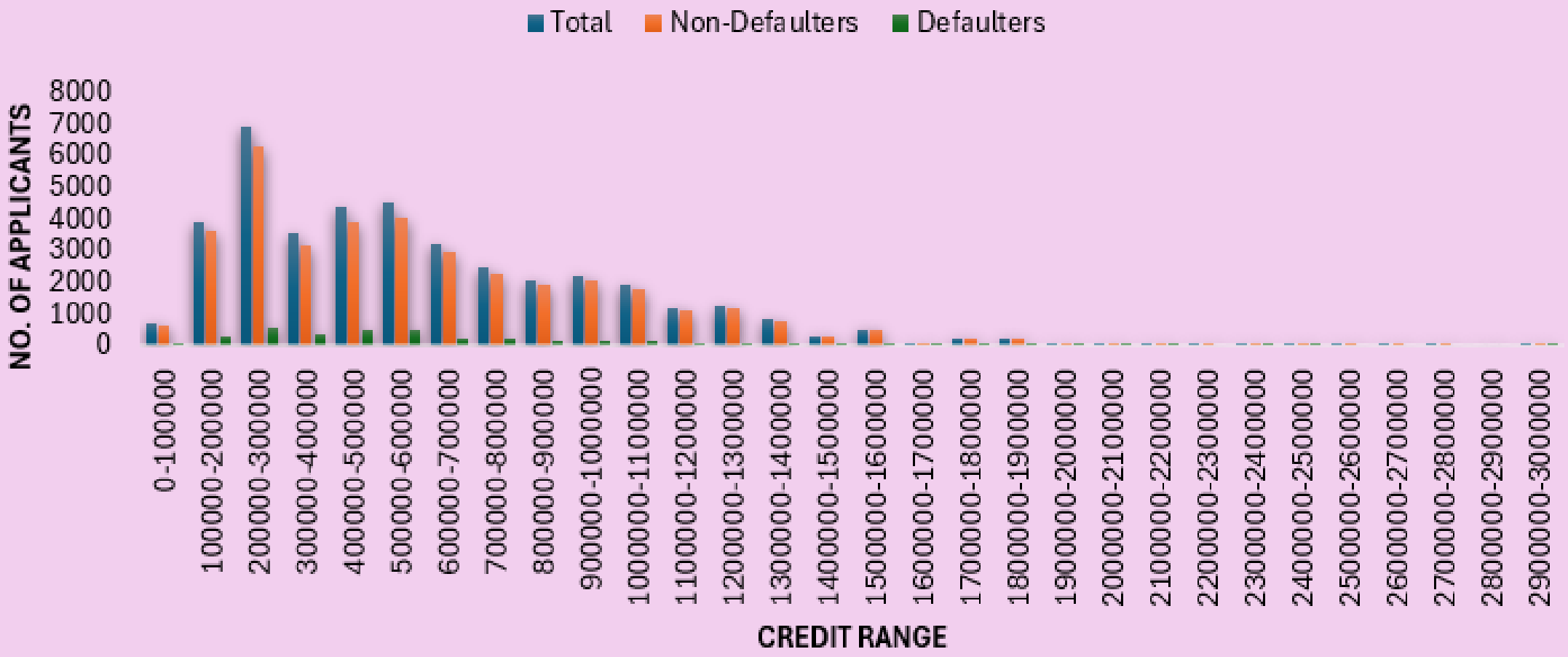
D) Perform Univariate, Segmented Univariate and Bivariate


Income categories vs Loan applicats
NO. OF APPLICANTS
14000
12000
10000
8000
6000
4000
2000
0
0-50000
50000-100000
100000-150000
150000-200000
200000-250000
250000-300000
300000-350000
350000-400000
400000-450000
450000-500000
500000-550000
550000-600000
600000-650000
650000-700000
700000-750000
750000-800000
800000-850000
850000-900000
900000-950000
950000-1000000
>1000000
INCOME RANGE

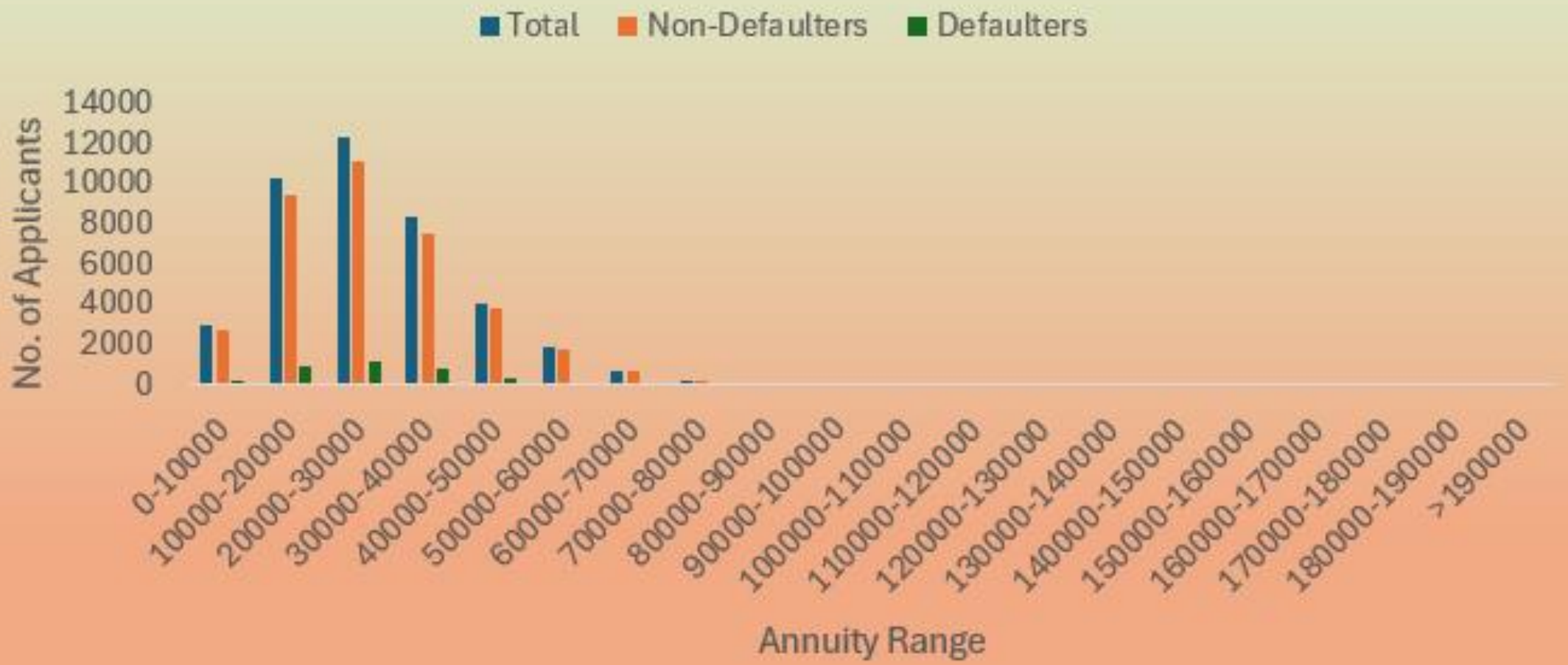

Income Categories Vs Defaulters&Non-Defaulters
14000
12000
10000
Vertical (Value) Axis
6000
4000
2000
0
0-50000
50000-100000
100000-150000
150000-200000
200000-250000
250000-300000
300000-350000
350000-400000
400000-450000
450000-500000
500000-550000
550000-600000
600000-650000
650000-700000
700000-750000
750000-800000
800000-850000
850000-900000
900000-950000
950000-1000000
>1000000
Total
Non-Defaulters
Defaulters

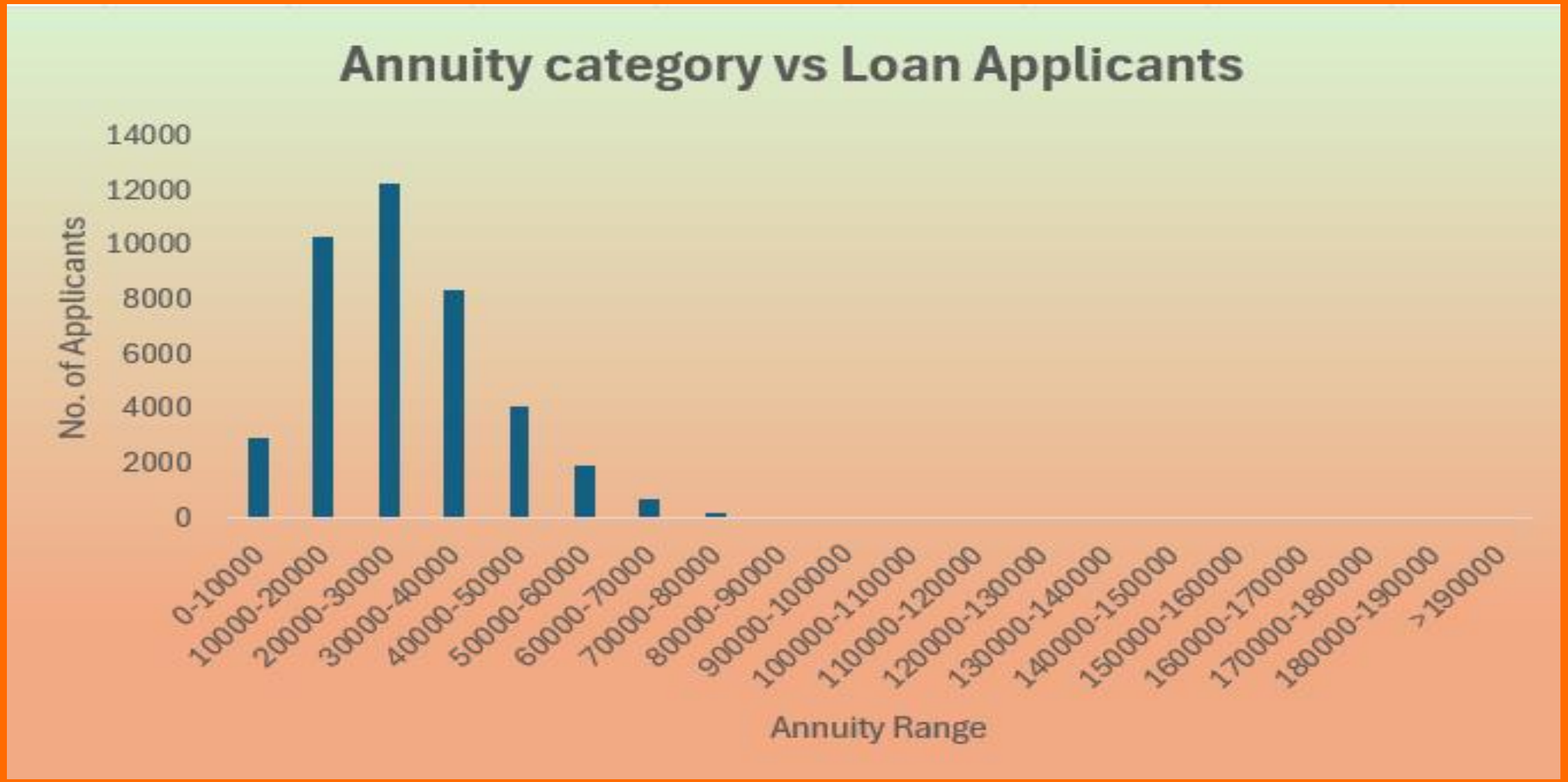
Credit Category vs Loan Applicants

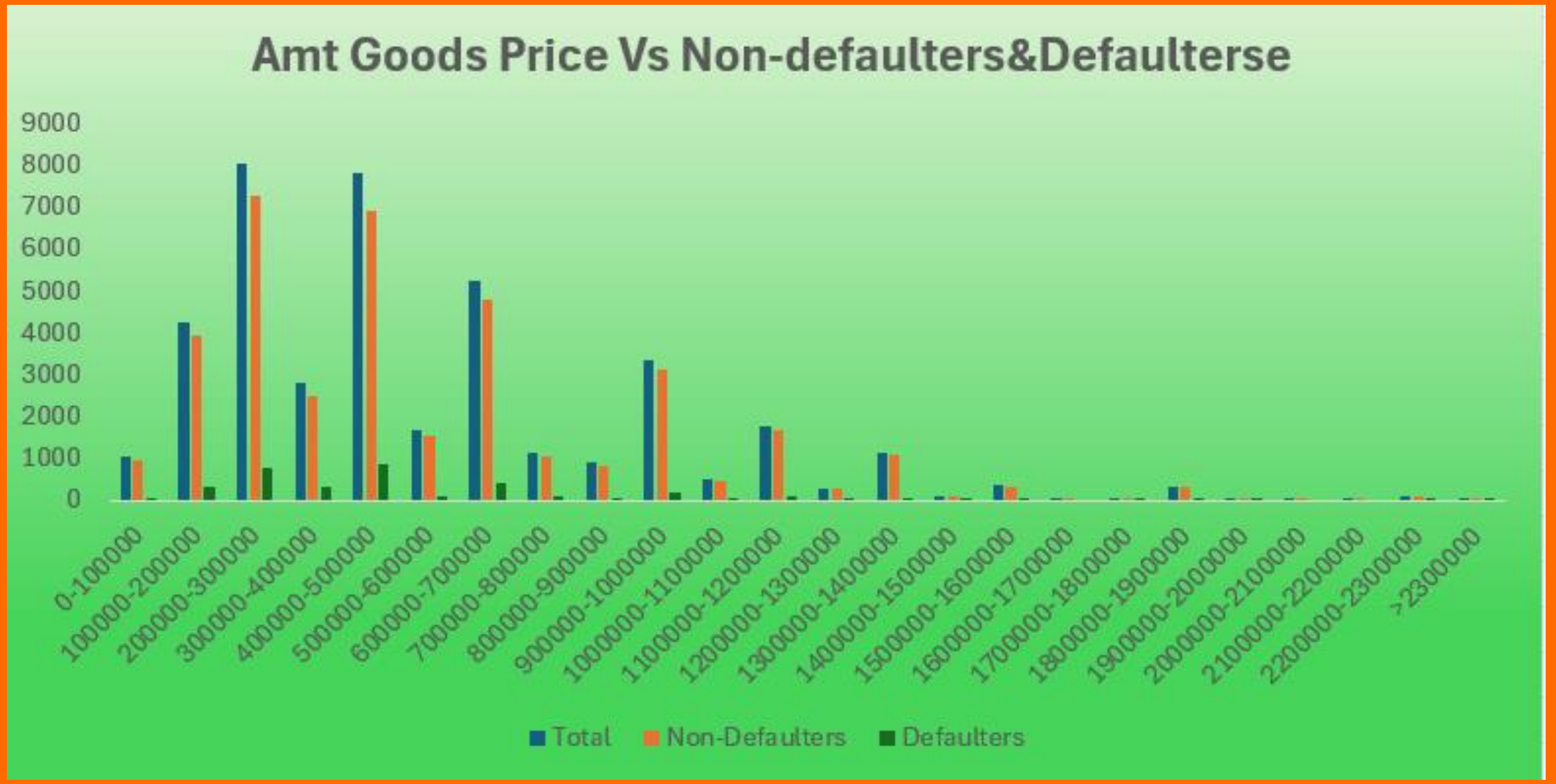
Credit Category Vs Defaulters&Non-Defaulters



Annuity Category Vs Defaulters & Non-Defaulters

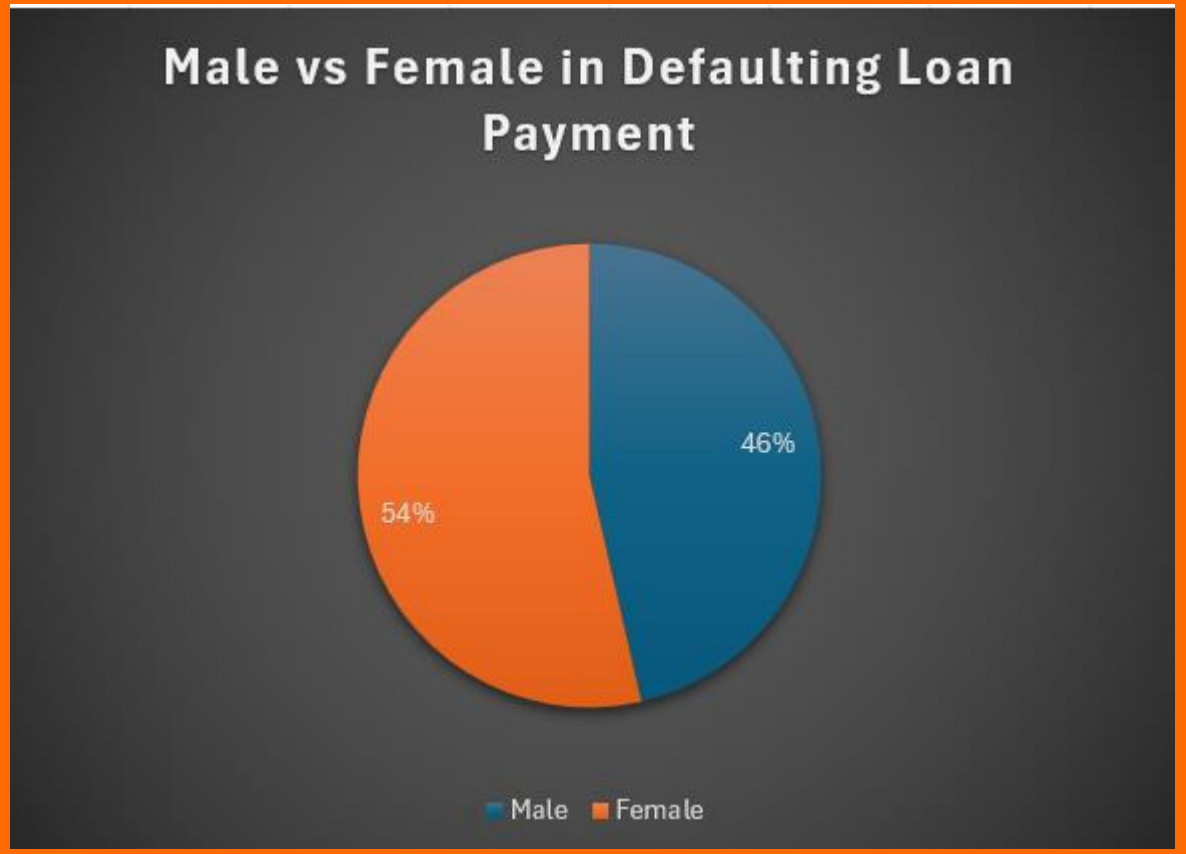



Annuity category vs Loan Applicants
No. of Applicants
14000
12000
10000
8000
6000
4000
2000
0
0-10000
10000-20000
20000-30000
30000-40000
40000-50000
50000-60000
60000-70000
70000-80000
80000-90000
90000-100000
100000-110000
110000-120000
120000-130000
130000-140000
140000-150000
150000-160000
160000-170000
170000-180000
180000-190000
>190000
Annuity Range


Amt Goods Price Vs Non-defaulters&Defaulterse
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0-100000
100000-200000
200000-300000
300000-400000
400000-500000
500000-600000
600000-700000
700000-800000
800000-900000
900000-1000000
1000000-1100000
1100000-1200000
1200000-1300000
1300000-1400000
1400000-1500000
1500000-1600000
1600000-1700000
1700000-1800000
1800000-1900000
1900000-2000000
2000000-2100000
2100000-2200000
2200000-2300000
>2300000
Total
Non-Defaulters
Defaulters

Gender Ratio Defaulters

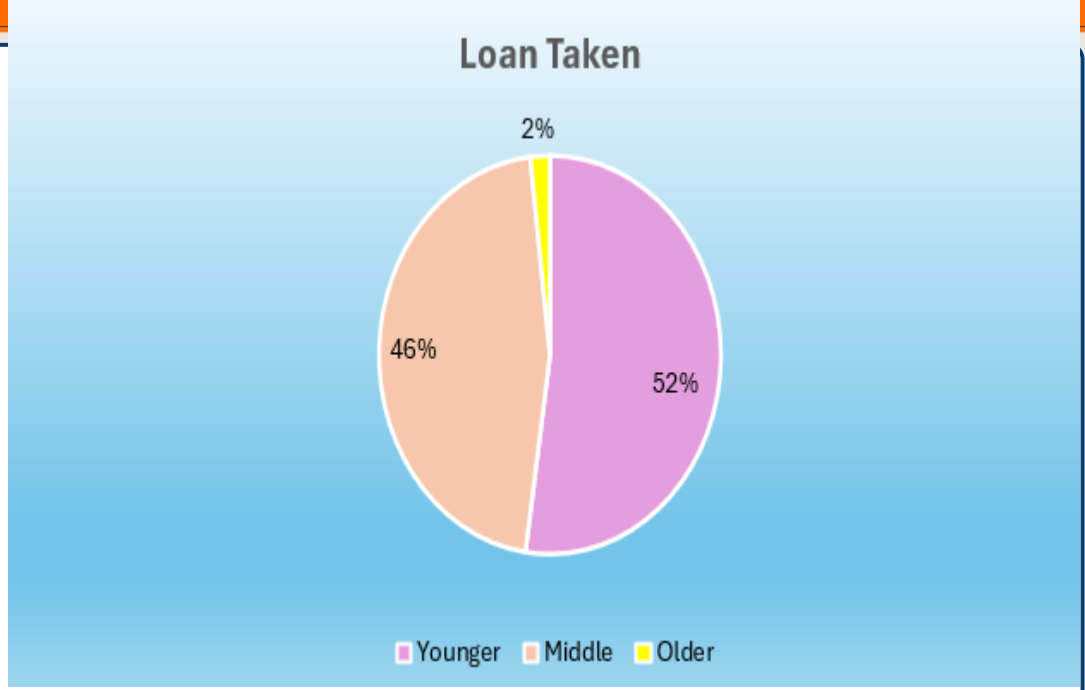
MALE			Female	
NON-DEFAULTER	DEFAULTER		Non-Defaulter	Defaulter
23648	1633		23648	1886
94%	6%		93%	7%
ratio of male to female				
Male	Female		Total	
1633	1886		3519	
46%	54%			

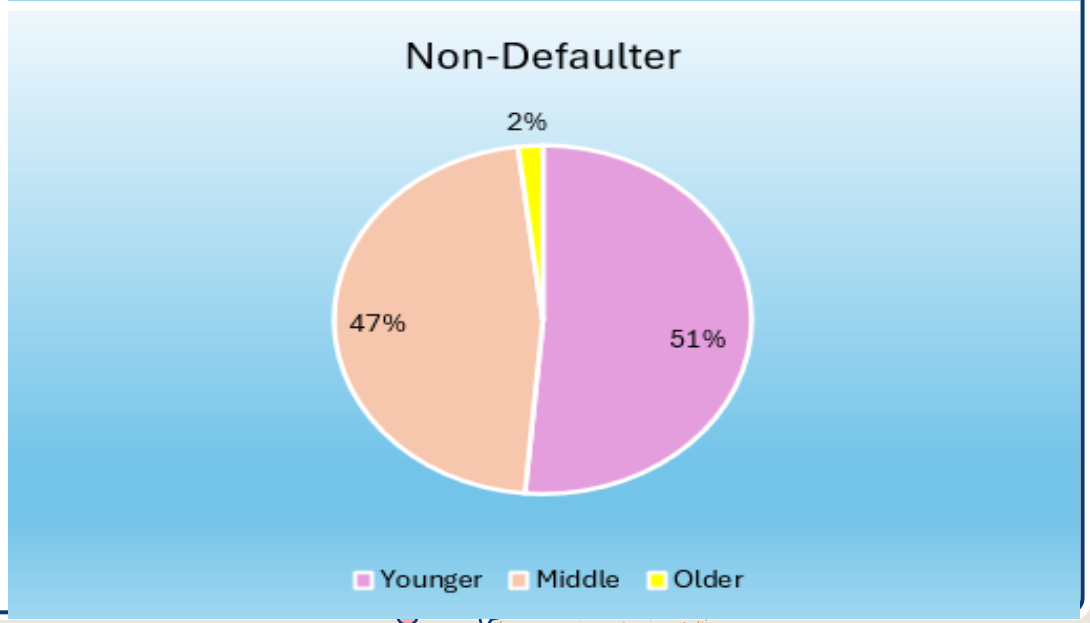


Age Factor

Age		Loan Taken	Non-Defaulter	Defaulter
20-40	Younger	20615	18502	2113
41-60	Middle	18073	16819	1254
>61	Older	728	688	40

- Most loans takers are of age 20-40
- Consider age as a factor in assessing loan default risk, with younger individuals potentially requiring closer scrutiny.
- Company might consider tailoring its loan term or interest rates based on age group to manage default risks effectively.

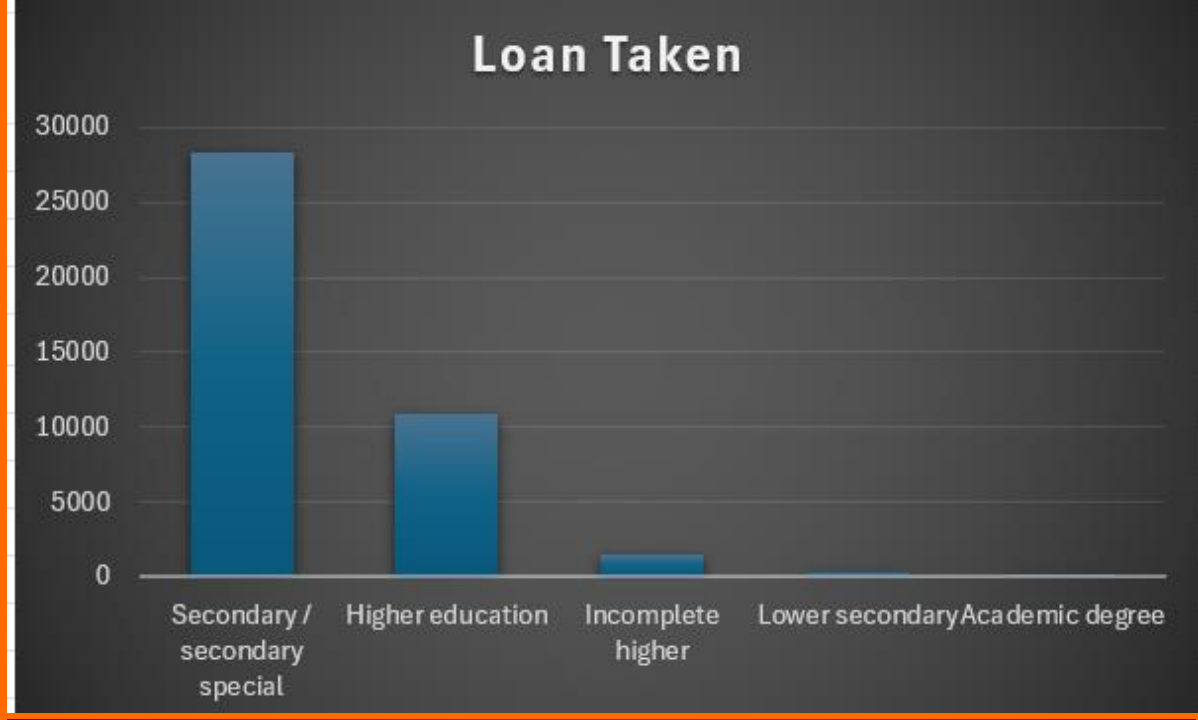


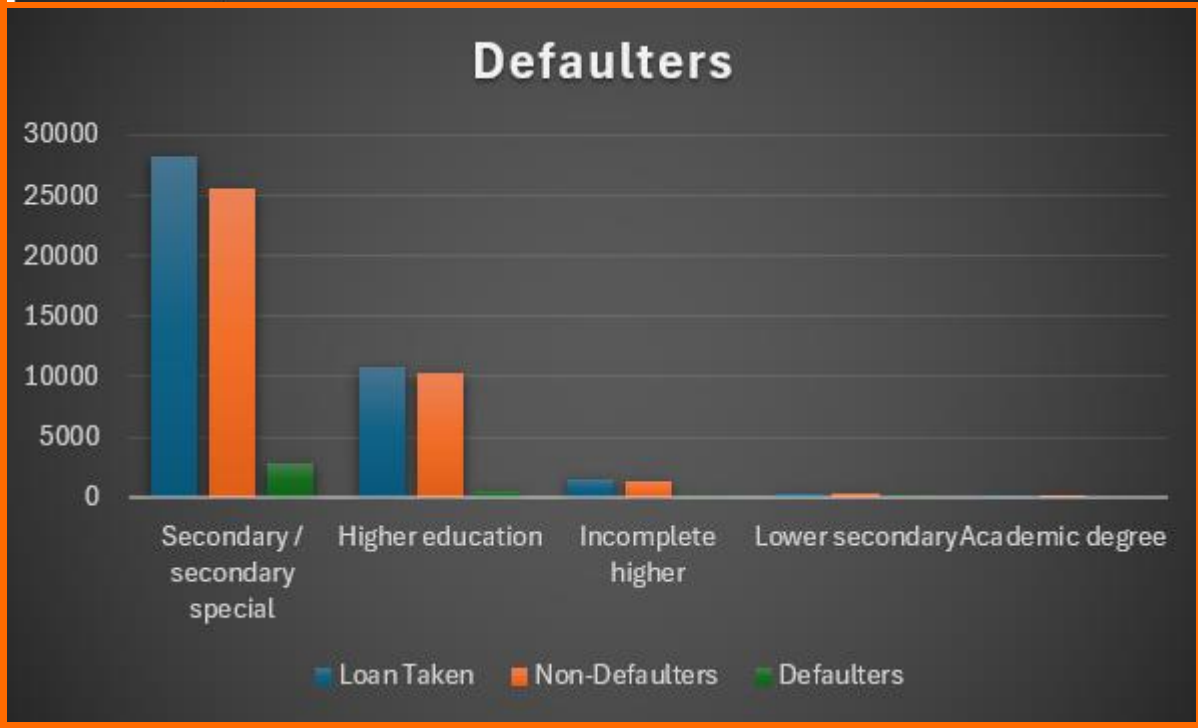


Education Factor

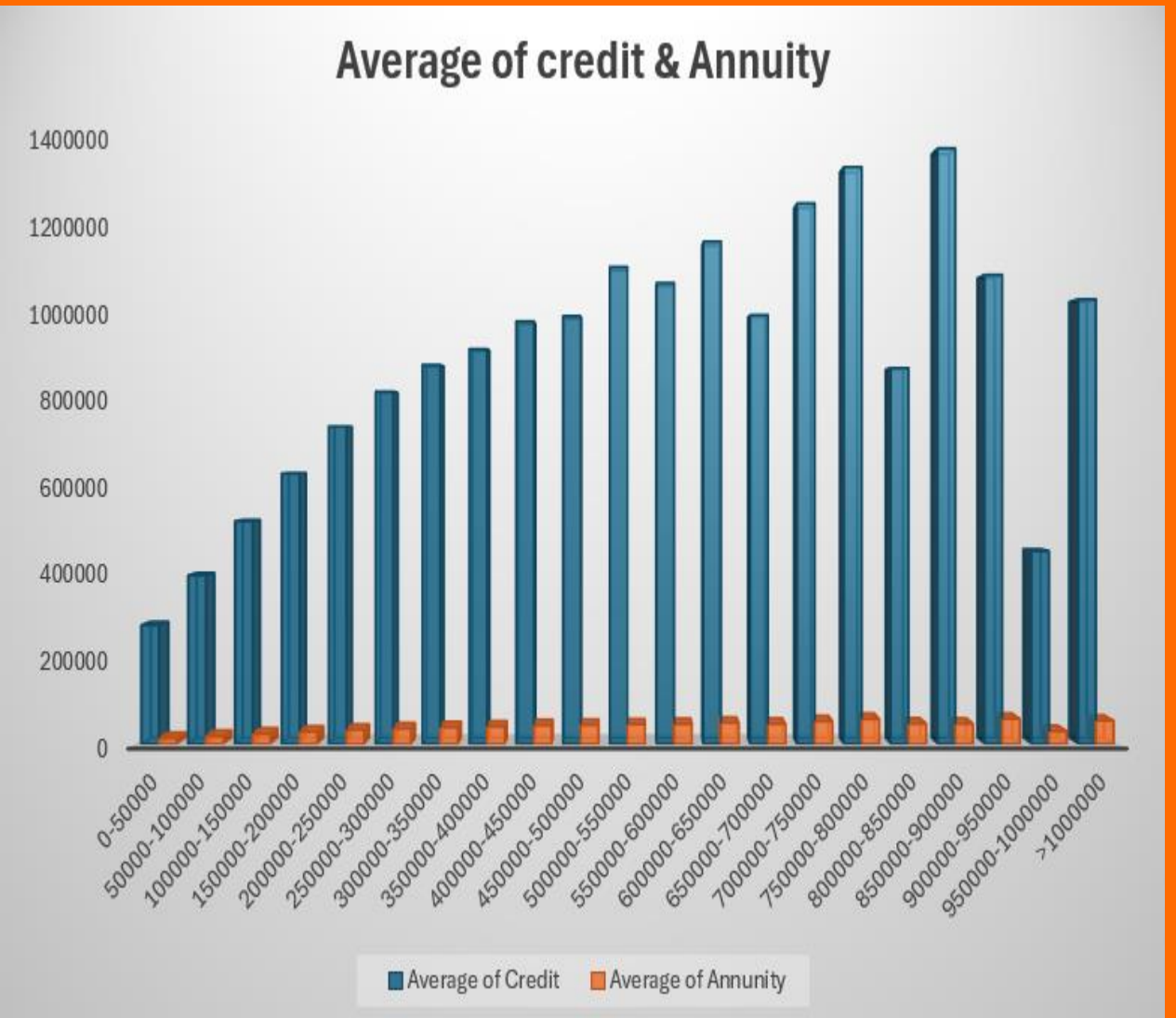
Education	Loan Take	Non-Defau	Defaulters
Secondary / secondary s	28322	25541	2781
Higher education	10828	10284	544
Incomplete higher	1532	1396	136
Lower secondary	368	310	58
Academic degree	17	17	0

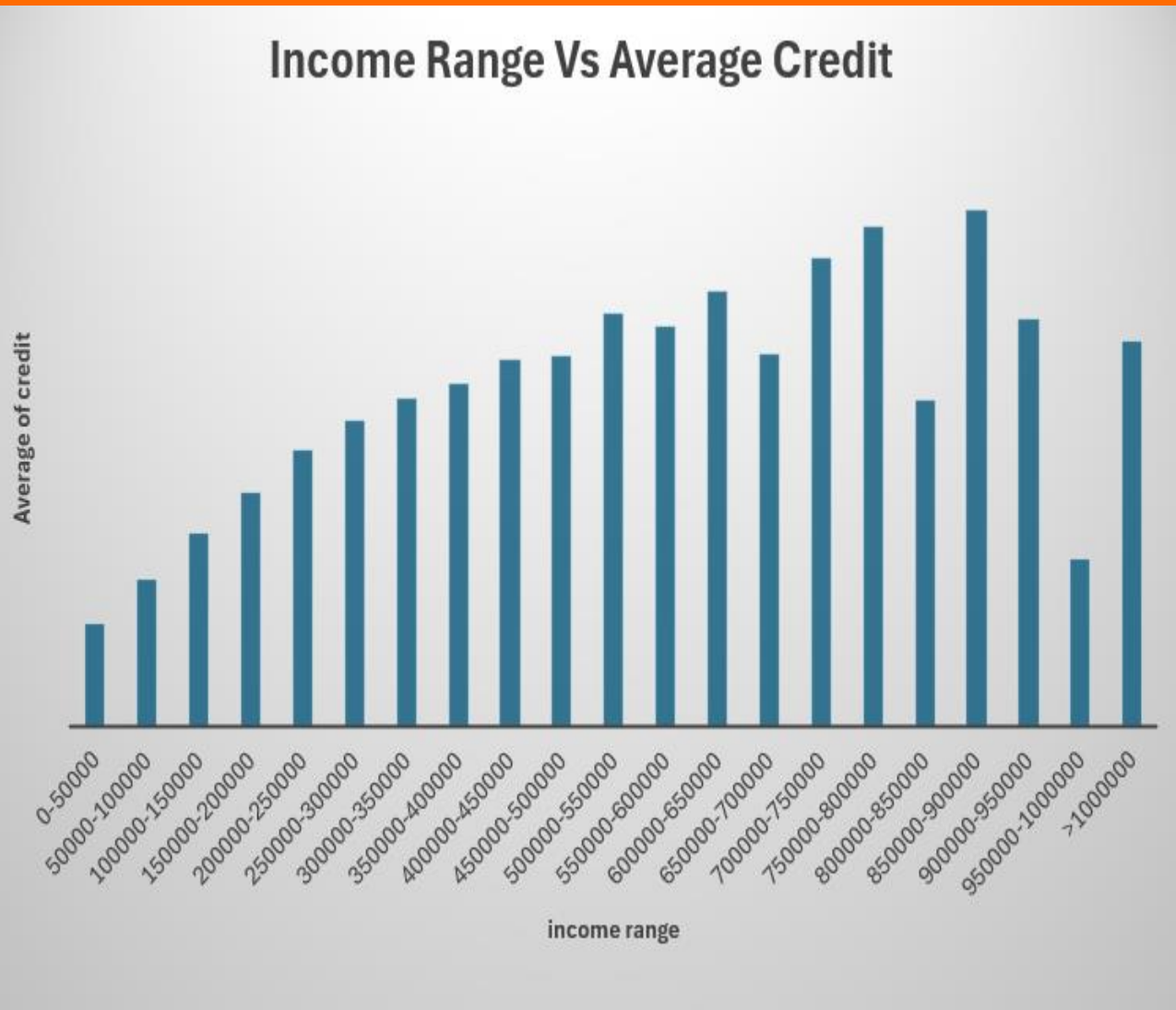
- Borrowers with an “Academic degree” show a 0% default rate.
- “Higher education” and “Incomplete higher” category have relatively low default rates, at category





Bivariate Analysis





E. IDENTIFY TOP CORRELATIONS :

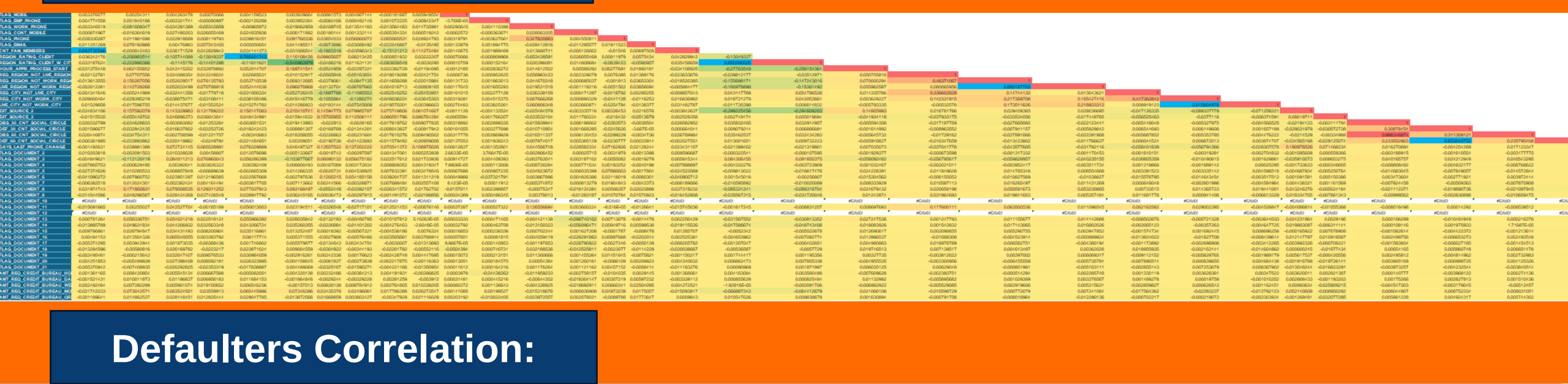
Top Correlation of Defaulters

Rank	Variable 1	Variable 2	Correla
1	AMT_GOODS_PRICE	AMT_CREDIT	0.981928
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.948021
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.8956
4	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891467
5	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.805583
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.773107
7	AMT_ANNUITY	AMT_GOODS_PRICE	0.746422
8	AMT_ANNUITY	AMT_CREDIT	0.745132

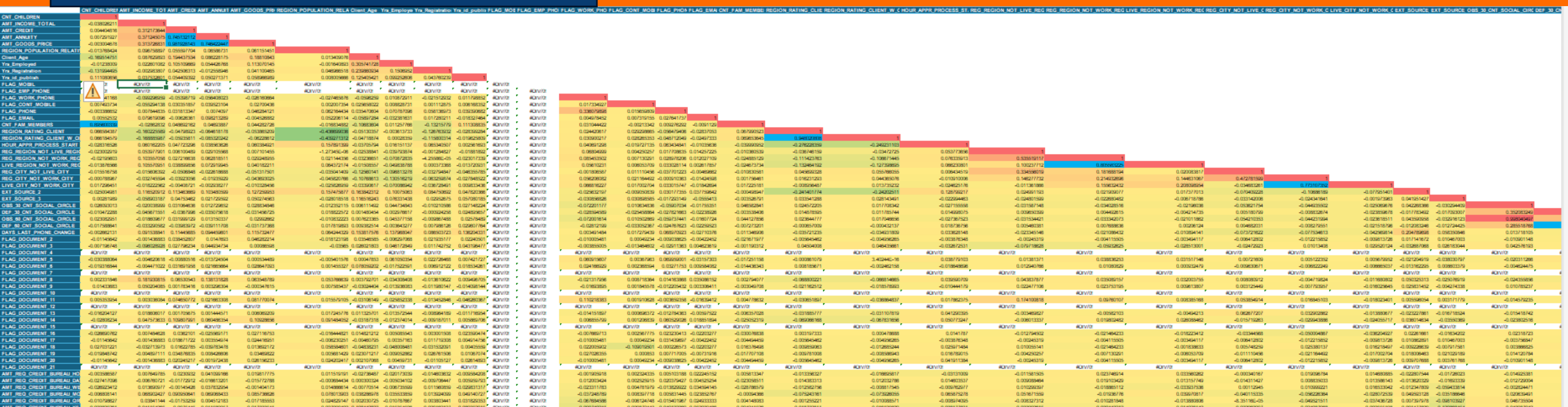
Top correlation of Non-Defaulters

Rank	Variable 1	Variable 2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.98635817
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950286525
3	CNT_CHILDREN	CNT_FAM_MEMBERS	0.893735596
4	REG_REGION_NOT_WORK_REGION -	LIVE_REGION_NOT_WORK_REGION	0.860167703
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.853040752
6	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.815604978
7	REGION_RATING_CLIENT	AMT_GOODS_PRICE	0.765201743
8	AMT_ANNUITY	AMT_GOODS_PRICE	0.765201743
9	AMT_CREDIT	AMT_ANNUITY	0.760827873

Non-Defaulters Correlation:



Defaulters Correlation:



Insights :

- Defaulting decreases with the age and experience, therefore bank give more priority to the older and experienced applicants.
- Educated clients tend to default less frequently.
- Also, clients with more than two children default more frequently.
- Male clients tend to default more than female clients.

Here is the link to Excel Sheet

Result:

In this project I applied the EDA(Exploratory data analysis) Using Excel to analyze patterns in the dataset. Before starting this project, I conducted a brief research on risk analytics in banking and finance services.
Through this project, I learned to identify and handle missing data, detect outliers, data imbalance. The result of this analysis highlighted key factors influencing loan defaults, such as income level, loan amount, and credit history. We discovered patterns indicating higher default rates among certain customer segments, leading to data-driven decision rules for approving or modifying loans.


Thank
You
PROJECT MADE BY
----GITANJALI PEKAMWAR