BANK LOAN CASE STUDY

Problem Statement:

The finance company gives loans to people living in cities and often extends credit to customers with limited or insufficient credit histories. Creditworthy applicants who would repay the loan are denied, resulting in lost revenue. High-risk applicants without repayment capacity are approved, causing financial losses through defaults.

The company wants to understand:

- What kind of customers are more likely to **default** (not pay on time)?
- What kind of customers are **safe** to give loans to?

Description

This case study focuses on identifying patterns that indicate whether a client is likely to face difficulties in repaying loan installments. These insights can help the company make decisions such as denying loans, reducing loan amounts, or charging higher interest rates for risky applicants.

The dataset provides information about clients at the time of their loan application and includes two types of scenarios:

1. **Clients with payment issues** – Those who made late payments exceeding a certain number of days on at least one of the first few installments.
2. **Clients who paid on time** – All other cases where the payment was made within the expected time.

When a client applies for a loan, one of the following four decisions may be taken by the client or the company:

- **Approved**
- **Cancelled**
- **Refused**
- **Unused Offer**

In this study, we'll perform Exploratory Data Analysis (EDA) to explore how customer characteristics and loan features influence the likelihood of default.

Approach:

- Download the datasets: **application_data** and **previous_data.**
- Understand the data.

- **Data Cleaning:**
    - Remove columns with more than 40% null values.
    - Replace missing values in other columns using the **mean** or **median**.
    - Drop columns that are not relevant.
    - Convert values in columns like DAYS_EMPLOYED, DAYS_BIRTH, DAYS_ID_PUBLISH, and DAYS_REGISTRATION from days to years using the formula ROUND(ABS()) in Excel, and store them in new columns.
- Detect outliers in the data.
- Create charts to visualize the data.

Tech stack used:

- MS Excel 2016
- MS WORD

The Datasets provided :

- **Applications_data** : Contains all the information regarding the client, if the client has payment difficulty or not.
- **Previous_data** : Contains the information about clients previous loan. If the loan was approved, cancelled, refused or unused.
- **Columns_description**: It contains information regarding all the columns in the dataset.
- **Important_notes** : It contains information regarding how to approach this case study.

Applications_data:

Total Columns: 132

Columns with null values >40%: 48

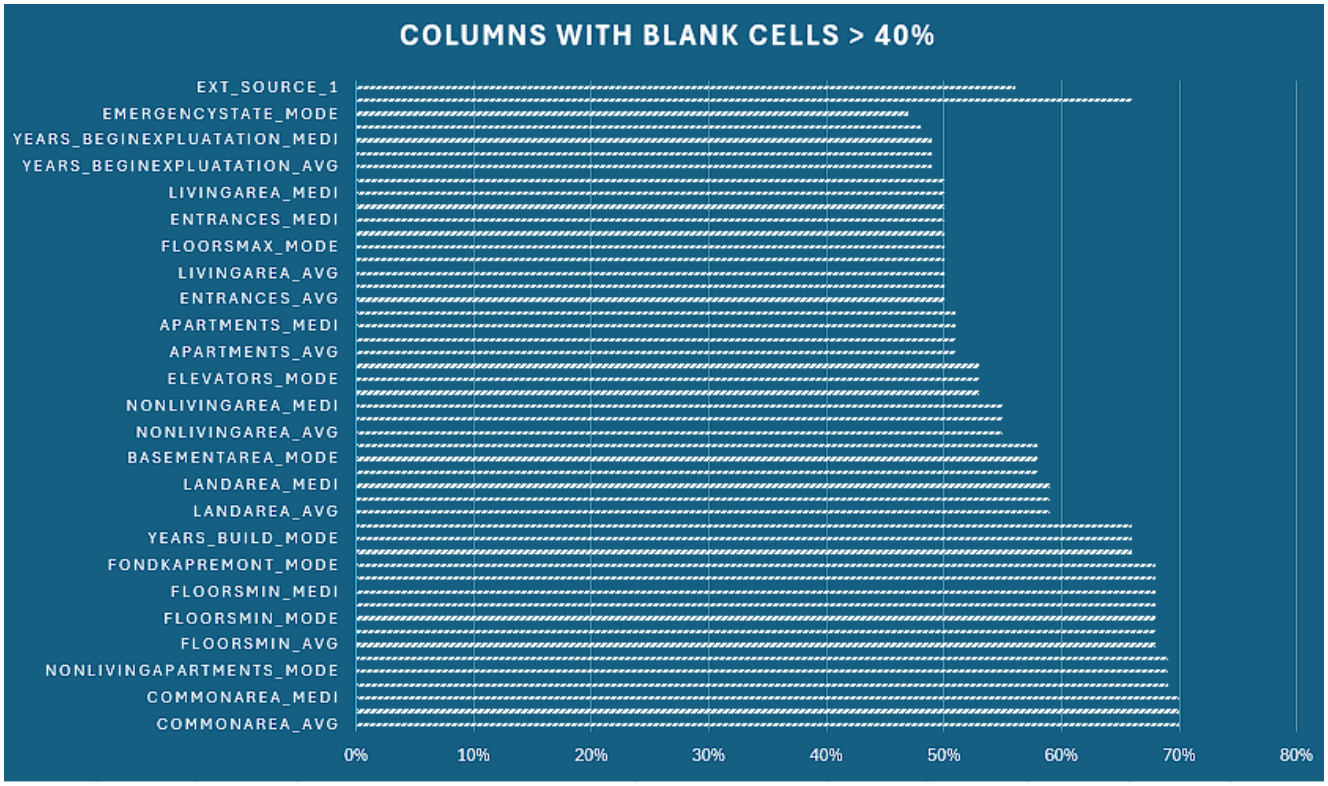
Total Rows: 50000

A. **Identify Missing Data and Deal with it Appropriately:**

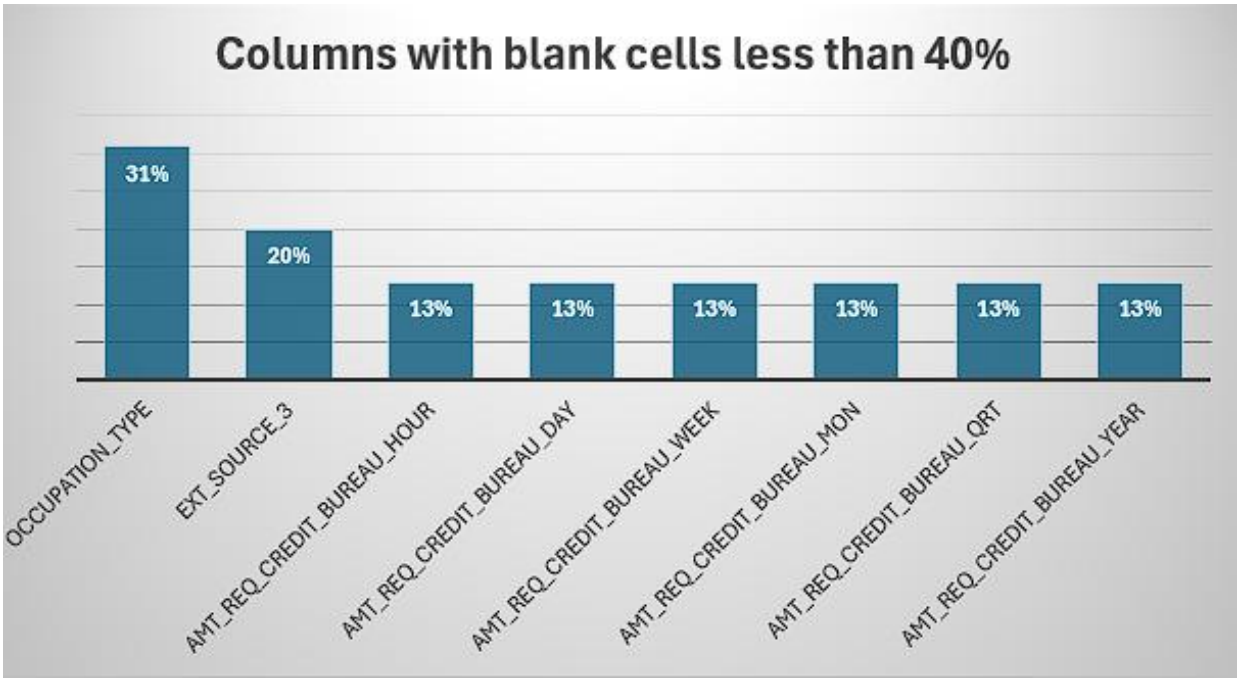
Task: Identify the missing data in the data set and decide on an appropriate method to deal with it using Excel built-in functions and features.

Solution:

- Performed Exploratory Data Analysis **(EDA)** to identify missing values in the dataset.
- Used the =COUNTBLANK () function in Excel to **count the number of blank cells** in each column.
- Calculated the **proportion of blank cells** using the formula:
  = (COUNTBLANK (column)/COUNT (column)) *100



Removed columns that had **more than 40% missing values**, as they carried insufficient data.



For columns with less than 40% missing values, handled them by using:

- Median imputation for numerical columns.
- Mode imputation for categorical columns.

B. **Identify Outliers in the Dataset:**

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

Solution:

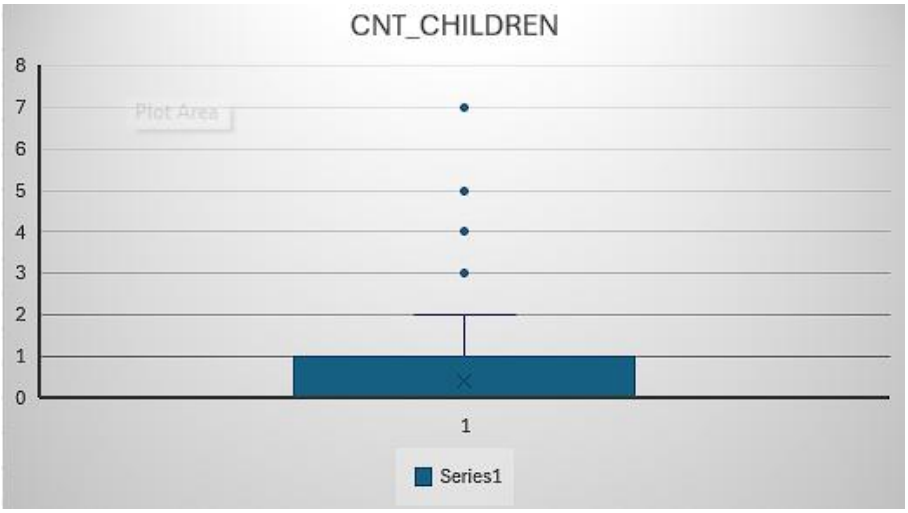
To identify outliers in the dataset, I used the **Interquartile Range (IQR)** method in Excel, focusing on numerical variables such as **CNT_CHILDREN**, **AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE,** and **REGION_POPULATION_RELATIVE.**

- First, Quartile 1 (Q1) and Quartile 3 (Q3) were calculated for each column using Excel's QUARTILE.INC () function.
- The IQR was then calculated as the difference between Q3 and Q1.

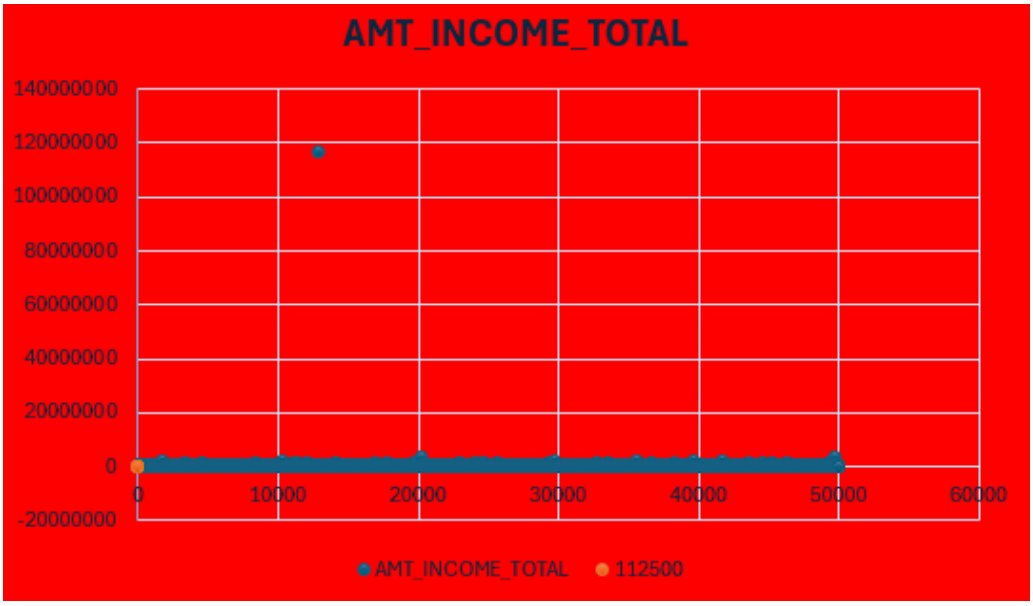
IQR = Q3 – Q1

- Using this, the **upper limit** was calculated as Q3 + 1.5 × IQR, and the **lower limit** was calculated as Q1 – 1.5 × IQR.
- Any value beyond these limits was considered an outlier
- This process helped in identifying and understanding unusually high or low values in the data, which may affect model performance if not treated properly.

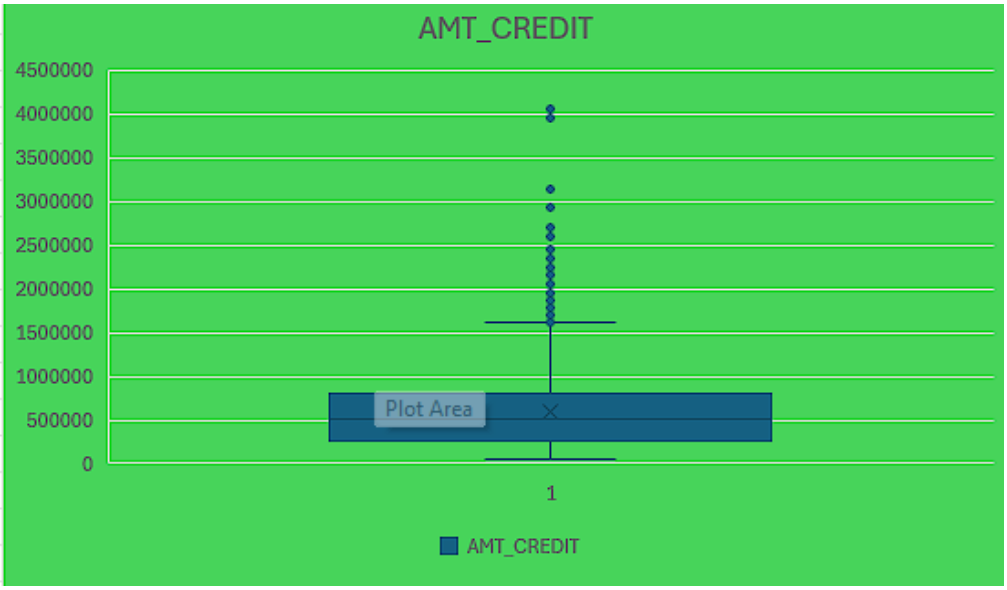
Column	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE
Quartile1	0	112500	270000	16456.5	238500	0.010006
Quartile3	1	202500	808650	34596	679500	0.028663
IQR	1	90000	538650	18139.5	441000	0.018657
Upper Limit	2.5	337500	1616625	61805.25	1341000	0.0566485
Lower Limit	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795



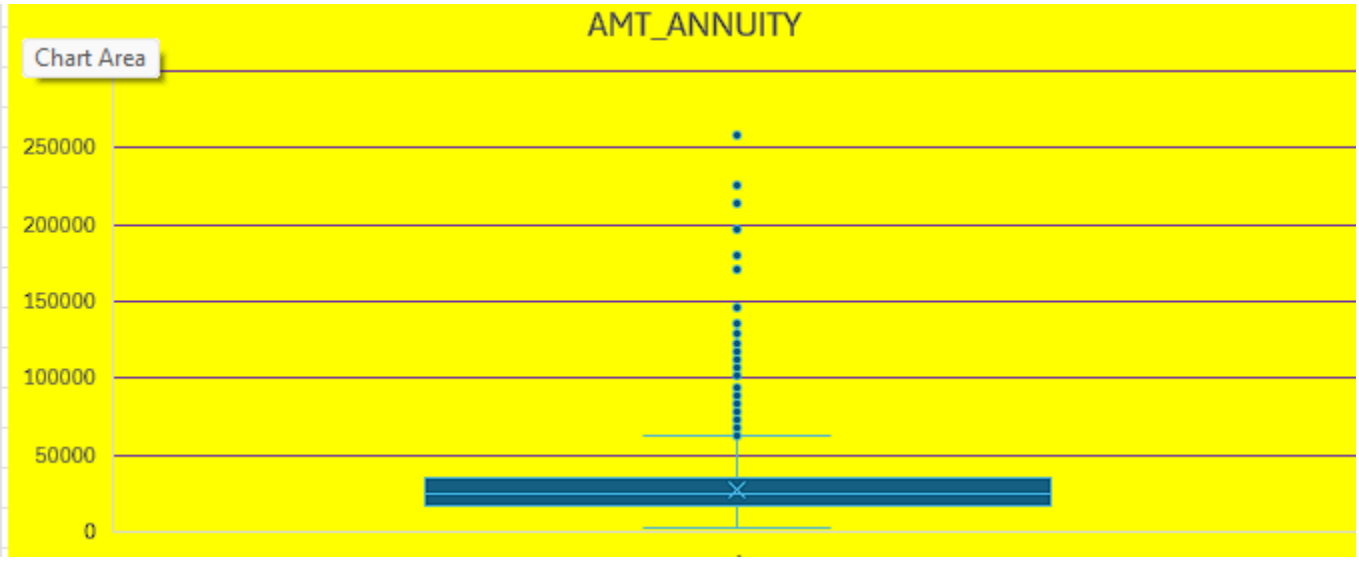
The central box shows the **interquartile range (IQR)**—the middle 50% of the data—ranging from 0 to 2 children. Values like 3, 4, 5, and 7 are detected as outliers.



The scatter plot of AMT_INCOME_TOTAL reveals a few extreme outliers, including an unusually high-income value around **1.2 crore**, which is far beyond the typical income range.

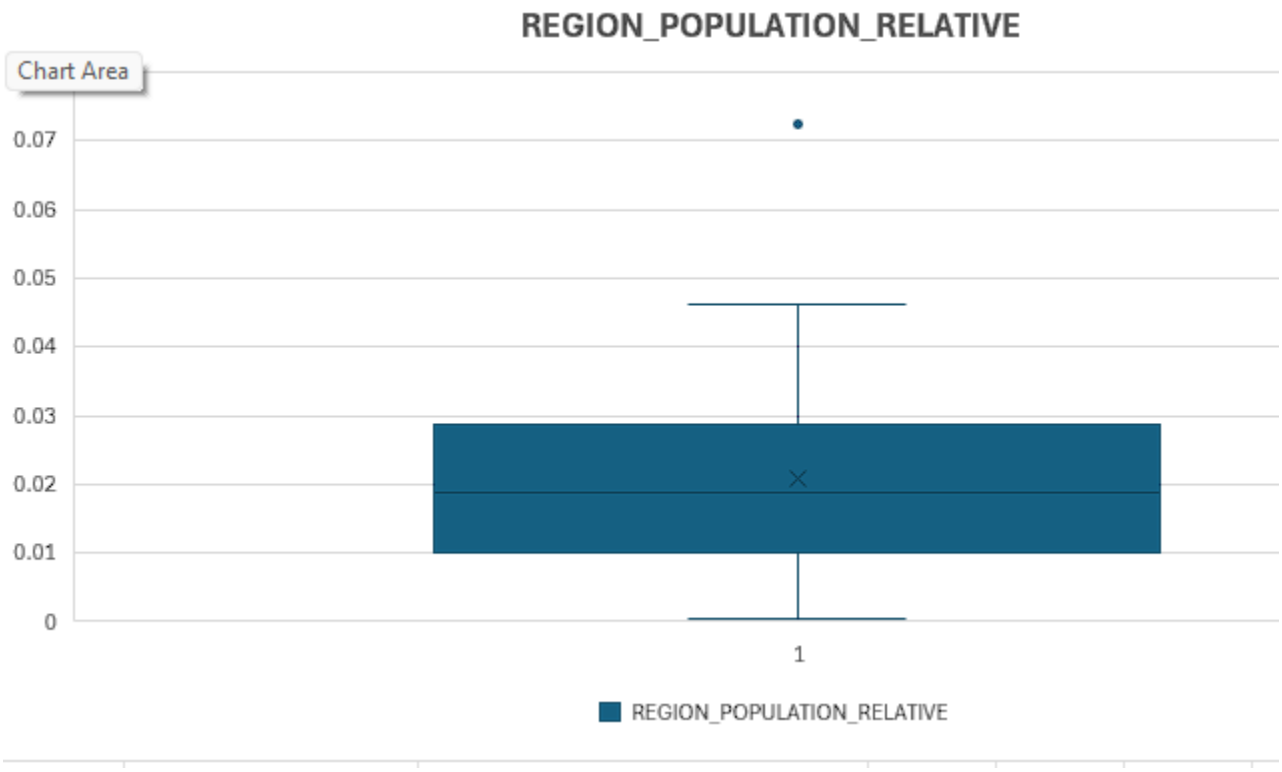


This shows multiple outliers, with credit amounts going beyond **4 million**. These extreme values were identified using the IQR (Interquartile Range) method.



This reveals several outliers above **150,000,** indicating unusually high annuity amounts. Using the IQR method, these outliers were identified and addressed by capping them at the upper whisker limit,

Several outliers above 2 million, indicating expensive goods. Outliers were kept since they reflect actual high purchases.



One outlier above 0.07, possibly a densely populated area.

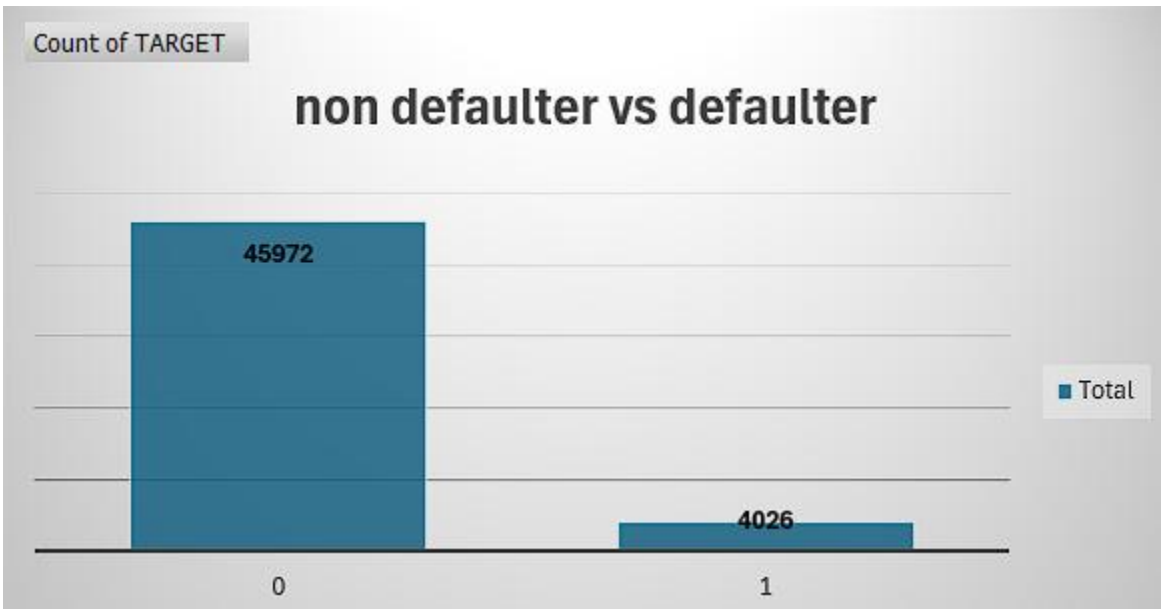
C. **Analyze Data Imbalance**

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

Solution: (used pivot table)

Loan Type	% of clients
Cash loans	91%
Revolving loans	9%

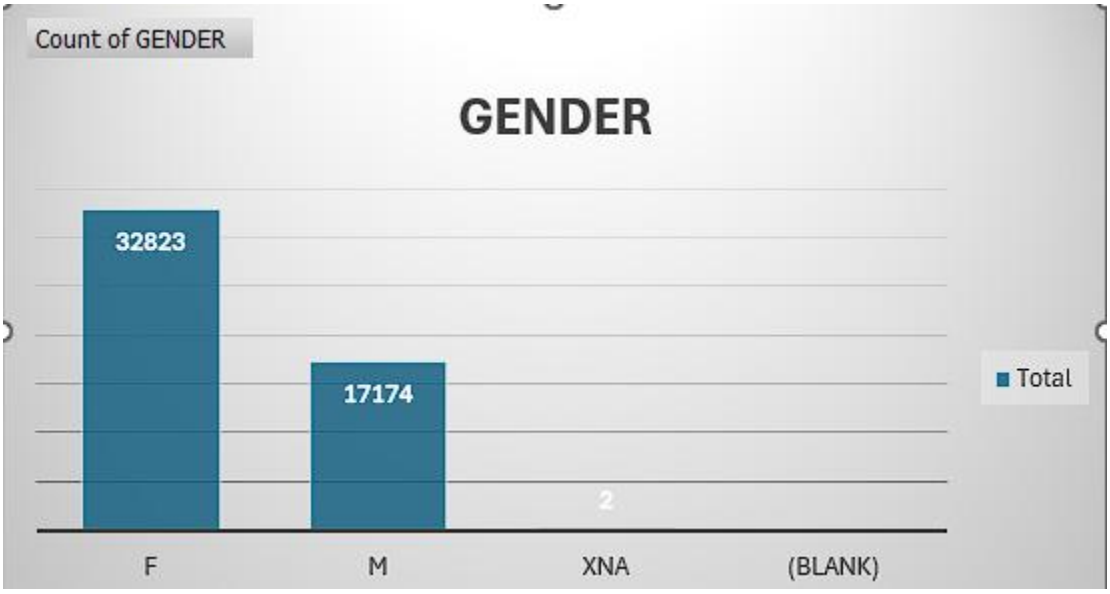
- Cash **loans** dominate the dataset with **45,276 records (91%)**, whereas **Revolving loans** represent only **4,723 records (9%)**.
- The **imbalance ratio** between the two classes is approximately **0.10**, highlighting a significant skew toward cash loans.



Target	% of Target
Defaulters	8%
Non_Defaulters	92%

Non-defaulters (label = 0): 45,972 records (**92%**)

Defaulters (label = 1): 4,026 records (**8%**). This **imbalance ratio of 0.08** indicates that default cases are underrepresented.



Gender	%
F	66%
M	34%
XNA	0%

Female applicants constitute **66% (32,823 records)**, while **Male applicants** make up **34% (17,174 records)**.

D. **Perform Univariate, Segmented Univariate, and Bivariate Analysis**

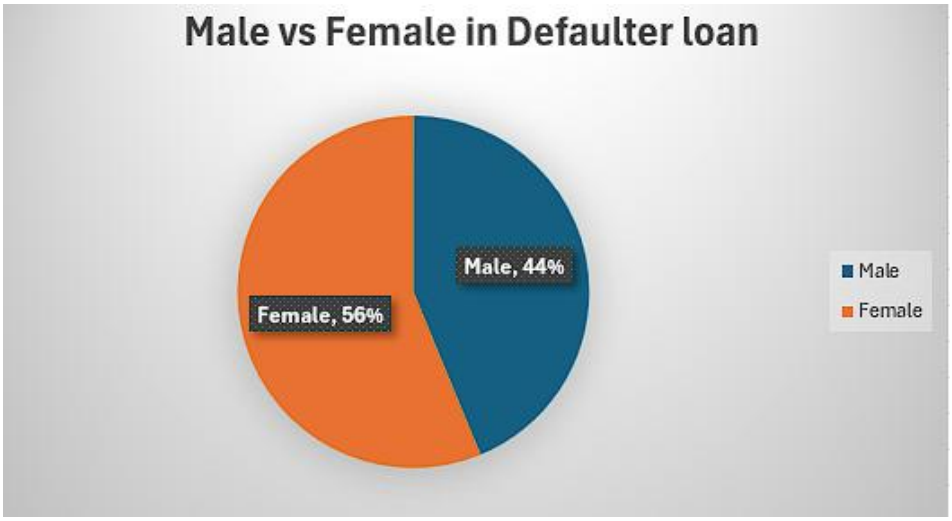
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate Analysis:

1. **Gender**

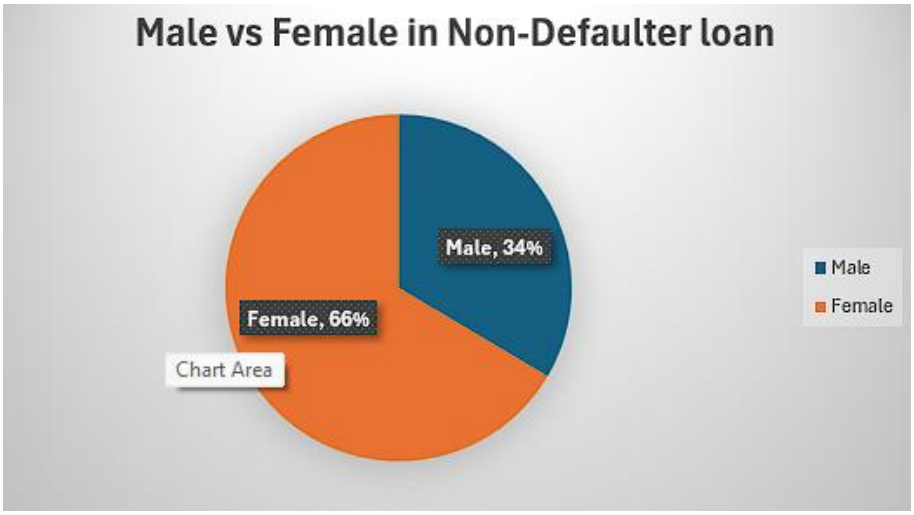
Gender	Defaulters	Non-Defaulters
Male	1762	15412
Female	2264	30559
total	4026	45971

- A univariate analysis of the 'Gender' and 'Target' columns revealed that the dataset is significantly **imbalanced.**
- Females form the majority in both defaulter and non-defaulter groups.
- This could indicate better loan performance among females overall.



56% of defaulters are female, suggesting higher loan participation.

Male defaulters account for 44% despite being fewer in total.

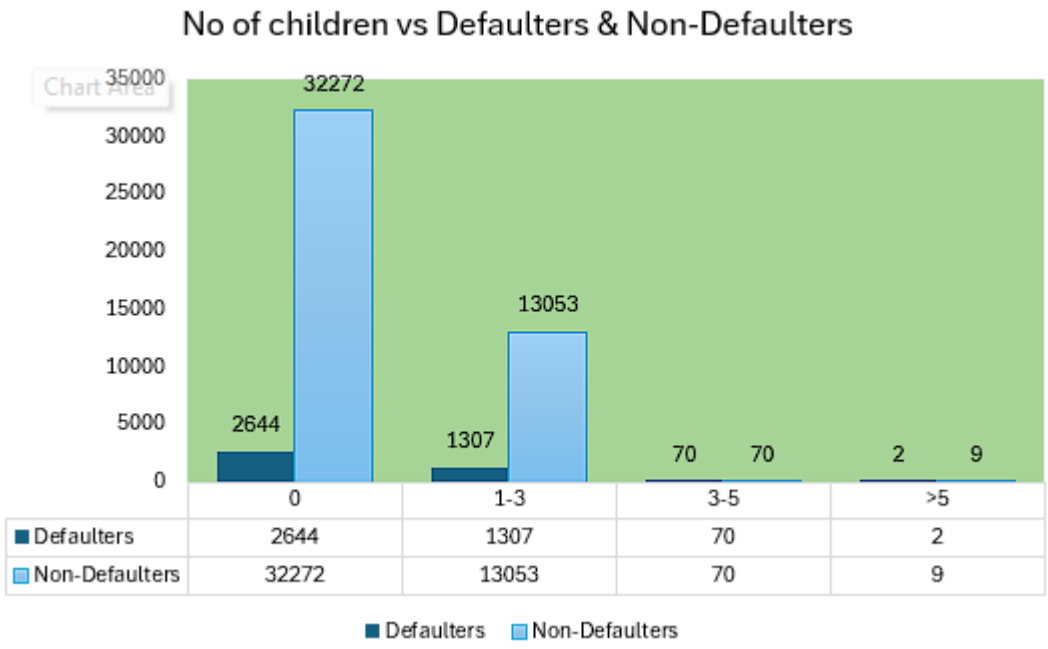


Among non-defaulters, females dominate with **66%** share.

2. **Number of Children:**

No of children	Loan Taken	Defaulters	Non-Defaulters	%Defaulters
0	34916	2644	32272	8%
1-3	14360	1307	13053	9%
3-5	699	70	70	10%
>5	11	2	9	18%

The analysis shows a clear trend between the number of children and the likelihood of loan default.



- **Majority of applicants (34,916)** had **no children**, with a default rate of **8%**.
- Applicants with **1–3 children** totaled **14,360**, showing a slightly higher default rate of **9%**.
- A smaller group with **3–5 children** (699 applicants) had a **10%** default rate, indicating a further increase.
- Although very few applicants had **more than 5 children** (only 11), the default rate here peaked at **18%** more than double that of applicants with no children.

Insight: As the number of children increases, so does the default rate. This may be due to increased financial burden, impacting repayment capacity.

3. **Family Status:**

Family Status	Loan Taken	Defaulters	Non-Defaulter	%Defaulter
Single / not married	7306	729	6577	10%
Married	32094	2395	29699	7%
Civil marriage	4859	482	4377	10%
Widow	2597	148	2449	6%
Separated	3142	272	2870	9%
Unknown	1	0	1	0%

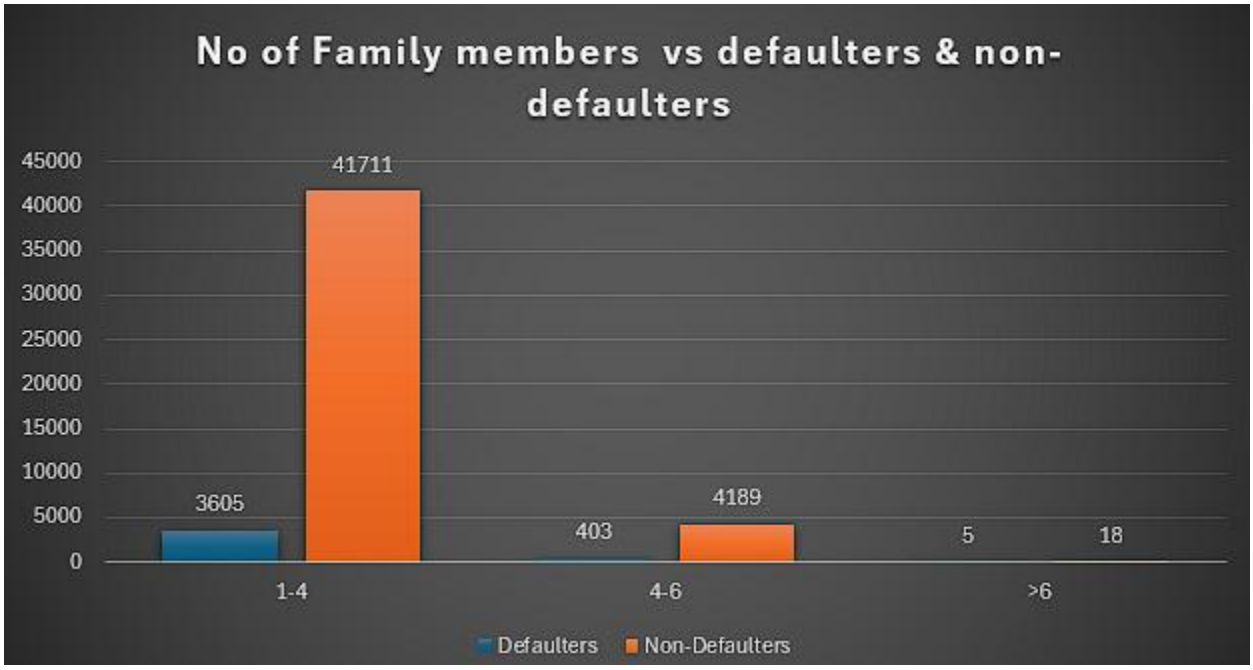
- **Married** individuals form the largest group of borrowers, taking **32,094** loans with a **default rate of 7%**, showing relatively stable repayment behavior.
- **Single / Not Married** and those in **Civil Marriage** both show a **10% default rate**, indicating a slightly higher risk segment.
- **Widows** have the lowest default rate at **6%**, though their loan count is comparatively small (**2,597**).
- **Separated** borrowers show a **9% default rate**, slightly above the overall average.

4. **Number of Family Members:**

No of Family Members	Loan Taken	Defaulters	Non-Defaulters	%Defaulters
1-4	49902	3605	41711	7%
4-6	4592	403	4189	9%
>6	23	5	18	22%

Smaller families appear to manage loans more effectively.

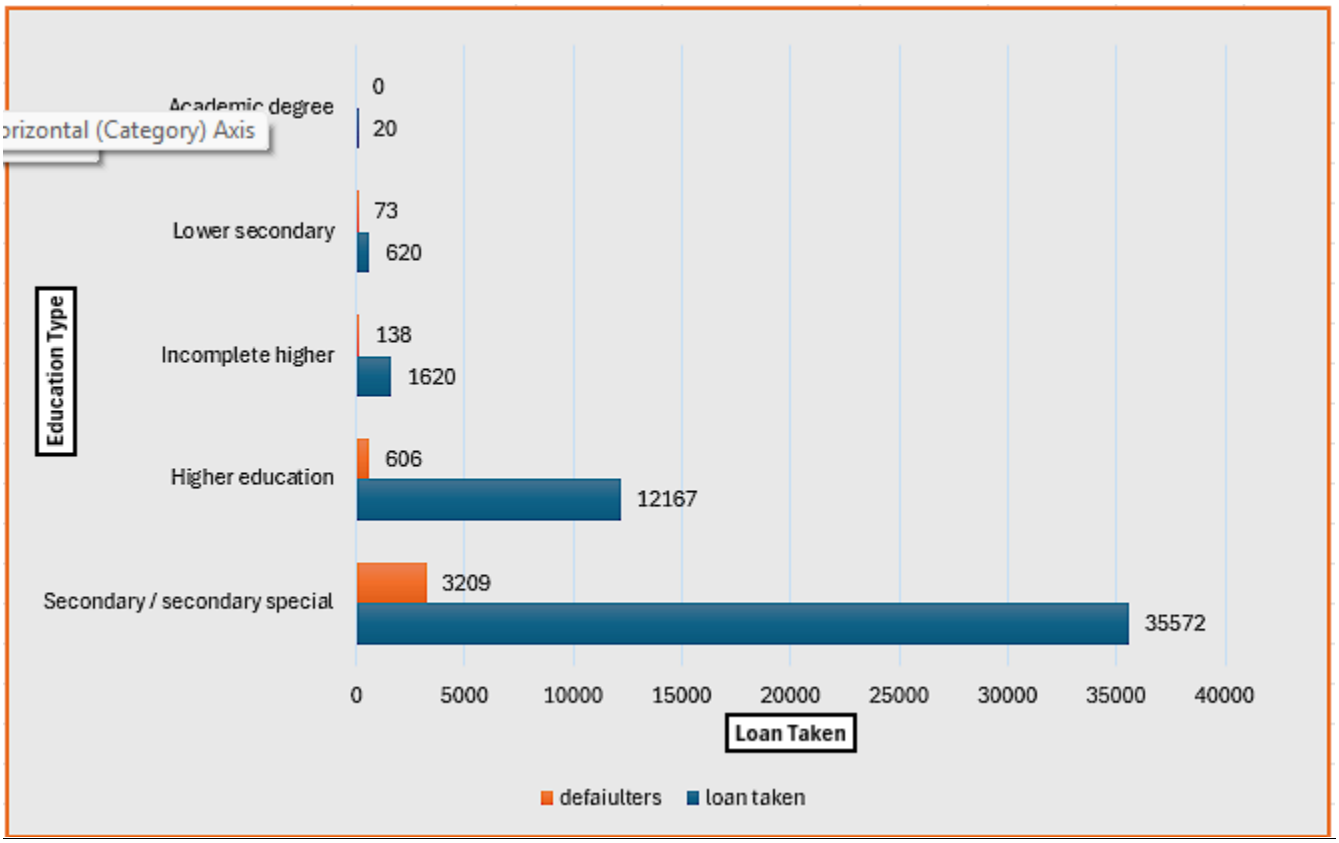
Higher default rate in large families may be due to increased financial pressure, but more data is needed for reliability.



- **Most borrowers (49,902)** belong to families with **1–4 members**. Their default rate is relatively low at **7%**.
- Borrowers with **4–6 members** show a slightly higher default rate of **9%**.
- Borrowers from families with **more than 6 members** have the **highest default rate (22%)**, although the sample size is **very small (23 loans)**.

5. **Education Type:**

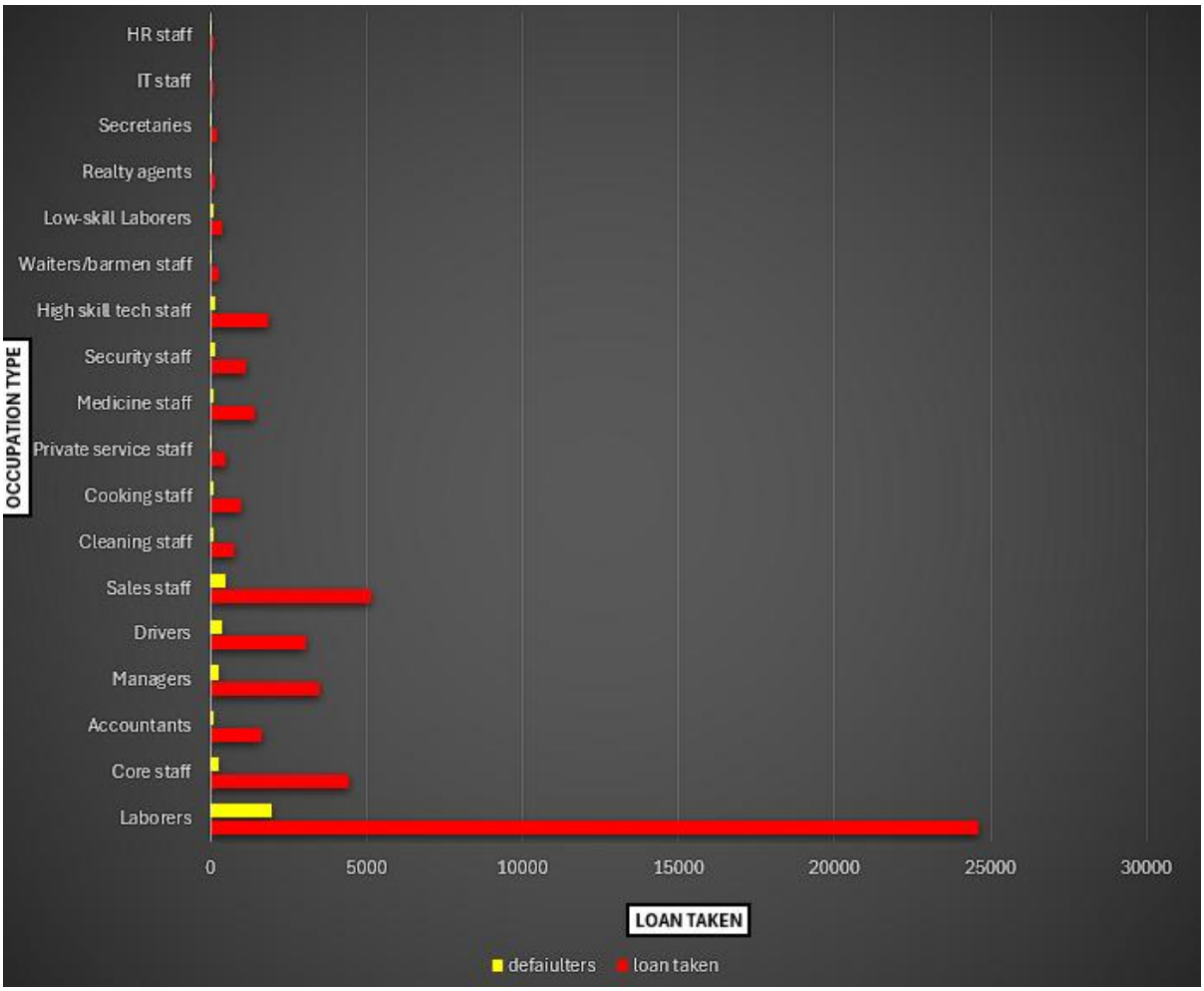
Education_type	loan taken	defaiulter	non-defaulters
Secondary / secondary special	35572	3209	32363
Higher education	12167	606	11561
Incomplete higher	1620	138	1482
Lower secondary	620	73	547
Academic degree	20	0	20



- Higher Education holders have the lowest default rate (~5%), showing better repayment reliability.
- Lower secondary education borrowers show the highest default rate (~12%).
- Most defaulters are from the secondary/secondary special category due to volume.
- Academic degree holders had no defaults, but sample size is too small (only 20 loans).

6. Occupation Type:

OCCUPATION_TYPE	loan taken	defaiulters	non-defaulters
Laborers	24606	1946	22660
Core staff	4434	250	4184
Accountants	1621	81	1540
Managers	3489	243	3246
Drivers	3044	338	2706
Sales staff	5160	492	4668
Cleaning staff	739	68	671
Cooking staff	963	101	862
Private service staff	447	37	410
Medicine staff	1403	106	1297
Security staff	1140	125	1015
High skill tech staff	1852	118	1734
Waiters/barmen staff	228	25	203
Low-skill Laborers	357	61	296
Realty agents	123	13	110
Secretaries	212	9	203
IT staff	80	4	76
HR staff	101	9	92

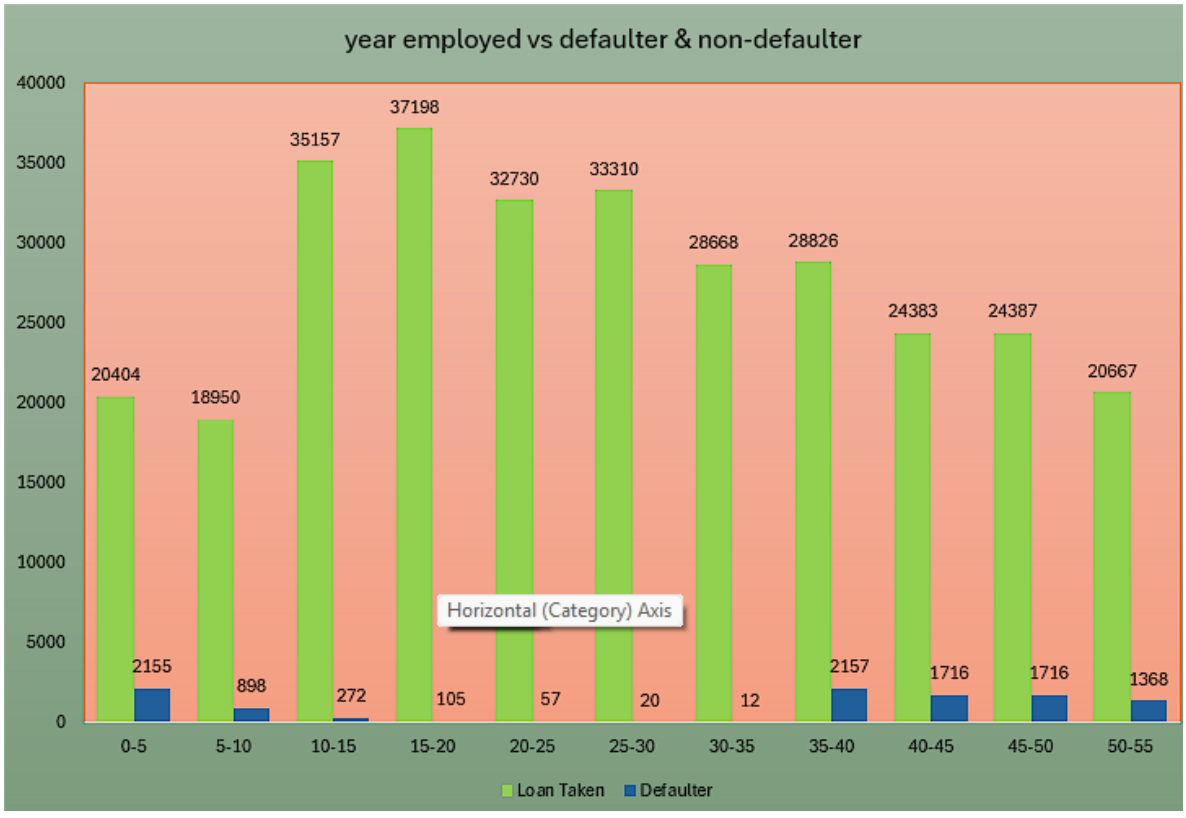


- Laborers have the highest loan count and the most defaults.
- Sales staff and Core staff follow in loan volume with moderate default levels.
- Drivers and Managers show higher default rates relative to loans taken.
- Specialized roles like IT, HR, and Secretaries show low loan activity and minimal risk.

7. Year Employed

Year Employed	Loan Taken	Defaulter	Non-Defaulter
0-5	20404	2155	18249
5-10	18950	898	10649
10-15	35157	272	4556
15-20	37198	105	1936
20-25	32730	57	1084
25-30	33310	20	560
30-35	28668	12	321
35-40	28826	2157	26669
40-45	24383	1716	22667
45-50	24387	1716	22671
50-55	20667	1368	26515

Mid-career individuals tend to repay more reliably than both early-career and late-career applicants.



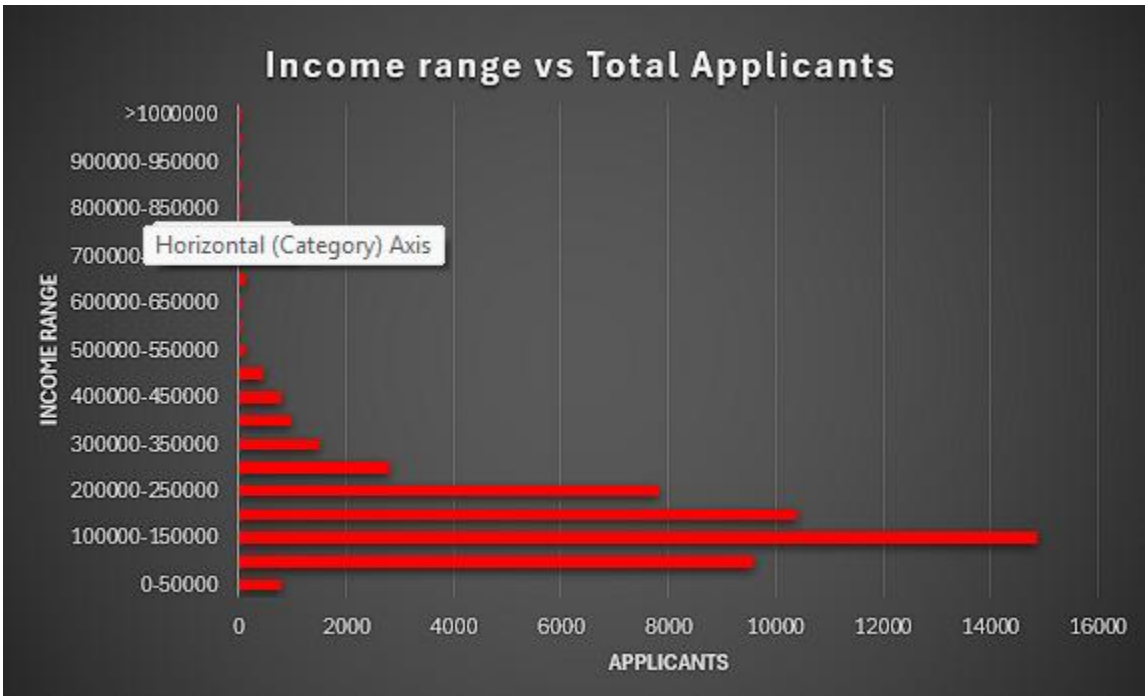
- **Highest Risk**: New employees (0–5 years) have the most defaulters — over *2,100 cases* from *20K+ loans*.
- **Low Risk Zone**: Employees with **10–35 years** of experience show **remarkably low** default counts despite high loan volumes.
- **Spike Again**: Default cases increase for those with **35–55 years**, notably at **35–40 years** with **2,157 defaulters**.

Segmented Analysis:

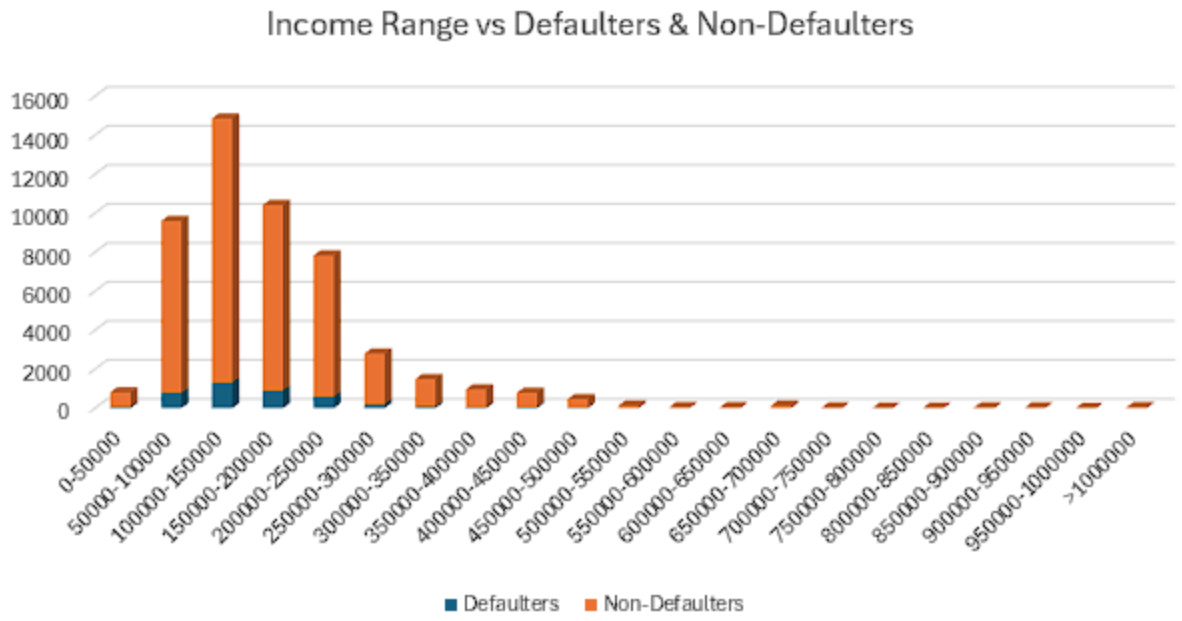
1. **Income Range:**

Concatenate function used to create Income Range.

Countifs() used to count the values in a range



- 100000–150000 income range has the highest number of applicants, approaching 15,000.
- There's a sharp decline in applicant numbers beyond the 300000–350000 range.
- Most applicants fall in the low to middle-income brackets (100000–300000).



150,000–200,000 income group has the **most applicants (~15,000)**.

- The majority are **non-defaulters**.
- Around **2,000 are defaulters**.

100,000–300,000 range:

- Contains the **bulk of applicants and defaulters**.
- Default rate is noticeable but **non-defaulters dominate**.

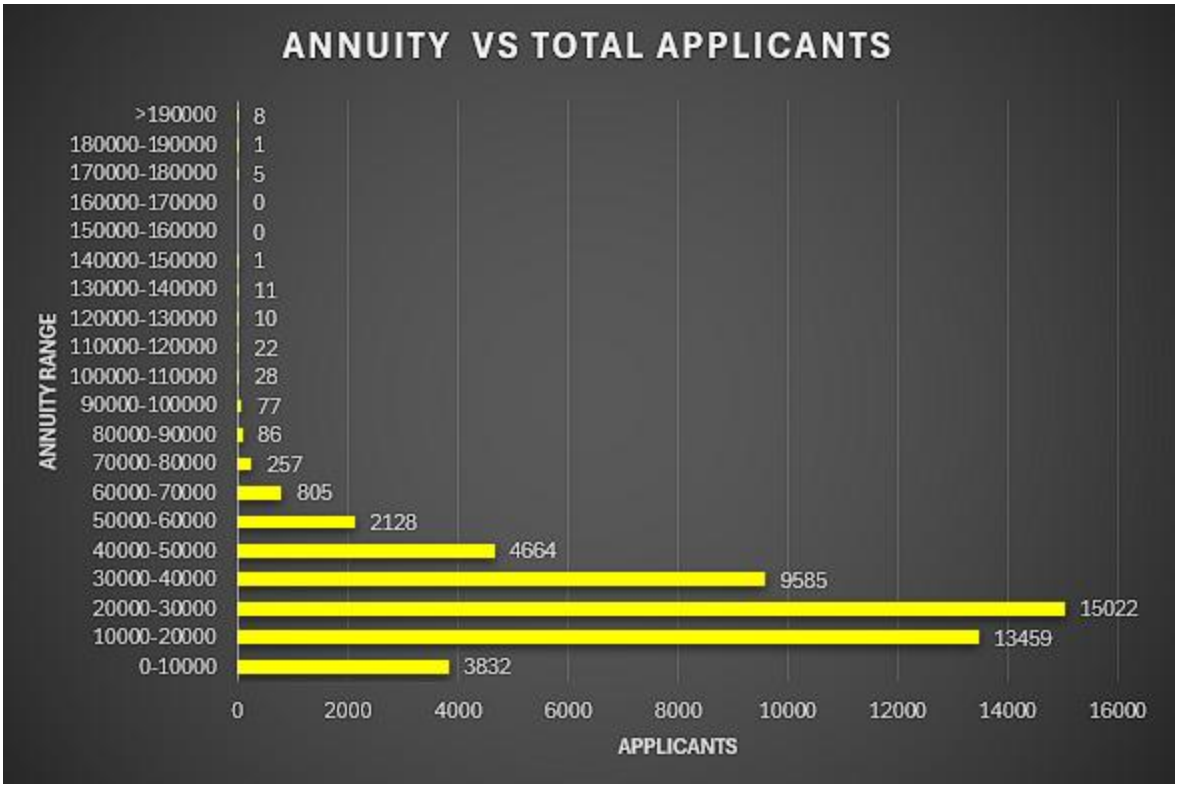
>300,000 income:

- **Very few applicants**, almost **no defaulters**.

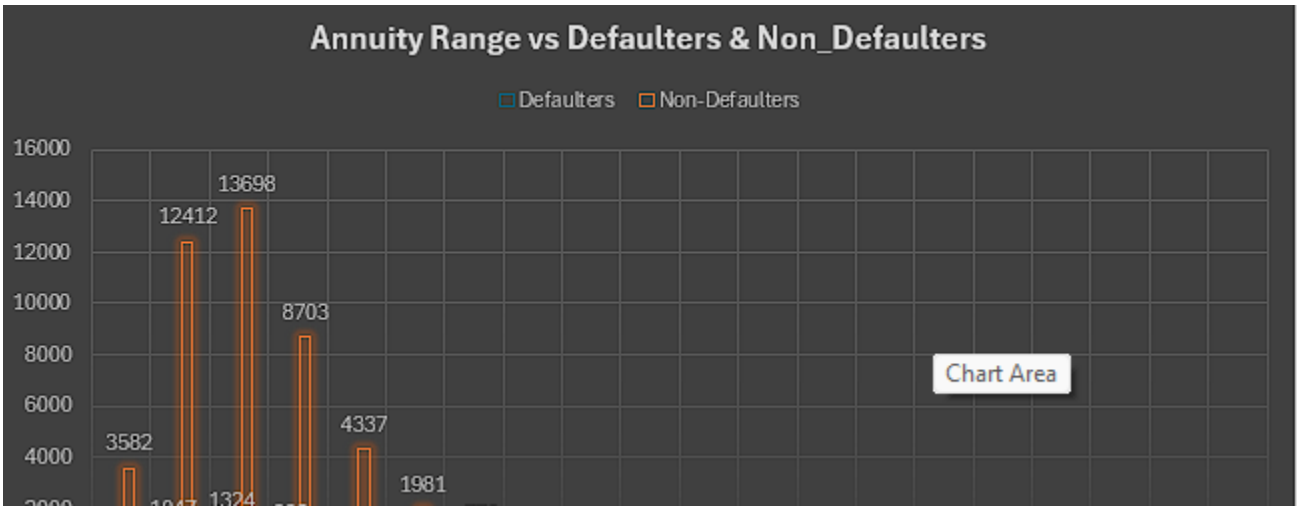
2. ANNUITY

Used the CONCATENATE () function in Excel to merge relevant fields and derive the annuity range for applicants.

Applied IF(ISNUMBER(FIND ())) to identify and count the total number of applicants matching specific criteria.

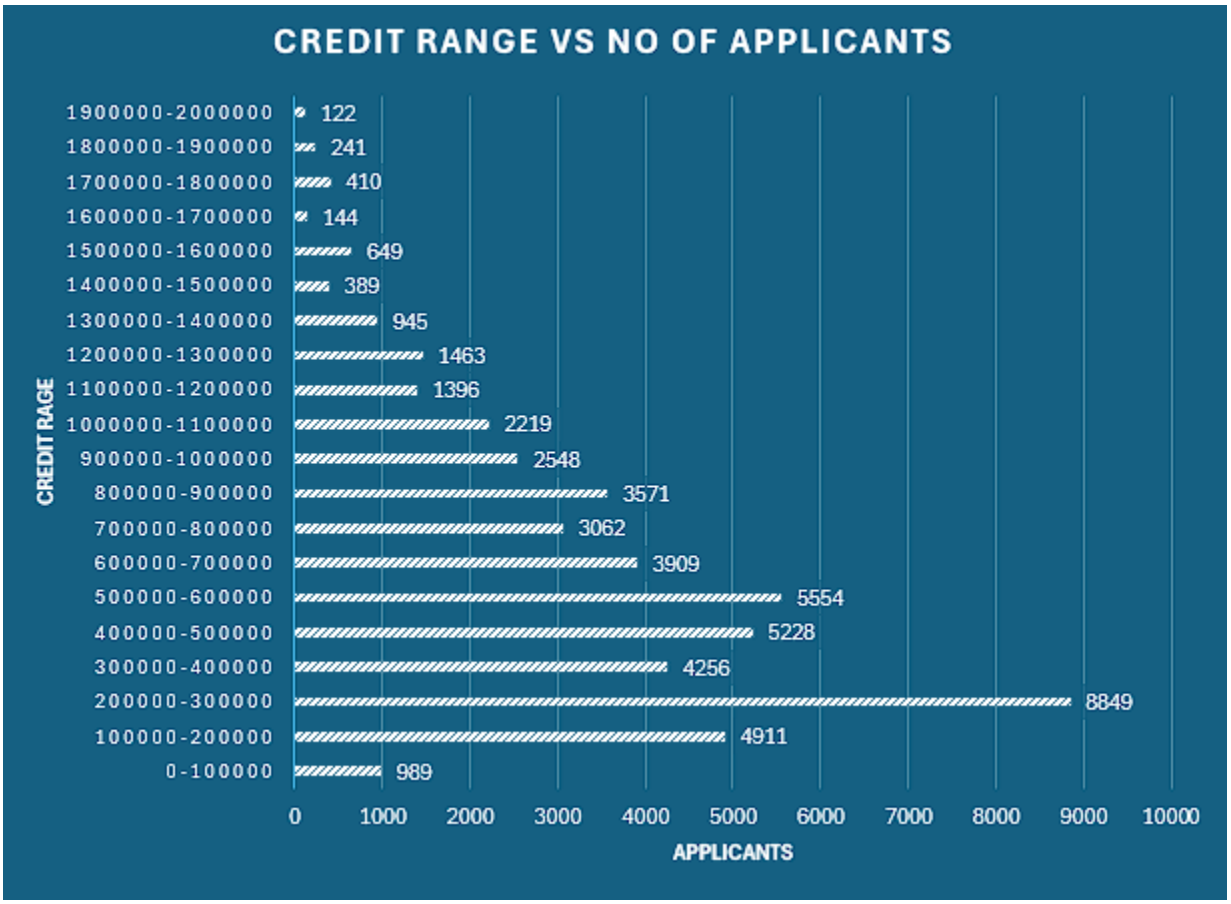


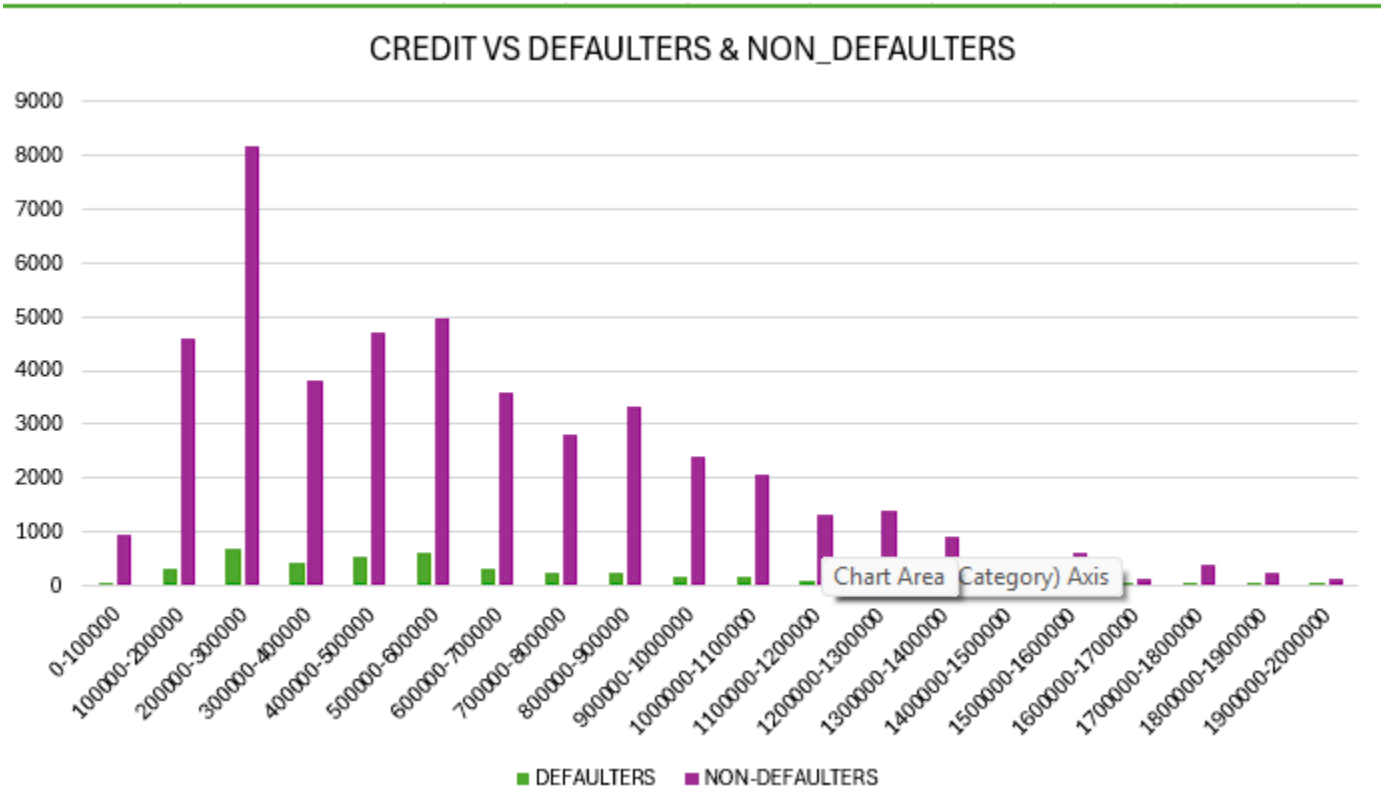
- Over 50% of applicants have annuities between **10,000–30,000**, making this the dominant segment.
- Applicant numbers fall sharply after **30,000.**
- Very few applicants have annuities above **90,000.**



Most defaulters and non-defaulters fall in the **20,000–30,000 annuity range** (≈13,698 non-defaulters, 12,412 defaulters).

3. CREDIT

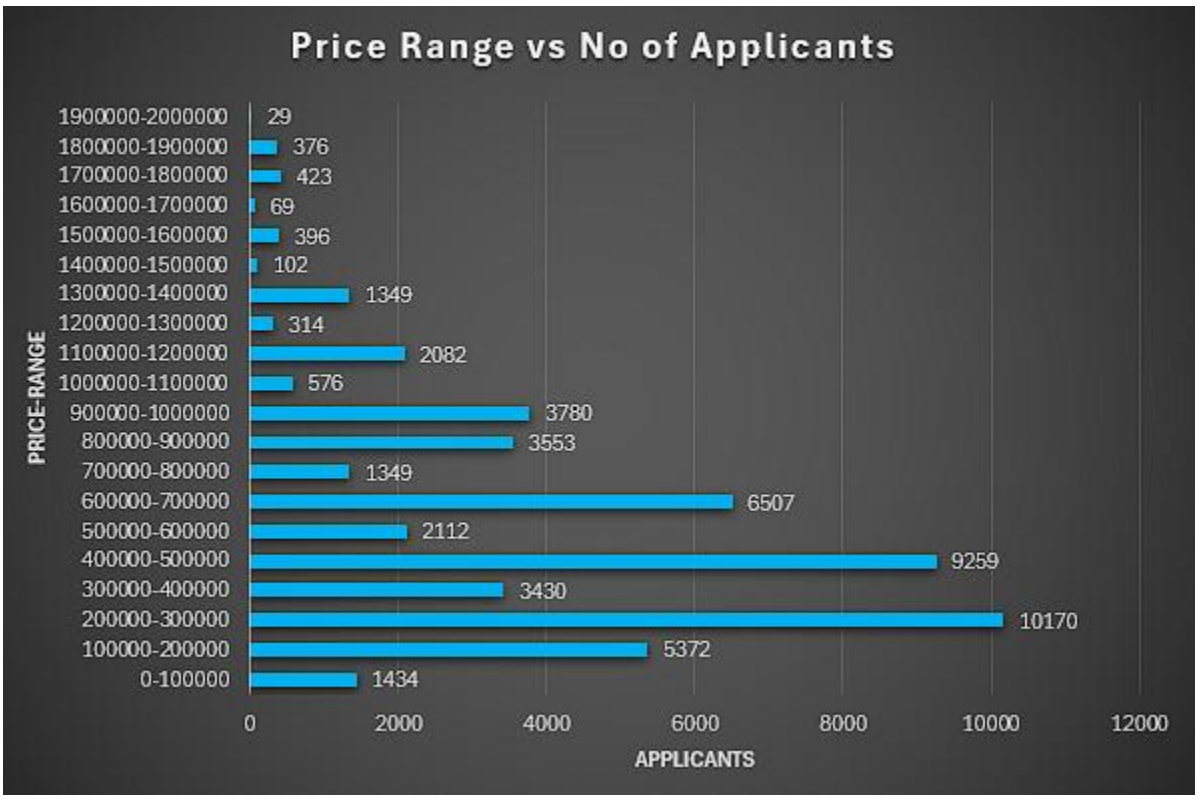




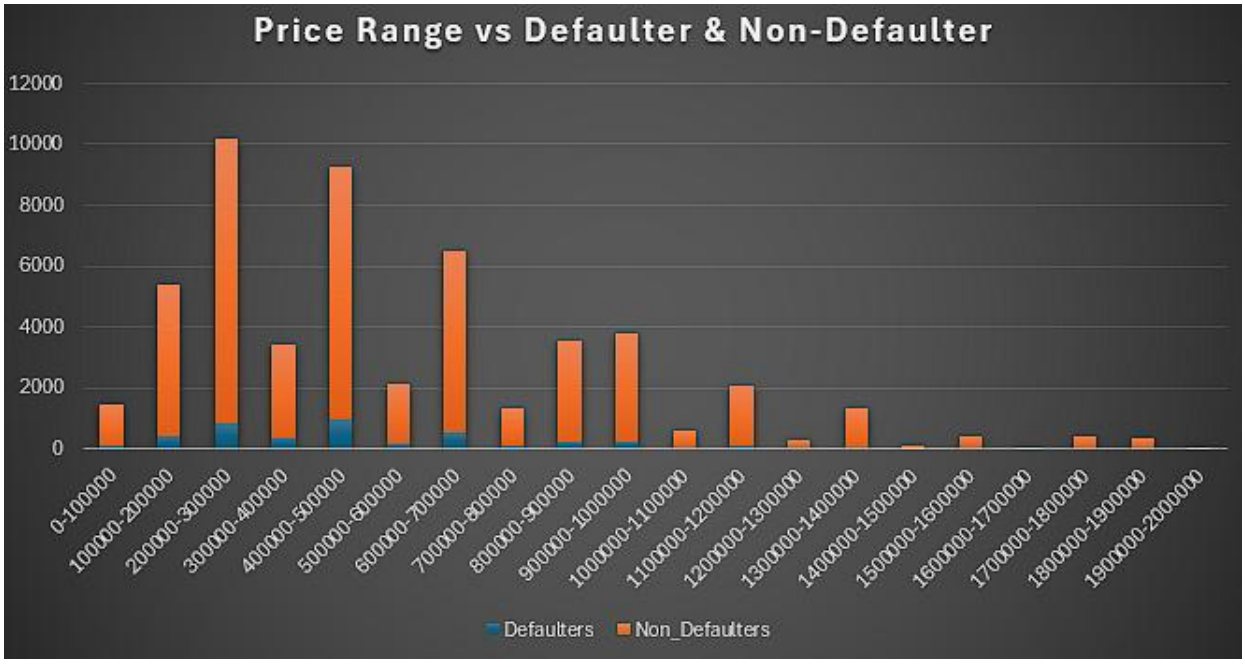
- The **highest concentration** of applicants falls in the **300,000–400,000 credit range** (8,849 applicants
- Very high credit ranges (≥1,500,000) have **minimal applicants** (<300 each).

- **Defaulters are more prevalent in mid-credit ranges** (200,000–600,000), which also hold the largest applicant volumes higher absolute default counts
- In lower ranges (0–100,000), the default rate is comparatively smaller in absolute terms, likely due to smaller loan sizes.

4. GOODS PRICE



Most applicants fall within the **200,000–400,000** and **500,000–700,000** price ranges, with smaller concentrations in very high or very low ranges.



The highest number of applicants falls in the **200,000–300,000** and **500,000–600,000** price ranges, with most being **non-defaulters**.

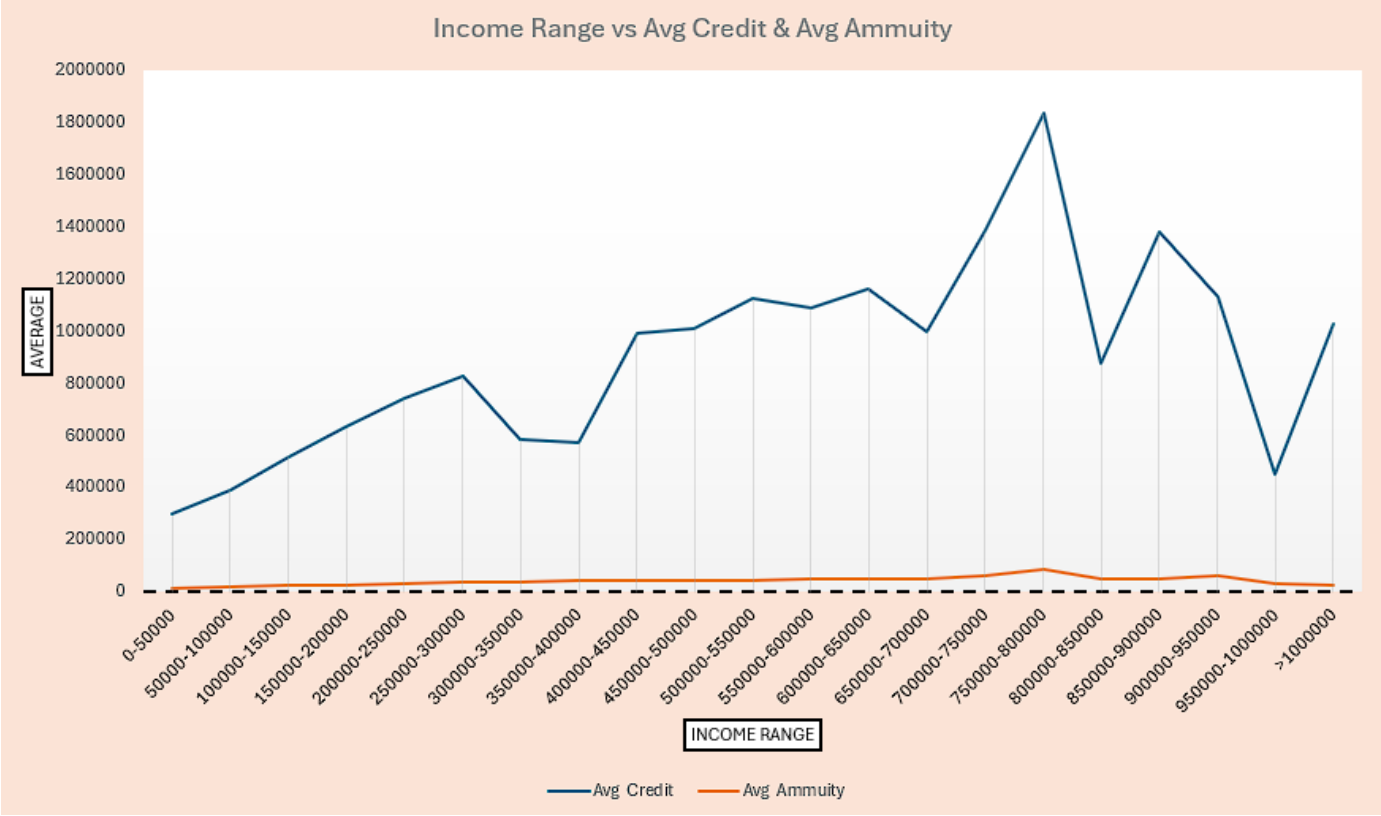
Default rates are relatively low across all price ranges, but slightly higher in the **200,000–400,000** segment.

Bivariate Analysis:

The bivariate analysis highlights a clear positive relationship between income and credit, with a moderate relationship between income and annuity.

Using COUNTIFS () to segment Total Income into predefined ranges

AVERAGEIFS () to calculate the average Credit and average Annuity for each income range.



- positive relationship between **Income Range** and **Average Credit** — higher income brackets tend to have significantly larger credit amounts, with a notable peak in the 800,000–850,000 range.
- **Average Annuity** values remain relatively low and stable across income groups, showing only minor increases in higher ranges.

Income Rang	Avg Credit	Avg Ammuity
0-50000	297752.0765	14086.09701
50000-100000	393033.3365	18478.04427
100000-150000	520073.6603	23850.11907
150000-200000	632290.9047	28489.22761
200000-250000	741970.0033	32838.22276
250000-300000	826106.6508	36130.31438
300000-350000	587857.629	39291.98244
350000-400000	574707.4484	41170.55172
400000-450000	993467.8397	44015.24427
450000-500000	1011895.836	45523.59868
500000-550000	1124198.819	46640.68548
550000-600000	1091708.163	47158.95349
600000-650000	1165433.738	50893.2
650000-700000	1001836.269	48305
700000-750000	1386633.682	60653.45455
750000-800000	1836769.091	84841.36364
800000-850000	876760.0714	47799.85714
850000-900000	1380720.6	51605.1
900000-950000	1132219.8	60760.35
950000-1000000	450000	30073.5
>1000000	1030422.713	26194.5

E. **Correlations Analysis:**

> **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.

Segmented the dataset into two groups — clients with payment difficulties (defaulters) and clients who pay on time (non-defaulters) — using the FILTER () function.

Subsequently, calculated the correlation between relevant variables within each group using the CORREL () function to identify relationships and trends specific to each client segment.

Non-Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CRE	AMT_ANN	REGION_F	DAYS_EMP	DAYS_REG	DAYS_ID_F	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.036319722	0.005705	0.026384	-0.02491	-0.01703	-0.18275	0.032535	0.021288992
AMT_INCOME_TOTAL	0.036319722	1	0.377966	0.451135	0.181941	-0.00946	-0.06896	-0.03207	-0.205031899
AMT_CREDIT	0.005705458	0.377965752	1	0.770773	0.095539	-0.00281	-0.00785	0.007965	-0.102556478
AMT_ANNUITY	0.02638396	0.451135167	0.770773	1	0.117279	0.00886	-0.03443	-0.00965	-0.129920896
REGION_POPULATION_RELATIVE	-0.024912809	0.181941261	0.095539	0.117279	1	0.001295	0.058355	0.002304	-0.539333113
DAYS_EMPLOYED(YRS)	-0.017027849	-0.009459684	-0.00281	-0.00886	0.001295	1	0.00679	0.01294	0.003640381
DAYS_REGISTRATION(YRS)	-0.182749601	-0.068956548	-0.00785	-0.03443	0.058355	0.00679	1	0.103719	-0.08248045
DAYS_ID_PUBLISH(Yrs)	0.032534853	-0.032065618	0.007965	-0.00965	0.002304	0.01294	0.103719	1	0.007512774
REGION_RATING_CLIENT	0.021288992	-0.205031899	-0.10256	-0.12992	-0.53933	0.00364	-0.08248	0.007513	1

Chart Area (Category) Axis

correlation for applicants with payment mode on time

- AMT_CREDIT and AMT_ANNUITY show a very high positive correlation (**0.77**)
- AMT_INCOME_TOTAL is moderately correlated with both AMT_CREDIT (**0.38**) and AMT_ANNUITY (**0.45**), suggesting that higher income applicants tend to have higher credit and annuity amounts.
- REGION_RATING_CLIENT is negatively correlated with REGION_POPULATION_RELATIVE (**-0.54**), meaning applicants from densely populated regions tend to have lower client ratings.
- DAYS_REGISTRATION(YRS) shows a mild negative correlation with CNT_CHILDREN (**-0.18**), implying applicants with more children tend to have shorter registration history.

Defaulters:

	CNT_CHILDREN	AMT_INC	AMT_CRE	AMT_ANN	REGION_F	DAYS_EM	DAYS_RE	DAYS_ID_PUBLISH(Yrs)	REGION_RATING_CLIENT
CNT_CHILDREN	1								
AMT_INCOME_TOTAL	0.010110177	1							
AMT_CREDIT	0.007601905	0.015271	1						
AMT_ANNUITY	0.029172977	0.018005	0.749665	1					
REGION_POPULATION_RELATIVE	-0.020359154	-0.00618	0.067776	0.073124	1				
DAYS_EMPLOYED(YRS)	-0.019027881	-0.00683	0.011312	0.009295	-0.00245	1			
DAYS_REGISTRATION(YRS)	-0.151818175	0.010368	0.042527	-0.0217	0.04647	0.016482	1		
DAYS_ID_PUBLISH(Yrs)	0.043481876	0.009176	0.044479	0.021496	0.005551	-0.02206	0.091495	1	
REGION_RATING_CLIENT	0.055515557	-0.01285	-0.04502	-0.06158	-0.43003	-0.00731	-0.1164	-0.028208994	1

correlation for applicants with payment difficulty

- AMT_CREDIT and AMT_ANNUITY again show a high positive correlation (**0.75**), indicating that larger credit amounts are tied to proportionally higher annuity amounts
- REGION_POPULATION_RELATIVE has a small but positive correlation (**0.073**) with AMT_ANNUITY

- REGION_RATING_CLIENT and REGION_POPULATION_RELATIVE are moderately negatively correlated (**-0.43**), meaning applicants from more populated regions tend to have lower client ratings

Top Correlations:

Top Correlations(non-defaulters)		
variable 1	variable 2	correlation
AMT_INCOME_TOTAL	CNT_CHILDREN	0.03632
AMT_ANNUITY	AMT_INCOME_TOTAL	0.451135
AMT_ANNUITY	AMT_CREDIT	0.770773
REGION_POPULATION_RELATIVE	AMT_ANNUITY	0.117279
DAYS_REGISTRATION(YRS)	REGION_POPULATION	0.058355
DAYS_ID_PUBLISH(Yrs)	DAYS_EMPLOYED(YRS	0.01294

Top Correlations(defaulters)		
variable1	variable2	correlation
REGION_RATING_CLIENT	CNT_CHILDREN	0.055516
AMT_ANNUITY	AMT_INCOME_TOTAL	0.018005
AMT_ANNUITY	AMT_CREDIT	0.749665
REGION_POPULATION_RELATIVE	AMT_ANNUITY	0.073124
DAYS_REGISTRATION(YRS)	REGION_POPULATION_RELATIVE	0.04647