Bank Loan Case Study

Project Description

This project aims to analyze a bank loan application dataset to uncover patterns and variables influencing loan default behaviour. The objective is to clean the data, handle missing values and outliers, evaluate data imbalance, and conduct univariate, bivariate, and correlation analysis to identify key risk factors. Insights derived will support data-driven decision-making for credit risk management.

Approach

The project was executed using a structured data analysis pipeline:

- **Data Cleaning & Preprocessing**
  - Identified missing values using Excel functions COUNTBLANK.
  - Calculated missing percentages for each column.
  - Applied imputation strategies for missing numerical data using AVERAGE functions.
  - Categorical missing values were handled using Mode imputation.
  - Delete unwanted columns.
  - Cleaned Data with outliers saved in a new sheet.
- **Outlier Detection**
  - Used the IQR method via Excel QUARITLE functions and calculated:

    Lower bound = Q1 − 1.5 × IQR

    Upper bound = Q3 + 1.5 × IQR
  - Outliers were detected and highlighted with conditional formatting for visual validation.
- **Data Imbalance Check**
  - Required variables was analyzed using PIVOT TABLE to evaluate distribution.
  - A bar chart and pie chart were created to visualize imbalance.
- **Univariate, Segmented, and Bivariate Analysis**
  - Univariate analysis explored variable distributions (e.g., income, credit amount).
  - Segmented univariate analysis compared distributions across categories (e.g., gender, education level).
  - Bivariate analysis correlated independent variables with the target using PivotTables.

➢ **Correlation Analysis**

- Calculated Pearson correlation coefficients using CORREL function in Excel.
- Segmented analysis done on defaulting and non-defaulting applicants.

Tech-Stack Used

In this project, we have used-

- Microsoft Excel 2024 – Used for initial data cleaning, data exploration, applying formulas (COUNTIF, AVERAGE, MODE, COUNTBLANK, CORREL) for statistical calculations, using graphs for visualization and summarizing data.
- Google Drive – Hosting and sharing reports.

For application_data.xlsx Dataset:

Data Analytics Tasks

A. Identify Missing Data and Deal with it Appropriately:

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

Percentage	Action
<40%	fill missing values with column mean value in case of numerical values
<40%	fill missing values with column median value in case of categorical values
>40%	Delete column if not needed
	Delete column if it is irrelevant or not necessary for our analysis

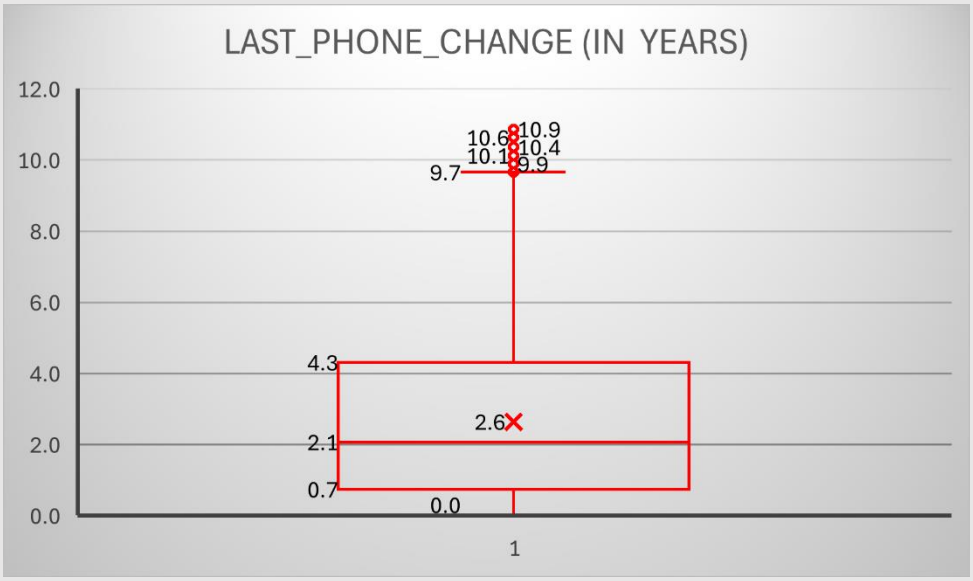
Observations:

- Many essential columns such as SK_ID_CURR, TARGET, and most demographic and regional columns have 0% missing values. These columns are already clean and ready for modelling.
- For columns like AMT_ANNUITY, AMT_GOODS_PRICE, EXT_SOURCE_2, EXT_SOURCE_3, CNT_FAM_MEMBERS, etc. we will Fill them with mean or mode values depending on type.
- Columns like OWN_CAR_AGE, EXT_SOURCE_1, all *_AVG, *_MODE, *_MEDI columns are not required for our analysis. These are being dropped to reduce noise and avoid imputation bias.

B. Identify Outliers in the Dataset:

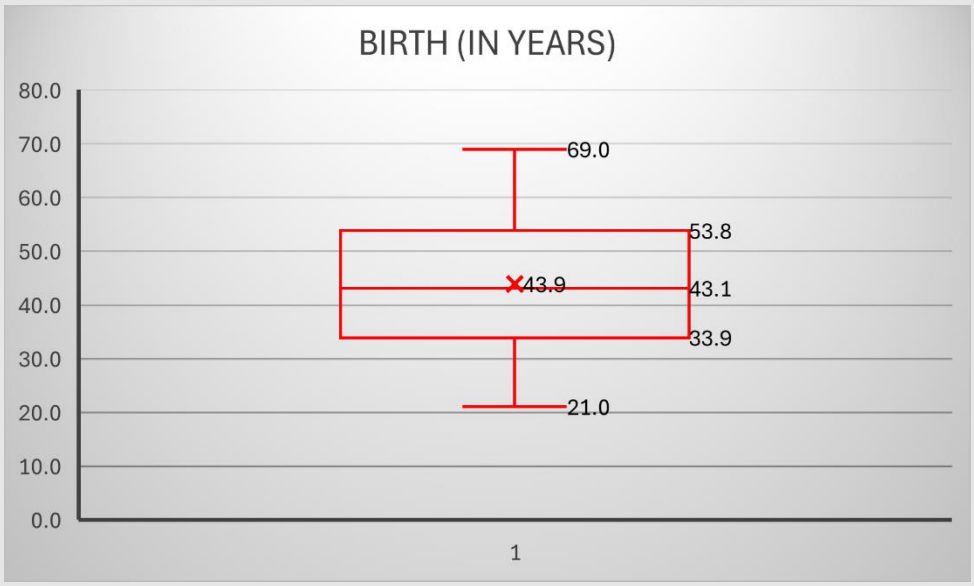
Task: Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

Columns	Outliers Counts
CNT_CHILDREN	723
AMT_INCOME_TOTAL	2295
AMT_CREDIT	1063
AMT_ANNUITY	1188
AMT_GOODS_PRICE	2387
DAYS_BIRTH	0
DAYS_EMPLOYED	9082
DAYS_REGISTRATION	96
ID_PUBLISH (IN YEARS)	0
LAST_PHONE_CHANGE (IN YEARS)	63

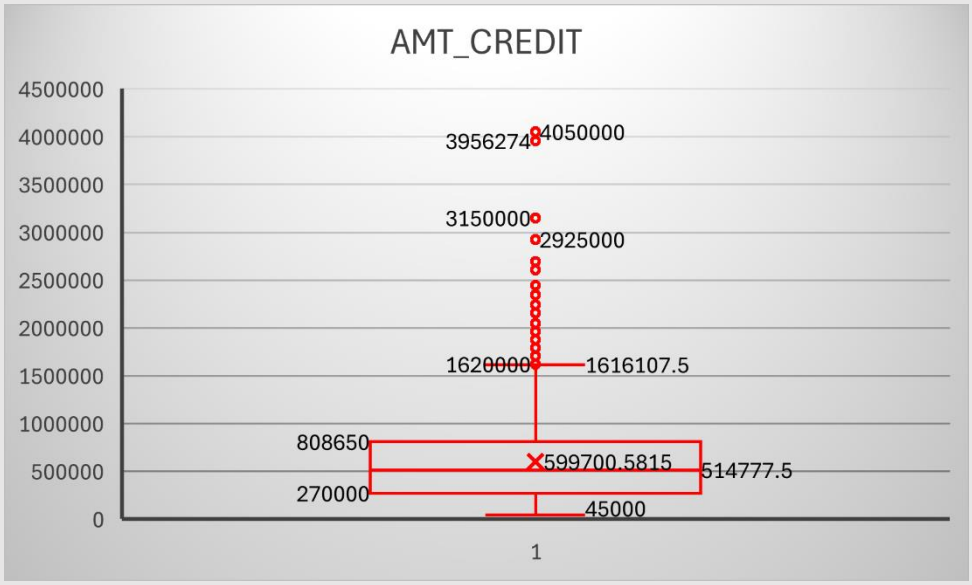


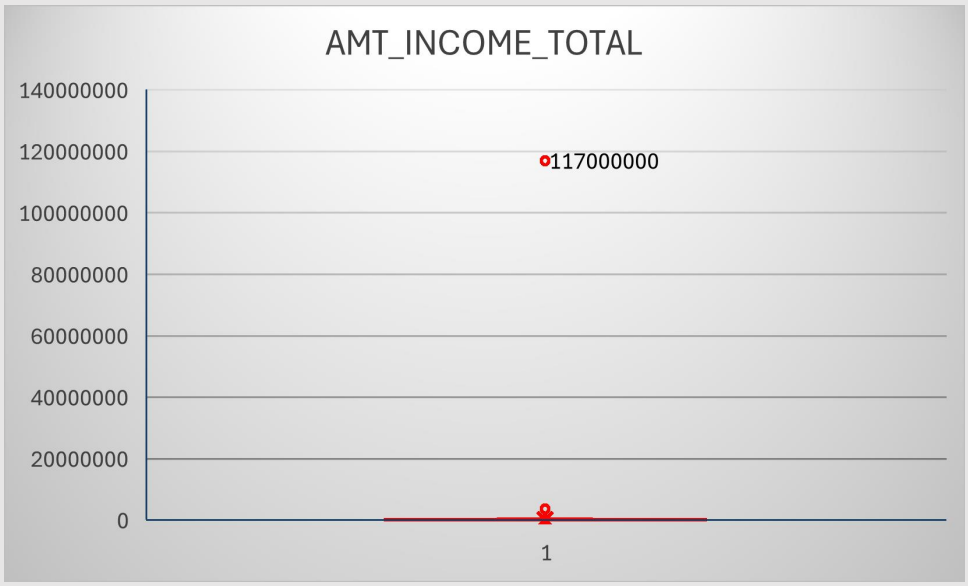
Here our Outliers are 10.9, 10.6 years etc. which is no. of years of client last change his phone and that can be possible.

Here Outliers are 184, 1000.7 etc. years and it is not possible that a person is employed more 40-45 years. Hence, this data is not valid.

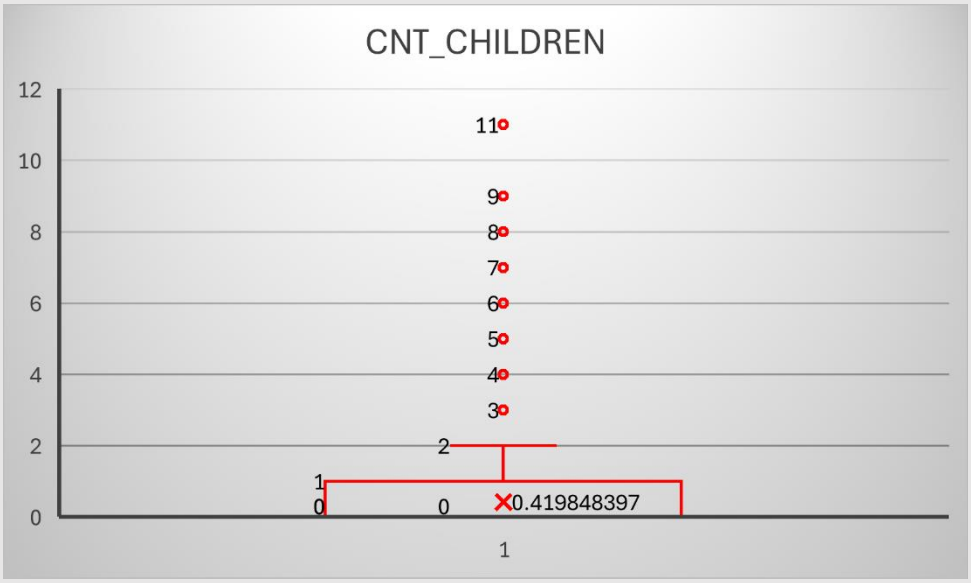


There is no Outliers in this data





Person having total income around Rs. 11,70,00,000 are possibly not required to take loan.



Children count more than 2-3 are very unusual hence, it is considered as Outlier.

Observations:

No outliers were found in BIRTH (IN YEARS) and ID_PUBLISH (IN YEARS).
High outliers in:

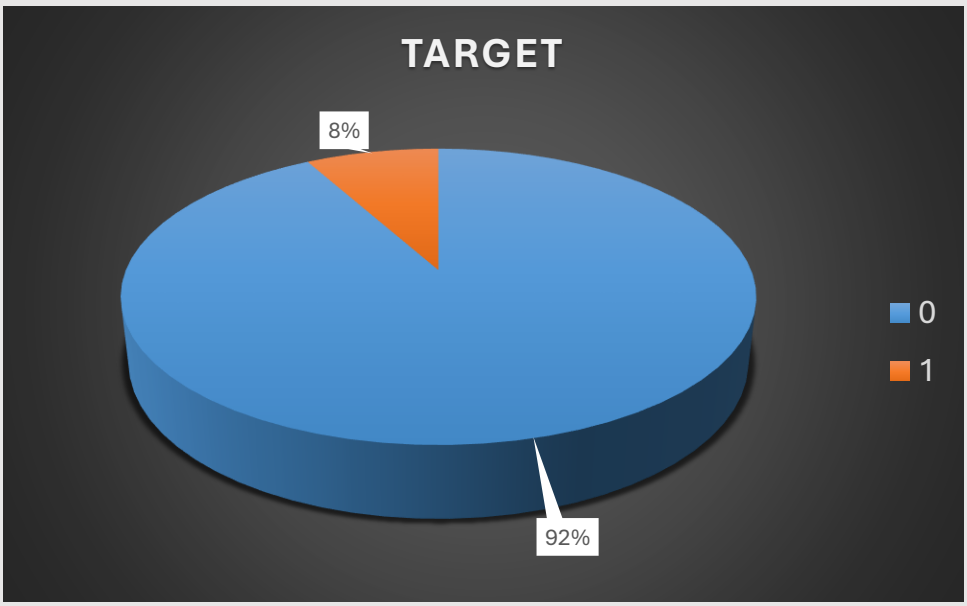
- EMPLOYED (IN YEARS) **(9,082)**: Possibly due to data entry errors or unrealistic employment durations (e.g., above 45 years).
- AMT_GOODS_PRICE **(2,387)** and AMT_INCOME_TOTAL **(2,295)**: Could be due to either very high-income individuals or incorrect entries.
- CNT_CHILDREN **(723)**: May include extreme family sizes or data input issues.

C. Analyze Data Imbalance

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

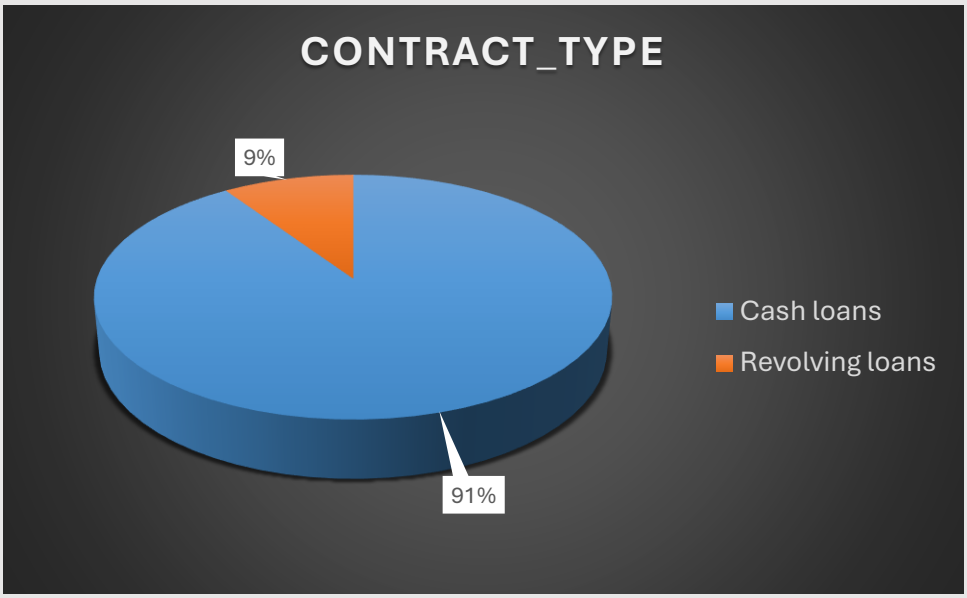
TARGET	COUNT
0	45973
1	4026

TARGET	% SHARE
0	91.95%
1	8.05%



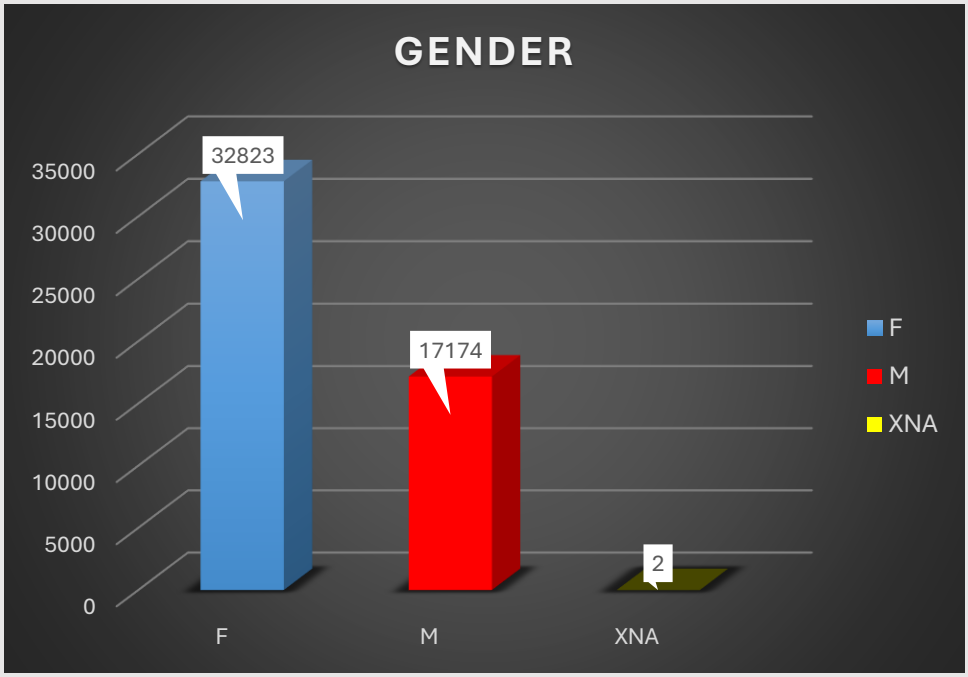
CONTRACT_TYPE	COUNT
Cash loans	45276
Revolving loans	4723

CONTRACT_TYPE	% SHARE
Cash loans	90.55%
Revolving loans	9.45%



GENDER	COUNT
F	32823
M	17174
XNA	2

GENDER	% SHARE
F	65.65%
M	34.35%
XNA	0.00%



Observations:

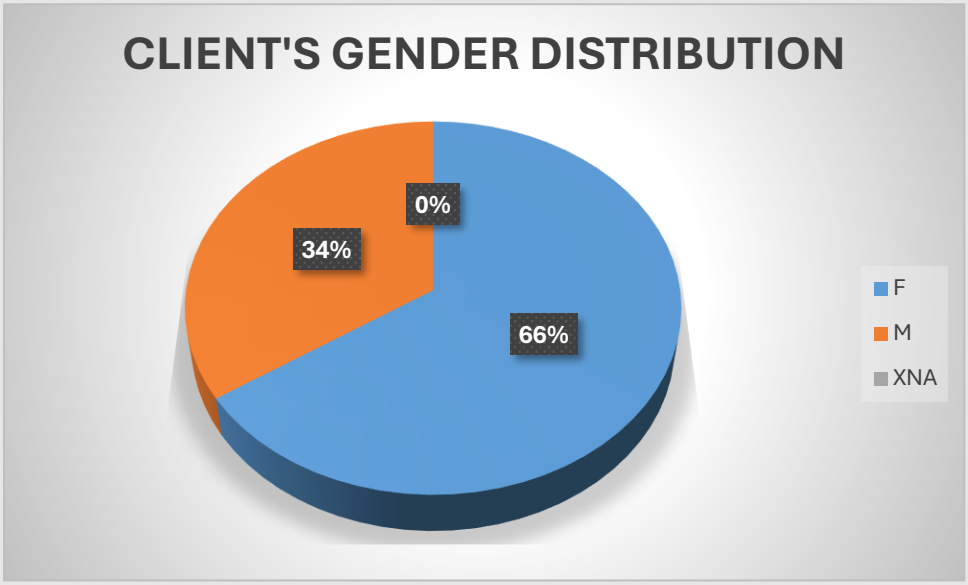
- In TARGET column, class imbalance exists with 91.95% of values being "0" (non-default) and only 8.05% being "1" (default).
- In NAME_CONTRACT_TYPE column, most contracts are "Cash loans" (90.55%), while only 9.45% are "Revolving loans", showing a significant imbalance.
- In CODE_GENDER column, majority are "F" (65.65%), while "M" represents 34.35%. An unusual value "XNA" appears with a very low proportion (0.004%), indicating possible data entry error.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

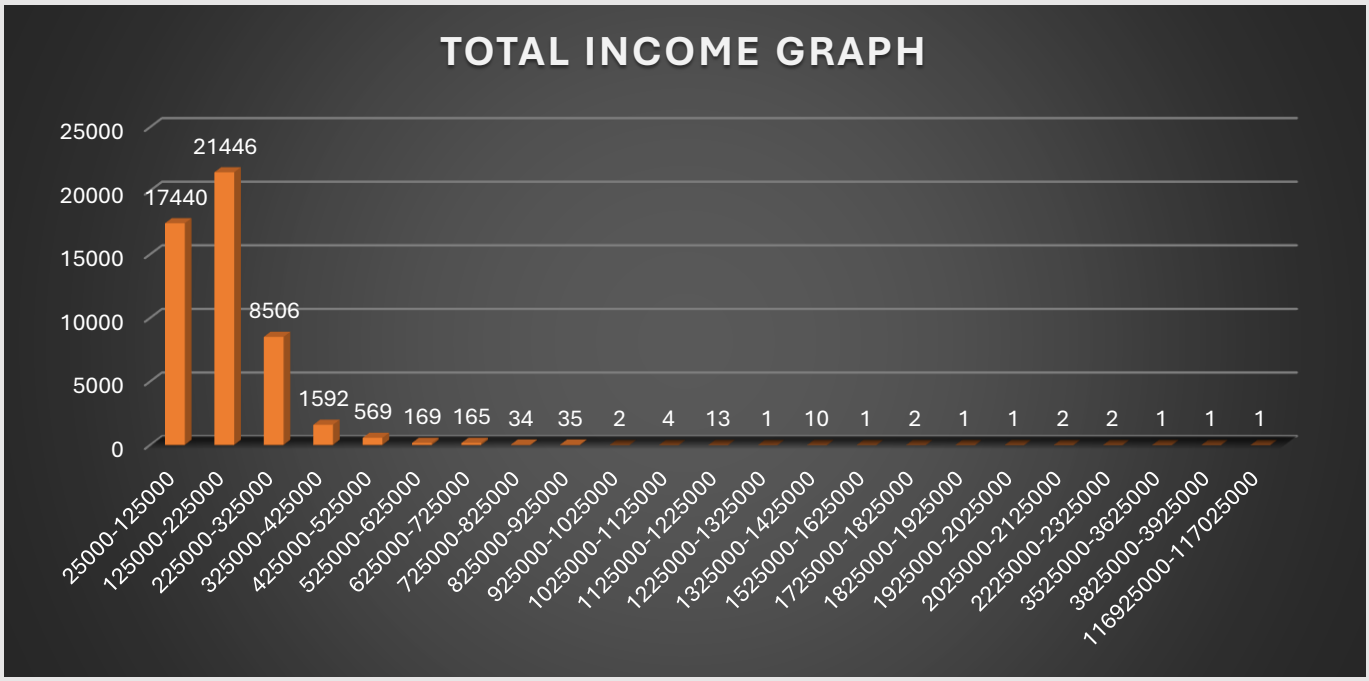
Task: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.

Univariate Analysis

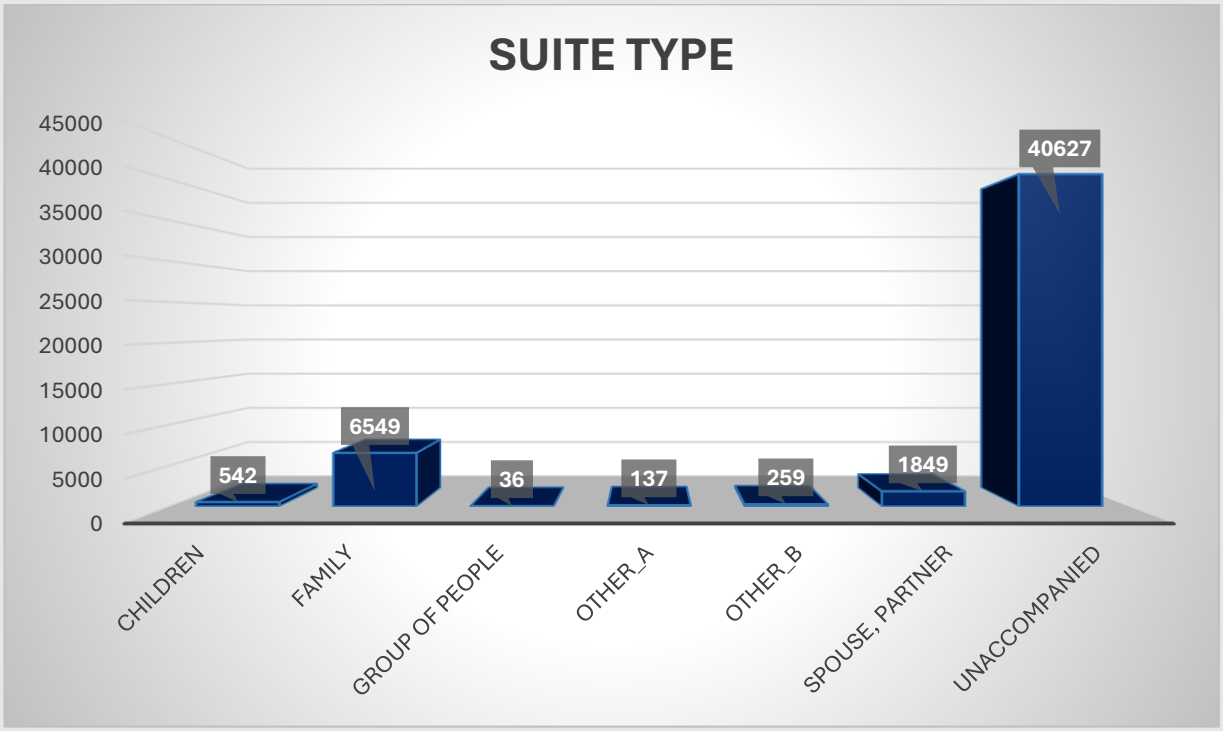
GENDER	COUNT
F	32823
M	17174
XNA	2



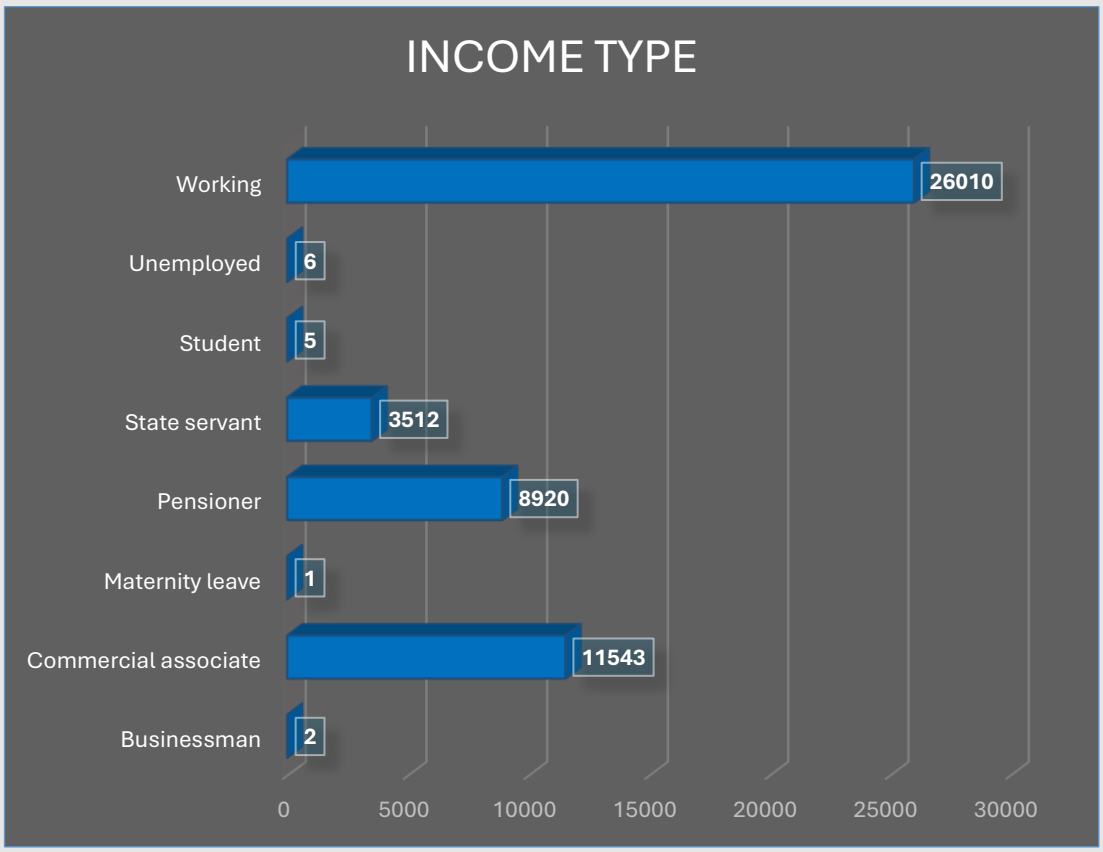
INCOME RANGE	COUNT
25000-125000	17440
125000-225000	21446
225000-325000	8506
325000-425000	1592
425000-525000	569
525000-625000	169
625000-725000	165
725000-825000	34
825000-925000	35
925000-1025000	2
1025000-1125000	4
1125000-1225000	13
1225000-1325000	1
1325000-1425000	10
1525000-1625000	1
1725000-1825000	2
1825000-1925000	1
1925000-2025000	1
2025000-2125000	2
2225000-2325000	2
3525000-3625000	1
3825000-3925000	1
116925000-117025000	1



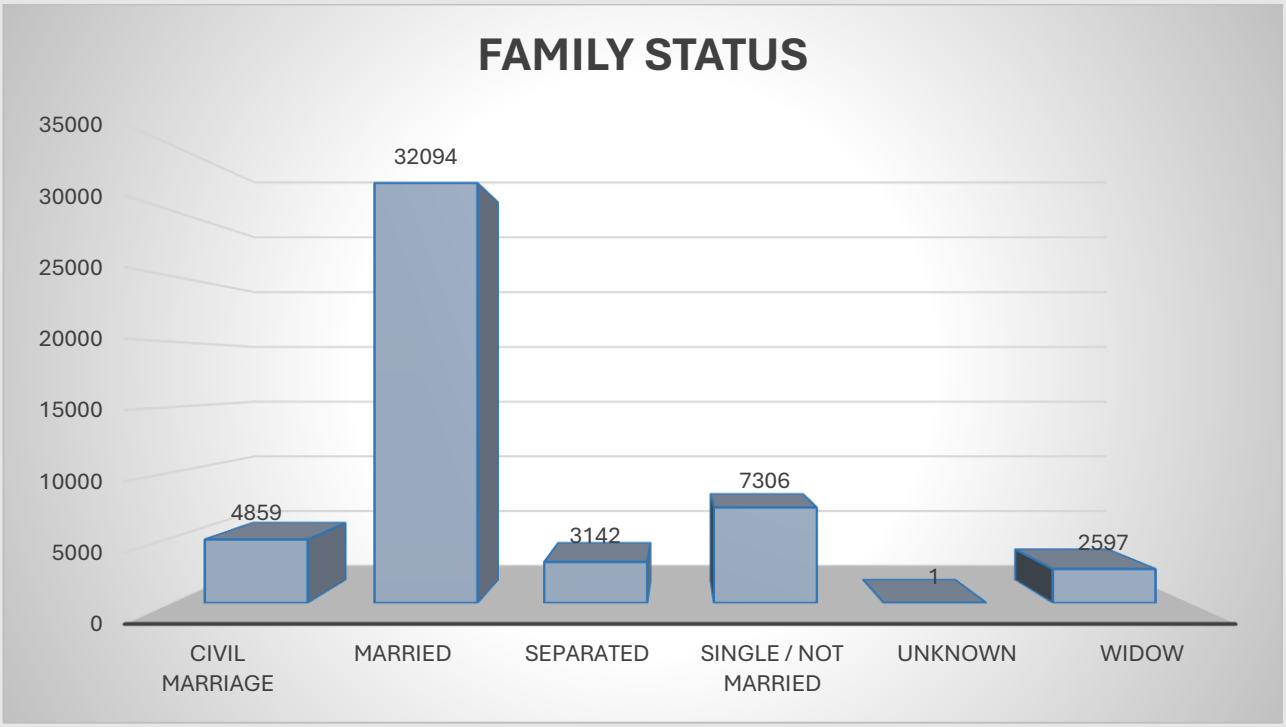
SUITE TYPE	COUNT
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40627



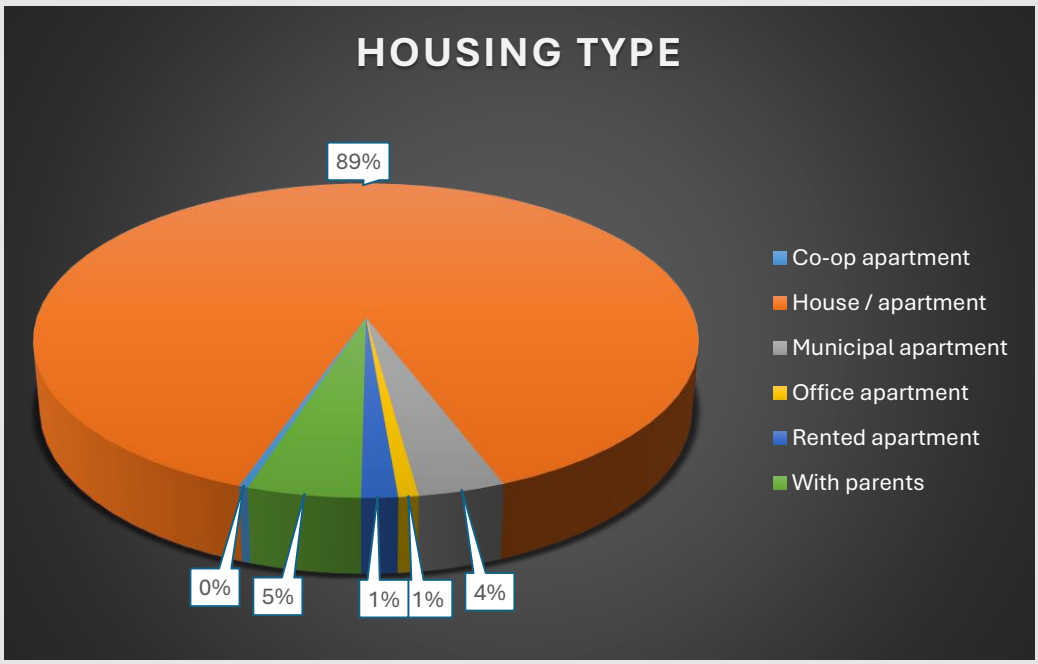
INCOME TYPE	COUNT
Businessman	2
Commercial associate	11543
Maternity leave	1
Pensioner	8920
State servant	3512
Student	5
Unemployed	6
Working	26010



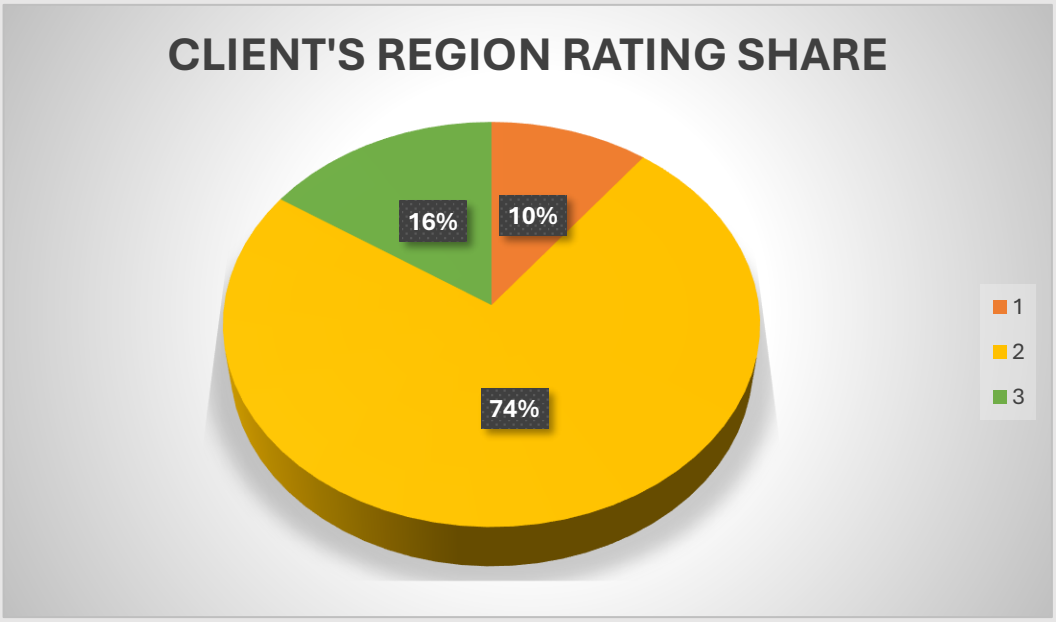
FAMILY STATUS	COUNT
Civil marriage	4859
Married	32094
Separated	3142
Single / not married	7306
Unknown	1
Widow	2597



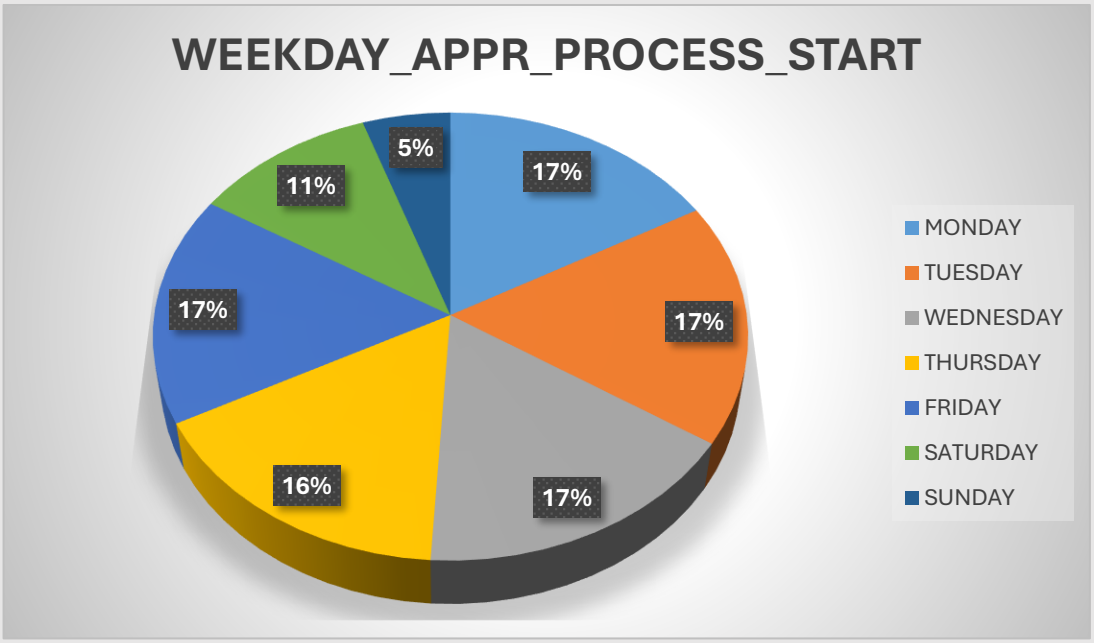
HOUSING TYPE	COUNT
Co-op apartment	191
House / apartment	44368
Municipal apartment	1845
Office apartment	427
Rented apartment	769
With parents	2399



CLIENT'S REGION RATING	COUNT
1	5226
2	36964
3	7809



WEEKDAY_APPR_PROCESS_START	COUNT
MONDAY	8385
TUESDAY	8741
WEDNESDAY	8355
THURSDAY	8149
FRIDAY	8286
SATURDAY	5467
SUNDAY	2616

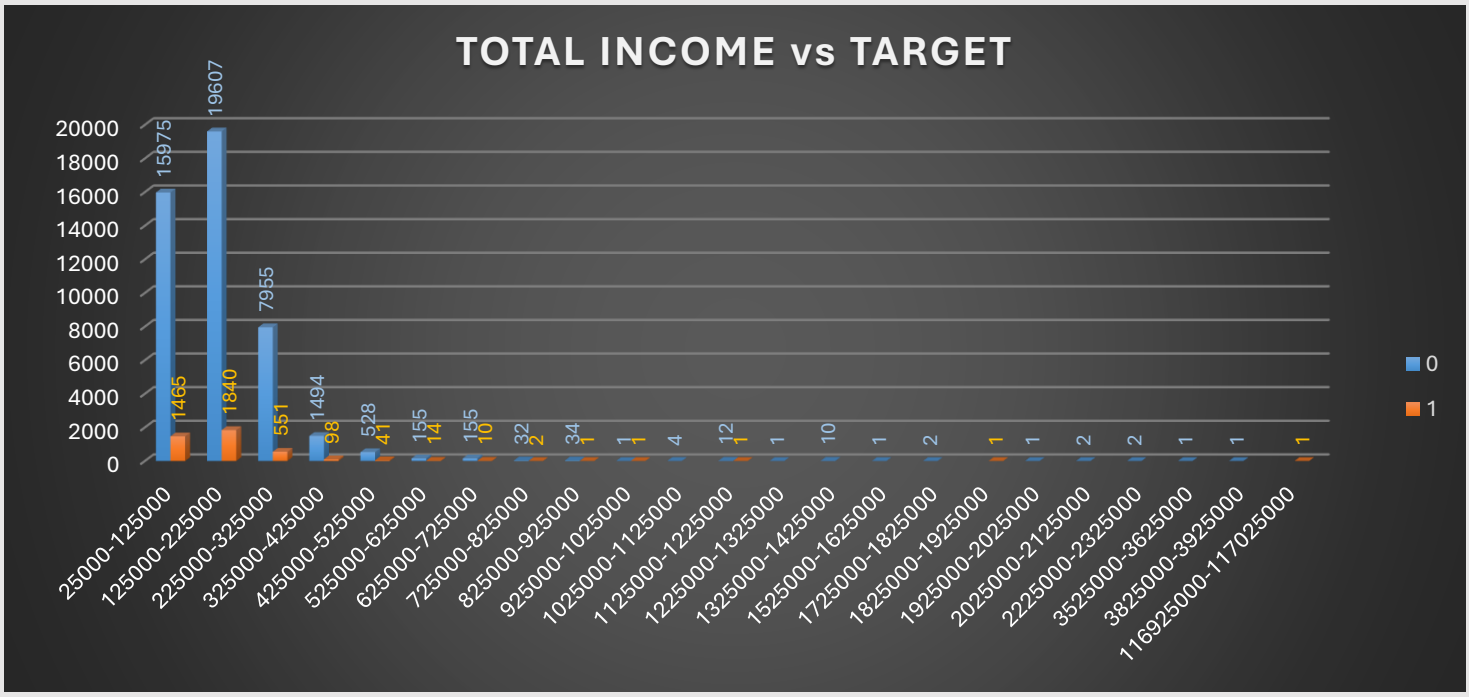


Observations:

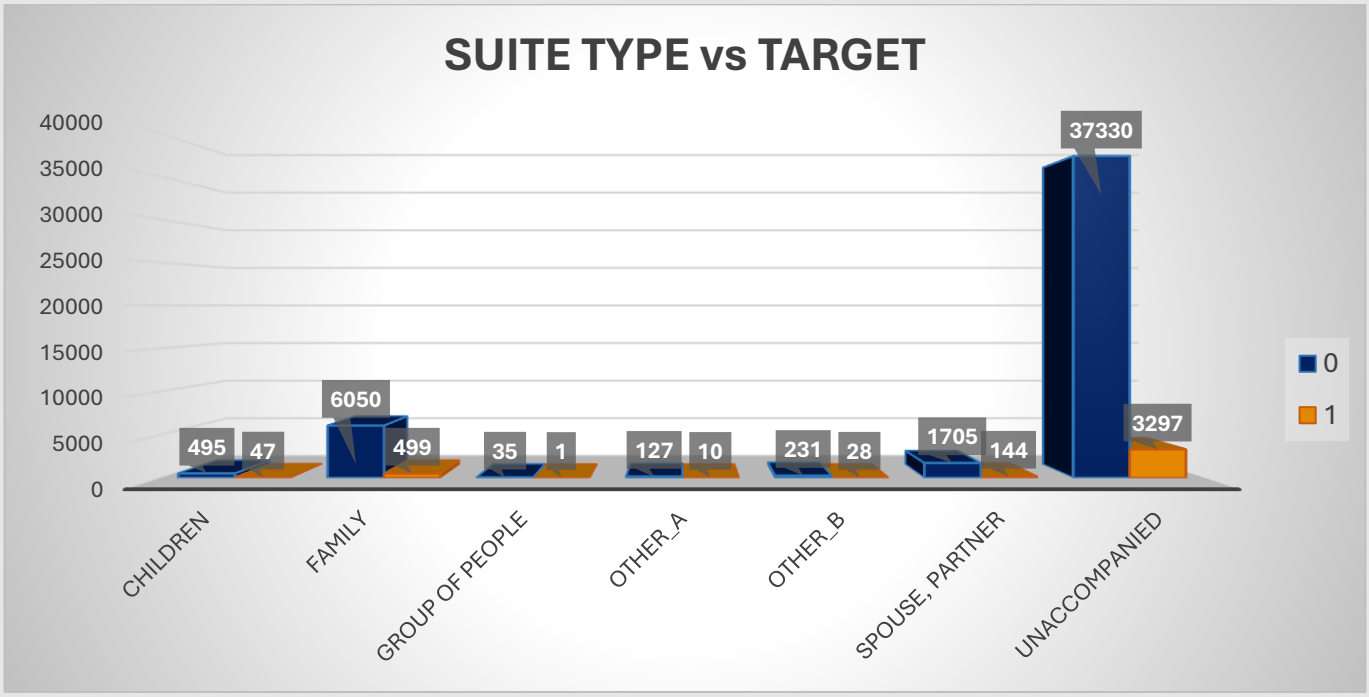
- Majority of the clients are **female**. The presence of "**XNA**" indicates data entry errors.
- The vast majority of loans are concentrated in the "**25,000 - 225,000**" range, which accounts for **38,886** loans.
- As income increases beyond **225,000**, the number of loans drops significantly.
- Beyond **525,000**, the loan counts become **extremely low,** with most ranges having fewer than **200** loans.
- Most clients apply for loans **alone (Unaccompanied)**. **Family** is the second most common option.
- The majority of applicants are working individuals **52.02%**. The "**Businessman**" and "**Maternity Leave**" categories are rare.
- **Married** applicants are in majority.
- Most applicants live in a **house/apartment**. Very few live in **co-op or office apartments.**
- Most regions are rated around **2**.

Bivariate Analysis

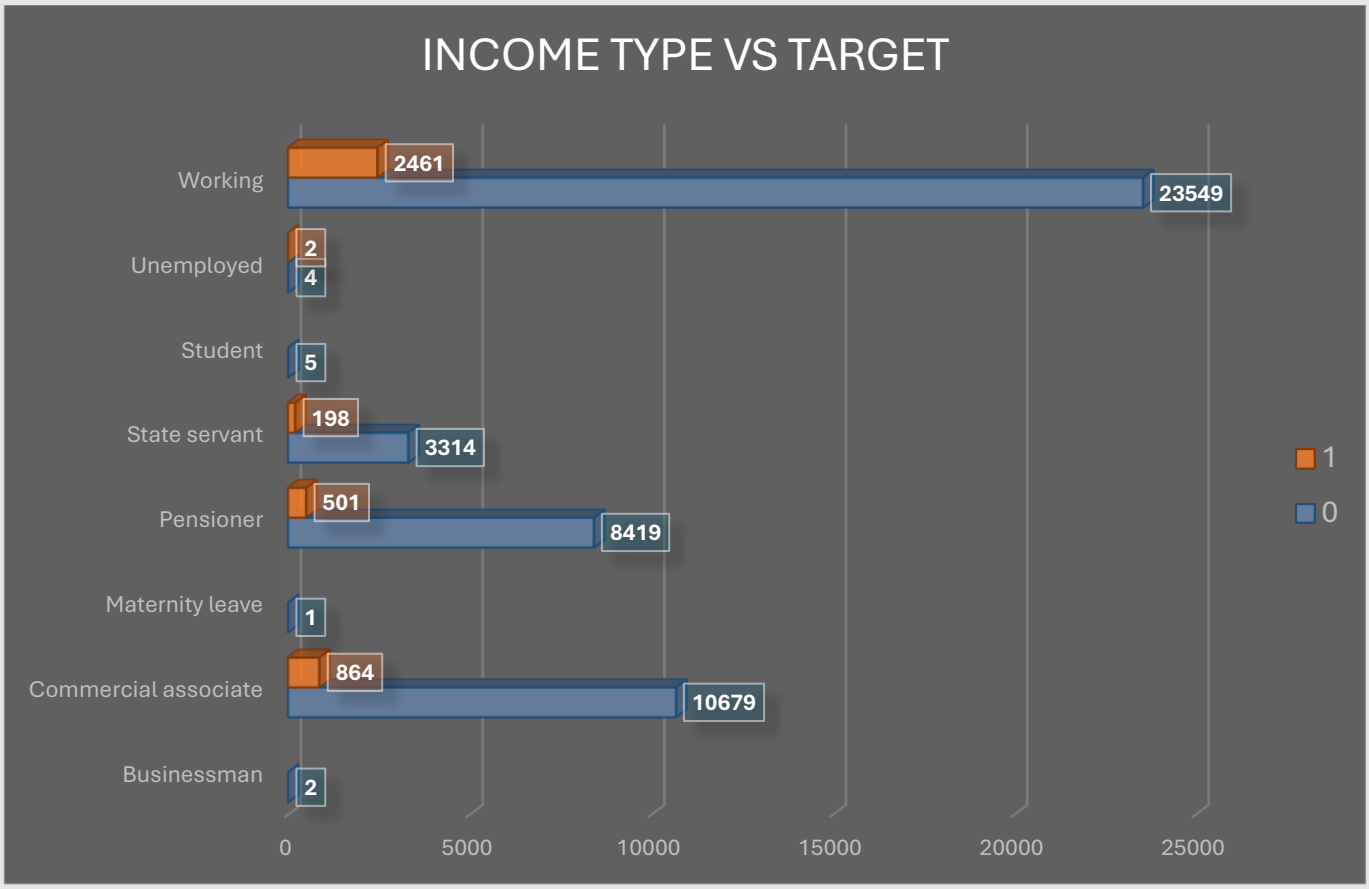
INCOME RANGE	TARGET COUNT	
	0	1
25000-125000	15975	1465
125000-225000	19607	1840
225000-325000	7955	551
325000-425000	1494	98
425000-525000	528	41
525000-625000	155	14
625000-725000	155	10
725000-825000	32	2
825000-925000	34	1
925000-1025000	1	1
1025000-1125000	4	
1125000-1225000	12	1
1225000-1325000	1	
1325000-1425000	10	
1525000-1625000	1	
1725000-1825000	2	
1825000-1925000		1
1925000-2025000	1	
2025000-2125000	2	
2225000-2325000	2	
3525000-3625000	1	
3825000-3925000	1	
116925000-117025000		1



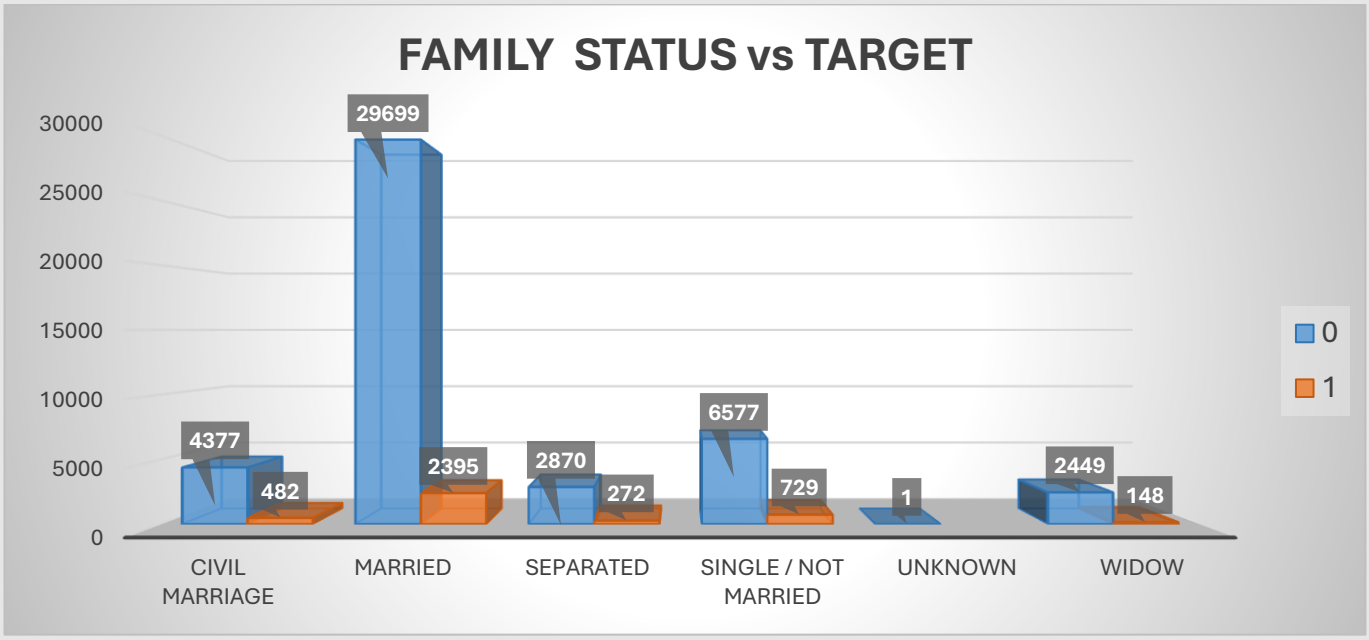
SUITE TYPE	TARGET COUNT	
	0	1
Children	495	47
Family	6050	499
Group of people	35	1
Other_A	127	10
Other_B	231	28
Spouse, partner	1705	144
Unaccompanied	37330	3297



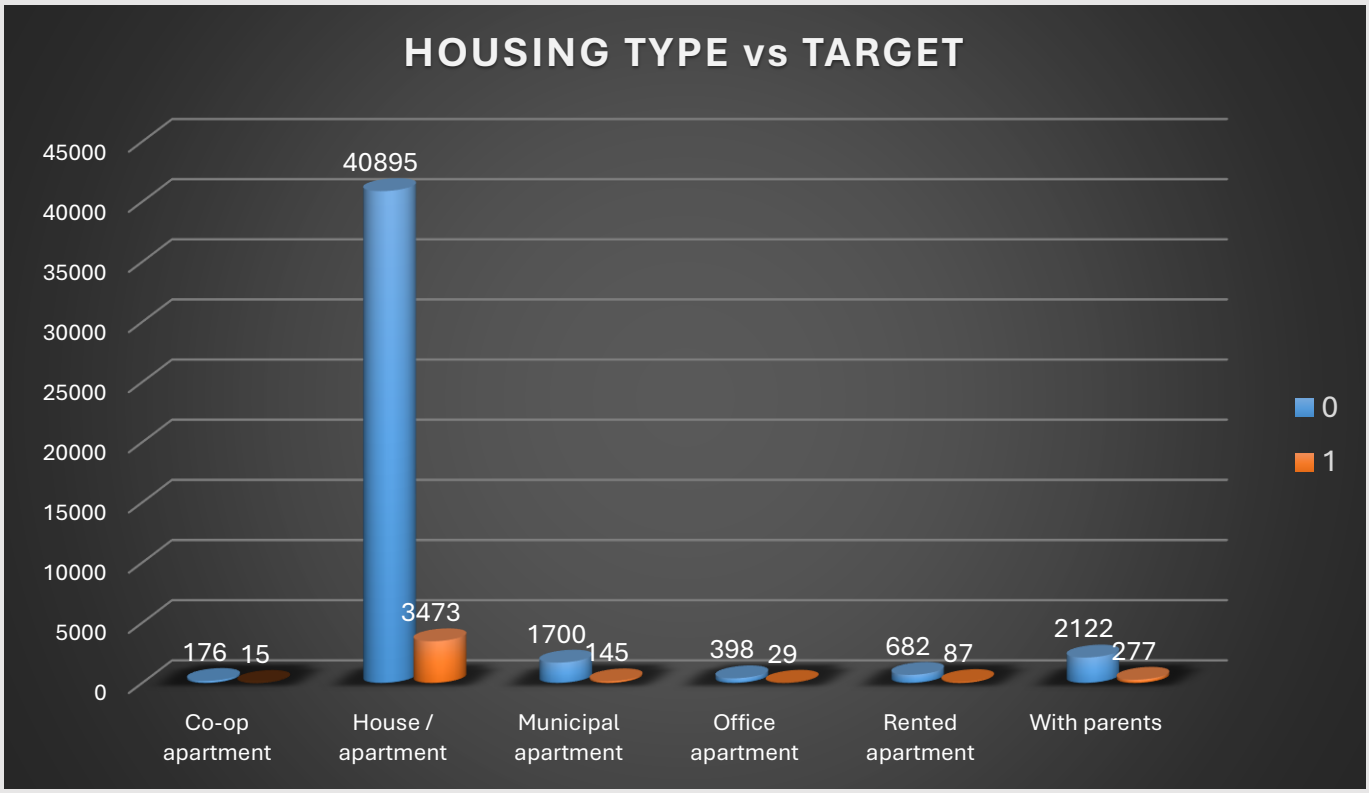
INCOME TYPE	TARGET COUNT	
	0	1
Businessman	2	
Commercial associate	10679	864
Maternity leave	1	
Pensioner	8419	501
State servant	3314	198
Student	5	
Unemployed	4	2
Working	23549	2461



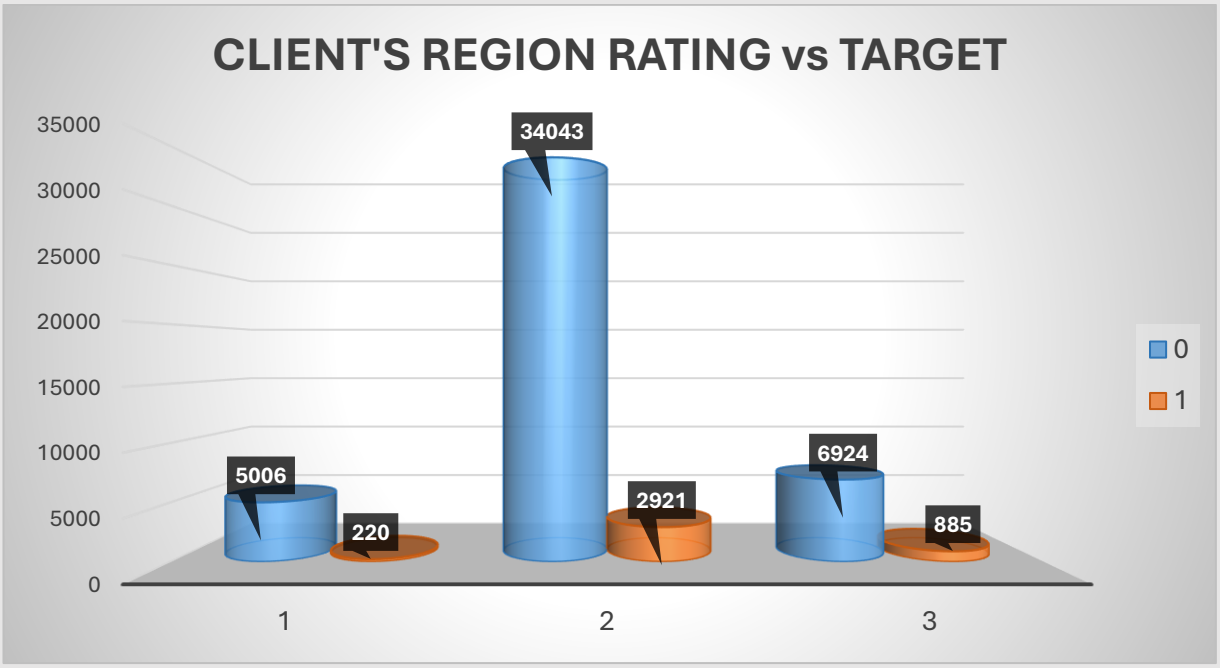
FAMILY STATUS	TARGET COUNT	
	0	1
Civil marriage	4377	482
Married	29699	2395
Separated	2870	272
Single / not married	6577	729
Unknown	1	
Widow	2449	148



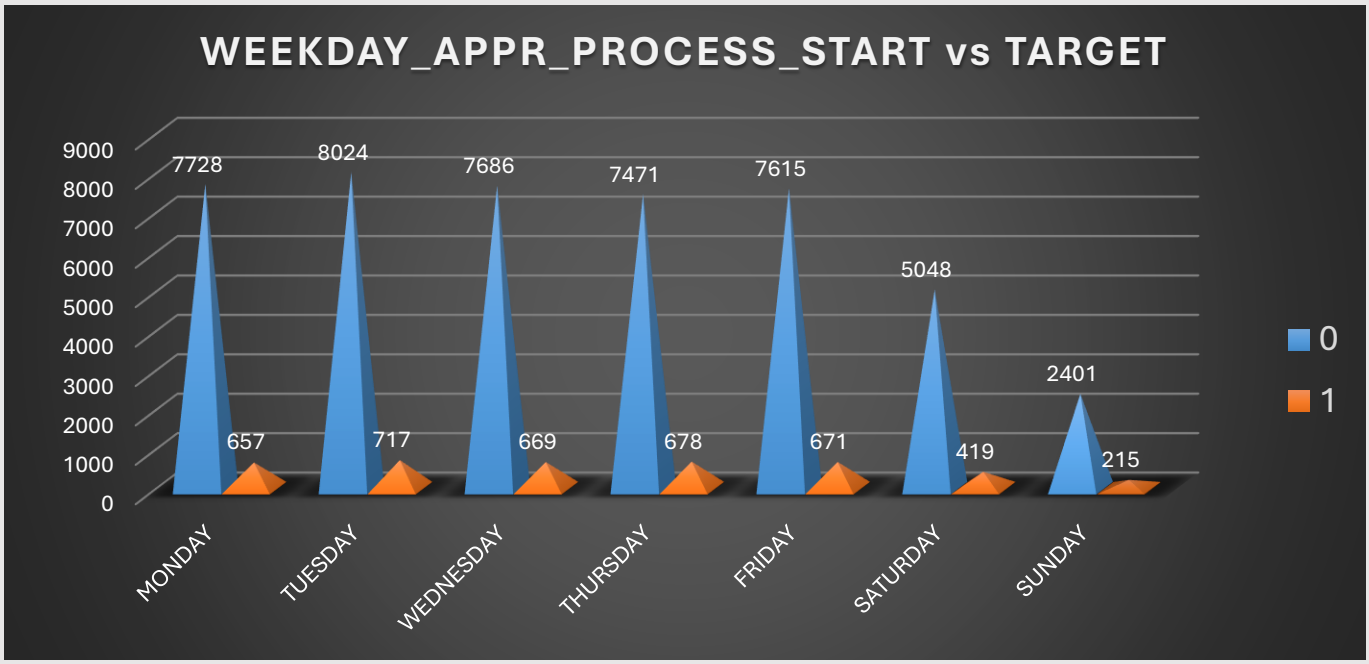
HOUSING TYPE	TARGET COUNT	
	0	1
Co-op apartment	176	15
House / apartment	40895	3473
Municipal apartment	1700	145
Office apartment	398	29
Rented apartment	682	87
With parents	2122	277



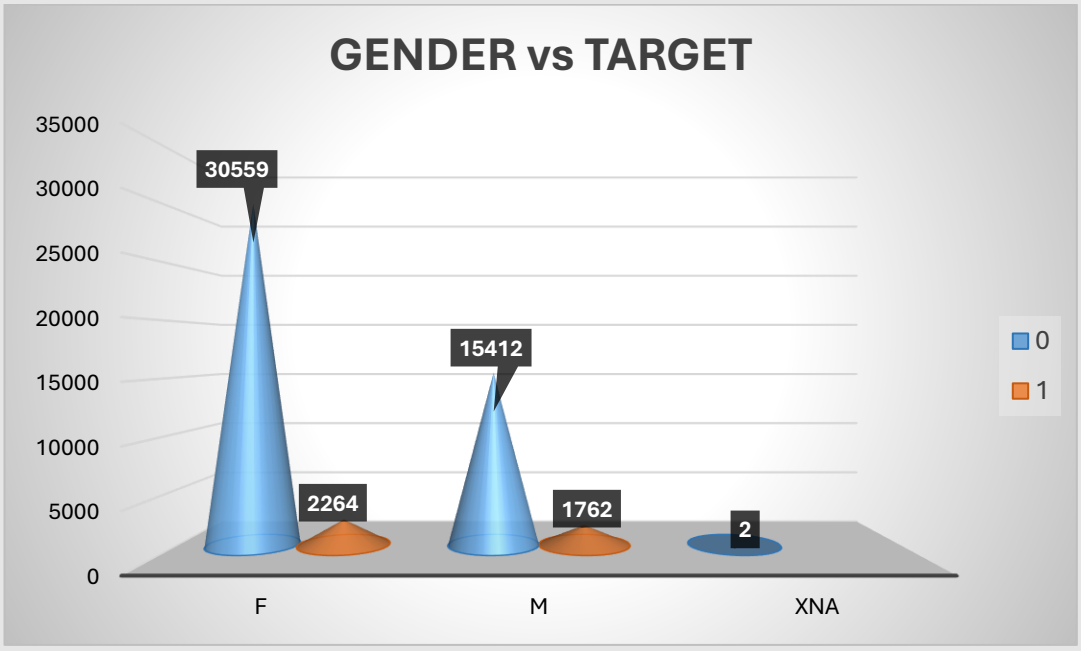
CLIENT'S REGION RATING	TARGET COUNT	
	0	1
1	5006	220
2	34043	2921
3	6924	885



WEEKDAY_APPR_PROCESS_START	TARGET COUNT	
	0	1
MONDAY	7728	657
TUESDAY	8024	717
WEDNESDAY	7686	669
THURSDAY	7471	678
FRIDAY	7615	671
SATURDAY	5048	419
SUNDAY	2401	215



GENDER	TARGET COUNT	
	0	1
F	30559	2264
M	15412	1762
XNA	2	



Observations:

- **Males** have a higher default rate than **females**.
- Default rate= default cases / total no. of cases in particular range or category
- The **highest default** rates are seen in the lower income ranges **(25,000 - 225,000).**
- The default rate decreases as income increases, but this is partially due to the drastic reduction in the number of loans in higher ranges.
- Clients applying **alone** or with **children** show slightly higher default rates.
- The "**Other_B**" group has the h**ighest default rate**.
- **Working** and **Commercial associate** have the highest default cases, indicating high risk. While **Pensioners** and **State Servants** show the lowest risk.
- **Single** and **Civil Marriage** clients have the **highest default rates**. But **default cases** are higher among **Married** clients.
- Clients living with **parents** or in **rented apartments** are at higher risk of default.

E. Identify Top Correlations for Different Scenarios

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Top 5 Correlations (Target=0)	
AMT_ANNUITY & AMT_INCOME_TOTAL	0.451135629
AMT_ANNUITY & AMT_CREDIT	0.770771802
AMT_GOODS_PRICE & AMT_CREDIT	0.986904954
AMT_GOODS_PRICE & AMT_ANNUITY	0.775727492
EMPLOYED (IN YEARS) & BIRTH (IN YEARS)	0.623474675

Observations:

- Strong correlations between AMT_GOODS_PRICE, AMT_CREDIT, and AMT_ANNUITY indicate that non-defaulters tend to have consistent financial behaviour. Their loan amounts, goods prices, and annuity payments are closely tied, reflecting prudent financial decisions.
- The strong correlation between EMPLOYED (IN YEARS) and BIRTH (IN YEARS) shows that older clients typically have longer employment histories, which is natural. It reflects that age and work experience are connected for non-defaulters.
- A strong correlation between AMT_ANNUITY and AMT_CREDIT indicates that higher credit amounts are associated with higher annuity payments, which is expected since larger loans generally have higher repayments.

Top 5 Correlations (Target=1)	
AMT_ANNUITY & AMT_CREDIT	0.749665201
AMT_GOODS_PRICE & AMT_CREDIT	0.982130206
AMT_GOODS_PRICE & AMT_ANNUITY	0.74932991
EMPLOYED (IN YEARS) & BIRTH (IN YEARS)	0.588242824
ID_PUBLISH (IN YEARS) & BIRTH (IN YEARS)	0.247896571

Observations:

- Similar to non-defaulters, defaulters also show a very strong relationship between AMT_GOODS_PRICE & AMT_CREDIT i.e. the value of goods purchased and the credit amount. However, despite this strong correlation, these clients still default and struggle with repayment.
- A weak but positive correlation between ID publication age and birth indicates that older defaulters are more likely to have older identification documents.

Insights

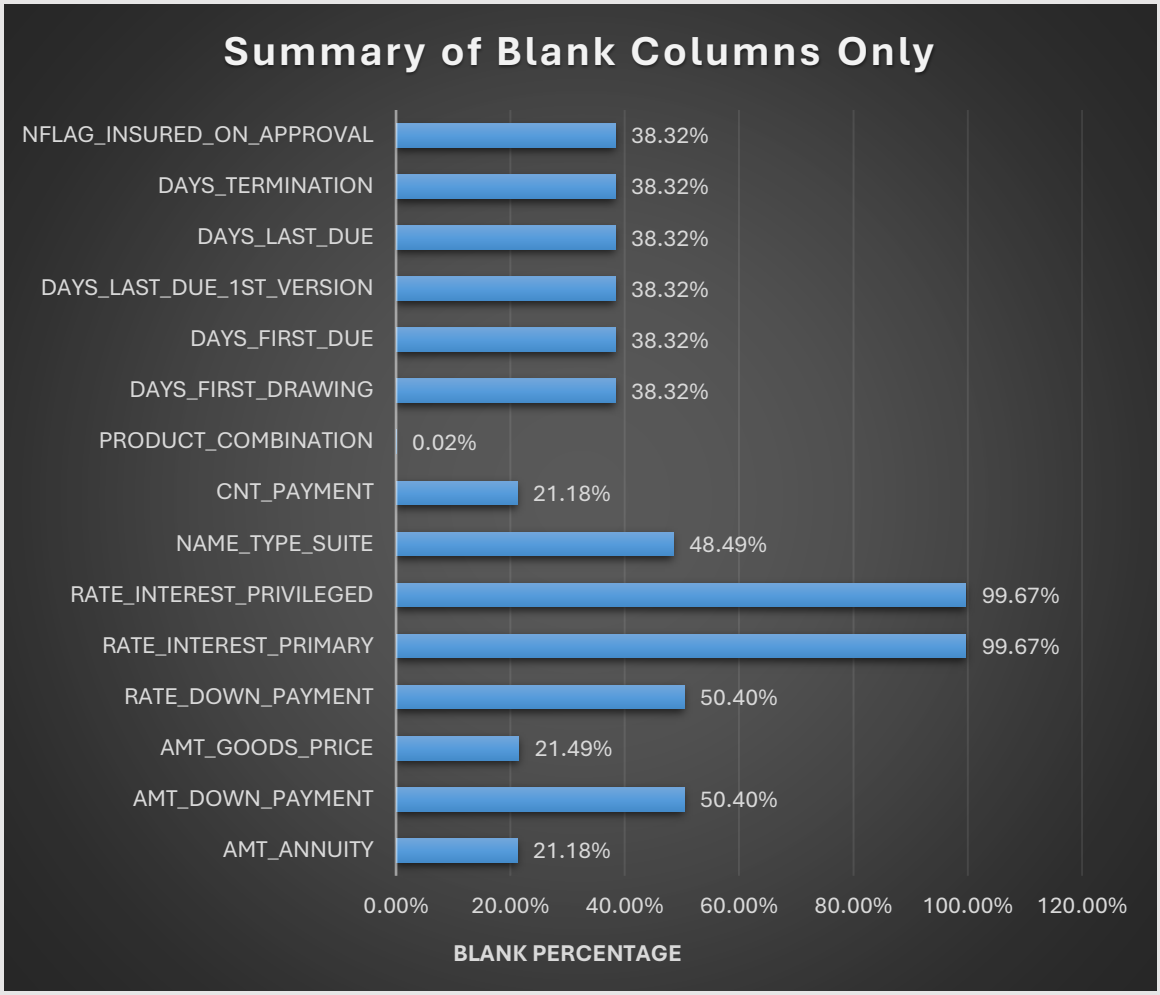
- Clean and complete demographic data (0% missing) ensures strong model performance without additional preprocessing.
- Filling missing values in financial columns with mean/mode maintains data consistency without heavy bias.
- Dropping columns like OWN_CAR_AGE and EXT_SOURCE_1 avoids noise from highly missing data, improving model reliability.
- High outliers in EMPLOYED (IN YEARS), AMT_GOODS_PRICE, and CNT_CHILDREN indicate potential data entry errors, which may skew model predictions if not treated.
- Severe class imbalance in TARGET (92% non-default) suggests a need for balancing techniques (SMOTE, class weighting) to prevent bias toward non-defaults.
- Majority of loans being cash loans and most clients being female reflect the product's primary target group.
- Loan distribution (25,000 - 225,000) shows most clients are in the lower income range, where default rates are higher, suggesting higher-risk clients dominate.
- Higher default rates among males, singles, and clients applying alone highlight vulnerable groups needing stricter credit assessment.
- Strong positive correlation between AMT_GOODS_PRICE, AMT_CREDIT, and AMT_ANNUITY suggests consistent spending and repayment patterns among non-defaulters, but defaulters lack this discipline.

For previous_application.xlsx Dataset

Data Analytics Tasks

A. Identify Missing Data and Deal with it Appropriately

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



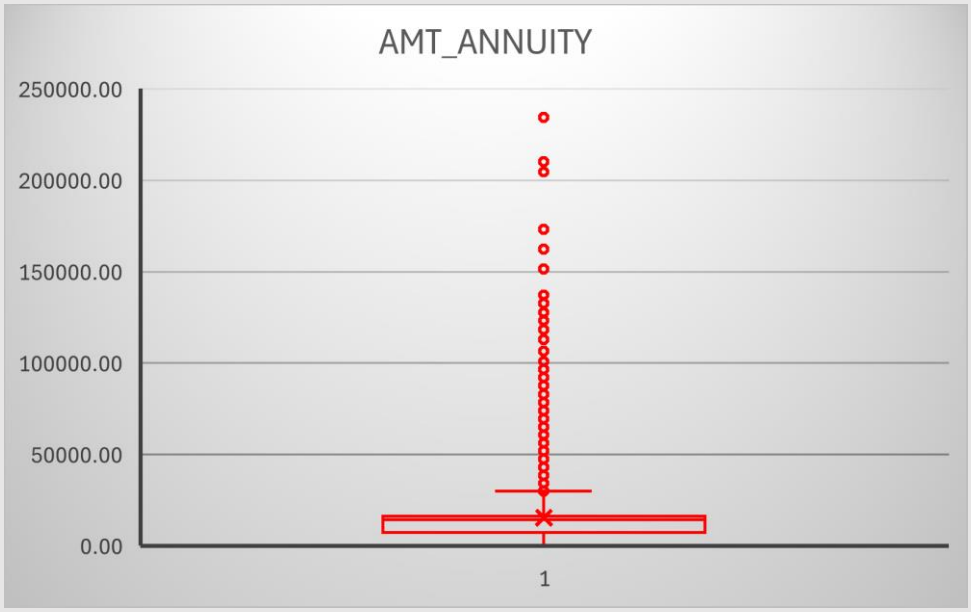
Observations:

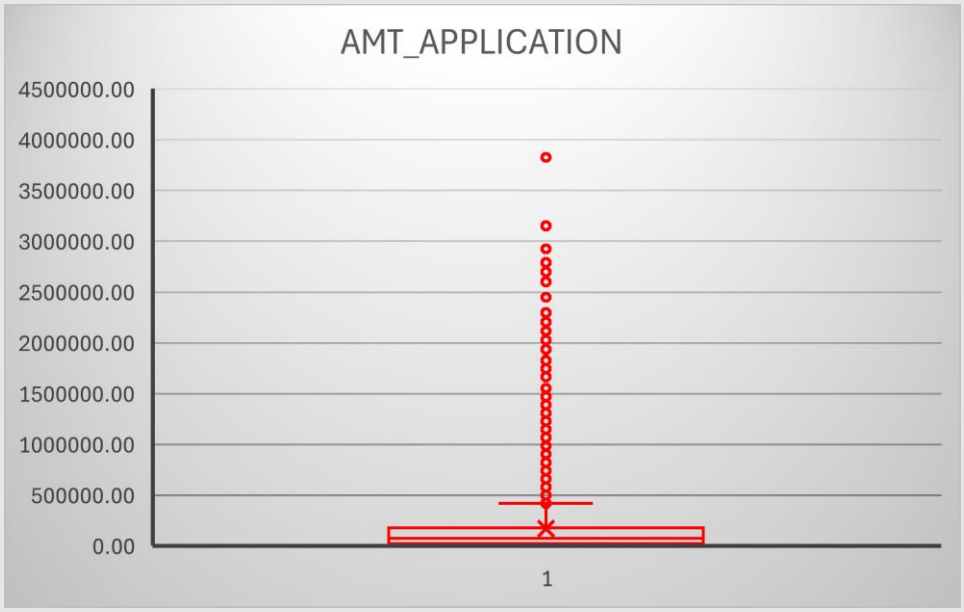
- Key columns like SK_ID_CURR, SK_ID_PREV, and most categorical columns have 0% missing values and are ready for modelling.
- Columns such as AMT_ANNUITY, AMT_GOODS_PRICE, CNT_PAYMENT, and several DAYS_* fields will be imputed using mean or mode, as they have moderate missingness **(20–38%)** but carry useful information.
- Columns like AMT_DOWN_PAYMENT, RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, and RATE_INTEREST_PRIVILEGED are dropped due to high missing rates **(over 40–99%),** reducing potential noise.
- Categorical fields like NAME_TYPE_SUITE and PRODUCT_COMBINATION will be retained using mode imputation.
- Columns like NFLAG_LAST_APPL_IN_DAY and NAME_YIELD_GROUP are excluded despite no missing data, likely due to low relevance.

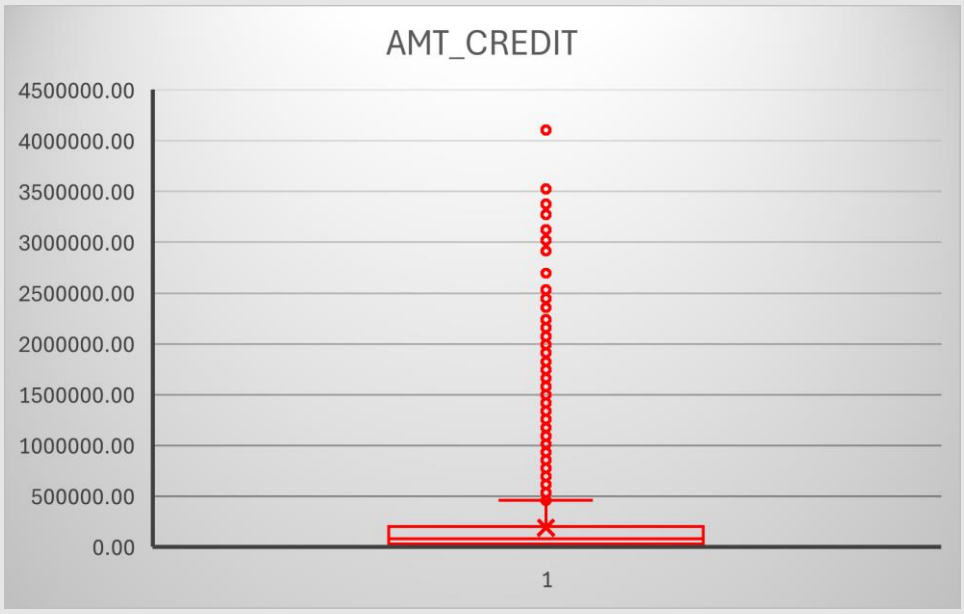
B. Identify Outliers in the Dataset

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

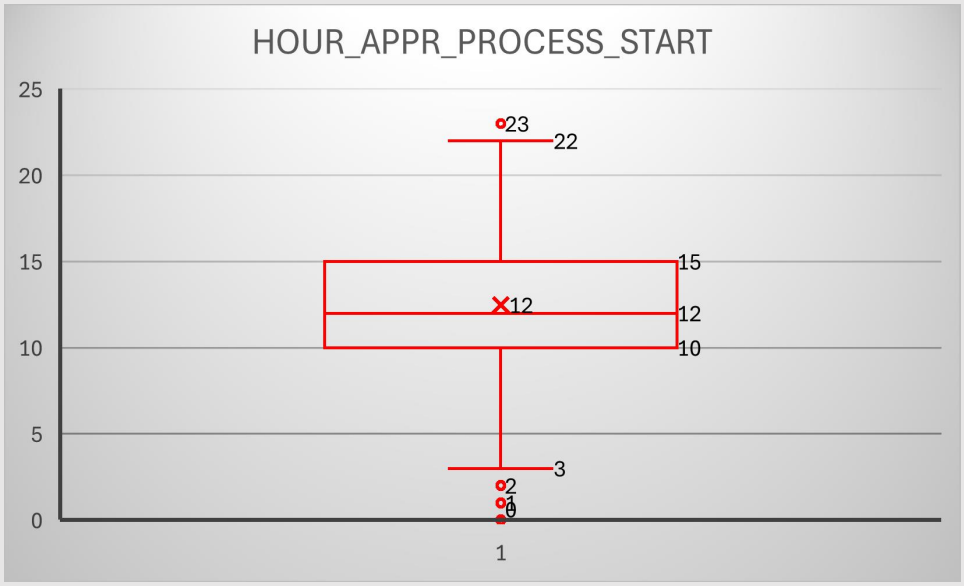
Columns	Outliers Counts
AMT_ANNUITY	4922
AMT_APPLICATION	5792
AMT_CREDIT	5648
AMT_GOODS_PRICE	5674
HOUR_APPR_PROCESS_START	40
DECISION (IN YEARS)	127
CNT_PAYMENT	14011
FIRST_DUE (IN YEARS)	1290
LAST_DUE_1ST_VERSION (IN YEARS)	2715
LAST_DUE (IN YEARS)	6545
TERMINATION (IN YEARS)	6960

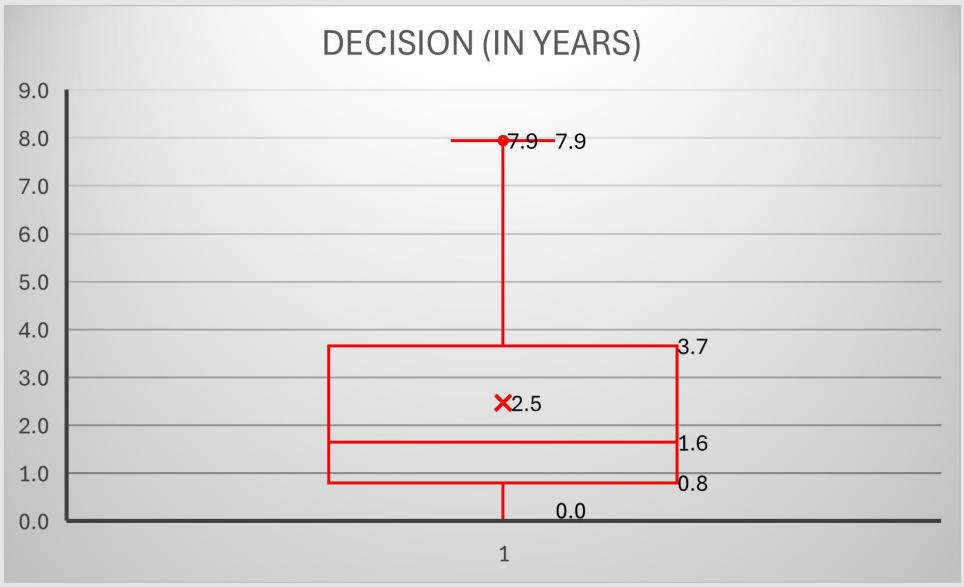


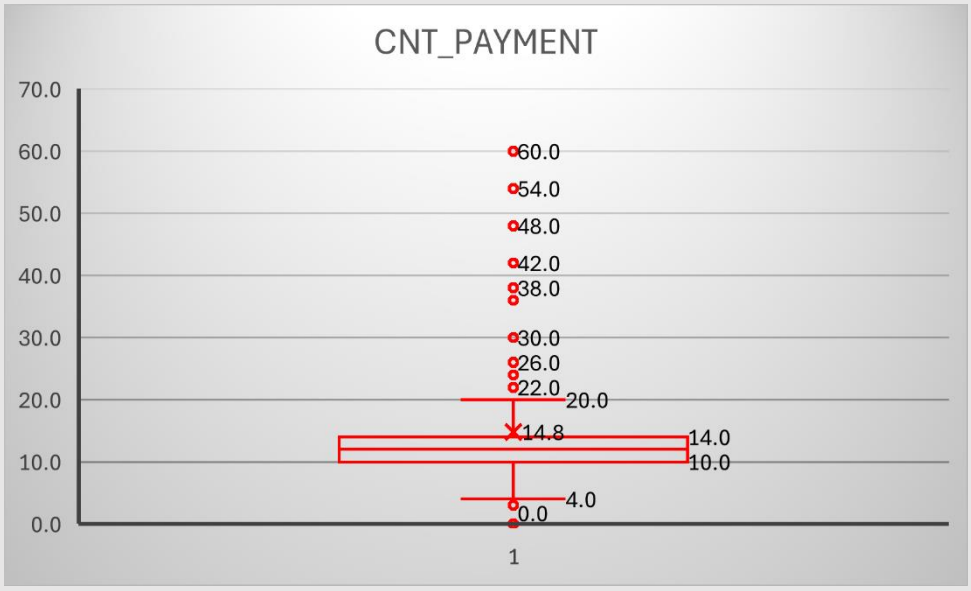


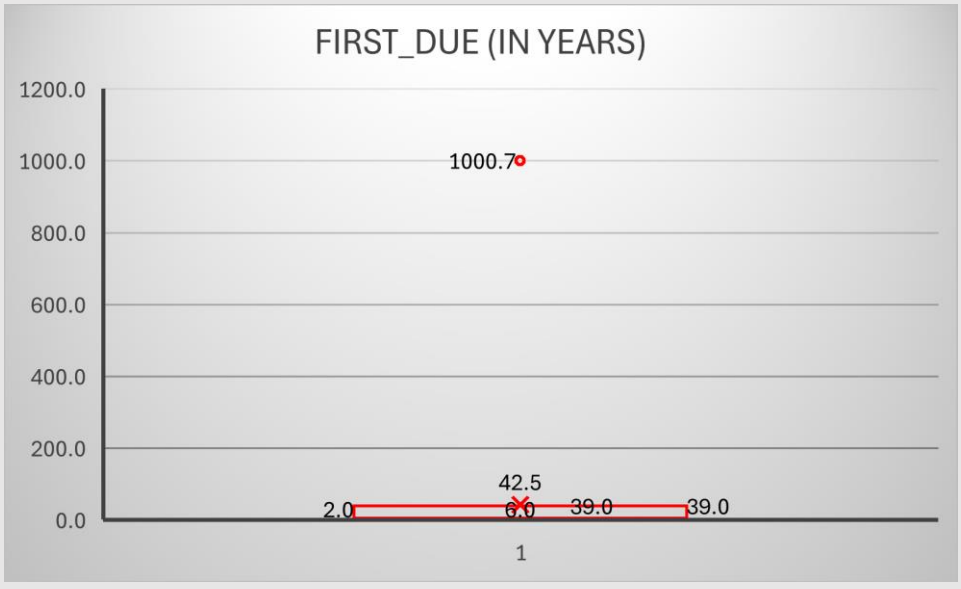

AMT_CREDIT
4500000.00
4000000.00
3500000.00
3000000.00
2500000.00
2000000.00
1500000.00
1000000.00
500000.00
0.00
1

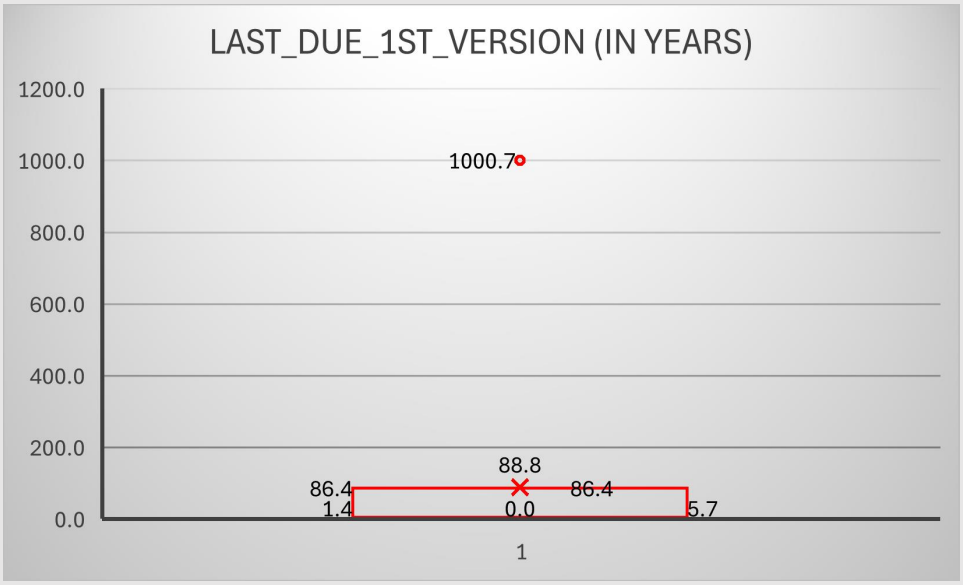

AMT_GOODS_PRICE
4500000.00
4000000.00
3500000.00
3000000.00
2500000.00
2000000.00
1500000.00
1000000.00
500000.00
0.00
1


HOUR_APPR_PROCESS_START
25
20
15
10
5
0
23
22
15
12
10
3
2
1
0
1


DECISION (IN YEARS)
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
7.9
7.9
3.7
2.5
1.6
0.8
0.0
1


CNT_PAYMENT
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
60.0
54.0
48.0
42.0
38.0
30.0
26.0
22.0
20.0
14.8
14.0
10.0
4.0
0.0
1


FIRST_DUE (IN YEARS)
1200.0
1000.0
800.0
600.0
400.0
200.0
0.0
1000.7
42.5
2.0
6.0
39.0
39.0
1



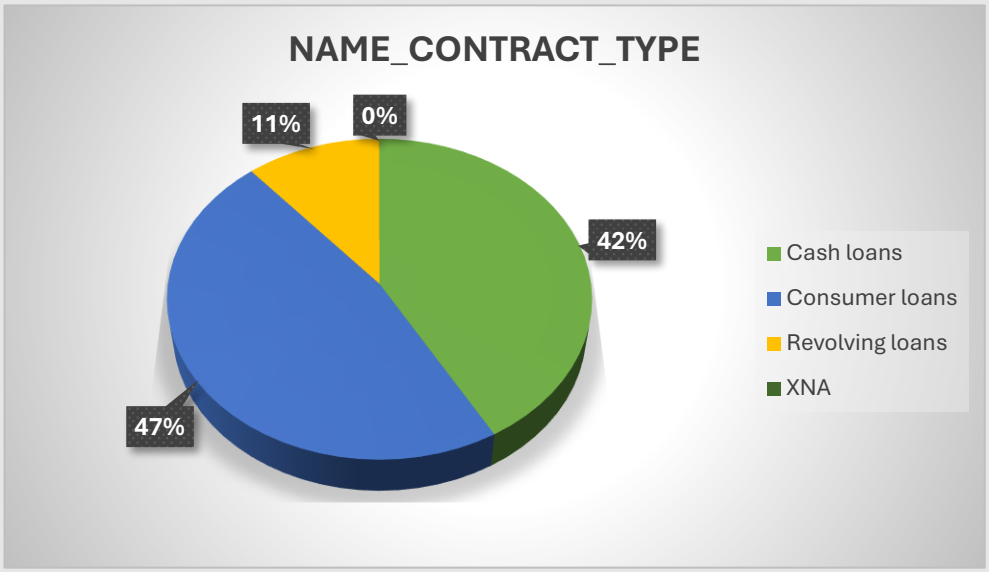
Observations:

- Many columns related to **loan amount and payment duration** (e.g., AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, CNT_PAYMENT) show a significant number of outliers, indicating large variation in loan sizes and repayment terms.
- Time-based fields such as LAST_DUE, TERMINATION, and LAST_DUE_1ST_VERSION contain **extreme and possibly unrealistic values**, including very high numbers.
- CNT_PAYMENT has the **highest number of outliers (14,000+)**, suggesting anomalies in instalment plans or special loan types.

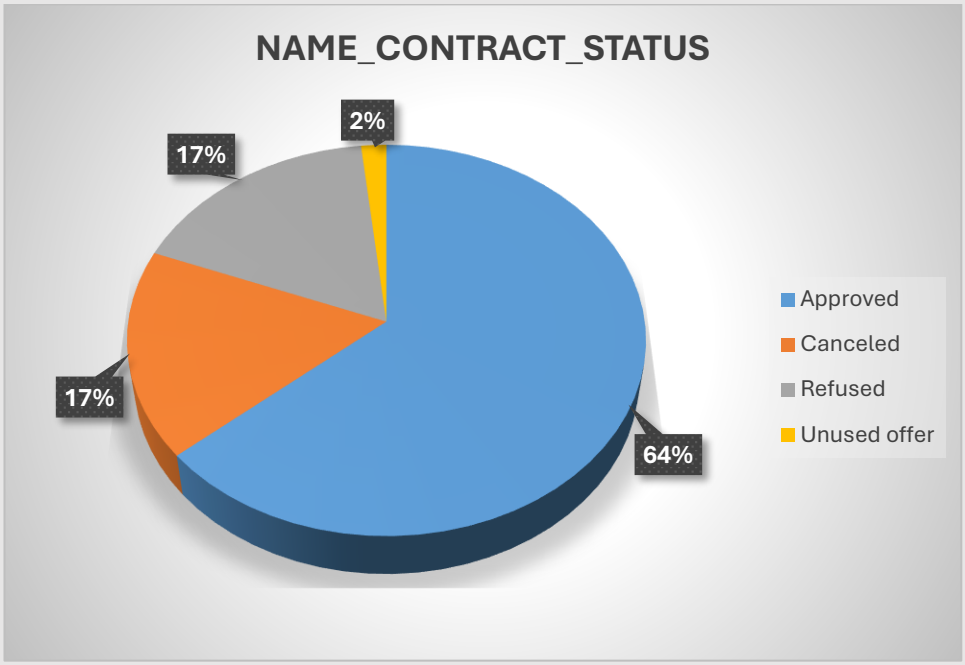
C. Analyze Data Imbalance

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

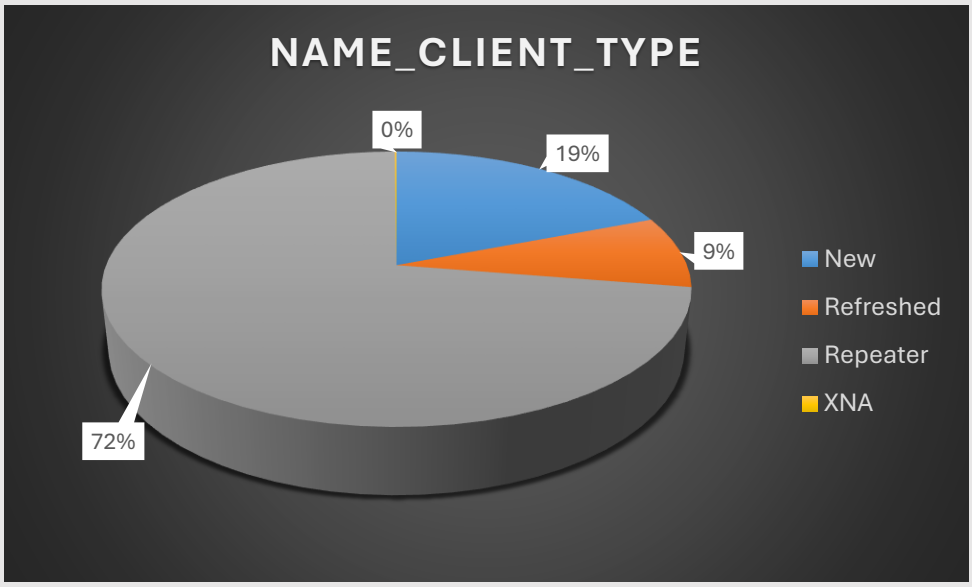
NAME_CONTRACT_TYPE	COUNT	PERCENTAGE
Cash loans	20856	41.71%
Consumer loans	23510	47.02%
Revolving loans	5625	11.25%
XNA	8	0.02%



NAME_CONTRACT_STATUS	COUNT	PERCENTAGE
Approved	31885	63.77%
Cancelled	8595	17.19%
Refused	8660	17.32%
Unused offer	859	1.72%
Grand Total	**49999**	**100.00%**

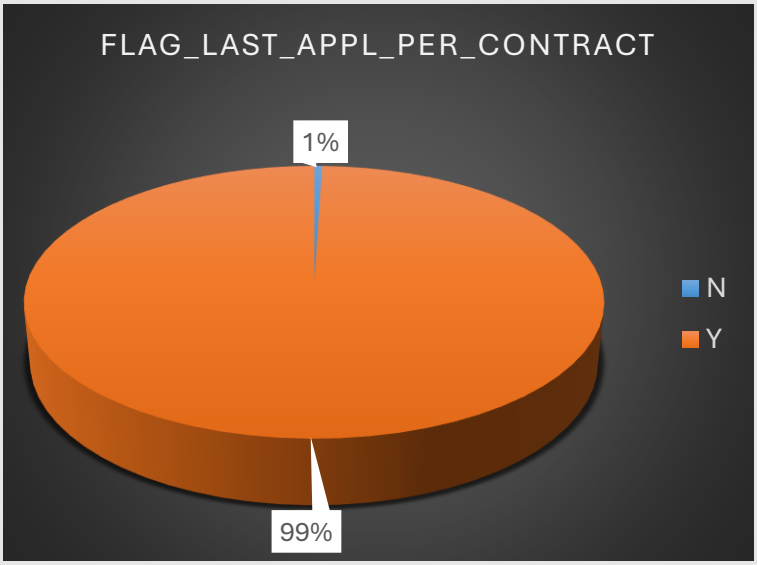


NAME_CLIENT_TYPE	COUNT	PERCETAGE
New	9548	19.10%
Refreshed	4227	8.45%
Repeater	36167	72.34%
XNA	57	0.11%
Grand Total	**49999**	**100.00%**



FLAG_LAST_APPL_PER_CONTRACT	COUNT
N	252
Y	49747
Grand Total	**49999**

FLAG_LAST_APPL_PER_CONTRACT	% SHARE
N	0.50%
Y	99.50%
Grand Total	**100.00%**



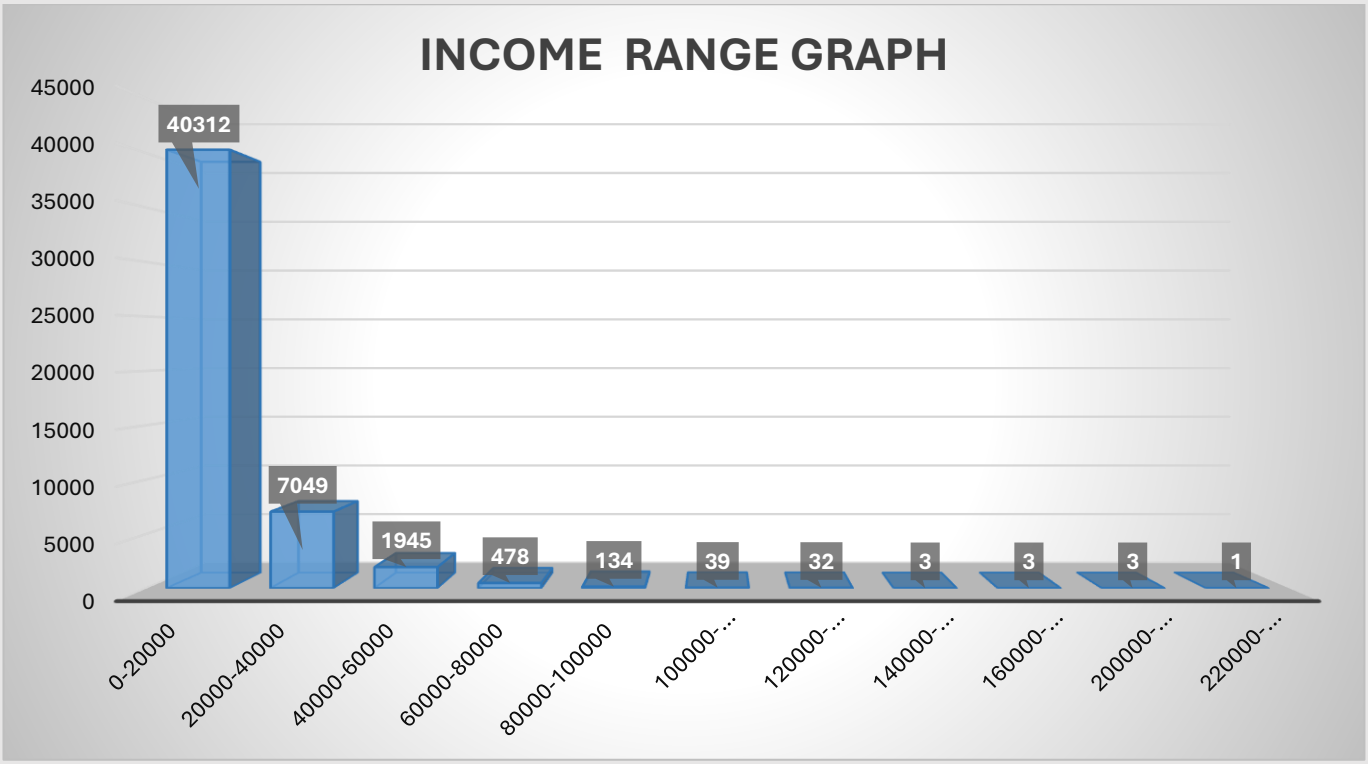
Observations:

- Consumer and Cash loans dominate, while Revolving loans are significantly fewer. The presence of 8 records with an unknown type (XNA) may indicate data quality issues.
- The majority of the contracts are Approved, while Cancelled and Refused have a balanced distribution. Unused offers are rare.
- Repeater clients are dominant, with a small proportion of new and refreshed clients.

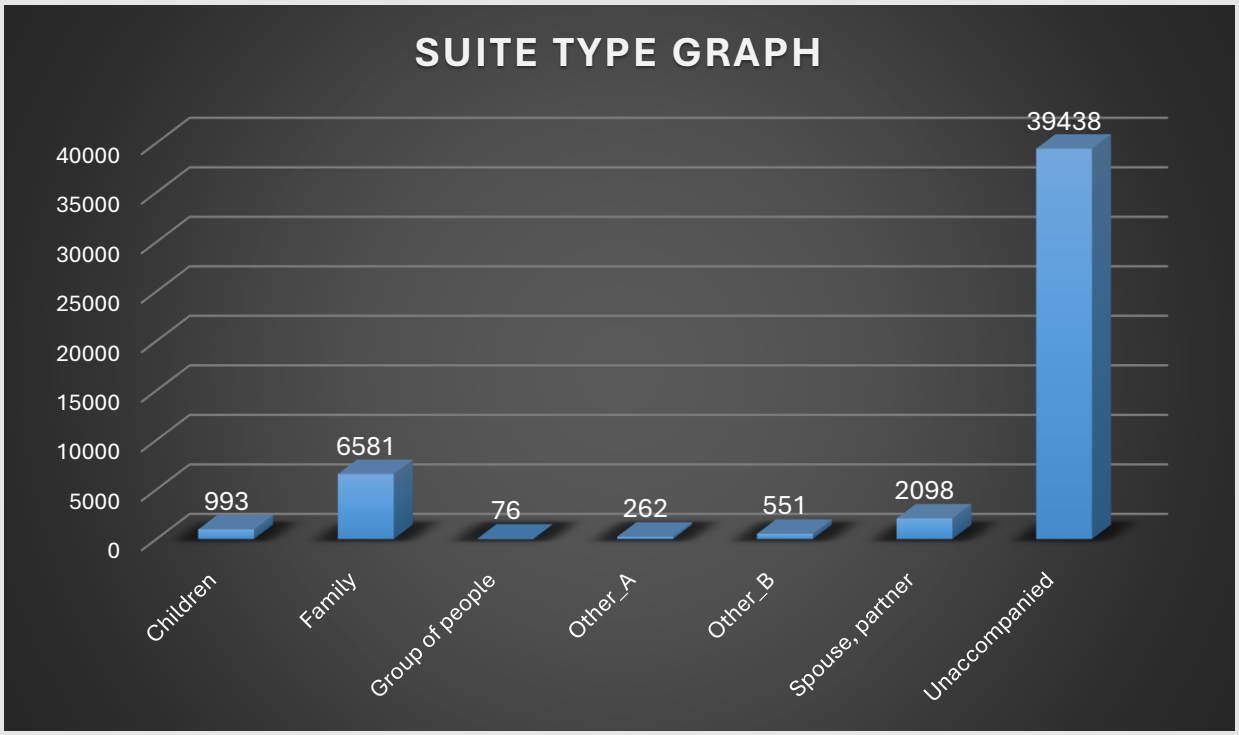
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

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.

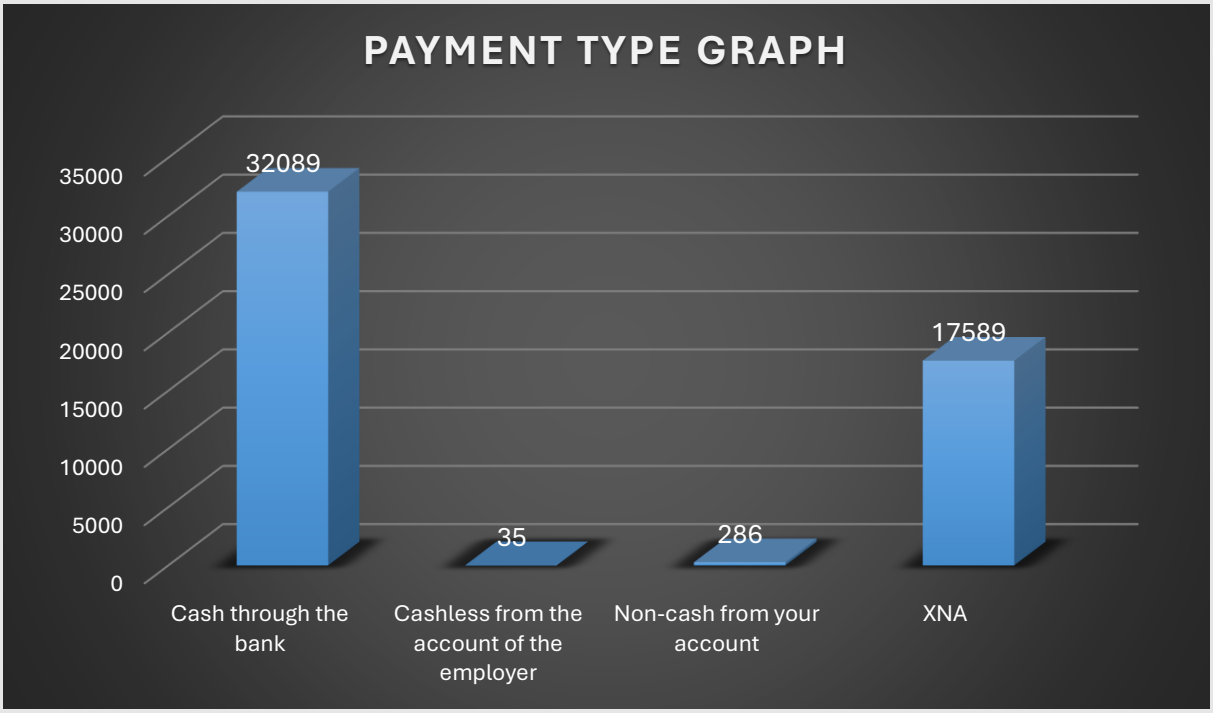
INCOME RANGE	COUNT
0-20000	40312
20000-40000	7049
40000-60000	1945
60000-80000	478
80000-100000	134
100000-120000	39
120000-140000	32
140000-160000	3
160000-180000	3
200000-220000	3
220000-240000	1



SUITE TYPE	COUNT
Children	993
Family	6581
Group of people	76
Other_A	262
Other_B	551
Spouse, partner	2098
Unaccompanied	39438



PAYMENT TYPE	COUNT
Cash through the bank	32089
Cashless from the account of the employer	35
Non-cash from your account	286
XNA	17589



Observations:

- The majority of values (over 80%) fall in the **0–20,000** income range, indicating a heavily left-skewed distribution.
- There's a steep decline in count from **0–20,000 (40,312) to 20,000–40,000 (7,049),** suggesting income is concentrated in the lower range.
- "**Unaccompanied**" is the dominant category. Very few entries belong to rare categories like "**Group of people**" or "**Other_A, Other_B**".
- A significant portion of entries (over 35%) have "**XNA**".
- "**Cash through the bank**" is the most common valid payment method.
- Other payment types are extremely rare.

E. Identify Top Correlations for Different Scenarios

Task: Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.

Top 3 Correlations	
AMT_CREDIT & AMT_APPLICATION	0.975771049
AMT_GOODS_PRICE & AMT_APPLICATION	0.949626505
AMT_GOODS_PRICE & AMT_CREDIT	0.942491794

Observations:

- Correlation between AMT_APPLICATION and AMT_CREDIT indicates that the amount applied for is almost equal to or very close to the amount credited. This suggests a high approval rate or minimal adjustment between requested and sanctioned amounts.
- Strong correlation between AMT_APPLICATION and AMT_GOODS_PRICE suggests that the amount applied for is closely tied to the price of goods being purchased.
- Strong correlation between AMT_CREDIT and AMT_GOODS_PRICE shows that loans are granted primarily to finance the cost of goods, rather than for liquidity or mixed purposes.

Insights

- SK_ID_CURR and SK_ID_PREV support reliable data merging.
- Financial fields with missing values (e.g., AMT_ANNUITY) are suitable for mean/mode imputation.
- Dropping highly missing columns (e.g., RATE_INTEREST_PRIMARY) improves model quality.
- Outliers in loan amounts (AMT_CREDIT, CNT_PAYMENT) need treatment to avoid bias.
- Unrealistic date values (LAST_DUE, TERMINATION) require correction.
- Loan types are imbalanced; Revolving loans and XNA entries are rare.
- "Approved" dominates contract status; others support multi-class modelling.
- Repeat borrowers dominate and need careful validation splitting.
- Income is left-skewed, with most clients earning under 20,000.
- Rare categories (e.g., NAME_TYPE_SUITE) should be grouped or removed.
- Loan amount features are strongly correlated, reflecting consistent approval behaviour.

Results

- Through this project, I effectively cleaned and pre-processed the loan dataset, ensuring data quality by addressing missing values, removing noisy columns, and handling outliers.
- I gained a deep understanding of the dataset's structure, including demographic and financial factors affecting loan defaults.
- The analysis revealed key insights, such as the high-risk characteristics of defaulters (low-income, single, and male clients) and the strong correlation between loan amounts and payment behaviour.
- These findings not only enhanced my data analysis skills but also improved my ability to identify critical risk factors in financial datasets, providing a solid foundation for building predictive models.

Google Drive Links-

Cleaned datasets (it is advised to view this dataset in **Excel**)

Application_data.xlsx | Link

previous_application.xlsx | Link