Bank Loan Case Study

BY

SHNGAIN KUPAR SHULLAI

07052024

Brief Overview of the Project

Approach

This work I am doing on improving how the bank decides who gets a loan. I'm looking closely at the data I have to find any patterns or issues that might affect Bank's decisions. I use different techniques to look at the data, like charts and graphs, to spot any missing information or unusual numbers. When I find missing data, I decide how to fill it in carefully, depending on what makes sense for the situation. I also look for any outliers, which are numbers that stand out from the rest, and make sure they won't affect our decisions unfairly.

I don't just look at the numbers, though. I also examine each piece of information on its own to see if there's anything strange about it. I divide the data into groups based on categories, like age or income, to see if there are any important differences between them.

And I don't stop there - I also look for connections between different pieces of information. By doing this, I hope to help the bank make better decisions about who should get a loan. It's important to make sure everyone who can handle a loan gets a fair chance, while still protecting the bank from people who might not be able to pay it back.

Through the lens of Exploratory Data Analysis (EDA), I'm delving deep into our dataset, armed with techniques to uncover hidden patterns and ensure fairness in the decision-making.

Using Quartiles functions, box plots, histograms, and other tools, I detect these anomalies and deal with them appropriately, always mindful of the context in which they appear. And as I navigate through the data, I keep a keen eye on balance, ensuring that the proportions of defaulted loans versus non-defaulted ones remain in check.

Segmentation becomes key of the data based on categorical variables, unveiling insights that might otherwise stay hidden.

Scatter plots, correlation matrices, and heatmaps become my companions as I seek out connections between variables, honing in on those that hold the most predictive power.

By conducting a thorough EDA, addressing missing data, outliers, and imbalances, I can refine loans approval process. It's about more than just minimizing risk; it's about ensuring that every capable applicant gets a fair shot, while safeguarding against defaults from those who might lack the necessary credit history.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

TECH-STACK USED

- I use MS Word 2021, MS Excel 2021, LightShot, , Google Drive and Notepad.
- I am deeply passionate about honing my skills with Excel and other related software tools.
- My goal is to steadily progress and refine my abilities until I reach a professional level of proficiency.

Data Analytics Tasks:

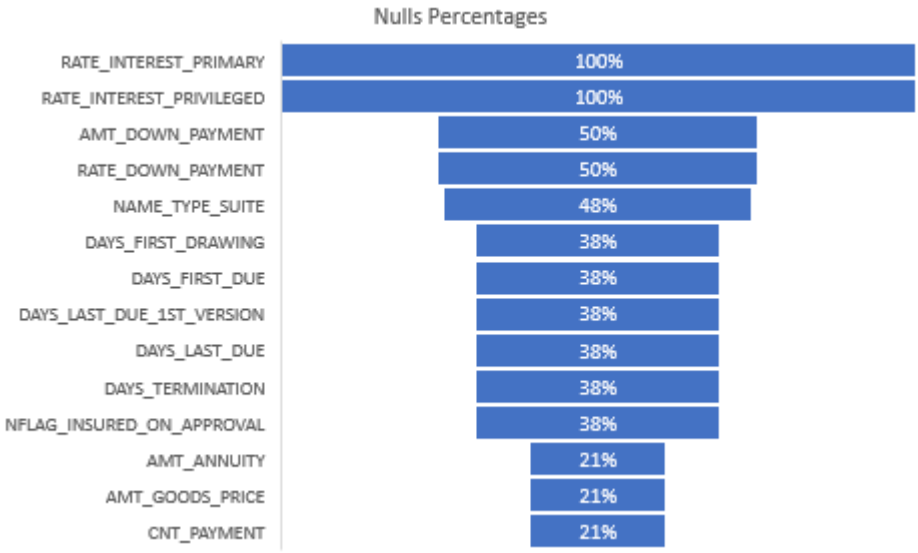
A. Identify Missing Data and Deal with it Appropriately:

***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.

Hint: Utilize Excel functions like COUNT, ISBLANK, and IF to identify missing data. Consider using functions like AVERAGE or MEDIAN for imputation or other appropriate methods available in Excel.

Graph suggestion: Create a bar chart or column chart to visualize the proportion of missing values for each variable.

Files: a. previous_application.csv:



Click Here to view the M S Excel file of this Case Study file1

[Click Here to view the M S Excel file of this Case Study file2]

1. I removed certain columns from the dataset because they either had more than 40% missing values or weren't relevant to my analysis. These columns are: **AMT_DOWN_PAYMENT, WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START, FLAG_LAST_APPL_PER_CONTRACT, NFLAG_LAST_APPL_IN_DAY, RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED, NAME_TYPE_SUITE, and PRODUCT_COMBINATION.**

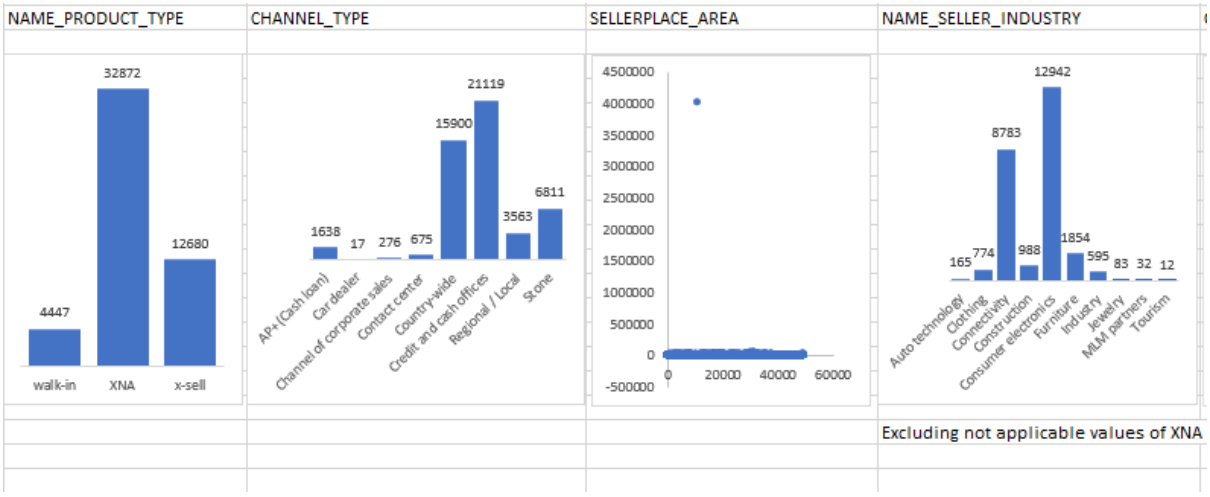
2. I've applied several features and functions from Excel to meet the requirements for all columns:

   1st Quartile: This is the value below which 25% of the data falls.
   2nd **Quartile**: Also known as the median, it's the midpoint of the data set.
   3rd **Quartile**: This marks the value below which 75% of the data falls.
   IQR (**Interquartile** Range): The range between the 1st and 3rd quartiles, indicating the spread of the middle 50% of the data.
   Lower ***Outlier*** Threshold: The lowest value considered not an outlier.
   Upper ***Outlier*** Threshold: The highest value considered not an outlier.
   Outliers: Data points that fall outside the lower and upper outlier thresholds.
   Max: The highest value in the dataset.
   Median function: Returns of the median, or the middle value, of a dataset.
   Average (mode) function: Calculates the average, or mode, of the dataset.

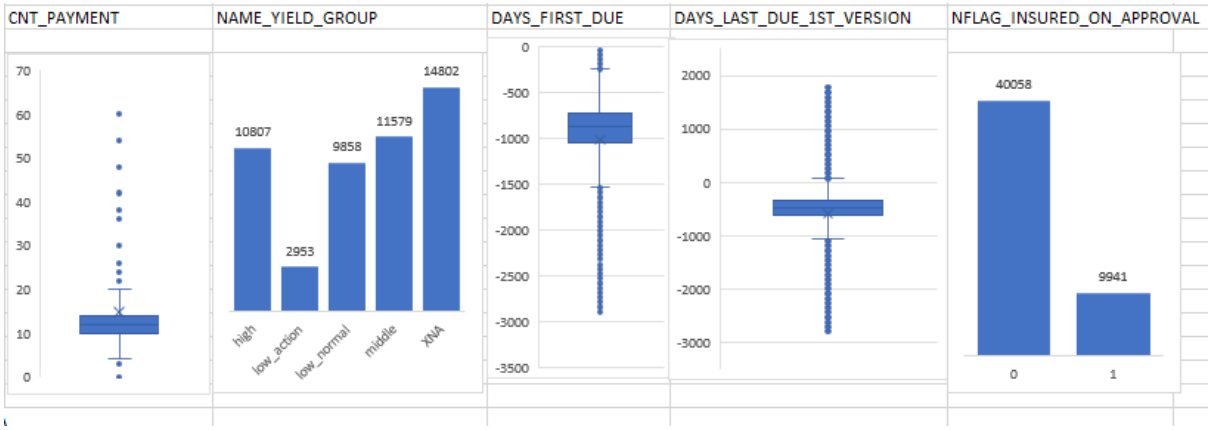
3. I filled in blank cells in the columns for **"Amt annuity," "Amt goods price," and "CNT_PAYMENT"** by using the median value.

4. I filled in the blank cells in the column "**NFLAG_INSURED_ON_APPROVAL**" by using the mode, which is the most frequently occurring value in that column.

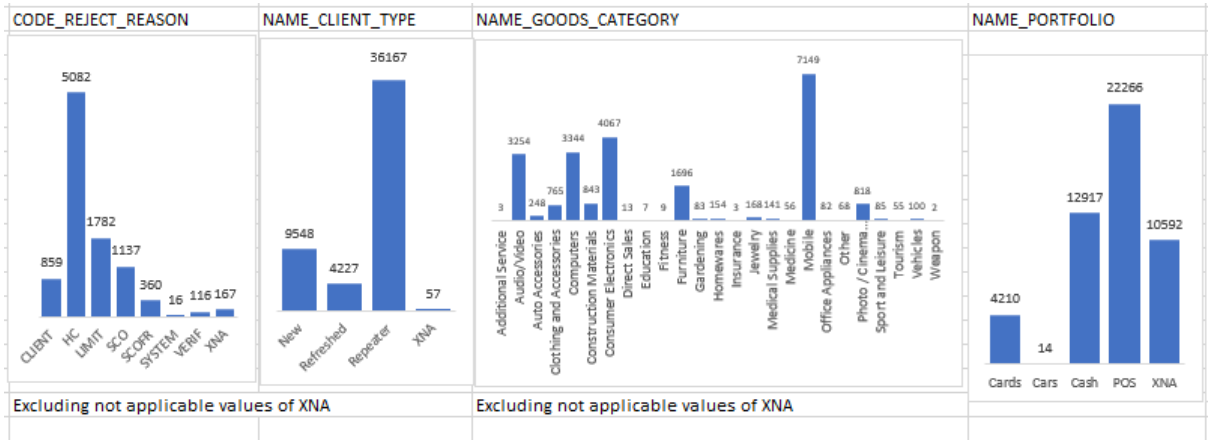
5. I replaced the outlier cells in the columns "**DAYS_DECISION," "Amt annuity," "Amt Application," "Amt credit," and "Amt goods price**" with the median values of each respective column.



6. I removed the columns **"DAYS_FIRST_DRAWING," "DAYS_LAST_DUE," and "DAYS_TERMINATION**" from the dataset. Initially, they had 38% null values, and after imputing their outliers with null values, the percentage of null values in these columns exceeded 53%. Hence, I decided to delete them from the dataset.

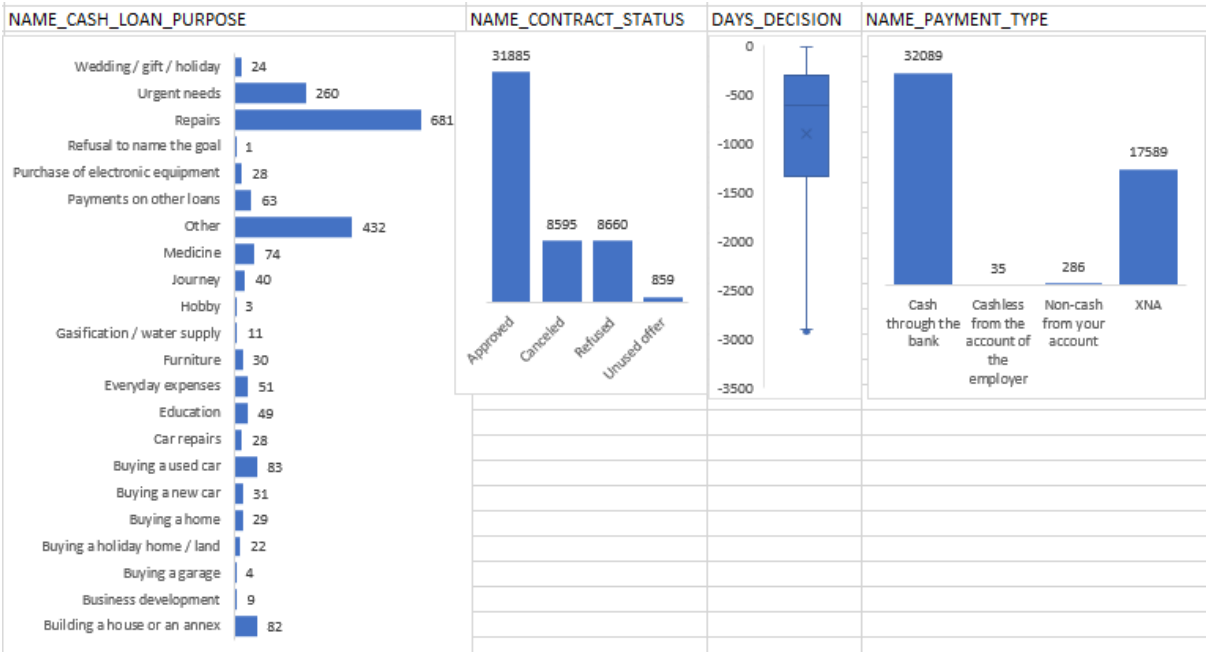


[Click Here to view the M S Excel file of this Case Study]



7. Please refer to the **Excel sheet** for a detailed list of all the outliers that were retained even after the imputation process.

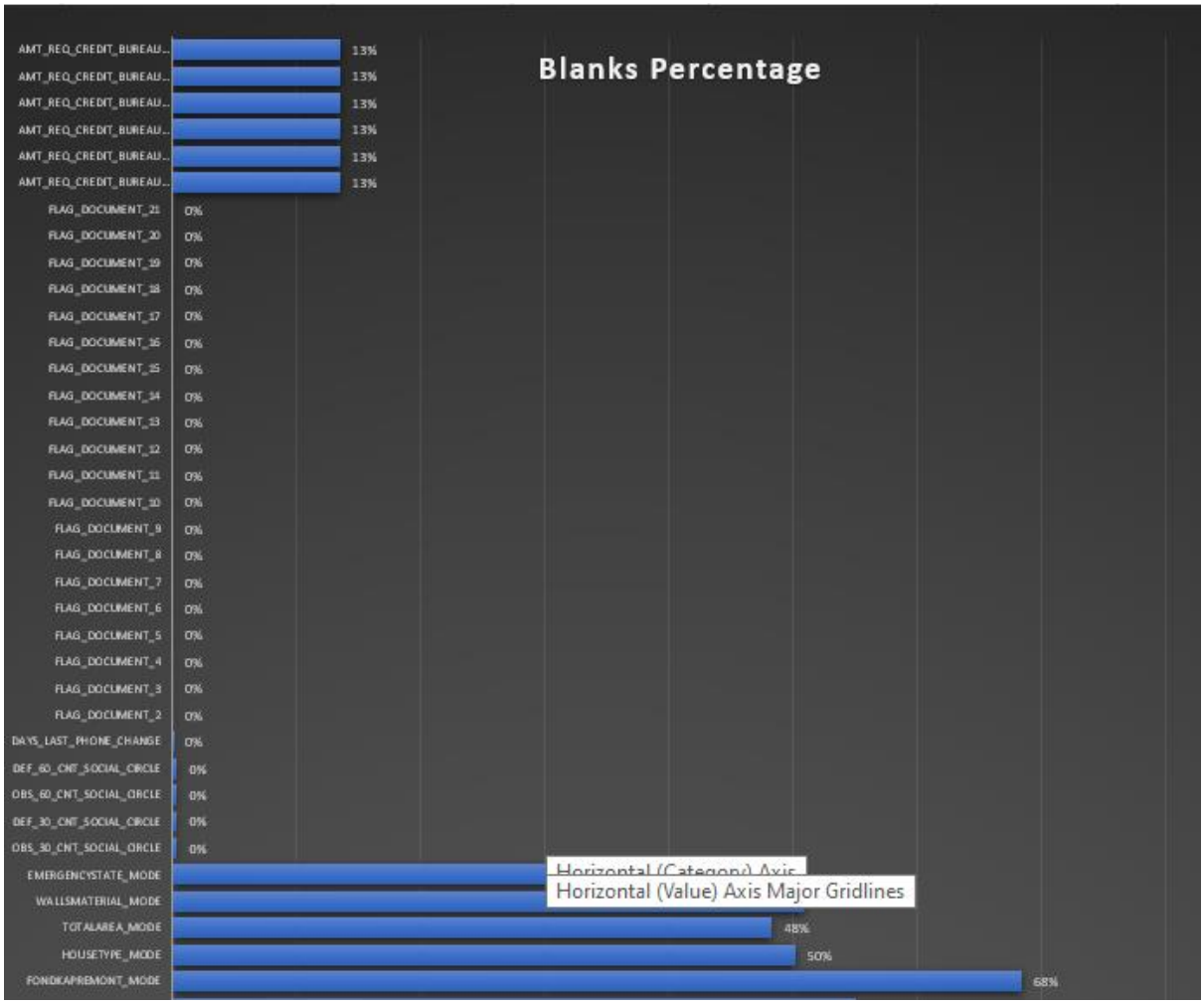
[Click Here to view the M S Excel file of this Case Study file1]
[Click Here to view the M S Excel file of this Case Study file2]



[Click Here to view the M S Excel file of this Case Study file1]

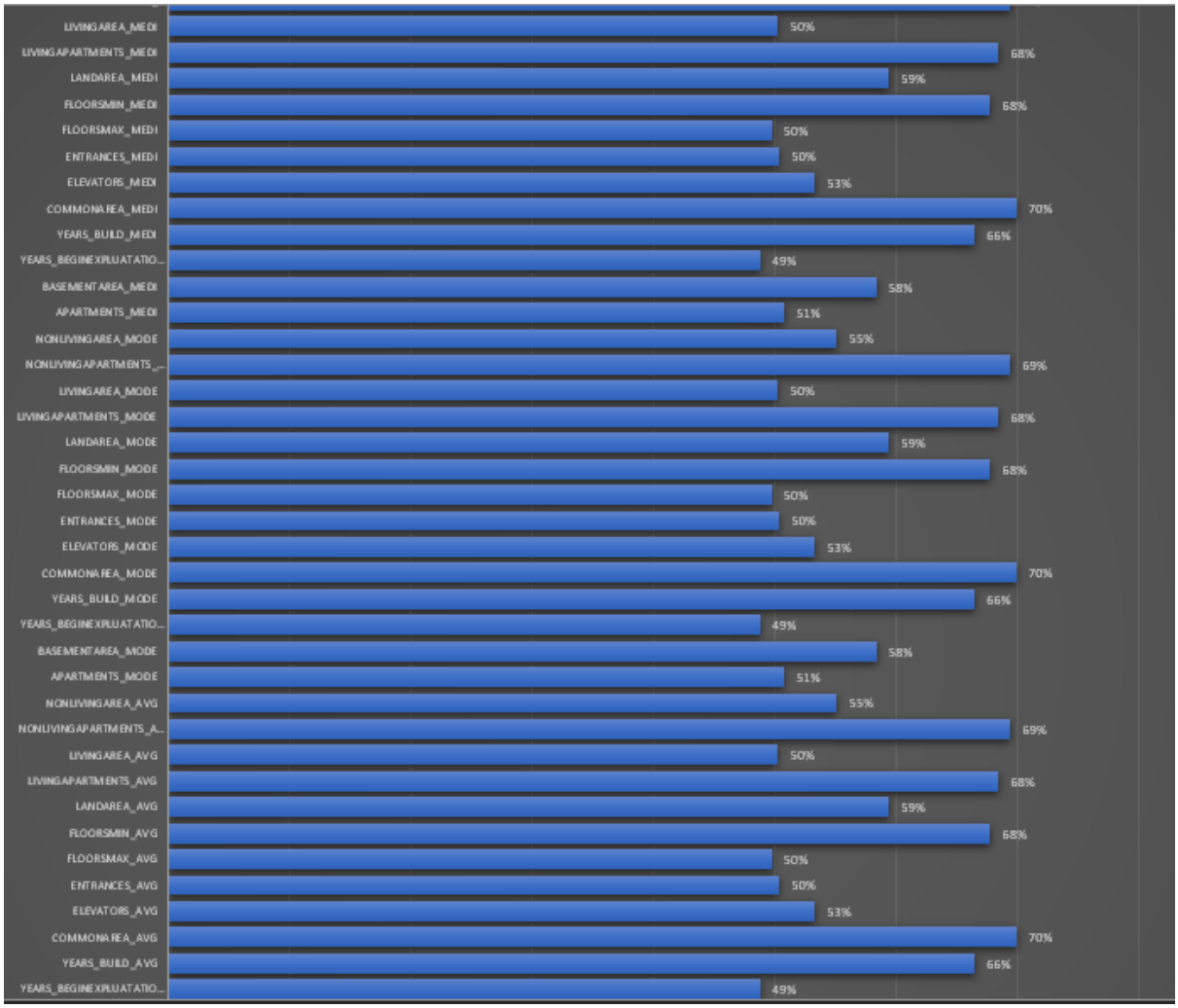
[Click Here to view the M S Excel file of this Case Study file2]

Files: b. app_data.csv:



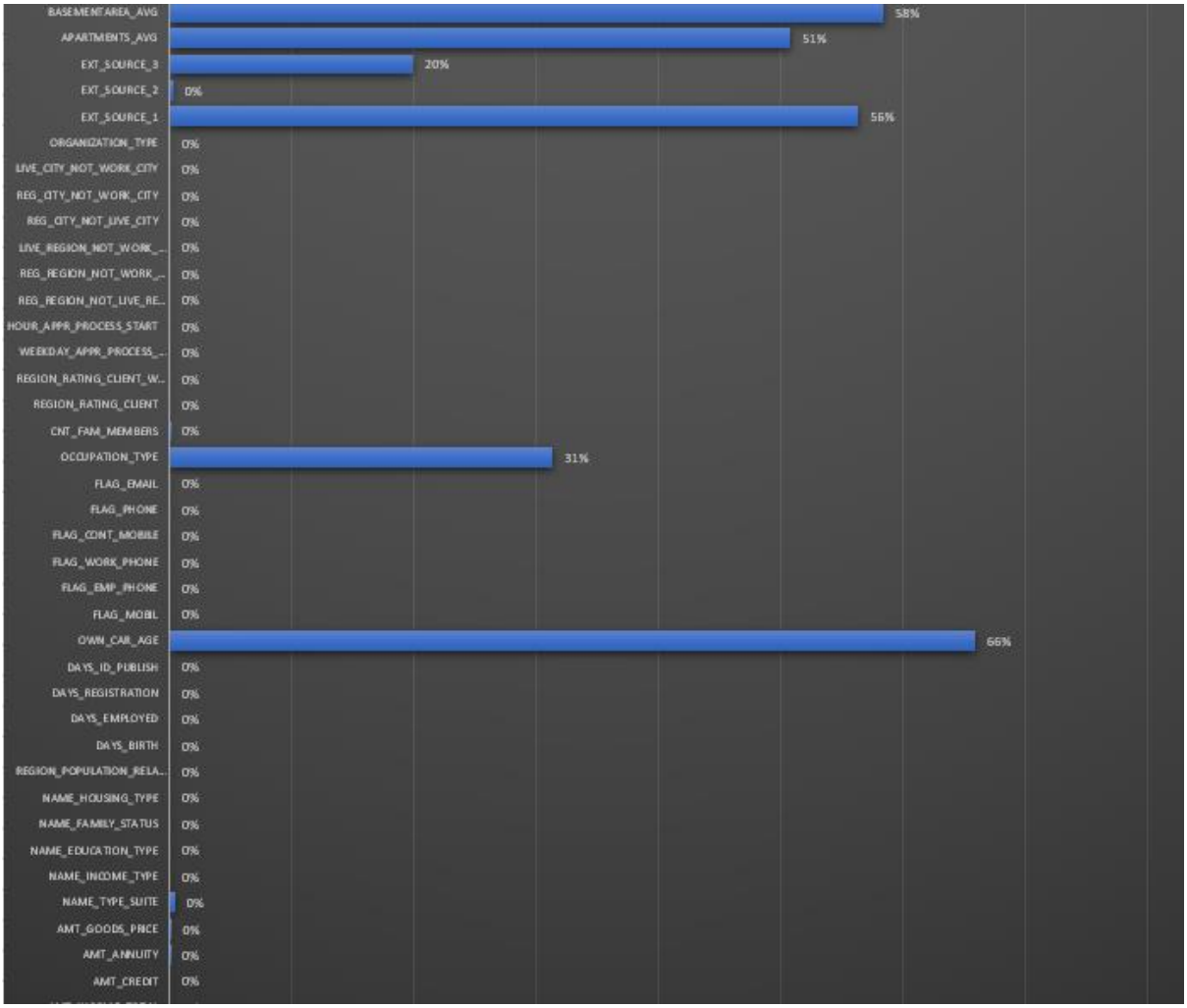
Click Here to view the M S Excel file of this Case Study file1

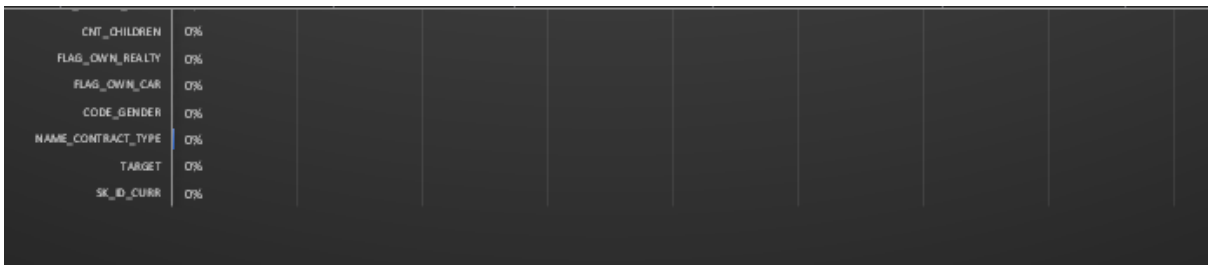
Click Here to view the M S Excel file of this Case Study file2



Click Here to view the M S Excel file of this Case Study file1

Click Here to view the M S Excel file of this Case Study file2





1. Based on the screenshots provided, I observed that many columns in the raw data file contain a significant percentage of **blank or null cells.**

2. In line with best practices, I decided to delete all columns containing more than **35% blank or null values.**

3. Additionally, I identified the **"FLAG_MOBIL"** column as not containing valuable data, so I removed it from the dataset as well.

B. Identify Outliers in the Dataset:

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

Hint: Utilize Excel functions like QUARTILE, IQR, and conditional formatting to identify potential outliers. Consider applying thresholds or business rules to determine if the outliers are valid data points or require further investigation.

Graph suggestion: Create box plots or scatter plots to visualize the distribution of numerical variables and highlight the outliers.

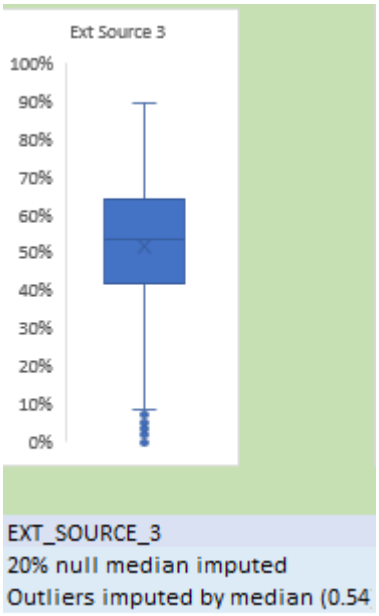
1. I utilized several features and functions from Excel to meet the requirements for all columns:
   1st **Quartile**: The value below which 25% of the data falls.
   2nd **Quartile**: Also known as the median, it's the midpoint of the dataset.
   3rd **Quartile**: The value below which 75% of the data falls.
   IQR (**Interquartile** Range): The range between the 1st and 3rd quartiles, indicating the spread of the middle 50% of the data.
   Lower **Outlier** Threshold: The lowest value considered not an outlier.
   Upper **Outlier** Threshold: The highest value considered not an outlier.
   Outliers: Data points that fall outside the lower and upper outlier thresholds.
   Max: The highest value in the dataset.
   Median function: Returns the median, or the middle value, of a dataset.
   Average (mode) function: Calculates the average, or mode, of the dataset.:

AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR	
0	0	0	1	1st Quartile
0	0	0	2	2nd Quartile
0	0	0	3	3rd Quartile
0	0	0	2	IQR
0	0	0	-2	<<< Lower Outlier Threshold
0	0	0	6	<<< Upper Outlier Threshold
0	0	0	0	Lower Outliers
1313	7137	8128	1169	Upper Outliers
1313	7137	8128	1169	<<< Total Outliers
6	24	8	25	<<< Max

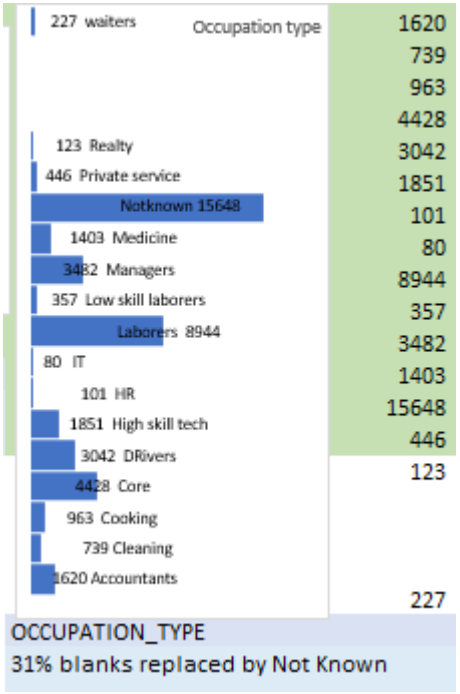
2. For the "**EXT_SOURCE_3**" column, which had **20%** null values, I filled in those missing values using the **median**.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

Additionally, I replaced any **outliers** in this column with the **median** value, which was 0.54.



3. In the "**OCCUPATION**_TYPE" column, where **31**% of the values were **null**, I replaced these missing values with the term "**Not Known**."

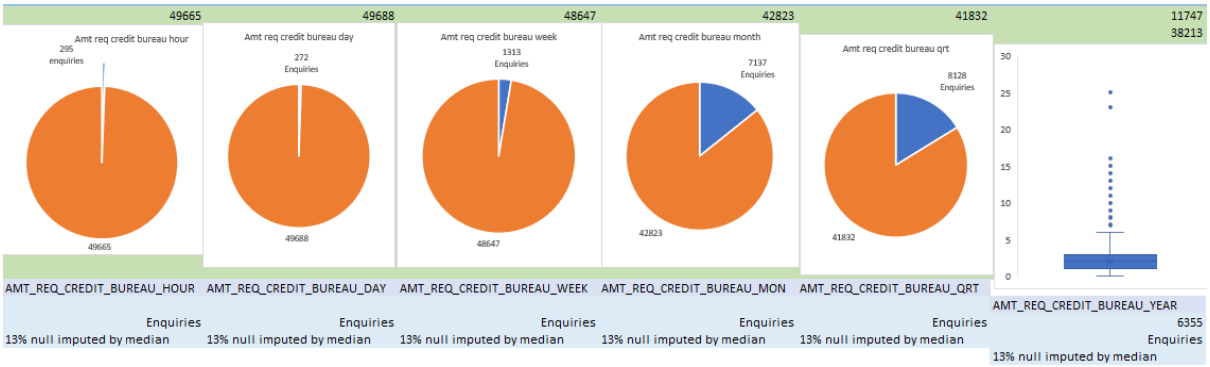


4. The columns **"AMT_REQ_CREDIT_BUREAU_HOUR," "AMT_REQ_CREDIT_BUREAU_DAY," "AMT_REQ_CREDIT_BUREAU_WEEK," "AMT_REQ_CREDIT_BUREAU_MON," "AMT_REQ_CREDIT_BUREAU_QRT," and "AMT_REQ_CREDIT_BUREAU_YEAR"** each had **13% null** values. I addressed these missing values by imputing them with the **median** value of each respective column.

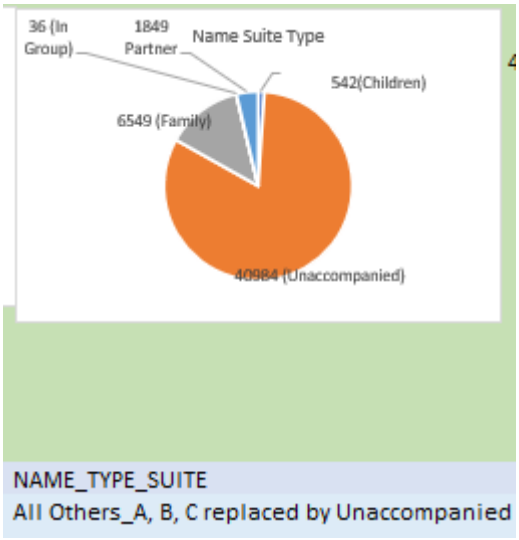
Click Here to view the M S Excel file of this Case Study file1

Click Here to view the M S Excel file of this Case Study file2

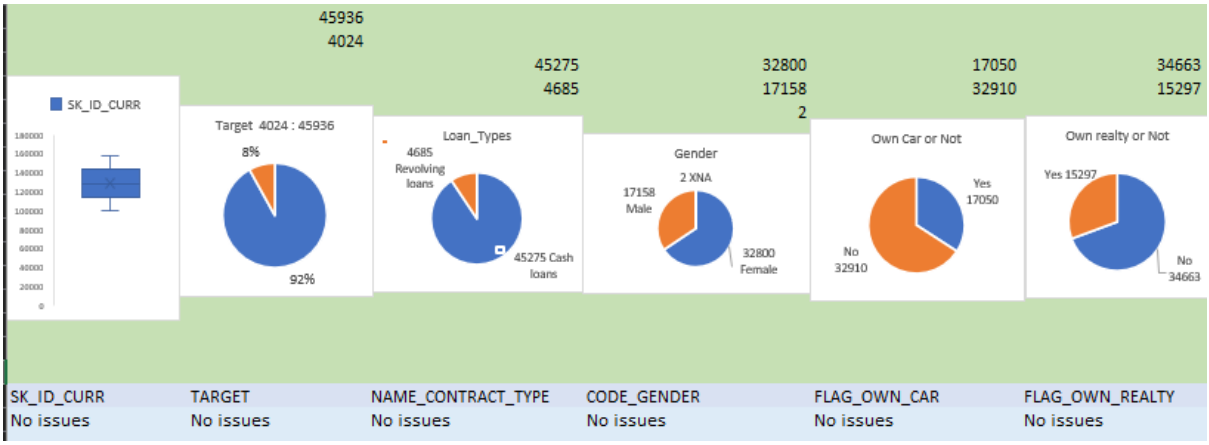
5.

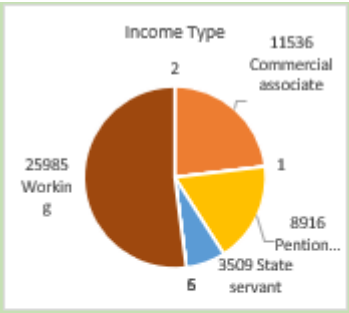


6. In the "**NAME_TYPE_SUITE**" column, where entries were listed as **Others_A, Others_B, and Others_C**, I replaced these with the most common entry in the column, which is "**Unaccompanied**"



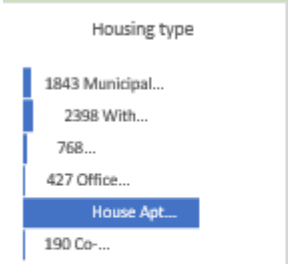
7. Columns with **fewer** issues



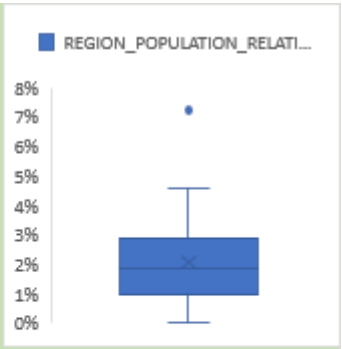


2
11536
1
8916
3509
5
6
25985

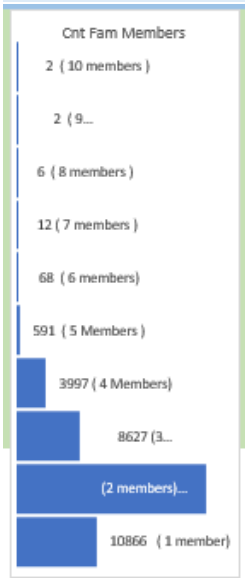
NAME_INCOME_TYPE
No issues



190
44334
427
768
2398
1843



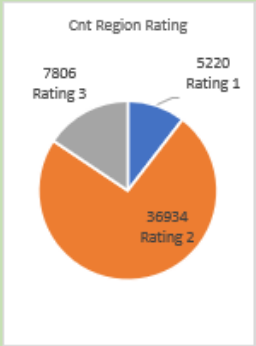
NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE
No issues	No issues

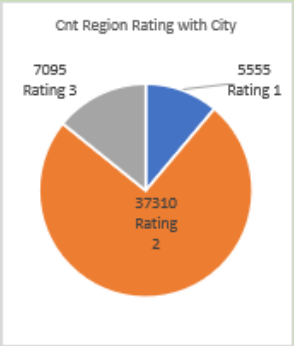


10866
25788
8627
3997
591
68
12
6
2
2
1

5220
36934
7806

5555
37310
7095



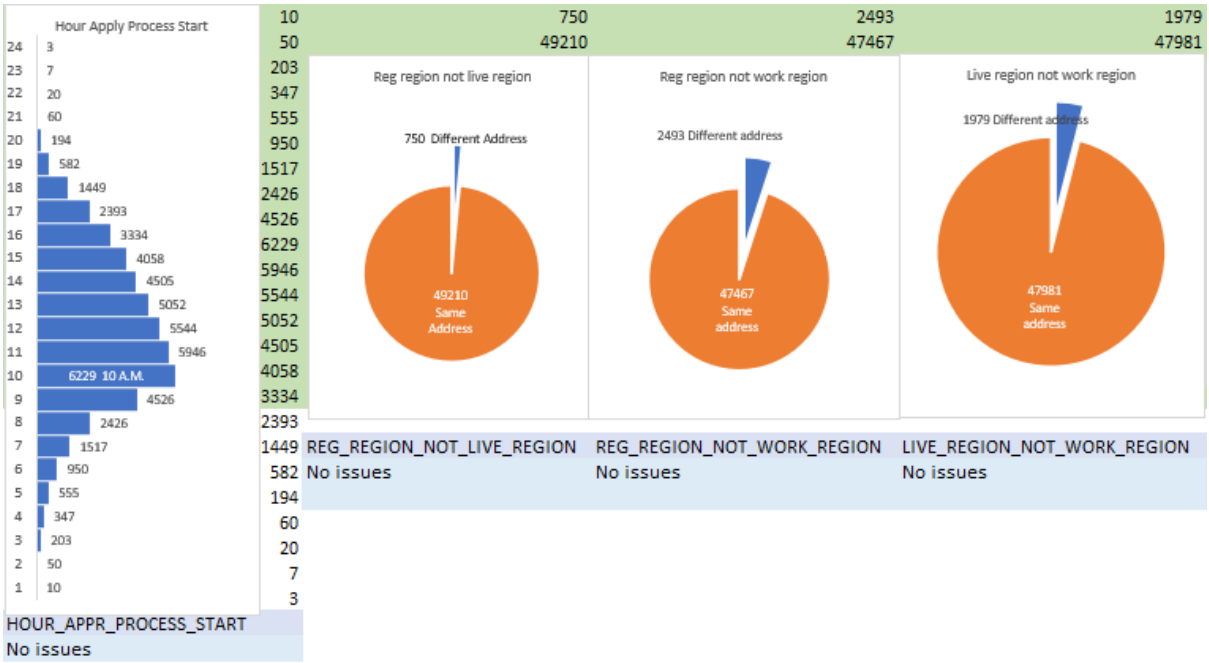


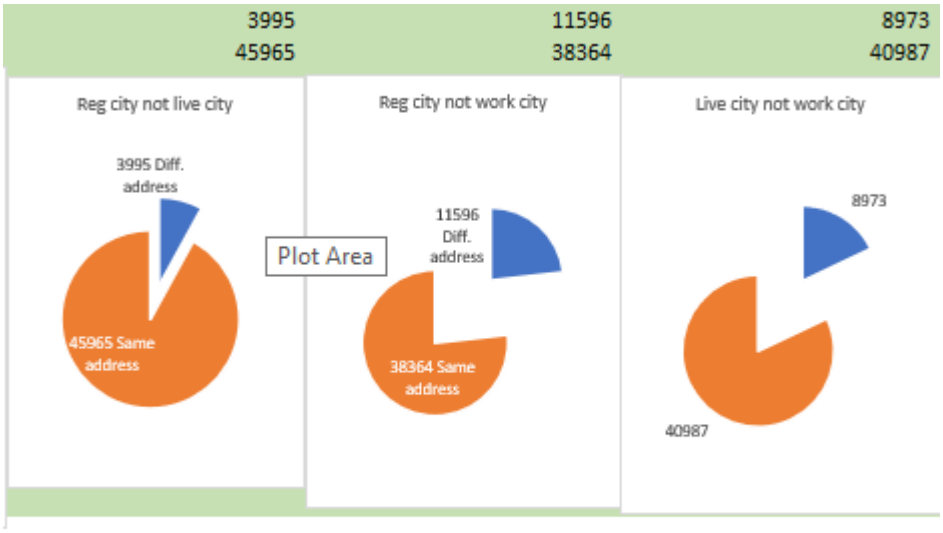
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	WEEKDAY_APPR_PROCESS_START
No issues	No issues	No issues

CNT_FAM_MEMBERS
No issues

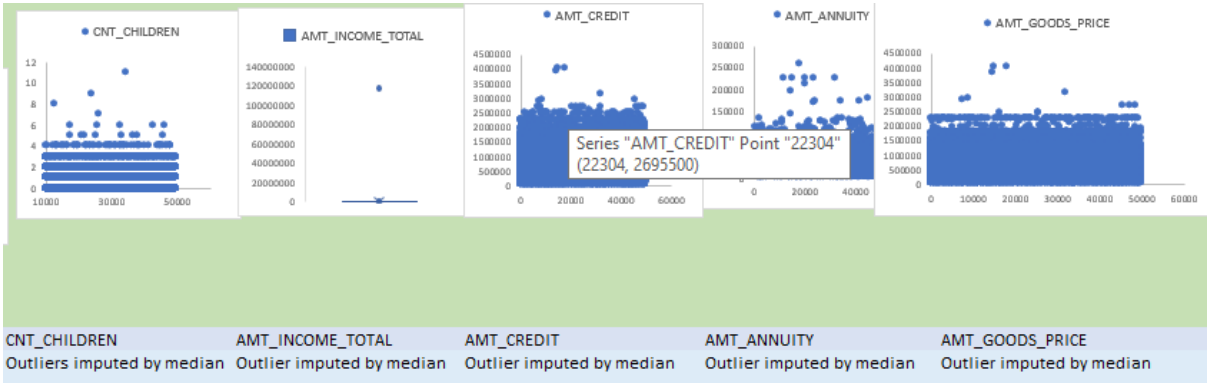
[Click Here to view the M S Excel file of this Case Study file1]

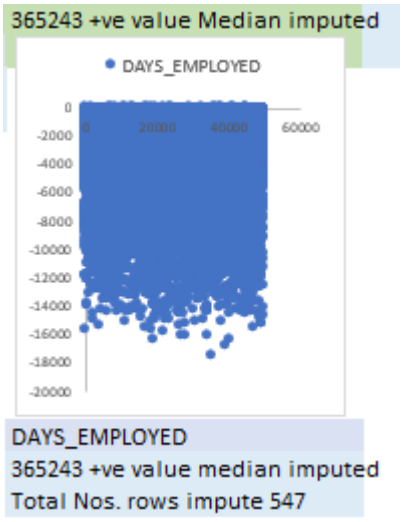
[Click Here to view the M S Excel file of this Case Study file2]



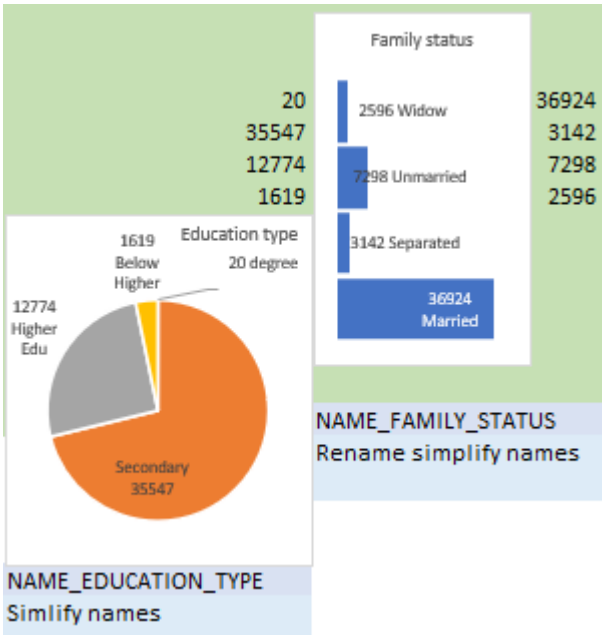


8. **Outliers** were addressed by imputing them with the **median** value.

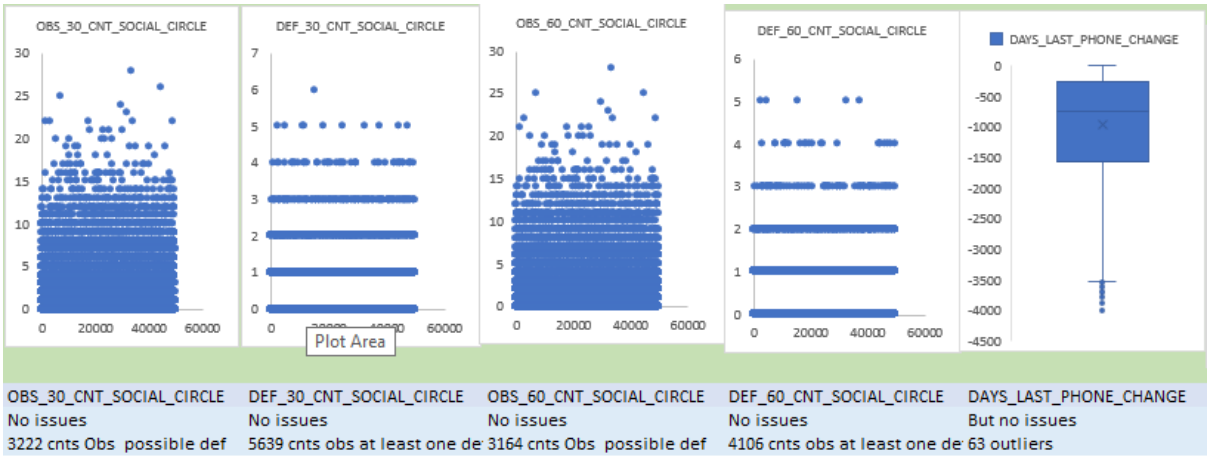




9. **Simplify** names

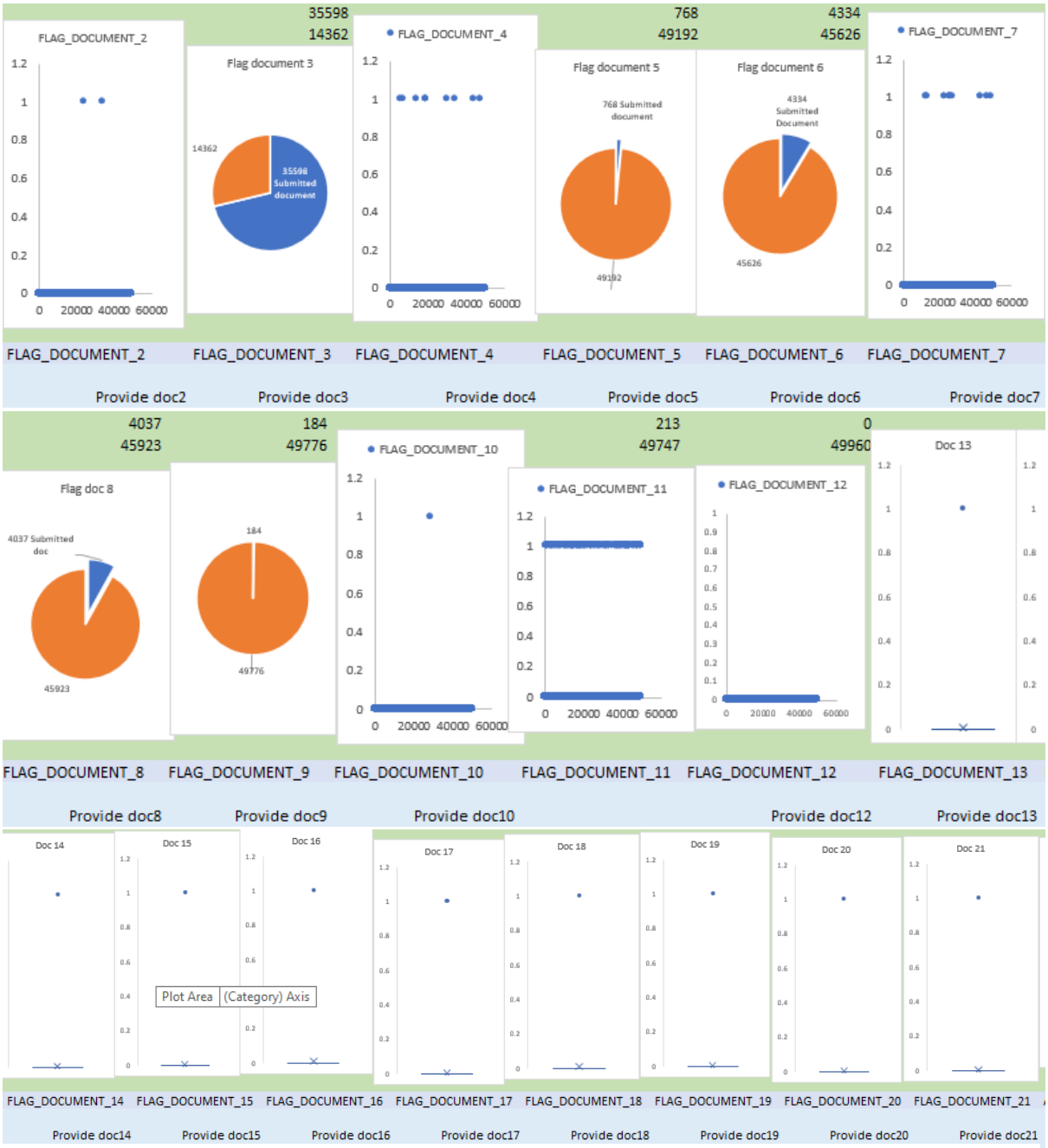


Outliers were left unchanged as they did **not** pose any issues.



Click Here to view the M S Excel file of this Case Study file1

Click Here to view the M S Excel file of this Case Study file2



Conclusion

Here's how I implemented steps in Excel:

I calculated Quartiles and IQR by using the QUARTILE function to calculate Q1, Q2, and Q3 for each numerical variable.

Subtract Q1 from Q3 to calculate the IQR.

Then determine Outlier Thresholds and Multiply the IQR by 1.5 to determine the outlier thresholds.

I visualize these Outliers with Box Plots or Scatter Plots.

By following these steps, I effectively identify outliers in the loan application dataset and visualize their impact on the analysis using Excel's statistical functions and features.

C. Analyze Data Imbalance:

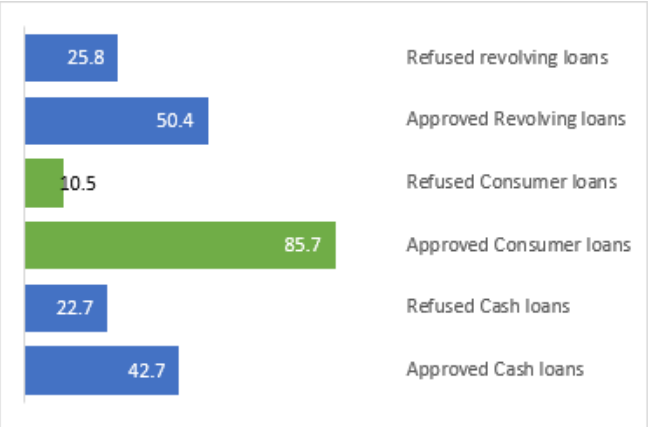
Task: Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
Hint: Utilize Excel functions like COUNTIF and SUM to calculate the proportions of each class. Compare the class frequencies to assess data imbalance.
Graph suggestion: Create a pie chart or bar chart to visualize the distribution of the target variable and highlight the class imbalance.

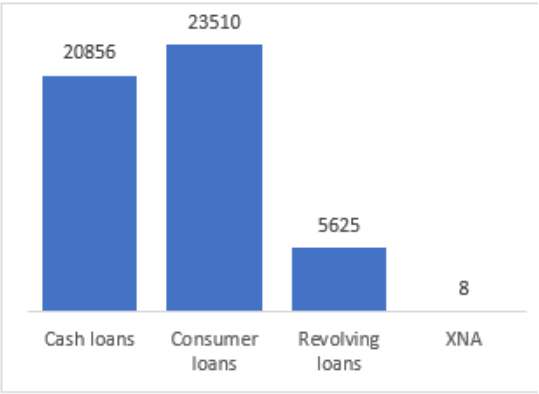
Analyzing data imbalance is crucial for binary classification problems, especially when building reliable models.

(i) **NAME_CONTRACT_STATUS**: The approval rates for consumer loans are significantly higher than other loan types, indicating a higher likelihood of approval for this category.

Cash loans also have a relatively high approval rate, making them another favorable option **Revolving loans** have moderate approval rates, suggesting they may be slightly less referred than cash and consumer loans.

This **imbalance** in approval rates could be indicative of varying risk profiles associated with different loan types.

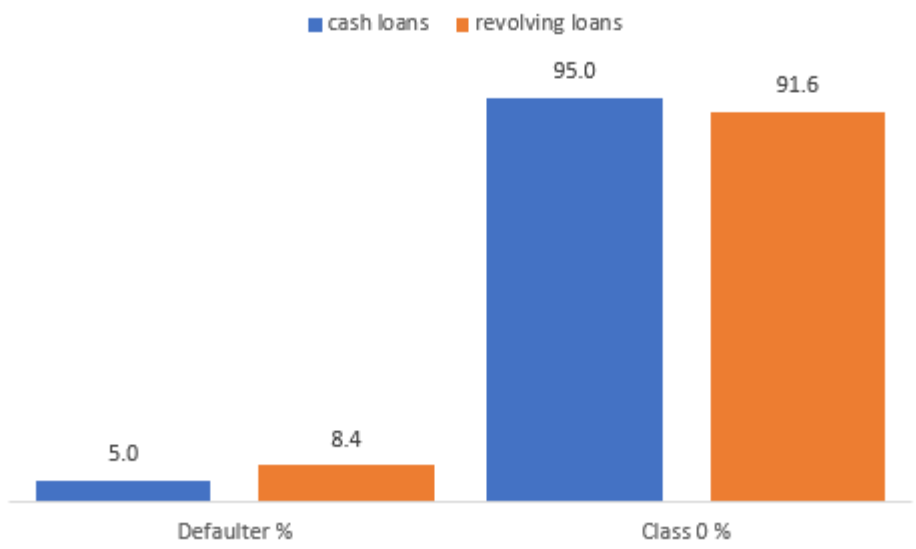




Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

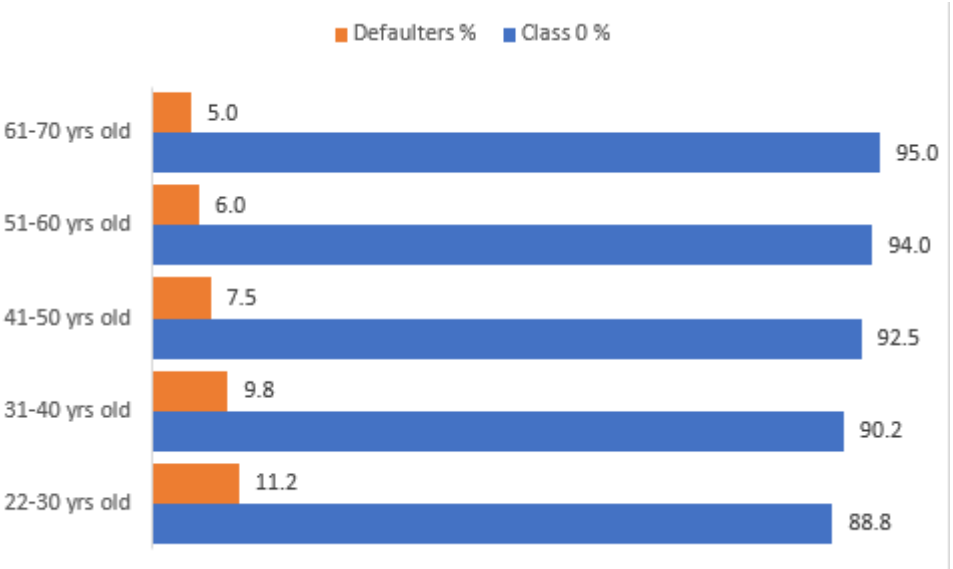
(ii) **NAME_CONTRACT_TYPE:** This column specifies the type of contract, which could be indicative of different risk levels associated with various loan types.

Analyzing the distribution of contract types across the target variable revealed that **Revolving loans** are more prone to **default**.



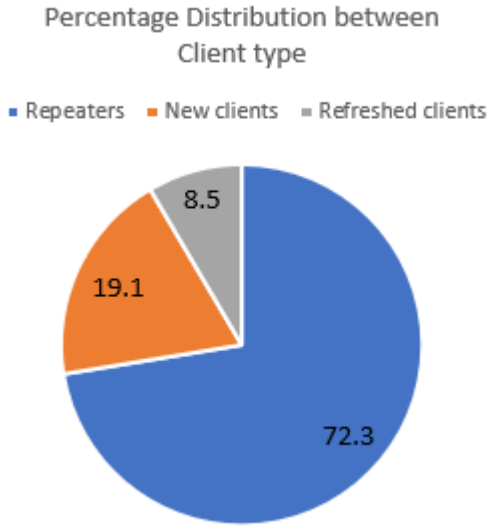
(iii) CODE_GENDER: Gender could potentially influence the likelihood of default.

Analyzing the distribution of **genders** across the **target** variable can help identify any gender-based **imbalances** in **default** rates.
I suggest that as **age increases**, the proportion of **defaulters decreases**, indicating a potential **correlation between older age groups and lower default rates**.



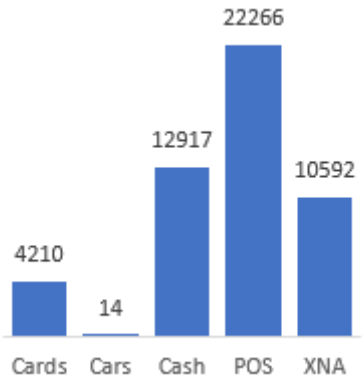
(iv) NAME_CLIENT_TYPE: This column specifies whether the client was a new or existing client when applying for the previous application.

Analyzing the distribution of client types can highlight any differences in default rates between **new and existing clients**.
The predominant presence of **repeat clients**, accounting for 72.3% of the distribution, indicates a substantial level of customer loyalty or satisfaction,
potentially offering a competitive advantage in terms of customer retention compared to **new or refreshed clients**.



(v) **NAME_PORTFOLIO**: The portfolio type of the previous application, such as **cash loans, POS loans, or car loans,** are having different **default** rates.

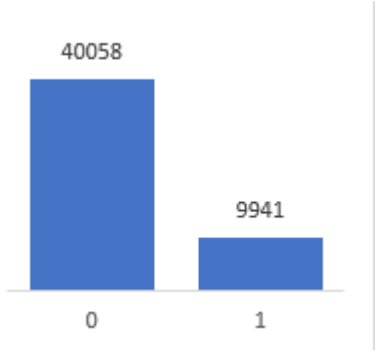
Distribution of portfolio types showing **imbalance** between different loan types.



Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

(vi) **NFLAG_INSURED_ON_APPROVAL**: Indicates whether the client requested insurance during the previous application.

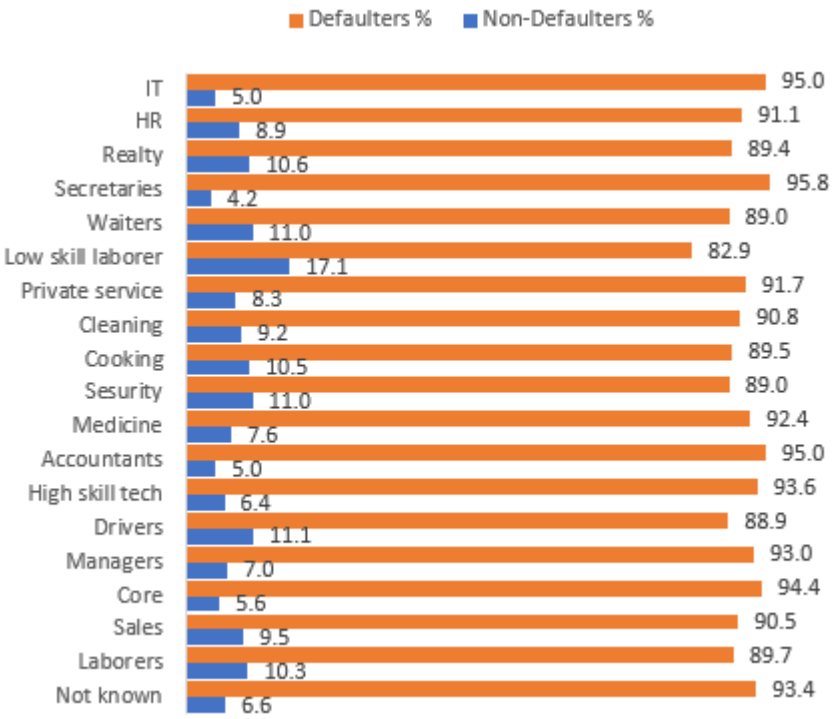
Distribution of insurance requests highlighted differences in default rates between less **insured (1) and uninsured (0)** applications.
Lenders may require some clients to obtain **insurance** as a condition of loan approval. This ensures that the lender's interests are protected.



(vii) OCCUPATION_TYPE: Occupation could indicate income stability and borrowing behavior.

The distribution of **occupation** types across the **target** variable identified occupational disparities in **default** rates.
Defaulters are more prevalent among occupations such as **Laborers, Drivers, and Low-skill Laborers**, while those in occupations like **Accountants**,
Secretaries, and **IT professionals** exhibit **lower default** rates.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2



Conclusion

Based on our analysis, it's evident that certain loan types, client demographics, and risk factors significantly influence default rates.

To mitigate risks and optimize loan portfolio performance, the bank should implement targeted risk assessment measures, consider demographic trends in loan approval processes, and prioritize customer retention strategies based on loyalty indicators.

By aligning lending practices with these insights, the bank can enhance risk management practices, ensure sustainable growth, and maintain competitive advantage in the market.

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.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

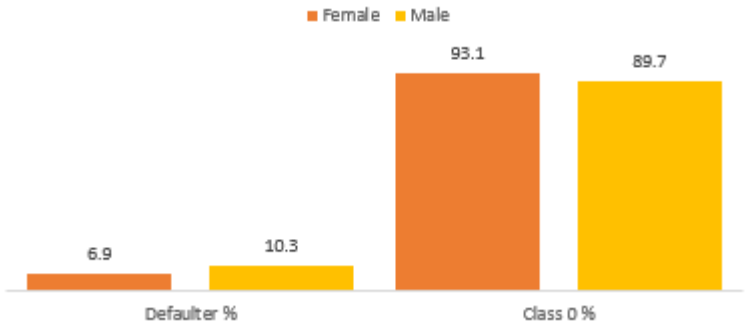
Hint: Utilize Excel functions like COUNT, AVERAGE, MEDIAN, and statistical functions for descriptive analysis. Utilize Excel features like filters, sorting, and pivot tables for segmented and bivariate analysis.

Graph suggestion: Create histograms, bar charts, or box plots to visualize the distributions of variables. Create stacked bar charts or grouped bar charts to compare variable distributions across different scenarios. Create scatter plots or heatmaps to visualize the relationships between variables and the target variable.

Columns from dataset that I consider for univariate analysis:

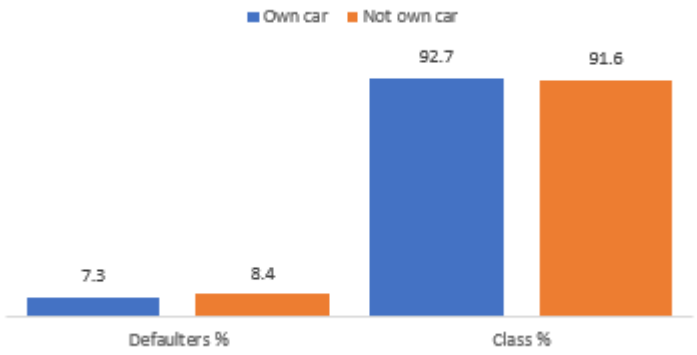
(i) Client Information: CODE_GENDER: Gender of the client.

The majority of **non-defaulters** belong to the **female** category, with a percentage of **93**.1%, compared to **89**.7% for **males**.This insight emphasizes the significance of **gender segmentation** in assessing credit risk, with **female** clients demonstrating a more favorable repayment profile compared to their **male** counterparts.



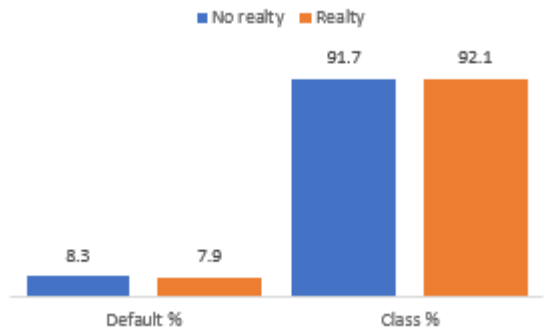
(ii) Client Information: FLAG_OWN_CAR: Flag indicating if the client owns a car.

Car ownership may correlate with a slightly lower likelihood of **default**.



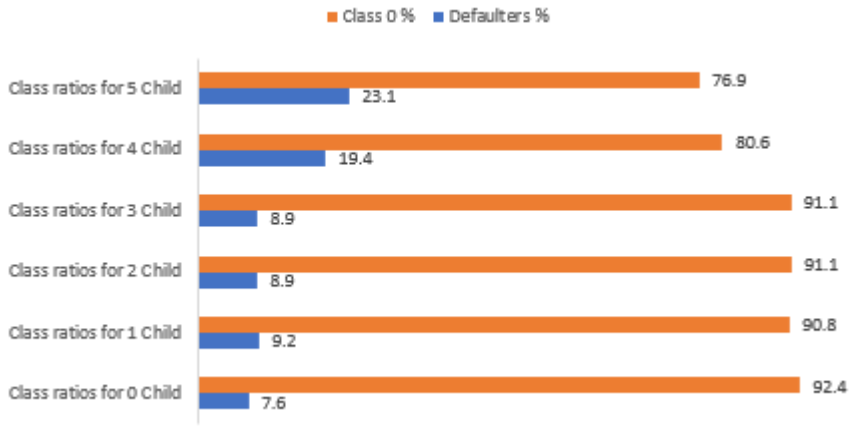
(iii) Client Information: FLAG_OWN_REALTY: Flag indicating if the client owns real estate.

The **percentage** distribution of clients who belong to the **non-defaulters** class shows a slightly higher representation among those with "**realty** (92.1%) compared to those without **realty** (**91**.7%). I suggest a subtle distinction in default behavior based on **realty** ownership, albeit with relatively minor variations in the distribution of **non-defaulters**."



(iv) Client Information: CNT_CHILDREN: Number of children the client has.

Clients with **no children** have a relatively lower default rate of **7.6**%, which gradually increases with the number of children. Notably, clients with four or more children exhibit significantly higher **default** rates, reaching as high as **23**.1% for clients with **five** children.



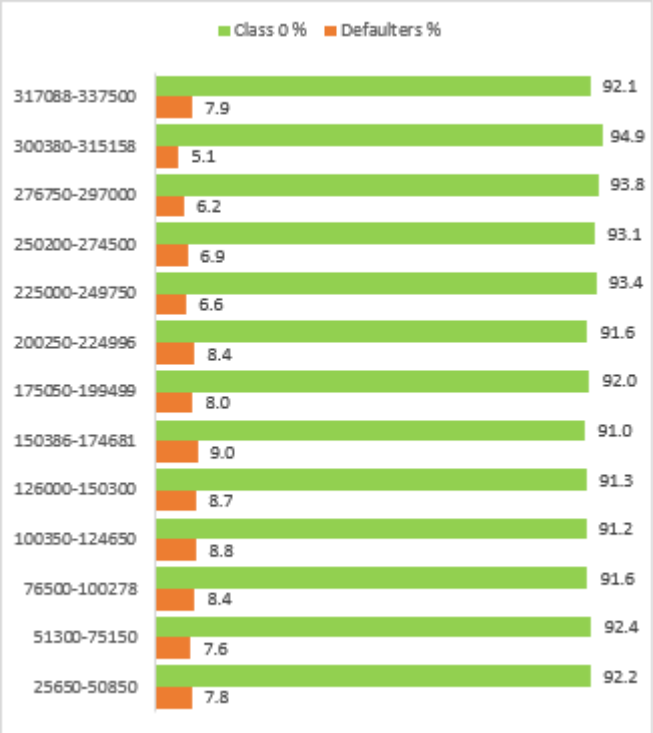
(v) Client Information: AMT_INCOME_TOTAL: Income of the client.

Those earning between **25,650 and 50,850**, exhibit a **default** rate of 7.8%, which remains relatively **stable for income** ranges up to 150,300.

However, default rates begin to increase for **higher** income brackets, reaching 9.0% for clients earning between 150,386 and 174,681.

There is a slight **decrease** in default rates for clients earning between **225,000 and 249,750**, where the default rate **drops to 6.6%.**

This suggests a nuanced relationship between **income** level and **default** risk, highlighting the importance of thorough analysis when assessing borrower **creditworthiness**.

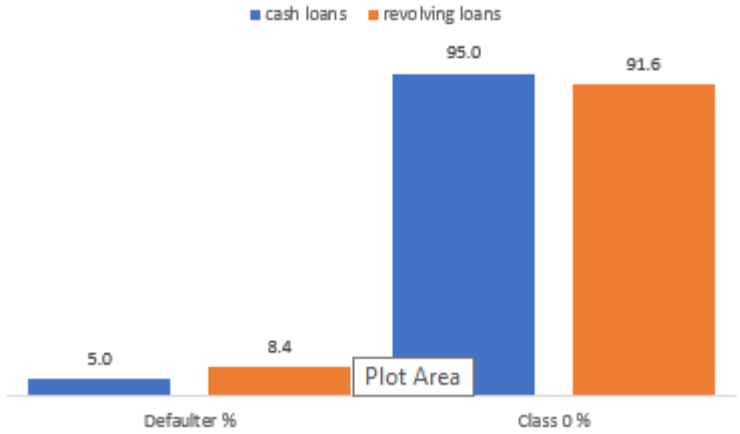


Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

(vi) **Contract Information: NAME_CONTRACT_TYPE**: Type of loan contract (**cash loan, revolving loan**).

Lenders may consider adjusting their **risk** assessment strategies, offering more favorable terms or **lower interest** rates for **cash** loans to **attract low-risk** borrowers.

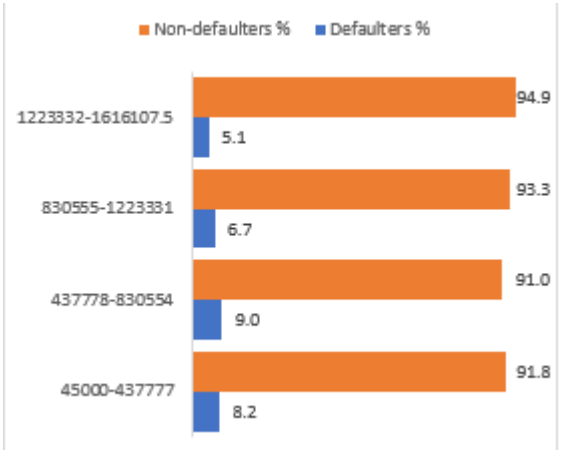
Additional scrutiny or stricter eligibility criteria may be warranted for **revolving** loans to mitigate the **higher** default risk associated with this loan type.



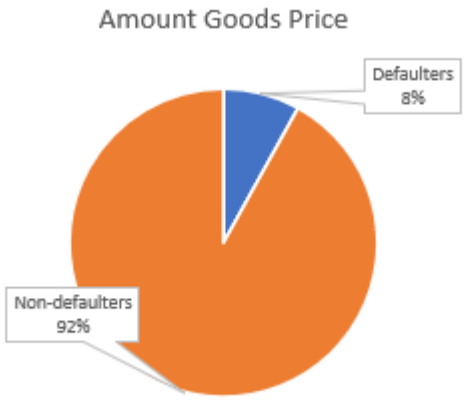
(vii) **Contract Information: AMT_CREDIT**: Credit amount of the loan.

The **higher the amount of loans** credited the **lesser the default rate.**

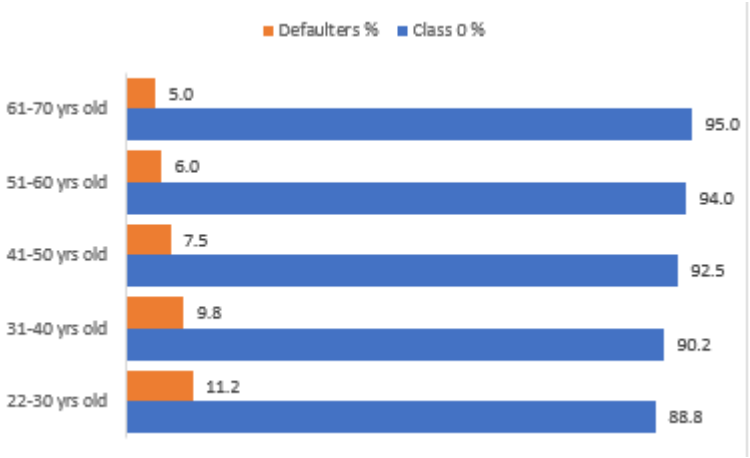
Lenders may implement stricter risk assessment measures for borrowers seeking **mid-range credit** amounts, where default rates are relatively higher, while offering more favorable terms to borrowers in the **lower and higher** credit amount ranges, where **default rates** are comparatively lower.



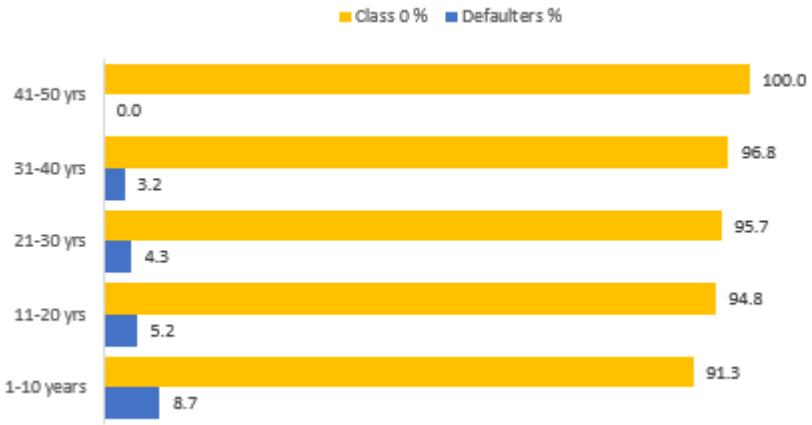
(viii) **Contract Information:AMT_GOODS_PRICE:** Price of the goods for which the loan is given (**for consumer loans).**



(ix) **Client's Age and Employment: DAYS_BIRTH:** Client's age in days at the time of application.



(x) **Client's Age and Employment: DAYS_EMPLOYED:** Number of days before the application the person started current employment.

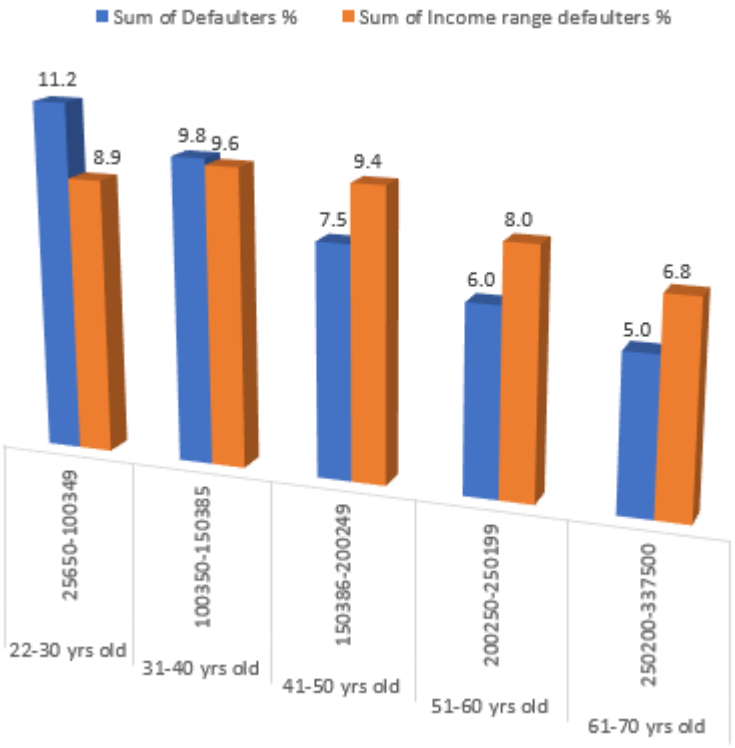


<u>Bivariate analysis involves examining the relationship between two variables to determine if there is a correlation or association between them.</u>

In the context of **analyzing** the distribution of individual variables, I typically look at how each **variable** interacts with or influences other variables.

To identify which columns are suitable for bivariate analysis to understand the distribution of **individual** variables, I consider the following:

(i) **Age vs. Income:** Analyzing how income levels vary across different age groups.



(ii) **Education Level vs. Income:** Exploring how income levels vary based on the education level of individuals.

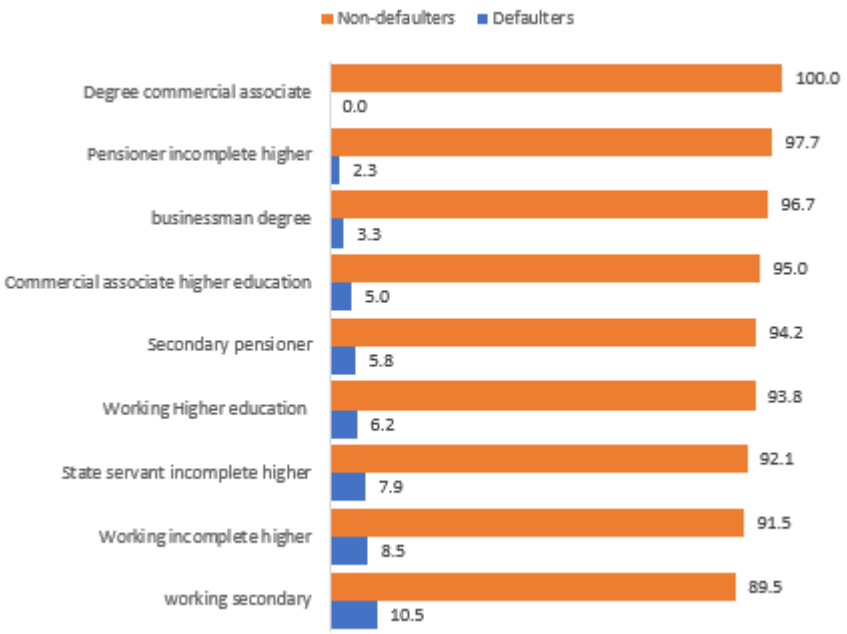
Incomplete higher education may face more challenges in **loan repayment** compared to those with a **secondary education**.

But **Pensioners with incomplete higher education** have a high likelihood of repaying their loans. This could be attributed to their **stable income** from pensions and pensioners higher ages.

Individuals with **incomplete** higher education show varying **levels** of default rates depending on their employment **status**,

indicating that factors such as **income** stability and **employment** status play a significant role in loan repayment behavior.

The findings suggest that **higher education** levels and **aged businessman** or **pensioners** are associated with **lower default** risk.



Conclusion

Leveraging these insights can inform the bank's decision-making processes, enabling them to refine risk assessment methodologies, tailor loan products to specific client segments, and ultimately enhance portfolio performance and mitigate default risks.

By incorporating these recommendations into the bank's loan system, stakeholders can expect improved risk management practices, more tailored loan products, and ultimately, enhanced portfolio performance with reduced default risks.

This proactive approach ensures the bank remains competitive while fostering responsible lending practices and maintaining customer satisfaction.

Incorporate gender-based segmentation and asset ownership considerations specifically for defaulters, while analyzing family dynamics, income patterns, and loan contract types to refine risk mitigation strategies.

Additionally, leverage insights from defaulters' age, employment duration, education level, and employment status to enhance risk assessment frameworks, facilitating more precise risk profiling and targeted loan interventions.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

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.

Hint: Utilize Excel functions like CORREL to calculate correlation coefficients between variables and the target variable within each segment. Rank the correlations to identify the top indicators of loan default for each scenario.

Graph suggestion: Create correlation matrices or heatmaps to visualize the correlations between variables within each segment. Highlight the top correlated variables for each scenario using different colors or shading.

	CNT_CHILDREN	Amt income total	amt credit	amt annuity	amt goods price	REGION_POPU	Days Birth	DAYS_REGIST	DAYS_ID_PUBL	FLAG_EMP_PH	FLAG_WORK_P
CNT_CHILDREN	1	0.013815532	0.008645039	0.03592808	0.008233629	-0.02521935	0.330209118	0.181648362	-0.03267306	0.241310747	0.05539575
Amt income total	0.013815532	1	0.213654371	0.25269011	0.184321123	0.046745298	0.071241054	0.043060992	0.030006188	0.134839718	-0.045545267
amt credit	0.008645039	0.213654371	1	0.61655732	0.753416436	0.052017414	-0.05349302	-0.002219	-0.005653626	0.059770405	-0.020884622
amt annuity	0.035928082	0.252690107	0.616557318	1	0.52493399	0.060886302	0.013340924	0.027949632	0.009770489	0.099839252	-0.02366584
amt goods price	0.008233629	0.184321123	0.753416436	0.52493399	1	0.040010582	-0.04197186	-0.00734103	0.002308613	0.057595579	-0.011871877
REGION_POPULATION_RELA	-0.025219349	0.046745298	0.052017414	0.0608863	0.040010582	1	-0.03259251	-0.05932284	-0.004431877	0.004254548	-0.016767146
Days Birth	0.330209118	0.071241054	-0.053499015	0.01334092	-0.04197186	-0.03259251	1	0.333753253	0.270316515	0.617588025	0.1751858
DAYS_REGISTRATION	0.181648362	0.043060992	-0.002219	0.02794963	-0.00734103	-0.05932284	0.333753253	1	0.10429114	0.206886083	0.059893449
DAYS_ID_PUBLISH	-0.03267306	0.030006188	-0.005653626	0.00977049	0.002308613	-0.00443188	0.270316515	0.10429114	1	0.271719531	0.05020355
FLAG_EMP_PHONE	0.241310747	0.134839718	0.059770405	0.09983925	0.057595579	0.004254548	0.617588025	0.206886083	0.271719531	1	0.232318376
FLAG_WORK_PHONE	0.05539575	-0.045545267	-0.020884622	-0.02366584	-0.01187188	-0.01676715	0.1751858	0.059893449	0.05020355	0.232318376	1
FLAG_CONT_MOBILE	-0.002867017	0.000523972	0.022337118	0.02231854	0.017898205	-0.00504479	-0.01198801	-0.00029098	-0.004034344	-0.015168376	0.022425103
FLAG_PHONE	-0.030502964	-0.018672308	0.0106746	-0.00478214	0.013184299	0.094342977	-0.04467237	-0.07244175	-0.033872173	-0.023645381	0.297537706
FLAG_EMAIL	0.027057995	0.055223151	0.000739784	0.0533258	0.000924939	0.03922941	0.092056295	0.029954694	0.033060809	0.066873719	-0.008463918
CNT_FAM_MEMBERS	0.877801587	0.018656211	0.058420035	0.07847339	0.049101536	-0.02308122	0.277117785	0.170179236	-0.026130803	0.230274079	0.067434972
REGION_RATING_CLIENT	0.026406926	-0.099344352	-0.056025162	-0.06954184	-0.03638423	-0.53286479	0.017023193	0.087529173	-0.002027262	-0.034380405	0.006189288
REGION_RATING_CLIENT_W	0.023210834	-0.110532008	-0.059612305	-0.07845342	-0.0370551	-0.53060769	0.014713778	0.079844297	-0.007002337	-0.036937283	0.011454571
HOUR_APPR_PROCESS_ST	-0.006871551	0.054976698	0.038069875	0.03216704	0.038761303	0.167827319	0.090110232	-0.00793583	0.033761669	0.088117173	0.034785093
REG_REGION_NOT_LIVE_RE	-0.010621289	0.040473305	0.012983843	0.02012968	0.011732199	-0.00357084	0.059106022	0.027237505	0.032836307	0.037363393	0.06550294
REG_REGION_NOT_WORK_F	0.012290733	0.082542939	0.028607213	0.05488034	0.01982442	0.060040938	0.093894967	0.033192558	0.047257521	0.10685707	0.071438661
LIVE_REGION_NOT_WORK_F	0.019918984	0.071116381	0.030000933	0.05469038	0.01777437	0.085700873	0.067977623	0.022285871	0.033195329	0.094694829	0.04355711
REG_CITY_NOT_LIVE_CITY	0.01959191	0.013991121	-0.023695732	-0.00255808	-0.01729678	-0.04634191	0.18220512	0.06826042	0.075885466	0.09488078	0.053148243
REG_CITY_NOT_WORK_CITY	0.069845204	0.037877048	-0.014347482	0.01645295	-0.00500827	-0.04045687	0.237820221	0.094194997	0.10255396	0.256348144	0.122674897
LIVE_CITY_NOT_WORK_CITY	0.067425799	0.033068824	0.003447862	0.02394103	0.007129747	-0.01355302	0.150327122	0.063134638	0.062698648	0.21816435	0.109141273
EXT_SOURCE_2	-0.017642432	0.08886703	0.098809809	0.08823312	0.073229148	0.200991652	-0.09373636	-0.06105848	-0.047417811	0.024714064	-0.018530726
EXT_SOURCE_3	-0.040859487	-0.05709512	0.01894811	-0.00319437	0.00827843	-0.00660855	-0.18154776	-0.10080865	-0.110863209	-0.101948739	-0.050856039
OBS_30_CNT_SOCIAL_CIRC	0.016462933	-0.003757749	0.006790877	0.00633176	0.007394041	-0.01837853	0.011375119	0.010370873	-0.012460985	-0.004948808	-0.020155842
DEF_30_CNT_SOCIAL_CIRCL	-0.00338096	-0.016328844	-0.009440992	-0.01325735	-0.00753071	0.008648682	0.001878241	0.004888719	0.00112625	-0.016093708	-0.014233274
OBS_60_CNT_SOCIAL_CIRC	0.016353577	-0.003646141	0.007014877	0.0067088	0.00760171	-0.017267	0.011265457	0.01064567	-0.012658886	-0.004881686	-0.020589815
DEF_60_CNT_SOCIAL_CIRCL	-0.004603934	-0.017933329	-0.013622535	-0.01738469	-0.00898212	0.003402776	0.002764857	0.006778154	0.001473871	-0.014984609	-0.01182143
DAYS_LAST_PHONE_CHANG	-0.00230436	-0.044141496	-0.077676892	-0.06932856	-0.07029429	-0.04777481	0.080143041	0.052186713	0.091375827	-0.025515818	-0.041379229

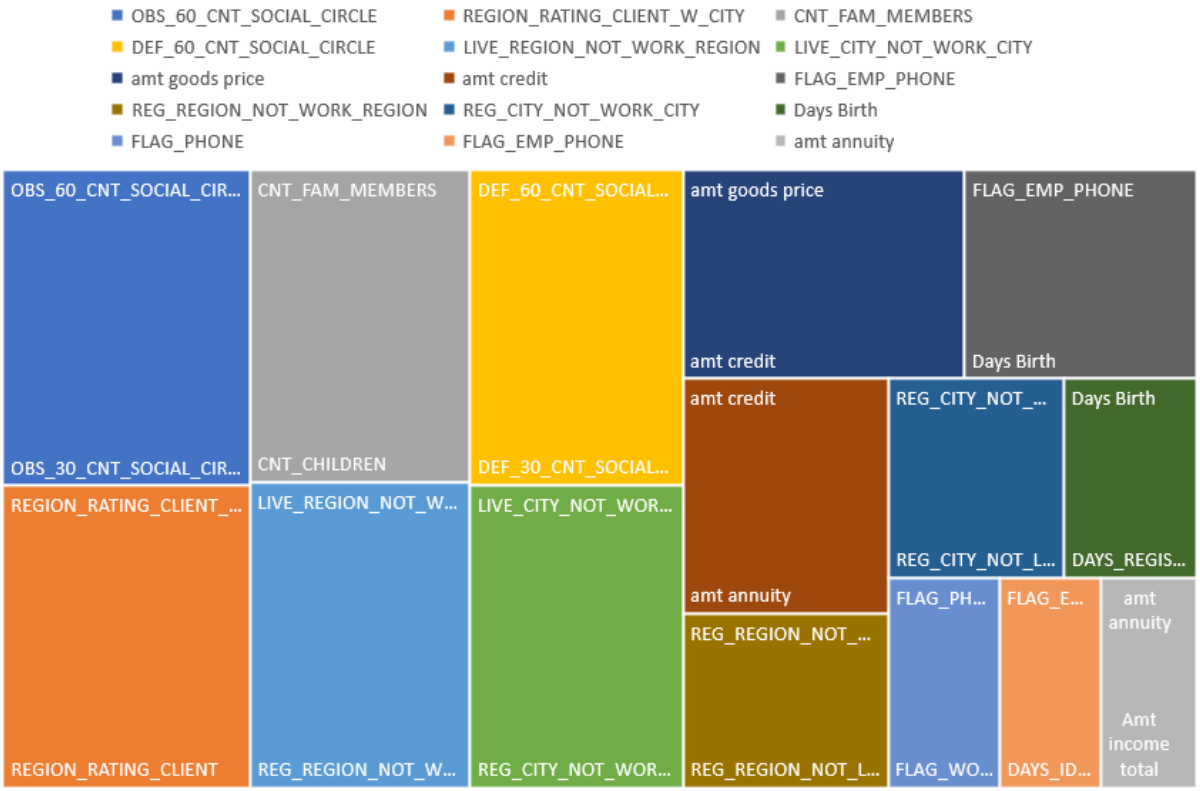
[Click Here to view the M S Excel file of this Case Study file1]

[Click Here to view the M S Excel file of this Case Study file2]

-0.002867017	-0.030502964	0.027057995	0.8778016	0.026406926	0.02321083	-0.00687155	-0.01062129	0.012290733	0.019916984	0.01959191	0.069845204
0.000523972	-0.018672308	0.055223151	0.0186562	-0.09934435	-0.110532	0.054976698	0.04047331	0.082542999	0.071116381	0.01399112	0.037877048
0.022337118	0.0106746	0.000739784	0.05842	-0.05602516	-0.0596123	0.038069875	0.01298384	0.028607213	0.030000933	-0.0236957	-0.01434748
0.022318543	-0.004782142	0.053325796	0.0784734	-0.06954184	-0.0784534	0.032167038	0.02012968	0.054880338	0.05469038	-0.0025581	0.016452953
0.017898205	0.013184299	0.000924939	0.0491015	-0.03638423	-0.0370551	0.038761303	0.0117322	0.01982442	0.01777437	-0.0172968	-0.00500827
-0.005044788	0.094342977	0.03922941	-0.0230812	-0.53286479	-0.5306077	0.167827319	-0.00357084	0.060040938	0.085700873	-0.0463419	-0.04045687
-0.011988011	-0.044672371	0.092056295	0.2771178	0.017023193	0.01471378	0.090110232	0.05910602	0.093894967	0.067977623	0.18220512	0.237820221
-0.000290984	-0.072441751	0.029354694	0.1701792	0.087529173	0.0798443	-0.00793583	0.0272375	0.033192558	0.022285871	0.06826042	0.094194397
-0.004034344	-0.033872173	0.033060809	-0.0261308	-0.00202726	-0.0070023	0.033761669	0.03283631	0.047257521	0.033195329	0.07588547	0.10255396
-0.015168376	-0.023645381	0.066873719	0.2302741	-0.03438041	-0.0369373	0.088117173	0.03736339	0.10685707	0.094694829	0.09488078	0.256348144
0.022425103	0.297537706	-0.008463918	0.067435	0.006189288	0.01145457	0.034785093	0.06550294	0.071438661	0.04355711	0.05314825	0.122674897
1	0.004988427	-0.01046405	-0.0048677	0.014235691	0.01429664	0.003314909	-0.00177291	-0.00196496	-0.00228293	-0.0031604	-0.00164382
0.004988427	1	0.015657941	-0.016841	-0.08856028	-0.0832428	0.061194926	0.00661586	-0.00213824	-0.00432315	-0.0413708	-0.04257912
-0.01046405	0.015657941	1	0.0220901	-0.05584662	-0.0533229	0.020759313	0.01743874	0.04024021	0.040261418	0.00991044	0.005787619
-0.004867688	-0.016841037	0.022090113	1	0.026209548	0.02533588	-0.01169098	-0.01212914	0.007055611	0.014932472	0.01346875	0.073293487
0.014235691	-0.088560277	-0.055846622	0.0262095	1	0.95071378	-0.28387121	-0.0410944	-0.14170207	-0.14776355	0.03838231	0.010338872
0.014296644	-0.083242826	-0.053322939	0.0253359	0.950713779	1	-0.2625608	-0.03707308	-0.13468281	-0.14159871	0.04802486	0.030637802
0.003314909	0.061194926	0.020759313	-0.011691	-0.28387121	-0.2625608	1	0.05073449	0.073624663	0.060179979	0.01714387	0.02320361
-0.001772911	0.006615862	0.017438737	-0.0121291	-0.0410944	-0.0370731	0.050734493	1	0.456288339	0.082107207	0.33561352	0.14345092
-0.001964957	-0.002138242	0.04024021	0.0070556	-0.14170207	-0.1346828	0.073624663	0.45628834	1	0.856960859	0.15442748	0.235461479
-0.002282934	-0.004323147	0.040261418	0.0149325	-0.14776355	-0.1415987	0.060179979	0.08210721	0.856960859	1	0.02147181	0.180769405
-0.003160396	-0.041370834	0.00991044	0.0134687	0.038382306	0.04802486	0.017143873	0.33561352	0.154427477	0.02147181	1	0.445162522
-0.001643817	-0.042579125	0.005787619	0.0732935	0.010338872	0.0306378	0.02320361	0.14345092	0.235461479	0.180769405	0.44516252	1
0.000162495	-0.024295429	0.001110565	0.0781902	-0.01595739	0.00014642	0.011617431	0.00312919	0.189729412	0.231672126	0.0256617	0.821527389
-0.003814635	0.061658575	0.023546685	0.0026293	-0.29526595	-0.2907282	0.15673624	0.016929	0.029056499	0.02787533	-0.0470761	-0.07794317
0.003559022	-0.004035558	-0.051543824	-0.0246633	-0.00902794	-0.0079772	-0.03889825	-0.03901377	-0.03805235	-0.02251685	-0.0620577	-0.06218926
0.006663695	-0.034921528	-0.003483504	0.0256791	0.035205228	0.03282846	-0.00900773	-0.01651648	-0.02594895	-0.02036742	-0.0090787	-0.00891262
0.000284307	-0.027756489	-0.003956238	-0.0027575	0.011047461	0.00935277	-0.00169588	-0.00600273	-0.00783366	-0.00695468	0.0068993	0.001229043
0.006581738	-0.034696916	-0.003039427	0.0256992	0.034880605	0.03239071	-0.00899664	-0.01653427	-0.02607867	-0.02051515	-0.0093174	-0.00887227
0.002375419	-0.026493523	-0.00378143	-0.0046251	0.013083785	0.01139562	-0.00494815	-0.00716425	-0.01169989	-0.01093507	0.00689504	0.003474436
-0.024578136	-0.06731675	-0.018036035	-0.0223569	0.02737055	0.02685799	-0.01785395	0.03160712	0.034483138	0.024025689	0.05378625	0.046830654
FLAG_CONT_M(	FLAG_PHONE	FLAG_EMAIL	CNT_FAM_M	REGION_RATIN	REGION_RATI	HOUR_APPR_	REG_REGION	REG_REGION_	LIVE_REGION_	REG_CITY_NC	REG_CITY_NC

0.067425799	-0.017642492	-0.04085949	0.016482933	-0.00338096	0.016353577	-0.00460393	-0.00230436	
0.033068824	0.08886703	-0.05709512	-0.003757749	-0.016328844	-0.003646141	-0.01793333	-0.044141496	
0.003447862	0.098809809	0.01894811	0.006790877	-0.009440992	0.007014877	-0.01362254	-0.077676892	
0.02394103	0.088233123	-0.00319437	0.006331756	-0.013257346	0.0067088	-0.01738469	-0.069328557	
0.007129747	0.073229148	0.00827843	0.007394041	-0.007530714	0.00760171	-0.00898212	-0.07029429	
-0.01355302	0.200991652	-0.00660855	-0.018378531	0.008648682	-0.017266999	0.003402776	-0.047774814	
0.150327122	-0.093736364	-0.18154776	0.011375119	0.001878241	0.011265457	0.002764857	0.080143041	
0.063134638	-0.061058484	-0.10080865	0.010370873	0.004888719	0.01064567	0.006778154	0.052186713	
0.062698648	-0.047417811	-0.11086321	-0.012460985	0.001126245	-0.012658886	0.001473871	0.091375827	
0.21816435	0.024714064	-0.10194874	-0.004948808	-0.016093708	-0.004881686	-0.01498461	-0.025515818	
0.109141273	-0.018530726	-0.05085604	-0.020155842	-0.014233274	-0.020589815	-0.01182143	-0.041379229	
0.000162495	-0.003814635	0.003559022	0.006663695	0.000284307	0.006581738	0.002375419	-0.024578136	
-0.02429543	0.061658575	-0.00403556	-0.034921528	-0.027756489	-0.034696916	-0.02649352	-0.06731675	
0.001110565	0.023546685	-0.05154382	-0.003483504	-0.003956238	-0.003039427	-0.00378143	-0.018036035	
0.07819019	0.002629275	-0.02466326	0.025679118	-0.002757522	0.025693201	-0.00462514	-0.02235688	
-0.01595739	-0.295265951	-0.00902794	0.035205228	0.011047461	0.034880605	0.013083785	0.02737055	
0.000146423	-0.290728158	-0.00797715	0.032828456	0.009352773	0.032390715	0.011395619	0.026857993	
0.011617431	0.15673624	-0.03889825	-0.009007731	-0.001695881	-0.008996638	-0.00494815	-0.017853948	
0.003129188	0.016929001	-0.03901377	-0.016516477	-0.006002733	-0.016534269	-0.00716425	0.031607123	
0.189729412	0.029056499	-0.03805235	-0.025948951	-0.007833658	-0.026078669	-0.01169989	0.034483138	
0.231672126	0.02787533	-0.02251685	-0.020367423	-0.006954679	-0.020515146	-0.01093507	0.024025689	
0.025661698	-0.047076109	-0.0620577	-0.009078692	0.006899305	-0.009317382	0.00689504	0.053786246	
0.821527389	-0.077943169	-0.06218926	-0.008912624	0.001229043	-0.008872268	0.003474436	0.046830654	
1	-0.060649766	-0.03237115	-0.006824654	-0.003640879	-0.006633134	-0.00156363	0.021990657	
-0.06064977	1	0.082683642	-0.018289152	-0.033037608	-0.0177233	-0.03668486	-0.192395002	
-0.03237115	0.082683642	1	-0.004419439	-0.034249211	-0.004888988	-0.03200566	-0.062980243	
-0.00682465	-0.018289152	-0.00441944	1	0.311728581	0.99833113	0.236000205	-0.013959209	
-0.00364088	-0.033037608	-0.03424921	0.311728581	1	0.314135954	0.856275241	0.005042925	
-0.00663313	-0.0177233	-0.00488899	0.99833113	0.314135954	1	0.238189895	-0.014760964	
-0.00156363	-0.036684856	-0.03200566	0.236000205	0.856275241	0.238189895	1	0.006115366	
0.021990657	-0.192395002	-0.06298024	-0.013959209	0.005042925	-0.014760964	0.006115366	1	
LIVE_CITY_NO	EXT_SOURCE_:	EXT_SOURCE_	OBS_30_CNT_	DEF_30_CNT_S	OBS_60_CNT_S	DEF_60_CNT_S	DAYS_LAST_PH(	

Top Most correlation columns		
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.856
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.857
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.822
amt goods price	amt credit	0.753
amt credit	amt annuity	0.617
FLAG_EMP_PHONE	Days Birth	0.618
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.456
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.445
Days Birth	DAYS_REGISTRATION	0.334
FLAG_PHONE	FLAG_WORK_PHONE	0.298
FLAG_EMP_PHONE	DAYS_ID_PUBLISH	0.272
amt annuity	Amt income total	0.253



(i) OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE:

The **high correlation** suggests that clients who have a **higher number of observations of social surrounding**s with observable default in **both 60 days and 30** days are more likely to default on their loans.
This indicates that **social** factors may play a role in loan default behavior.

(ii) **REGION_RATING_CLIENT_W_CITY and REGION_RATING_CLIENT:**

The **strong correlation** suggests that clients who rate the region with and without the city factor similarly are more **likely** to **default** on their loans.
This could imply that clients living in regions with similar ratings across different factors may face similar economic conditions leading to **loan default**.

(iii) **CNT_FAM_MEMBERS and CNT_CHILDREN:**

The correlation indicates that **larger families** tend to have **more children**, and this could impact their **financial** stability.
Families with **more** children may face **higher** expenses, potentially increasing the likelihood of **loan defaults**.

(iv) **DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE**:

Clients with a **higher** number of **social** surroundings **defaulted** on in both **60** days and **30** days are more likely to **default** on their **loans**.
This suggests that clients with **social** connections who have **defaulted** may influence **others within** their **social circles** to **default** as well.

(v) LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION:

Clients whose **living** and **working** regions do **not** match are more likely to **default** on their **loans**.
This could indicate potential **instability** or **commuting** challenges, **impacting** their financial situation and ability to **repay loans**.

(vi) **LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY**:

Similar to the previous correlation, clients whose living and working cities do not match are more likely to default on their loans.
This discrepancy may reflect geographical mobility or economic disparities between residential and employment areas.

(vii) **AMT_GOODS_PRICE and AMT_CREDIT:**

Clients who borrow higher amounts relative to the price of goods they are purchasing are more likely to default on their loans.
This suggests that clients who overextend themselves financially may struggle to meet their repayment obligations.

(viii) **AMT_CREDIT and AMT_ANNUITY:**

Clients with higher credit amounts and corresponding annuities are more likely to default on their loans.
This could indicate that clients with larger loans may face challenges in managing their debt burden and meeting regular payment obligations.

(ix) **FLAG_EMP_PHONE and Days Birth:**

There is a moderate positive correlation between having an employer phone and the client's age.
This suggests that older clients who are still actively employed may be more likely to default on their loans, potentially due to financial instability or retirement-related factors.

(x) **REG_REGION_NOT_WORK_REGION and REG_REGION_NOT_LIVE_REGION:**

Clients whose regions of registration do not match with their regions of work or living are more likely to default on their loans.
This discrepancy may reflect broader economic disparities or challenges in establishing stable residency or employment.

Conclusion

These insights highlight various socio-economic, demographic, and geographical factors that may influence loan default behavior, providing valuable information for risk assessment and decision-making in lending operations.

Overall Results

Through this project, I used Excel functions like COUNT, ISBLANK, IF, AVERAGE, MEDIAN, QUARTILE, IQR, COUNTIF, SUM, and CORREL to analyze a Bank Loan Case Study.

I identified missing data, imputed values using appropriate methods, detected potential outliers, assessed data imbalance, and performed descriptive and segmented analysis.

By utilizing these functions and techniques, I gained insights into the patterns and relationships within the loan dataset.

I could identify key indicators of loan default and understand how various factors such as income, credit history, and loan types influence the likelihood of default.

This project enhanced my understanding of how data analysis can contribute to informed decision-making in the banking sector.

It allowed me to identify potential risks, improve loan approval processes, and ensure fairness in lending practices.

Moreover, by exploring correlations and trends, I gained valuable insights that can aid in risk assessment and improve overall business strategies for the bank.

Click Here to view the M S Excel file of this Case Study file1
Click Here to view the M S Excel file of this Case Study file2

TheEnd