Trainity Project

BANK LOAN CASE STUDY

Oct 2024

Created By: Ankit Bodar

LOAN DEFAULT RISK ANALYSIS USING EDA

Business Objectives:

The key objective of this project is to identify patterns that can predict loan default risks. Understanding these patterns will allow the company to:

- Deny loans to high-risk customers.
- Reduce loan amounts or offer higher interest rates to applicants with a higher risk of default.
- Ensure that capable applicants are not unfairly rejected. The ultimate aim is to optimize loan approval decisions, minimize defaults, and enhance profitability.

Risk Scenarios:

The company faces two critical risks when processing loan applications:

1. Business Loss: If a customer is capable of repaying the loan but is not approved, the company loses a potential source of revenue.
2. Financial Loss: If a customer is unlikely to repay the loan and is approved, the company risks financial loss due to default.

Dataset Information:

The dataset contains information about loan applications and includes two types of scenarios:

1. Customers with Payment Difficulties: Those who experienced late payments on one or more installments.
2. Customers with On-Time Payments: Those who made timely payments.

When a customer applies for a loan, the outcomes fall into four categories:

- Approved: Loan application was approved.
- Cancelled: Customer cancelled the loan application.
- Refused: Loan application was rejected.
- Unused Offer: The loan was approved, but the customer did not use it.

Expected Outcomes:

By performing thorough EDA, the company will be able to:

- Pinpoint critical customer and loan attributes that are linked to defaults.
- Create a data-driven approach to loan approvals.
- Mitigate risks by improving decision-making in the loan approval process.

APPROCH:

1. Data Understanding:

- Begin by understanding the dataset. This includes examining the attributes in the datasets such as customer demographics, loan application details, and loan statuses.
- The dataset consists of:
    - Approved, cancelled, refused, and unused loan offers.
    - Customers with payment difficulties and customers who paid on time.

2. Data Cleaning:

- Handle missing data: Identify columns with missing values and decide whether to impute, fill, or drop them based on their importance.
- Outlier detection and handling: Use statistical methods (IQR, Z-scores) to identify and cap or remove outliers, especially for numerical features like loan amount or credit scores.

3. Exploratory Data Analysis (EDA)

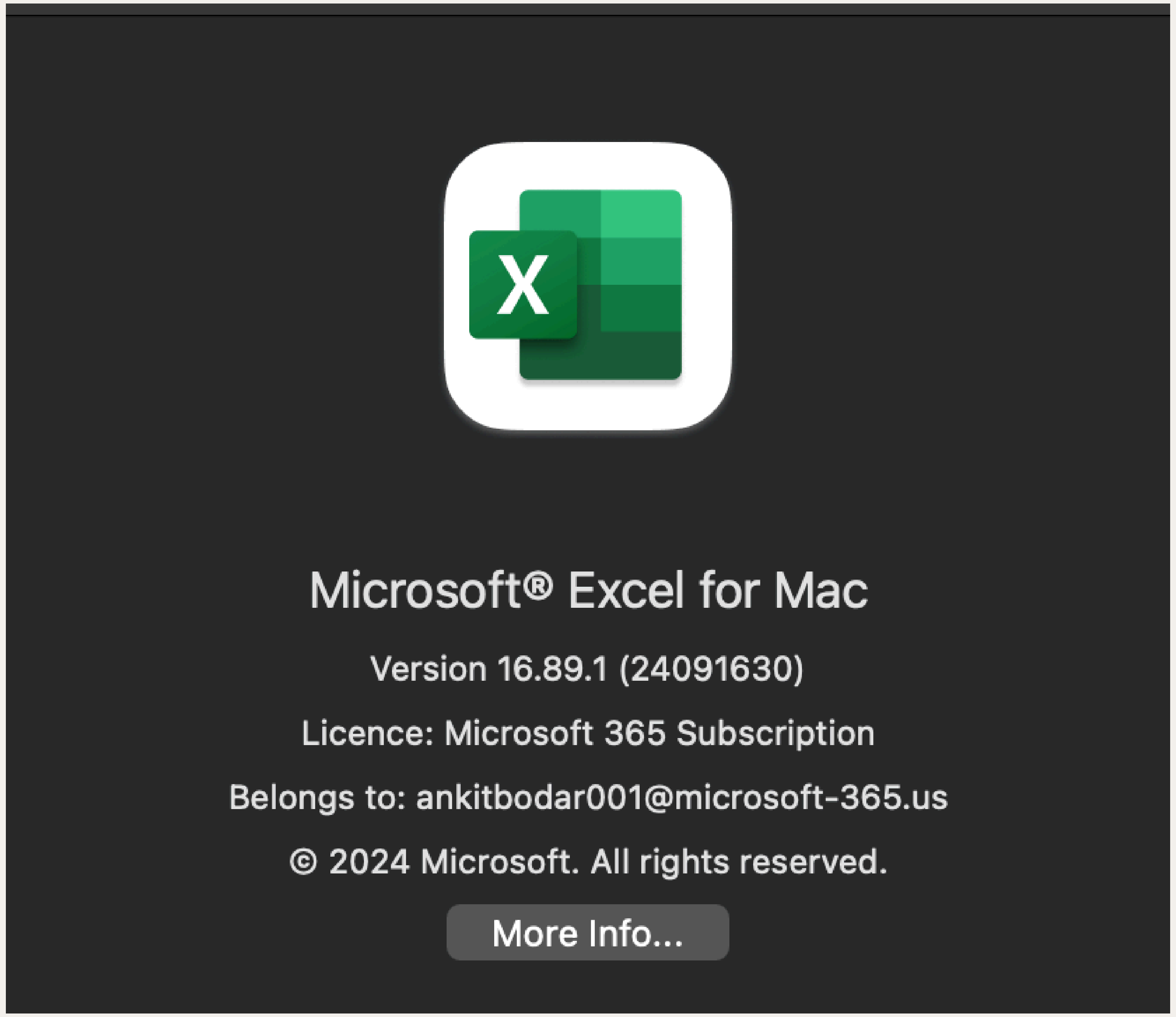
- Univariate Analysis: Focus on single variables, such as loan amount, age, income, etc., to observe general trends in the data.
- Bivariate Analysis: Analyze the relationship between key variables (e.g., loan amount vs. repayment status, income vs. loan approval) to find patterns influencing default.
- Correlation Analysis: Investigate how loan attributes (interest rate, loan amount) and customer attributes (income, credit score) correlate with default risk.

4. Insights:

- Identify high-risk applicants and potential causes of default.
- Identify critiacal parameters related to customer background (age, eduction, income, family members, etc.) which effect the loan repayment.

This structured approach will allow you to derive insights, improve loan approval decisions, and minimize default risks.

Tech-Stack:



WHY EXCEL?

A **Easy to Create Visualizations:** Excel's Quick Analysis tool allows you to quickly generate charts, graphs, and other visualizations with a single click. This makes it easy to interpret data and uncover trends, even for users with minimal technical skills

B **Built-in Statistical Formulas:** Excel provides a wide range of statistical formulas such as AVERAGE(), STDEV(), and CORREL() that are essential for analyzing patterns and making data-driven decisions

C **Pivot Tables for In-depth Analysis:** Excel's PivotTable feature is invaluable for summarizing and analyzing large datasets. You can quickly group, filter, and calculate data, making it easier to draw insights and explore different dimensions of your project

DATA ANALYTICS TASKS:

A. IDENTIFY MISSING DATA AND DEAL WITH IT APPROPRIATELY

IMPORTANCE OF DATA CLEANING:

DATA CLEANING IS CRUCIAL TO ENSURE THAT ANALYSES AND BUSINESS DECISIONS ARE BASED ON ACCURATE, CONSISTENT, AND RELEVANT DATA. IT REMOVES ERRORS AND INCONSISTENCIES THAT CAN DISTORT RESULTS, LEADING TO FLAWED INSIGHTS. CLEAN DATA IMPROVES PRODUCTIVITY BY MINIMIZING TIME SPENT CORRECTING ERRORS LATER AND INCREASES TRUST IN THE ANALYTICS OUTPUT. IN THE CONTEXT OF BUSINESS, RELIABLE DATA IS KEY TO UNDERSTANDING CUSTOMER BEHAVIOR, IMPROVING OPERATIONS, AND MAKING DATA-DRIVEN STRATEGIC DECISIONS .

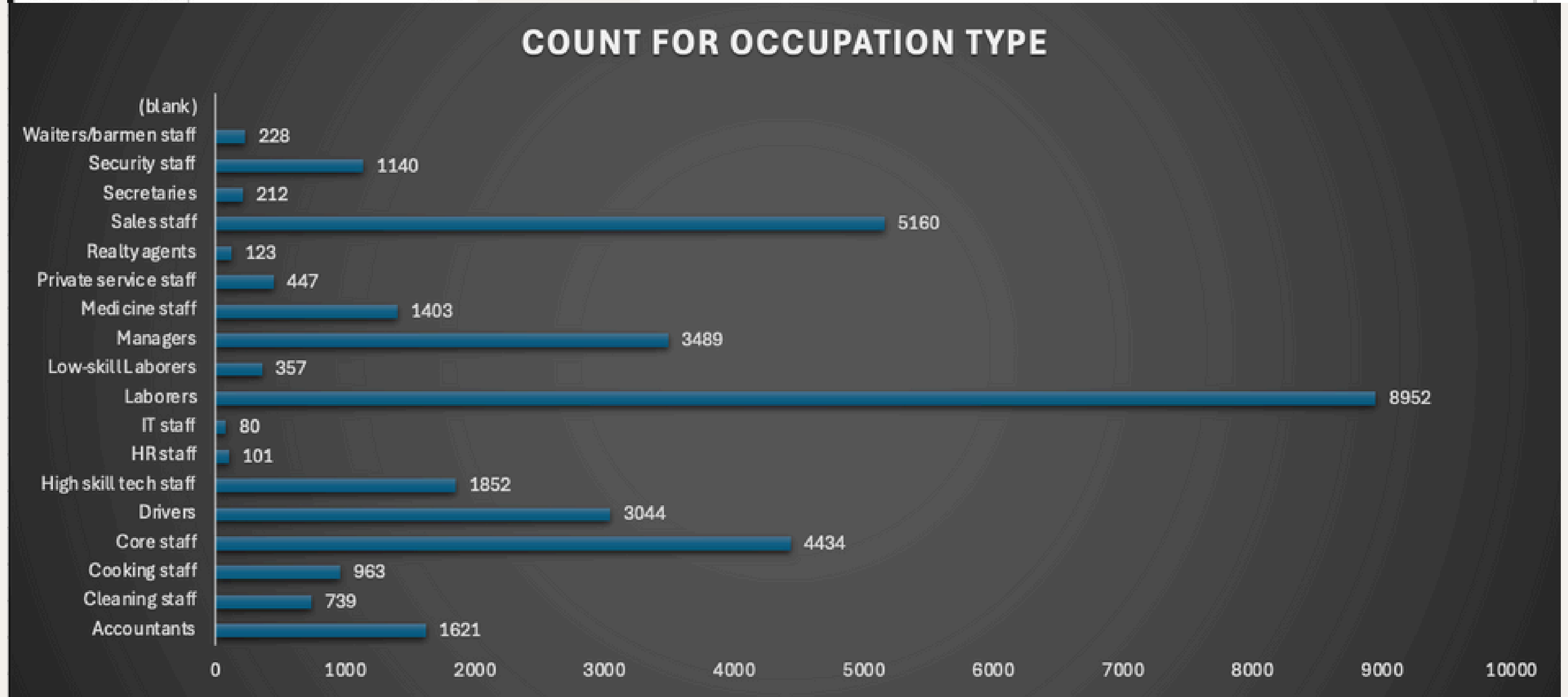
STEPS TAKEN TO CLEAN PROVIDED DATA:

1. WE CONDUCTED A COMPREHENSIVE AUDIT OF THE DATASET USING THE COUNTBLANK FUNCTION TO IDENTIFY COLUMNS WITH MISSING DATA, ENSURING TRANSPARENCY IN DATA QUALITY.
2. WE APPLIED THRESHOLDS OF 30%, 40%, AND 50% TO ASSESS THE IMPACT OF MISSING DATA ON THE DATASET'S INTEGRITY, HELPING US PRIORITIZE COLUMNS FOR FURTHER ACTION BASED ON THE PERCENTAGE OF MISSING VALUES.
3. ALL COLUMNS WITH MORE THAN 40% MISSING DATA WERE REMOVED TO MAINTAIN THE RELEVANCE AND RELIABILITY OF THE DATASET, PREVENTING SKEWED OR INCOMPLETE INSIGHTS FROM AFFECTING BUSINESS DECISIONS.
4. FOR FIELDS WITH FEWER MISSING VALUES, WE APPLIED CONTEXTUAL REPLACEMENTS TO ENSURE CONSISTENCY. FOR INSTANCE, BLANK VALUES IN NAME_TYPE_SUITE WERE REPLACED WITH "UNACCOMPANIED" AND MISSING OCCUPATION DATA WAS LABELED AS "NOT PROVIDED." NUMERIC FIELDS WERE IMPUTED WITH THE MEDIAN TO RETAIN THE DATASET'S STATISTICAL PROPERTIES.
5. WE ALSO CORRECTED ANOMALIES LIKE NEGATIVE VALUES FOR AGE AND WORK EXPERIENCE BY CONVERTING THESE FIELDS INTO MEANINGFUL METRICS (I.E., YEARS), ENSURING ACCURATE AND INTERPRETABLE DATA FOR STRATEGIC DECISION-MAKING.

Row Labels	Count of NAME_TYPE_SUITE
Children	542
Family	6549
Group of people	36
Other_A	137
Other_B	259
Spouse, partner	1849
Unaccompanied	40435
(blank)	
Grand Total	**49807**
Total Null Values	192

Row Labels	Count of OCCUPATION_TYPE
Accountants	1621
Cleaning staff	739
Cooking staff	963
Core staff	4434
Drivers	3044
High skill tech sta	1852
HR staff	101
IT staff	80
Laborers	8952
Low-skill Laborer:	357
Managers	3489
Medicine staff	1403
Private service st	447
Realty agents	123
Sales staff	5160
Secretaries	212
Security staff	1140
Waiters/barmen s	228
(blank)	
Grand Total	**34345**
Total Null Values	15654

Values	Average	Median
AMT_REQ_CREDIT_BUREAU_YEAR	2	1
AMT_REQ_CREDIT_BUREAU_QRT	0	0
AMT_REQ_CREDIT_BUREAU_MON	0	0
AMT_REQ_CREDIT_BUREAU_WEEK	0	0
AMT_REQ_CREDIT_BUREAU_DAY	0	0
AMT_REQ_CREDIT_BUREAU_HOUR	0	0
DAYS_LAST_PHONE_CHANGE	-964	-755
DEF_60_CNT_SOCIAL_CIRCLE	0	0
OBS_60_CNT_SOCIAL_CIRCLE	1	0
DEF_30_CNT_SOCIAL_CIRCLE	0	0
OBS_30_CNT_SOCIAL_CIRCLE	1	0
CNT_FAM_MEMBERS	2	2
AMT_GOODS_PRICE	539060	450000
AMT_ANNUITY	27107	24939
EXT_SOURCE_2	0.513885412	0.565601484
EXT_SOURCE_3	0.511881408	0.53527625

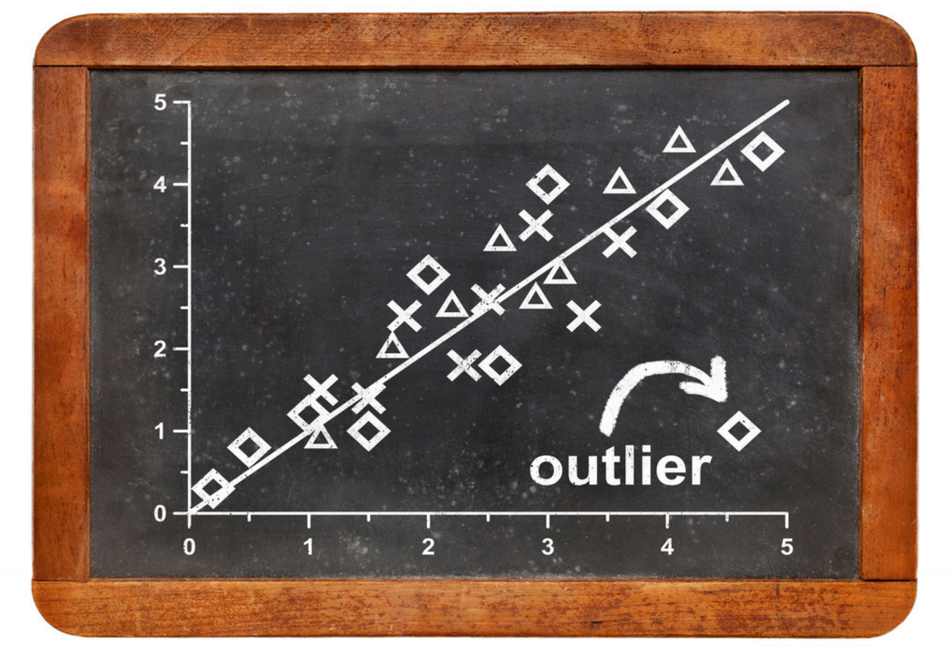


DATA ANALYTICS TASKS:

B. IDENTIFY OUTLIERS IN THE DATASET

IMPORTANCE OF REMOVING/HANDLING OUTLIERS:

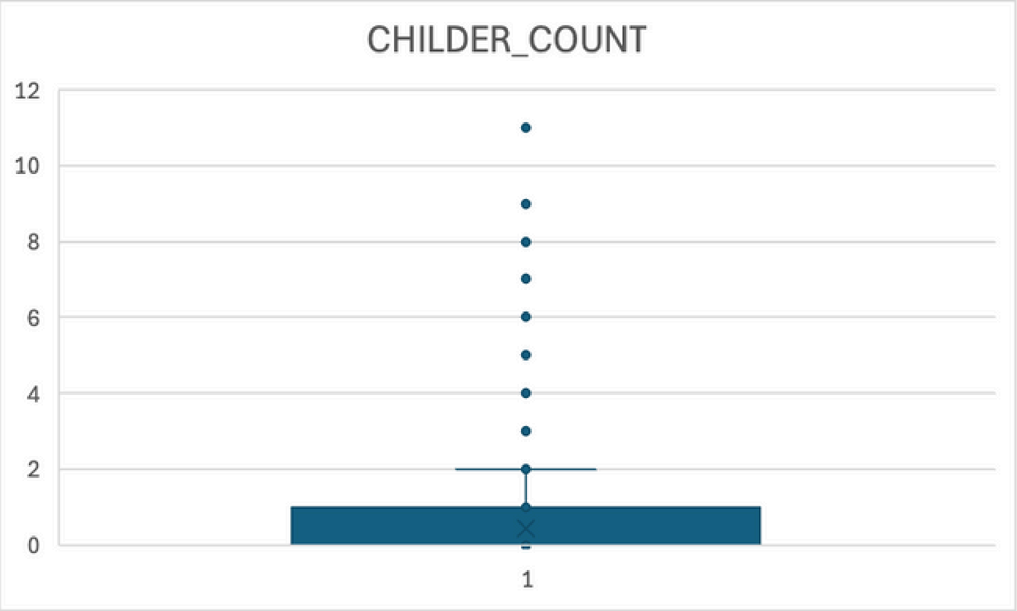
OUTLIERS CAN SIGNIFICANTLY DISTORT THE RESULTS OF DATA ANALYSIS BY SKEWING MEAN VALUES, AFFECTING CORRELATIONS, AND LEADING TO INCORRECT CONCLUSIONS. HANDLING OUTLIERS ENSURES THAT THE ANALYSIS IS BASED ON DATA THAT ACCURATELY REFLECTS THE MAJORITY OF THE DATASET, MAKING INSIGHTS MORE RELIABLE. PROPER MANAGEMENT OF OUTLIERS HELPS IN MAINTAINING THE INTEGRITY OF THE DATASET WITHOUT LOSING KEY INFORMATION. IN A BUSINESS CONTEXT, HANDLING OUTLIERS IS ESSENTIAL TO PREVENT FLAWED DECISION-MAKING, ESPECIALLY IN FINANCIAL RISK ASSESSMENTS, CUSTOMER BEHAVIOR ANALYSIS, AND OPERATIONAL FORECASTING .

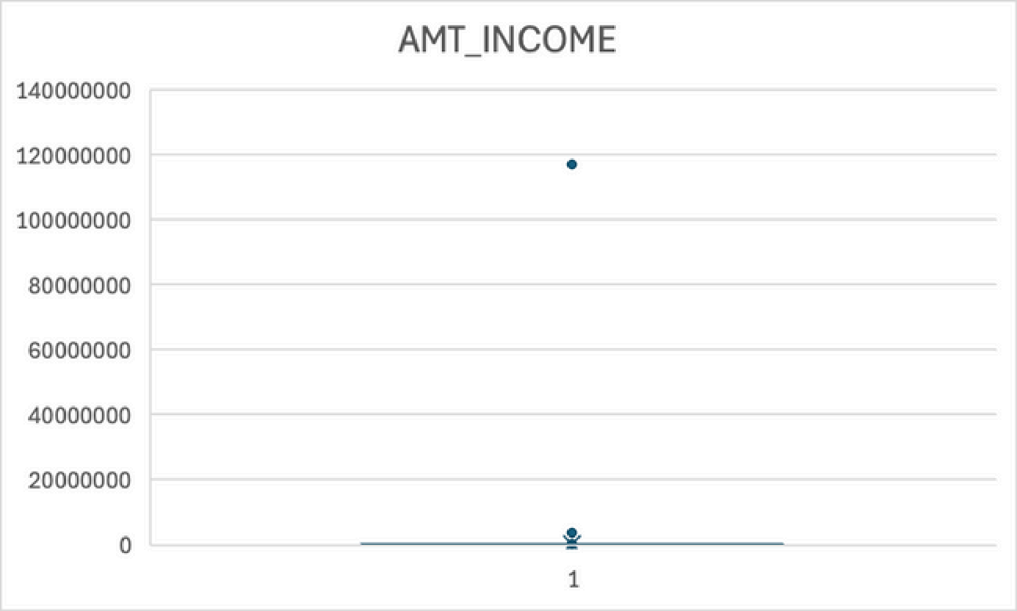


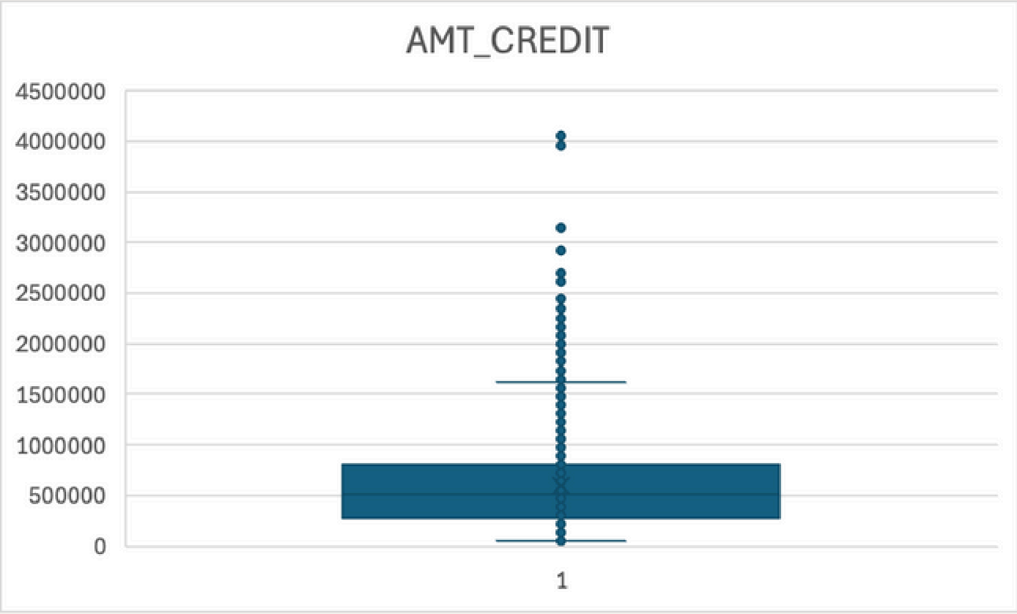
STEPS TAKEN TO REMOVE OUTLIERS FROM THE DATASET:

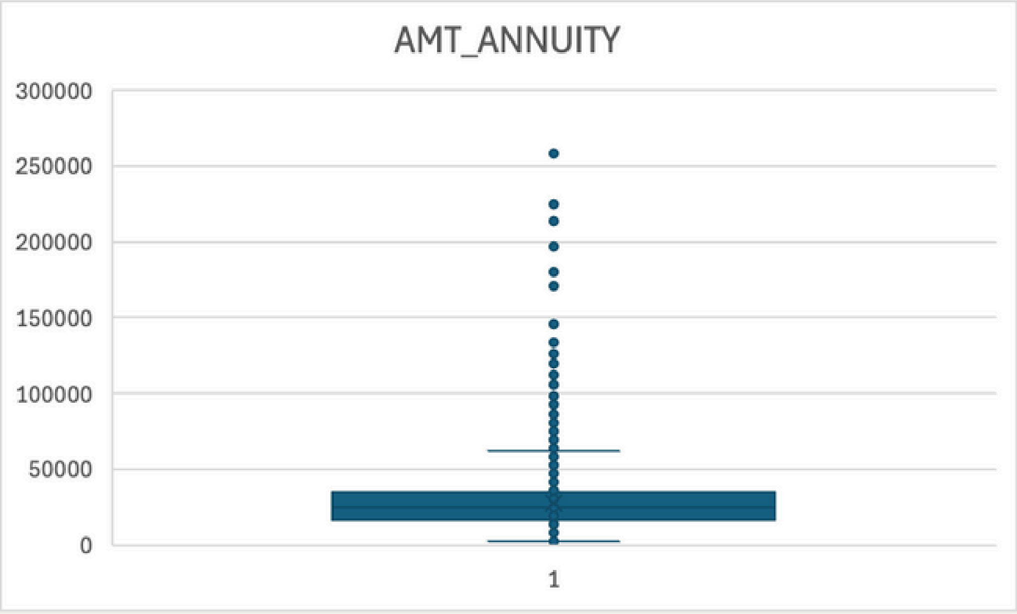
1. I UTILIZED STATISTICAL MEASURES SUCH AS MINIMUM, MAXIMUM, MEAN, AND QUARTILE FUNCTIONS TO CALCULATE RELEVANT METRICS NEEDED FOR OUTLIER DETECTION, ENSURING DATA-DRIVEN ACCURACY IN THE ANALYSIS.
2. BY LEVERAGING THESE VALUES, I IDENTIFIED UPPER AND LOWER BOUNDS (USING THE IQR METHOD) TO SYSTEMATICALLY DETECT OUTLIERS IN EACH NUMERIC COLUMN, ENSURING THAT EXTREME VALUES DON'T DISTORT THE ANALYSIS.
3. TO VISUALLY REPRESENT THESE OUTLIERS, I CREATED BOX AND WHISKER PLOTS, WHICH HELPED IN BETTER UNDERSTANDING THE DISTRIBUTION AND EXTREMITIES OF THE DATA.
4. WHILE EVERY COLUMN HAD OUTLIERS, REMOVING ALL OUTLIERS WOULD HAVE RESULTED IN SIGNIFICANT DATA LOSS. THEREFORE, I STRATEGICALLY FOCUSED ON KEY FIELDS FROM THE 'PREVIOUS_APPLICATION' DATASET TO REMOVE OUTLIERS, PRESERVING THE DATA'S RELEVANCE FOR BUSINESS INSIGHTS.
5. AFTER CLEANING, THE DATASET NOW CONTAINS 36,010 ROWS, PROVIDING A MORE REFINED AND RELIABLE BASIS FOR THE NEXT STEPS IN THE ANALYSIS.

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	AGE_OF_CLIENT	YEARS_EMPLOYED	YEARS_REGISTRATION	YEARS_ID_PUBLISH
Min	0	25650	45000	2052	45000	0.000533	21.04109589	0	0	0
Q1	0	112500	270000	16456.5	238500	0.010006	33.91369863	2.556164384	5.473972603	4.717808219
Mean	0	145800	514777.5	24939	450000	0.01885	43.09863014	6.071232877	12.30136986	8.934246575
Q3	1	202500	808650	34596	679500	0.028663	53.81917808	15.66575342	20.44794521	11.77260274
Max	11	3825000	4050000	258025.5	4050000	0.072508	68.99726027	1000.665753	61.34794521	17.0739726
Upper limit	2.5	337500	1616625	61805.25	1341000	0.0566485	83.67739726	35.33013699	42.90890411	22.35479452
Lower limit	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795	4.055479452	-17.10821918	-16.9869863	-5.864383562
Total Outliers	723	2295	1063	1188	2387	1329	0	9082	96	0

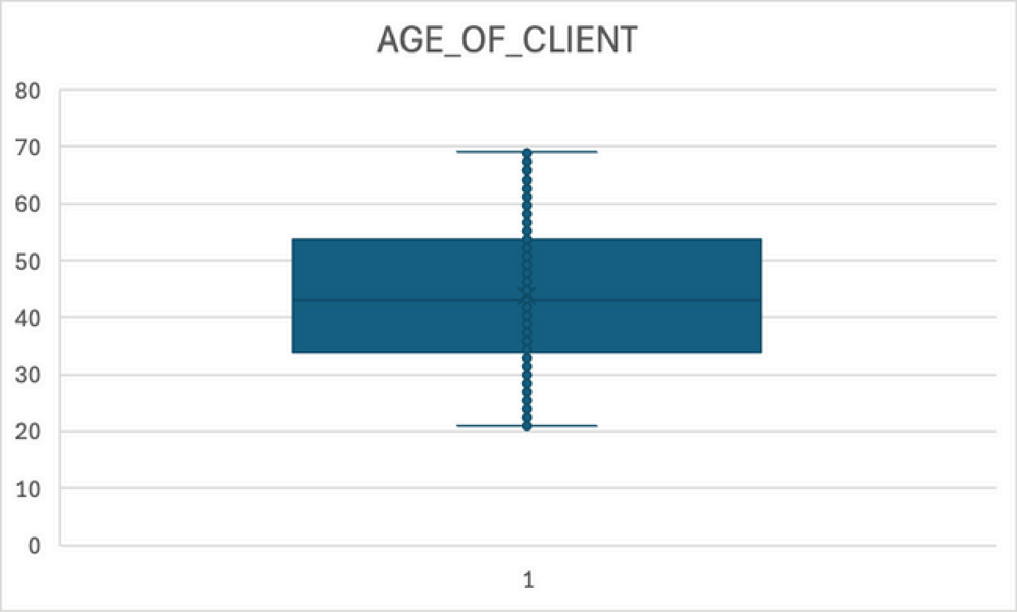

CHILDER_COUNT
12
10
8
6
4
2
0
1


AMT_INCOME
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1


AMT_CREDIT
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1


AGE_OF_CLIENT
80
70
60
50
40
30
20
10
0
1

DATA ANALYTICS TASKS:

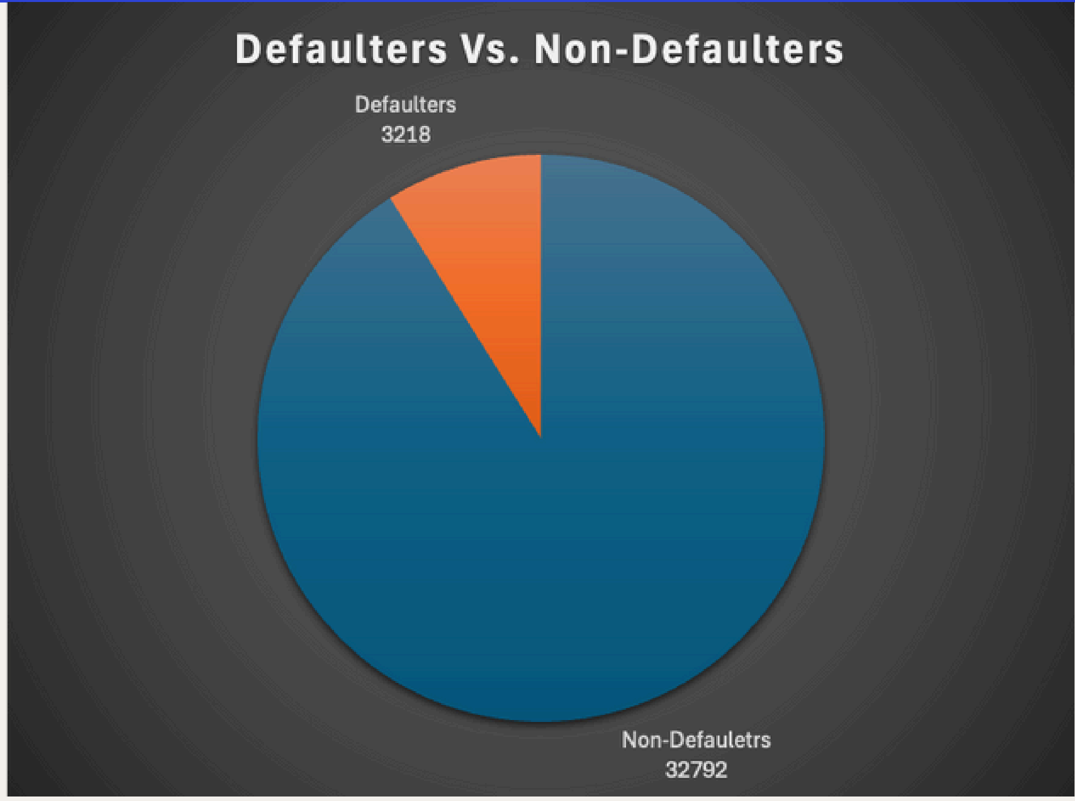
C. ANALYZE DATA IMBALANCE

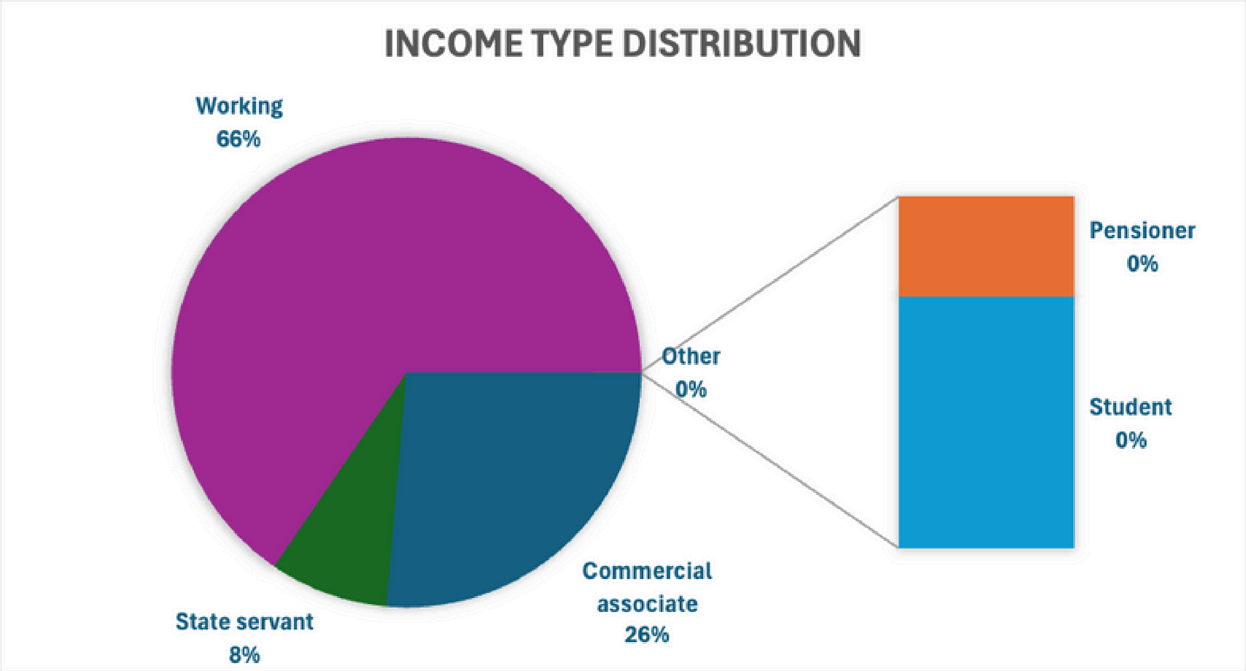
IMPACT OF IMBALANCED DATASETS ON DATA ANALYSIS (ESPECIALLY FOR BINARY CLASSIFICATION):

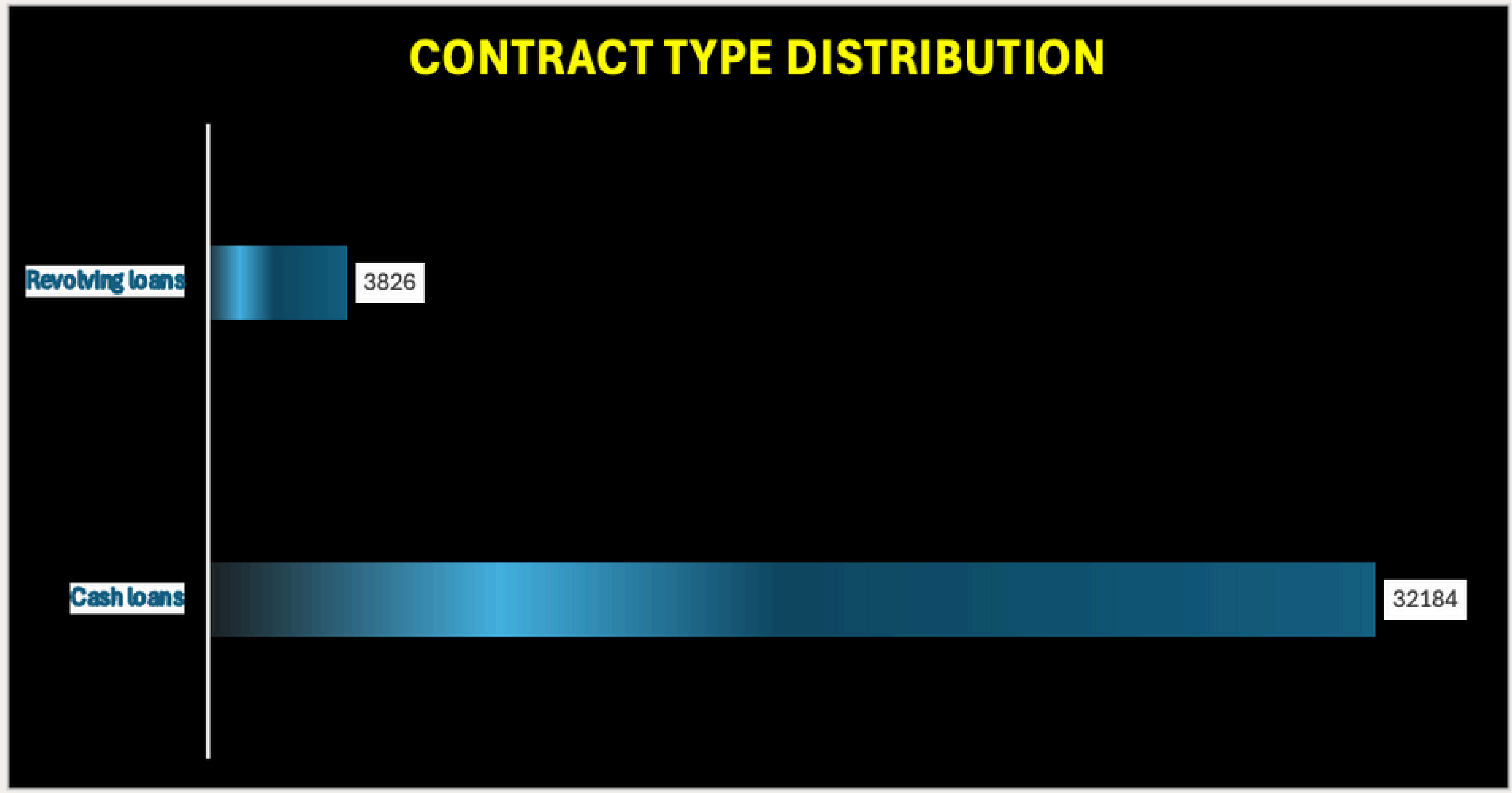
IMBALANCED DATASETS CAN SIGNIFICANTLY AFFECT THE ACCURACY AND RELIABILITY OF MODELS, PARTICULARLY IN BINARY CLASSIFICATION PROBLEMS. WHEN ONE CLASS DOMINATES THE DATASET, THE MODEL TENDS TO OVERFIT TO THE MAJORITY CLASS, FAILING TO RECOGNIZE THE MINORITY CLASS EFFECTIVELY. THIS CAN LEAD TO BIASED PREDICTIONS AND POOR PERFORMANCE IN IDENTIFYING IMPORTANT TRENDS, SUCH AS DEFAULTING CUSTOMERS IN A LOAN DATASET. IMBALANCE ALSO COMPLICATES THE EVALUATION PROCESS SINCE ACCURACY METRICS MAY LOOK FAVORABLE WHILE THE MODEL PERFORMS POORLY ON THE MINORITY CLASS. IN BUSINESS CONTEXTS, THIS CAN LEAD TO INCORRECT DECISIONS, MISALLOCATION OF RESOURCES, OR MISSED OPPORTUNITIES TO MITIGATE RISKS.

STEPS FOLLOWED TO IDENTIFY THE IMBALANCE IN DATASET:

1. VISUALIZING DATA DISTRIBUTION THROUGH PIE/BAR CHARTS ALLOWED US TO QUICKLY IDENTIFY POTENTIAL IMBALANCES IN KEY BUSINESS PARAMETERS SUCH AS TARGET (DEFAULT STATUS), GENDER, INCOME TYPE, AND LOAN CONTRACTS.
2. IDENTIFYING IMBALANCES IN THESE KEY FIELDS INDICATED POTENTIAL SKEWED INSIGHTS. FOR INSTANCE, 91% OF THE DATA IN THE TARGET FIELD REPRESENTS NON-DEFAULTERS, LIMITING OUR ABILITY TO IDENTIFY PATTERNS RELATED TO DEFAULTERS, WHICH ARE CRUCIAL FOR RISK MITIGATION.
3. THE GENDER IMBALANCE SHOWS 64% OF THE DATA REPRESENTS FEMALE CUSTOMERS, MAKING IT DIFFICULT TO CAPTURE TRENDS AND BEHAVIORS OF MALE LOAN APPLICANTS, POTENTIALLY LEADING TO BIASED INSIGHTS.
4. FOR THE LOAN CONTRACT TYPE, WE OBSERVED A SIGNIFICANT IMBALANCE, WITH 32,184 ROWS FOR CASH LOANS VERSUS ONLY 3,826 FOR REVOLVING LOANS. THIS COULD SUGGEST A BUSINESS TREND WHERE THE COMPANY ISSUES MORE CASH LOANS OR MAY HIGHLIGHT AN IMBALANCED DATASET THAT NEEDS ATTENTION.







DATA ANALYTICS TASKS:

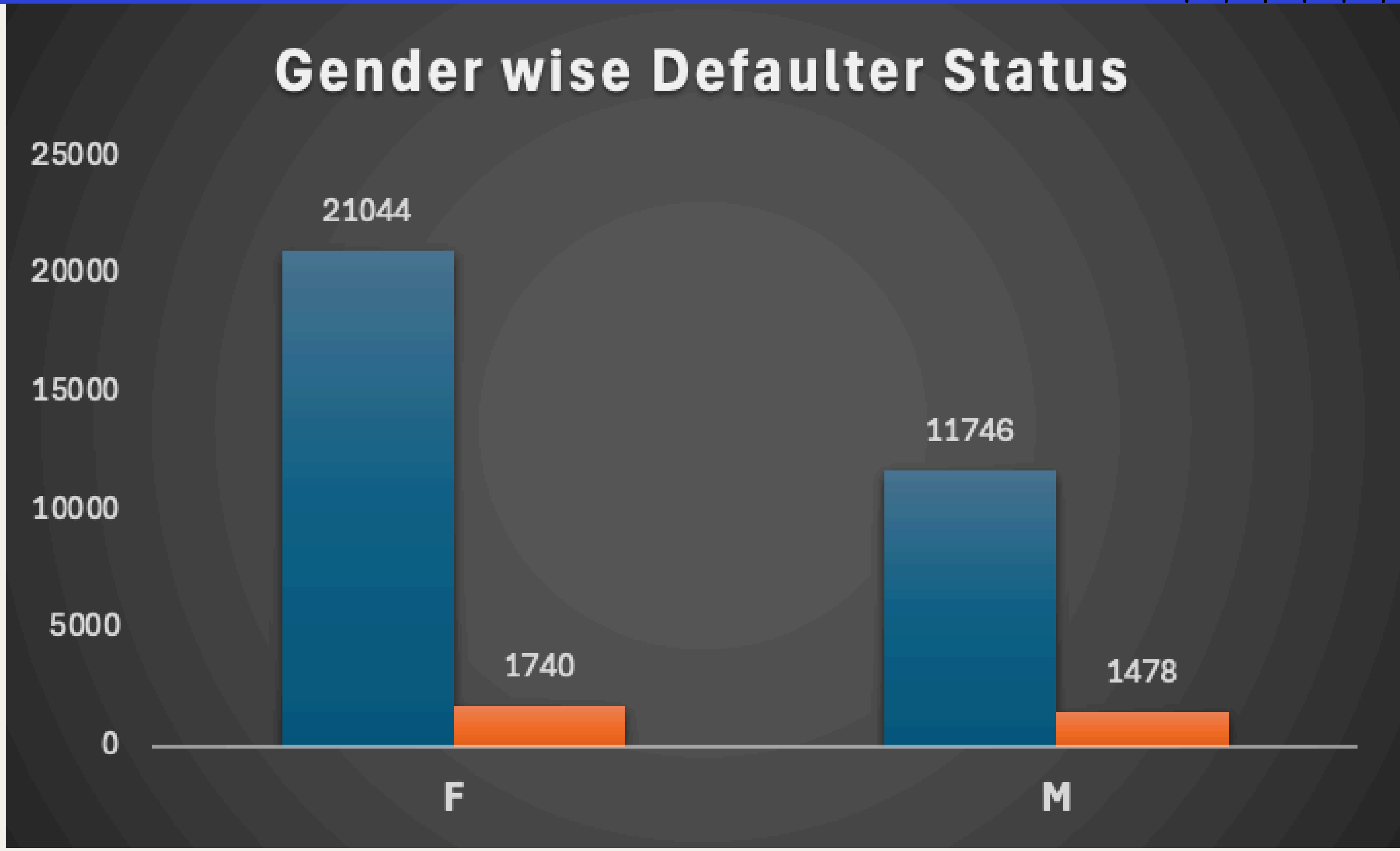
D. PERFORM UNIVARIATE, SEGMENTED UNIVARIATE, AND BIVARIATE ANALYSIS

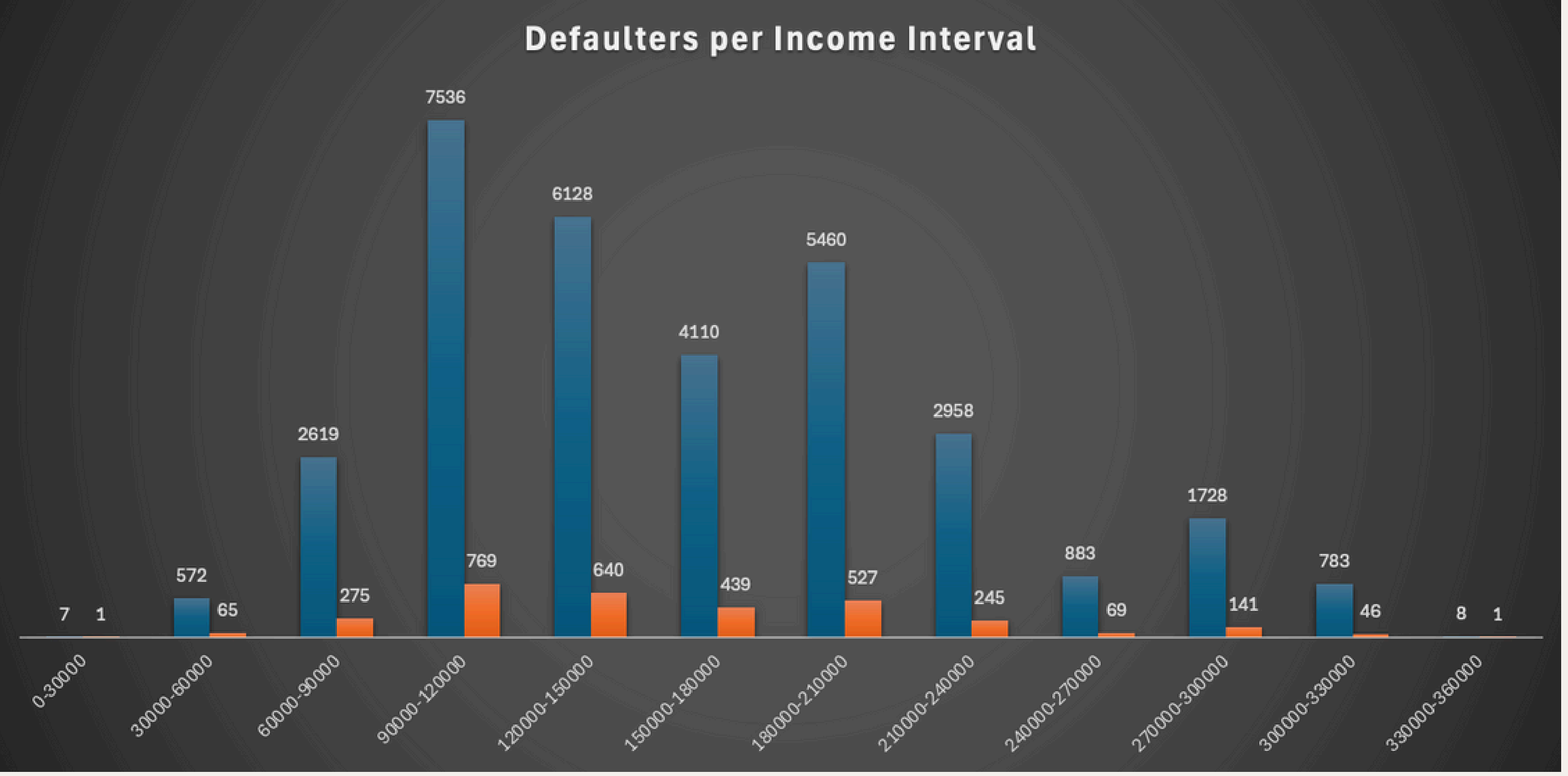
IMPORTANCE OF SEGMENTED UNIVARIATE AND BIVARIATE ANALYSIS:

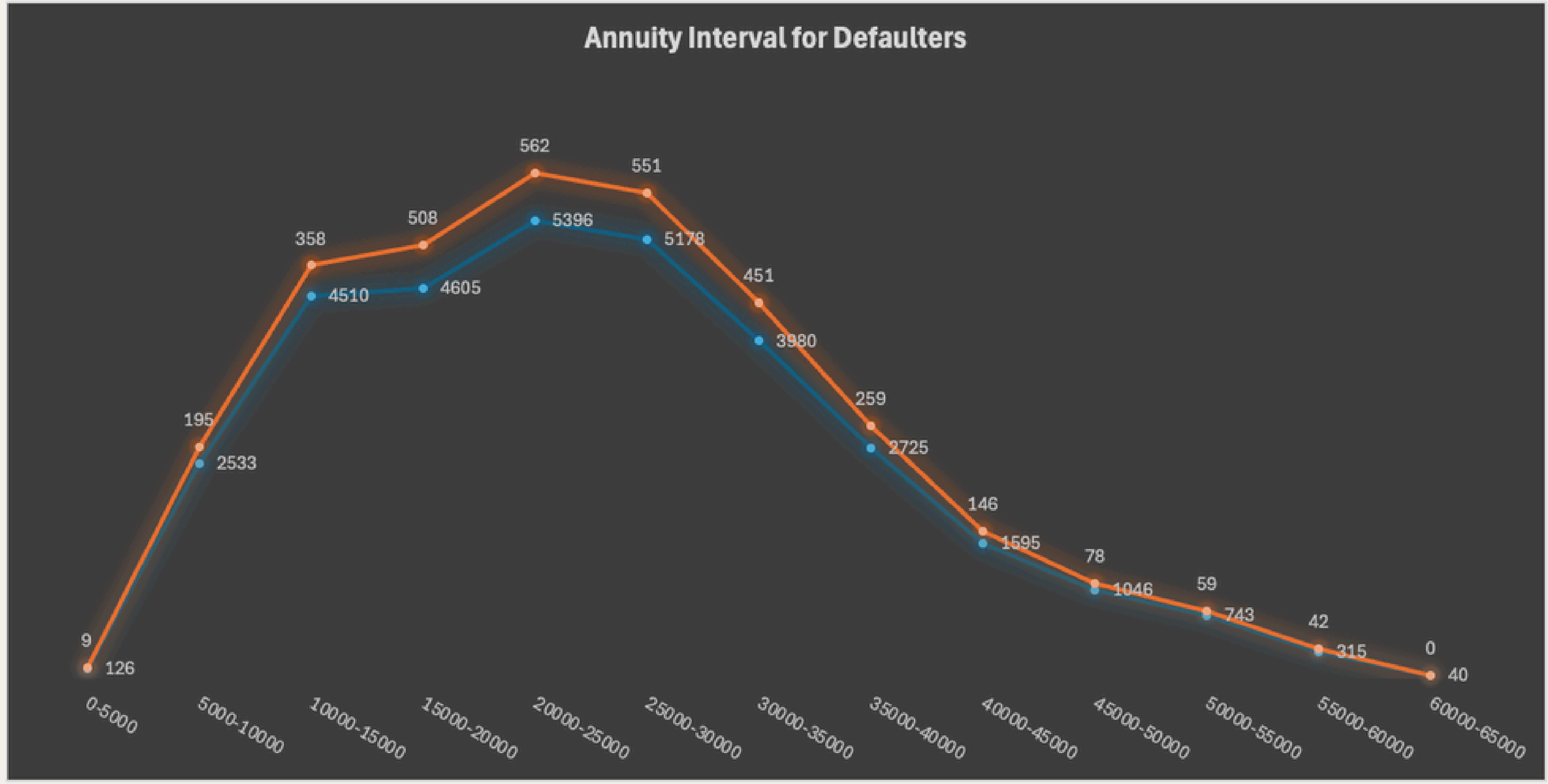
SEGMENTED UNIVARIATE ANALYSIS HELPS BUSINESSES BREAK DOWN INDIVIDUAL VARIABLES AND EXPLORE PATTERNS WITHIN DIFFERENT SEGMENTS, PROVIDING A GRANULAR UNDERSTANDING OF HOW SPECIFIC FACTORS LIKE INCOME OR FAMILY STATUS RELATE TO BUSINESS OUTCOMES SUCH AS LOAN DEFAULTS. IT ALLOWS FOR A MORE FOCUSED EXAMINATION OF KEY VARIABLES, REVEALING TRENDS THAT MAY OTHERWISE GO UNNOTICED. BIVARIATE ANALYSIS, ON THE OTHER HAND, EXAMINES RELATIONSHIPS BETWEEN TWO VARIABLES, ENABLING COMPANIES TO UNDERSTAND INTERACTIONS AND DEPENDENCIES, SUCH AS HOW A CUSTOMER'S INCOME RELATES TO CREDIT OR LOAN APPROVAL. THIS AIDS IN MORE ACCURATE PREDICTIONS, RISK ASSESSMENTS, AND INFORMED DECISION-MAKING.

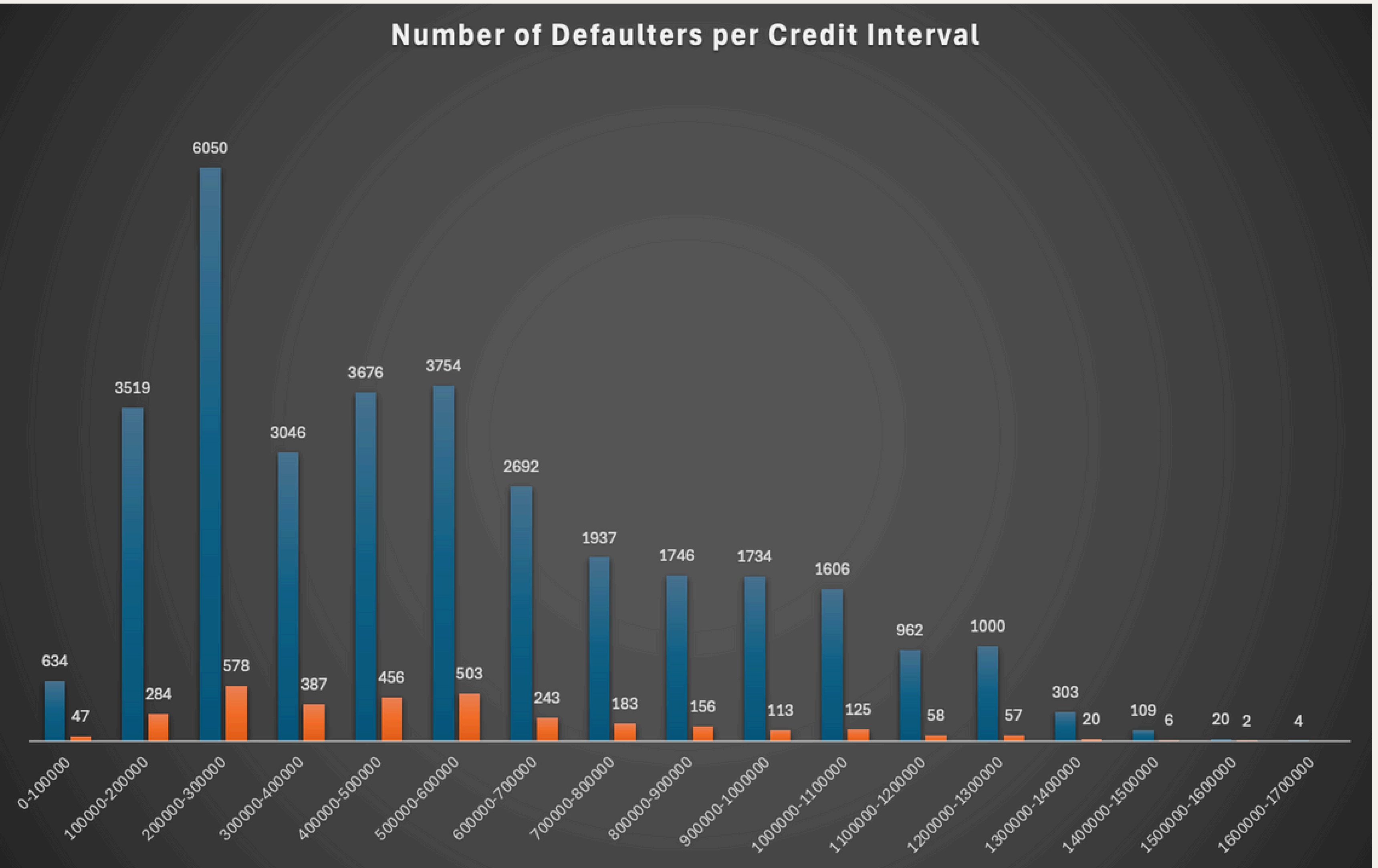
STEPS FOLLOWED TO PERFORM UNIVARIATE AND BIVARIATE ANALYSIS:

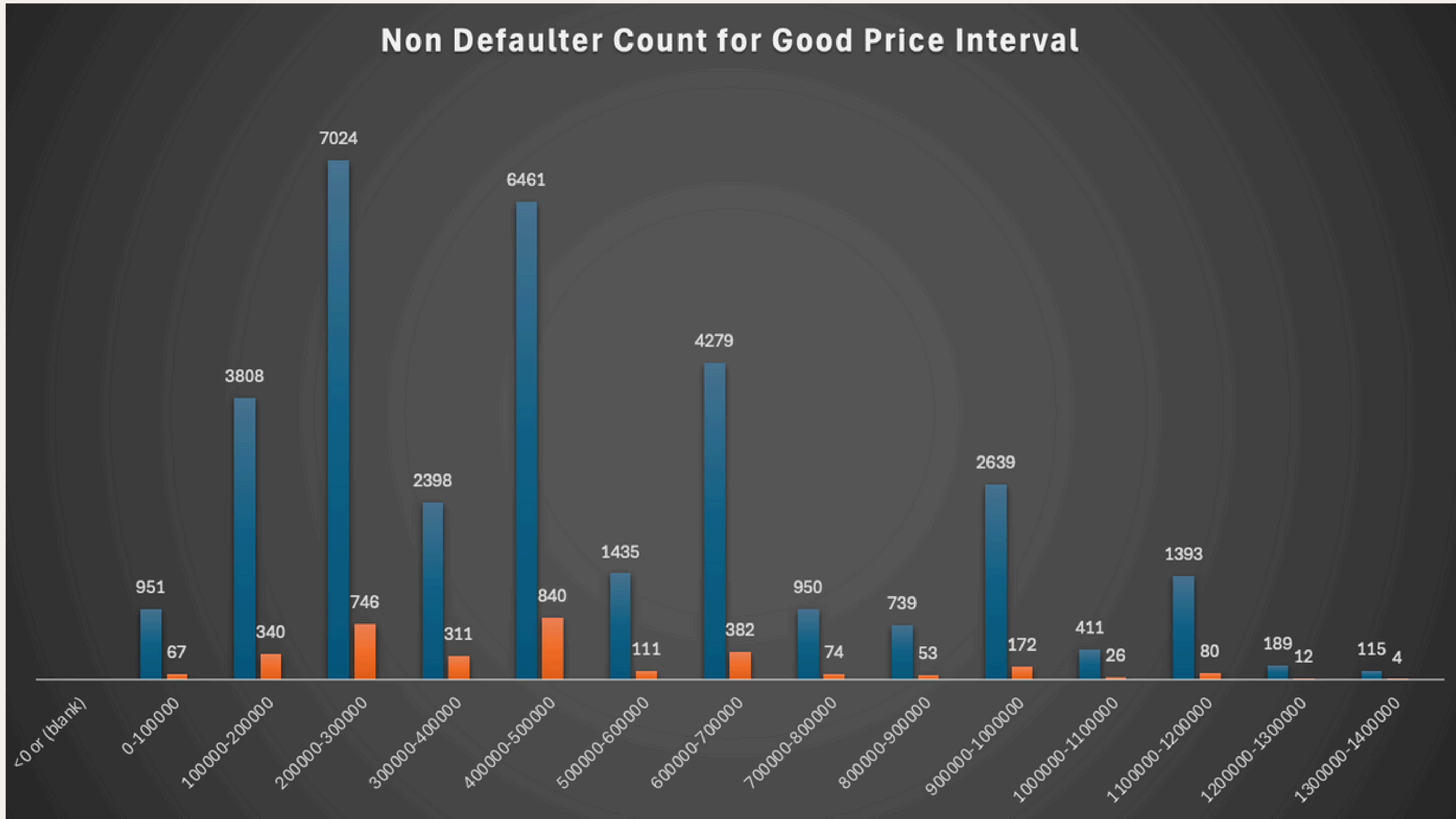
1. VISUALIZING DEFAULT TRENDS USING BAR, PIE, LINE, AND DOUGHNUT CHARTS PROVIDED INSIGHTS INTO HOW KEY FACTORS LIKE INCOME, CREDIT, ANNUITY, FAMILY STATUS, AND OCCUPATION INFLUENCE DEFAULT RISK, HELPING US PINPOINT HIGH-RISK CUSTOMER SEGMENTS.
2. FOR QUANTITATIVE PARAMETERS SUCH AS INCOME, CREDIT, AND ANNUITY, WE USED BAR AND LINE CHARTS TO CLEARLY DISTINGUISH THE DISTRIBUTION BETWEEN DEFAULTERS AND NON-DEFAULTERS, OFFERING A CLEAR PICTURE OF WHERE THE RISK LIES.
3. WE ANALYZED CUSTOMER FAMILY AND OCCUPATIONAL BACKGROUNDS USING VARIOUS CHART TYPES (PIE, BAR, DOUGHNUT) TO VISUALLY REPRESENT DEFAULTER COUNTS, UNCOVERING SOCIAL OR OCCUPATIONAL TRENDS THAT MIGHT AFFECT DEFAULT RISK.
4. IN THE BIVARIATE ANALYSIS, WE COMPARED CUSTOMER INCOME WITH AVERAGE CREDIT AND ANNUITY AMOUNTS, HELPING US OPTIMIZE LOAN APPROVAL AMOUNTS BASED ON CUSTOMER AFFORDABILITY.
5. FINALLY, WE EXAMINED THE RELATIONSHIP BETWEEN CREDIT GIVEN AND CUSTOMER WORK EXPERIENCE, PROVIDING INSIGHTS INTO HOW EXPERIENCE INFLUENCES CREDITWORTHINESS AND SHAPING MORE TAILORED CREDIT POLICIES.

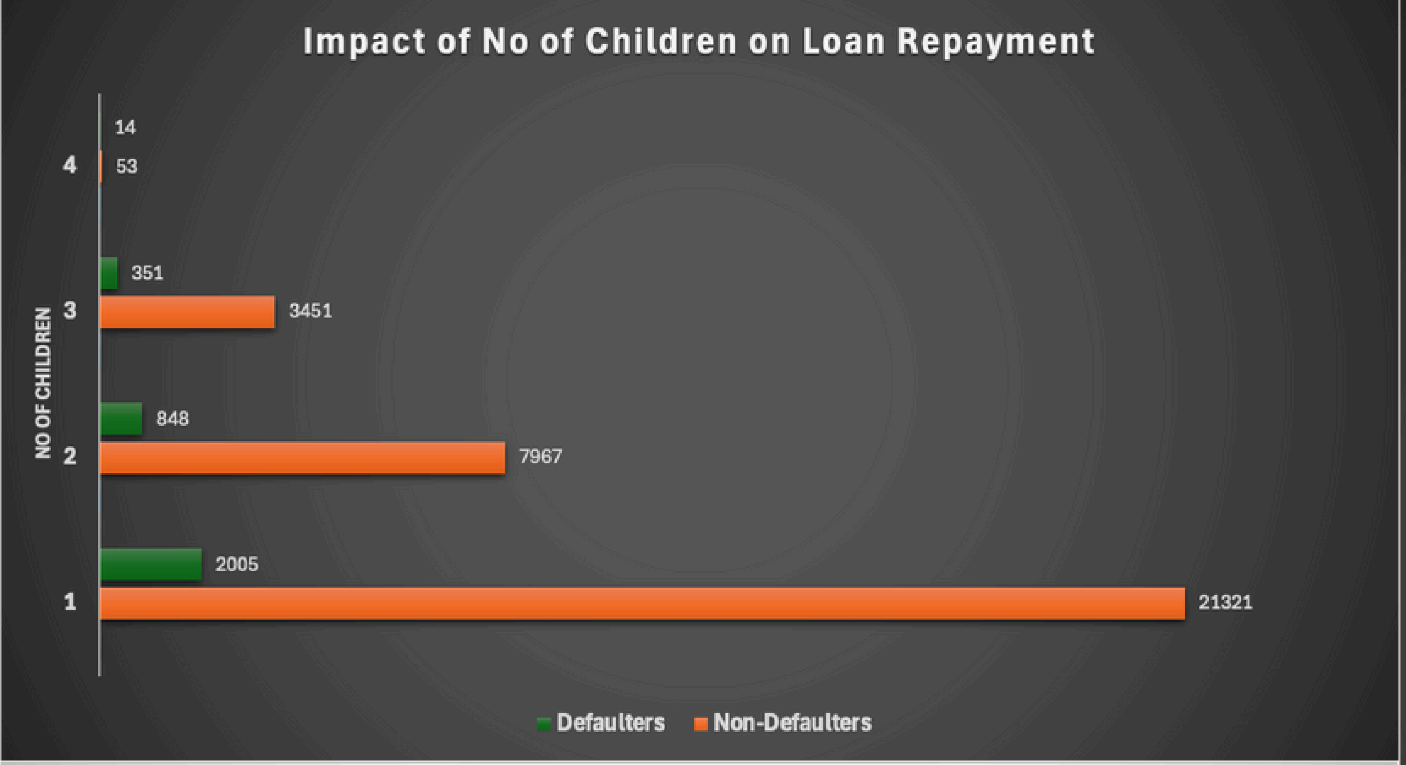


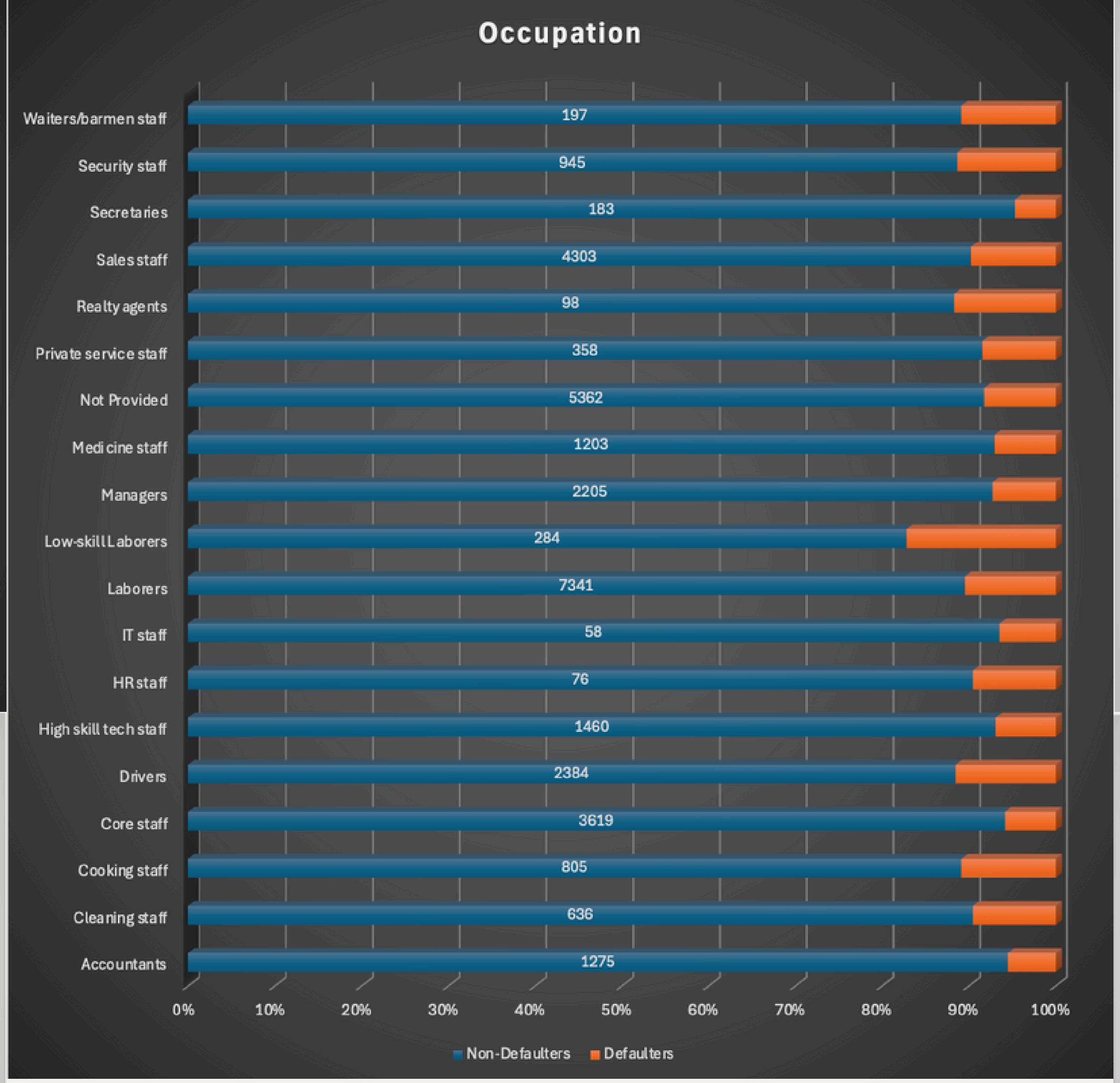

Defaulters per Income Interval
0-30000
7
1
30000-60000
572
65
60000-90000
2619
275
90000-120000
7536
769
120000-150000
6128
640
150000-180000
4110
439
180000-210000
5460
527
210000-240000
2958
245
240000-270000
883
69
270000-300000
1728
141
300000-330000
783
46
330000-360000
8
1

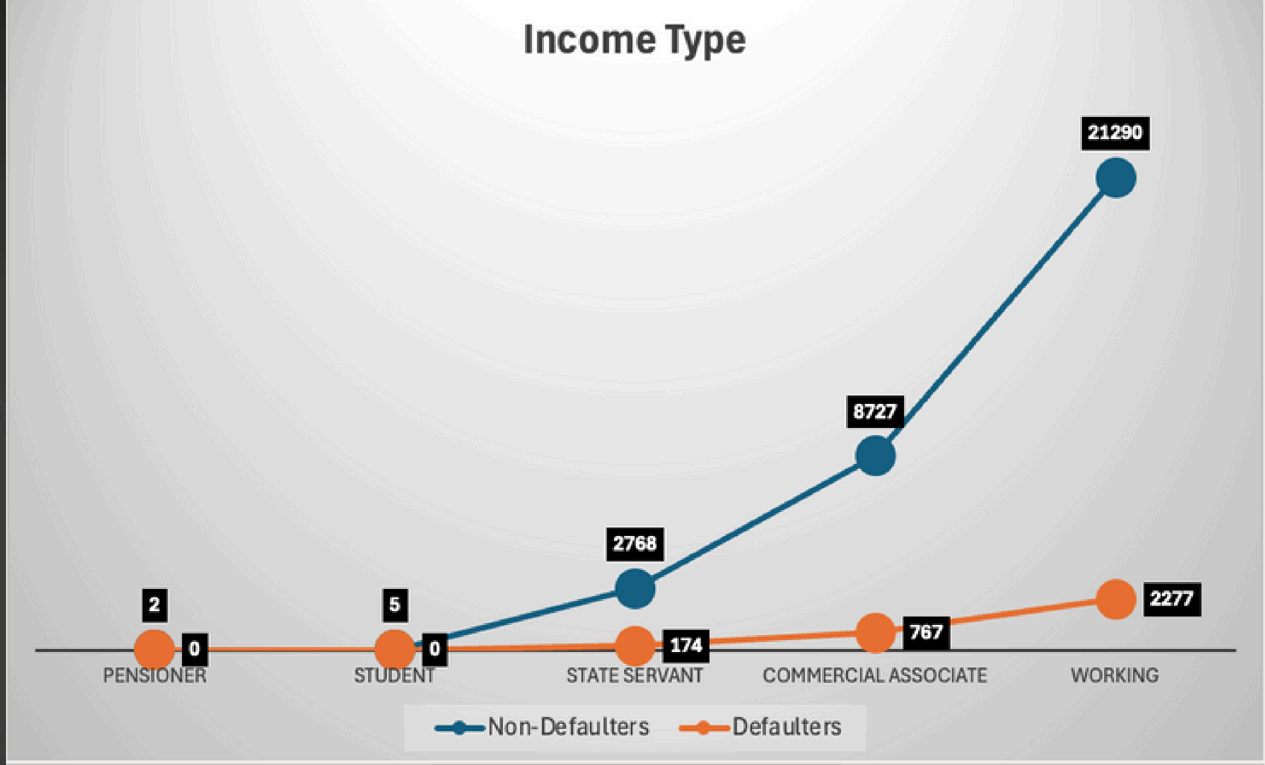

Annuity Interval for Defaulters
0-5000
9
126
5000-10000
195
2533
10000-15000
358
4510
15000-20000
508
4605
20000-25000
562
5396
25000-30000
551
5178
30000-35000
451
3980
35000-40000
259
2725
40000-45000
146
1595
45000-50000
78
1046
50000-55000
59
743
55000-60000
42
315
60000-65000
0
40

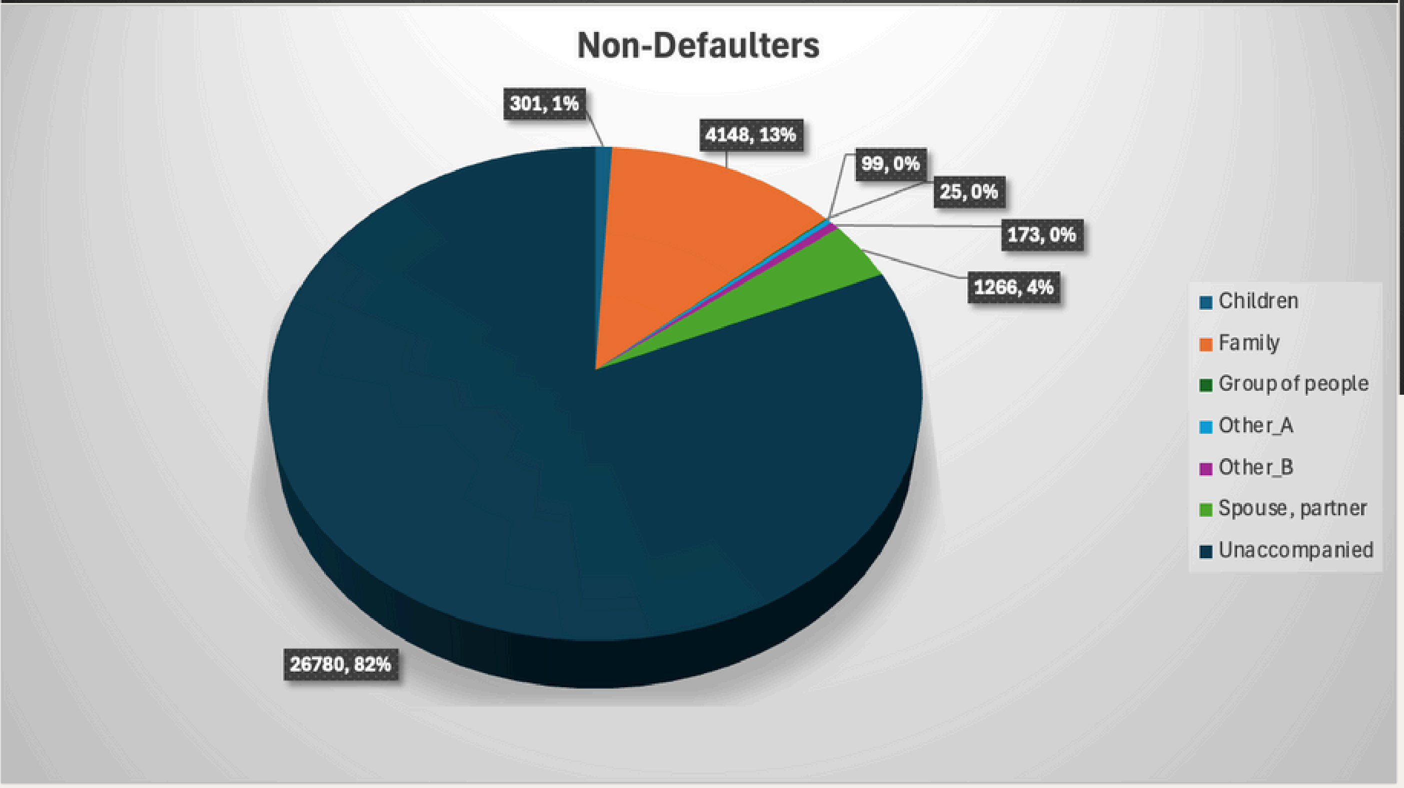

Number of Defaulters per Credit Interval
0-100000
634
47
100000-200000
3519
284
200000-300000
6050
578
300000-400000
3046
387
400000-500000
3676
456
500000-600000
3754
503
600000-700000
2692
243
700000-800000
1937
183
800000-900000
1746
156
900000-1000000
1734
113
1000000-1100000
1606
125
1100000-1200000
962
58
1200000-1300000
1000
57
1300000-1400000
303
20
1400000-1500000
109
6
1500000-1600000
20
2
1600000-1700000
4

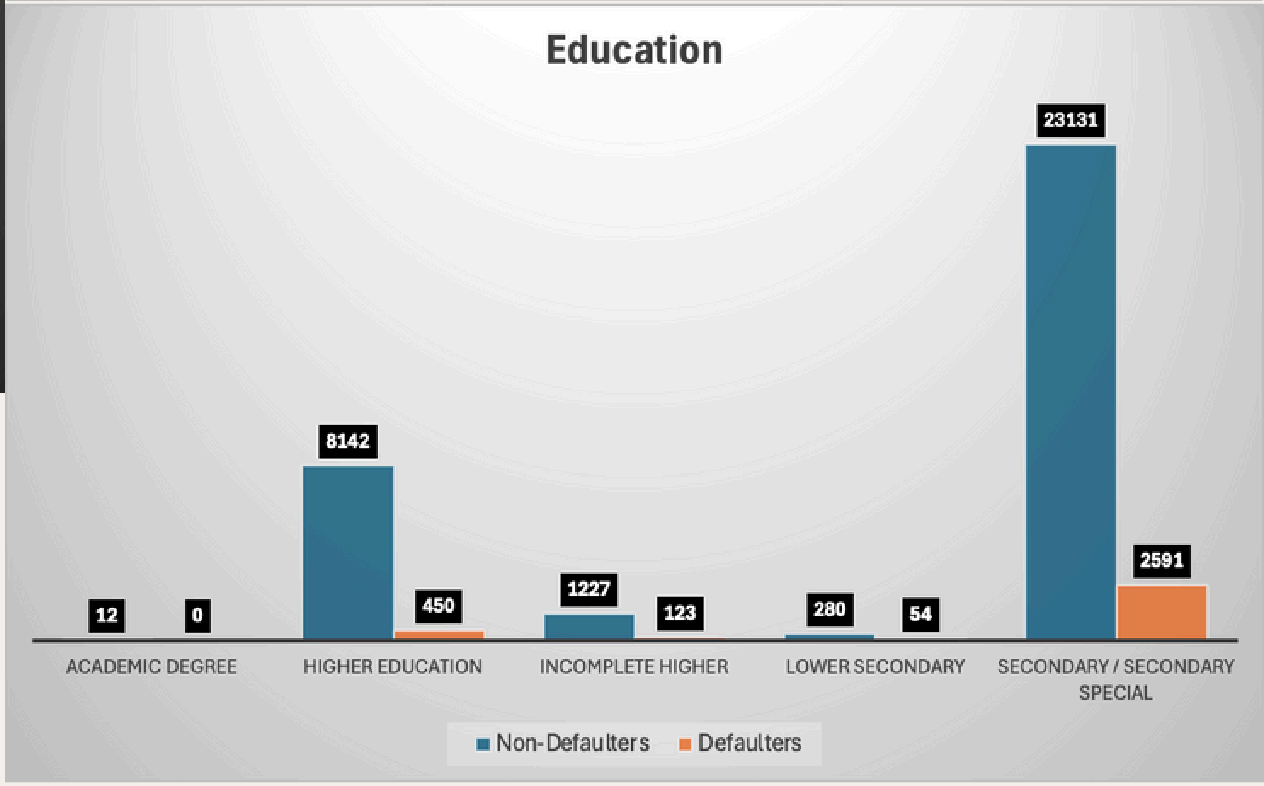

Non Defaulter Count for Good Price Interval
<0 or (blank)
0-100000
951
67
100000-200000
3808
340
200000-300000
7024
746
300000-400000
2398
311
400000-500000
6461
840
500000-600000
1435
111
600000-700000
4279
382
700000-800000
950
74
800000-900000
739
53
900000-1000000
2639
172
1000000-1100000
411
26
1100000-1200000
1393
80
1200000-1300000
189
12
1300000-1400000
115
4

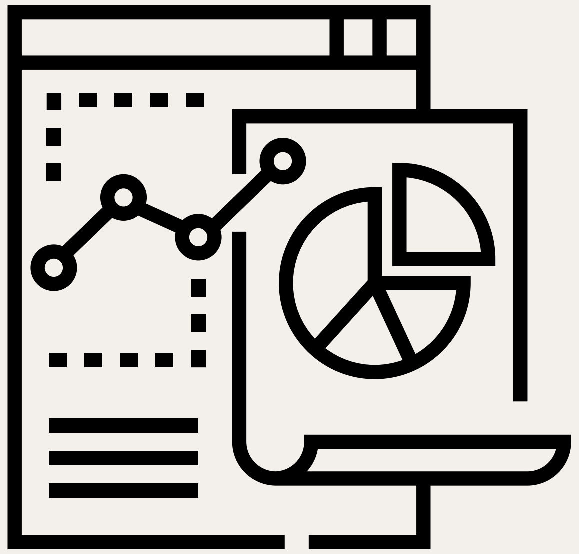

Impact of No of Children on Loan Repayment
NO OF CHILDREN
4
14
53
3
351
3451
2
848
7967
1
2005
21321
Defaulters
Non-Defaulters

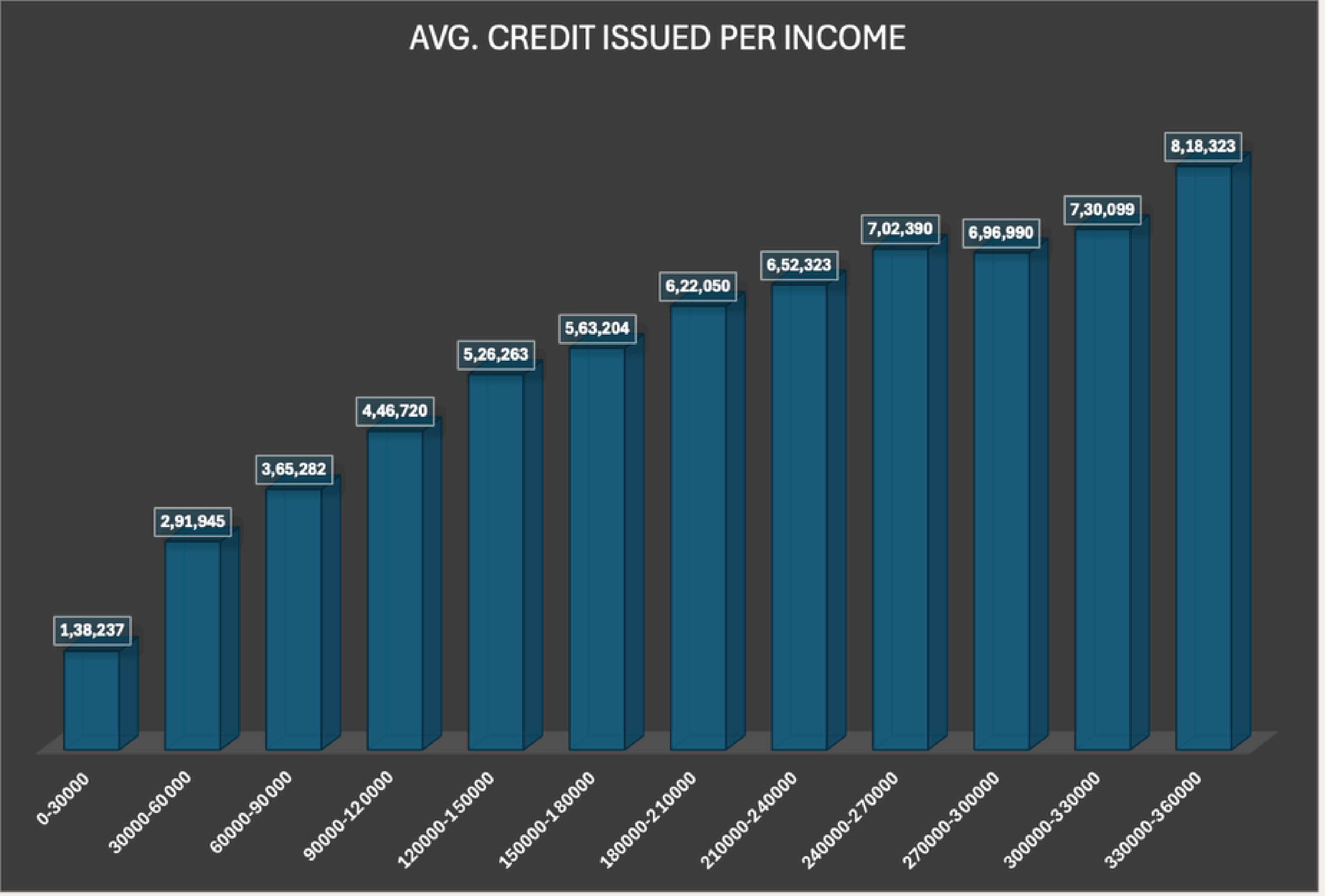

Occupation
Waiters/barmen staff
197
Security staff
945
Secretaries
183
Sales staff
4303
Realty agents
98
Private service staff
358
Not Provided
5362
Medicine staff
1203
Managers
2205
Low-skill Laborers
284
Laborers
7341
IT staff
58
HR staff
76
High skill tech staff
1460
Drivers
2384
Core staff
3619
Cooking staff
805
Cleaning staff
636
Accountants
1275
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Non-Defaulters
Defaulters

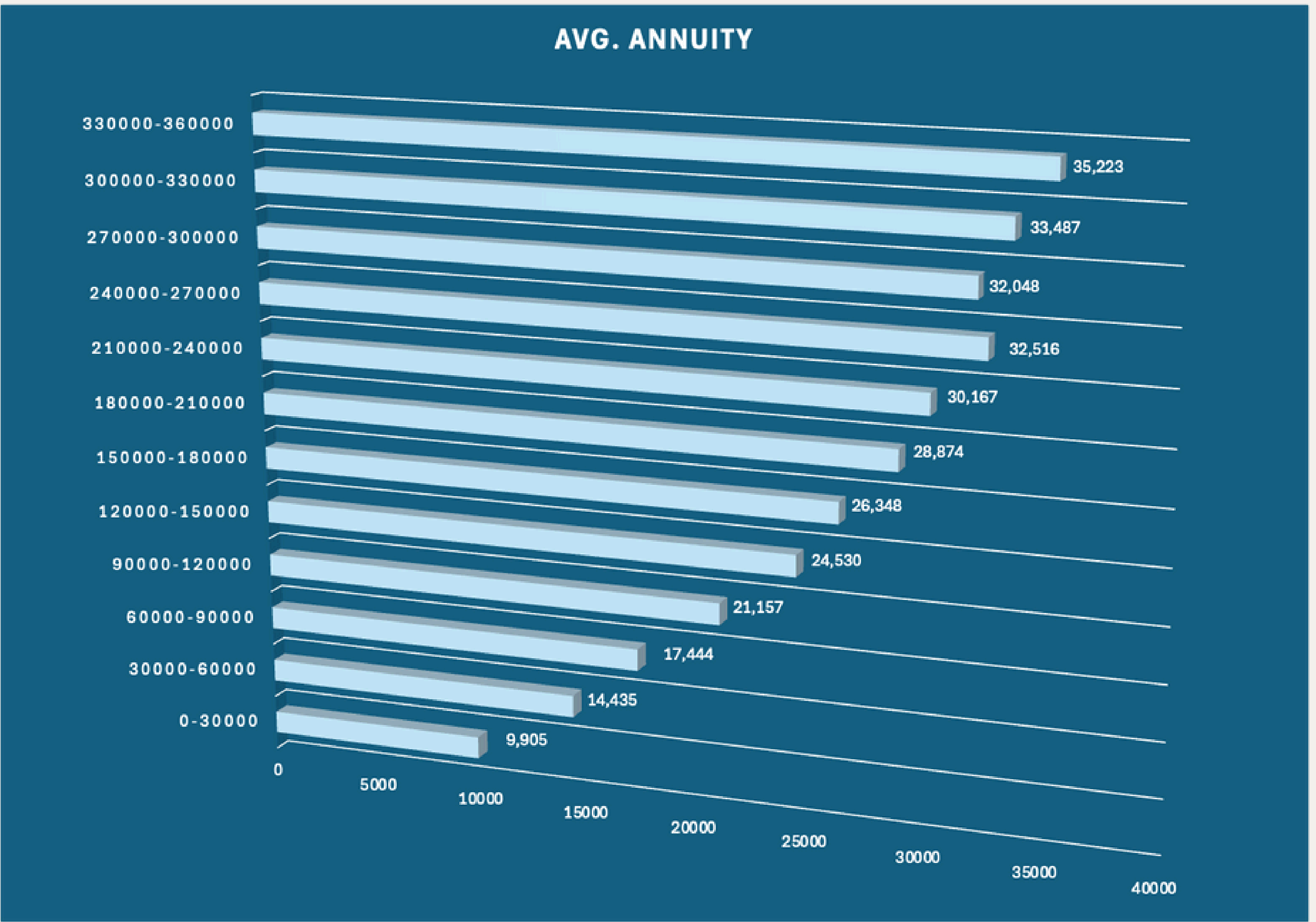

Income Type
2
0
PENSIONER
5
0
STUDENT
2768
174
STATE SERVANT
8727
767
COMMERCIAL ASSOCIATE
21290
2277
WORKING
Non-Defaulters
Defaulters

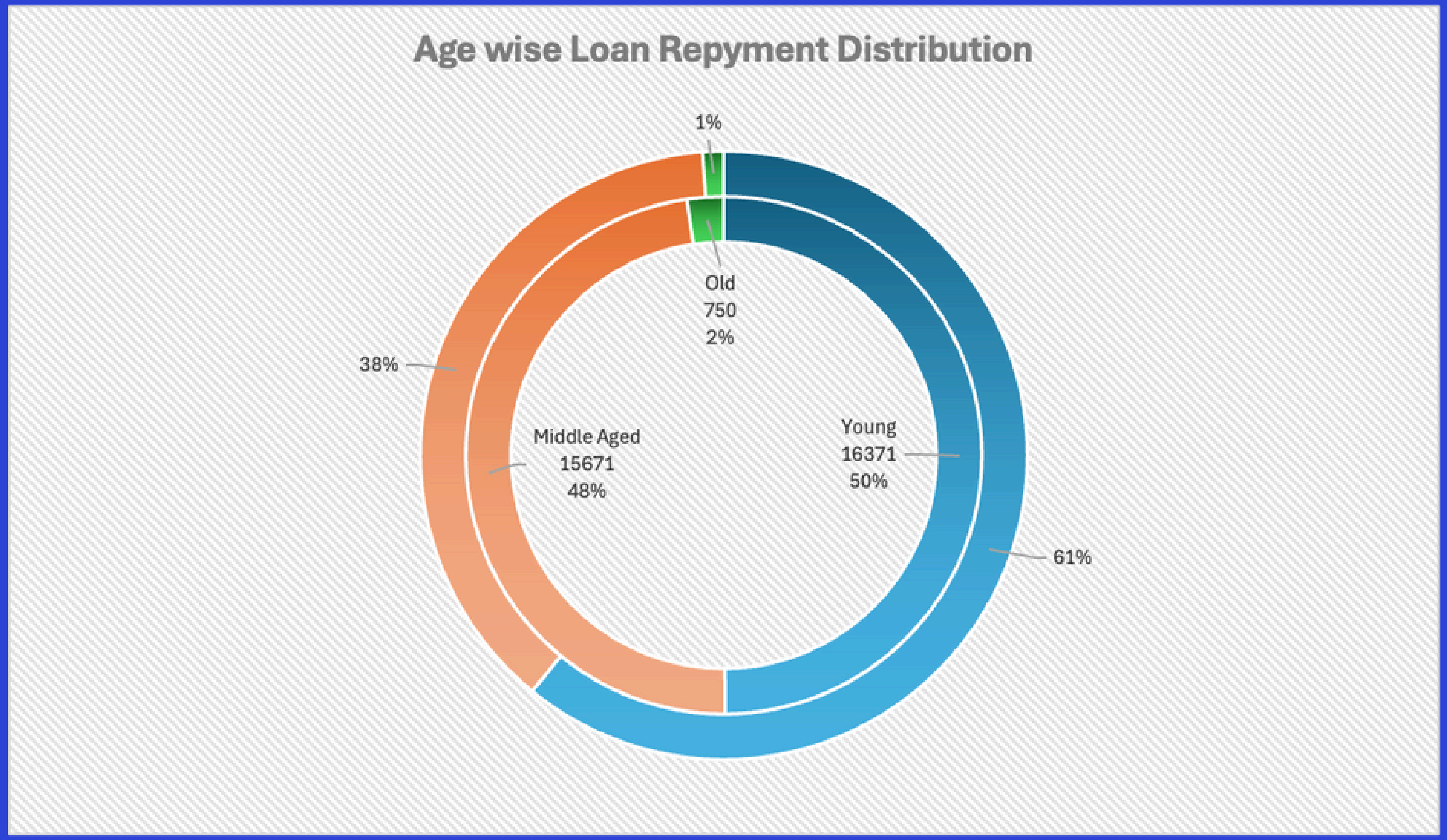

Non-Defaulters
301, 1%
4148, 13%
99, 0%
25, 0%
173, 0%
1266, 4%
26780, 82%
Children
Family
Group of people
Other_A
Other_B
Spouse, partner
Unaccompanied


Education
12
0
ACADEMIC DEGREE
8142
450
HIGHER EDUCATION
1227
123
INCOMPLETE HIGHER
280
54
LOWER SECONDARY
23131
2591
SECONDARY / SECONDARY SPECIAL
Non-Defaulters
Defaulters




AVG. CREDIT ISSUED PER INCOME
1,38,237
2,91,945
3,65,282
4,46,720
5,26,263
5,63,204
6,22,050
6,52,323
7,02,390
6,96,990
7,30,099
8,18,323
0-30000
30000-60000
60000-90000
90000-120000
120000-150000
150000-180000
180000-210000
210000-240000
240000-270000
270000-300000
300000-330000
330000-360000


AVG. ANNUITY
330000-360000
35,223
300000-330000
33,487
270000-300000
32,048
240000-270000
32,516
210000-240000
30,167
180000-210000
28,874
150000-180000
26,348
120000-150000
24,530
90000-120000
21,157
60000-90000
17,444
30000-60000
14,435
0-30000
9,905
0
5000
10000
15000
20000
25000
30000
35000
40000


Age wise Loan Repyment Distribution
1%
Old
750
2%
38%
Middle Aged
15671
48%
Young
16371
50%
61%

DATA ANALYTICS TASKS:

E. IDENTIFY TOP CORRELATIONS FOR DIFFERENT SCENARIOS

IMPORTANCE OF CORRELATION IN DATA ANALYSIS:

CORRELATION BETWEEN VARIABLES IS CRUCIAL IN DATA ANALYSIS BECAUSE IT HELPS IDENTIFY RELATIONSHIPS AND DEPENDENCIES BETWEEN DIFFERENT FACTORS. UNDERSTANDING THESE CORRELATIONS ALLOWS BUSINESSES TO PINPOINT KEY DRIVERS AND HOW THEY INFLUENCE OUTCOMES. FOR EXAMPLE, IN FINANCIAL ANALYSIS, HIGH CORRELATIONS BETWEEN INCOME, CREDIT SCORE, AND LOAN REPAYMENT BEHAVIOR CAN REVEAL CUSTOMER PROFILES THAT ARE MORE LIKELY TO DEFAULT. ADDITIONALLY, STRONG CORRELATIONS HELP IN PREDICTIVE MODELING BY FOCUSING ON THE MOST IMPACTFUL VARIABLES, REDUCING NOISE FROM IRRELEVANT DATA, AND IMPROVING DECISION-MAKING ACCURACY.

STEPS TO ANALYSE THE CORRELATION:

1. I UTILIZED THE 'CORREL' FUNCTION TO IDENTIFY RELATIONSHIPS BETWEEN KEY FINANCIAL PARAMETERS, HELPING UNCOVER PATTERNS THAT DIRECTLY IMPACT CUSTOMER BEHAVIOR.
2. THE DATASET WAS SEGMENTED INTO DEFAULTERS AND NON-DEFAULTERS, WITH NON-CRITICAL FIELDS REMOVED TO ENHANCE THE PRECISION OF THE ANALYSIS AND FOCUS ON IMPORTANT FACTORS.
3. FOR THIS ANALYSIS, I USED A 0.70 CORRELATION THRESHOLD TO IDENTIFY FIELDS WHERE STRONG RELATIONSHIPS EXIST, AS THOSE PARAMETERS ARE LIKELY TO INFLUENCE A CUSTOMER'S LOAN REPAYMENT ABILITY.

Non Defaulters

Field_1	Field_2	Correlation
AMT_ANNUITY	AMT_CREDIT	0.750801772
AMT_GOODS_PRICE	AMT_CREDIT	0.981088097
AMT_GOODS_PRIC	AMT_ANNUITY	0.749525827
CNT_FAM_MEMBERS	CNT_CHILDREN	0.874975254
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.855252665
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.813217826
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.856538392
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.997037494
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951686246

Defaulters

Field_1	Field_2	Correlation
AMT_ANNUITY	AMT_CREDIT	0.731387145
AMT_GOODS_PRICE	AMT_CREDIT	0.977083611
AMT_GOODS_PRICE	AMT_ANNUITY	0.732738761
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878348798
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.888812352
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.766677989
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.797983006
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.997446815
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949217866

INSIGHTS

- The majority of approved credit limits fall within the range of $200,000 to $300,000, making this the optimal bracket for evaluating loan risk and customer affordability.
- Married individuals demonstrate higher likelihood of loan acceptance and repayment, indicating a stable demographic for financial institutions to target.
- Customers with fewer dependents are considered more financially stable, reducing the risk of loan default and making them attractive borrowers.
- Lower levels of education correlate with higher loan default rates, suggesting a need for stricter credit assessments for applicants with limited educational backgrounds.
- Younger individuals are more likely to apply for and secure loans, presenting a growing customer segment that financial institutions should monitor and cater to.
- Low-skilled workers tend to struggle with loan repayment, highlighting the importance of tailored lending criteria or risk management strategies for this demographic.

RESULT

- This project significantly enhanced my Exploratory Data Analysis (EDA) skills, allowing me to derive meaningful insights from complex datasets and improve decision-making processes.
- I gained hands-on experience in handling large datasets with missing values, applying various imputation techniques and ensuring data integrity for accurate analysis.
- Researching banking and financial services prior to the project provided a deeper understanding of the financial landscape, enabling me to better interpret the dataset and draw relevant business insights.
- The project improved my data visualization abilities, using tools like bar charts, pie charts, and histograms to effectively communicate trends and findings to stakeholders.
- I developed a strong command of advanced Excel functions, such as VLOOKUP, pivot tables, and data analysis tools, which will be invaluable in future data-driven projects.

THANK YOU!