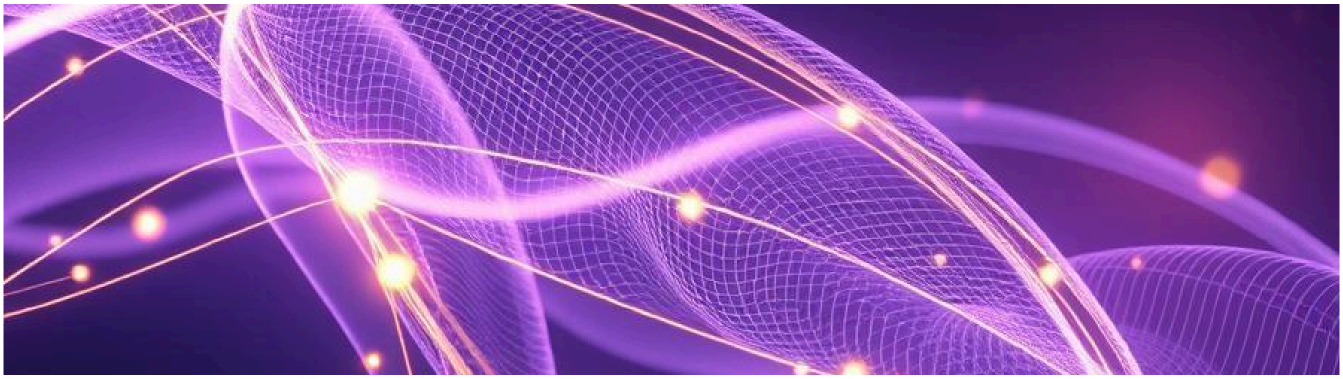


Bank Loan Default Risk Analysis

Case Study

01. Project Description

The goal of this case study is to perform a comprehensive Exploratory Data Analysis (EDA) on a loan application dataset to identify risk patterns and enhance the decision-making process for loan approvals. Lending institutions need a strong risk assessment framework to ensure applicants who are likely to repay are approved, while minimizing financial exposure from likely defaulters.

The dataset contains various applicant attributes such as income, employment history, credit history, loan amount, and demographics. These variables help assess whether an applicant belongs to the high-risk or low-risk segment.

Business Objectives

- Improve lending decisions by identifying high-risk applicants.
- Use data to develop strategies that optimize loan approvals.
- Minimize the rejection of creditworthy customers.
- Identify trends and features most correlated with default risks.

Approach

1. Identify and handle missing data
2. Detect and analyze outliers
3. Check for data imbalance
4. Perform univariate, segmented univariate, and bivariate analysis
5. Identify top correlations for different scenarios
6. Clean and prepare final dataset for insights

Tech Stack Used

- **Microsoft Excel 2022**
  - Functions: `IF`, `ISBLANK`, `COUNT`, `COUNTIF`, `CORREL`, `QUARTILE`, `MEDIAN`
  - Charts: Bar, Pie, Scatter, Box Plot, Heatmap
  - Features: Pivot Tables, Conditional Formatting, Filters
- **Google Docs** for creating a full report.

Result

- Cleaned and analyzed full dataset.
- Identified top risk indicators.
- Provided strategic insights for loan approval processes.

Google Drive Link

[Click for Google Drive Link Here]

Contains:

- Excel workbook with analysis and visualizations

Project Tasks -

---

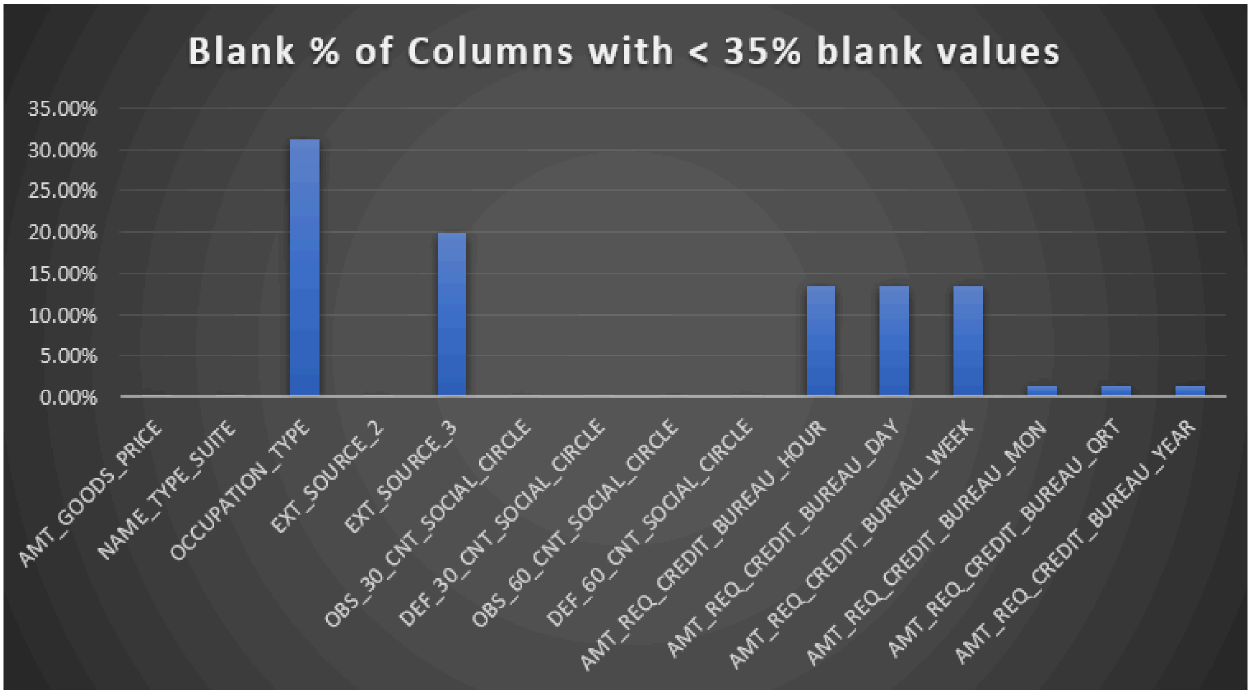
1. Identify Missing Data and Deal with It Appropriately

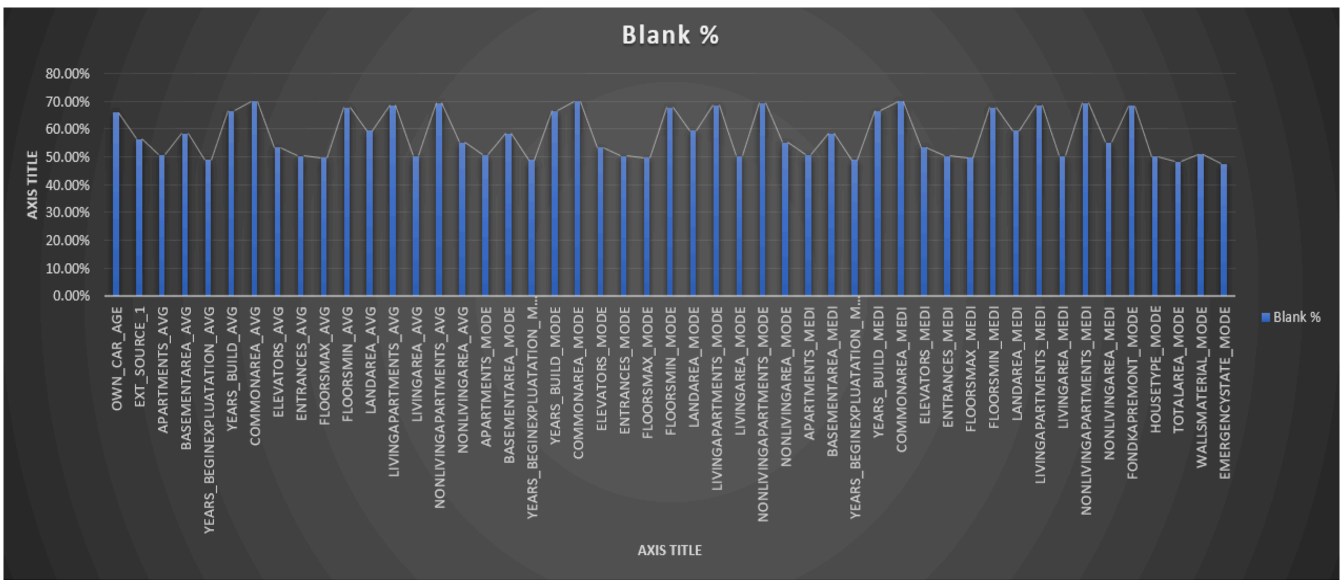
Objective: Handle incomplete records to ensure the analysis remains unbiased and reliable.

Methodology:

- **Detection:** Used Excel functions such as `ISBLANK()`, `COUNTBLANK()`, and `IF()` to identify columns and rows with missing values.
- **Imputation Strategy:**
  - For numerical variables: Used `MEDIAN()` imputation to reduce the influence of outliers.
  - For categorical variables: Used the most frequent category where appropriate.
- **Outcome:** Reduced data sparsity and ensured complete cases for further analysis.

Visualization:





---

2. Identify Outliers in the Dataset

Objective: Detect extreme values that can skew the results of statistical analysis.

Methodology:

- Applied the Interquartile Range (IQR) method:
  - Q1 = `=QUARTILE.EXC(A2:A50000,1)`

- Q3 = `=QUARTILE.EXC(A2:A50000,3)`
- INTER QUARTILE= Q3-Q1
- Upper Limit= Q3+(1.5*IQR)
- Lower Limit=Q1-(1.5*IQR)

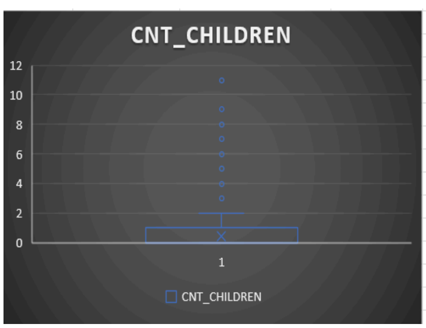
- Used Conditional Formatting to highlight outliers for quick visual identification.

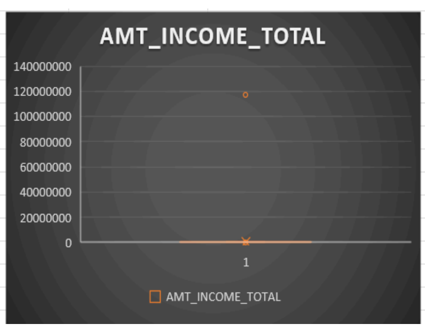
- Focused particularly on fields like `CNT_CHILDREN,AMT_INCOME_TOTAL,`

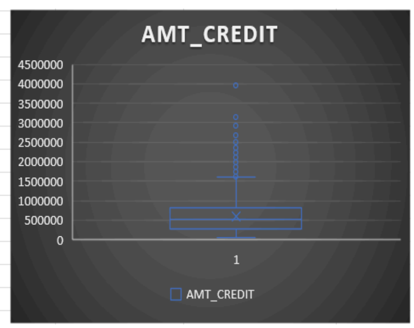
`AMT_GOODS_PRICE,YEARS_BIRTH etc`

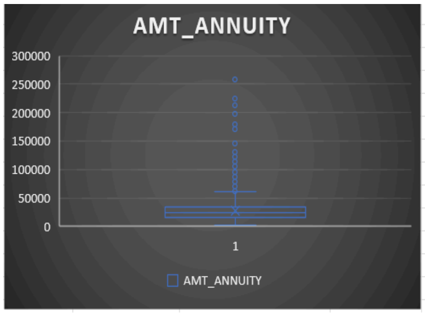
Visualization: Box plots and scatter plots of numeric variables.

Columns	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION
Q1	0	112500	270000	16456.5	238500	0.010006	33.91232877	2.556164384	5.473972603
Q3	1	202500	808650	34596	679500	0.028663	53.81917808	15.66575342	20.44931507
IQR	1	90000	538650	18139.5	441000	0.018657	19.90684932	13.10958904	14.97534247
Upper Limit	2.5	337500	1616625	61805.25	1341000	0.0566485	83.67945205	35.33013699	42.91232877
Lower Limit	-1.5	-22500	-537975	-10752.75	-423000	-0.0179795	4.052054795	-17.10821918	-16.9890411
Discriptive analysis									
Mean	0.419848397	170767.5905	599700.5815	27107.33399	538992.3491	0.020798283	43.8960057	184.0008887	13.63639086
Standard Error	0.003238031	2378.391081	1799.674528	65.12748183	1653.458318	6.15398E-05	0.053438274	1.702588554	0.043196956
Median	0	145800	514777.5	24939	450000	0.01885	43.09863014	6.071232877	12.30136986
Mode	0	135000	450000	9000	450000	0.035792	36.79178082	1000.665753	0.008219178
Standard Deviation	0.724038548	531819.0951	402415.4339	14562.80203	369720.8225	0.013760581	11.94904184	380.7065674	9.659036452
Sample Variance	0.524231818	2.82832E+11	1.61938E+11	212075202.9	1.36693E+11	0.000189354	142.7796008	144937.4905	93.29698519
Kurtosis	4.673335403	46582.52582	1.917459058	9.412285897	2.491524273	3.267863428	-1.04298699	0.818713082	-0.304458318
Skewness	1.877689555	212.0777967	1.223668739	1.688550535	1.348777751	1.48358065	0.118848404	1.678426236	0.59916788
Range	11	116974350	4005000	255973.5	4005000	0.071975	47.95616438	1000.665753	61.34794521
Minimum	0	25650	45000	2052	45000	0.000533	21.04109589	0	0
Maximum	11	117000000	4050000	258025.5	4050000	0.072508	68.99726027	1000.665753	61.34794521
Sum	20992	8538208758	29984429376	1355339592	26949078464	1039.893349	2194756.389	9199860.436	681805.9068
Count	49999	49999	49999	49999	49999	49999	49999	49999	49999

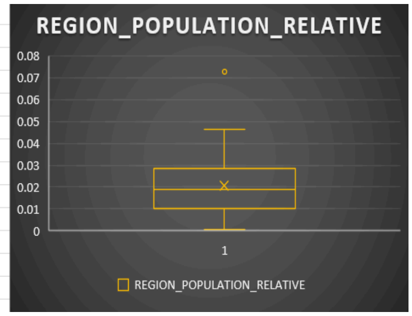


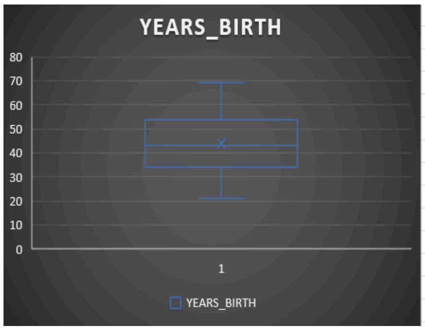




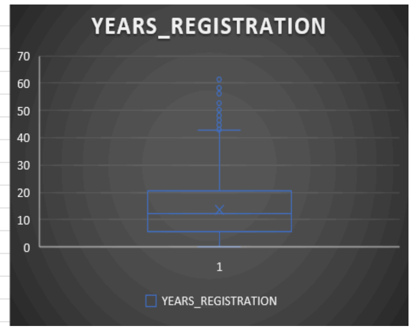

AMT_ANNUITY
300000
250000
200000
150000
100000
50000
0
1
AMT_ANNUITY


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


REGION_POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1
REGION_POPULATION_RELATIVE


YEARS_BIRTH
80
70
60
50
40
30
20
10
0
1
YEARS_BIRTH


YEARS_EMPLOYED
1200
1000
800
600
400
200
0
1
YEARS_EMPLOYED


YEARS_REGISTRATION
70
60
50
40
30
20
10
0
1
YEARS_REGISTRATION

3. Analyze Data Imbalance

Objective: Evaluate whether there is an unequal representation of classes in the target variable (e.g., defaulters vs non-defaulters).

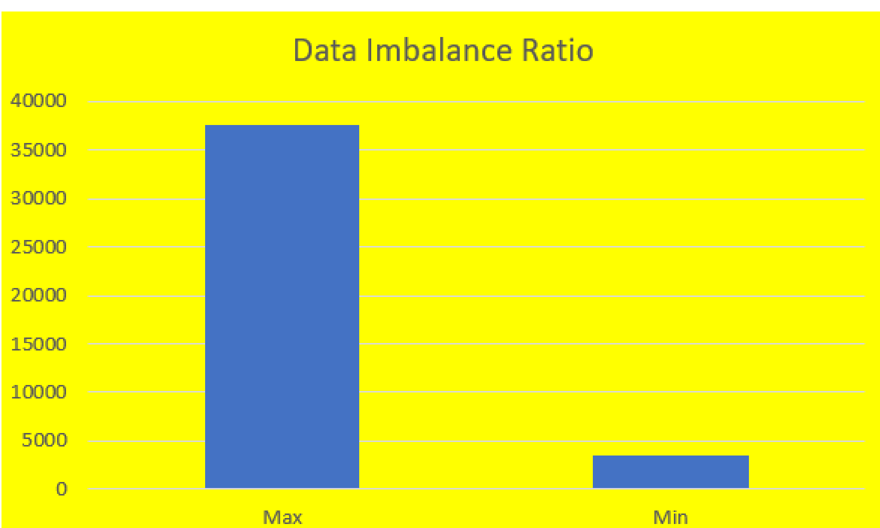
Methodology:

- Used `COUNTIF()` and `COUNT()` to count values for each class.
- Calculated the proportion of defaulters (TARGET = 1) to non-defaulters (TARGET = 0).
- Determined if imbalance might impact further statistical or predictive modeling.

Observation: Noted a high imbalance in favor of non-defaulters (~92%).

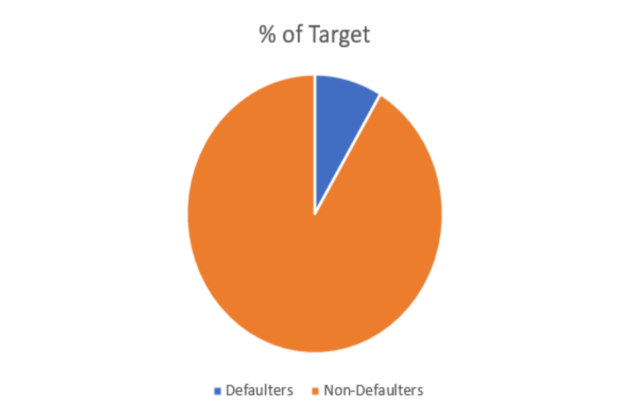
Visualization: Pie chart and bar chart showing target distribution.

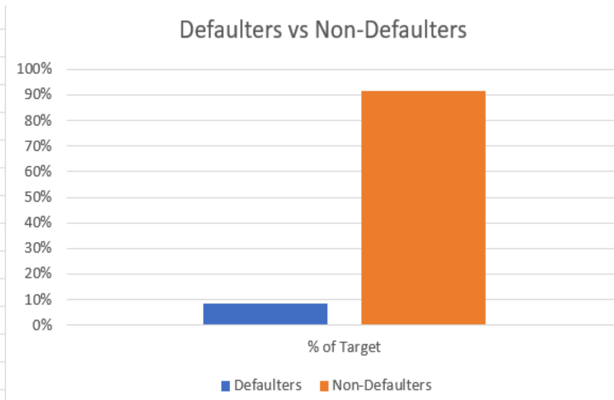
Max	37549
Min	3520
Ratio of Data Imbalance(Min/Max)	0.09




% Client
Cash loans
Revolving loans


Cash Loan vs Revolving Loan
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% Client
Cash loans
Revolving loans


% of Target
Defaulters
Non-Defaulters


Defaulters vs Non-Defaulters
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% of Target
Defaulters
Non-Defaulters

4. Perform Univariate, Segmented Univariate, and Bivariate Analysis

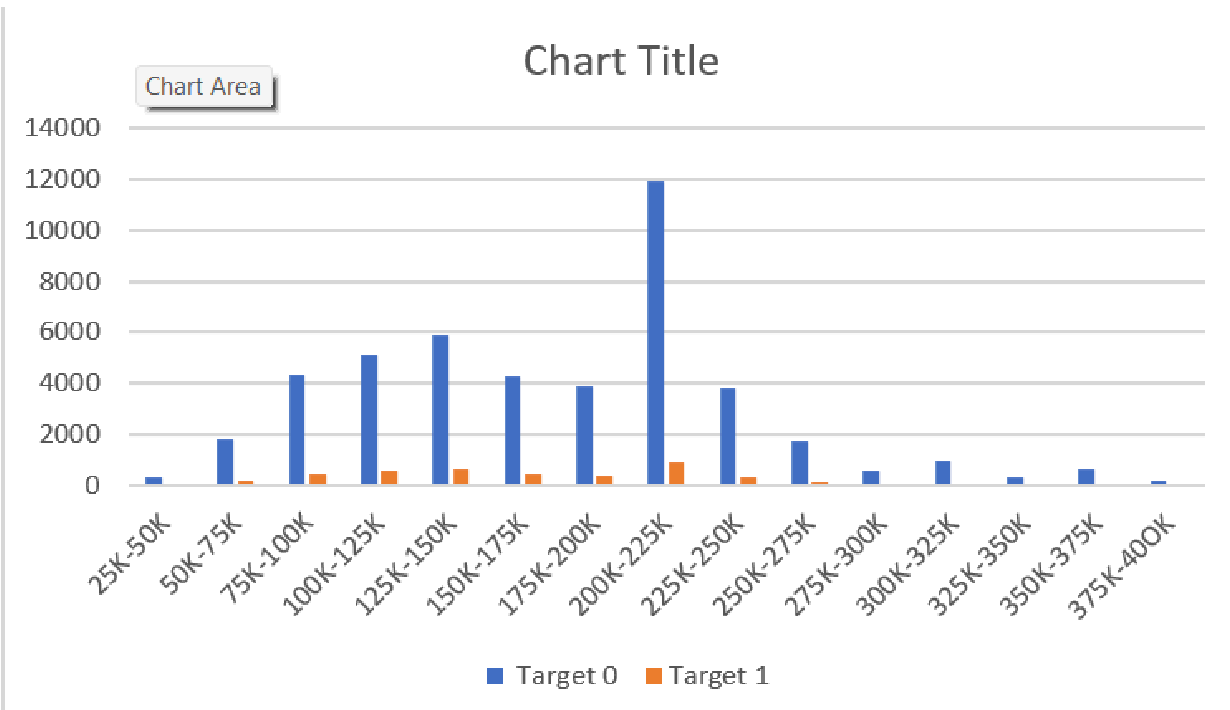
Objective: Understand the distribution and relationships between variables and the target variable.

Univariate Analysis:

- Examined individual variables such as `AMT_INCOME_TOTAL`.
- Tools used: `AVERAGE`, `MEDIAN`, `MODE`, histograms.

MAX	MIN
3825000	25650

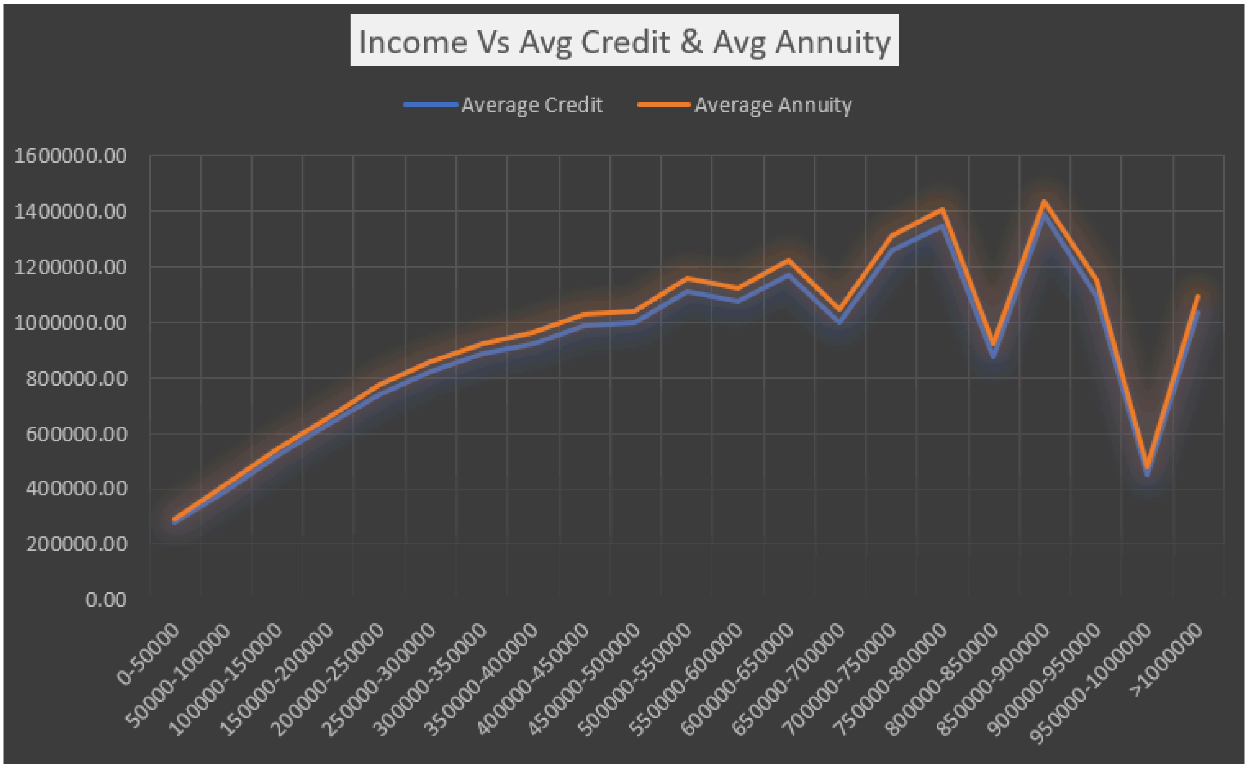
Income bins	Target 0	Target 1
25K-50K	310	37
50K-75K	1805	184
75K-100K	4304	433
100K-125K	5101	536
125K-150K	5905	603
150K-175K	4282	445
175K-200K	3880	366
200K-225K	11960	916
225K-250K	3821	287
250K-275K	1724	130
275K-300K	589	41
300K-325K	980	54
325K-350K	293	22
350K-375K	661	32
375K-400K	170	13



Bivariate Analysis:

- Used scatter plots and pivot tables to examine correlations.
- Example: Relationship between Income Vs Avg Credit & Avg Annuity.

Income Range	Average Credit	Average Annuity
0-50000	277298.05	13640.21
50000-100000	393349.85	18704.74
100000-150000	519709.47	24009.44
150000-200000	630878.77	28654.87
200000-250000	740833.61	33007.89
250000-300000	821826.37	36100.31
300000-350000	884090.21	39301.63
350000-400000	920791.10	40810.46
400000-450000	985704.88	44097.51
450000-500000	997944.62	44784.76
500000-550000	1112433.21	45984.46
550000-600000	1074844.07	46788.96
600000-650000	1171325.88	49857.69
650000-700000	1000031.84	47379.41
700000-750000	1259983.35	53482.28
750000-800000	1344940.00	58803.00
800000-850000	876760.07	47799.86
850000-900000	1388400.75	47631.38
900000-950000	1093675.66	58976.22
950000-1000000	450000.00	30073.50
>1000000	1037054.80	54840.04



5. Identify Top Correlations for Different Scenarios

Objective: Find the strongest predictors of loan default.

Methodology:

- Segmented the data based on TARGET (1 or 0).
- Used `CORREL()` to compute correlation between numeric variables and TARGET.
- Identified top predictors for defaulters vs non-defaulters.

TARGET 0

	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_ID_PUBLISH	REGION_RATING_CLIENT
AMT_INCOME_TOTAL	1						
AMT_CREDIT	0.360011781	1					
REGION_POPULATION_RELATIVE	0.188785867	0.09654213	1				
YEARS_BIRTH	0.049536299	0.160878908	0.048987416	1			
YEARS_EMPLOYED	0.036221572	0.094943177	-0.005606334	0.352389434	1		
YEARS_ID_PUBLISH	0.023115928	0.044246818	0.004355924	0.107692262	0.08250215	1	
REGION_RATING_CLIENT	-0.206983514	-0.10574409	-0.544721325	-0.04962383	0.015487867	-0.006932647	1

TARGET 1

	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_ID_PUBLISH	REGION_RATING_CLIENT
AMT_INCOME_TOTAL	1						
AMT_CREDIT	0.312173644	1					
REGION_POPULATION_RELATIVE	0.096758897	0.0555977	1				
YEARS_BIRTH	0.087629893	0.19443753	0.013409076	1			
YEARS_EMPLOYED	0.022601082	0.10510967	-0.001640893	0.305741728	1		
YEARS_ID_PUBLISH	0.037532601	0.05440939	0.008005666	0.125405421	0.099252606	1	
REGION_RATING_CLIENT	-0.160225589	-0.04798923	-0.436699036	-0.05130357	-0.003613733	-0.028399284	1

Top 5 Correlation (Non-Defaulters)		
Variable 1	Variable 2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.986
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.948
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.853

Top 5 Correlation (Defaulters)		
Variable 1	Variable 2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
AMT_GOODS_PRICE	AMT_CREDIT	0.982
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.806

---

6. Insights and Recommendations

- Applicants with low `EXT_SOURCE` scores are more likely to default.
- Long employment history is associated with lower default risk.
- Applicants with extremely high or low credit amounts are at higher risk.
- Consider using `EXT_SOURCE_2`, `DAYS_EMPLOYED`, and `AMT_CREDIT` as key variables in risk scoring models.

Prepared By: Anjali sahni