Bank Loan Case Study

Prepared by: Indhu Reddy KR

---

1. Project Description

This case study aims to address loan defaults among customers with limited credit history. By using Exploratory Data Analysis (EDA), patterns in loan approvals and defaults are analyzed to improve the loan approval process. The goal is to reduce the rejection of eligible applicants and minimize default risks.

Key Scenarios Faced by Banks:

- Loss of business when eligible customers are not approved.
- Financial loss when ineligible customers are approved.

---

2. Outcomes of Loan Applications

1. **Approved:** Loan applications successfully granted.
2. **Cancelled:** Customers cancel applications during approval.
3. **Refused:** Applications rejected by the bank.
4. **Unused Offer:** Approved loans not utilized by customers.

---

3. Tech-Stack Used

- **Microsoft Excel 365:** Data cleaning, analysis, and visualization.
- **PDF:** Presentation of project insights.
- **Microsoft Clipchamp:** Creating video presentations for detailed project explanations.

---

4. Approach

1. Importing and understanding the dataset.
2. Cleaning and imputing missing values.
3. Performing EDA to extract insights.
4. Visualizing data with graphs and charts.
5. Summarizing findings into a report and video explanation.

Datasets Used:

- **Application Data:** 126 columns, 50,000 rows, 67 columns with null values.
- **Previous Application Data:** 37 columns, 50,000 rows, 15 columns with null values.

Steps Taken for Cleaning:

- Dropped columns with more than 40% null values.
- Categorical columns filled with mode; numerical columns filled with median.

---

5. Data Challenges and Analysis

- **Data Imbalance:**
  - Class 0 (Non-Defaulters): 91%.
  - Class 1 (Defaulters): 9%.
  - Challenge: Accurately predicting defaulters due to minority representation.
- **Outliers Detection:** Identified and managed.

6.Data Analytics Tasks:

A. Identify Missing Data and Deal with it Appropriately

- **Process:**
  - Reviewed datasets to identify columns with missing values.
  - Dropped columns with more than **40% missing values** to focus on more reliable data.
  - For remaining columns:
    - **Categorical columns:** Filled missing values with the mode (most frequent value). For example:
      - Columns like `NAME_TYPE_SUITE` and `OCCUPATION_TYPE` were filled with "Unknown."
    - **Numerical columns:** Filled missing values with the median to avoid skewing the data. For example:
      - Columns like `AMT_ANNUITY`, `EXT_SOURCE_2`, and `AMT_REQ_CREDIT_BUREAU_YEAR`.

B. Identify Outliers in the Dataset

- **Process:**
    - Performed outlier detection to identify extreme values that could distort analyses.
    - Although specific method IQR, the emphasis was on identifying values that deviated significantly from the dataset norms. Also used Box plots to visualize the outliers
    - Outliers were likely handled appropriately to ensure they did not affect model performance or insights derived.

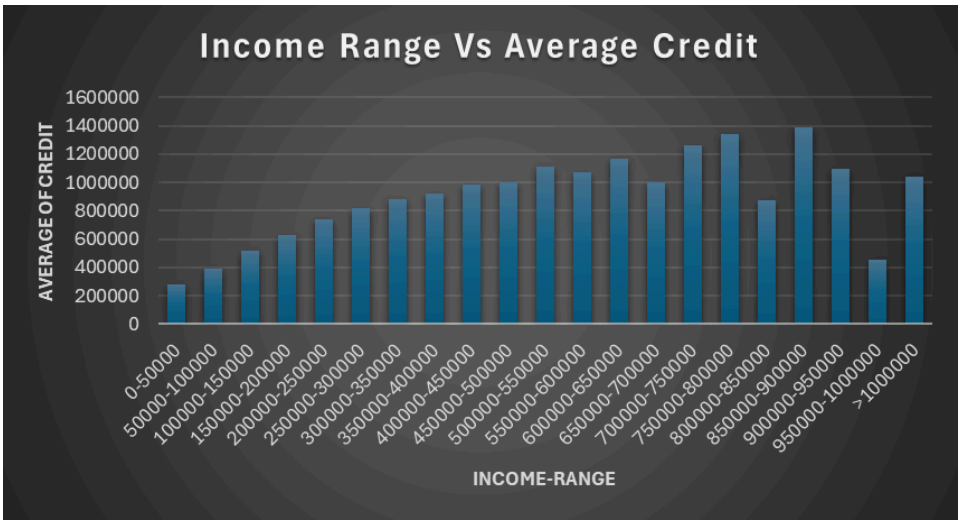
C. Analyze Data Imbalance

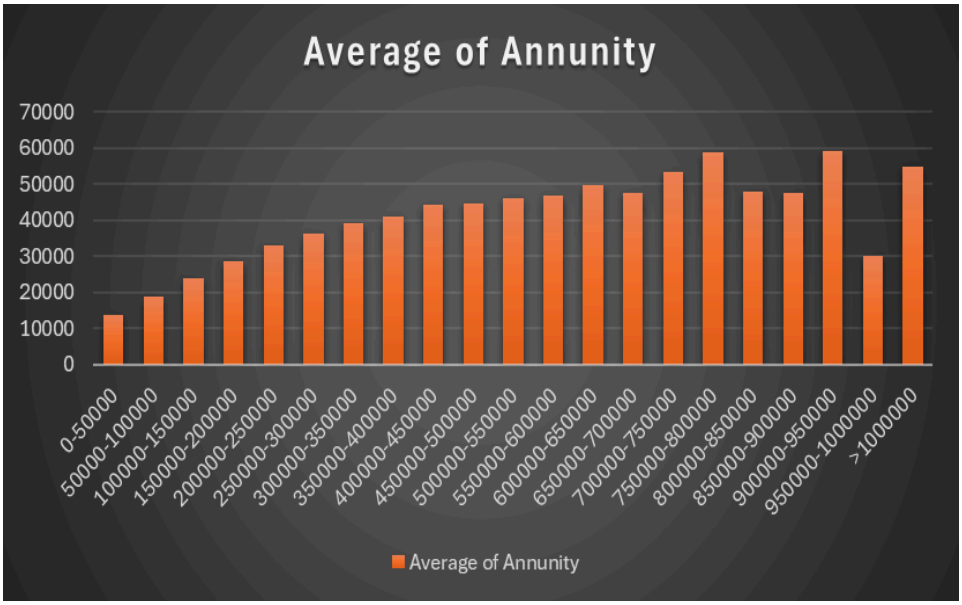
- **Observations:**
    - The dataset was highly imbalanced:
        - **Non-Defaulters (Class 0):** 91% of the data.
        - **Defaulters (Class 1):** 9% of the data.
    - **Challenge:** The small proportion of defaulters makes it harder to predict this class accurately.
- **Approach:**
    - Acknowledged the need for techniques like resampling (e.g., oversampling minority class, undersampling majority class) or using algorithms designed to handle imbalanced data for better performance.

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis

- **Univariate Analysis:**
    - Explored individual attributes to identify distributions and trends:
        - **Age:** Most loan takers were aged 20-40, while younger individuals posed higher default risks.
        - **Gender:** we checked which gender had the highest defaulters, while females were slightly higher in defaulters.
        - **Family Status:** Married individuals had the highest loan participation. Single individuals had a lower default rate.
        - **Education:** Borrowers with higher education levels showed lower default rates.
        - **Number of Children:** Families with more than 5 children showed lower default risks.
- **Segmented Univariate Analysis:**
    - Examined subcategories within individual attributes:
        - Default rates based on **age groups, education, and family size** provided deeper insights into borrower behavior.

- **Bivariate Analysis:**
    - Explored relationships between two variables:
        - For instance, analyzing the correlation between `AMT_CREDIT` (loan amount) and `AMT_INCOME_TOTAL` (income) to understand repayment capacity.





E. Identify Top Correlations for Different Scenarios

- **Process:**

  Investigated correlations between key features in **Application Data** to uncover factors most strongly associated with loan defaults.

  These correlations helped identify:

  - High-risk customer segments.
  - Key loan attributes influencing default likelihood.

Top Correlation of Defaulters

Rank	Variable 1	Variable 2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.981928143
2	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.948020808
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895600339
4	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.891467244
5	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.805583225
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.773107352
7	AMT_ANNUITY	AMT_GOODS_PRICE	0.746422447
8	AMT_ANNUITY	AMT_CREDIT	0.745132112

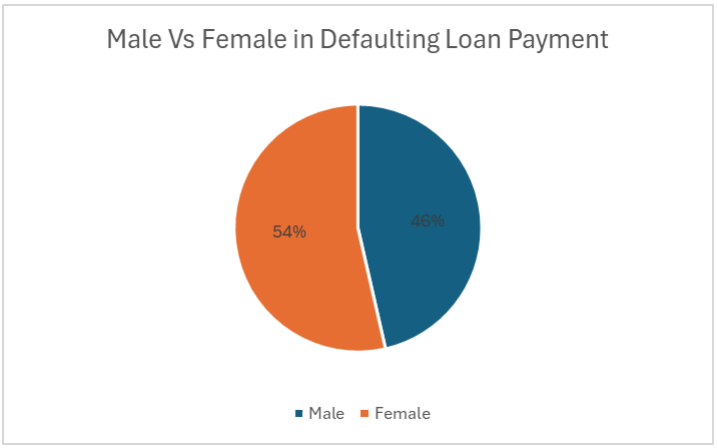
Top correlation of Non-Defaulters

Rank	Variable 1	Variable 2	Correlation
1	AMT_GOODS_PRICE	AMT_CREDIT	0.98635817
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950286525
3	CNT_CHILDREN	CNT_FAM_MEMBERS	0.893735596
4	REG_REGION_NOT_WORK_REGION -	LIVE_REGION_NOT_WORK_REGION	0.860167703
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.853040752
6	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.815604978
7	REGION_RATING_CLIENT	AMT_GOODS_PRICE	0.765201743
8	AMT_ANNUITY	AMT_GOODS_PRICE	0.765201743
9	AMT_CREDIT	AMT_ANNUITY	0.760827873

7. Key Insights: Univariate, Segmented Univariate, and Bivariate Analysis

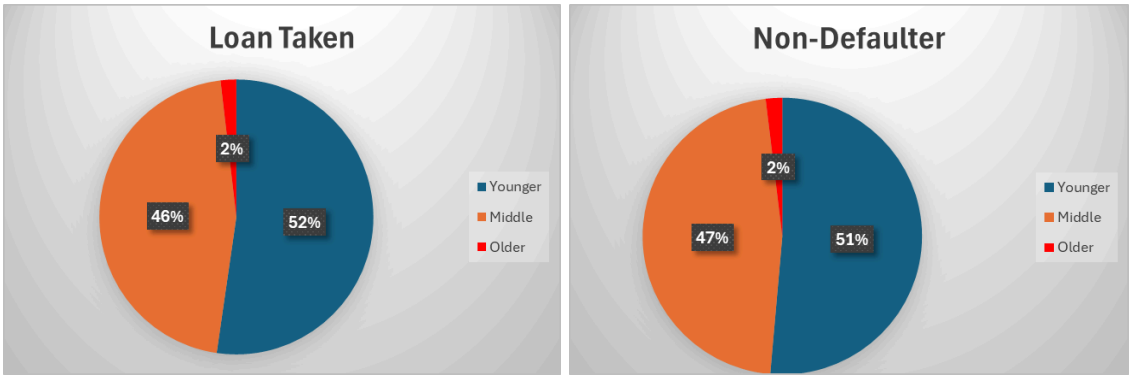
Demographic Factors:

- **Gender:**



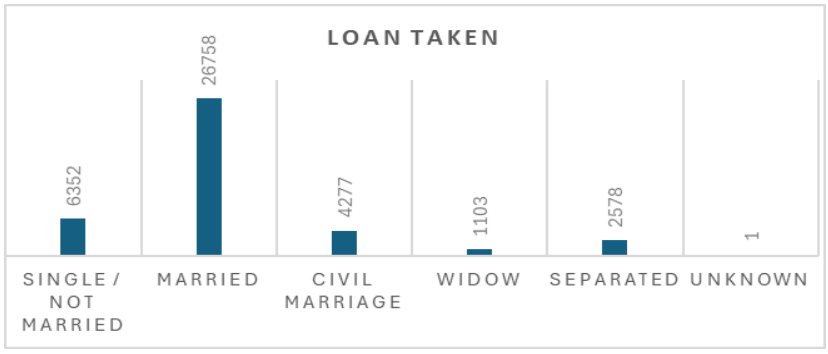
Compared to males,females were the ones with the slightest higher Defaulter.

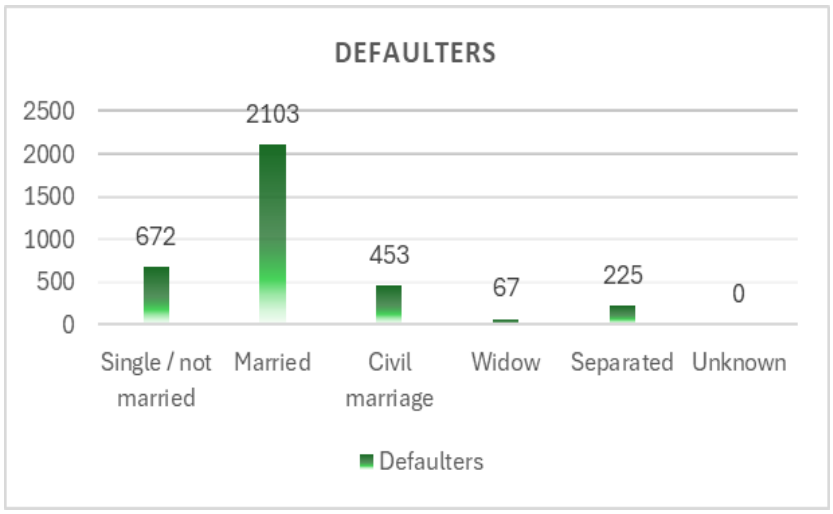
- **Age:**



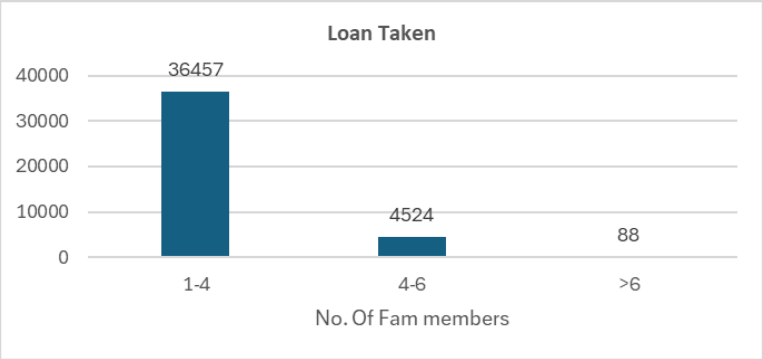
  - Most loan takers are aged 20-40.
  - Younger individuals might require tailored terms to reduce default risks.

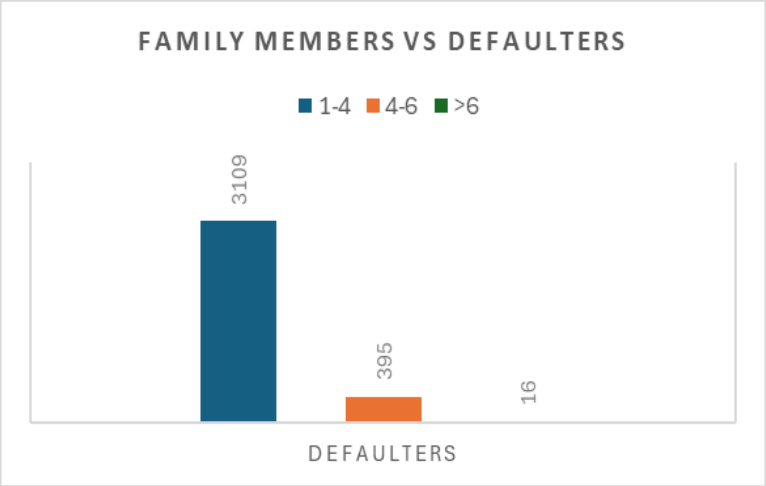
- **Family Status:**





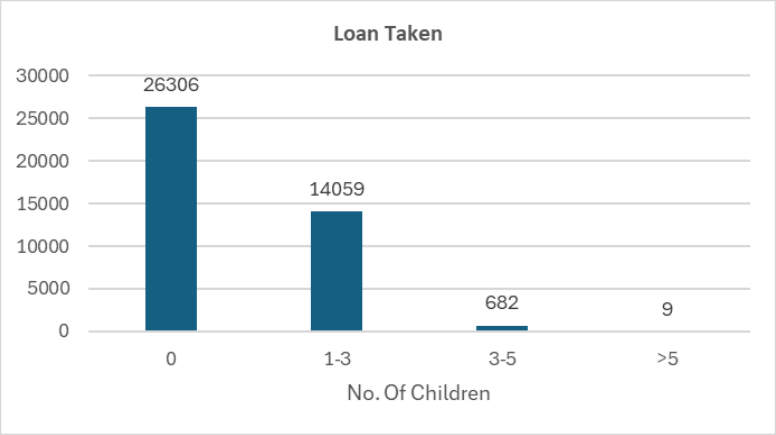
Family members;

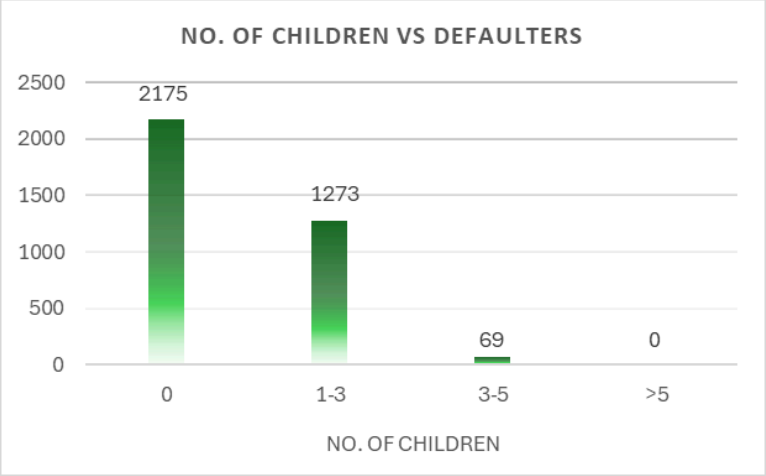




- Married individuals: Highest loan participation.
- Widows and separated individuals: Lower default risks.

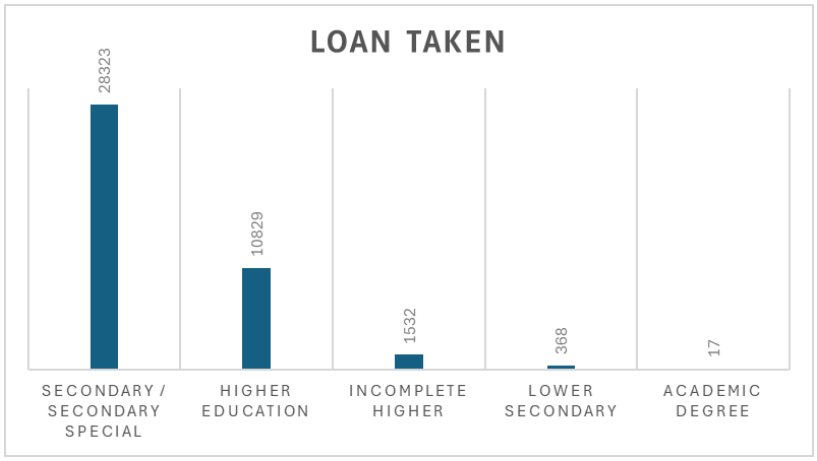
- **Number of Children:**

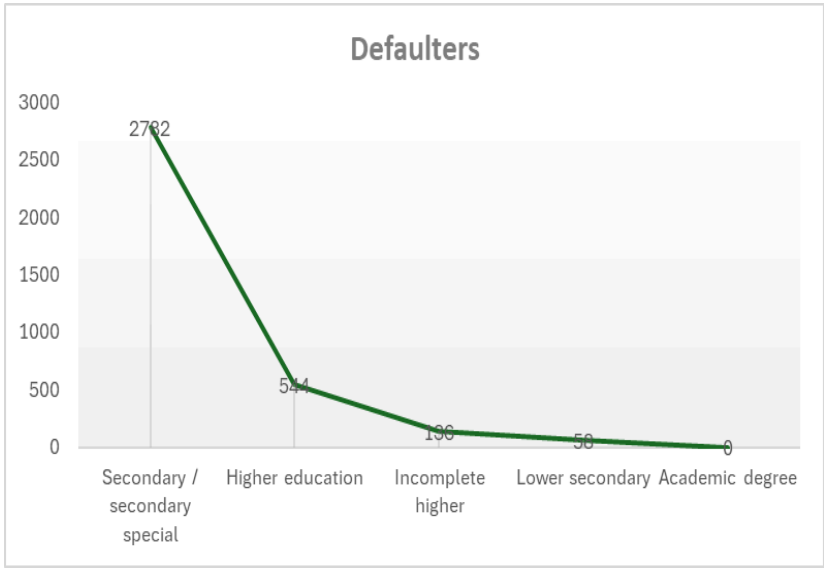




  - Larger families (>5 children): Lower default risks.
  - Families with >6 members: Higher default rate at 18%.

- **Education:**



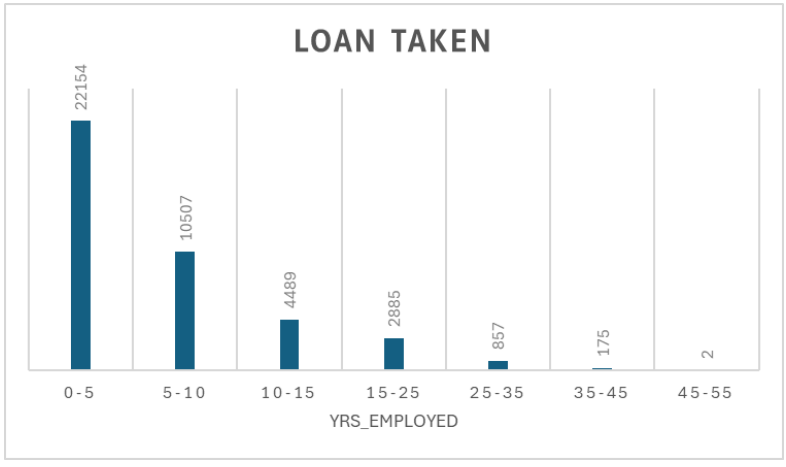


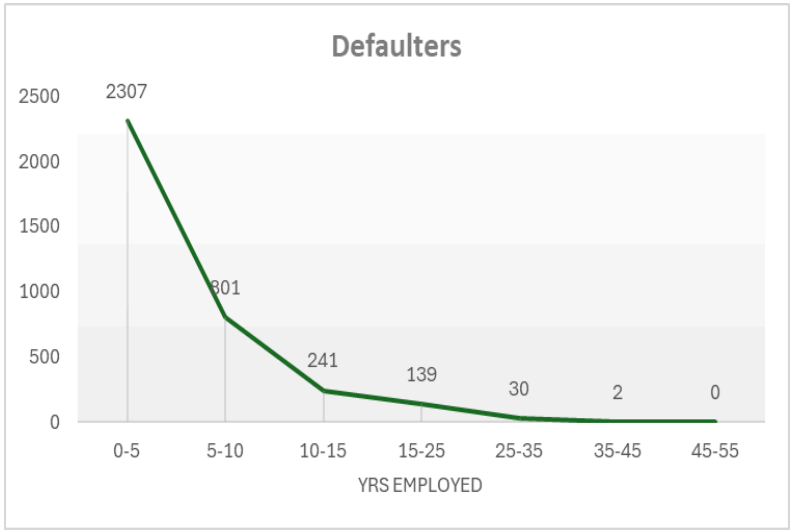
  - Academic degree holders: 0% default rate.
  - Borrowers with "Secondary education": Higher default rate at 10%.

- **Occupation:**

Occupation	Loan Taken	Non-Defaulters	Defaulters	Defaulters %
Laborers	8951	8032	919	10%
Core staff	4434	4184	250	6%
Accountants	1621	1540	81	5%
Managers	3485	3243	242	7%
Drivers	3044	2706	338	11%
Sales staff	5159	4668	491	10%
Cleaning staff	739	671	68	9%
Cooking staff	963	862	101	10%
Unknown	6730	6207	523	8%
Private service staff	447	410	37	8%
Medicine staff	1403	1297	106	8%

Security staff	**1140**	**1015**	**125**	**11%**
High skill tech staff	**1852**	**1734**	**118**	**6%**
Waiters/barmen staff	**228**	**203**	**25**	**11%**
Low-skill Laborers	**357**	**296**	**61**	**17%**
Realty agents	**123**	**110**	**13**	**11%**
Secretaries	**212**	**203**	**9**	**4%**
IT staff	**80**	**76**	**4**	**5%**
HR staff	**101**	**92**	**9**	**9%**





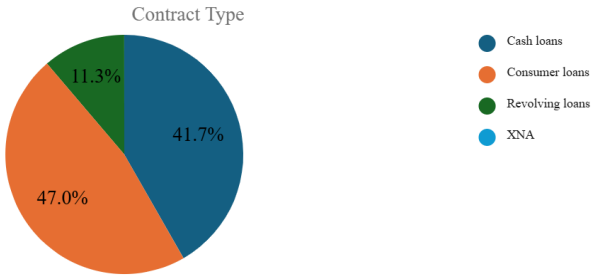
- High-risk occupations include Drivers, Low-skill Laborers, and Waiters.

Contract Types and Status:

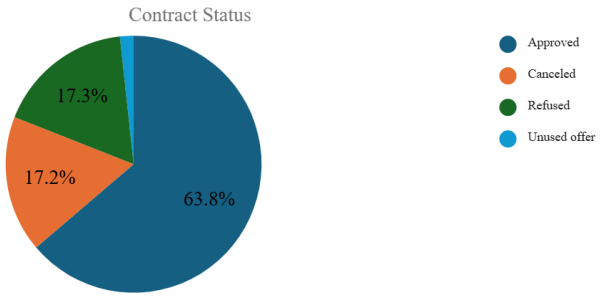
- Previous applications analyzed for segmented insights.

Previous Application Data – Univariate, Segmented Univariate Analysis

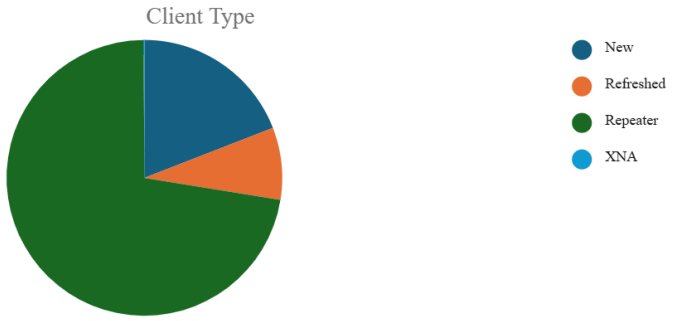
Contract Type



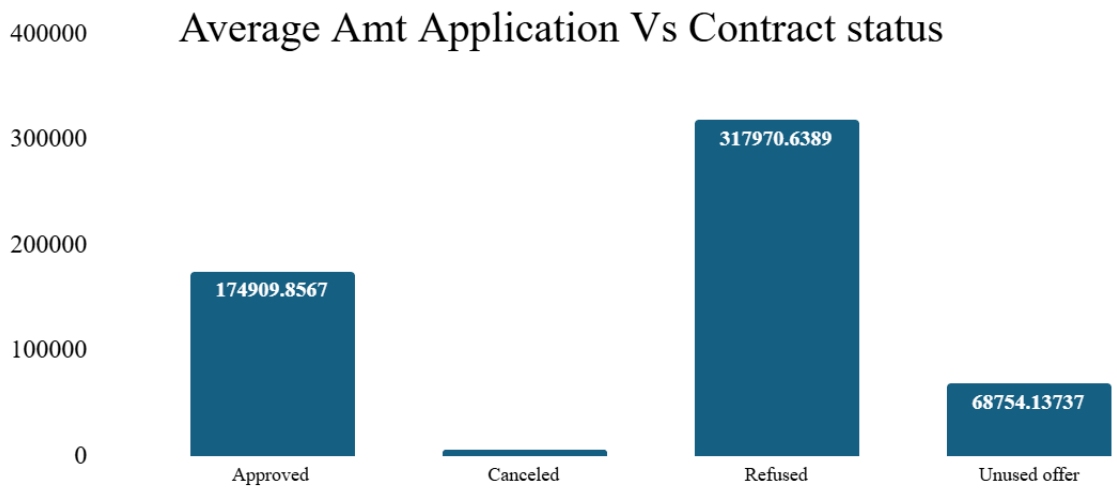
Contract Status

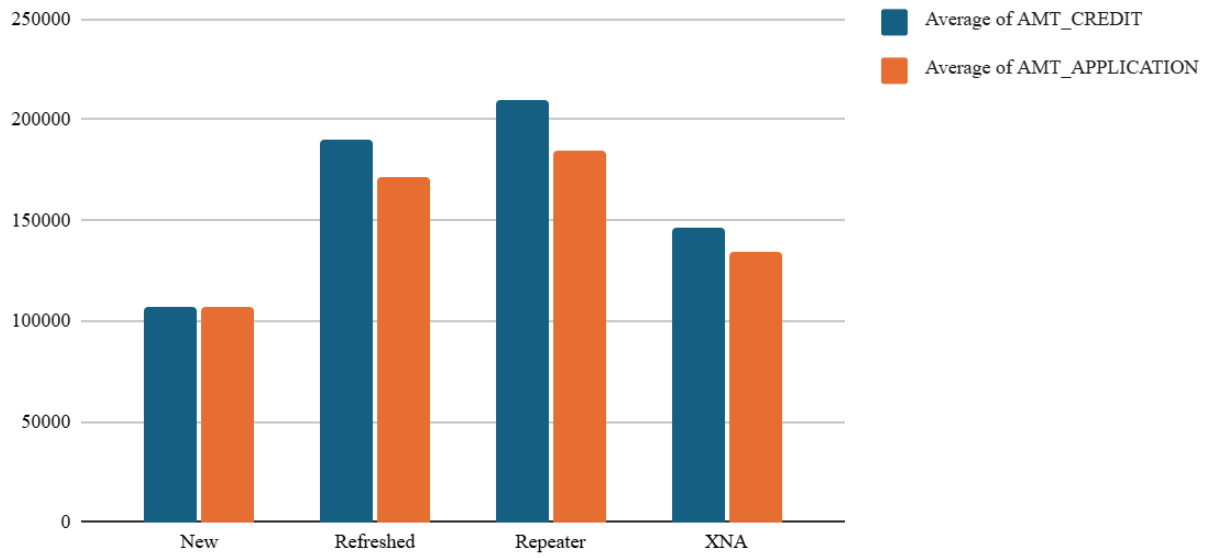


Client Type



New	9548
Refreshed	4227
Repeater	36167
XNA	57





---

8. Results

The project demonstrated the ability to apply EDA to large datasets, uncovering key patterns influencing loan defaults. By addressing data imbalance and managing outliers, critical insights were extracted, supporting risk analytics in banking.

---

9. Conclusion

This case study enhanced my ability to analyze financial data, interpret trends, and draw actionable conclusions. It has also deepened my understanding of risk management in banking, particularly in loan approval scenarios.
