SCALETRIX.AI

BFSI - Risk Analysis

AICPA SOC

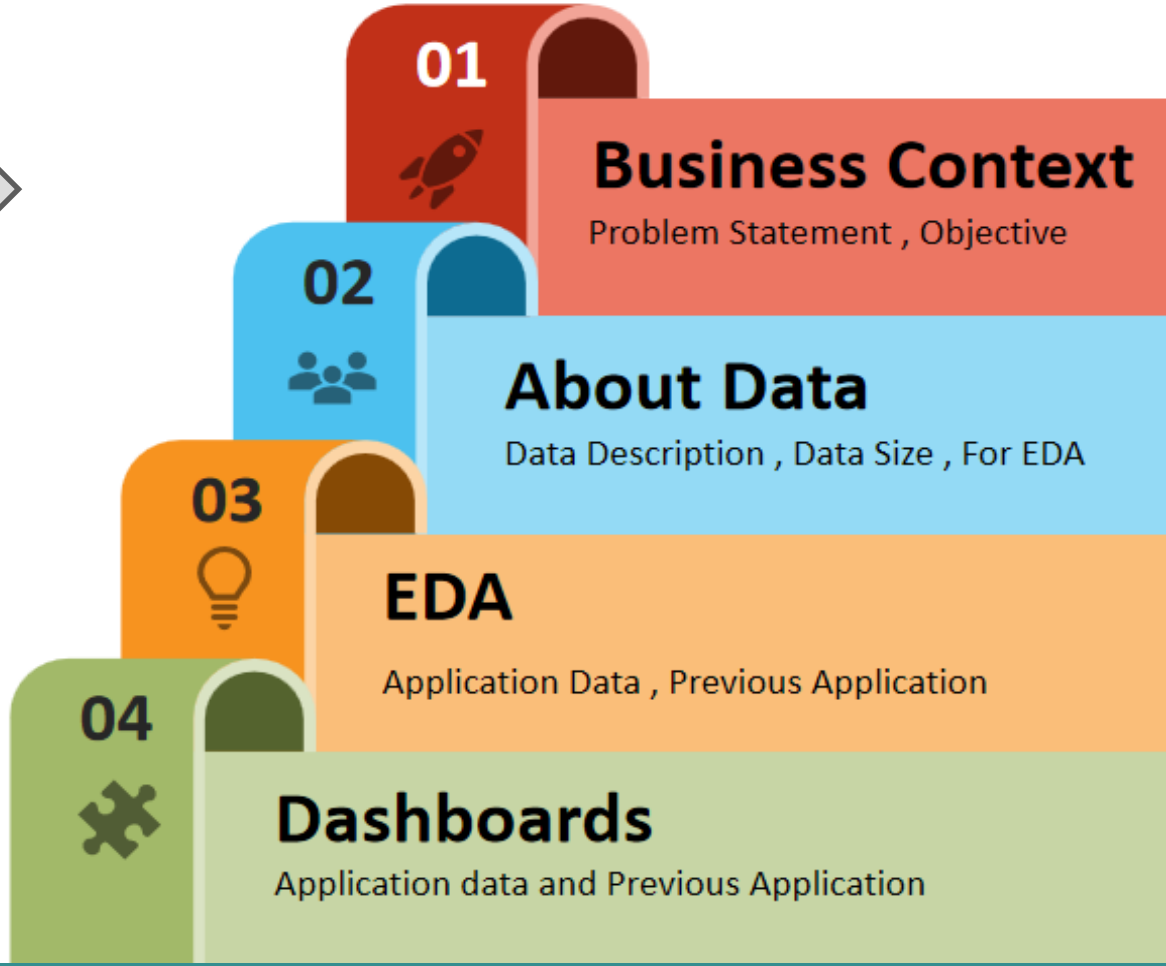
HIPAA COMPLIANT

GDPR COMPLIANT

Unleash Growth with Digital Transformation & Analytics

• NEW DELHI • BANGALORE • UNITED STATES

Agenda


01
Business Context
Problem Statement , Objective
02
About Data
Data Description , Data Size , For EDA
03
EDA
Application Data , Previous Application
04
Dashboards
Application data and Previous Application

Business Context

Problem Statement -

In the banking and financial services industry, managing the risk associated with lending is crucial. Lenders must evaluate the likelihood that a borrower will default on their loan, as defaults can result in significant financial losses. The goal is to identify patterns that indicate a higher risk of default, which can inform actions such as denying the loan, reducing the loan amount, or charging higher interest rates to risky applicants. By doing so, lenders can better ensure that only creditworthy consumers receive loans.

Objective -

The objective of this analysis is to apply exploratory data analysis (EDA) and predictive modelling to identify key factors contributing to loan defaults. By examining borrower demographics, financial history, and loan details, we aim to uncover insights that enhance risk assessment and support better lending strategies to reduce defaults and improve profitability.

About Data

Data Description-

1. **application_data.csv** It contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. **previous_application.csv** It contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Data Size -

1. **application_data.csv** It contain about 122 columns and 307511 records.
2. **previous_application.csv** It contain about 37 columns and 1670214 records.

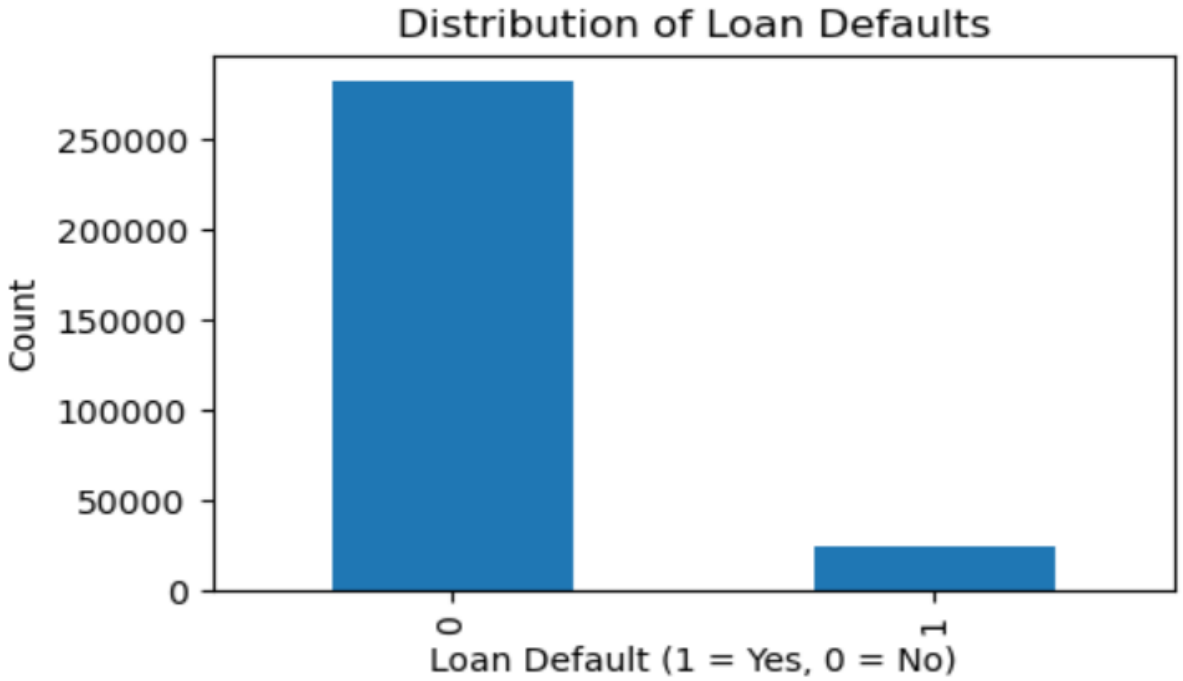
For exploratory data analysis (EDA) -

From both application_data.csv and previous_application.csv ignoring all the columns where null value percentage is greater than 40%

Exploratory Data Analysis (EDA)

Analysis On Application Data

Distribution of loan defaults

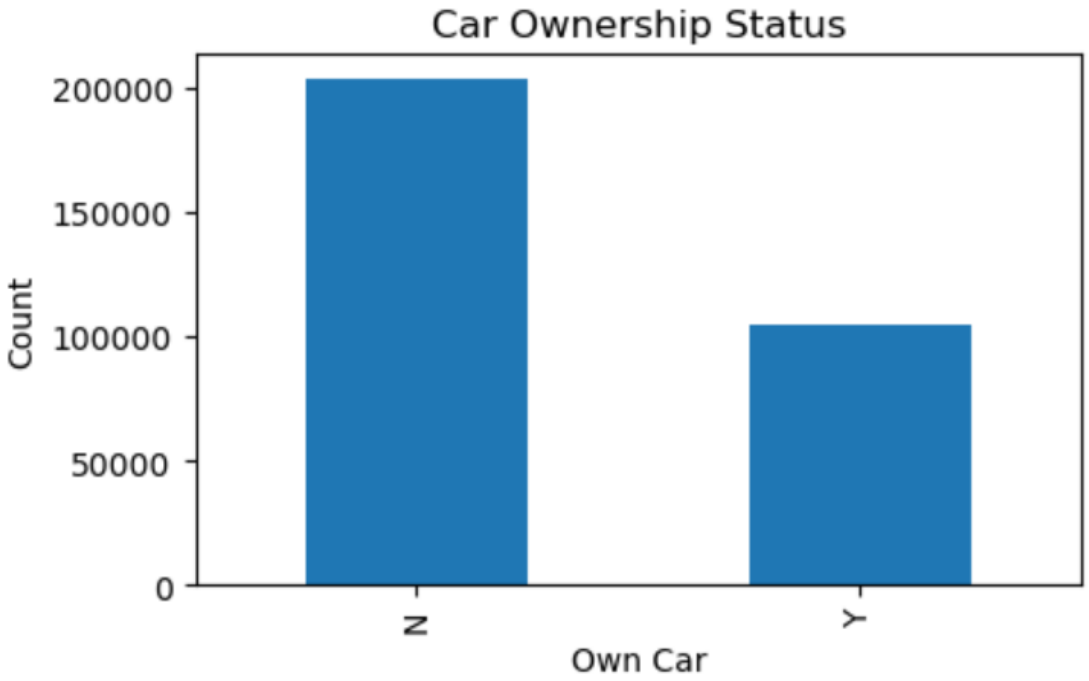


Class Imbalance:
- **Total Applications**: 307,511
- **Non-Defaulting Borrowers (TARGET = 0)**: 282,686 (91.8%)
- **Defaulting Borrowers (TARGET = 1)**: 24,825 (8.2%)
- **Insight**: Significant imbalance in class distribution; majority of applicants do not default.

Gender Distribution

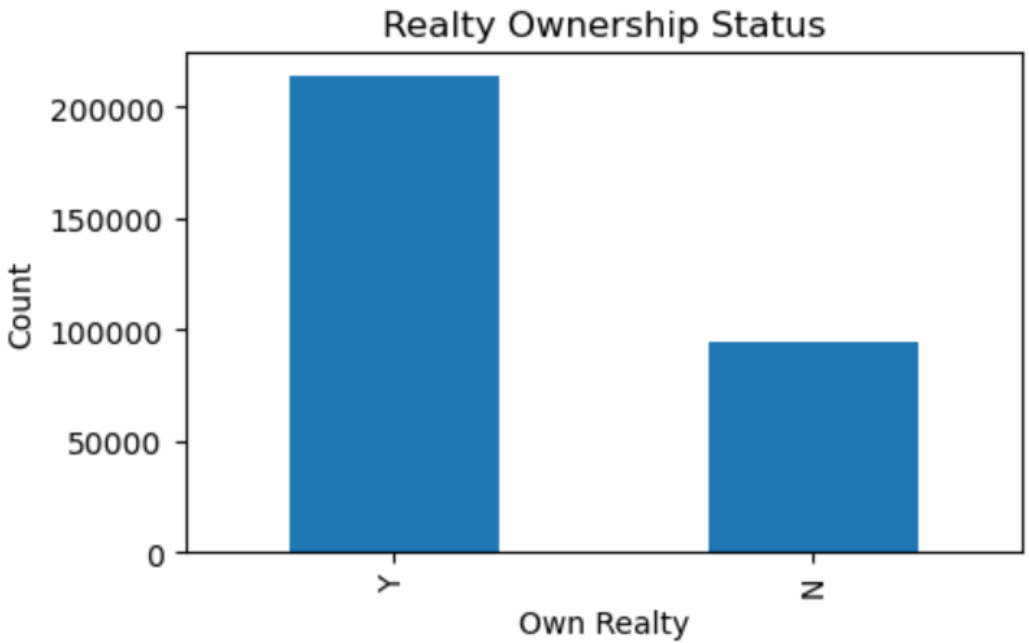
- **Gender Representation**:
  Female Borrowers (F): 202,448 (65.8%)
  Male Borrowers (M): 105,059 (34.2%)
  Unknown Gender (XNA): 4 (negligible)
- **Female Majority**:
  Insight: The dataset shows a significant majority of female borrowers, indicating potential differences in borrowing behavior or default risk profiles compared to male borrowers.

Car Ownership Status



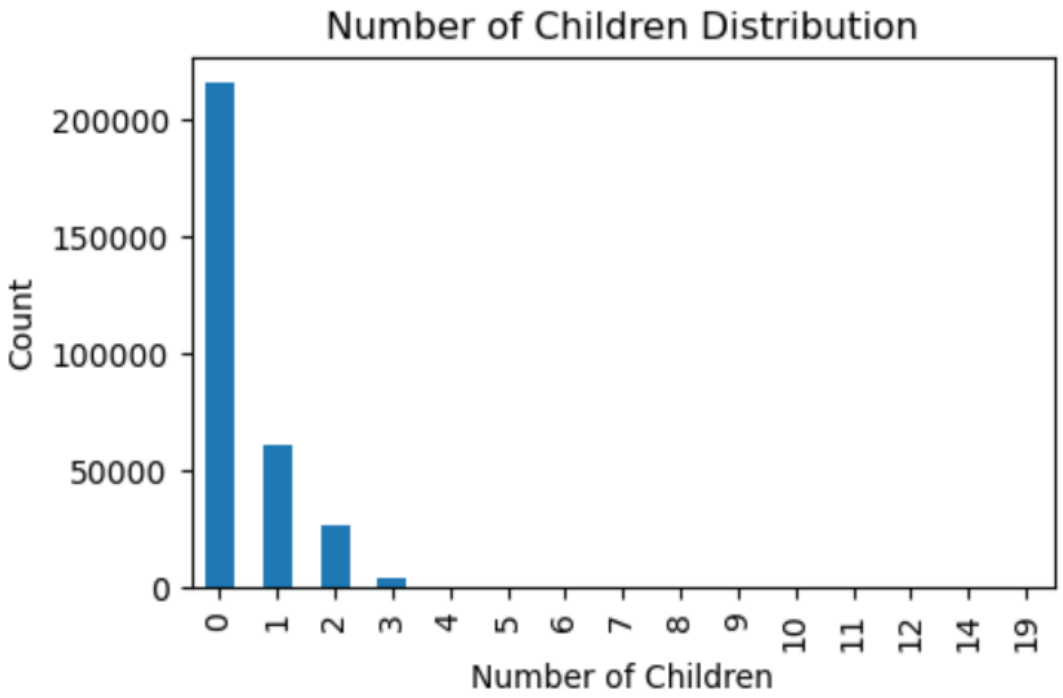
- **Car Ownership Distribution**:
  Non-Car Owners (N): 202,924 (65.9%)
  Car Owners (Y): 104,587 (34.1%)
- **Majority Without Car Ownership**:
  Insight: A significant portion of applicants do not own a car, which may influence their financial behaviors and borrowing needs.

Realty Ownership Status



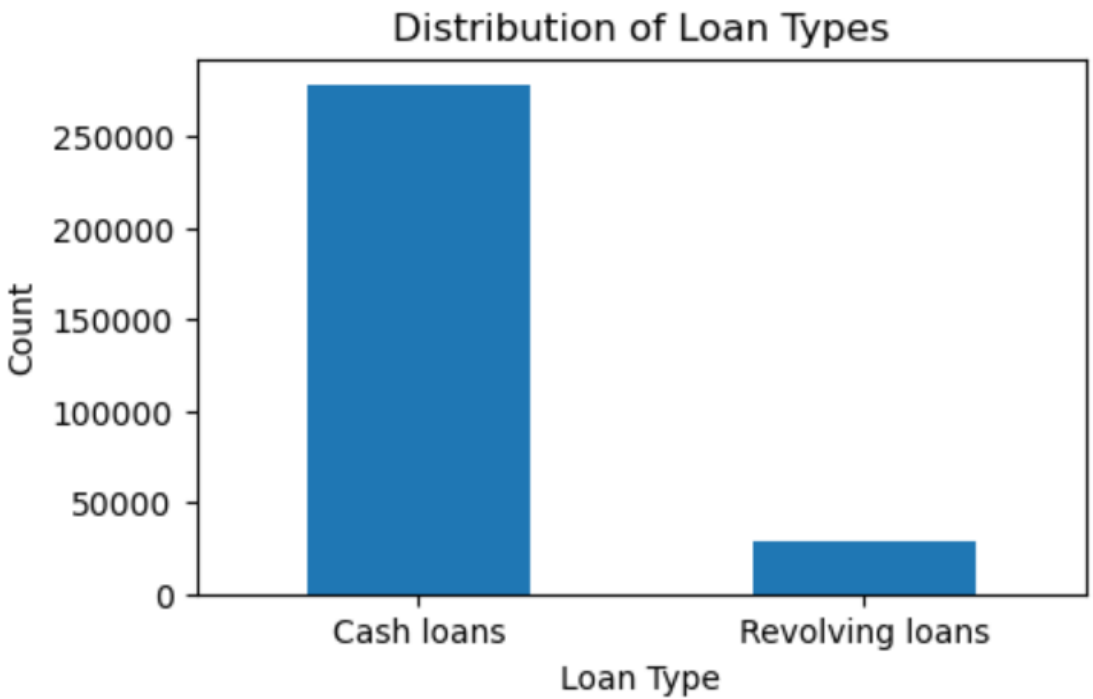
- **Realty Ownership Distribution**:
  Owners (Y): 213,312 (69.2%)
  Non-Owners (N): 94,199 (30.8%)
- **Predominance of Realty Owners**:
  Insight: A significant majority of applicants own real estate, suggesting potential stability And a lower likelihood of default.

Number of Children Distribution



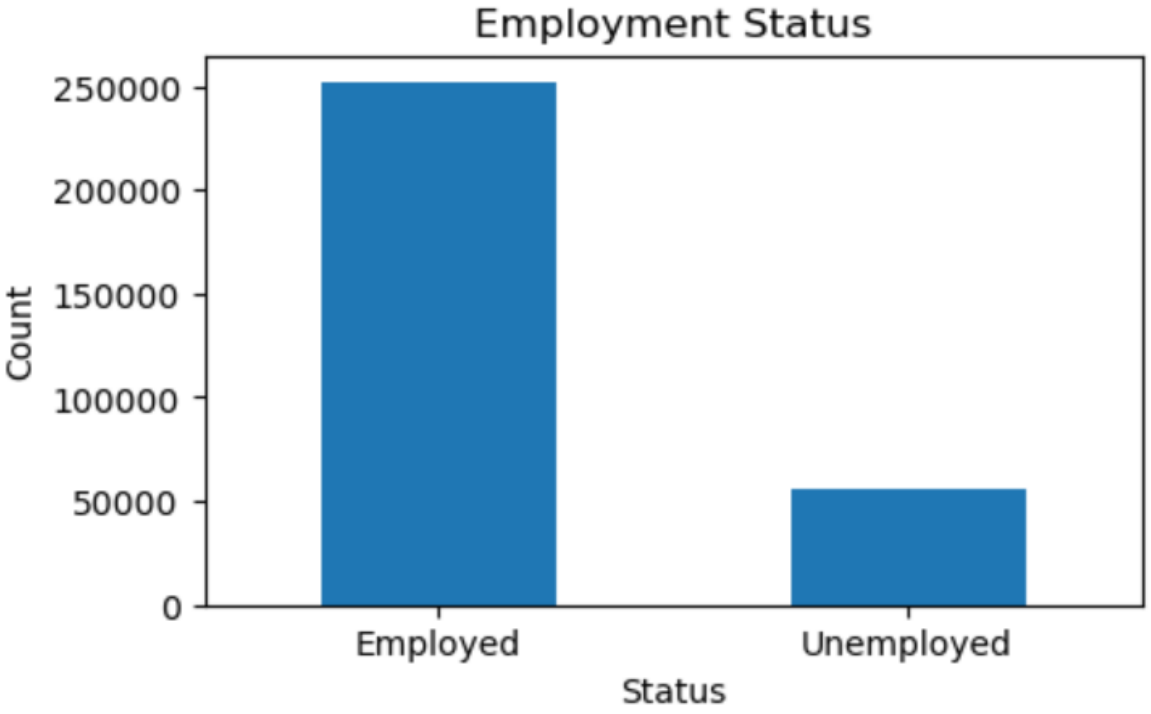
- **Children Count Distribution**:
  No Children (0): 215,371 (69.9%)
  One Child (1): 61,119 (19.9%)
  Two Children (2): 26,749 (8.7%)
  Three Children (3): 3,717 (1.2%)
  Four or More Children: 4 (negligible)
- **Majority Without Children**:
  Insight: A significant majority of applicants (69.9%) do not have children, which may affect their financial responsibilities and capacity to repay loans.

Distribution of Loan Types



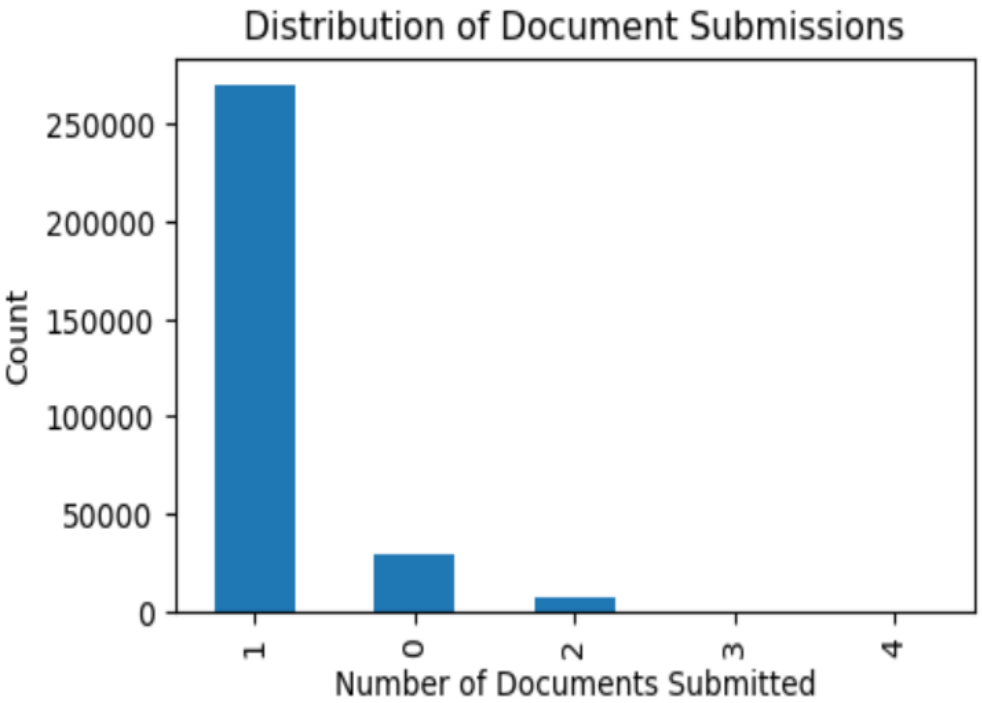
- **Loan Type Distribution**:
  Cash Loans: 278,232 (90.4%)
  Revolving Loans: 29,279 (9.6%)
- **Predominance of Cash Loans**:
  Insight: The overwhelming majority of loan applications are for cash loans, indicating strong preference or need for this type of financing among borrowers.

Employment Status



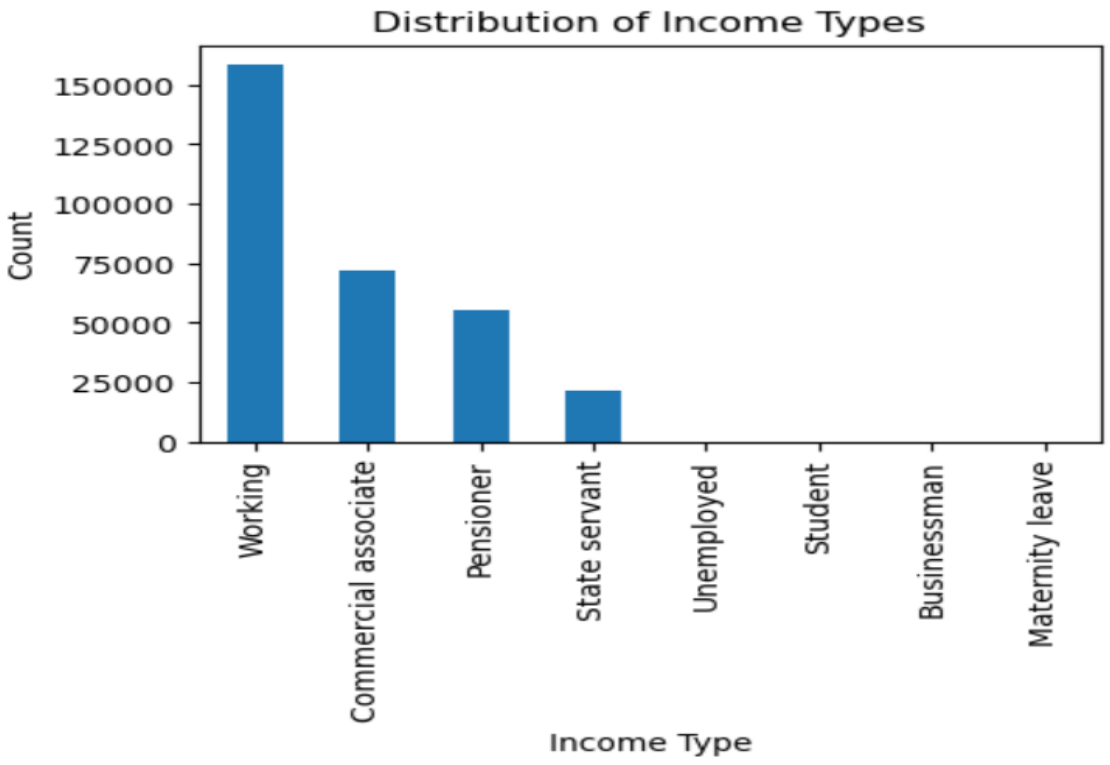
- **Employment Status Distribution**:
  Employed: 252,135 (81.1%)
  Unemployed: 55,376 (18.9%)
- **High Employment Rate**:
  Insight: A significant majority of loan applicants are employed, indicating a potentially more stable financial situation and capacity to repay loans.

Distribution of Document Submissions



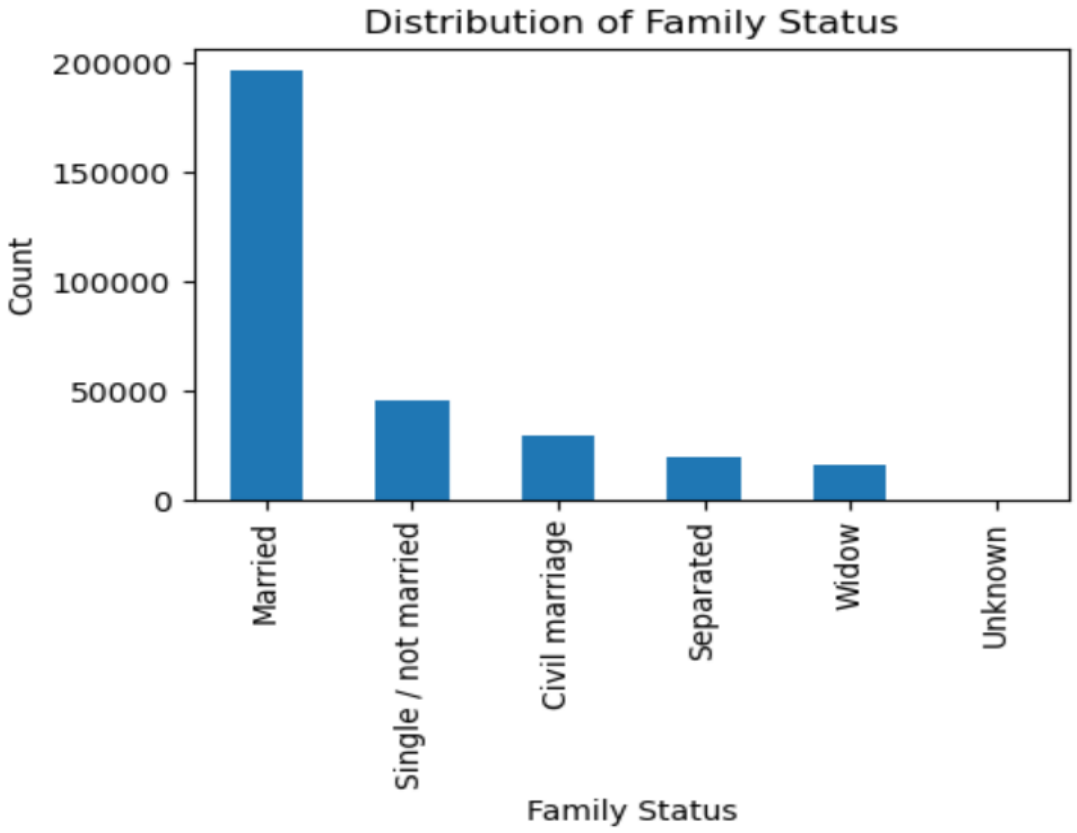
- **Document Submission Distribution**:
- **1 Document Provided**: 270,056 (89.4%)
- **No Documents Provided**: 29,549 (9.8%)
- **2 Documents Provided**: 7,742 (2.6%)
- **3 Documents Provided**: 163 (0.1%)
- **4 Documents Provided**: 1 (negligible)
- **Majority Submit Minimal Documentation**:
- **Insight**: A large portion of applicants (71%) only provides DOCUMENT_3, which may raise concerns regarding the completeness of their application and potential risk.

Distribution of Income Types



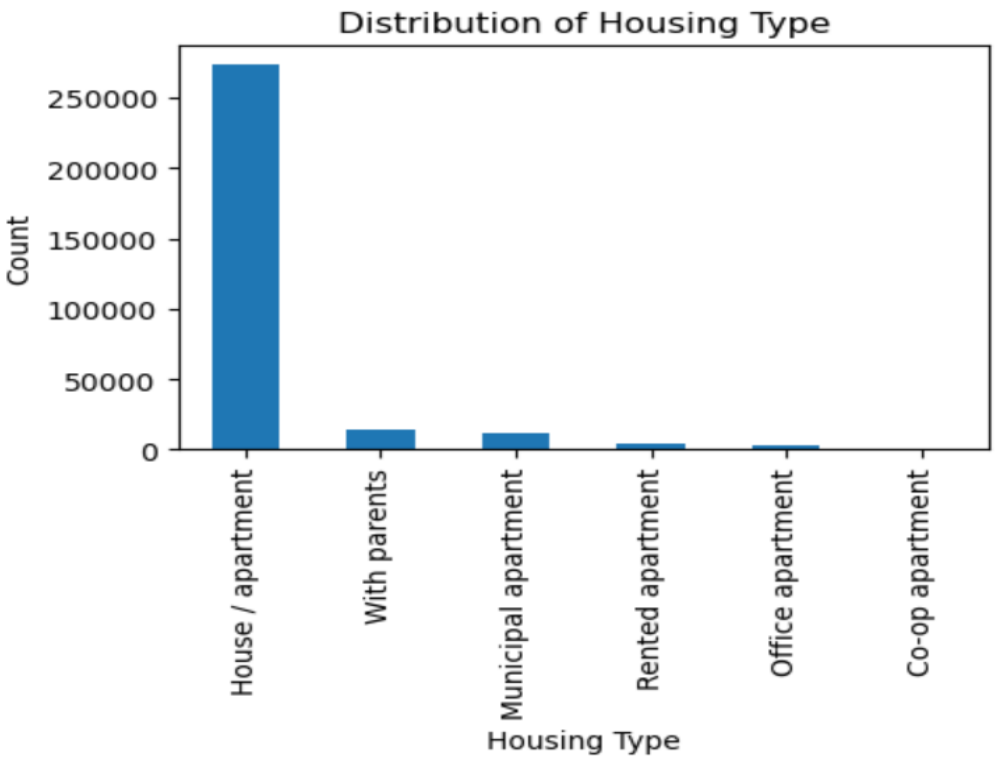
- **Income Type Distribution**:
  Working: 158,774 (51.3%)
  Commercial Associate: 71,617 (23.3%)
  Pensioner: 55,362 (18.0%)
  State Servant: 21,703 (7.1%)
- **Predominance of Employment**:
  Insight: The majority of applicants are employed, highlighting a stable income base that may contribute positively to their creditworthiness

Distribution of Family Status



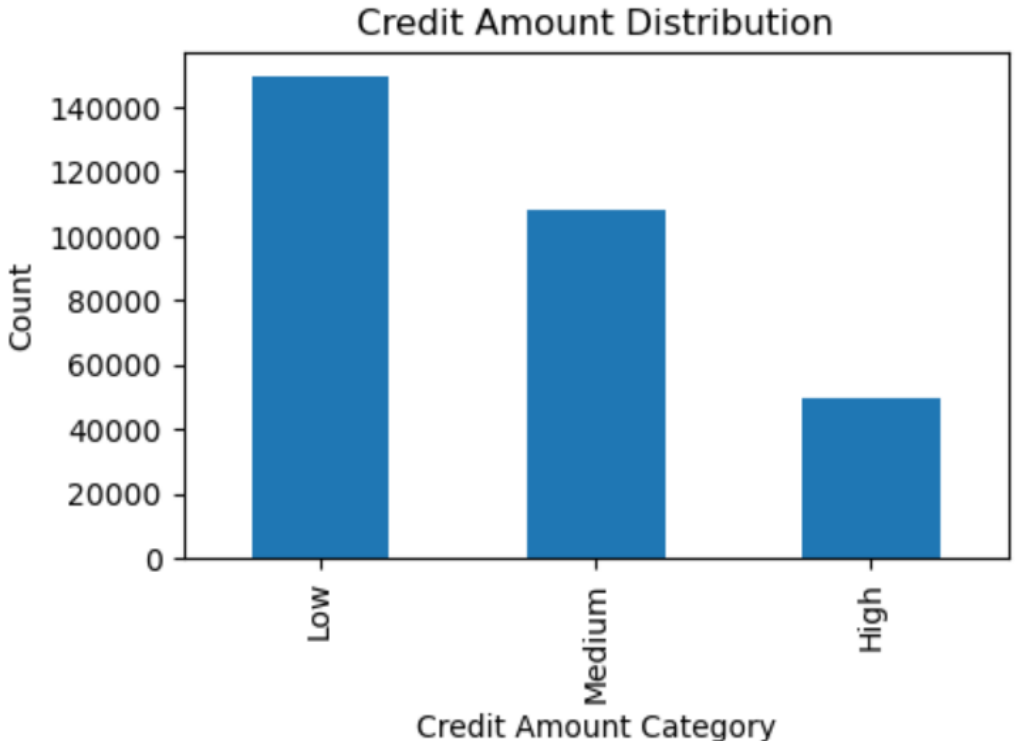
- **Family Status Distribution**:
  Married: 196,432 (64.1%)
  Single / Not Married: 45,444 (14.8%)
  Civil Marriage: 29,775 (9.7%)
  Separated: 19,770 (6.5%)
  Widow: 16,088 (5.2%)
- **Majority are Married**:
  Insight: A significant portion of applicants are married, which may correlate with greater financial stability and responsibility, potentially impacting their loan repayment behavior positively

Distribution of Housing Type



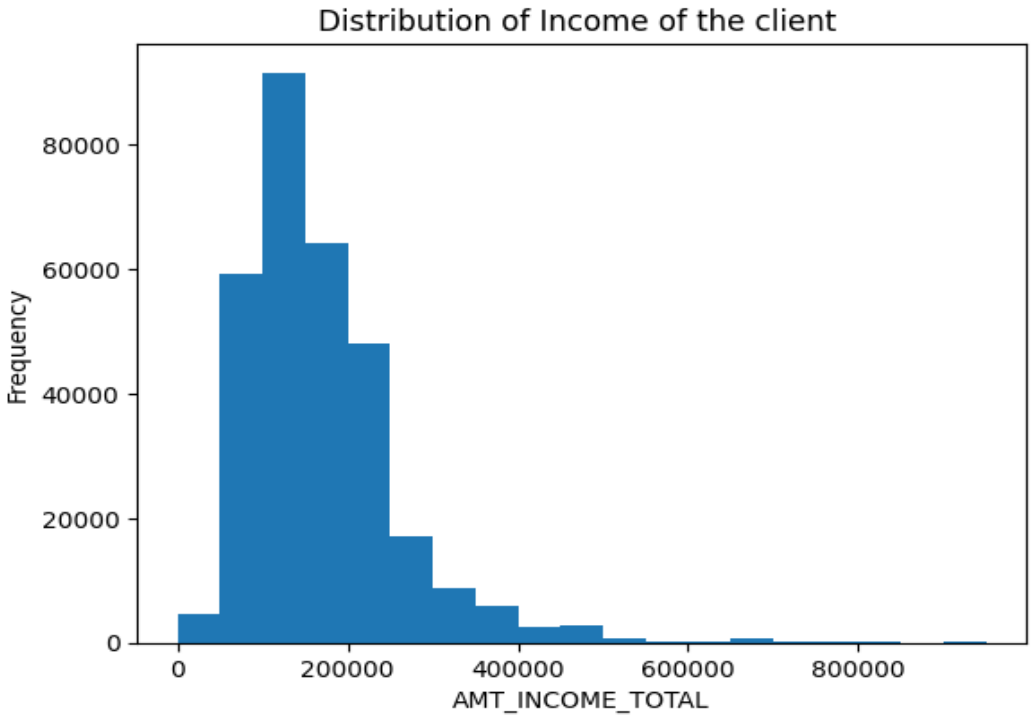
- **Housing Type Distribution**:
  House / Apartment: 272,868 (85.8%)
  With Parents: 14,840 (4.7%)
  Municipal Apartment: 11,183 (3.5%)
  Rented Apartment: 4,881 (1.5%)
  Office Apartment: 2,617 (0.8%)
  Co-op Apartment: 1,122 (0.4%)
- **Predominance of Home Ownership**:
  Insight: A significant majority of applicants (85.8%) reside in their own house or apartment, which often indicates greater financial stability and may correlate with lower default rates.

Credit Amount Distribution



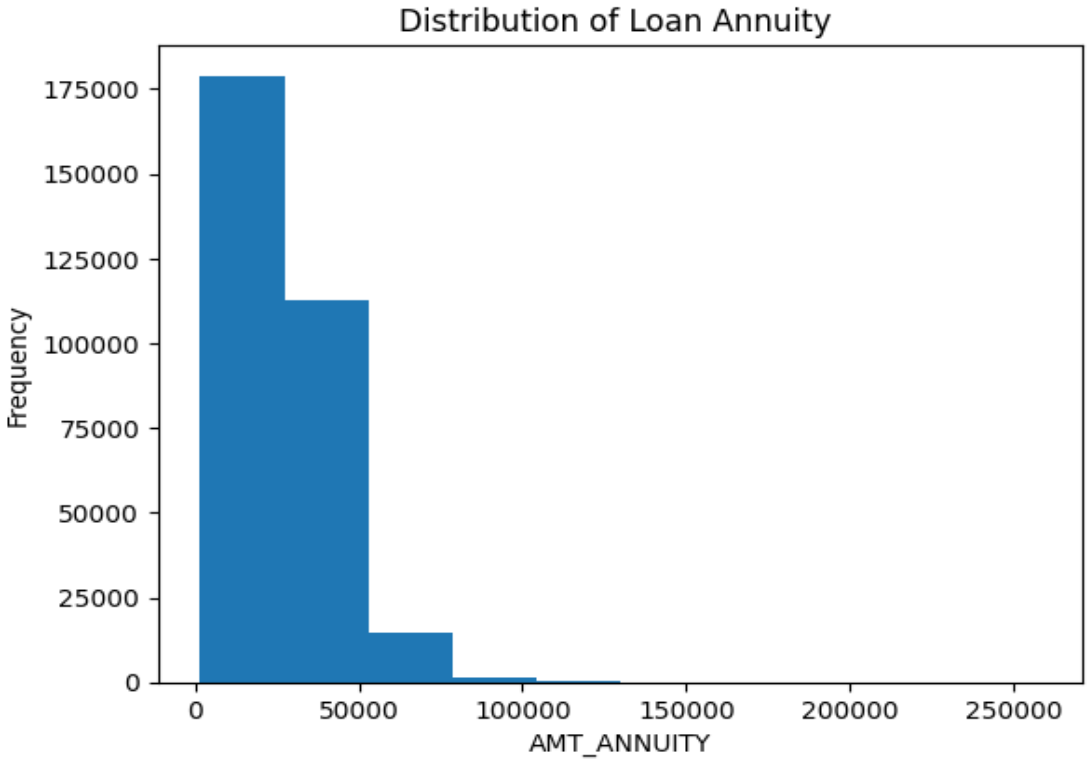
- **Credit Amount Categories**:
  Low Credit Amount: 149,333 (57.6%)
  Medium Credit Amount: 108,193 (41.4%)
  High Credit Amount: 49,985 (1.0%)
- **Dominance of Low Credit Amounts**:
  Insight: The majority of applicants (57.6%) are classified under low credit amounts, indicating a preference or need for smaller loan sizes, possibly reflecting cautious borrowing behavior or financial constraints.

Income Distribution Among Clients



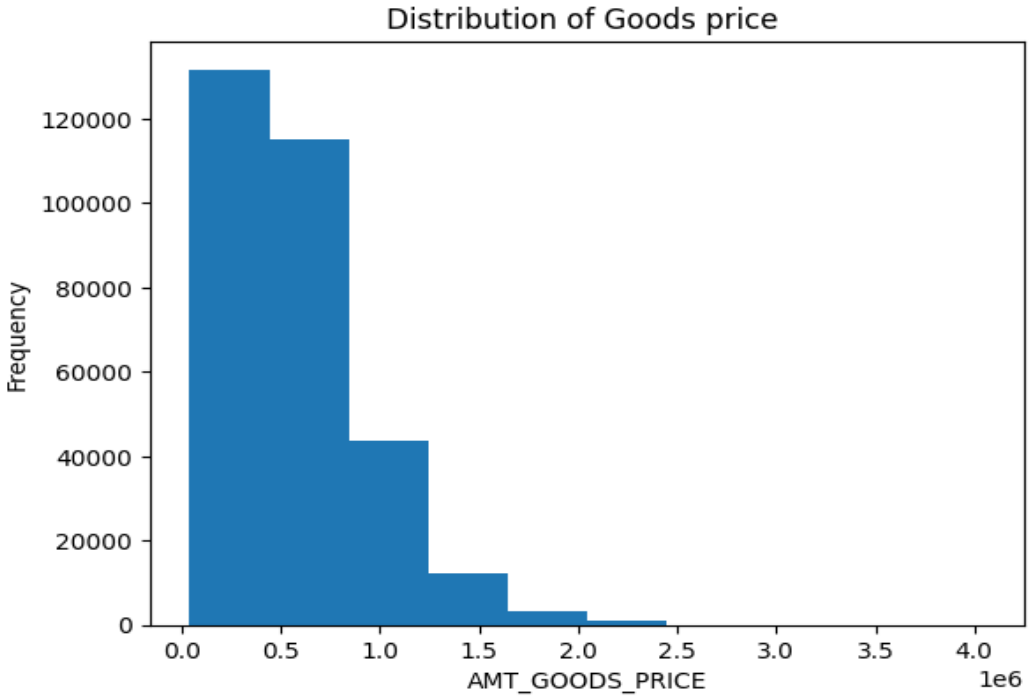
1. **High Concentration of Low Incomes**: Most clients have a total income between 0 and 200,000, with the highest frequency in this range.
2. **Decrease with Higher Incomes**: The frequency decreases as the income increases, indicating fewer clients with higher incomes.
3. **Minimal High Incomes**: Very few clients have an income above 600,000.

Loan Annuity Distribution Among Clients



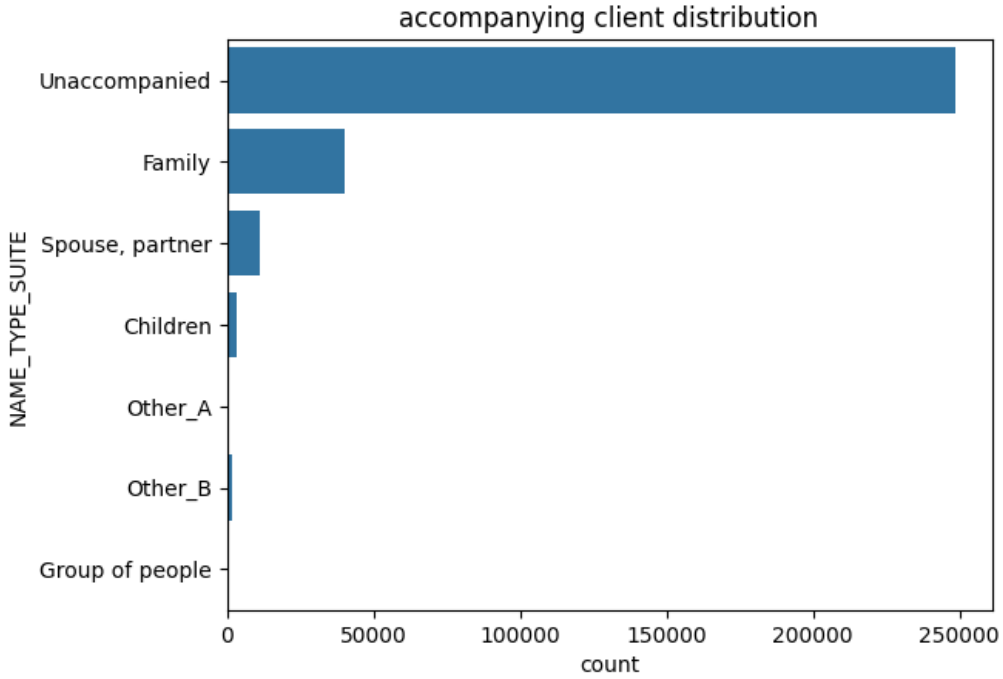
1. **Concentration of Low Annuities**: Most loan annuities are concentrated between 0 and 50,000.
2. **Steep Decline**: There's a sharp decline in the frequency of annuities above 50,000.
3. **Minimal High Annuities**: Very few data points exist beyond 100,000.

Goods Price Distribution



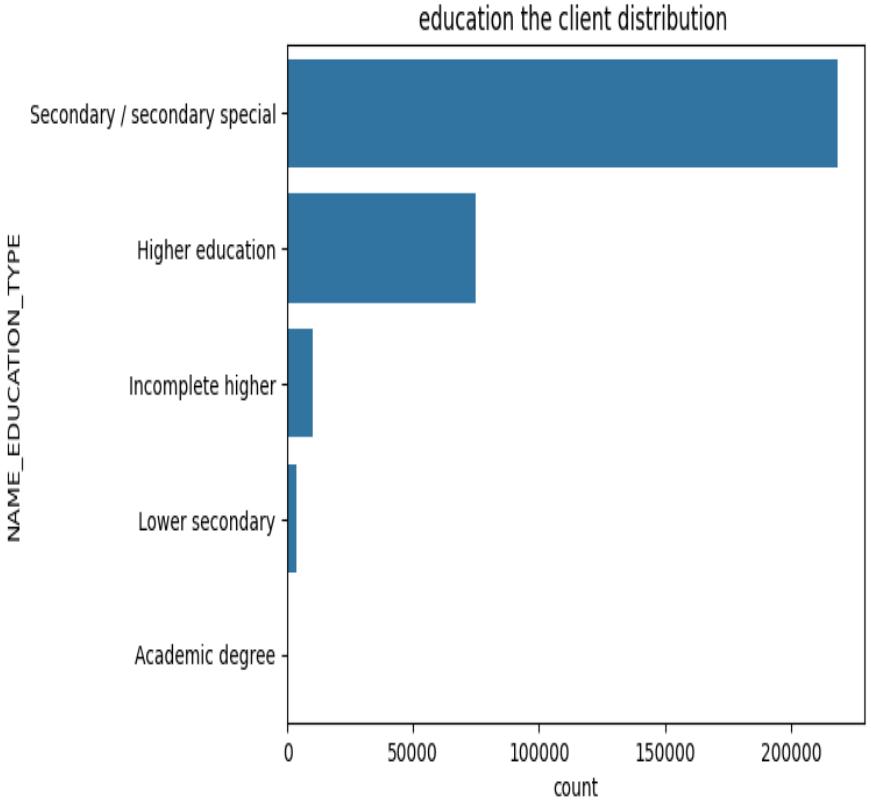
1. **Clustered at Low Prices**: The majority of goods are priced between 0 and 1,000,000.
2. **Rapid Drop-Off**: There's a noticeable decline in frequency as prices increase beyond 1,000,000.
3. **Few High-Priced Goods**: Very few goods are priced above 2,000,000.

Accompanying Client Distribution



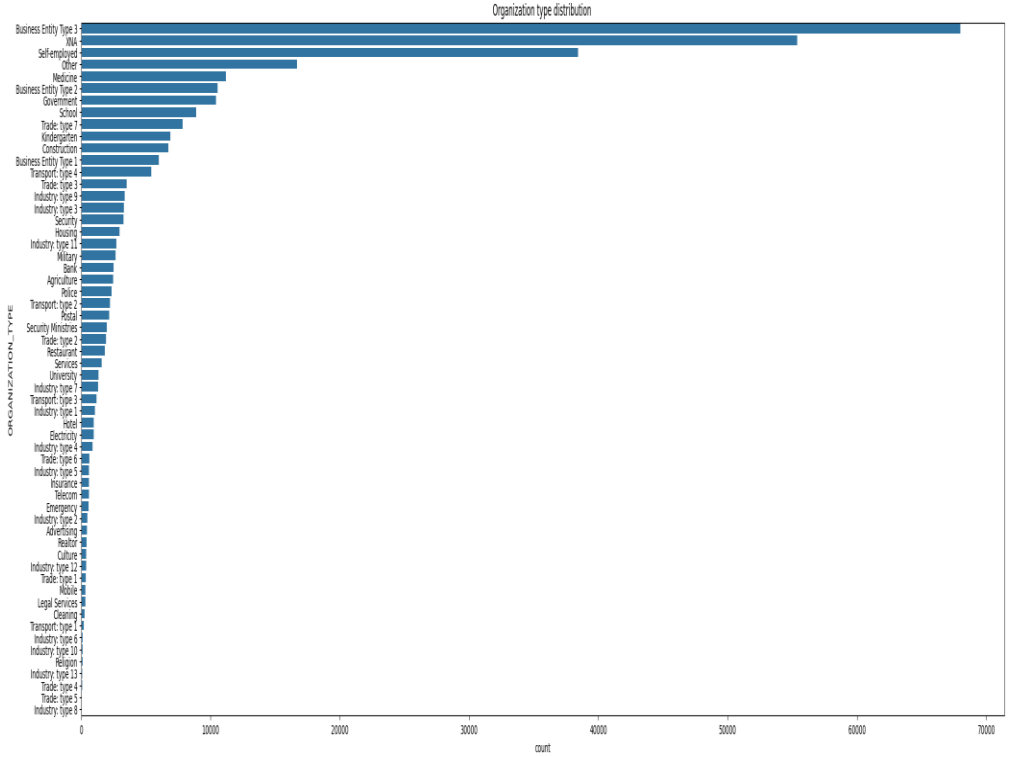
1. **Majority Unaccompanied**: Nearly 250,000 clients are unaccompanied.
2. **Family Comes Second**: Around 50,000 clients are accompanied by family.
3. **Minimal Other Categories**: Significant drop in numbers for spouse/partner (10,000), children (2,000), and others, with “Group of people” barely represented.

Education of Client Distribution



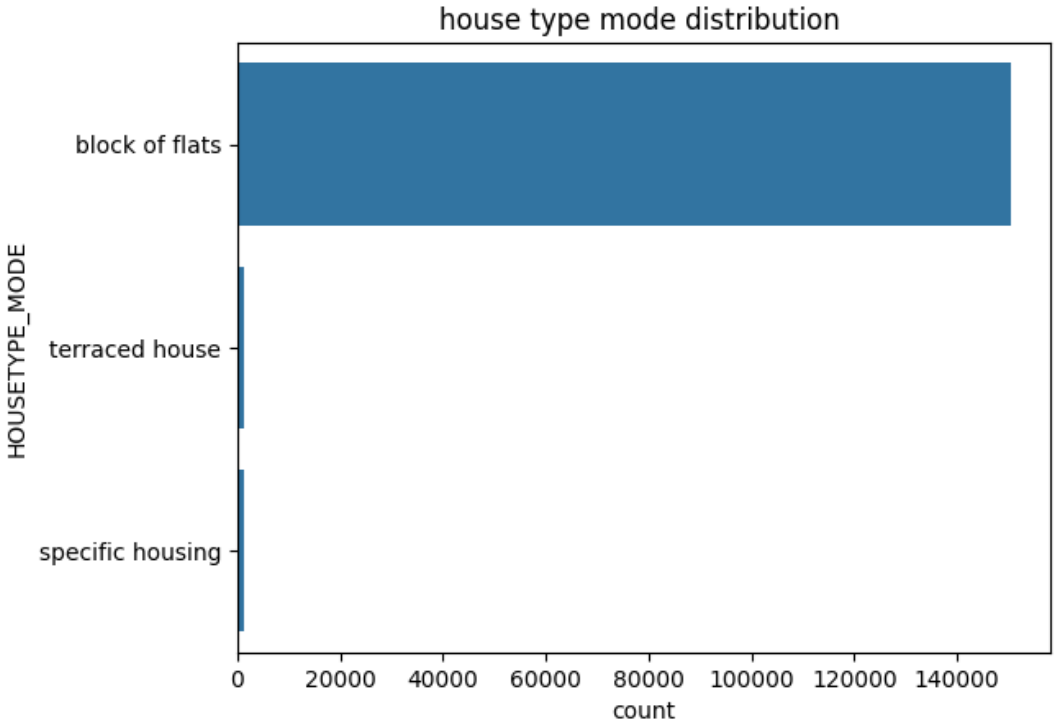
1. **Dominance of Secondary Education**: The vast majority of clients have a "Secondary / secondary special" education level, with over 200,000 clients.
2. **Significant Higher Education Group**: "Higher education" follows, with around 100,000 clients.
3. **Few with Incomplete or Lower Secondary Education**: Both "Incomplete higher" and "Lower secondary" education levels have significantly fewer clients, around 20,000 and slightly above 5,000 respectively.
4. **Minimal Academic Degrees**: Barely any clients hold an "Academic degree," the count being close to zero.

Organization Type Distribution



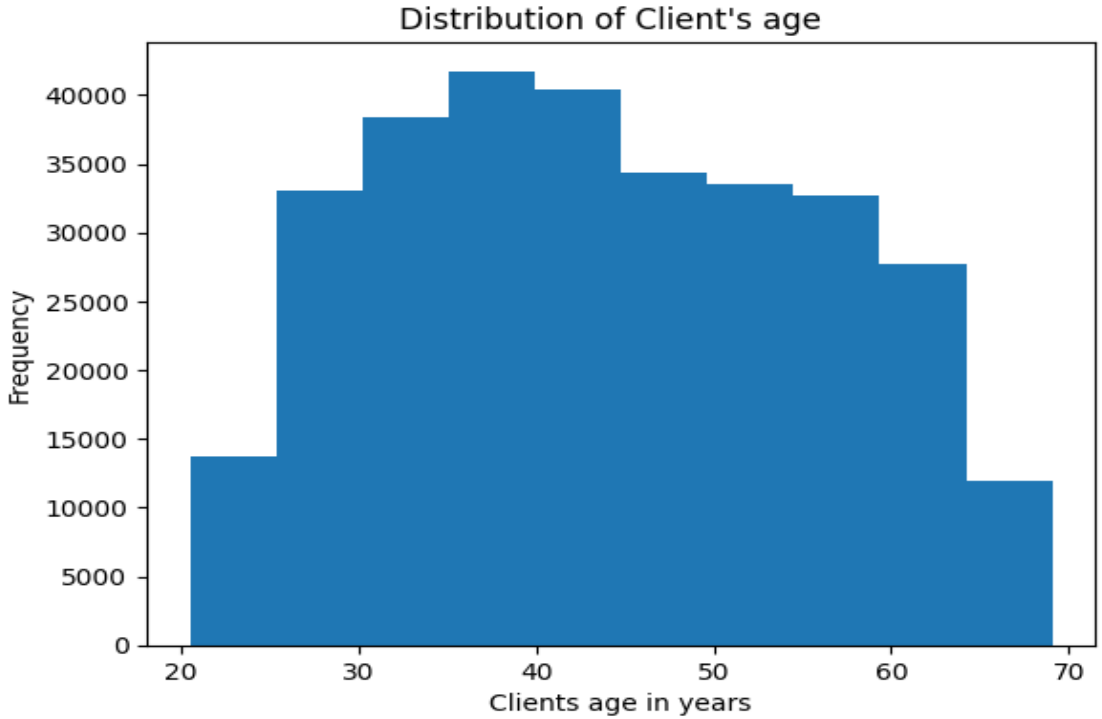
1. **Dominance of Business Entity Type 3**: It's the most common organization type, with a count approaching 70,000.
2. **High Counts for Others**: "Other" and "Self-employed" categories are also significant, each exceeding 40,000.
3. **Drop in Counts**: After the top three categories, there's a notable drop; "Medicine" and "Business Entity Type 2" count around 20,000.

House Type Distribution



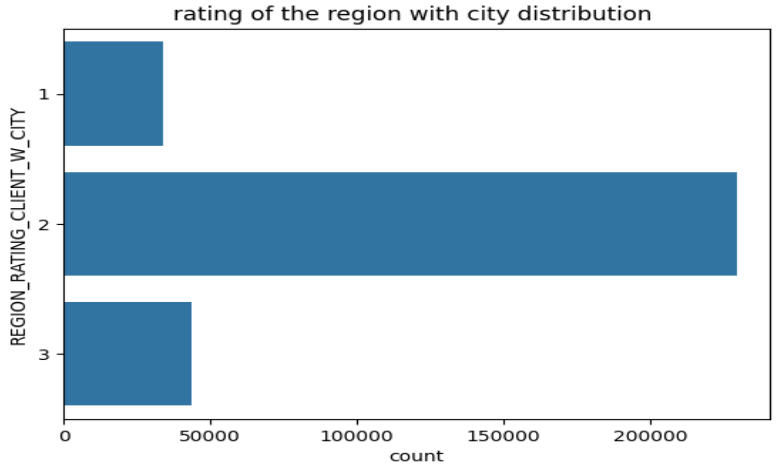
1. **Predominance of Flats**: The "Block of Flats" category towers above the others, making it the most common house type by a substantial margin.
2. **Scarcity of Other Types**: Both "Terraced House" and "Specific Housing" have minimal representation compared to "Block of Flats."
3. **Clear Preference Trend**: The data indicates a strong preference or prevalence for living in flats.

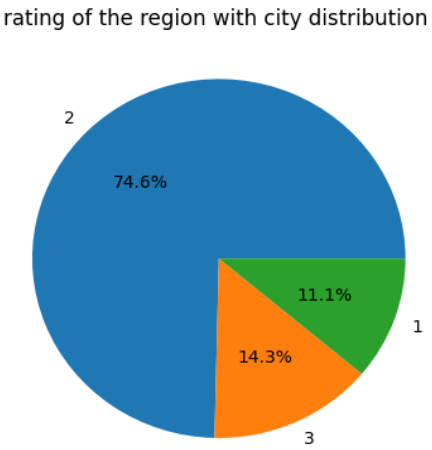
Client Age Distribution



1. **Symmetrical Distribution**: Clients' ages, in days, are symmetrically distributed around the 40-year mark.
2. **Peak at 40 Years**: The hig hest frequency of clients occurs at around 40 year of age.
3. **Balanced Tapering**: The f requency of clients decre ases evenly on both sides from the peak.

Rating of the Region City Distribution

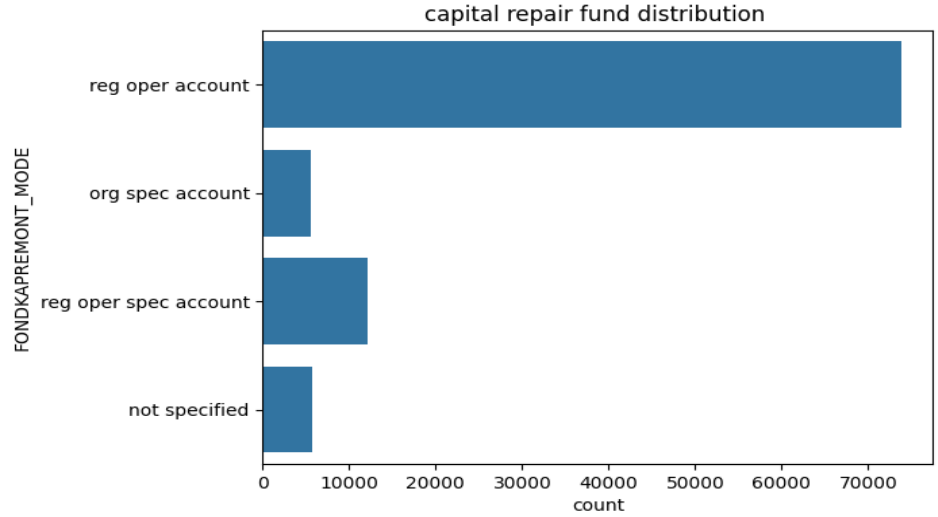


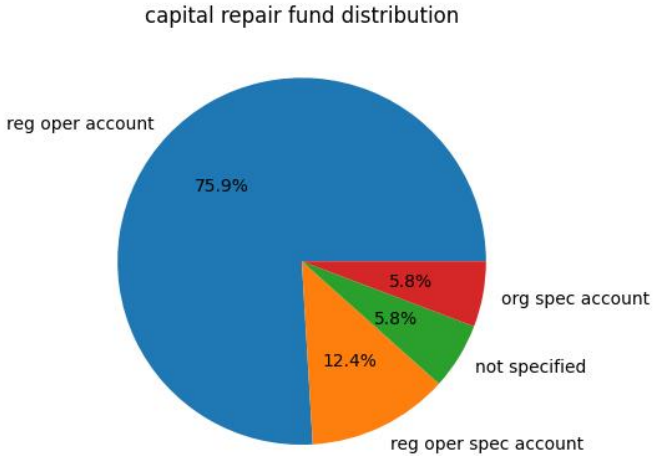


1. **Rating 2 Dominates**: Rating 2 has the highest count, exceeding 200,000.
2. **Significant Gap**: Rating 3 follows but is much lower, a little above 50,000.
3. **Rating 1 is Lowest**: Rating 1 has the lowest count, slightly above 25,000.
4. **Heavy Skew**: The distribution is heavily skewed towards Rating 2, indicating most regions fal l into this category.

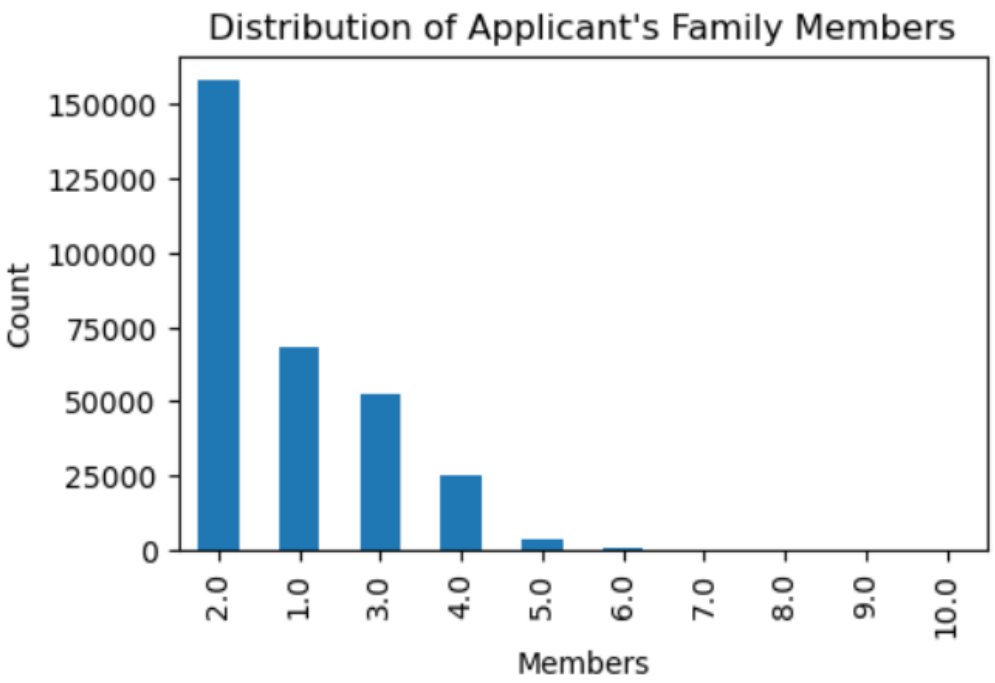
Capital Repair Fund Distribution

1. **Dominance of 'Reg Oper Account'**: This mode has the highest count, approximately 7 0,000.
2. **Moderate Counts for 'Org Spec Account' & 'Reg Oper Spec Account'**: They stand at around 10,000 and 15,000, respectively.
3. **Lowest Count for 'Not Specified'**: This mode has the lowest count, around 5,000.



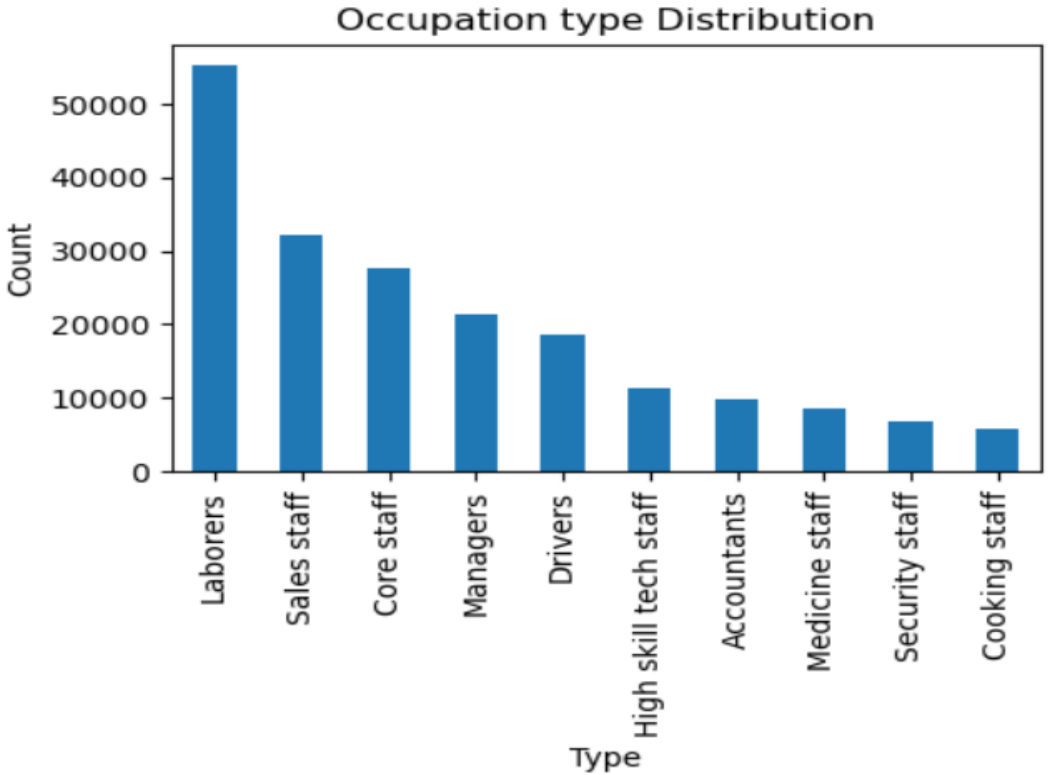


Applicant's Family Members Distribution



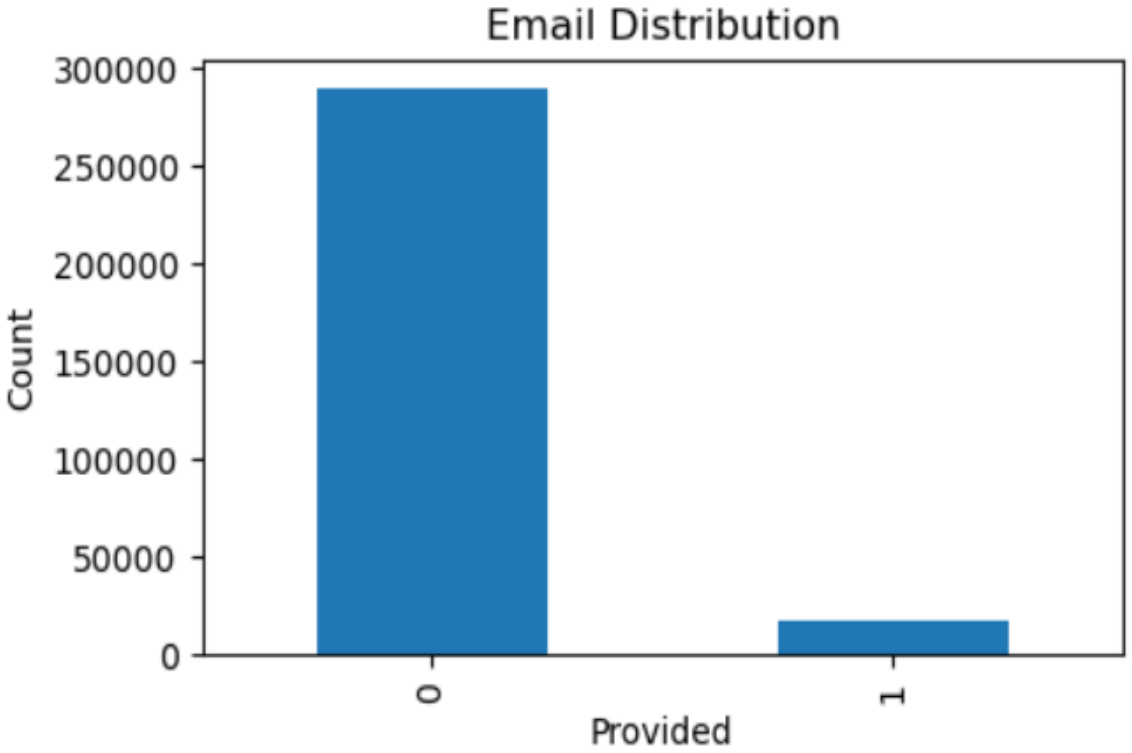
- **Most applicants have small Families:** The majority of applicants have 2 family members 51%, followed by 1 family member 22% and 3 family members 17%.
- **Larger Families Are Less Common:** Families with 4 or more members are much less frequent, with a significant drop
- **Very Large Families Are Rare**: Families with 6 or more members are extremely rare, indicating that most applicants come from smaller family units.

Occupation Type Distribution



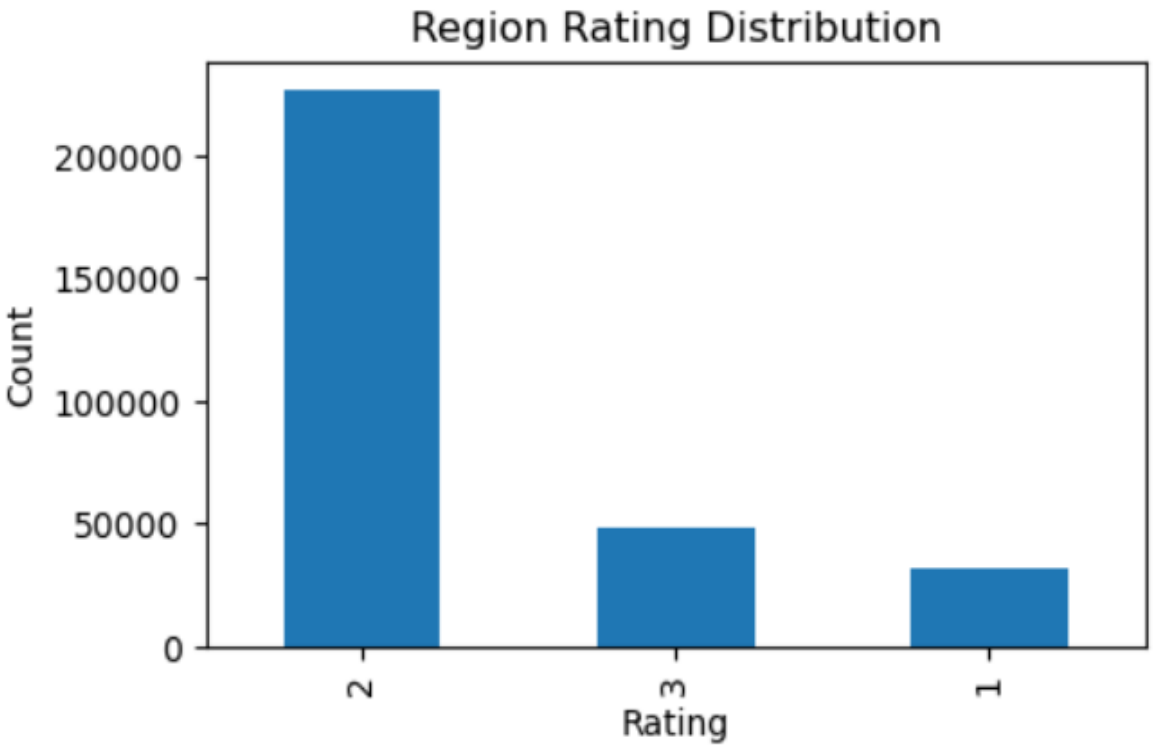
- **Laborers** are the most common occupation, with over 26% individuals.
- **Sales staff** and **Core staff** follow as the second and third largest groups with 15% & 13%.
- **Managers** and **Drivers** are moderately represented, each with significant percent.
- **High skill tech staff**, **Accountants**, **Medicine staff, Security staff** , **Cooking staff** make up smaller portions, indicating these are less common occupations.

Applicant's Email Distribution



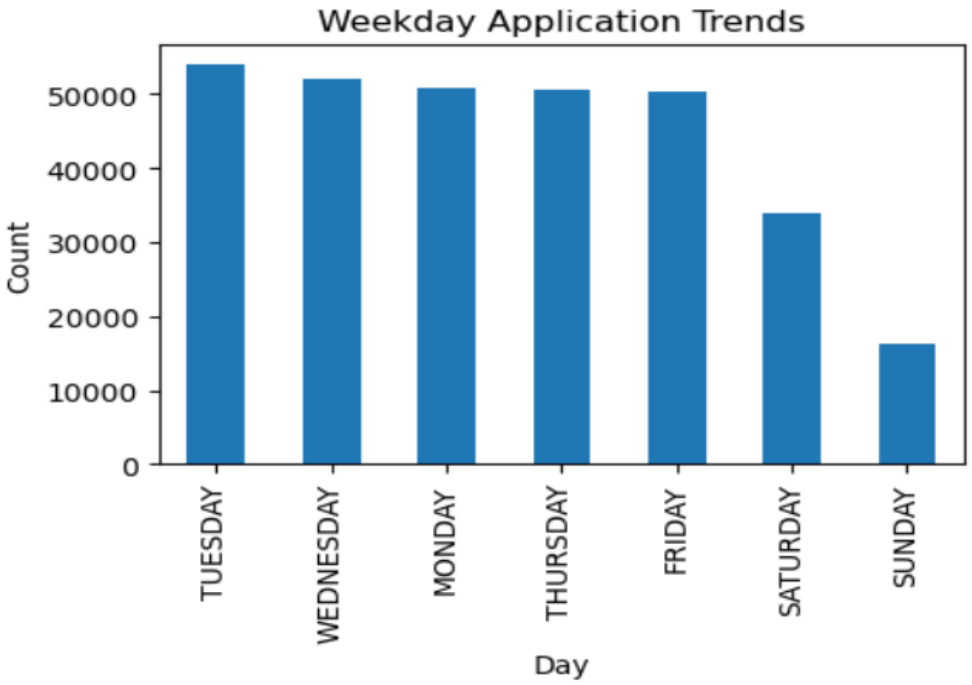
- **Majority Missing Emails**: A large portion of the data shows that emails are not provided, with over 250,000 (94%) instances.
- **Few Emails Available**: Only a small number of cases have emails provided (6%), indicating a limited collection of email contact information.

Region Rating Distribution



- **Most Regions Rated 2**: The majority of regions with 74% have a rating of 2.
- **Moderate Count for Rating 3**: A smaller group16% has a rating of 3,
- **Fewest Regions Rated 1**: The lowest count of regions, under 10%, falls into the rating 1 category

Weekday Application Trend



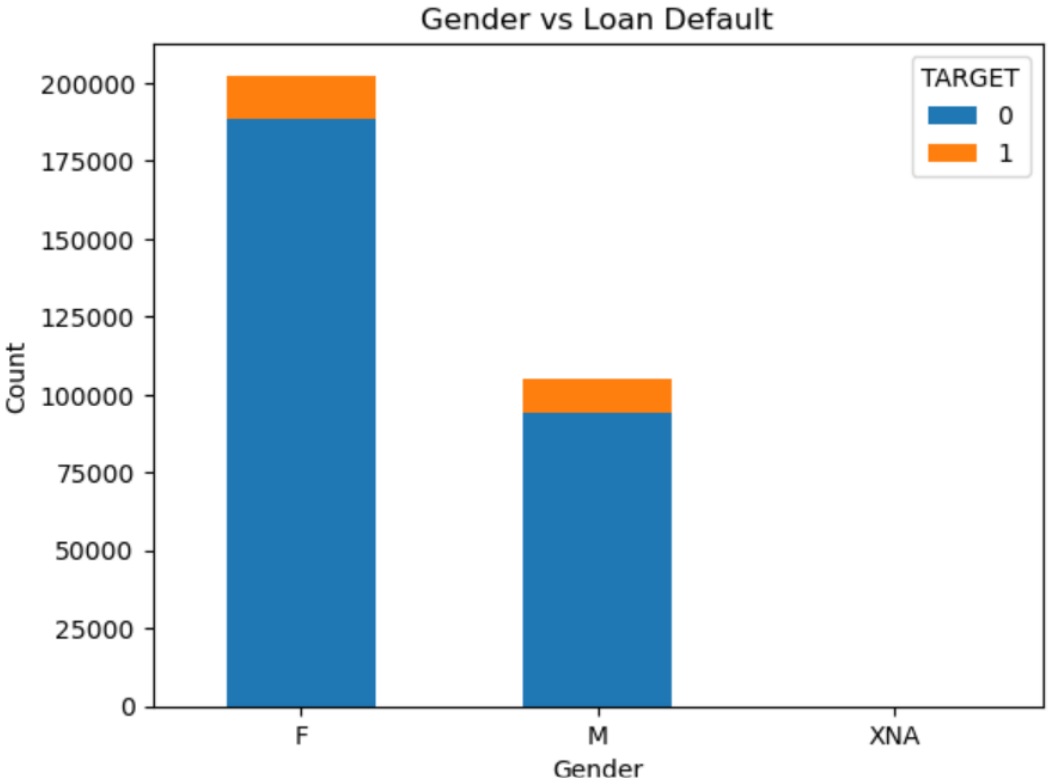
Weekday Trend:

- Tuesday & Wednesday has the highest number of applications 18% each.
- On Wednesday, Thursday and Friday, the number of applications is fairly consistent, 16% each.

Weekend Drop:

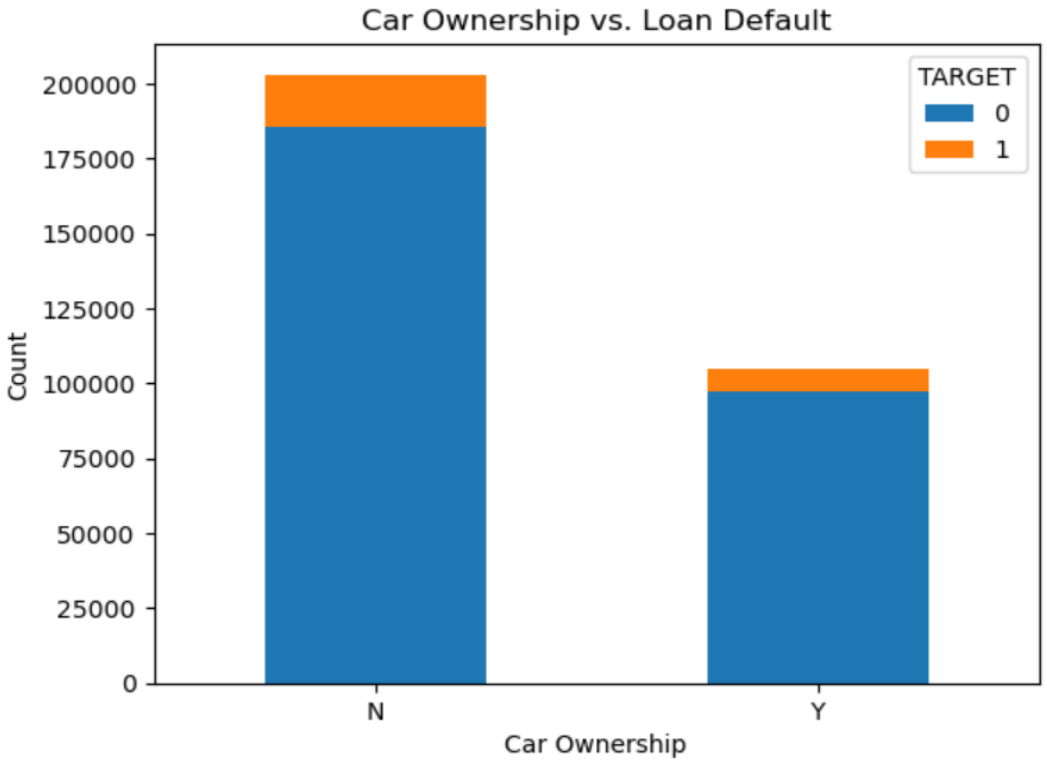
- Saturday's count is notably lower than the weekdays, at about 11%
- Sunday has the lowest count, with applications dropping to 5%

Gender vs Loan Default



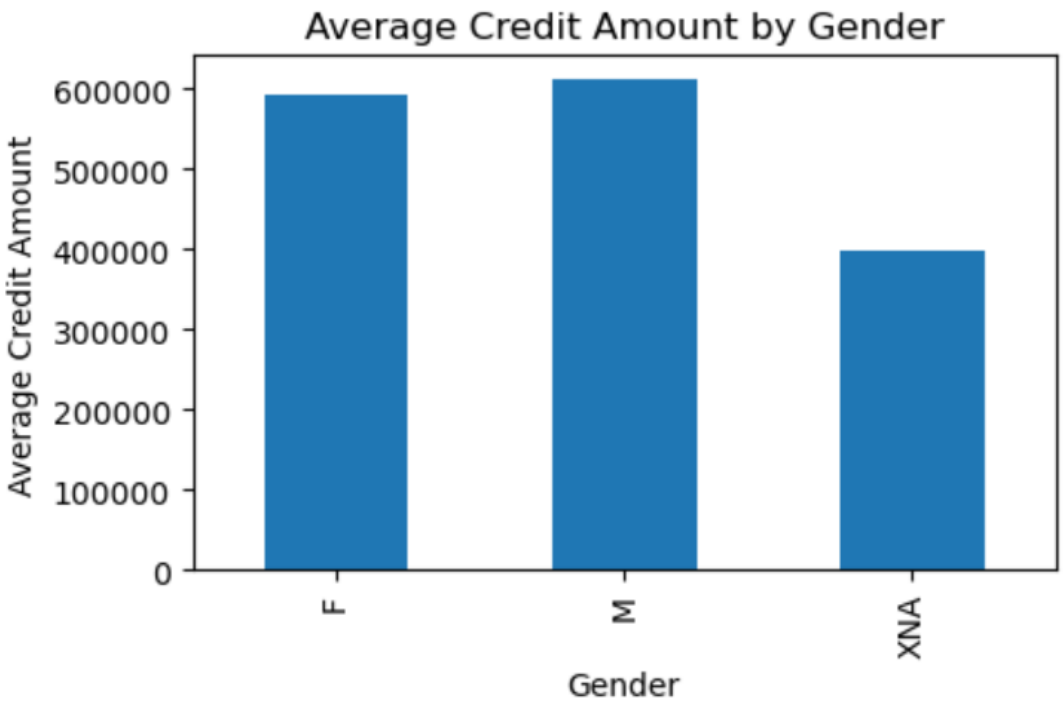
- **Female Default Rate(1)**: 7.0%
- **Male Default Rate(1)**: 10.1%
- Males exhibit a higher default rate compared to females, indicating that gender may be a factor in assessing loan repayment risk

Car Ownership vs. Loan Default



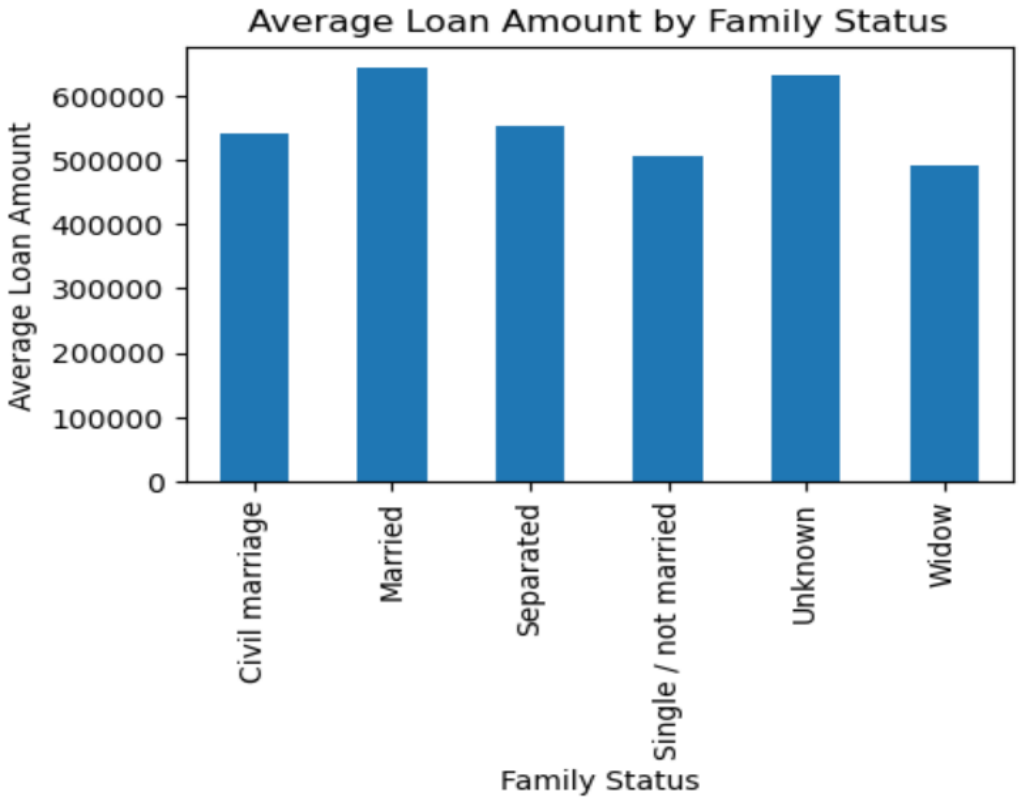
- **No Car Default Rate(1)**: 8.5%
- **Car Ownership Default Rate(1)**: 7.2%
- Borrowers who own a car have a slightly lower default rate compared to those who do not, suggesting that car ownership might be associated with more stable financial situations.

Average Credit Amount by Gender



- **Average Credit Amount**:
  Female Borrowers (F): $592,766.72
  Male Borrowers (M): $611,095.20
  Unknown Gender (XNA): $399,375.00
- **Gender Comparison**:
  Insight: Male borrowers receive a higher average credit amount compared to female borrowers, with a difference of approximately $18,328.48. This may suggest different lending behaviors or risk perceptions between genders.

Average Loan Amount by Family Status

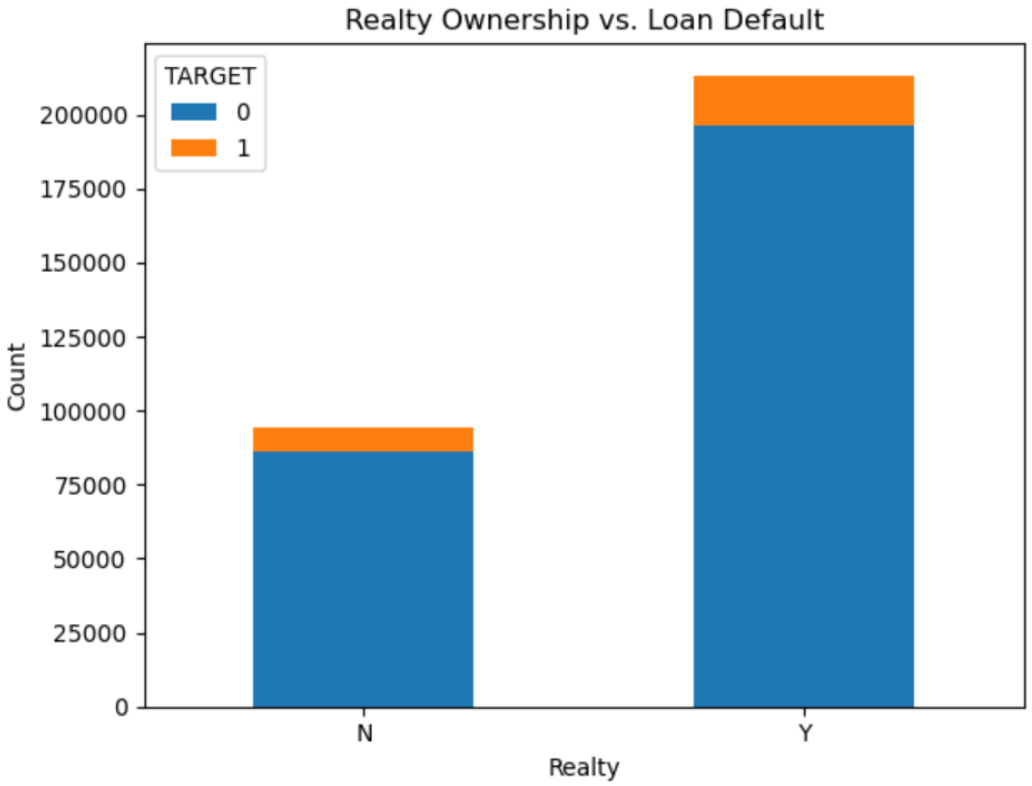


- **Average Loan Amounts by Family Status**:
  Married: $642,999.79
  Civil Marriage: $541,573.46
  Separated: $552,113.82
  Single / Not Married: $505,350.18
  Widow: $490,695.91
  Unknown: $630,000.00
- **Married Applicants Have the Highest Loan Amount**:
  Insight: Married applicants receive the highest average loan amount, suggesting they may be viewed as lower risk due to potential dual incomes and financial stability.

Amount Credit vs Amount Goods Price

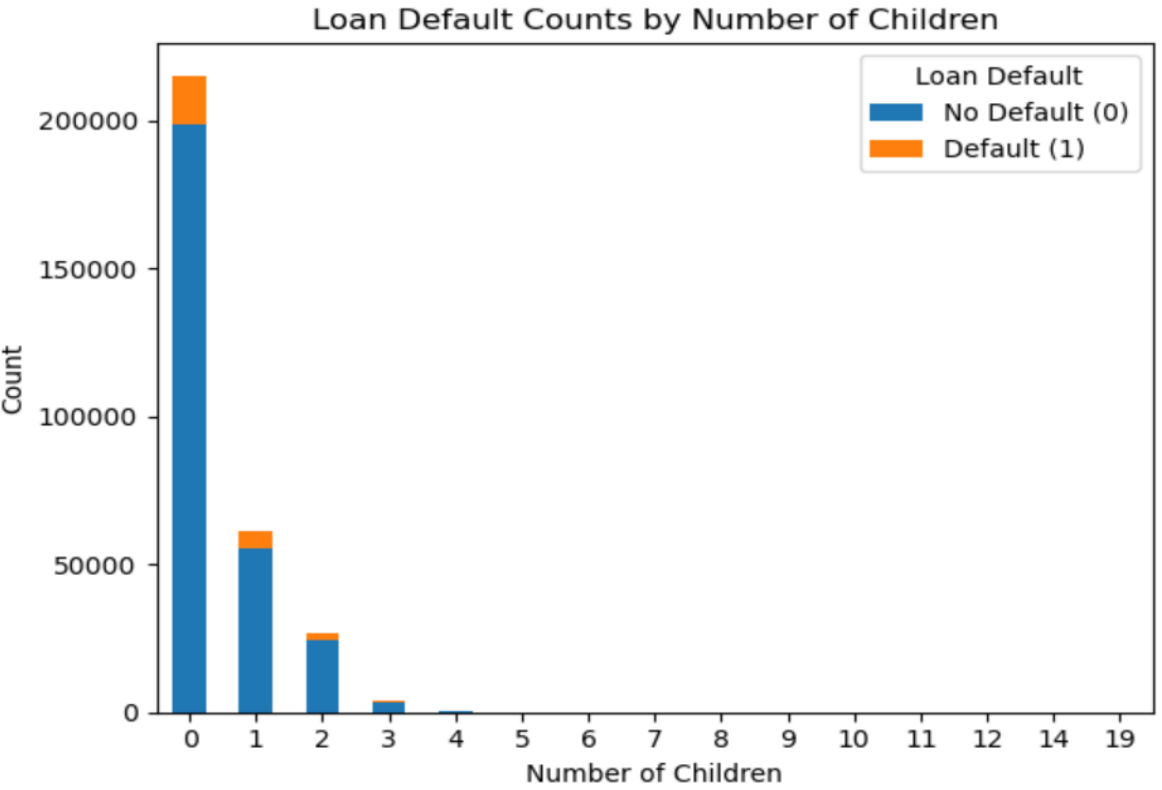
1. **Strong Positive Relationship**: The amount of credit increases proportionally with the amount of goods price.
2. **Dense Clustering at Lower Values**: Most transactions occur at lower values of both credit and goods price.
3. **Outliers Present**: Some data points significantly deviate from the general trend, with goods prices either much lower or higher than the corresponding amount of credit.

Realty Ownership vs. Loan Default



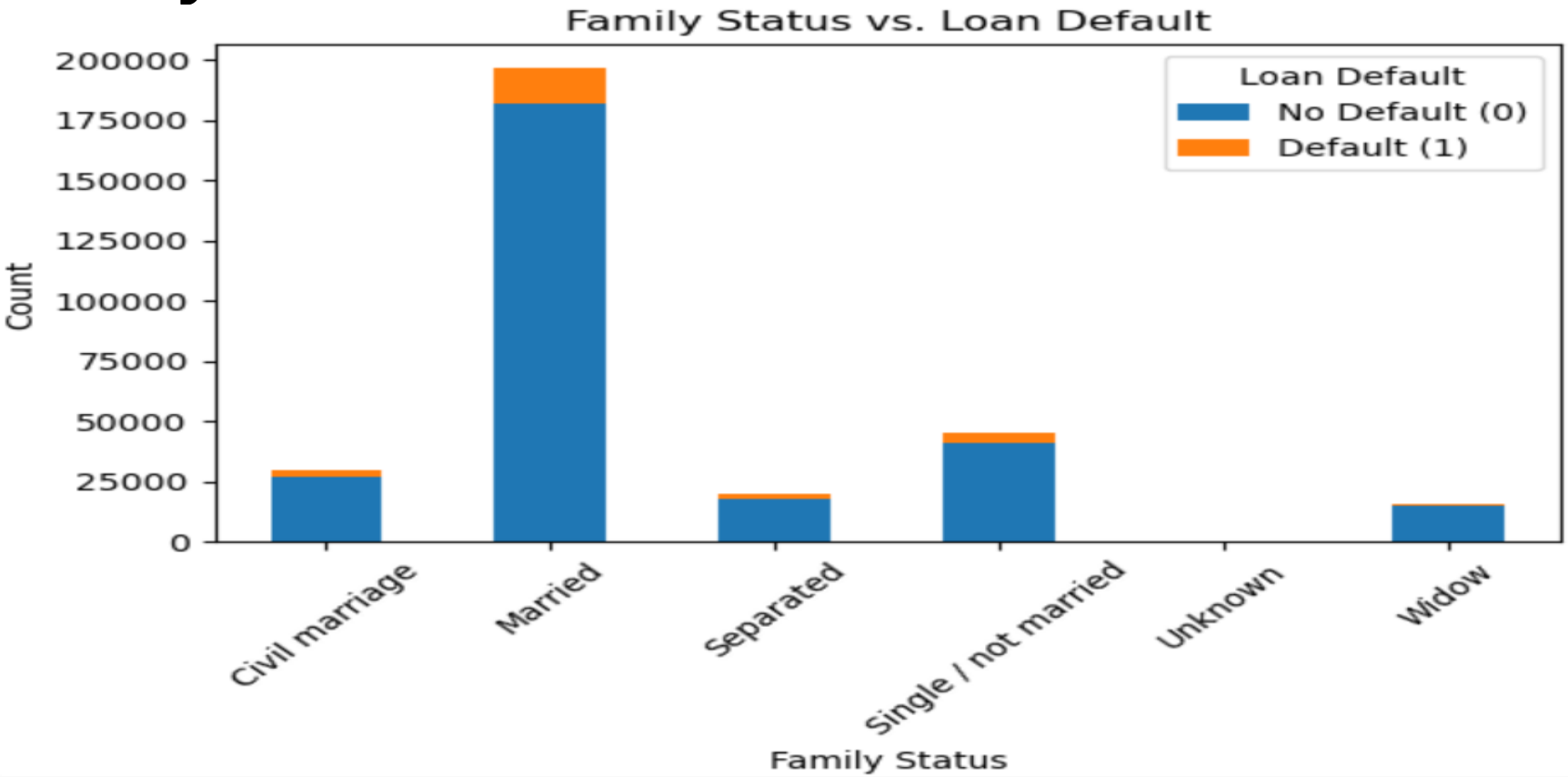
- Clients who own realty (homes) have a lower default rate. Out of **196,329** homeowners, **16,983** defaulted (about **8.6%**), compared to **7,842** defaults from **86,357** non-homeowners (about **9.1%**).
- Homeowners show a significantly higher number of non-defaults (**196,329**) compared to non-homeowners (**86,357**), indicating greater financial stability.

Number of Children vs. Loan Default



- **Clients with 0 children** have a default rate of approximately **8.36%**.
- **Clients with 1 child** show a slightly higher default rate of around **9.80%**.
- This trend continues, with higher default rates observed for borrowers with **3 or more children** (over **10%** for 3 children).

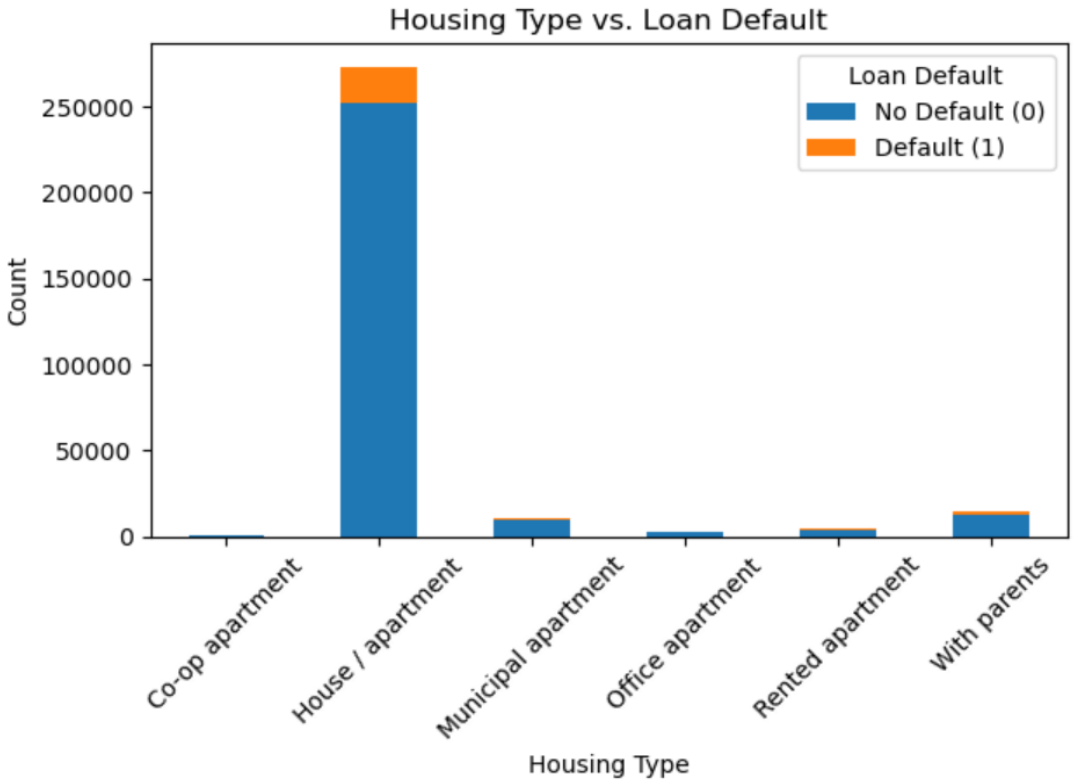
Family Status vs. Loan Default



Family Status vs. Loan Default

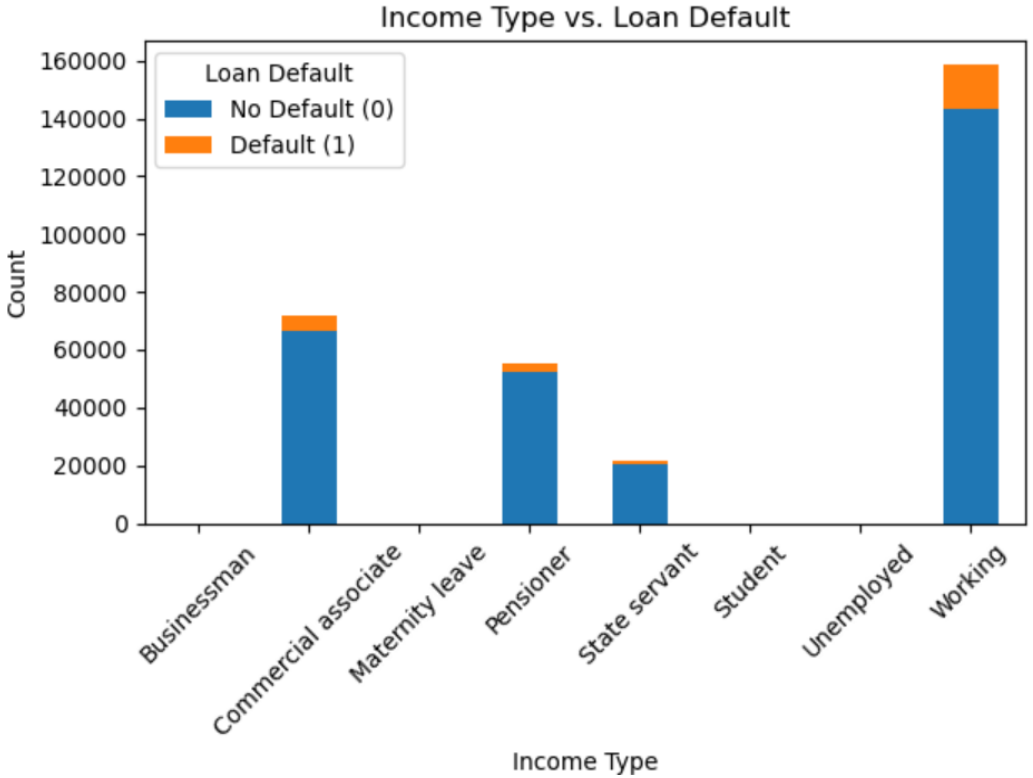
- **Married** borrowers have the lowest default rate at about **7.6%** (14,850 out of 181,582).
- **Civil marriage** follows with a default rate of approximately **10%** (2,961 out of 26,814).
- **Single/not married** borrowers show a default rate of about **9.8%** (4,457 out of 40,987).
- **Separated** individuals have a higher default rate of around **8.2%** (1,620 out of 18,150).
- **Widows** also display a notable default rate of **6.2%** (937 out of 15,151).

Housing Type vs. Loan Default



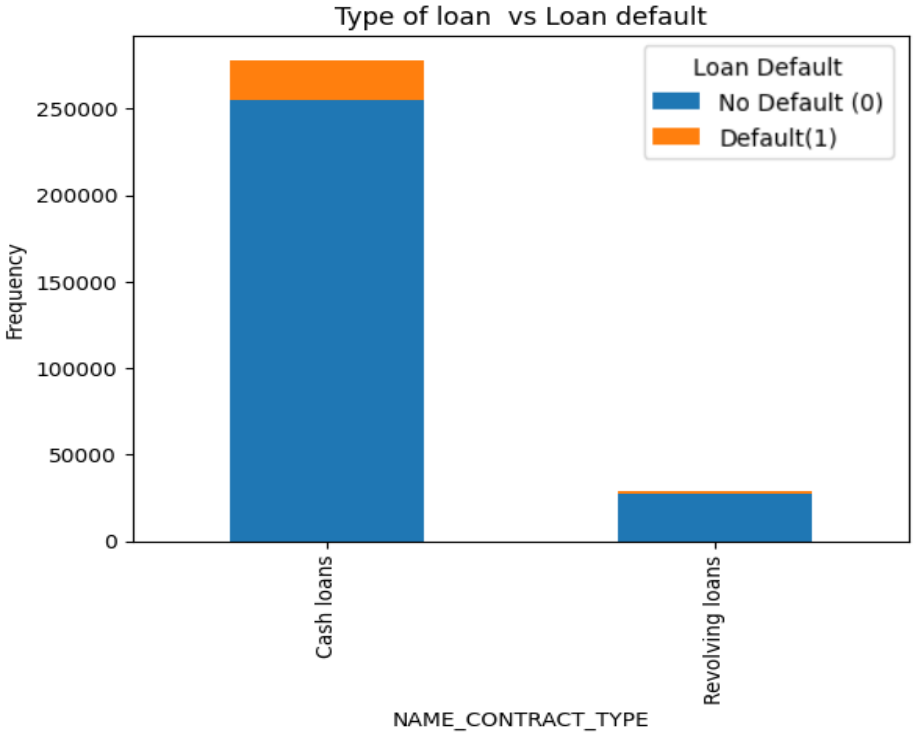
- **Higher Default Rates**: Borrowers living **with parents** (13.23%) and in **rented apartments** (14.04%) show the highest default rates, indicating financial vulnerability.
- **Moderate Risk**: Those in **municipal apartments** (9.34%) and **co-op apartments** (8.59%) also exhibit elevated default rates, though lower than renters and those living with parents.
- **Lower Default Rate**: The **house/apartment** category has the largest number of borrowers but a more moderate default rate of **7.8**, suggesting greater stability.

Income Type vs. Loan Default



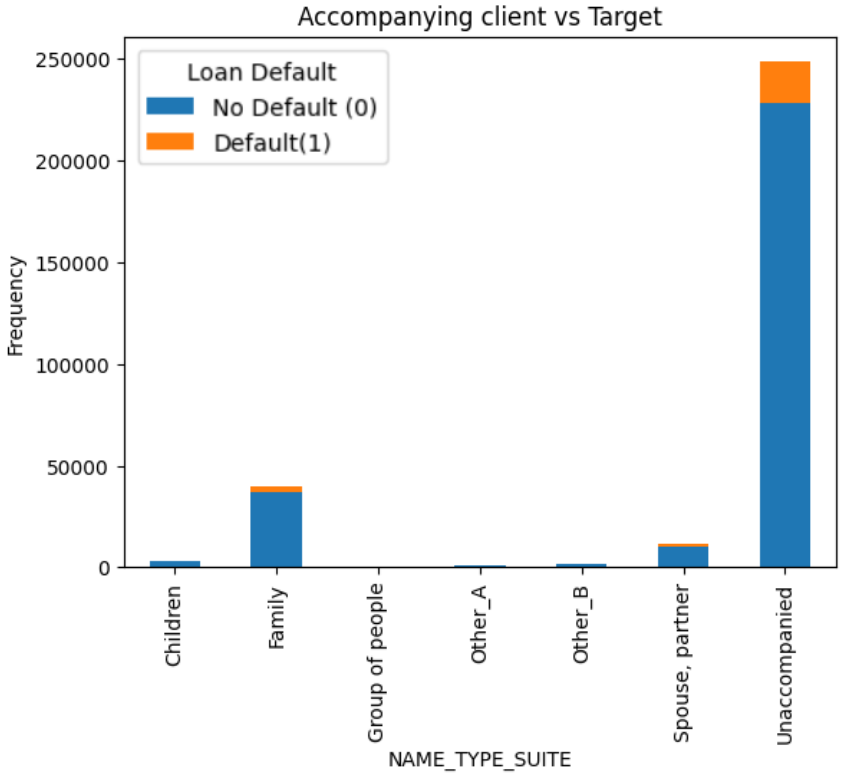
- **Highest Risk**: **Unemployed** borrowers have a default rate of **36.36%**, indicating significant financial vulnerability.
- **Moderate Risk**: **Pensioners** (5.38%) and **state servants** (6.11%) show moderate default rates, suggesting a potential risk.
- **Relatively Stable**: **Commercial associates** have a default rate of **7.46%**, while **businessmen** and **students** show minimal defaults due to low sample sizes.

Type of Loan vs Loan Default



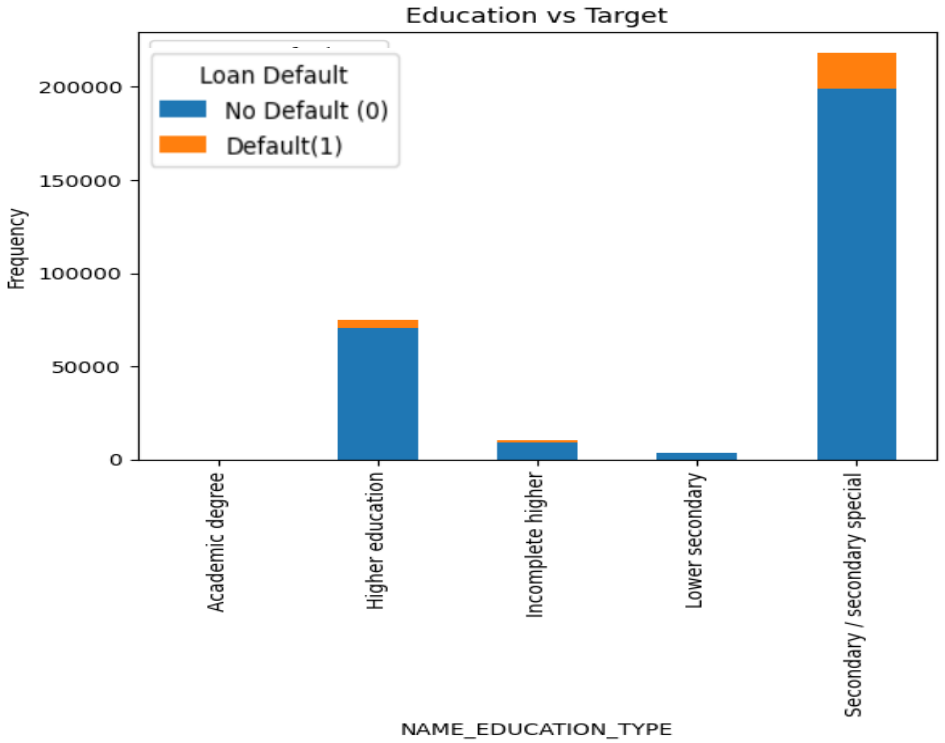
1. Overall, cash loans make up the vast majority of the loan data.
2. Default rate in cash loan is 8.34% and in revolving loan is 5.47%
3. Defaults are present in both loan types, but are more prominent in terms of volume in cash loans simply because there are many more of them.

Accompanying Client vs Loan Default



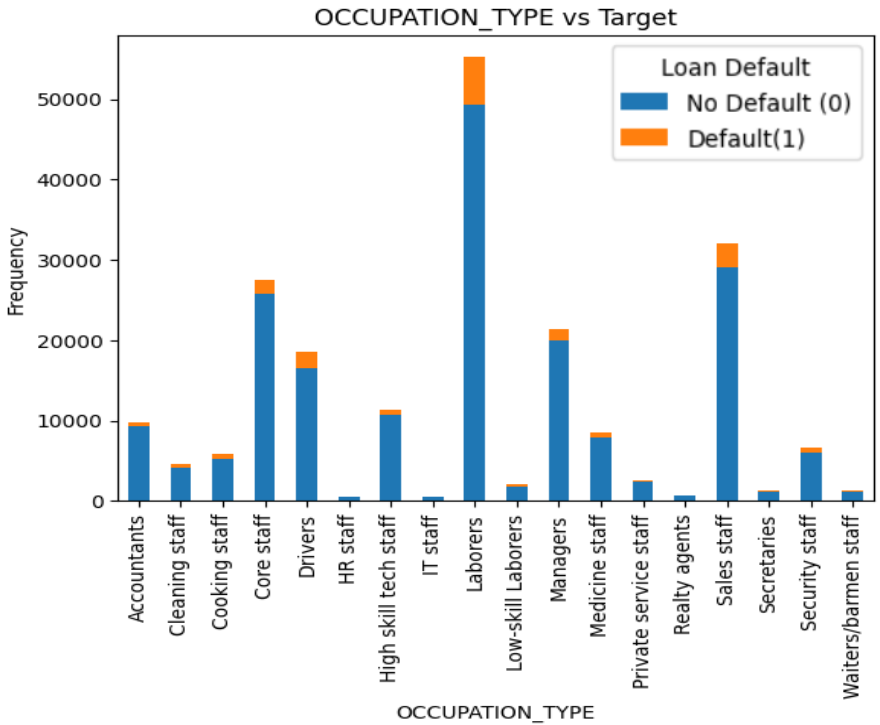
1. Overall, unaccompanied make up the vast majority of the loan data.
2. Default rate for Other_B is 9.83% followed by Other_A with 8.77% and Lowest in Children 7.37%
3. Defaults are present in all cases, but are more prominent in terms of volume in unaccompanied simply because there are many more of them.

Education vs Loan Default



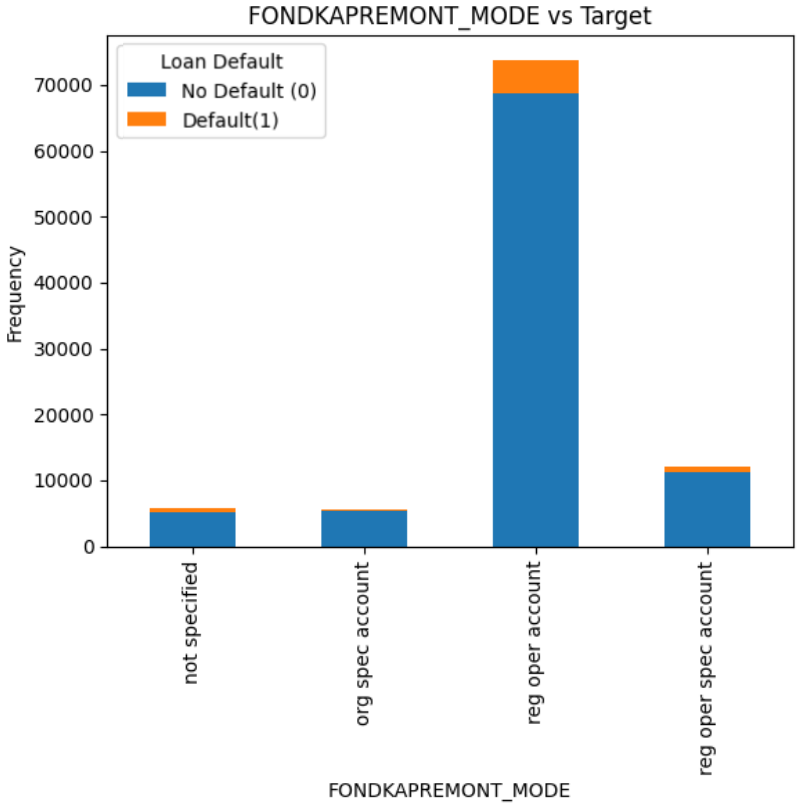
1. Default rate in Lower Secondary is 10.92% followed by secondary 8.93% and least in Academic Degree 1.82%
2. Individuals with lower educational qualifications (secondary education) are more likely to take loans, and also more likely to default.
3. Higher education appears to correlate with a lower default rate, despite the relatively high number of borrowers.

Occupation Type vs Loan Default



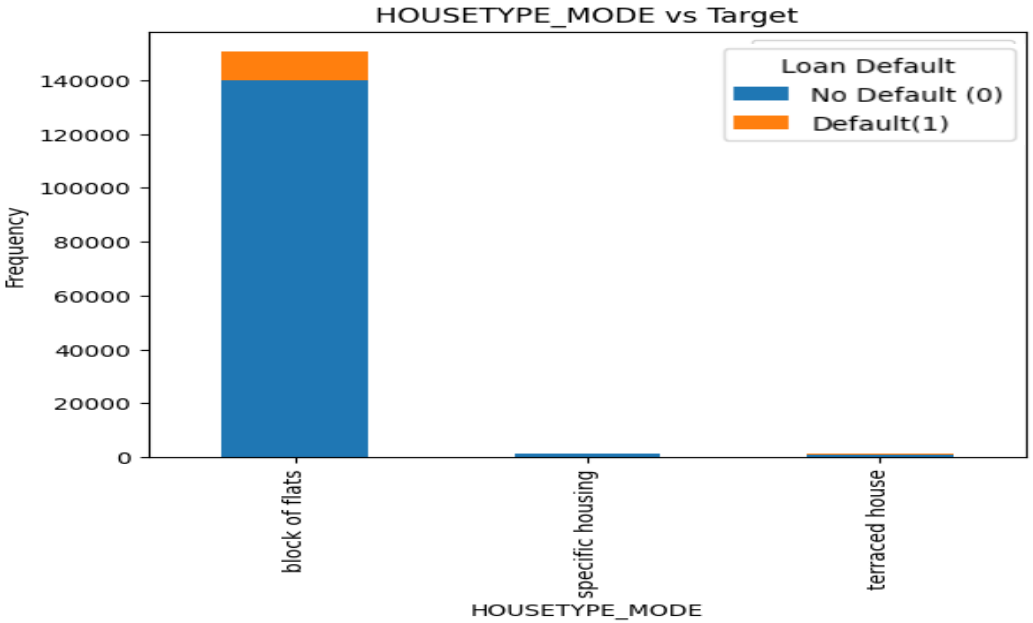
1. Default rate in Low-skill Laborers is 17.15% followed by Drivers is 11.32%
- lowest in Accountants is 4.83%
1. Laborers are mostly likely to take loan and mostly likely to default
2. High skill tech staff appears to have low default rate

Capital Repair Fund vs Loan Default



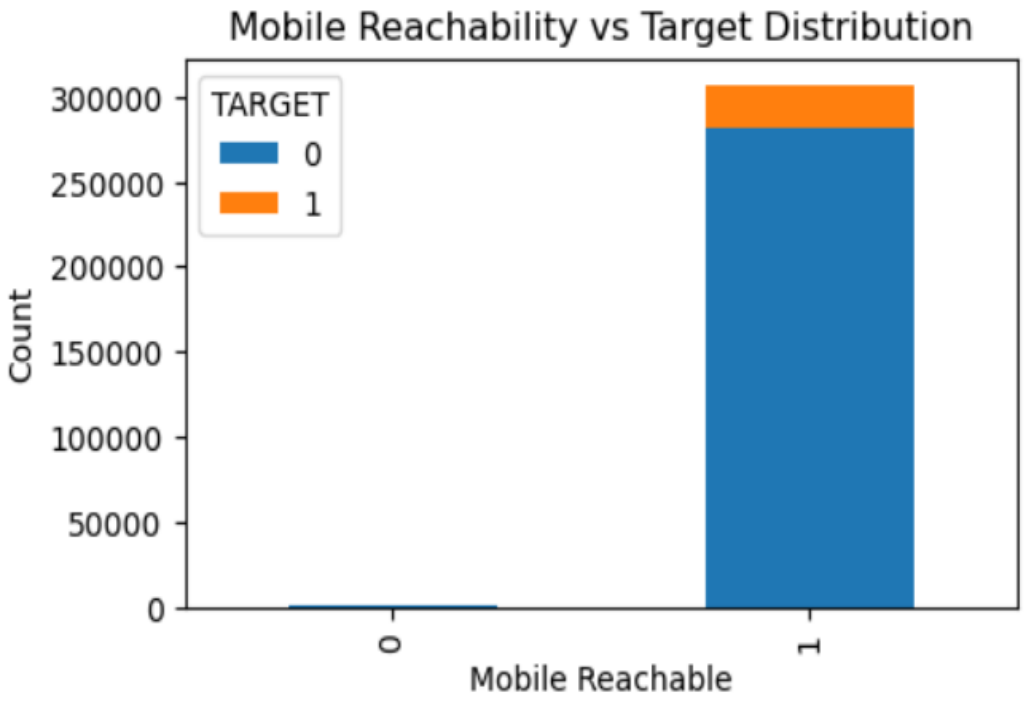
- Default rate for not specified is 7.54% followed by reg oper account 6.97% and
- lowest is in org spec account 5.81%
- Regular Operational Account is mostly present in the data with significant default rate due to its high presence

House Type Mode vs Loan Default



1. Most of client lives in the block of flats
2. Default rate for specific housing is 10.14% followed by terrace housing 8.49%
3. lowest in block of flats 6.94%

Mobile Reachability vs Loan Default

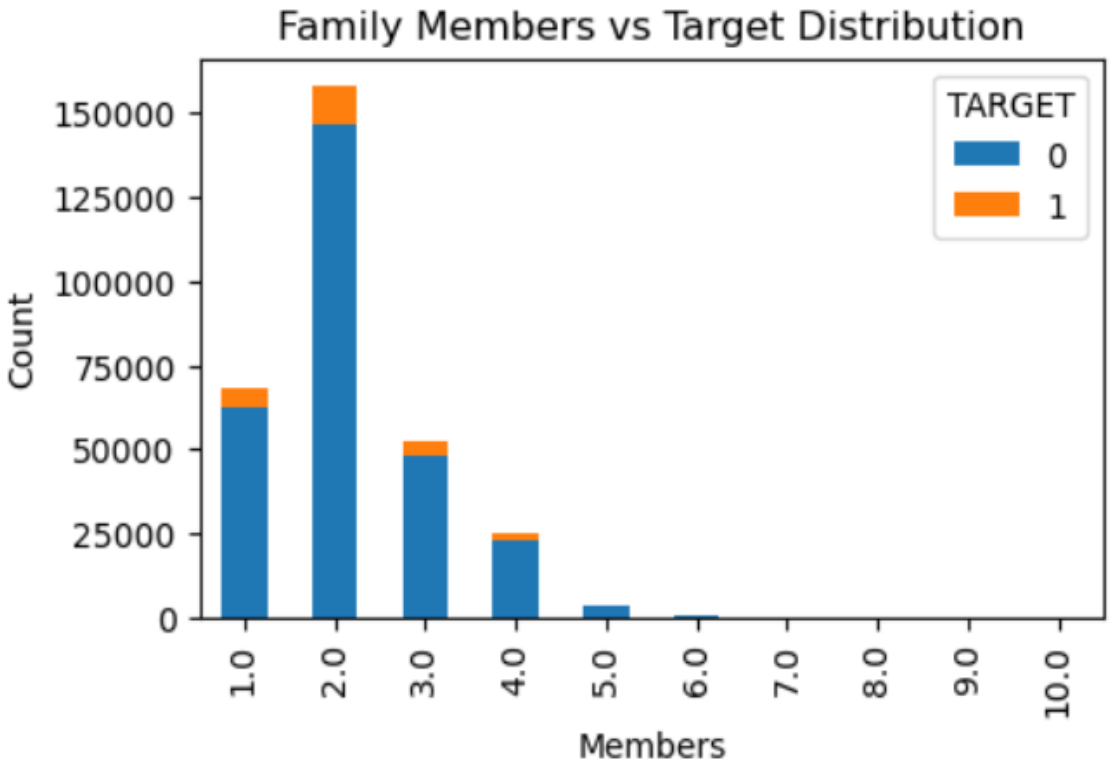


- **Mobile Default Group:**
- Represents a significant portion of defaulters 24,780 , accounting for approximately **8%** of the total population.
- **Non-Mobile Default Group:**
- Represents a much smaller portion of defaulters 45, accounting for approximately 0**.01%** of the total population.

Email Ownership vs Loan Default

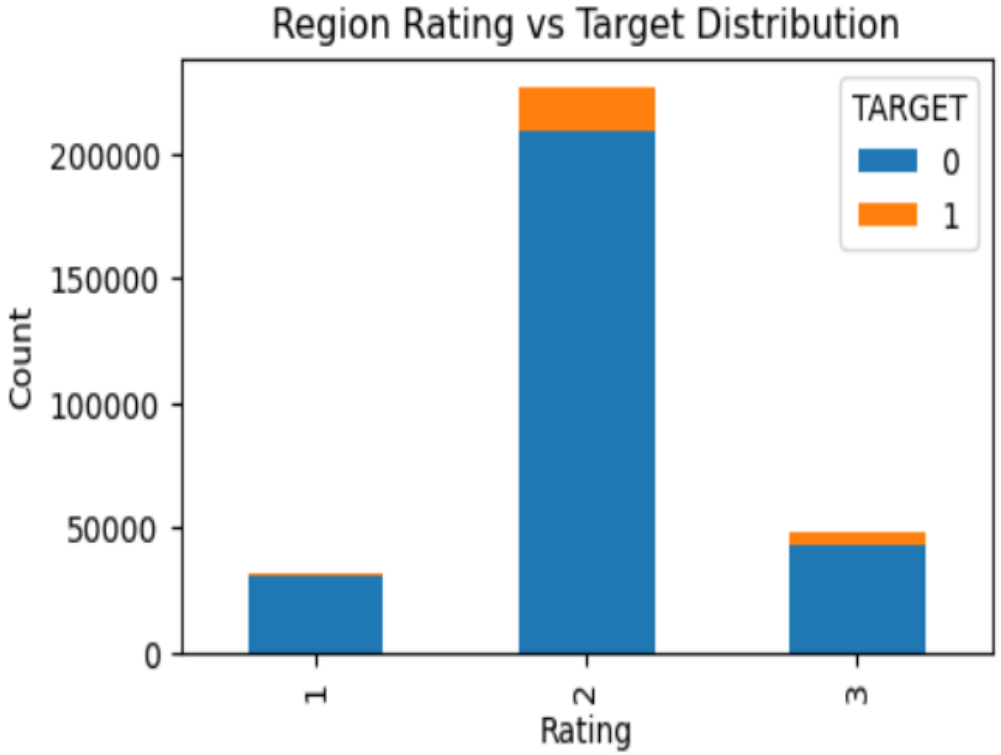
- **Non-Email Users:** 8.0%, 23,451, a significant number of non-email users default, indicating potential challenges in repayment.
- **Email Users:** 7.86%, 1374, the default rate for email users is slightly lower than that for non-email users.
- The data indicates that email communication may play a positive role in borrower engagement and repayment.
- **Overall Default Rate:**8.77%

Family Members vs Loan Default



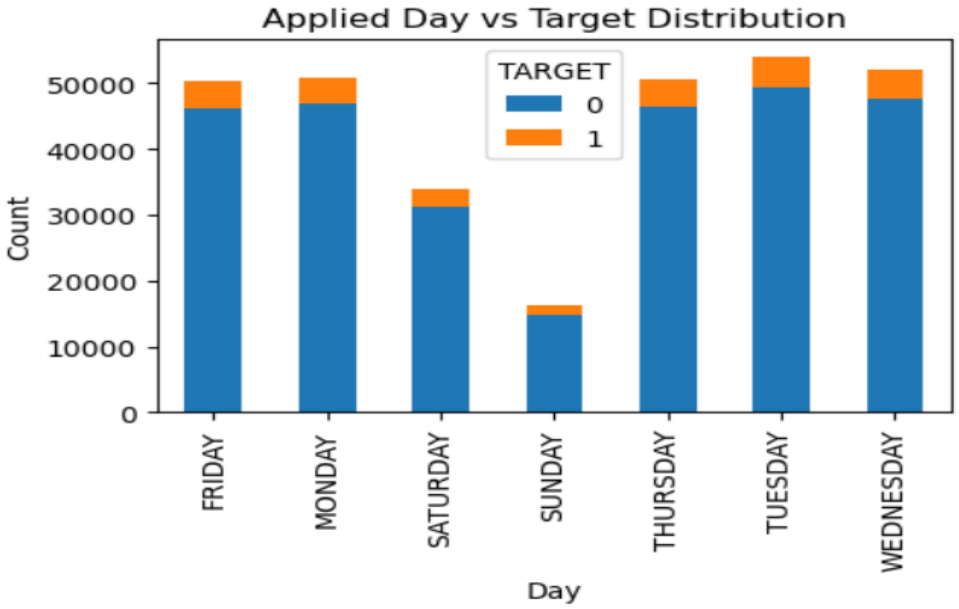
- The default rates for families with **1** and **2 members** are relatively lower, indicating that smaller families may have fewer financial burdens or greater income stability.
- There is a slight increase in the default rate for families with **3** to **5 member s**.
- The default rate peaks at **9.4%** for families with **5 members**,
- **Overall Default Rate** 8.07%

Region Rating vs Loan Default



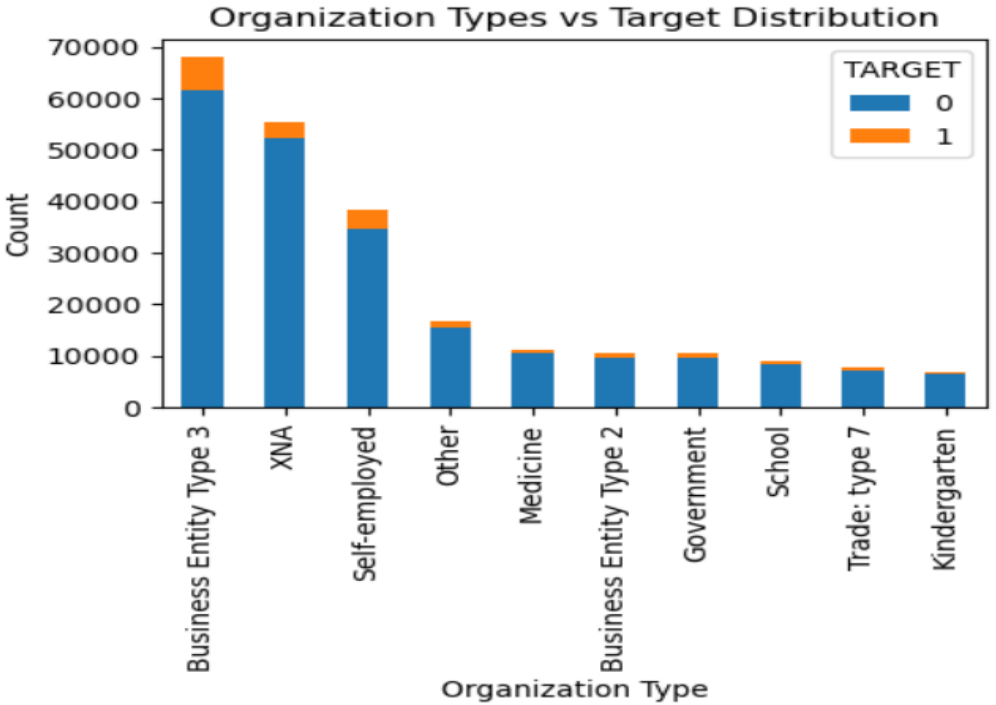
- Clients with **Region Rating 1** show the lowest default rate 5%, suggesting that borrowers in this region may have better financial stability
- Clients with **Region Rating 3** have the highest default rate **11%**, indicating that this group is at a higher risk for defaults
- **Overall Default Rate** 8.77%

Applied Day vs Loan Default



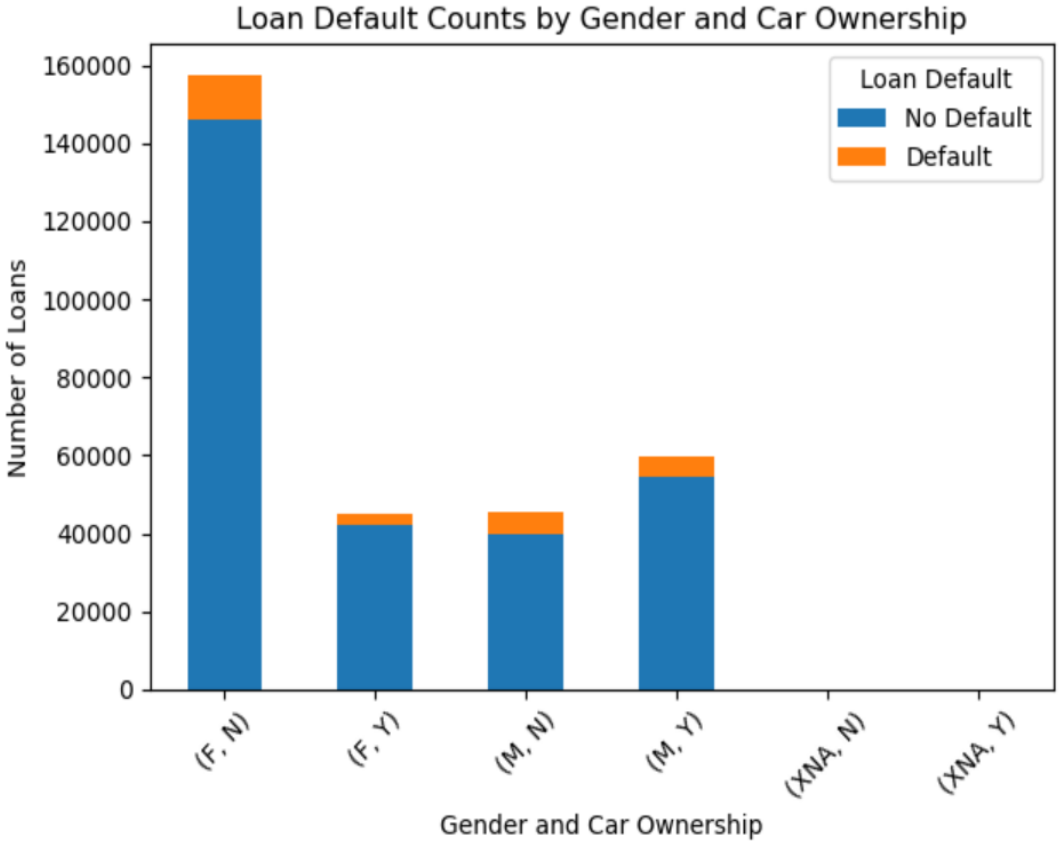
- **Highest Default Rate:** Sunday (8.60%)
- **Lowest Default Rate:** Monday (7.76%)
- Default rates are **relatively consistent across weekdays**, with some fluctuation.

Organization Type vs Loan Default



- **Self-employed:** Highest default rate at **10%**, indicating financial instability.
- **Business Entity Type 3, Trade: Type 7 and Business Entity Type 2:** High default rate of **9**%, showing significant risk.
- **Government & Medicine Sector:** Relatively low default rate of **7.00%**, indicating reliable repayment behavior.
- **XNA and School:** Lowest default rates at **5%**, suggesting stable finances.

Gender, Car Ownership, and Loan Default



Females (8.63% default rate) and **males** (8.34% default rate) show similar default rates, with females slightly higher.

Car Owners:

Females: 5.78% default rate (2,603 defaults)

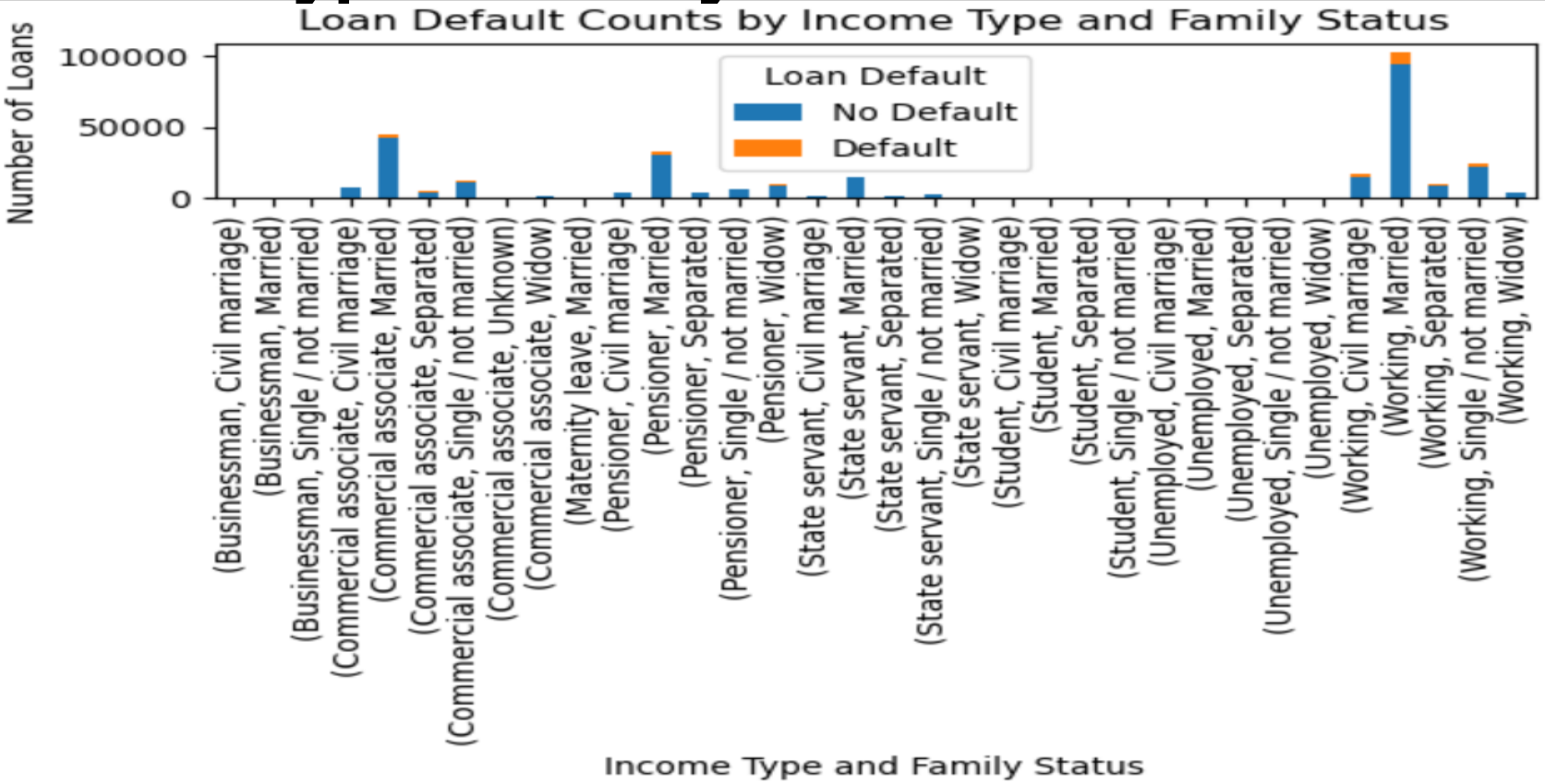
Males: 8.34% default rate (4,973 defaults)

Non-Car Owners:

Females: 7.36% default rate (11,567 defaults)

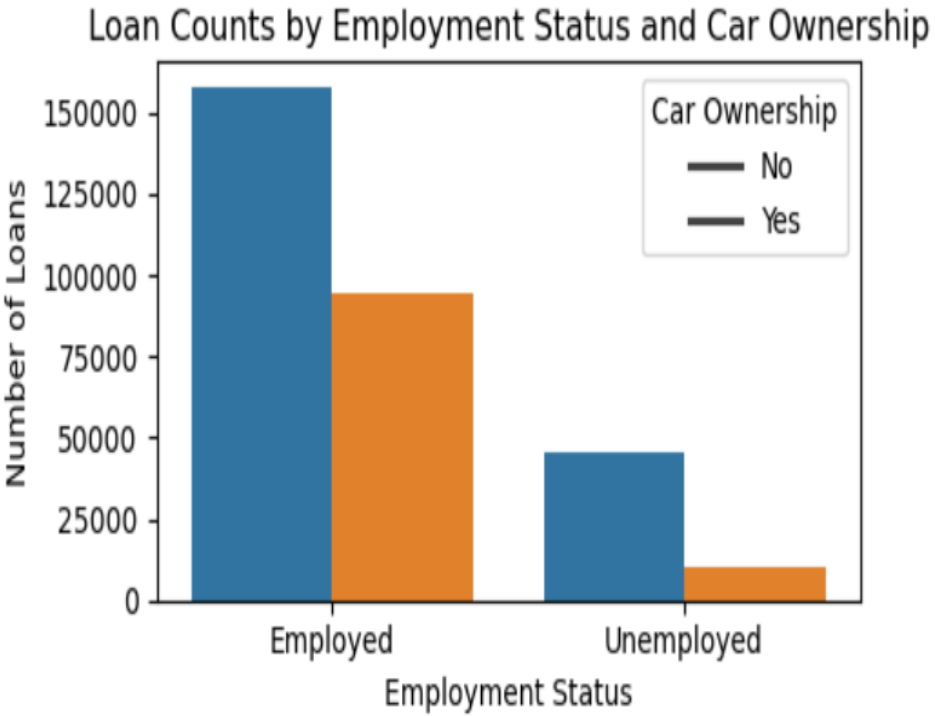
Males: 12.53% default rate (5,682 defaults)

Income Type , Family Status, and Loan Default



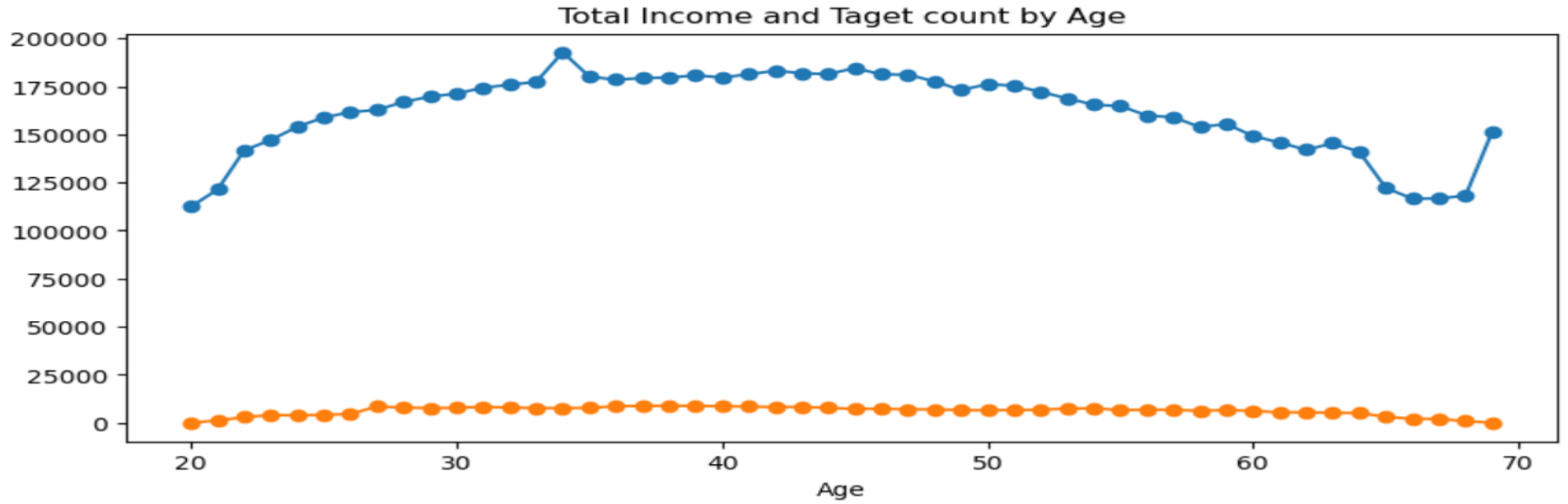
- **Commercial Associates** and **Pensioners** display notable default rates:
  - **Commercial Associates**: Out of 70,35 total loans in civil marriage status, **682 defaults** indicate a default rate of approximately **9.7%**.
  - **Pensioners**: With 33,50 loans in civil marriage status, **227 defaults** lead to a default rate of about **6.8%**.
- **Stable Categories**:
  - **Businessmen** and **Students** show no defaults across any family status, indicating strong creditworthiness in these groups.
  - **Working individuals** demonstrate the highest total loan counts and varying default rates, highlighting their crucial role in the lending landscape.
- **General Trends**:
  - Married individuals generally have lower default rates across all income types, indicating stability.
  - The **"Single / not married"** category shows a mix of low defaults in Businessman and high defaults in Working, suggesting varied financial behaviors.

Employment Status, Car Ownership, and Loan Default



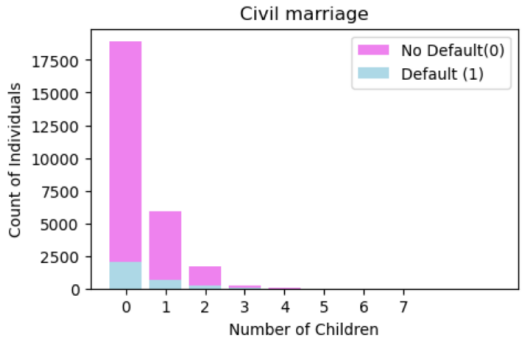
- **Employed Individuals**:
  - **Without Car Ownership**:
    - Total Count: **157,717**
  - **With Car Ownership**:
    - Total Count: **94,418**
- Employed individuals are more likely to have car ownership, suggesting greater financial stability.
- **Unemployed Individuals**:
  - **Without Car Ownership**:
    - Total Count: **45,207**
  - **With Car Ownership**:
    - Total Count: **10,169**
- A significantly lower number of unemployed individuals own cars, indicating potential financial constraints.

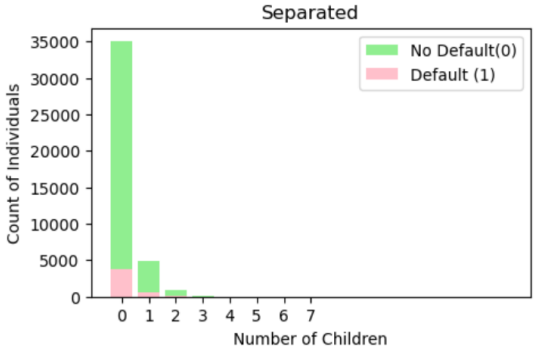
Age, Income, and Loan Default

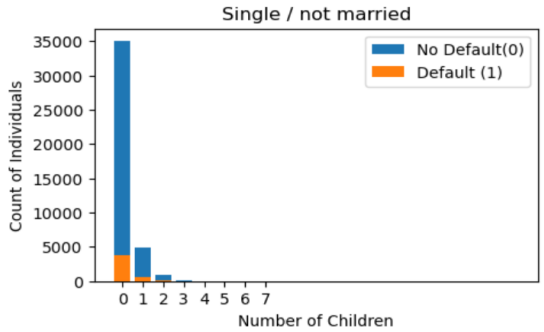


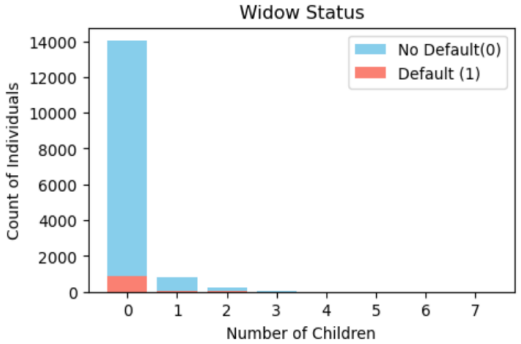
- **Loan Defaults and Age**: Default rates rise through the 30s, peaking at around age 34, while younger individuals (20s) show lower default rates.
- **Income Trends**: Average income increases with age, with individuals in their 30s earning significantly more (up to **$192,746** at age 34).

Family Status, Number of Children and Loan Default



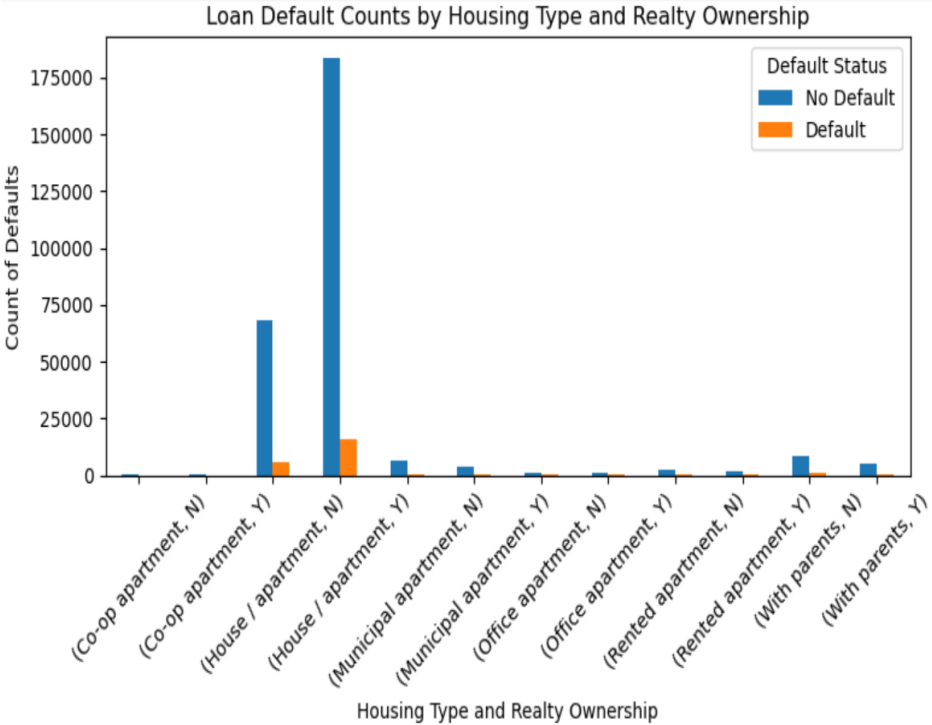






Across all family statuses, the number of defaults tends to decrease as the number of children increases, th ough certain groups (e.g., married) stil l experience substantial default rates even with more children.

Housing Type, Realty Ownership and Loan Default



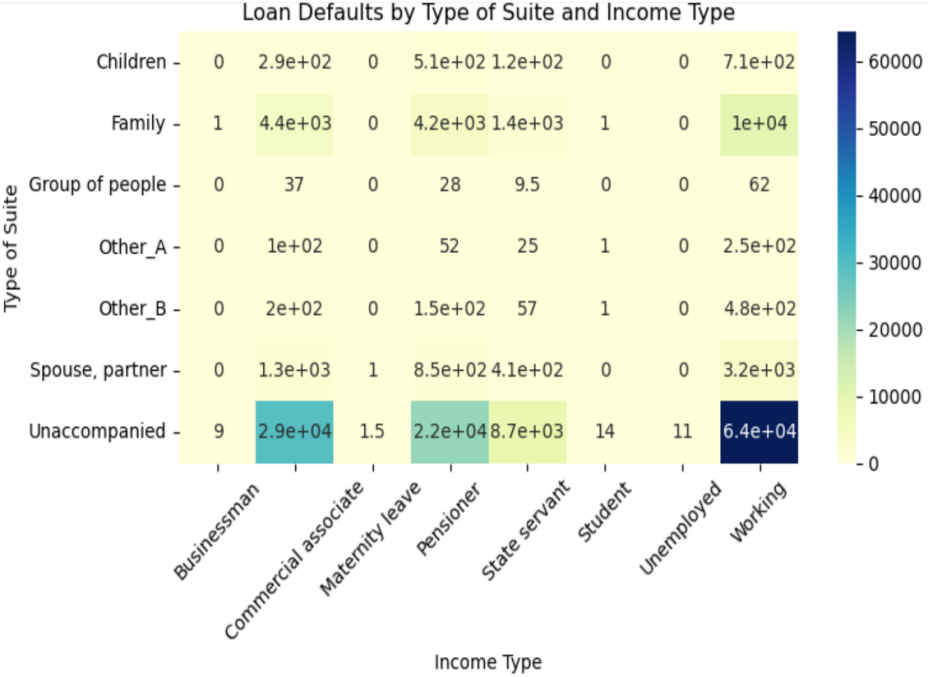
House/Apartment Owners: Default rate of **7.84%** (15,585 defaults) for owners vs. **7.7%** (5,687 defaults) for non-owners, showing minimal difference.

Co-op Apartments: Owners have a higher default rate (**7.77%**) compared to non-owners (**8.25%**), indicating financial risk is present for both.

Rented Apartments show a default rate of **14.76%** for non-owners and **13.06%** for owners, indicating higher vulnerability among renters.

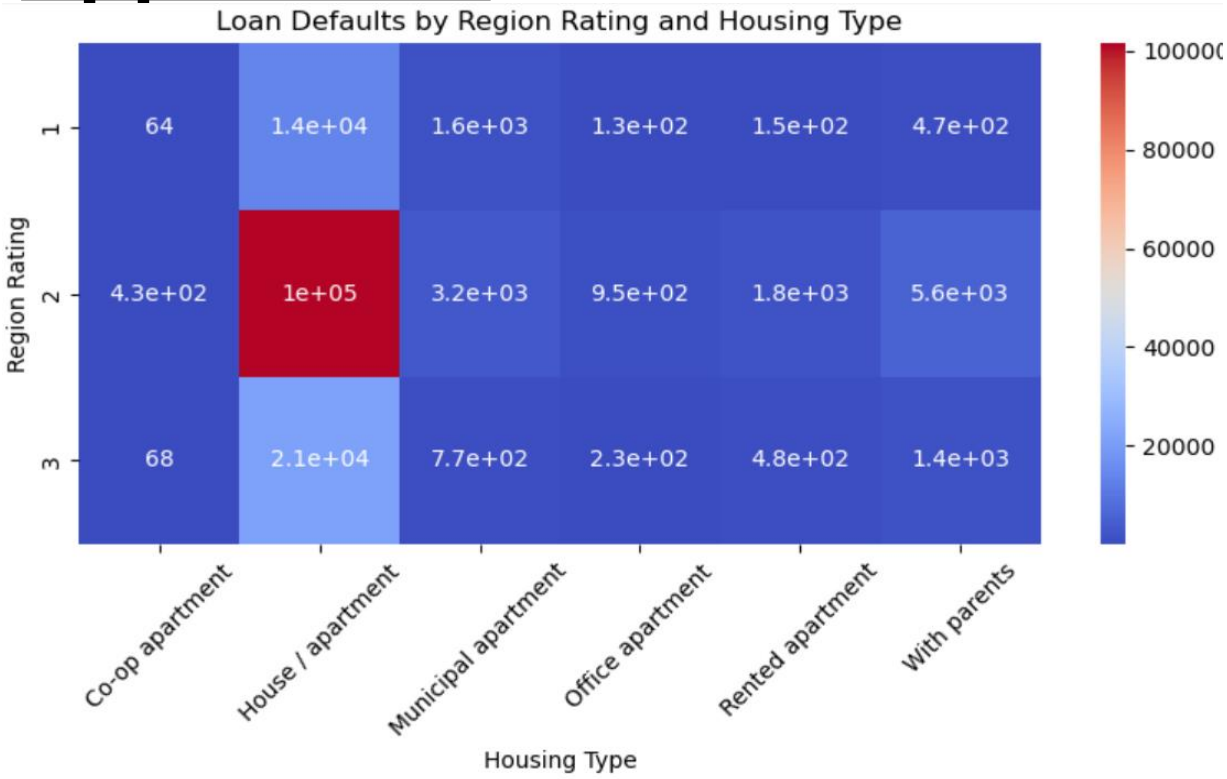
Living with Parents: Non-owners face a default rate of **12.99%**, which is notable.

Name Type Suite, Income Type and Loan Application



- **Income Type Variability**: Income types like "Commercial associate" and "Working" show significant counts across various suite types, especially for "Unaccompanied," indicating a higher risk associated with these categories.
- **Family Structure Impact**: The "Family" suite has high counts for "Commercial associate," but notably lower for "Maternity leave," suggesting that certain family structures may correlate with more stable financial situations.

Region Rating, Housing Type and Loan Application



- The "House / apartment" category consistently shows the highest counts across all region ratings, particularly in Region 2, suggesting that it's the most common housing type among clients.
- In Region 1, "Co-op apartment" has a relatively low count, indicating it might be less associated with lower financial risk

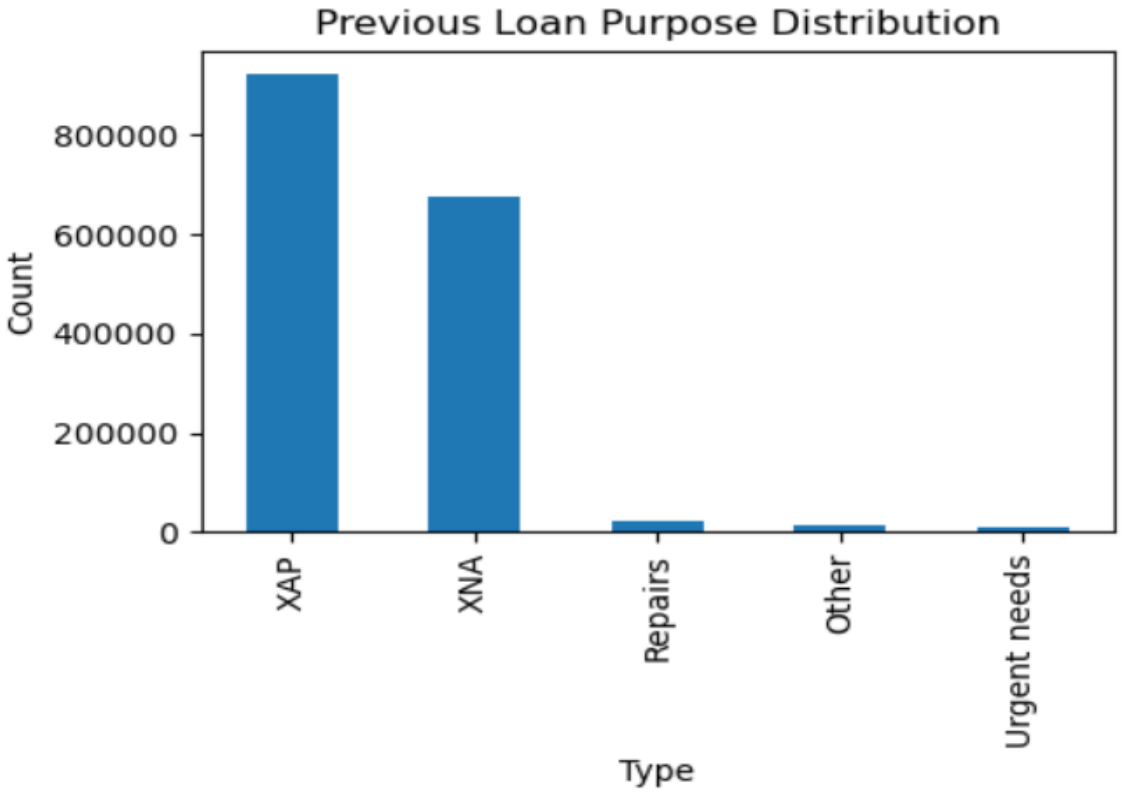
Document Count and Loan Default

	CNT_FAM_MEMBERS	TARGET	FLAG_DOCUMENT_2	count
0	1.0	0	0	62171
1	1.0	0	1	1
2	1.0	1	0	5673
3	1.0	1	1	2
4	2.0	0	0	146343
5	2.0	0	1	5
6	2.0	1	0	12007
7	2.0	1	1	2
8	3.0	0	0	47993
9	3.0	1	0	4608
10	4.0	0	0	22558
11	4.0	0	1	3
12	4.0	1	0	2136
13	5.0	0	0	3151

•**Document Importance**:
The presence of documents generally correlates with lower loan default rates, particularly with larger family sizes, suggesting that documentation may be a factor in securing loans effectively.

Analysis On Previous Application

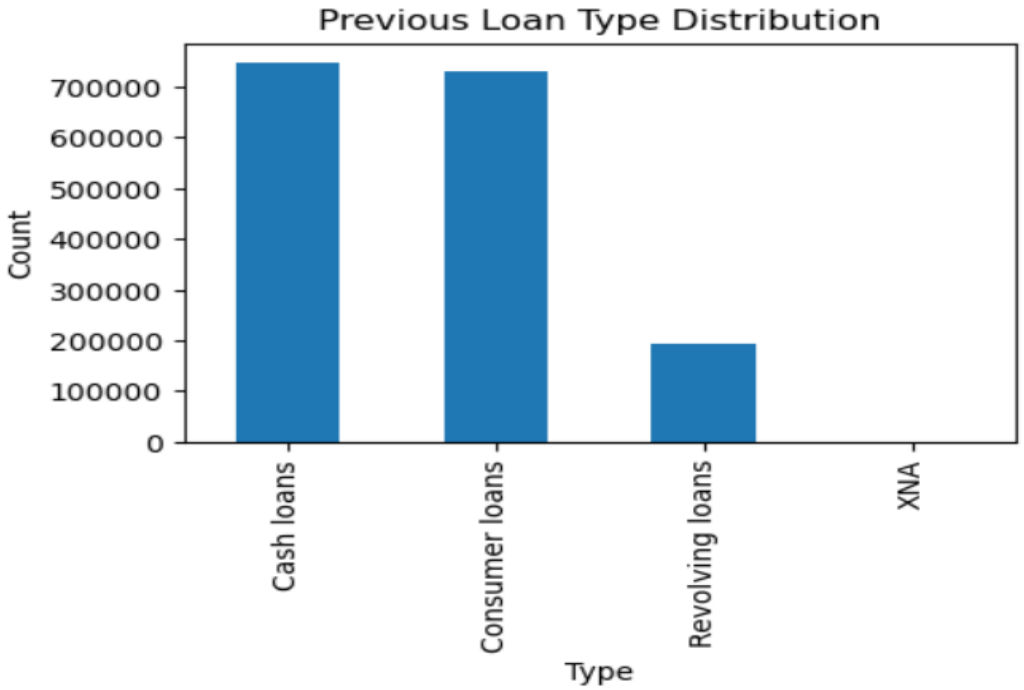
Previous Loan Purpose Distribution



- "XAP" and "XNA" dominate the loan purposes, with "XAP" having 55%, and "XNA" close behind with 40%
- .Other categories like "Repairs," "Other," and "Urgent needs" have very small applicants in comparison,

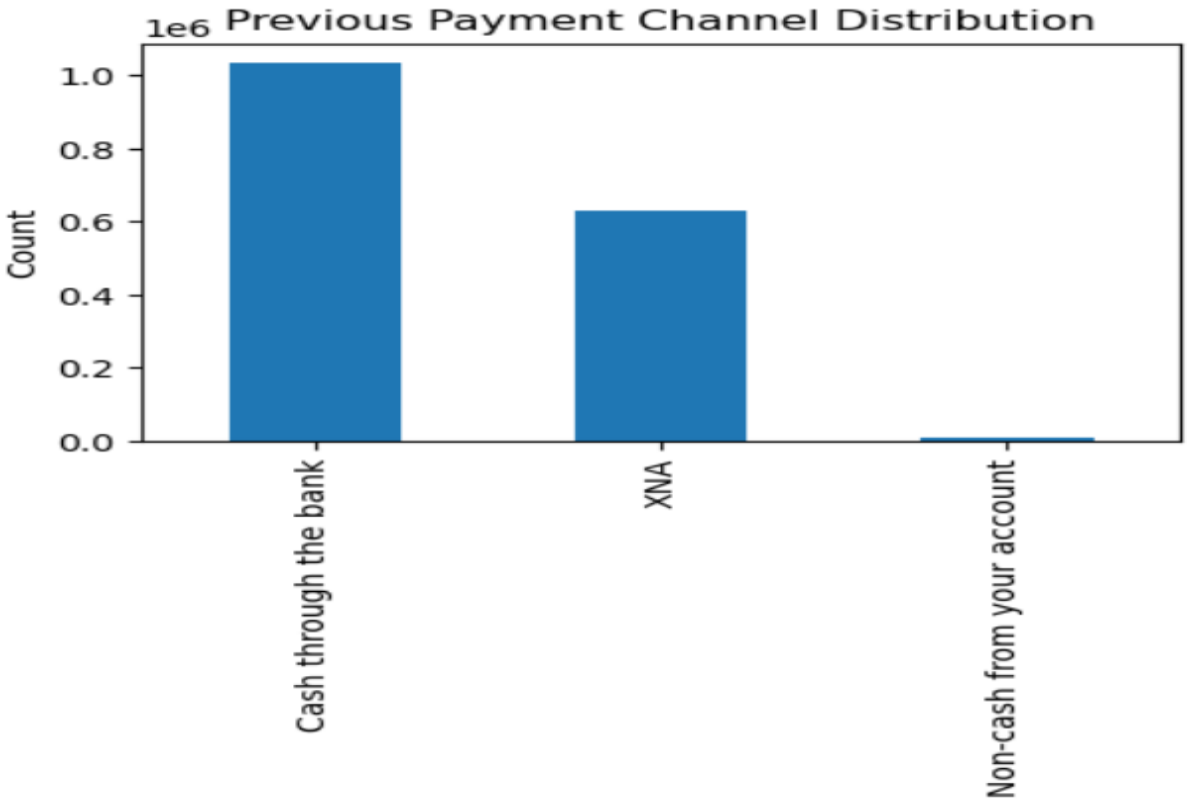
Most previous loans fall under the categories **"XAP"** and **"XNA,"**

Previous Loan Type Distribution



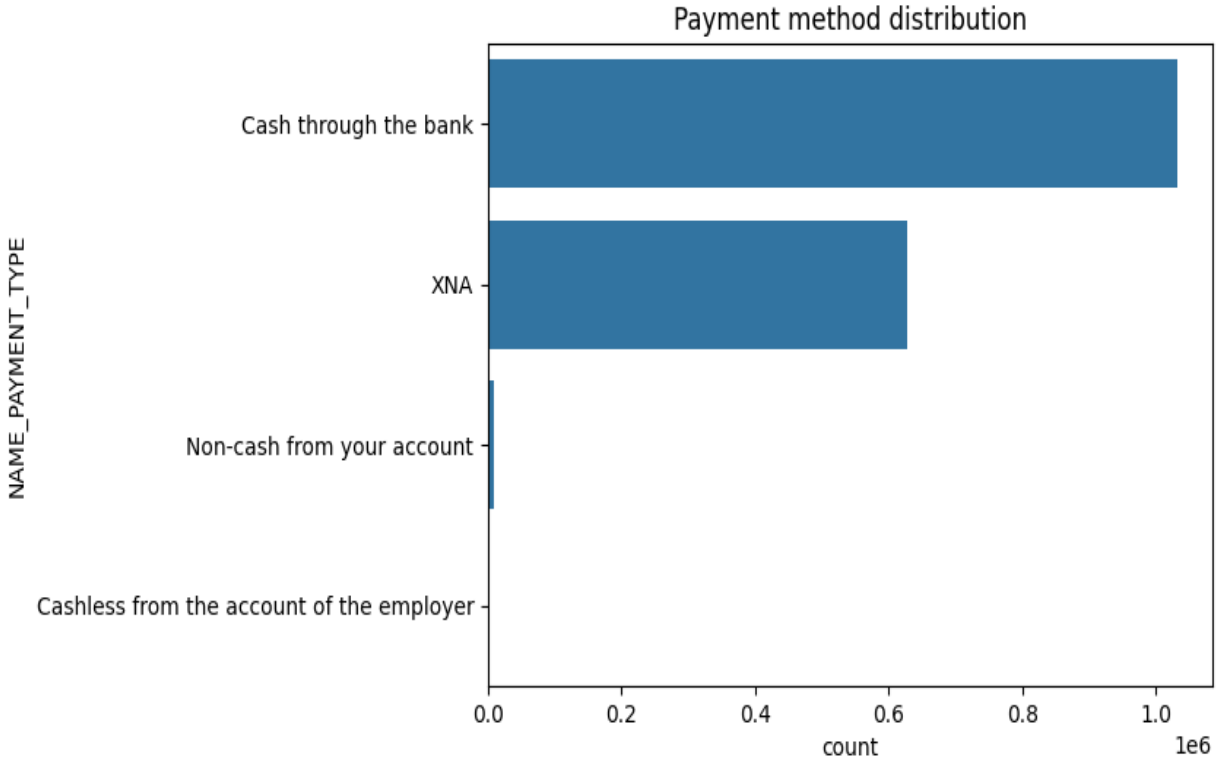
- Cash loans are the most common type, with 45% applicants, indicating that they are highly preferred.
- Consumer loans follow closely with 44% applicants, suggesting a similar demand as cash loan.
- Revolving loans, with 12% applicants, are significantly less common than cash and consumer loans.

Previous Payment Channel Distribution



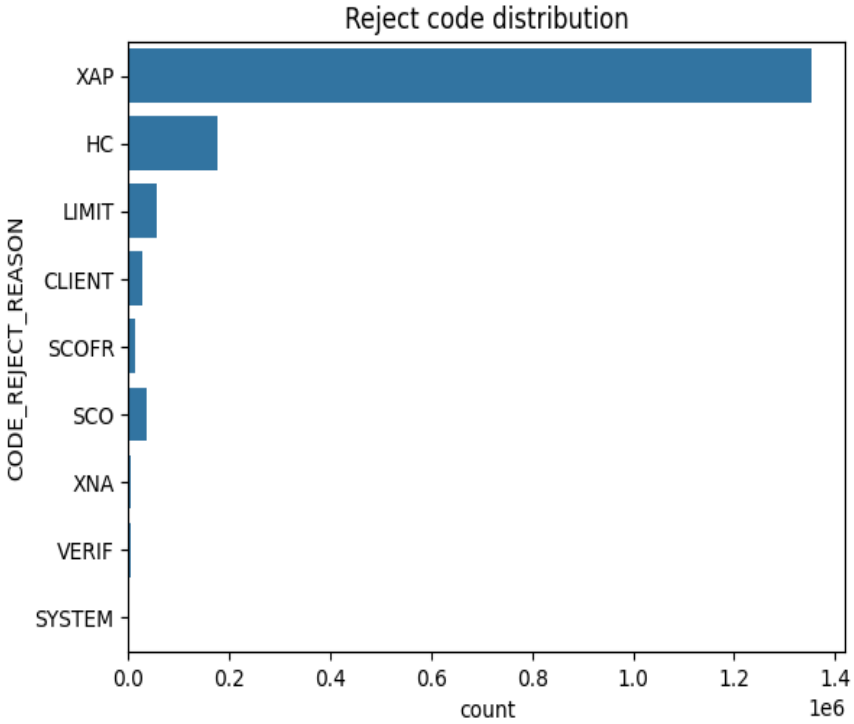
- **Cash through the bank** is the most common payment method, with 62% transactions, indicating a strong preference.
- **XNA** follows closely with 37% transactions, showing it's a popular alternative.
- **Non-cash from your account**, with 8,193, (0.5%) transactions, is far less common.

Payment method distribution



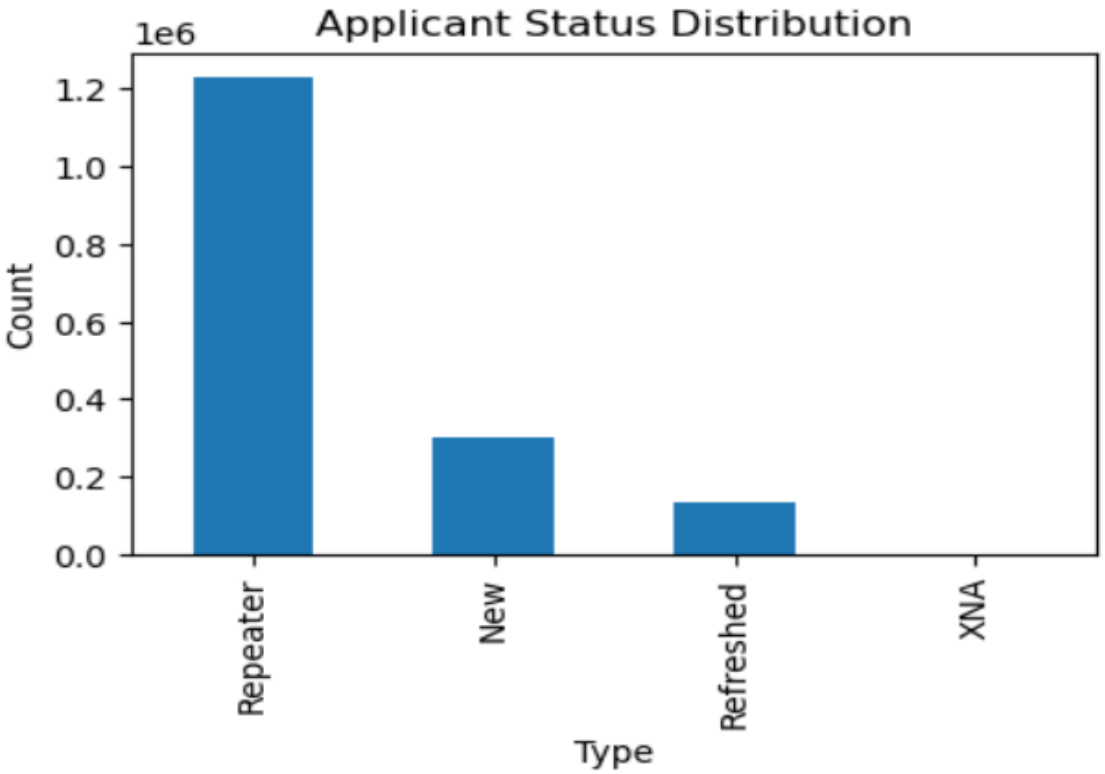
- Cash through bank is the most prominent method 61.9% used
- Payment method is not available for 37.6%

Loan Rejection Reason Distribution



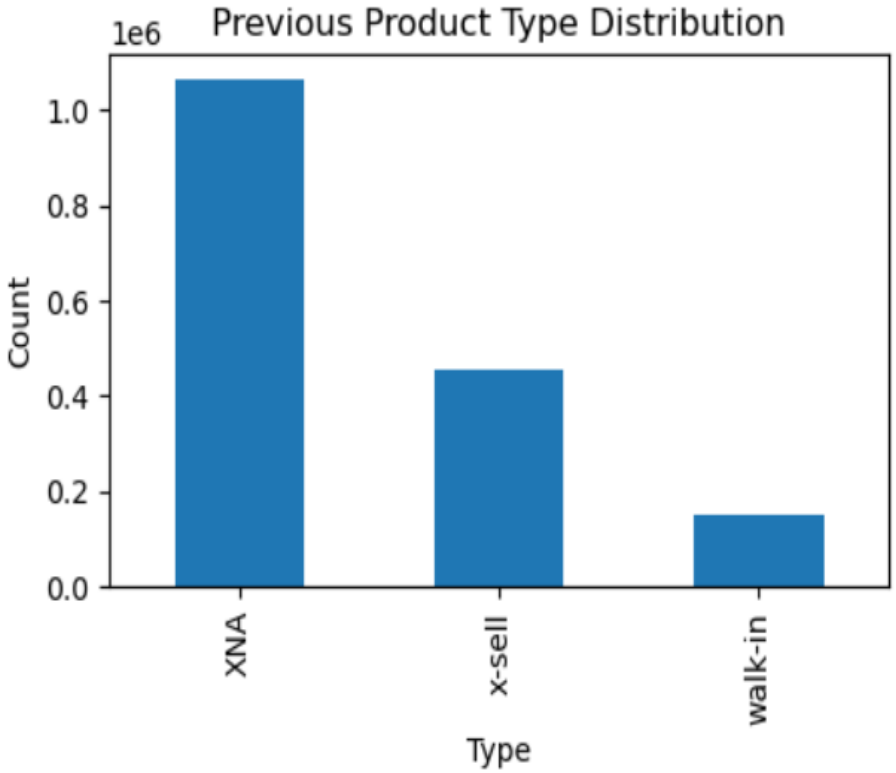
- **XAP** Reject code is not applicable for 81% of application i.e for approved and cancelled application
- The second largest category is **HC**, with 10.5% applicants, possibly indicating loans reviewed under specific high-priority or high-value conditions.
- **LIMIT** has 3.3%, while **SCO** has 2.2%. These might represent loans approved with specific limits or scored under certain criteria, showing moderate representation

Applicant Status Distribution



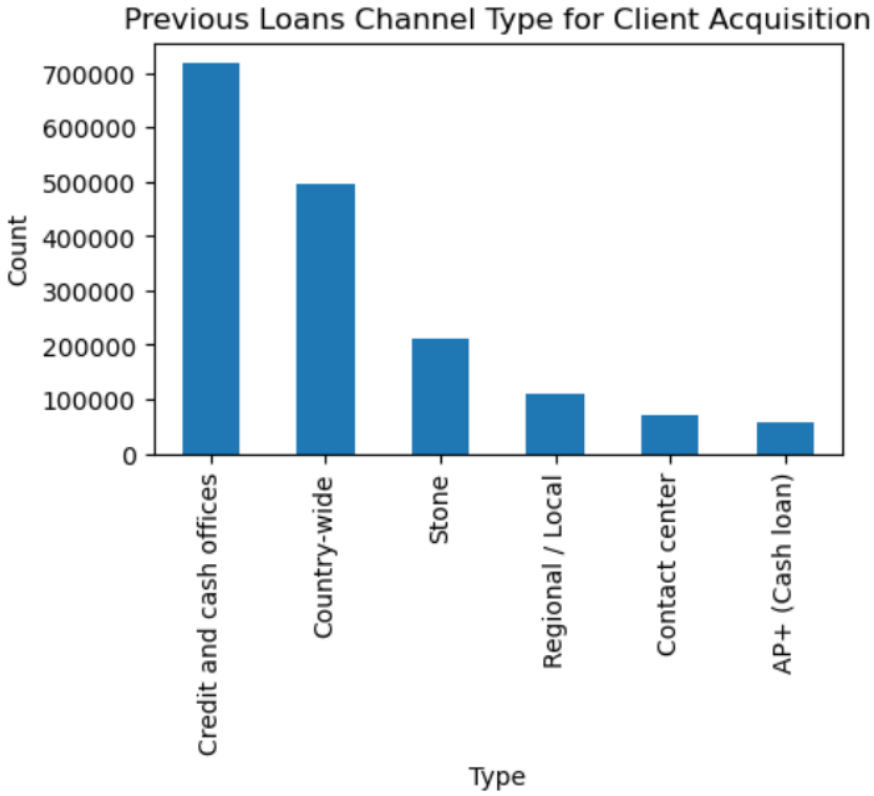
- **Repeater 74%:** Highest value, suggests primary focus or significance.
- **New 18%:** Lower, indicates a niche area or less frequent occurrence.
- **Refreshed 8%:** Intermediate, may represent updates or modifications.
- **XNA 0.1%:** Lowest, likely the least significant term.

Previous Product Type Distribution



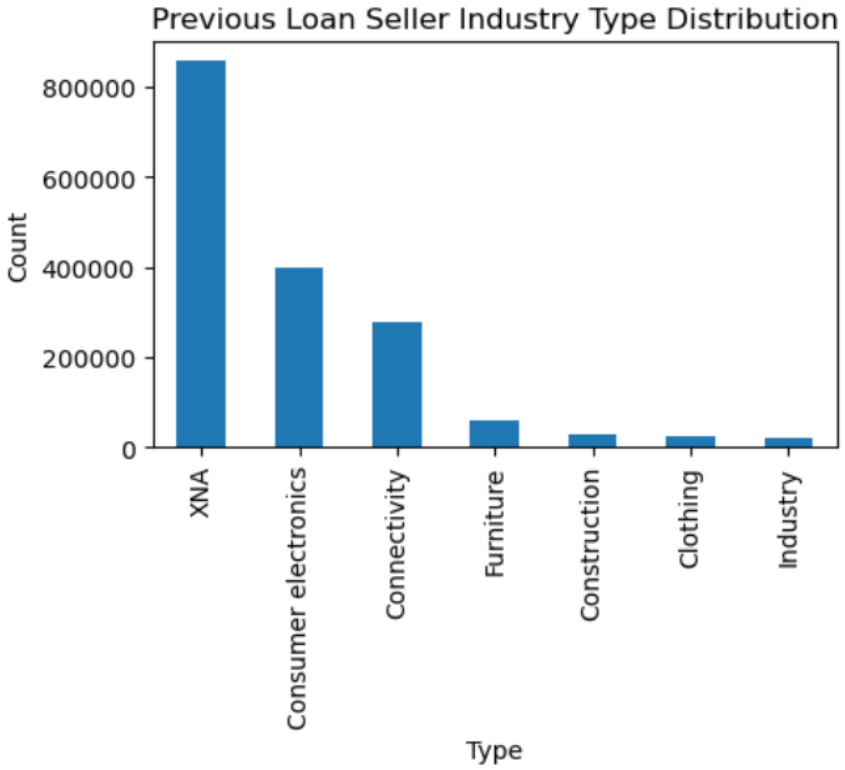
- **XNA 64%**
- This is the highest value, indicating that "XNA" has become a significant category.
- **X-sell 27%**
- This term is notably lower but substantial, suggesting it represents a key area, possibly related to upselling or cross-selling.
- **Walk-in 9%**
- This is the lowest value, indicating that "Walk-in" might represent a less frequent occurrence

Previous Acquisition Channel Distribution



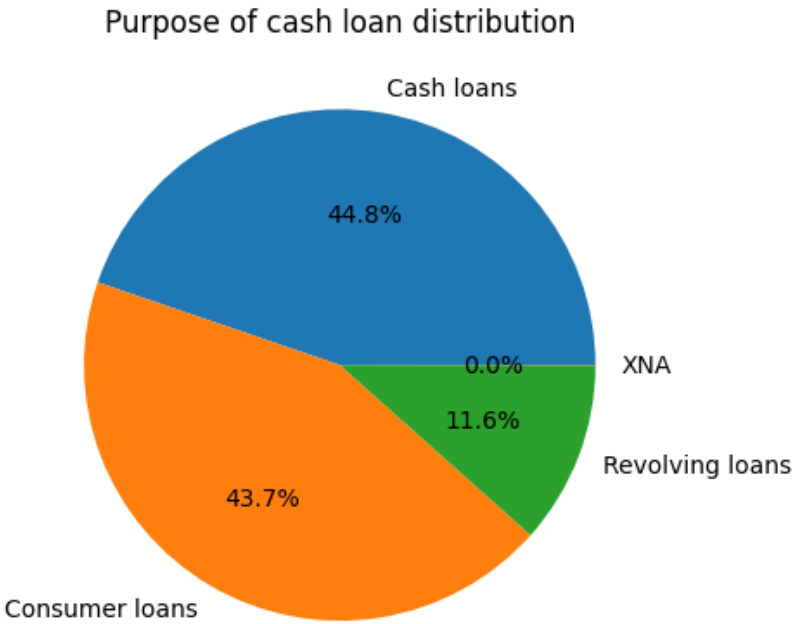
- **Credit and cash offices (45%)**: Dominant channel, crucial for traditional in-person transactions.
- **Country-wide (31%)**: Significant national reach, driving broad accessibility.
- **Stone (13%)**: Notable but may need contextual relevance to enhance impact.
- **Regional / Local (7%)**: Supports local operations, complementing national efforts.
- **Contact center & AP+ (Cash Loan) (4%)**: Moderate contributor with potential for growth.

Previous Seller Industry Distribution



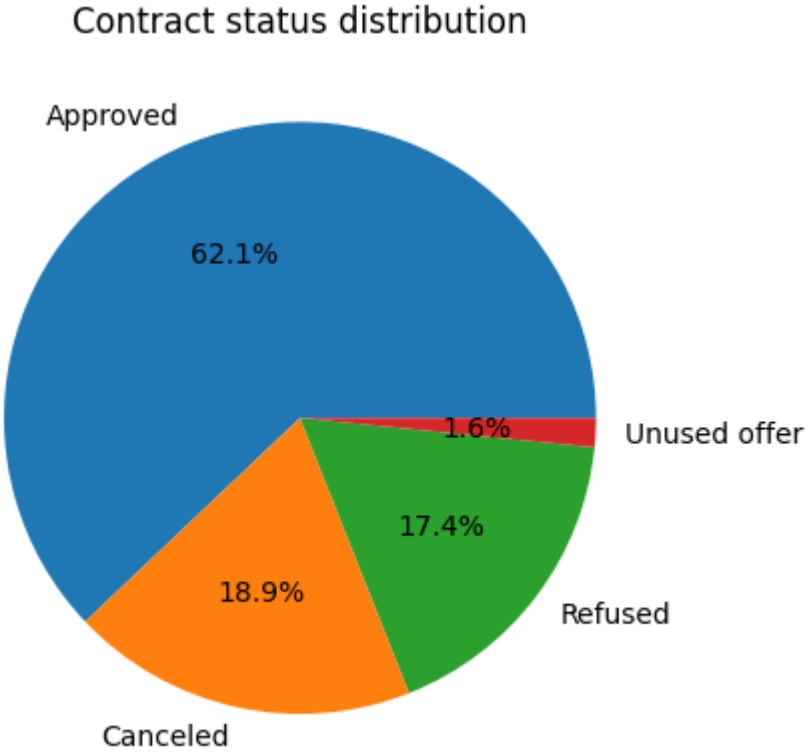
- **XNA 51%:** Highest value; primary focus area with significant importance.
- **Consumer Electronics 24%:** Substantial category; likely a key revenue driver.
- **Connectivity 17% :** Important segment; suggests growing demand for connectivity solutions.
- **Furniture 4% & Construction 2%:** Lower emphasis; could represent an area for growth.
- **Clothing & Industry1% :** Less prominent; further investigation could identify opportunities.

Types of Loan Distribution



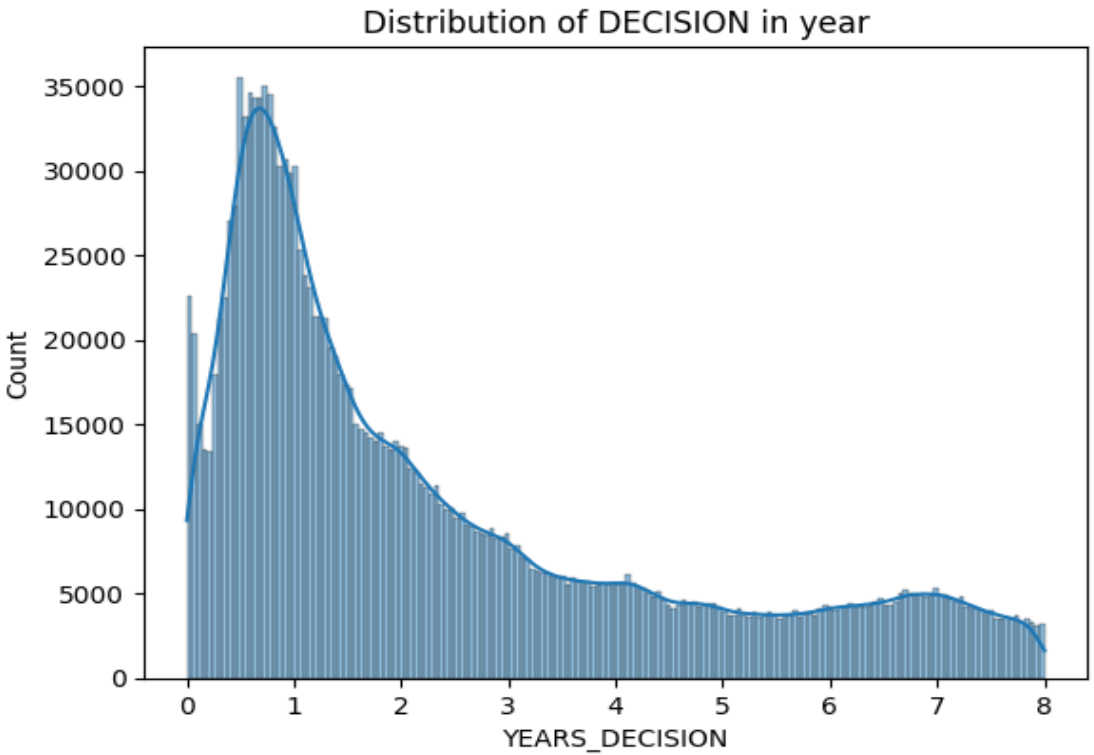
- Cash loans are the most common type of loan, with a count 747,553
- Consumer loans follow closely behind, with a count 729,151
- Revolving loans are much less common in comparison, having count of 193,164

Contract Status Distribution



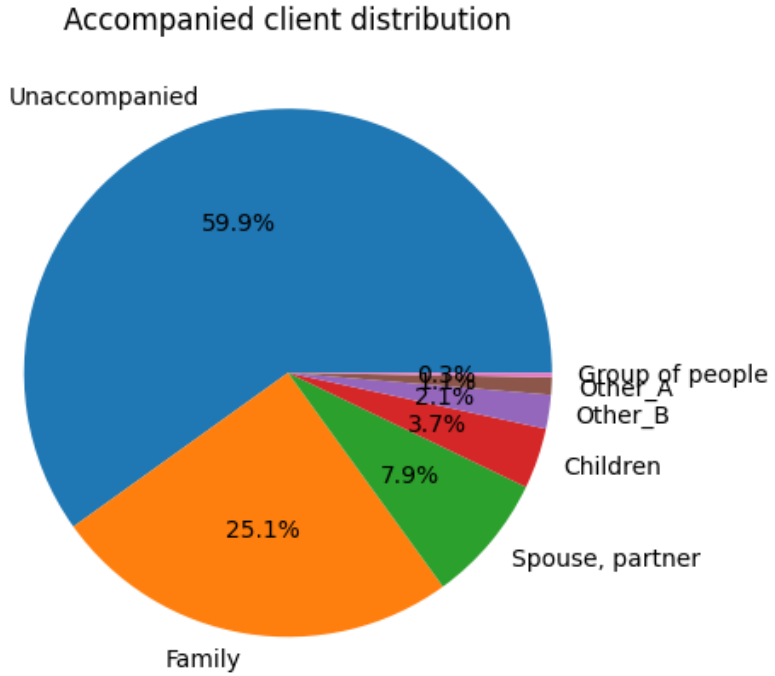
- Approved application were 62.1% with a count of 1,036,781
- Cancelled application were 18.9% with a count of 316,319
- Refused application were 17.4% with a count of 290,678
- Unused offer were 1.6% with a count of 26,436

Decision Made Distribution



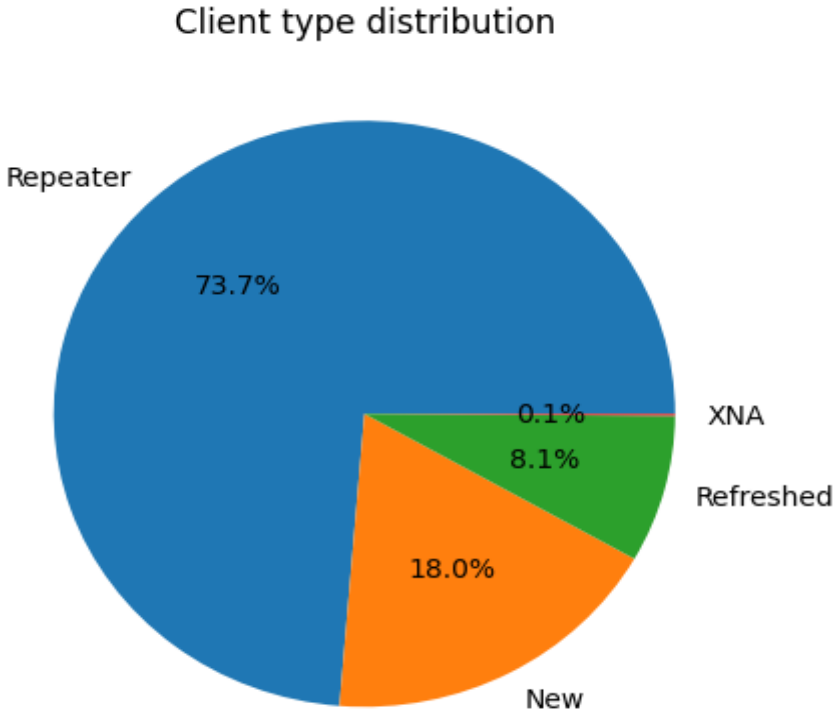
Most of the decision were taken in first two year

Accompanied Client Distribution



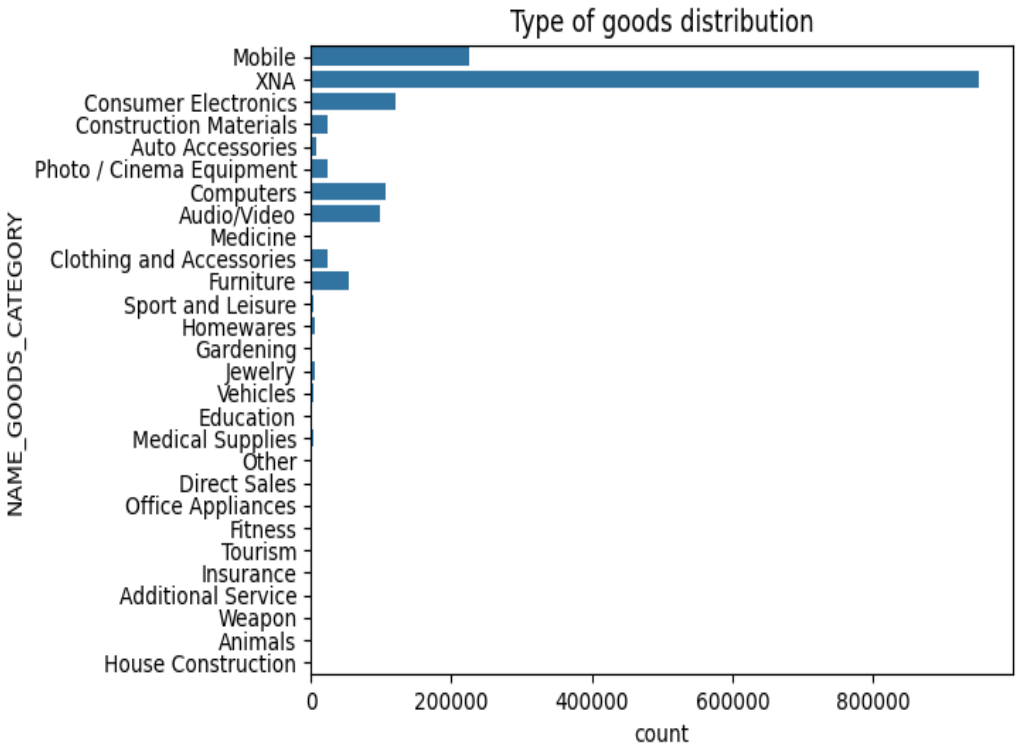
- Application that were unaccompanied was 59.9% with count of 508,970
- Application that were accompanied with Family was 25.1% with count of 213,263
- Application that were accompanied with Spouse, partner was 7.9% with count of 67,069
- Application that were accompanied with children, partner was 3.7% with count of 31,566

Client Type Distribution



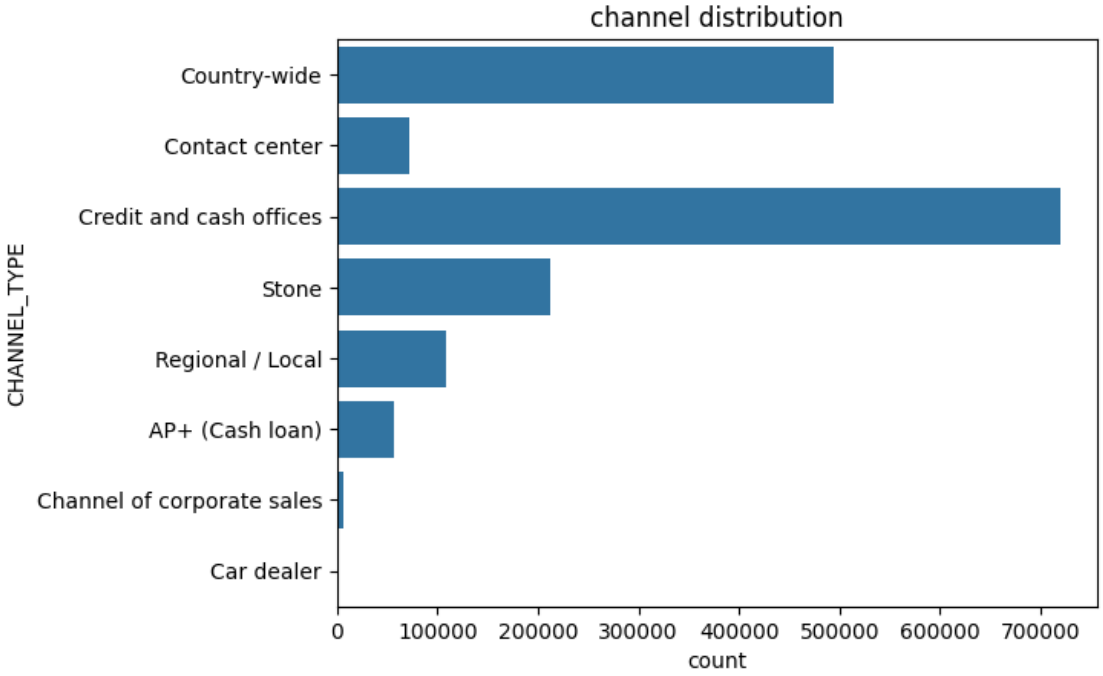
- We have 73.7% repeater type client with a count of 1,231,261
- New client are 18% with a count of 301,363
- Refreshed client are 8.1 with a count of 135,649

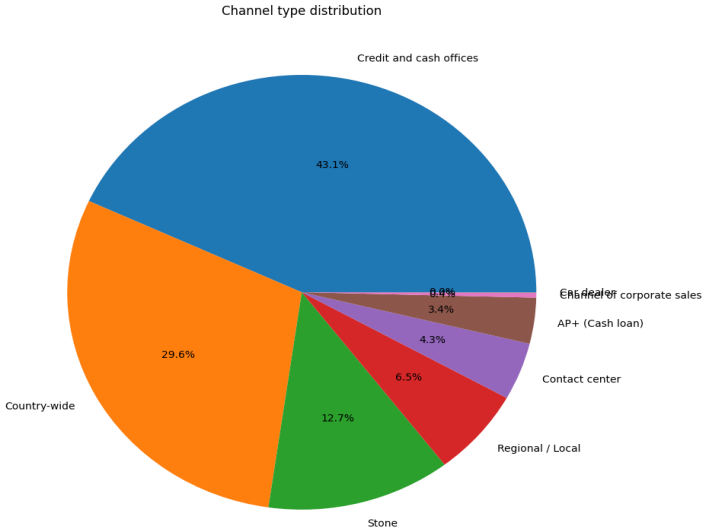
Type of Goods Distribution



- What Kind of goods did the client apply loan for is not available for 56.9%
- For mobile 13.5% of client applied loan
- For consumer electronics 7.3% of client applied
- For computers 6.3% ,Audio/Video 6.0% ,Furniture 3.2%, Photo/Cinema Equipment 1.5%

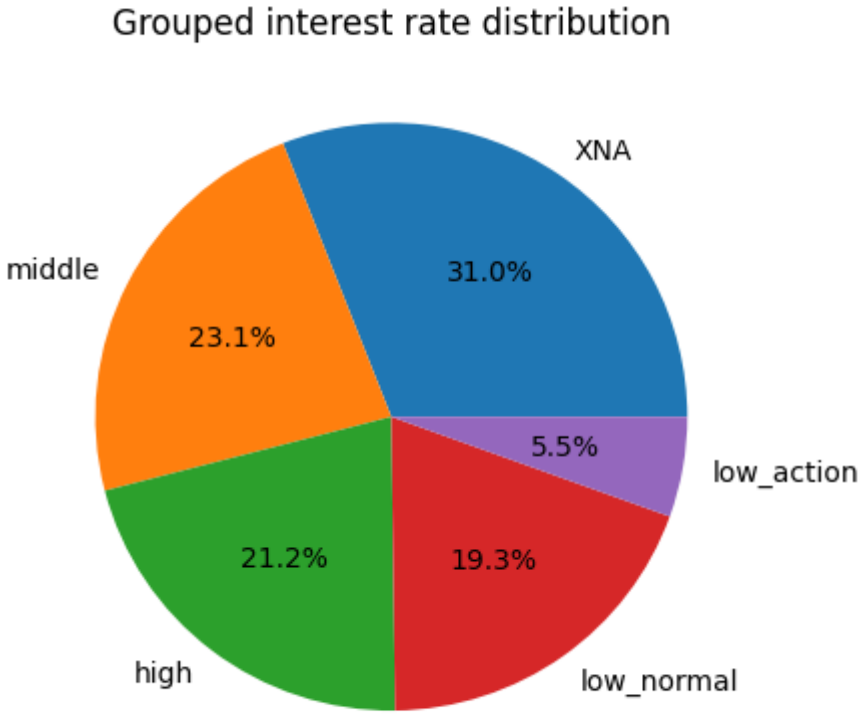
Channel Distribution





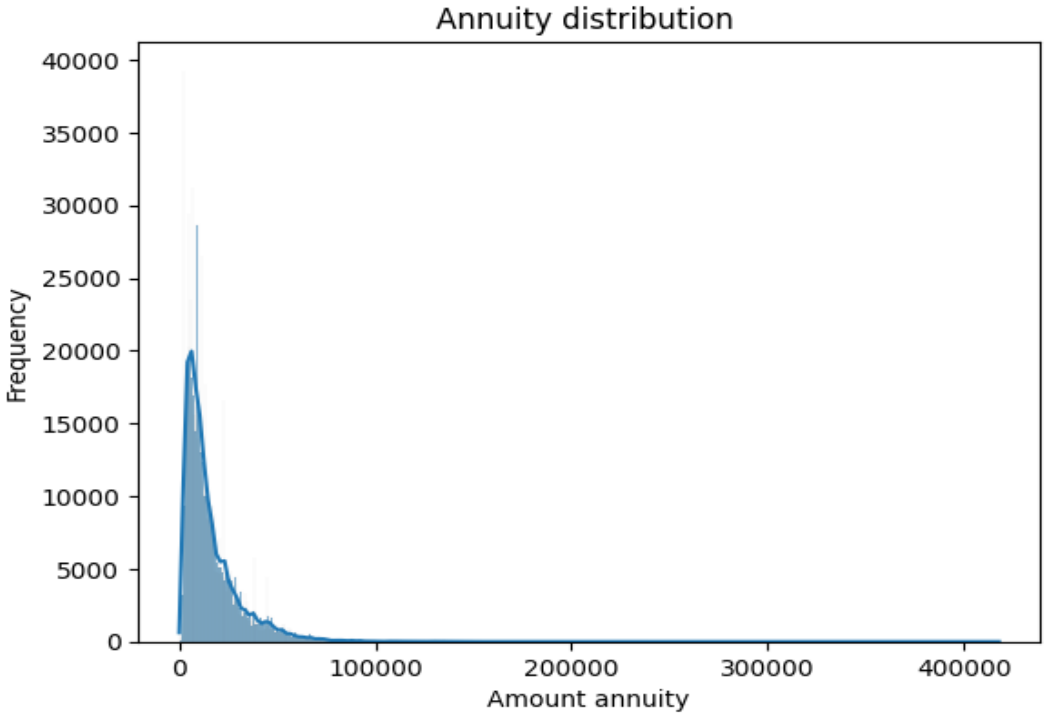
- Top Channel is Credit and cash offices which acquired 43.1% of client
- Next is Country Wide which acquired 29.6% of client

Grouped Interest Rate Distribution



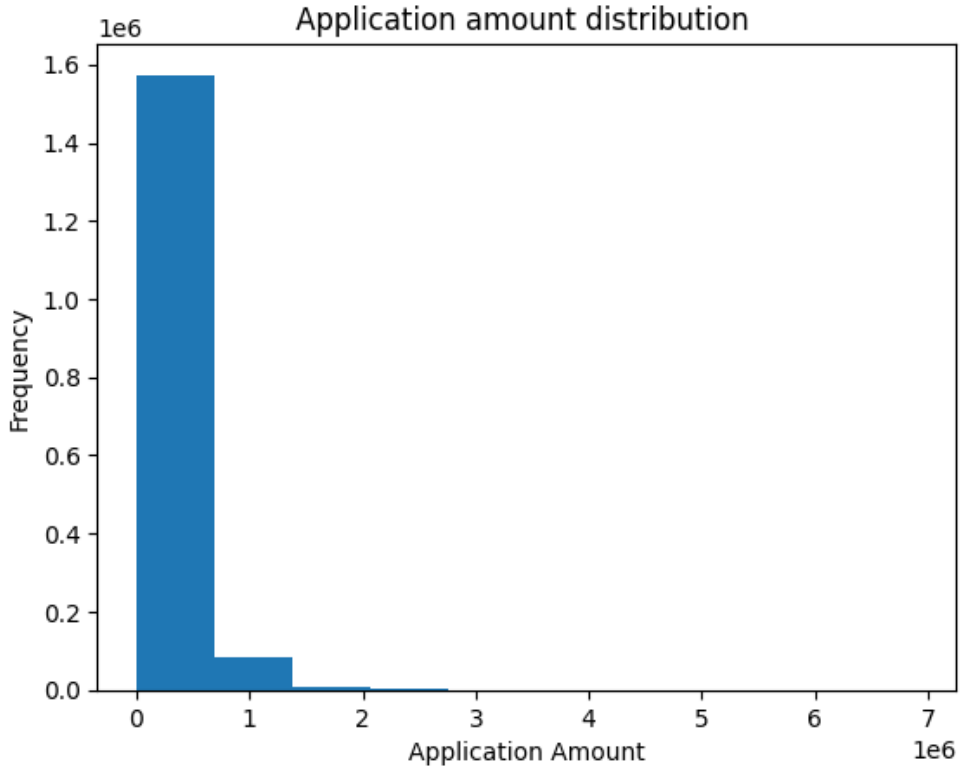
- Grouped interest rate is not available for 31% of application
- Middle interest were applied on 23.1% of application
- High interest rate is applied on 21.2% of application

Previous Application Annuity Distribution



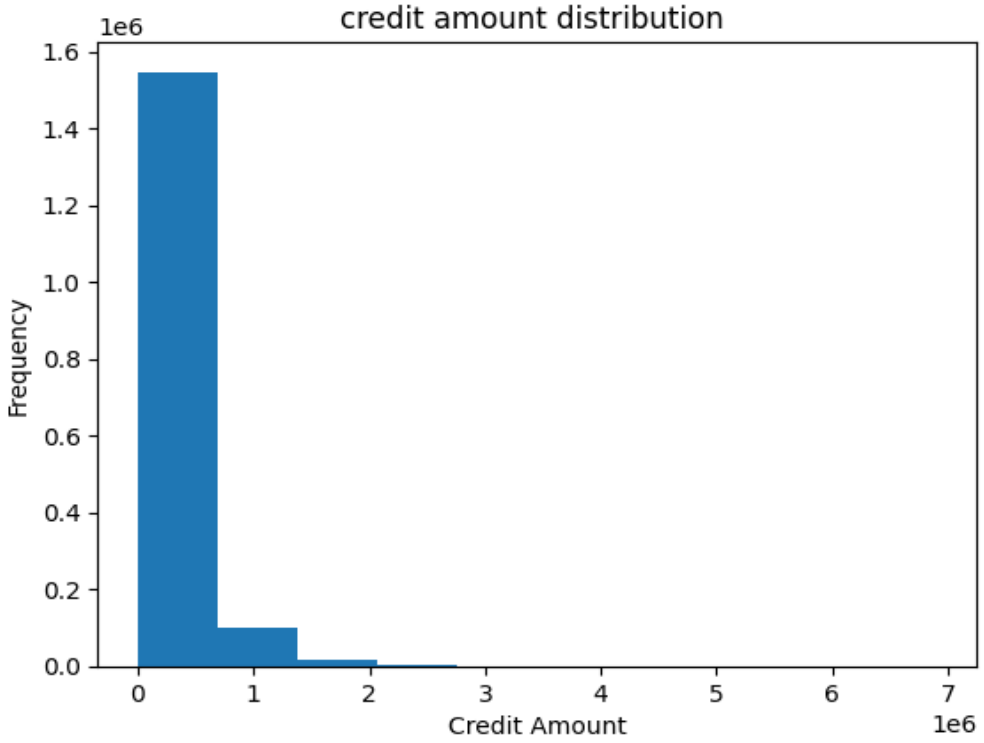
- **Highly Skewed Distribution**: The annuity distribution is heavily right-skewed, indicating that most annuity amounts are quite low, with a few larger annuities stretching out to the right.
- **Majority of Annuities Under 100,000**: The bulk of the annuities seem to be clustered below 100,000, with a massive spike around very low values, suggesting that small annuities are far more common.

Previous Application Amount Distribution



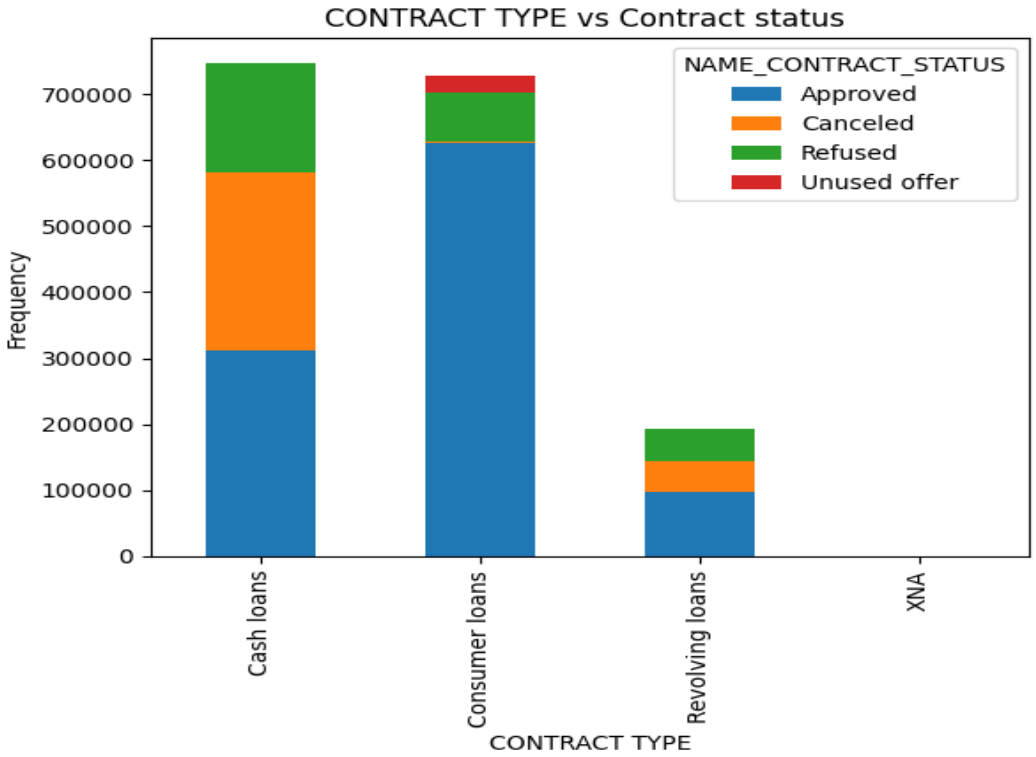
- **Highly Skewed Distribution**: The amount distribution is heavily right-skewed, indicating that most amounts are quite low, with a few larger amount stretching out to the right.
- **Majority of Amount Under 1,000,000**: The bulk of the amount seem to be clustered below 1,000,000, with a massive spike around very low values, suggesting that small amount are far more common.

Previous Credit Amount Distribution



- **Highly Skewed Distribution**: The credit amount distribution is heavily right-skewed, indicating that most credit amounts are quite low, with a few larger amount stretching out to the right.
- **Majority of Credit Amount Under 1,000,000**: The bulk of the amount seem to be clustered below 1,000,000, with a massive spike around very low values, suggesting that small amount are far more common.

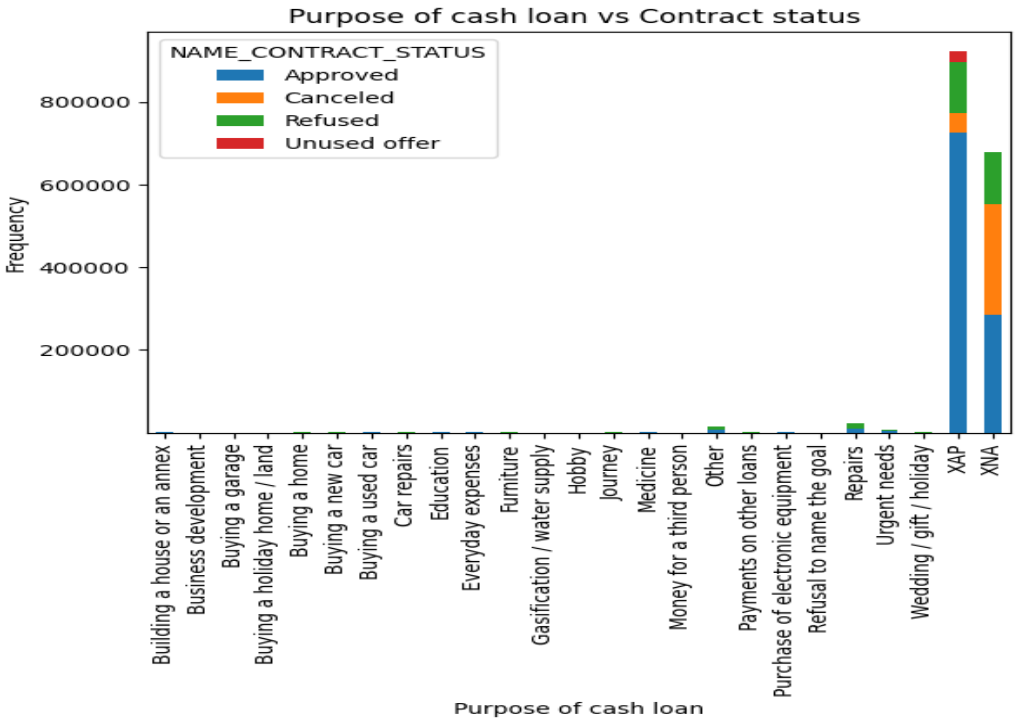
Contract Type Vs Contract Status



- Approval rate for Consumer loans is 85.91% followed by Revolving loans is 50.65% and least is for Cash Loans 41.80%

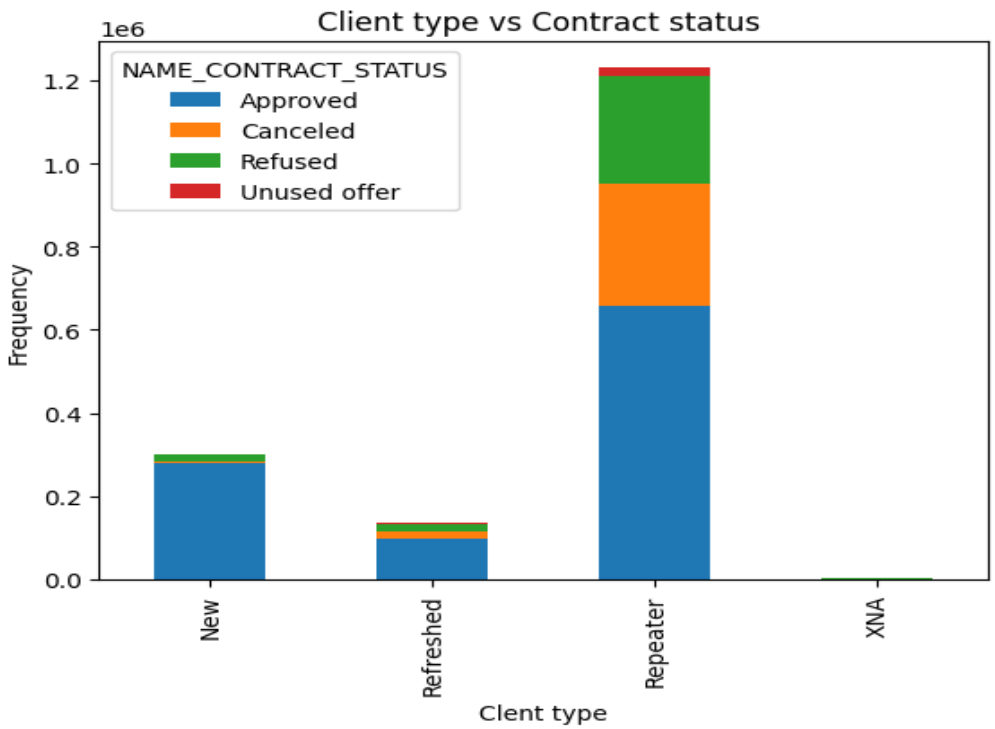
- Most approved contract type is Consumer loan
- Most unused contract type is consumed loan
- Most cancelled contract type is cash loan followed by revolving loan

Purpose of Cash Loan Vs Contract Status



- Approval rate for XAP is 78.49% followed by Purchase of electronic equipment is 55.41% and least is for Payments on other loans 15.74%
- The purpose of Cash Loan is not applicable or not available of majority of data
- Approval rate for XAP is more than XNA

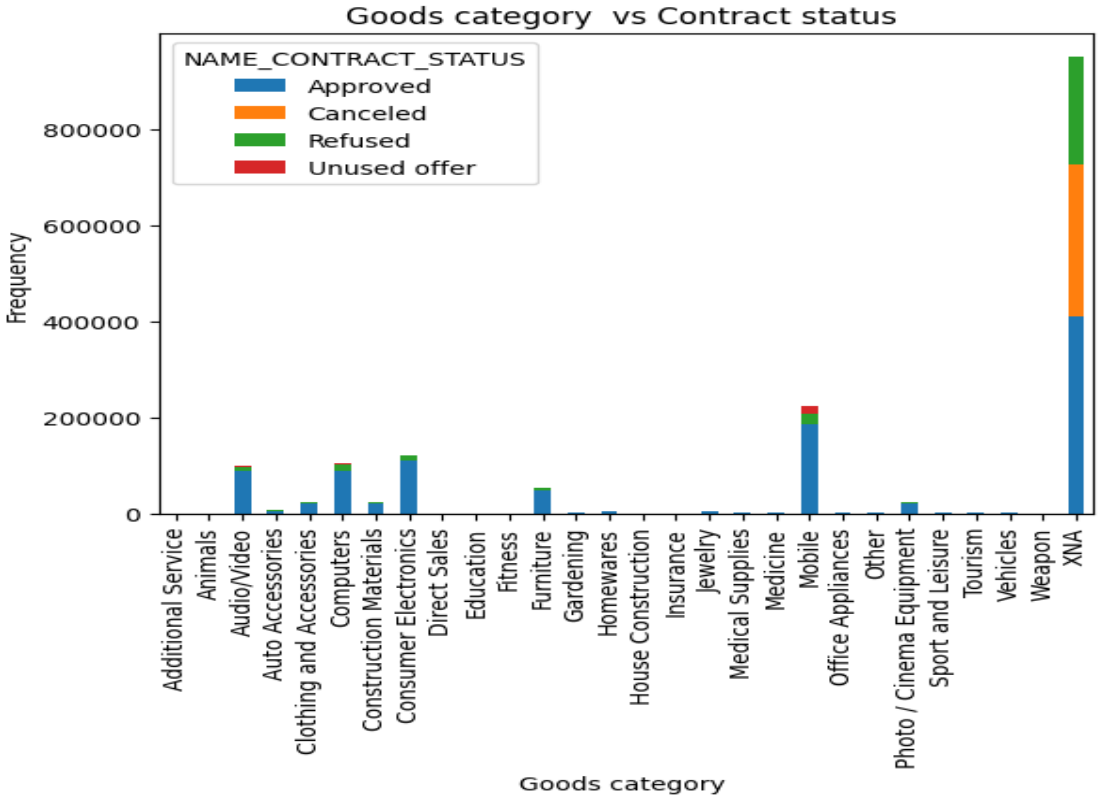
Client Type Vs Contract Status



- Approval rate for New is 93.32% followed by Refreshed is 71.58% and least is for XNA 29.77%

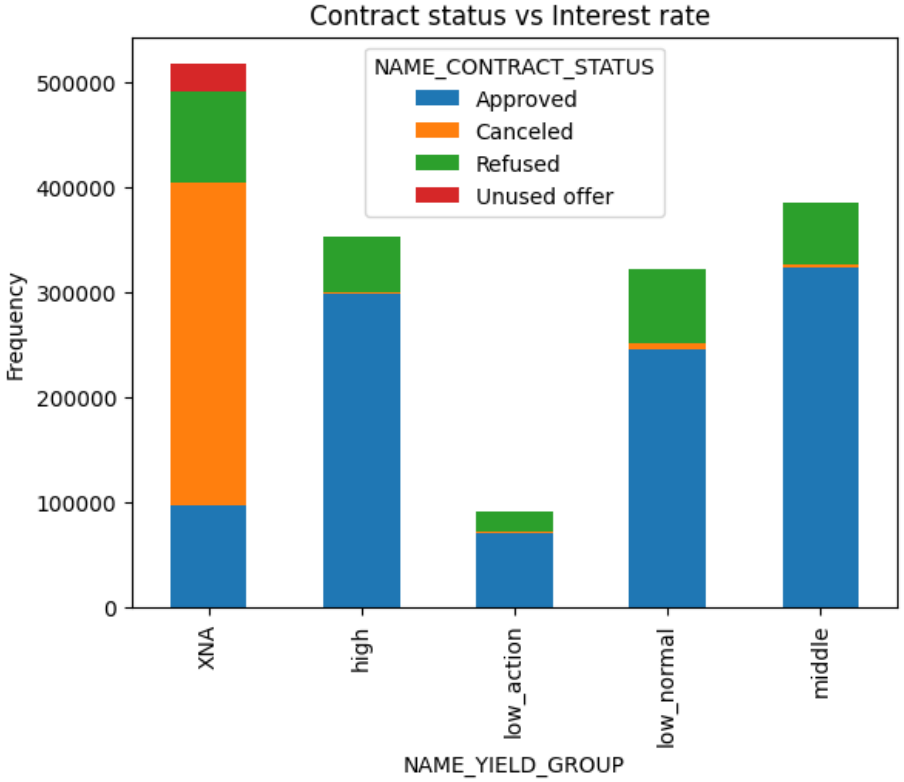
- As we know 73.7% of client is repeat customer
- Hence for Repeat customer we see prominent approval , cancelation and refusal rate
- Negligible Cancellation/Refusal rate for New and Refreshed client due to their low proportion in the data

Goods Category Vs Contract Status



- Approval rate for Animals is 100% followed by Fitness is 99.04% and least is for House Construction 0%
- What kind of goods did the client applied for is not available and there contribution is very high and hence they have higher cancellation, refusal and approval rate

Interest Rate Vs Contract Status



- Approval rate for High is 84.62% followed by middle is 83.78% and least is for XNA is 18.90%

- We see very few low_action interest rate gets approved
- Some approved contract does not have interest rate defined
- High, middle, low normal have nearly equal contribution in approved contract

Dashboards

Application Dashboard

308K	25K	$165.61K	$168.80K	3.56	3.37
Total Application Count	Defaulted Application	Average Defaulter Income	Average Income Amount	Credit to Income (Non Defaulter)	Credit to Income (Defaulter)

NAME_EDUCATION_TY...
- Academic degree
- Higher education
- Incomplete higher
- Lower secondary
- Secondary / secondar...

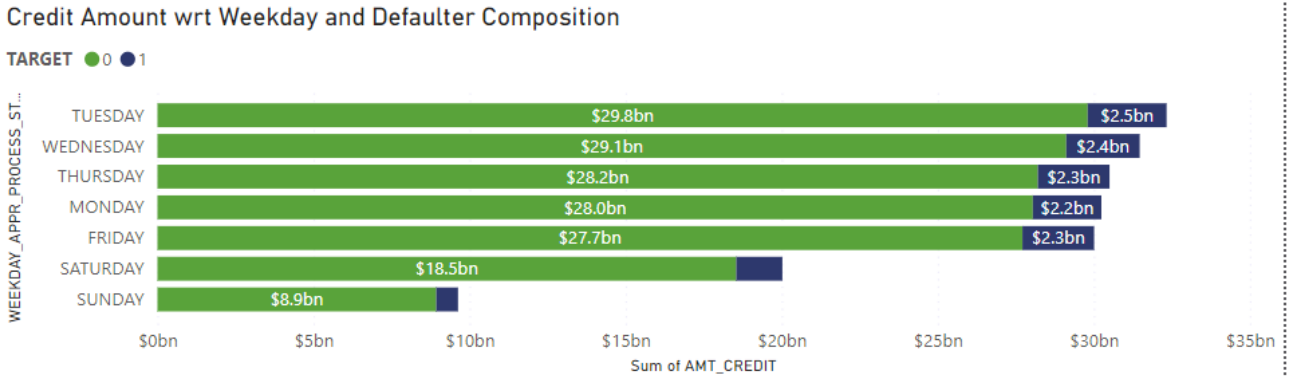
NAME_HOUSING_TYPE
- Co-op apartment
- House / apartment
- Municipal apartment
- Office apartment
- Rented apartment

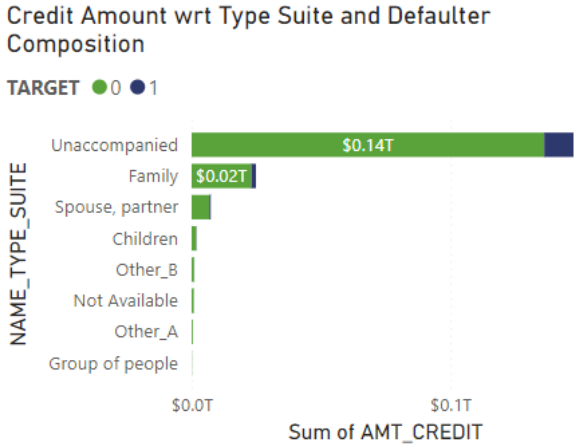
NAME_FAMILY_STATUS
- Civil marriage
- Married
- Separated
- Single / not married
- Unknown
- Widow

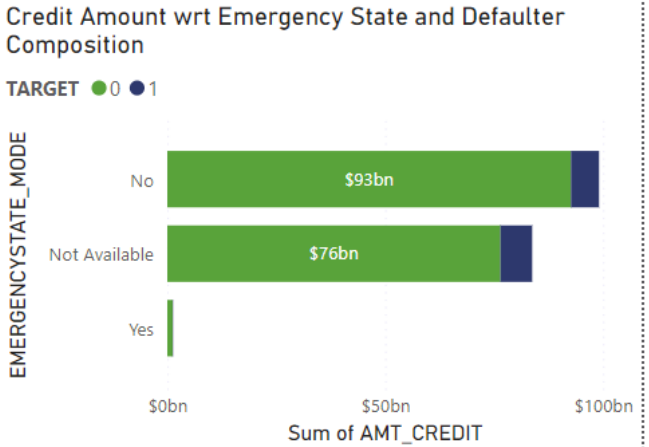
NAME_CONTRAC...
- Cash loans
- Revolving loans

NAME_INCOME_T...
- Businessman
- Commercial asso...
- Maternity leave
- Pensioner
- State servant
- Student
- Unemployed
- Working

NAME_TYPE_SUI...
- Children
- Family
- Group of people
- Not Available
- Other_A
- Other_B







Financial Dashboard

$8.34bn
Total Annuity Amount

$184.21bn
Total Credit Amount

$13.85bn
Defaulted Credit Amount

NAME_EDUCATION_TYPE
- [ ] Academic degree
- [ ] Higher education
- [ ] Incomplete higher
- [ ] Lower secondary
- [ ] Secondary / secondar...

NAME_HOUSING_TYPE
- [ ] Co-op apartment
- [ ] House / apartment
- [ ] Municipal apartment
- [ ] Office apartment
- [ ] Rented apartment
- [ ] With parents

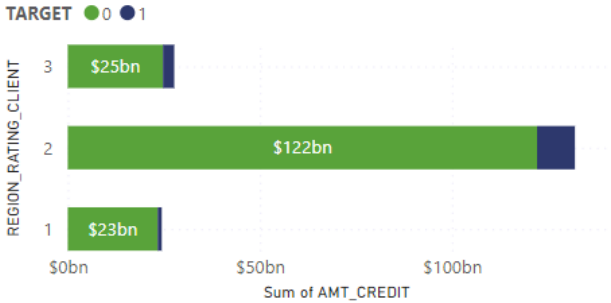
NAME_FAMILY_STATUS
- [ ] Civil marriage
- [ ] Married
- [ ] Separated
- [ ] Single / not married
- [ ] Unknown
- [ ] Widow

NAME_CONTRAC...
- [ ] Cash loans
- [ ] Revolving loans

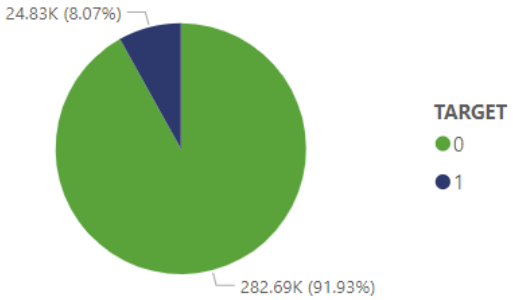
NAME_INCOME_T...
- [ ] Businessman
- [ ] Commercial asso...
- [ ] Maternity leave
- [ ] Pensioner
- [ ] State servant
- [ ] Student
- [ ] Unemployed
- [ ] Working

NAME_TYPE_SUI...
- [ ] Children
- [ ] Family
- [ ] Group of people
- [ ] Not Available
- [ ] Other_A
- [ ] Other_B

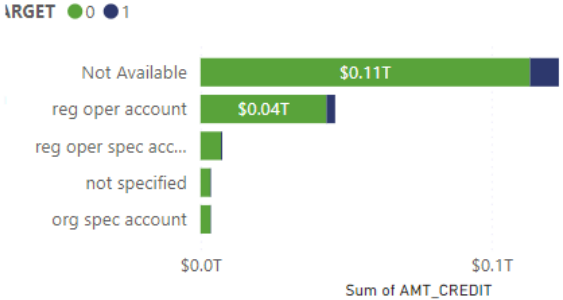
Credit Amount wrt Region Rating and Defaulter Composition



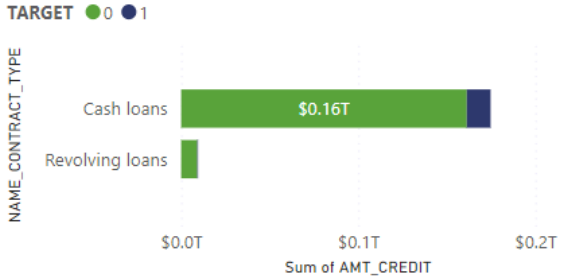
Defaulter Non-Defaulter Composition



redit Amount wrt Capital Repair Fund and Defaulter omposition



Credit Amount wrt Contract type and Defaulter Composition



Previous Applicant's Dashboard

1.67M
Total Applicant's

15.96K
Average Annuity Amount

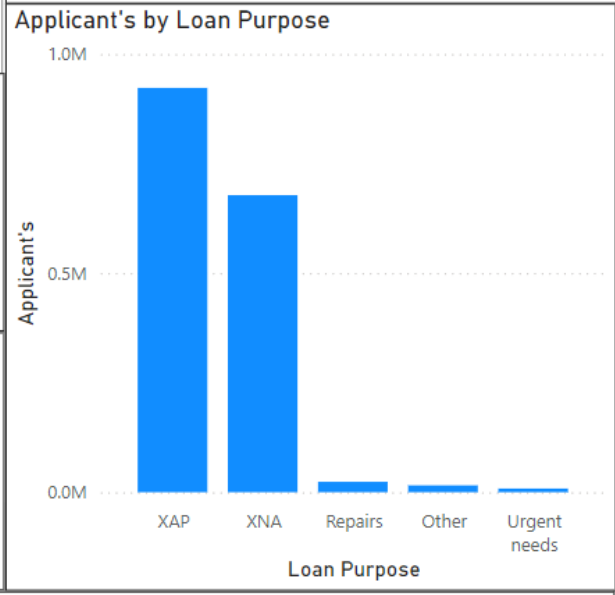
175.23K
Average Applied Amount

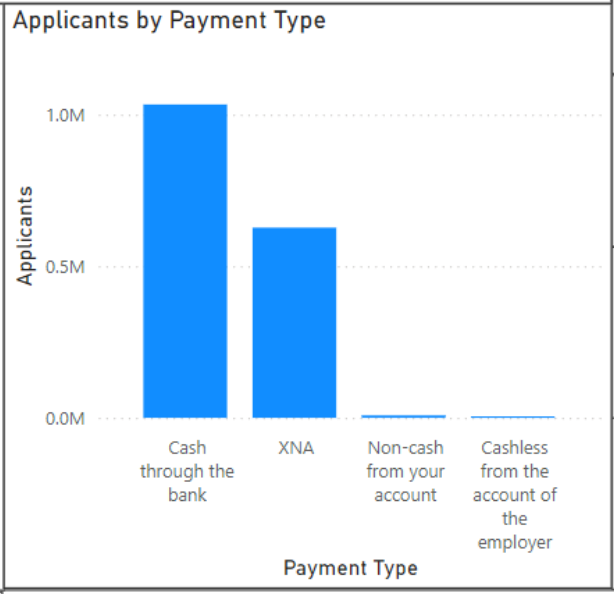
196.11K
Average Credit Amount

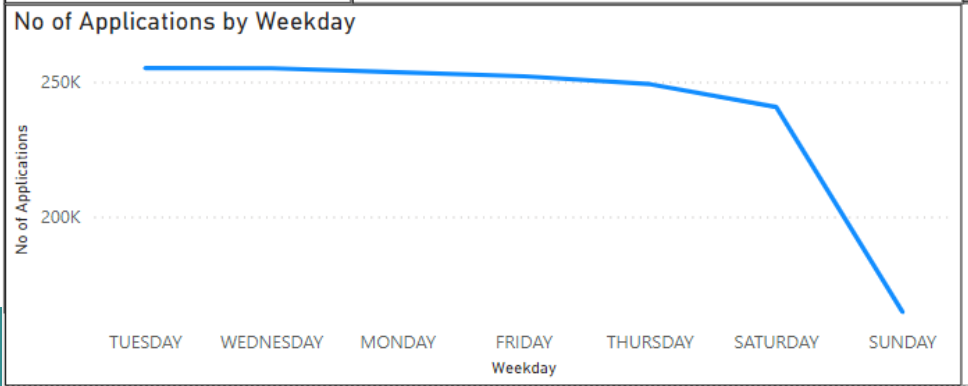
227.85K
Average Goods Price Am...

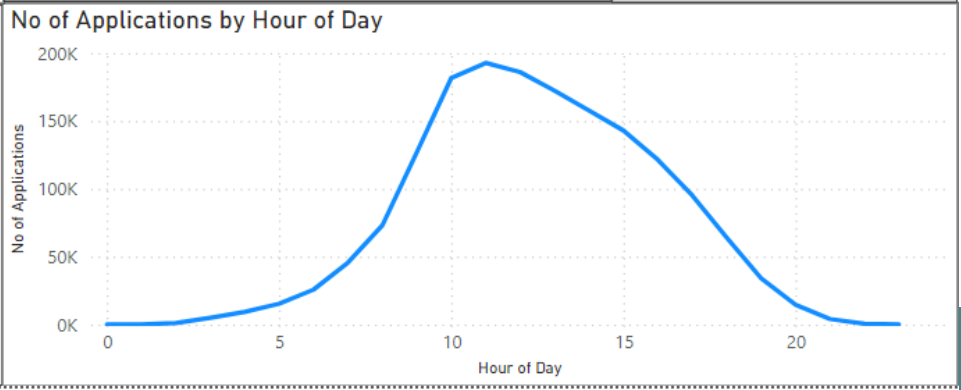
NAME_CONTRACT_TYPE
- [ ] Cash loans
- [ ] Consumer loans
- [ ] Revolving loans
- [ ] XNA

NAME_CONTRACT_STATUS
- [ ] Approved
- [ ] Canceled
- [ ] Refused
- [ ] Unused offer





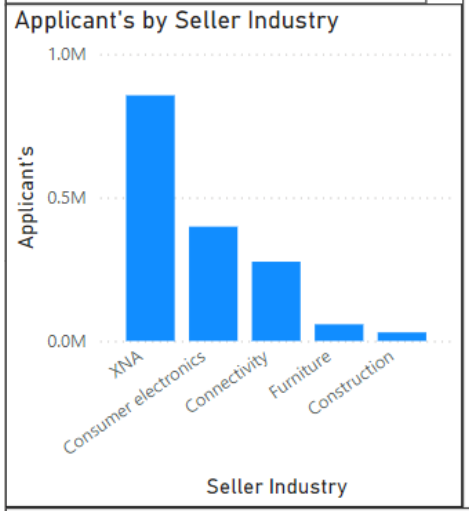


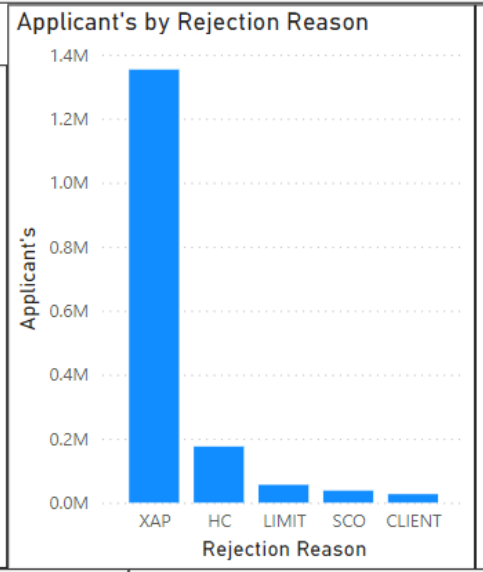


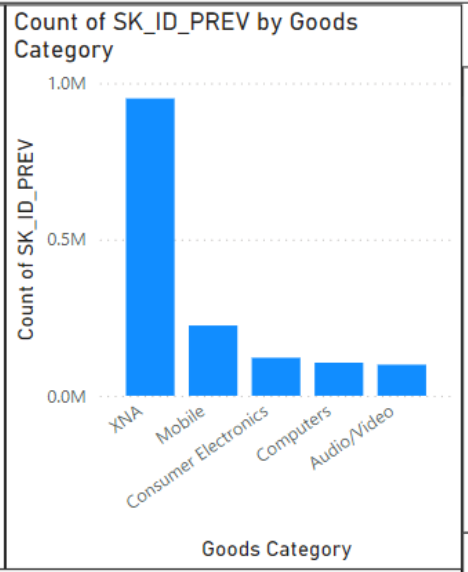
Previous Applicant's Dashboard

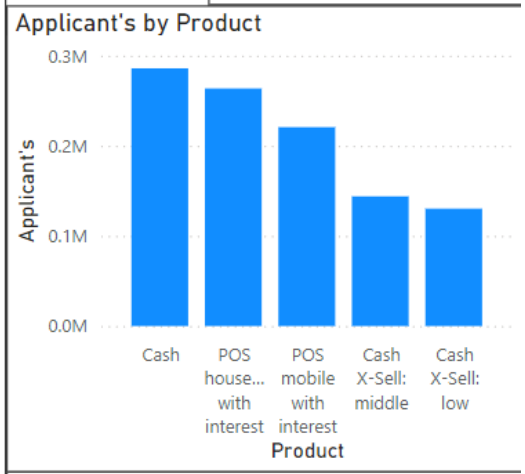
NAME_CLIENT_TYPE
- [ ] New
- [ ] Refreshed
- [ ] Repeater

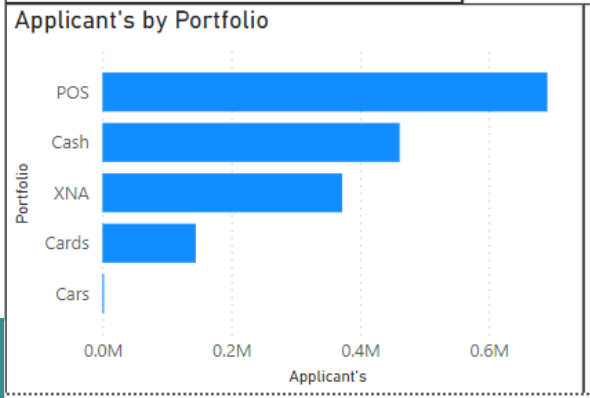
NAME_PRODUCT_TYPE
- [ ] walk-in
- [ ] XNA
- [ ] x-sell

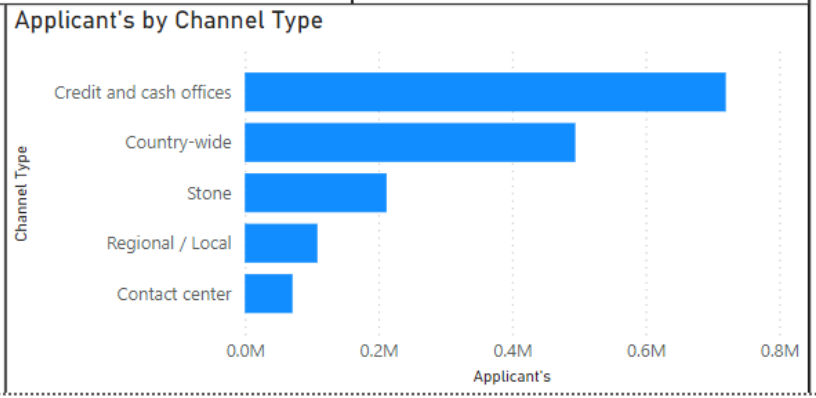


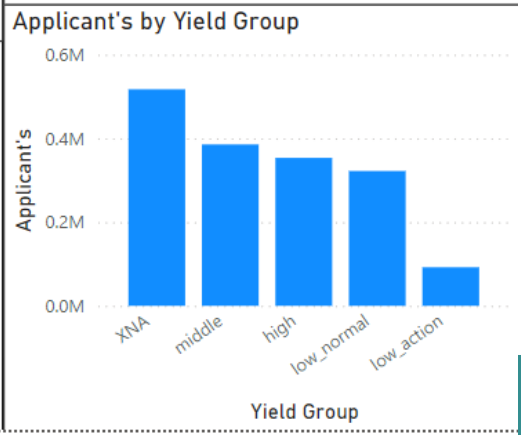






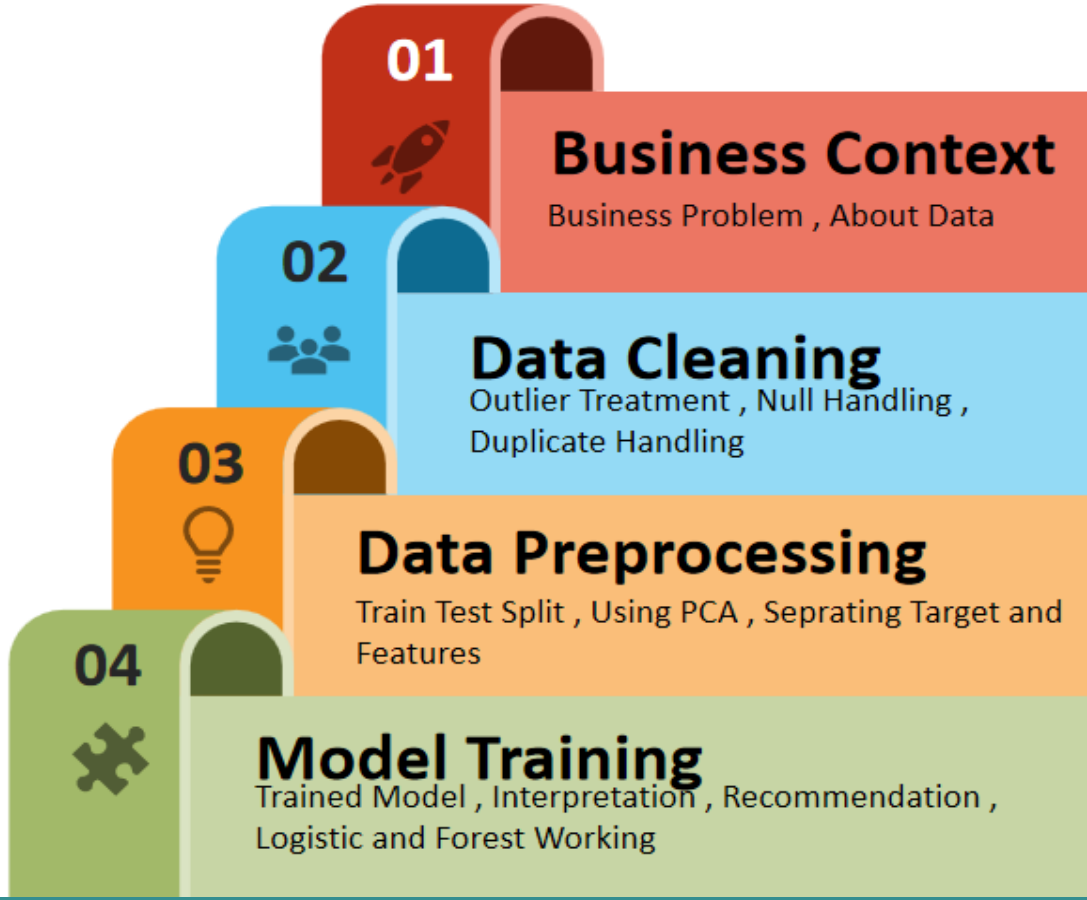






Predictive Modelling

Agenda


01
Business Context
Business Problem , About Data
02
Data Cleaning
Outlier Treatment , Null Handling , Duplicate Handling
03
Data Preprocessing
Train Test Split , Using PCA , Seprating Target and Features
04
Model Training
Trained Model , Interpretation , Recommendation , Logistic and Forest Working

Business Problem -

The goal is to identify patterns that indicate a higher risk of default, which can inform actions such as denying the loan, reducing the loan amount, or charging higher interest rates to risky applicants with the help of predictive modelling. By doing so, lenders can better ensure that only creditworthy consumers receive loans.

About Data -

application_data.csv It contain about 122 columns and 307511 records and contain all information that are related to client. (like - Loan Information,Client Demographics,Document Information,Credit Bureau Inquiries etc.)

Data Cleaning-

Null Handling

Removing all the columns where the null value percentage is greater than 20 with the help of slicing and dicing.

01

02

Filling all the records where the null value precentage is less than and equal to 20 with the help of mean(for continuous variable) and mode(for categorical variables).

Null Handling

Duplicate Handle

Checking for duplicate records but there is no duplicate records in the table.

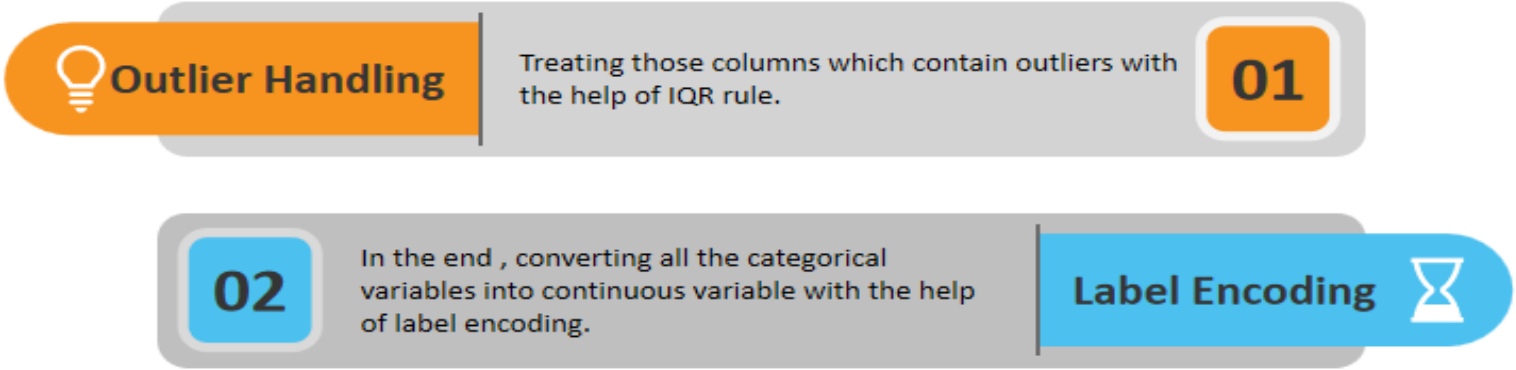
03

04

Checking correlation for the numeric data with the help of heatmap.Removing all the columns where correlation is greater than and equal to 0.80 or less than and equal to 0.80.

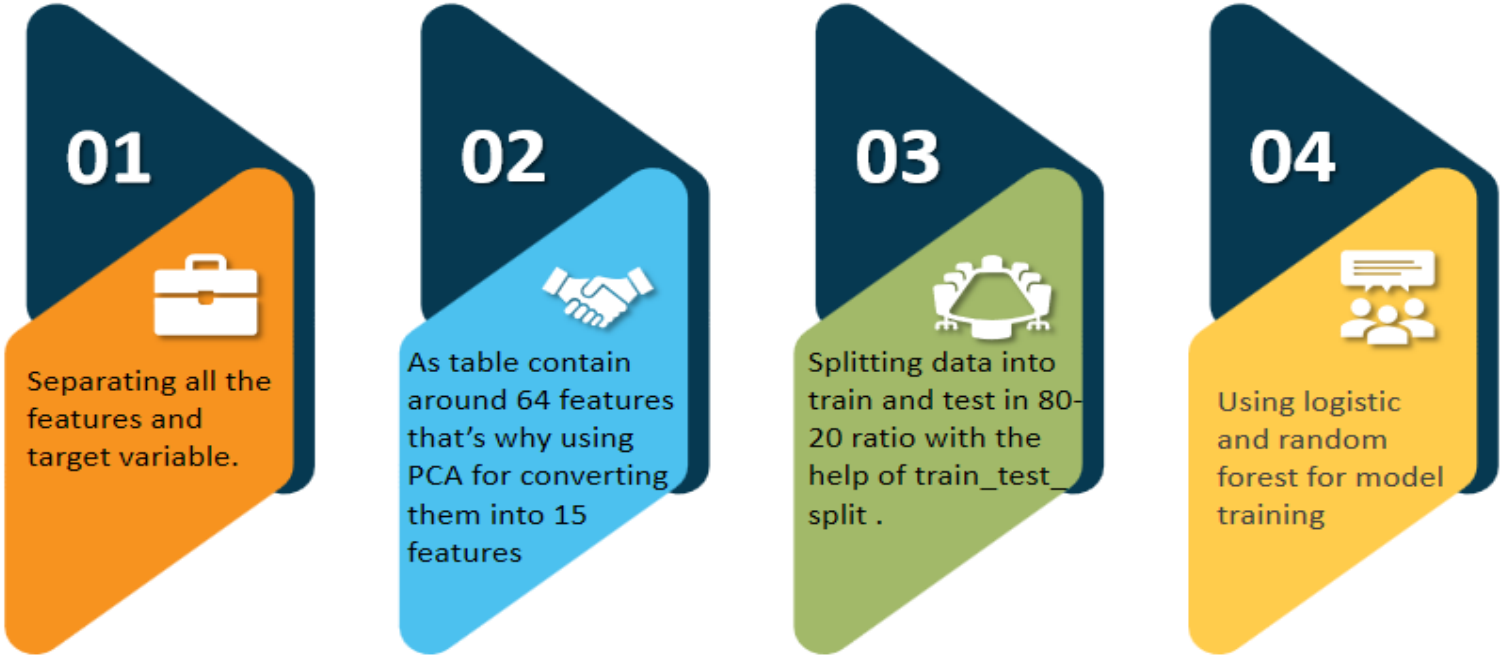
Correlation Check

Data Cleaning-



Note : After cleaning , The final data in table contains about 64 columns and 307511 records.

Data Preprocessing -



Logistic Regression-

$$\log\left[\frac{p(x)}{1-p(x)}\right] = z$$

$$\log\left[\frac{p(x)}{1-p(x)}\right] = w \cdot X + b$$

$$\frac{p(x)}{1-p(x)} = e^{w \cdot X + b} \quad \cdots \text{Exponentiate both sides}$$

$$p(x) = e^{w \cdot X + b} \cdot (1 - p(x))$$

$$p(x) = e^{w \cdot X + b} - e^{w \cdot X + b} \cdot p(x))$$

$$p(x) + e^{w \cdot X + b} \cdot p(x)) = e^{w \cdot X + b}$$

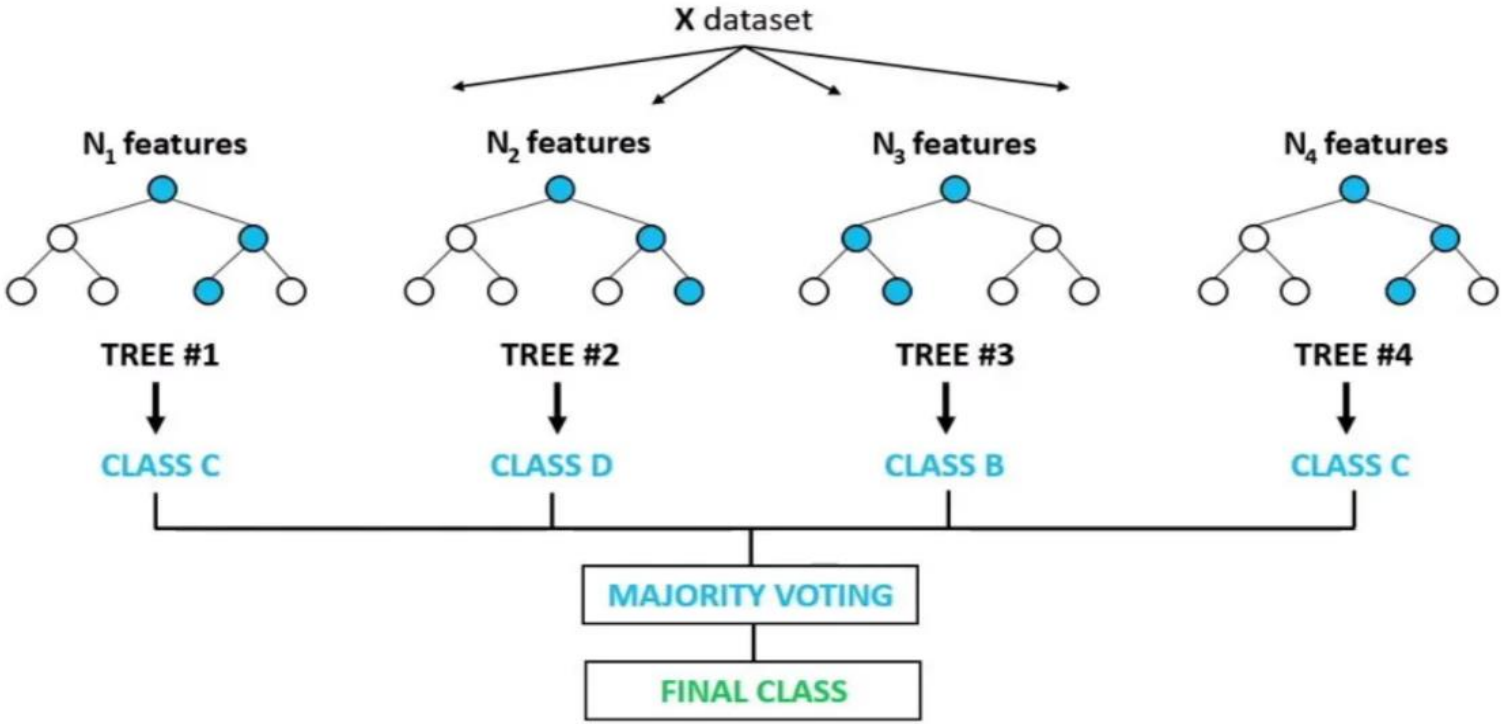
$$p(x)(1 + e^{w \cdot X + b}) = e^{w \cdot X + b}$$

$$p(x) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}}$$

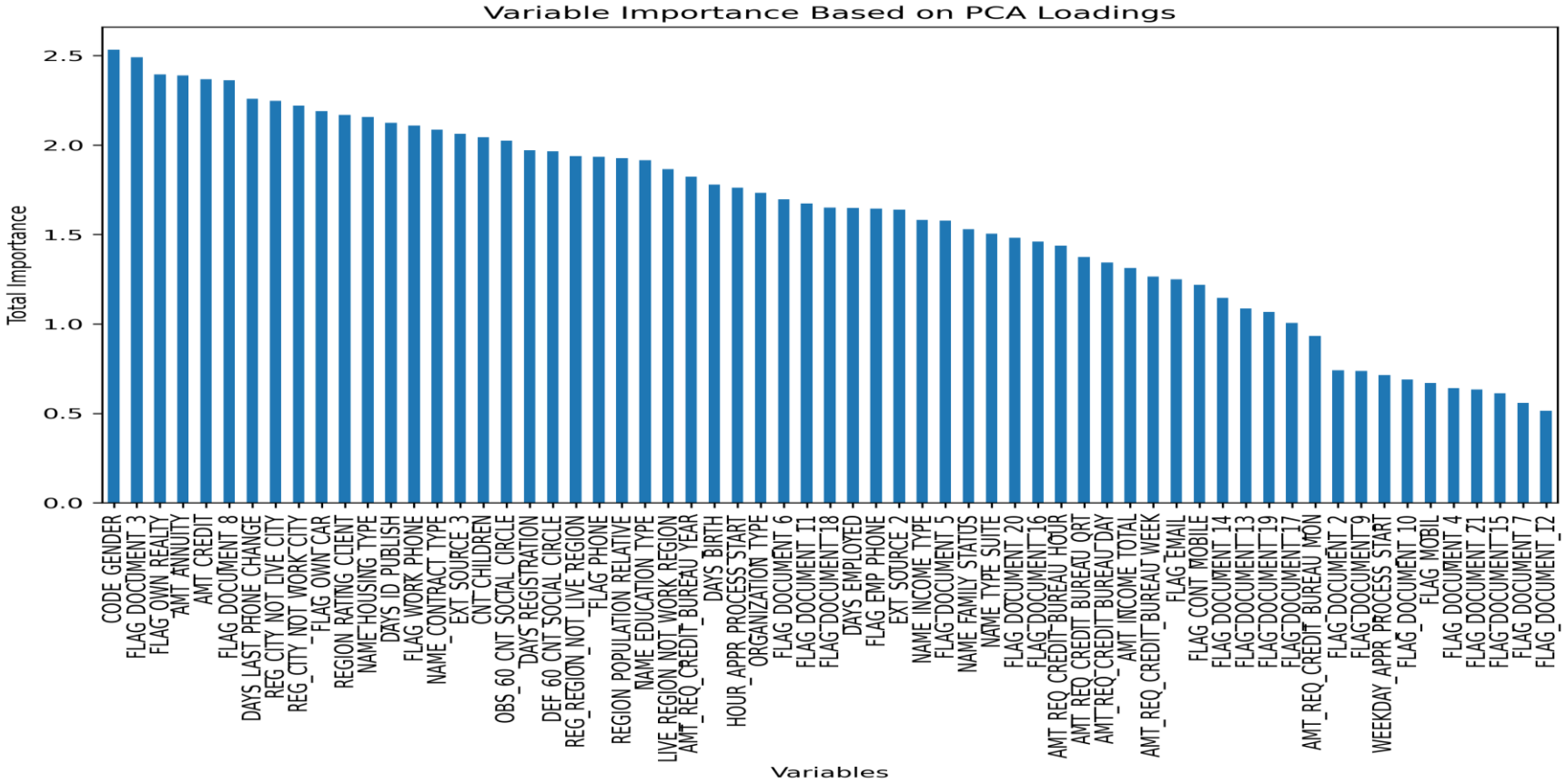
then the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}}$$

Random Forest Classifier -



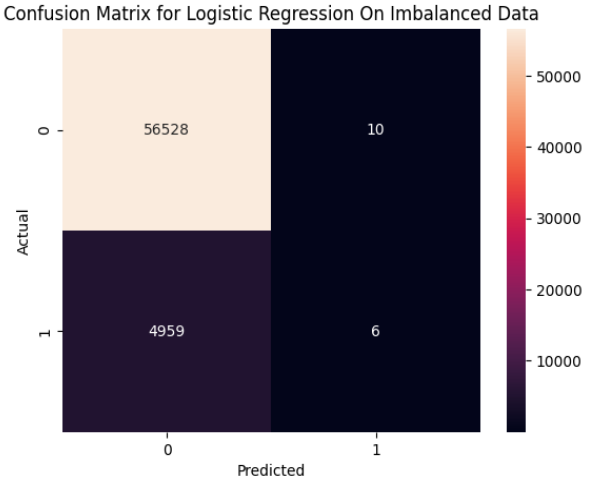
Importance Of Variables -



Model-01 -

With Logistic Regression Classification Report -

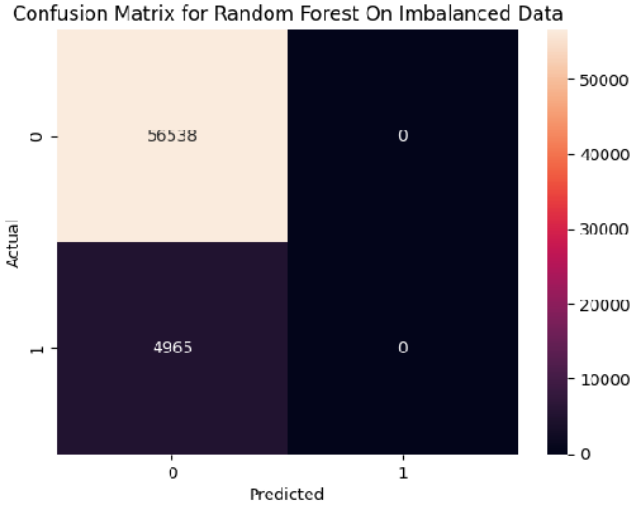
	precision	recall	f1-score	support
0	0.92	1.00	0.96	56538
1	0.22	0.00	0.00	4965
accuracy			0.92	61503
macro avg	0.57	0.50	0.48	61503
weighted avg	0.86	0.92	0.88	61503



Model-02 -

With Random Forest Classification Report -

```
# Checking accuracy score at diffrent-diffrent depth
accuracy_scores = []
max_depths = []

for i in range(5, 1, -1):
    rf = RandomForestClassifier(max_depth=i)
    rf.fit(xtrain, ytrain)
    pred1 = rf.predict(xtest)

    acc = accuracy_score(ytest, pred1)

    accuracy_scores.append(acc)
    max_depths.append(i)
```




```
accuracy_scores,max_depths
```

```
([0.9192722306228964,
  0.9192722306228964,
  0.9192722306228964,
  0.9192722306228964],
 [5, 4, 3, 2])
```

Interpretation For Model 01 & 02 -

- **High Performance for Class 0**: The model performs very well in identifying non default cases (Class 0), with high precision, recall, and F1-score.
- **High Performance for Class 0**: The model performs very well in identifying non-default cases (Class 0), with high precision, recall, and F1-score.
- **Poor Performance for Class 1**: The model fails to identify any default cases (Class 1), as indicated by the precision, recall, and F1-score being zero. This indicates a potential issue with model bias, likely favoring the majority class.
- **Overall Accuracy**: While the overall accuracy is high (92%), it can be misleading due to the class imbalance in the dataset.

Using Random Sampling Approach -

Using random sampling because the dataset are present into imbalance form . Random sampling will pick the same number of records from class 0

```
app['TARGET'].value_counts()
```

```
TARGET
0    282686
1     24825
Name: count, dtype: int64
```

Before Random Sampling

```
app1['TARGET'].value_counts()
```

```
TARGET
1    24825
0    24825
Name: count, dtype: int64
```

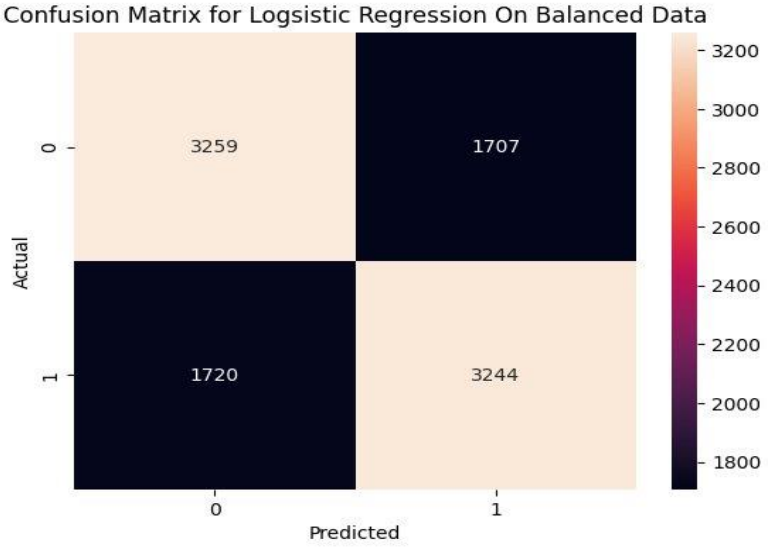
After Random Sampling

Model-03 -

With Logistic Regression Classification Report -

```
print(classification_report(ytest,pred))
```

	precision	recall	f1-score	support
0	0.65	0.64	0.65	4966
1	0.65	0.65	0.65	4964
accuracy			0.65	9930
macro avg	0.65	0.65	0.65	9930
weighted avg	0.65	0.65	0.65	9930



Model-04 -

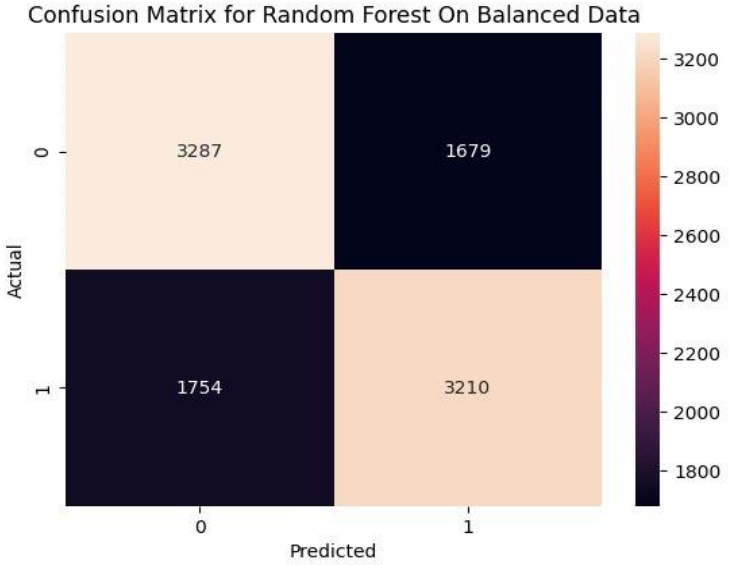
With Random Forest Classification Report -

```
# Checking accuracy score at diffrent-diffrent depth
accuracy_scores = []
max_depths = []

for i in range(20, 1, -1):
    rf = RandomForestClassifier(max_depth=i)
    rf.fit(xtrain, ytrain)
    pred1 = rf.predict(xtest)

    acc = accuracy_score(ytest, pred1)

    accuracy_scores.append(acc)
    max_depths.append(i)
```

```
accuracy_scores,max_depths
```

```
([0.6468277945619335,
  0.6430010070493454,
  0.6473313192346425,
  0.6462235649546828,
  0.6503524672708962,
  0.6479355488418933,
  0.6510574018126888,
  0.6504531722054381,
  0.6529707955689829,
  0.6522658610271903,
  0.6470292044310171,
  0.649546827794562,
  0.6484390735146022,
  0.6467270896273918,
  0.6452165156092649,
  0.641792547834844,
  0.641289023162135,
  0.6415911379657603,
  0.6302114803625377],
 [20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2])
```

	precision	recall	f1-score	support
0	0.65	0.66	0.66	4966
1	0.66	0.65	0.65	4964
accuracy			0.65	9930
macro avg	0.65	0.65	0.65	9930
weighted avg	0.65	0.65	0.65	9930



Interpretation For Model 03 & 04 -

- **Balanced Performance**: The model demonstrates balanced performance across both classes, with precision, recall, and F1-scores all at 0.65 for both Class 0 and Class 1.
- **Moderate Accuracy**: The overall accuracy is also 0.65, indicating that the model is correctly predicting around 65% of the instances.

Recommendation -

- **Threshold Optimization**: Experiment with adjusting the classification threshold to improve precision or recall based on specific business needs.


THANK YOU