LOAN APPROVAL DATASET

EDA Using SQL

BY

SANDRA GANESHAN

Introduction

The lending industry plays a critical role in personal finance and economic stability. Understanding the dynamics of loan applicants, including their demographics, income levels, and loan purposes, is essential for optimizing lending practices and improving customer service. This project aims to perform an in-depth Exploratory Data Analysis (EDA) on the loan dataset to extract meaningful insights that can help enhance loan approval processes, tailor financial products, and improve overall business performance.

Objective:

- Demographic Analysis: Understanding the demographic distribution of loan applicants, including age, gender, and education levels.
- Income Insights: Analyzing income levels and categorizing them to understand borrower profiles better.
- Loan Purpose Analysis: Investigating the primary reasons borrowers seek loans and their impact on approval rates.
- Loan Amount and Interest Rates: Studying variations in loan amounts and interest rates to identify trends and potential areas for adjustment.
- Debt-to-Income Ratio: Exploring how applicants' debt-to-income ratios affect loan approval and risk assessments.

Aim:

- Optimize Lending Strategies: By identifying key demographics and income levels, lenders can tailor their loan offerings to meet customer needs.
- Enhance Customer Understanding: Analyzing loan purposes and income categories provides insights into borrower motivations, helping institutions refine their marketing and customer service approaches.
- Assess Risk Factors: Understanding the correlation between income, loan amounts, and debt-to-income ratios can help lenders evaluate risks associated with loan approvals.
- Inform Policy Changes: Insights gained from the analysis can guide policy adjustments to improve loan offerings and reduce defaults.

Data Overview

Dataset Structure:

- **Total Rows**: 45000

Column Description

person_age	Age of the applicant
person_gender	Gender of the applicant
person_education	Education level of the applicant
person_income	Annual income of the applicant
person_emp_exp	Years of employment experience
person_home_ownership	Home ownership status (Rent/Own)
loan_amnt	Loan amount requested
loan_intent	Purpose of the loan (e.g., Education, Home, etc.)
loan_int_rate	Interest rate for the loan
loan_percent_income	Percentage of income that goes towards loan repayment
cb_person_cred_hist_length	Length of credit history
credit_score	Credit score of the applicant
previous_loan_defaults_on_file	Count of previous loan defaults
loan_status	Status of the loan (Approved/Denied)

person_age	person_gen	person_ed	person_inc	person_em	person_ho	loan_amnt	loan_inten	loan_int_r	loan_perce	cb_person	credit_sco	previous_l	loan_status
22	female	Master	71948	0	RENT	35000	PERSONAL	16.02	0.49	3	561	No	1
21	female	High Schoc	12282	0	OWN	1000	EDUCATIO	11.14	0.08	2	504	Yes	0
25	female	High Schoc	12438	3	MORTGAG	5500	MEDICAL	12.87	0.44	3	635	No	1
23	female	Bachelor	79753	0	RENT	35000	MEDICAL	15.23	0.44	2	675	No	1
24	male	Master	66135	1	RENT	35000	MEDICAL	14.27	0.53	4	586	No	1
21	female	High Schoc	12951	0	OWN	2500	VENTURE	7.14	0.19	2	532	No	1
26	female	Bachelor	93471	1	RENT	35000	EDUCATIO	12.42	0.37	3	701	No	1
24	female	High Schoc	95550	5	RENT	35000	MEDICAL	11.11	0.37	4	585	No	1
24	female	Associate	100684	3	RENT	35000	PERSONAL	8.9	0.35	2	544	No	1
21	female	High Schoc	12739	0	OWN	1600	VENTURE	14.74	0.13	3	640	No	1
22	female	High Schoc	102985	0	RENT	35000	VENTURE	10.37	0.34	4	621	No	1
21	female	Associate	13113	0	OWN	4500	HOMEIMP	8.63	0.34	2	651	No	1
23	male	Bachelor	114860	3	RENT	35000	VENTURE	7.9	0.3	2	573	No	1
26	male	Master	130713	0	RENT	35000	EDUCATIO	18.39	0.27	4	708	No	1
23	female	Associate	138998	0	RENT	35000	EDUCATIO	7.9	0.25	4	583	No	0
23	female	Master	600891	5	MORTGAG	30000	DEBTCONS	10.65	0.05	3	670	Yes	0
23	male	Bachelor	144943	0	RENT	35000	EDUCATIO	7.9	0.24	4	663	No	0
23	female	High Schoc	111369	0	RENT	35000	MEDICAL	20	0.31	4	694	No	1
22	male	Bachelor	126628	0	RENT	35000	DEBTCONS	18.25	0.26	4	709	No	1

Data Cleaning & Preparation

To ensure the dataset is suitable for accurate analysis, a series of data cleaning and preparation steps were undertaken to eliminate inconsistencies, improve clarity, and structure the data for efficient querying.

---

1. Duplicate Checking: *Duplicate records can lead to inaccurate analysis by inflating the number of entries for restaurants or skewing the results. To resolve this:*

- All columns were checked for duplicate values, confirming that no critical fields contained redundant data, ensuring data completeness for analysis.

---

2. Column Renaming: *For consistency and better readability, column names were standardized:*

1. **Income Category:** Categorized applicants' income into 'LOW', 'MODERATE', and 'HIGH'.
2. **Credit Score Category:** Categorized applicants' credit scores into 'LOW', 'FAIR', and 'HIGH'.
3. **Debt-to-Income Ratio (DTI):** Calculated as the percentage of income that goes toward loan repayment.
4. **Age Category:** Categorized applicants into 'Young Adult', 'Adult', 'Middle-aged', and 'Senior'.

---

3. Null Value Checking: *Null values can cause errors in data analysis or skew results if left unchecked. Therefore:*

- All columns were checked for null values, confirming that no critical fields contained missing data. This step ensured data completeness for analysis.

---

4. Outlier Handling:

- Outliers in age (e.g., ages over 100) were removed to maintain the integrity of the analysis.

DELETE FROM loan_data WHERE person_age > 100;

SELECT COUNT(*) FROM LOAN_DATA; #44993

Univariate Analysis

#1.Income statistics

```
SELECT
   MIN(person_income) AS min_income,  #8000
   MAX(person_income) AS max_income,  #2448661
   AVG(person_income) AS avg_income,  #79908.44758073478
   STDDEV(person_income) AS stddev_income  #63321.42865625254
FROM loan_data;
```

#2.Distribution by income ranges

```
SELECT CASE
      WHEN person_income < 20000 THEN 'Under 20K'
      WHEN person_income BETWEEN 20000 AND 50000 THEN '20K-50K'
      WHEN person_income BETWEEN 50000 AND 100000 THEN '50K-100K'
      ELSE 'Over 100K'
   END AS income_range,COUNT(*) AS count
FROM loan_data
GROUP BY income_range
ORDER BY count DESC;
```

RESULT:

- Most people comes under 50 to 100k range. and least in under 20k range

---

#3.Loan_Intent

```
SELECT LOAN_INTENT,COUNT(*) AS COUNT_OF_EACH
FROM LOAN_DATA
GROUP BY LOAN_INTENT
ORDER BY COUNT_OF_EACH DESC;
```

RESULT:

- So most people took loan for 'Education' and least for 'Home Improvement'.

#4.Person_age

```
SELECT PERSON_AGE,COUNT(*) AS AGE_COUNT FROM LOAN_DATA GROUP BY
PERSON_AGE ORDER BY AGE_COUNT desc;
```

```
SELECT AGE_CATEGORY,COUNT(*) AS AGE_COUNT FROM LOAN_DATA GROUP BY
AGE_CATEGORY ORDER BY AGE_COUNT desc;
```

RESULT:

- Highest age_group-23(5254). so mostly the age range is in b/w 20 and 30
- Age range-20 to 144(So there are outliers)
- Highest->Adult-27227

#5.Gender:

SELECT PERSON_GENDER,COUNT(*) AS GENDER_COUNT
FROM LOAN_DATA GROUP BY PERSON_GENDER ORDER BY GENDER_COUNT DESC;

RESULT:

- Male-24841
- Female-20159

#6.Education

SELECT PERSON_EDUCATION,COUNT(*) AS EDU_COUNT
FROM LOAN_DATA
GROUP BY PERSON_EDUCATION
ORDER BY EDU_COUNT DESC;

RESULT:

- Most of them 'Bachelors' degree holder and least Doctorate.
- There are a total of 5categories

#7.Loan Amount Statistics

SELECT
  MIN(loan_amnt) AS min_amnt, #500
  MAX(loan_amnt) AS max_amnt, #35000
  AVG(loan_amnt) AS avg_amnt, #9583.157555555556
  STDDEV(loan_amnt) AS stddev_amnt #6314.816524743697
FROM loan_data;

#8.Employee experience count

SELECT PERSON_EMP_EXP,COUNT(*) AS EXP_COUNT
FROM LOAN_DATA
GROUP BY PERSON_EMP_EXP
ORDER BY EXP_COUNT DESC;

RESULT

- Most of the people who took loan have '0' years experience.

#9.Home Ownership

SELECT person_home_ownership,COUNT(*) AS ownership_count
FROM loan_data
GROUP BY person_home_ownership
ORDER BY ownership_count DESC;

RESULT

- Most people lives for 'Rent'

#10.Loan Interest Rate Analysis

```
SELECT
   MIN(loan_int_rate) AS min_int_rate,  -- Lowest interest rate--5.42
   MAX(loan_int_rate) AS max_int_rate,  -- Highest interest rate--20
   AVG(loan_int_rate) AS avg_int_rate,  -- Average interest rate--11.00644789189424
   STDDEV(loan_int_rate) AS stddev_int_rate  -- Std deviation of int_rate--2.9789523311281356
  FROM loan_data;
```

#11.Credit History Length Analysis

```
SELECT
   MIN(cb_person_cred_hist_length) AS min_hist_length, #2
   MAX(cb_person_cred_hist_length) AS max_hist_length, #30
   AVG(cb_person_cred_hist_length) AS avg_hist_length, #5.866557019980886
   STDDEV(cb_person_cred_hist_length) AS stddev_hist_length #3.8771238852626158
FROM loan_data;
```

#12. Loan Amount to Income Ratio Analysis

```
SELECT
   MIN(loan_percent_income) AS min_percent_income, #0
   MAX(loan_percent_income) AS max_percent_income, #0.66--66% of their income goes to loan
   AVG(loan_percent_income) AS avg_percent_income, #0.14
   STDDEV(loan_percent_income) AS stddev_percent_income #0.09
FROM loan_data;
```

#13.Debt-to-income ratio

```
SELECT
   MIN(dti_ratio) AS min, #0.0657612
   MAX(dti_ratio) AS max, #66.4186
   AVG(dti_ratio) AS avg, #14
   STDDEV(dti_ratio) AS stddev #8.71
FROM loan_data;

#for categorization

SELECT CASE
      WHEN dti_ratio < 20 THEN 'Under 20%'
      WHEN dti_ratio BETWEEN 20 AND 35 THEN '20%-35%'
      WHEN dti_ratio BETWEEN 35 AND 50 THEN '35%-50%'
      ELSE 'Over 50%'
   END AS dti_range,COUNT(*) AS count
FROM loan_data
GROUP BY dti_range
ORDER BY count DESC;
```

RESULT:

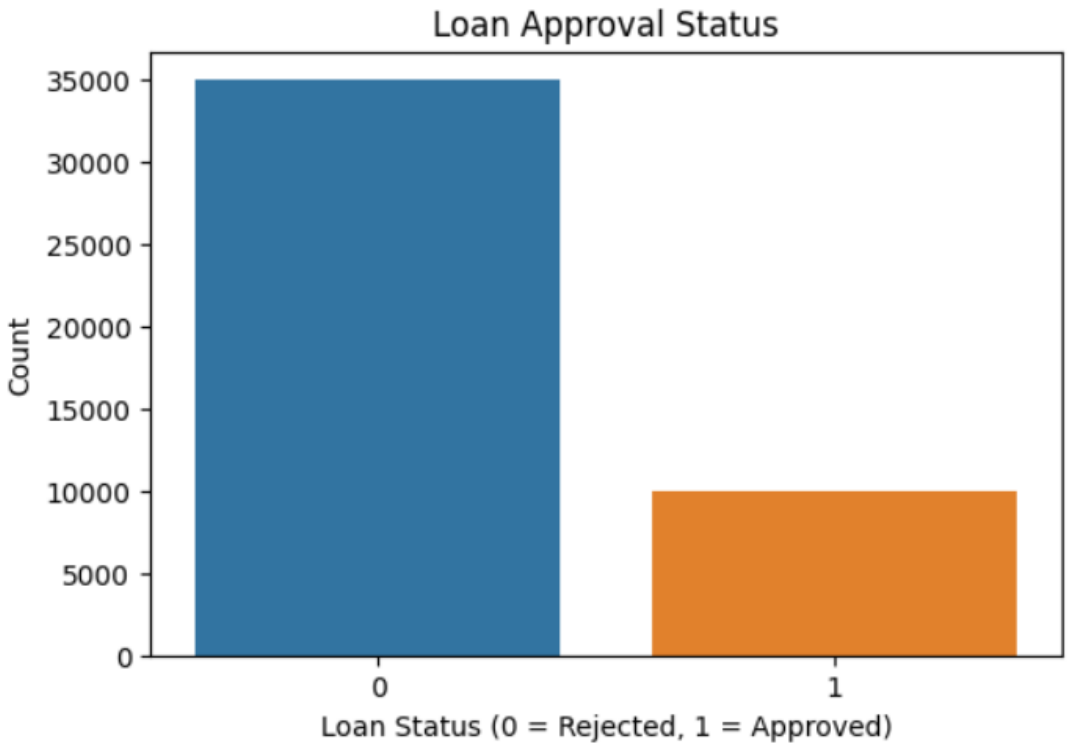
- Highest in Under 20%

#14.Loan Status Analysis

SELECT loan_status, COUNT(*) AS count
FROM loan_data
GROUP BY loan_status
ORDER BY count DESC;

RESULT:

- 0(Rejected)-34993,1(Approved)-10000



Bivariate Analysis

1.Home Ownership Analysis by Loan Intent(To explore if loan purpose varies with homeownership

SELECT PERSON_HOME_OWNERSHIP,LOAN_INTENT,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY PERSON_HOME_OWNERSHIP,LOAN_INTENT
ORDER BY COUNT DESC;

RESULT:

- **Renters:** Primarily borrow for medical, education, and debt consolidation, suggesting higher financial needs.
- **Mortgage Holders:** Frequently seek loans for education, personal needs, ventures, and home improvement.
- **Fully Owned Homes:** Lower loan demand, focusing on ventures and education, indicating more financial stability.
- **Other Homeownership:** Least loan activity, with a slight focus on ventures and personal needs.

2.Analyzing Income Category vs. Average Loan Amount

SELECT INCOME_CATEGORY,AVG(LOAN_AMNT) AS AVG_LOAN_AMT,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY INCOME_CATEGORY
ORDER BY COUNT DESC;

INCOME_CATEGORY	AVG_LOAN_AMT	COUNT
MODERATE	8686.182726287518	21223
HIGH	12486.446632124353	16405
LOW	5701.1349626612355	7365

RESULT:

- **Moderate-Income Category** has the highest count of loan requests, despite having a lower average loan amount compared to the **High-Income Category.**
- The lower average loan amount in the moderate category compared to the high category may indicate a more cautious approach to borrowing. Moderate-income earners may be more conservative in their financial planning, seeking smaller loans that are easier to repay.
- With a high number of loans from moderate earners, lenders might view this segment as a stable customer base, whereas high-income individuals might have fewer loan requests due to existing wealth or alternative financing options.

3.Loan Purposes by Income Category

SELECT INCOME_CATEGORY,LOAN_INTENT,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY INCOME_CATEGORY, LOAN_INTENT

ORDER BY COUNT DESC;

RESULT:

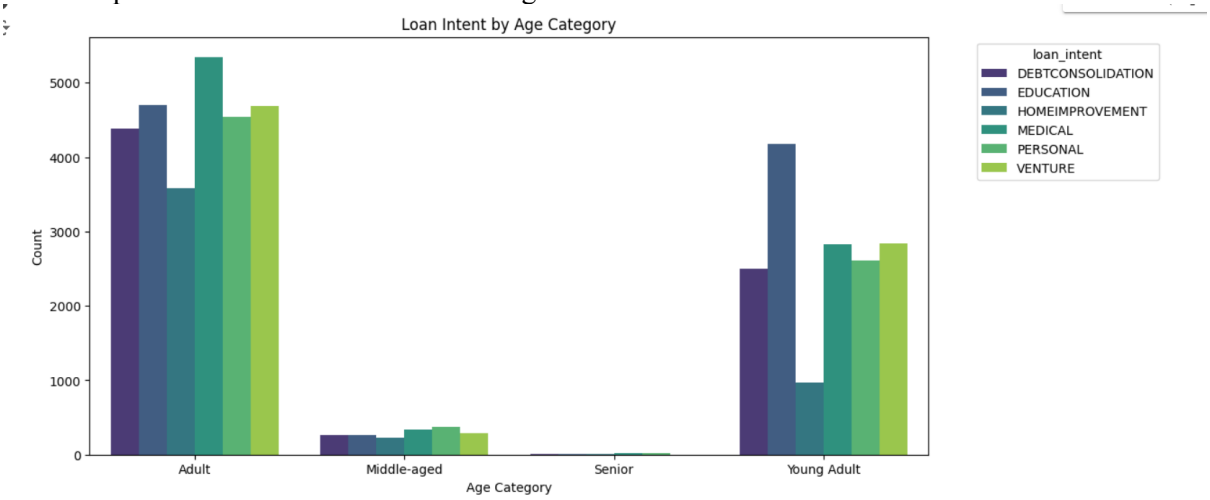
- **Moderate Income Category**: Leads in loan counts for Education (4382) and Medical (4366).Indicates prioritization of education and healthcare expenses.
- **High Income Category**: Significant loans for Education (3286) and Venture (2978).Higher earners invest in education and entrepreneurial ventures.
- **Low Income Category**: Lower counts for all loan intents, with Medical (1702) and Education (1483) being the most common. Reflects limited financial access; loans are primarily for essential needs.
- Medical loans are a common need across all categories, with moderate income having the highest requests.
- Home improvement loans are less frequently requested, especially in the low income category (634), suggesting prioritization of basic needs.

4.Age Category vs. Loan Intent

SELECT AGE_CATEGORY, LOAN_INTENT, COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY AGE_CATEGORY, LOAN_INTENT
ORDER BY COUNT DESC;

RESULT:

- The Adult age category shows the highest count of loan intents, especially for MEDICAL and EDUCATION purposes.
- Both Adults and Young Adults show strong interest in EDUCATION loans, suggesting investment in education to enhance career prospects.
- MEDICAL loans are sought by both Adults and Young Adults, reflecting health-related financial needs across life stages.
- Significant counts for VENTURE and PERSONAL loans among Adults and Young Adults suggest a desire to start businesses or address personal finances.
- Middle-aged and Seniors show much lower counts for loan intents, indicating different financial priorities or conservative borrowing habits.



#5.Loan Status vs. Income Category

SELECT INCOME_CATEGORY, LOAN_STATUS, COUNT(*) AS COUNT
FROM LOAN_DATA

GROUP BY INCOME_CATEGORY, LOAN_STATUS
ORDER BY INCOME_CATEGORY, LOAN_STATUS;

INCOME_CATEGORY	LOAN_STATUS	COUNT
HIGH	0	14488
HIGH	1	1917
LOW	0	3989
LOW	1	3376
MODERATE	0	16516
MODERATE	1	4707

RESULT:

- Rejections dominate across all income categories.
- Moderate-income individuals face the most significant hurdles.
- Low-income applicants show some potential for approval compared to moderate and high-income applicants.

#6.LOAN_INTENT vs. CREDIT_SCORE_CATEGORY

SELECT LOAN_INTENT, CREDIT_SCORE_CATEGORY, COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY LOAN_INTENT, CREDIT_SCORE_CATEGORY
ORDER BY LOAN_INTENT, COUNT DESC;

RESULT:

- The **FAIR** credit score category has the highest counts across all loan intents, suggesting that individuals with fair credit scores are more likely to apply for loans in various categories.
- The **EDUCATION** loan intent has the highest number of applicants (6,367), indicating that education loans are a popular choice, likely reflecting the need for funding for educational expenses.
- Significant drop in applicants with **LOW** and **HIGH** credit scores compared to **FAIR**, suggesting selective borrowing.
- The **DEBTCONSOLIDATION** intent has a substantial number of applicants, especially in the **FAIR** and **LOW** categories, which might indicate that individuals with fair or low credit scores are seeking to consolidate debt to improve their financial standing.
- Lower counts across all credit score categories for **HOMEIMPROVEMENT** loans suggest less demand or stricter approvals.

#7.Loan_Intent vs. Loan_Status

SELECT LOAN_INTENT, LOAN_STATUS, COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY LOAN_INTENT, LOAN_STATUS
ORDER BY LOAN_INTENT, LOAN_STATUS;

RESULT:

- **Productive Loans (Education and Personal)**: These are associated with lower default rates, reflecting improved borrower stability. Lenders may prioritize these loans as safer investments.
- **Essential Expense Loans (Medical and Debt Consolidation)**: Higher risk categories, showing the financial strain borrowers face in meeting urgent needs.

- **Investment Loans (Venture)**: High-risk, indicating the need for lenders to assess business potential closely or consider additional collateral.
- **Overall Trend**: Productive or future-oriented loans like education show the highest repayment success, whereas emergency-related or high-stakes investments, such as medical and venture loans, carry higher default rates.

#8.Gender vs.Loan Approval rate

SELECT PERSON_GENDER,LOAN_STATUS,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY PERSON_GENDER,LOAN_STATUS
ORDER BY PERSON_GENDER;

RESULT:

PERSON_GENDER	LOAN_STATUS	COUNT
female	0	15672
female	1	4485
male	0	19321
male	1	5515

Approval Rate Proportion:By calculating the approval rate (approved loans / total applications for each gender):

- **Female Approval Rate**: 4,485 / (15,672 + 4,485) ≈ 22.2%
- **Male Approval Rate**: 5,515 / (19,321 + 5,515) ≈ 22.2%

```
180 •   SELECT PERSON_GENDER,
181         SUM(CASE WHEN LOAN_STATUS = '1' THEN 1 ELSE 0 END) AS Approved_Loans,
182         SUM(CASE WHEN LOAN_STATUS = '0' THEN 1 ELSE 0 END) AS Denied_Loans,
183         COUNT(*) AS Total_Loans,
184         ROUND(SUM(CASE WHEN LOAN_STATUS = '1' THEN 1 ELSE 0 END) * 100.0 / COUNT(*), 2) AS Approval_Rate_Percentag
185     FROM LOAN_DATA GROUP BY PERSON_GENDER ORDER BY PERSON_GENDER;
186
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

PERSON_GENDER	Approved_Loans	Denied_Loans	Total_Loans	Approval_Rate_Percentage
female	4485	15672	20157	22.25
male	5515	19321	24836	22.21

- Approval rates are consistent across genders, suggesting no gender bias in loan approval.
- Males have a higher number of both approvals and denials, which may indicate higher loan application rates among males.
- Both genders appear to be assessed similarly in terms of loan risk and approval probability.

#9.Loan status and Credit Score

SELECT CREDIT_SCORE_CATEGORY,LOAN_STATUS,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY CREDIT_SCORE_CATEGORY, LOAN_STATUS
ORDER BY LOAN_STATUS, COUNT DESC;

RESULT:

CREDIT_SCORE_CATEGORY	LOAN_STATUS	COUNT
FAIR	0	24309
LOW	0	8562
HIGH	0	2122
FAIR	1	6943
LOW	1	2487
HIGH	1	570

Approval Rate=Total Loans (Approved + Rejected)Approved Loans×100

```
190 •   SELECT CREDIT_SCORE_CATEGORY,SUM(CASE WHEN LOAN_STATUS = 1 THEN 1 ELSE 0 END) * 100.0 / COUNT(*) AS APPROVAL_RATE
191     FROM LOAN_DATA GROUP BY CREDIT_SCORE_CATEGORY;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

CREDIT_SCORE_CATEGORY	APPROVAL_RATE
LOW	22.50882
FAIR	22.21618
HIGH	21.17385

- The approval rate for applicants with a **low credit score** is the highest at **22.51%**, indicating that these applicants have a fair chance of being approved, despite their lower credit status.
- **Fair credit score** applicants have a close approval rate of **22.22%**.
- **High credit score** applicants have the lowest approval rate at **21.17%**, which is somewhat unexpected as high credit scores typically correlate with higher approval likelihood.

#10. Gender and Loan Purpose

SELECT PERSON_GENDER,LOAN_INTENT,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY PERSON_GENDER,LOAN_INTENT
ORDER BY PERSON_GENDER,LOAN_INTENT;

RESULT:

PERSON_GENDER	LOAN_INTENT	COUNT
female	DEBTCONSOLIDATION	3171
female	EDUCATION	4078
female	HOMEIMPROVEMENT	2140
female	MEDICAL	3885
female	PERSONAL	3393
female	VENTURE	3490
male	DEBTCONSOLIDATION	3974
male	EDUCATION	5073
male	HOMEIMPROVEMENT	2643
male	MEDICAL	4663
male	PERSONAL	4158
male	VENTURE	4325

- Males generally have higher counts for almost all loan purposes compared to females, except for education loans, which are very close.
- Higher applications for medical and personal loans among males could reflect greater financial needs in these areas.
- More males are applying for debt consolidation loans compared to females, which may suggest differences in debt management strategies or existing debt levels.

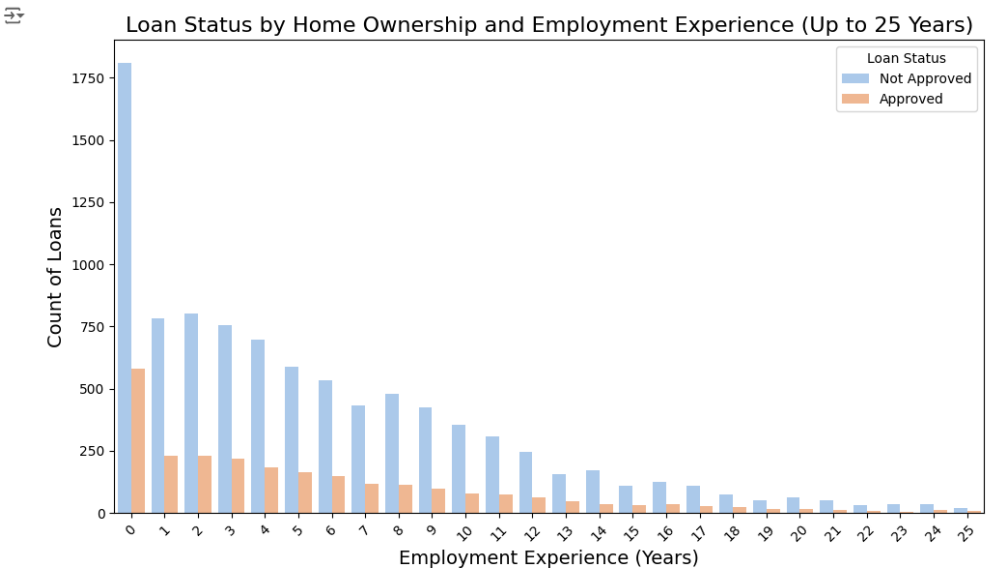
Multivariate Analysis

#1.Loan Status by Home Ownership and Employment Experience:

SELECT LOAN_STATUS,PERSON_HOME_OWNERSHIP,PERSON_EMP_EXP,COUNT(*) AS COUNT
FROM LOAN_DATA
GROUP BY PERSON_HOME_OWNERSHIP,PERSON_EMP_EXP,LOAN_STATUS
ORDER BY LOAN_STATUS,COUNT DESC;

RESULT:

- Loans were more likely to be approved for individuals with low employment experience (0–3 years), regardless of home ownership status.
- Renters with 0 years of employment experience had the highest number of loans rejected, followed by those with mortgages and similar employment levels.
- Both loan approvals and rejections were most common among renters.
- Mortgaged homeowners had a more balanced distribution of approvals and rejections across different employment experience levels compared to renters.



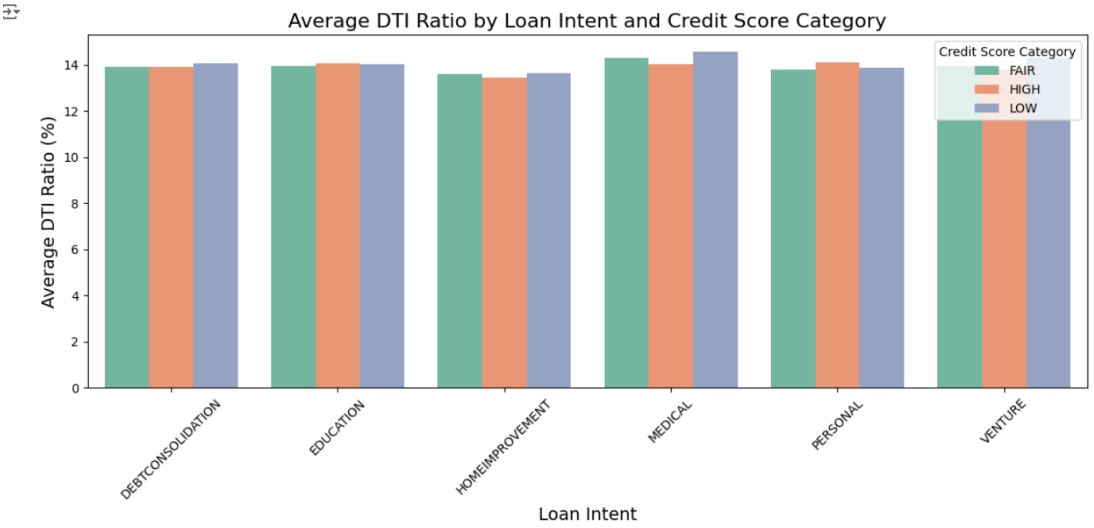
#2.Debt-to-Income (DTI) Ratio Analysis by Loan Intent and Credit Score

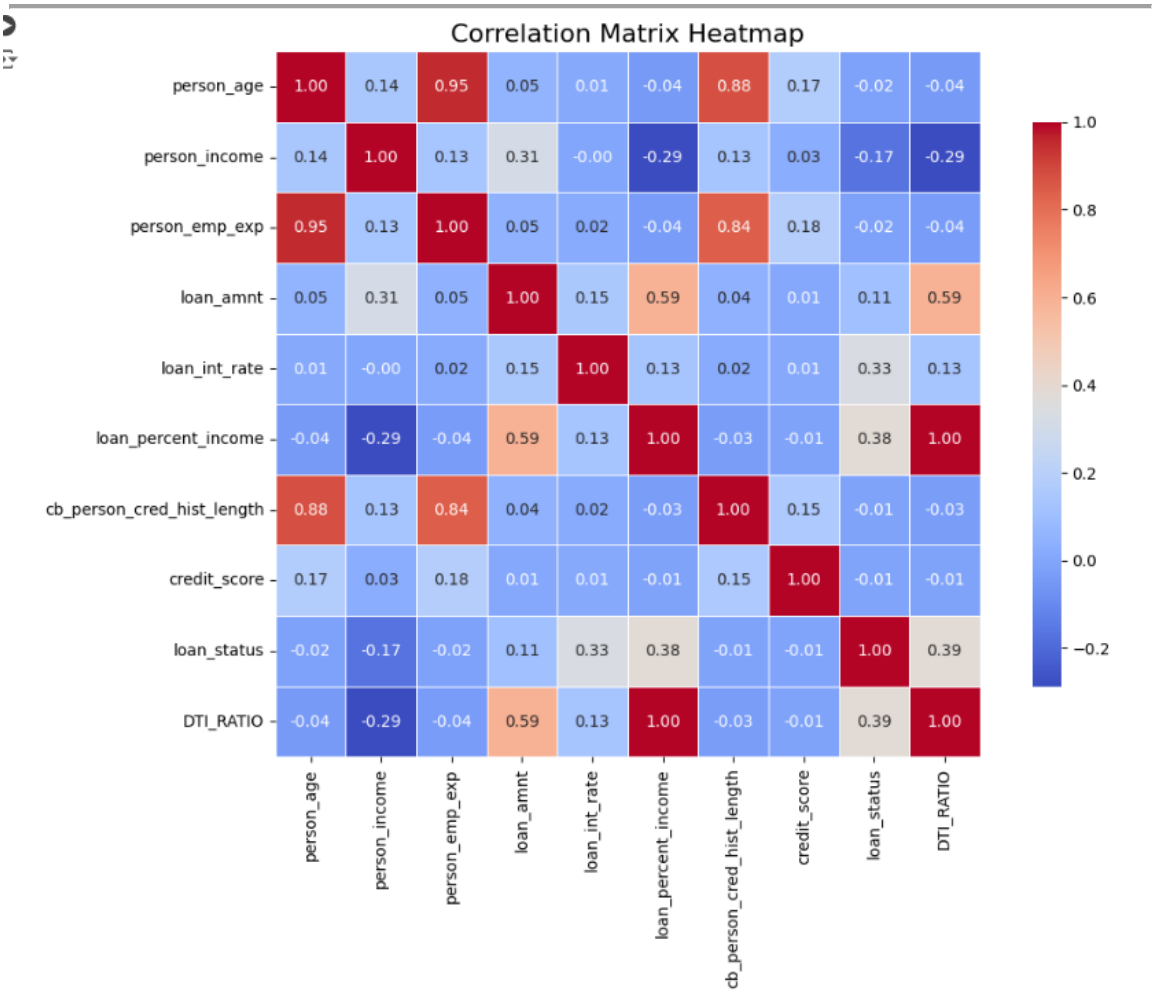
SELECT loan_intent,credit_score_category,AVG(dti_ratio) AS avg_dti_ratio
FROM loan_data
GROUP BY loan_intent, credit_score_category
ORDER BY avg_dti_ratio DESC;

RESULT:

- Medical and venture loans have the highest DTIs for low-credit borrowers, showing significant debt burden.
- Consistently high DTIs across all credit levels, highlighting education as a common, high-debt purpose.
- Borrowers with low credit also take on high DTI in personal and venture loans, reflecting discretionary or business needs.

- Lower DTIs, especially among high-credit borrowers, suggesting more conservative borrowing.
- High DTIs in medical and education loans reflect common, high-debt essentials, especially risky for lower-credit borrowers.





FOR NUMERIC COLUMNS ONLY

Conclusion

The exploratory data analysis of the loan dataset has provided valuable insights into the applicant demographics and their loan characteristics. Key findings include:

1. **Demographic Distribution:** A diverse range of income levels exists among applicants, with most falling within the ₹50,000 to ₹100,000 range. The majority of applicants are young adults, with a notable number seeking loans for education.
2. **Gender Insights:** The dataset shows a slight male predominance among applicants.
3. **Loan Intent and Amounts:** Education remains the primary purpose for loans, with varying amounts requested, suggesting tailored financial products could enhance lending strategies.
4. **Interest Rates:** Average interest rates indicate potential areas for adjustment in loan offerings based on risk assessment.
5. **Debt-to-Income Insights:** Understanding the debt-to-income ratios will aid in evaluating borrower risk profiles and improving loan approval processes.

Overall, these insights can guide financial institutions in optimizing their lending strategies, enhancing customer service, and reducing defaults by focusing on the identified trends and areas for improvement.