📌 https://turquoise-constancy-43.tiiny.site/

Title:

🎯 Project Objective

The objective of this project is to develop a forward-looking, interpretable Behaviour Score model for Bank A that predicts the likelihood of a credit card customer defaulting on their payment in the following month. By leveraging historical behavioral data from over 30,000 customers, the project aims to:

- Build an accurate and robust classification model to identify potential defaulters in advance.
- Enhance the bank's credit risk management by enabling proactive decision-making.
- Provide financial interpretability to understand the drivers of default behavior.
- Support early warning systems and risk-based customer strategies to reduce overall credit exposure and financial loss.

🎯 EDA Objective

- Explore the distribution and interrelationships of customer features in relation to default behavior.
- Uncover key behavioral patterns and early risk indicators that can inform predictive modeling.
- Generate actionable insights to improve credit risk management, support early intervention strategies, and guide risk-based decision-making.
- Lay the groundwork for a model that not only predicts defaults but also provides financial interpretability, helping the bank understand and proactively manage credit exposure.

The dataset contains information on various demographic and financial attributes of credit card holders, including their past payment behavior. Each row represents a unique customer, and the columns are as follows:

Column Name	Description
Customer_Id	Unique identifier for each customer
marriage	Marital status (1 = Single, 2 = Married, 3 = Others)
sex	Gender (1 = Male, 0 = Female)

Column Name	Description
education	Education level (1 = Graduate, 2 = University, 3 = High School, 4 = Others)
LIMIT_BAL	Credit limit assigned to the customer
Age	Age in years
PAY_0 to PAY_6	Payment status for each of the past 6 months
BILL_AMT1 to BILL_AMT6	Total bill amount at the end of each month
PAY_AMT1 to PAY_AMT6	Payment made in each month toward previous month's bill
AVG_Bill_amt	Average bill amount over 6 months
PAY_TO_BILL_ratio	Ratio of total payment to total bill over 6 months
next_month_default	Target variable: 1 if defaulted, 0 otherwise

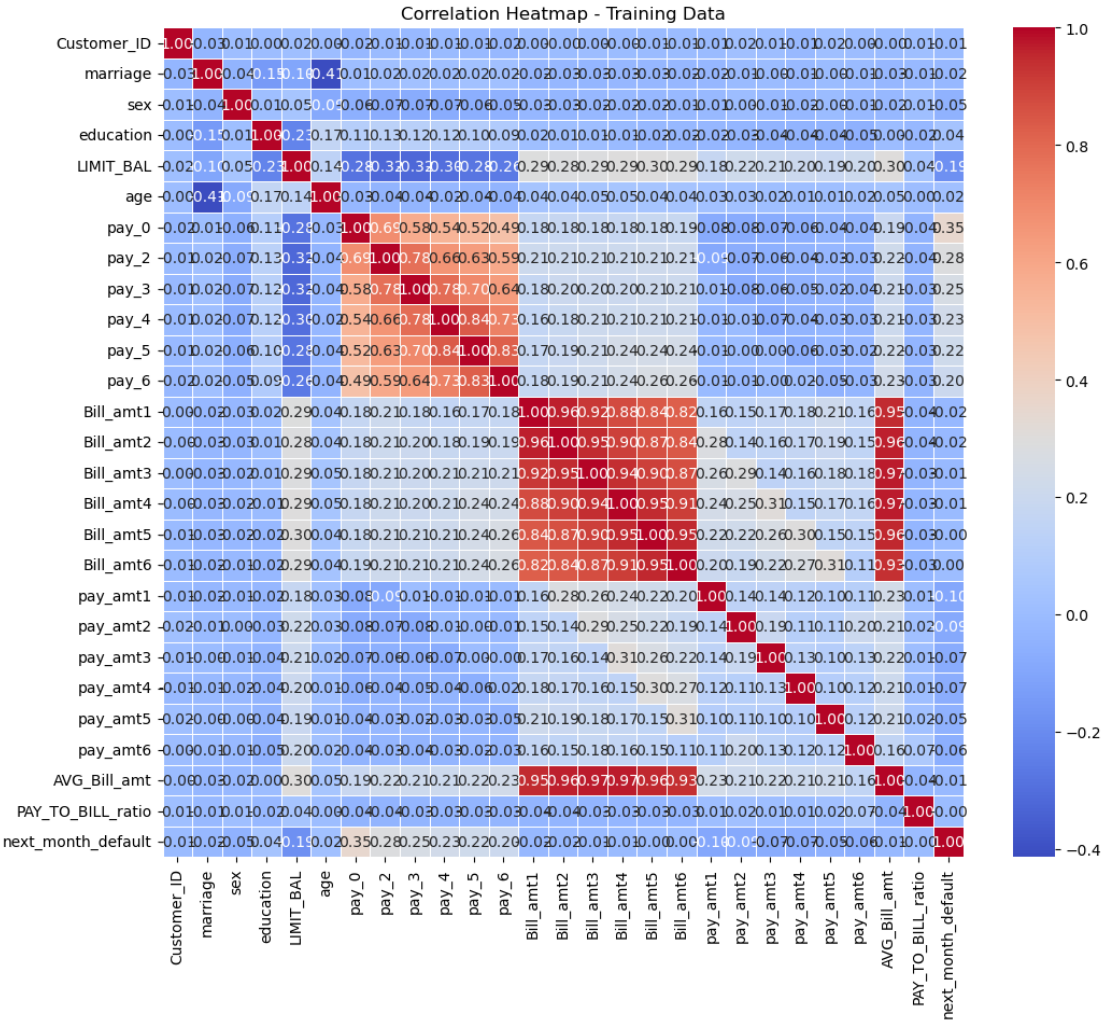
With this understanding, we now proceed to explore each feature in detail.

Training data info

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Customer_ID      25247 non-null  int64
 1   marriage         25247 non-null  int64
 2   sex              25247 non-null  int64
 3   education        25247 non-null  int64
 4   LIMIT_BAL        25247 non-null  int64
 5   age              25121 non-null  float64
 6   pay_0            25247 non-null  int64
 7   pay_2            25247 non-null  int64
 8   pay_3            25247 non-null  int64
 9   pay_4            25247 non-null  int64
 10  pay_5            25247 non-null  int64
 11  pay_6            25247 non-null  int64
 12  Bill_amt1        25247 non-null  float64
 13  Bill_amt2        25247 non-null  float64
 14  Bill_amt3        25247 non-null  float64
 15  Bill_amt4        25247 non-null  float64
 16  Bill_amt5        25247 non-null  float64
 17  Bill_amt6        25247 non-null  float64
 18  pay_amt1         25247 non-null  float64
 19  pay_amt2         25247 non-null  float64
 20  pay_amt3         25247 non-null  float64
 21  pay_amt4         25247 non-null  float64
```

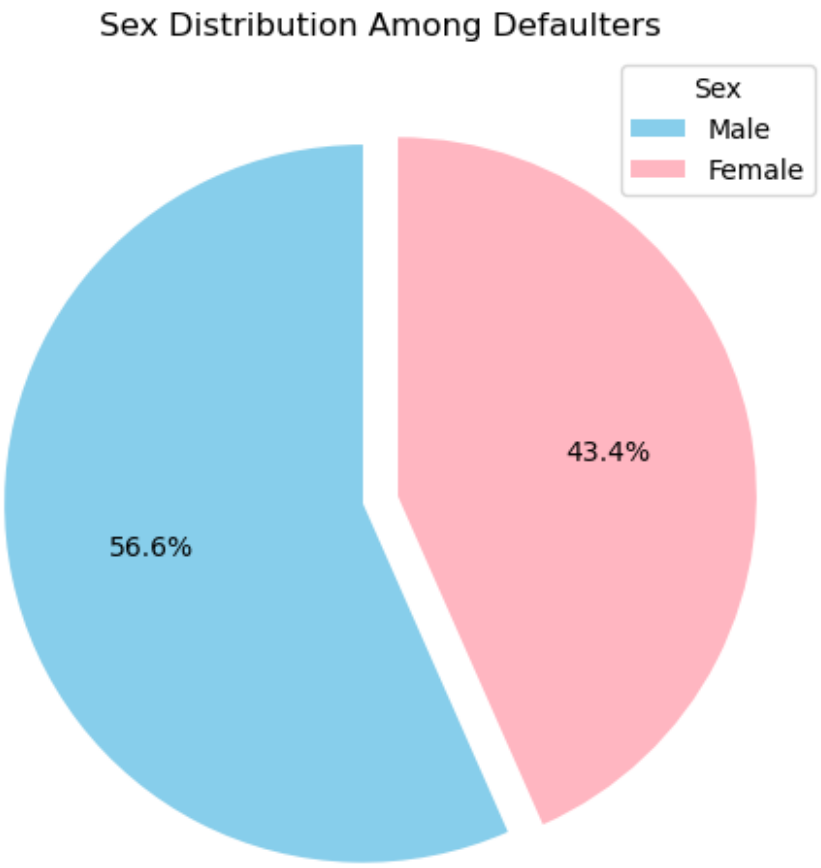
22 pay_amt5 25247 non-null float64
23 pay_amt6 25247 non-null float64
24 AVG_Bill_amt 25247 non-null float64
25 PAY_TO_BILL_ratio 25247 non-null float64
26 next_month_default 25247 non-null int64

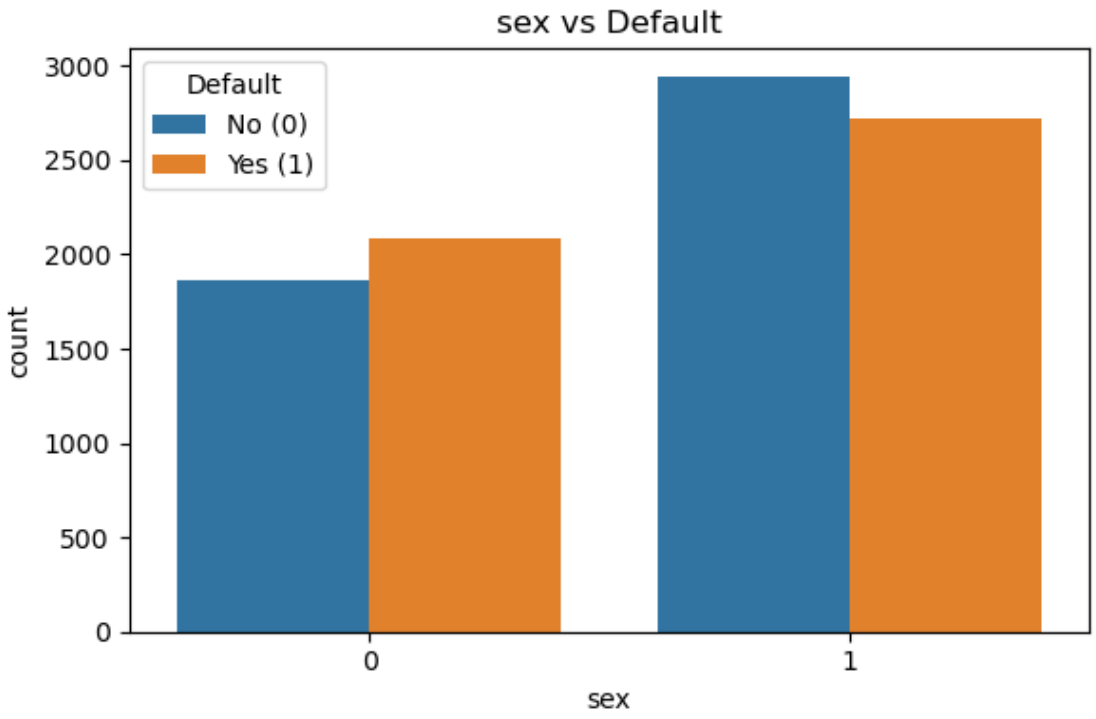
Plotting Correlation Matrix



EDA of categorical cells

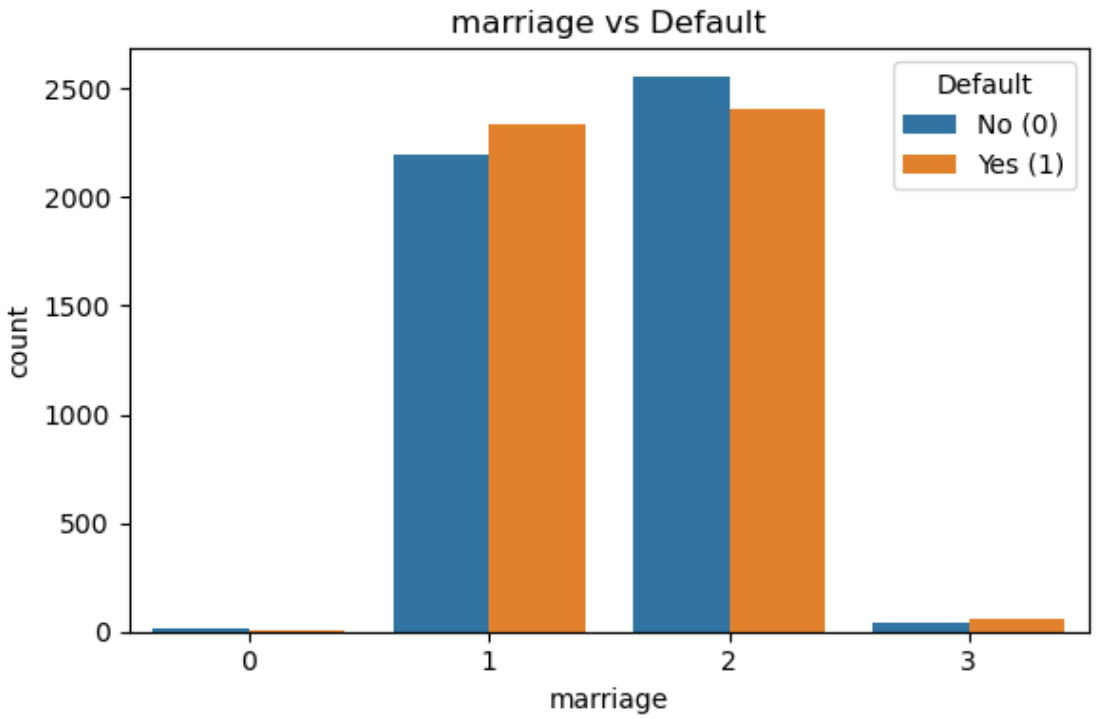
1.on the basis of sex

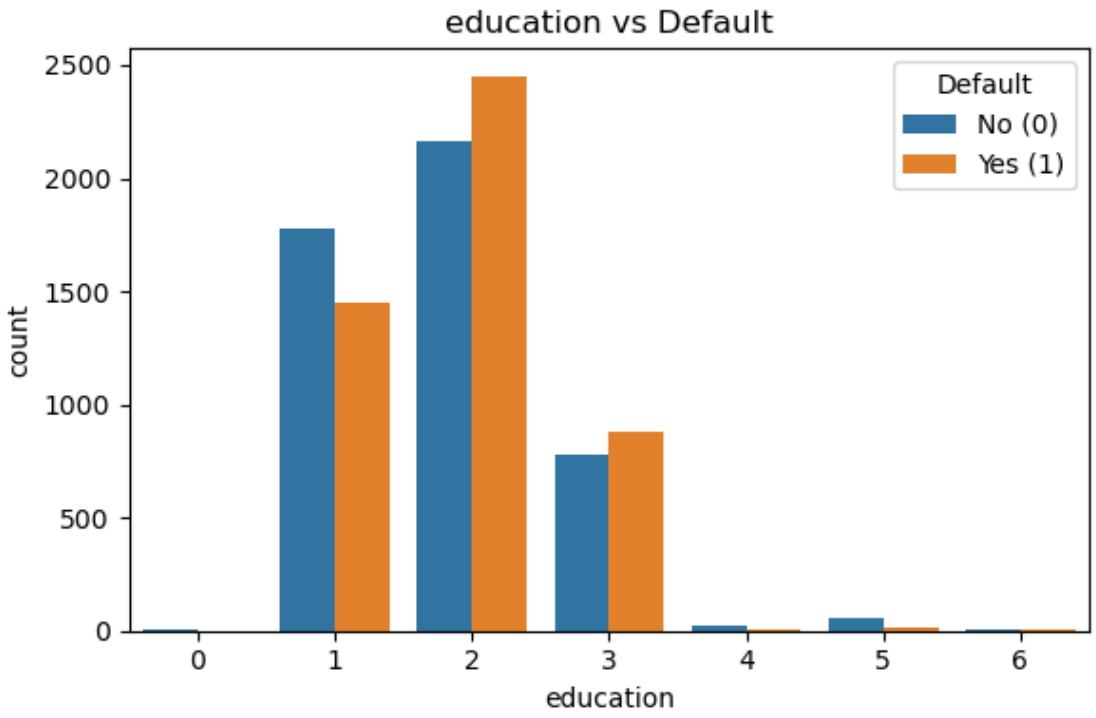


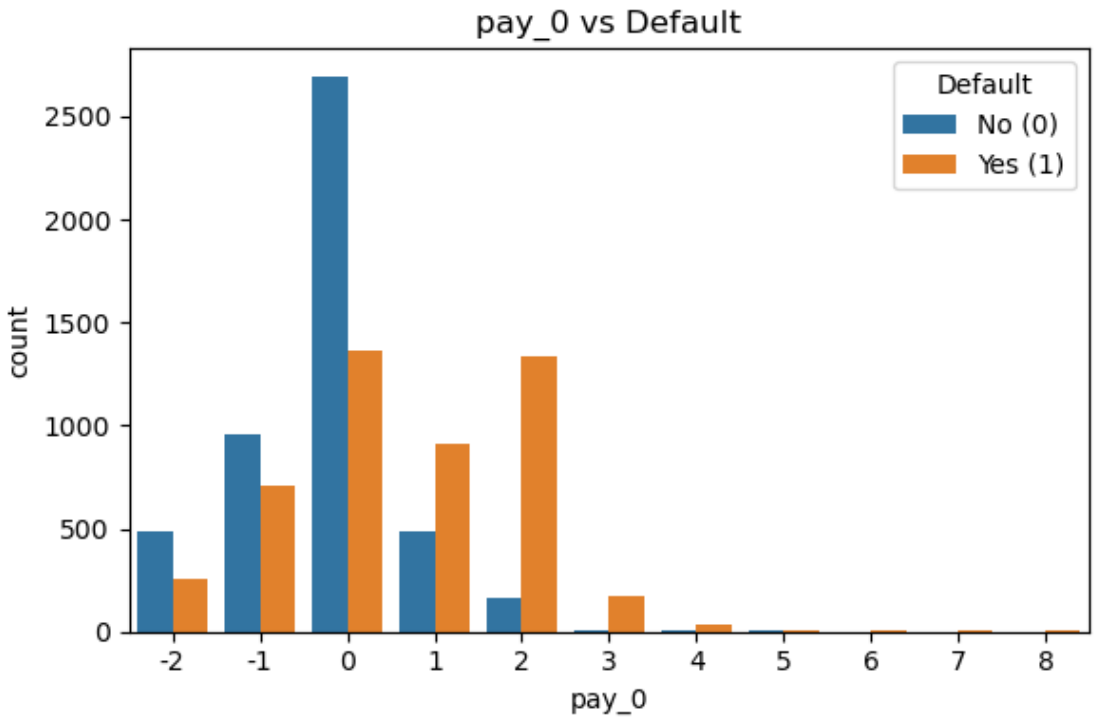


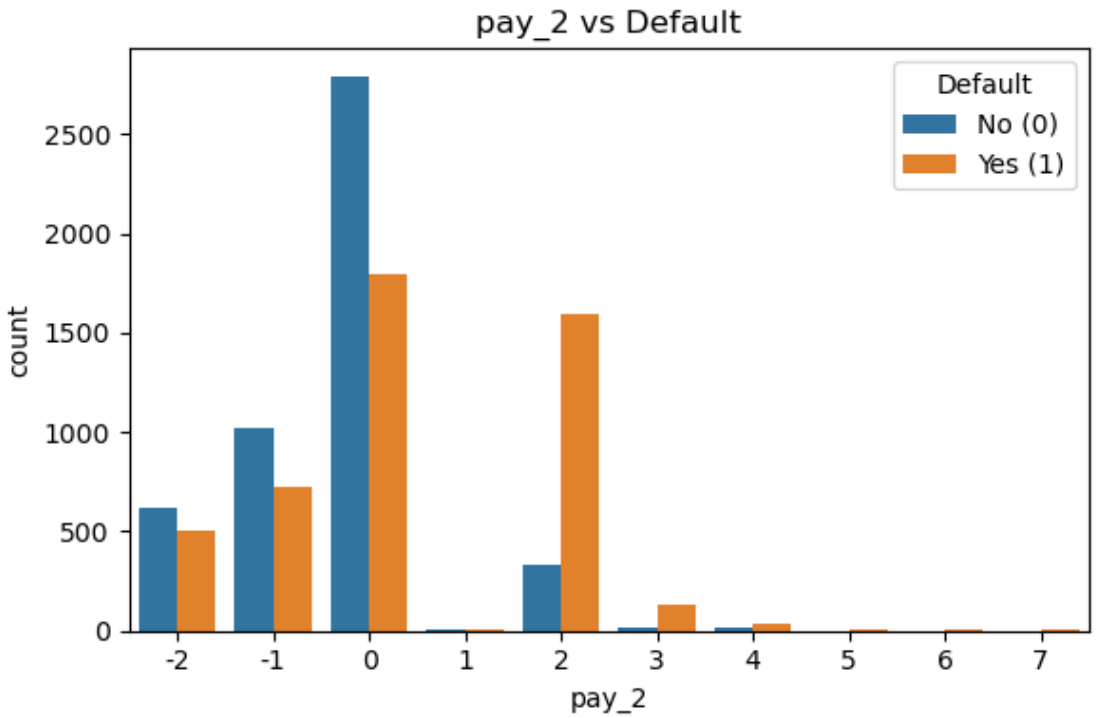
Insight: Gender and Default Behavior

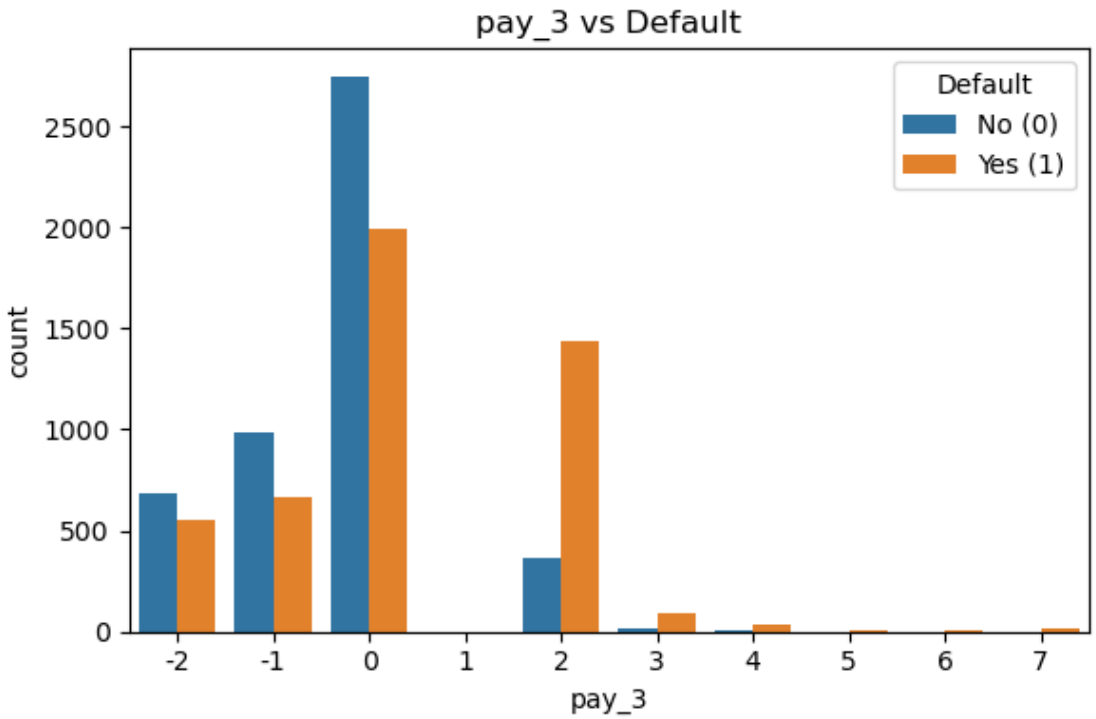
- **Female customers have a higher default rate** compared to male customers.
- However, **male customers make up a larger share of total defaulters** due to their higher representation in the overall customer base.
- The **proportion of males among defaulters** is only slightly lower than their overall proportion in the dataset, indicating that gender alone may not be a strong predictor.
- The **mean credit limit (LIMIT_BAL) among male and female defaulters is nearly the same**, suggesting credit limit is not significantly skewed by gender within defaulters.

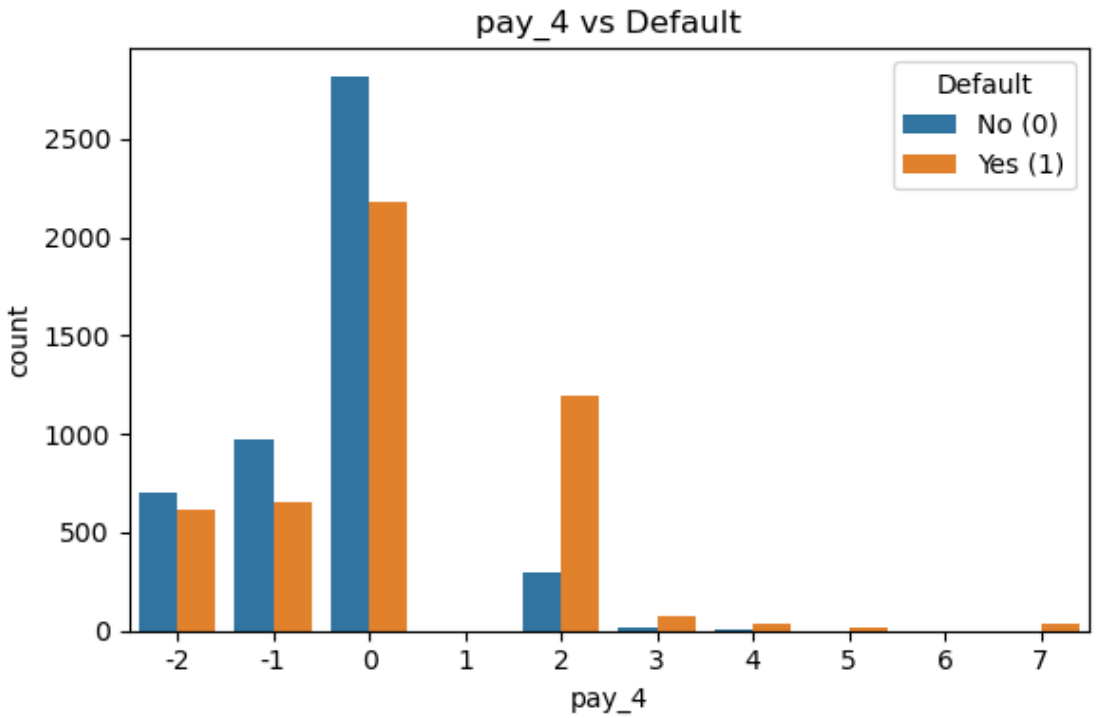

marriage vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
0
1
2
3
marriage

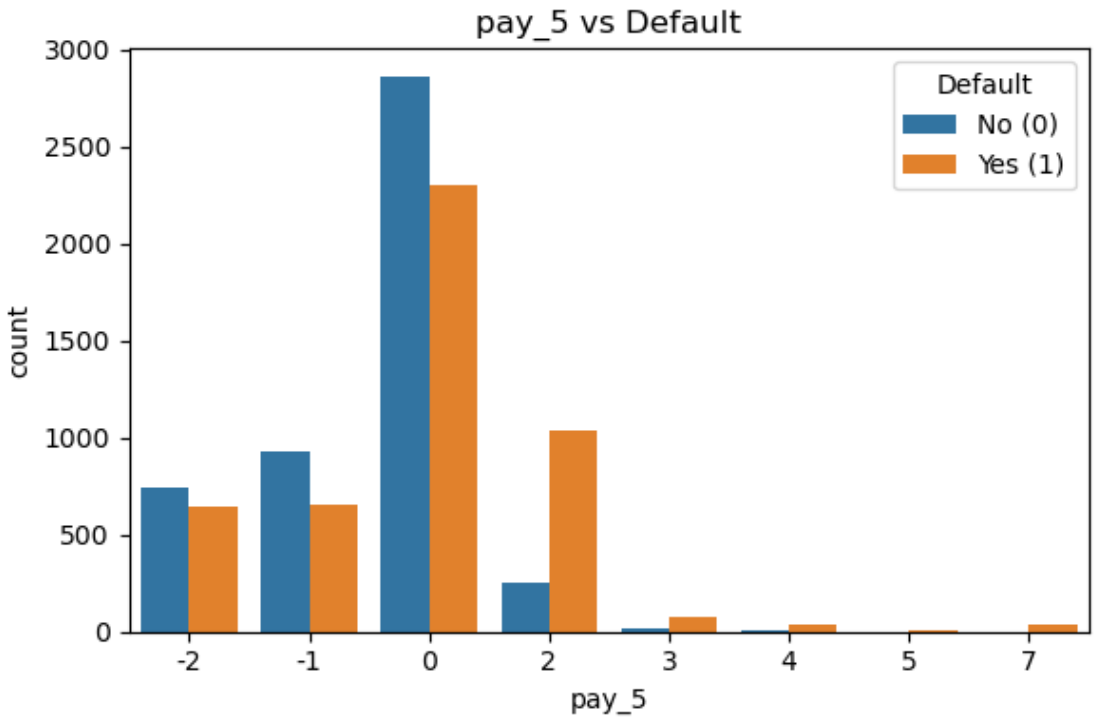

education vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
education

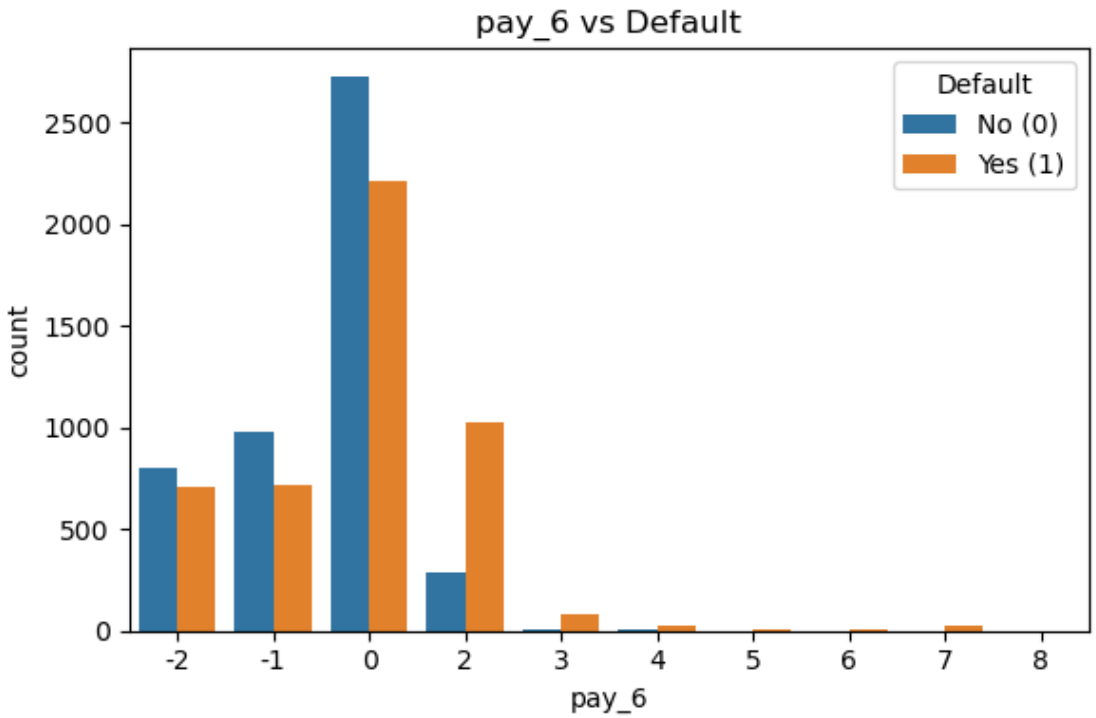

pay_0 vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
-2
-1
0
1
2
3
4
5
6
7
8
pay_0


pay_2 vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
-2
-1
0
1
2
3
4
5
6
7
pay_2


pay_3 vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
-2
-1
0
1
2
3
4
5
6
7
pay_3


pay_4 vs Default
Default
No (0)
Yes (1)
count
2500
2000
1500
1000
500
0
-2
-1
0
1
2
3
4
5
6
7
pay_4





Analysis by Features

1. pay_0 vs Default

- **Distribution**:

  - Most non-defaulters (0) have pay_0 values clustered around -2 to 0.
  - Defaulters (1) show higher frequencies for pay_0 values ≥ 1.
- **Insight**: Late payments (pay_0 ≥ 1) correlate with higher default rates.

---

2. pay_6 vs Default

- **Trend**: Similar to pay_0, but with fewer extreme values (6–8).
- **Key Observation**: Consistent pattern of defaults increasing with positive pay_* values.

---

3. sex vs Default

- **Counts**:
  - Non-defaulters (0): Dominant group (e.g., 12,000 counts).
  - Defaulters (1): Significantly fewer (e.g., 2,000 counts).

---

4. marriage vs Default

- **Categories**: Likely encoded as 0, 1, 2, 3 (e.g., single, married, etc.).

---

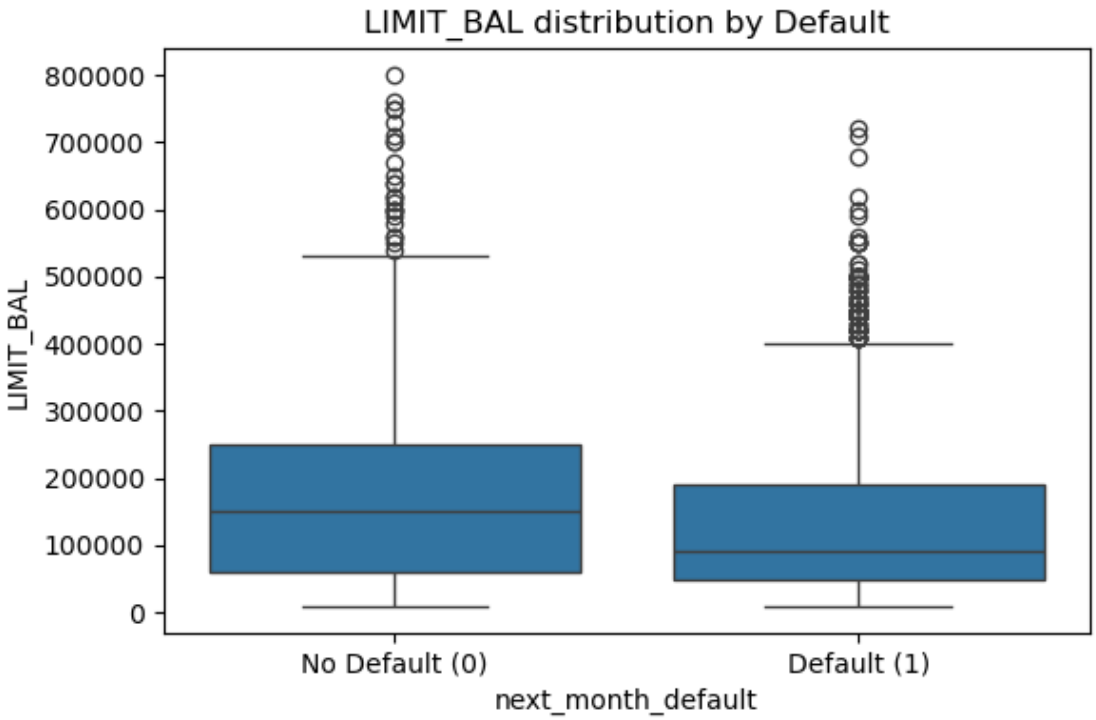
5. education vs Default

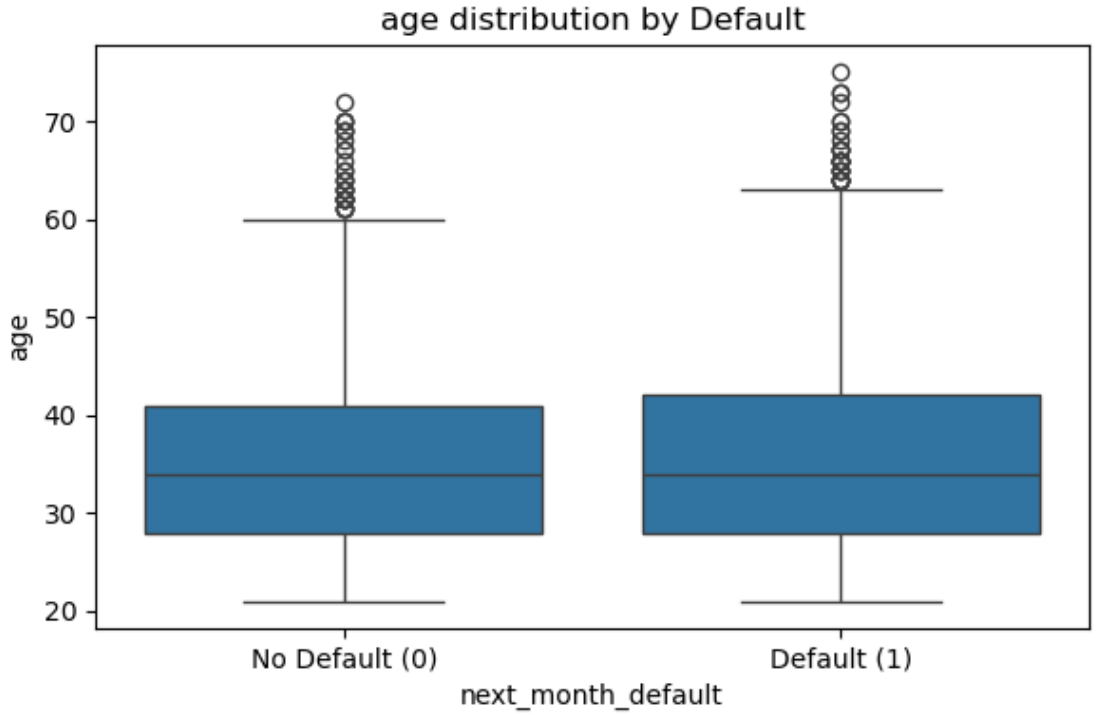
- **Distribution**:
  - Education levels (0–6) show varying default rates.
  - Highest counts in lower education levels (e.g., 0 or 1).
- **Insight**: Lower education may correlate with higher default risk.

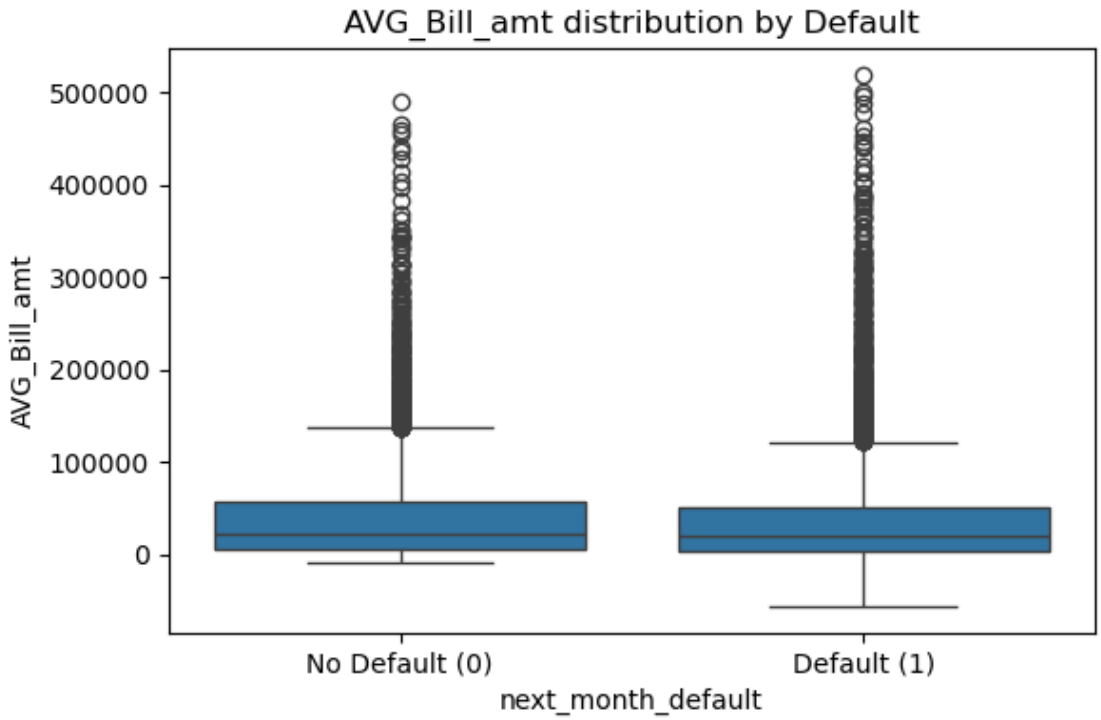
---

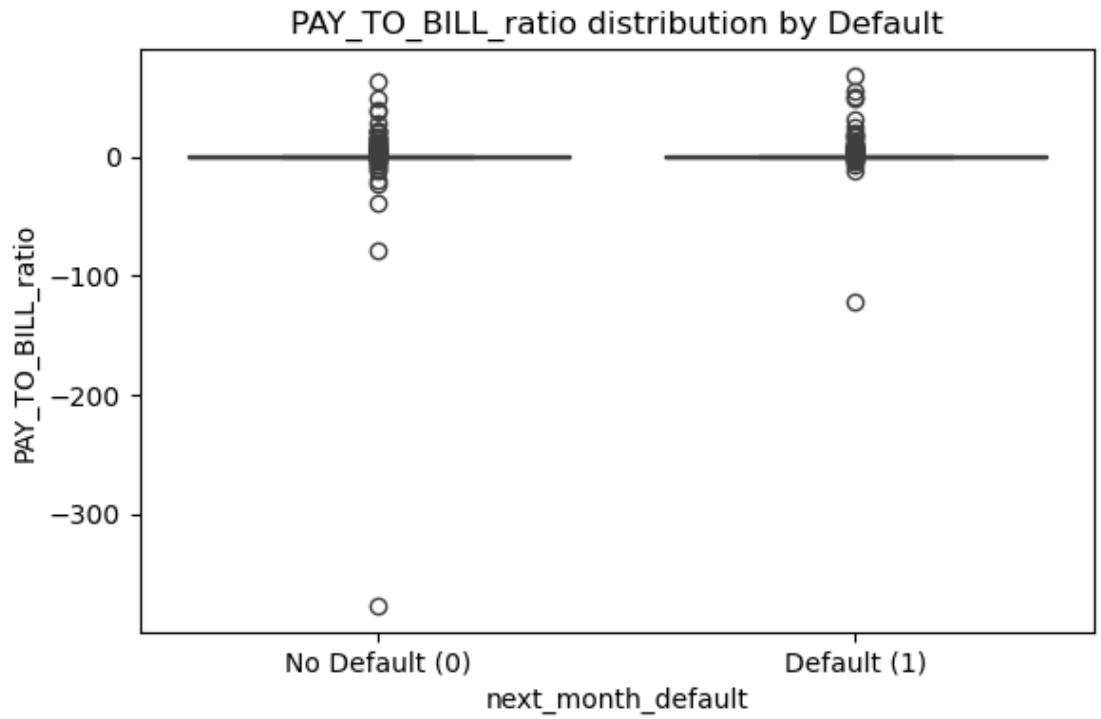
6. pay_2 vs Default

- **Pattern**: Mirrors pay_0 and pay_6 trends.
- **Takeaway**: Repeated late payments (pay_2 ≥ 1) are strong default predictors.

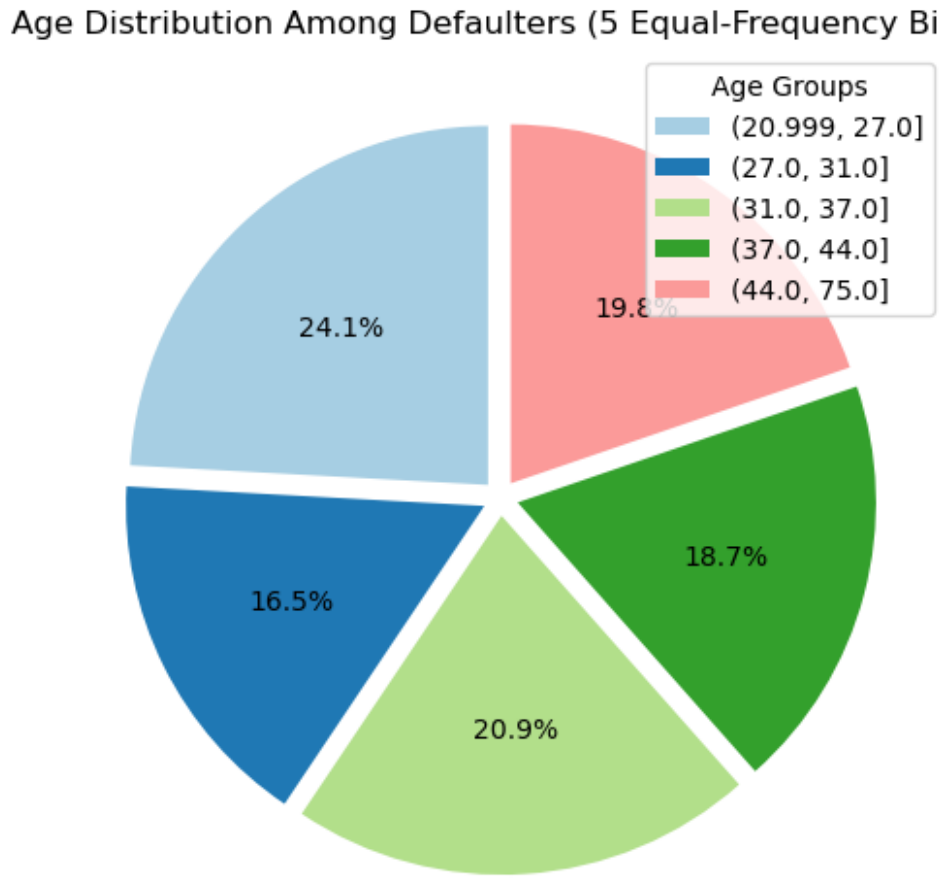

LIMIT_BAL distribution by Default
LIMIT_BAL
800000
700000
600000
500000
400000
300000
200000
100000
0
No Default (0)
Default (1)
next_month_default


age distribution by Default
age
70
60
50
40
30
20
No Default (0)
Default (1)
next_month_default

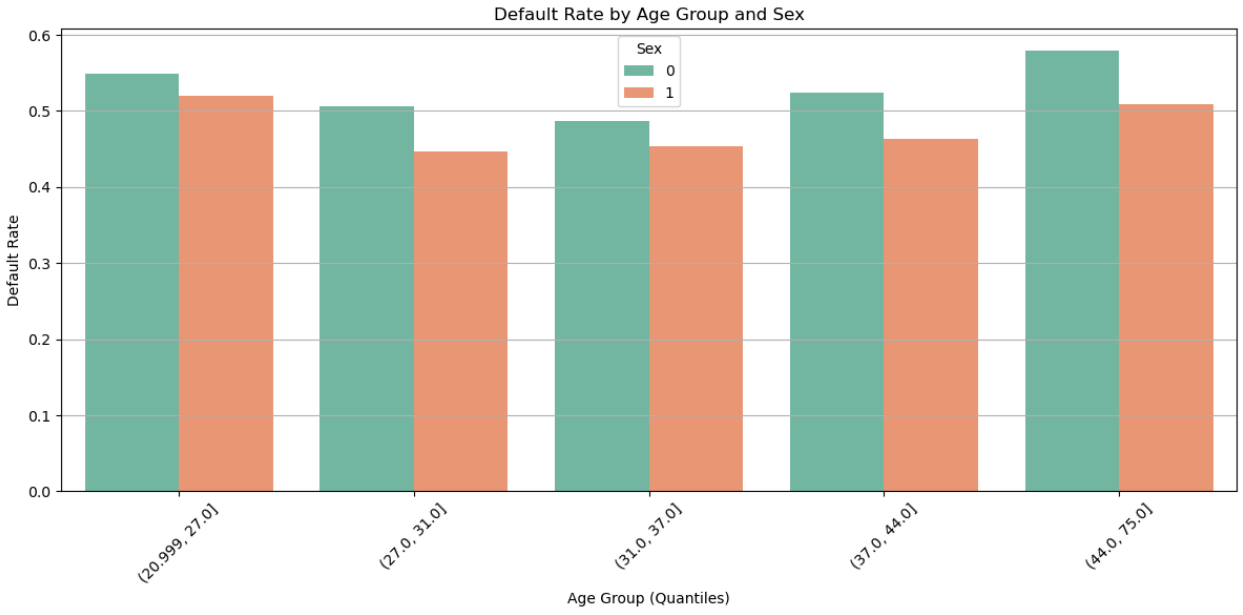




On the basis of Age



taking age and sex at same time



Insight: Marital Status, Age, Gender, and Default Behavior

- **Married individuals have a higher default rate** overall.
- Among defaulters, **single individuals make up the largest proportion**, likely due to their higher population share.
- **Female default rates are higher than male default rates across all marital statuses**.
- Within **marital status = 3 (Others)**, **females show the highest default rate**, but this group has a very small population, so generalizing from it isn't reliable.
- Among males, **married men** have the highest default counts.
- Among females, **single women** are the largest group of defaulters.

Age-Marriage-Gender Interaction:

- In the **43–79 age group with marital status = married**, the number of male and female defaulters is nearly equal.
- In the **27–31 age group with marital status = single**, defaulters are balanced between genders.
- For **marital status = other**, defaulters are evenly distributed across all age groups.
- The **21–27 age group** shows a significant gender gap: married and single **female defaulters outnumber males**, and they also have the **highest default rate** when ignoring the outlier group (marital status = 3).

Limit Balance Insight:

- **Defaulters generally have a lower mean LIMIT_BAL than non-defaulters**.
- However, the **mean LIMIT_BAL among defaulters is fairly consistent across categories**, though influenced by outliers.

📌 Conclusion: While marital status and age reveal several default risk patterns, the **most at-risk group appears to be married females aged 21–27**, excluding small-population anomalies.


Education: 0 | Sex: 0
Education: 0 | Sex: 1
Education: 1 | Sex: 0
Education: 1 | Sex: 1
Education: 2 | Sex: 0
Education: 2 | Sex: 1
Education: 3 | Sex: 0
Education: 3 | Sex: 1
Education: 4 | Sex: 0
Education: 4 | Sex: 1
Education: 5 | Sex: 0
Education: 5 | Sex: 1
Education: 6 | Sex: 0
Education: 6 | Sex: 1
Default Rate (%)
Age Group (Quantiles)
Marriage Status
0
1
2
3

Insight: Education, Age, Gender, and Default Behavior

Overall Education-Level Trends:

- **Highest default rate** is observed in the **High School** education group, followed closely by **University**.
- **University-educated customers make up the largest share of defaulters**, due to their high representation in the dataset.
- Across all education levels, the **female default rate is higher than male**, consistent with earlier findings.

Gender Distribution Across Education & Age:

- In most combinations of education and age groups, **male defaulters outnumber female defaulters**, except in three specific cases:
    a. **High School**, Age group: **27–31**
    a. **Graduate School**, Age group: **37–43**
    b. **Graduate School**, Age group: **43–79**
- The **female default rate is higher across almost all education-age combinations**, except in three:
    a. **Graduate School**, Age group: **21–27**
    a. **Education: Other**, Age group: **21–27**
    b. **Education: Other**, Age group: **37–43**

Highest Risk Subgroups:

- **Female defaulters with High School education, married, aged 21–27** show the **highest default rate** (excluding outliers from small-population categories).
- Among males, **those with High School education, aged 37–43, single** have the **highest default rate**.
- In terms of absolute numbers:
    - **Male, University-educated, aged 21–27, single** have the **highest number of defaulters**.
    - **Female, University-educated, aged 21–27, single** have the **highest number of defaulters among women**.

📌 These findings can help in designing targeted risk-based strategies based on education level, age, gender, and marital status combinations.

MODEL EVALUATION:

model	roc_auc	f1	F2	precision	recall
RandomForest	0.777513	0.689903	0.660706	0.744920	0.642604
GradientBoosting	0.755534	0.677079	0.658035	0.711446	0.645933
LogisticRegression	0.764480	0.683723	0.649704	0.749305	0.628876
KNN	0.713797	0.659598	0.650436	0.675510	0.644480
SVM	0.751960	0.686509	0.661949	0.731927	0.646558

Among all the tried models i would prefer to go with " SVM" due to:

- -Among all the best F2 score it has second highest F2 score
- -Also along with second highest F2 score it has a good F1 score and a comparable accuracy which is important for a good model.
- -It falls under all criteria of Accuracy metrices and F2 score

Train Performance:

Accuracy class 0: 0.7631
Accuracy class 1: 0.6466
F1 Score class 0: 0.7211
F1 Score class 1: 0.6865

Presented By:-

Abhishek Yadav(22113007)