CREDIT CARD DEFAULTER EDA ANALYSIS

Business Understanding

- The Goal of this analysis is to explore and understand the factors influencing credit card default payments. By analyzing demographic and financial variables, we aim to uncover insights, relationships, and trends that can improve model performance in predicting defaults.

SECTION 1

DATA UNDERSTANDING
DATA CLEANING
DATA PROCESSING

Data Understanding

Dataset Source: Default of Credit Card Clients

Rows: 30,000
The data

Key Variables:

- Demographic: SEX, EDUCATION, MARRIAGE, AGE
- Financial: LIMIT_BAL, BILL_AMT1–6, PAY_AMT1–6
- Payment History: PAY_1–PAY_6
- Target: default payment next month

VARIABLE DESCRIPTION

- LIMIT_BAL
- Amount of credit assigned to the client (in NT dollars)
- SEX
- Gender of the client (1 = Male, 2 = Female)
- EDUCATION
- Level of education (1 = Graduate School, 2 = University, 3 = High School, 4 = Others)
- MARRIAGE
- Marital status (1 = Married, 2 = Single, 3 = Others)
- AGE
- Age of the client in years
- PAY_1 – PAY_6
- Repayment status for the last six months

VARIABLE DESCRIPTION

• BILL_AMT1 – BILL_AMT6	• Amount of bill statement for the last six months
• PAY_AMT1 – PAY_AMT6	• Amount of payment made in the last six months
• default payment next month	• Target variable (1 = Default, 0 = Non-default)

DATA CLEANING

- **Column Renaming:**

  All columns were renamed with meaningful titles to improve readability (e.g., PAY_0 → PAY_1, default.payment.next.month → default_payment_next_month).

- **Validation of Categorical Data:**

  Checked and corrected invalid entries in categorical columns — invalid codes in EDUCATION (0, 5, 6) and MARRIAGE (0) were replaced with “Others.”

- **Missing & Duplicate Values:**

  Verified dataset integrity — no missing or duplicate records found.

- **Outlier Detection & Treatment:**

  Identified extreme values in financial columns (LIMIT_BAL, BILL_AMT, PAY_AMT). Instead of removal, we understood the contex of those outliers and capped them at upper percentiles to retain valid but extreme financial behavior.

DATA CLEANING

- **Descriptive Mapping**:
  Temporary readable mappings (e.g., “Male”, “University”, “Married”) added for clearer visualizations.

- **Scaling & Normalization**:
Continuous variables were scaled to reduce large numeric disparities, ensuring balanced comparisons and ML readiness.

- **Final Verification:**
  Confirmed valid data types, consistent ranges, and accurate feature distributions. Dataset was fully clean, structured, and ready for analysis and modeling.

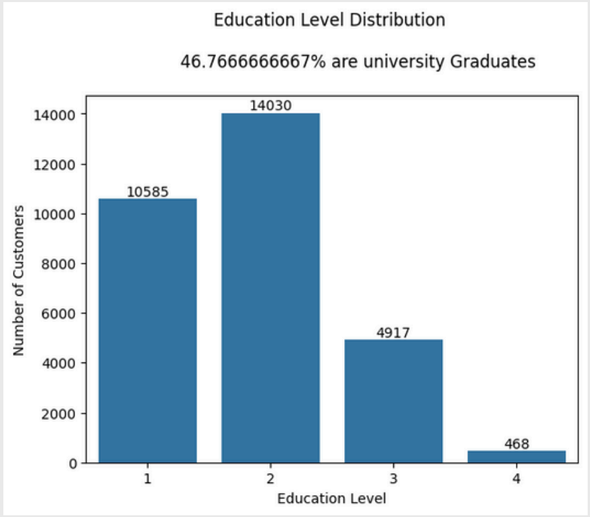
SECTION 2

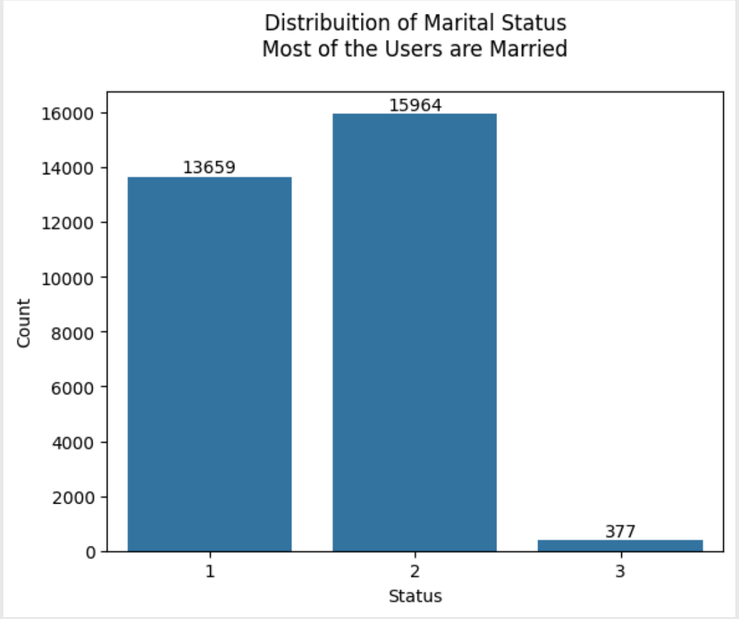
CATEGORICAL ANALYSIS
NUMERICAL ANALYSIS
TARGET VARIABLE UNIVARIATE ANALYSIS

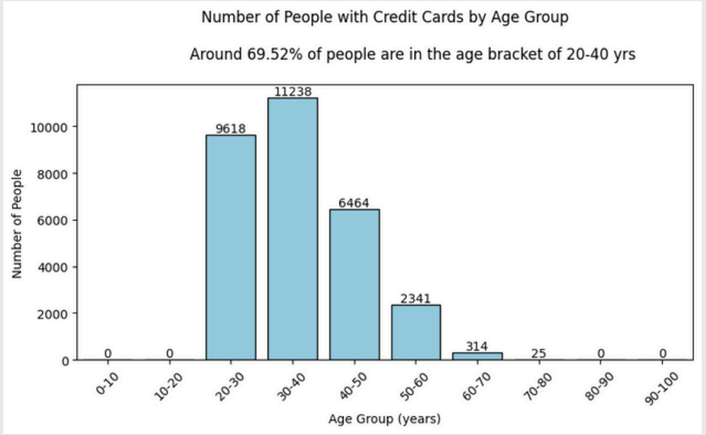
CATEGORICAL ANALYSIS

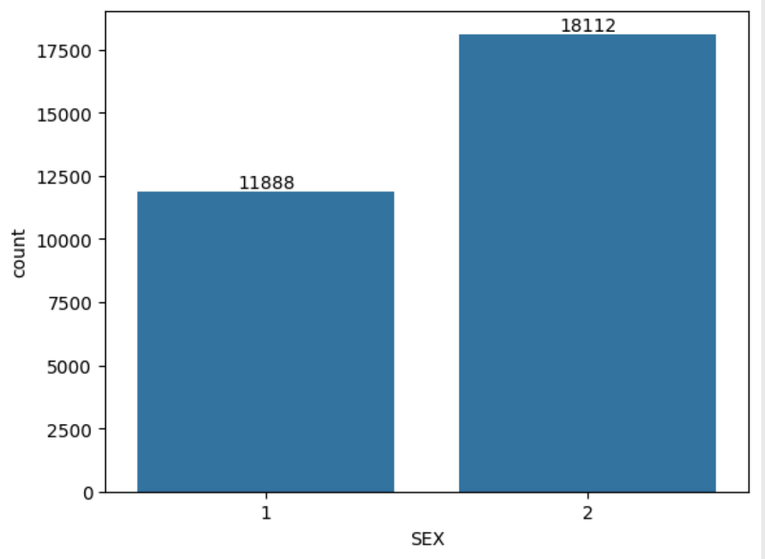
UNIVARIATE

- By just analysing single variables, we won't find much insights related to the defaulters, as here we will just have an idea which category of people are present in abundance
- Having said that we do need to look at them for context that may help us with bivariate analysis
- The number of females are slightly more than males
- Most users are married
- Education level of college is the most common and around >50% of the users are in 20-40 yrs of age







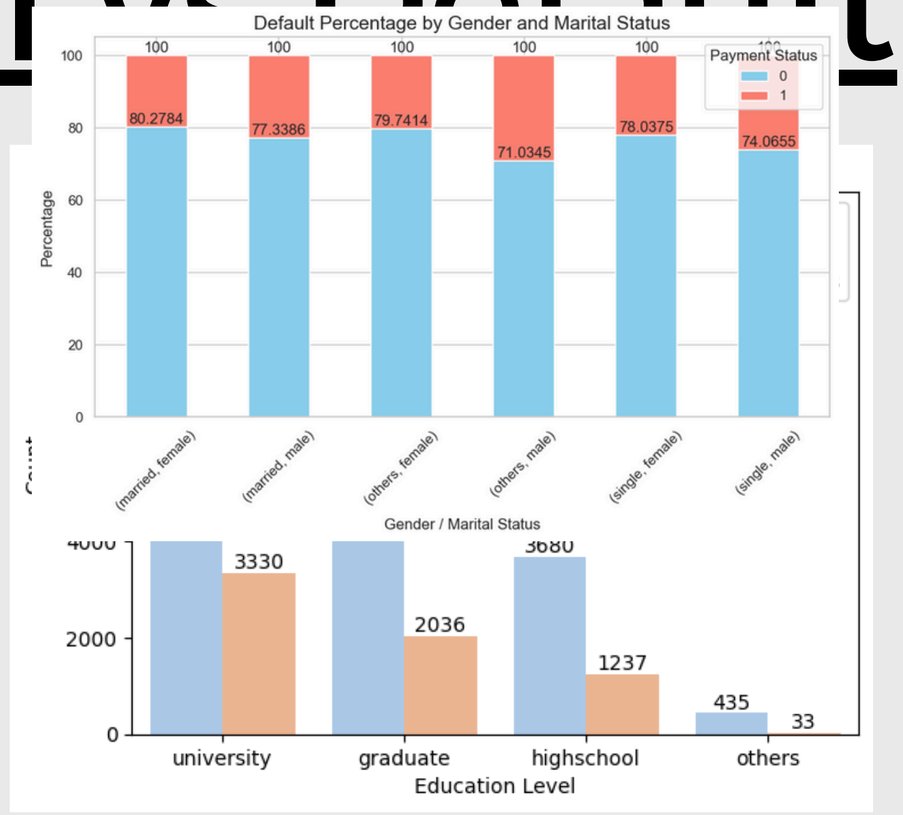


Education level vs Default

- UNIVERSITY IS THE MOST COMMON LEVEL OF EDUCATION PRESENT AMONG THE CARD HOLDERS

- DEFAULT RATE OF HIGH SCHOOLERS IS THE HIGHEST ¼ HIGHSCHOOLERS TEND TO DEFAULT, UNIVERSITY HAS THE 2ND MOST NUMBER OF DEFAULTERS AS WELL

- THIS CAN BE DUE TO VARIOUS REASONS SUCH AS LACK OF AWARENESS AT HIGHSCHOOL LEVEL AND INCREASED LEVEL OF TUITION FEES IN UNIVERSITY LEVELS

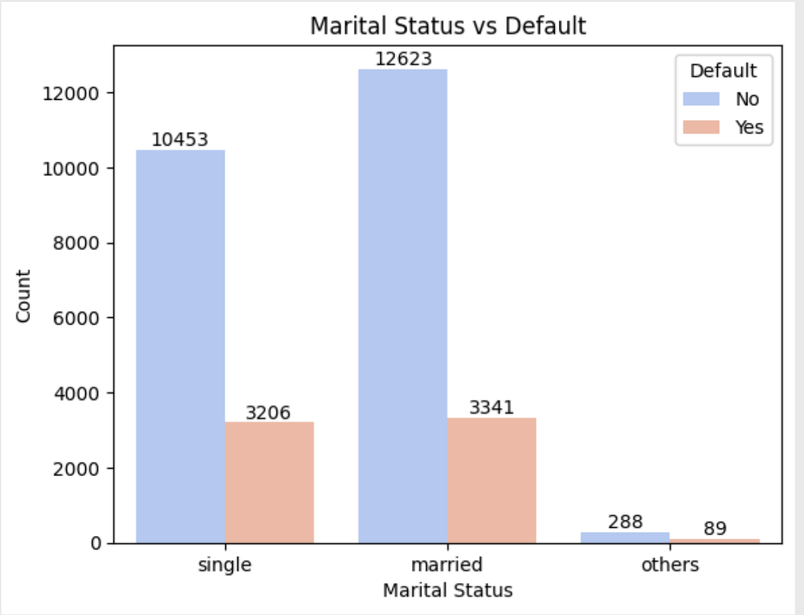


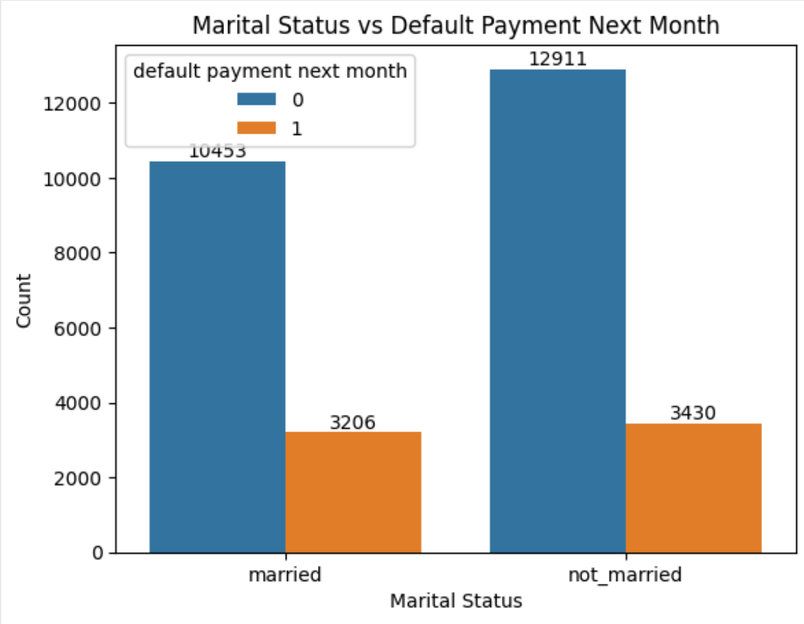
default payment next month	0	1
education_temp		
graduate	80.765234	19.234766
highschool	74.842384	25.157616
others	92.948718	7.051282
university	76.265146	23.734854

Marital status vs Default

- At first glance it seems that married people have least default rate but when we combine both single and other and make columns like married and not married we see the data is equally divided

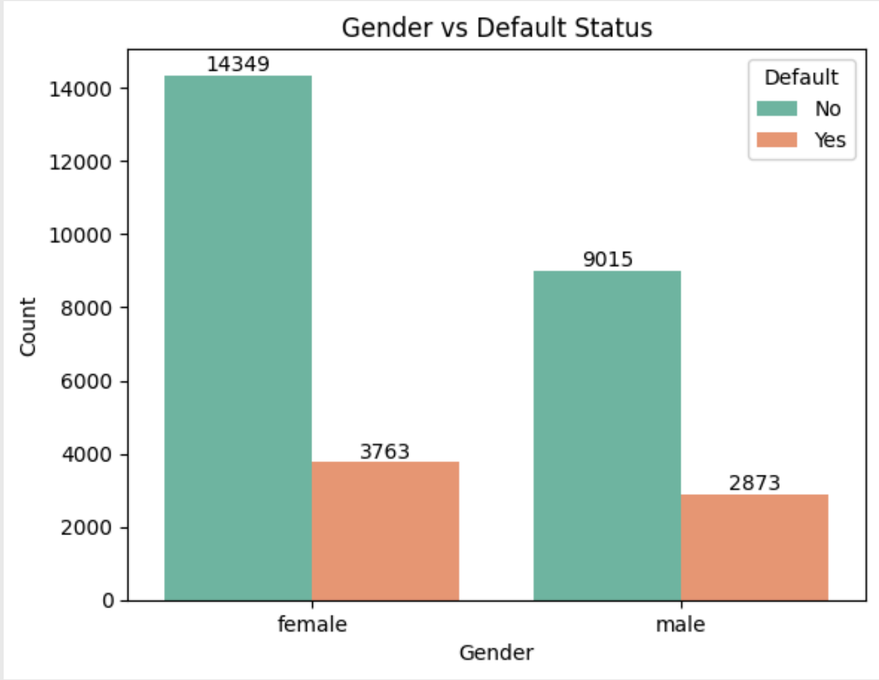
default payment next month	0	1
marriage_temp		
married	79.071661	20.928339
others	76.392573	23.607427
single	76.528296	23.471704





Gender Vs Default

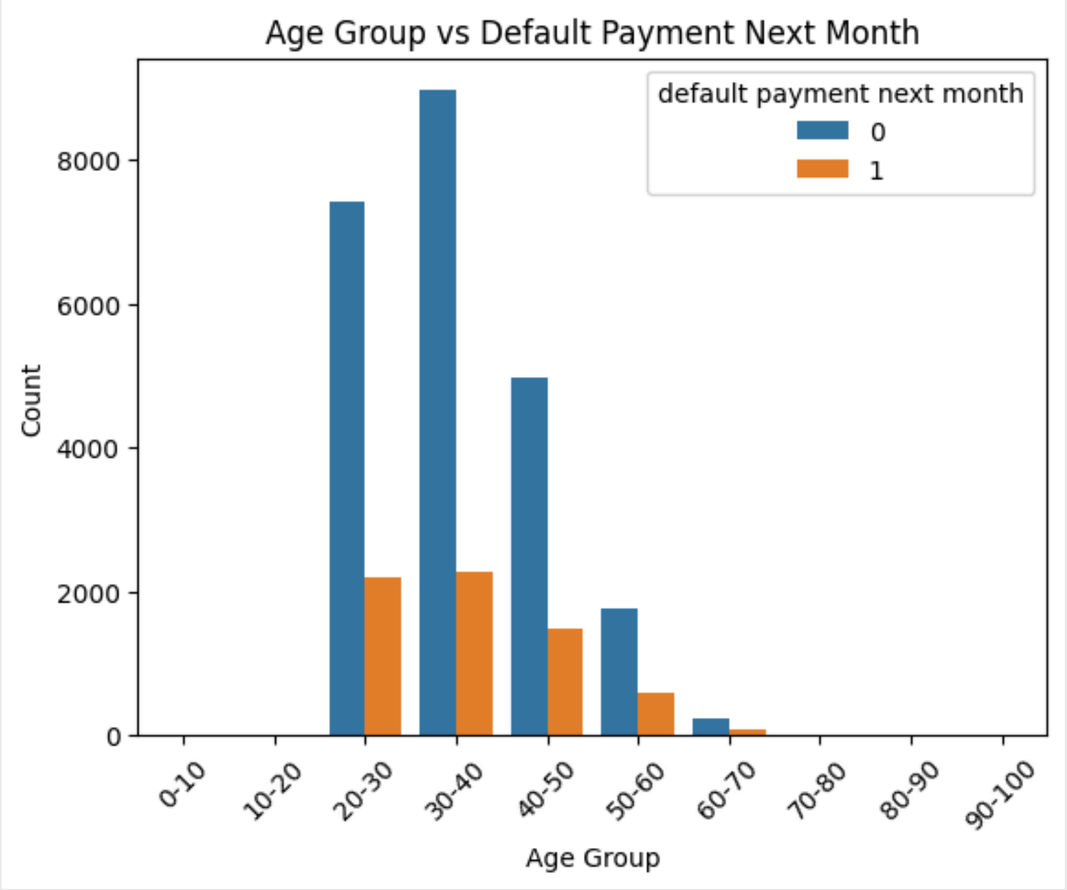
- While the number of females in more compared to the males ,
- The number of defaulters in male is much higher nearly 1 in 4 men are at risk of default
- reasons like males having multiple credit cards and onlythem beibgearning member of the family can be a good cause for this phenomenon



default payment next month	0	1
sex_temp		
female	79.223719	20.776281
male	75.832773	24.167227

AGE Vs Default

- Middle and yound adults (20-40) are the safest with minimal risk of defaulters
- senior citizens and retired professionals have higher risk
- reasons for this can be no or low income after retirement,bad retirement planning and even people forgetting to pay

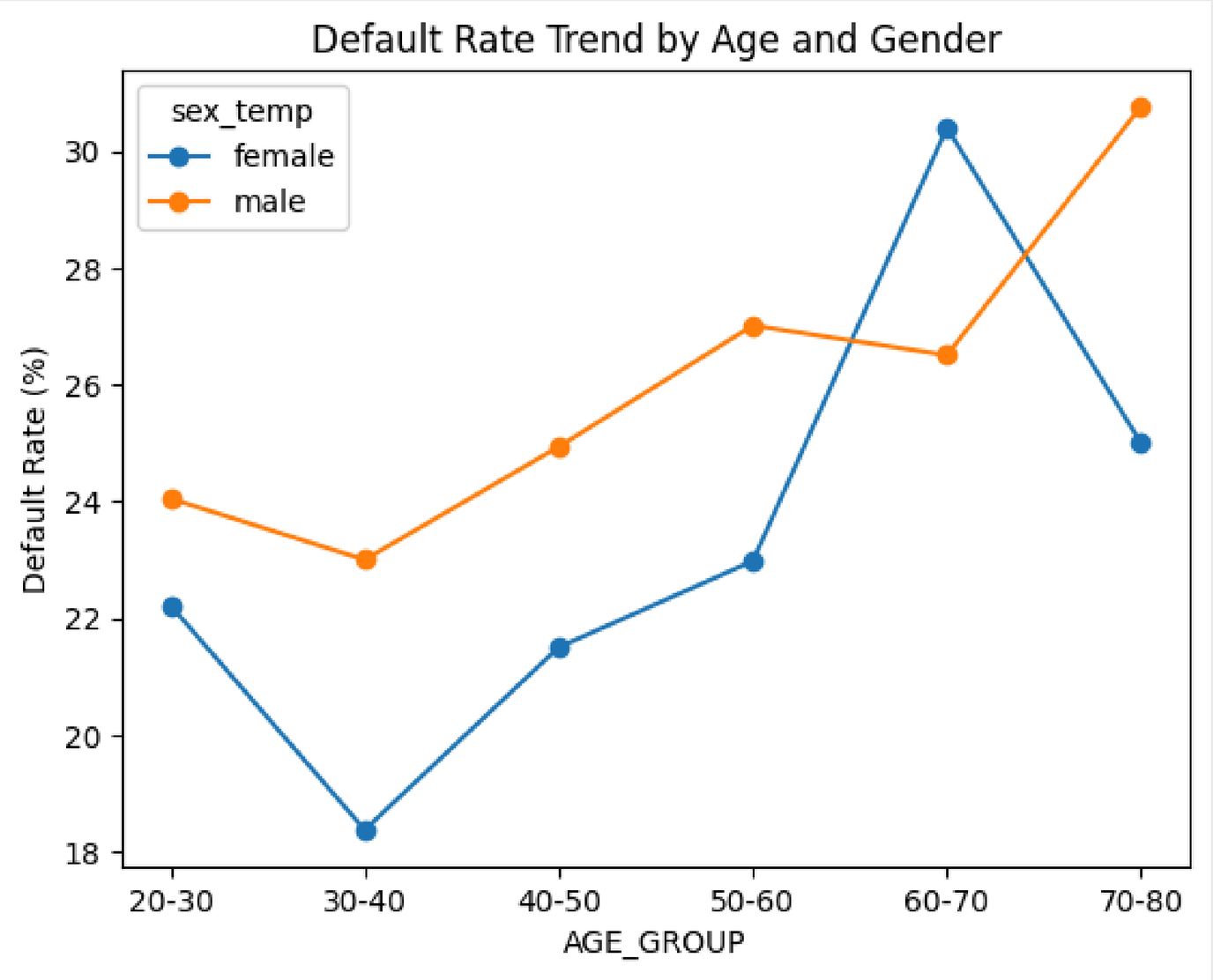


default payment next month	0	1
AGE_GROUP		
20-30	77.157413	22.842587
30-40	79.747286	20.252714
40-50	77.026609	22.973391
50-60	75.138830	24.861170
60-70	71.656051	28.343949
70-80	72.000000	28.000000

MULTIPLE VARIBALE ANALYSIS

AGE VS GENDER VS DEFAULT RATE

- As stated earlier senior citizens do tend to deafault a lot more
- 60-70 ages in female and 70-80 ages in male tend to notice a huge spike in %age of defaulters
- also for women default %age decreases after 70



LIMIT BALANCE

Most of the poeple have credit card limit of 200k with the average being at 167k

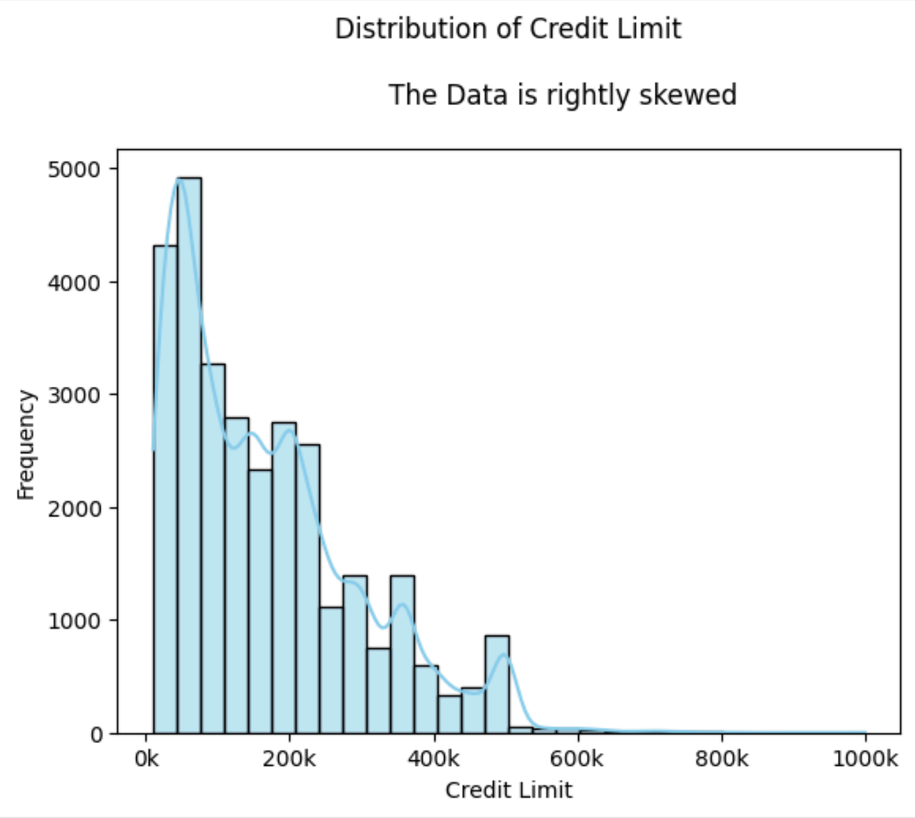
```
df['LIMIT_BAL'].describe()
```

```
count      30000.000000
mean      167484.322667
std       129747.661567
min        10000.000000
25%        50000.000000
50%       140000.000000
75%       240000.000000
max      1000000.000000
Name: LIMIT_BAL, dtype: float64
```

people with balance under 50k has the highest % of default

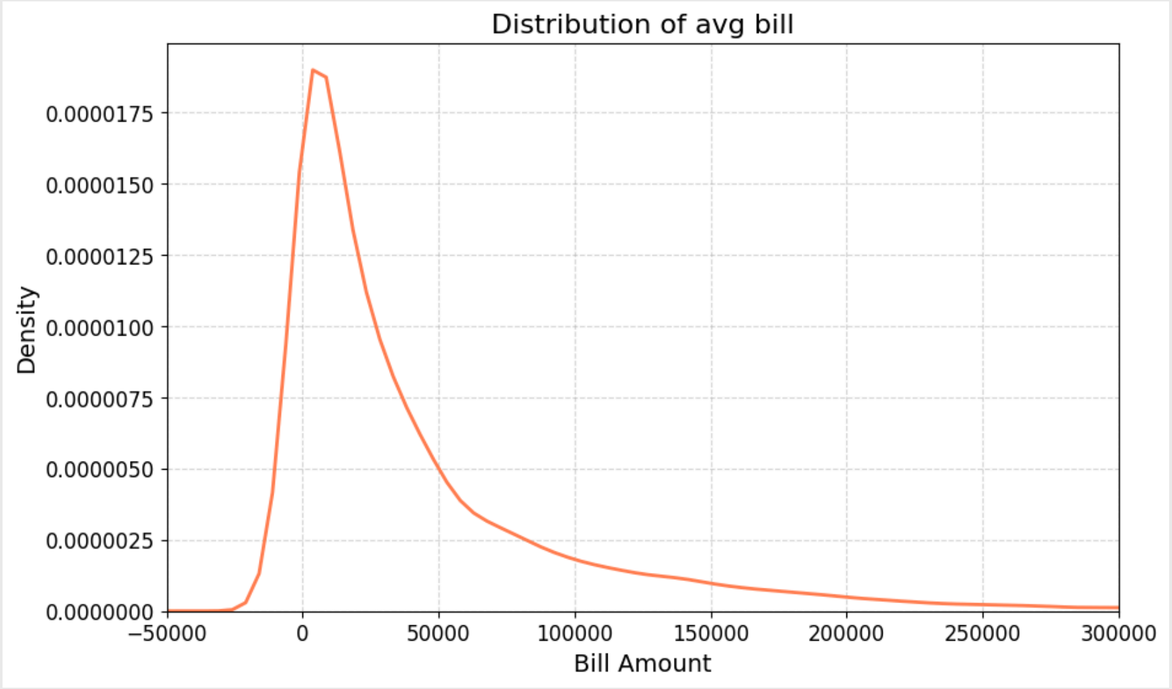
default payment next month	0	1
BALANCE_GROUP		
0-50000	63.929483	36.070517
50000-100000	73.987953	26.012047
100000-150000	76.302083	23.697917
150000-200000	82.387640	17.612360
200000-250000	82.880235	17.119765
250000-300000	85.230769	14.769231
300000-350000	85.611511	14.388489
350000-400000	84.666226	15.333774
400000-450000	87.847731	12.152269
450000-500000	86.853448	13.146552
500000-550000	88.931789	11.068211
550000-600000	83.928571	16.071429
600000-650000	90.000000	10.000000
650000-700000	92.857143	7.142857
700000-750000	85.714286	14.285714
750000-800000	100.000000	0.000000
800000-850000	100.000000	0.000000

only 15% of the population have more than 250k balance for their credit cards



Bill Amounts

- The average bill amount is less 50k per month
- Considering the credit limit for most people were 200k people are only utilizing ¼ of their credit limit

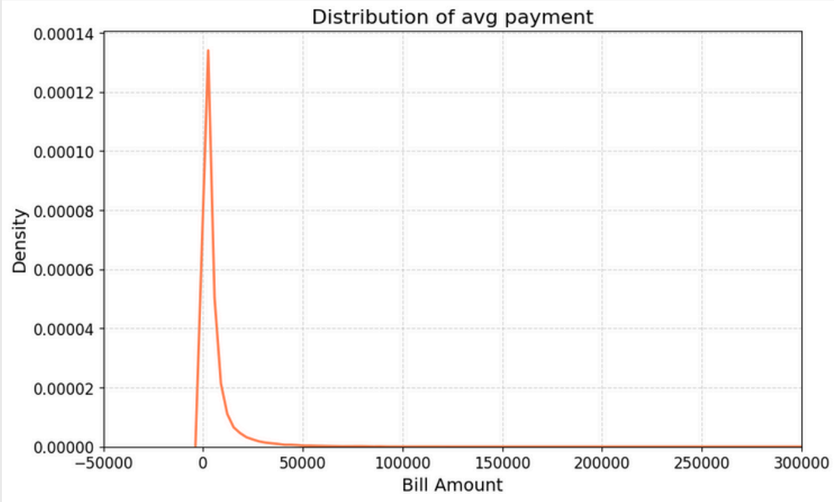



```
description of avg bill

count     30000.000000
mean      44976.945200
std       63260.721860
min      -56043.166667
25%        4781.333333
50%       21051.833333
75%       57104.416667
max      877313.833333
Name: avg_bill, dtype: float64
```

Pay Amounts

- the average payment amount for a user is 5k and capping at 62k
- the users are only paying 11% of their bill, this can cause default rates to go high



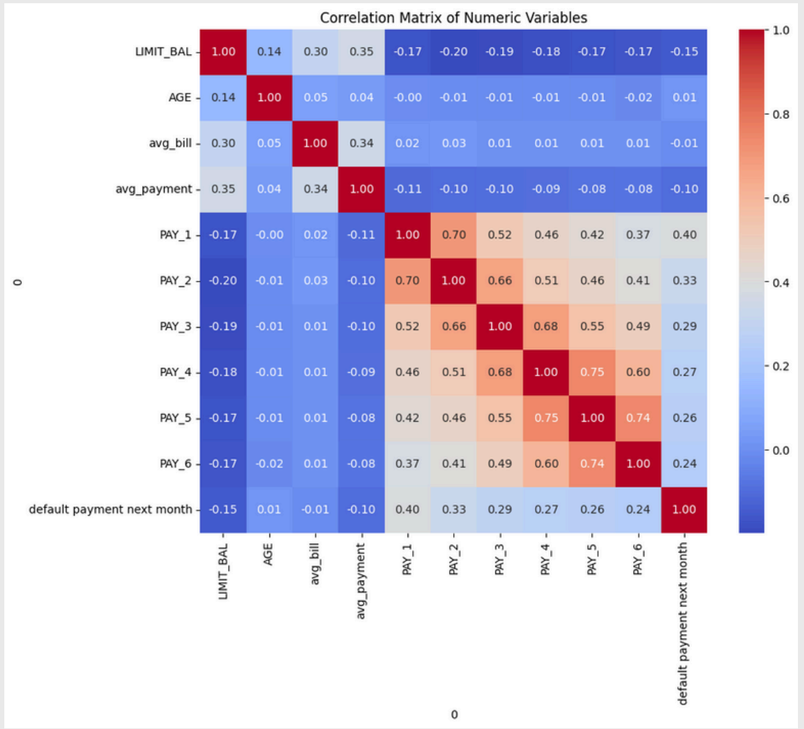

```
description of avg payment

count     30000.000000
mean       5275.232094
std       10137.946323
min           0.000000
25%        1113.291667
50%        2397.166667
75%        5583.916667
max      627344.333333
Name: avg_payment, dtype: float64
```

- People who miss payment for more than 2 months are sure to default

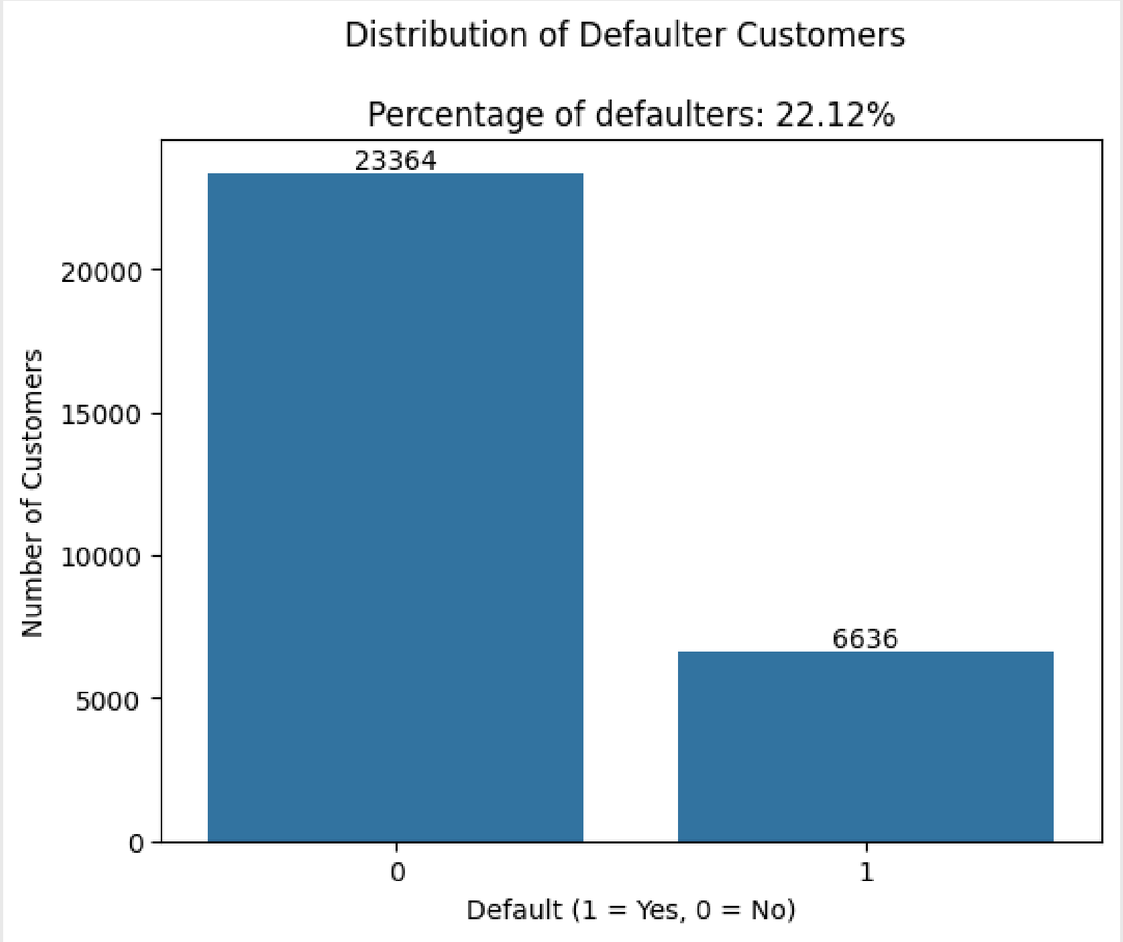
- also people who are defaulter in month 1 tend to continue their habit and forming a viscious cycle and end up becoming a defaulter

default payment next month	0	1
PAY_5		
0	81.551495	18.448505
2	45.811120	54.188880
3	36.516854	63.483146
4	39.285714	60.714286
5	41.176471	58.823529
6	25.000000	75.000000
7	17.241379	82.758621
8	0.000000	100.000000



Target variable

- The target variable is divided into nearly a 80 20 ratio making the data set imbalanced
- Although it is not ideal but it is acceptable and managable to make a good prediction model with this data set



SECTION 3

FEATURE BINNING
FEATURE ENGINEERING
SCALING

Future Goals

THE FUTURE GOAL OF THIS PROJECT IS TO TRANSFORM THE ORIGINAL CREDIT CARD DEFAULT DATASET INTO A CLEAN, STRUCTURED, AND MODEL-READY FORMAT. THE FINAL MACHINE LEARNING MODEL WILL PREDICT THE PROBABILITY OF DEFAULT FOR ANY CUSTOMER BASED ON KEY INPUTS SUCH AS AGE, MARITAL STATUS, GENDER, CREDIT LIMIT, LAST BILL AMOUNT, AND LAST PAYMENT AMOUNT. THIS ENABLES EARLY RISK DETECTION AND MORE INFORMED CREDIT-LENDING DECISIONS.

ENHANCEMENTS MADE TO THE DATASET

- CLEANED AND STANDARDIZED CATEGORICAL VARIABLES SUCH AS MARITAL STATUS, EDUCATION, AND GENDER TO ENSURE CONSISTENCY.
- CORRECTED INVALID VALUES AND CONVERTED CATEGORICAL FIELDS INTO MEANINGFUL LABELS.
- HANDLED OUTLIERS IN BILL AMOUNTS, PAYMENT AMOUNTS, AND PAYMENT DELAYS TO MAINTAIN REALISTIC DATA DISTRIBUTION.
- CREATED AGGREGATED BEHAVIOURAL FEATURES, INCLUDING AVERAGE BILL AMOUNT, AVERAGE PAYMENT AMOUNT, AND AVERAGE DELAY PATTERNS.
- BINNED CONTINUOUS VARIABLES INTO CATEGORIES (E.G., LOW/MEDIUM/HIGH SPENDING) TO SIMPLIFY PATTERN ANALYSIS.
- APPLIED ONE-HOT ENCODING TO CONVERT CATEGORICAL VARIABLES INTO MACHINE-READABLE BINARY COLUMNS.
- SCALED NUMERICAL FEATURES TO ENSURE BALANCED CONTRIBUTION DURING MODEL TRAINING.
- DROPPED IRRELEVANT FIELDS SUCH AS ID AND TEMPORARY PREPROCESSING COLUMNS TO KEEP THE DATASET FOCUSED AND EFFICIENT.

THANK YOU