Credit Card Analysis :

Unveiling Trends , Risks and Default Probability

Data Summary

There are 30k rows and 25 columns.

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PA
0	1	20000	2	2	1	24	2	2	0	0	...	0	0	0	0	
1	2	120000	2	2	2	26	0	2	0	0	...	3272	3455	3261	0	
2	3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	
3	4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	
4	5	50000	1	2	1	57	0	0	0	0	...	20940	19146	19131	2000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
29995	29996	220000	1	3	1	39	0	0	0	0	...	88004	31237	15980	8500	
29996	29997	150000	1	3	2	43	0	0	0	0	...	8979	5190	0	1837	
29997	29998	30000	1	2	2	37	4	3	2	0	...	20878	20582	19357	0	

PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
2	2	0	0	...	0	0	0	0	689	0	0	0	0	1
0	2	0	0	...	3272	3455	3261	0	1000	1000	1000	0	2000	1
0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
0	0	0	0	...	20940	19146	19131	2000	36681	10000	9000	689	679	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	...	88004	31237	15980	8500	20000	5003	3047	5000	1000	0
0	0	0	0	...	8979	5190	0	1837	3526	8998	129	0	0	0
4	3	2	0	...	20878	20582	19357	0	0	22000	4200	2000	3100	1

Problem Statement: A Card issuing Bank has over issued its cash and credit card in-order-to increase its market share, even to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash–card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders. From the perspective of risk control, estimating the probability of default will be more meaningful than classifying customers into the binary results – risky and non-risky.

Description of the Data:

This research employed a binary variable – default payment, Y (Yes = 1, No = 0), as the response variable. This study used the following 23 variables as explanatory variables: X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. X2: Gender (1 = male; 2 = female). X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). X4: Marital status (1 = married; 2 = single; 3 = others). X5: Age (year). X6–X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above. X12–X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September,2005; X13 = amount of bill statement in August, 2005; X17 = amount of bill statement in April, 2005. X18–X23: Amount of previous payment (NT dollar). X18 = amount paid in September,2005; X19 = amount paid in August,2005; X23 = amount paid in April 2005.

Renaming Columns According to Data Description

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_SEPT	PAY_AUG	PAY_JUL	PAY_JUN	...	BILL_AMT_APR	PAY_AMT_SEPT	PAY_AMT_AUG
0	1	20000	female	University	married	24	2	2	0	0	...	0	0	689
1	2	120000	female	University	single	26	0	2	0	0	...	3261	0	1000
2	3	90000	female	University	single	34	0	0	0	0	...	15549	1518	1500
3	4	50000	female	University	married	37	0	0	0	0	...	29547	2000	2019
4	5	50000	male	University	married	57	0	0	0	0	...	19131	2000	36681

PAY_AMT_JUL	PAY_AMT_JUN	PAY_AMT_MAY	PAY_AMT_APR	default payment next month
0	0	0	0	yes
1000	1000	0	2000	yes
1000	1000	1000	5000	no
1200	1100	1069	1000	no
10000	9000	689	679	no

ANALYTICS TASKS:

1.Identify the errors in dataset.

2.Clean those errors in the dataset.

3.Analyze the trend on outstanding amount for the bank

4.Is there any relationship between in outstanding amount / trend with respect to age, education, marriage, credit

5.Does outstanding amount / trend affect the default behavior in next month.

6..Try to Apply Statistics like EDA, Confidence Interval, Probability Distribution & Hypothesis.

1.Identifying errors in dataset

```
data.isnull().sum()
```

```
ID                          0
LIMIT_BAL                   0
SEX                         0
EDUCATION                   0
MARRIAGE                    0
AGE                         0
PAY_0                       0
PAY_2                       0
PAY_3                       0
PAY_4                       0
PAY_5                       0
PAY_6                       0
BILL_AMT1                   0
BILL_AMT2                   0
BILL_AMT3                   0
BILL_AMT4                   0
BILL_AMT5                   0
BILL_AMT6                   0
PAY_AMT1                    0
PAY_AMT2                    0
PAY_AMT3                    0
PAY_AMT4                    0
PAY_AMT5                    0
PAY_AMT6                    0
default payment next month  0
dtype: int64
```

```
data['SEX'].unique()
```

```
array([2, 1], dtype=int64)
```

```
data['EDUCATION'].unique()
```

```
array([2, 1, 3, 5, 4, 6, 0], dtype=int64)
```

```
data['MARRIAGE'].unique()
```

```
array([1, 2, 3, 0], dtype=int64)
```

```
data['default payment next month'].unique()
```

```
array([1, 0], dtype=int64)
```

2.CLEANING THE ERRORS

```
data['SEX'].replace(1,'male',inplace = True)
```

```
data['SEX'].replace(2,'female',inplace = True)
```

```
data['SEX'].unique()
```

```
array(['female', 'male'], dtype=object)
```

```
data['EDUCATION'].replace(1,'graduate school',inplace = True)
```

```
data['EDUCATION'].replace(2,'University',inplace = True)
```

```
data['EDUCATION'].replace(3,'High school',inplace = True)
```

```
data['EDUCATION'].replace(4,'others',inplace = True)
```

```
data['EDUCATION'].replace(5,'others',inplace = True)
```

```
data['EDUCATION'].replace(6,'others',inplace = True)
```

```
data['EDUCATION'].replace(0,'others',inplace = True)
```

```
data['EDUCATION'].unique()
```

```
array(['University', 'graduate school', 'High school', 'others'],
      dtype=object)
```

```
data['MARRIAGE'].replace(1, 'married' , inplace = True)
```

```
data['MARRIAGE'].replace(2, 'single' , inplace = True)
```

```
data['MARRIAGE'].replace(3, 'others' , inplace = True)
```

```
data['MARRIAGE'].replace(0, 'others' , inplace = True)
```

```
data['MARRIAGE'].unique()
```

```
array(['married', 'single', 'others'], dtype=object)
```

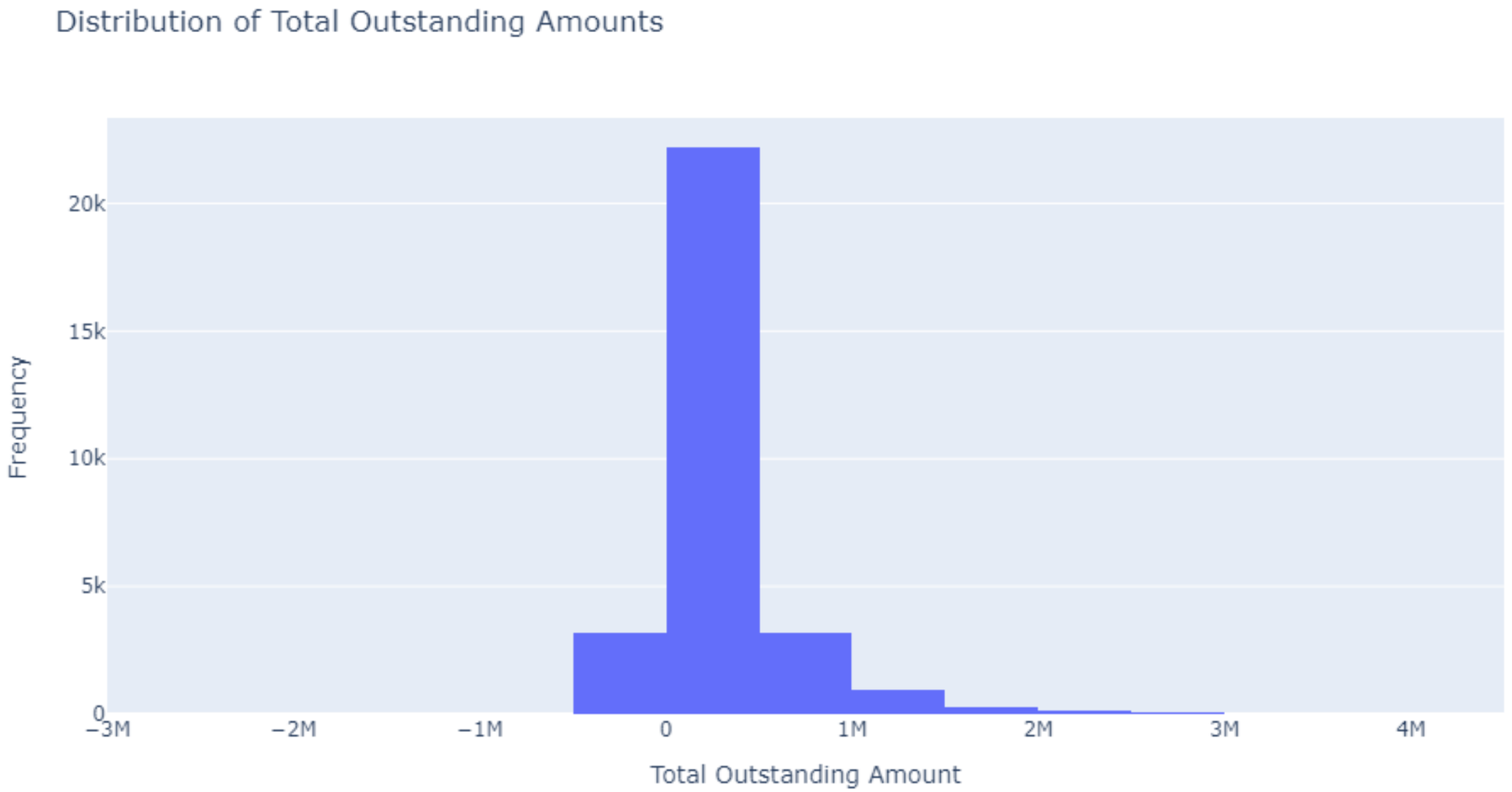
```
data['default payment next month'].replace(0,'no', inplace = True)
```

```
data['default payment next month'].replace(1,'yes', inplace = True)
```

```
data['default payment next month'].unique()
```

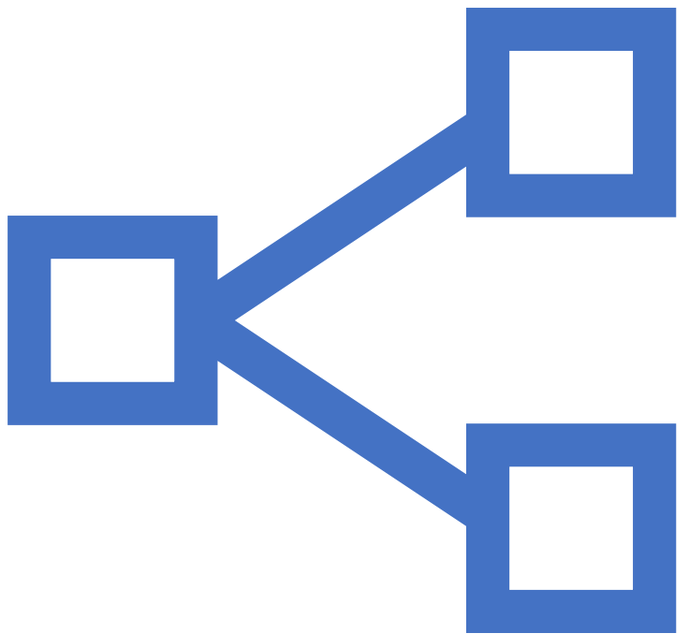
```
array(['yes', 'no'], dtype=object)
```

3.ANALYZE TREND ON OUTSTANDING AMOUNT



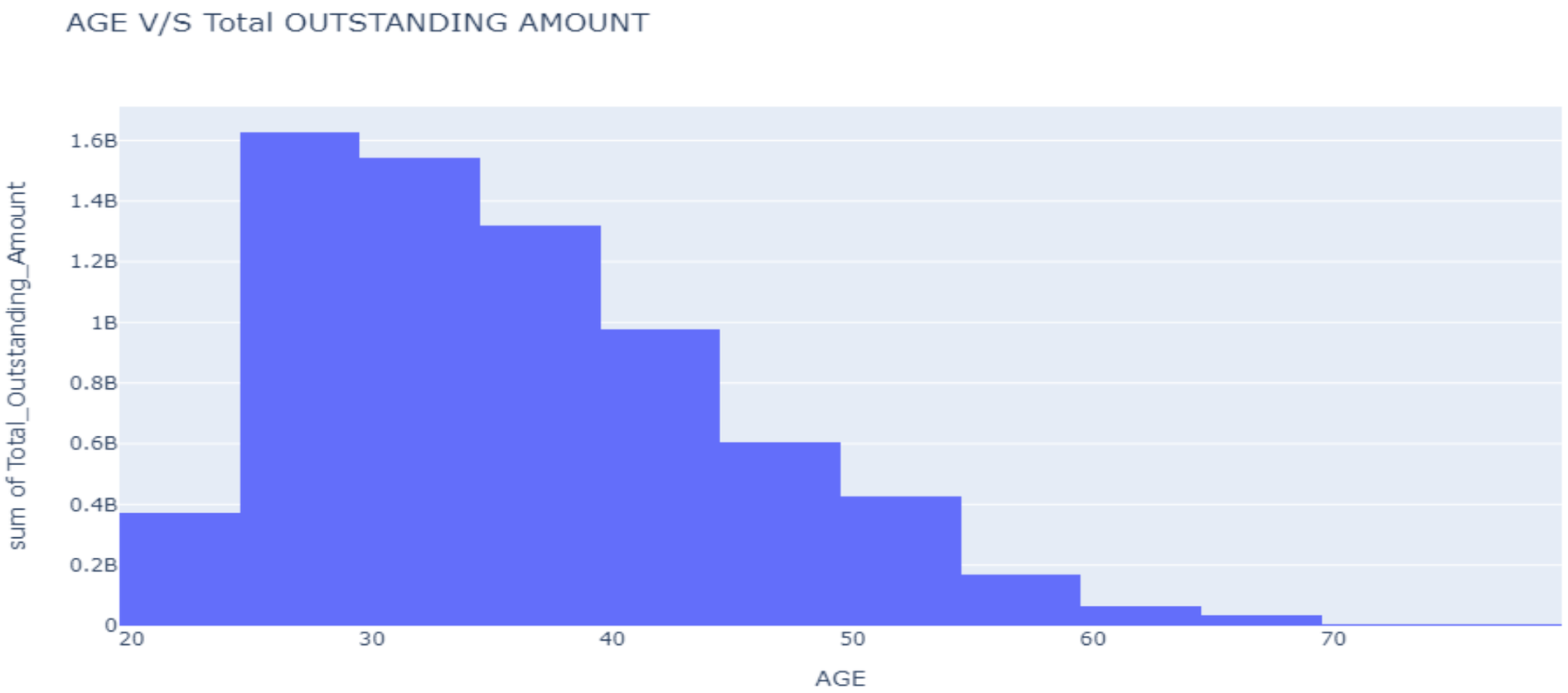
Here Total Outstanding Amount = Total Amount Owed – Total Payments

TOTAL PAYMENT AMOUNT AND TOTAL AMOUNT OWED



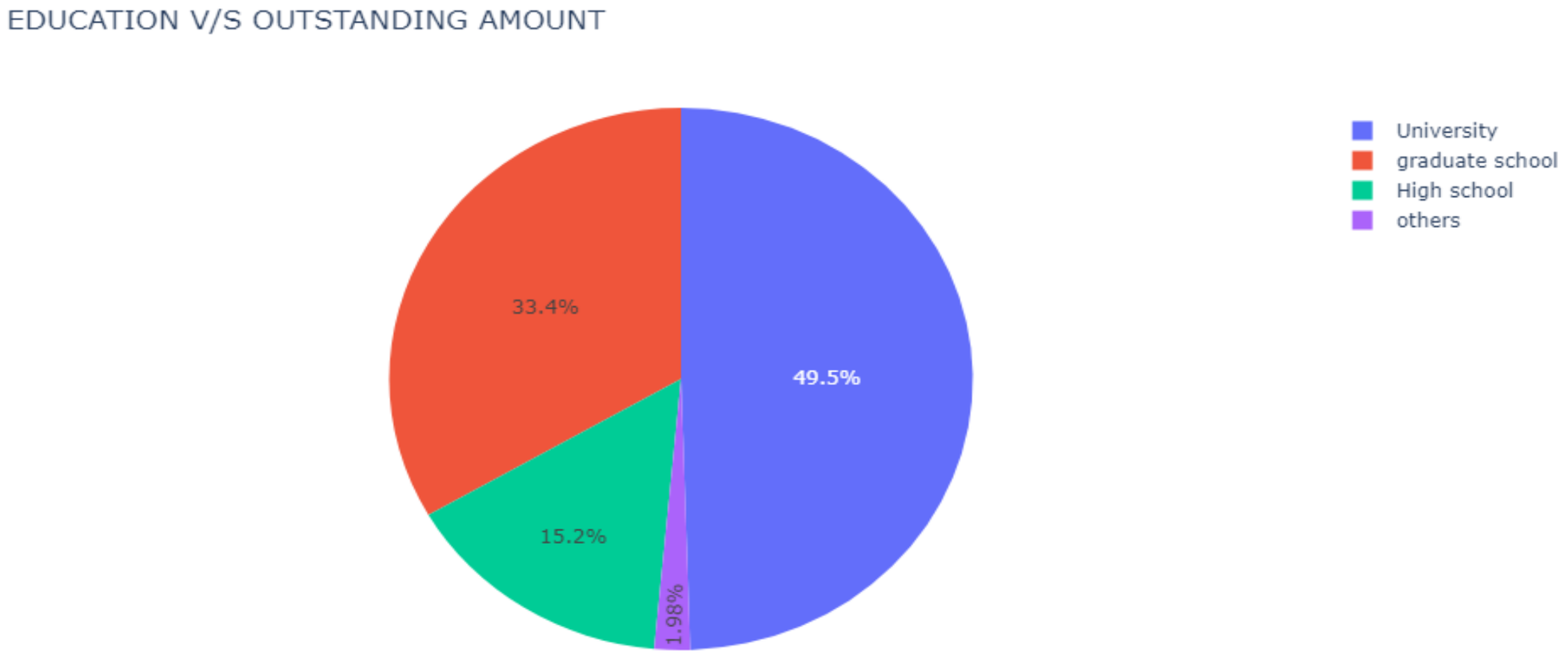
4. Is there any relationship between in outstanding amount / trend with respect to age, education, marriage, credit limit

RELATIONSHIP B/W AGE AND TOTAL OUTSTANDING AMOUNT



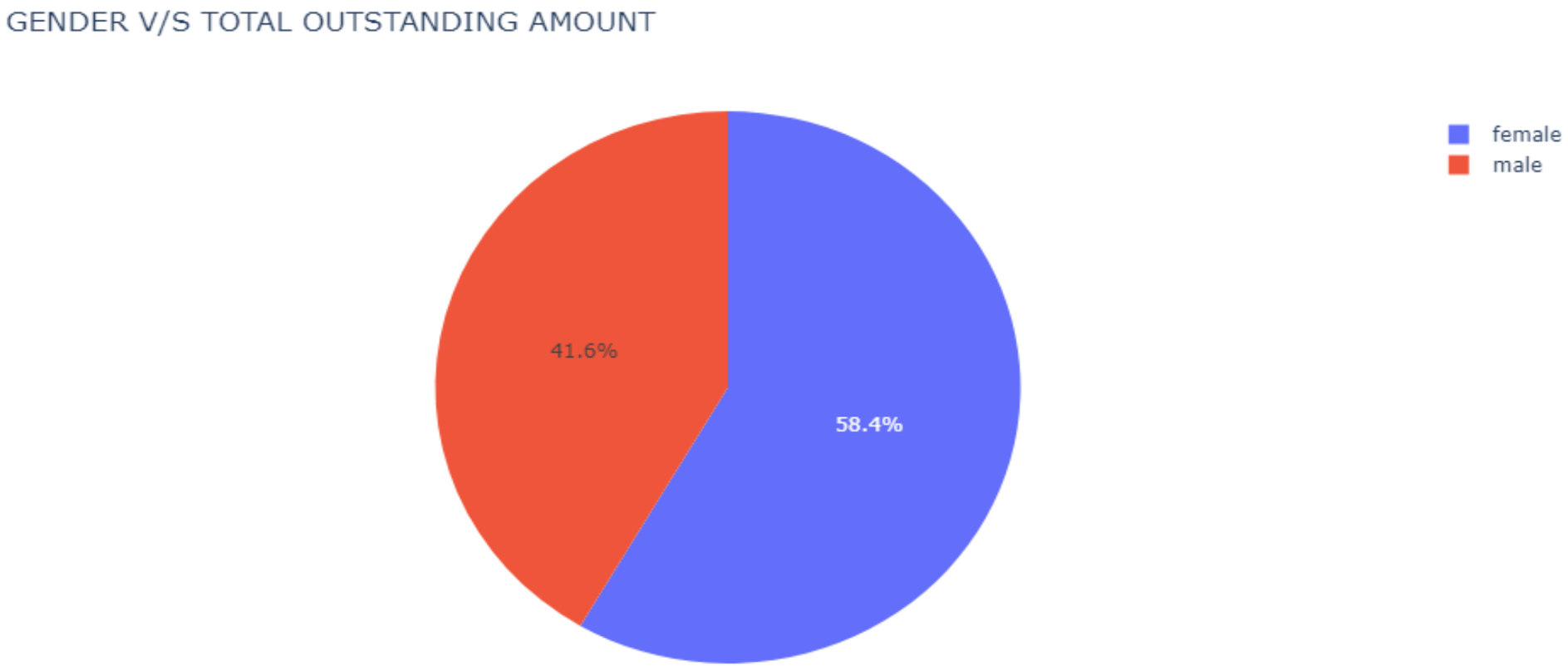
The total outstanding amount of ages b/w 25-29 is more followed by 30-34 .

RELATIONSHIP B/W EDUCATION AND TOTAL OUTSTANDING AMOUNT



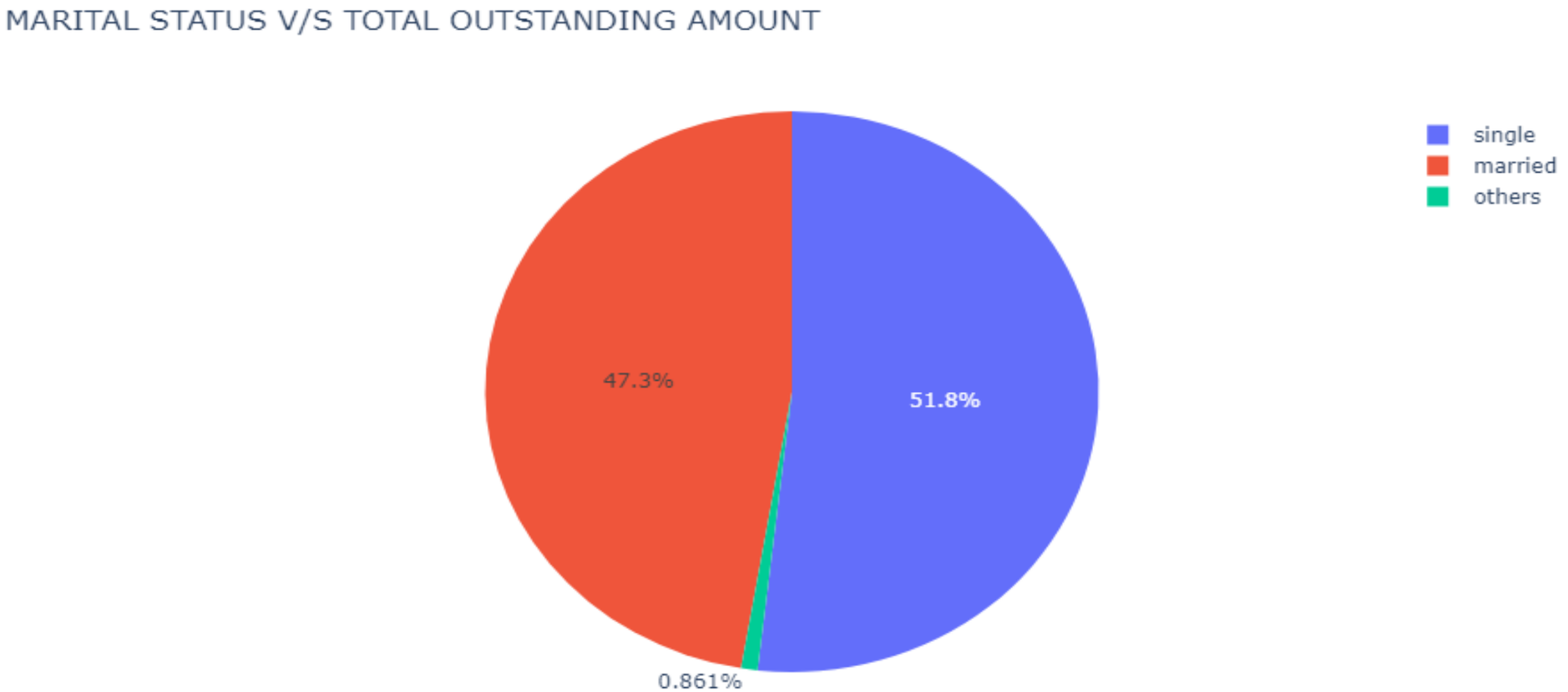
University students are having more total outstanding amount followed by graduates.

RELATIONSHIP B/W GENDER AND TOTAL OUTSTANDING AMOUNT



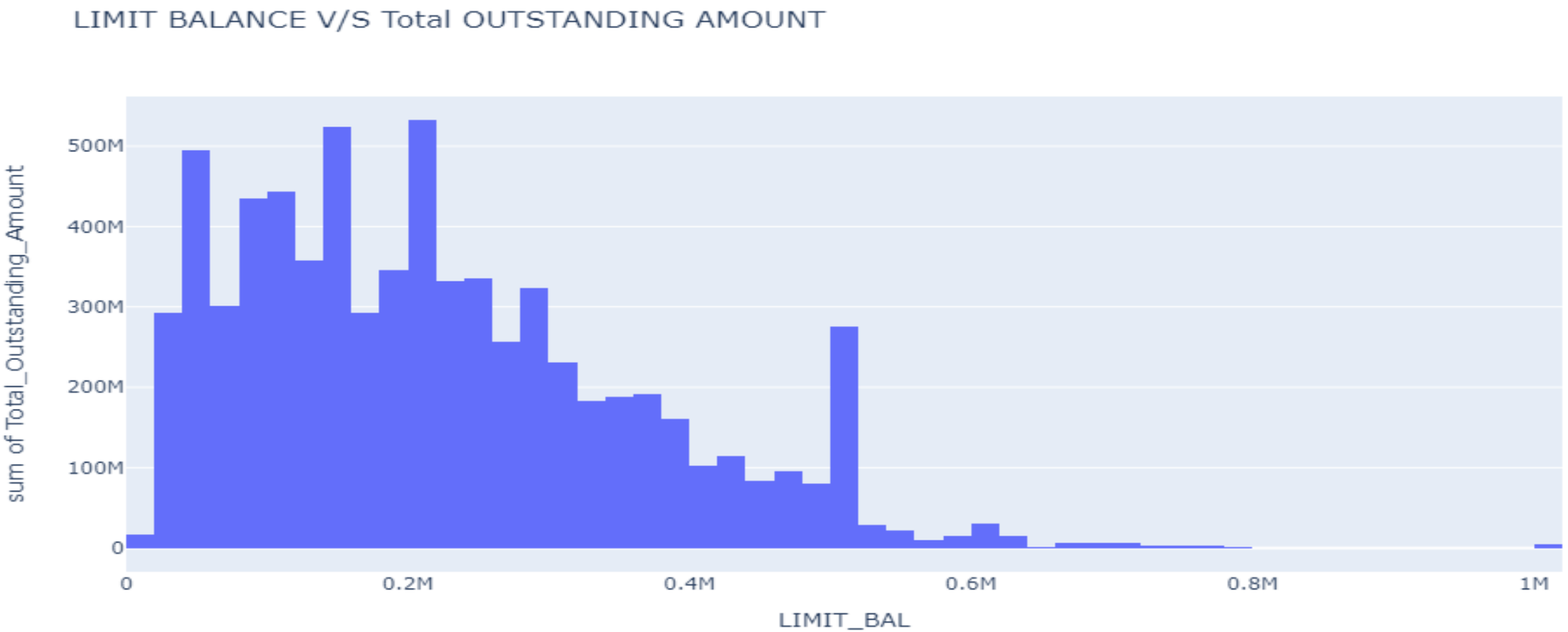
Females are having high total outstanding amount.

RELATION B/W MARITAL STATUS AND TOTAL OUTSTANDING AMOUNT



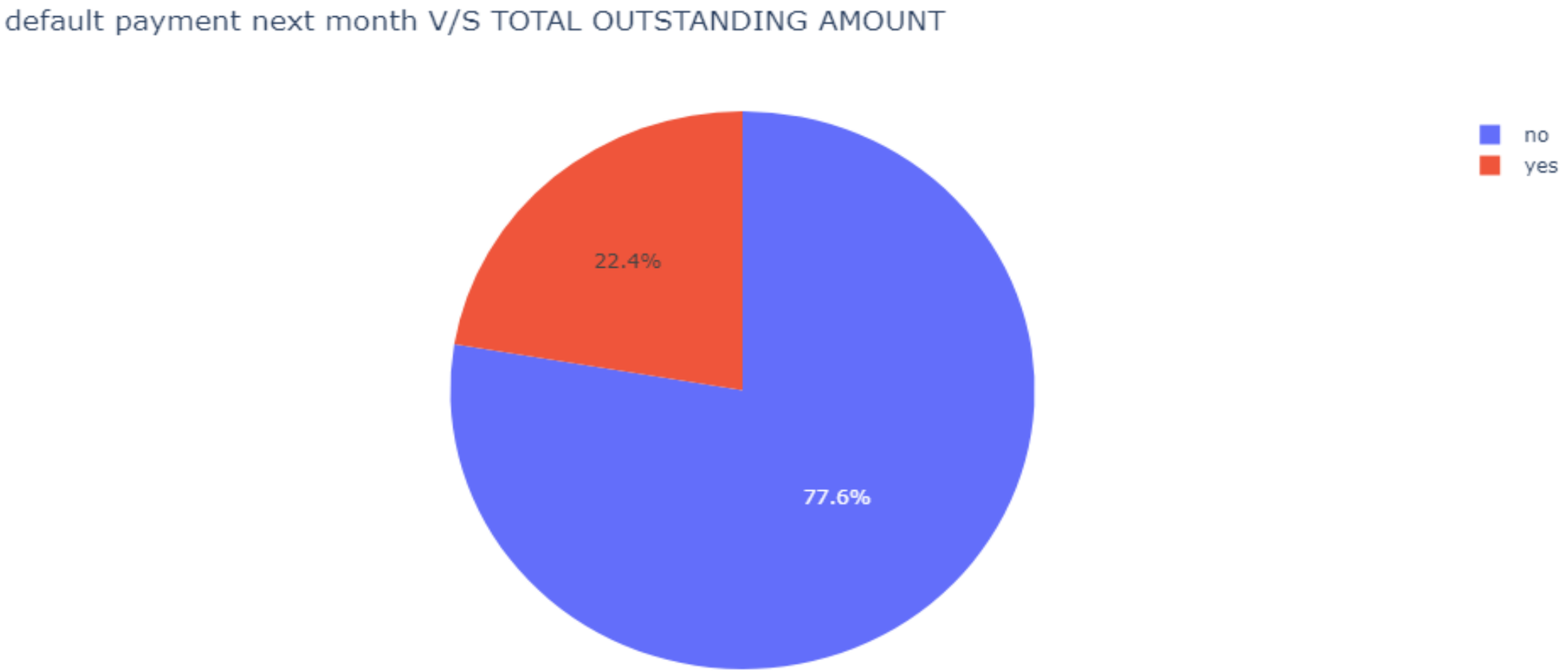
Singles are having more outstanding amount followed by Married ones.

RELATIONSHIP B/W LIMIT BALANCE & TOTAL OUTSTANDING AMOUNT



People with 200K-219k limit balance are having more total outstanding amount (I . E. 532.1907M), followed by 140K-149K limit balance people with total outstanding amount (524.7013M).

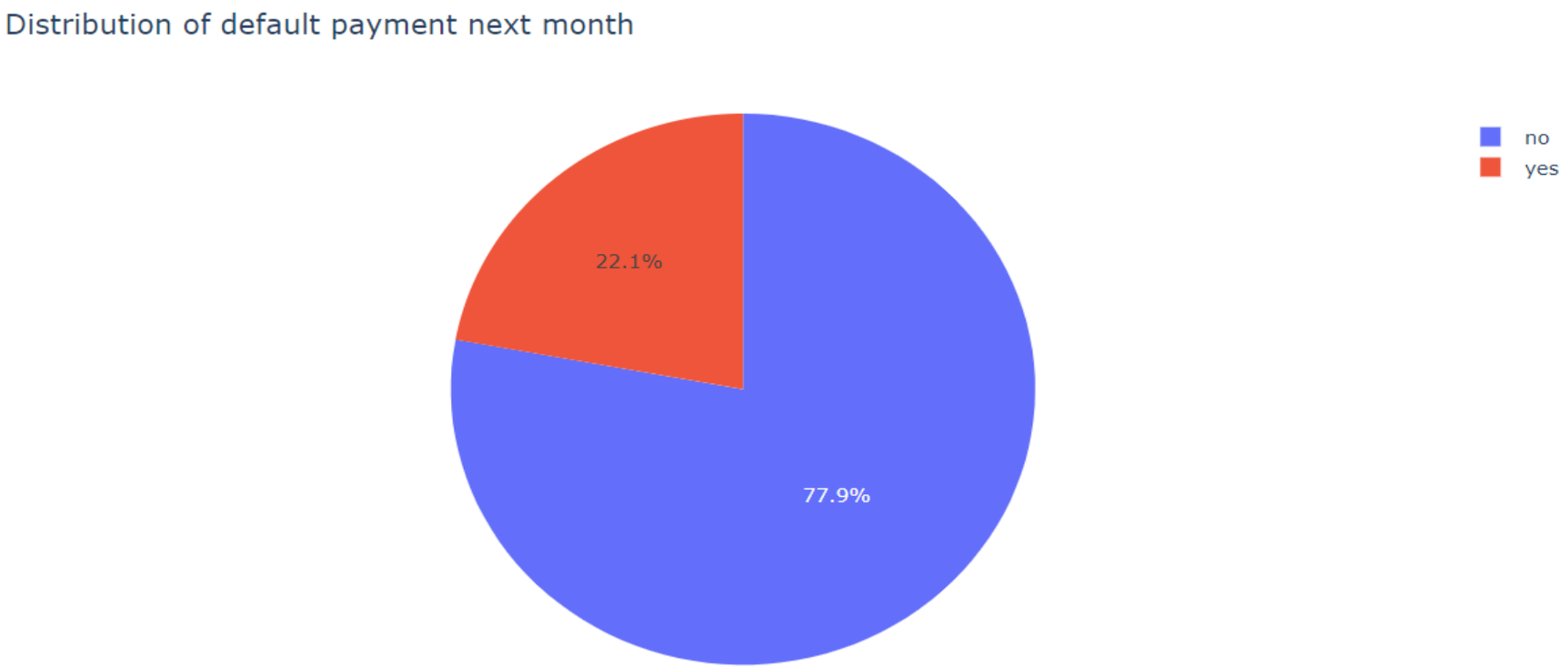
5.EFFECT OF TOTAL OUTSTANDING AMOUNT & DEFAULT PAYMENT NEXT MONTH



As we can see from the above pie that there are 77.6% non-defaulters are present & defaulters are up to 22.4%.

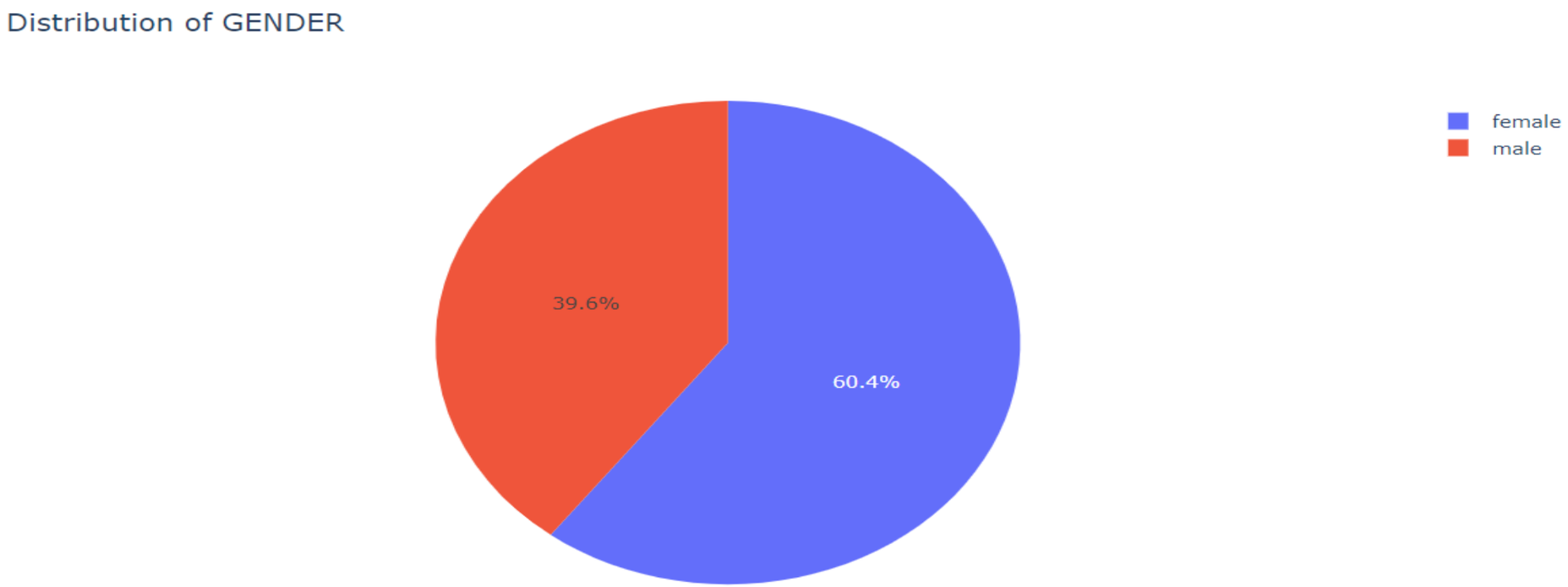
6.Exploratory Data Analysis

ANALYSIS OF DEPENDENT VARIABLE : DEFAULT PAYMENT NEXT MONTH



There are 23364 (77.9%) Non- defaulters and 6636 (22.1%) defaulters are present .

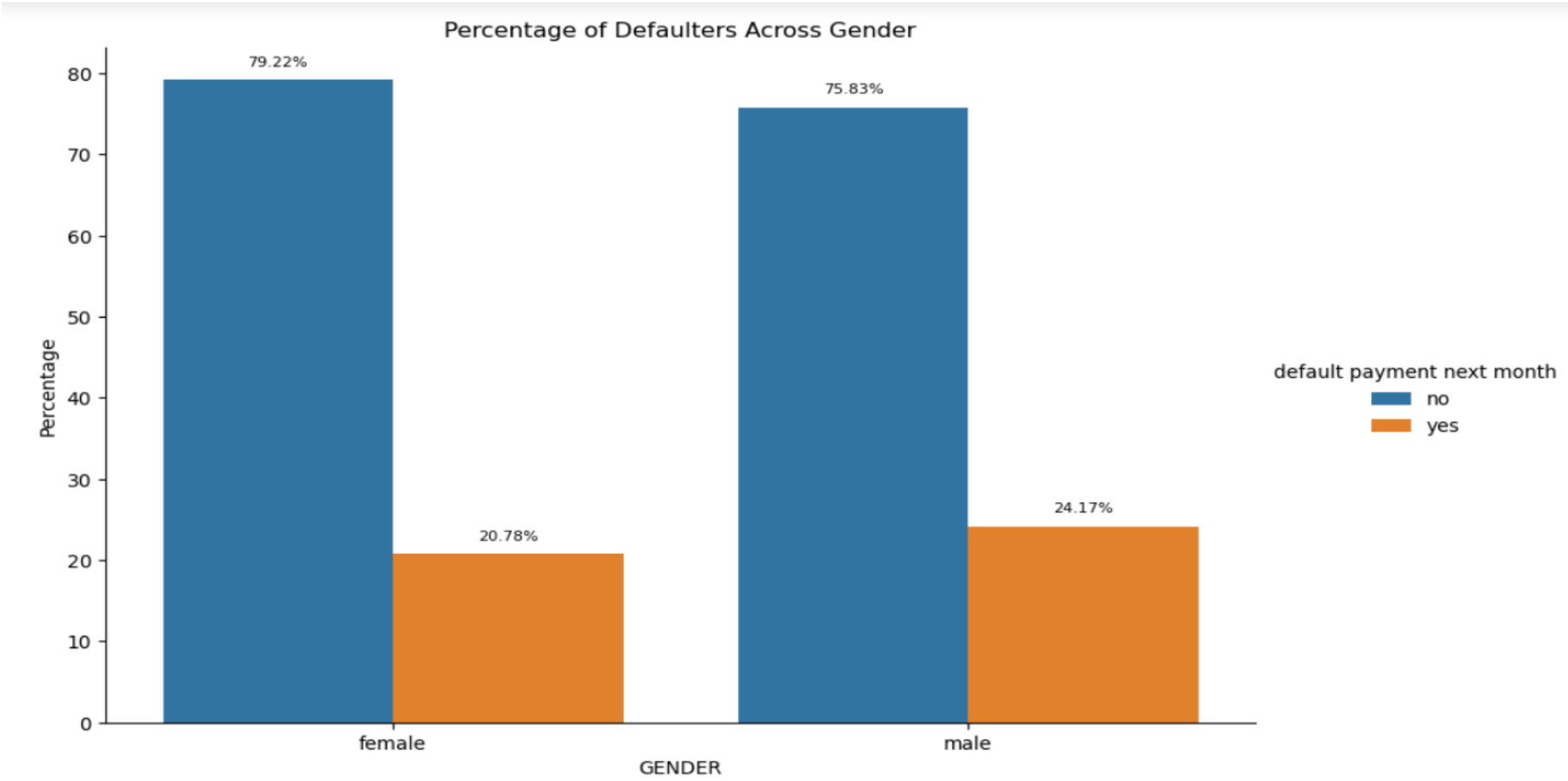
ANALYSIS OF SEX VARIABLE



There are more female credit card holders

SEX	
female	18112
male	11888

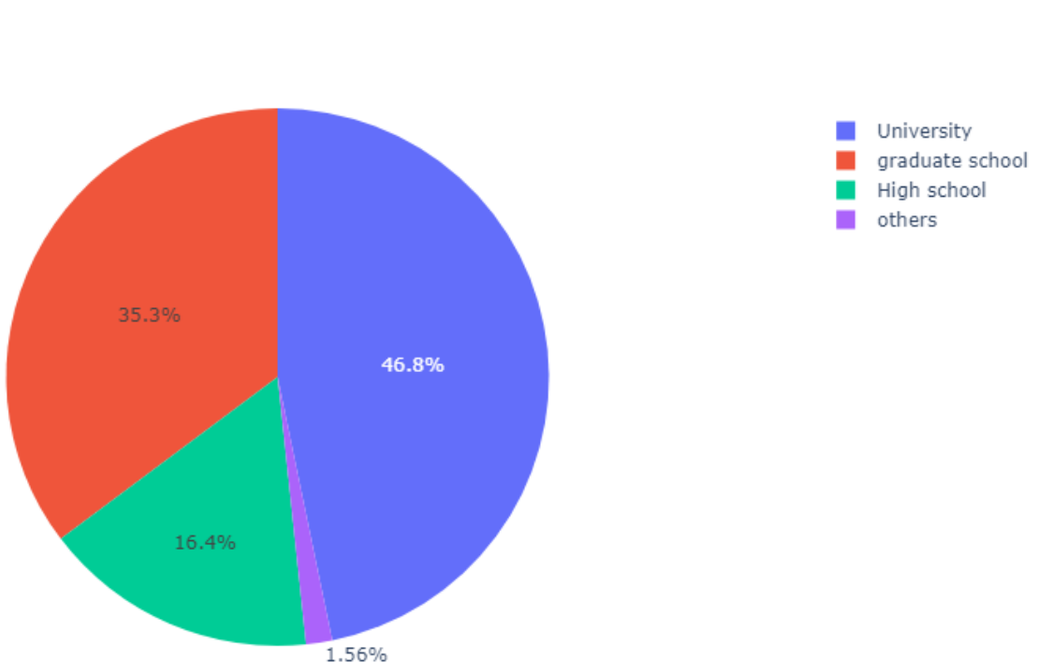
PROPORTION OF DEFAULTERS AMONG GENDER



- It is evident from graph that the number of defaulter have high proportion of males I . e 24.17%.

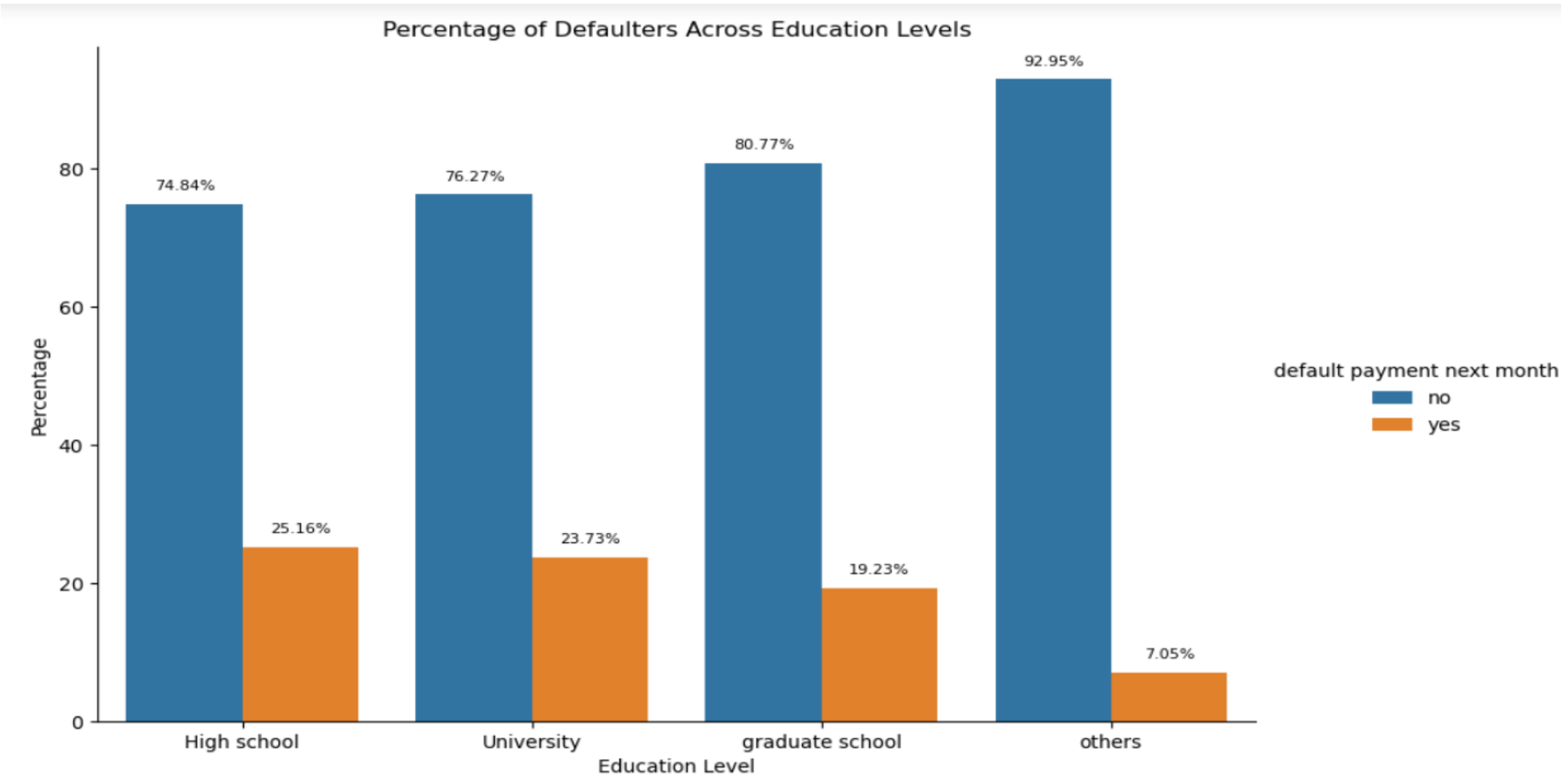
ANALYSIS OF EDUCATION VARIABLE

- **EDUCATION**
- **University 14030**
- **graduate school 10585**
- **High school 4917**
- **others 468**



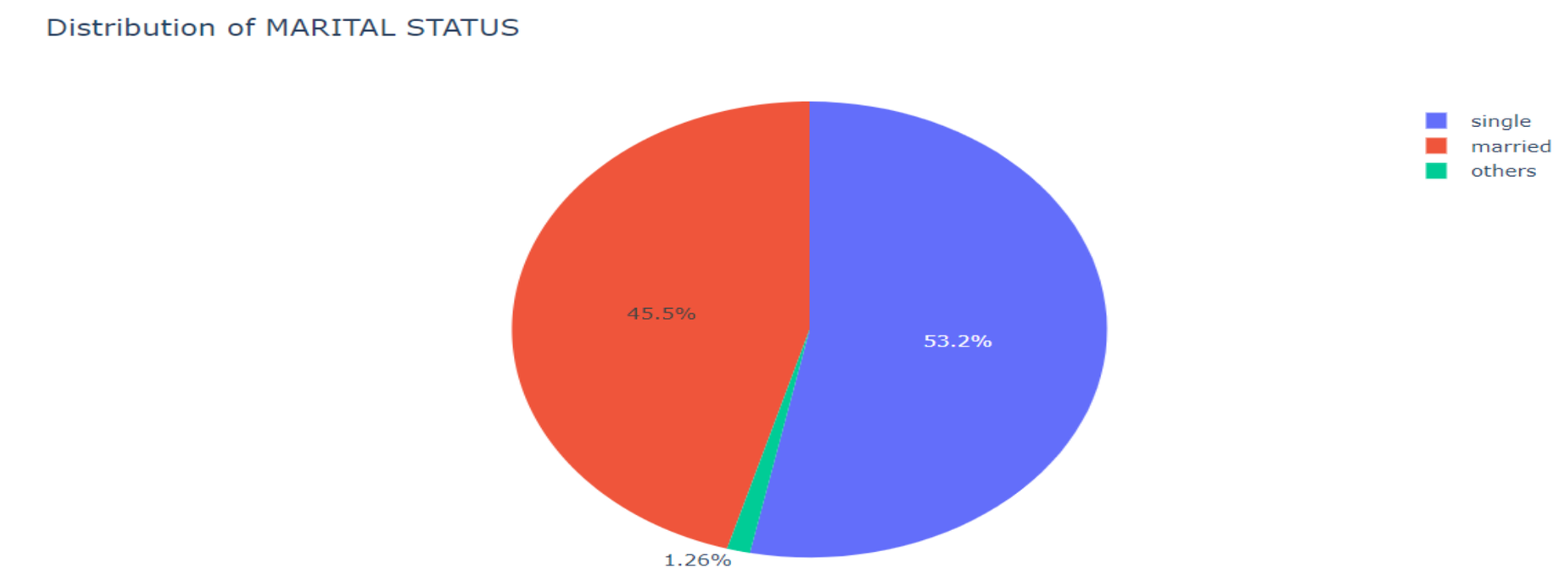
From the above pie we can say that more number of credit card holders are university students followed by graduates & then high school students.

DEFAULTERS AMONG EDUCATION LEVELS



There are more defaulters among High school students followed by university students.

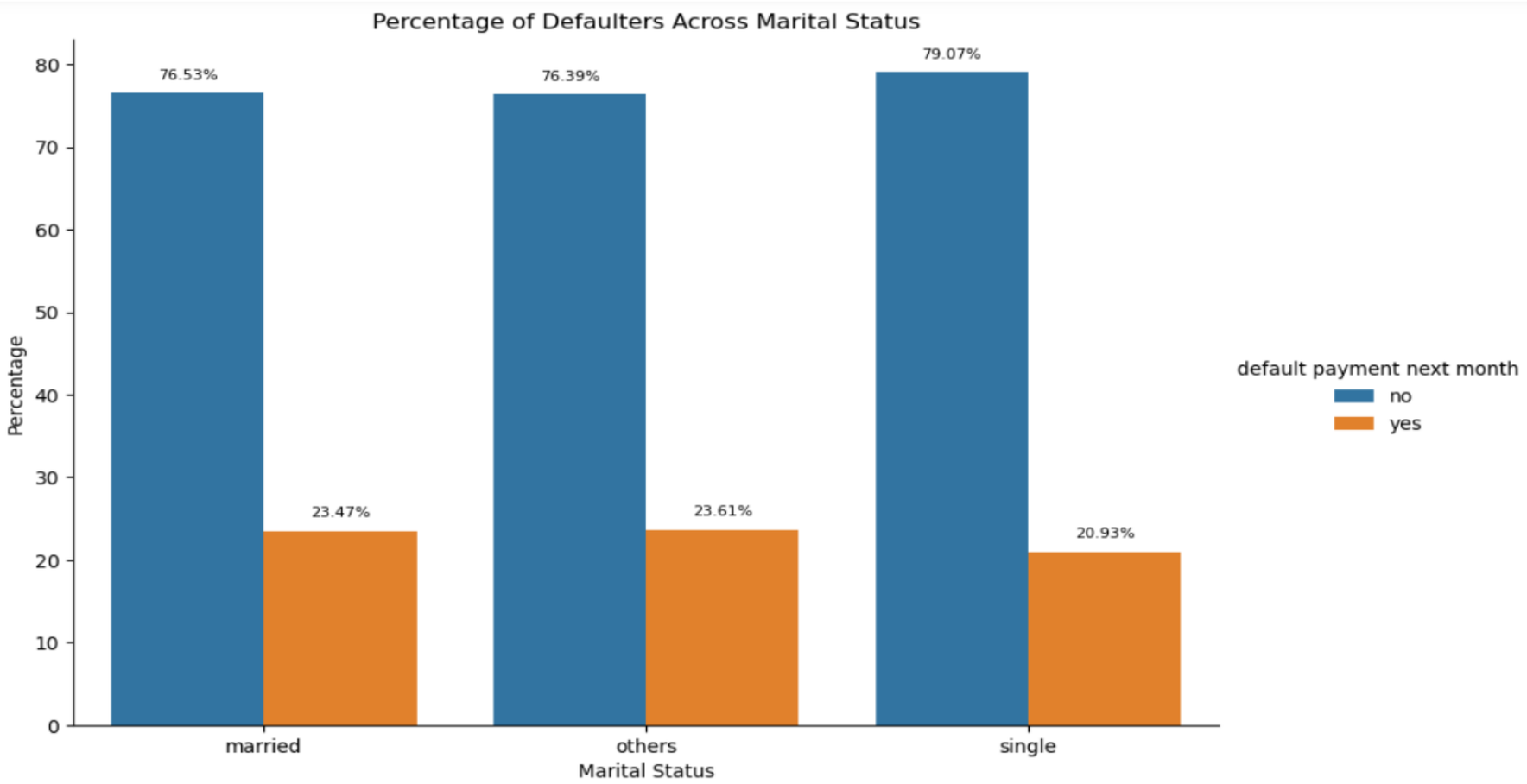
ANALYSIS OF MARRIAGE VARIABLE



From the above pie we can observe that more credit card holders are single.

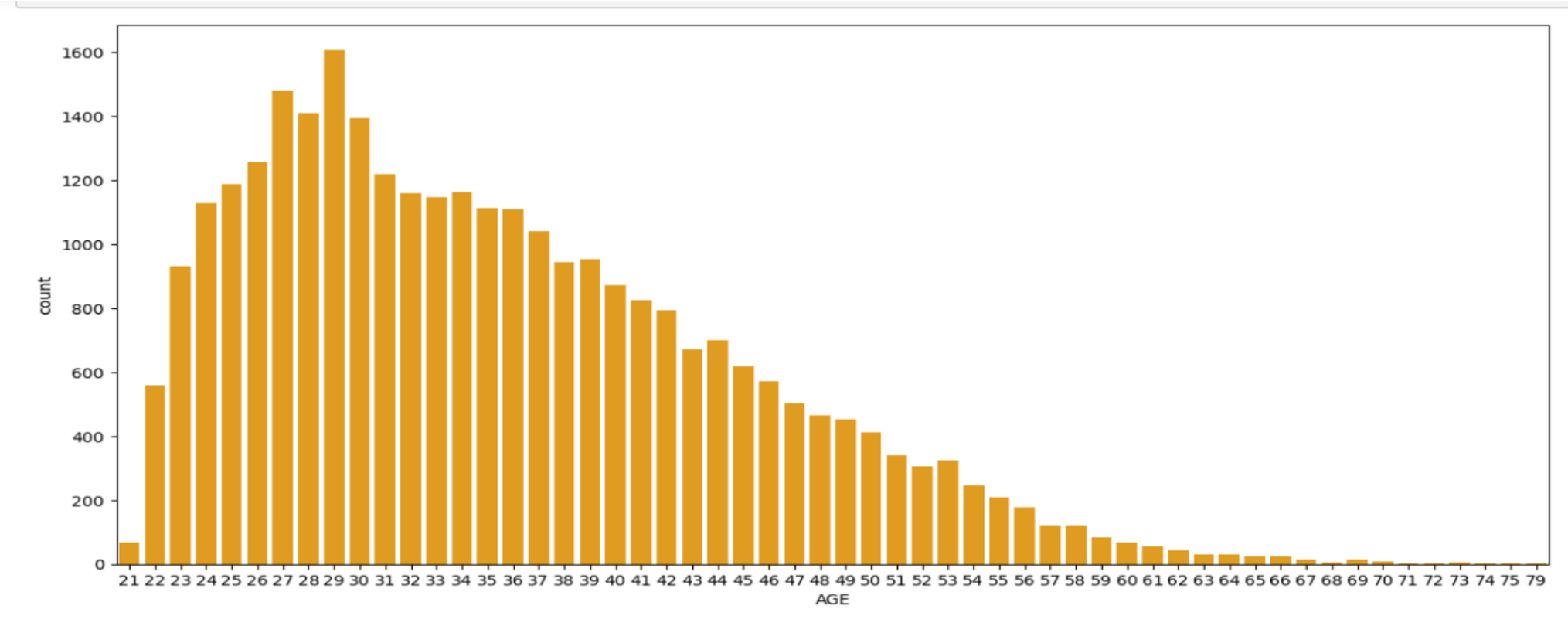
MARRIAGE	
single	15964
married	13659
others	377

DEFAULTERS AND MARITAL STATUS



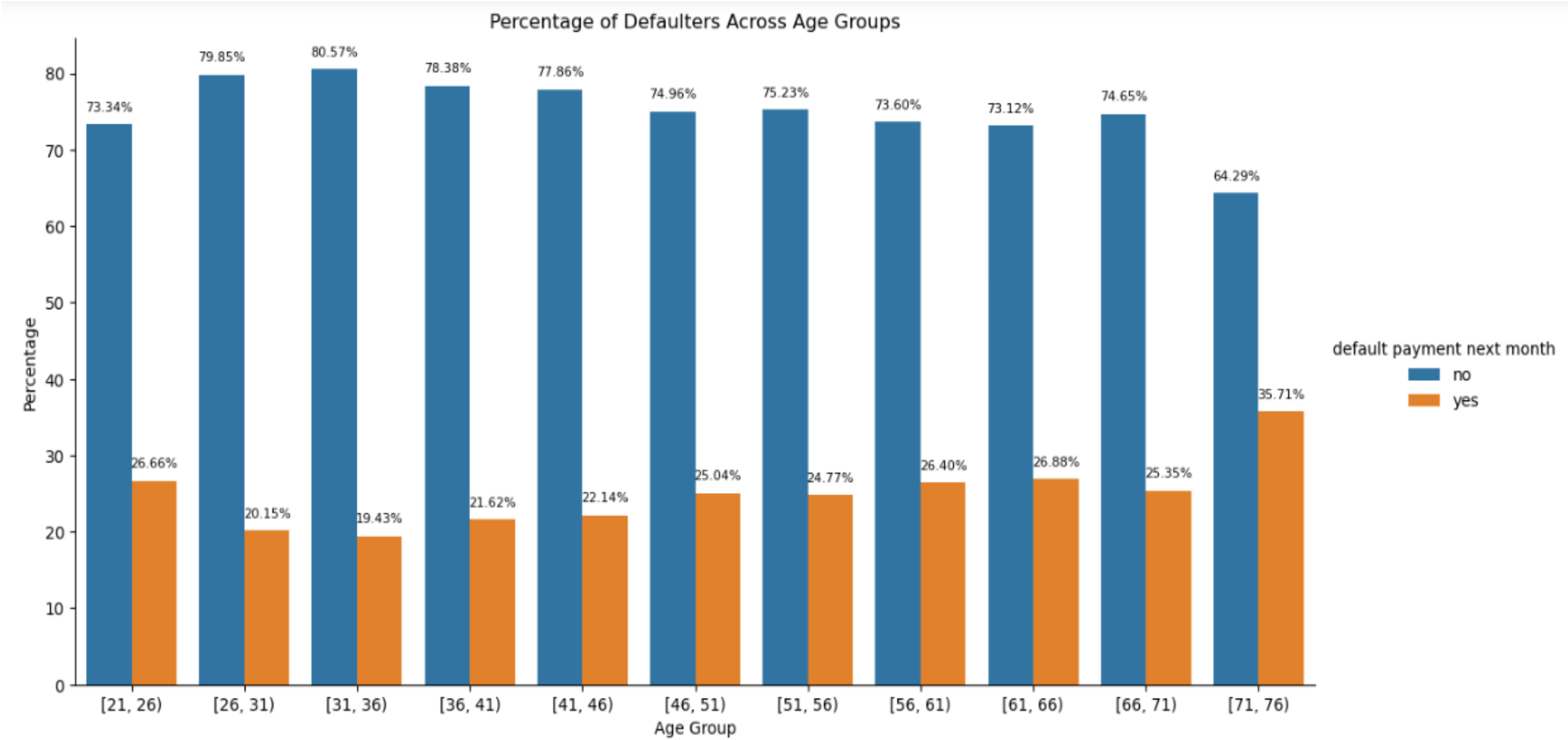
From the right plot we see that high defaulters are Others followed by Married ones.

ANALYSIS OF AGE VARIABLE



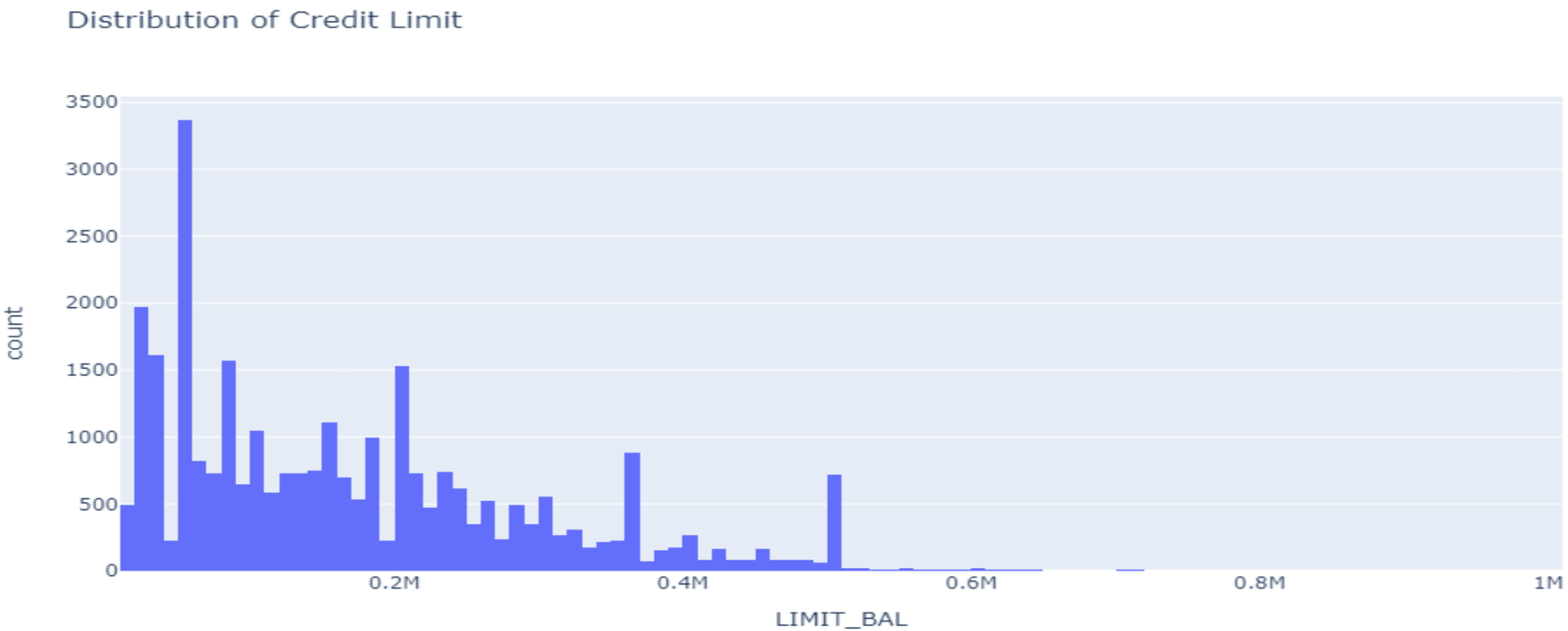
- From this count plot we can say that
- The more credit card holders are aged between 26-30.
- Age above 60 years old rarely uses credit card.

Defaulters And Non-Defaulters AMONG AGE GROUPS



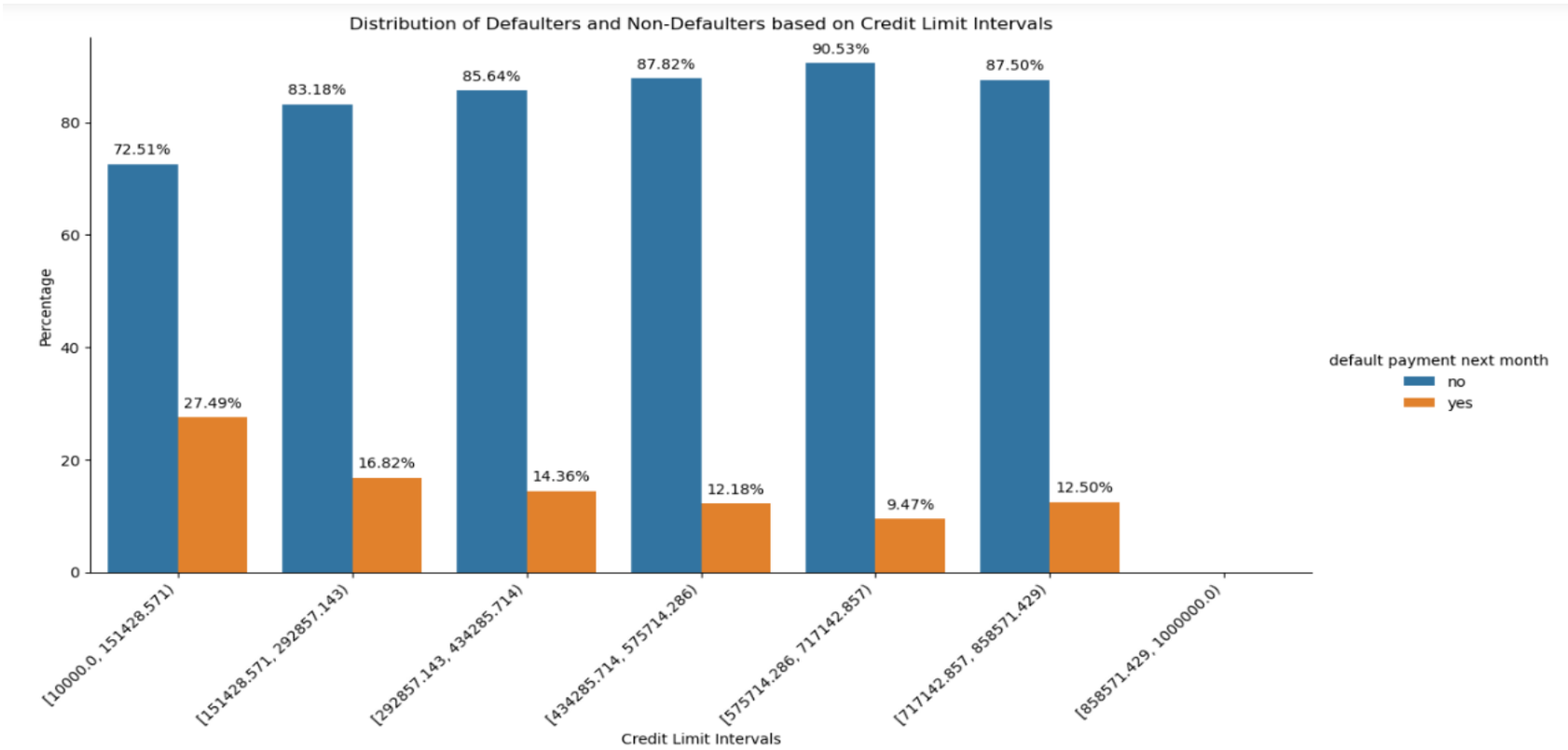
Those who default are 60 years and older ,that may be they don't use their card frequently.

ANALYSIS OF LIMIT BALANCE VARIABLE



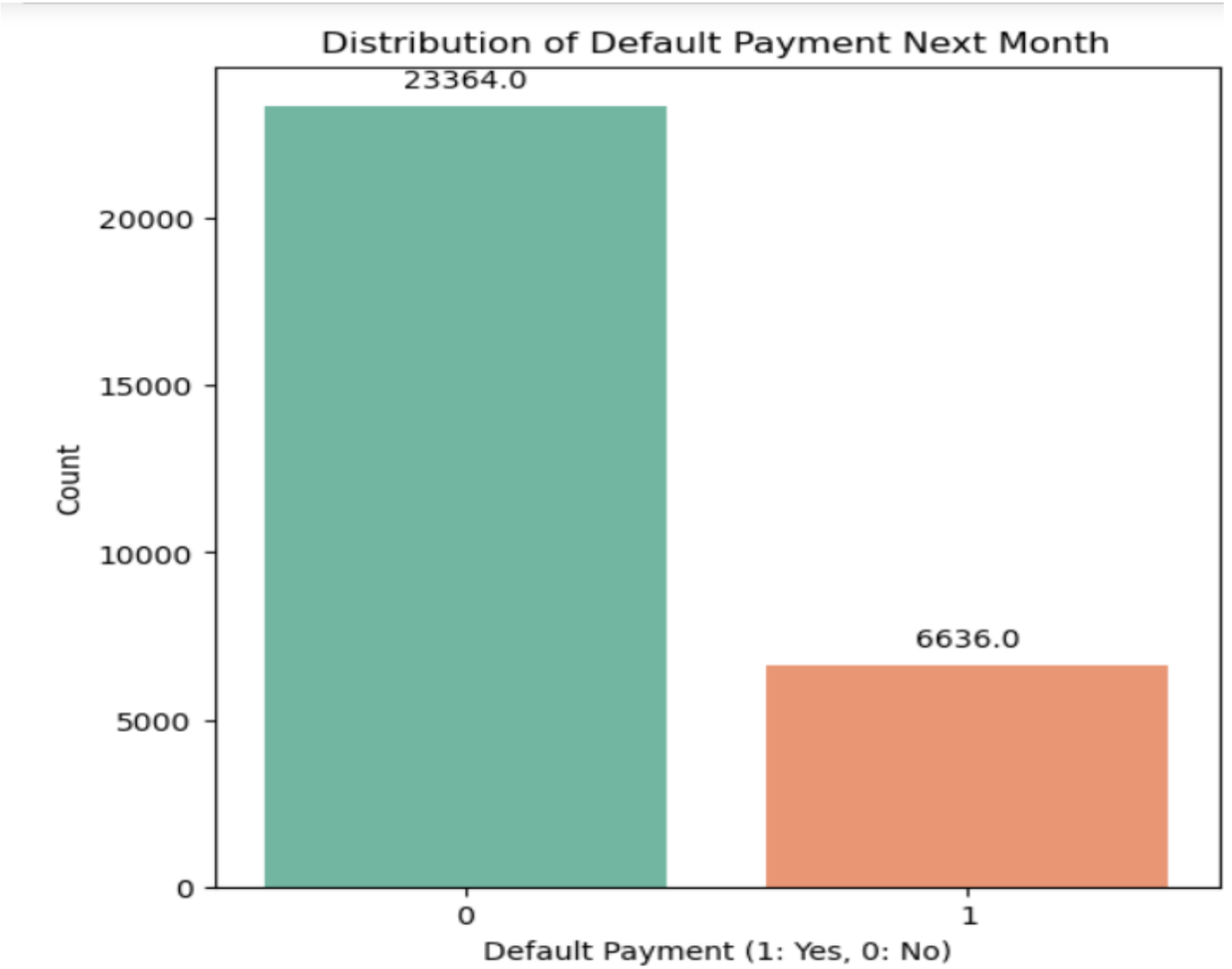
From the above plot analysis we can say that
Maximum amount of given credit in NT dollars is 50-59k followed by 20-29k.

DEFAULTERS W.R.T LIMIT BALANCE

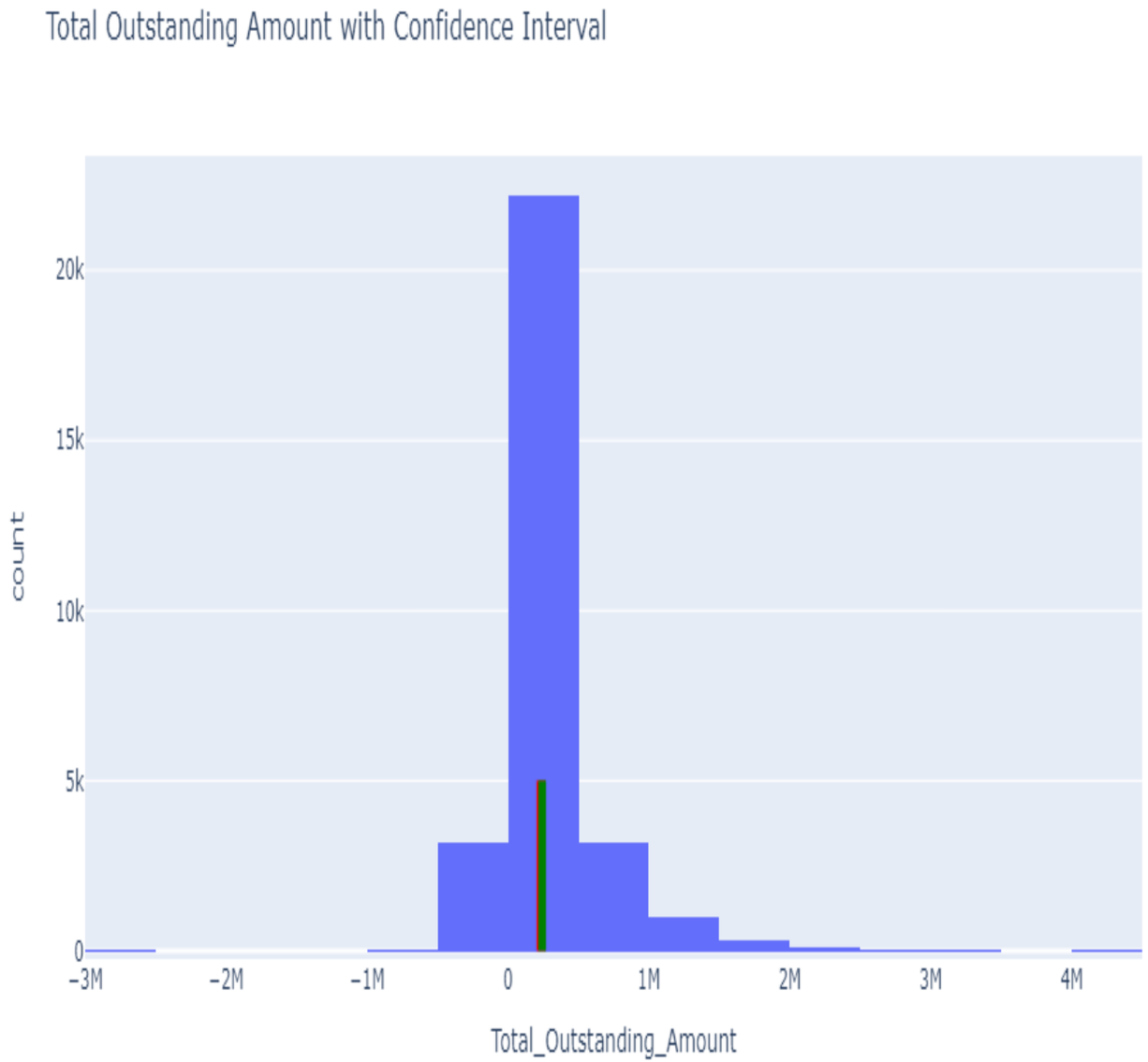


There are more defaulters among the limit balance between 10,000 – 151428.571.

APPLYING PROBABILITY DISTRIBUTION



Here we applied Bernoulli distribution for default payment next month.



CONFIDENCE INTERVAL

- The red & green lines in the histogram indicate (confidence interval lower & upper) the range within which we are 95% confident that the true population mean of the 'Total outstanding Amount lies.
- Confidence Interval for Total Outstanding Amount Mean: (234100.75458933282, 242319.80267733385)

HYPOTHESIS TESTING

HYPOTHESIS TESTING: TWO SAMPLE PROPORTION TEST

```
import statsmodels.api as sm
import pandas as pd
# Create a contingency table
contingency_table = pd.crosstab(data['SEX'], data['default payment next month'])

# Extract the counts for males (assuming 1 corresponds to male)
male_default_count = contingency_table.loc[1, 1]
male_total_count = contingency_table.loc[1].sum()

# Extract the counts for females (assuming 2 corresponds to female)
female_default_count = contingency_table.loc[2, 1]
female_total_count = contingency_table.loc[2].sum()

# Perform two-sample proportion test
z_stat, p_value = sm.stats.proportions_ztest([male_default_count, female_default_count],
                                             [male_total_count, female_total_count],
                                             alternative='larger')

# Print the results
print(f'Z-statistic: {z_stat:.4f}')
print(f'P-value: {p_value:.4f}')

# Check if the p-value is less than the significance level (e.g., 0.05)
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: There is evidence that males are going to default more than females.")
else:
    print("Fail to reject the null hypothesis: There is not enough evidence that males are going to default more than females.")
```

```
Z-statistic: 6.9214
P-value: 0.0000
Reject the null hypothesis: There is evidence that males are going to default more than females.
```

❑ STEPS INVOLVED IN HYPOTHESIS TESTING

1.Formulate Hypotheses:

Null Hypothesis (H0): The proportion of males defaulting on payment next month is equal to or less than the proportion of females.

Alternative Hypothesis (H1): The proportion of males defaulting on payment next month is greater than the proportion of females.

2.Choose Significance Level (α): The code doesn't explicitly set the significance level, but it is common to use α = 0.05.

3.Select the Test Statistic : The code uses a two-sample proportion test (z-test) as implemented in the (sm.stats.proportions_ztest) function from the stats models library.

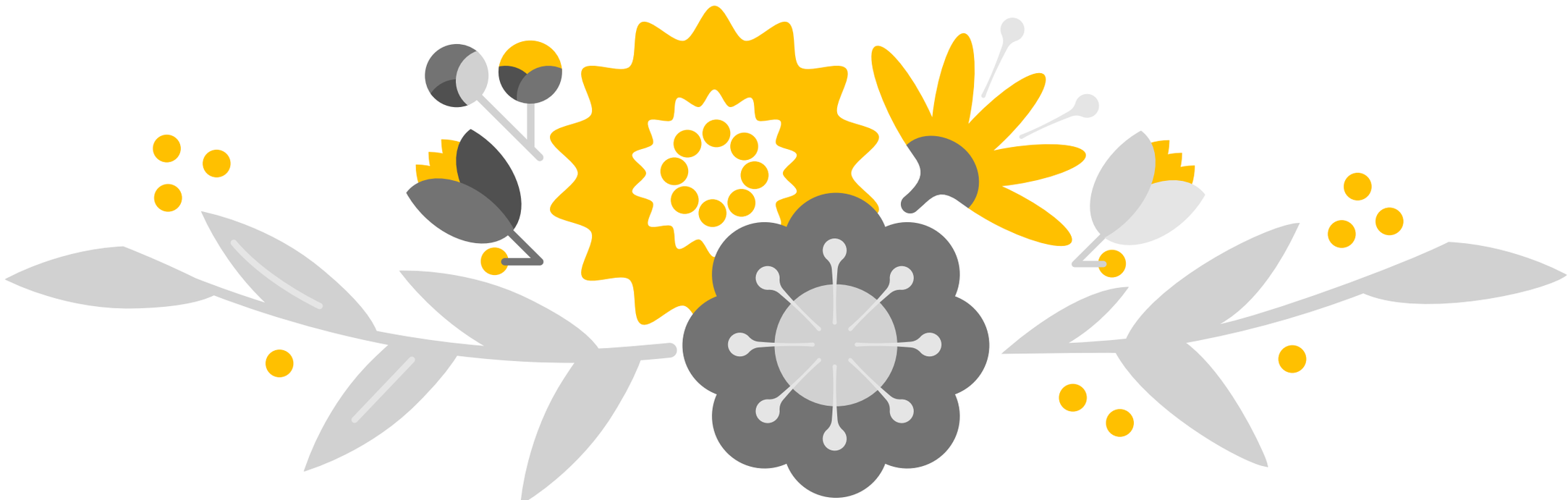
4.Collect and Analyze Data : The code extracts the counts of defaulters and total counts for both males and females from the provided data frame.

5.Calculate the P-value : The sm.stats.proportions_ ztest function calculates the z-statistic and the corresponding p-value.

6.Compare P-value to Significance Level : The code compares the calculated p-value to a hypothetical significance level. If p-value < α, the null hypothesis is rejected.

7.Draw Conclusions : **Reject null hypothesis : There is evidence that males are going to default more than females.**

- ✓ **Which is true ,we have evidenced this result in EDA step where we analyzed the defaulters & non-defaulters among gender.**



THANK YOU!