CREDIT CARD DEFAULT REPORT

Project of Data Analysis And Visualisation(DSC)

Sameep Gupta

Bachelors Of Science(Hons.)

Computer Science
Semester-3rd
Roll Number-22053500018

INTRODUCTION

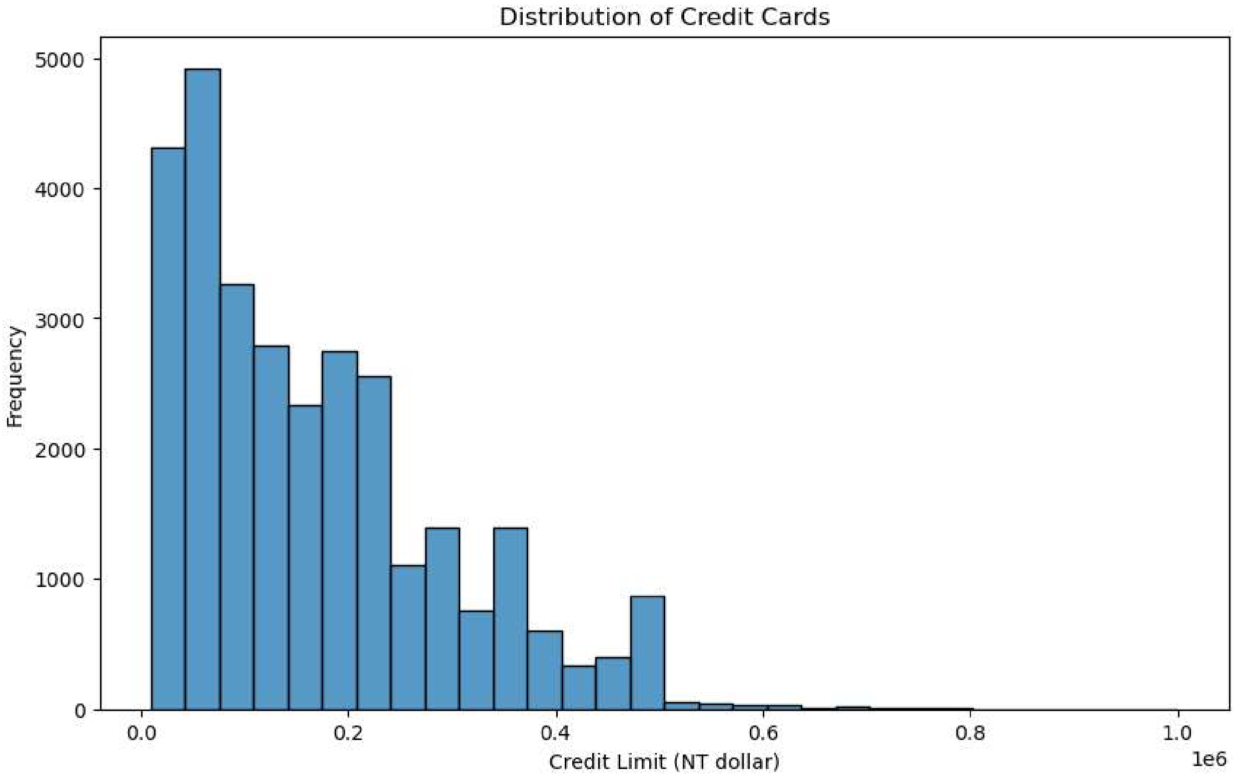
The objective of the project is to analyze the Credit_Card_Default dataset and gain insights into the factors that contribute to credit card default payments. The dataset contains information on demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset also indicates whether the customer defaulted on their payment the following month. The project aims to use this dataset to build predictive models that can help to identify customers who are at risk of defaulting on their payments and take appropriate measures to mitigate the risk. Additionally, the project can be used to gain insights into customer behavior and preferences, which can be used to develop targeted marketing strategies and improve customer satisfaction.

DATASET INFORMATION

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

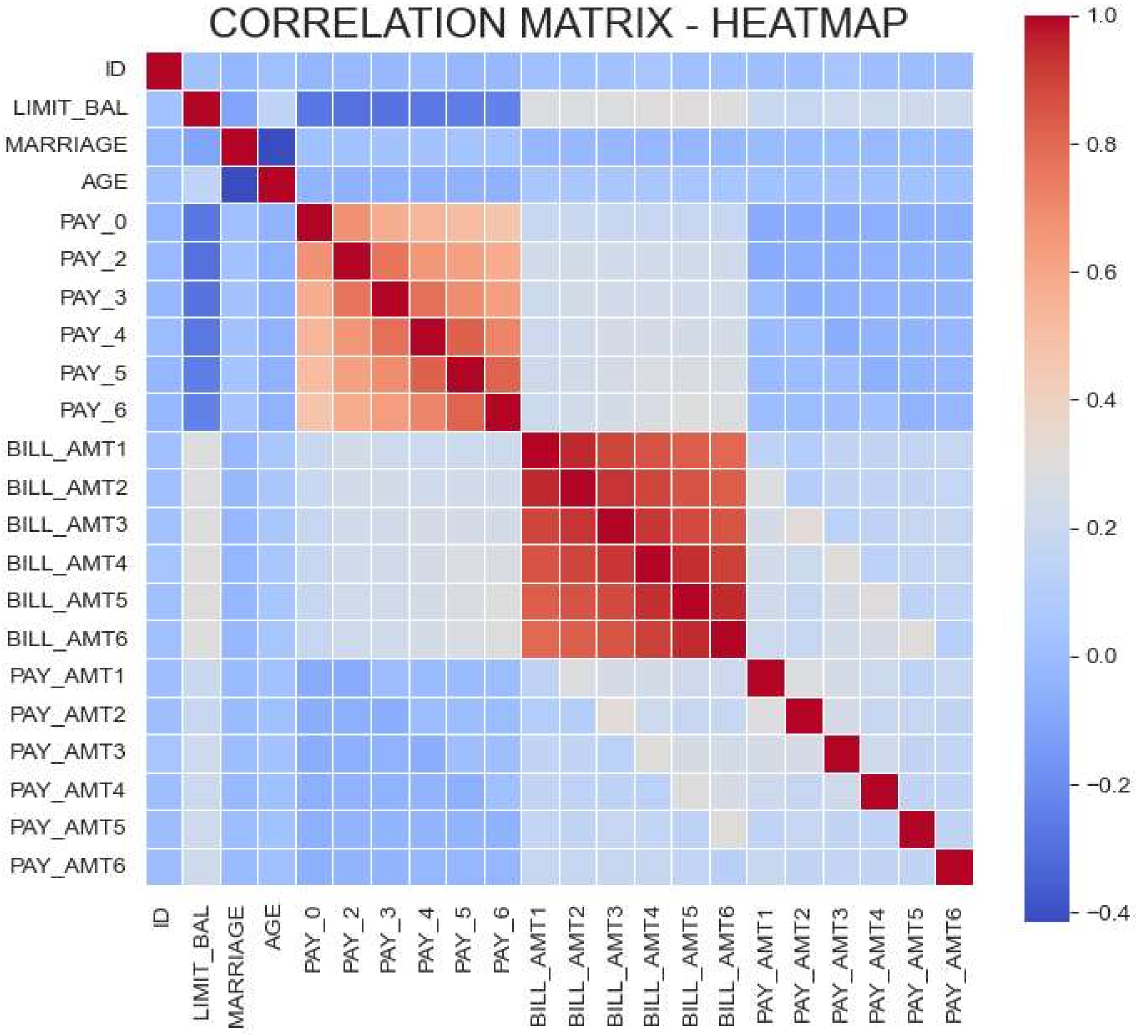
VARIABLES

- **ID**: ID of each client
- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-1=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)
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default.payment.next.month**: Default payment (1=yes, 0=no)



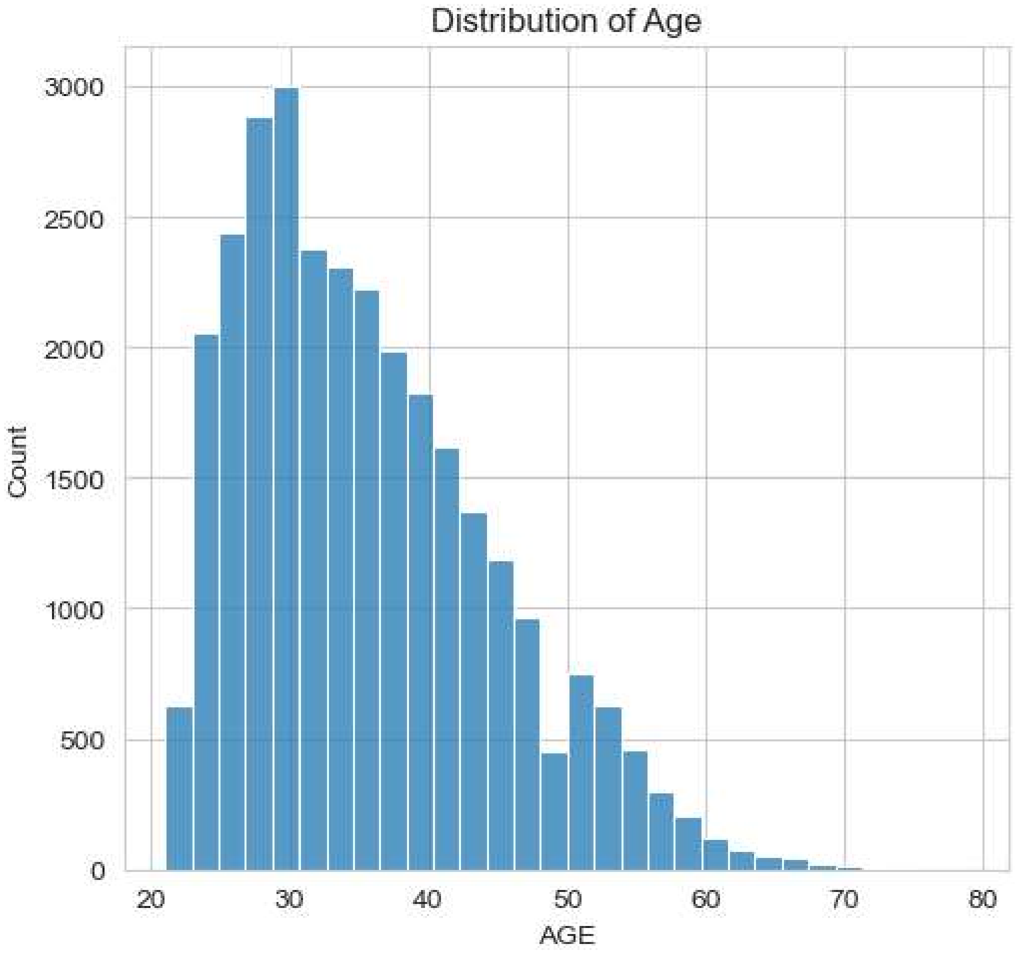
Based on the plot, we can observe the following:

- The distribution of credit card limits is not normal, as it has a skewed shape, with a long tail on the right side.
- There are some outliers in the data, as a few customers have credit limits significantly higher than the majority of the customers.
- The plot can be used to identify customers with potentially high credit risk, as they are likely to have credit limits lower than the average.



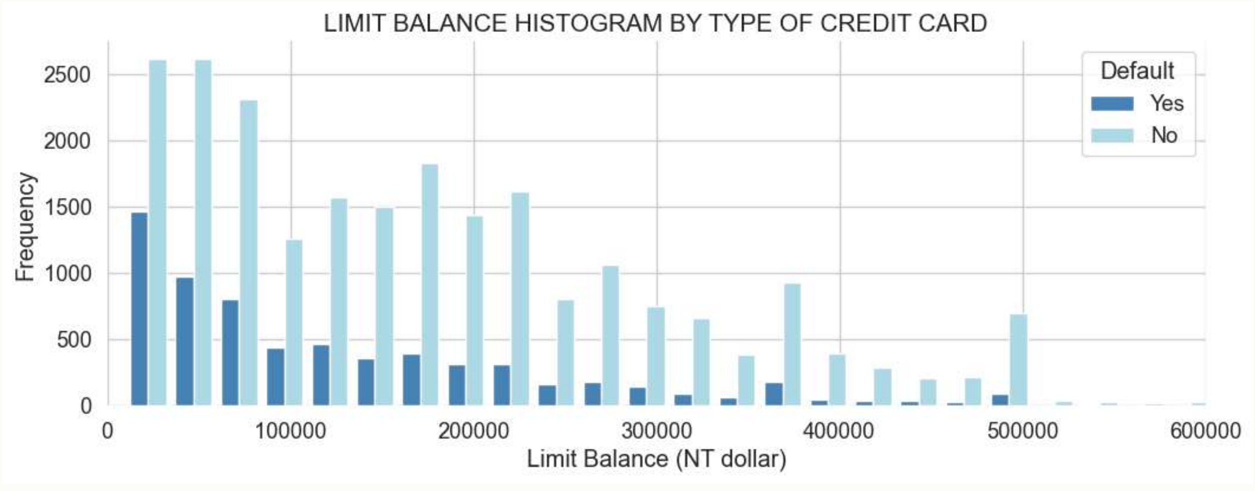
The heatmap is useful for visualizing the correlation between different features in the dataset. The color intensity represents the strength and direction of the correlation, with cooler colors indicating negative correlations and warmer colors indicating positive correlations. The annotations on the heatmap provide the exact values of the correlation coefficients, allowing for a more detailed analysis of the relationships between the features

.



1. The distribution of credit limits (`LIMIT_BAL`) varies across the dataset, with some individuals having higher credit limits and others having lower limits. This variation could be due to factors such as income, credit history, and other financial factors.
2. The distribution of the default payment next month (`default.payment.next.month`) shows that a significant number of individuals have a default payment next month. This could indicate that a large portion of the dataset consists of individuals who have not met their payment obligations.

Key observation: People around the age of 30 are high in credit-seekers count.



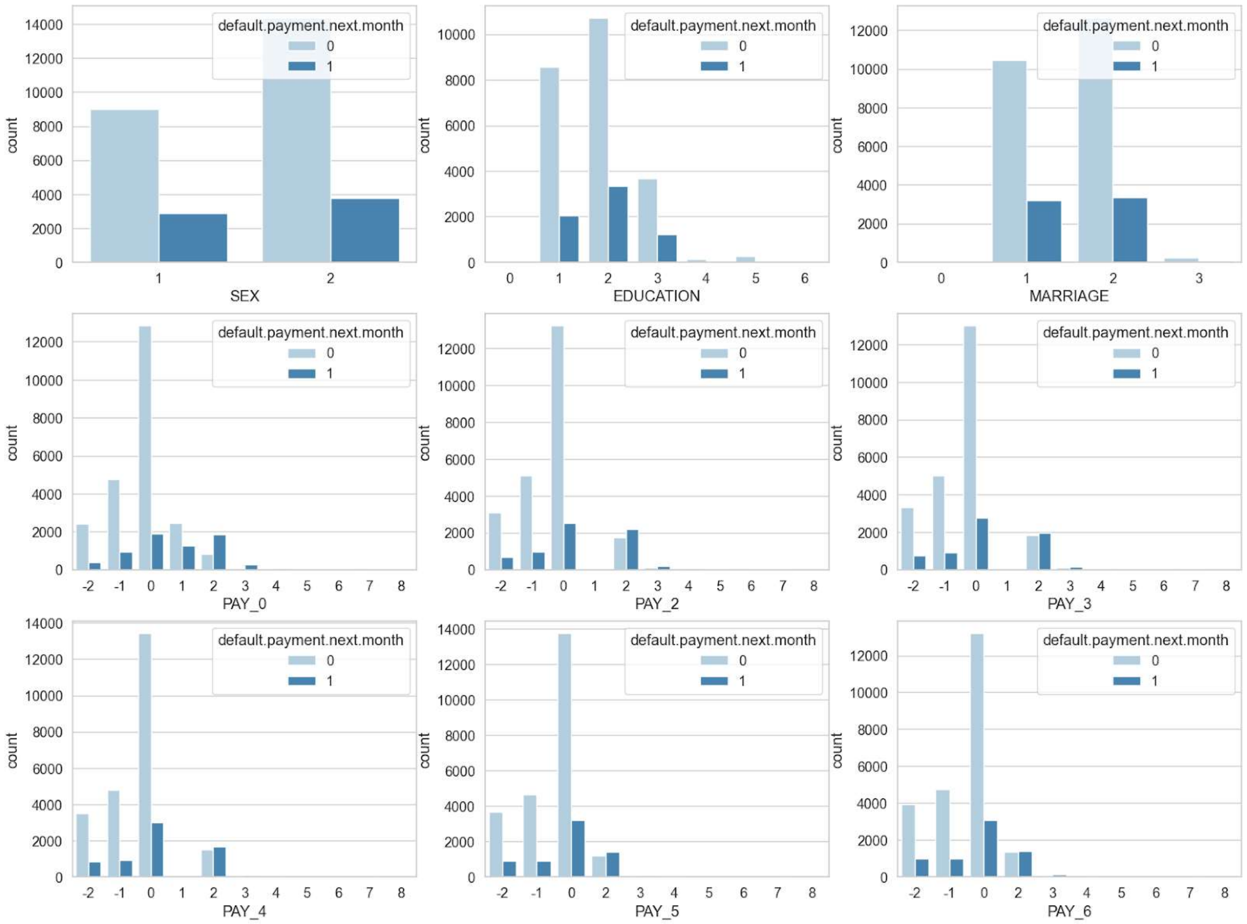
The histogram provides a visual comparison of the distribution of credit limits for two groups: those who defaulted on their payment next month (labeled as 'Yes') and those who did not default (labeled as 'No'). The x-axis represents the credit limit in New Taiwan Dollars, and the y-axis represents the frequency of individuals with that credit limit.

Key observations from the histogram:

- The distribution of credit limits for those who defaulted and those who did not default appears to be different.
- The 'No' group (those who did not default) shows a more uniform distribution across credit limits.
- The 'Yes' group (those who defaulted) seems to have a higher frequency of lower credit limits compared to the 'No' group.

This analysis provides an initial understanding of how credit limits are distributed among those who defaulted and those who did not. It suggests that lower credit limits may be associated with a higher likelihood of default, but further statistical analysis would be needed to confirm this observation.

FREQUENCY OF CATEGORICAL VARIABLES (BY TARGET)

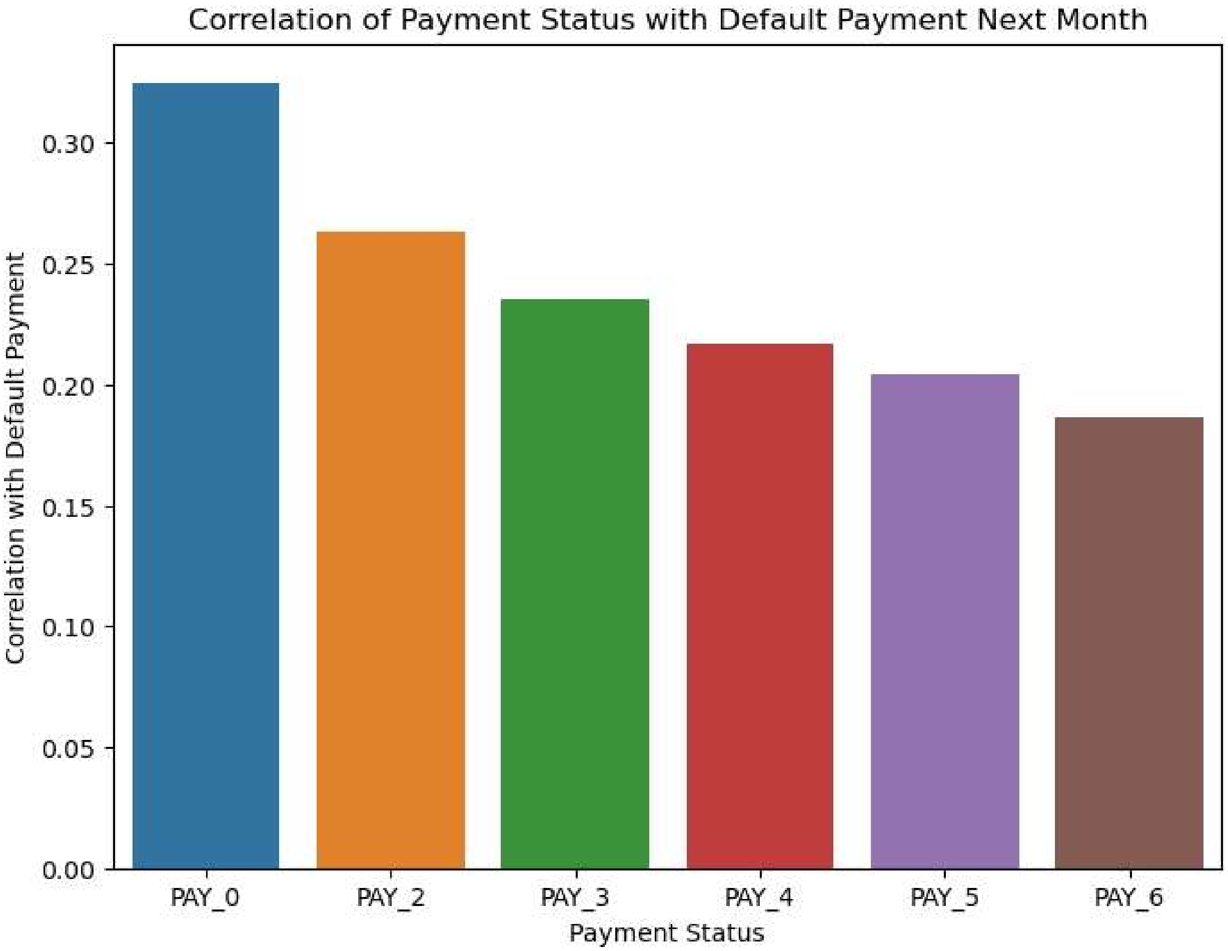


Key observations from the count plots:

- Sex (Gender): The first plot (top-left) compares the frequency of default and non-default cases for each gender. It provides insights into whether gender is associated with a higher likelihood of default.
- Education: The second plot (top-center) compares the frequency of default and non-default cases across different education levels. This can help in understanding the relationship between education and default behavior.
- Marriage: The third plot (top-right) compares the frequency of default and non-default cases for different marital status categories. It provides

insights into the impact of marital status on default behavior.

- Payment Status (PAY_0 to PAY_6): The remaining plots in the grid compare the frequency of default and non-default cases for the payment status across the previous six months. This can help in understanding the relationship between payment behavior and the likelihood of default.



The provided code calculates the correlation between PAY_0 to PAY_6 and the default payment next month for a dataset related to credit card payments. The analysis is based on the following aspects:

1. Correlation Matrix: The code computes the correlation matrix for the payment status columns and the default payment next month column. It then extracts the correlations with the default payment next month and plots the correlations using a bar plot.
2. Correlation Values: The code prints the correlation values between the payment status columns and the default payment next month column.
3. Analysis: Based on the correlation values, it is possible to observe the strength and direction of the relationship between the payment status columns and the default payment next month column. A positive correlation indicates that as one variable increases, the other variable also

increases, while a negative correlation indicates that as one variable increases, the other variable decreases. The correlation values can help in understanding the patterns and trends in the data, which can be useful for further analysis or decision-making.

1.