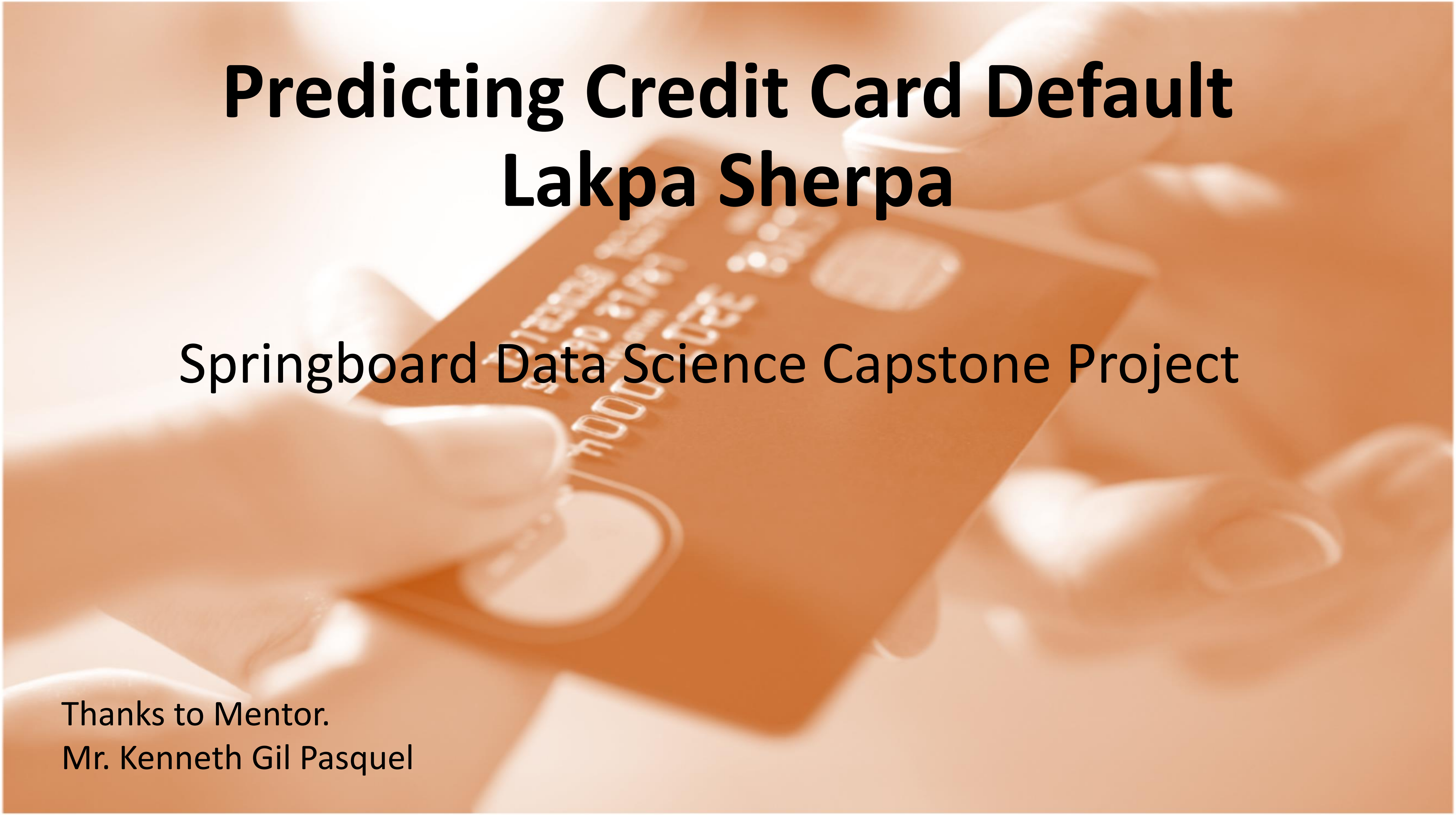


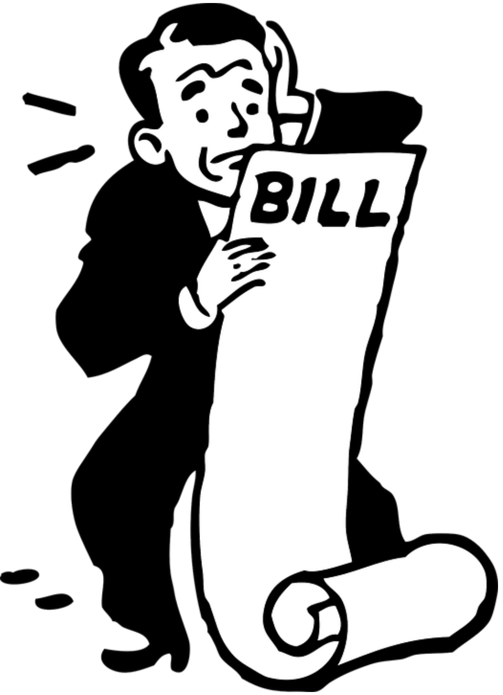
Predicting Credit Card Default
Lakpa Sherpa

Springboard Data Science Capstone Project

Thanks to Mentor.
Mr. Kenneth Gil Pasquel

Introduction

- Credit Card default can be very costly.
- Clients also suffer.
- Precaution: predict if the client will default on a credit card loan.



What factors affect credit card default?
Can we predict the likelihood if a client will default?

Who might care?

Banks and Financial Institutions.

AIA
American International Assurance Company, Limited
AFFINBANK
ALLIANCE BANK
AmBank Group
Bumiputra-Commerce
BANKRAKYAT
Bank Islam
Bank Muamalat
Bank Simpanan Nasional
citibank
EON BANK
Hong Leong
HSBC
ING INSURANCE
Maybank
MBSB
Malaysia Building Society Berhad
A Subsidiary of EPF
OCBC Bank
PUBLIC BANK
大众银行
RHB BANK
Scotiabank
Standard Chartered
SOUTHERN Bank Berhad
UNITED OVERSEAS BANK · MALAYSIA
direct access
ABN·AMRO Bank
Bank of America.
中国银行马来西亚
BANK OF CHINA (MALAYSIA) BERHAD (511251-V)
Cagamas Berhad
THE NATIONAL MORTGAGE CORPORATION
BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA
Employees PROVIDENT FUND
POST BY HOMELINK
LHDN
LEMBAGA HASIL DALAM NEGERI

Data Information

- Data acquired from UCI Machine Learning Repository
- Clients information: 30,000 credit card clients of Taiwanese Bank
- Number of fields: 24

What Factor might affect Default Payment?

- Limit Balance
- Gender
- Education
- Marital Status
- Age
- History of past payment

Data Wrangling

- Acquiring and Cleaning or preparing Data
- Imported Data
- Renamed Columns
- Dealing with the Outliers
- Dealing with Bad Data

- Approach 1:

I updated and modified bad data and outliers.

- Approach 2:

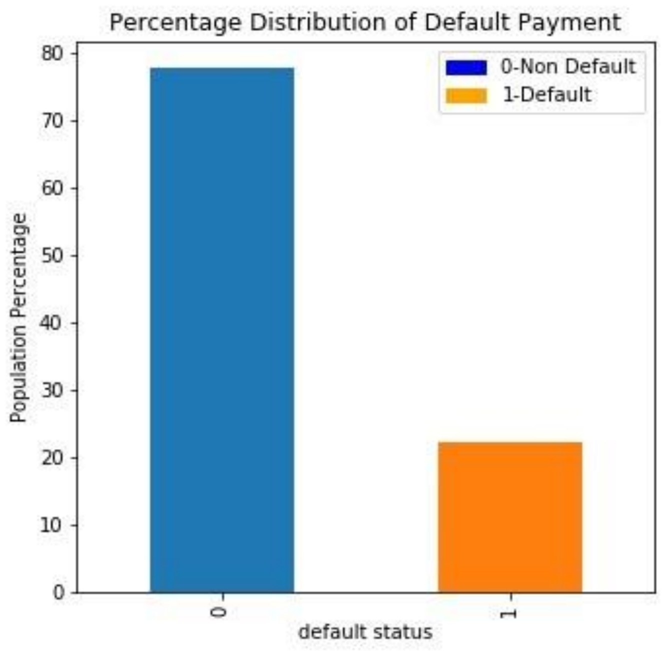
I dropped bad data and outliers.

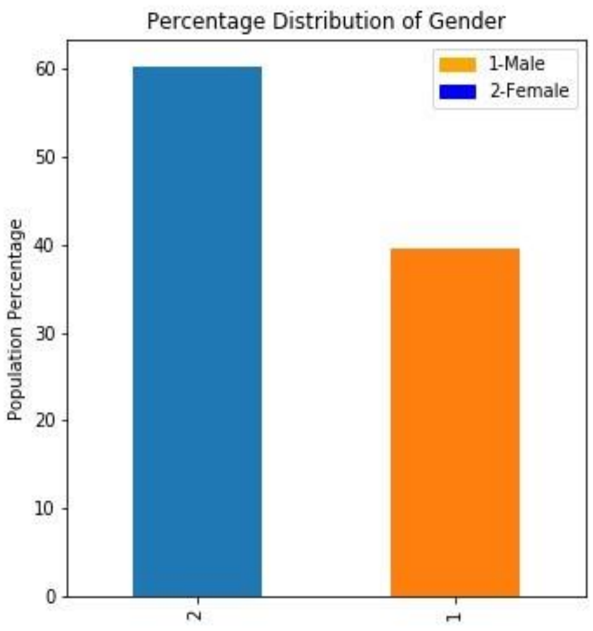
- Approach 3:

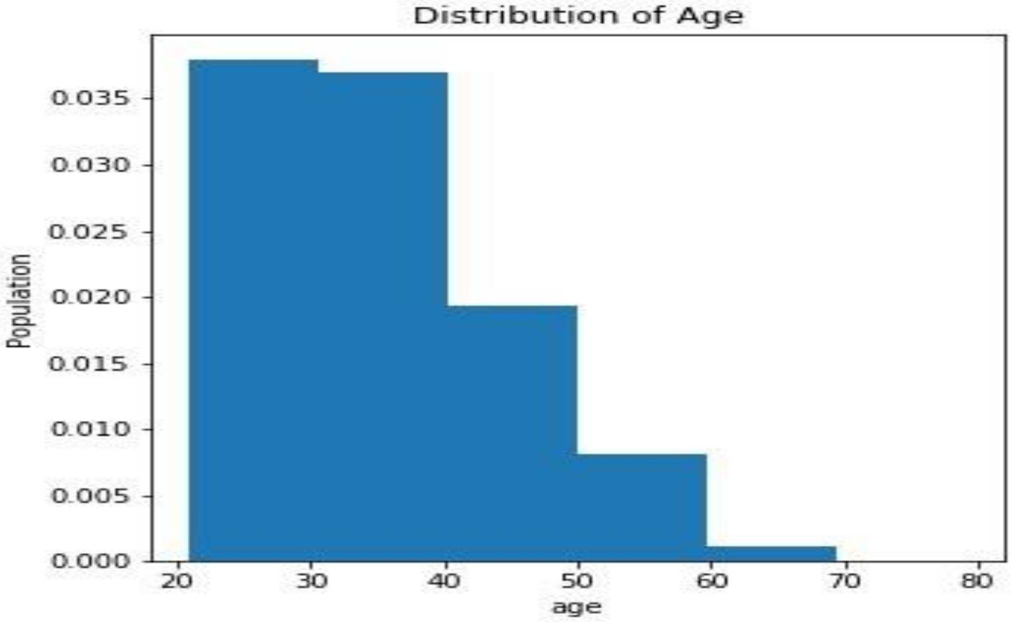
I replaced bad data with mode value and outliers with median value.

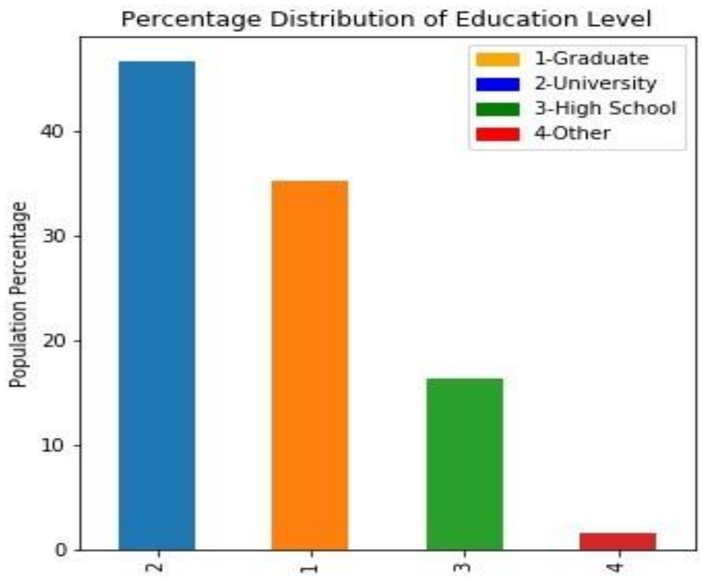
Data Story Approach 1

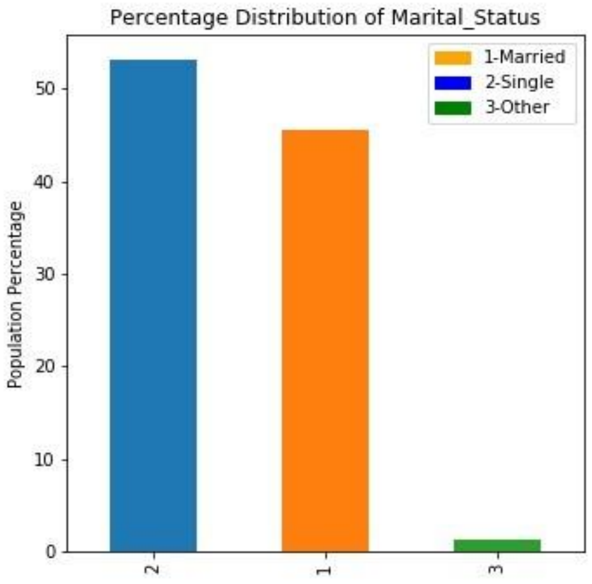
- About 77% of card holder did not default and 22.12% of card holder defaults
- We have 60% female clients and 40% male clients.
- About 35% with Graduate level education, 46% with University level, 16% with High School level, 1.5% with Others level.
- About 53% married, 45% single and rest others.
- Most of the clients are of age group 20 to 40.





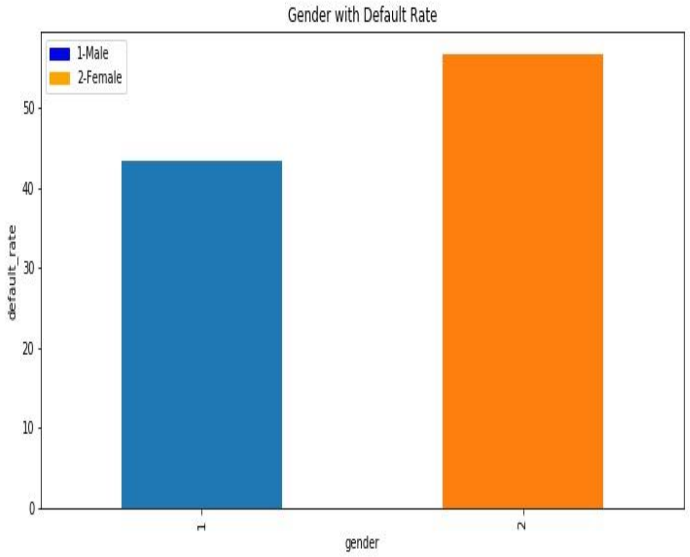


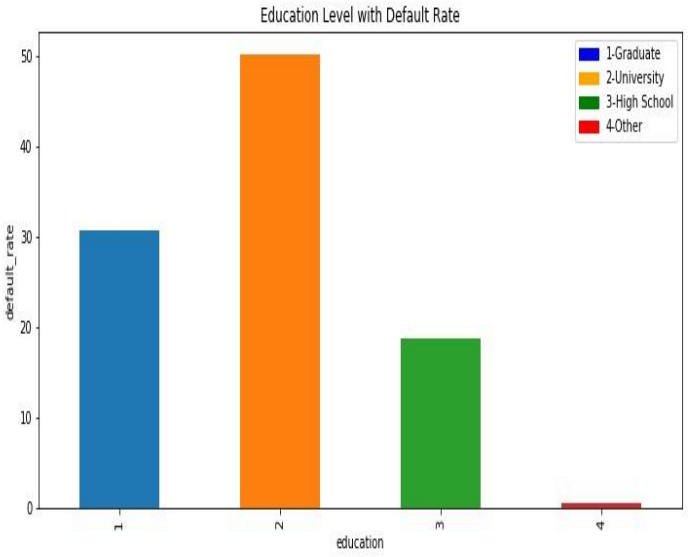


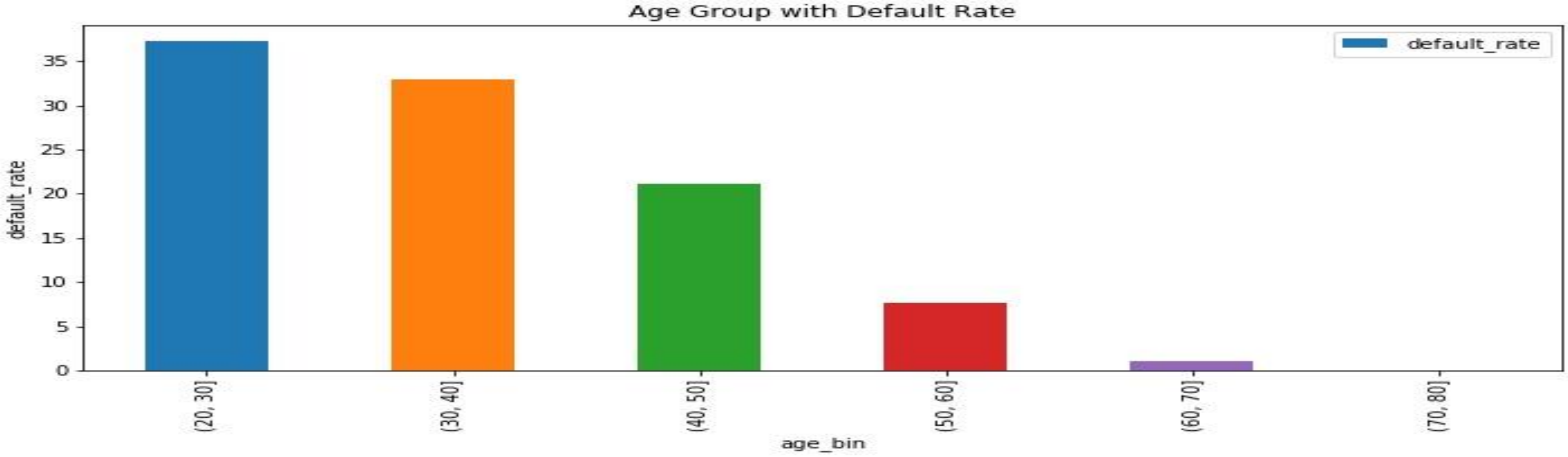


Who defaults more?

- More female defaults than male clients.
- Card holder with University as highest level of education default more than other education level.
- Age group 20-30 defaults more than other age groups.





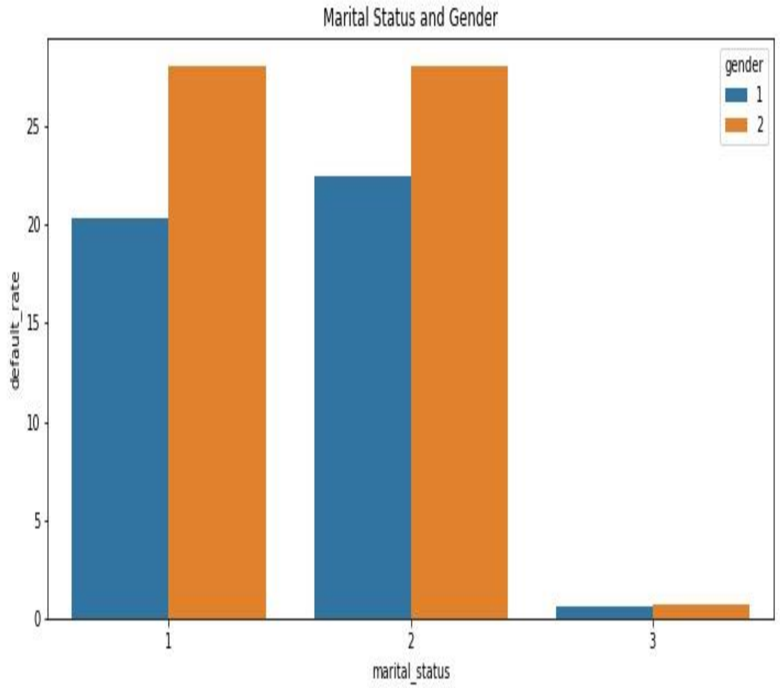


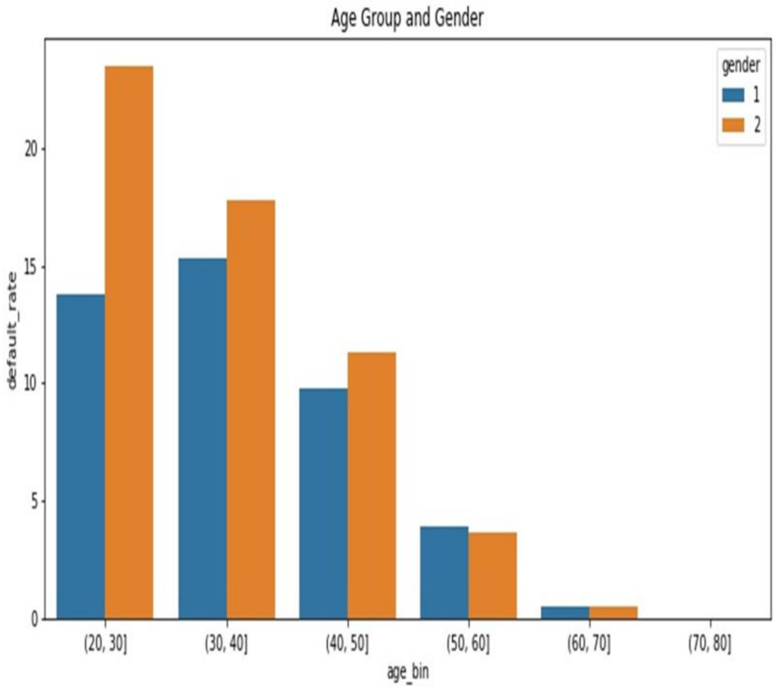
Grouping Gender with other features

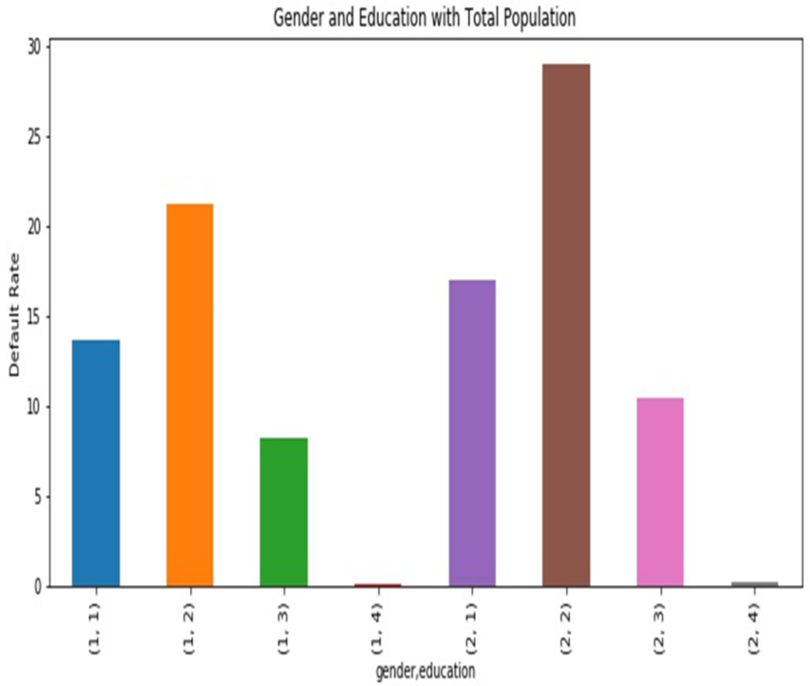
More Female default than male for any marital status.

More Female default than male for all age group expect for 50-60.

Most of the defaulter are female with university level education.





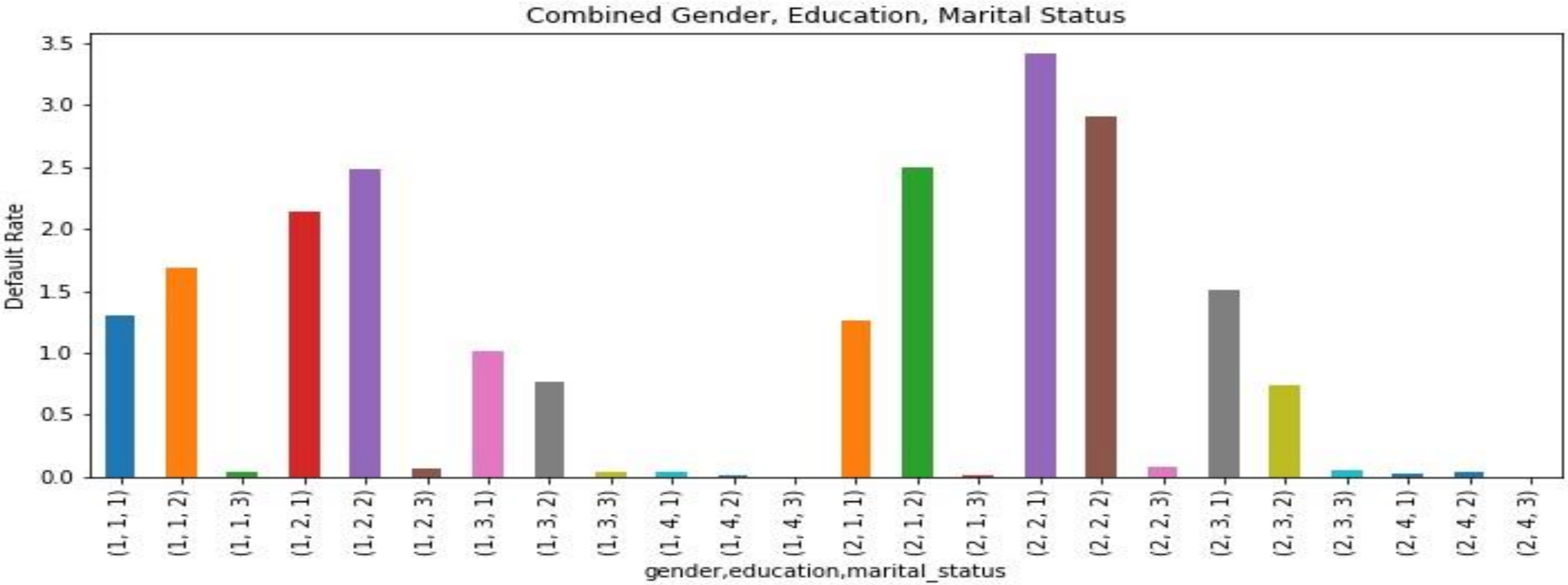


Group Gender, Education and Marital status together.

(2,2,1) = Female, University level, and Married

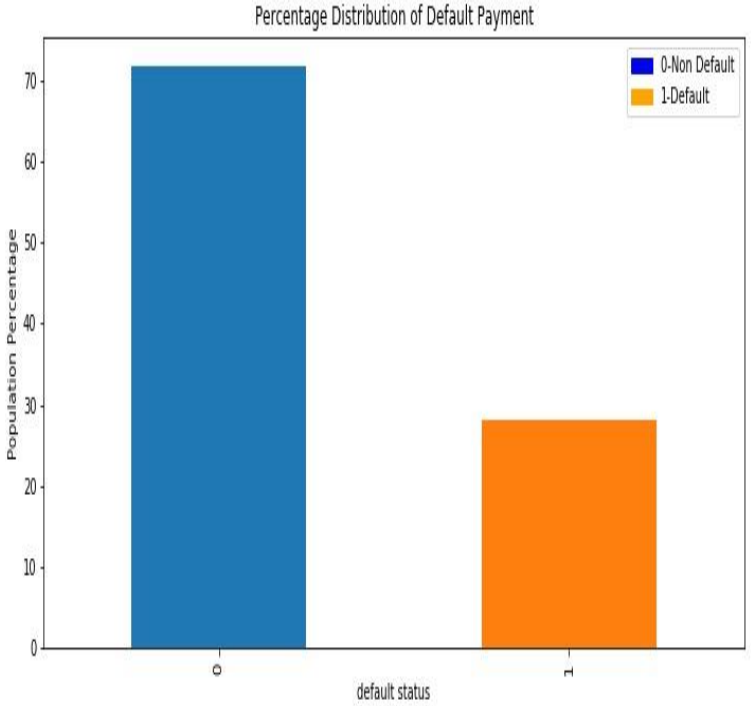
(2,2,2) = Female, University level, and Single

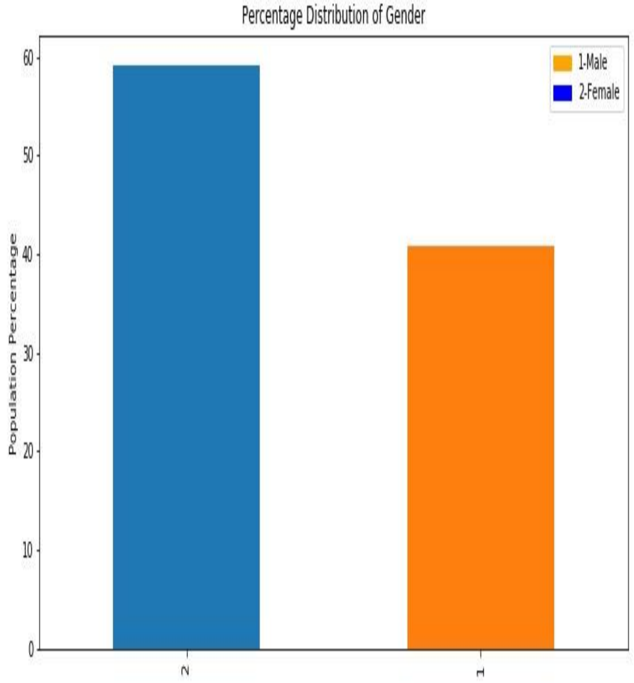
(2,1,2) = Female, Graduate level, and single

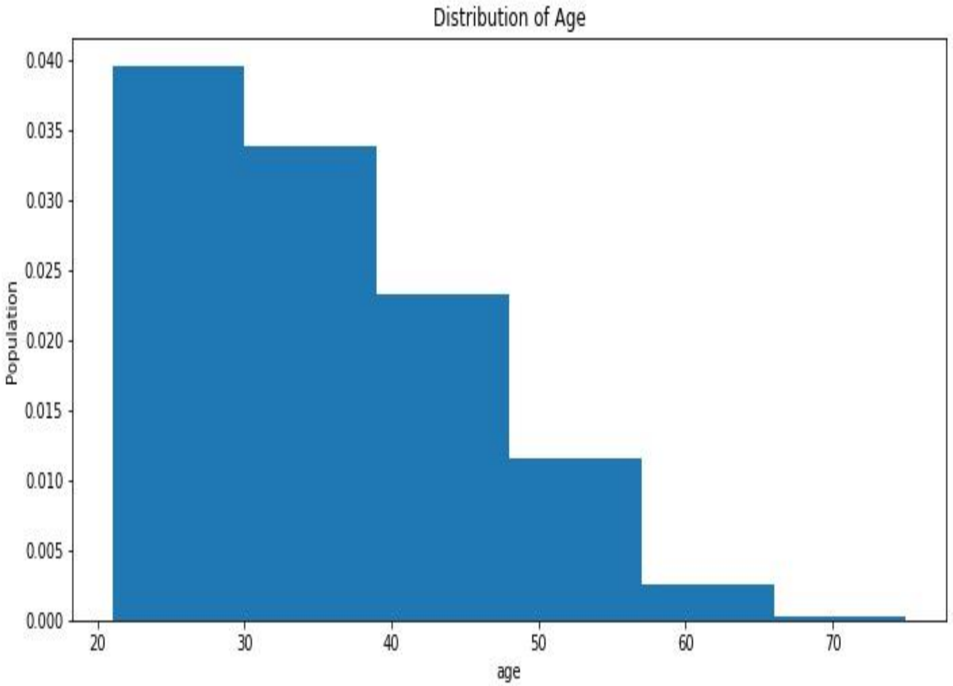


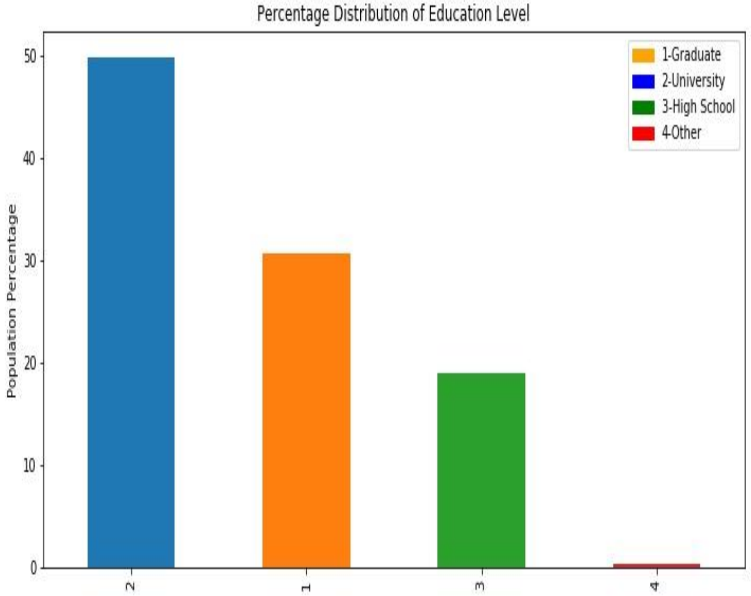
Data Story Approach 2

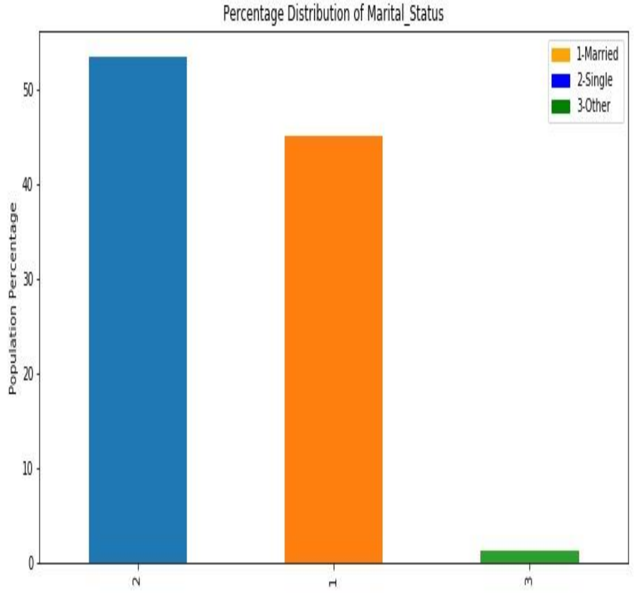
- We found that about 72% of the cardholders(10,315) did not default and about 28% of the cardholders (4,042) default.
- We have 60% female and 40% male clients.
- We have 30% with Graduate level education, 49% with University level, 19% with High School level, and 0.33% with Others as level of education.
- We have 53% married, 45% single, and rest as others.
- Most of our cardholders are of age group 20 to 40.


Percentage Distribution of Default Payment


Percentage Distribution of Gender

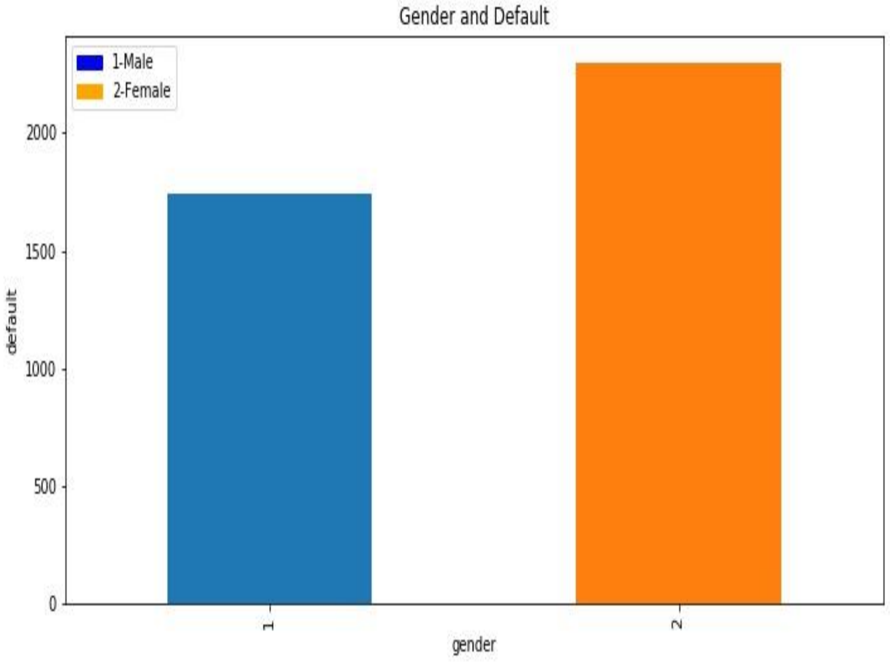

Distribution of Age

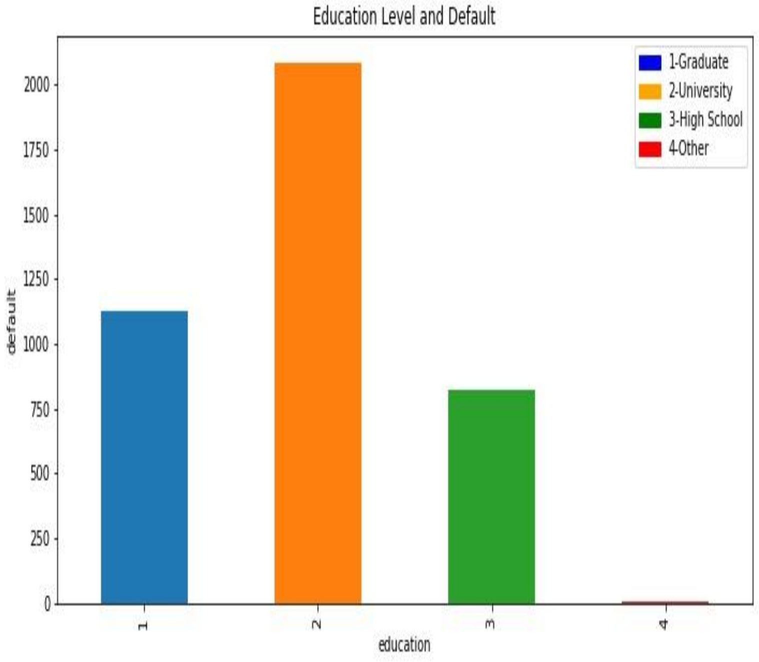

Percentage Distribution of Education Level

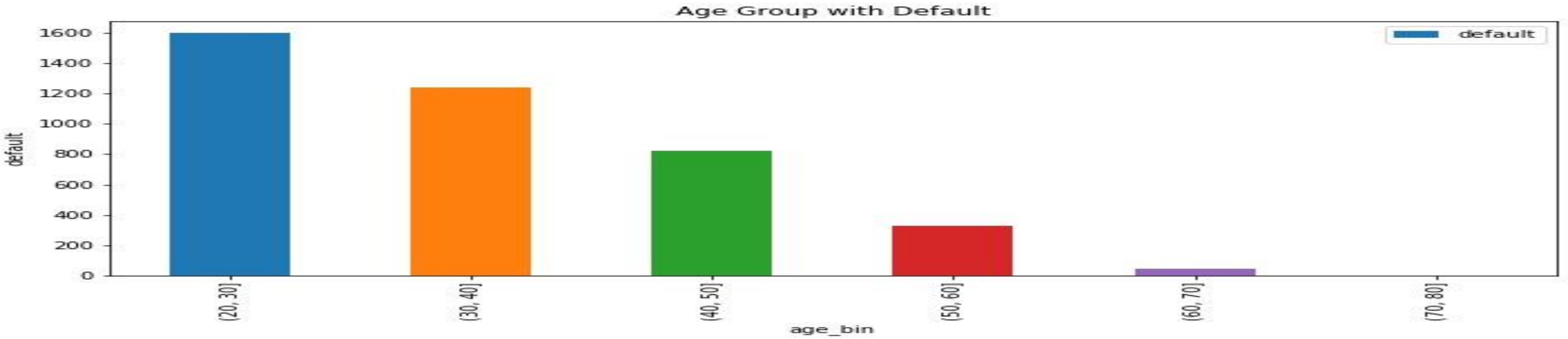

Percentage Distribution of Marital_Status

Who defaults more?

- More female defaults than male clients.
- Card holder with University as highest level of education default more than other education level.
- Age group 20-30 defaults more than other age groups.


Gender and Default


Education Level and Default

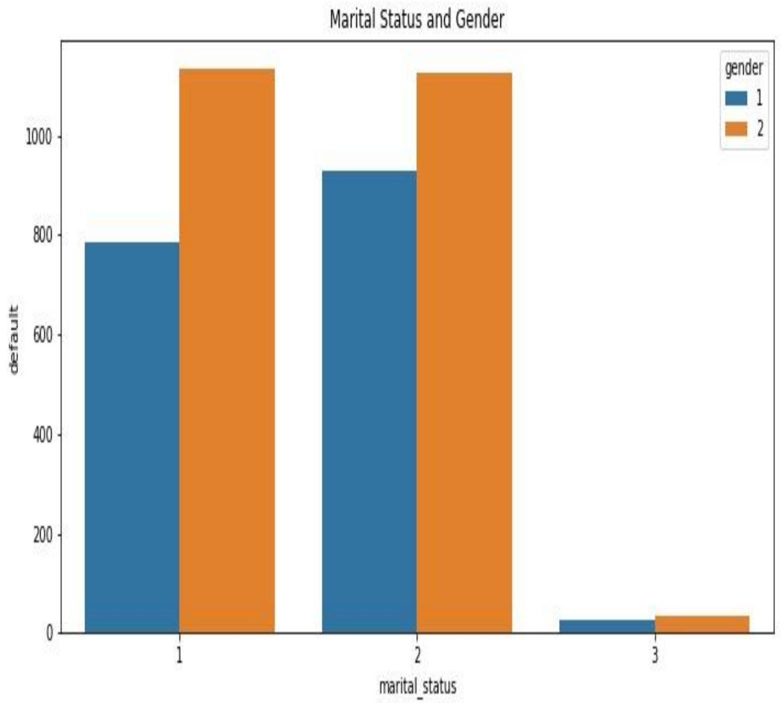

Age Group with Default

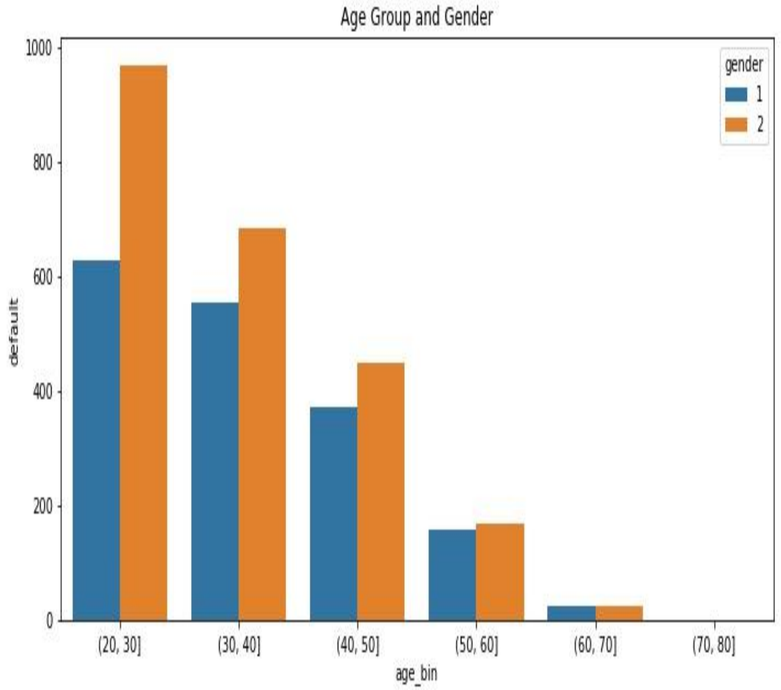
Grouping Gender with other features

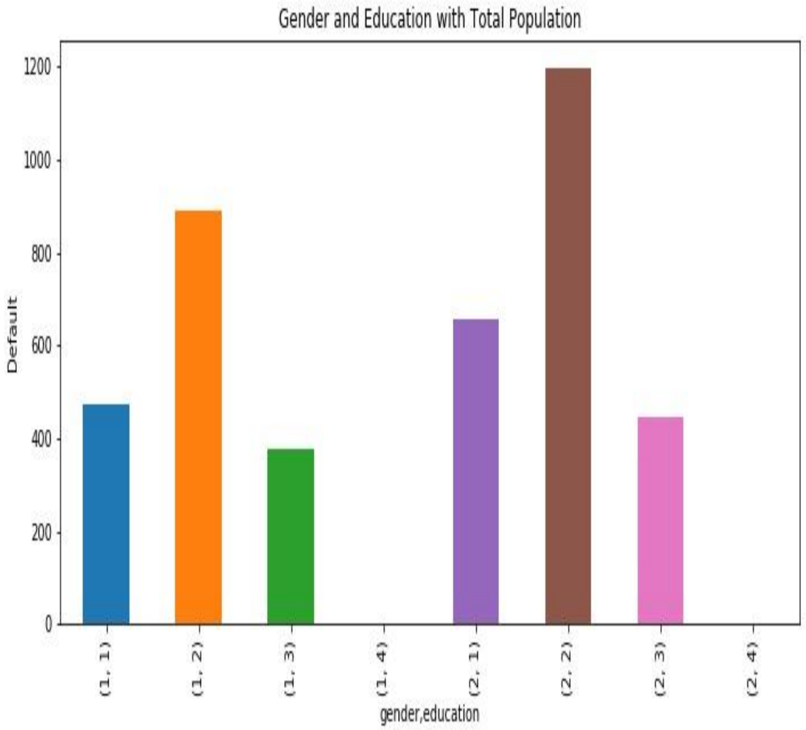
More Female default than male for any marital status.

More Female default than male for all age group expect for 50-60.

Most of the defaulter are female with university level education.



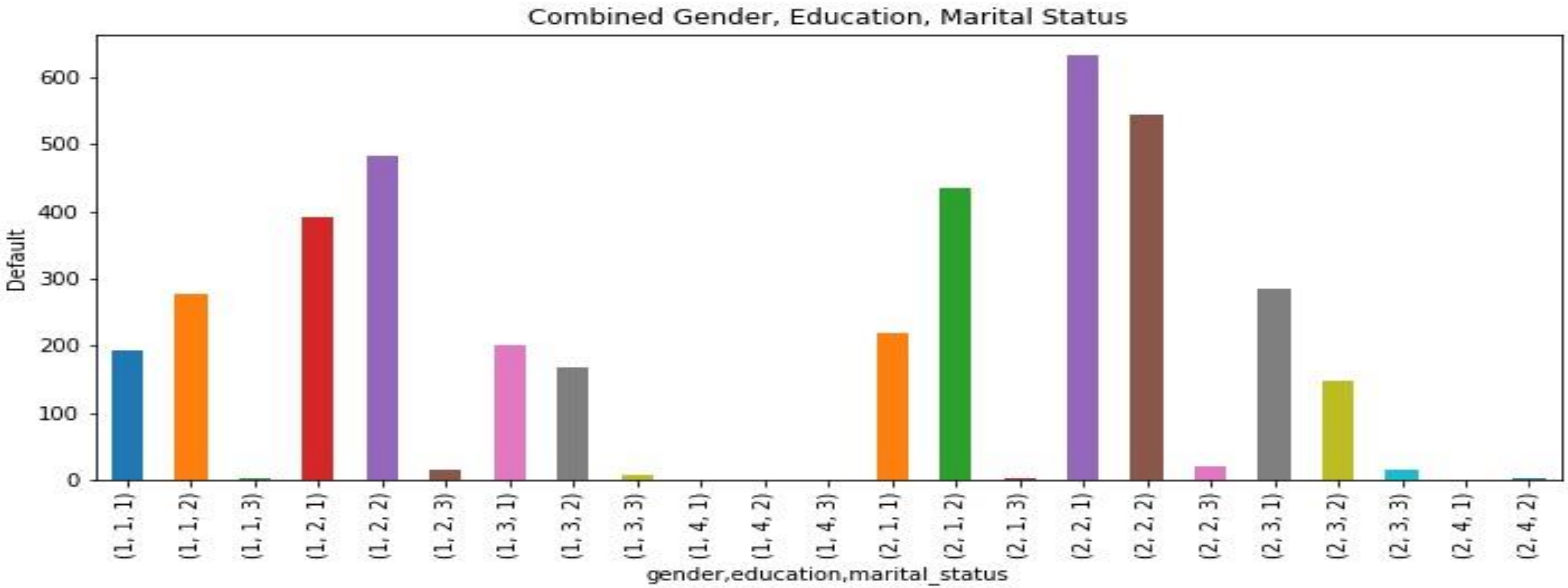




Group Gender, Education and Marital status together.

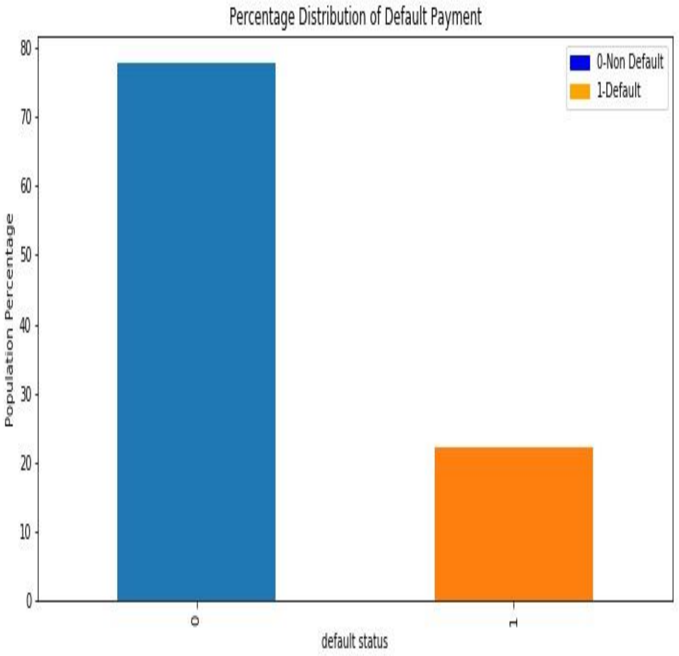
(2,2,1) = Female, University level, and Married

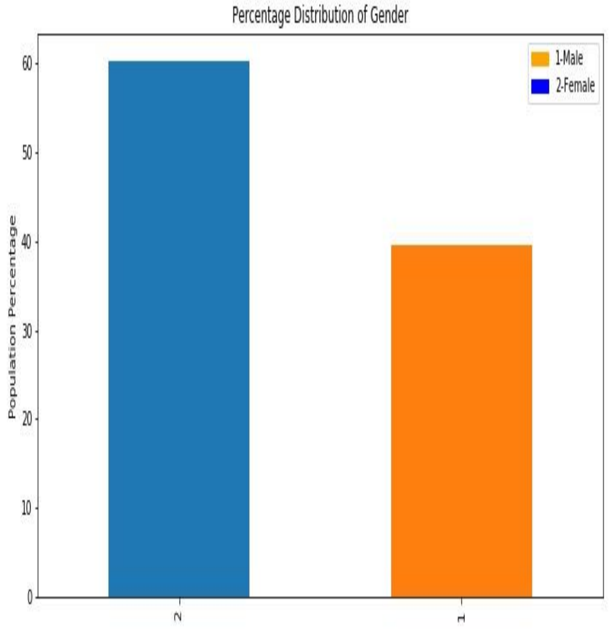
(2,2,2) = Female, University level, and Single

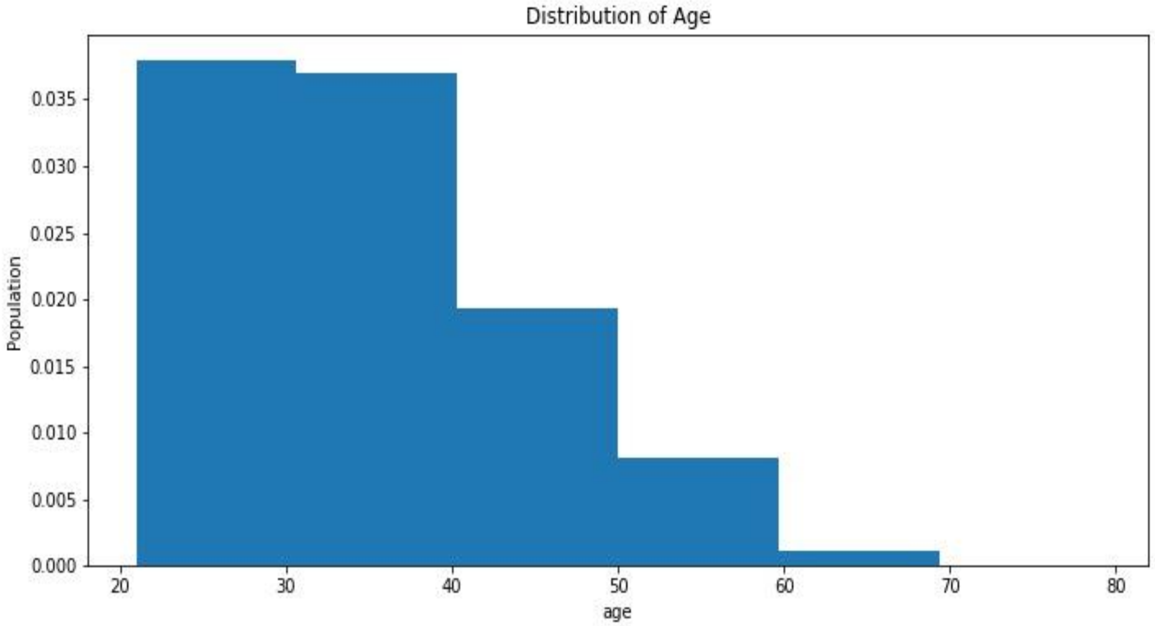


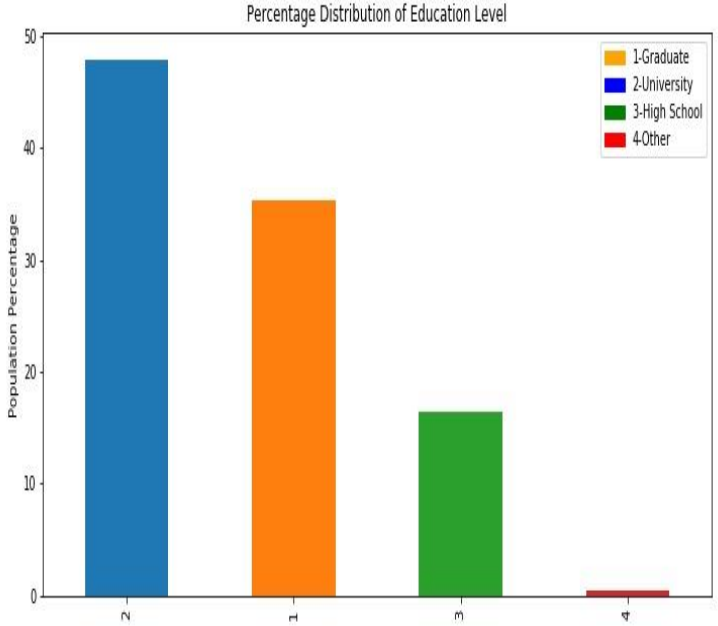
Data Story Approach 3

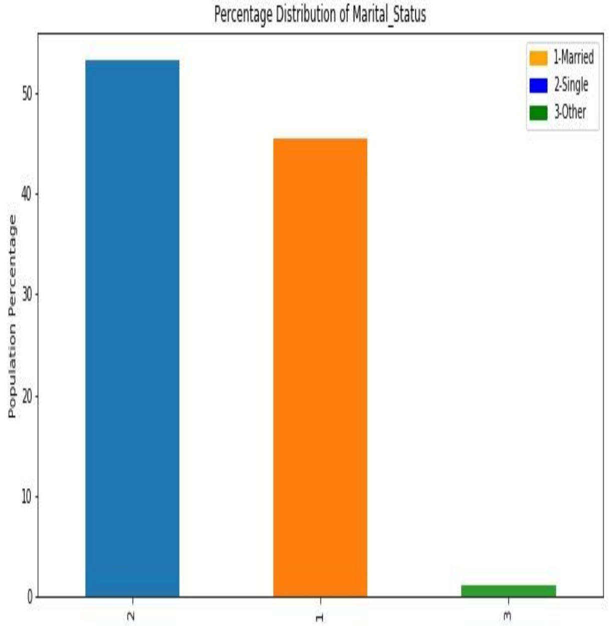
- About 77% of card holder did not default and 22.12% of card holder defaults
- We have 60% female clients and 40% male clients.
- About 35% with Graduate level education, 47% with University level, 16% with High School level, 0.4% with Others level.
- About 53% married, 45% single and rest others.
- Most of the clients are of age group 20 to 40.





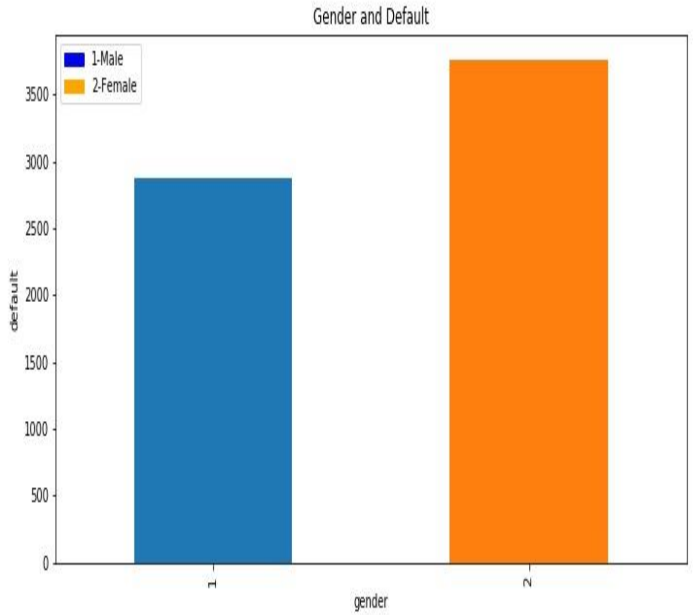


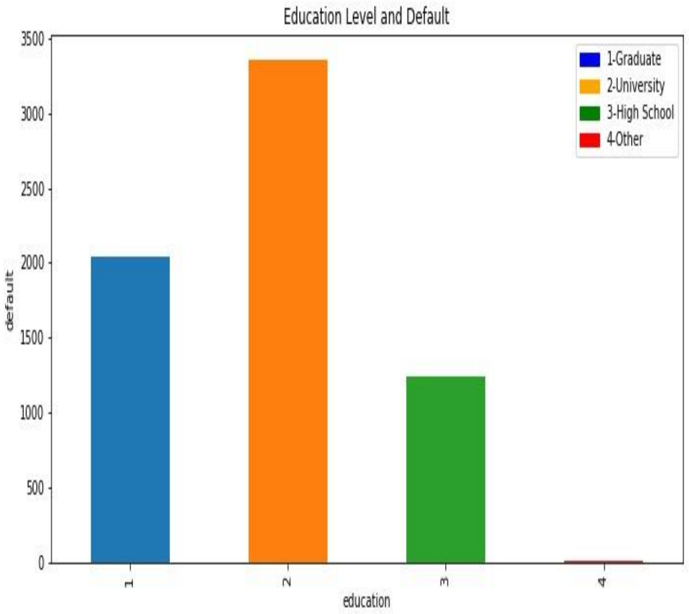


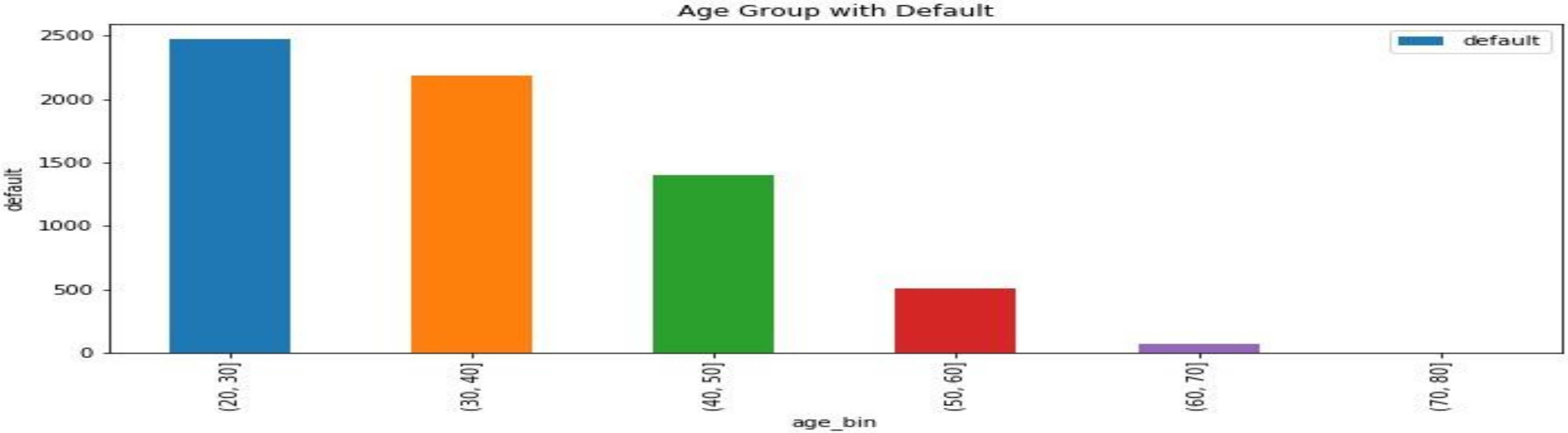


Who defaults more?

- More female defaults than male clients.
- Card holder with University as highest level of education default more than other education level.
- Age group 20-30 defaults more than other age groups.

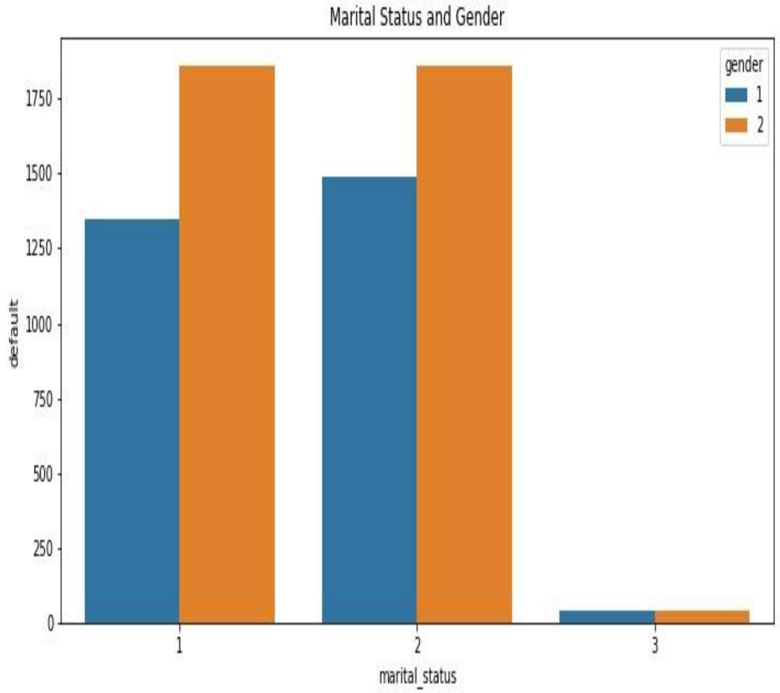




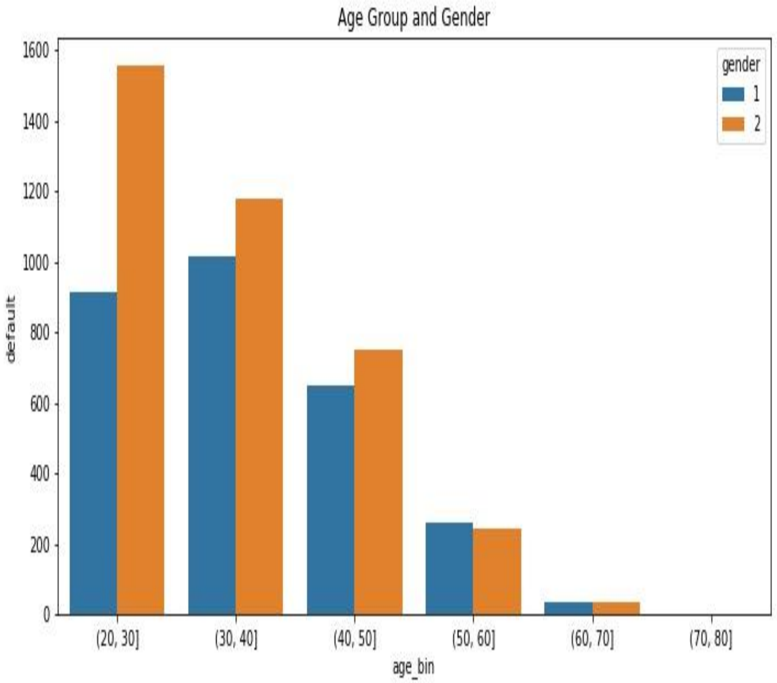


Grouping Gender with other features

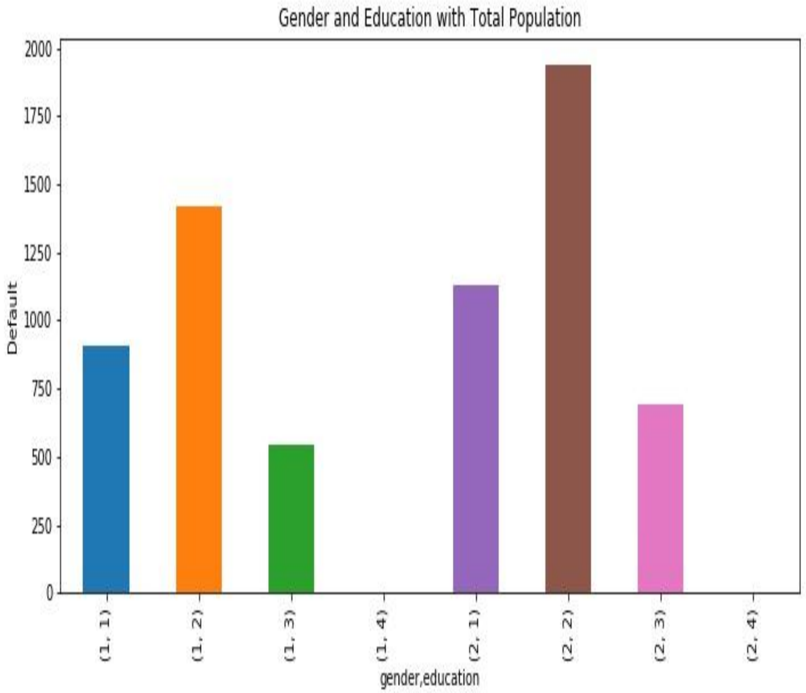
More Female default than male for any marital status.



More Female default than male for all age group expect for 50-60.



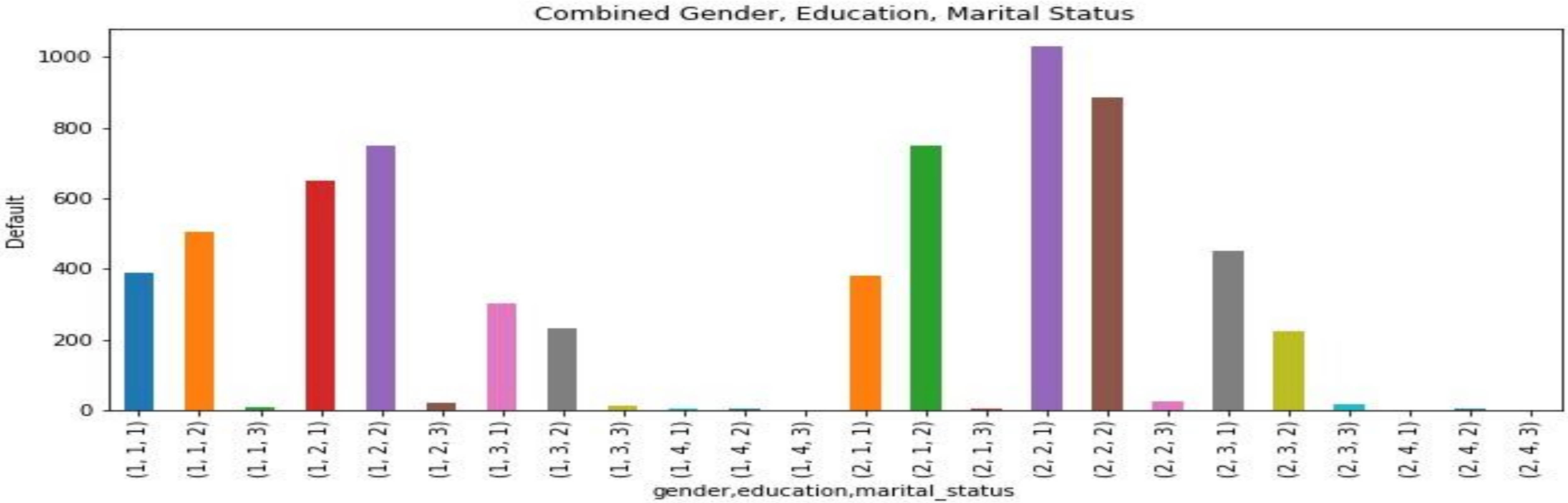
Most of the defaulter are female with university level education.



Group Gender, Education and Marital status together.

(2,2,1) = Female, University level, and Married

(2,2,2) = Female, University level, and Single



Inferential Statistics

- Hypothesis testing
- Used t-Statistics and bootstrap method
- Chi-squared Test

Machine Learning

- Logistic Regression
- Decision Tree
- Gaussian Naïve Bayes Classifier
- Random Forest Classifier
- Support Vector Machine

Build model

- Create Dummy Variables
- Create X input features and y labels
- Create training and testing data

Handle Class Imbalance

- Balancing the Class weight
- Resampling the dataset
- Ensemble method

Compare Handling Class Imbalance

- Balancing Class Weight
- Random Under sampling
- NearMiss1
- NearMiss2
- Edited Nearest Neighbors
- Repeated Edited Nearest Neighbors
- Tomek Links
- Random Over Sampling
- SMOTE
- SMOTETOMEK
- SMOTEENN

Approach 1:
According to the F1 Score of the classification model, the SMOTE+ENN method to handle the class imbalance performed the best. So we will use SMOTE+ENN to handle class imbalance, then compare different models with each other.

Approach 2:
According to the F1 Score of the classification model, the Edited Nearest Neighbors method to handle the class imbalance performed the best. So we will use ENN to handle class imbalance, then compare different models with each other.

Approach 3:
According to the F1 Score of the classification model, the Edited Nearest Neighbors method to handle the class imbalance performed the best. So we will use ENN to handle class imbalance, then compare different models with each other.

Comparing Different Classification Model

- Logistic Regression Classification
- Decision Tree Classification
- GaussianNB
- Random Forest Classification
- Support Vector Machine

Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default. Let's try to tune all our model to see if we can improve the performance and then decide which classification model would be the best.

Hyperparameter Tuning for model Performance

Approach 1:
After tuning the hyperparameter of all different model, we can see that the f1 score for Decision Tree had decreased after hyperparameter tuning, and the f1 score for Random Forest has improved and increased. Logistic Regression, GaussianNB, and SVM has not changed even after tuning.
Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default.

Approach 2:
Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

Approach 3:
Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

Conclusion:

Random Forest Classification

Recommendation:

I would like to recommend **Approach 2** Random Forest Classification model to predict credit card default because it is better than random guess and performs better than other classification models.

Thank You!!
Lakpa Sherpa