Exploring Categorical Quantities

This lesson will focus on how to explore relationships between different categorical variables in the dataset with examples.

WE'LL COVER THE FOLLOWING

Exploratory Data Analysis is all about exploring relationships in the dataset that might be hidden or might not be easy to spot just by looking at the dataset. We will try to explore these kinds of relationships in the Default of Credit Card Clients Dataset. We will use the cleaned version of the dataset from the lesson Inconsistent Data. The details of individual columns are mentioned below.

```
# Default of Credit Card Clients Dataset
# There are 25 variables:

# ID: ID of each client
# LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementa
# GENDER: Gender (male,female)
# EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others)
# MARRIAGE: Marital status (married, single, others)
# AGE: Age in years
# PAY_1: Repayment status in September, 2005 (0=pay duly, 1=payment delay for one month, 2=pa
# 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 (yes,no)
```

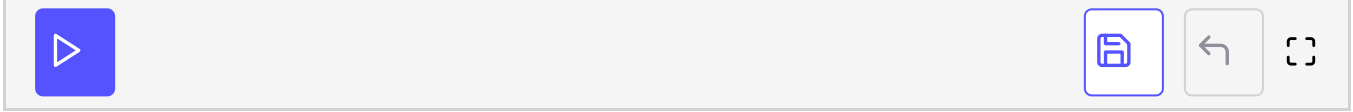
More specifically, we are interested in finding out how the variable `default.payment.next.month` is affected by other variables.

Grouping

As we saw in Chapter 3 of this course, grouping data can give us very useful insights. Let's see how the categorical variables `GENDER`, `EDUCATION`, and `MARRIAGE` are related to `default.payment.next.month`.

`GENDER`

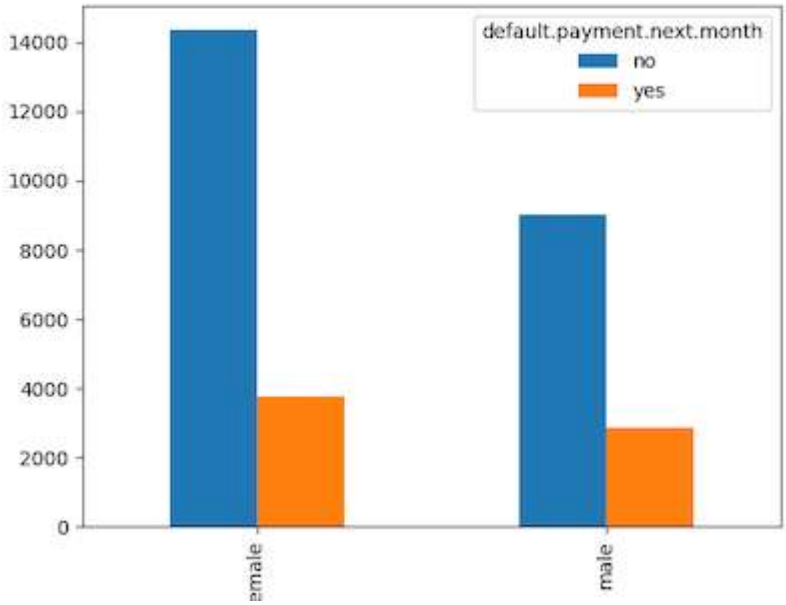
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('credit_card_cleaned.csv')

# Group data
grouped_df = df.groupby(['GENDER','default.payment.next.month']).size()
grouped_df = grouped_df.unstack()
print(grouped_df)

# Plot
grouped_df.plot(kind='bar')

# Calculate probabilities
grouped_df['prob_default'] = grouped_df['yes'] / (grouped_df['no'] + grouped_df['yes'])

print('\n\n',grouped_df[['prob_default']])
```



We group the data by `EDUCATION` and `default.payment.next.month` on **line 6** and use the function `size` to retrieve the number of males and females. We then use the function `unstack` in the next line. The function `unstack` performs two steps here:

- It changes the table into a dataframe
- It names the columns `no` and `yes`, the two categories of the variable `default.payment.next.month`.

We can see the resultant dataframe in the output of **line 8**. We plot the dataframe `grouped_df` in **line 11**. We see in the produced bar plot the number of males and females for each category of `default.payment.next.month`.



A natural question that arises after looking at the bar plot is that out of females and males, which category is more likely to default the next month since the number of females and males in our dataset is not equal? To answer this, we can calculate the probabilities of each gender defaulting the next month. We can calculate the probability of a male defaulting by dividing the number of males defaulting by the total number of males. We can do the same for females. Therefore, we divide the column `yes` with the sum of both `yes` and `no`. We save the probabilities in a new column in the dataframe and name the column `prob_default` in **line 14**. From the output of **line 16**, we see that the:

- probability of a *female* defaulting is approximately 0.20
- probability of a *male* defaulting is approximately 0.24

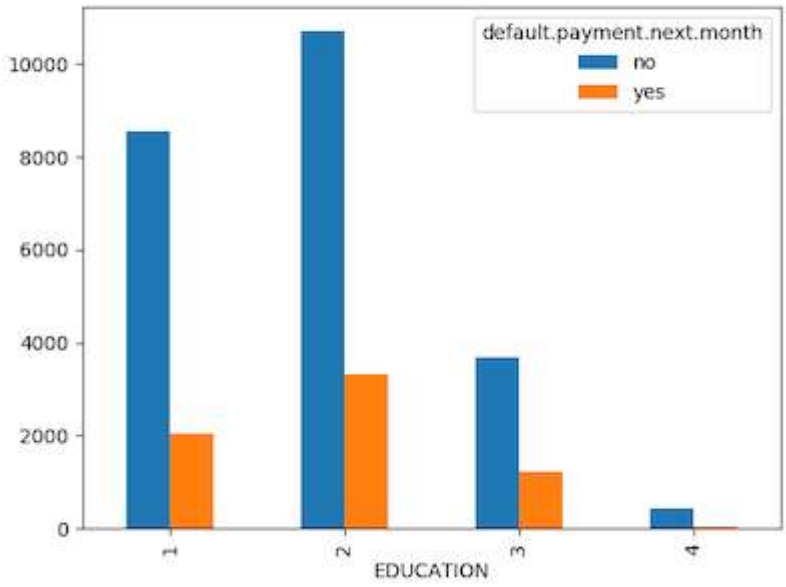
This means that a male is more likely to default according to this dataset.

`EDUCATION`

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('credit_card_cleaned.csv')
```

```
# Group Data
grouped_df = df.groupby(['EDUCATION','default.payment.next.month']).size()
grouped_df = grouped_df.unstack()
print(grouped_df)

# Plot
grouped_df.plot(kind='bar')

# Calculate probabilities
grouped_df['prob_default'] = grouped_df['yes'] / (grouped_df['no'] + grouped_df['yes'])

print('\n\n',grouped_df[['prob_default']])
```

We have written the same code that we did above except that we have replaced `GENDER` with `EDUCATION` in **line 6**. We get a bar plot in **line 11** in which we have counts of people in each category of education. The colors indicate whether or not they defaulted.



We calculate the probability of a person defaulting in each category of education by using the same formula that we used above for calculating the probabilities of males and females defaulting. We find out that the:

- probability of a postgraduate defaulting is approximately 0.19.
- probability of a university graduate defaulting is approximately 0.23.
- probability of a high school graduate defaulting is approximately 0.25.

This gives us a general trend in the data that as people get more educated they

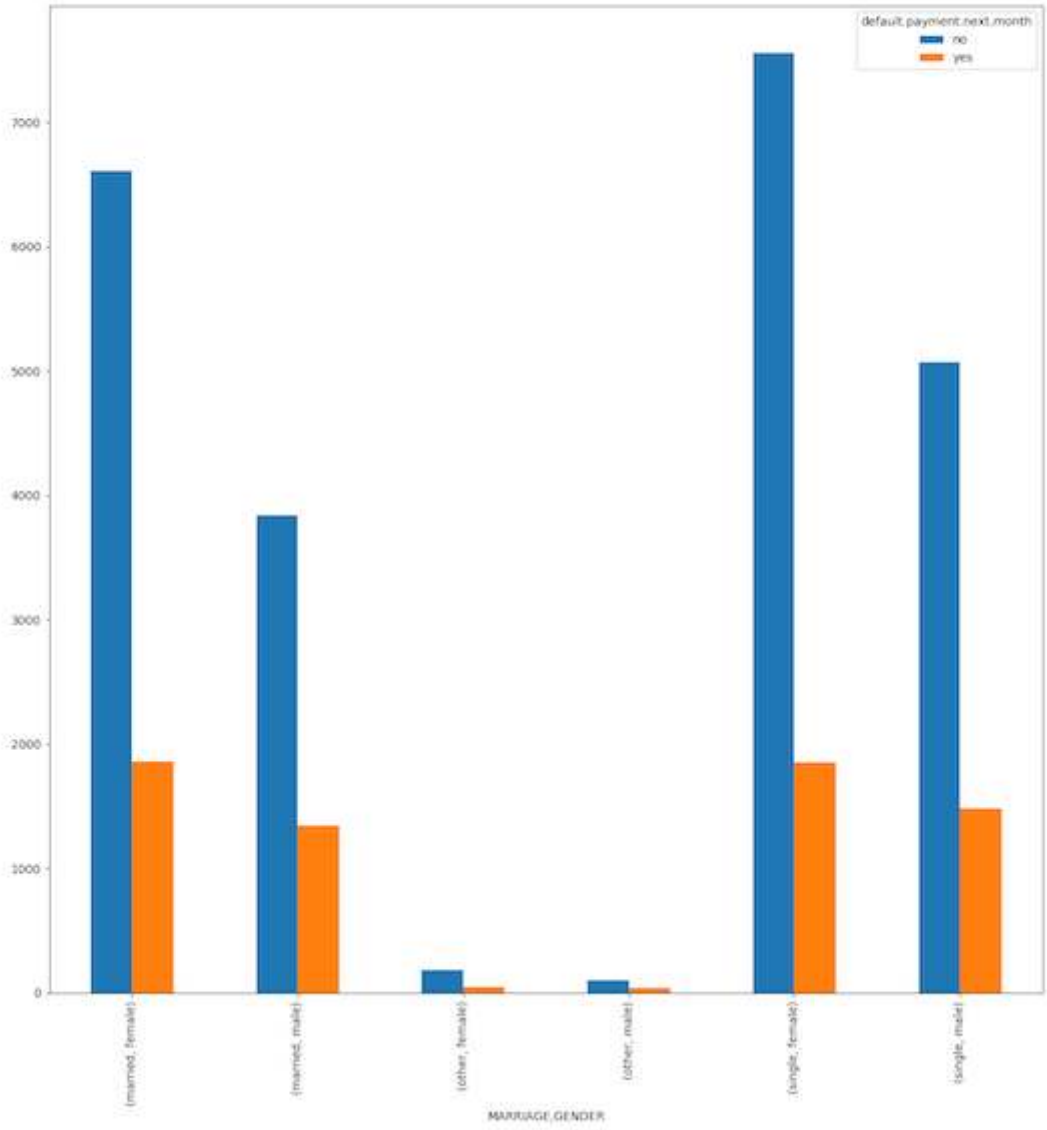
are less likely to default.

`MARRIAGE` with `GENDER`

We have calculated above the probability for defaulting of males, females, married people, and singles according to our dataset. But we might want to go a level deeper and find how likely single males or single females are to default? So, let's see how we can do that.

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('credit_card_cleaned.csv')

grouped_df = df.groupby(['MARRIAGE','GENDER','default.payment.next.month']).size()
grouped_df = grouped_df.unstack()
print(grouped_df)

grouped_df.plot(kind='bar',figsize = (15,15))

grouped_df['prob_default'] = grouped_df['yes'] / (grouped_df['no'] + grouped_df['yes'])

print('\n\n',grouped_df[['prob_default']])
```

We have written the same code that we did above except that we have added three variables (`MARRIAGE,` `GENDER` and `default.payment.next.month`) by which we group by in **line 5**. We get a bar plot in **line 9** in which we have counts of people in each category of education. The colors indicate whether they defaulted or not.



We calculate the probability of a male and female defaulting in each category of marriage status by using the same formula that we used above for calculating the probabilities of males and females defaulting. We find out that

- A single female is the least likely to default with a probability of 0.19.
- A married male has a probability of almost 0.26 to default.

Similarly, we can make other combinations using categorical variables and draw plots and calculate probabilities to find out general trends or patterns in the dataset.

Quiz

You have been given a quiz on the `MARRIAGE` column below. You are also

You have been given a quiz on the `MARRIAGE` column below. You are also provided an empty code window. You have to answer the quiz questions by writing code and finding answers to the questions.

```
import pandas as pd
df = pd.read_csv('credit_card_cleaned.csv')

# Write code here
```

1 How many married persons have defaulted in our dataset?

COMPLETED 0%

1 of 4

In the next lesson, we will learn how to explore relationships between numerical quantities.