Analysis of Credit Card Defaulters

Professor Ernesto Lee

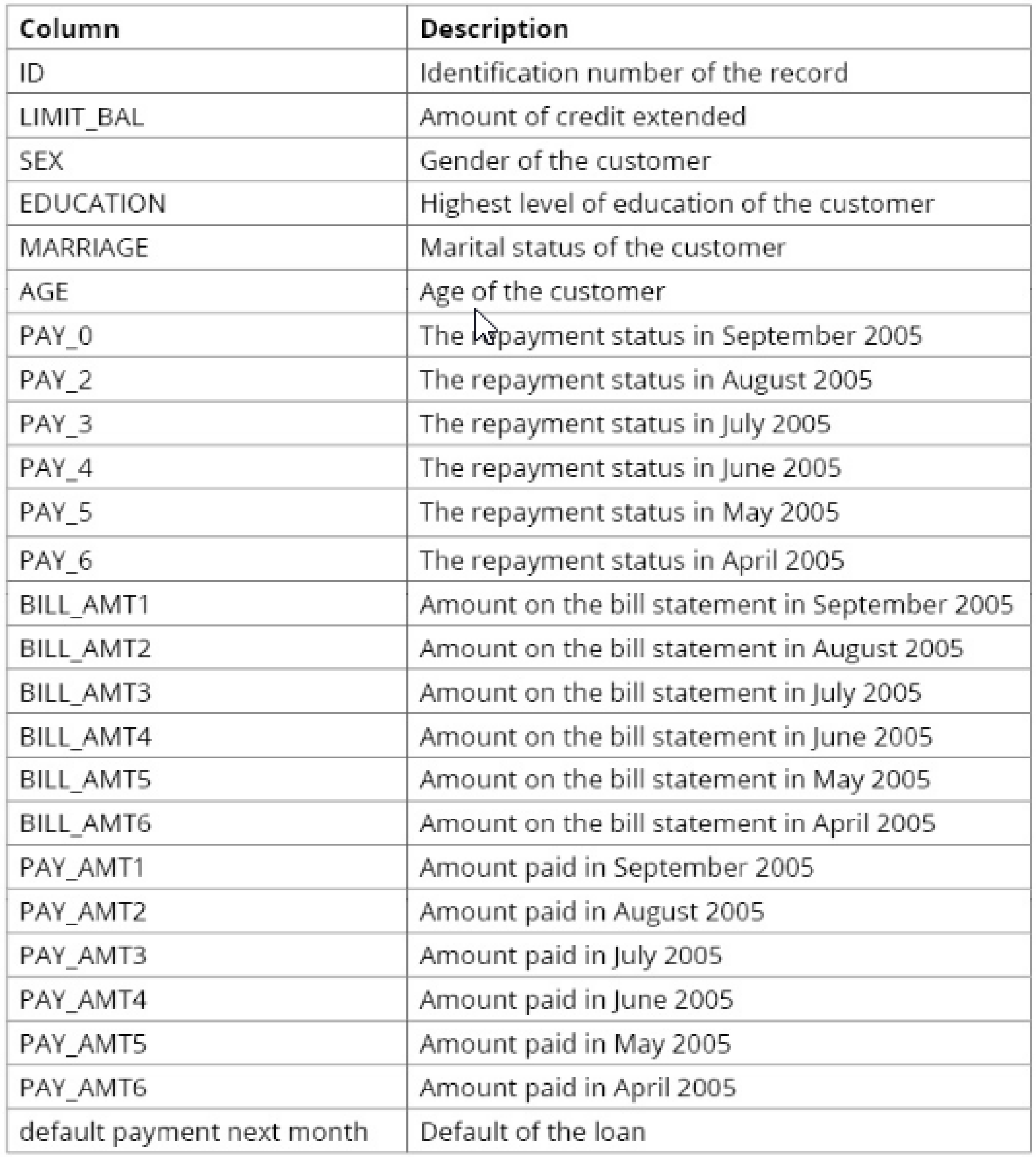
Detour: Why Learn this Stuff?

- Frederick Douglas
- 1817-1895
- Only the educated are free…

What we will learn

Access the data

- https://bit.ly/38default
- Download and open the data in Excel or Tableau
- Read the actual Data Dictionary:
  - https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#
- Take a look at each column and identify the data type for every column

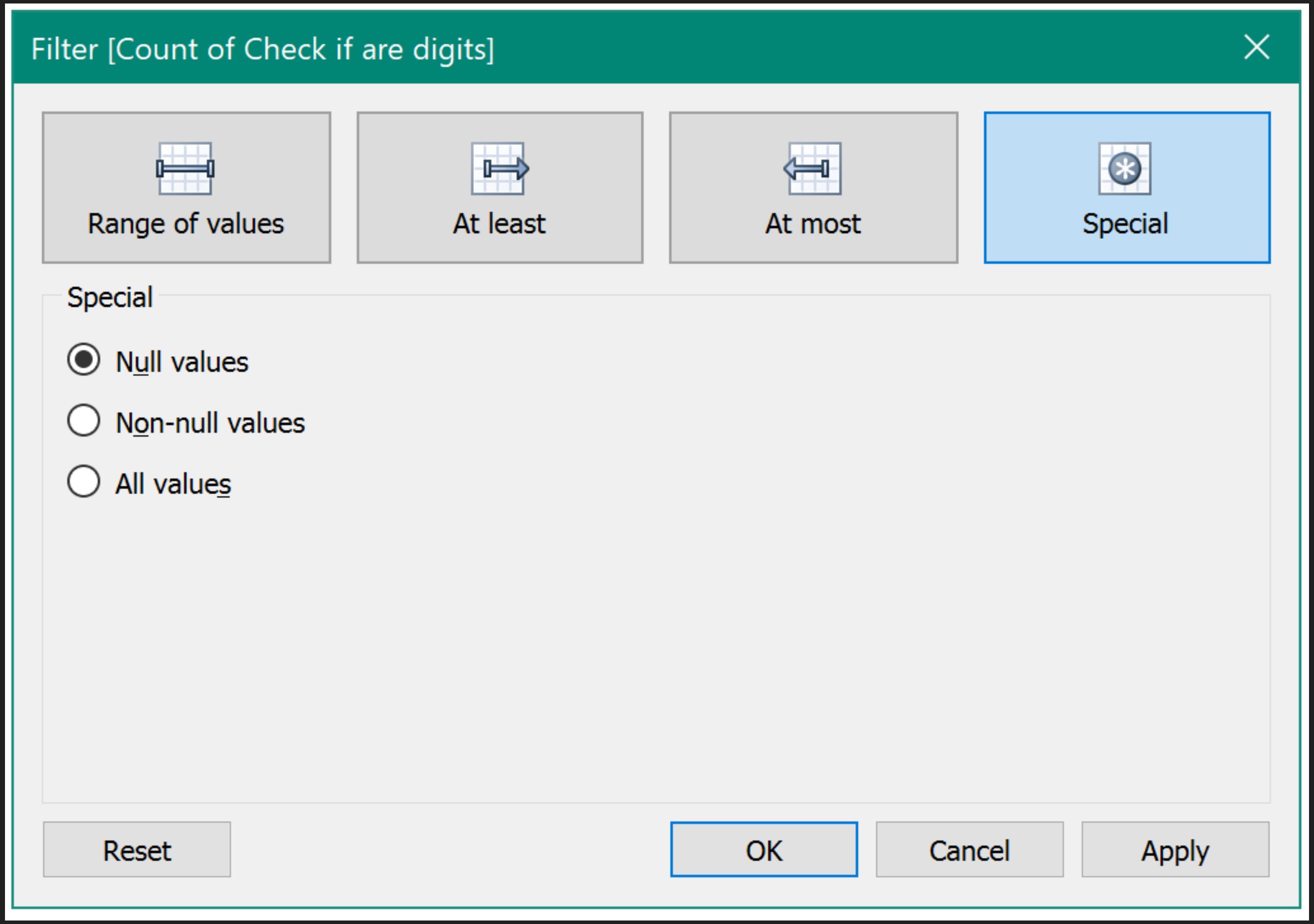


Column	Description
ID	Identification number of the record
LIMIT_BAL	Amount of credit extended
SEX	Gender of the customer
EDUCATION	Highest level of education of the customer
MARRIAGE	Marital status of the customer
AGE	Age of the customer
PAY_0	The repayment status in September 2005
PAY_2	The repayment status in August 2005
PAY_3	The repayment status in July 2005
PAY_4	The repayment status in June 2005
PAY_5	The repayment status in May 2005
PAY_6	The repayment status in April 2005
BILL_AMT1	Amount on the bill statement in September 2005
BILL_AMT2	Amount on the bill statement in August 2005
BILL_AMT3	Amount on the bill statement in July 2005
BILL_AMT4	Amount on the bill statement in June 2005
BILL_AMT5	Amount on the bill statement in May 2005
BILL_AMT6	Amount on the bill statement in April 2005
PAY_AMT1	Amount paid in September 2005
PAY_AMT2	Amount paid in August 2005
PAY_AMT3	Amount paid in July 2005
PAY_AMT4	Amount paid in June 2005
PAY_AMT5	Amount paid in May 2005
PAY_AMT6	Amount paid in April 2005
default payment next month	Default of the loan

Data Preprocessing

- Use the DA Template to define the descriptive statistics for all columns
- Check for NULL values (hint – there are none in this dataset)
- You would usually start by finding the unique values in ALL columns but now, please find the unique values in the following columns:
  - SEX
  - EDUCATION
  - Marriage
  - Pay
  - Default*
- Ask yourself, is this data categorical or numeric?
- For consistency:
  - Rename PAY_0 to PAY_1 and default_payment_next_month to DEFAULT

Check for NULL values



Filter [Count of Check if are digits]

Special

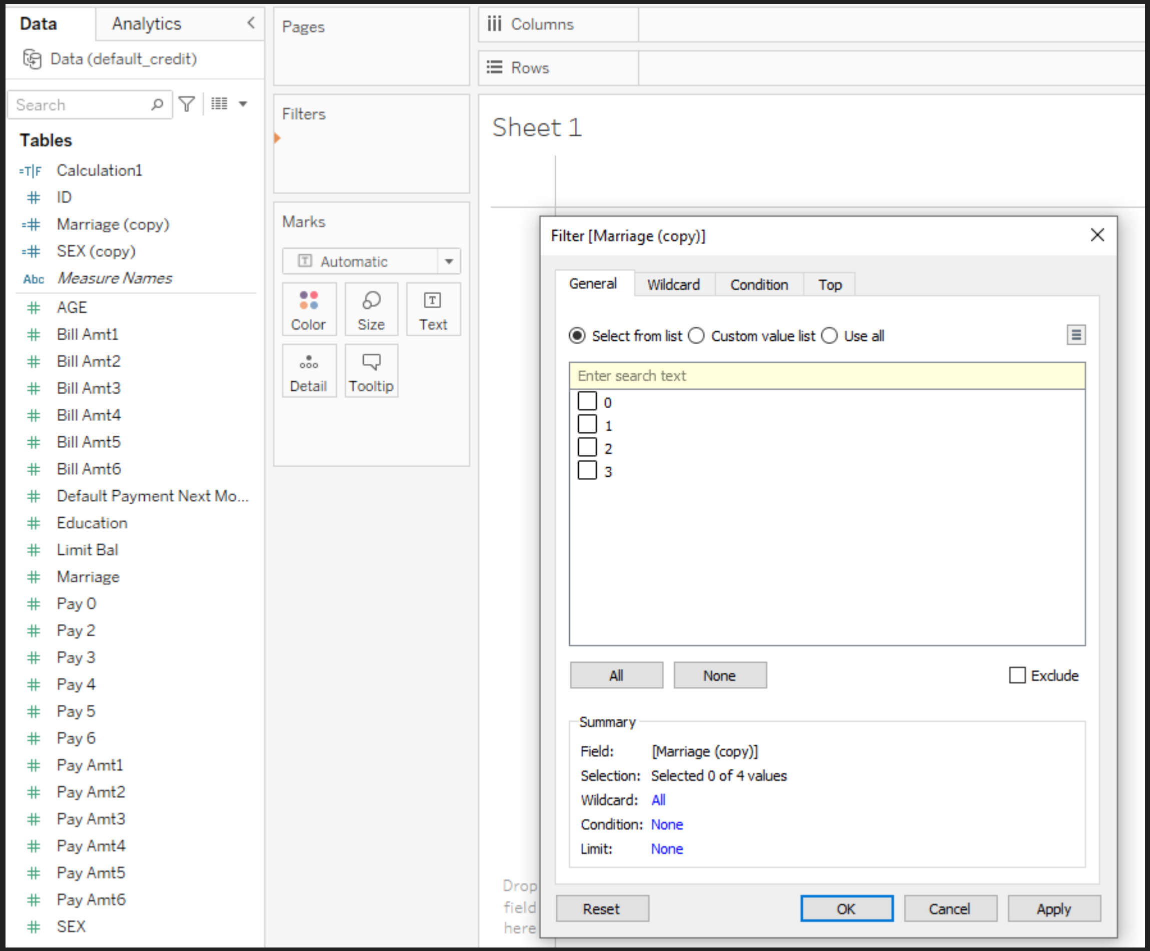
- ● Null values
- ○ Non-null values
- ○ All values

- You can filter like you see here
- Or you can create another column where all null values are set to 1 and anything else to 0, then count that. It should be something like

- **COUNT(IF "null" THEN 1 ELSE 0)**

Find Unique Values in Tableau

- 1) Drag dimension/measure to filter
  - I like to convert the measure to a dimension (make a copy first)
  - Then drop the dimension on the filter
- 2) Go to the General tab and it will show you the categories of data and the values.



● Select from list ○ Custom value list ○ Use all

□ 0
□ 1
□ 2
□ 3

□ Exclude

Summary

Field: [Marriage (copy)]
Selection: Selected 0 of 4 values
Wildcard: All
Condition: None
Limit: None

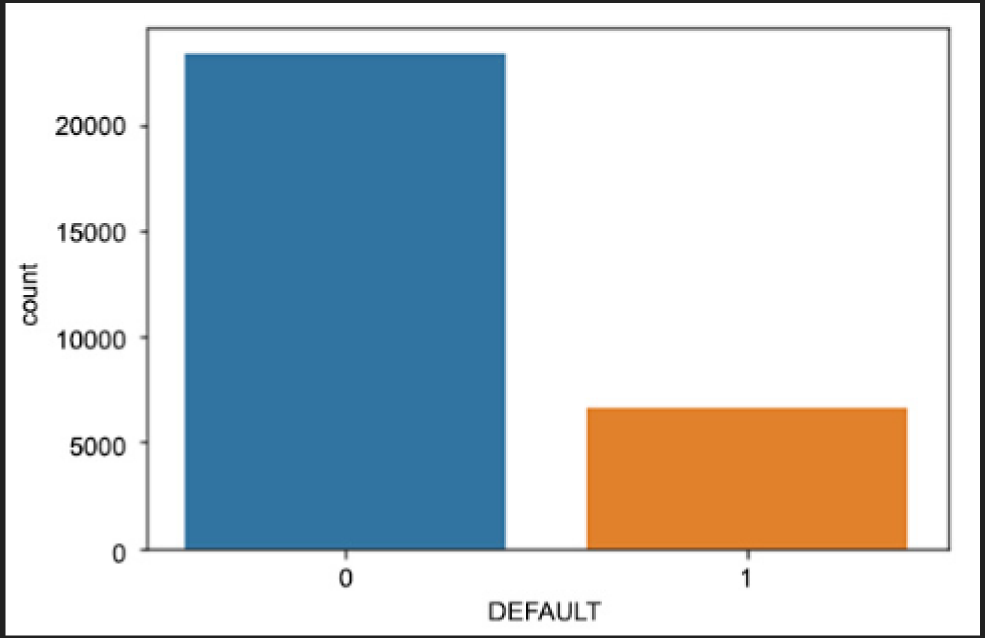
Exploratory Data Analysis

- Univariate analysis
- Bivariate analysis
- Correlation

Univariate Analysis

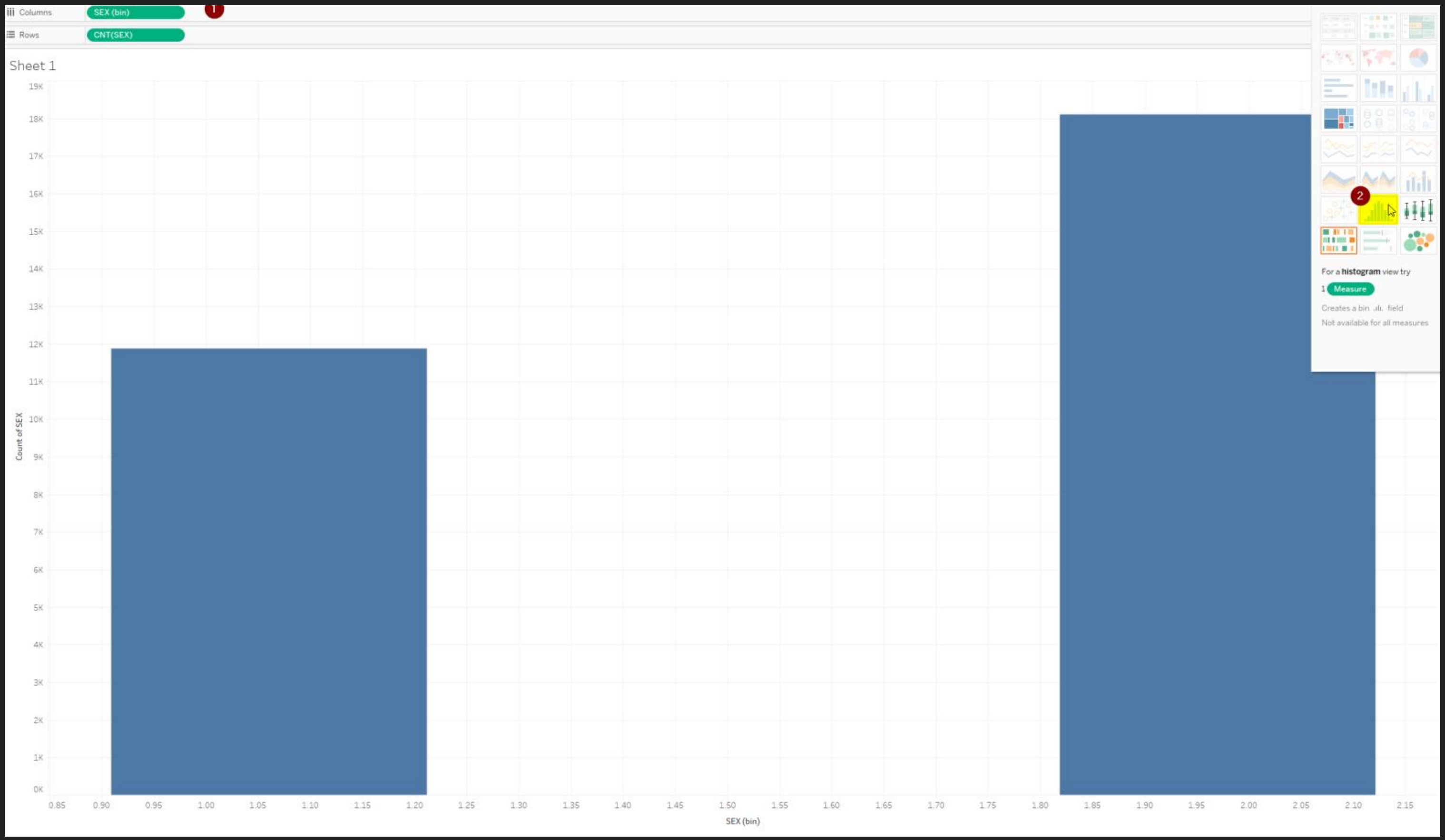
- Univariate analysis is the simplest form of analysis where we analyze each feature (that is, each column) and try to uncover the pattern or distribution of the data.

- In univariate analysis, we will be analyzing the categorical columns (DEFAULT, SEX, EDUCATION, and MARRIAGE) to mine useful information about the data.
- Create a histogram for: Default, Sex, Education and Marriage.
- Answer the following questions:
  - What percentage of folks defaulted?
  - How many males and females?
  - What is the count of each level of education?
  - How many are married, single, or divorced?



Create a Histogram for key columns

- Drag and Drop a measure onto the palette
- Select the histogram from "Show Me"



For a **histogram** view try

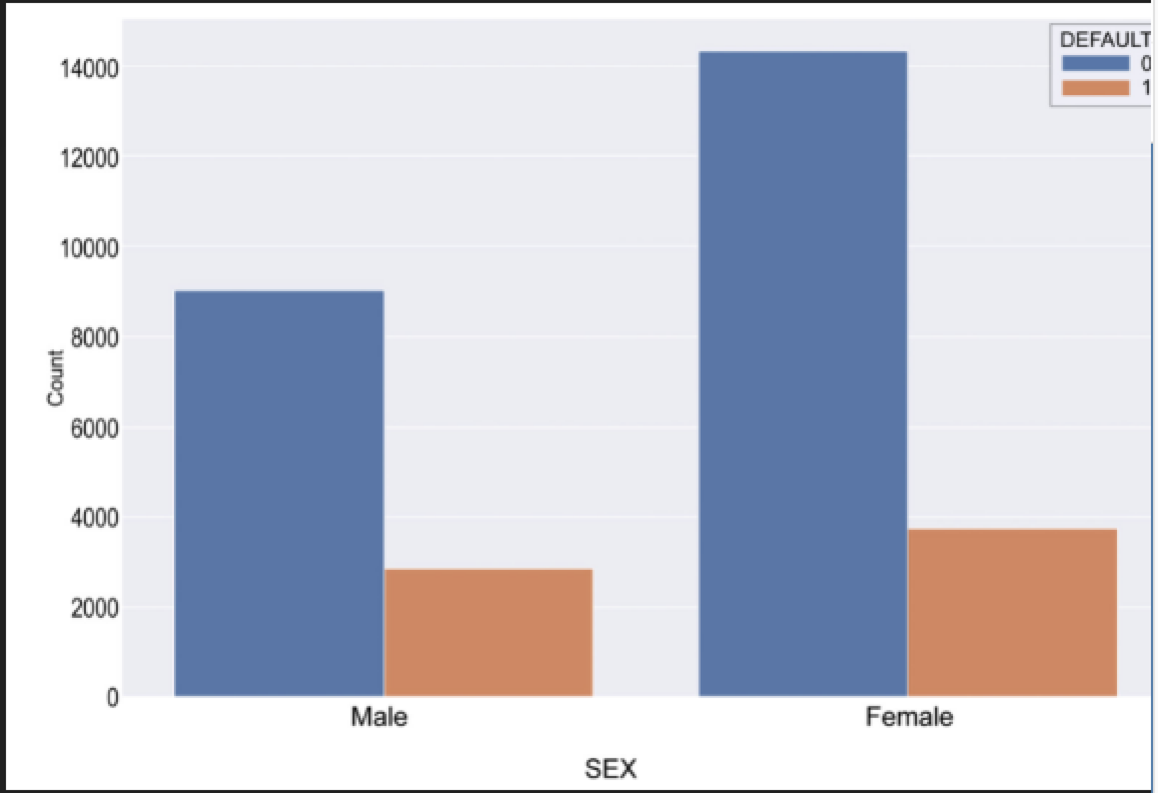
1 Measure

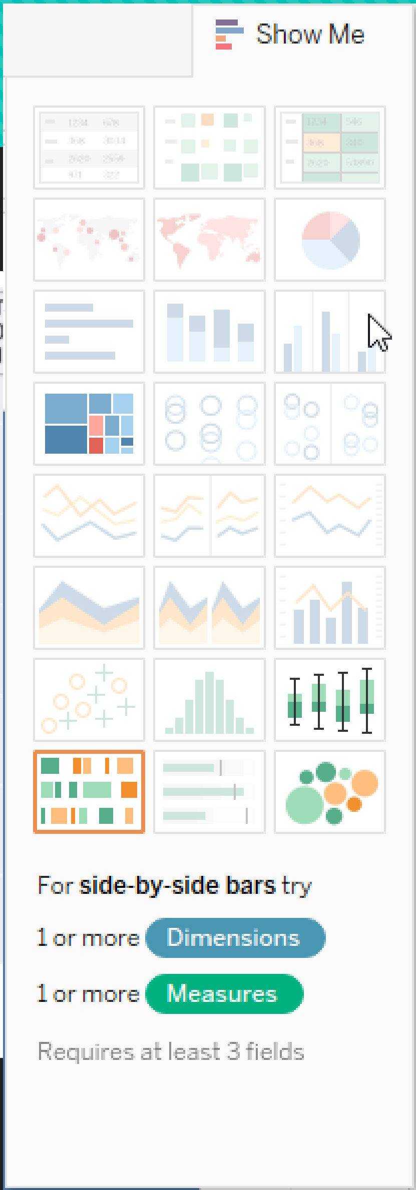
Creates a bin field

Not available for all measures

Bivariate Analysis

- Bivariate analysis is performed between two variables to look at their relationship.
- In this section, you will consider the relationship between the DEFAULT column and other columns in the dataset with the help of Pivot Tables and visualization techniques.
- Create BAR Charts to analyze default vrs:
  - Sex (demo on slide)
  - Education, Marriage
  - Payments, Balance
  - Age
  - *in all of these, consider the TOTAL numbers and also the percentage of the sub-group!! Very important.
  - Additionally, ask yourself, who is most likely to default based on these features and you will build out a profile***





For **side-by-side bars** try

1 or more Dimensions

1 or more Measures

Requires at least 3 fields

Correlation (Excel is easier for this)

- In this section, we will cover correlation – what does correlation mean, and how do we check the correlation between the DEFAULT column and other columns in our dataset?

- Correlation measures the degree of dependency between any two variables. Say, for example, we have two variables, A and B. If the value of B increases when the value of A is increased, we say the variables are positively correlated. On the other hand, if the value of B decreases when we increase the value of A, we say the variables are negatively correlated. There could also be a situation where an increase in the value of A doesn't affect the value of B, for which we say the variables are uncorrelated.

- The value of a correlation coefficient can vary between -1 to 1, with 1 being a strong positive correlation and -1 a strong negative correlation.

- By studying the correlation between the DEFAULT column and other columns with the help of a heatmap, we can figure out which column/variable has a high impact on the DEFAULT column.

What is your final profile?

- A male customer is more likely to default than a female customer.
- People with a relationship status of other are more likely to default than married or single people.
- A customer whose highest educational qualification is a high-school diploma is more likely to default than a customer who has gone to graduate school or university.
- A customer who has delayed payment for 2 consecutive months has a higher probability of default.
- A customer who is 22 years of age has a higher probability of defaulting on payments than any other age group.