ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

INTRODUCTION TO DATA

PART 1

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

What Is Statistics?

Statistics

- Is a discipline that focuses on the collection, analysis, and interpretation of data.

- Wide range of applications, spanning all areas where data is available or can be collected, including:
  - "Hard" Science (*physics, chemistry, biology, etc.)*
  - Social Sciences (*politics, economics, education, etc.)*
  - Medicine

What Is Data?

Data

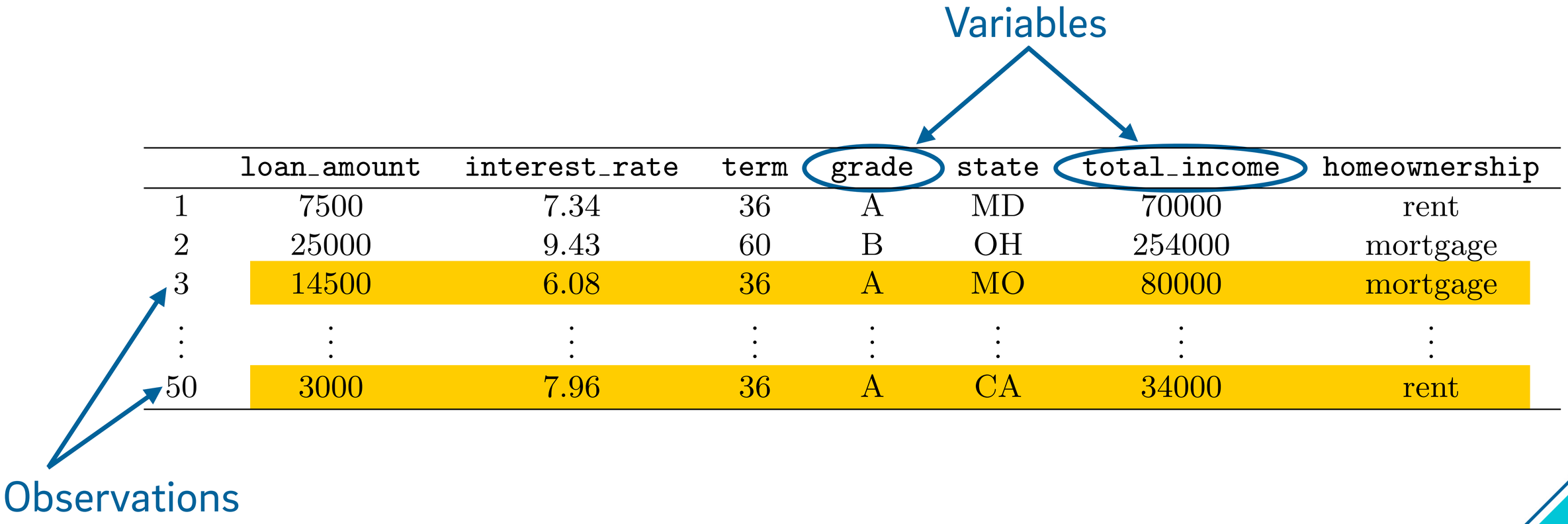
- Individual facts, observations, or pieces of information.

- Can be:
  - numerical *(e.g., height, weight, wealth, test scores)*
  - discrete *(e.g., number of people in city)*
  - continuous *(e.g., a person's weight)*
  - categorical *(e.g., Democrat/Republican, yes/no, etc.)*

- Can be collected:
  - via observations in the field
  - by careful design of experiments

Basics of Data

Table from Open-intro Statistics textbook, Chapter 1



	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Relationships Between Variables

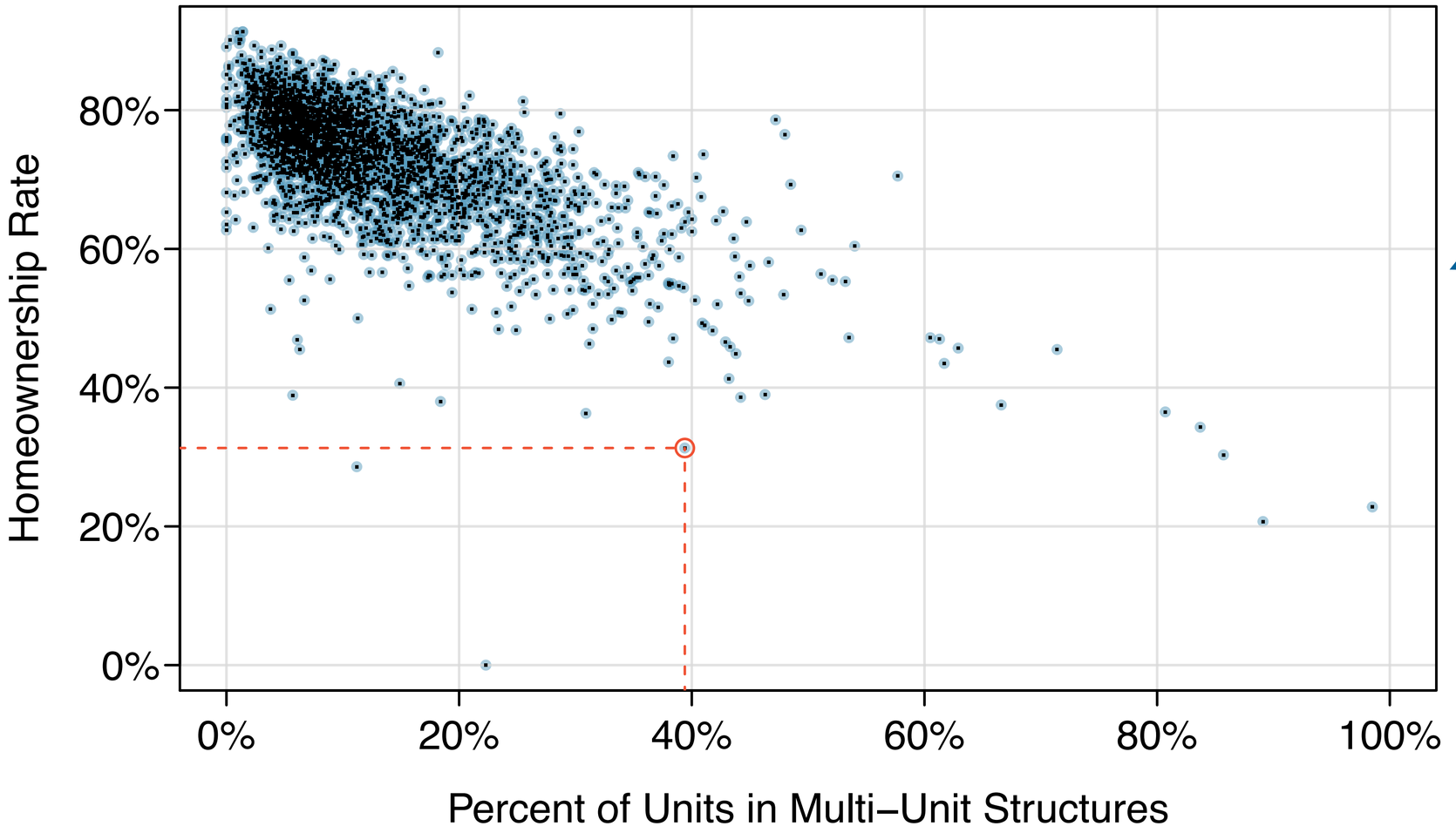
Often we are interested in the relationships between variables.

Example

- Suppose we have a data set containing observations across many variables in US counties: unemployment rate, population, state, home ownership rates, percent of units in multi-unit buildings, etc.
- Suppose we want to understand the relationship between homeownership rate and the percentage of units in multi-unit buildings.

Relationships Between Variables

Table from Open-intro Statistics textbook, Chapter 1



Scatterplot

Relationships Between Variables

- The scatterplot shows that there is a discernible pattern relating the two variables.
- We call these variables **associated** variables, or **dependent** variables.
- The scatterplot shows a downward trend: counties with a high percent of multi-unit structures, are associated with low rate of homeownership *(and vice-versa)*.
- These variables are then said to **negatively associated**. *(A **positive association** is defined analogously)*
- If no pattern is discernible, we say the variables are **independent**.

Fact: No pair of variables is both associated and independent.