

Introduction to Data Science

Numerical vs. Categorical Data

Joanna Bieri DATA101

Welcome

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 [Click Here for Joanna's Schedule](#)

Today's Big Ideas

- Numerical data and categorical data, and why the difference matters
- Visualizing numerical data with histograms
- Comparing two numerical variables
- Visualizing categorical data with bar plots

Types of Data

Big Idea: Know What Kind of Data You Have

- The number of variables you are looking at changes the analysis: **univariate** (one), **bivariate** (two), **multivariate** (many)
- **Numerical** variables are numbers you can do math on. They are **continuous** (any value in a range, like price) or **discrete** (only specific values, like birth year)
- **Categorical** variables are labels. They are **ordinal** if there is a natural order (gold, silver, bronze) or **nominal** if there is not (blood type, hair color)
- The type of the variable decides which graph makes sense

Check Your Understanding

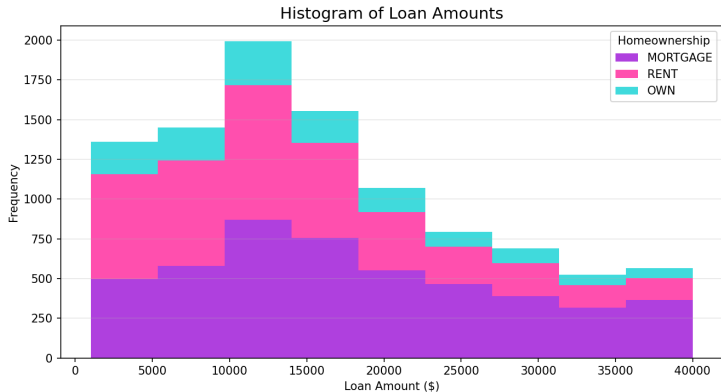
- **Q1.** Is `loan_amount` numerical or categorical? Continuous or discrete?
- **Q2.** The loans data has a `grade` column with values A through G. Which type is that, and is it ordinal or nominal?
- **Q3.** A zip code is stored as a number. Is it really numerical data? Why or why not?
- **Q4.** Give an example of a bivariate question you could ask about the loans data.

Visualizing Numerical Data

Big Idea: Histograms Show Shape

- `.describe()` gives you summary statistics, but only for the numerical columns. It quietly skips the categorical ones
- A **histogram** sorts the data into bins and counts how many land in each one
- The number of bins is your choice, and it changes the story: too few hides the shape, too many turns it into noise
- $\text{range} = \text{max} - \text{min}$, then $\text{bin width} = \text{range} / \text{number of bins}$
- Adding `marginal="box"` or `marginal="violin"` puts the distribution right alongside the histogram

Example: Loan Amounts by Homeownership



Check Your Understanding

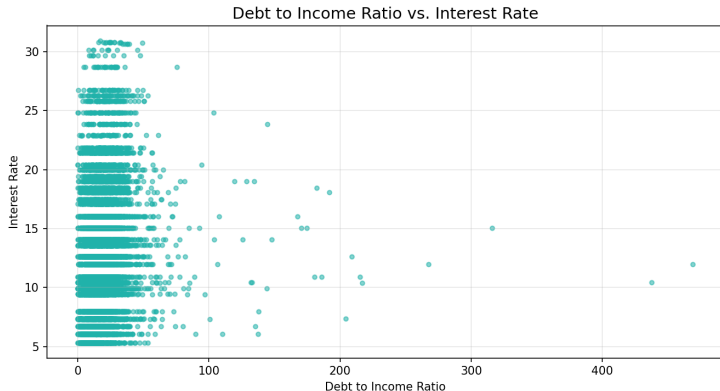
- **Q5.** Looking at the graph, which homeownership group takes the largest loans?
- **Q6.** What would happen to this graph if we set `nbins=2`? What about `nbins=200`?
- **Q7.** Why does `.describe()` leave out the `homeownership` column?

Comparing Two Numerical Variables

Big Idea: Scatter Plots and Heat Maps

- A **scatter plot** puts one dot per observation, so you can see whether two variables move together
- With 10,000 observations the dots pile up on top of each other and you lose the detail
- A **heat map** fixes that: it bins the x and y values and colors each rectangle by how many points fell inside
- Same two variables, two different views. Pick the one that answers your question

Example: Debt to Income vs. Interest Rate



Check Your Understanding

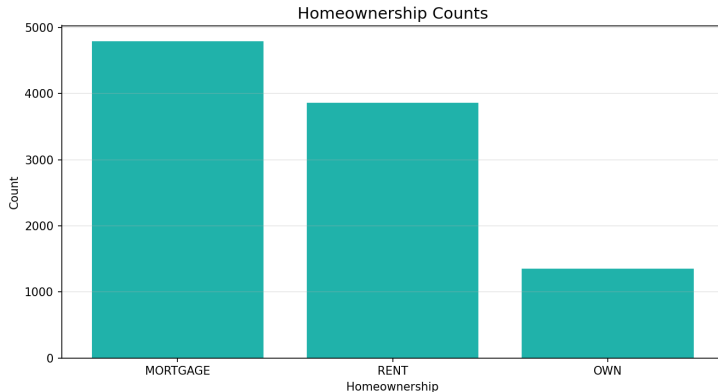
- **Q8.** Does a higher debt to income ratio seem to go with a higher interest rate?
- **Q9.** Where on this graph are the dots so dense that you cannot tell how many there are?
- **Q10.** What would a heat map show you here that the scatter plot cannot?

Visualizing Categorical Data

Big Idea: Bar Plots Count Categories

- Summary statistics do not work on categorical data, so we count instead
- `value_counts()` gives you the count for each category
- A **bar plot** is like a histogram, except the bins are the categories themselves
- You choose the direction: put the categorical variable on x for vertical bars, on y for horizontal
- Adding `color=` fills the bars by a second category, and `facet_col=` splits them into separate plots

Example: Homeownership Counts



Check Your Understanding

- **Q11.** Why can we not just run `.describe()` on the homeownership column?
- **Q12.** What is the difference between a histogram and a bar plot?
- **Q13.** If you colored these bars by grade, what new question could you answer?

Wrap-Up

Start Simple, Build Up

Start with a plain graph, then add one thing at a time.

Questions?