Scatter Plots

BOOTSTRAP DATA SCIENCE

@BootstrapWorld

Making Scatter Plots

- Do you think that younger animals get adopted faster? Why or why not?
- What kind of data is `age`? What kind of data is `weeks`?
- What kind of display would help us analyze the relationship between age and adoption time?

Pie and **bar charts** help us see the *frequency* of values in a *categorical* column. **Histograms** and **box plots** help us explore the *distribution* of values in a *quantitative* column.

Making Scatter Plots

What we really want is a display that will help us search for **a relationship between two quantitative columns**, and that's exactly what scatter plots do.

Scatter plots reveal the relationship between two columns by plotting one on the x-axis and the other on the y-axis.

Making Scatter Plots

Before we can draw a ***scatter plot***, we have to make an important decision: which variable is ***explanatory*** and which is the ***response***?

In this case, are we suspecting that an animal's weight might explain how long it takes to be adopted? Or that adoption time can explain how much an animal weighs?

The first one makes sense, and reflects our suspicion that age plays a role in adoption time.

Making Scatter Plots

We will produce our scatter plot by graphing each animal's `age` and `weeks` values as a point on the x and y axes.

Complete Creating a Scatter Plot, to get a feel for making scatter plots by hand.

Making Scatter Plots

When you created the scatter plot by hand, you started with a Table. Then you plotted a series of dots, using one column for your x's, one column for your y's, and the `name` column to provide a label for each dot.

The `scatter-plot` function works exactly the same way: it starts with a table, and then needs to know which columns to use for labels, xs, and ys. Here's the contract:

```
scatter-plot :: (t::Table, ls::String, xs::String, ys::String)
```

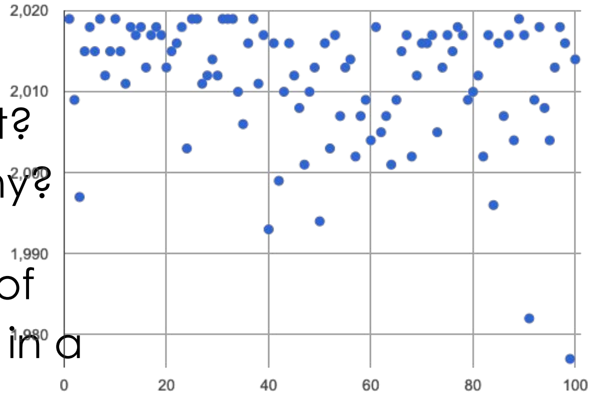
Making Scatter Plots

- Open your saved Animals Starter File, or make a new copy.
- Make a scatter plot that displays the relationship between `age` and adoption time (`weeks`).
- Are there any patterns or trends that you see here?

Making Scatter Plots

Scatter plots show us a collection of points, arranged along two axes. If there's a relationship between these axes, we'll see clumps and clouds of points in the graph.

- What pattern do you see in *your* scatter plot?
- Are there any points that seem unusual? Why?



Suppose we plotted the age and adoption ime of four random animals, and found that they all fell in a line. Is this enough to determine that there's a relationship between the variables?

The Data Cycle

Is age the only factor that determines how long it takes for an animal to get adopted?

Many apartment buildings do not allow large breeds of dogs, and have a limit on how heavy a tenant's dog can be. Bigger dogs are not welcome in many apartments.

*Perhaps the **weight** of an animal influences the adoption time!*

The Data Cycle

Take a look at the animals dataset, either in your workbook or on the spreadsheet (Google).

Do you think there's a relationship between `pounds` and `weeks` in this table? Why or why not?

The Data Cycle

Complete the first Data Cycle on Data Cycle: Relationships in the Animals Dataset.

- What did you find when you looked at the scatter-plot?
- Does there appear to be a pattern or trend?
- What might be problematic about including every species in the same scatter plot of weight?
- What follow-up questions do you have?

The Data Cycle

Write your follow-up question in the second Data Cycle on Data Cycle: Relationships in the Animals Dataset, and complete the Data Cycle for your new question.

The Data Cycle

We’ve got a lot of tools in our toolkit that help us think about an entire *column* of a dataset:

- We have ways to find measures of center and spread for a given quantitative column.
- We have visualizations that let us see the shape of values in a quantitative column.
- We have visualizations that let us see frequencies or percentages in a categorical column.

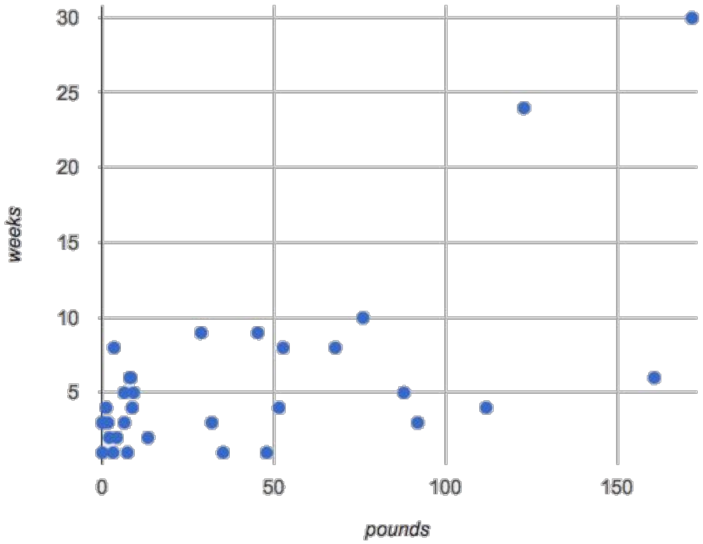
The Data Cycle

Now we also have a tool that lets us think about two columns at the same time!

What new questions did the Data Cycle lead you to ask? What did you find?

Looking for Trends

Shown below is a scatter plot of the relationships between the animals' `pounds` and the number of `weeks` it takes to be adopted. **Do you see a trend?**

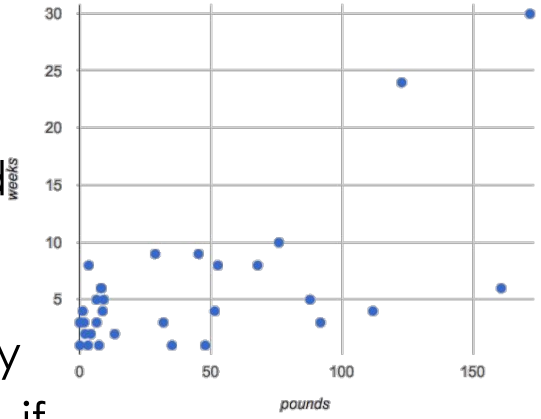


Looking for Trends

A straight-line pattern in the cloud of points suggests a linear relationship between two columns.

If we can find a line around which the points cluster (as we'll do in a future lesson), it would be useful for **making predictions**.

For example, our line might predict how many `weeks` a new dog would wait to be adopted, if it weighs 68 `pounds`.



Looking for Trends

Do any data points seem unusually far away from the main cloud of points? Which animals are those?

These points are called *unusual observations*. Unusual observations in a scatter plot are like outliers in a histogram, but more complicated because it's the *combination* of x and y values that makes them stand apart from the rest of the cloud.

Looking for Trends

Unusual observations are *always* worth thinking about!

- Sometimes they're *just random*. Felix seems to have been adopted quickly, considering how much he weighs. Maybe he just met the right family early, or maybe we find out he lives nearby, got lost and his family came to get him. In that case, we might need to do some deep thinking about whether or not it's appropriate to remove him from our dataset.

Looking for Trends

Unusual observations are *always* worth thinking about!

- Sometimes they can give you a *deeper insight* into your data. Maybe Felix is a special, popular (and heavy!) breed of cat, and we discover that our dataset is missing an important column for breed!

Looking for Trends

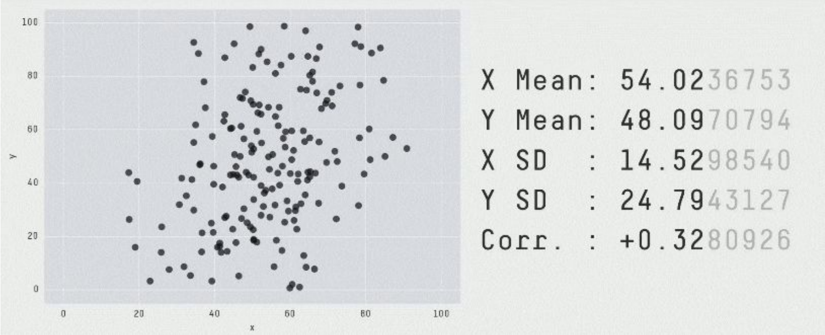
Unusual observations are *always* worth thinking about!

- Sometimes unusual observations are *the points we are looking for*! What if we wanted to know which restaurants are a good value, and which are rip-offs? We could make a scatter plot of restaurant reviews vs. prices, and look for an observation that's high above the rest of the points. That would be a restaurant whose reviews are *unusually good* for the price. An observation way below the cloud would be a really bad deal.

Looking for Trends

It's not just about the numbers!

These numbers and scatter plot come from different datasets. The patterns in the scatter plot vary wildly, but the numbers that summarize dataset barely change at all!



```
X Mean: 54.0236753
Y Mean: 48.0970794
X SD  : 14.5298540
Y SD  : 24.7943127
Corr. : +0.3280926
```

Looking for Trends

Data Scientists and Statisticians use their eyes all the time. Sometimes there's a pattern hiding in the data, which can't be seen just by focusing on numbers and measures.

Until we really look at the *shape* of the data, we aren't seeing the whole picture.

Looking for Trends

For practice, consider each of the following relationships. First think about what you *expect*, then make the scatter plot to see if it supports your hunch.

- How are the `pounds` of an animal related to its `age`?
- How are the number of `weeks` it takes for an animal to be adopted related to its number of `legs`?
- How are the number of `legs` an animal has related to its `age`?
- Do you see a linear (straight-line) relationship in any of these?
- Are there any unusual observations?

Looking for Trends

It might be tempting to go straight into making a scatter plot to explore how weeks to adoption may be affected by age. But different animals have very different lifespans!

Why does that matter?

Looking for Trends

A 5-year-old tarantula is still really young, while a 5-year-old rabbit is fully grown. With differences like this, it doesn’t make sense to put them all on the same scatter plot. By mixing them together, we may be *hiding* a real relationship, or creating the illusion of a relationship that isn’t really there!

Looking for Trends

It would be nice if the dots in our scatter plot were different colors or shapes, depending on the species. That would give us a much better picture of what's really going on in the data. *But making a special image for every single row in the table would take a very long time!* If only there was a function with a contract like:

```
species-dot :: (r :: Row) -> Image
```

This function could take in a row from the animals table, and draw a special dot depending on the species. Unfortunately, no such function exists...yet!

Data Exploration Project (Scatter Plots)

Let’s review what we have learned about making and interpreting scatter plots.

- Does a scatter plot display categorical or quantitative data? How many columns of data does a scatter plot display?
- What do scatter plots show us about a dataset?

Data Exploration Project (Scatter Plots)

Let's connect what we know about scatter plots to your chosen dataset.

- Open your chosen dataset starter file in Pyret.
- Choose two quantitative columns from your dataset whose relationship you want to explore, and another column that makes sense to use as labels for your points.
- What question does your display answer?
- Write down that question in the top section of Data Cycle: Relationships in Your Dataset.

Data Exploration Project (Scatter Plots)

- Complete the rest of the data cycle, recording how you considered, analyzed and interpreted the question.
- Repeat this process for at least one other pair of quantitative columns.

Data Exploration Project (Scatter Plots)

It's time to add to your Data Exploration Project.

- Copy/paste at least two scatter plots. Be sure to also add any interesting questions that you developed while making and thinking about your scatter plots.
- Which relationships did you look for?
- Do you see any possible relationships or trends?
- What do those findings mean?
- What new questions come up for you?

Data Exploration Project (Scatter Plots)

Share your findings!

Were the relationships you investigated stronger or weaker than they expected?

What questions did the scatter plots raise about your dataset?

What, if any, outliers did you discover when making scatter plots?

Were there any surprises when you compared your findings with other students? (For instance: Did everyone find outliers? Was there more or less similarity than expected?)