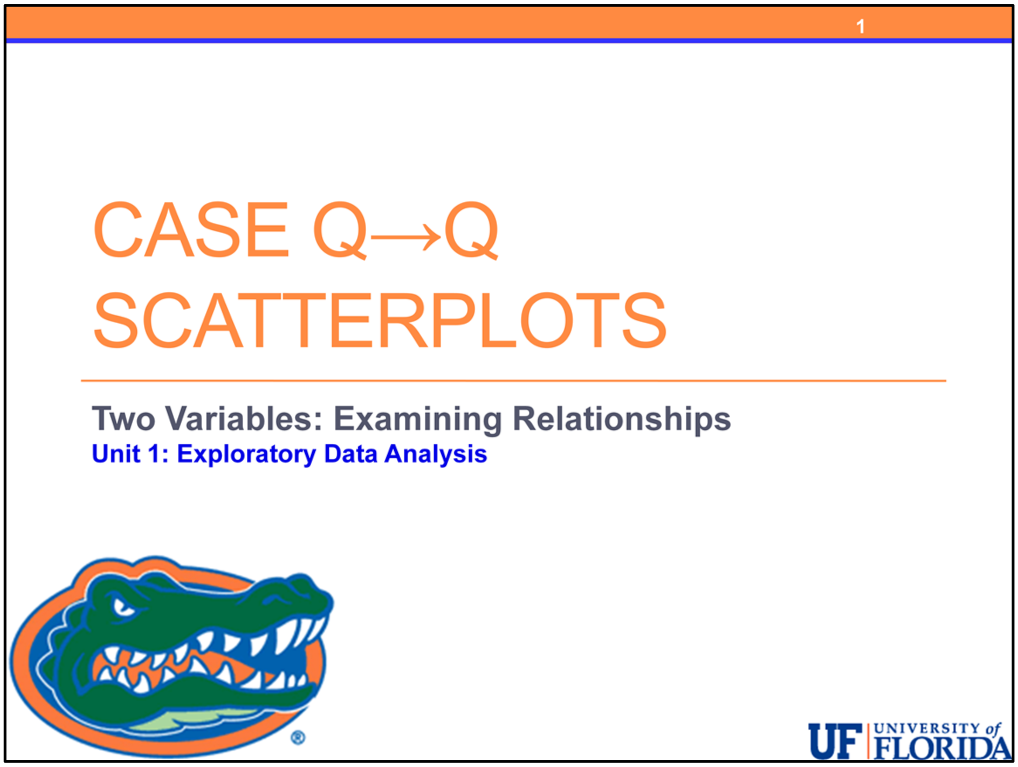


CASE Q→Q SCATTERPLOTS

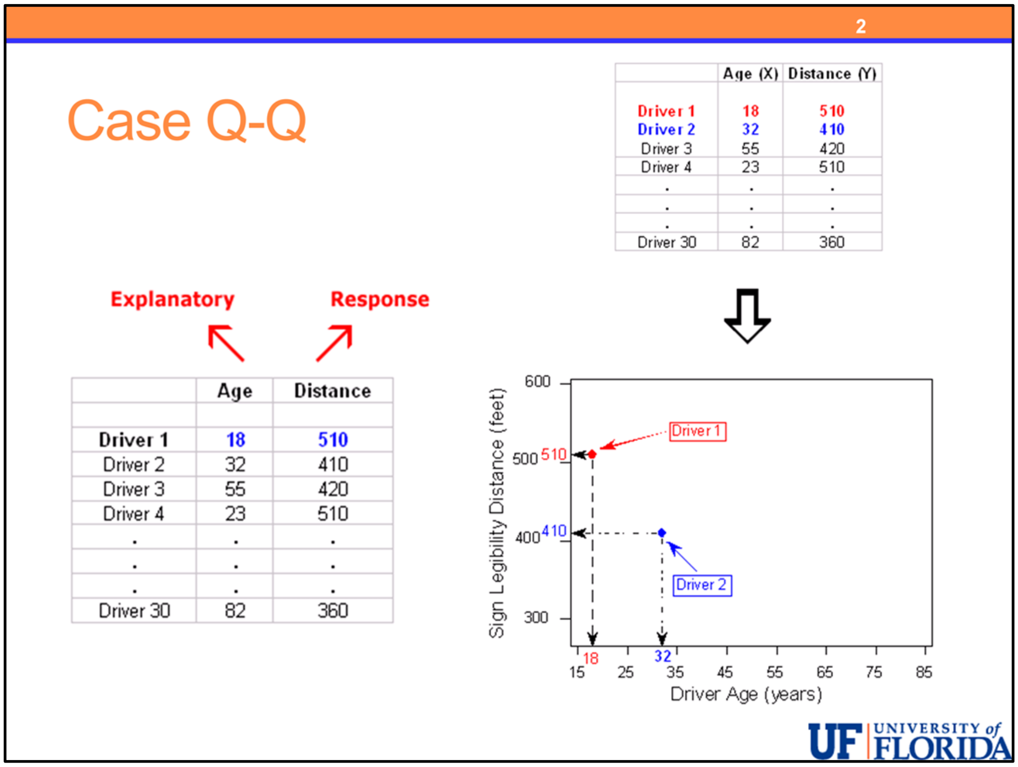
Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

In most situations involving two quantitative variables, a scatterplot is the appropriate visual display.

Remember our assumption that the observations in our dataset be independent. If individuals appear more than once or are related to other observations in any of the displays or numerical summaries we have been using for one or two variables, the results could easily be misleading or completely incorrect.



2

Case Q-Q

	Age (X)	Distance (Y)
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

Explanatory

Response

	Age	Distance
Driver 1	18	510
Driver 2	32	410
Driver 3	55	420
Driver 4	23	510
.	.	.
.	.	.
.	.	.
Driver 30	82	360

We will have one variable on the X-axis which is the explanatory variable. And another variable on the Y-axis, which is the response variable.

If there is no clear distinction of the Role of the variables in the scenario, either choice can be made.

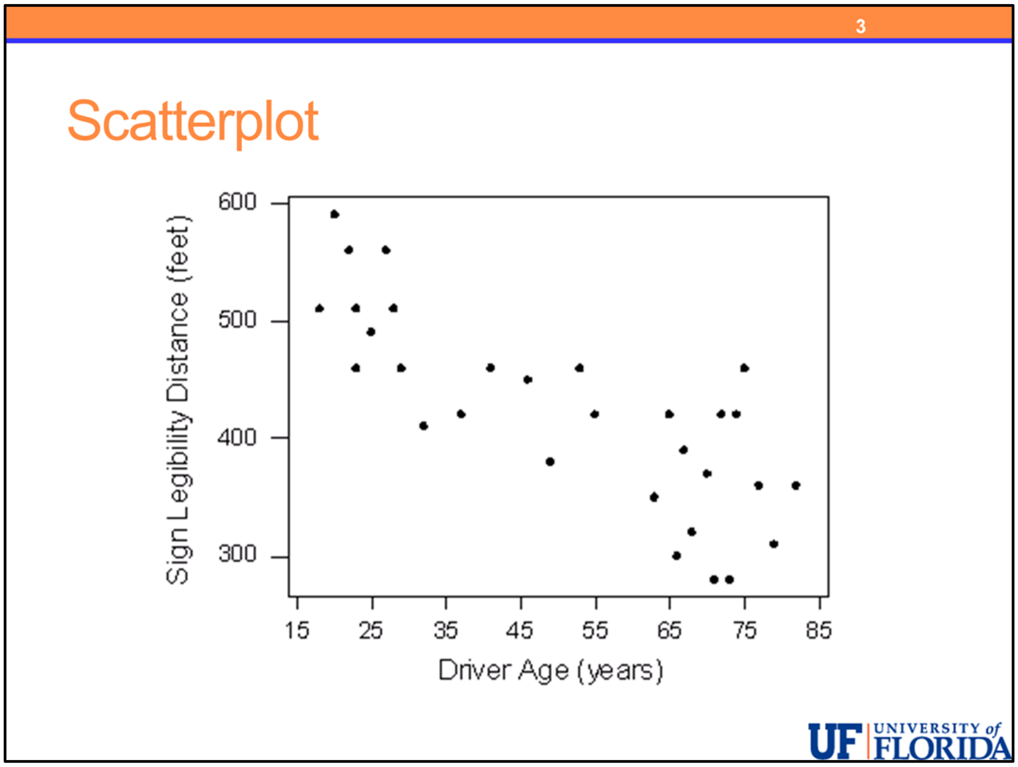
Here we're asking: Does age predict how far away you can comfortably read a highway sign?

In this case, it is fairly obvious that age isn't going to be what we are PREDICTING, age is not something we wish to alter or predict, instead we are interested in determining if age tells me something about how far I expect the individual to be able to clearly see a highway sign.

To create the scatterplot, for each individual, we locate their age on the X-axis and their distance on the Y-axis and place a point at the intersection between that X and Y location in two-dimensional space.

In this plot you can see we have graphed driver 1 (in red in the upper left of the plot) and driver 2 (in blue).

After we plot the points for all of the data, we get something like this:



3

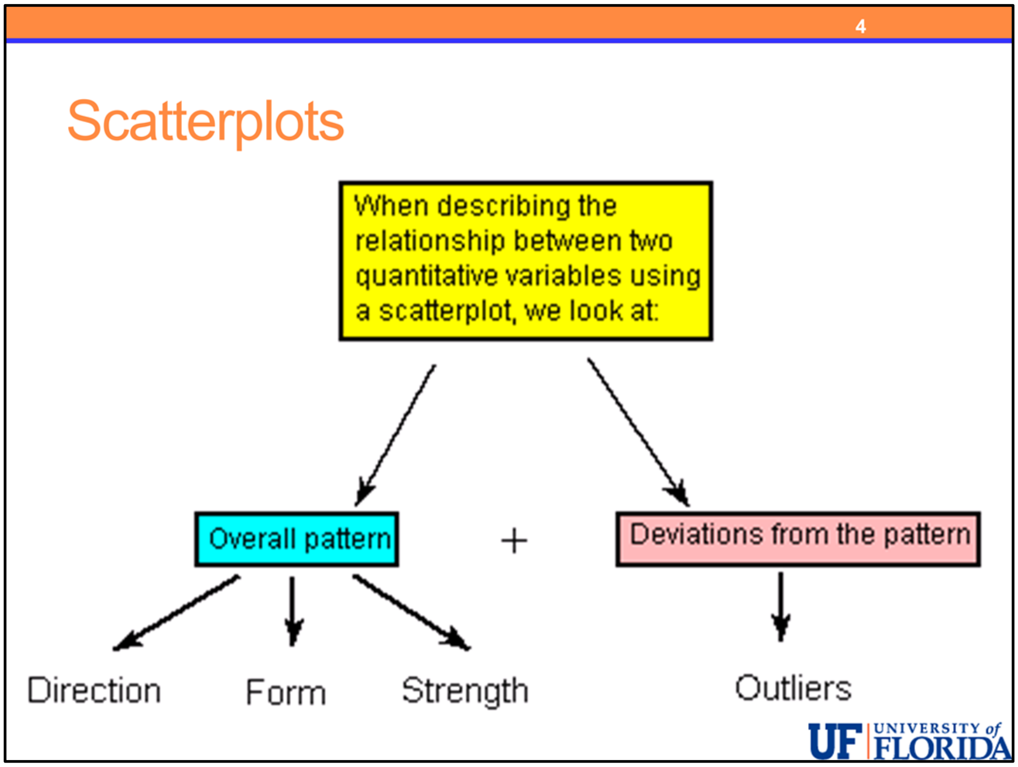
Scatterplot

UF UNIVERSITY of FLORIDA

In this data, older individuals tended to have sign legibility distances which were shorter than did individuals who were younger. We could say: “as age increases, sign legibility distance tends to decrease.”

We might want to measure how much does the distance decrease per year or 5 years or 10 years? The slope of the regression line will help us address those questions.

And we would like to quantify: How tight or strong is the relationship? How good is the prediction?



4

Scatterplots

When describing the relationship between two quantitative variables using a scatterplot, we look at:

Overall pattern + Deviations from the pattern

Direction

Form

Strength

Outliers

UF UNIVERSITY of FLORIDA

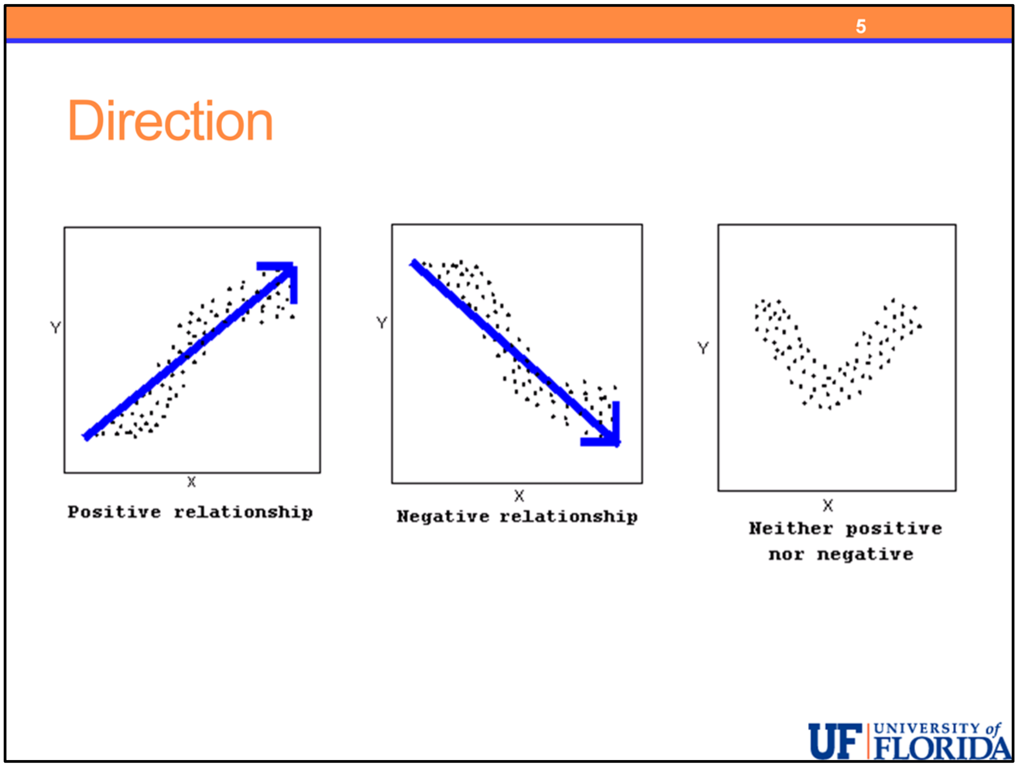
When we look at a scatterplot we want to describe the relationship we see. To do this we need to describe the overall pattern and deviations from that pattern.

When we have two quantitative variables, the overall pattern involves:
- The direction of the relationship (positive for increasing trends, negative for decreasing trends, or neither)
- The form of the relationship (linear or curvilinear/non-linear)
- The strength of the relationship (how closely the data points fit the overall pattern)

If we have deviations from that pattern, we identify them as outliers.

Before beginning our discussion of classifying or describing relationships, it is possible that there is “no relationship.” If so, there is no need to discuss the direction, form, or strength of what you see.

Before describing a relationship, you should generally see some sort of relationship in your scatterplot worth describing.



5

Direction

Positive relationship

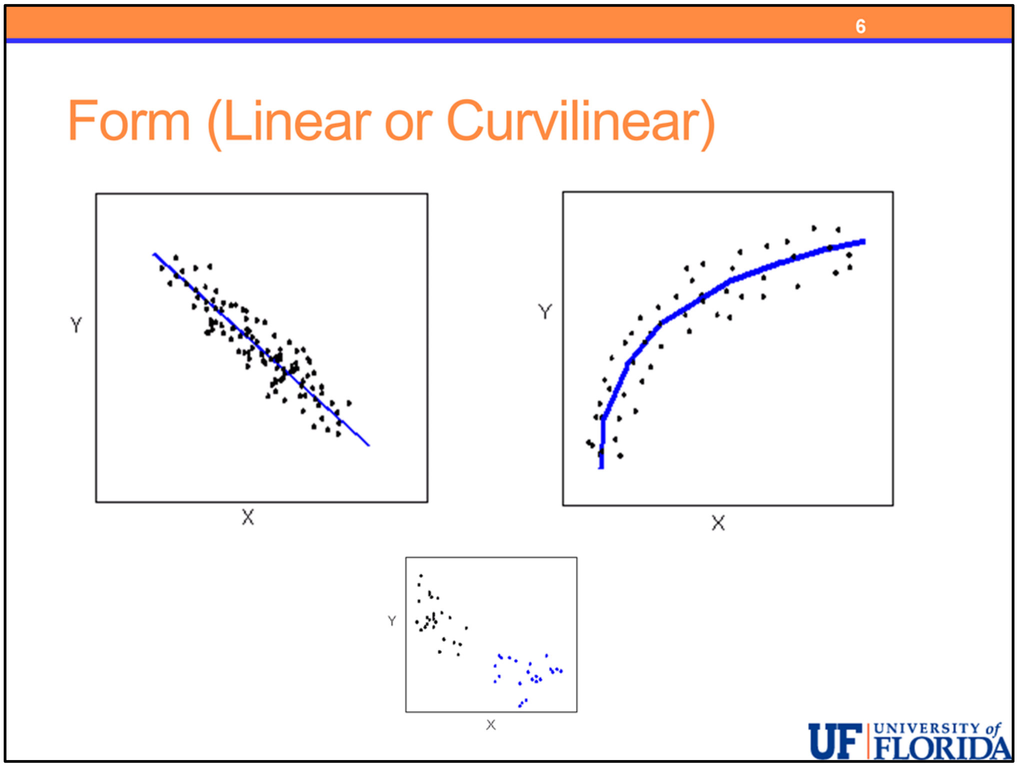
Negative relationship

Neither positive nor negative

For direction, we have

- Positive relationship – an overall increasing trend of any kind
- Negative relationship – an overall decreasing trend of any kind
- Neither Positive nor Negative – relationship has both increasing and decreasing components

Notice that both of the relationships drawn here for positive and negative are NOT linear. It doesn't have to be linear to be positive or negative as long as it is always increasing or always decreasing.


6

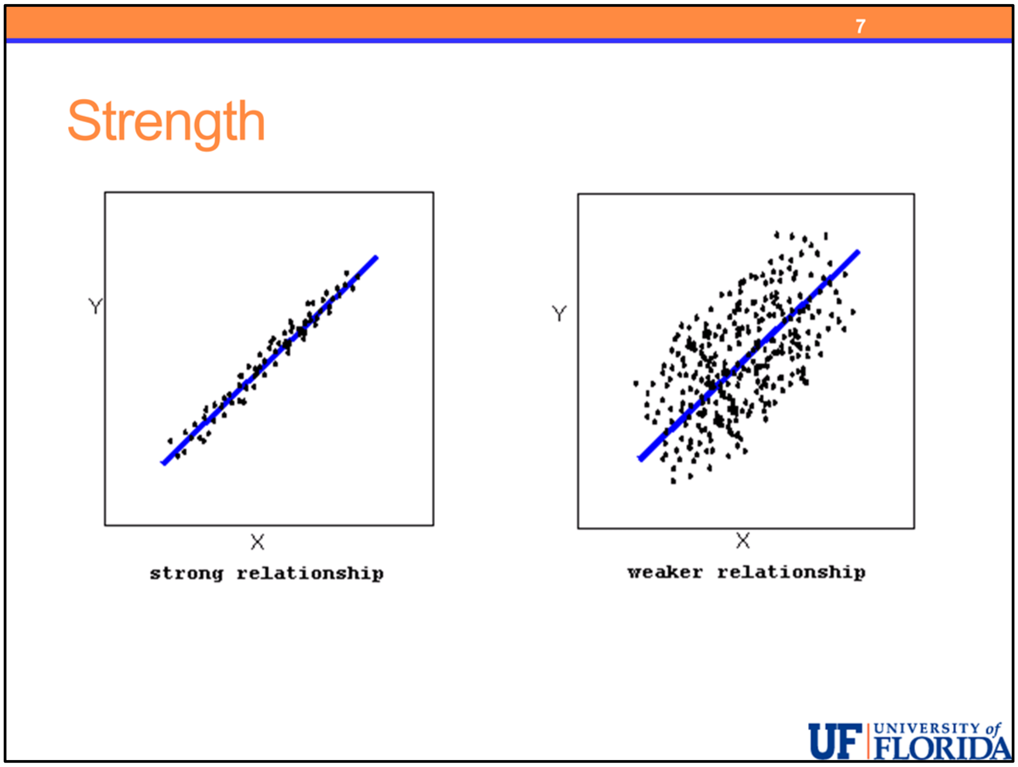
Form (Linear or Curvilinear)

The form of a relationship – in this class, we mostly want to distinguish between linear and not linear – which we will call curvilinear or non-linear.

We might also be interested in noting any odd patterns or groupings and attempting to explain the patterns we see.

The perfect graph will look similar to the linear graph displayed here – it will have a nice random scatter around a clear linear trend. The tighter the points are to the line, the better our predictions will be but linearity is the most important factor to being able to conduct most of the analyses we will learn in this course.

When relationships are not linear, often, we can perform transformations or use more complex models to provide better understanding and predictions. In such cases, the interpretability can become more difficult and any transformations or models used should fit into the theoretical framework of the topic under study.

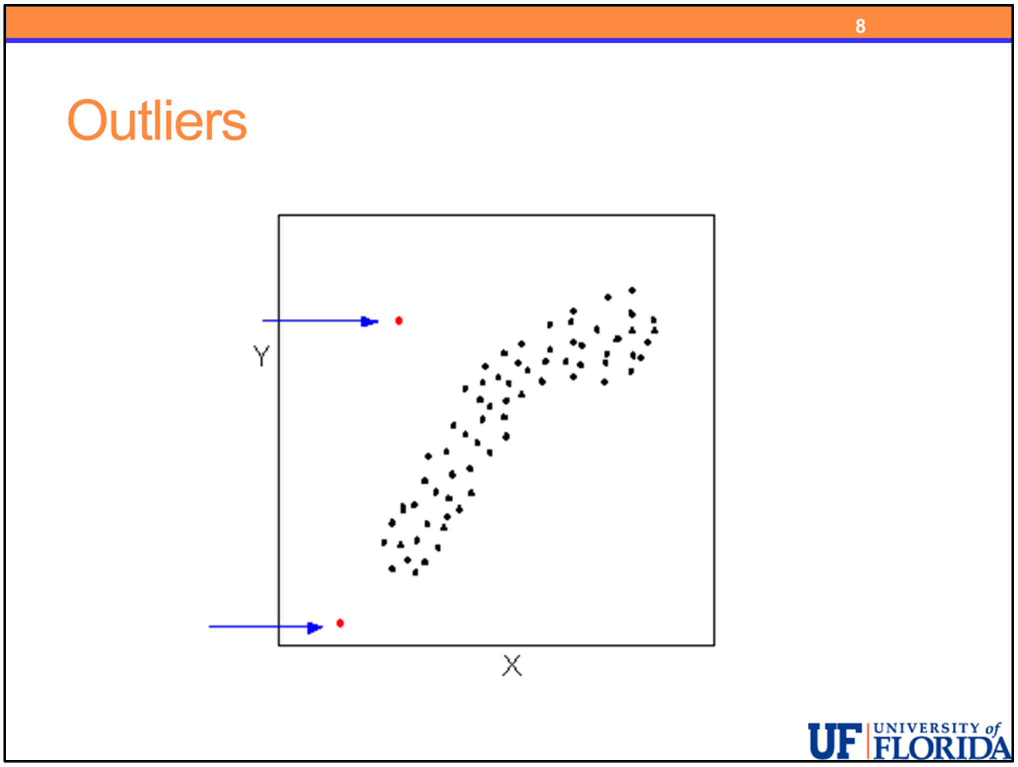


Strength

In addition to the direction (positive, negative, or neither) and the form (linear or non-linear), we wish to describe and measure the strength.

When given a comparison, it is usually easy to distinguish a stronger relationship from a weaker one. But we still will desire a measure of this strength.

Although the relationship on the right is weaker, this still would likely be of interest to many researchers. There is clearly an increasing trend. It is often simply a fact of the variation in nature that we see weak relationships in practice. Our interest is still to understand the underlying process and to try to explain as much of that variation as possible.



8

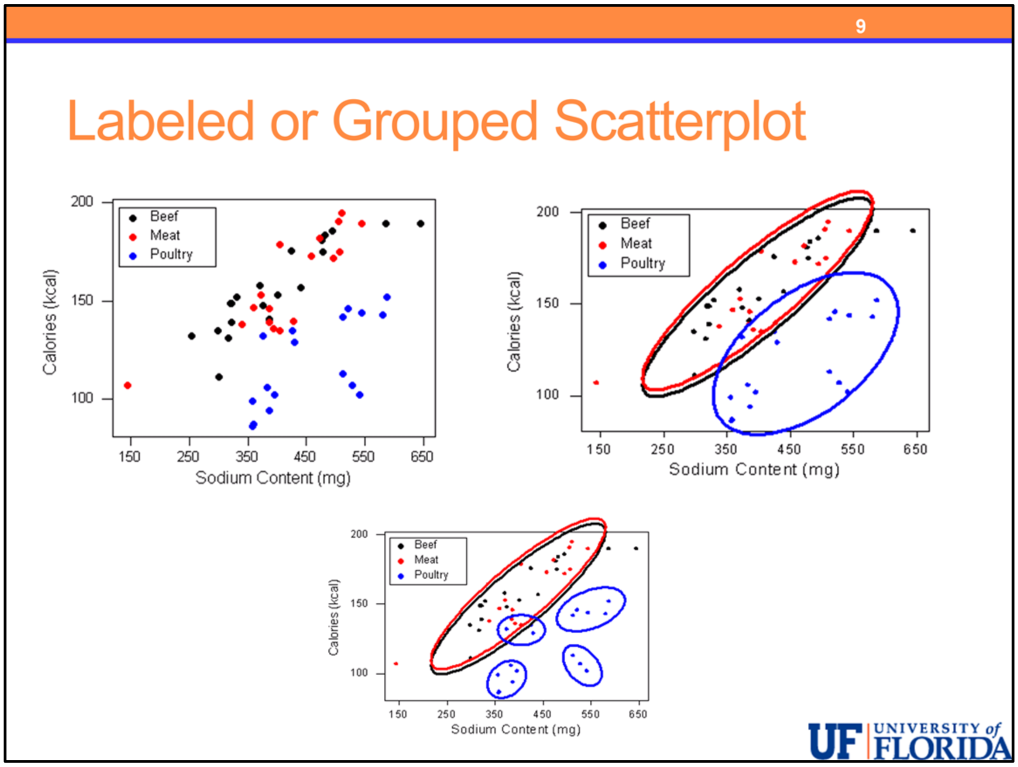
Outliers

Outliers are observations which fall outside of the pattern we see in our scatterplot. At this stage, we simply note them and possibly check to be sure they are not errors in our original data.

Values can be unusual in two main ways:

- Observations that fall in line with the trend but are unusually low or high for the dataset.
- Observations that don't even seem to fit the trend at all

From a scatterplot, we have no method of numerically identifying outliers but we can identify clearly unusual observations and locate them in the dataset ourselves or with some assistance from software.



9

Labeled or Grouped Scatterplot

Here, with the hot dogs we looked at earlier, now we are graphing the sodium content on the x-axis and calories on the y-axis and we have the different types (beef, other, poultry) denoted by colors.

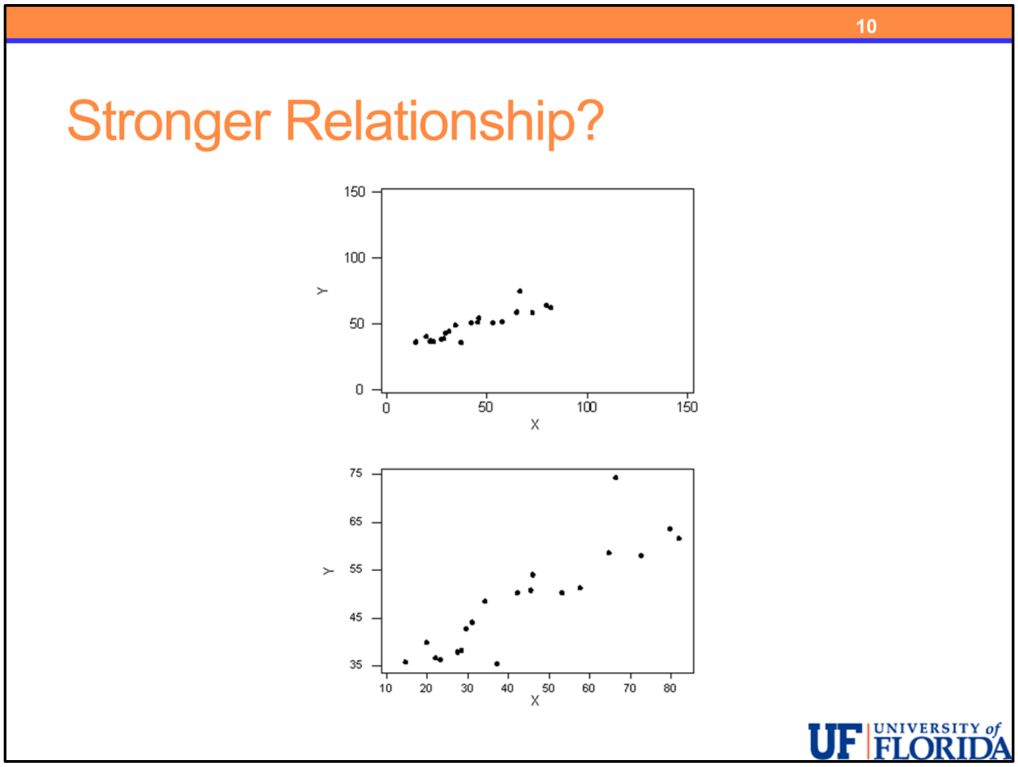
We call this a grouped or labeled scatterplot. This adds a third variable to our analysis. Here we can see a similar trend in the beef and “other meat” groups but the poultry group looks different. It is overall lower on the y-axis – which we know from our previous analysis. The sodium content in all three is a similar range but we would need boxplots and/or summaries to compare the sodium levels for the different types of hot dogs more closely.

The relationship for poultry is weaker than that of Beef and Meat. All three seem to indicate an overall increasing trend. Increased sodium tends to be associated with increased calories.

The apparent clustering in the poultry data points could be an anomaly or some sort of additional hidden pattern in the data.

We will delve deeper into the idea of lurking variables later, however, it is a good time to note that adding, considering, or adjusting for another variable in our analysis can change the interpretations considerably. It often takes a good grasp of statistics and a theoretical knowledge of the research question under study to have a good chance of determining the true causal relationships.

Finding associations is easy, explaining them carefully is a much more difficult task.

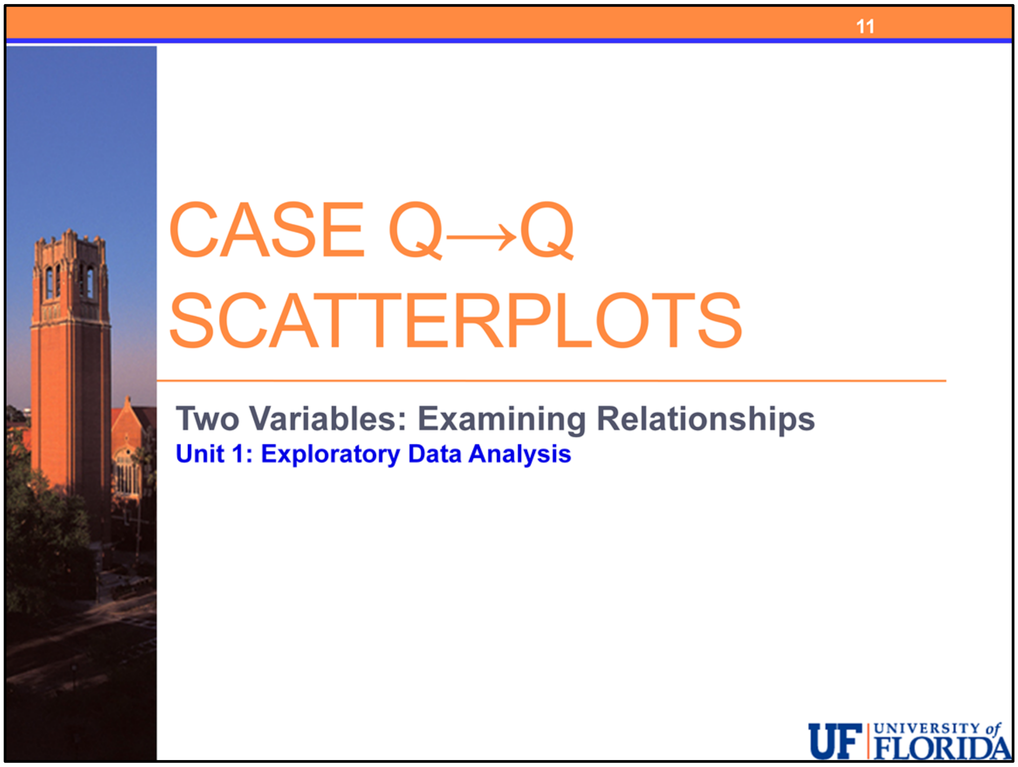


10

Stronger Relationship?

We end with a question. Which of these two is stronger? You can find the answer in the materials under Case Q-Q: Correlation.

{Web link: http://bolt.mph.ufl.edu/2012/12/24/learn-by-doing-strength-of-correlation/ }



CASE Q→Q SCATTERPLOTS

Two Variables: Examining Relationships

Unit 1: Exploratory Data Analysis

In Case Q-Q – our visual display is a scatterplot.

From this plot we can visualize the direction, form, and strength of a relationship that exists as well as any deviations from the overall trend of the relationship.

In the special case of linear relationships, we will discuss two methods of numerically summarizing data. Correlation and Regression.