Communicating Association Between Quantitative Variables

Dr Austin R Brown

Kennesaw State University

Introduction

- So far, we have generally been learning about methods of visualization for single variables.
- However, there are often situations in which the questions we are asking involve the association of variables.
- For instance, what if I wanted to visually assess the association between Major League Baseball (MLB) team runs scored and homeruns hit during the 2022 regular season?
  - Nothing we have learned so far can do this in an effective way.
- Instead, a tool called a "scatterplot" would be more effective.

Scatterplots

- A scatterplot, in general, is a visualization which exists on a Cartesian plane, like what we learned about back in high school algebra.
- It is common to plot individual points, which contain quantitative x and y coordinates.
- With this visualization, we can get a sense of the relationship or association which may exist between the two quantitative variables we are interested in!
  - So let's take a look at our own example!

Scatterplots

- Of course, we must first get our data into the right format prior to visualization:

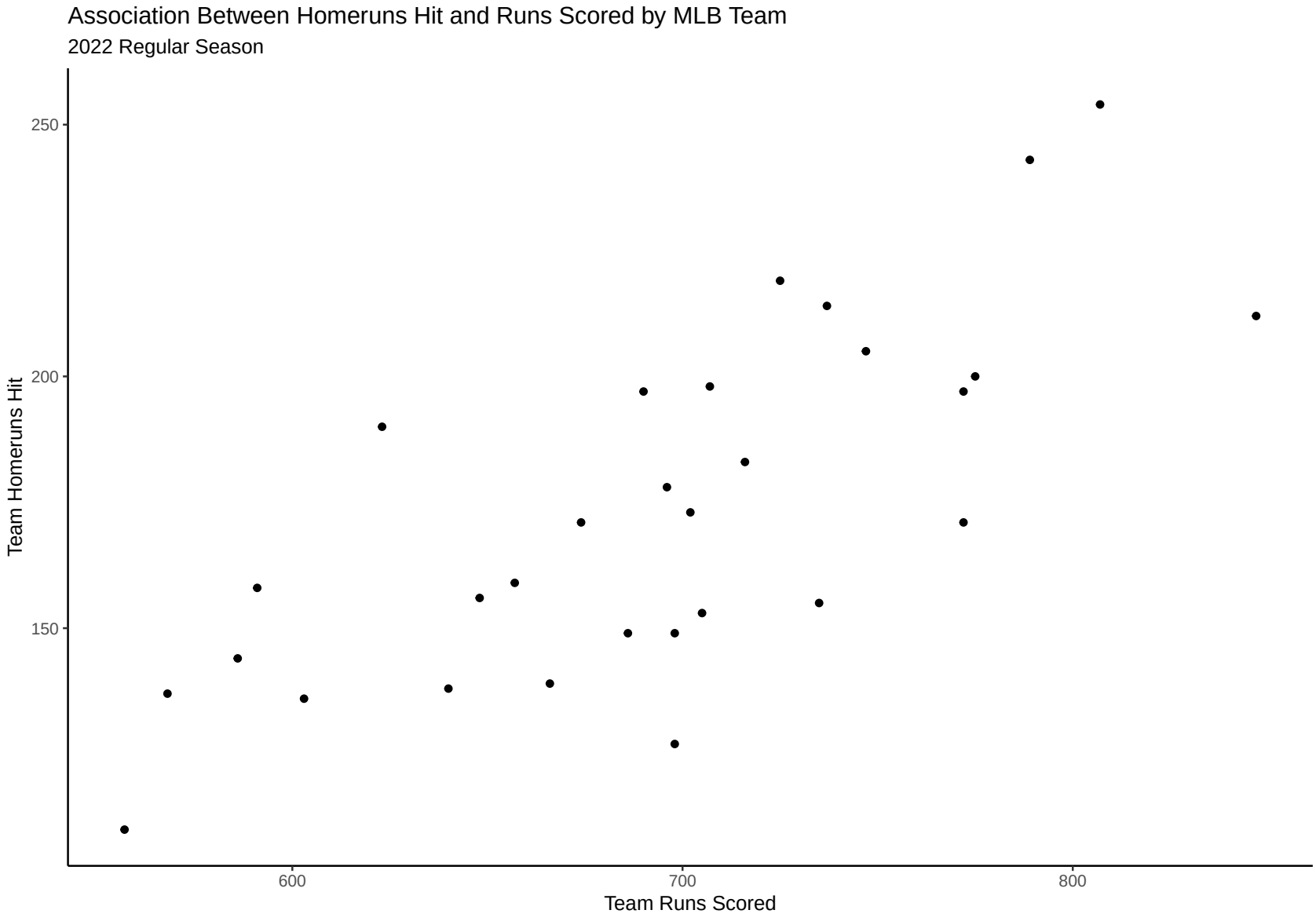
```
library(tidyverse)
mlb <- Lahman::Batting |>
  select(yearID,teamID,R,HR) |>
  filter(yearID == 2022) |>
  group_by(teamID) |>
  summarize(R = sum(R,na.rm=T),
            HR = sum(HR,na.rm=T))
```

Scatterplots

- Now, we can use `geom_point` in order to visualize this association:

```
mlb |>
  ggplot(aes(x=R,y=HR)) + geom_point() +
  labs(x="Team Runs Scored",
       y="Team Homeruns Hit",
       title="Association Between Homeruns Hit and Runs Scored by MLB Team",
       subtitle="2022 Regular Season") +
  theme_classic()
```

Scatterplots



Scatterplots: Interpretations

- Okay great! We probably know how to generally interpret a scatterplot. Here, it seems as the number of runs scored increases, intutitively homeruns hit also increases.

- But we can be a little bit more specific in our interpretations of a scatterplot by answering the below questions:
  1. What is the direction of the relationship?
  2. What is the form of the relationship?
  3. What is the strength of the form of the relationship?
  4. What unusual characteristics are exhibited?

- Let's talk more specifically about how to answer these questions.

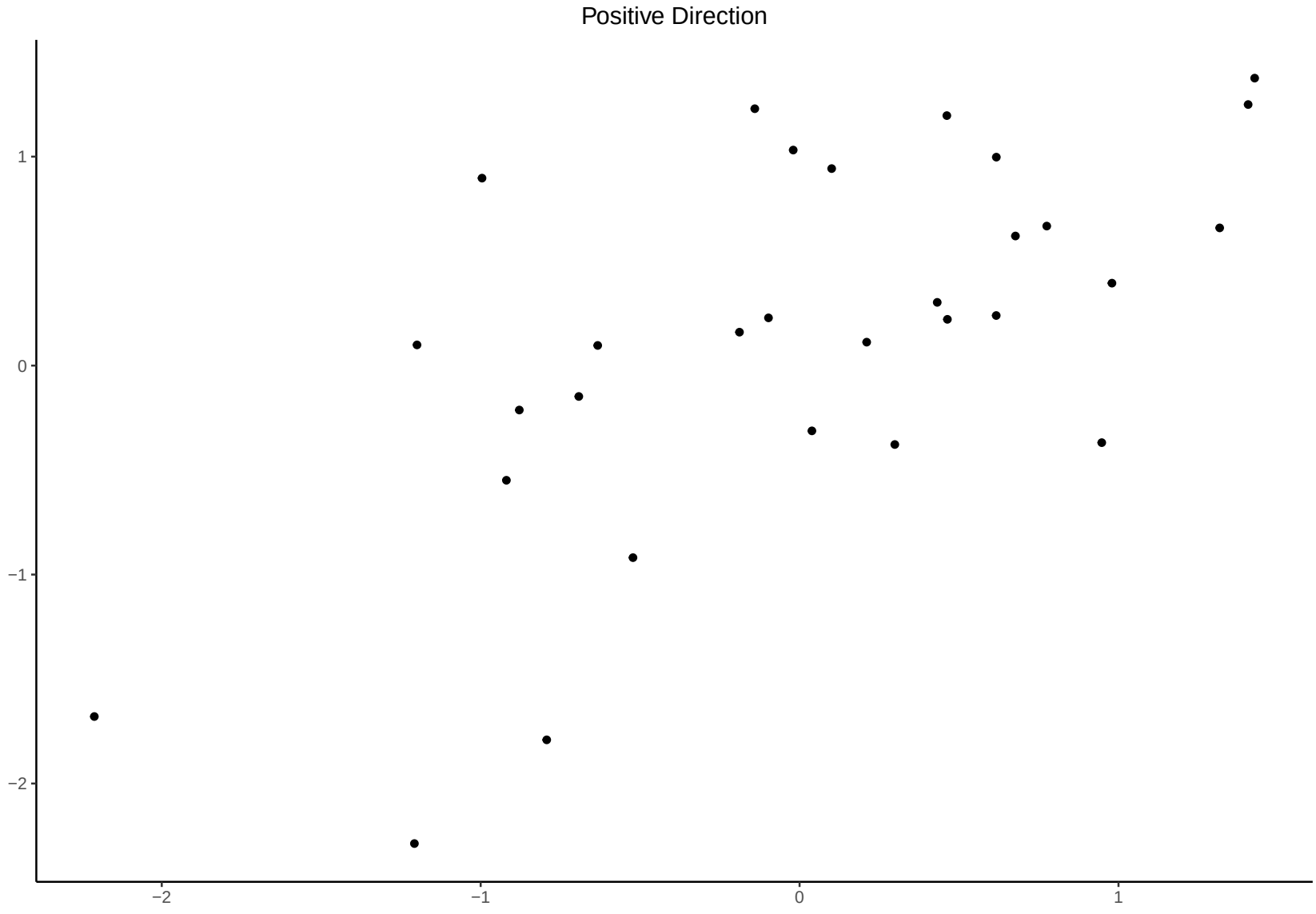
Scatterplots: Direction

- **What is the direction of the relationship?**

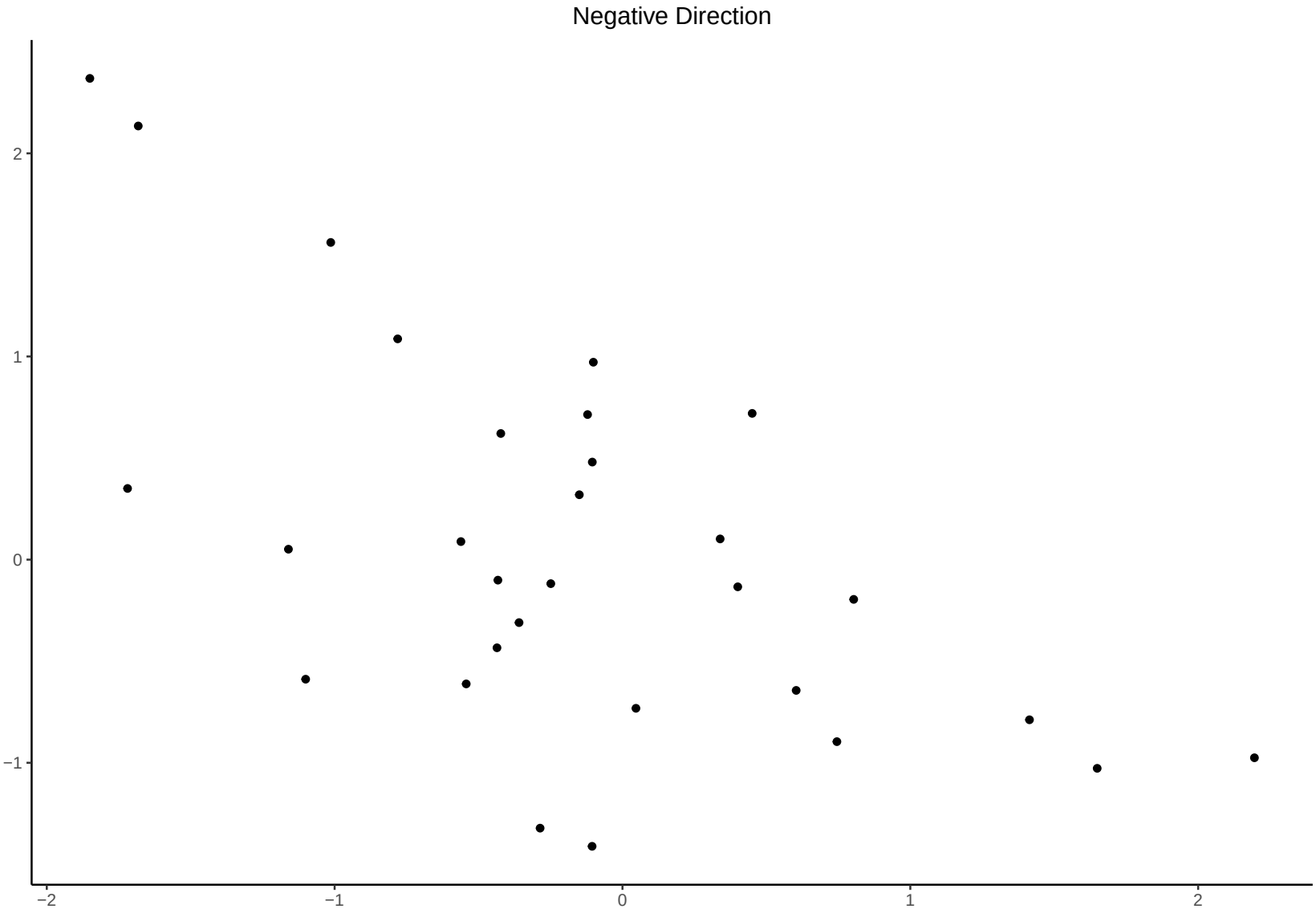
- "Direction" refers to how the points "move" together. If as the values on the x-axis increase, the values on the y-axis also increase, meaning that we have a general upward direction moving left to right across the graph, then we say the direction is "positive".

- If, as the values on the x-axis increase, the values on the y-axis decrease, meaning that we have a general downward direction moving left to right across the graph, then we say the direction is "negative".

Scatterplots: Direction



Scatterplots: Direction



Scatterplots: Form

- **What is the form of the relationship?**
- When we talk about the "form" of the relationship, we are referring to the general pattern the points follow.
- For me, I usually refer to two main "forms": Linear and Non-Linear

Scatterplots: Form

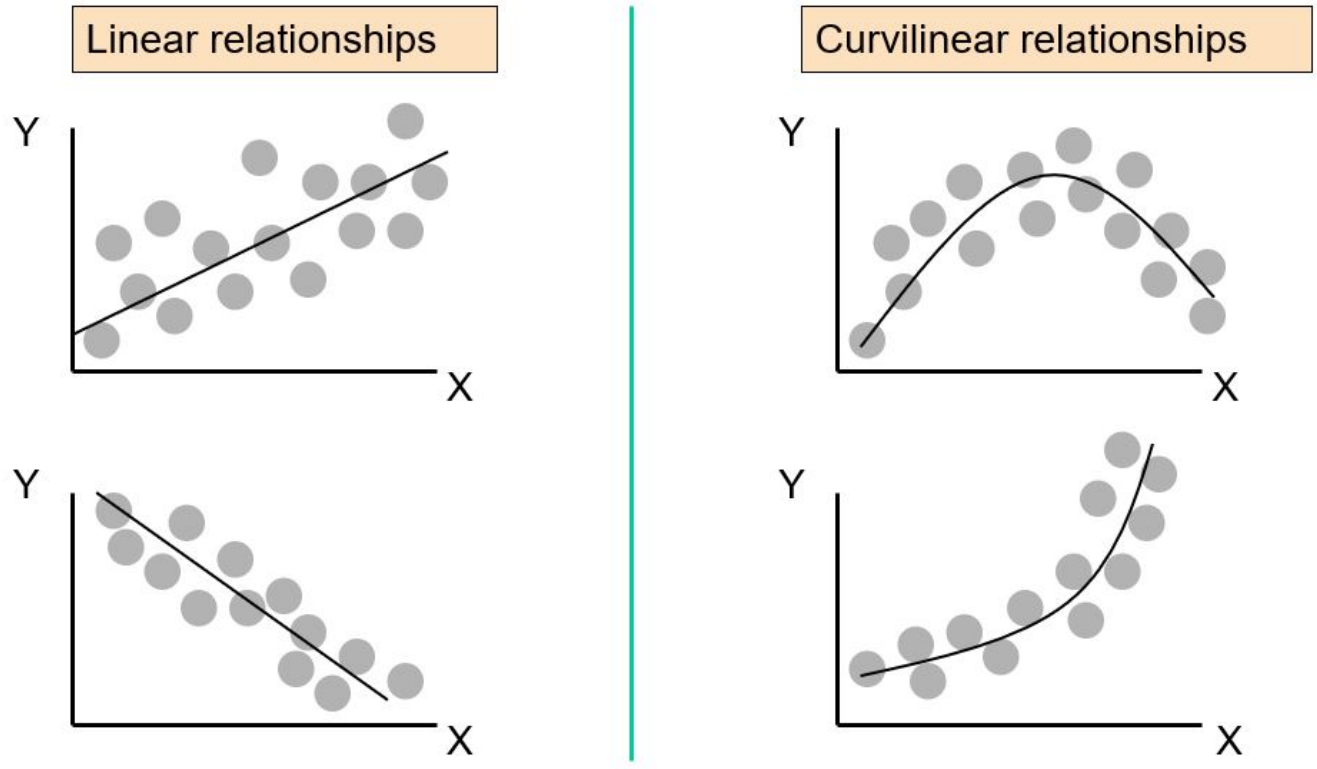


Figure 1: From: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall, c/o Dr. Taasoobshirazi

Scatterplots: Strength

- **What is the strength of the form of the relationship?**
- What we're talking about with "strength" is how close the points fall to the general form of the relationship identified in the prior question.
- We can use adjectives like "weak", "moderate", and "strong" to describe the strength.

Scatterplots: Stregnth

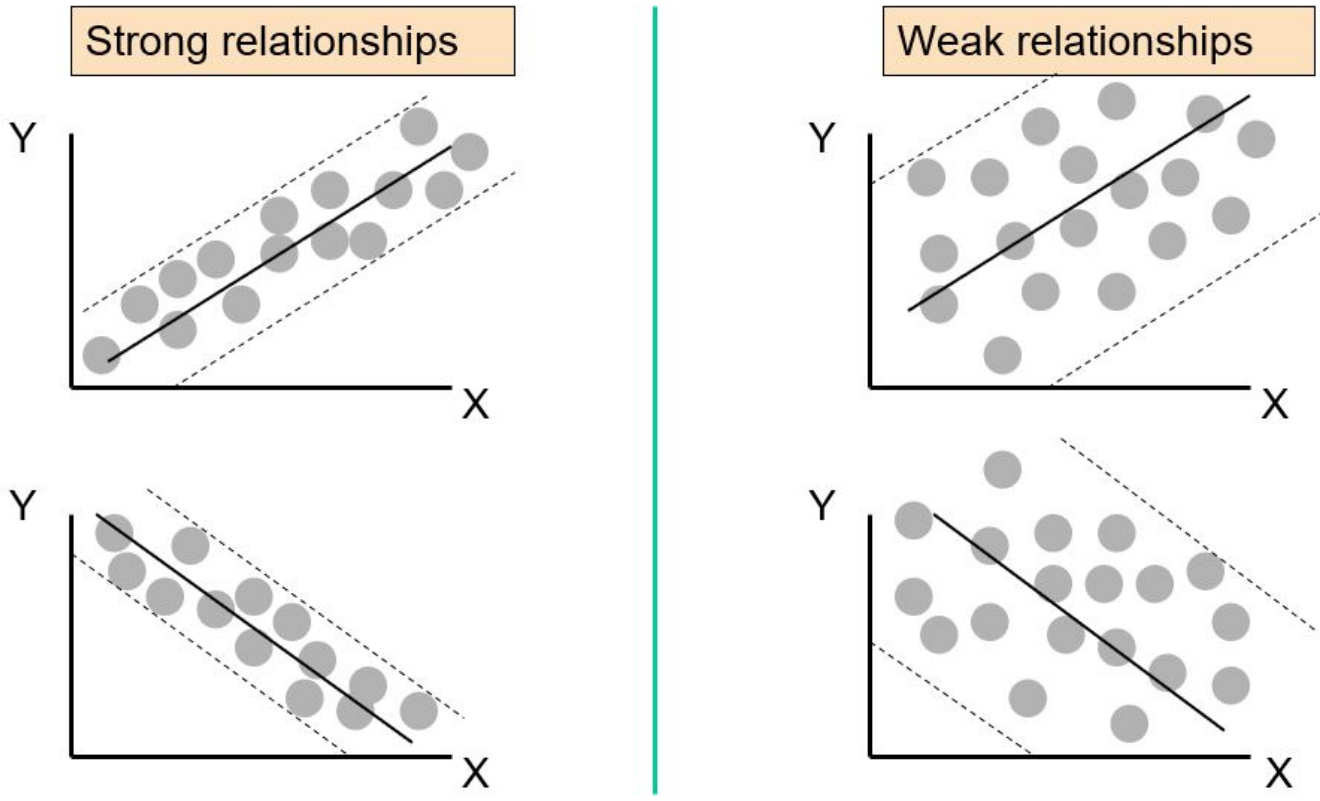
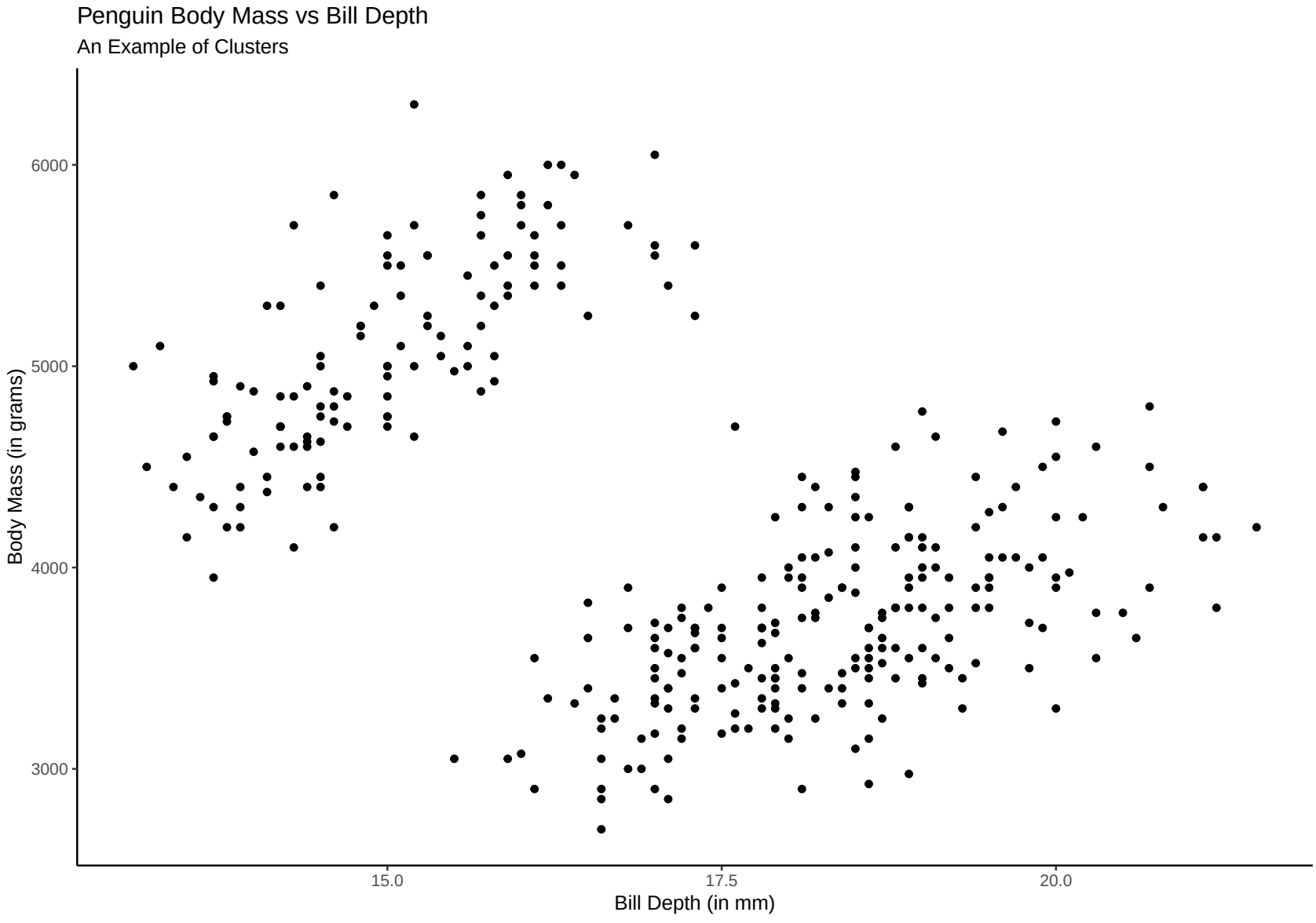


Figure 2: From: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall, c/o Dr. Taasoobshirazi

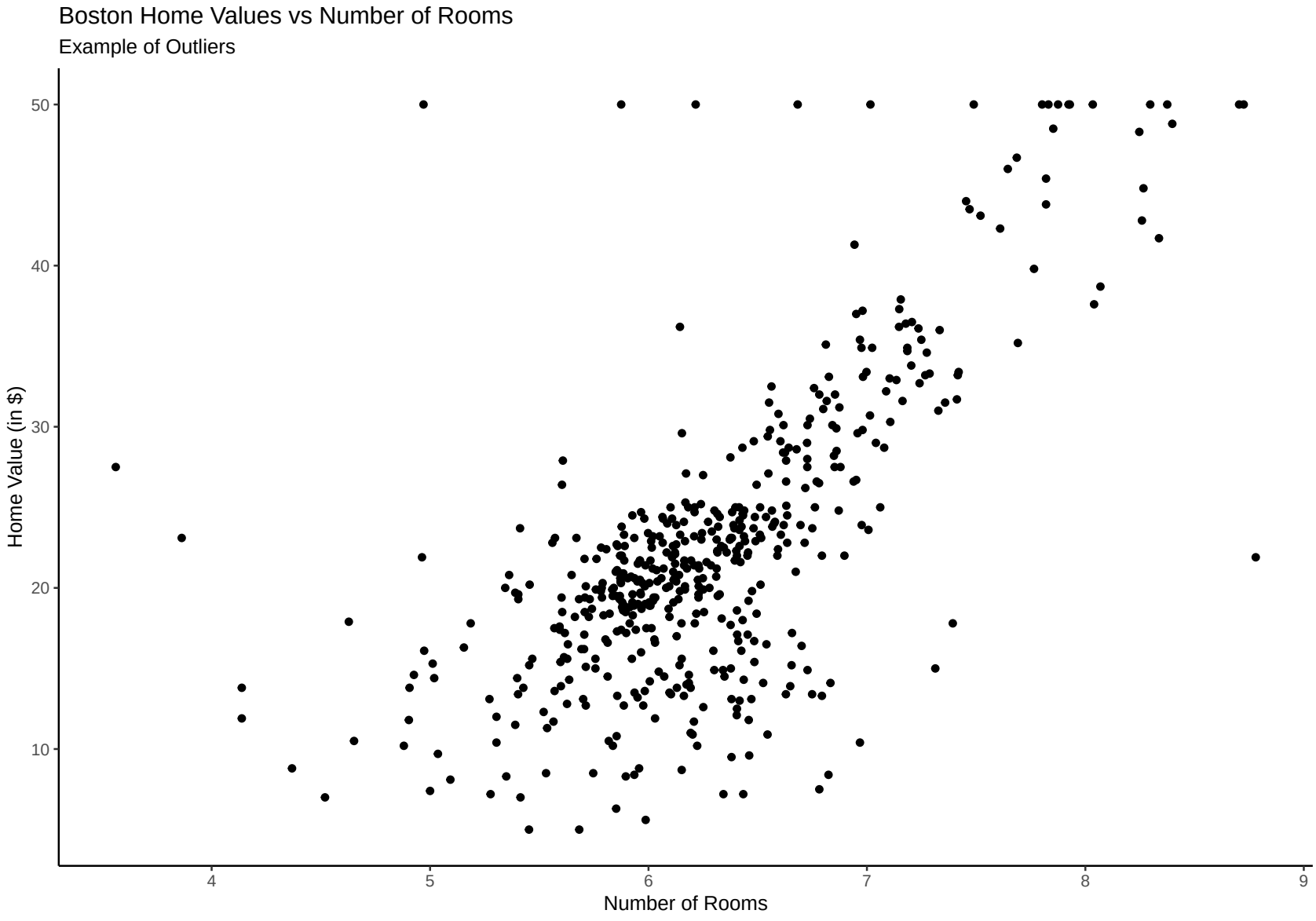
Scatterplots: Unusual Characteristics

- **What unusual characteristics are exhibited?**
- What we mean by unusual characterstics is really anything that just visually appears odd.
- Generally these are things like clusters or outliers, but could be anything really.
  - Let's look at a clustering example using the Penguins data and an outlier example using the Boston data.

Scatterplots: Unusual Characteristics

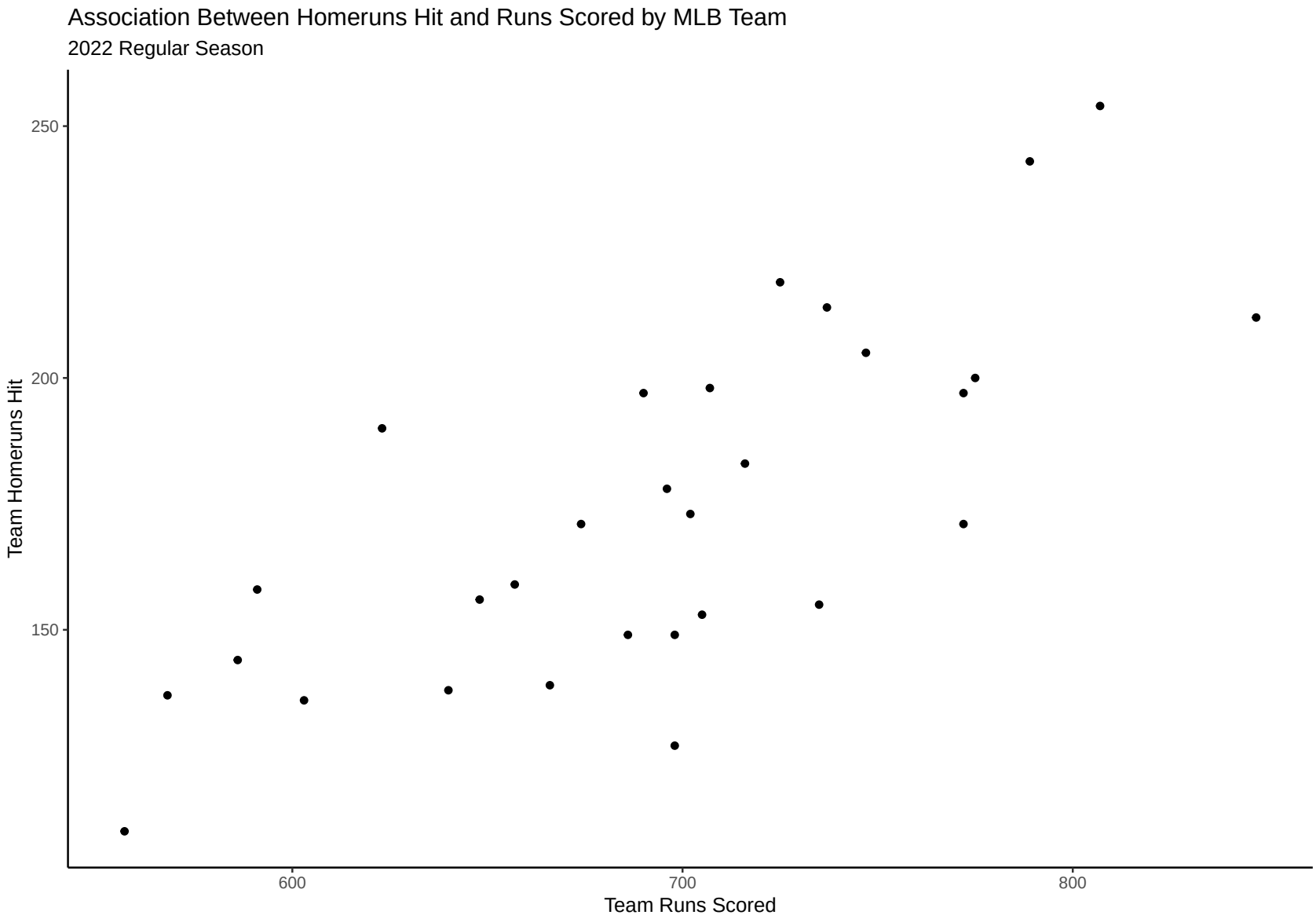


Scatterplots: Unusual Characteristics



Scatterplots: Baseball Example

- Consider again our scatterplot:



Scatterplots: Baseball Example

- If I were to interpret this scatterplot, I would say we have evidence for a positive, linear relationship between runs scored and homeruns hit of moderately strong strength with no clear unusual characterstics.

- Obviously there is a good deal of subjectivity in these interpretations, so on assignments where I ask you for interpretations, there aren't necessarily right or wrong answers.

Scatterplots: Adding Regression Line & Regression Equation

- At the beginning of the semester, an early lecture on general `ggplot2` functionality showed you how to modify certain elements of a scatterplot, including changing point size and color, adding annotations, and general things of the sort!

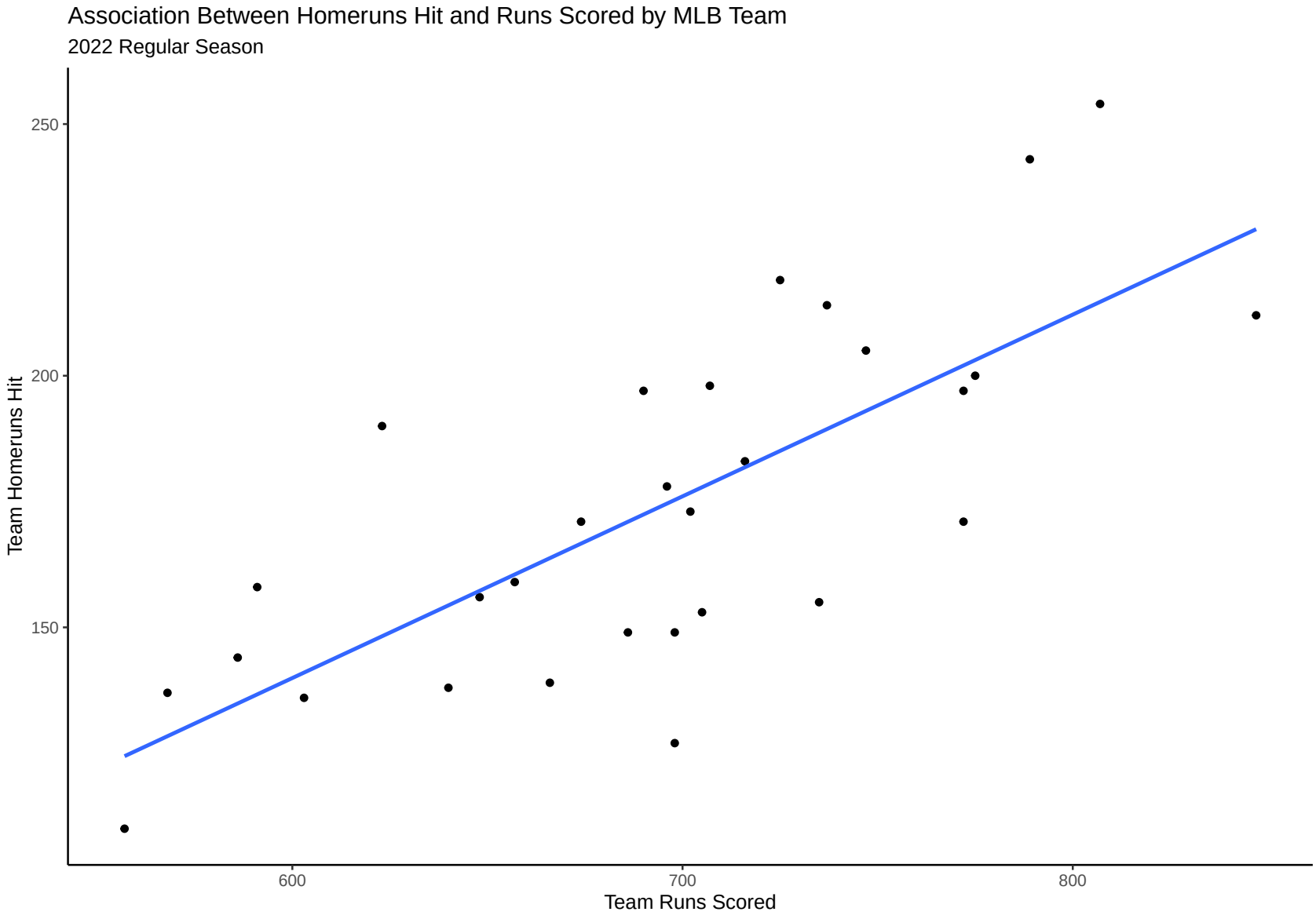
- Very commonly when we are generating scatterplots, it is often of interest to us to determine an equation for a line which best explains the relationship we are visually interpreting.
  - This "line" is called a "simple linear regression (SLR)" line.

- How can we include this line and the equation itself on our plots?
  - Let's take a look!

Scatterplots: Adding Regression Line & Regression Equation

```
mlb |>
  ggplot(aes(x=R,y=HR)) + geom_point() +
  geom_smooth(method='lm',se=F) +
  labs(x="Team Runs Scored",
       y="Team Homeruns Hit",
       title="Association Between Homeruns Hit and Runs Scored by MLB Team",
       subtitle="2022 Regular Season") +
  theme_classic()
```

Scatterplots: Adding Regression Line & Regression Equation



Scatterplots: Adding Regression Line & Regression Equation

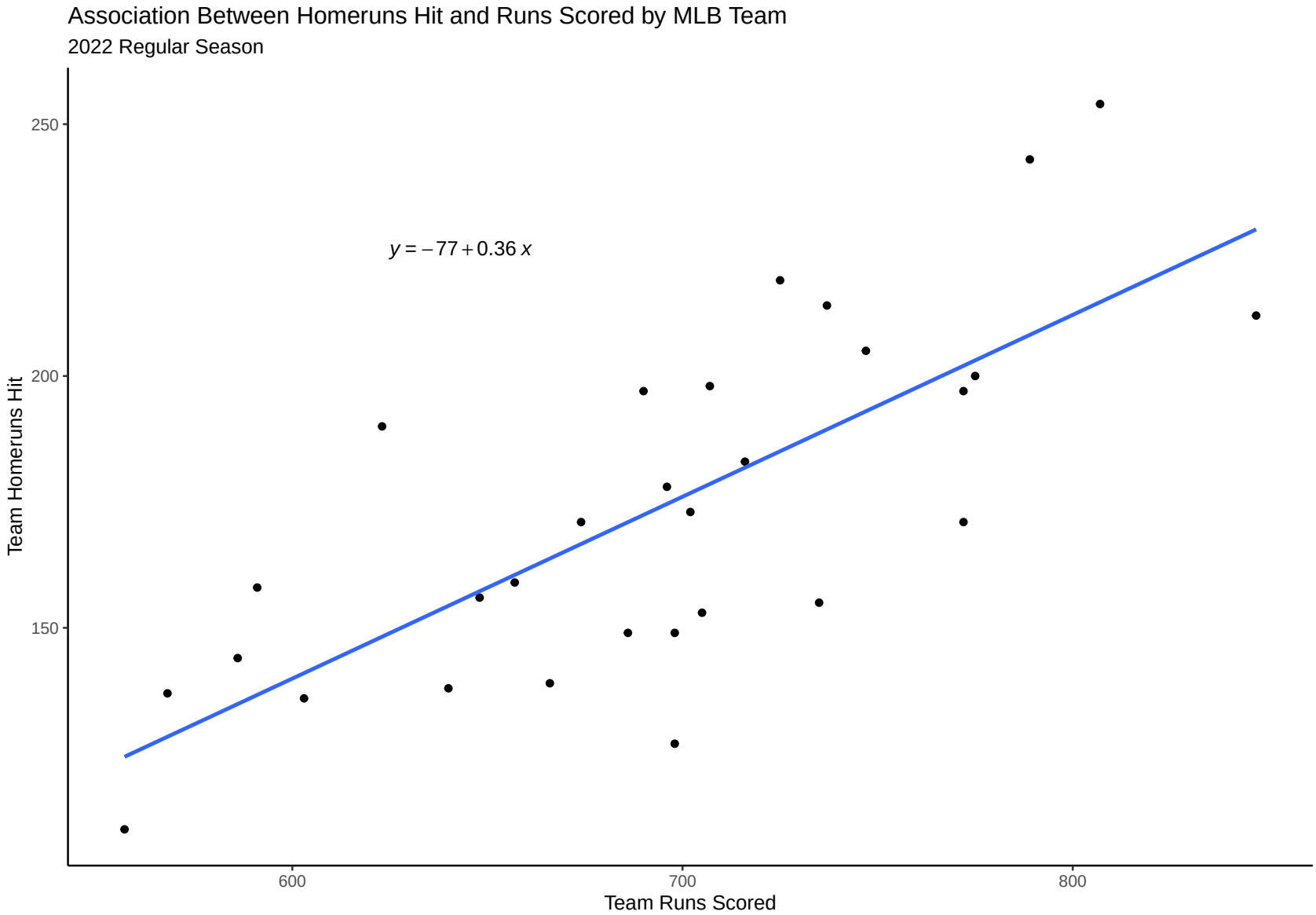
- In the code, notice we added a new geom called `geom_smooth` which allows us to visualize lines or curves around data points.
  - Note, `method='lm'` refers to the line to be fit, which is a linear model in this case. `se=F` means that we don't want the standard errors of the line to be rendered on the visualization.

- Okay this looks great! But now how do we get the equation of that line to also render on the graph?
  - Here, we can use a nice function called `stat_regline_equation` which is part of the helper package `ggpubr`

Scatterplots: Adding Regression Line & Regression Equation

```
library(ggpubr)
mlb |>
  ggplot(aes(x=R,y=HR)) + geom_point() +
  geom_smooth(method='lm',se=F) +
  stat_regline_equation(label.x = 625, label.y = 225) +
  labs(x="Team Runs Scored",
       y="Team Homeruns Hit",
       title="Association Between Homeruns Hit and Runs Scored by MLB Team",
       subtitle="2022 Regular Season") +
  theme_classic()
```

Scatterplots: Adding Regression Line & Regression Equation



Scatterplots: Adding Regression Line & Regression Equation

- Notice that the `label.x` and `label.y` arguments control the position of the regression equation.
- We can also control the size and color of the text as well!

Faceting Scatterplots

- Obviously with the `Lahman` data, there are lots of different ways that we can slice the data to answer different questions.
- For instance, in the prior visualization, we observed the relationship between homeruns hit and runs scored for a single, static season.
  - What if we wanted to see how that relationship changed from say, 2019 to 2022?
- We can use either `facet_wrap` or `facet_grid` to help us here!