Using the Grammar of Graphics in Applied Economic Research

Andrew J. Van Leuven

September 30, 2022

DEPARTMENT OF
AGRICULTURAL ECONOMICS

Today's Objectives

- Why visualize data?
- Data visualization basics
- How to visualize data using the ggplot2 package in R

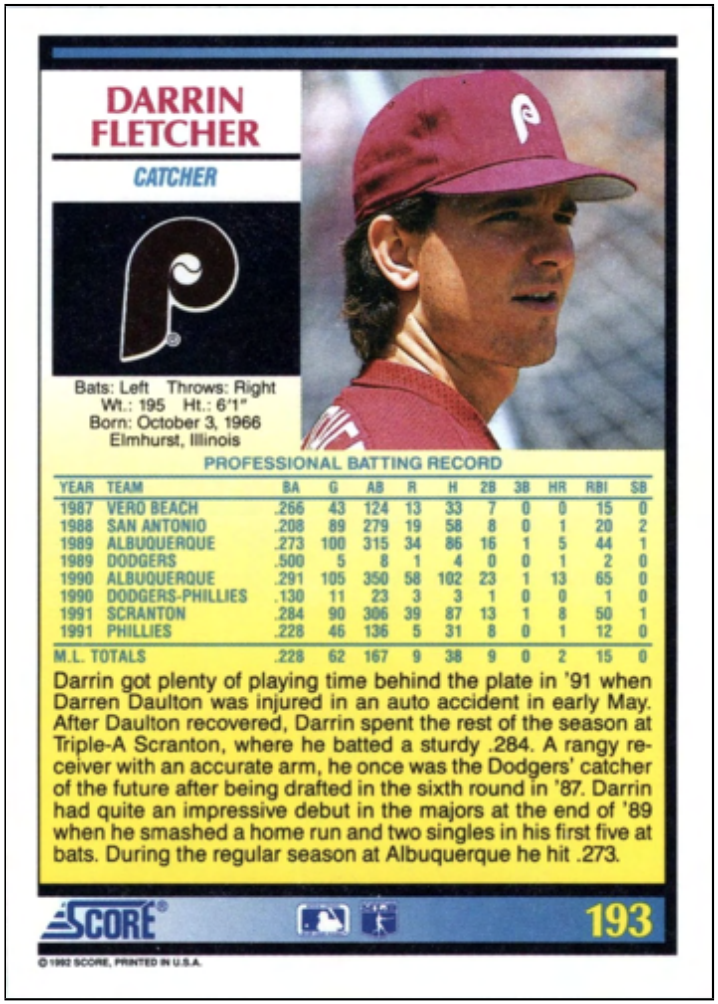
1 Why Visualize Data?

2 Some Data Visualization Basics

3 Data Visualization Using the R Statistical Language

DEPARTMENT OF
AGRICULTURAL ECONOMICS

What is Data?



DARRIN FLETCHER

CATCHER

Bats: Left Throws: Right
Wt.: 195 Ht.: 6'1"
Born: October 3, 1966
Elmhurst, Illinois

PROFESSIONAL BATTING RECORD

YEAR	TEAM	BA	G	AB	R	H	2B	3B	HR	RBI	SB
1987	VERO BEACH	.266	43	124	13	33	7	0	0	15	0
1988	SAN ANTONIO	.208	89	279	19	58	8	0	1	20	2
1989	ALBUQUERQUE	.273	100	315	34	86	16	1	5	44	1
1989	DODGERS	.500	5	8	1	4	0	0	1	2	0
1990	ALBUQUERQUE	.291	105	350	58	102	23	1	13	65	0
1990	DODGERS-PHILLIES	.130	11	23	3	3	1	0	0	1	0
1991	SCRANTON	.284	90	306	39	87	13	1	8	50	1
1991	PHILLIES	.228	46	136	5	31	8	0	1	12	0
M.L. TOTALS		.228	62	167	9	38	9	0	2	15	0

Darrin got plenty of playing time behind the plate in '91 when Darren Daulton was injured in an auto accident in early May. After Daulton recovered, Darrin spent the rest of the season at Triple-A Scranton, where he batted a sturdy .284. A rangy receiver with an accurate arm, he once was the Dodgers' catcher of the future after being drafted in the sixth round in '87. Darrin had quite an impressive debut in the majors at the end of '89 when he smashed a home run and two singles in his first five at bats. During the regular season at Albuquerque he hit .273.

SCORE 193

- Data is plural for *datum*, which refers to a piece of information
- We use data to answer questions or to have a better idea of what is going on
- Data are rectangular:
  - Rows are observations (horizontal)
  - Columns are variables (vertical)

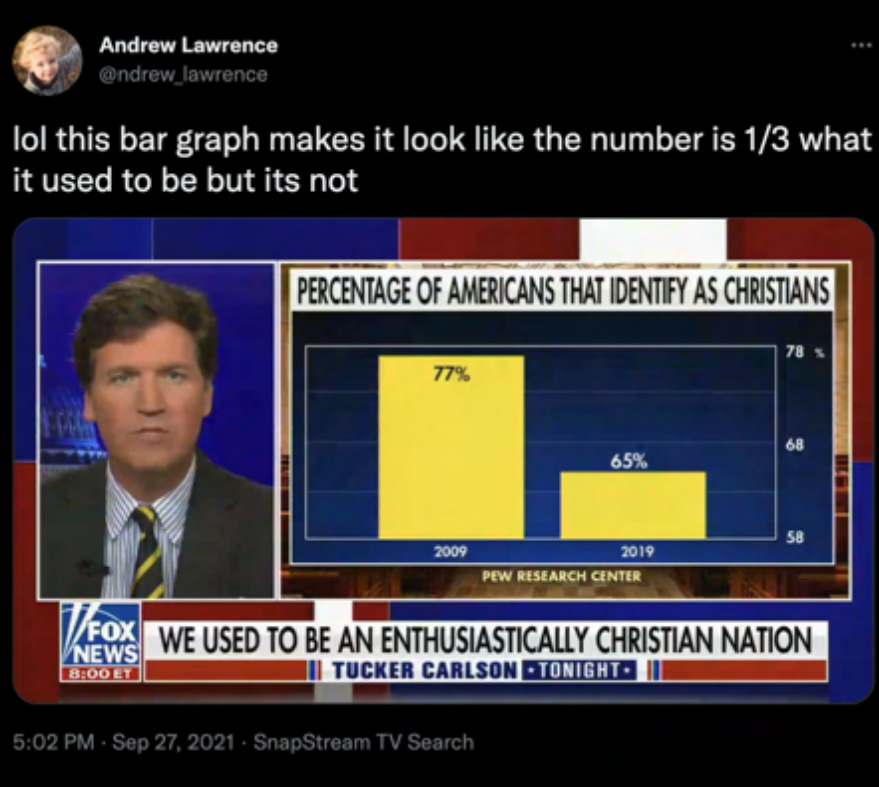
DEPARTMENT OF AGRICULTURAL ECONOMICS

Types of Data

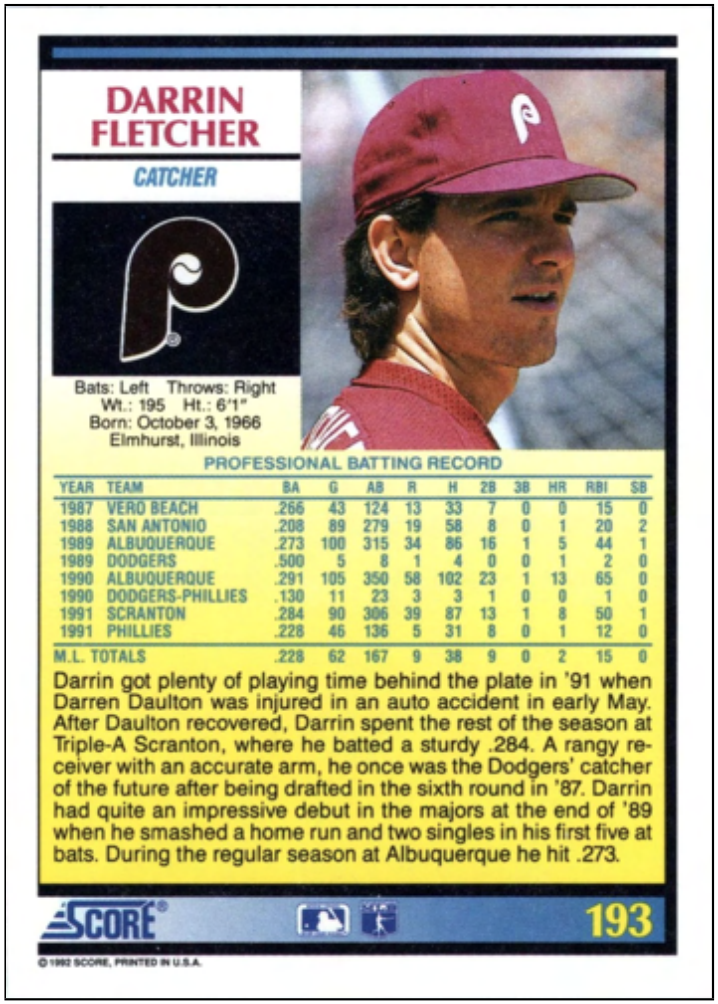
- **Quantitative Data**: can be measured using numbers; answers questions like "how many?" and "how often?"
  - **Continuous**: fractions make sense (1.5 hot dogs, 3.14 slices of pie)
  - **Discrete**: fractions don't make sense (1.5 dogs, 3.14 employees working at the pie stand)
- **Qualitative Data**: describes the attributes or properties that an object possesses
  - **Nominal**: uses names or categories (public or private, male or female)
  - **Ordinal**: uses a ranking or ordering system (A+/B/C- or always/sometimes/never)

DEPARTMENT OF AGRICULTURAL ECONOMICS

Why Visualize Data?



So, Why Visualize Data?



DARRIN FLETCHER

CATCHER

Bats: Left Throws: Right
Wt.: 195 Ht.: 6'1"
Born: October 3, 1966
Elmhurst, Illinois

PROFESSIONAL BATTING RECORD

YEAR	TEAM	BA	G	AB	R	H	2B	3B	HR	RBI	SB
1987	VERO BEACH	.266	43	124	13	33	7	0	0	15	0
1988	SAN ANTONIO	.208	89	279	19	58	8	0	1	20	2
1989	ALBUQUERQUE	.273	100	315	34	86	16	1	5	44	1
1989	DODGERS	.500	5	8	1	4	0	0	1	2	0
1990	ALBUQUERQUE	.291	105	350	58	102	23	1	13	65	0
1990	DODGERS-PHILLIES	.130	11	23	3	3	1	0	0	1	0
1991	SCRANTON	.284	90	306	39	87	13	1	8	50	1
1991	PHILLIES	.228	46	136	5	31	8	0	1	12	0
M.L. TOTALS		.228	62	167	9	38	9	0	2	15	0

Darrin got plenty of playing time behind the plate in '91 when Darren Daulton was injured in an auto accident in early May. After Daulton recovered, Darrin spent the rest of the season at Triple-A Scranton, where he batted a sturdy .284. A rangy receiver with an accurate arm, he once was the Dodgers' catcher of the future after being drafted in the sixth round in '87. Darrin had quite an impressive debut in the majors at the end of '89 when he smashed a home run and two singles in his first five at bats. During the regular season at Albuquerque he hit .273.

SCORE 193

© 1992 SCORE, PRINTED IN U.S.A.

- Tables are great! But it can be hard to look for patterns.
- One data point is a statistic. Many data points are a story.
- Visualizing data helps tell that story.

DEPARTMENT OF AGRICULTURAL ECONOMICS

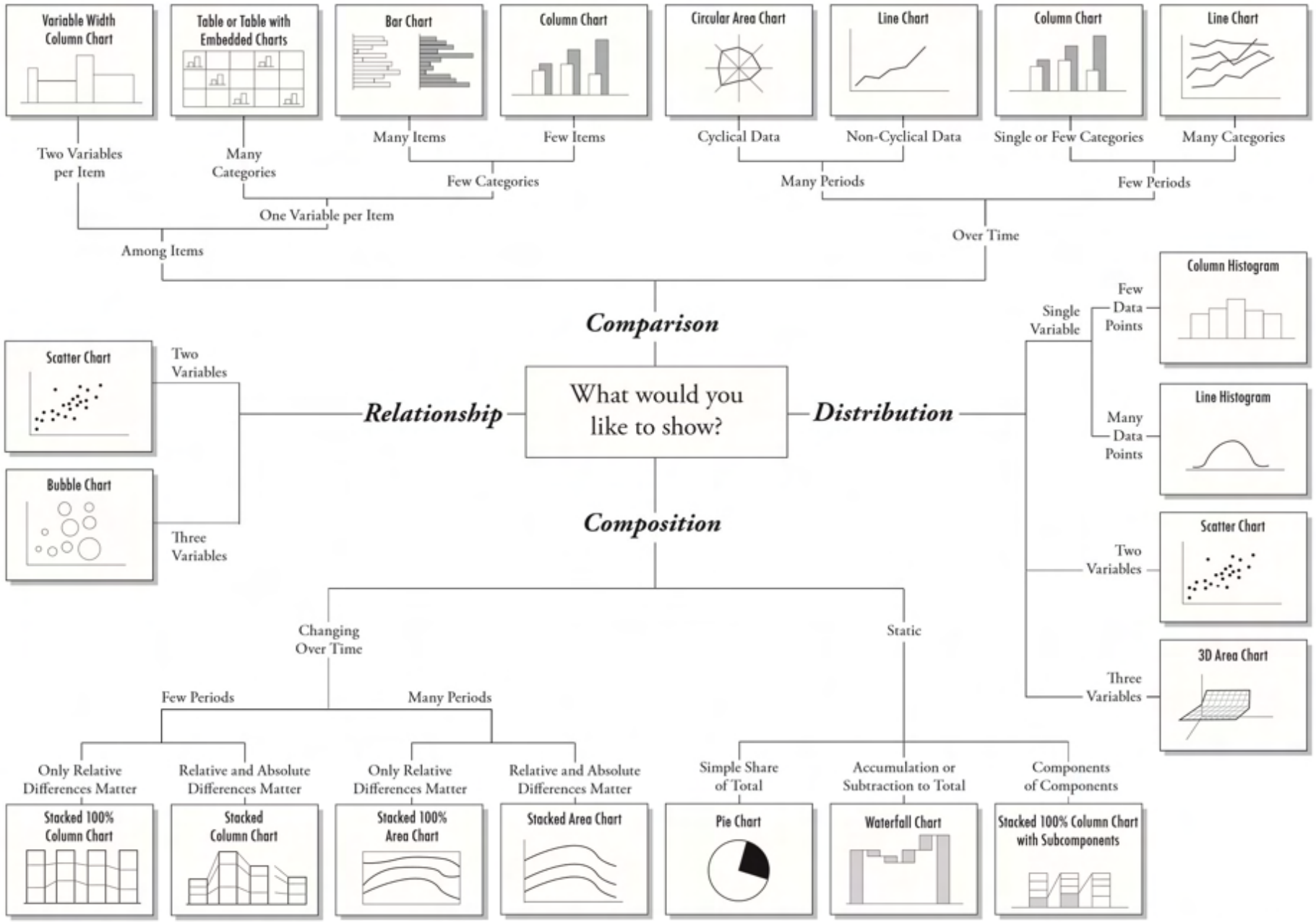
1 Why Visualize Data?

2 Some Data Visualization Basics

3 Data Visualization Using the R Statistical Language

DEPARTMENT OF
AGRICULTURAL ECONOMICS

Chart Suggestions—A Thought-Starter



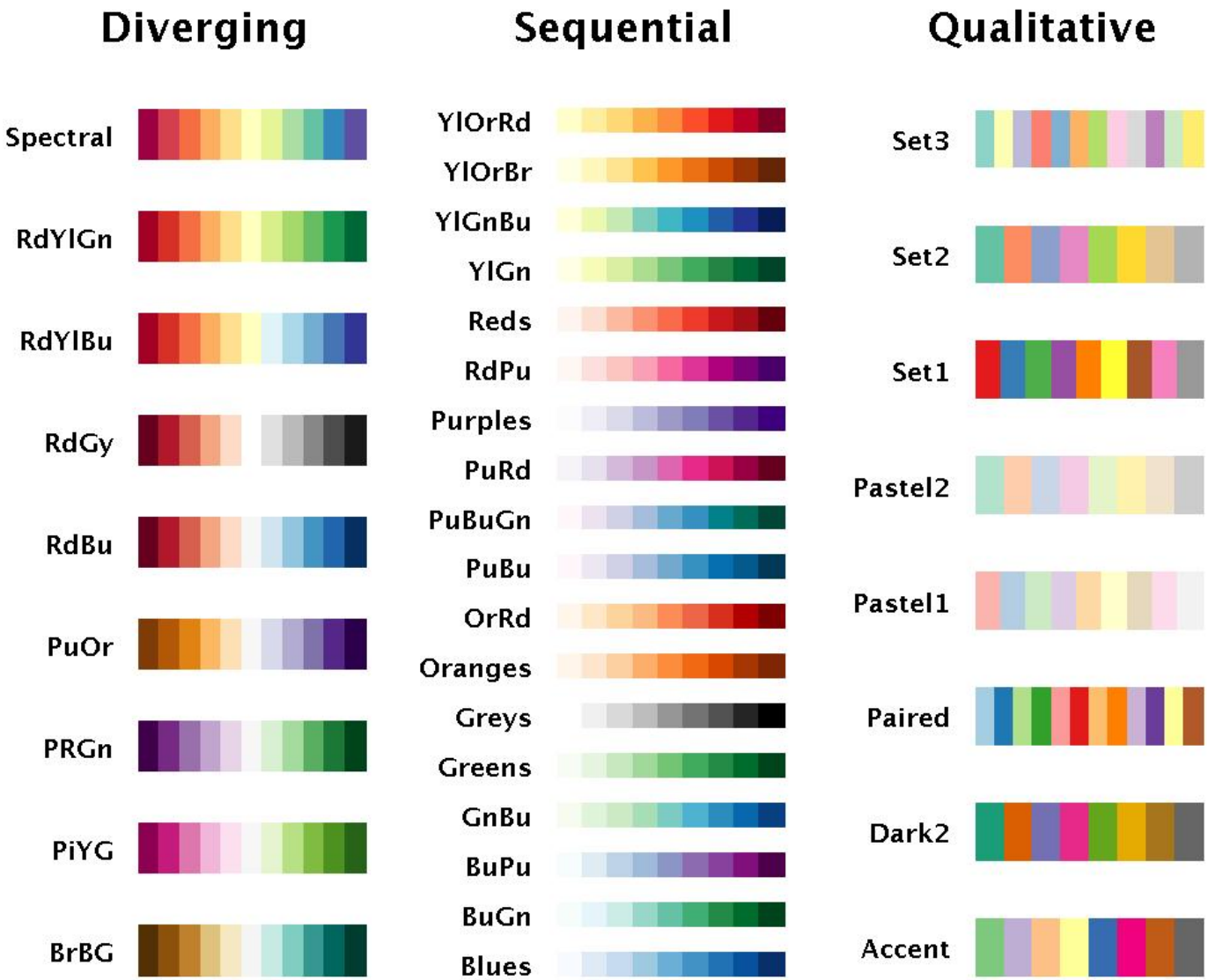
Visualizing Data: Where to Start?

Deciding how to visually represent data depends on the answer to the question:

"What would you like to show?"

- **Distribution**: shows all the possible values of the data and quantifies the relative frequency (how often they occur)
- **Relationship**: shows two or more variables and their properties (such as correlation)
- **Comparison**: shows how changes in one data point relate to changes in another data point(s)
- **Composition**: show how individual parts make up the whole
- **Connection**: show how people, things, or organizations relate to one another
- **Location**: show spatial distributions and relationships (maps)

Thinking About Color



1. Why Visualize Data?
2. Some Data Visualization Basics
3. Data Visualization Using the R Statistical Language

DEPARTMENT OF AGRICULTURAL ECONOMICS

What is R?

- Open-source (free) statistical programming language
- Mostly used by statisticians and "data scientists"
- R $\neq$ RStudio. R is just a command-line console, while RStudio is an IDE (integrated development environment) that makes R more user-friendly.
- The "base" R syntax is decades old. A newer subset (dialect?) of the language known as "tidy" R has risen in popularity. The `tidyverse` shares "an underlying design philosophy, grammar, and data structures."
- The `ggplot2` package is part of the tidyverse

DEPARTMENT OF AGRICULTURAL ECONOMICS

What is GGPlot2?

- An open-source data visualization package for R
- A replacement for R's "base" graphics plotting option
- Created by Hadley Wickham in 2005, `ggplot2` is an implementation of Leland Wilkinson's Grammar of Graphics—a general scheme for data visualization which breaks up graphs into semantic components such as scales and **layers**.
- From Cédric Scherer's online course:
  - the components follow a consistent syntax
  - each ggplot needs at least data, some aesthetics, and a layer
  - users set constant properties outside `aes()`
  - users map data-related properties inside `aes()`

DEPARTMENT OF
AGRICULTURAL ECONOMICS

Let's Set Some Expectations

- If you've never used R (or even "tidy" R), you will most likely *not* come away from this hour having learned the R statistical language! But I still believe it will be a helpful exercise.
- I will [try to] answer any questions you have today, but I'm not afraid to say "I don't know." Please don't hesitate to email me and I promise I will try to answer any questions you have.
- Finally, please know that the way I use R is not "the right way" to use R. It's not the wrong way either! It's just the way that works for me. I encourage you to adapt my code to whatever system works for you!

DEPARTMENT OF
AGRICULTURAL ECONOMICS

Thank You!

Andrew J. Van Leuven

andrew.vanleuven@okstate.edu