# Using the Grammar of Graphics in Applied Economic Research

Andrew J. Van Leuven

September 30, 2022



- Why visualize data?
- Data visualization basics
- How to visualize data using the  ${\color{black} {ggplot2}}$  package in R



2 Some Data Visualization Basics

3 Data Visualization Using the R Statistical Language



# What is Data?

FLETCHER			6	100	-		6			
CATCHER			100	1	/	100			1	>
Bats: Left Throws: Rig W:: 195 H:: 01* Brm: October 3, 1966 Effemburgt Illingis	ht		R			No. of the second se	State of the state		C XX V	Nor North
PROFE	SSIO	NAL	BATT	NG	RECO	ORD	_			
YEAR TEAM	BA	6	AB	8	H	28	38	HR	RBI	\$8
1987 VEHU BEACH	.266	43	124	13	33	7		0	15	0
1989 ALBUOUFROUF	273	100	315	34	86	16	1	5	44	1
1989 DODGERS	.500	5	8	1	4	0	- Ó	1	2	Ó.
1990 ALBUQUERQUE	.291	105	350	58	102	23	1	13	65	0
	.130	11	23	3	3	1	0	0	1	0
1990 DODGERS-PHILLIES	784	90	300	39	31	13		1	12	
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES	228	45				0		-	14	-
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M I TOTALS	.228	40	150	0	38	9		2	15	
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M.L. TOTALS	.228 .228	46 62	167	9 bin	38 d th	9 9 DI	0 ote	2 in '1	15 D1 wi	hen
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M.L. TOTALS Darrin got plenty of p Darren Daulton was i	.228 .228 laying	46 62 g tim	167 167 an a	9 ehin	38 d th acc	9 e pl	0 ate	z in 's n ea	15 91 wi riv N	nen lav.
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M.L. TOTALS Darren Daulton was i After Daulton recover	.228 .228 laying njure ed, D	46 62 g tim d in arrin	167 ne be an a	9 hin auto nt t	38 d th acc he re	9 e pl ide	ate nt in of th	2 in '9 n ea 10 St	91 wi rly N easor	nen lay. h at
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M.L. TOTALS Darren Daulton vas i After Daulton recovert Triple-A Scranton, wh	.228 .228 laying njure ed, D ere h	d in arrin	167 ne be an a spe atted	9 auto at ti a s	38 d th acc he re turd	9 e pl side est o y .2	ate nt in of th 84.	in 's n ea le se A n	91 wi rly N easor angy	hen lay. h at re-
1990 DODGERS-PHILLIES 1991 SCRAMTON 1991 PHILLIES M.L. TOTALS Darrien got plenty of p Darren Daulton was i After Daulton recovert Triple-A Scranton, wh ceiver with an accurat	.228 .228 laying njure ed, D ere h e arm	d in arrin le ba	167 ne be an a spe atted	9 auto at ti a s e wa	38 d th acc he re turd as th	9 e pl eider est o y .2 ie D	ate nt in of th 84. odg	in 's n ea le se A n jers'	91 will riy N eason angy catc	hen lay. n at re- her
1990 DODGERS-PHILLIES 1991 SCRANTON 1991 PHILLIES M.L. TOTALS Darren Daulton was i After Daulton recovers Triple-A Scranton, wh ceiver with an accurat of the future after bein	.228 .228 laying njure ed, D ere h e arm ig dra	45 62 g tim d in arrin ie ba n, he afted	167 ne be an a spe atted onc	9 hin auto nt ti a s e wa he si	38 d th acc he re turd as th ixth	9 e pl idei est o y .2 ie D roui	ate nt in of th 84. odg	in 's	91 wi riy N easo angy catc 7. Da	hen lay. n at re- her rrin
1990 DODGERS-PHILLES 1991 SCRANTON 1991 PHILLES M.L TOTALS Darrin got plenty of p Darren Daulton recover After Daulton recover Triple-A Scranton, wh ceiver with an accurat of the future after bein had quite an impressi	.228 .228 laying njure ed, Di ere h e arm ig dra ive de	45 62 g tim arrin ie ba arrin ie ba afted ebut	167 ne be an a spe atted onco in th in th	9 ahin auto nt ti a s e wa he si he n	38 d thi acc he re turd as thi ixth najo	9 e pl idei est o y .2 ie D roui rs a	ate nt in of th 84. odg nd i	in 's n ea le se A n jers' n '8 e en	15 P1 wi rly N eason angy catc 7. Da id of	nen lay. n at re- her rrin '89
1990 DODGERS-PHILLES 1991 SCRANTON 1991 PHILLES Darrin got plenty of p Darren Daulton was i After Daulton recover Triple-A Scranton, wh ceiver with an accurat of the future after bein had quite an impress when he smashed a h	.228 .228 laying njure ed, D ere h e arm ig dra ive de ome	45 62 g tim d in arrin ie ba n, he afted ebut run a	167 167 an a spe atted onco in th in th and t	9 ehin auto nt ti a s e wa he si he n wo s	38 d the acc he re turd as the ixth najo sing	9 e pl idei est o y .2 ie D roui rs a les i	ate nt in of th 84. odg nd i t th n hi	in 's n ea n ea A n jers' n '8 e en is fir	91 will riy N eason angy catc 7. Da id of st fiv	nen lay. n at re- her rrin '89 e at
1990 DODDERS-PHILLES 1991 SCRANTON 1991 PHILLES Darrin got plenty of p Darren Daulton was After Daulton recover Triple-A Scranton, wh ceiver with an accurat of the future after bein had quite an impress when he smashed a h bats. During the regul	.228 .228 laying njure ed, D ere h e arm ig dra ive de ome ar se	45 62 g tim d in arrin ie ba n, he afted ebut run a ason	167 167 an a spe atted onco in th in th and t h at A	9 auto at ti a s e wa he si he n wo s	38 d thi acc he re turd as th ixth najoi sing que	9 e pl idei est o y .2 ie D roui rs a les i rque	ate nt in of th 84. odg nd i t th n hi e he	in 's n ea le se A n jers' n '8 e en is fir hit	15 P1 wi Pasoi angy catc 7. Da id of st fiv .273.	hen lay. re- her rrin '89 e at
1990 DODARS-PHILLES 1991 SCRATON 1991 PHILLES Darrin got plenty of p Darren Daulton was i After Daulton recover Tiriple-A Scranton, wh Ceiver with an accurat of the future after bein and quite an impress when he smashed a h bats. During the regul	.228 .228 laying njure ed, Di ere h e arm ig dra ive de ome ar se	45 62 g tim d in arrin ie ba n, he afted ebut run a ason	130 167 ne be an a spe atted onco in th in th and t n at A	9 auto at ti a s e wa he si he n WO s	38 d thi acc he re turd as thi ixth najoi sing que	9 e pl idei est o y .2 ie D roui rs a les i rque	ate nt in of th 84. odg nd i t th n hi e he	in 's n ea e se A n jers' n '8 e en is fir hit	15 P1 will riy N eason angy catc 7. Da id of st fiv .273.	hen lay. n at re- her rrin '89 e at

- 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)



- 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)





I would like to nominate CNN for the worst data visualization of 2022.



9:48 AM - 5/16/22 - Twitter Web App



lol this bar graph makes it look like the number is 1/3 what it used to be but its not



5:02 PM · Sep 27, 2021 · SnapStream TV Search

Bits Let Thorse Step Disc. Control Contro Control Control		DARRIN FLETCHER CATCHER			~			AN A	9	Contraction of the second		AN NON
INF VEX <th>Bat</th> <th>s: Left Throws: Rig Wt.: 195 Ht.: 61" Jorn: October 3, 1966 Elimhurst, Illinois PROFI</th> <th>pht 6 ESSION BA</th> <th>ALE</th> <th>BATTI</th> <th>NG</th> <th>RECO</th> <th>ORD 28</th> <th></th> <th>HR</th> <th>RBI</th> <th>58</th>	Bat	s: Left Throws: Rig Wt.: 195 Ht.: 61" Jorn: October 3, 1966 Elimhurst, Illinois PROFI	pht 6 ESSION BA	ALE	BATTI	NG	RECO	ORD 28		HR	RBI	58
Image Automotion 228 65 79 95 86 8 7 70   70	1987	VERO BEACH	.266	43	124	13	33	7	0	0	15	0
Image Description 273 18 34 64 15 16 34 16 17 16 17 16 16 17 16 17 17 18 17 18 17 18 17 18 18	1988	SAN ANTONIO	.208	89	279	19	58	8	Ő	1	20	2
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	1989	ALBUQUERQUE	.273	100	315	34	85	16	1	5	44	1
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	1989	DODGERS	.500	5	8	_1	4	0	0	1	2	0
The sector in the sector is a sector	1990	ALBUQUERQUE	.291	105	350	58	102	Z3	1	13	65	0
Initial Parallelia 228 44 158 8 8 1 12 64   MLTORAS 228 64 108 5 15 6 1 12 6 107 7 58 9 2 15 6 Darin cput control of the set of the se	1001	SCRANTON	284	00	206	30	87	13	1		50	1
ALL INTACT 200 42 47 5 48 4 5 6 6 6 1 5 6 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 6 7 5 9 6 7 6 7 5 9 6 7 6 7 6 7 6 7 6 7 6 7 6 7 6 7 6 7 6	1001	PHILLIES	228	46	136	5	31	8		1	12	
Derrin qut plenty of playing time behind the plate in '91 when Derren Daution was injured in an auto accident in early May. After Daution recovered, Darrin spent the rest of the season at Tigle-A Scrandto, where he batterd a sturdy 284. A rangy re- of the future after being drafted in the sixth round in '87. Darrin that quite an impresive debut in the majors at then of '89 when he smashed a home run and two singles in his first tive at stas. During the regular season at Rougiarque he ht. 273.		OTALS	0.00	00	167	-	0.0	0	-	-	12	-
	Darri	in got plenty of p	aying	tim	e be	hin	d th	e pl	ate	in 's	91 wi rly N	hen lay.

- 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.



### 2 Some Data Visualization Basics

### 3 Data Visualization Using the R Statistical Language



#### Chart Suggestions—A Thought-Starter



Andrew J. Van Leuven

## 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**



#### Sequential YIOrRd YIOrBr YIGnBu YIGn Reds RdPu Purples PuRd PuBuGn PuBu OrRd Oranges Greys Greens GnBu BuPu BuGn Blues

#### Qualitative



2 Some Data Visualization Basics

### **3** Data Visualization Using the R Statistical Language



- 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



- 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()



- 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!



# Thank You!

Andrew J. Van Leuven

## andrew.vanleuven@okstate.edu

