Introduction to the Grammar of Graphics

Nate Wells

Math 141, 1/29/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Motivate intentional data visualization
- Discuss the Grammar of Graphics
- Decompose particular graphics using the GG paradigm

Section 1

Data Visualization

Why construct a graph?

- Even small data sets are big. We need to summarize the data.

Why construct a graph?

- Even small data sets are big. We need to summarize the data.
- Humans are best at finding patterns in visual media.

Why construct a graph?

- Even small data sets are big. We need to summarize the data.
- Humans are best at finding patterns in visual media.
- Graphs allow us to compare and explore relationships between variables.

Why construct a graph?

- Even small data sets are big. We need to summarize the data.
- Humans are best at finding patterns in visual media.
- Graphs allow us to compare and explore relationships between variables.
- Most importantly, graphs **tell a compelling story.**

Graphs Gone Awry: The Challenger Disaster

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.

Graphs Gone Awry: The Challenger Disaster

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.
  - Believed cold weather impacted the o-rings that held the rockets together.
  - Used 13 charts in their argument.

Graphs Gone Awry: The Challenger Disaster

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.
  - Believed cold weather impacted the o-rings that held the rockets together.
  - Used 13 charts in their argument.
- After a two hour conference call, the engineer's recommendation was overruled due to lack of persuasive evidence and the launch proceeded.

Graphs Gone Awry: The Challenger Disaster

- On January 27th, 1986, engineers from Morton Thiokol recommended NASA delay launch of space shuttle *Challenger* due to cold weather.
  - Believed cold weather impacted the o-rings that held the rockets together.
  - Used 13 charts in their argument.
- After a two hour conference call, the engineer's recommendation was overruled due to lack of persuasive evidence and the launch proceeded.
- The Challenger exploded 73 seconds into launch.

The Challenger Charts

- Here is one of those charts:

The Challenger Charts

- Here is one of those charts:

BLOW BY HISTORY
SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

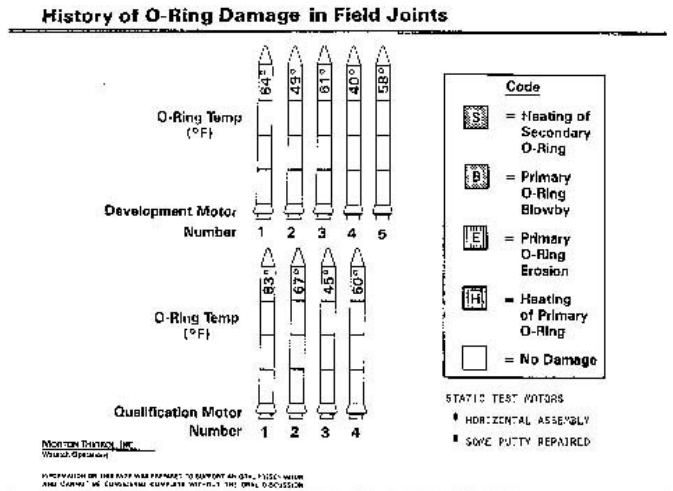
MOTOR	MBT	AMB	O-RING	WIND
DM-1	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

The Challenger Charts

- Here is another of those charts:

The Challenger Charts

- Here is another of those charts:

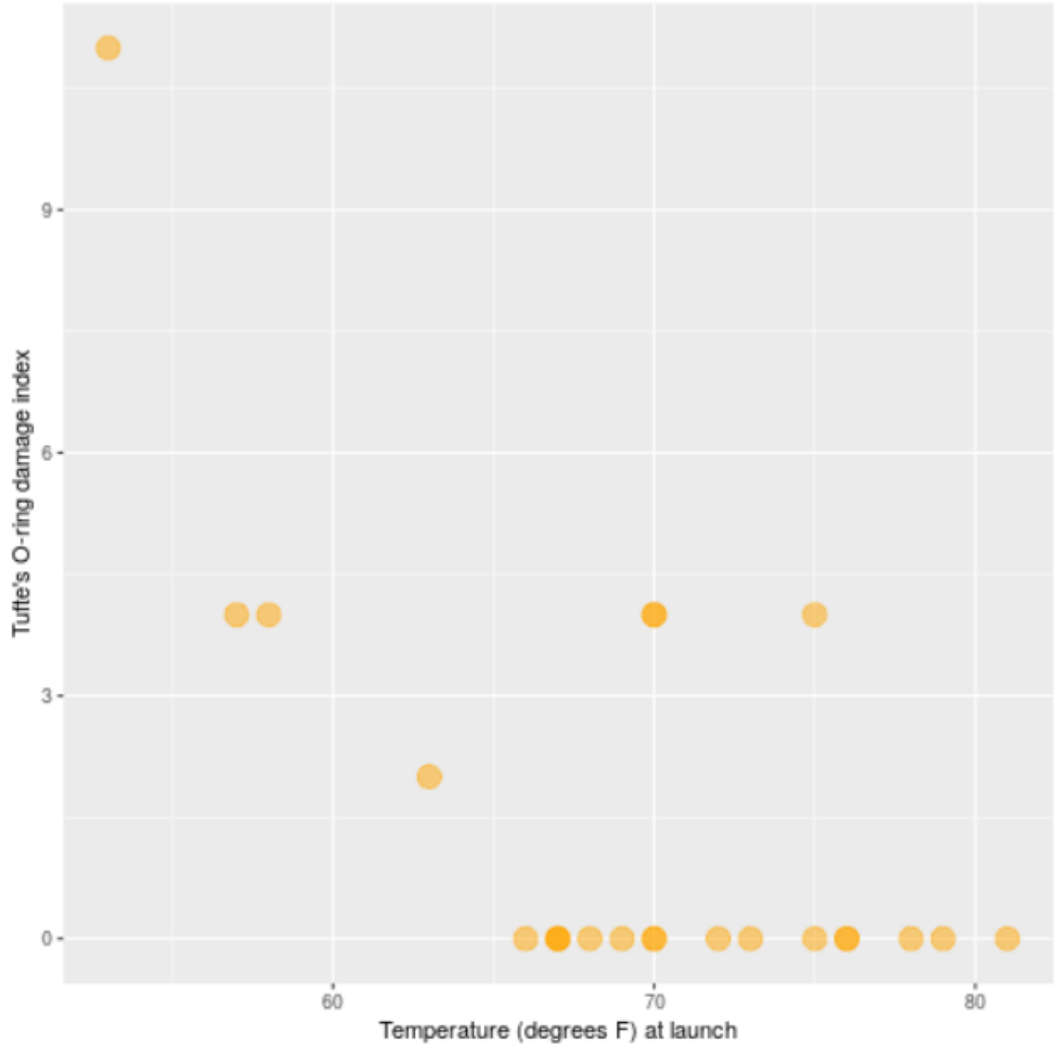


A Better Graph?

- The following is a graphic created in RStudio from Edward Tufte's data.

A Better Graph?

- The following is a graphic created in RStudio from Edward Tufte's data.

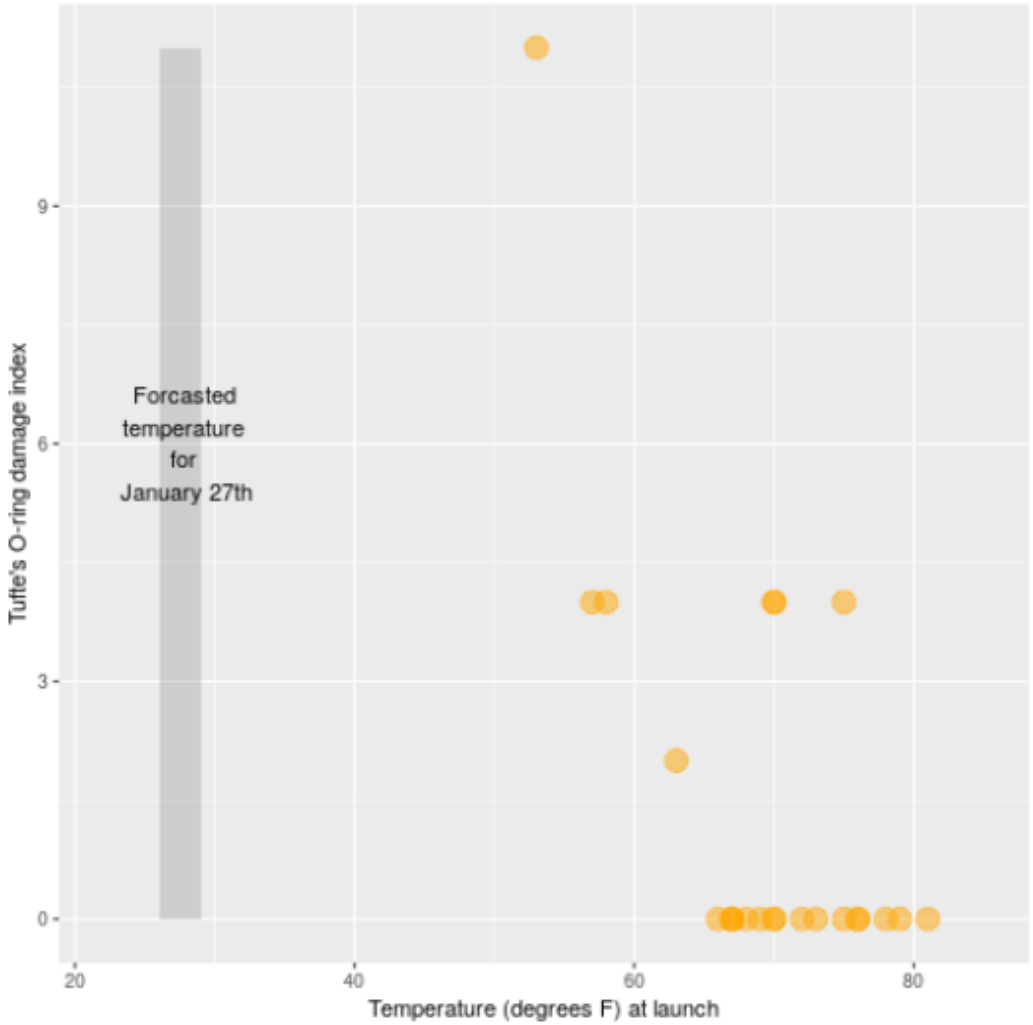


With Context

- And this graphic further emphasizes the direction of the trend.

With Context

- And this graphic further emphasizes the direction of the trend.



Section 2

The Grammar of Graphics

The Guiding Principle

A statistical graphic is a mapping of `data` variables to `aesthetic` attributes of `geometric` objects.

The Guiding Principle

A statistical graphic is a mapping of `data` variables to `aesthetic` attributes of `geometric` objects.

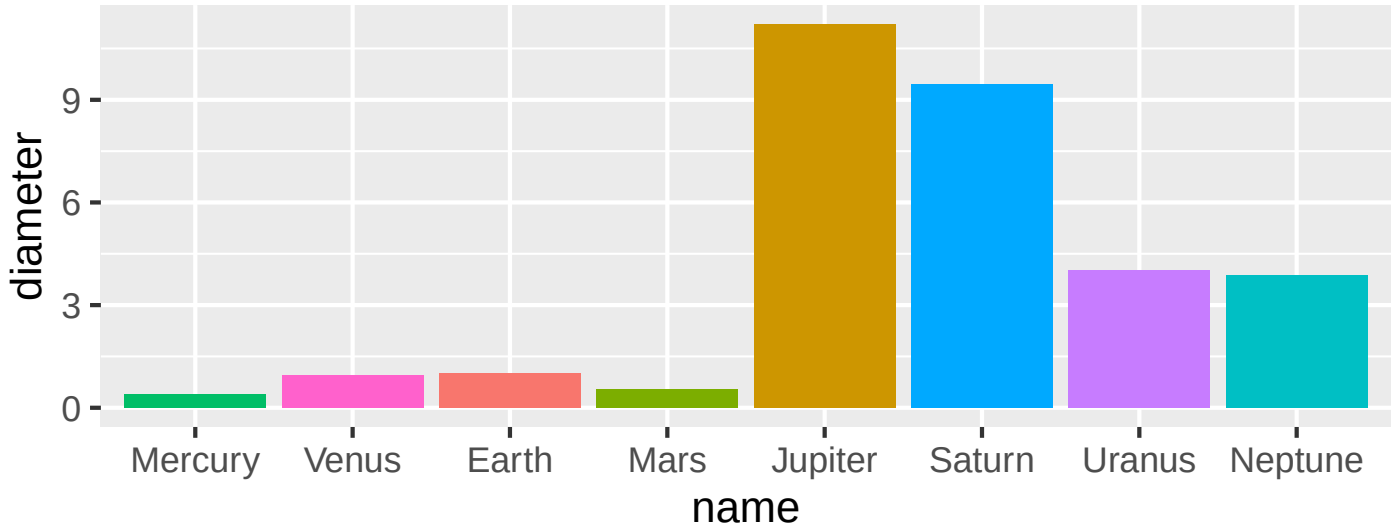
data	aesthetics	geometric object
Planet Name	x position	bar
Planet Diameter	y height	bar
Planet Name	color	bar

The Guiding Principle

A statistical graphic is a mapping of data variables to aesthetic attributes of geometric objects.

data	aesthetics	geometric object
Planet Name	x position	bar
Planet Diameter	y height	bar
Planet Name	color	bar

```
ggplot(data = planets_df) +
  geom_bar(stat = "identity",  mapping = aes(x = name, y = diameter, fill  = name)
```



GG Definitions

- **data**: data frame that contains the raw data and variables of interest

GG Definitions

- **data**: data frame that contains the raw data and variables of interest
- **geom**: geometric shape that the data are mapped to. Can be points, lines, bars, etc.

GG Definitions

- **data**: data frame that contains the raw data and variables of interest
- **geom**: geometric shape that the data are mapped to. Can be points, lines, bars, etc.
- **aesthetic**: visual properties of the **geom** object, like x position, y position, color, fill, shape

GG Definitions

- **data**: data frame that contains the raw data and variables of interest
- **geom**: geometric shape that the data are mapped to. Can be points, lines, bars, etc.
- **aesthetic**: visual properties of the **geom** object, like x position, y position, color, fill, shape
- **scale**: controls how data are mapped to the visual values of the aesthetic.
  i.e. specifying particular colors or shapes

GG Definitions

- **data**: data frame that contains the raw data and variables of interest
- **geom**: geometric shape that the data are mapped to. Can be points, lines, bars, etc.
- **aesthetic**: visual properties of the **geom** object, like x position, y position, color, fill, shape
- **scale**: controls how data are mapped to the visual values of the aesthetic.
  i.e. specifying particular colors or shapes
- **guide**: a legend to help user convert visual display back to the data

Plotting the Planets

Consider the planets data frame, `planets_df`:

name	type	diameter	rotation	rings	distance
Mercury	Terrestrial planet	0.382	58.64	FALSE	0.4
Venus	Terrestrial planet	0.949	-243.02	FALSE	0.7
Earth	Terrestrial planet	1.000	1.00	FALSE	1.0
Mars	Terrestrial planet	0.532	1.03	FALSE	1.5
Jupiter	Gas giant	11.209	0.41	TRUE	5.2
Saturn	Gas giant	9.449	0.43	TRUE	9.5
Uranus	Gas giant	4.007	-0.72	TRUE	19.2
Neptune	Gas giant	3.883	0.67	TRUE	30.1

Plotting the Planets

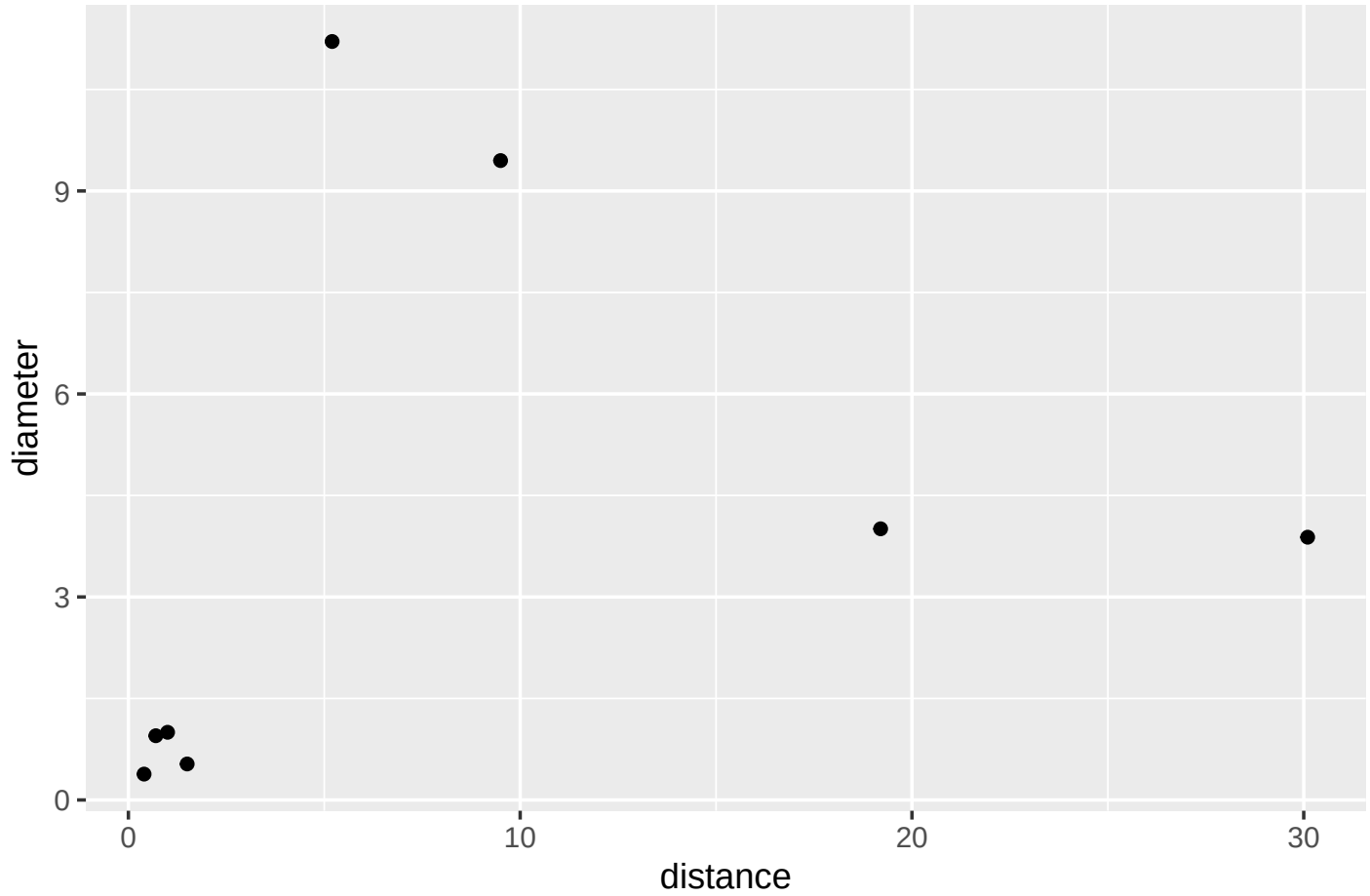
Consider the planets data frame, `planets_df`:

name	type	diameter	rotation	rings	distance
Mercury	Terrestrial planet	0.382	58.64	FALSE	0.4
Venus	Terrestrial planet	0.949	-243.02	FALSE	0.7
Earth	Terrestrial planet	1.000	1.00	FALSE	1.0
Mars	Terrestrial planet	0.532	1.03	FALSE	1.5
Jupiter	Gas giant	11.209	0.41	TRUE	5.2
Saturn	Gas giant	9.449	0.43	TRUE	9.5
Uranus	Gas giant	4.007	-0.72	TRUE	19.2
Neptune	Gas giant	3.883	0.67	TRUE	30.1

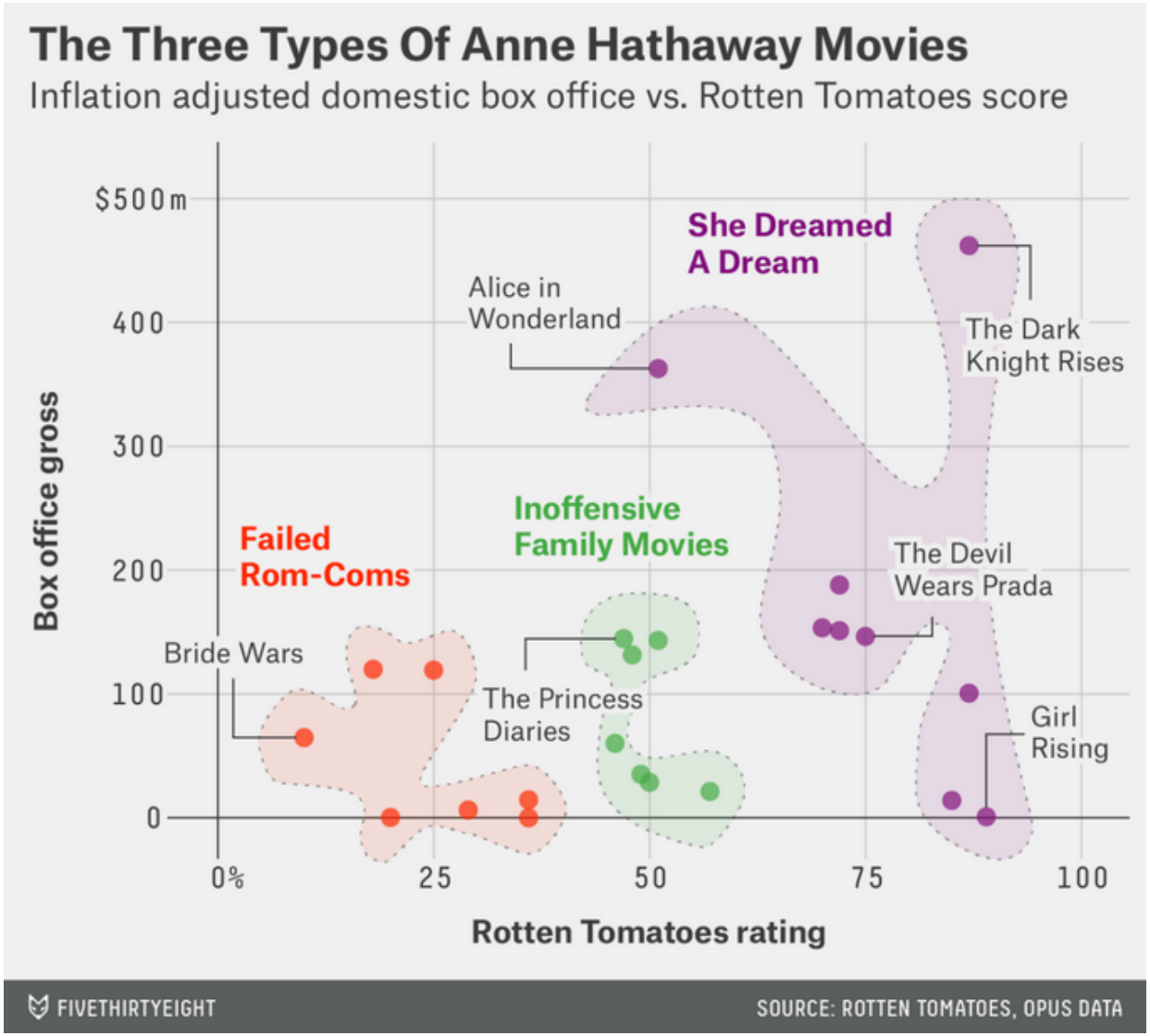
Describe how to create a plot of distance vs. diameter.

Plotting the Planets

```
ggplot(data = planets_df, mapping = aes(x = distance, y = diameter)) +
  geom_point()
```

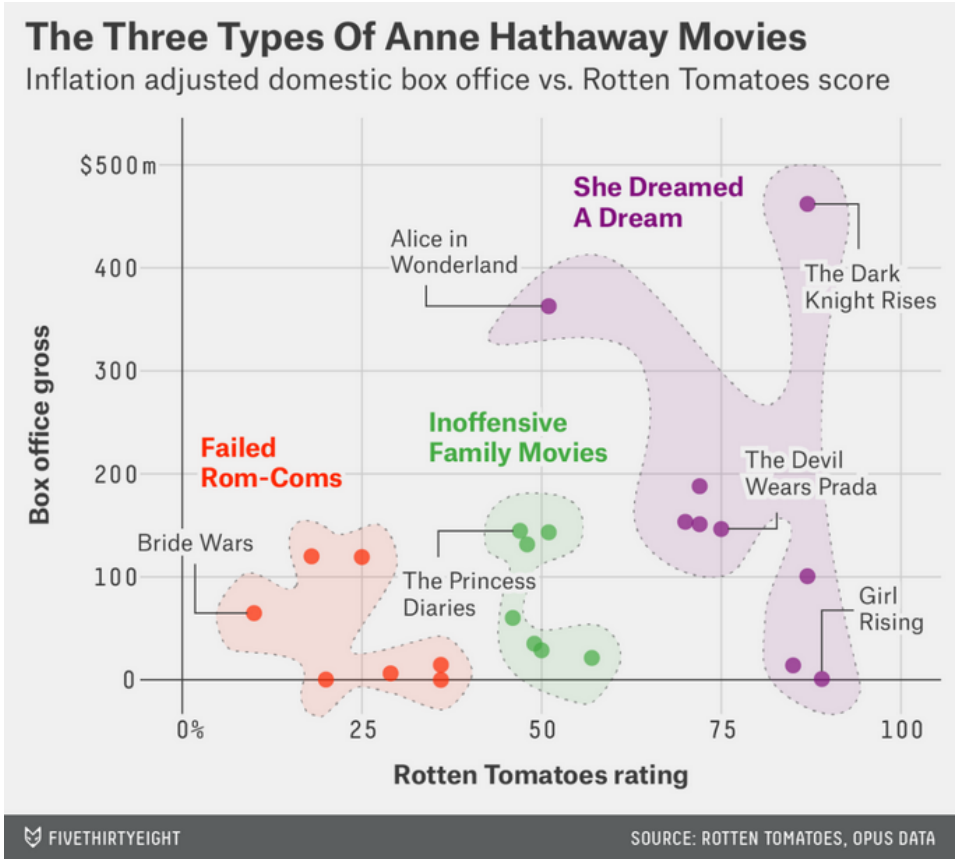


Example 1 Graphic



Example 1

1. What is the story this graphic is telling?
2. What are the variables here?
3. What **geom** are the variables mapped to?
4. What are the **aesthetics** of the **geom**? Which variable sets the value of that **aesthetic**?
5. What additional context does this graphic provide?



Example 2 Graphic

Sexual harassment charges, by industry

Among charges filed by women, fiscal years 2005-2015

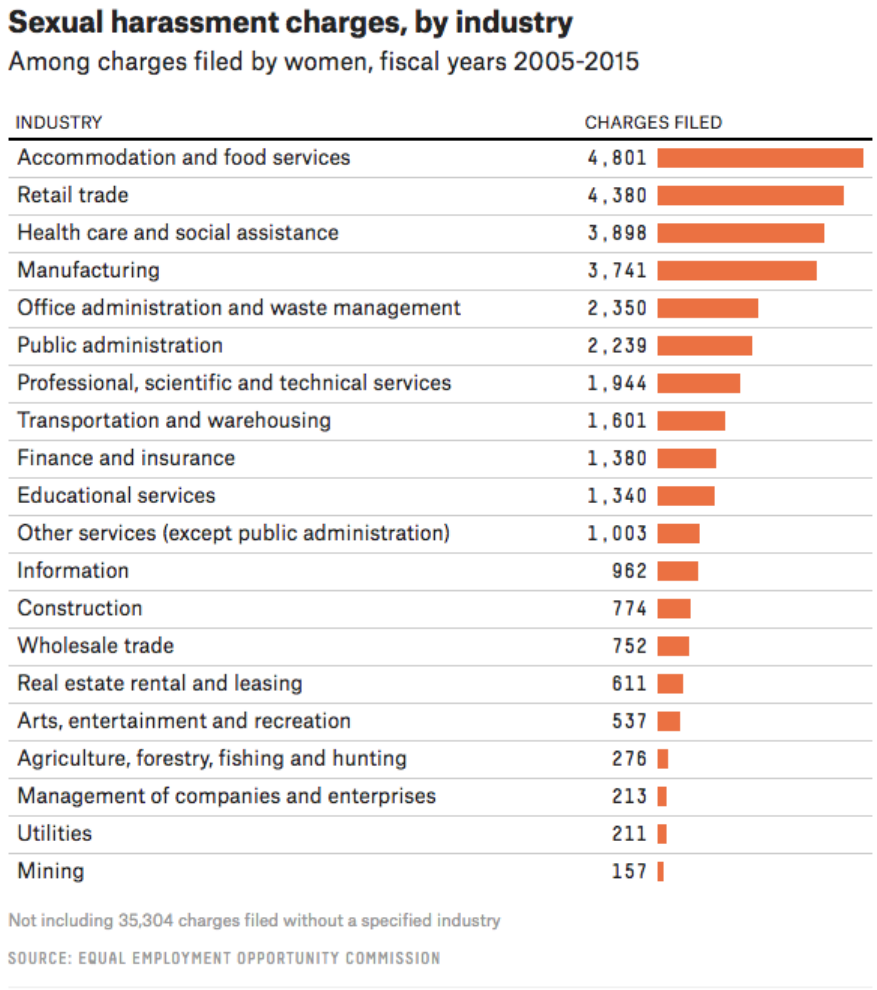
INDUSTRY	CHARGES FILED
Accommodation and food services	4,801
Retail trade	4,380
Health care and social assistance	3,898
Manufacturing	3,741
Office administration and waste management	2,350
Public administration	2,239
Professional, scientific and technical services	1,944
Transportation and warehousing	1,601
Finance and insurance	1,380
Educational services	1,340
Other services (except public administration)	1,003
Information	962
Construction	774
Wholesale trade	752
Real estate rental and leasing	611
Arts, entertainment and recreation	537
Agriculture, forestry, fishing and hunting	276
Management of companies and enterprises	213
Utilities	211
Mining	157

Not including 35,304 charges filed without a specified industry

SOURCE: EQUAL EMPLOYMENT OPPORTUNITY COMMISSION

Example 2

1. What is the story this graphic is telling?
2. What are the variables here?
3. What **geom** are the variables mapped to?
4. What are the **aesthetics** of the **geom**? Which variable sets the value of that **aesthetic**?
5. What additional context does this graphic provide?



Sexual harassment charges, by industry

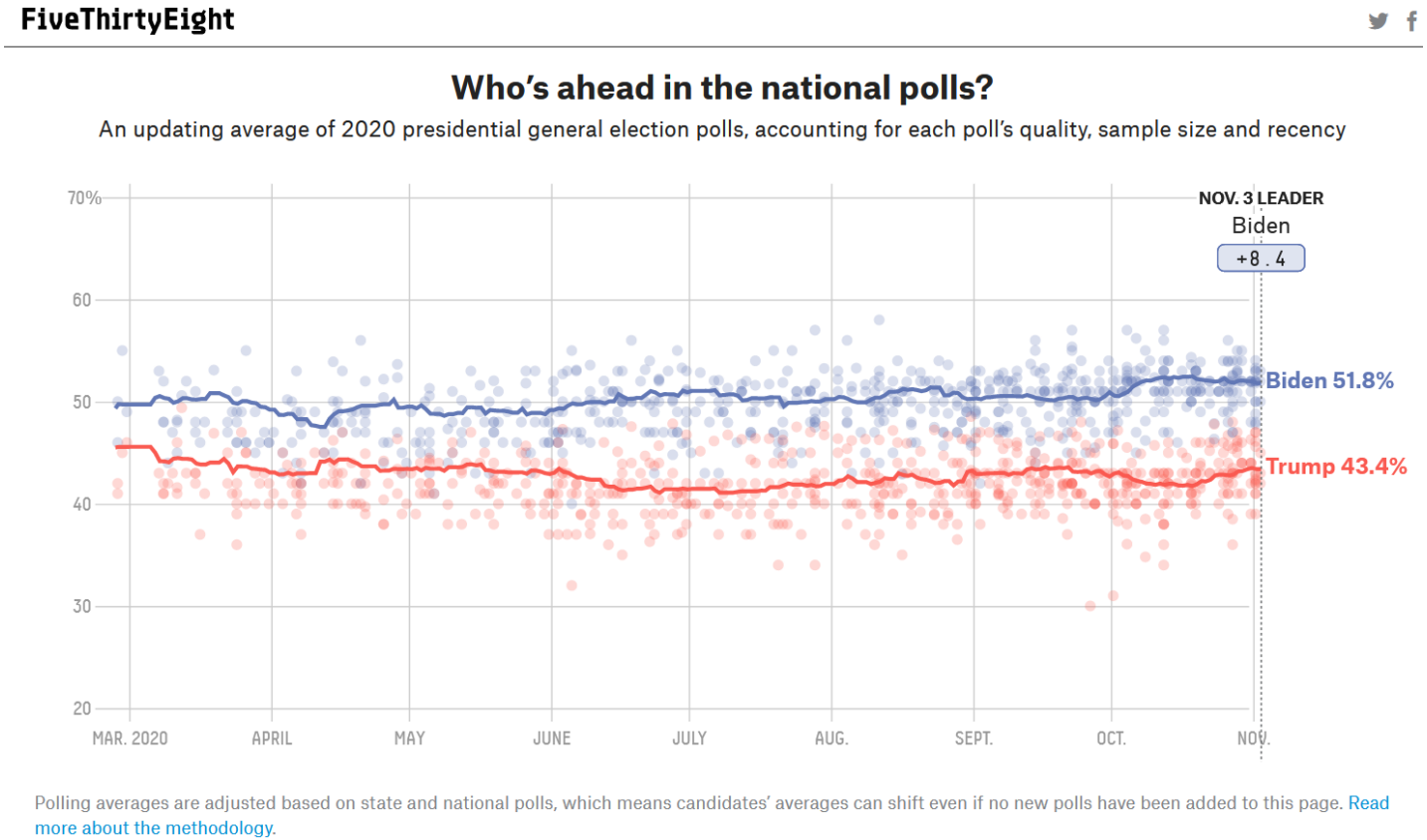
Among charges filed by women, fiscal years 2005-2015

INDUSTRY	CHARGES FILED
Accommodation and food services	4,801
Retail trade	4,380
Health care and social assistance	3,898
Manufacturing	3,741
Office administration and waste management	2,350
Public administration	2,239
Professional, scientific and technical services	1,944
Transportation and warehousing	1,601
Finance and insurance	1,380
Educational services	1,340
Other services (except public administration)	1,003
Information	962
Construction	774
Wholesale trade	752
Real estate rental and leasing	611
Arts, entertainment and recreation	537
Agriculture, forestry, fishing and hunting	276
Management of companies and enterprises	213
Utilities	211
Mining	157

Not including 35,304 charges filed without a specified industry

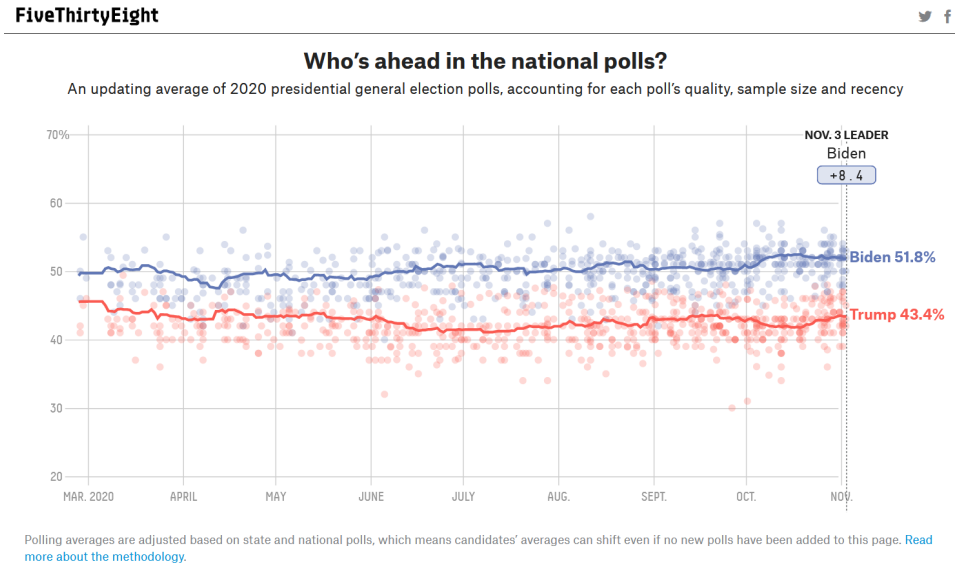
SOURCE: EQUAL EMPLOYMENT OPPORTUNITY COMMISSION

Example 3 Graphic

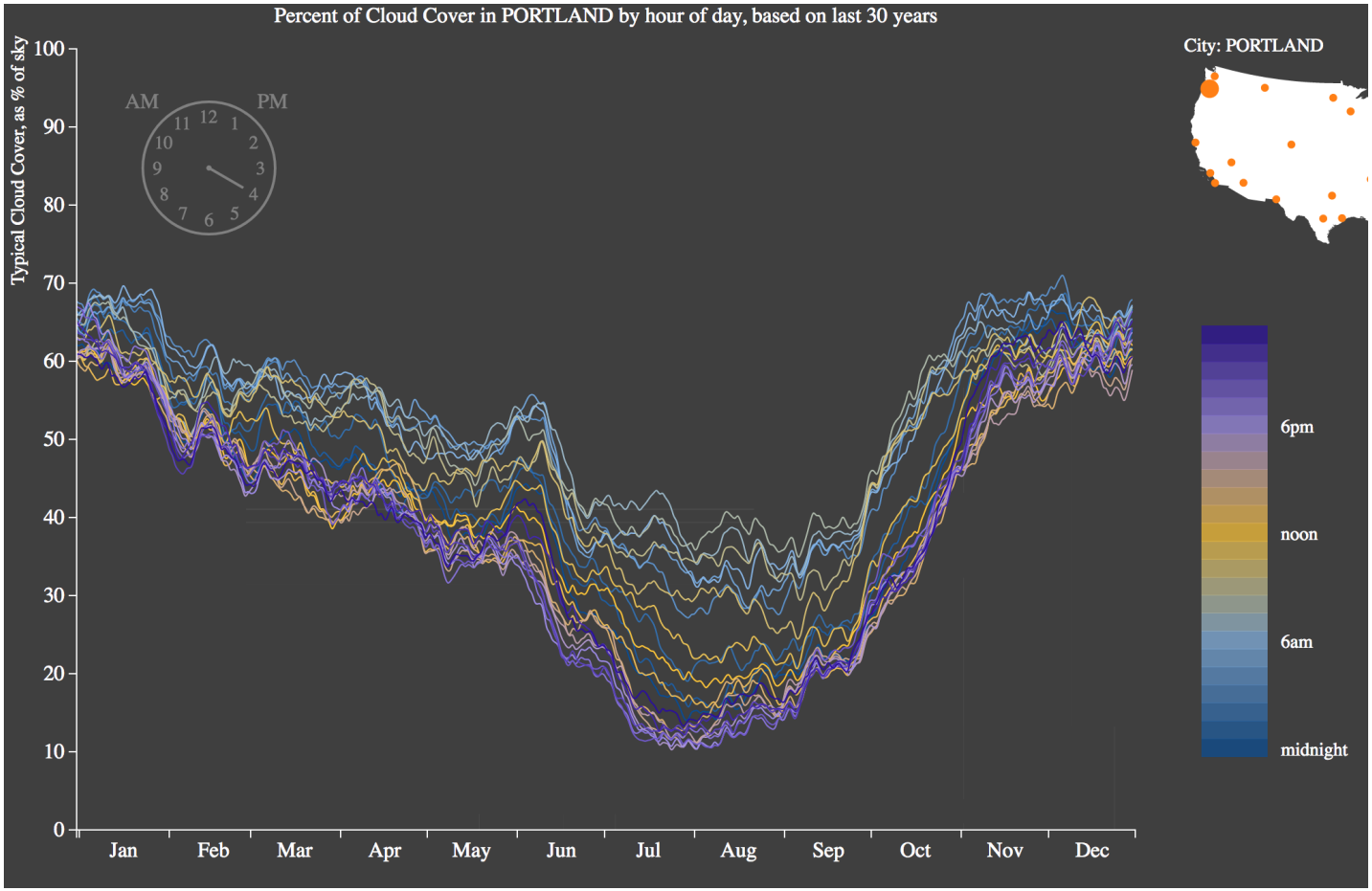


Example 3

1. What is the story this graphic is telling?
2. What are the variables here?
3. What **geom** are the variables mapped to?
4. What are the **aesthetics** of the **geom**? Which variable sets the value of that **aesthetic**?
5. What additional context does this graphic provide?

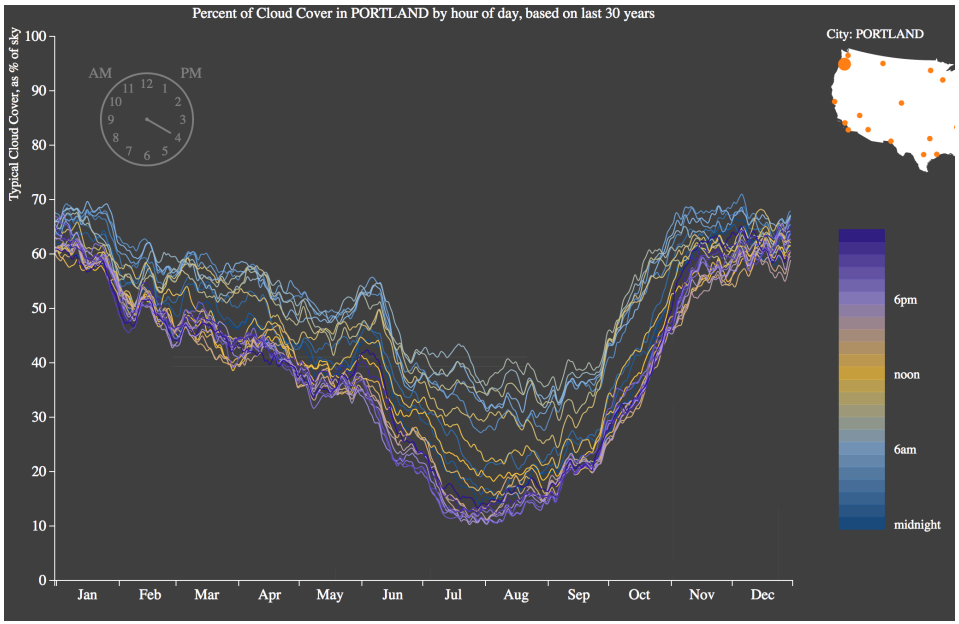


Example 4 Graphic



Example 4

1. What is the story this graphic is telling?
2. What are the variables here?
3. What **geom** are the variables mapped to?
4. What are the **aesthetics** of the **geom**? Which variable sets the value of that **aesthetic**?
5. What additional context does this graphic provide?



Best Practices for Graphics

- Determine what story your graphic should tell. And ensure that all aspects of your graphic serve this purpose.

Best Practices for Graphics

- Determine what story your graphic should tell. And ensure that all aspects of your graphic serve this purpose.
- Label axes and units for improved clarity.

Best Practices for Graphics

- Determine what story your graphic should tell. And ensure that all aspects of your graphic serve this purpose.
- Label axes and units for improved clarity.
- Include a legend to translate from aesthetics to variables.

Best Practices for Graphics

- Determine what story your graphic should tell. And ensure that all aspects of your graphic serve this purpose.
- Label axes and units for improved clarity.
- Include a legend to translate from aesthetics to variables.
- Specify your data source for reproducibility/verification.

Best Practices for Graphics

- Determine what story your graphic should tell. And ensure that all aspects of your graphic serve this purpose.
- Label axes and units for improved clarity.
- Include a legend to translate from aesthetics to variables.
- Specify your data source for reproducibility/verification.
- Minimize/eliminate extraneous elements that do not serve main purpose.