Exploring Data

Prof. Wells

STA 209, 1/25/22

Outline

In this class, we will. . .

Outline

In this class, we will...

- Investigate the structure of data
- Explore data frames in R
- Lab: Introduce RStudio

Section 1

Structure of Data

What is data?

Data discussion questions. Spend 2 - 3 minutes thinking individually about these questions. Then discuss with a partner in class.

What is data?

Data discussion questions. Spend 2 - 3 minutes thinking individually about these questions. Then discuss with a partner in class.

- When we use the word *data* in everyday converstation, what do we mean?
- Is there a difference between *data* and *information*?
- What is an example of data?
- What do *you* use data for?
- What do others use data about you for?
- Does data have to be numeric? If not, what are some examples?

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.
  - Each step in the data collection and distribution chain requires decisions.

What is data?

- In order to perform *any* meaningful statistical analysis or inference, we need transmittable and tractable real world observations.
- **Data** are characteristics or measurements describing some individuals in a population.
- It may be tempting to suggest that data is objective.
  - But all data must be gathered, collated, organized, classified, shaped.
  - Each step in the data collection and distribution chain requires decisions.
  - Data tells a story.

Data Frames

- For convenience, data is often stored in a spreadsheet format (like a google sheet)

Data Frames

- For convenience, data is often stored in a spreadsheet format (like a google sheet)
- In R, these spreadsheet-type data sets are called **data frames** (also tibbles)

Data Frames

- For convenience, data is often stored in a spreadsheet format (like a google sheet)
- In R, these spreadsheet-type data sets are called **data frames** (also tibbles)
- Example data frame:

Table 1: The Planets

name	type	diameter	distance	rings
Mercury	Terrestrial planet	0.382	0.390	FALSE
Venus	Terrestrial planet	0.949	0.723	FALSE
Earth	Terrestrial planet	1.000	1.000	FALSE
Mars	Terrestrial planet	0.532	1.524	FALSE
Jupiter	Gas giant	11.209	5.203	TRUE
Saturn	Gas giant	9.449	9.539	TRUE
Uranus	Gas giant	4.007	19.180	TRUE
Neptune	Gas giant	3.883	30.060	TRUE

Data Organization

- An **observation** or **case** is a single entity about which we've recorded characteristics

Data Organization

- An **observation** or **case** is a single entity about which we've recorded characteristics
- A **variable** is a characteristic of the entity.

Data Organization

- An **observation** or **case** is a single entity about which we've recorded characteristics
- A **variable** is a characteristic of the entity.
- A data frame is **tidy** when each row corresponds to an observation and each column corresponds to a variable

Table 2: The Planets

name	type	diameter	distance	rings
Mercury	Terrestrial planet	0.382	0.390	FALSE
Venus	Terrestrial planet	0.949	0.723	FALSE
Earth	Terrestrial planet	1.000	1.000	FALSE
Mars	Terrestrial planet	0.532	1.524	FALSE
Jupiter	Gas giant	11.209	5.203	TRUE
Saturn	Gas giant	9.449	9.539	TRUE
Uranus	Gas giant	4.007	19.180	TRUE
Neptune	Gas giant	3.883	30.060	TRUE

Data Organization

- An **observation** or **case** is a single entity about which we've recorded characteristics
- A **variable** is a characteristic of the entity.
- A data frame is **tidy** when each row corresponds to an observation and each column corresponds to a variable

Table 2: The Planets

name	type	diameter	distance	rings
Mercury	Terrestrial planet	0.382	0.390	FALSE
Venus	Terrestrial planet	0.949	0.723	FALSE
Earth	Terrestrial planet	1.000	1.000	FALSE
Mars	Terrestrial planet	0.532	1.524	FALSE
Jupiter	Gas giant	11.209	5.203	TRUE
Saturn	Gas giant	9.449	9.539	TRUE
Uranus	Gas giant	4.007	19.180	TRUE
Neptune	Gas giant	3.883	30.060	TRUE

- What are the observations and variables in the preceding data set?

Tidy Data

Consider the following two data frames containing swim times for the prelims and finals races of the 400 yard indidividual medley in the 2008 Beijing Olympics.

swimmer	prelims	finals
Phelps	4:07.82	4:03.84
Cseh	4:09.26	4:06.16
Lochete	4:10.33	4:08.09

swimmer	race	time
Phelps	prelims	4:07.82
Phelps	finals	4:03.84
Cseh	prelims	4:09.26
Cseh	finals	4:06.16
Lochete	prelims	4:10.33
Lochete	finals	4:08.09

- In what ways are these two data frames similar? Different?

Tidy Data

Consider the following two data frames containing swim times for the prelims and finals races of the 400 yard indidividual medley in the 2008 Beijing Olympics.

swimmer	prelims	finals
Phelps	4:07.82	4:03.84
Cseh	4:09.26	4:06.16
Lochete	4:10.33	4:08.09

swimmer	race	time
Phelps	prelims	4:07.82
Phelps	finals	4:03.84
Cseh	prelims	4:09.26
Cseh	finals	4:06.16
Lochete	prelims	4:10.33
Lochete	finals	4:08.09

- In what ways are these two data frames similar? Different?
- What are advantages of each form?

Observational Unit Example

Consider the following data frame. What are the observational units?

show_id	type	title	date_added	rating	duration	listed_in
s4230	Movie	Mortified Nation	February 1, 2018	TV-MA	84 min	Documentaries
s6918	Movie	The Surrounding Game	August 30, 2018	TV-14	98 min	Documentaries
s550	TV Show	Anthony Bourdain: Parts Unknown	NA	TV-PG	5 Seasons	Docuseries
s6063	Movie	The Adderall Diaries	July 15, 2018	R	87 min	Dramas, Thrillers
s1283	TV Show	Charlie's Colorforms City	March 22, 2019	TV-Y	1 Season	Kids' TV
s4803	TV Show	Pawn Stars	September 15, 2019	TV-14	1 Season	Reality TV
s1031	Movie	Bolt	July 22, 2018	PG	99 min	Children & Family Movies, Comedies
s1496	TV Show	Cooked with Cannabis	April 20, 2020	TV-MA	1 Season	Reality TV
s7220	TV Show	Trial By Media	May 11, 2020	TV-MA	1 Season	Crime TV Shows, Docuseries
s3968	TV Show	Marvel's Jessica Jones	June 14, 2019	TV-MA	3 Seasons	Crime TV Shows, TV Action & Adventure, TV Dramas
s1632	Movie	Dave Chappelle: Sticks & Stones	August 26, 2019	TV-MA	66 min	Stand-Up Comedy
s1773	TV Show	Dirty John	November 25, 2019	TV-MA	1 Season	Crime TV Shows, TV Dramas, TV Mysteries

Observational Unit Example

Consider the following data frame. What are the observational units?

show_id	type	title	date_added	rating	duration	listed_in
s4230	Movie	Mortified Nation	February 1, 2018	TV-MA	84 min	Documentaries
s6918	Movie	The Surrounding Game	August 30, 2018	TV-14	98 min	Documentaries
s550	TV Show	Anthony Bourdain: Parts Unknown	NA	TV-PG	5 Seasons	Docuseries
s6063	Movie	The Adderall Diaries	July 15, 2018	R	87 min	Dramas, Thrillers
s1283	TV Show	Charlie's Colorforms City	March 22, 2019	TV-Y	1 Season	Kids' TV
s4803	TV Show	Pawn Stars	September 15, 2019	TV-14	1 Season	Reality TV
s1031	Movie	Bolt	July 22, 2018	PG	99 min	Children & Family Movies, Comedies
s1496	TV Show	Cooked with Cannabis	April 20, 2020	TV-MA	1 Season	Reality TV
s7220	TV Show	Trial By Media	May 11, 2020	TV-MA	1 Season	Crime TV Shows, Docuseries
s3968	TV Show	Marvel's Jessica Jones	June 14, 2019	TV-MA	3 Seasons	Crime TV Shows, TV Action & Adventure, TV Dramas
s1632	Movie	Dave Chappelle: Sticks & Stones	August 26, 2019	TV-MA	66 min	Stand-Up Comedy
s1773	TV Show	Dirty John	November 25, 2019	TV-MA	1 Season	Crime TV Shows, TV Dramas, TV Mysteries

- This dataset consists of tv shows and movies available on Netflix as of 2021 The dataset is collected from Flixable which is a third-party Netflix search engine.

Types of Variables

Variables come in two general types: quantitative and categorical

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**
- Variables taking non-numeric values are called **categorical**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**
  - Some categorical variables can be ordered (but not averaged). These are called **ordinal** variables

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**
  - Some categorical variables can be ordered (but not averaged). These are called **ordinal** variables
  - Categorical variables that cannot be ordered are called **nominal**

Types of Variables

Variables come in two general types: quantitative and categorical

- Variables taking numeric values are called **quantitative**
  - Quantitative variables can be ordered and averaged.
  - Not every variable involving numbers in quantitative!
  - Quantitative variables that can take any range of values are called **continuous**
  - While those that can only take particular (often whole number) values are called **discrete**
- Variables taking non-numeric values are called **categorical**
  - The values that a categorical variable can take are called its **levels**
  - Some categorical variables can be ordered (but not averaged). These are called **ordinal** variables
  - Categorical variables that cannot be ordered are called **nominal**
- In some data frames, certain variables (called **identification variables**) uniquely specify each observation

Classifying Variables

Label the variables in the following data frame as: quantitative, categorical, or id.

- What is one continuous variable that could be added?

Classifying Variables

Label the variables in the following data frame as: quantitative, categorical, or id.

- What is one continuous variable that could be added?

show_id	type	title	date_added	rating	duration	listed_in
s4230	Movie	Mortified Nation	February 1, 2018	TV-MA	84 min	Documentaries
s6918	Movie	The Surrounding Game	August 30, 2018	TV-14	98 min	Documentaries
s550	TV Show	Anthony Bourdain: Parts Unknown	NA	TV-PG	5 Seasons	Docuseries
s6063	Movie	The Adderall Diaries	July 15, 2018	R	87 min	Dramas, Thrillers
s1283	TV Show	Charlie's Colorforms City	March 22, 2019	TV-Y	1 Season	Kids' TV
s4803	TV Show	Pawn Stars	September 15, 2019	TV-14	1 Season	Reality TV
s1031	Movie	Bolt	July 22, 2018	PG	99 min	Children & Family Movies, Comedies
s1496	TV Show	Cooked with Cannabis	April 20, 2020	TV-MA	1 Season	Reality TV
s7220	TV Show	Trial By Media	May 11, 2020	TV-MA	1 Season	Crime TV Shows, Docuseries
s3968	TV Show	Marvel's Jessica Jones	June 14, 2019	TV-MA	3 Seasons	Crime TV Shows, TV Action & Adventure, TV Dramas
s1632	Movie	Dave Chappelle: Sticks & Stones	August 26, 2019	TV-MA	66 min	Stand-Up Comedy
s1773	TV Show	Dirty John	November 25, 2019	TV-MA	1 Season	Crime TV Shows, TV Dramas, TV Mysteries

Relationships between Variables

Of chief interest to statisticians and data scientists are the relationships between variables in a data set.

Relationships between Variables

Of chief interest to statisticians and data scientists are the relationships between variables in a data set.

- Consider the Netflix data set:
  - How has the *proportion of films added* changed over the past 5 *years*?
  - Do the typical number of letters in a show's title vary by TV rating?
  - What is the predicted number of movies and shows that will be added in 2022?

Relationships between Variables

Of chief interest to statisticians and data scientists are the relationships between variables in a data set.

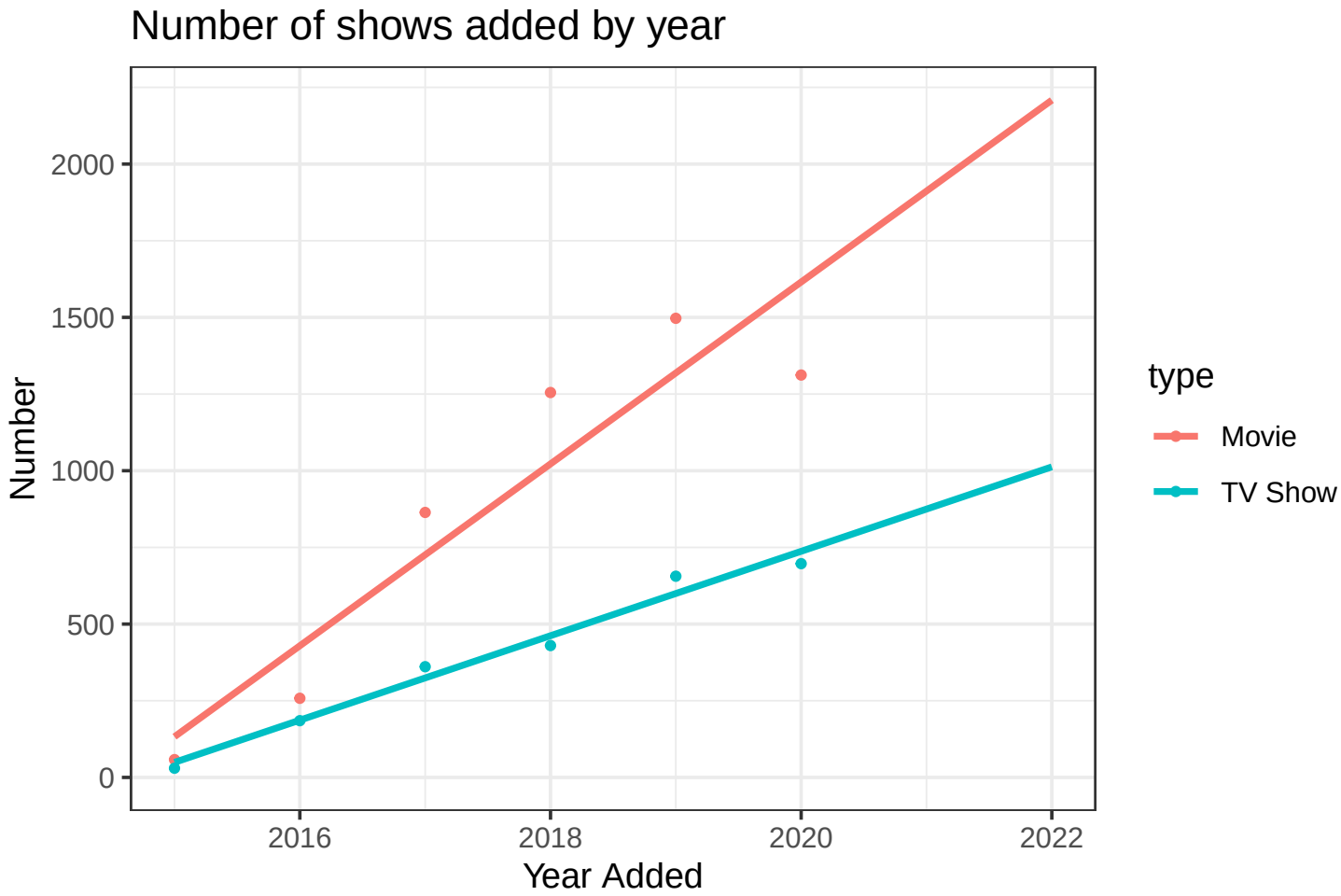
- Consider the Netflix data set:
  - How has the *proportion of films added* changed over the past 5 *years*?
  - Do the typical number of letters in a show's title vary by TV rating?
  - What is the predicted number of movies and shows that will be added in 2022?
- To answer these questions, we'll need...
  - **data visualizations** that provide a holistic overview of the data
  - **summary statistics** that capture the essential attributes of data in a few numeric values
  - **statistical models** that allow us to predict the value of one variable given another

Data Visualization

Summary Statistics

```
## # A tibble: 15 x 4
##    rating   mean_number_letters first_quartile third_quartile
##    <chr>                  <dbl>          <dbl>          <dbl>
##  1 TV-Y7-FV                28             15.2           33.5
##  2 <NA>                    27.3           17.5           34.5
##  3 TV-Y7                   24.5           14.5           32.5
##  4 TV-Y                    23.5           13             32
##  5 TV-G                    21.6           14             27.8
##  6 G                       20.8           15             26
##  7 TV-PG                   19.1           11             24
##  8 PG                      18.6           11             25.5
##  9 NC-17                   17.7           14             22
## 10 UR                      17.6           13             23
## 11 TV-MA                   17.3           10             22
## 12 TV-14                   16.4           10             21
## 13 PG-13                   16.2           10             20
## 14 NR                      16.1            9             21
## 15 R                       14.7           10             18
```

Statistical Models



$$\text{Number of TV Shows Added} = 97 \cdot \text{ Years since 2015 } + 380$$