IS 6489: Statistics and Predictive Analytics

Class 2

Jeff Webb

Tonight's agenda

- Details

Tonight's agenda

- Details
- Questions about the course or the material?

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review
- Review: tidy data

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review
- Review: tidy data
- Review: Why EDA?

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review
- Review: tidy data
- Review: Why EDA?
- EDA workflow

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review
- Review: tidy data
- Review: Why EDA?
- EDA workflow
- .Rmd script on tidy data and EDA workflow; mini-project.

Tonight's agenda

- Details
- Questions about the course or the material?
- Live poll review
- Review: tidy data
- Review: Why EDA?
- EDA workflow
- .Rmd script on tidy data and EDA workflow; mini-project.
- Note: My assumption tonight is that you've studied the lecture and tutorial videos for week 2 and are ready to practice the concepts and techniques covered there.

Details

Details

- My office hours: 9:30 AM - 10:30 AM Tuesday or by appointment.

Details

- My office hours: 9:30 AM - 10:30 AM Tuesday or by appointment.
- TA Ali Samanazari: ali.samanazari at utah.edu. Ali will conduct weekly tutorial sessions on Mondays, 5 - 6 PM in SFEBB 5163. Please also feel free to email him with any questions about R programming or other course content.

Details

- My office hours: 9:30 AM - 10:30 AM Tuesday or by appointment.
- TA Ali Samanazari: ali.samanazari at utah.edu. Ali will conduct weekly tutorial sessions on Mondays, 5 - 6 PM in SFEBB 5163. Please also feel free to email him with any questions about R programming or other course content.
- Homework coming up is *Introduction to the Tidyverse* at Datacamp. If you want more to do I would suggest additional courses in dplyr, ggplot2 or tidyverse.

Questions

Questions

- Any questions on the course or the material so far that I can address?

Live poll review

live poll review

Go to PollEv.com/jeffwebb768

Tidy data

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.
- Nevertheless, here are some guidelines for tidy data:

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.
- Nevertheless, here are some guidelines for tidy data:
  1. Each variable must have its own column.

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.
- Nevertheless, here are some guidelines for tidy data:
  1. Each variable must have its own column.
  2. Each observation must have its own row (meaning that each value must have its own cell).

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.
- Nevertheless, here are some guidelines for tidy data:
  1. Each variable must have its own column.
  2. Each observation must have its own row (meaning that each value must have its own cell).
  3. A table should be dedicated to same observational unit.

Messy data

- As noted: all messy datasets are messy in their own way, which makes it hard to generalize about how to fix them.
- Nevertheless, here are some guidelines for tidy data:
  1. Each variable must have its own column.
  2. Each observation must have its own row (meaning that each value must have its own cell).
  3. A table should be dedicated to same observational unit.
- Adapted from *R for Data Science* by Wickham and Grolemund.

Messy data

Tuberculosis cases and population by country and year. What's messy here?

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases            37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases            80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases           212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases           213766
## 12 China        2000 population 1280428583
```

Messy data

What's messy here?

```
## # A tibble: 6 x 3
##   country      year rate
## * <chr>       <int> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

Messy data

What's messy here?

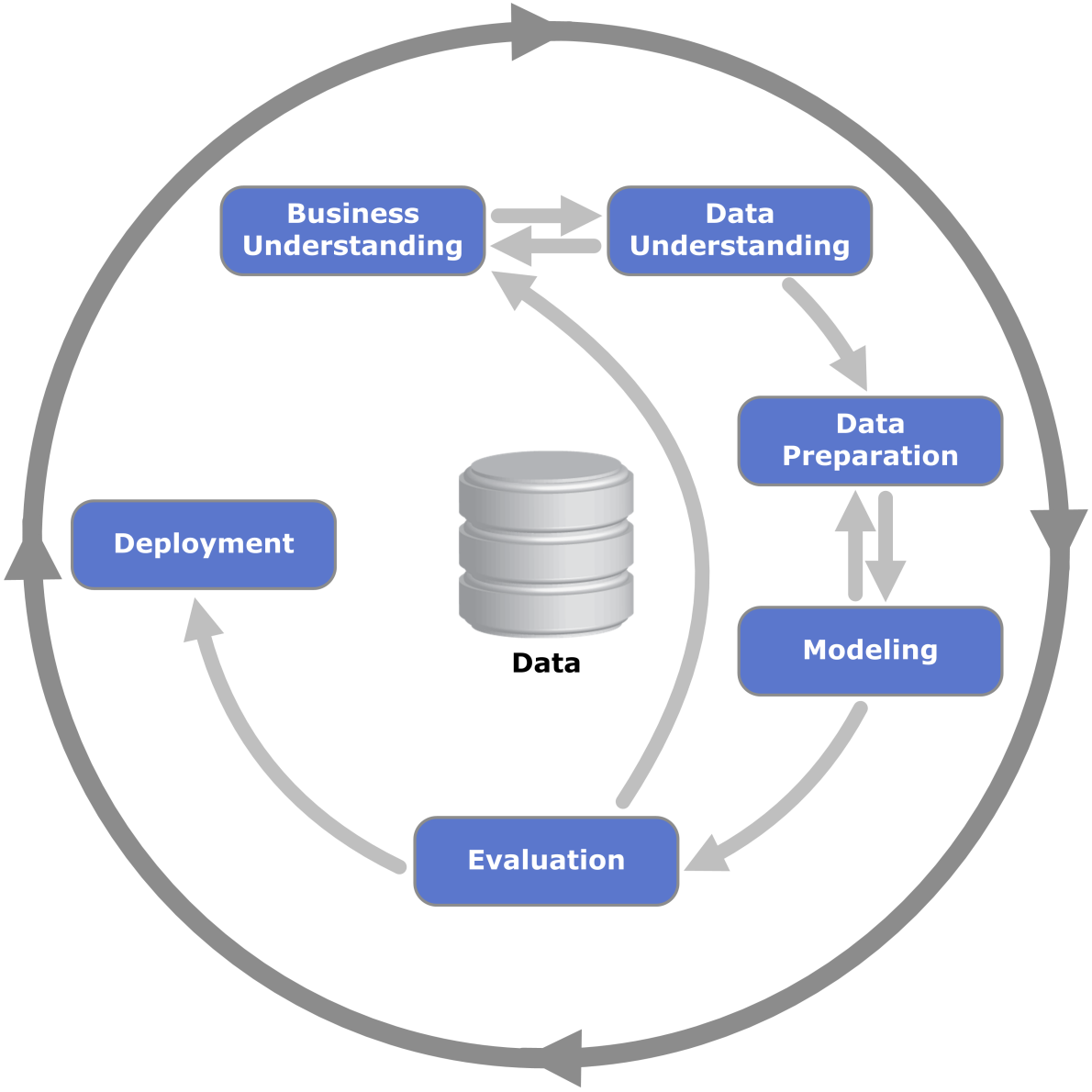
```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

Tidy data

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Why EDA?

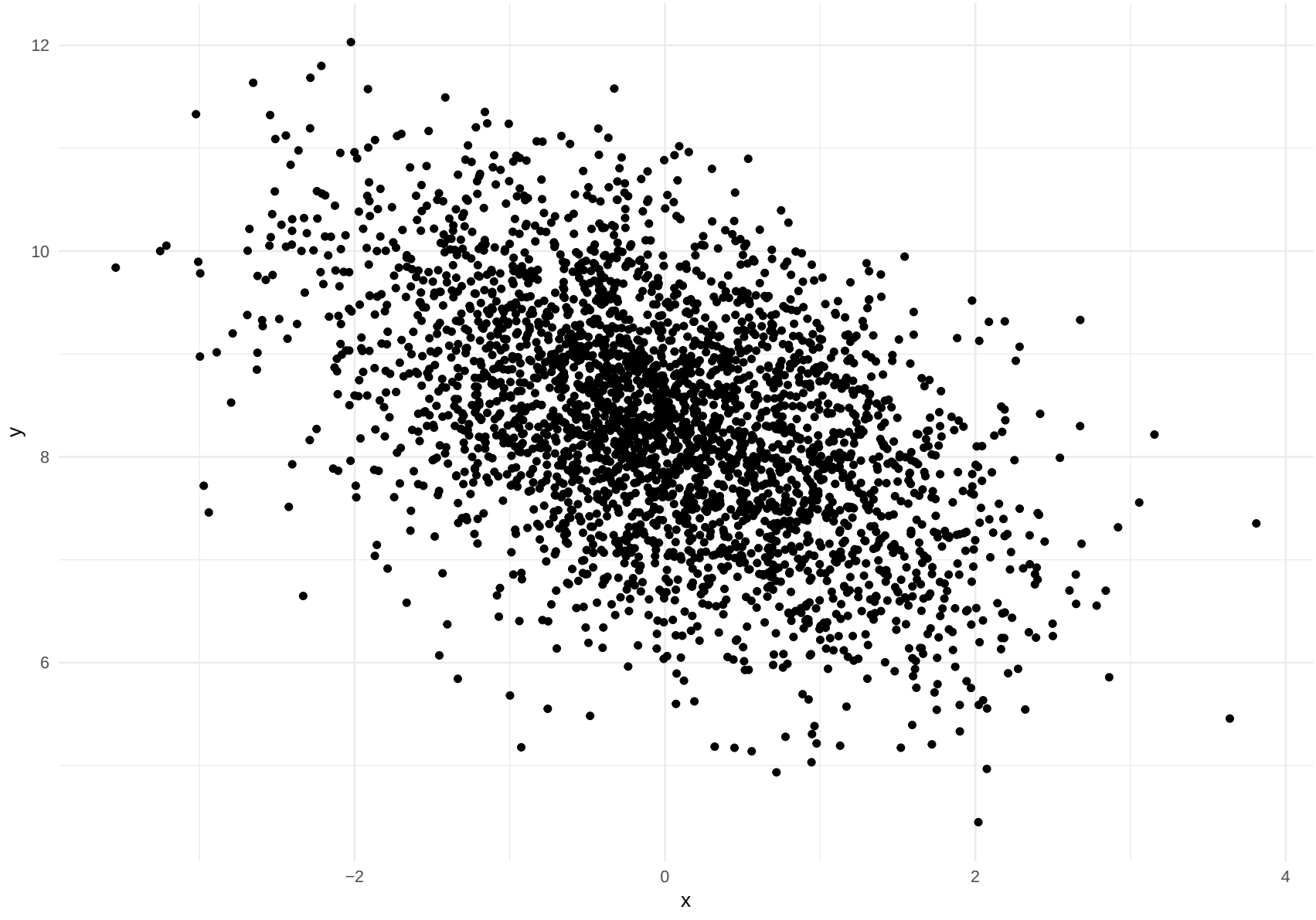
Reminder: Cross Industry Standard Process for Data Mining (CRISP-DM)



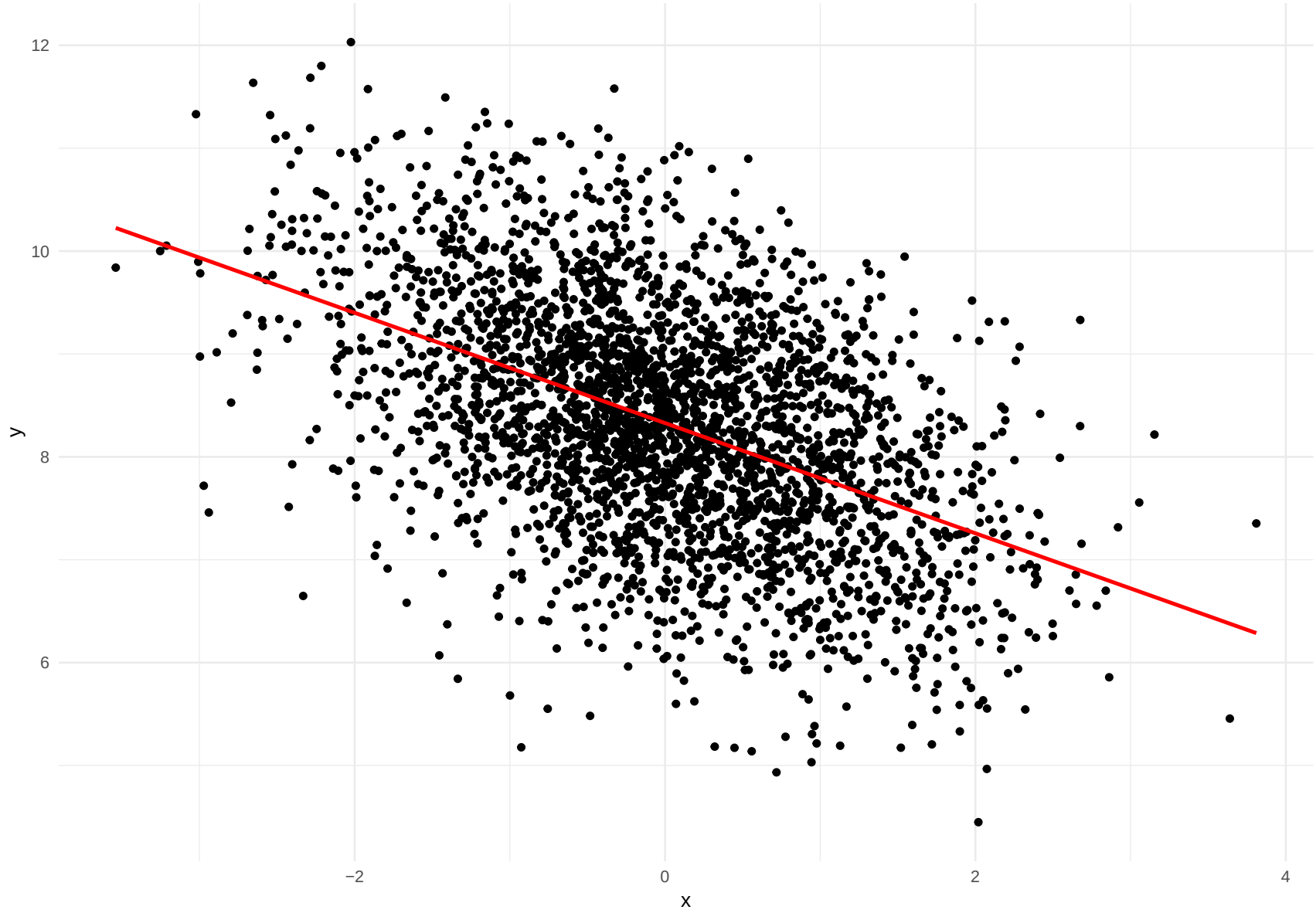
Simpson's paradox

x	y	group
-0.6264538	9.420478	grp07
0.1836433	8.617918	grp06
-0.8356286	10.093969	grp08
1.5952808	6.643508	grp03
0.3295078	6.547648	grp04
-0.8204684	9.722400	grp07
0.4874291	9.868629	grp07
0.7383247	7.866649	grp05
0.5757814	9.899366	grp06
-0.3053884	10.081377	grp07
1.5117812	7.995739	grp04
0.3898432	8.546668	grp06
-0.6212406	7.412821	grp05
-2.2146999	11.799955	grp10
1.1249309	8.040340	grp05

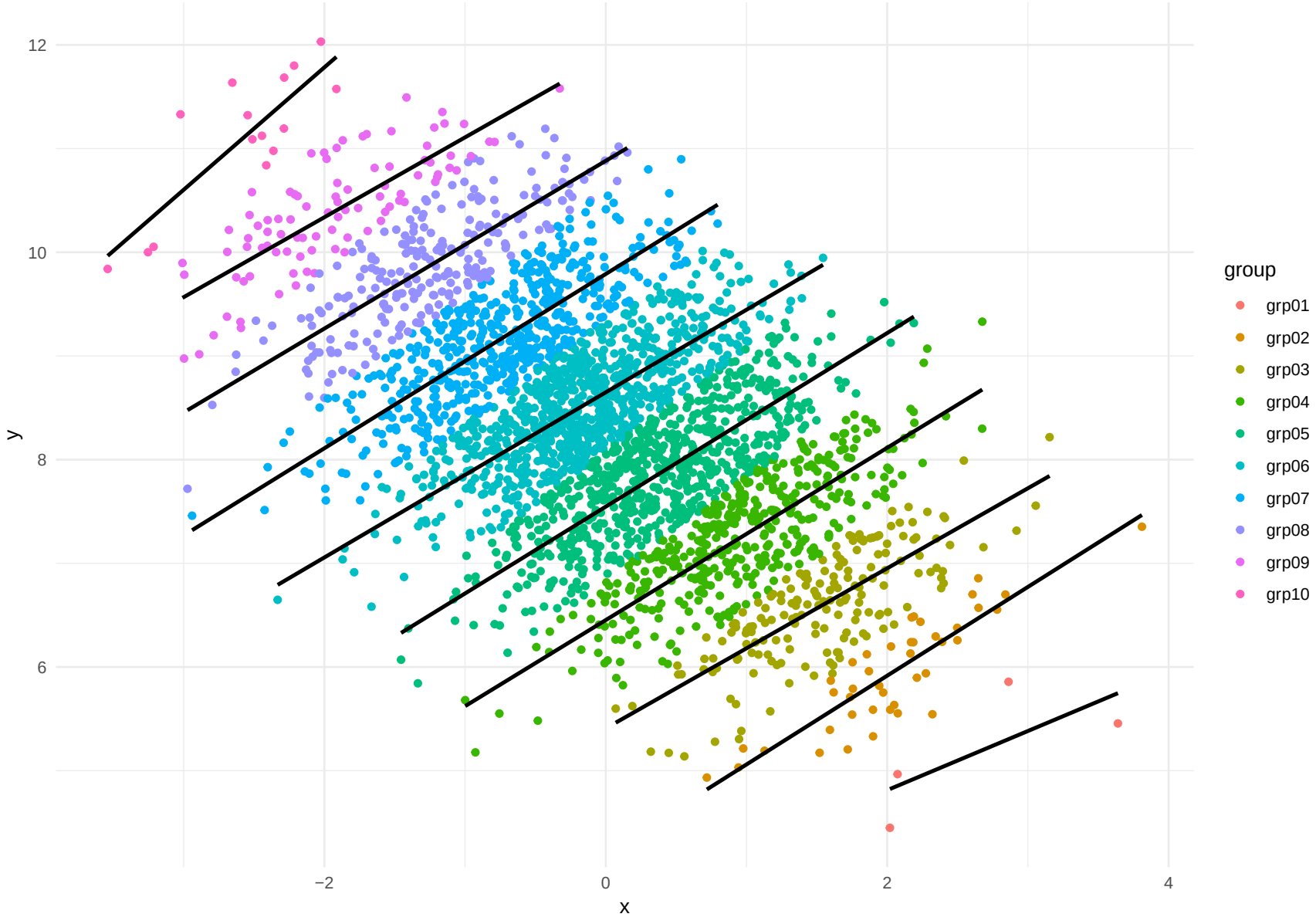
Simpson's paradox



Simpson's paradox



Simpson's paradox



Why EDA?

- We explore the data prior to fitting a model so that we understand the idiosyncrasies of the data and can make informed modelling decisions.

Why EDA?

- We explore the data prior to fitting a model so that we understand the idiosyncrasies of the data and can make informed modelling decisions.
- In the case of Simpson's paradox data we may be interested in the relationship between x and y, but through EDA we (hopefully) learn that we need to examine the relationship between x and y *within* each group.

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data
6. Summarize the data

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data
6. Summarize the data
7. Try the easy solution first

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data
6. Summarize the data
7. Try the easy solution first
8. Challenge your solution

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data
6. Summarize the data
7. Try the easy solution first
8. Challenge your solution
9. Follow up questions

EDA workflow

After understanding the business context and the motivating business problem for the analysis:

1. Formulate a question
2. Read in your data
3. Check the packaging
4. Inspect dataset: str(), glimpse(), View()
5. Look at the top and the bottom of your data
6. Summarize the data
7. Try the easy solution first
8. Challenge your solution
9. Follow up questions

- Adapted from *Exploratory Data Analysis* by Roger Peng.