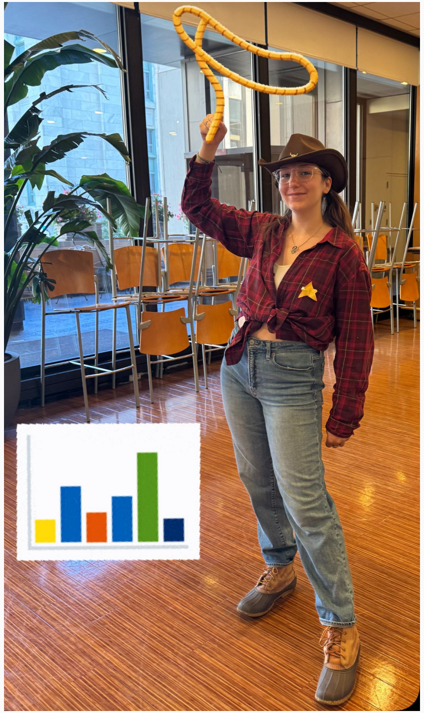
BST 219
Core Principles of Data Science

Lecture 19: TableOne continued and Maps
November 5, 2024

Recipes of the Day!

Election Night Comfort Foods

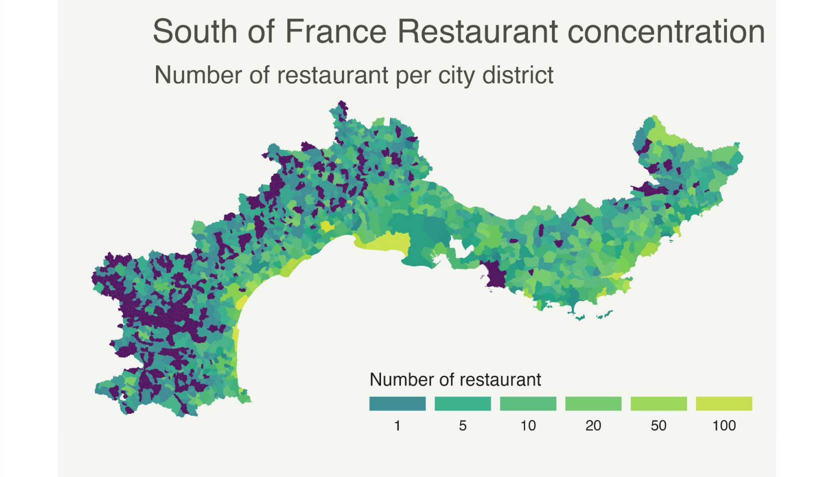
Biostats Halloween Fun: **Ranger Dee Plyer**, the best data wrangler in Boston. No dataset is too tough for a cowgirl like me to handle!



Agenda

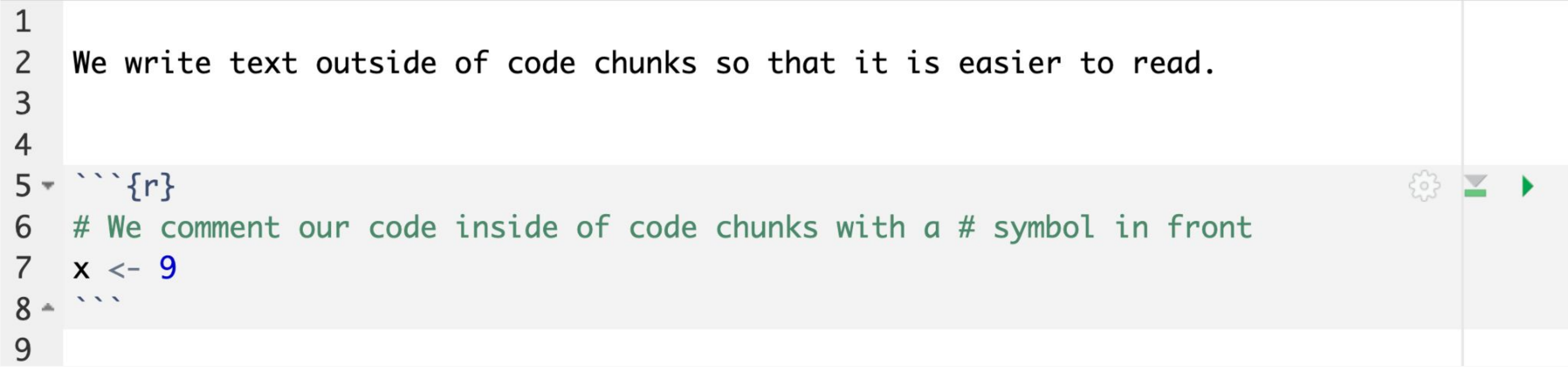
- Announcements
  - No lab this week
  - Homework 3 is due 11/8
  - Midterm 11/8 - 11/17
- Review common HW2 mistakes
- Finish TableOne
- Start maps
- Practice problems
  - Exercises at the end of each chapter in R for Data Science
  - Solutions are available here

Source


South of France Restaurant concentration
Number of restaurant per city district
Number of restaurant
1
5
10
20
50
100

Common Homework #2 Mistakes

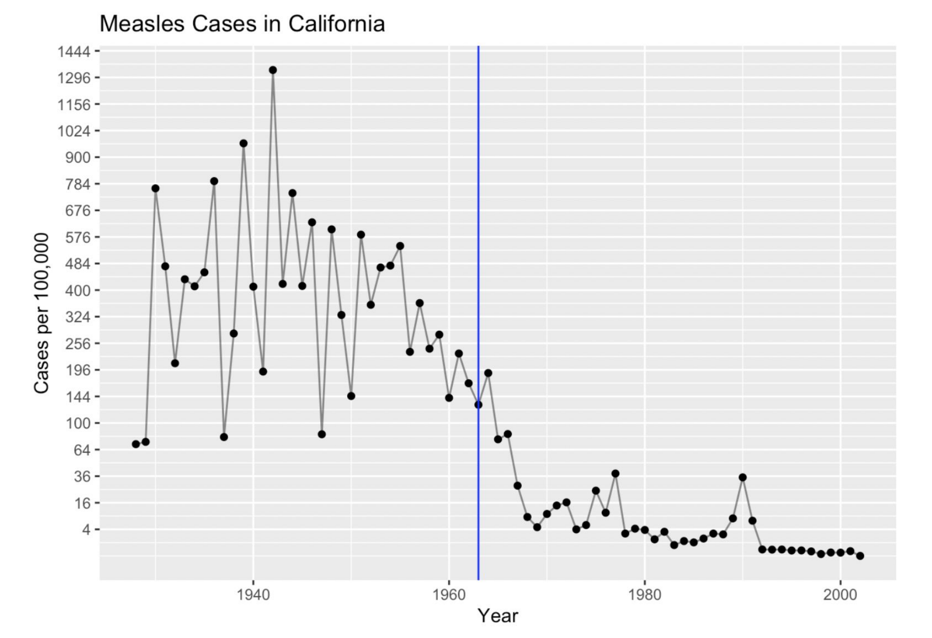
- Writing text answers inside code chucks
    - The reason we are using R Markdown files is so we can separate regular text from code!

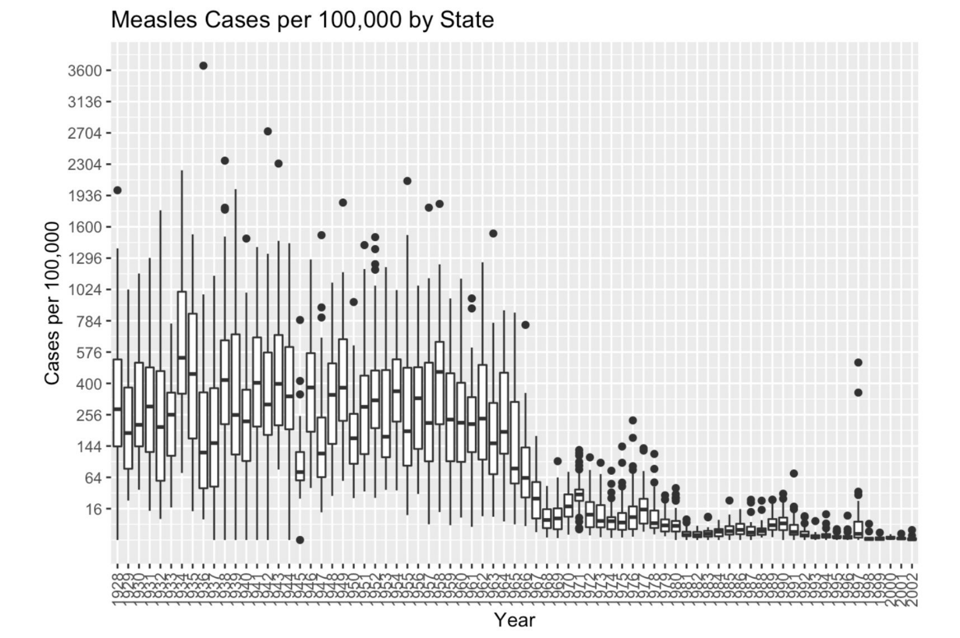


~~~

We write text outside of code chunks so that it is easier to read.


```{r}
# We comment our code inside of code chunks with a # symbol in front
x <- 9
```

~~~

Common Homework #2 Mistakes

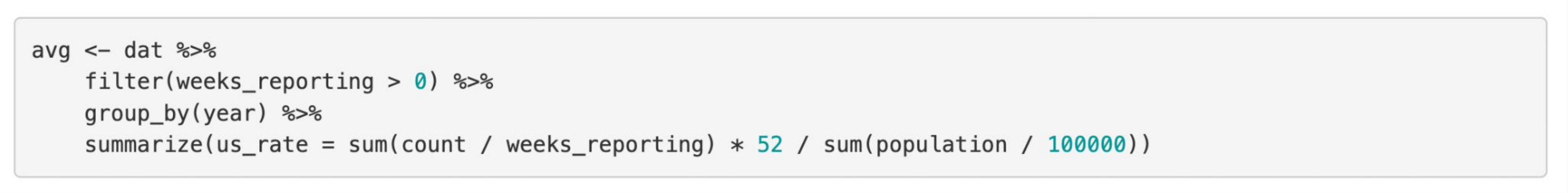
- Question 5: plotting one boxplot for each state rather than one boxplot for each year.
  - Use boxplots to get an idea of the distribution of rates for **each year**, and see if the pattern holds across states.





Common Homework #2 Mistakes

- Question 6: calculating the US rate
    - Taking the average of all of the individual state rates ignores that the population size differs for each state and treats each state the same in terms of contributing to the US rate
    - To calculate the US rate we need to incorporate the sum of **`count`** across states as well as the **`population`** across states - the formula is exactly like the formula used in Question 1, except with two sum functions



```
avg <- dat %>%
    filter(weeks_reporting > 0) %>%
    group_by(year) %>%
    summarize(us_rate = sum(count / weeks_reporting) * 52 / sum(population / 100000))
```

Special Announcements

- Heather’s **11/5** office hour will be moved to **12-1pm**

- Heather’s **11/12** office hour will be moved to **12-1pm**

- The 11/14 lecture will be moved to **11/13, 12:30-2pm via Zoom**

- Lecture on **11/26** will be held via **Zoom**
    - Will be on a special topic that will be stand alone and not part of an assignment

- Heather’s office hour on **11/26** will be **Zoom only**

Coding Question of the Day!

A link to a dataset available on GitHub was sent out in a Canvas announcement. The data comes from the Johns Hopkins University COVID-19 data repository. This particular dataset is called "who_covid_19_sit_rep_time_series.csv" and contains reported COVID-19 case counts from countries around the globe from **January 22, 2020** to **March 23, 2020**.

Read in the data using the code provided in the Canvas announcement. This dataset is currently in wide format. Convert the data frame to long format, with a column named `date` and a `column` named `cases`. Then, make the `date` column values into date objects.

Bonus challenge: rename the `Country/Region` column `country`. Create a new data frame with country name, date, and the total number of cases in the country on each day.