

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Lecture 03: Visualizing Data

Lecture 03: Visualizing Data

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Learning objectives for today:

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Visualizing your data:

1. Making lovely plots using ggplot in R
 - ▶ Visualization of categorical data: use ggplot's `geom_bar()`
 - ▶ Visualization of continuous data: use ggplot's `geom_histogram()`
2. Describe distributions based on shape, center, spread

Visualization of categorical data

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

- ▶ What is the best way to visualize one categorical variable at a time?

Visualization of categorical data

Introducing ggplot

Visualizing quantitative
variables

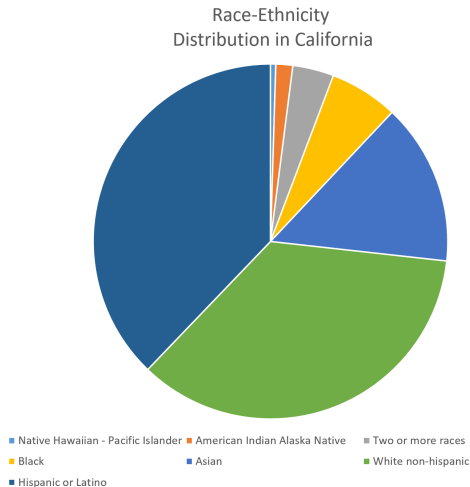
Describing your distribution
based on shape, center and
spread

Time plots

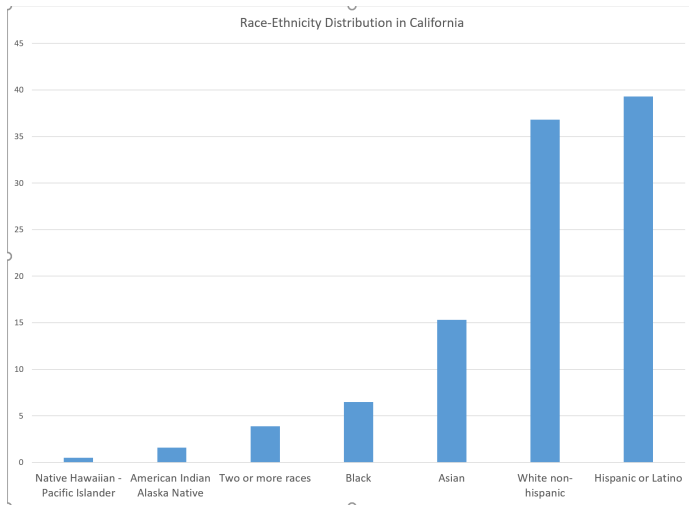
- ▶ Generally speaking, it is not a good idea to use pie charts

Visualziation of categorical data

Can you judge the area of the slices?



Visualziation of categorical data



Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots

Visualization of categorical data

- ▶ We prefer **bar graphs** (also called **bar charts**) for the display of categorical data.
- ▶ Bar charts display the number or percent of data for each level of the categorical variable being plotted

Example: infectious disease data

- ▶ Task: Make a bar chart of the percent of cases on infectious disease for each category of disease.
- ▶ First, read and view the infectious disease data from Baldi and Moore:

```
id_data <- read_csv("Ch01_ID-data.csv")
```

```
## Rows: 7 Columns: 4
## -- Column specification -----
## Delimiter: ","
## chr (2): disease, type
## dbl (2): number_cases, percent_cases
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this m
```

Introducing ggplot

Visualizing quantitative
variablesDescribing your distribution
based on shape, center and
spread

Time plots

Example: infectious disease data

```
id_data
```

```
## # A tibble: 7 x 4
##   disease          type    number_cases percent_cases
##   <chr>          <chr>         <dbl>         <dbl>
## 1 Chlamydia      STI           174557        66.4
## 2 Gonorrhea      STI           44974         17.1
## 3 Pertussis      Pertussis     11219         4.27
## 4 Campylobacteriosis Foodborne     7919         3.01
## 5 Early syphilis STI           7191         2.74
## 6 Salmonellosis  Foodborne     5361         2.04
## 7 Other          Other        11559         4.40
```

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Example: infectious disease data

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

- ▶ Note the variables `number_cases` and `percent_cases`
- ▶ What do you want the bar chart to display?
- ▶ What are the x and y variables for a bar chart?

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Introducing ggplot

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

First step to building a `ggplot()`: set up the canvas

- ▶ The first line of code below pulls in the `ggplot` package
- ▶ The second line of code below specifies the data set and what goes on the x and y axes

```
library(ggplot2) ggplot(id_data, aes(x = disease, y = percent_cases))
```

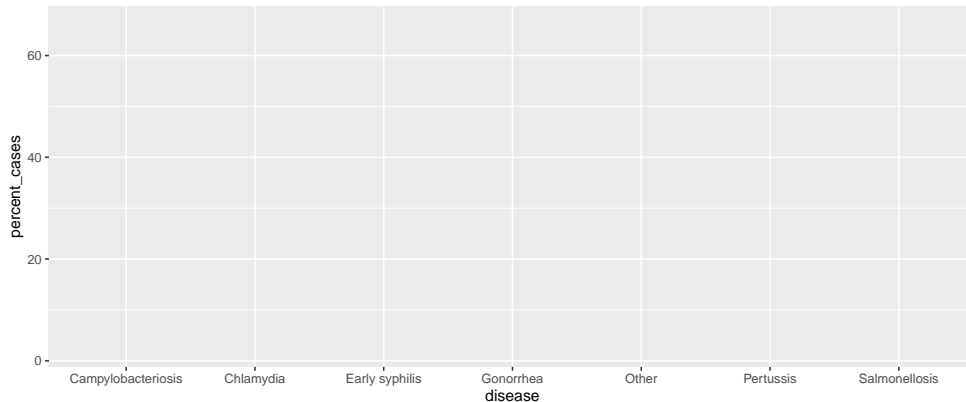
First step to building a `ggplot()`: set up the canvas

Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots



Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Next choose a function

- ▶ We will use a `geom_` function to create our chart

`ggplot()`'s `geom_bar()` makes a bar chart

Syntax for bar charts

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

```
ggplot(id_data, aes(x = disease, y = percent_cases)) +  
geom_bar(stat = "identity")
```

`stat = "identity"` tells `geom_bar` that we supplied a y variable that is exactly what we want to plot.

We do not need `geom_bar()` to calculate the number or percent for us.

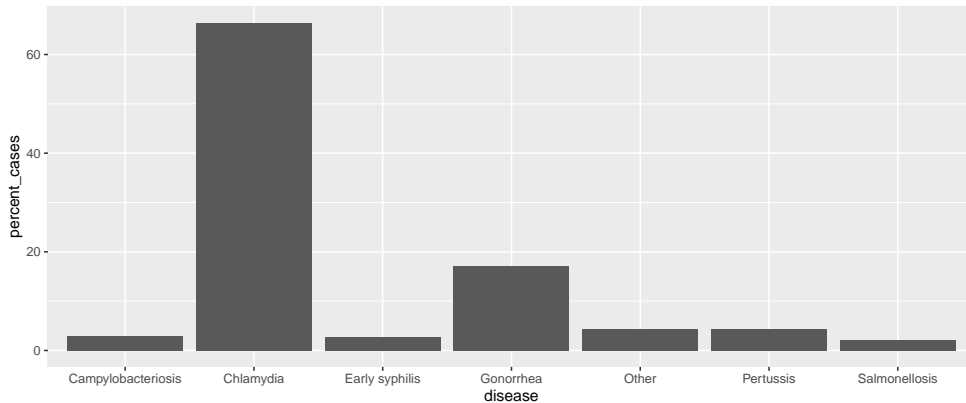
ggplot()'s geom_bar() makes a bar chart

Introducing ggplot

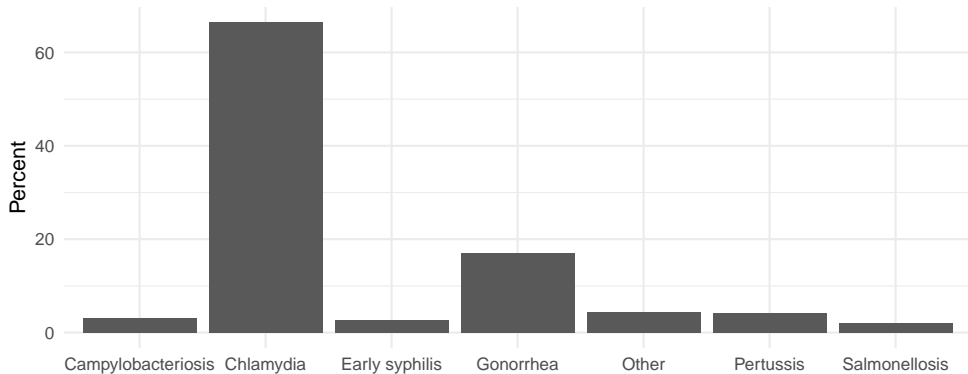
Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots



some additions to ggplot for style



`base_size` controls the font size on these plots

`theme_minimal` affects the “look” of the plot it removes the grey background and adds grey gridlines

Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots

fct_reorder reorders disease according to value of
percent_cases

Introducing ggplot

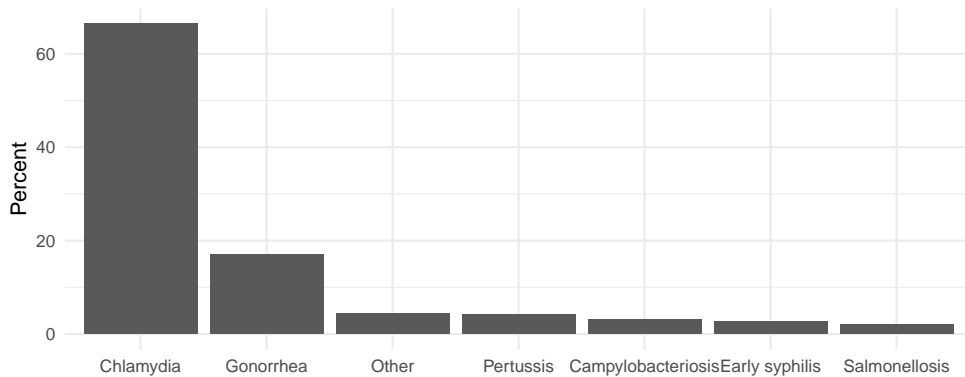
Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

```
id_data <- id_data %>%  
  mutate(disease_ordered = fct_reorder(disease, percent_cases, .desc = T))
```

Re-ordered plot



Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots

Use `aes(fill = type)` to link the bar's fill to the disease type

Introducing ggplot

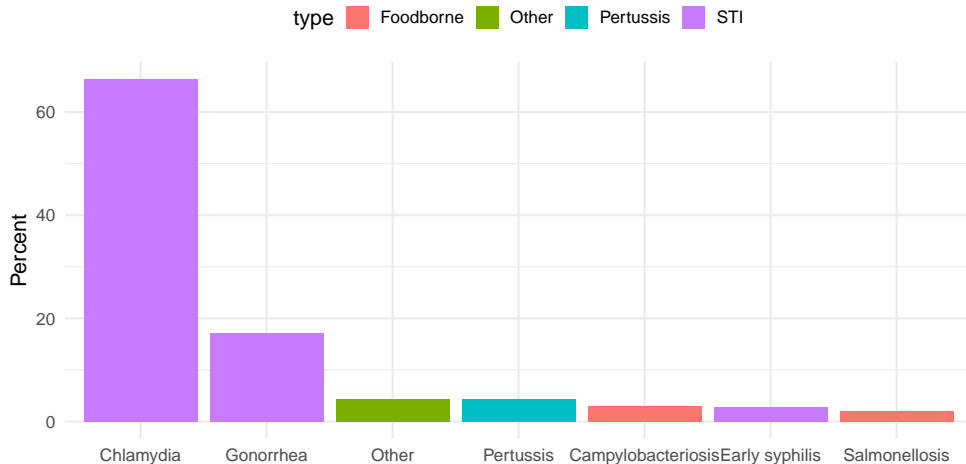
Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

```
geom_bar(stat = "identity", aes(fill = type)) +  
theme(legend.position = "top")
```

Use `aes(fill = type)` to link the bar's fill to the disease type



Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Introducing ggplot

**Visualizing quantitative
variables**

Describing your distribution
based on shape, center and
spread

Time plots

Visualizing quantitative variables

Visualize quantitative variables using histograms

- ▶ Histograms look a lot like bar charts, except that the bars touch because the underlying scale is continuous and the order of the bars matters
- ▶ In order to make a histogram, the underlying data needs to be **binned** into categories and the number or percent of data in each category becomes the height of each bar.
- ▶ the **bins** divide the entire range of data into a series of intervals and counts the number of observations in each interval
- ▶ the intervals must be consecutive and non-overlapping and are almost always chosen to be of equal size

Example: opioid state prescription rates

- ▶ The textbook gives an example using data from 2012.
- ▶ In the data folder, there is updated data from 2018. It came from the paper: “Opioid Prescribing Rates by Congressional Districts, United States, 2016”, by Rolheiser et al. [link](#)

Example: opioid state prescription rates

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Problem: To determine the extent to which opioid prescribing rates vary across US congressional districts.

Plan: In an observational cross-sectional framework using secondary data, they constructed 2016 congressional district-level opioid prescribing rate estimates using a population-weighted methodology.

Data: In the data structure we have State as the unit of analysis, and measured prescription rates as the variable of interest

Example: opioid state prescription rates

```
opi_data <- read.csv("Ch01_opioid-data.csv")  
head(opi_data)
```

##	Rank	State	Mean	Median	SD	Min	Max	Num_Districts
## 1	1	AL	121.31	113.09	21.87	105.58	166.69	7
## 2	2	AR	115.22	115.13	8.59	104.80	125.79	4
## 3	3	TN	108.12	108.26	19.16	73.60	133.00	9
## 4	4	MS	105.64	106.25	17.36	83.90	126.14	4
## 5	5	LA	98.38	98.88	10.34	83.22	112.65	6
## 6	6	KY	98.13	85.76	26.72	77.62	147.00	6

- Mean provides the mean prescribing rate per 100 individuals. Thus, a mean of 121.31 implies that in Alabama, there were 121.31 opioid prescriptions per 100 persons, an average across the 7 congressional districts.

Introducing ggplot

Visualizing quantitative
variablesDescribing your distribution
based on shape, center and
spread

Time plots

Histogram of opioid prescription rates

Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots

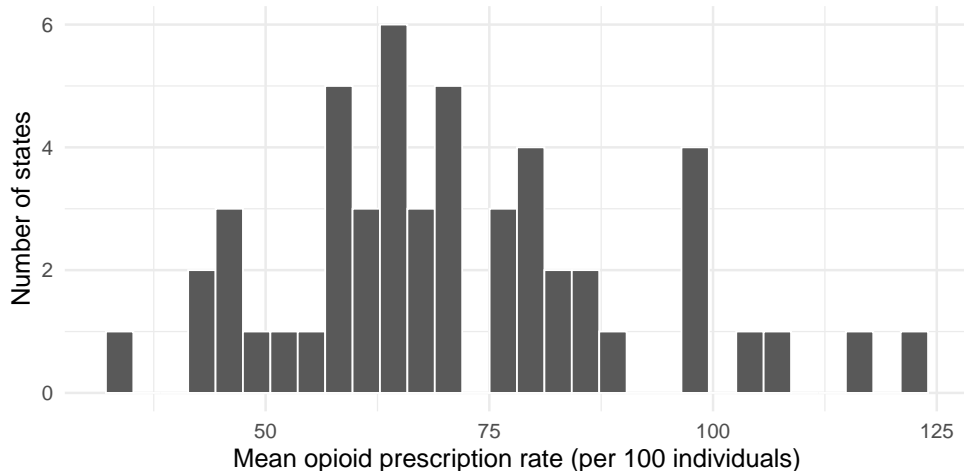
- ▶ Task: Make a histogram of the average prescribing rates across US states
- ▶ What is the x variable? What is the y variable?
- ▶ What geom should be used?

Histogram of opioid prescription rates - default is 30 bins

```
ggplot(data = opi_data, aes(x = Mean)) +  
  geom_histogram(col = "white") +  
  labs(x = "Mean opioid prescription rate (per 100 individuals)",  
       y = "Number of states") +  
  theme_minimal(base_size = 15)
```

Histogram of opioid prescription rates

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'

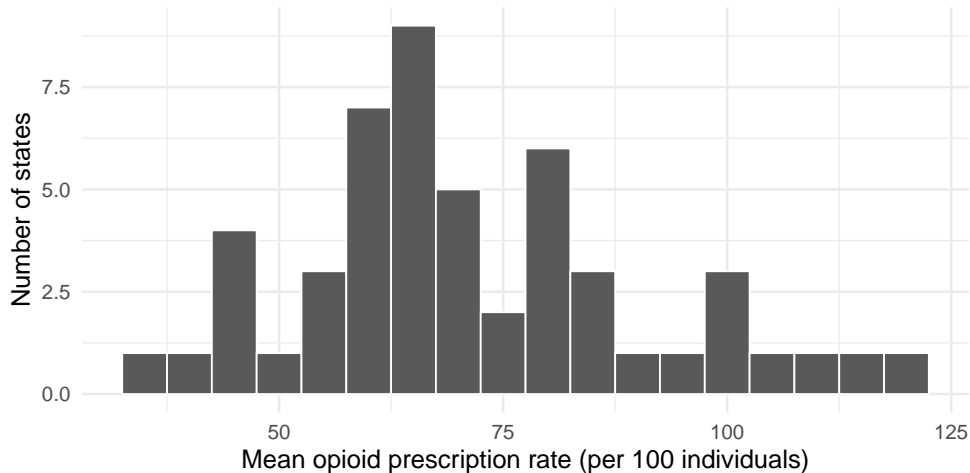


Introducing ggplot
Visualizing quantitative variables
Describing your distribution based on shape, center and spread
Time plots

same graph, change the bins `geom_histogram(binwidth = 5)`

```
ggplot(data = opi_data, aes(x = Mean)) +  
  geom_histogram(col = "white", binwidth = 5) +  
  labs(x = "Mean opioid prescription rate (per 100 individuals)",  
        y = "Number of states") +  
  theme_minimal(base_size = 15)
```

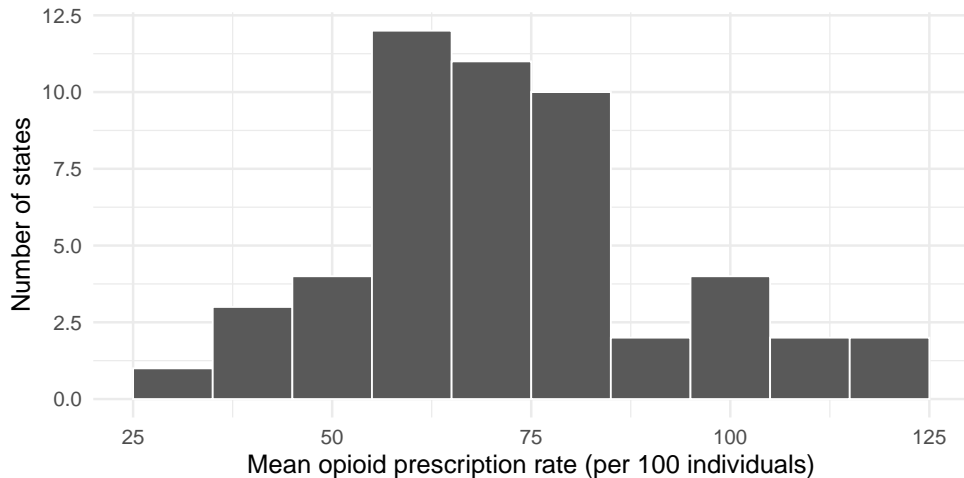
same graph, change the bins `geom_histogram(binwidth = 5)`



change the bins again `geom_histogram(binwidth = 10)`

```
ggplot(data = opi_data, aes(x = Mean)) +  
  geom_histogram(col = "white", binwidth = 10) +  
  labs(x = "Mean opioid prescription rate (per 100 individuals)",  
        y = "Number of states") +  
  theme_minimal(base_size = 15)
```

change the bins again `geom_histogram(binwidth = 10)`



Introducing ggplot

Visualizing quantitative
variables

**Describing your distribution
based on shape, center and
spread**

Time plots

Describing your distribution based on shape, center and spread

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

- ▶ When we examine histograms, we can make comments on a distribution's:
 - ▶ **Shape**: Is the distribution **symmetric** or **skewed** to the left or right?
 - ▶ **Center**: Does the histogram have one peak (unimodal), or two (bimodal) or more?
 - ▶ **Spread**: How spread out are the values? What is the range of the data?
 - ▶ **Outliers**: Do any of the measurements fall outside of the range of most of the data points?

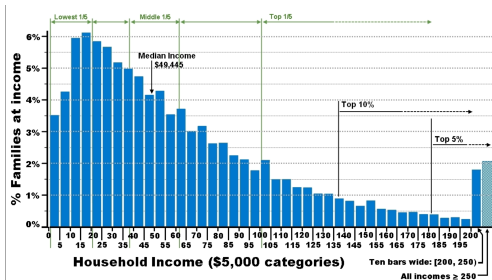
Is this skewed left or skewed right?

Introducing ggplot

Visualizing quantitative
variables

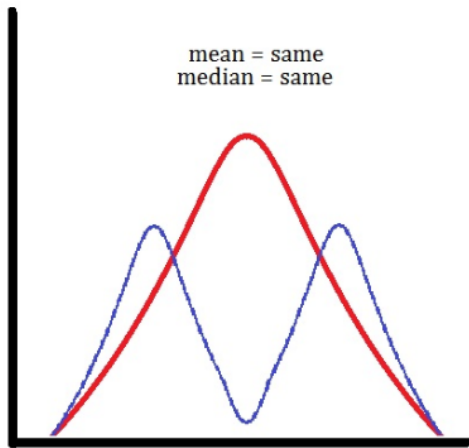
Describing your distribution
based on shape, center and
spread

Time plots



Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

Center - one hump or two?



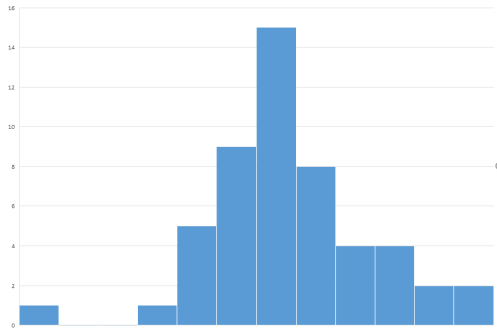
Outlier

Introducing ggplot

Visualizing quantitative
variables

**Describing your distribution
based on shape, center and
spread**

Time plots



Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

Time plots

Visualize quantitative variables over time using time plots

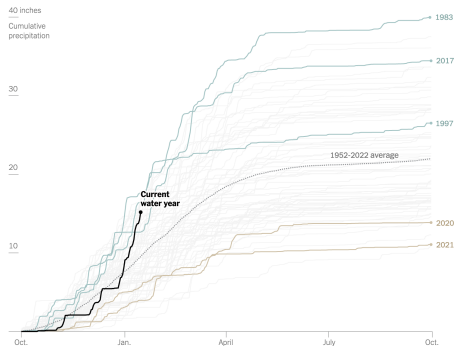
- ▶ **Time plots** are a specific subset of plots where the x variable is time.
- ▶ Unlike the previous plots, the time plot shows a relationship between two variables:
 - i) a quantitative variable
 - ii) time
- ▶ Often times, these plots can be used to look for cycles (e.g., seasonal patterns that recur each year) or trends (e.g., overall increases or decreases seen over time).

Time plot

► from nytimes.com 19 Jan 2023 article:

California's Storms in Context

Cumulative rainfall across the state is above average so far this winter, but other years have been even wetter.



Note: The California water year starts in October, aligning with the typical beginning of the rainy season. Current water year data as of Jan. 15, 2023. Source: NOAA NClimGrid By Mira Rojanasakul

Introducing ggplot

Visualizing quantitative variables

Describing your distribution based on shape, center and spread

Time plots

Life expectancy for White men in California

Make a scatter plot of the life expectancy for White men in California over time.

Since the dataset contains 39 states across two genders and two races, first use a function to subset the data to contain only White men in California.

Which function from last lecture do we need?

► `mutate()`, `select()`, `filter()`, `rename()`, or `arrange()`?

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

dplyr's filter() to select a subset of rows

```
wm_cali <- le_data %>% filter(state == "California",  
                               sex == "Male",  
                               race == "white")
```

#this is equivalent:

```
wm_cali <- le_data %>% filter(state == "California" & sex == "Male" & race ==
```

Here we use `geom_point` to make a graph with dots

```
ggplot(data = wm_cali, aes(x = year, y = LE)) +  
  geom_point() +  
  labs(title = "Life expectancy in white men in California, 1969-2013",  
  
        y = "Life expectancy",  
  
        x = "Year",  
  
        caption = "Data from Riddell et al. (2018)")
```

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

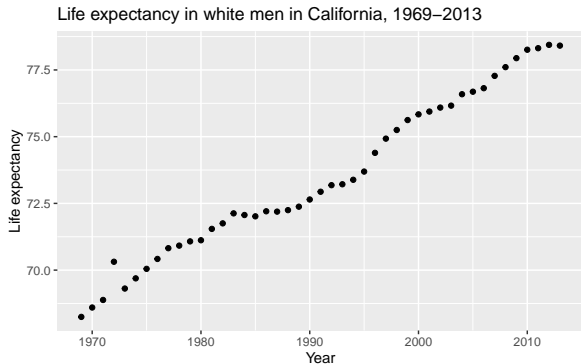
Here we use `geom_point` to make a graph with dots

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots



Data from Riddell et al. (2018)

geom_line() to make a line plot

```
ggplot(data = wm_cali, aes(x = year, y = LE)) +  
  geom_line(col = "blue") +  
  labs(title = "Life expectancy in white males in California, 1969-2013",  
  
        y = "Life expectancy",  
  
        x = "Year",  
  
        caption = "Data from Riddell et al. (2018)")
```

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

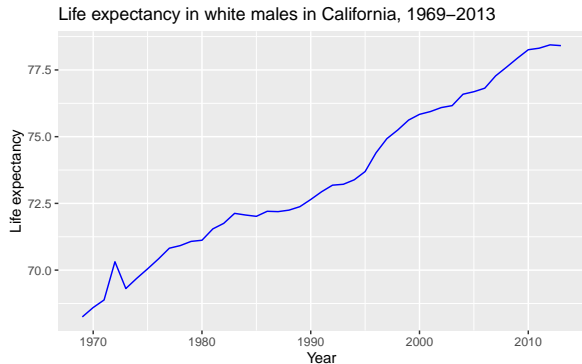
geom_line() to make a line plot

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots



Data from Riddell et al. (2018)

R Recap: new code?

1. `'ggplot'` to set up a canvas for graphics
2. `geom_bar(stat = "identity")` to make a bar chart when you specify the y variable
3. `geom_histogram()` to make a histogram for which ggplot needs to calculate the count
4. `fct_reorder(var1, var2)` to reorder a categorical variable (`var1`) by a numeric variable (`var2`)
 - ▶ from the `forcats` package
5. `geom_point()` to make a plot with dots
6. `geom_line()` to make a plot with lines

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

- ▶ Ask questions during labs, GSI office hours, or on Piazza discussion forum. Use the appropriate thread!
- ▶ Develop your online search skills. For example if you have a `ggplot2` question, begin your google search with “r ggplot” and then describe your issues, e.g., “r ggplot how do I make separate lines by a second variable”.
- ▶ The most common links that will appear are:
 - ▶ <https://stackoverflow.com>: Crowd-sourced answers that have been upvoted. The top answer is often the best one.
 - ▶ <https://ggplot2.tidyverse.org/>: The official ggplot2 webpage is very helpful.
 - ▶ <https://community.rstudio.com/>: The RStudio community page.
 - ▶ <https://rpubs.com/>: Web pages made by R users that often contain helpful tutorials.

We only skimmed the surface!

Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

- ▶ Here is some extra material for those of you who love data visualization. This material won't be tested.
 - ▶ RStudio ggplot2 cheatsheet
 - ▶ Kieran Healy's data visualization book

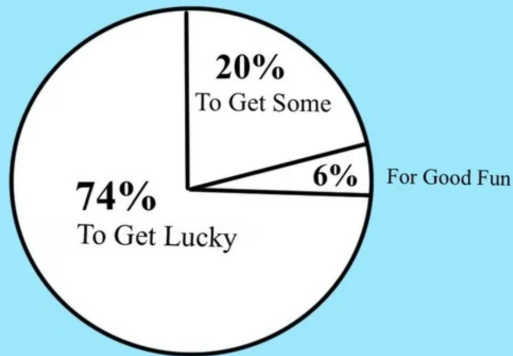
Introducing ggplot

Visualizing quantitative
variables

Describing your distribution
based on shape, center and
spread

Time plots

REASONS WE'RE UP ALL NIGHT



Source: Daft Punk (research assistance by Pharrell Williams)