

Grammar of Graphics intro

VISUALIZATION BEST PRACTICES IN R



Nick Strayer
Instructor

What is this course?

What you will learn

How to make better visualizations by thinking deeply about the data at hand.



How you will learn it

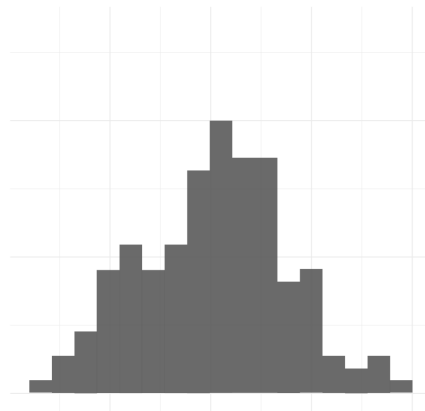
- Overviews of different data types
- Standard visualizations
- Alternatives

Course layout

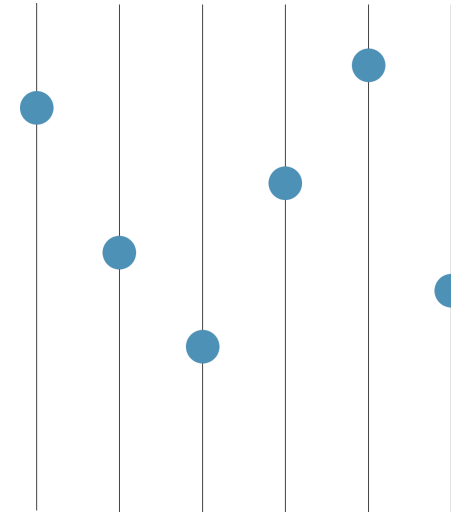
Chapter 1: Proportions of a whole



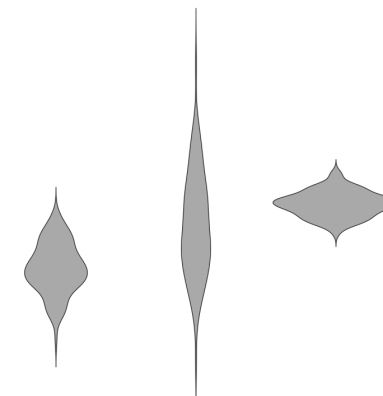
Chapter 3: Single distributions



Chapter 2: Point data



Chapter 4: Multiple(or conditional) distributions



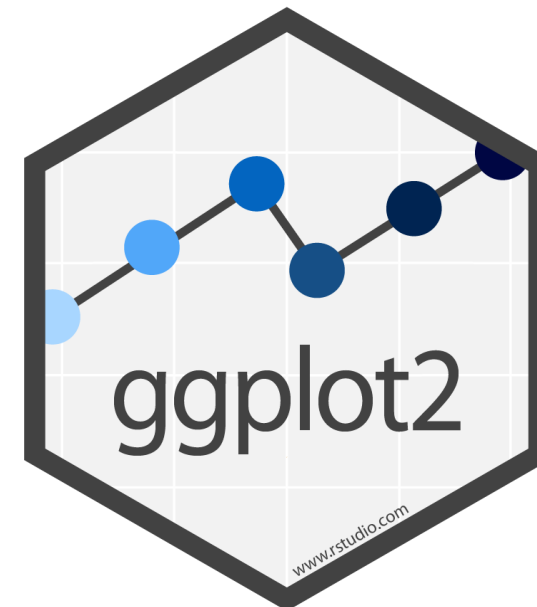
Warning!



- Topics here are not as cut and dry as other programming topics
- Every rule will have exceptions
- An emphasis on thinking through each problem is given to help you deal with these cases when you get to them

Tools used

- R
- The 'Tidyverse'
- Ggplot2



Data used

- Comes from the World Health Organization (WHO)

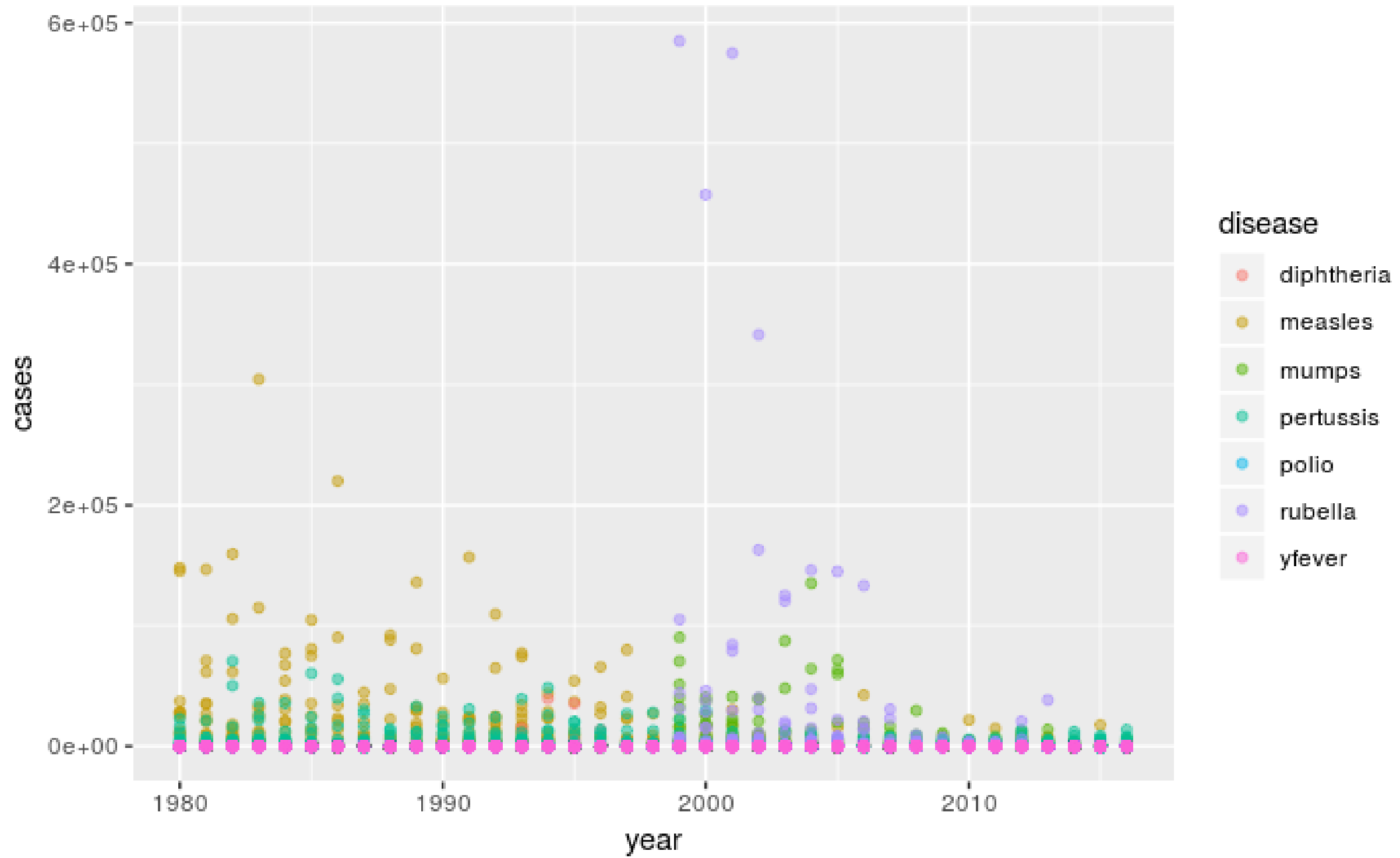
```
who_disease
```

```
# A tibble: 43,262 x 6
  region countryCode country      disease year  cases
  <chr>   <chr>      <chr>      <chr>   <int> <dbl>
1 EMR     AFG        Afghanistan measles  2016  638
2 EUR     ALB          Albania   measles  2016  17.0
3 AFR     DZA          Algeria    measles  2016  41.0
4 EUR     AND          Andorra    measles  2016    0
5 AFR     AGO          Angola     measles  2016  53.0
6 AMR     ATG  Antigua and Barbuda measles  2016    0
7 AMR     ARG          Argentina measles  2016    0
8 EUR     ARM          Armenia    measles  2016   2.00
# ... with 43,254 more rows
```

WHO disease data

```
# Filter to AMR region
amr_region <- who_disease %>%
  filter(region == 'AMR')

# Map x to year and y to cases, color by disease
ggplot(amr_region, aes(x = year, y = cases, color = disease)) +
  geom_point(alpha = 0.5)
```



Let's practice!

VISUALIZATION BEST PRACTICES IN R

The pie chart and its friends

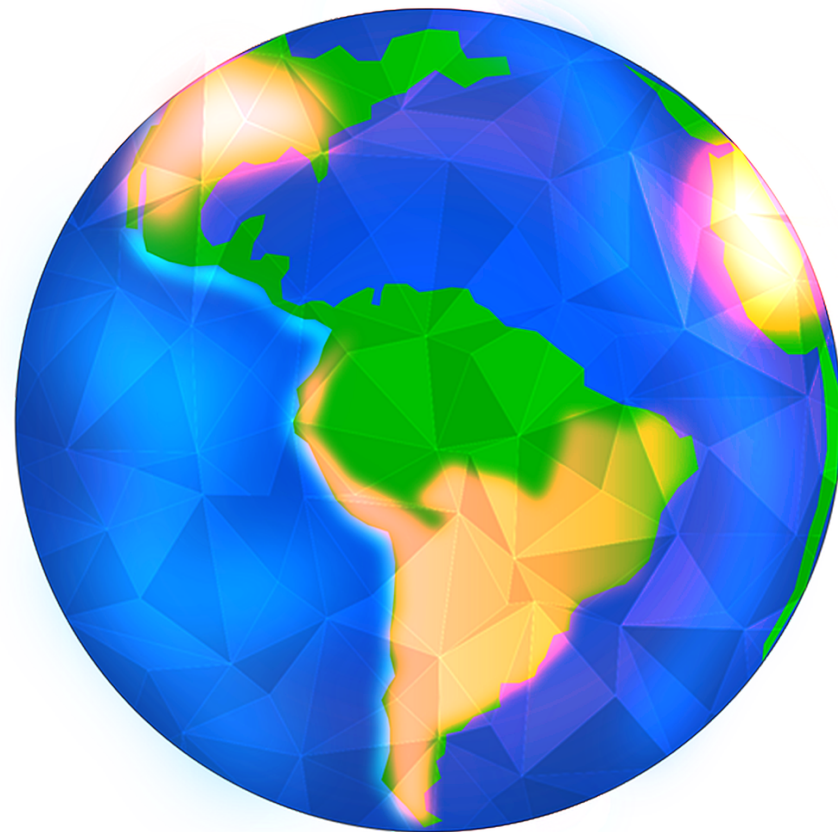
VISUALIZATION BEST PRACTICES IN R



Nick Strayer
Instructor

What is a proportion?

- Parts making up a whole
- Often used to understand population



The pie chart

- Often the first technique people learn
- Also, the first technique people learn to dislike
- Dislike is not *entirely* warranted

A sour pie

- Pie charts are not very precise
 - Data encoded in angles
- Doesn't handle lots of classes well
 - After three slices it becomes hard to compare

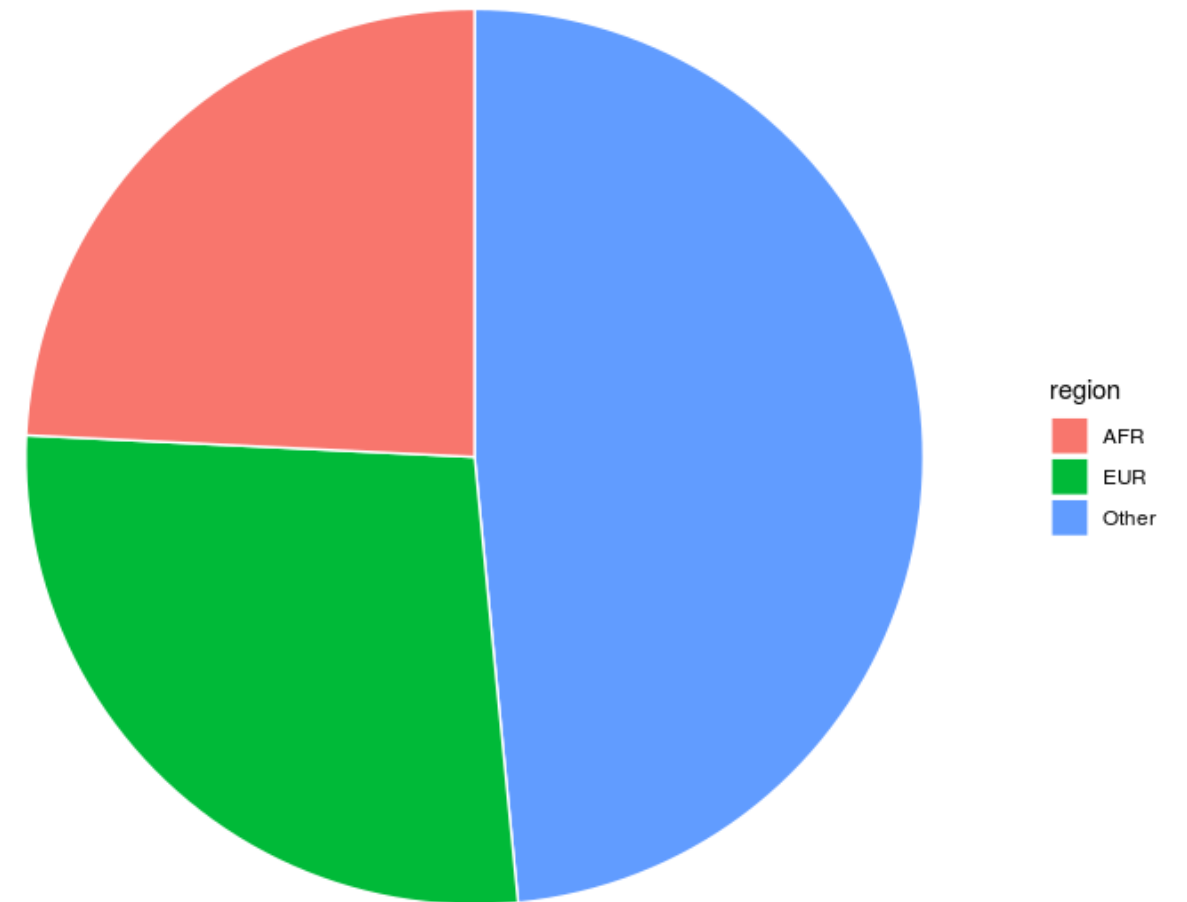


A sweet pie

- Intuitive and compact

```
who_disease %>%  
  mutate(  
    region = ifelse(  
      region %in% c('EUR', 'AFR'),  
      region, 'Other')  
  ) %>%  
  ggplot(aes(x = 1, fill = region)) +  
    geom_bar(color = 'white') +  
    coord_polar(theta = "y") +  
    theme_void()
```

Proportion of observations by region.



The waffle chart

- More precise than pie charts
- Encode data in area, not angles

```
obs_by_region <- who_disease %>%  
  group_by(region) %>% summarise(num_obs = n()) %>%  
  mutate(percent = round(num_obs/sum(num_obs)*100))  
  
# Array of rounded percentages  
percent_by_region <- obs_by_region$percent  
names(percent_by_region) <- obs_by_region$region  
  
# Send array of percentages to waffle plot function  
waffle::waffle(percent_by_region, rows = 5)
```

The waffle chart

Proportion of observations by region.



Let's practice!

VISUALIZATION BEST PRACTICES IN R

When to use bars

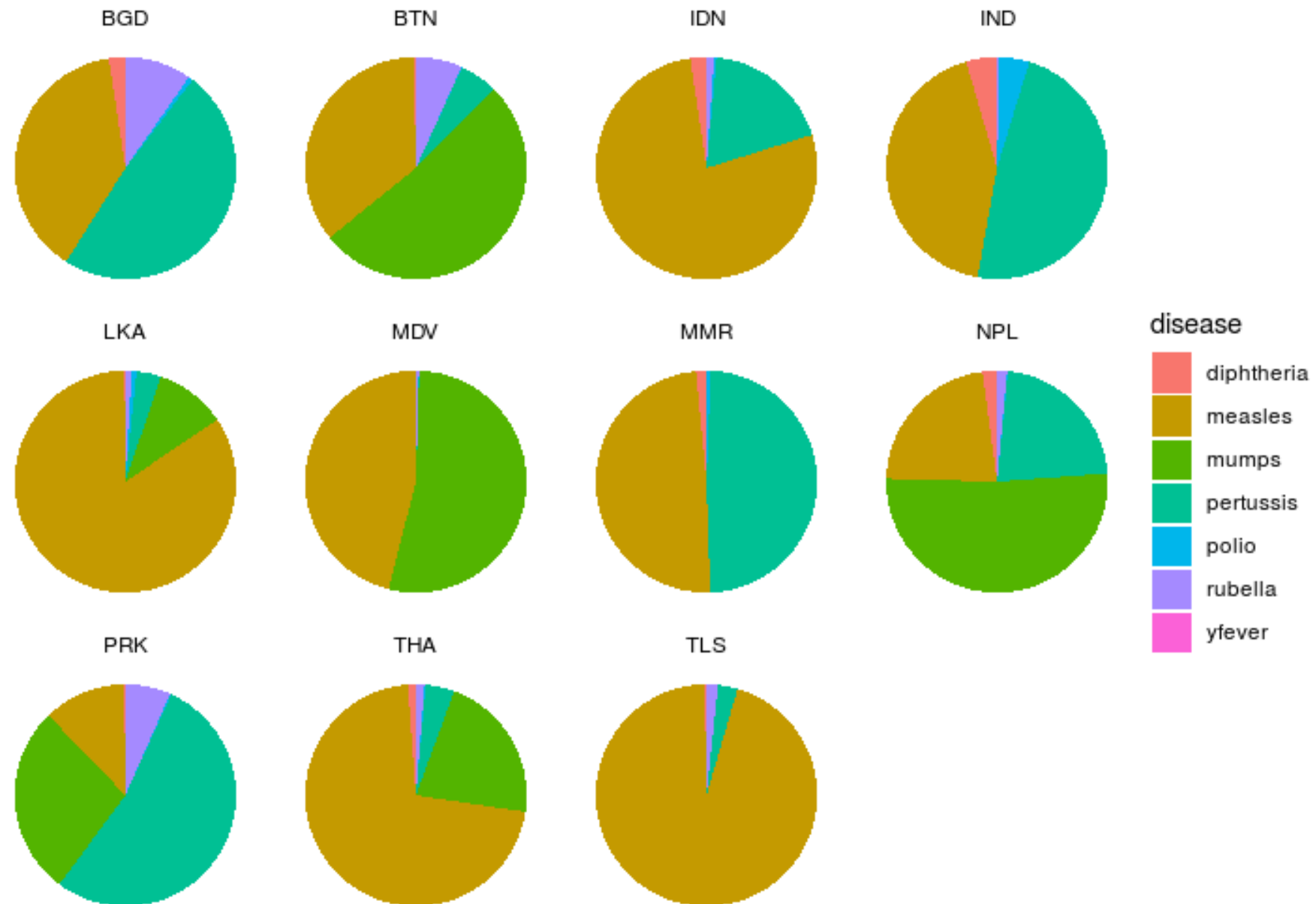
VISUALIZATION BEST PRACTICES IN R



Nick Strayer
Instructor

Why not use faceting?

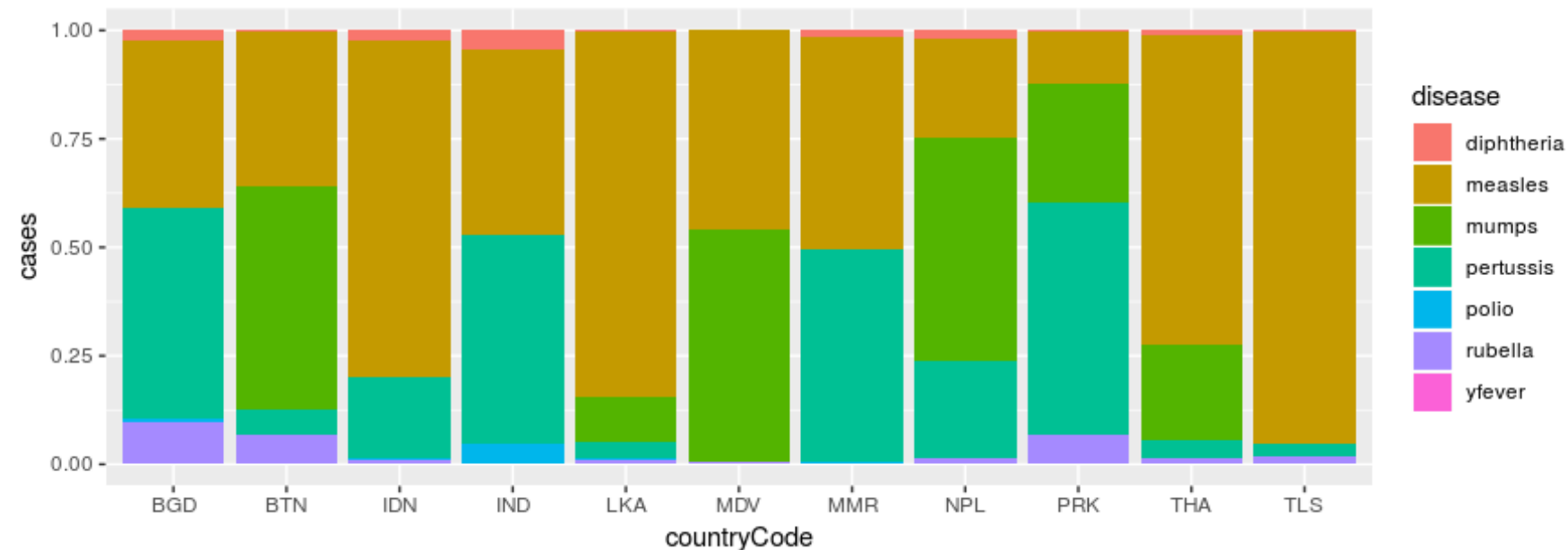
- Almost impossible to compare



The stacked bar chart

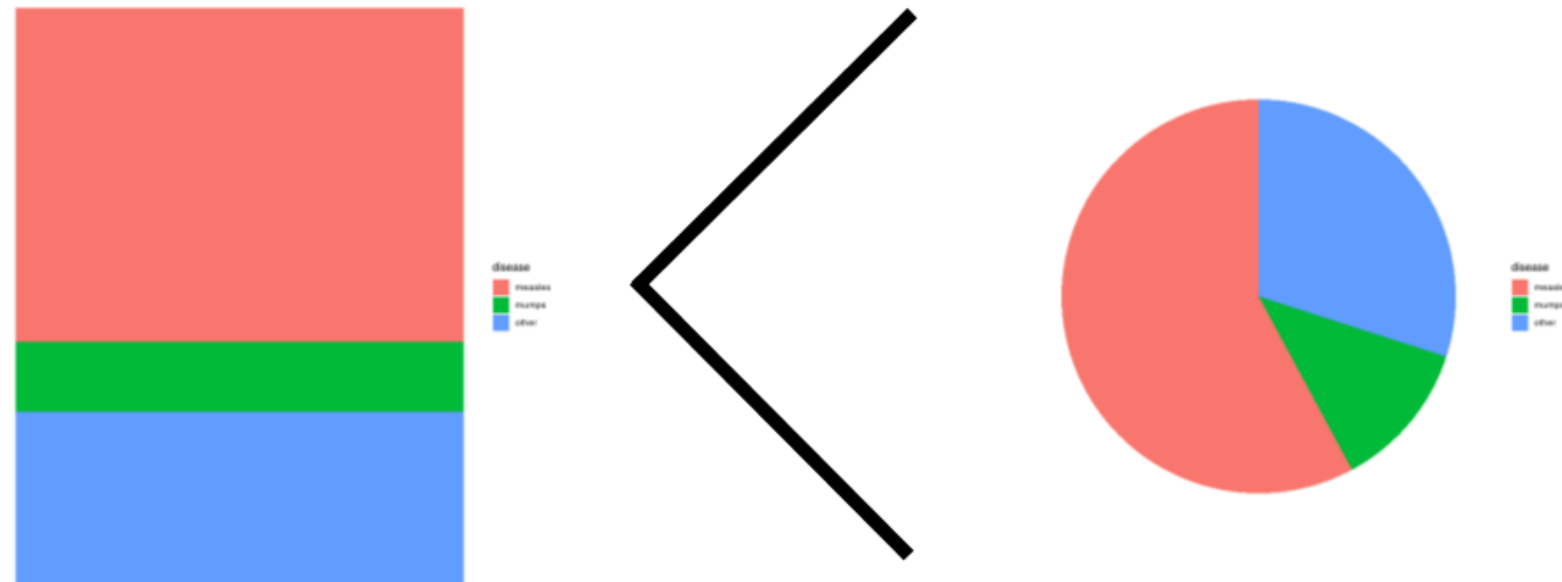
- Allow each population to share the same y-axis
- Enables easier comparisons based on vertical position/size

```
who_disease %>%  
  filter(region == 'SEAR') %>%  
  ggplot(aes(x = countryCode, y = cases, fill = disease)) +  
    geom_col(position = 'fill')
```



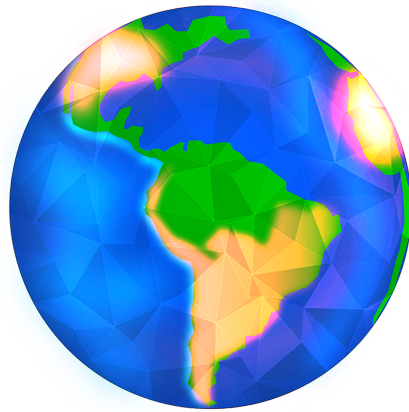
Caveats

- Worse in isolation than pie or waffle charts
- Accuracy degrades rapidly after 3 classes



Chapter recap

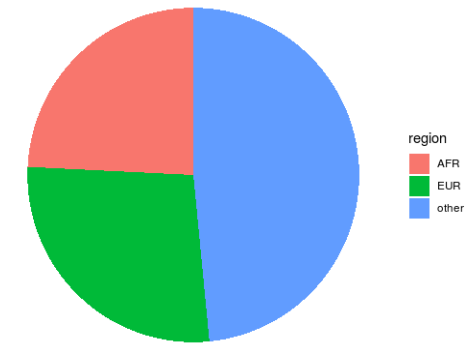
Proportions:



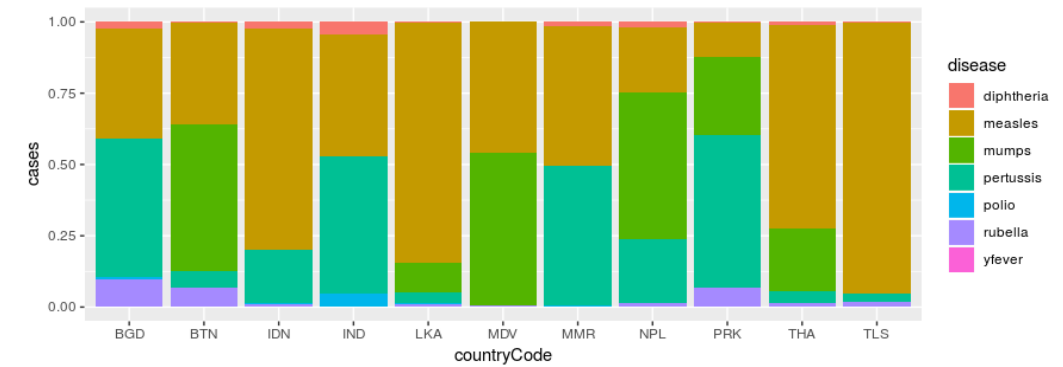
Waffle charts:



Pie charts:



Stacked bars:



Let's practice!

VISUALIZATION BEST PRACTICES IN R