

STAT 220

Data Viz with Q

Who am I?

- Husband of Professor Kurtz
- Father of Glenn
- Player of Video Games
- Watcher of Anime
- Maker of Maple Syrup



Why Me?

- Group member who always prettied up slides
- Data Visualization Developer at Mayo Clinic
- Varied background
 - Mechanical Engineering -> Software Support -> MBA -> Viz Developer
- Sister is a Carleton Alum

Why Do We Visualize?

- Human brains spent millennia evolving the ability to see patterns in the natural world
- Data helps us understand the world, but is unnatural to our brains
- Visualizations let us trick our brain's natural pathways into understanding data

Viz Contexts

Investigation

- **Complex** visuals that help us analyze the data and discover patterns
- Important when **analyzing** data

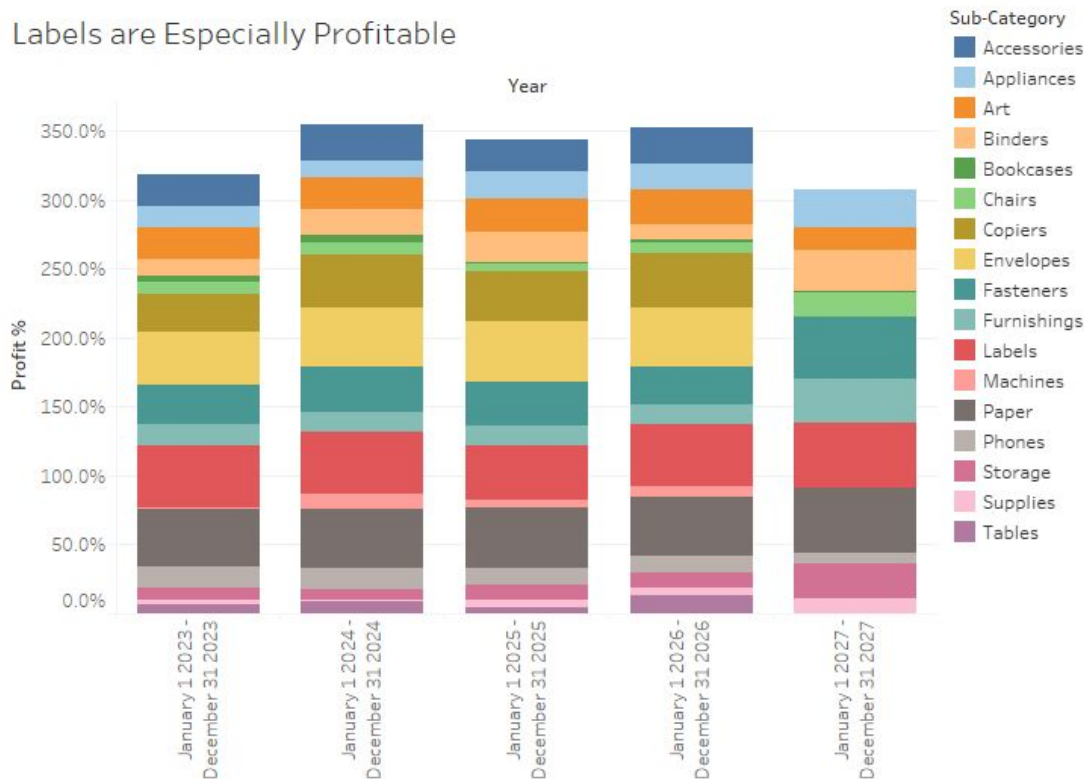
Storytelling

- **Clean** visuals that make the important parts obvious
- Important when **sharing** with others

Roast a Viz

- What feels bad about this?

Labels are Especially Profitable



Viz Principles

- Use Preattentive Attributes sparingly & deliberately
- Design for colorblindness
- Reduce mental burden
- Pick the right chart for the task
- Keep it honest
- Keep it simple

Preattentive Attributes



Orientation



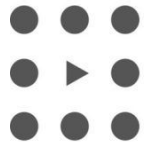
Length



Width



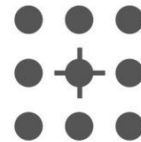
Size



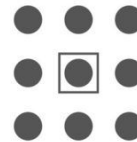
Shape



Curvature



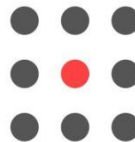
Added Marks



Enclosure



Contrast



Colour

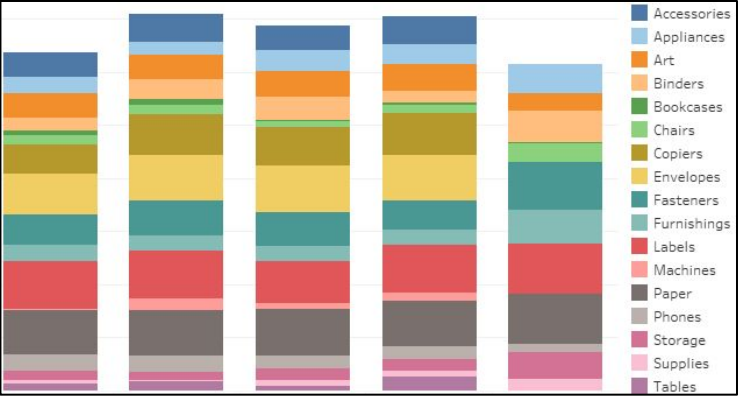


Position

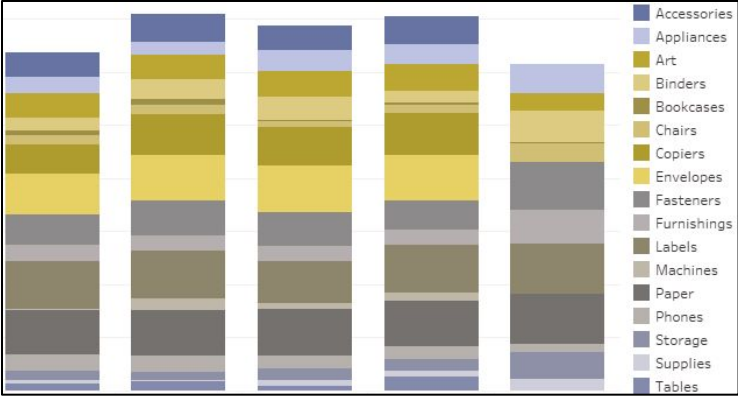


Spatial Grouping

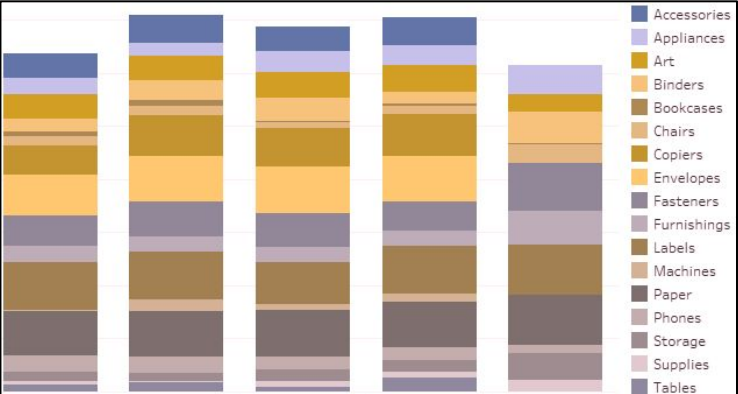
Design for Colorblindness



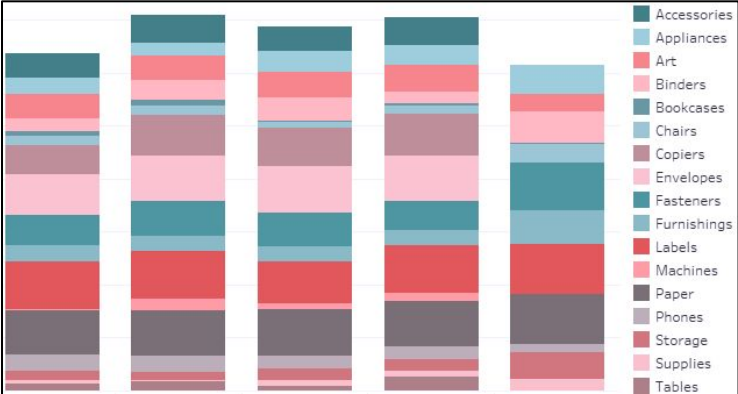
Standard



Red-Blind



Green-Blind



Blue-Blind

Colorblindness Resources

Tips

- Avoid using red and green in the same viz
- Add white borders to visually separate colors
- Using different brightnesses of the same color is usually safe
 - E.g. light blue, blue, and dark blue
- Use other Preattentive Attributes in addition to color (e.g. shape)

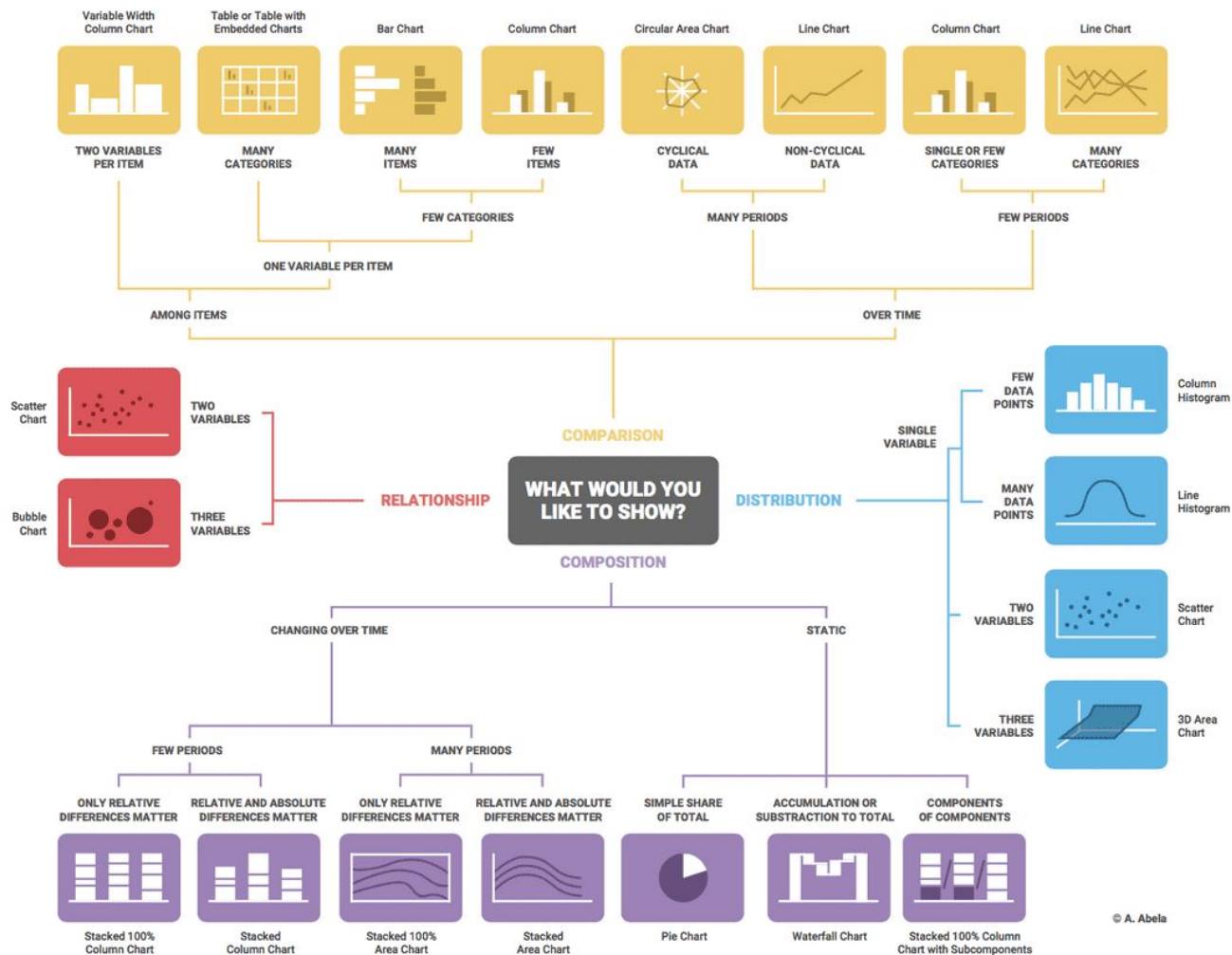
Resources

- <https://www.color-blindness.com/coblis-color-blindness-simulator/>
- <https://davidmathlogic.com/colorblind/>

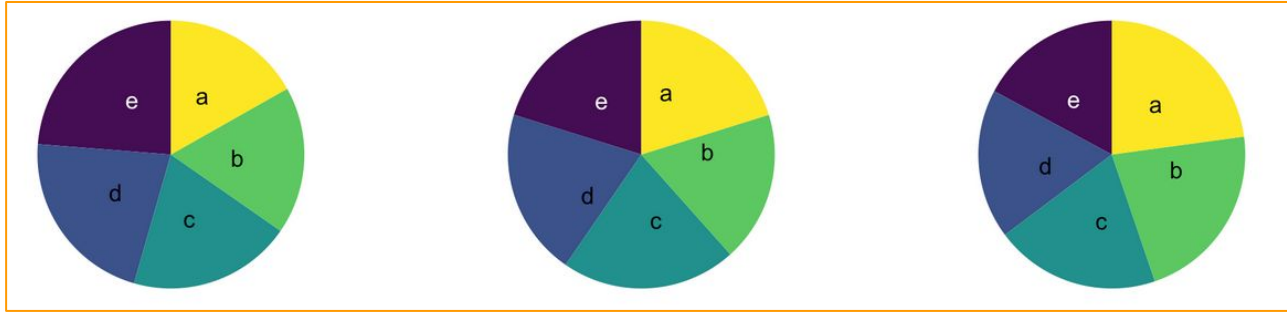
Reduce Mental Burden

- Look for ways to streamline the user's experience
- Examples:
 - Direct labels on lines rather than legends
 - Keep text horizontal

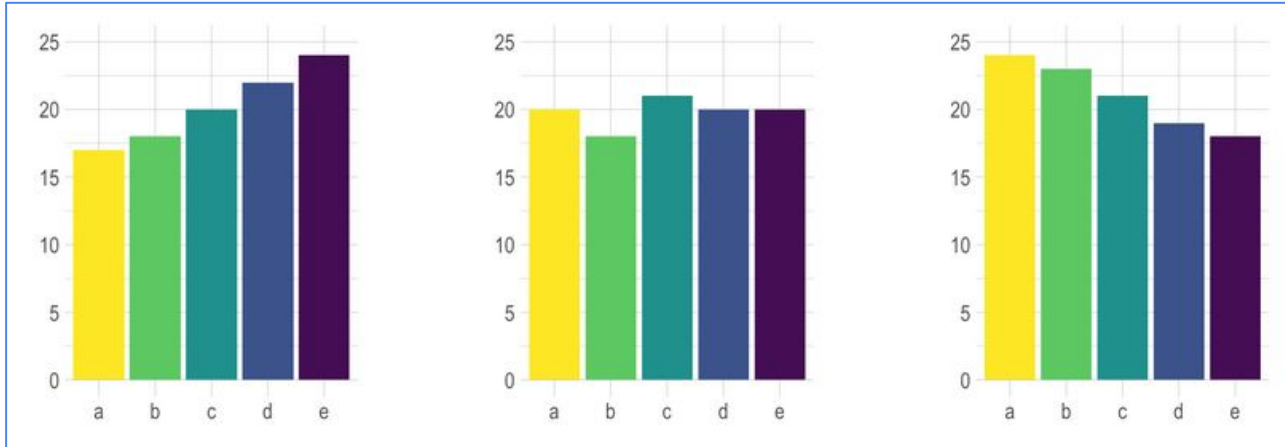
Pick the Right Chart



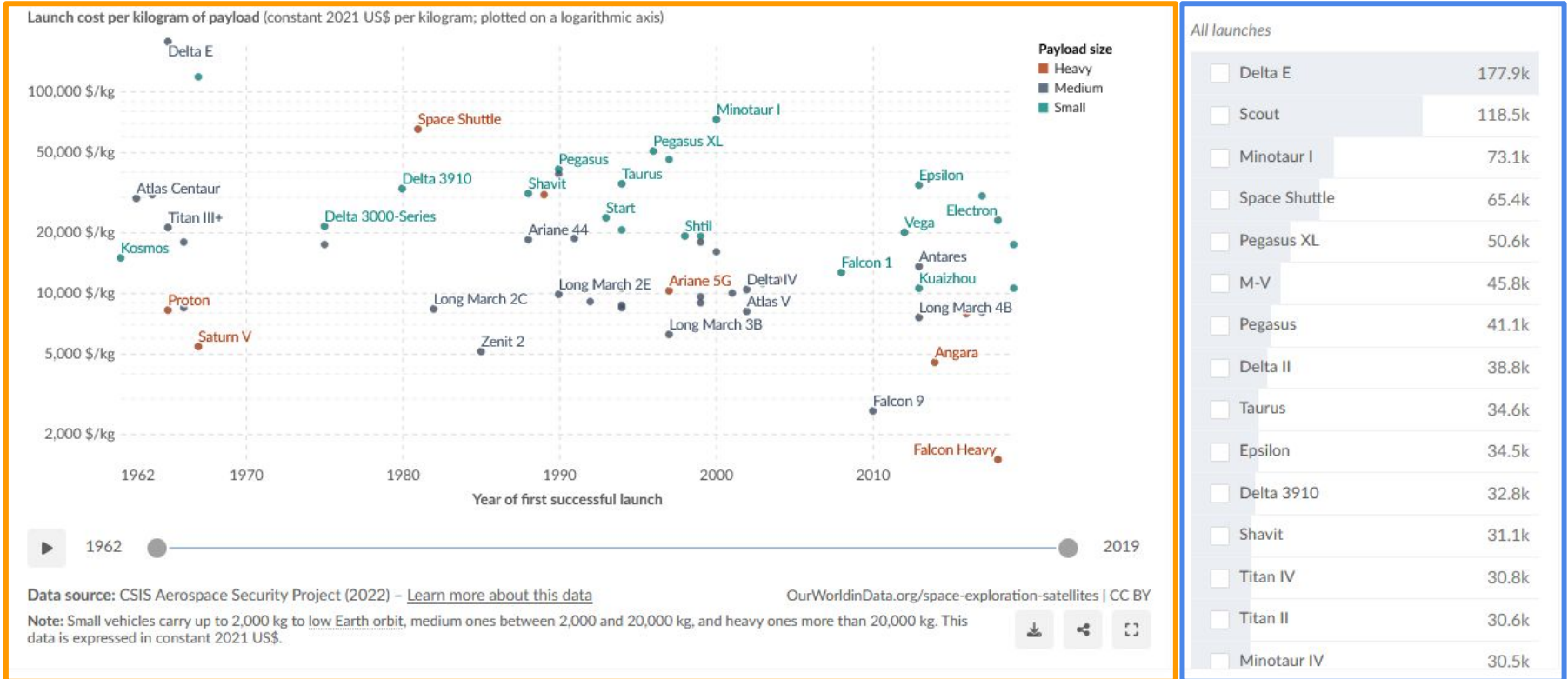
Pick the Right Chart: Pie Example



VS

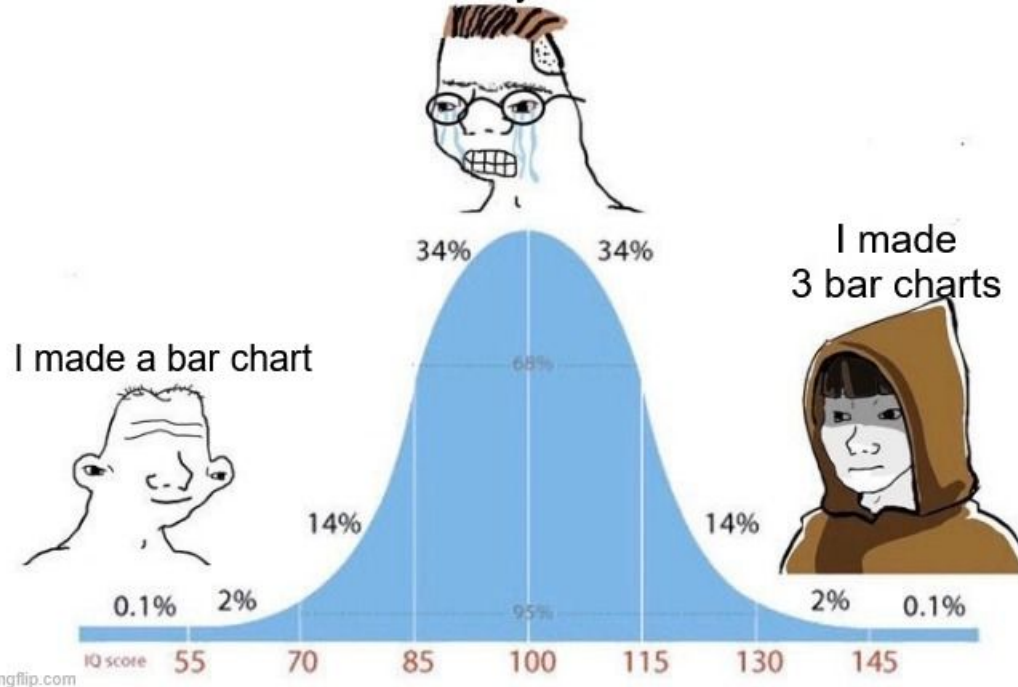


Keep it Honest



Keep it Simple

I designed a new chart
that can fit all my data at once



Draw a Viz

Imagine you have a dataset that shows historical Tuberculosis (TB) data across the world. It has the following columns:

- **Country** - the name of the country
- **Year** - the year the data represents
- **Population** - the population of the country in that year
- **Incidents Per 100k** - the number of TB cases per 100k people
- **Deaths Per 100k** - the number of TB deaths per 100k people
- **Fatality Ratio** - the ratio (fraction) of incidents that ended in death

Draw a visual that you think might tell an interesting story about this data.

Customizing R Plots

By the end of today, you should understand a few more layers of ggplot:

- Setting aesthetics
- Using facets
- Changing scales
- Changing coordinates
- Changing themes
- Adding annotations

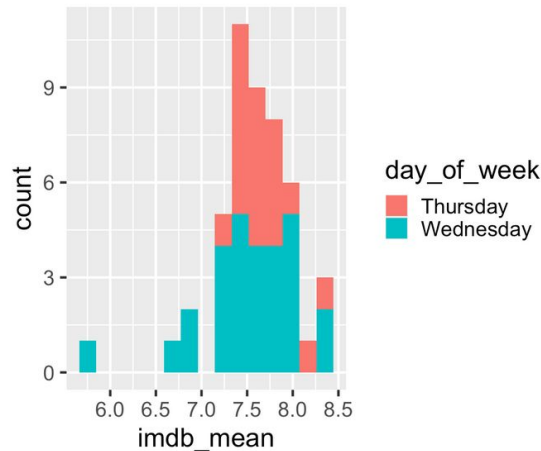
Starting Context

We'll be using images from a starting visual to demonstrate how each of these layers can change it

The intent of this visual is to show the difference in ratings for Survivor seasons that aired episodes on Wednesdays vs Thursdays.

Starting Code:

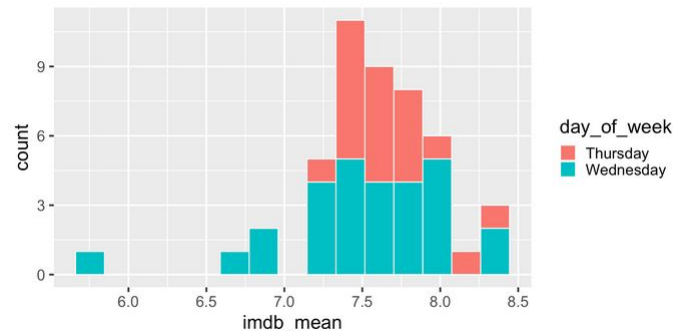
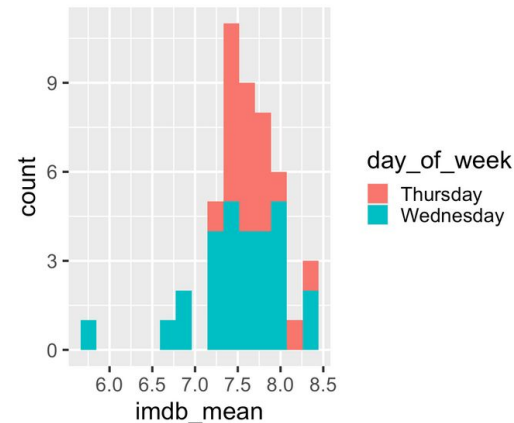
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15  
  )
```



Aesthetics

The general shape and contents of the plot

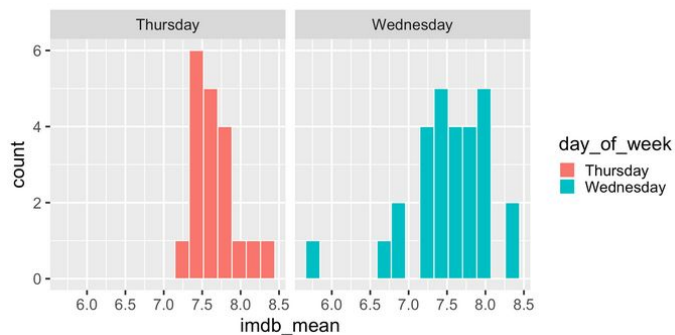
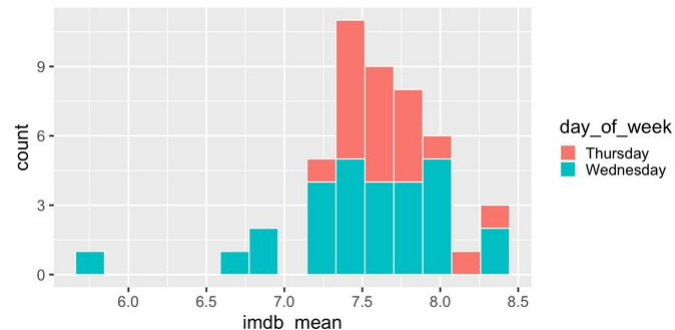
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  )
```



Facets

Multiple charts at once

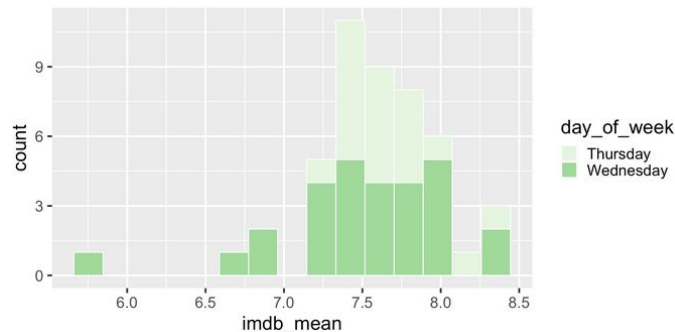
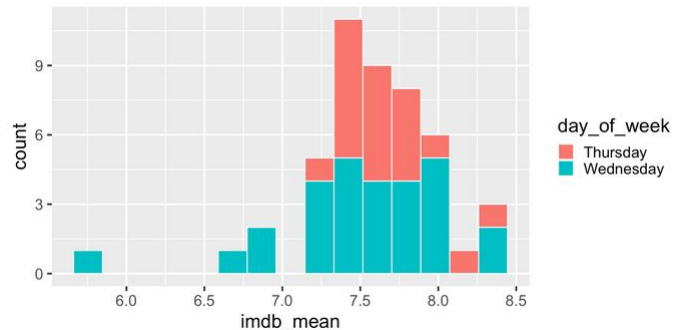
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  facet_wrap(vars(day_of_week))
```



Scales

Distinguishing one category from another using things like color and shape

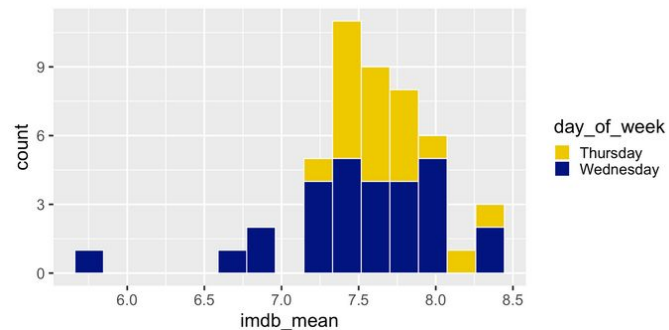
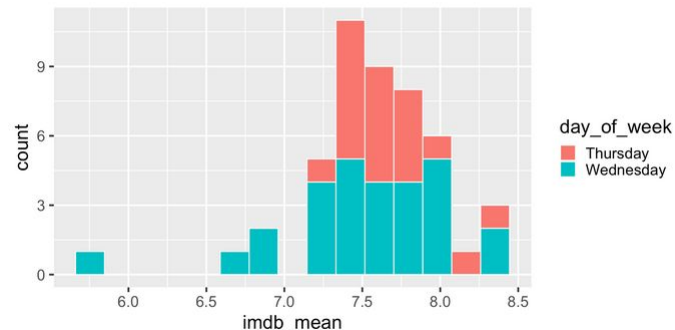
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_brewer(palette = "Greens")
```



Scales

2nd example using manual fill

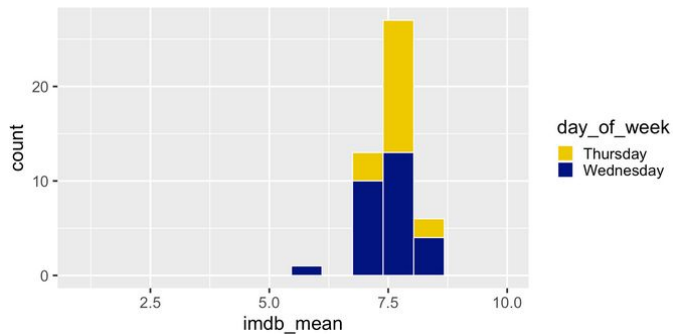
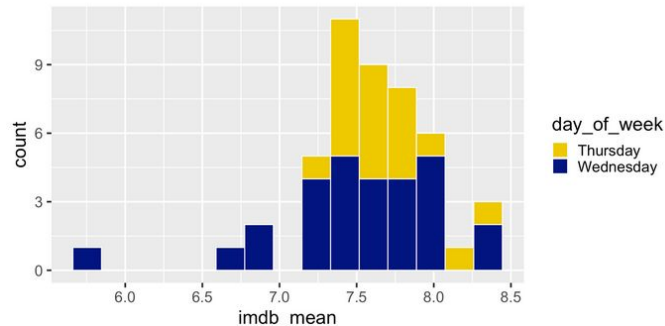
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_manual(values = c("gold2", "navyblue"))
```



Scales

Scale can also refer to the axis

```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_manual(values = c("gold2", "navyblue"))  
+  
  scale_x_continuous(limits = c(1,10))
```



Themes

Polish - the non-data parts of the viz, such as

- Background
- Tick marks
- Gridlines
- Font
- Legend Position & Appearance

Themes elements can be set individually or applied as a package

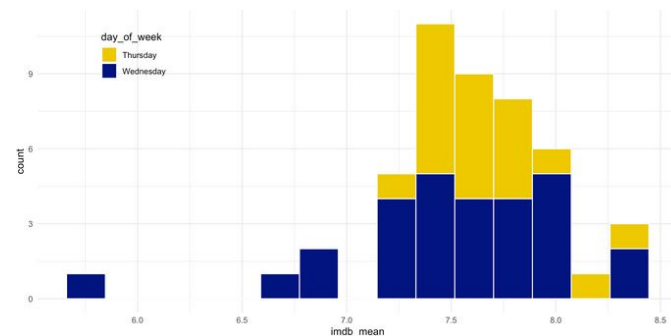
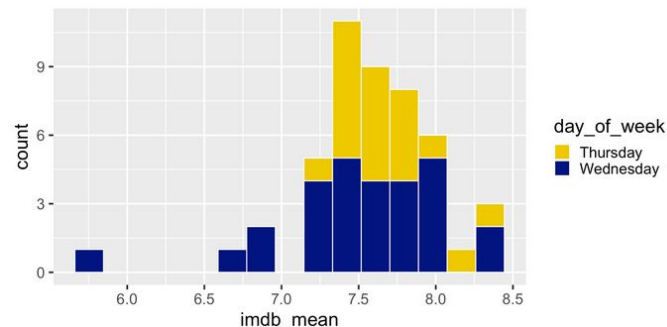
- Try applying `theme_minimal()` or `theme_clean()`

[Theme Cheat Sheet](#)

Themes

Apply preset and adjust legend

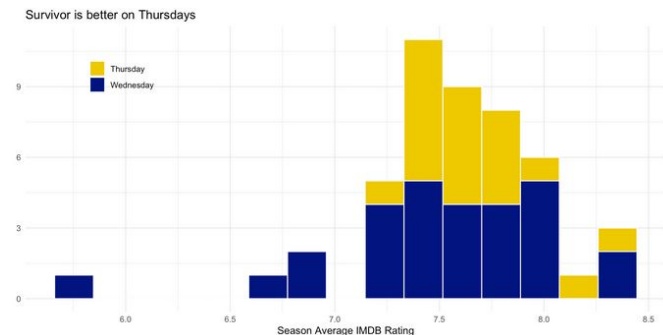
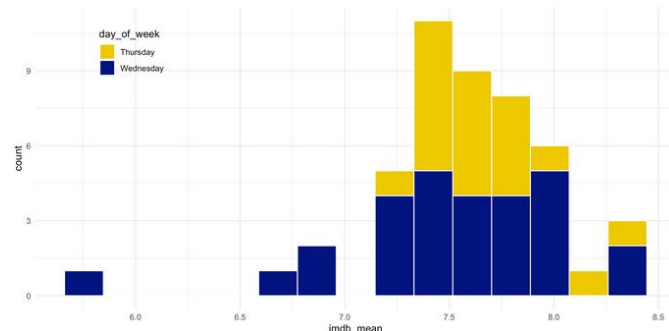
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_manual(values = c("gold2", "navyblue"))  
+  
  theme_minimal() +  
  theme(  
    legend.position = c(.15, .85),  
    legend.background = element_blank()  
  )
```



Themes

Improve labels

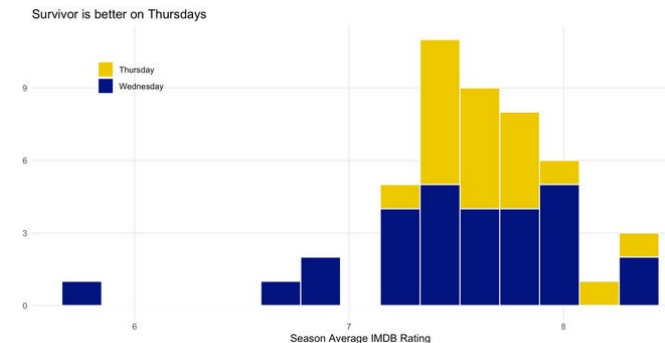
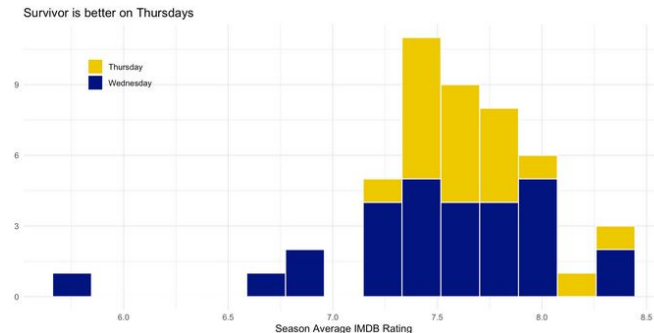
```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_manual(values = c("gold2", "navyblue"))  
+  
  theme_minimal() +  
  theme(  
    legend.position = c(.15, .85),  
    legend.background = element_blank()  
  ) +  
  labs(  
    x = "Season Average IMDB Rating",  
    y = "",  
    fill = "",  
    title = "Survivor is better on Thursdays"  
  )
```



Themes

Remove minor gridlines and 0.5 increments

```
ggplot(season_summary) +  
  geom_histogram(  
    aes(x = imdb_mean, fill = day_of_week),  
    bins = 15,  
    color = "white"  
  ) +  
  scale_fill_manual(values = c("gold2", "navyblue")) +  
  scale_x_continuous(breaks = c(6, 7, 8, 9)) +  
  theme_minimal() +  
  theme(  
    legend.position = c(.15, .85),  
    legend.background = element_blank(),  
    panel.grid.minor = element_blank()  
  ) +  
  labs(  
    x = "Season Average IMDB Rating",  
    y = "",  
    fill = "",  
    title = "Survivor is better on Thursdays"  
  )
```



And More...

Read the book!

I only covered a tiny fraction of what's possible

Code a Viz

Try to create the Tuberculosis (TB) viz you drew earlier in R

Column	Plain Name	Description
country	Country	The country associated with the measurements
g_whoregion	WHO Region	The World Health Organization region the country belongs to: AFR: Africa AMR: Americas EMR: Eastern Mediterranean EUR: Europe SEA: South-East Asia WPR: Western Pacific
year	Year	The year associated with the measurements
e_pop_num	Population	The population of the country in that year
e_inc_100k	Incidents per 100k	The number of TB cases per 100k people
e_mort_100k	Deaths per 100k	The number of TB deaths per 100k people
cfr	Fatality Ratio	The ratio (fraction) of incidents that ended in death