Data Visualization in the Tidyverse

The Great Tidy Plot Off

Alison Hill, PhD
Data Scientist & Professional Educator

apreshill
@apreshill
alison@rstudio.com

RStudio

Inspired by:

Flowing Data

tl;dr

- Tidy data is a place to start

Subgoals

- labeling!
- lose the defaults!

Packages first

I'll use all of the following:

```r
library(tidyverse)
library(bakeoff) # data + colors!
library(extrafont) # fonts!
```

Data second

```
library(bakeoff)
ratings ← ratings_seasons %>%
  mutate(series = as.factor(series))
```

Glimpse

```
Observations: 74
Variables: 10
$ series               <fct> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ...
$ episode              <int> 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 7, 8, ...
$ uk_airdate           <date> 2010-08-17, 2010-08-24, 2010-08-31, 2010 ...
$ viewers_7day         <dbl> 2.24, 3.00, 3.00, 2.60, 3.03, 2.75, 3.10, ...
$ viewers_28day        <dbl> 7, 3, 2, 4, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, ...
$ network_rank         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N ...
$ channels_rank        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N ...
$ bbc_iplayer_requests <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N ...
$ us_season            <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N ...
$ us_airdate           <date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,  ...
```

🎂

Recipe 1: Continuous Bar Chart

Recipe 1: Continuous Bar Chart

Recipe 1: Code for Bar Chart

```r
# some small wrangling
ratings_bonanza1 <- ratings %>%
  mutate(ep_id = row_number()) %>%
  select(ep_id, viewers_7day, series, episode)

# make the plot
ggplot(ratings_bonanza1, aes(x = ep_id, y = viewers_7day,
                             fill = series)) +
  geom_col(alpha = .9) +
  ggtitle("Series 8 was a Big Setback in Viewers",
          subtitle= "7-Day Viewers across All Series/Episodes") +
  theme(legend.position = "bottom",
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.x = element_blank()) +
  scale_fill_bakeoff() +
  scale_x_continuous(expand = c(0, 0)) +
  guides(fill = guide_legend(nrow = 1))
```

🎂

Recipe 1.2: Ribbons not Bars

Recipe 1.2: Ribbons not Bars

Recipe 1.2: Code for Ribbons

```r
ggplot(ratings_bonanza1, aes(x = ep_id, y = viewers_7day,
                             fill = series, color = series)) +
  geom_ribbon(aes(ymin = 0, ymax = viewers_7day), alpha = .75) +
  geom_line() +
  geom_text(data = filter(ratings_bonanza1,
                          series %in% c(1:2) & episode == 4),
            aes(y = 1.5, label = series),
            size = 3, color="white", family = "Lato") +
  geom_text(data = filter(ratings_bonanza1, series %in% c(3:8) & episode
            aes(y = 1.5, label = series),
            size = 3, color="white") +
  ggtitle("Series 8 was a Big Setback in Viewers",
          subtitle= "7-Day Viewers across All Series/Episodes") +
  theme(legend.position = "bottom",
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.x = element_blank()) +
  scale_fill_bakeoff() +
  scale_color_bakeoff() +
  scale_x_continuous(expand = c(0, 0)) +
  guides(fill = FALSE, color = FALSE)
```

🎂

What is going on with Series 8?

> "The eighth series of The Great British Bake Off began on 29 August 2017, with this being the first of The Great British Bake Off to be broadcast on Channel 4, after the production company Love Productions moved the show. It is the first series for new hosts Noel Fielding and Sandi Toksvig, and new judge Prue Leith." -- Wikipedia

Read:

Read:

No Mary Berry, no Mel, no Sue

🎂

Recipe 2: Lollipop Plot

Recipe 2: Lollipop Plot

Recipe 2: Code for Lollipop Plot

```r
ratings_bonanza2 <- ratings %>%
  group_by(series) %>%
  mutate(series_avg = mean(viewers_7day, na.rm = TRUE),
         diff_avg = viewers_7day - series_avg)%>%
  filter(max(episode) == 10) %>%
  mutate(episode = as.factor(episode)) %>%
  select(episode, viewers_7day, series, diff_avg, series_avg)

ggplot(ratings_bonanza2, aes(x = episode,
                             y = viewers_7day,
                             color = diff_avg)) +
  geom_hline(aes(yintercept = series_avg), alpha = .5) +
  geom_point() +
  geom_segment(aes(xend = episode, yend = series_avg)) +
  facet_wrap(~series) +
  scale_color_viridis_c(option="plasma", begin = 0,
                        end = .8, guide = FALSE) +
  ggtitle("Great British Bake Off Finales Get the Most Viewers",
          subtitle = "Way Higher than Series Average (for Series with 10
```

🎂

Recipe 3: Grouped Line Plot by Series

Recipe 3: Grouped Line Plot by Series

Recipe 3: Code for Series Grouped Line Plot

```r
ggplot(ratings, aes(x = as.factor(episode), y = viewers_7day,
                    color = fct_reorder2(series, episode, viewers_7day),
                    group = series)) +
  geom_line() +
  scale_color_bakeoff() +
  labs(color = "Series", x = "Episode")
```

🎂

Recipe 3.1: Redo Recipe 3

Facetted Series Grouped Line Plot

Recipe 3.1: Facetted Line Plot

Recipe 3.1: Code for Facetted Line Plot

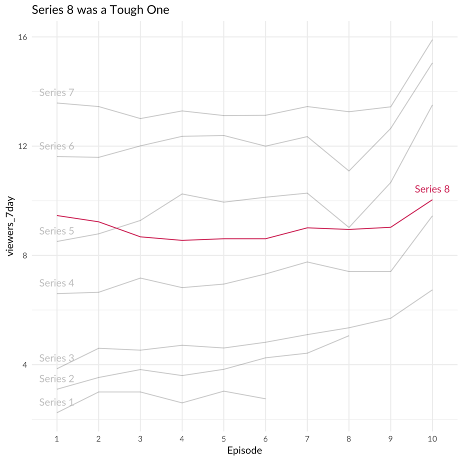
```r
ggplot(ratings, aes(x = as.factor(episode), y = viewers_7day,
                    color = fct_reorder2(series, episode, viewers_7day),
                    group = series)) +
  geom_line(lwd = 2) +
  scale_color_bakeoff() +
  labs(color = "Series", x = "Episode") +
  facet_wrap(~series) +
  guides(color = FALSE)
```

🎂

Recipe 3.2: Redo Recipe 3

Pop-Out Series Grouped Line Plot

Recipe 3.2: Redo Recipe 3



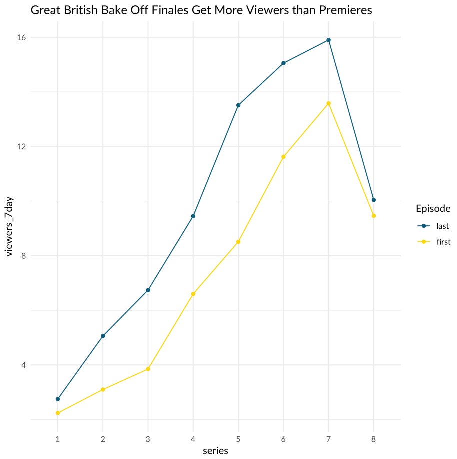
Recipe 3.2: Code for Redo Recipe 3

```
ggplot(ratings, aes(x = as.factor(episode), y = viewers_7day,
                    group = series)) +
  geom_line(data = filter(ratings, !series == 8), alpha = .25) +
  geom_line(data = filter(ratings, series == 8), color = "#CF2154") +
  labs(color = "Series", x = "Episode") +
  ggtitle("Series 8 was a Tough One") +
  geom_text(data = filter(ratings, episode == 1 & series %in% c(1:7)), c
            aes(label = paste0("Series ", series)), vjust = -1) +
  geom_text(data = filter(ratings, episode == 10 & series == 8), color =
            aes(label = paste0("Series ", series)), vjust = -1)
```

🎂

Recipe 4: Grouped Line Plot by Episode

Recipe 4: Grouped Line Plot by Episode



Recipe 4: Code for Grouped Episode Line Plot

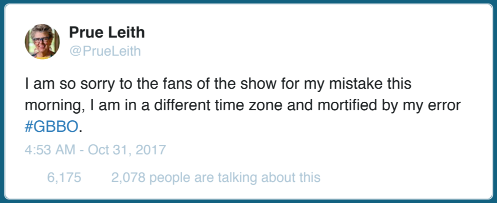
```
# some wrangling here
ratings_bonanza4 ← ratings %>%
  select(series, episode, viewers_7day) %>%
  group_by(series) %>%
  filter(episode == 1 | episode == max(episode)) %>%
  mutate(episode = recode(episode, `1` = "first", .default = "last")) %>
  ungroup()

# code for plot
ggplot(ratings_bonanza4, aes(x = series, y = viewers_7day,
                             color = fct_reorder2(episode, series, viewe
                             group = episode)) +
  geom_point() +
  geom_line() +
  scale_color_bakeoff() +
  ggtitle("Great British Bake Off Finales Get More Viewers than Premiere
  labs(color = "Episode")
```

🎂

What is going on with the Series 8 finale?

A tweet heard 'round the world



Prue Leith
@PrueLeith

I am so sorry to the fans of the show for my mistake this morning, I am in a different time zone and mortified by my error #GBBO.

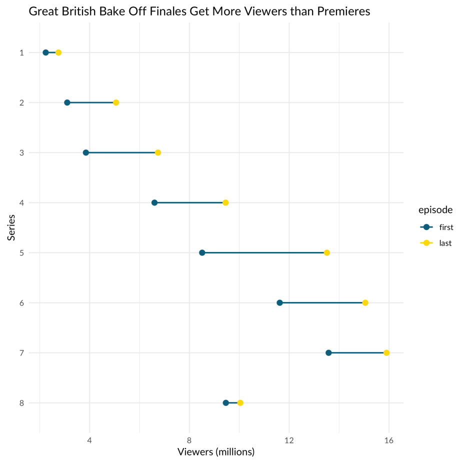
4:53 AM - Oct 31, 2017

6,175 2,078 people are talking about this

🎂

Recipe 5: Dumbbell Plot

Recipe 5: Dumbbell Plot



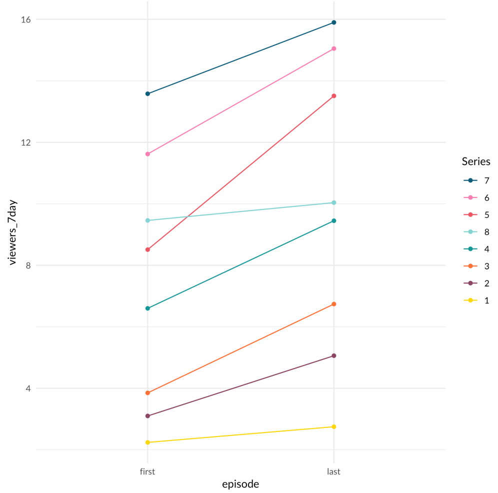
Recipe 5: Code for Dumbbell Plot

```
ggplot(ratings_bonanza4, aes(x = viewers_7day, y = fct_rev(series),
                             color = episode, group = series)) +
  geom_line(size = .75) +
  geom_point(size = 2.5) +
  scale_color_bakeoff() +
  labs(y = "Series", x = "Viewers (millions)") +
  ggtitle("Great British Bake Off Finales Get More Viewers than Premiere
```

🎂

Recipe 6: Slope Graph

Recipe 6: Slope Graph



Recipe 6: Code for Slope Graph

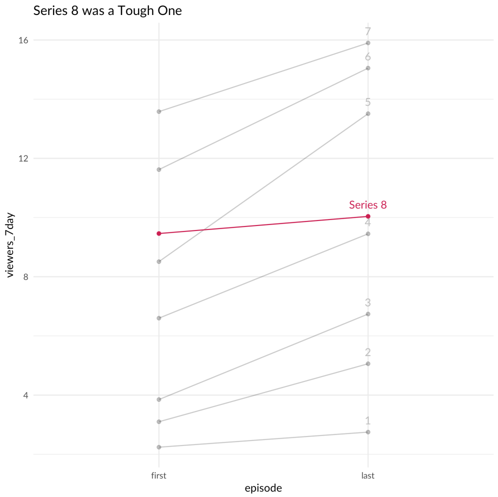
```
ggplot(ratings_bonanza4, aes(x = episode, y = viewers_7day,
                             color = fct_reorder2(series, episode, viewe
                             group = series)) +
  geom_point() +
  geom_line() +
  scale_color_bakeoff() +
  labs(color = "Series")
```

🎂

Recipe 6.1: Redo Recipe 6

Pop-Out Slope Graph

Recipe 6.1: Redo Recipe 6



Recipe 6.1: Redo Recipe 6

```r
ggplot(ratings_bonanza4, aes(x = episode, y = viewers_7day,
                             group = series)) +
  geom_point(data = filter(ratings_bonanza4, !series == 8), alpha = .25)
  geom_point(data = filter(ratings_bonanza4, series == 8),
             color = "#CF2154") +
  geom_line(data = filter(ratings_bonanza4, !series == 8), alpha = .25)
  geom_line(data = filter(ratings_bonanza4, series == 8),
            color = "#CF2154") +
  ggtitle("Series 8 was a Tough One") +
  geom_text(data = filter(ratings_bonanza4, episode == "last" & series %
            aes(label = series), vjust = -1, hjust = .5) +
  geom_text(data = filter(ratings_bonanza4, episode == "last" & series =
            aes(label = paste0("Series ", series)), vjust = -1)
```

🎂

Recipe 7: Bar Chart

Recipe 7: Bar Chart

Recipe 7: Code for Bar Chart

```r
# some more serious wrangling here
ratings_bonanza7 <- ratings %>%
  select(series, episode, viewers_7day) %>%
  group_by(series) %>%
  filter(episode == 1 | episode == max(episode)) %>%
  mutate(episode = recode(episode, `1` = "first", .default = "last")) %>
  spread(episode, viewers_7day) %>%
  mutate(finale_bump = last - first)

# plot
ggplot(ratings_bonanza7, aes(x = fct_rev(series),
                             y = finale_bump)) +
  geom_col(fill = bakeoff_cols("berry"), alpha = .8) +
  coord_flip() +
  labs(x = "Series", y = "Difference in Viewers for Finale from Premiere
  ggtitle("Finale 'Bumps' were Smallest for Series 1 and 8",
          subtitle= "Finale 7-day Viewers Relative to Premiere")
```

🎂

Recipe 8: % Change Bar Chart

Recipe 8: % Change Bar Chart

Recipe 8: Code for % Bar

```
# wrangling to calculate percent change
ratings_bonanza8 <- ratings %>%
  select(series, episode, viewers_7day) %>%
  group_by(series) %>%
  filter(episode == 1 | episode == max(episode)) %>%
  ungroup() %>%
  mutate(episode = recode(episode, `1` = "first", .default = "last")) %>
  spread(episode, viewers_7day) %>%
  mutate(pct_change = (last - first) / first)

# plot
ggplot(ratings_bonanza8, aes(x = fct_rev(series),
                             y = pct_change)) +
  geom_col(fill = bakeoff_cols("tangerine"), alpha = .5) +
  labs(x = "Series", y = "% Increase in Viewers from First to Last Episo
  ggtitle("Series 8 had a 6% Increase in Viewers from Premiere to Finale
          subtitle= "The Lowest Across All Series (Line is the Median)")
  geom_hline(aes(yintercept = median(pct_change, na.rm = TRUE)),
             color = bakeoff_cols("orange")) +
  scale_y_continuous(labels = scales::percent) +
  coord_flip()
```

🎂

Recipe 9: Bars Diverging from Median

Recipe 9: Bars Diverging from Median

Recipe 9: Bars from Median

```
# some more serious wrangling here
ratings_bonanza9 <- ratings %>%
  select(series, episode, viewers_7day) %>%
  group_by(series) %>%
  filter(episode == 1 | episode == max(episode)) %>%
  ungroup() %>%
  mutate(episode = recode(episode, `1` = "first", .default = "last")) %>
  spread(episode, viewers_7day) %>%
  mutate(pct_change = (last - first) / first,
         pct_change_diff = pct_change - median(pct_change),
         change_sign = if_else(pct_change_diff > 0, 1, 0))

# plot
ggplot(ratings_bonanza9, aes(x = fct_rev(series),
                             y = pct_change_diff,
                             fill = as.factor(change_sign))) +
  geom_col(alpha = .5) +
  labs(x = "Series",
       y = "% Change in Viewers from First to Last Episode, Relative to
  scale_fill_bakeoff(guide = FALSE) +
  ggtitle("Series 8 had the Most Disappointing Finale") +
  scale_y_continuous(labels = scales::percent) +
  coord_flip()
```

🎂

Recipe 10: Lollipop Plot, % Change

Recipe 10: Lollipop Plot, % Change

Recipe 10: Code for % Lollipop Plot

```
# plot
ggplot(ratings_bonanza9, aes(x = fct_rev(series),
                             y = pct_change)) +
  geom_point(color = bakeoff_cols("riptide"), size = 2) +
  geom_segment(aes(xend = fct_rev(series), yend = 0), color = bakeoff_co
  geom_text(aes(label = scales::percent(pct_change)), hjust = -.25) +
  labs(x = "Series", y = "% Change in Viewers from First to Last Episode
  ggtitle("Percent Increase in Viewers was the Smallest for Series 8",
          subtitle= "Finale 7-day Viewers Relative to Premiere") +
  scale_y_continuous(labels = scales::percent, limits = c(0, .85)) +
  coord_flip()
```

🎂

Recipe 11: Scatterplot

Recipe 11: Scatterplot

Recipe 11: Code for Scatterplot

```r
ggplot(ratings_bonanza7, aes(x = first, y = last)) +
  geom_point() +
  geom_smooth(se = FALSE, color = '#EBBFDD') +
  geom_abline(slope = 1, intercept = 0, color = "gray", alpha = .5) +
  geom_text(aes(label = series), hjust = -1) +
  labs(x = "Premiere Episode 7-day Viewers (millions)",
       y = "Finale Episode 7-day Viewers (millions)") +
  coord_equal(ratio=1)
```

🎂

Recipe 11.1: Pop-Out Scatterplot

Recipe 11.1: Pop-Out Scatterplot

Recipe 11.1: Code for Pop-Out Scatterplot

```r
ggplot(ratings_bonanza7, aes(x = first, y = last)) +
  geom_abline(slope = 1, intercept = 0, color = "gray", alpha = .5) +
  geom_smooth(se = FALSE, color = "#11B2E8") +
  geom_point(data = filter(ratings_bonanza7, series %in% c(1:7))) +
  geom_point(data = filter(ratings_bonanza7, series == 8), colour = "#CF
  geom_text(data = filter(ratings_bonanza7, series %in% c(1:7)),
            aes(label = series), hjust = -1) +
  geom_text(data = filter(ratings_bonanza7, series == 8),
            aes(label = series), hjust = -1, colour = "#CF2154") +
  labs(x = "Premiere Episode 7-day Viewers (millions)",
       y = "Finale Episode 7-day Viewers (millions)")
```

#GBBO
THE GREAT BRITISH
BAKE OFF