Intermediate Data Analysis in R

Derek Sollberger

2022-05-31

Contents


```
library("corrplot")
library("forcats")
library("ggsignif")
library("janitor")
library("infer")
library("skimr")
library("tidyverse")
```

Today's Data Set

"This survey is the result of a partnership between Axios and Harris Poll to gauge the reputation of the most visible brands in America, based on 20 years of Harris Poll research. From Trader Joe's to Disney, here's how this year's class stacks up.

"Methodology: The Axios Harris Poll 100 is based on a survey of 33,096 Americans in a nationally representative sample conducted March 11-April 3, 2022. The two-step process starts fresh each year by surveying the public's top-of-mind awareness of companies that either excel or falter in society.

"These 100 "most visible companies" are then ranked by a second group of Americans across the seven key dimensions of reputation to arrive at the ranking. If a company is not on the list, it did not reach a critical level of visibility to be measured."

Since we have CSV files (comma-separated values), we can use the `read_csv` function in the `readr` package to load the file into our current programming session.

```
reputation <- readr::read_csv("reputation.csv",
                              name_repair = janitor::make_clean_names)
reputation_wide <- readr::read_csv("reputation_wide.csv",
                                   name_repair = janitor::make_clean_names)
```

Reshaping Data Frames

You have noticed that for these examples, I have employed two slightly different data frames. Each had situations where it was then easier to employ R code. Those data frames were made with the `pivot_longer()` and `pivot_wider()`

```
# from reputation to reputation_wide
reputation_wide <- reputation |>
  select(-rank) |> #removed the 'rank' column
  pivot_wider(names_from = name,
              values_from = score) |>
  janitor::clean_names()
```

```
# from reputation_wide to reputation
reputation <- reputation_wide |>
  pivot_longer(cols = trust:culture,
               names_to = "name",
               values_to = "score")
```

Confidence Intervals

If our data is a *sample*, can we generalize to a larger *population*? The field of statistics employs **confidence intervals**.

For example, what is the true population mean for the ethics score for the food delivery industry? We can use the `infer` package to handle the simulations and computation.

```
reputation_wide |>
  filter(industry == "Food Delivery") |>
  specify(response = ethics) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "mean") |>
  get_ci(level = 0.95, type = "percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     70.5     74.6
```

We are 95 percent confident that the true population mean for the ethics score for the food delivery industry is in between 70.5 and 74.6 (on a scale from 0 to 100).

Hypothesis Testing

Example 1

- null hypothesis: the healthcare and media industries have the same reputation score
- alternative hypothesis: the healthcare and media industries have different reputation scores

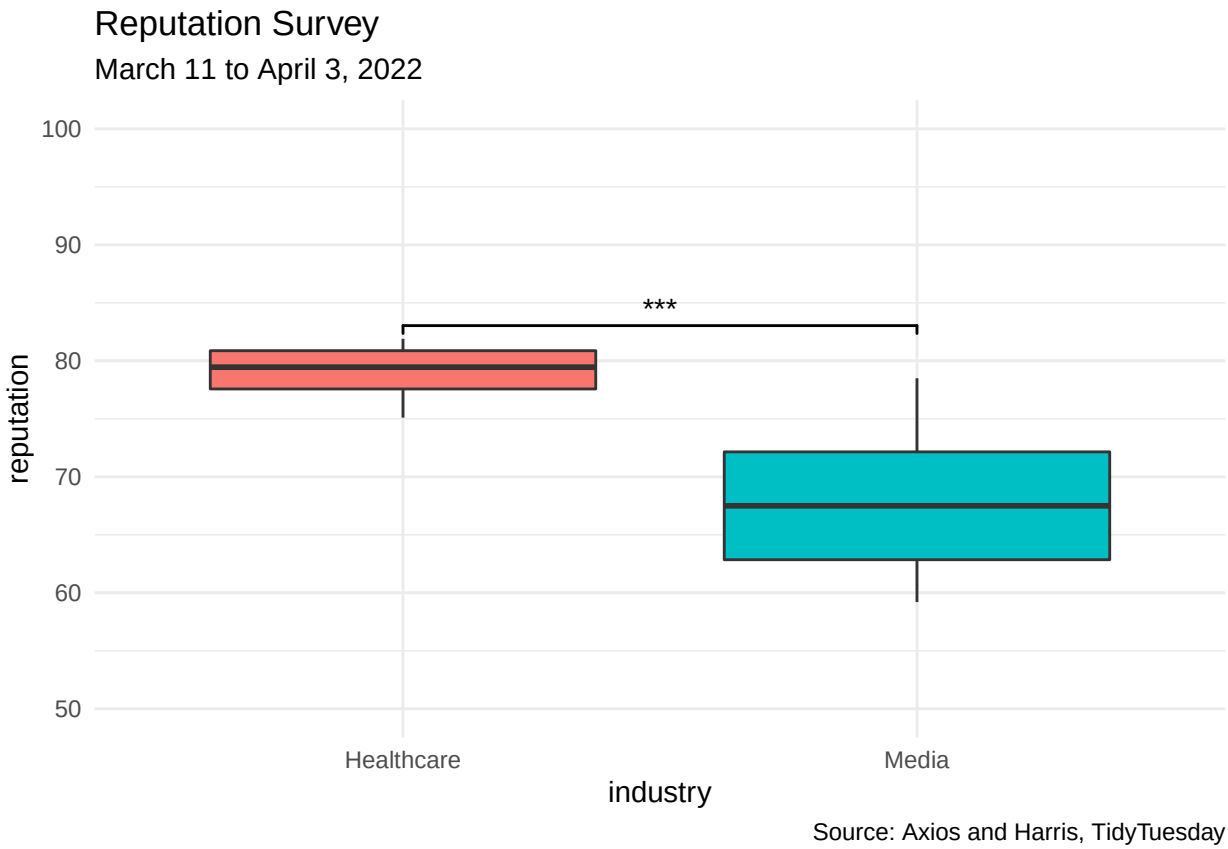
$$H_o : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

We can use the `ggsignif` package to indicate the significance level on our boxplot!

- p-value < 0.05: *
- p-value < 0.01: **
- p-value < 0.001: ***

```
reputation |>
  filter(industry %in% c("Healthcare", "Media")) |>
  ggplot(aes(x = industry, y = score,
             fill = industry)) +
  geom_boxplot() +
  geom_signif(
    comparisons = list(c("Healthcare", "Media")),
    map_signif_level = TRUE
  ) +
  labs(title = "Reputation Survey",
       subtitle = "March 11 to April 3, 2022",
       caption = "Source: Axios and Harris, TidyTuesday",
       x = "industry",
       y = "reputation") +
  ylim(50, 100) +
  theme_minimal() + #removes gray background
  theme(legend.position = "none")
```

```
## Warning in wilcox.test.default(c(79.9, 79, 81.9, 80.1, 76.8, 80.5, 81, 76.1, :
## cannot compute exact p-value with ties
```



Since the p-value < 0.05, we reject the claim of equal means between the healthcare and media industries.

Example 2

- null hypothesis: the pharmacy and retail industries have the same reputation score
- alternative hypothesis: the pharmacy and retail industries have different reputation scores

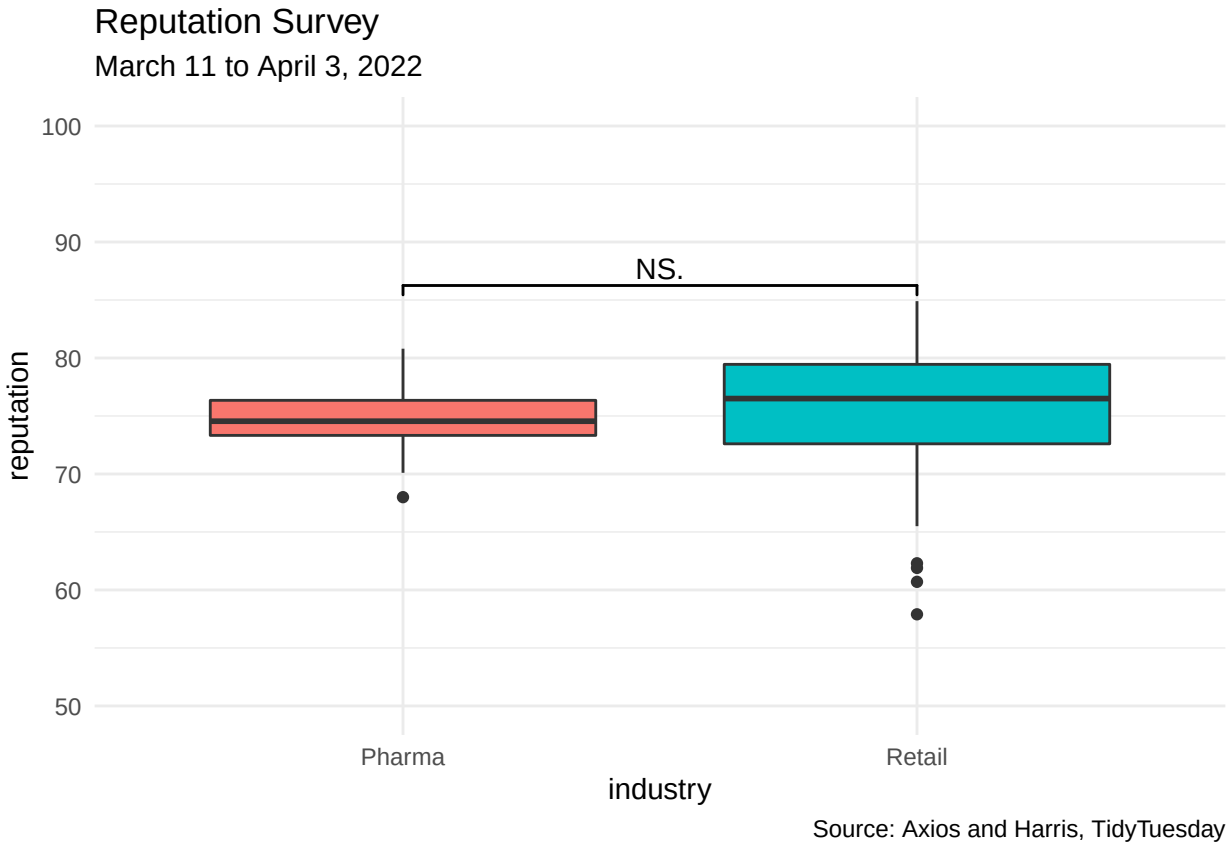
$$H_o : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

We can use the `ggsignif` package to indicate the significance level on our boxplot!

- p-value < 0.05: *
- p-value < 0.01: **
- p-value < 0.001: ***

```
reputation |>
  filter(industry %in% c("Pharma", "Retail")) |>
  ggplot(aes(x = industry, y = score,
             fill = industry)) +
  geom_boxplot() +
  geom_signif(
    comparisons = list(c("Pharma", "Retail")),
```

```
  map_signif_level = TRUE
) +
labs(title = "Reputation Survey",
     subtitle = "March 11 to April 3, 2022",
     caption = "Source: Axios and Harris, TidyTuesday",
     x = "industry",
     y = "reputation") +
ylim(50, 100) +
theme_minimal() + #removes gray background
theme(legend.position = "none")
```



Since the p-value > 0.05, we fail to reject the claim of equal means between the pharmacy and retail industries.

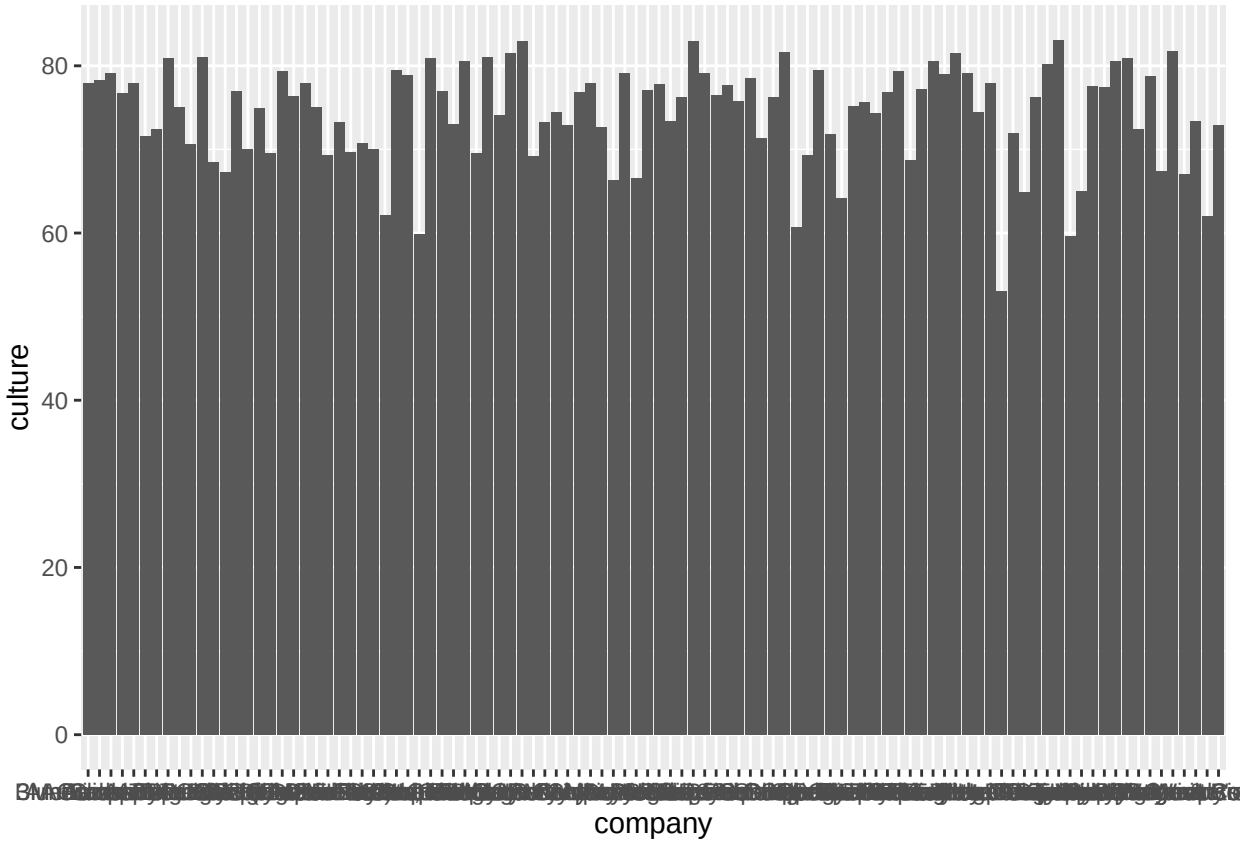
Ranking

In this longer example, I am going to show you an example of improving a graph iteratively for clarity.

- Query: What are the top 10 companies for workplace culture?

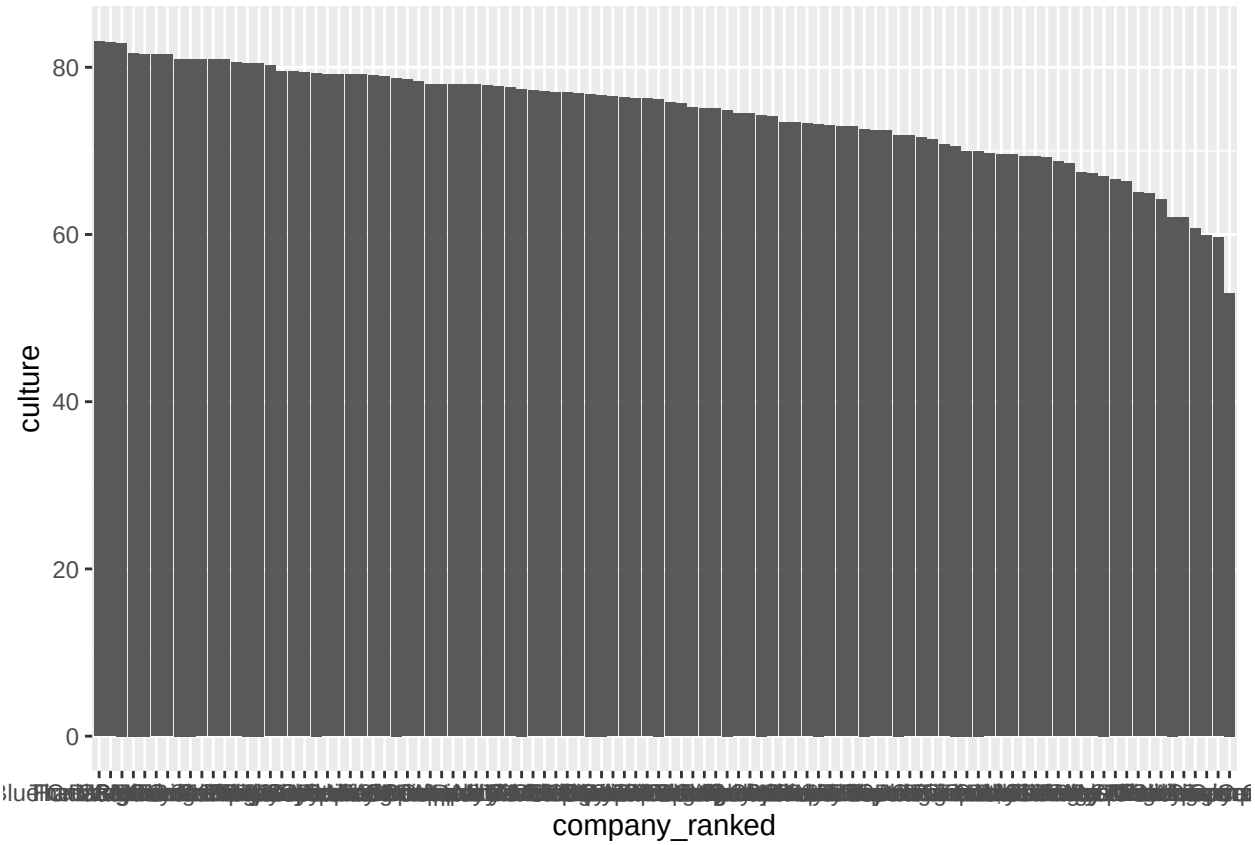
Let us start with the framework for a bar chart.

```
reputation_wide |>
  ggplot(aes(x = company, y = culture)) +
  geom_bar(stat = "identity")
```



Next, we can use the `forcats` package to help us reorder the data from highest to lowest.

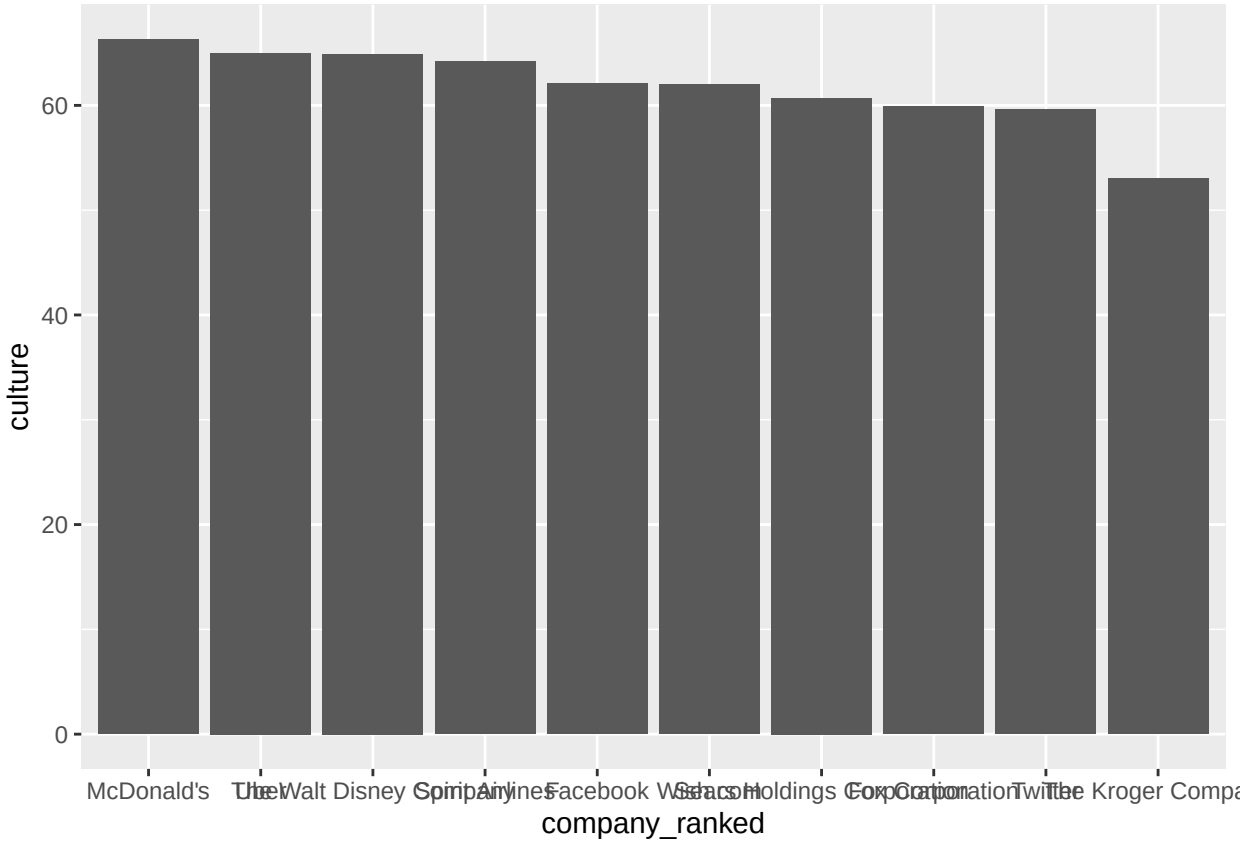
```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture, .desc = TRUE)) |>
  ggplot(aes(x = company_ranked, y = culture)) +
  geom_bar(stat = "identity")
```



The graph is still crowded, so let us focus on the top 10 companies.

```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture, .desc = TRUE)) |>
  top_n(10) |>
  ggplot(aes(x = company_ranked, y = culture)) +
  geom_bar(stat = "identity")
```

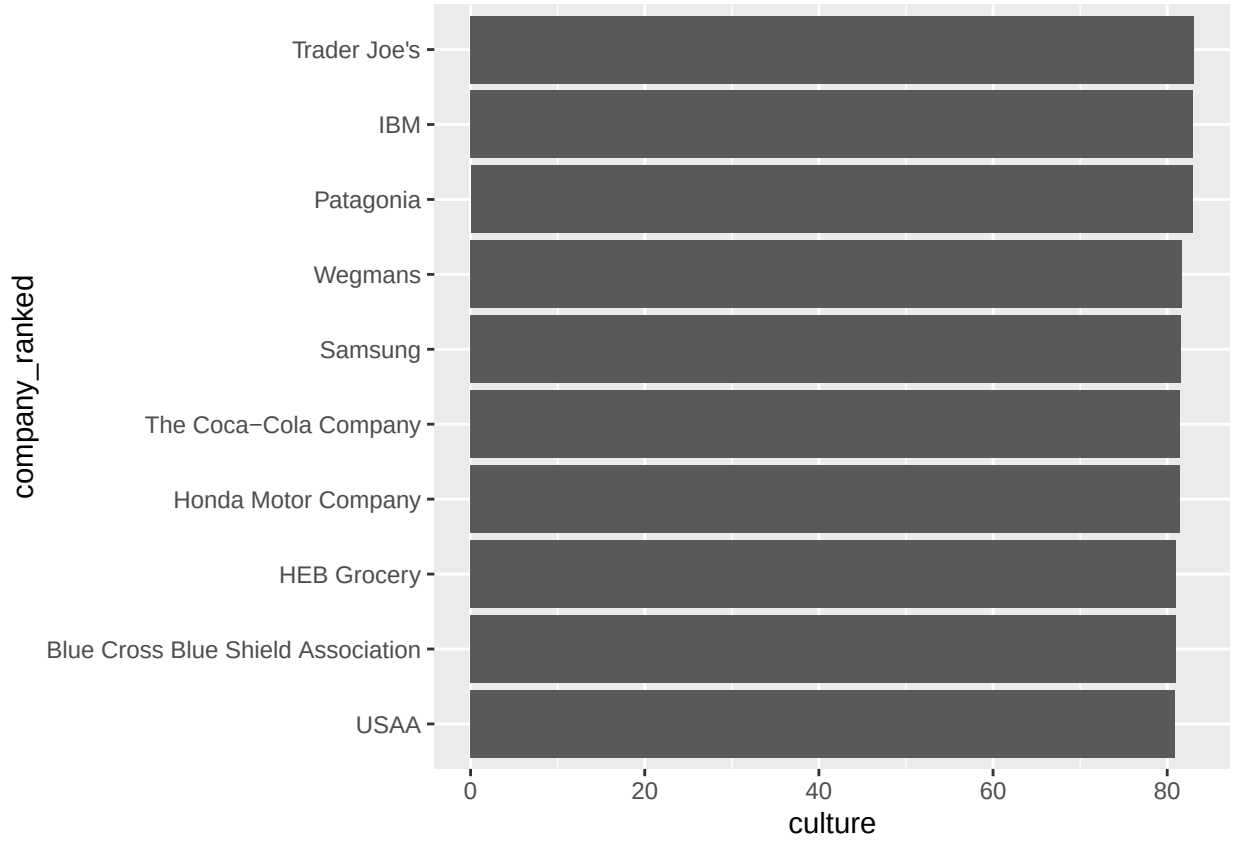
```
## Selecting by company_ranked
```



The labels for the company names are overlapping and still hard to read. One way to ease the congestion is to switch what are on the x- and y-axes.

```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture)) |>
  top_n(10) |>
  ggplot(aes(x = culture, y = company_ranked)) +
  geom_bar(stat = "identity")
```

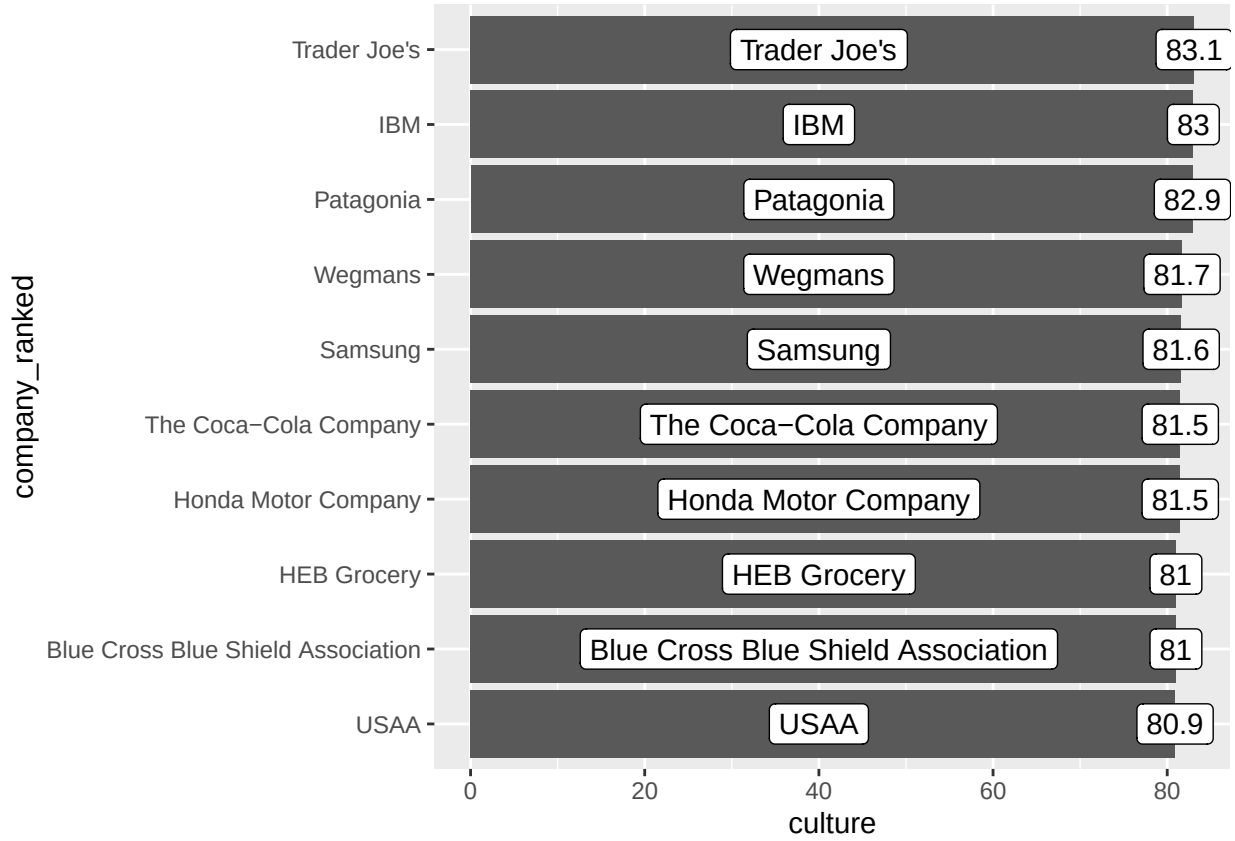
```
## Selecting by company_ranked
```



While the viewer can find a lot of information on this graph, we can ease the view by putting the company names on the bars themselves! Furthermore, we can specify the culture score values too.

```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture)) |>
  top_n(10) |>
  ggplot(aes(x = culture, y = company_ranked)) +
  geom_bar(stat = "identity") +
  geom_label(aes(x = 40, y = company_ranked, label = company_ranked)) +
  geom_label(aes(x = culture, y = company_ranked, label = culture))
```

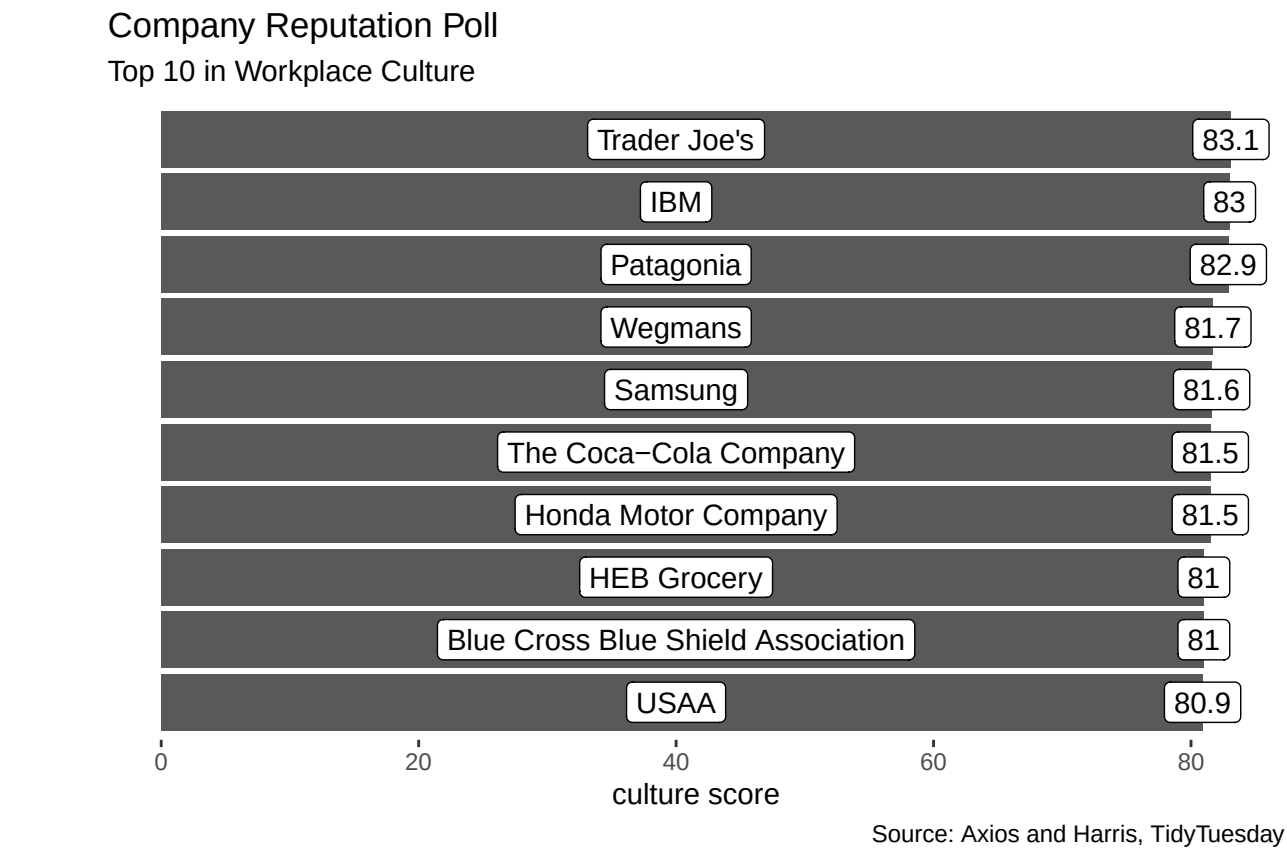
```
## Selecting by company_ranked
```



Now, the y-axis labels are redundant, so we can remove them and some of the other elements of the `theme`

```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture)) |>
  top_n(10) |>
  ggplot(aes(x = culture, y = company_ranked)) +
  geom_bar(stat = "identity") +
  geom_label(aes(x = 40, y = company_ranked, label = company_ranked)) +
  geom_label(aes(x = culture, y = company_ranked, label = culture)) +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        panel.background = element_blank()) +
  labs(title = "Company Reputation Poll",
       subtitle = "Top 10 in Workplace Culture",
       caption = "Source: Axios and Harris, TidyTuesday",
       x = "culture score",
       y = "")
```

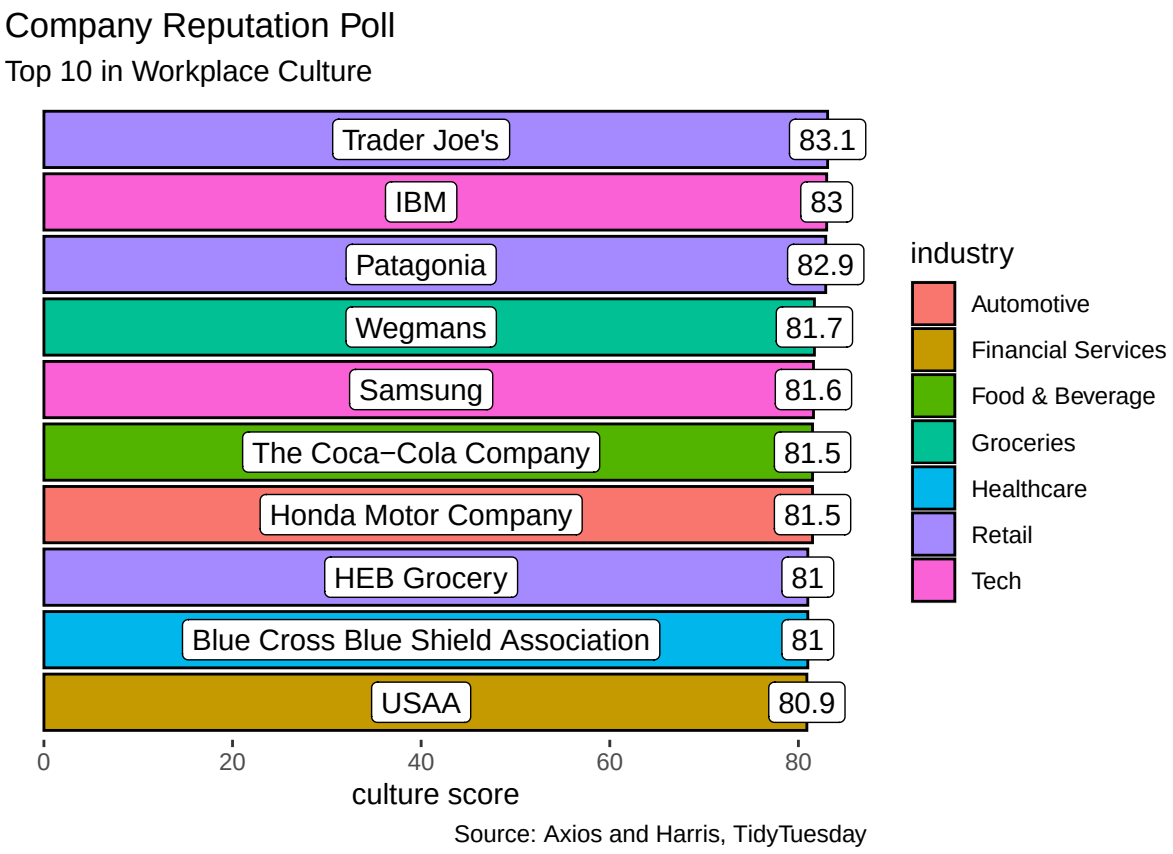
```
## Selecting by company_ranked
```



We can use the `fill` aesthetic to highlight another category. Let's see how the industries are represented in this top 10 list.

```
reputation_wide |>
  mutate(company_ranked = fct_reorder(company, culture)) |>
  top_n(10) |>
  ggplot(aes(x = culture, y = company_ranked)) +
  geom_bar(aes(fill = industry),
           color = "black",
           stat = "identity") +
  geom_label(aes(x = 40, y = company_ranked, label = company_ranked)) +
  geom_label(aes(x = culture, y = company_ranked, label = culture)) +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        panel.background = element_blank()) +
  labs(title = "Company Reputation Poll",
       subtitle = "Top 10 in Workplace Culture",
       caption = "Source: Axios and Harris, TidyTuesday",
       x = "culture score",
       y = "")
```

```
## Selecting by company_ranked
```


Company Reputation Poll
Top 10 in Workplace Culture
Trader Joe's
83.1
IBM
83
Patagonia
82.9
Wegmans
81.7
Samsung
81.6
The Coca−Cola Company
81.5
Honda Motor Company
81.5
HEB Grocery
81
Blue Cross Blue Shield Association
81
USAA
80.9
0
20
40
60
80
culture score
Source: Axios and Harris, TidyTuesday
industry
Automotive
Financial Services
Food & Beverage
Groceries
Healthcare
Retail
Tech