Intermediate Data Analysis in R

Derek Sollberger

2022-05-31

Table of 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 × 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

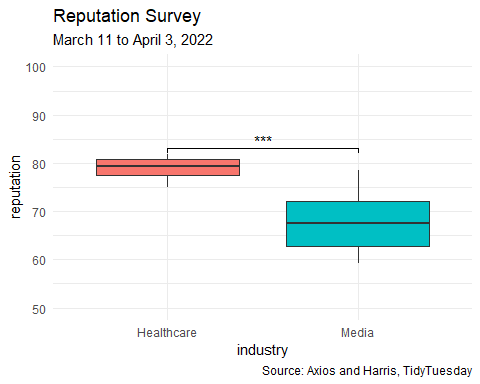
* null hypothesis: the healthcare and media industries have the same reputation score
* alternative hypothesis: the healthcare and media industries have different reputation scores

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.

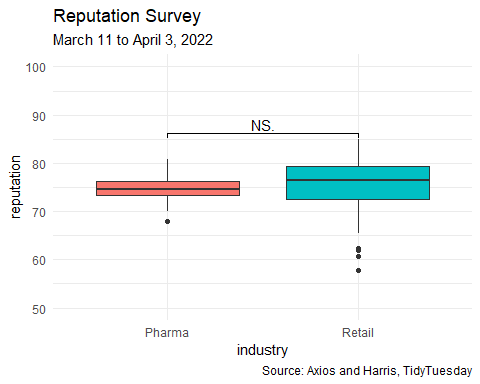
## Example 2

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

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.

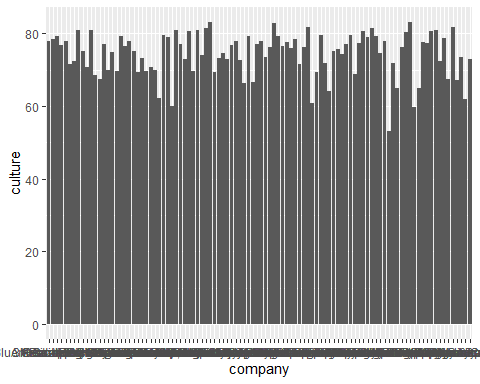
# Ranking

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?

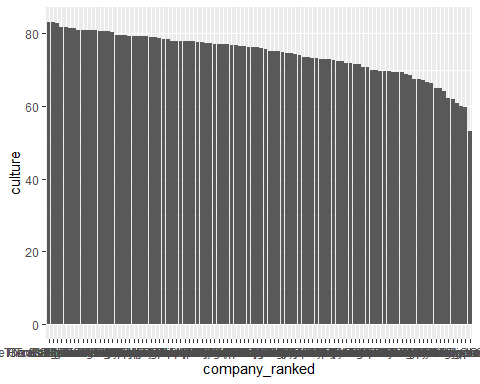
Let us start with the framework for a bar chart.

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.

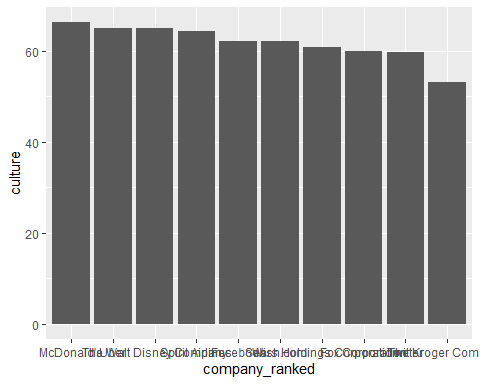
reputation\_wide |>  
 mutate(company\_ranked = fct\_reorder(company, culture, .desc = TRUE)) |>  
 ggplot(aes(x = company\_ranked, y = culture)) +  
 geom\_bar(stat = "identity")



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")

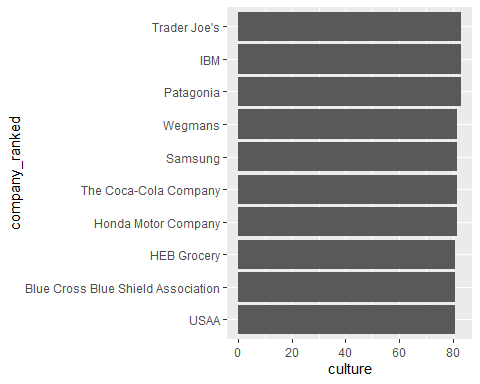
## Selecting by company\_ranked



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")

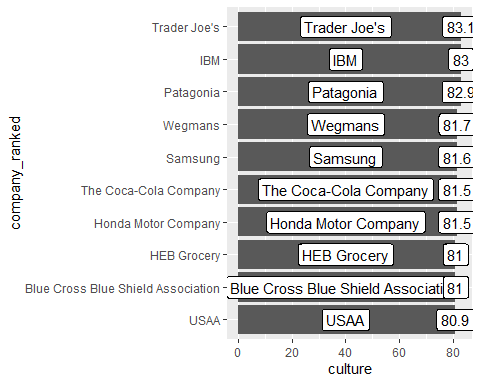
## Selecting by company\_ranked



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))

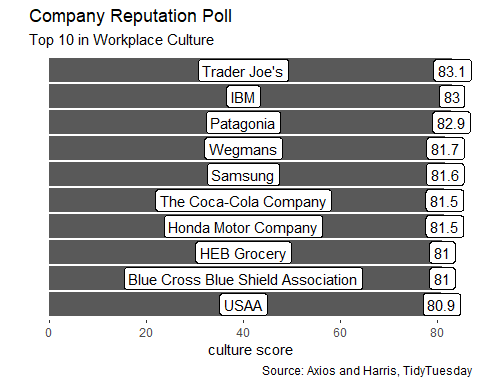
## Selecting by company\_ranked



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

