# PCA_2

Vandrade

2024-08-06

```
library(tidyverse)
```

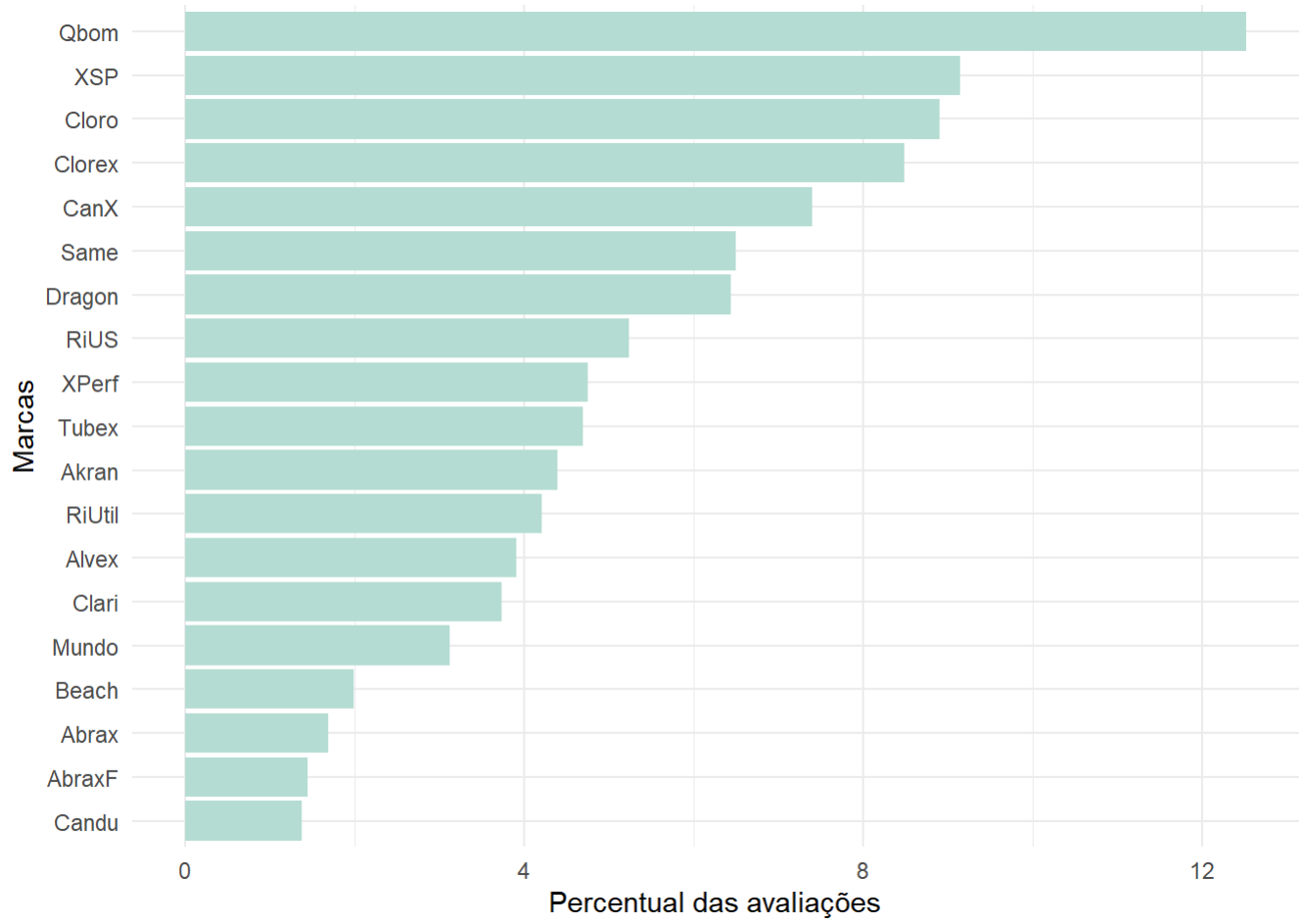
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(factoextra)
```

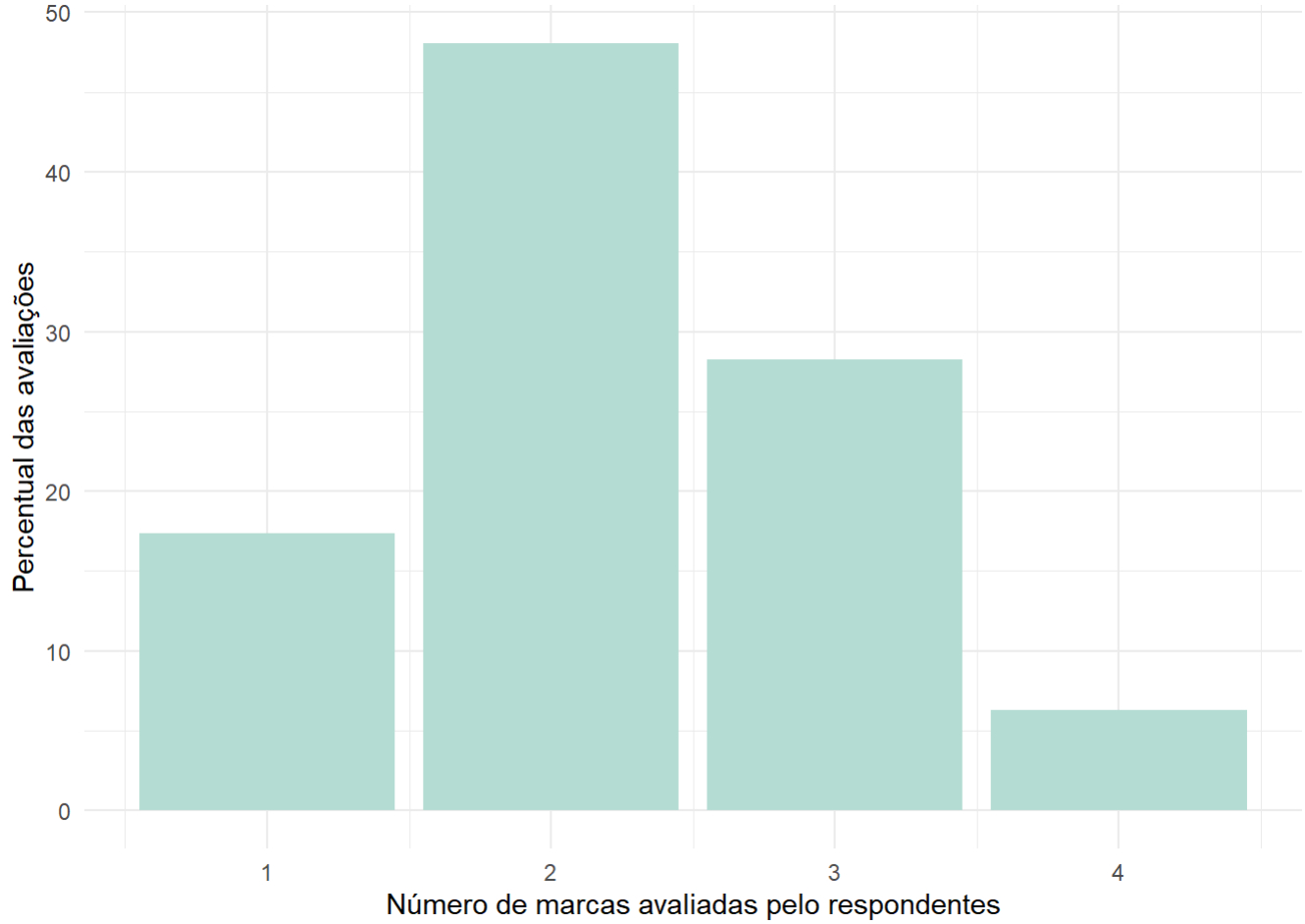
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
av <- read.csv("avaliacoes.csv")
```

```
av |>
  count(marca) |>
  mutate(pct_marca = prop.table(n)*100,
         marca = fct_reorder(marca, n)) |>
  ggplot(aes(x = marca, y=pct_marca)) +
  geom_bar(stat = "identity", fill =    "#b7ded2") +
  labs(x = 'Marcas', y = 'Percentual das avaliações') +
  theme_minimal() +
  coord_flip()
```




```
av |>
  group_by(respondente) |>
  summarize(marcas_avaliadas = n()) |>
  count(marcas_avaliadas) |>
  mutate(pct_marcas_avaliadas = prop.table(n)*100) |>
  ggplot(aes(x = marcas_avaliadas, y = pct_marcas_avaliadas)) +
  geom_bar(stat = "identity", fill = "#b7ded2") +
  theme_minimal() +
  labs(x = 'Número de marcas avaliadas pelo respondentes',
       y = 'Percentual das avaliações')
```




```
quest <- read_csv("questoes.csv")
```

```
## Rows: 29 Columns: 2
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (2): numero, pergunta
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pca <- av |>
  select(starts_with('Q')) |>
  prcomp(scale = TRUE)

# proportion of variance explained

pve <- cumsum((pca$sdev^2) / sum(pca$sdev^2))
```

Kaiser, Henry F. 1961. “A Note on Guttman’s Lower Bound for the Number of Common Factors.” British Journal of Statistical Psychology 14: 1-2.

An eigenvalue > 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained. This holds true only when the data are standardized.

```
pca |>
  get_eigenvalue() |>
  filter(eigenvalue >= 1)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  12.836409        44.263481                    44.26348
## Dim.2   2.978842        10.271869                    54.53535
## Dim.3   1.418355         4.890880                    59.42623
## Dim.4   1.039233         3.583564                    63.00979
```

```
Phi <- pca$rotation

sort( 100*Phi[,1]^2 / sum(Phi[,1]^2), decreasing = TRUE)
```

```
##      Q28      Q29      Q11      Q12      Q13      Q27      Q16      Q19
## 4.945283 4.731308 4.583622 4.488331 4.194313 4.112073 4.015795 4.007660
##      Q24      Q20      Q22      Q15       Q8       Q1      Q10       Q6
## 3.988114 3.967978 3.863419 3.843924 3.828314 3.798611 3.767432 3.717116
##       Q4      Q23      Q17       Q9      Q25       Q5       Q3      Q14
## 3.646803 3.525495 3.441287 3.327355 3.247392 3.021844 3.003838 2.938469
##      Q18      Q21       Q7      Q26       Q2
## 2.863425 2.331537 1.375088 1.071628 0.352545
```

a contribuição percentual de uma variável é 100 * o quadrado da carga correspodente dividido pela soma dos quadrados das cargas.

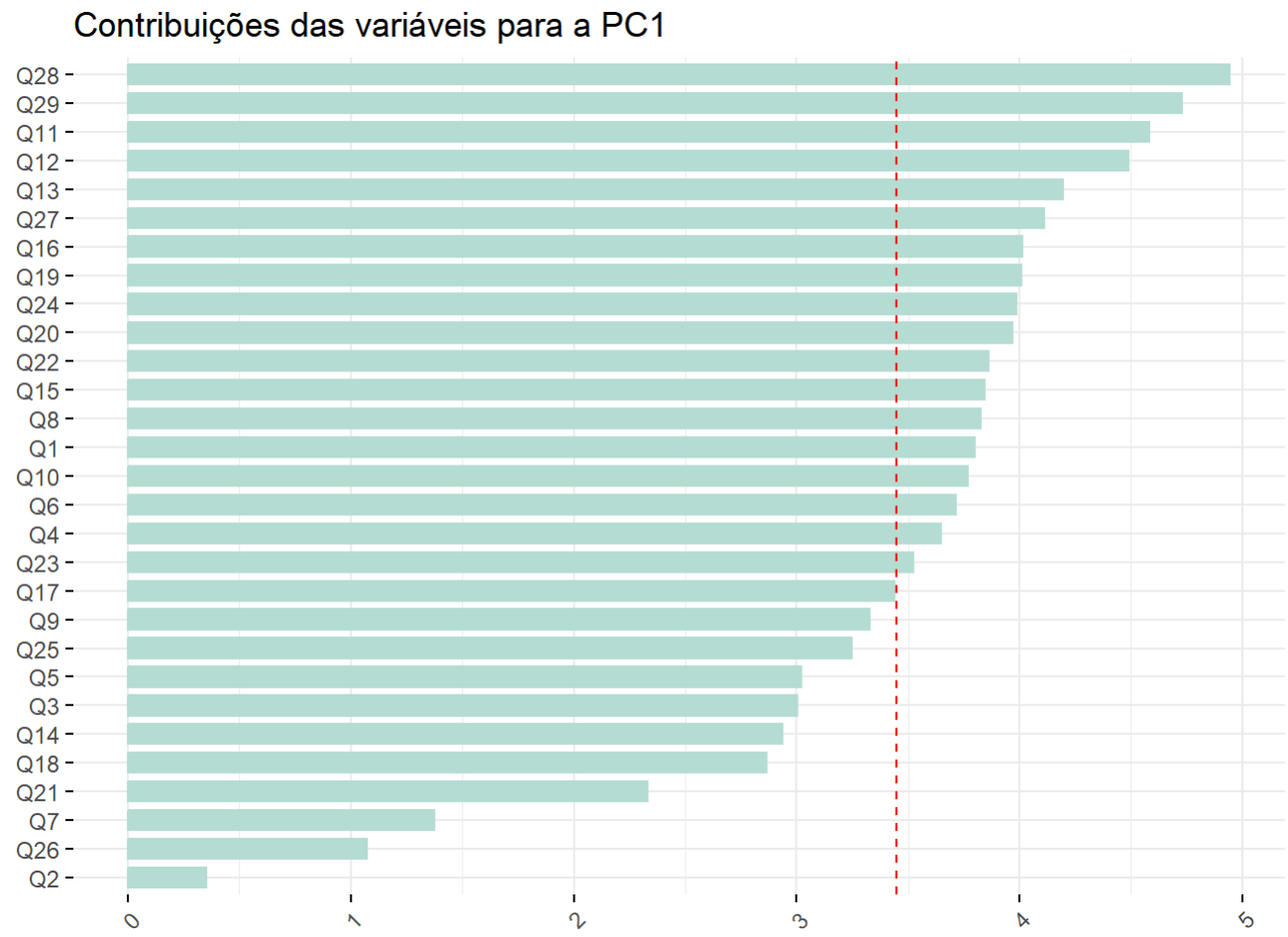
a linha tracejada seria o valor de uma contribuição percentual igual para cada variável; ou seja, 100 / número de variáveis

o termo “axes” determina a dimensao (PC) a ser observado

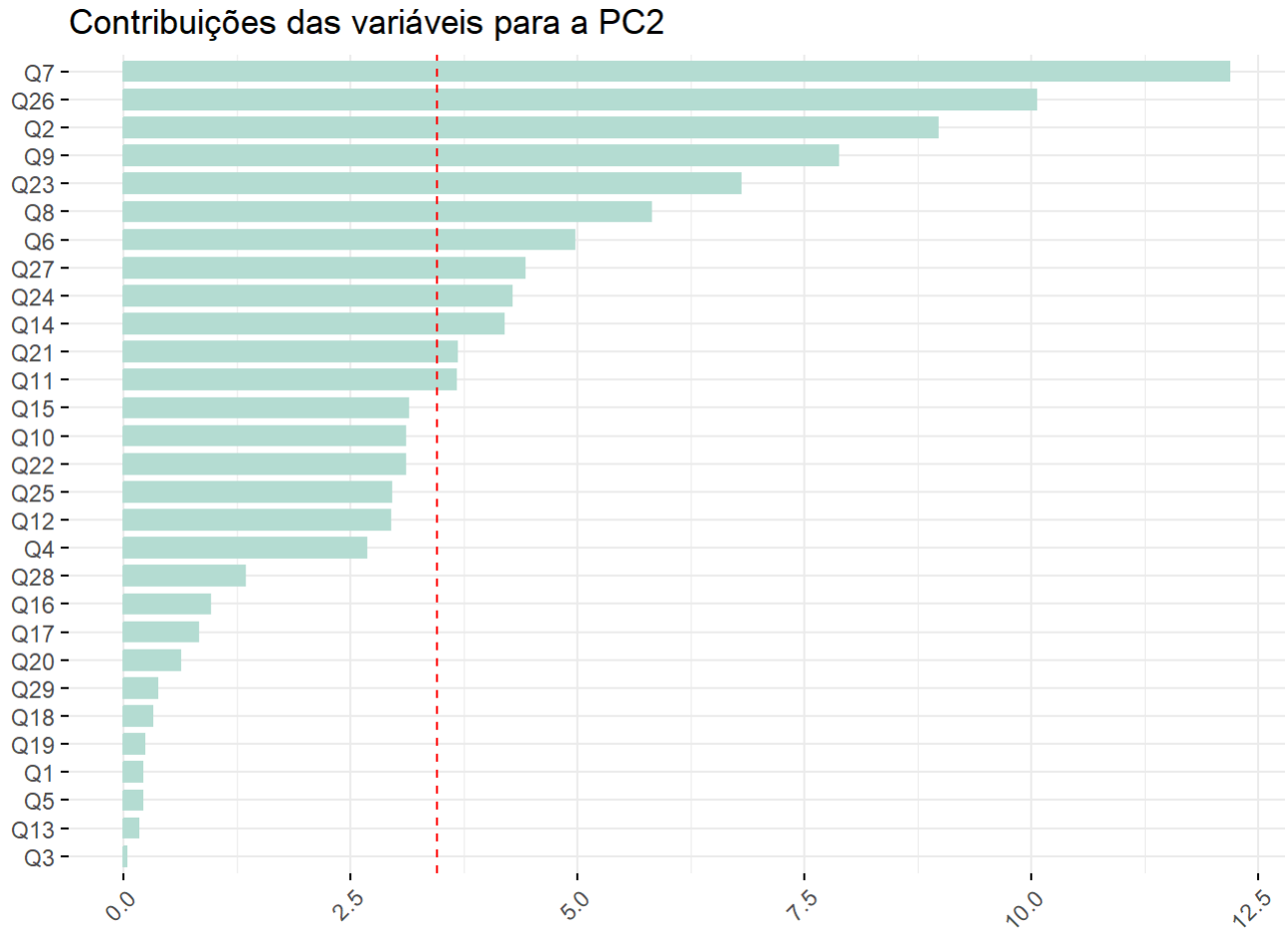
```
sort(100 * Phi[, 1]^2 / sum(Phi[, 1]^2), decreasing = TRUE)
```

```
##      Q28      Q29      Q11      Q12      Q13      Q27      Q16      Q19
## 4.945283 4.731308 4.583622 4.488331 4.194313 4.112073 4.015795 4.007660
##      Q24      Q20      Q22      Q15       Q8       Q1      Q10       Q6
## 3.988114 3.967978 3.863419 3.843924 3.828314 3.798611 3.767432 3.717116
##       Q4      Q23      Q17       Q9      Q25       Q5       Q3      Q14
## 3.646803 3.525495 3.441287 3.327355 3.247392 3.021844 3.003838 2.938469
##      Q18      Q21       Q7      Q26       Q2
## 2.863425 2.331537 1.375088 1.071628 0.352545
```

```
pca |>
  fviz_contrib("var", axes = 1, sort.val = "asc", fill = '#b7ded2', color = '#b7ded2') +
  labs(x = "", title = 'Contribuições das variáveis para a PC1') +
  coord_flip()
```




```
pca |>
  fviz_contrib("var", axes = 2, sort.val = "asc", fill = '#b7ded2', color = '#b7ded2') +
  labs(x = "", title = 'Contribuições das variáveis para a PC2') +
  coord_flip()
```




```
z <- pca$x[,1:3]

colnames(z) <- sprintf("driver_%d", 1:3)


get_driver <- function(Phi, quest, drv, top) {tibble(numero = rownames(Phi), carga = Phi[, dr
v]) %>%
        left_join(quest) %>%
        mutate(contribuicao = carga^2 / sum(carga^2)) %>%
        arrange(desc(contribuicao)) %>%
        head(n = top)}
```

os sinais relativos carga x score são importantes! abaixo, com os sinais originais, Q28 com escore mais negativo indicaria mais limpeza

```
driver_1 <- get_driver(Phi, quest, drv = 1, top = 6) # Limpeza
```

```
## Joining with `by = join_by(numero)`
```

ajustando os sinais relativos para tornar mais direta a interpretação da PC1

```
Phi[, 1] <- -Phi[, 1]
z[, 1] <- -z[, 1]

(driver_1 <- get_driver(Phi, quest, drv = 1, top = 6)) # Limpeza
```

```
## Joining with `by = join_by(numero)`
```

```
## # A tibble: 6 × 4
##   numero carga pergunta                                          contribuicao
##   <chr>  <dbl> <chr>                                                    <dbl>
## 1 Q28    0.222 É eficiente na limpeza da casa toda                     0.0495
## 2 Q29    0.218 Deixa a casa com aroma de limpeza                       0.0473
## 3 Q11    0.214 Facilita a tarefa da dona de casa na limpeza da casa    0.0458
## 4 Q12    0.212 Dá a melhor sensação de desinfecção                     0.0449
## 5 Q13    0.205 É a mais adequada para a lavagem de roupas              0.0419
## 6 Q27    0.203 Deixa um aroma agradável nos lugares onde foi usada     0.0411
```

```
(driver_2 <- get_driver(Phi, quest, drv = 2, top = 10)) # Suavidade
```

```
## Joining with `by = join_by(numero)`
```

```
## # A tibble: 10 × 4
##    numero  carga pergunta                                          contribuicao
##    <chr>   <dbl> <chr>                                                    <dbl>
##  1 Q7      0.349 É suave para as mãos                                    0.122
##  2 Q26     0.317 Não deixa um aroma forte e ruim nas mãos                0.101
##  3 Q2      0.299 É adequada para roupas coloridas                        0.0896
##  4 Q9      0.280 Deixa um aroma agradável nas roupas                     0.0787
##  5 Q23    -0.261 É eficiente para desinfetar vasos sanitários e ra…      0.0679
##  6 Q8     -0.241 É adequada para a limpeza pesada                        0.0581
##  7 Q6      0.223 Deixa um aroma agradável na casa                        0.0497
##  8 Q27     0.210 Deixa um aroma agradável nos lugares onde foi usa…      0.0442
##  9 Q24     0.207 Tem um aroma adequado para ser usada na casa toda       0.0427
## 10 Q14     0.205 É fácil de enxaguar                                     0.0419
```

```
(driver_3 <- get_driver(Phi, quest, drv = 3, top = 5)) # Intensidade
```

```
## Joining with `by = join_by(numero)`
```

```
## # A tibble: 5 × 4
##   numero  carga pergunta                                          contribuicao
##   <chr>   <dbl> <chr>                                                    <dbl>
## 1 Q3      0.291 É a melhor para a remoção de manchas de gordura          0.0846
## 2 Q20     0.277 É econômica no uso                                       0.0768
## 3 Q10    -0.266 É um produto para ser usado tanto na cozinha como …      0.0709
## 4 Q19     0.254 É a melhor marca de alvejante do mercado                 0.0646
## 5 Q16     0.254 Rende mais                                               0.0646
```

```
library(ggrepel)

tb <- tibble(marca = av$marca) |>
    bind_cols(as_tibble(z))


tb |>
  group_by(marca) |>
  summarise_all(mean) |>
  gather(key = 'driver', value = 'score_medio', driver_1:driver_3) |>
  ggplot(aes(x = driver, y = score_medio,
         group = marca, color = marca,
         label = ifelse(driver == "driver_1", marca, ""))) +
  geom_line(size = 1, alpha = .55) +
  geom_point(size = 2) +
  labs(x = "",
       y = "Escore Medio",
       title = "Posicionamento das marcas") +
  geom_label_repel(direction = "both") +
  scale_x_discrete(breaks = sprintf("driver_%d", 1:3),
                   labels = c("Limpeza",
                              "Suavidade",
                              "Intensidade")) +
  theme(legend.position = "none")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

## Posicionamento das marcas