Describing Data

Luis Francisco Gómez López

FAEDIS

2024-08-01

UNIVERSIDAD MILITAR
NUEVA GRANADA

Table of contents I

- This presentation is based on (Chapman and Feit 2019, chap. 3)

- Utilize descriptive statistics and single variable visualization techniques for summarizing and exploring a data set

- **storeNum**: store identifier
- **Year**: year identifier
- **Week**: week as it would appear in the ISO 8601 system (1-52)
- **p1sales**: units sold of product 1
- **p2sales**: units sold of product 2
- **p1price**: price of product 1
- **p2price**: price of product 2
- **p1prom**: whether product 1 was promoted (1) or not (0)
- **p2prom**: whether product 2 was promoted (1) or not (0)
- **country**: two-letter country codes defined in ISO 3166-1

- **Import data**

```
weekly_store <- read_csv(file = "http://goo.gl/QPDdMl")
weekly_store |> head(n=5)
```

```
# A tibble: 5 x 10
  storeNum  Year  Week p1sales p2sales p1price p2price p1prom p2prom country
     <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl> <chr>
1      101     1     1     127     106    2.29    2.29      0      0 US
2      101     1     2     137     105    2.49    2.49      0      0 US
3      101     1     3     156      97    2.99    2.99      1      0 US
4      101     1     4     117     106    2.99    3.19      0      0 US
5      101     1     5     138     100    2.49    2.59      0      1 US
```

- **Transform data**

```r
weekly_store <- weekly_store |>
  mutate(storeNum = factor(storeNum, ordered = FALSE),
         Year = factor(Year, levels = 1:2, ordered = TRUE),
         Week = factor(Week, levels = 1:52, ordered = TRUE),
         p1prom = as.logical(p1prom),
         p2prom = as.logical(p2prom))
weekly_store |> head(n=5)
```

```
# A tibble: 5 x 10
  storeNum Year  Week  p1sales p2sales p1price p2price p1prom p2prom country
  <fct>    <ord> <ord>   <dbl>   <dbl>   <dbl>   <dbl> <lgl>  <lgl>  <chr>
1 101      1     1         127     106    2.29    2.29 FALSE  FALSE  US
2 101      1     2         137     105    2.49    2.49 FALSE  FALSE  US
3 101      1     3         156      97    2.99    2.99 TRUE   FALSE  US
4 101      1     4         117     106    2.99    3.19 FALSE  FALSE  US
5 101      1     5         138     100    2.49    2.59 FALSE  TRUE   US
```

- **Inspect data: the base R way**

```
as.data.frame(weekly_store) |> str()
```

```
'data.frame':	2080 obs. of  10 variables:
 $ storeNum: Factor w/ 20 levels "101","102","103",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year    : Ord.factor w/ 2 levels "1"<"2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Week    : Ord.factor w/ 52 levels "1"<"2"<"3"<"4"<..: 1 2 3 4 5 6 7 8 9 10 ...
 $ p1sales : num  127 137 156 117 138 115 116 106 116 145 ...
 $ p2sales : num  106 105 97 106 100 127 90 126 94 91 ...
 $ p1price : num  2.29 2.49 2.99 2.99 2.49 2.79 2.99 2.99 2.29 2.49 ...
 $ p2price : num  2.29 2.49 2.99 3.19 2.59 2.49 3.19 2.29 2.29 2.99 ...
 $ p1prom  : logi  FALSE FALSE TRUE FALSE FALSE FALSE ...
 $ p2prom  : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
 $ country : chr  "US" "US" "US" "US" ...
```

- **Inspect data: the tidyverse way**

```
weekly_store |> glimpse()
```

```
Rows: 2,080
Columns: 10
$ storeNum <fct> 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 1~
$ Year     <ord> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ Week     <ord> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
$ p1sales  <dbl> 127, 137, 156, 117, 138, 115, 116, 106, 116, 145, 123, 169, 1~
$ p2sales  <dbl> 106, 105, 97, 106, 100, 127, 90, 126, 94, 91, 104, 73, 79, 10~
$ p1price  <dbl> 2.29, 2.49, 2.99, 2.99, 2.49, 2.79, 2.99, 2.99, 2.29, 2.49, 2~
$ p2price  <dbl> 2.29, 2.49, 2.99, 3.19, 2.59, 2.49, 3.19, 2.29, 2.29, 2.99, 2~
$ p1prom   <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,~
$ p2prom   <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE,~
$ country  <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "US", "~
```

- ## Summarize data: the R base way

```
weekly_store |> summary()
```

```
   storeNum     Year          Week          p1sales        p2sales     
 101     : 104   1:1040   1       :  40   Min.   : 73   Min.   : 51.0  
 102     : 104   2:1040   2       :  40   1st Qu.:113   1st Qu.: 84.0  
 103     : 104            3       :  40   Median :129   Median : 96.0  
 104     : 104            4       :  40   Mean   :133   Mean   :100.2  
 105     : 104            5       :  40   3rd Qu.:150   3rd Qu.:113.0  
 106     : 104            6       :  40   Max.   :263   Max.   :225.0  
 (Other):1456             (Other):1840                                 
    p1price          p2price        p1prom          p2prom       
 Min.   :2.190   Min.   :2.29   Mode :logical   Mode :logical  
 1st Qu.:2.290   1st Qu.:2.49   FALSE:1872      FALSE:1792     
 Median :2.490   Median :2.59   TRUE :208       TRUE :288      
 Mean   :2.544   Mean   :2.70                                  
 3rd Qu.:2.790   3rd Qu.:2.99                                  
 Max.   :2.990   Max.   :3.19                                  
                                                               
   country         
 Length:2080       
 Class :character  
 Mode  :character  
```

- **Summarize data: the skimr way**
  - Ups the table is really big!!! Try it in your console to see the complete table

```
weekly_store |> skim()
```

- **Count data: the R base way**

```
table(weekly_store$p1price)
```

```

2.19 2.29 2.49 2.79 2.99
 395  444  423  443  375
```

- **Count data: the tidyverse way**

```
weekly_store |> count(p1price)
```

```
# A tibble: 5 x 2
  p1price     n
    <dbl> <int>
1    2.19   395
2    2.29   444
3    2.49   423
4    2.79   443
5    2.99   375
```

- **Data visualization**

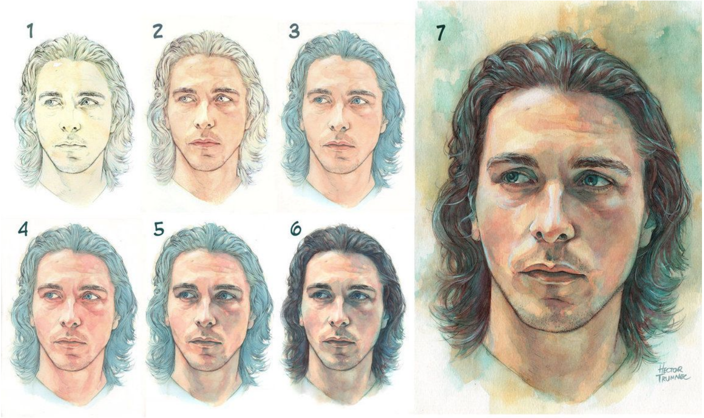
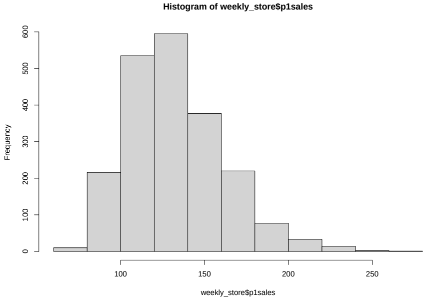


Figure 1: Analogy of data visualization as painting step by step (Watercolor portrait - Step by Step by Hector Trunnec (Valencia, Spain) 2015-03-03)

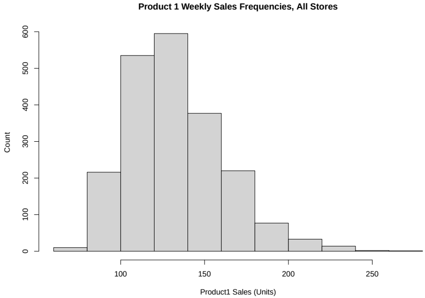
- **Histograms: the base R way**

```r
weekly_store$p1sales |> hist()
```



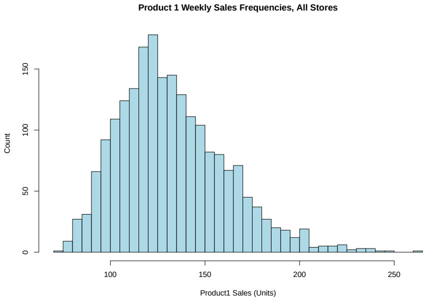
- **Histograms: the base R way**

```
weekly_store$p1sales |> hist(main = "Product 1 Weekly Sales Frequencies, All Stores",
                             xlab = "Product1 Sales (Units)" ,
                             ylab = "Count")
```



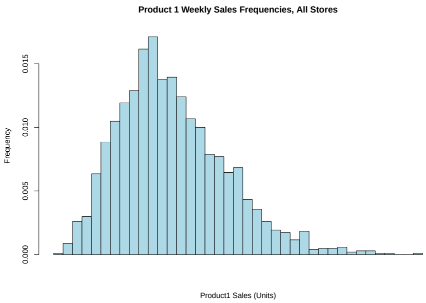
- **Histograms: the base R way**

```
weekly_store$p1sales |> hist(main = "Product 1 Weekly Sales Frequencies, All Stores",
                             xlab = "Product1 Sales (Units)" ,
                             ylab = "Count",
                             breaks = 30,
                             col =  "lightblue")
```



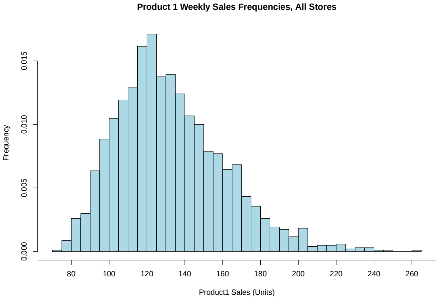
- **Histograms: the base R way**

```
weekly_store$p1sales |> hist(main = "Product 1 Weekly Sales Frequencies, All Stores",
                             xlab = "Product1 Sales (Units)" ,
                             ylab = "Frequency",
                             breaks = 30,
                             col =  "lightblue",
                             freq = FALSE,
                             xaxt = "n")
```



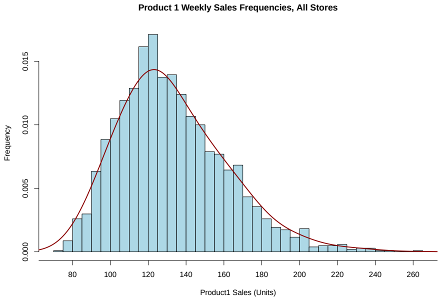
- **Histograms: the base R way**

```
weekly_store$p1sales |> hist(main = "Product 1 Weekly Sales Frequencies, All Stores",
                             xlab = "Product1 Sales (Units)" ,
                             ylab = "Frequency",
                             breaks = 30,
                             col =  "lightblue",
                             freq = FALSE,
                             xaxt = "n")
axis(side=1 , at=seq(from = 60, to = 300, by = 20))
```



- **Histograms: the base R way**

```
weekly_store$p1sales |> hist(main = "Product 1 Weekly Sales Frequencies, All Stores",
                             xlab = "Product1 Sales (Units)" ,
                             ylab = "Frequency",
                             breaks = 30,
                             col =  "lightblue",
                             freq = FALSE,
                             xaxt = "n")
axis(side=1 , at=seq(from = 60, to = 300, by = 20))
lines(x = density(weekly_store$p1sales, bw=10), type="l", col="darkred", lwd=2)
```

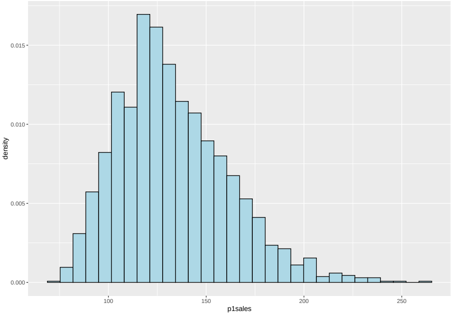


- **Histograms: the tidyverse way**

```
weekly_store |> ggplot()
```

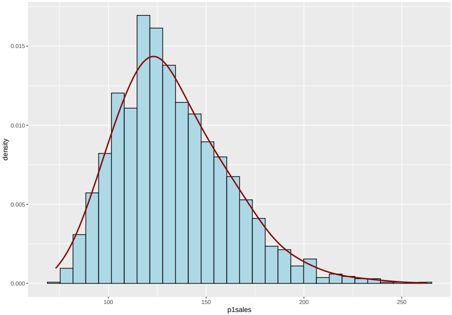
- **Histograms: the tidyverse way**

```r
weekly_store |> ggplot() +
  geom_histogram(aes(x = p1sales, y = after_stat(density)),
                 color = "black", fill = "lightblue", bins = 30)
```



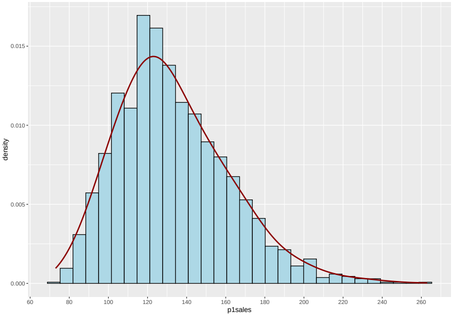
Histograms: the tidyverse way

```
weekly_store |> ggplot() +
  geom_histogram(aes(x = p1sales, y = after_stat(density)),
                 color = "black", fill = "lightblue", bins = 30) +
  geom_density(aes(x=p1sales),
               bw=10, color="darkred",
               linetype = "solid", linewidth = 1)
```



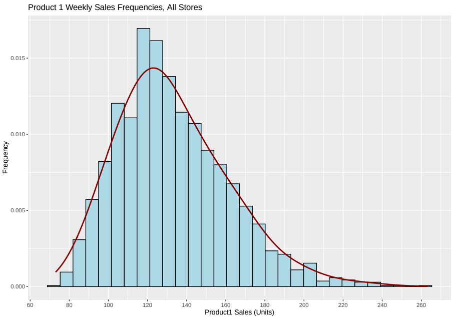
- **Histograms: the tidyverse way**

```
weekly_store |> ggplot() +
  geom_histogram(aes(x=p1sales, y = after_stat(density)),
                 color = "black", fill = "lightblue", bins = 30) +
  geom_density(aes(x=p1sales),
               bw=10, color="darkred", linetype="solid", linewidth=1) +
  scale_x_continuous(breaks = seq(from = 60, to = 300, by = 20))
```



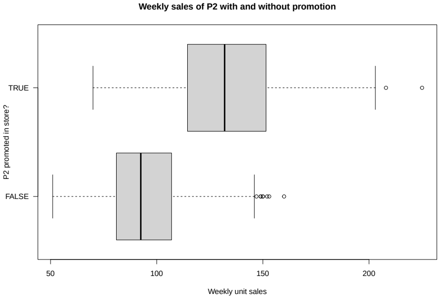
- **Histograms: the tidyverse way**

```
weekly_store |> ggplot() +
  geom_histogram(aes(x = p1sales, y = after_stat(density)),
                 color = "black", fill = "lightblue", bins = 30) +
  geom_density(aes(x = p1sales),
               bw = 10, color = "darkred", linetype = "solid", linewidth = 1) +
  scale_x_continuous(breaks = seq(from = 60, to = 300, by = 20)) +
  labs(x = "Product1 Sales (Units)", y = "Frequency",
       title = "Product 1 Weekly Sales Frequencies, All Stores")
```



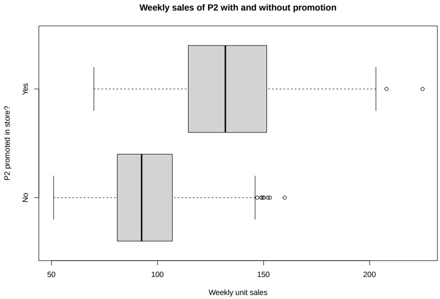
- **Boxplots: the base R way**
  - Boxplot product 2 sales by promotion

```
boxplot(weekly_store$p2sales ~ weekly_store$p2prom,
        main = "Weekly sales of P2 with and without promotion",
        xlab = "Weekly unit sales", ylab = "P2 promoted in store?",
        horizontal = TRUE, las = 1)
```



- **Boxplots: the base R way**
    - Boxplot product 2 sales by promotion

```r
boxplot(weekly_store$p2sales ~ weekly_store$p2prom,
        main = "Weekly sales of P2 with and without promotion",
        xlab = "Weekly unit sales", ylab = "P2 promoted in store?",
        horizontal = TRUE, las = 1, yaxt = "n")
axis(side = 2, at = c(1,2), labels = c("No", "Yes"))
```

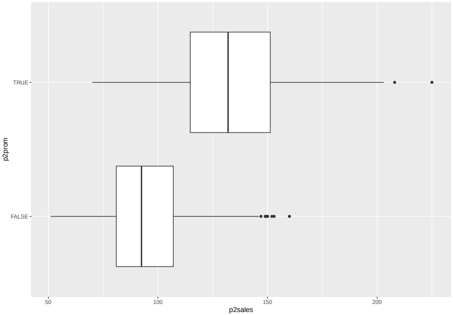


- **Boxplots: the tidyverse way**
  - Boxplot product 2 sales by promotion

```
weekly_store |> ggplot()
```

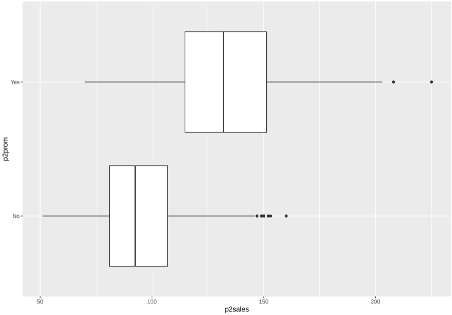
- **Boxplots: the tidyverse way**
    - Boxplot product 2 sales by promotion

```
weekly_store |> ggplot() +
  geom_boxplot(aes(x = p2sales, y = p2prom))
```



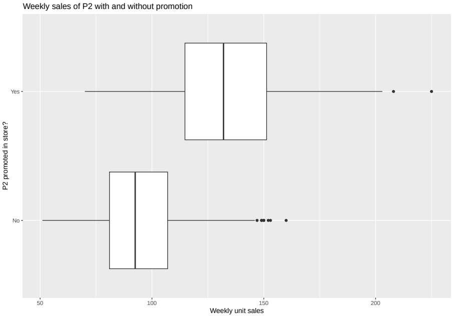
- **Boxplots: the tidyverse way**
  - Boxplot product 2 sales by promotion

```r
weekly_store |> ggplot() +
  geom_boxplot(aes(x = p2sales, y = p2prom)) +
  scale_y_discrete(labels = c("No", "Yes"))
```



- **Boxplots: the tidyverse way**
  - Boxplot product 2 sales by promotion

```
weekly_store |> ggplot() +
  geom_boxplot(aes(x = p2sales, y = p2prom)) +
  scale_y_discrete(labels = c("No", "Yes")) +
  labs(x = "Weekly unit sales", y = "P2 promoted in store?",
       title = "Weekly sales of P2 with and without promotion")
```



- **In what countries the company sell more units of product 2?**
  - Preparing the data

```
weekly_store_sales_by_country <- weekly_store |>
  group_by(country)
weekly_store_sales_by_country
```

```
# A tibble: 2,080 x 10
# Groups:   country [7]
   storeNum Year  Week  p1sales p2sales p1price p2price p1prom p2prom country
   <fct>    <ord> <ord>   <dbl>   <dbl>   <dbl>   <dbl> <lgl>  <lgl>  <chr>
 1 101      1     1           127     106    2.29    2.29 FALSE  FALSE  US
 2 101      1     2           137     105    2.49    2.49 FALSE  FALSE  US
 3 101      1     3           156      97    2.99    2.99 TRUE   FALSE  US
 4 101      1     4           117     106    2.99    3.19 FALSE  FALSE  US
 5 101      1     5           138     100    2.49    2.59 FALSE  TRUE   US
 6 101      1     6           115     127    2.79    2.49 FALSE  FALSE  US
 7 101      1     7           116      90    2.99    3.19 FALSE  FALSE  US
 8 101      1     8           106     126    2.99    2.29 FALSE  FALSE  US
 9 101      1     9           116      94    2.29    2.29 FALSE  FALSE  US
10 101      1     10          145      91    2.49    2.99 FALSE  FALSE  US
# i 2,070 more rows
```

- **In what countries the company sell more units of product 2?**
    - Preparing the data

```
weekly_store_sales_by_country <- weekly_store |>
  group_by(country) |>
  summarise(sum_p2sales = sum(p2sales))
weekly_store_sales_by_country
```

```
# A tibble: 7 x 2
  country sum_p2sales
  <chr>         <dbl>
1 AU             9934
2 BR            21362
3 CN            20911
4 DE            52263
5 GB            31264
6 JP            41344
7 US            31248
```

- **In what countries the company sell more units of product 2?**
    - Preparing the data

```r
weekly_store_sales_by_country <- weekly_store |>
  group_by(country) |>
  summarise(sum_p2sales = sum(p2sales)) |>
  mutate(country = fct_reorder(.f = country, .x = sum_p2sales))
weekly_store_sales_by_country
```

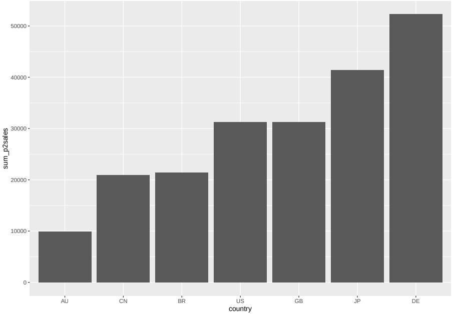
```
# A tibble: 7 x 2
  country sum_p2sales
  <fct>         <dbl>
1 AU             9934
2 BR            21362
3 CN            20911
4 DE            52263
5 GB            31264
6 JP            41344
7 US            31248
```

- **In what countries the company sell more units of product 2?**
  - Visualizing data

```
weekly_store_sales_by_country |> ggplot()
```

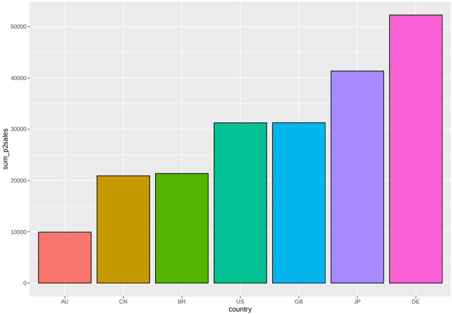
- **In what countries the company sell more units of product 2?**
  - Visualizing data

```
weekly_store_sales_by_country |> ggplot() +
  geom_col(aes(x = country, y = sum_p2sales))
```



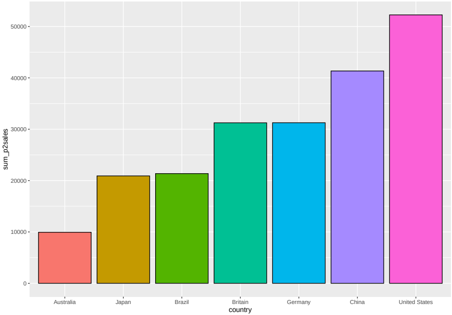
- **In what countries the company sell more units of product 2?**
    - Visualizing data

```
weekly_store_sales_by_country |> ggplot() +
  geom_col(aes(x = country, y = sum_p2sales, fill = country),
           color = "black", show.legend = FALSE)
```



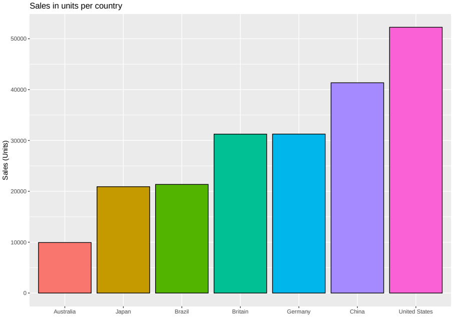
- **In what countries the company sell more units of product 2?**
    - Visualizing data

```r
weekly_store_sales_by_country |> ggplot() +
  geom_col(aes(x = country, y = sum_p2sales, fill = country),
           color = "black", show.legend = FALSE) +
  scale_x_discrete(labels = c("Australia", "Japan", "Brazil",
                              "Britain", "Germany", "China", "United States"))
```



- **In what countries the company sell more units of product 2?**
  - Visualizing data

```r
weekly_store_sales_by_country |> ggplot() +
  geom_col(aes(x = country, y = sum_p2sales, fill = country),
           color = "black", show.legend = FALSE) +
  scale_x_discrete(labels = c("Australia", "Japan", "Brazil",
                              "Britain", "Germany", "China", "United States")) +
  labs(x = NULL, y = "Sales (Units)",
       title = "Sales in units per country")
```



- To my family that supports me
- To the taxpayers of Colombia and the **UMNG students** who pay my salary
- To the **Business Science** and **R4DS Online Learning** communities where I learn **R** and π**-thon**
- To the **R Core Team**, the creators of **RStudio IDE**, **Quarto** and the authors and maintainers of the packages **tidyverse**, **skimr** and **tinytex** for allowing me to access these tools without paying for a license
- To the **Linux kernel community** for allowing me the possibility to use some **Linux distributions** as my main **OS** without paying for a license

References I

Chapman, Chris, and Elea McDonnell Feit. 2019. *R For Marketing Research and Analytics*. 2nd ed. 2019. Use R! Cham: Springer International Publishing : Imprint: Springer. https://doi-org.ezproxy.umng.edu.co/10.1007/978-3-030-14316-9.