Avocado Production by Country

Reconstruct Visualization of Avocado Production by Country

Introduction:

In this notebook, I will explain the problems represented in the data visualizations and how to solve them and reconstruct the right visual, which can lead to the correct insights and not distract the eyes of the audience from reality and the correct patterns. Our example is avocado production between the countries over time in metric tonnes. The chart shows which country was the largest producer of avocados until 2020, and how many tons of avocados this country produced. It also provides a comparison between the top-producing countries and shows the differences between their productions over time. The chart has a lot of issues; I will describe them below, and I will scrape the data from Wikipedia to reconstruct the chart and get the truth.

The issues are as follows:

- An issue with data integrity
- Perceived bias
- Color issues
- Deceptive method
- Visual design and quality

Original chart:

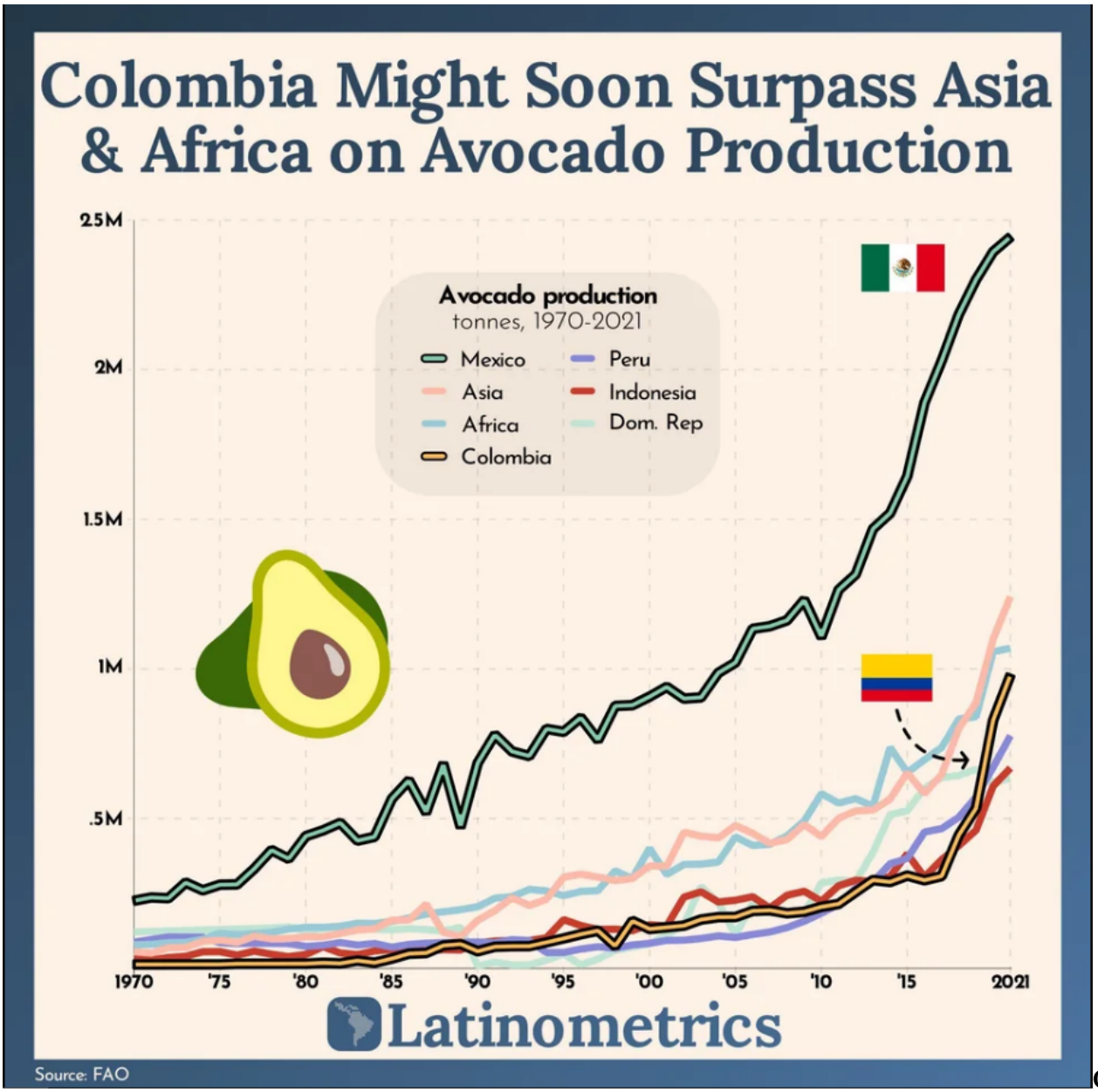


Chart Source:

- https://external-preview.redd.it/BoMxW7q1Pjnl4yLCD2xHk5Bkhz81ltsOXA8-MeT-Cao.jpgauto=webp&v=enabled&
- DataIsBeautiful Community on Reddit

Objective:

The publisher chooses the line chart to represent avocado production over time and plots multiple line charts to represent the countries and continents. He shows that Mexico is the top country that produces avocados, with almost 2.5 million metric tons in 2020. The second is Asia, with 1.3 million, followed by Africa and

Colombia. The publisher targets people who are interested in avocados and their market. He also publishes it in a community of data science learners on Reddit to learn data visualizations.

The visualization chosen had the following three main issues:

- **Low data-ink ratio:**

  The data-ink ratio is a concept introduced by Edward Tufte, a pioneer in the field of data visualization. It refers to the proportion of ink (or other visual elements) in a graph or chart that is used to represent the actual data being presented, as opposed to non-data elements such as labels, titles, and grid lines.

  The publisher used annotations like flags and the avocado picture and different colors for line charts; some of them have black borders and others do not; a yellow background; grid lines; and a blue outer frame. He also used the yellow frame for the legend. Inconsistent x-axis range, he moved between the years by a 5-year step, but in the right, he moved from 2015 to 2021 by a 6-year step, not 5.

- **Inconsistent insights (Bias):**

  The publisher didn't use a consistent representation. He compared Mexico and Colombia with continents like Africa and Asia. This led to significant bias because countries like Kenya were classified among the top ten avocado producers. So when he represented Africa, we knew that most of the production came from Kenya, so he should not compare the entities in this way.

  Another misleading issue is that he first represented Indonesia as a separate country from Asia, then Asia as a separate entity. This is an inconsistent representation of entities.

  Another issue he assumed from the chart is that Colombia will soon surpass Africa and Asia. But when we see the chart, it is not the truth because the line of Columbia is always below the Africa and Asia lines, and there is no strong evidence about that assumption. He inconsistently represented the data and derived unreliable insights.

- **Displaying too much data**

  The publisher plotted a lot of information on the graph. The lines are overlapping which is difficult to observe a particular value of any country at a specific time. He also used a large time frame, we must be interested in the relevant data and recent time frame like the last 5 years of production.

Code:

The code below was used to address the issues mentioned in the original chart. I will scrape the source data, clean it, organize it, and analyze it in order to provide accurate and unambiguous answers to the question:

`What countries have produced the most avocados over the last five years?`

Data collection:

First, we need to scrape the data from Wikipedia. I am interested in the last five years of the original graph because it is the recent time frame (2016-2020).

Data Reference:

- From Wikipedia, based on data from the **Food and Agriculture Organization Corporate Statistical Database** (**FAOSTAT**).
- URL: https://en.wikipedia.org/wiki/List_of_countries_by_avocado_production

```r
#scrape data from the Wikipedia page
library(rvest)

#URL
url <- 'https://en.wikipedia.org/wiki/List_of_countries_by_avocado_production'

#reads the HTML content of the page
html_scrape <- read_html(url)

#extracts the first table on the page
avocado_data <- html_table(html_scrape,fill = TRUE)[[1]]

#converts the resulting table into a data frame
avocado_data <- as.data.frame(avocado_data)

#displays the first 6 rows
head(avocado_data)
```

```
##            Country/region      2020      2019      2018      2017      2016
## 1 Mexico (Cultivation in) 2,393,849 2,300,889 2,184,663 2,029,886 1,889,354
## 2                Colombia   876,754   535,021   445,075   308,166   294,389
## 3      Dominican Republic   676,373   665,652   644,603   637,688   601,349
## 4                    Peru   660,003   571,992   504,840   466,796   455,394
## 5               Indonesia   609,049   461,613   410,084   363,157   304,938
## 6                   Kenya   322,556   264,032   233,933   217,688   176,045
```

```r
#Structure of the data
str(avocado_data)
```

```
## 'data.frame':    64 obs. of  6 variables:
##  $ Country/region: chr  "Mexico (Cultivation in)" "Colombia" "Dominican Republic" "Peru" ...
##  $ 2020          : chr  "2,393,849" "876,754" "676,373" "660,003" ...
##  $ 2019          : chr  "2,300,889" "535,021" "665,652" "571,992" ...
##  $ 2018          : chr  "2,184,663" "445,075" "644,603" "504,840" ...
##  $ 2017          : chr  "2,029,886" "308,166" "637,688" "466,796" ...
##  $ 2016          : chr  "1,889,354" "294,389" "601,349" "455,394" ...
```

- We observed that the number of countries that produce avocados is 64.
- The data have quality issues needed to be fixed.

Data preparation:

- The columns types should be converted to numerical type.
- Rename the “country/region” column to “country” for simplicity.
- Remove the “,” between the numbers.
- Remove the (Cultivation in) string from the first row in the country/region column.

```r
#data manipulation package
library(dplyr)

# Rename the first column of the data frame
colnames(avocado_data)[1] <- "Country"

# Convert all columns (except the first) to numeric and remove commas
avocado_data <- avocado_data %>%
  mutate(across(-1, ~as.numeric(gsub(",", "", .))))

#displays the first 6 rows
head(avocado_data)
```

```
##                   Country    2020    2019    2018    2017    2016
## 1 Mexico (Cultivation in) 2393849 2300889 2184663 2029886 1889354
## 2                Colombia  876754  535021  445075  308166  294389
## 3      Dominican Republic  676373  665652  644603  637688  601349
## 4                    Peru  660003  571992  504840  466796  455394
## 5               Indonesia  609049  461613  410084  363157  304938
## 6                   Kenya  322556  264032  233933  217688  176045
```

```r
# Remove the string " (Cultivation in)" from the first column
avocado_data[,1] <- gsub(" \\(Cultivation in\\)", "", avocado_data[,1])

#displays the first 6 rows
head(avocado_data)
```

```
##              Country    2020    2019    2018    2017    2016
## 1             Mexico 2393849 2300889 2184663 2029886 1889354
## 2           Colombia  876754  535021  445075  308166  294389
## 3 Dominican Republic  676373  665652  644603  637688  601349
## 4               Peru  660003  571992  504840  466796  455394
## 5          Indonesia  609049  461613  410084  363157  304938
## 6              Kenya  322556  264032  233933  217688  176045
```

Descriptive statistics:

```r
# Obtain summary statistics for each variable
summary(avocado_data)
```

```
##    Country               2020             2019             2018
##  Length:64          Min.   :     13   Min.   :     13   Min.   :     13
##  Class :character   1st Qu.:   1502   1st Qu.:   1503   1st Qu.:   1503
##  Mode  :character   Median :  12882   Median :  13941   Median :  11914
##                     Mean   : 125928   Mean   : 110372   Mean   : 105165
##                     3rd Qu.:  98281   3rd Qu.:  94433   3rd Qu.:  90252
##                     Max.   :2393849   Max.   :2300889   Max.   :2184663
##       2017              2016
##  Min.   :     13   Min.   :     12
##  1st Qu.:   1705   1st Qu.:   1508
```

```
##  Median :  12438   Median :  11452
##  Mean   :  96639   Mean   :  89418
##  3rd Qu.:  84308   3rd Qu.:  87463
##  Max.   :2029886   Max.   :1889354
```

- We noticed that avocado production was increased over time.
- 2020 has the largest mean across all years.
- There are 64 countries around the world that produce avocados until 2020.
- The largest amount of production of avocados in one year is 2393849 tons in 2020.

Data visualization (Reconstruction):

In this part, I will construct the chart to avoid the issues the publisher made.

First, we need to get the six most countries that have the largest production in 2020 and compare their production between 2016-2020.

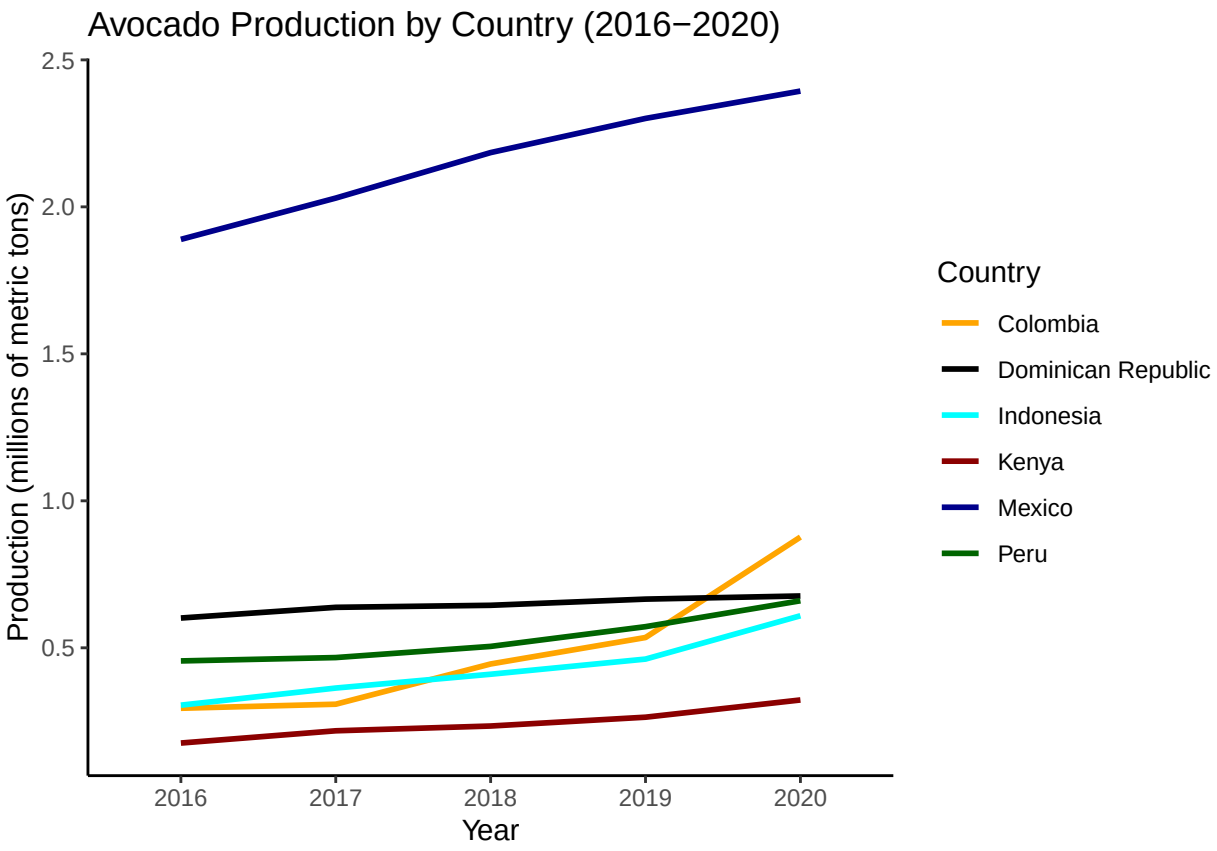
```
# Sort the data frame by the 2020 column in descending order
avocado_data_sorted <- arrange(avocado_data, desc(`2020`))

#select the top 6 countries
avocado_data_top <- avocado_data_sorted[1:6,]

# View the sorted data frame
avocado_data_top
```

```
##              Country    2020    2019    2018    2017    2016
## 1             Mexico 2393849 2300889 2184663 2029886 1889354
## 2           Colombia  876754  535021  445075  308166  294389
## 3 Dominican Republic  676373  665652  644603  637688  601349
## 4               Peru  660003  571992  504840  466796  455394
## 5          Indonesia  609049  461613  410084  363157  304938
## 6              Kenya  322556  264032  233933  217688  176045
```

The following plot fixes the main issues in the original plot.

```
#  data manipulation package for reshaping data between "wide" and "long" formats
library(tidyr)

#data visualization package
library(ggplot2)

# Reshape the data from wide to long format
avocado_data_long <- gather(avocado_data_top, "Year", "Production", `2016`:`2020`, factor_key = TRUE)

# Convert the Country column to a factor
avocado_data_long$Country <- factor(avocado_data_long$Country)

# Create a ggplot object with the data
ggplot(avocado_data_long, aes(x = Year, y = Production/1000000, group = Country, color = Country)) +
  # Add a line layer for each country with a thicker line width   and custom colors
```

```
  geom_line(size = 1) +
  scale_color_manual(values = c("orange", "black", "cyan",  "darkred", "darkblue", "darkgreen")) +
  # Remove the gridlines
  theme(
  # Hide panel borders and remove grid lines
  panel.border = element_blank(),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank()) +
  # Add a title and axis labels
  labs(title = "Avocado Production by Country (2016-2020)",
       x = "Year",
       y = "Production (millions of metric tons)") +
  # Use a classic theme with a white background
  theme_classic()
```



Mexico is the biggest avocado-producing country, with 2.4 million metric tons in 2020. Over time, it has dominated global avocado production. The Dominican Republic has now emerged as a reliable producer over time. Indonesia and Kenya now appear independently from their continents, representing clarity and avoiding bias. Colombia's production climbed from 0.4 million metric tons in 2016 to 0.8 million metric tons in 2020.

```
# Save the plot as a PNG file with a resolution of 300 dpi
ggsave("avocado_production.png", dpi = 300)
```

Conclusions:

I solved a lot of the problems in the new plot:

- Use a consistent format and reliable data from trustworthy sources.
- The comparison is fair; we have not introduced bias now.
- Plot the recent data that we need for this analysis.
- Remove unnecessary elements like grid lines, the background, annotations, and the frames.
- We can now observe that Colombia is the second-ranked country in avocado production in 2020.
- The colors are simple, and the plot is very informative.