Week 3

Essential Data Visualization

Agenda

- Data Visualisation - The good and the bad.
- The essential data visualizations
- Stylizing your plots

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

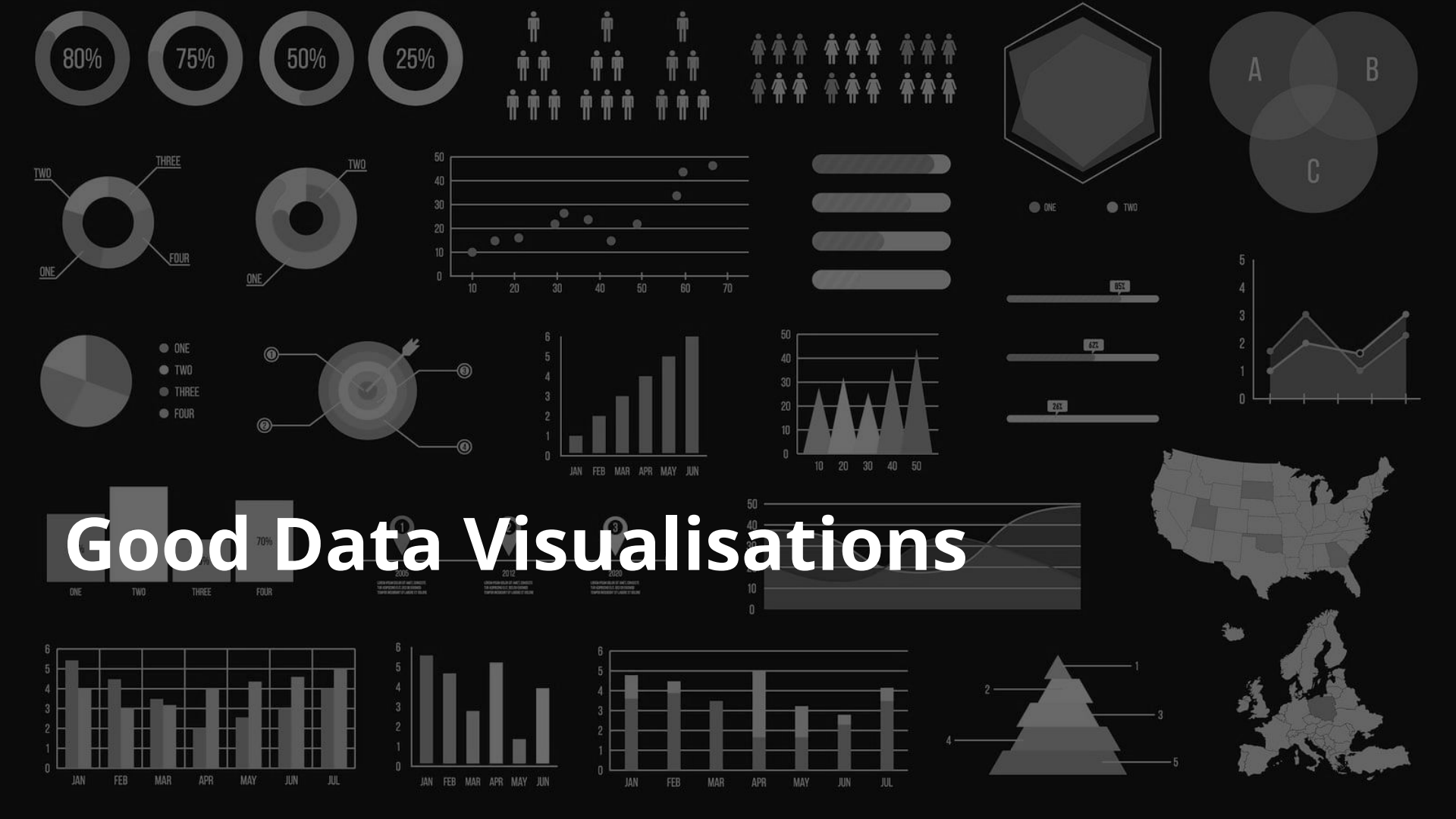
From the October 2012 Issue

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

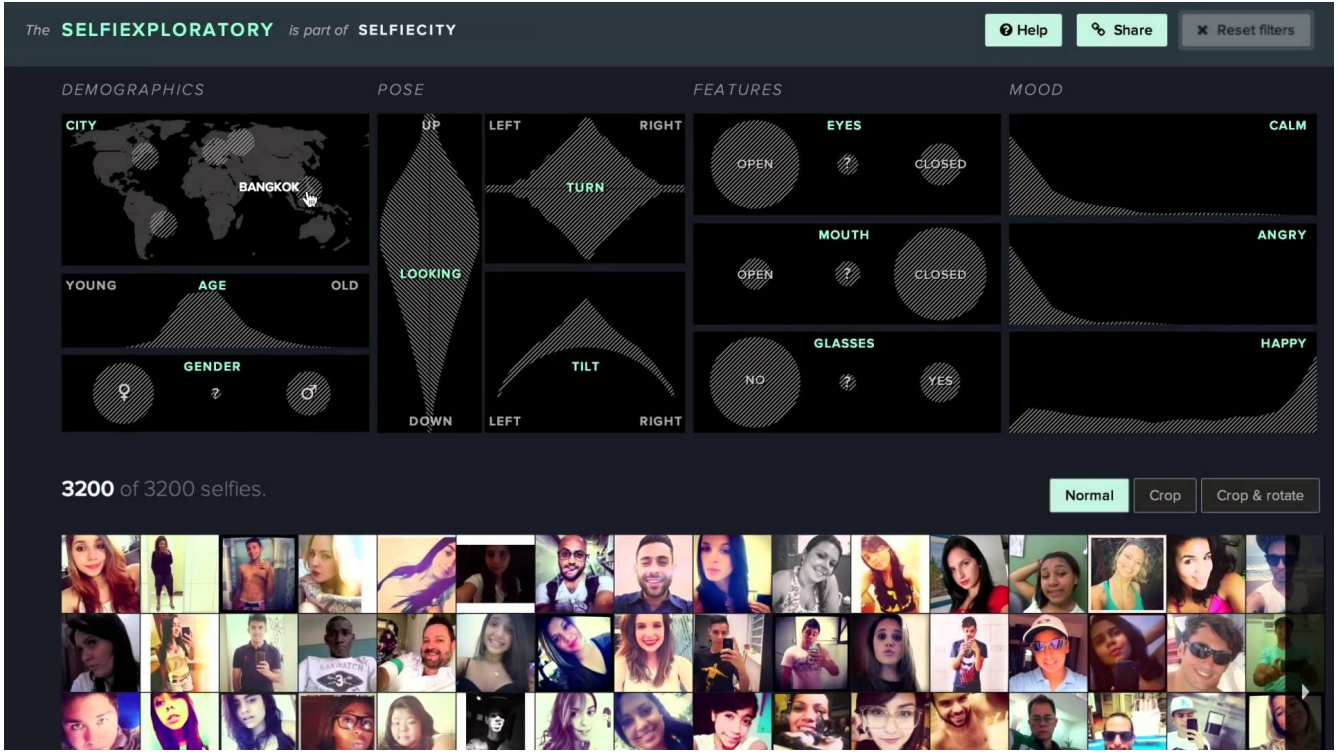
Sep 27, 2012, 10:49pm EDT

Data Scientists: The Definition of Sexy

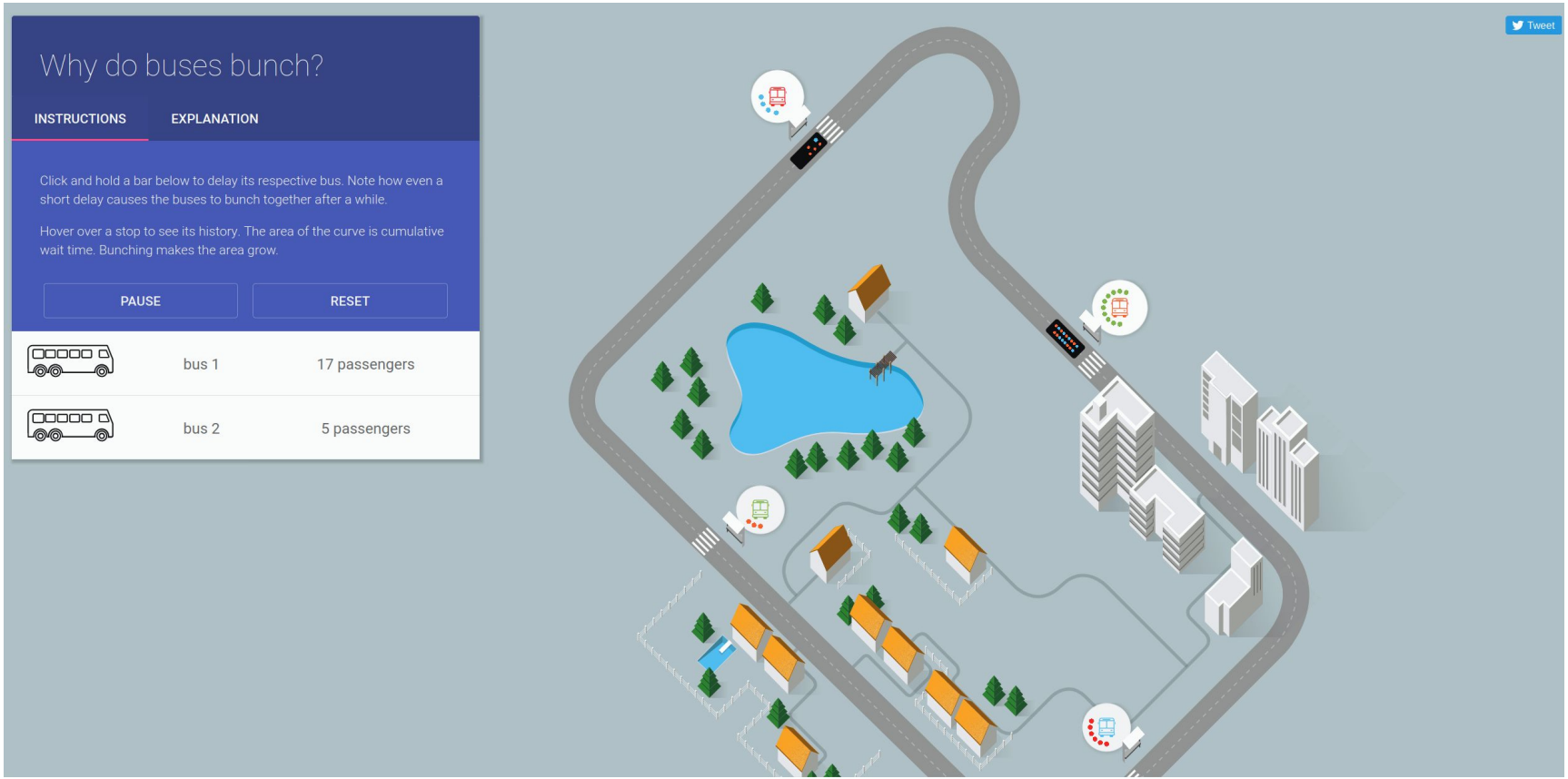
Gil Press Senior Contributor ⓘ
Enterprise & Cloud
I write about technology, entrepreneurs and innovation.



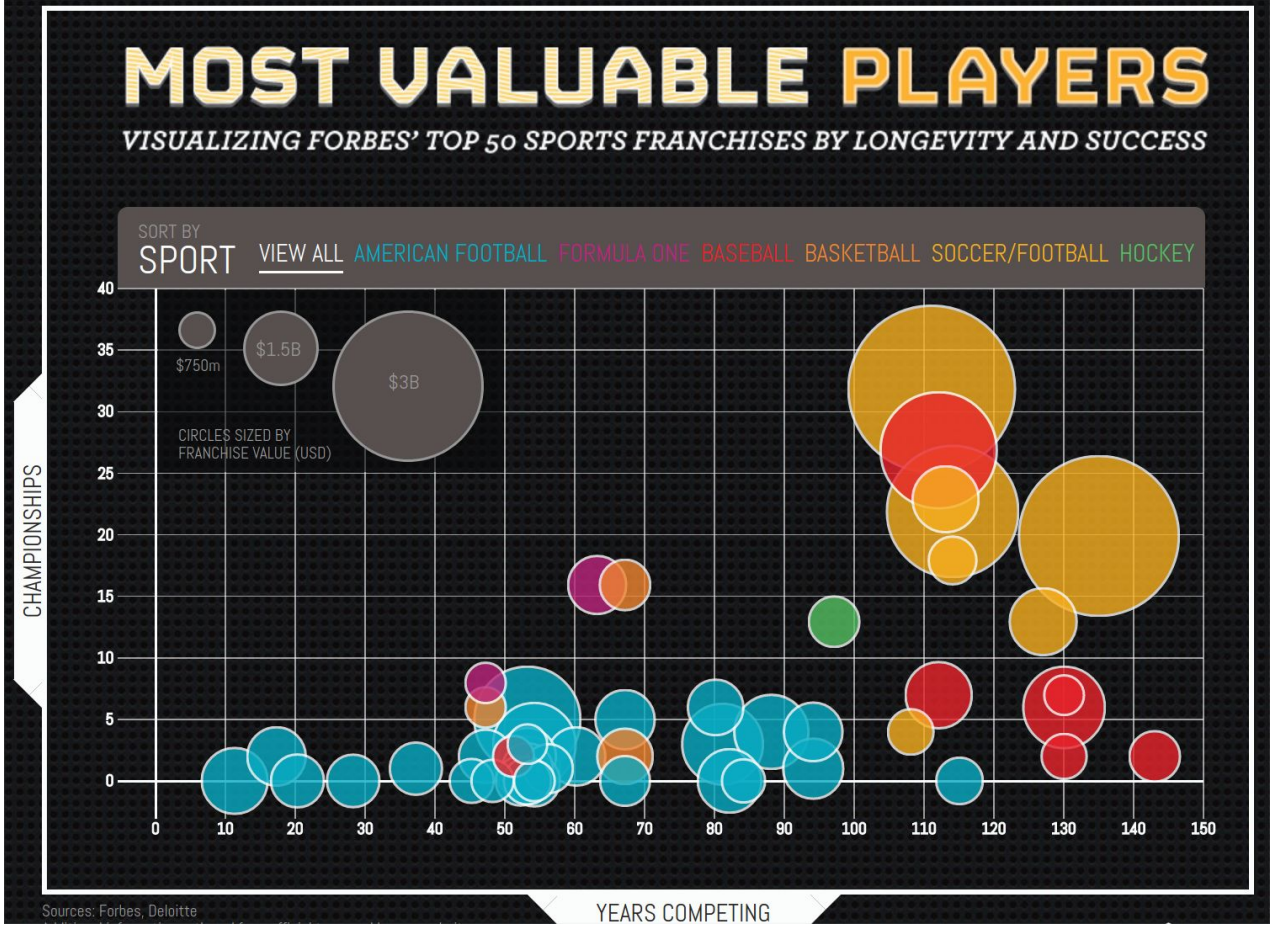
Good Data Visualisations



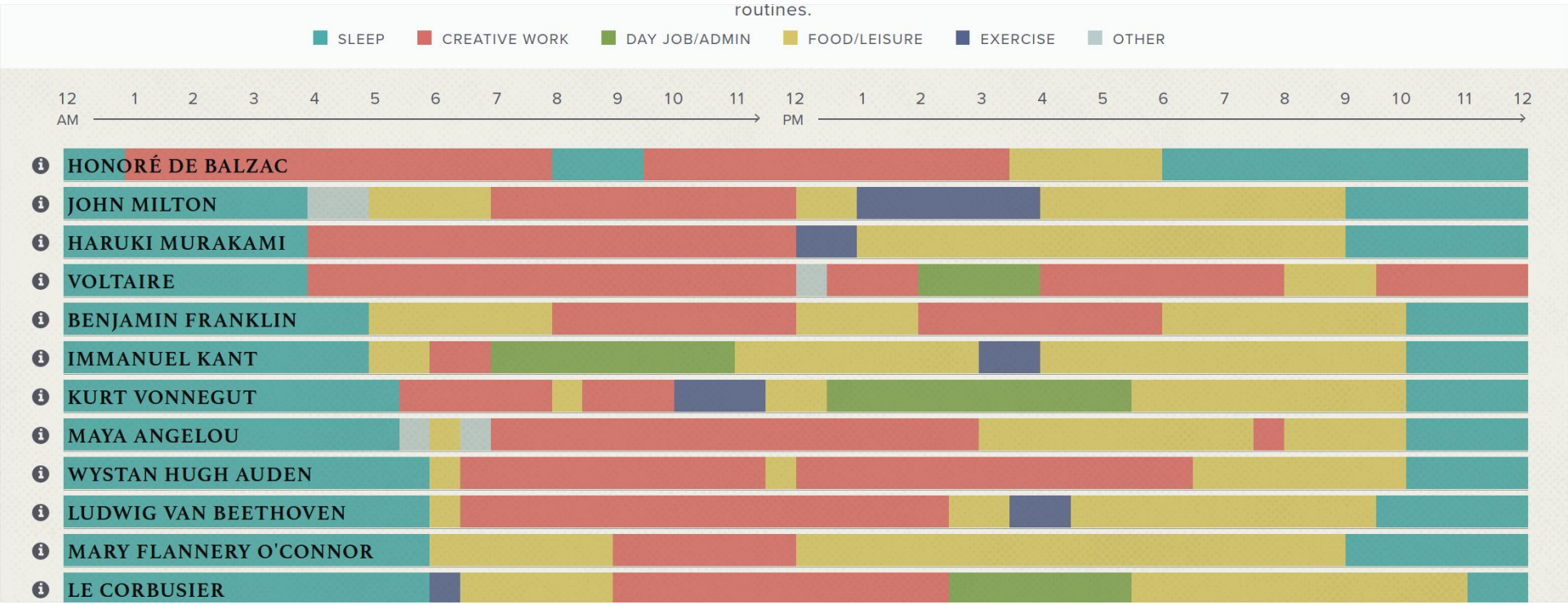
http://selfiecity.net/#selfiexploratory



https://setosa.io/bus/



http://mvp.columnfivemedia.com/



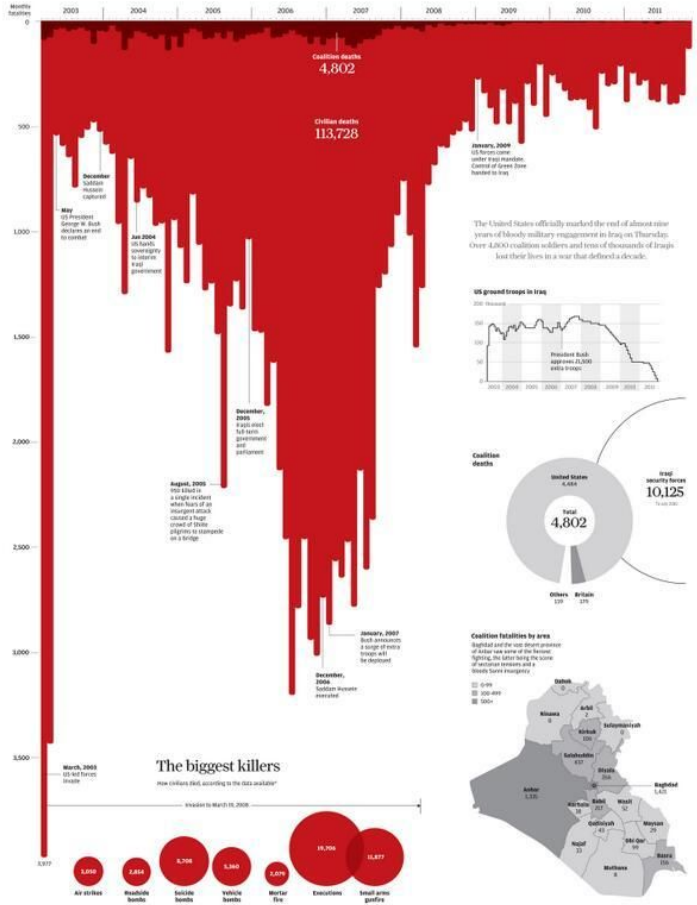
https://podio.com/site/creative-routines

Austria Solar Energy Report

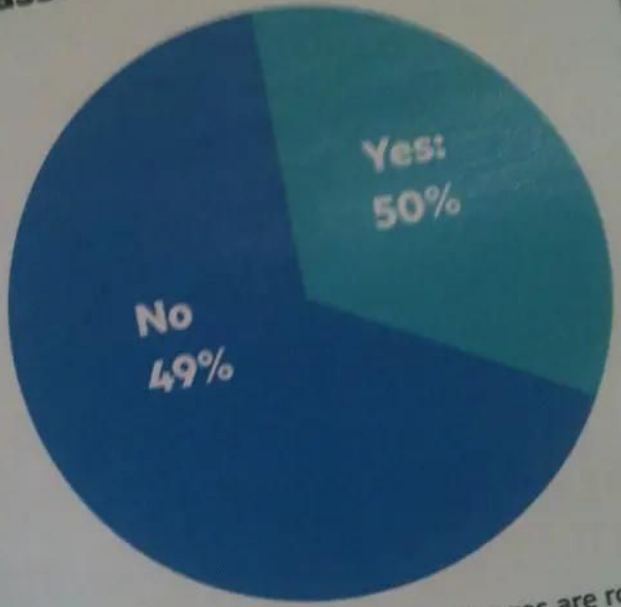
https://www.behance.net/gallery/2986075/The-Solar-Annual-Report-powered-by-the-sun

HOT OR NOT

Iraq's bloody toll



Last week, we asked whether A levels are becoming harder to pass. You said:

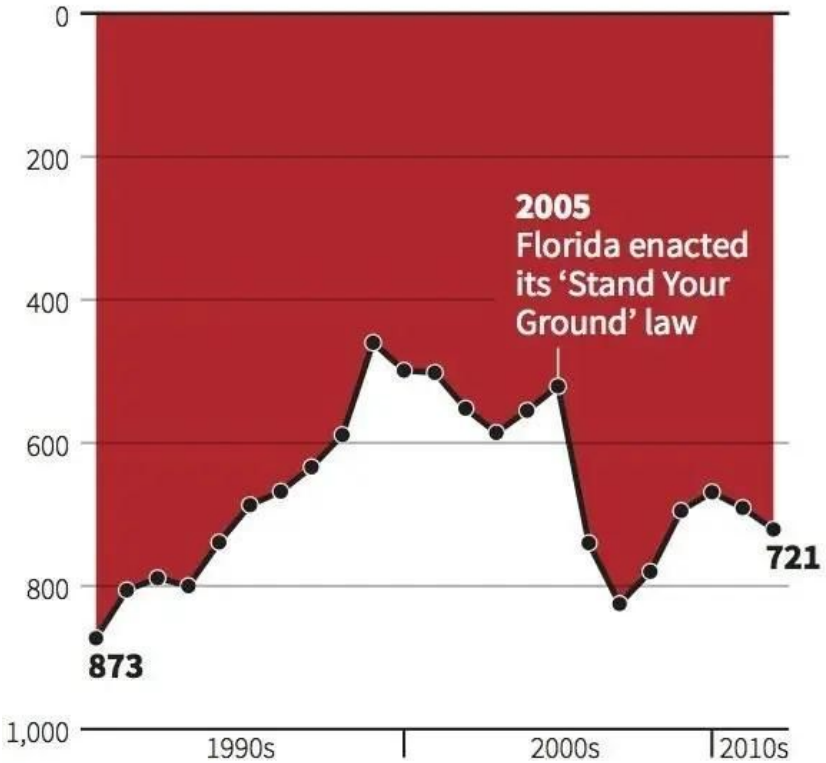


Figures are round

This week's question is:

Gun deaths in Florida

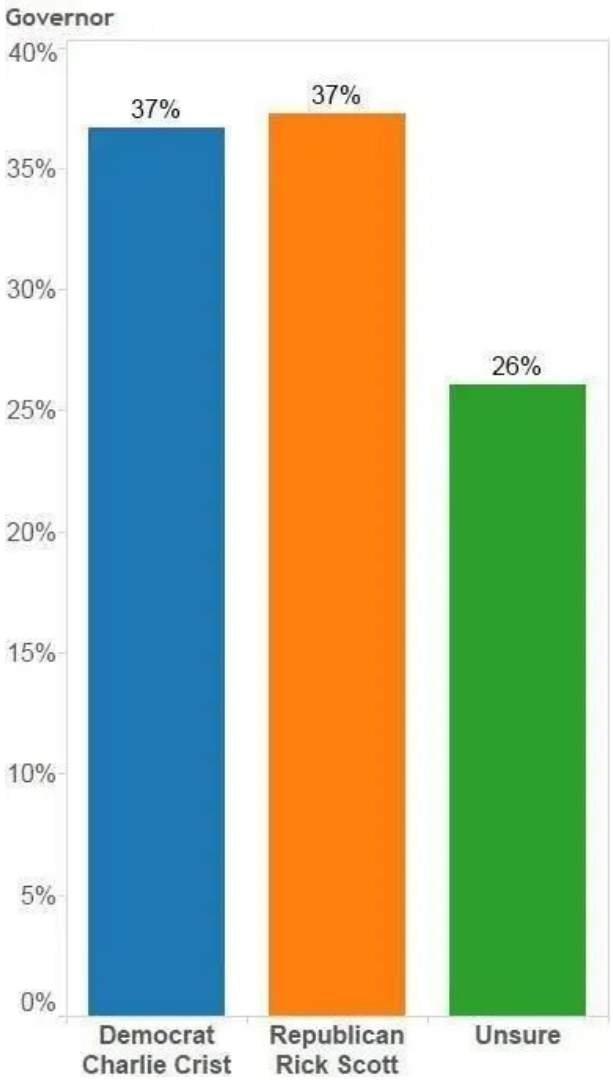
Number of murders committed using firearms

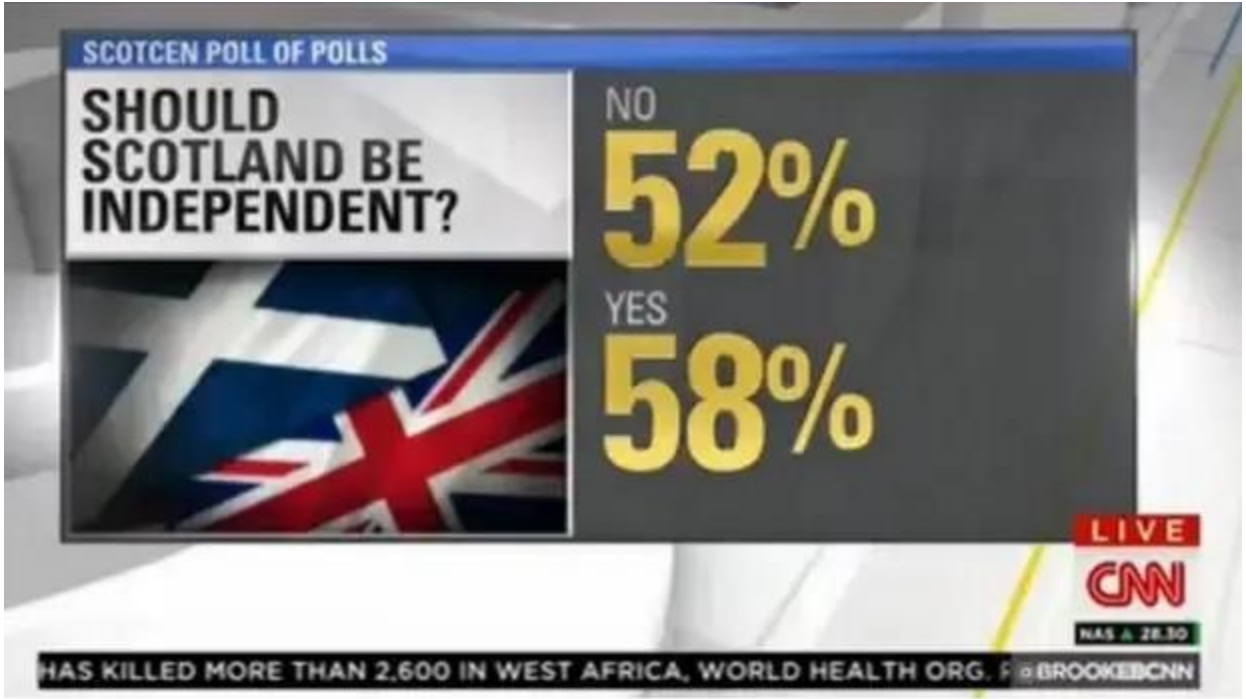


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

REUTERS


Governor
40%
35%
30%
25%
20%
15%
10%
5%
0%
37%
37%
26%
Democrat
Charlie Crist
Republican
Rick Scott
Unsure


SCOTCEN POLL OF POLLS
SHOULD SCOTLAND BE INDEPENDENT?
NO
52%
YES
58%
LIVE
CNN
HAS KILLED MORE THAN 2,600 IN WEST AFRICA, WORLD HEALTH ORG.
BROOKEBCNN

talks should happen soon.

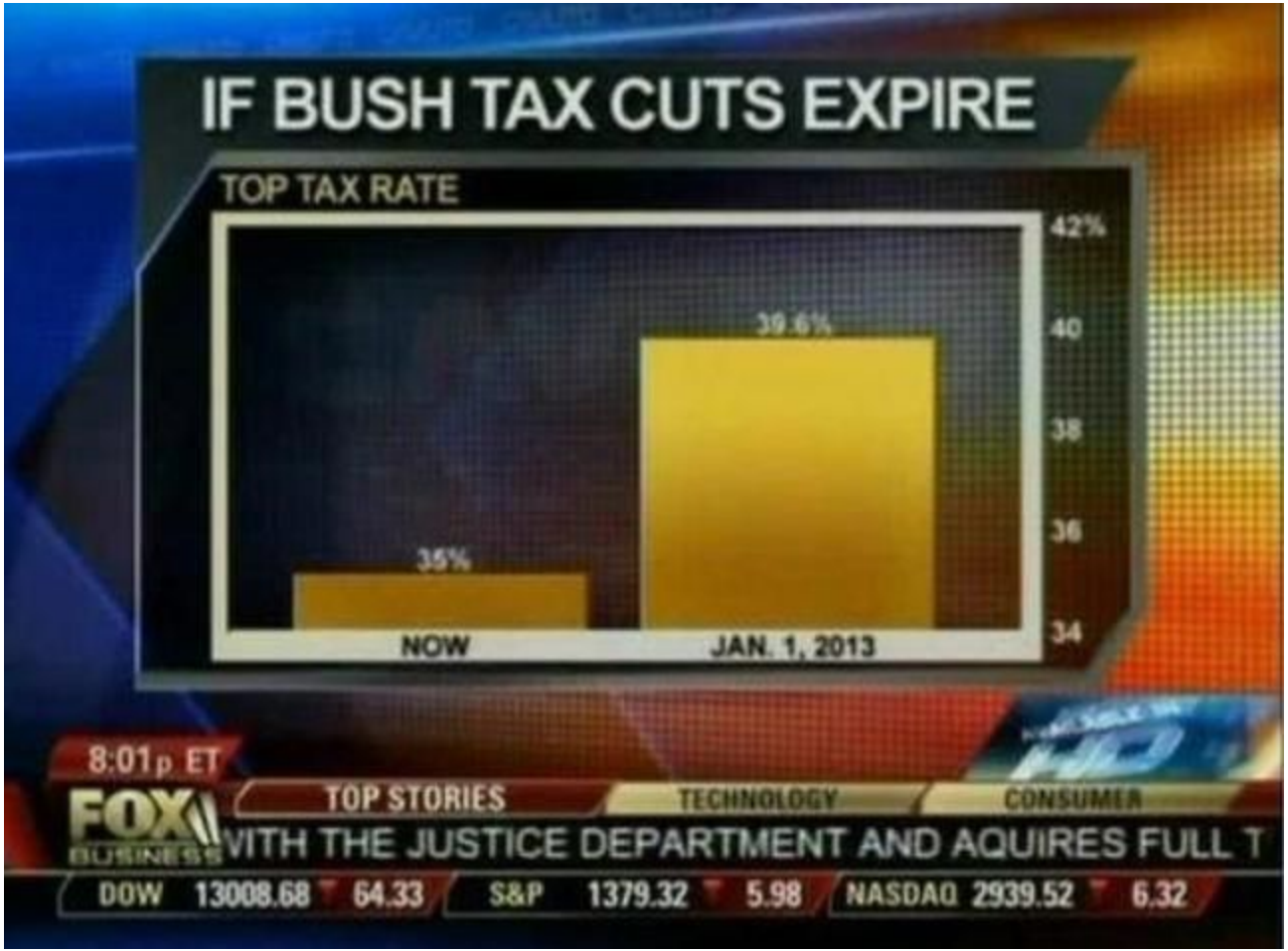
Question Of The Day

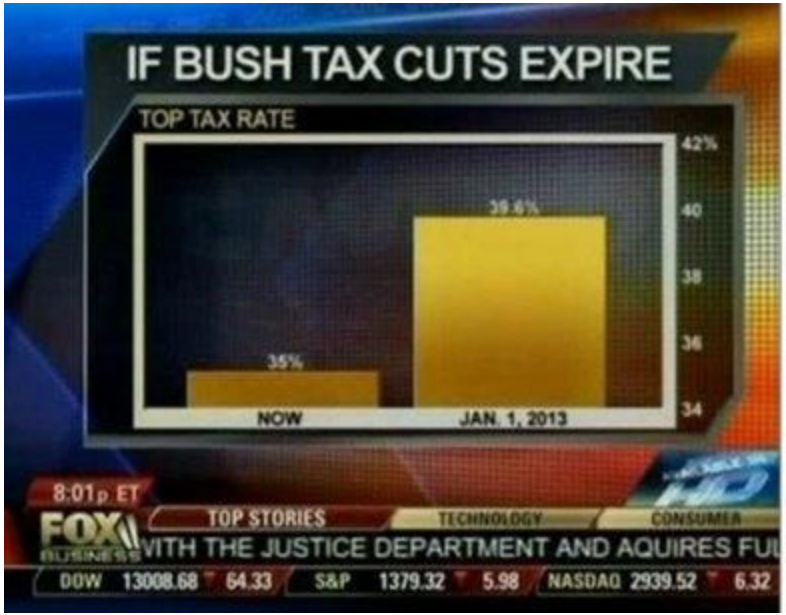
What should cost less: a gallon of gas or a gallon of milk?

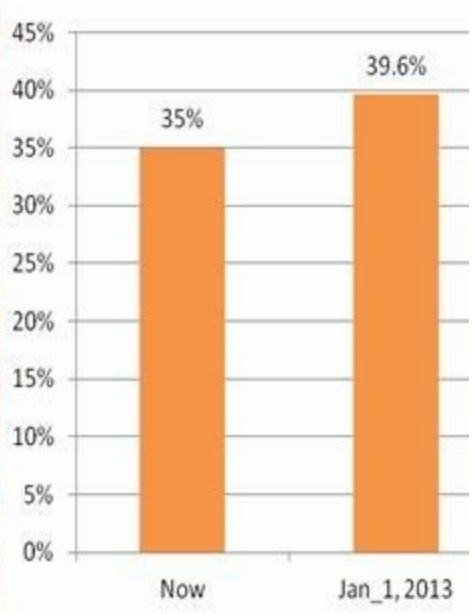
YES
43%

NO
57%

To record your answer to the Question of the Day and other polls about today's news, visit www.post-journal.com.

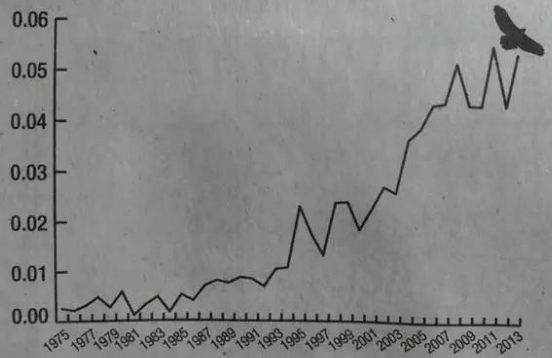

IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
VITH THE JUSTICE DEPARTMENT AND AQUIRES FU
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32


45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
39.6%
35%
Now
Jan_1, 2013

Cooper's hawk population soars

A deadly combination of shootings and a pesticide, DDT, caused the Cooper's hawk population in Illinois to stay at low levels throughout the 20th century. However, over the past few years, the raptor has made a strong comeback.



SOURCE: NATURAL HISTORY SURVEY

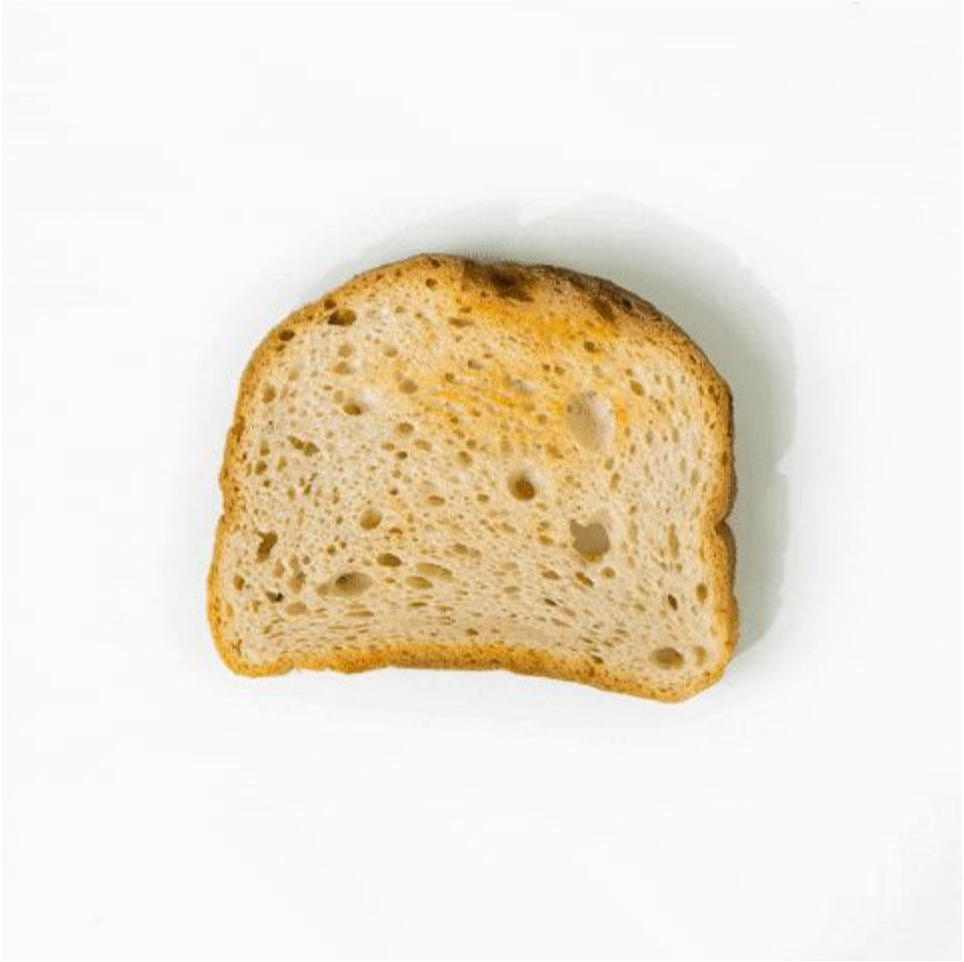
AUSTIN BAIRD THE DAILY ILLINI


Quote
“Only half of programming is coding. The other 90% is debugging.”
Anonymous


РОССИЯ 24
ПРЯМОЙ ЭФИР
ПО ДАННЫМ ЦИК
Ростовская область
1 Единая Россия 58,99%
2 КПРФ 32,96%
3 ЛДПР 23,74%
4 Справедливая Россия 19,41%
5 Яблоко 9,32%
6 Патриоты России 1,46%
7 Правое дело 0,59%
В Госдуму пройдут партии, набравшие более 7% голосов
22:58 ЛУКОЙЛ ММВБ 1704,00 -0,06%

Setup Google Collab


plotly




Dataset
1248
Avocado Prices
Historical data on avocado prices and sales volume in multiple US markets
Justin Kiggins • updated 2 years ago (Version 1)
Data
Tasks
Kernels (167)
Discussion (10)
Activity
Metadata
Download (2 MB)
New Notebook

Relevant columns: Date

- **Date** - The date of the observation
  - First: 2015-01-04
  - Last: 2018-03-25

Date
2015-12-27
2015-12-20
2015-12-13
2015-12-06
2015-11-29
2015-11-22
2015-11-15

Relevant columns: AveragePrice

- AveragePrice - the average price of a single avocado
  - Median: 1.37
  - Mean: 1.40
  - std: 0.40
  - min: 0.44
  - max: 3.2

AveragePrice
1.33
1.35
0.93
1.08
1.28
1.26
0.99

Relevant columns: Type

- **Type** - conventional or organic
  - Equal distribution of both types

type
organic
conventional
organic
conventional
organic
organic
conventional

QUIZ: What type of data is column ‘type’

Relevant columns: region

- **Region** - the city or region of the observation
  - 54 unique values

region
Chicago
PhoenixTucson
RaleighGreensboro
Southeast
Midsouth
WestTexNewMexico

Relevant columns: Total Volume

- **Total Volume** - Total number of avocados sold
  - mean: 850644.0
  - median: 107376.76
  - std: 3453545.0
  - min: 84.56
  - max : 2505650.0

Relevant columns: 4225

- **4225** - Total number of avocados with PLU 4225 sold

Relevant columns: 4770

- **4770** - Total number of avocados with PLU 4770 sold

Relevant columns: 4046

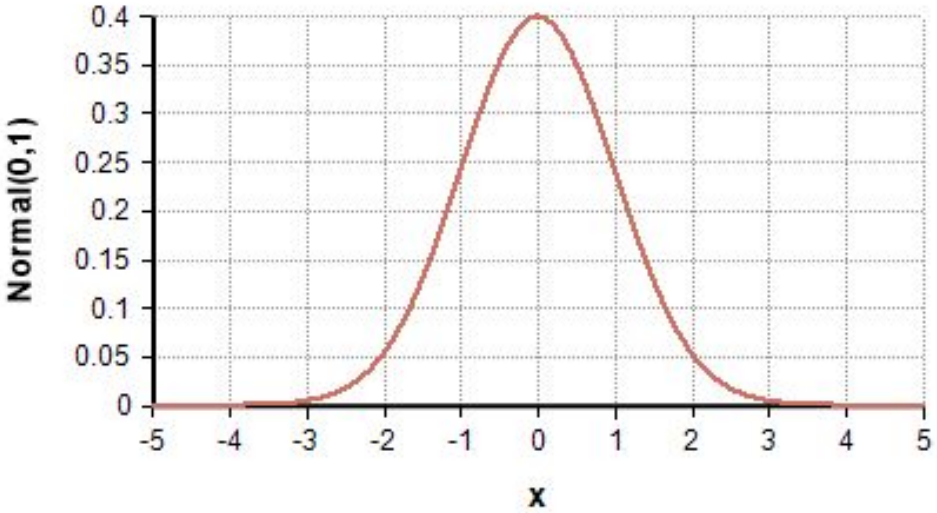
- **4046** - Total number of avocados with PLU 4046 sold


plotly

Plotly Express

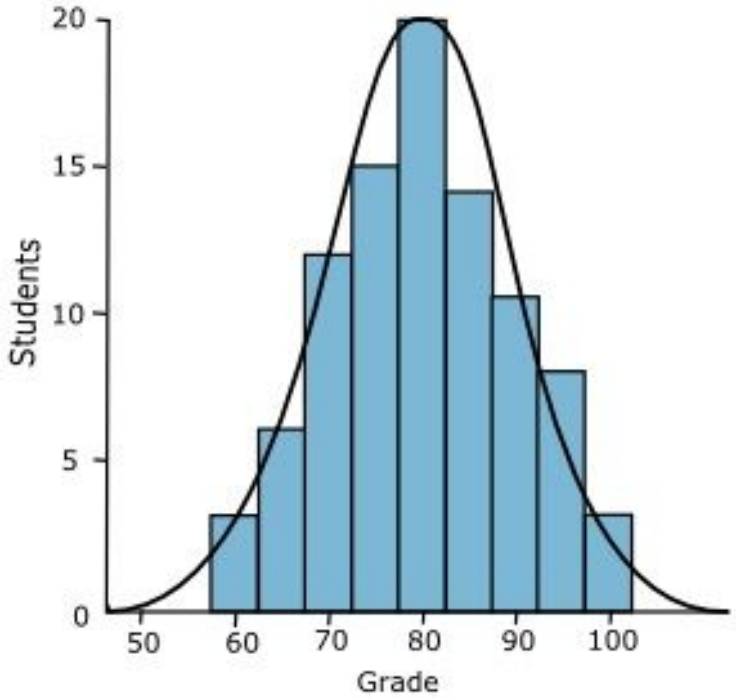
Plotly Express is the easy-to-use, high-level interface to Plotly, which operates on "tidy" data and produces easy-to-style figures.

Remember this?

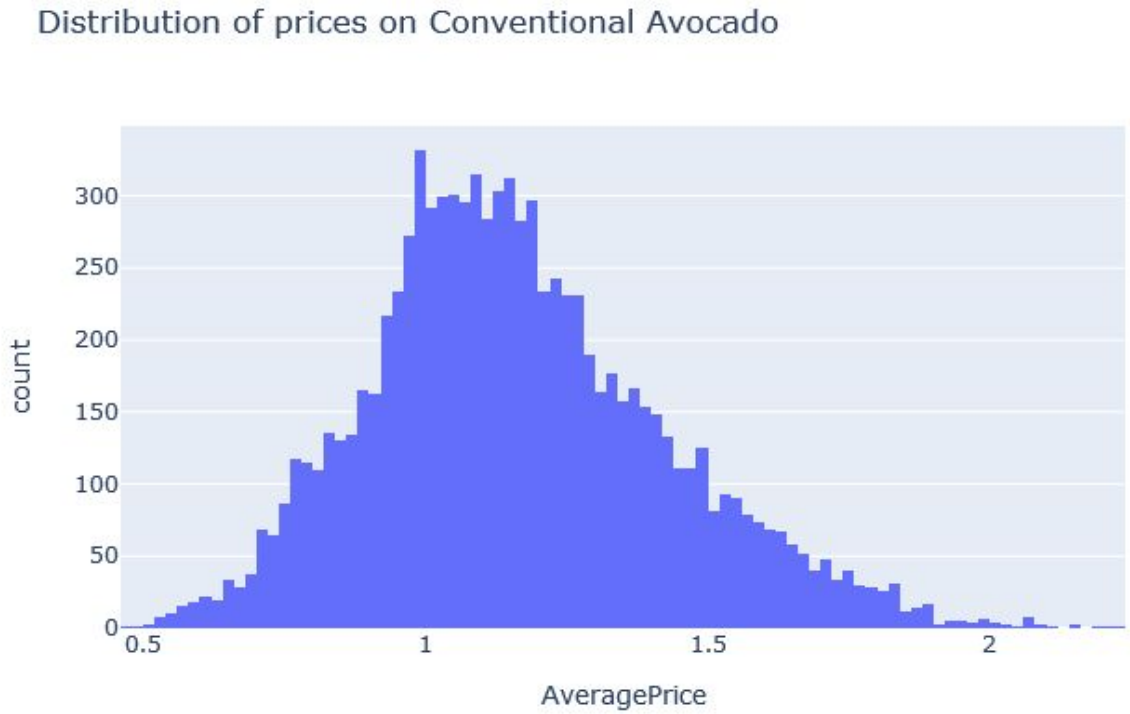


Histogram

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.



Histogram of Conventional Avocado Prices

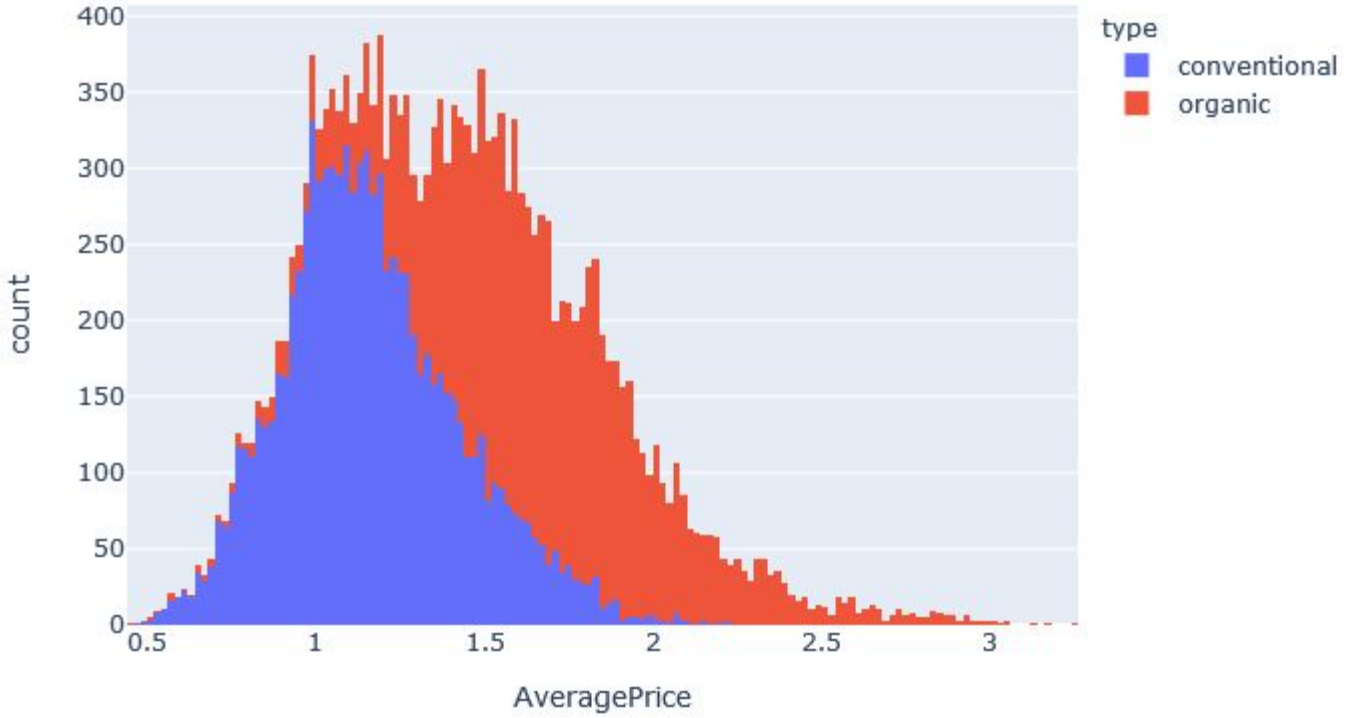


DEMO

What is price range of the highest count? (organic)

1.58-1.59

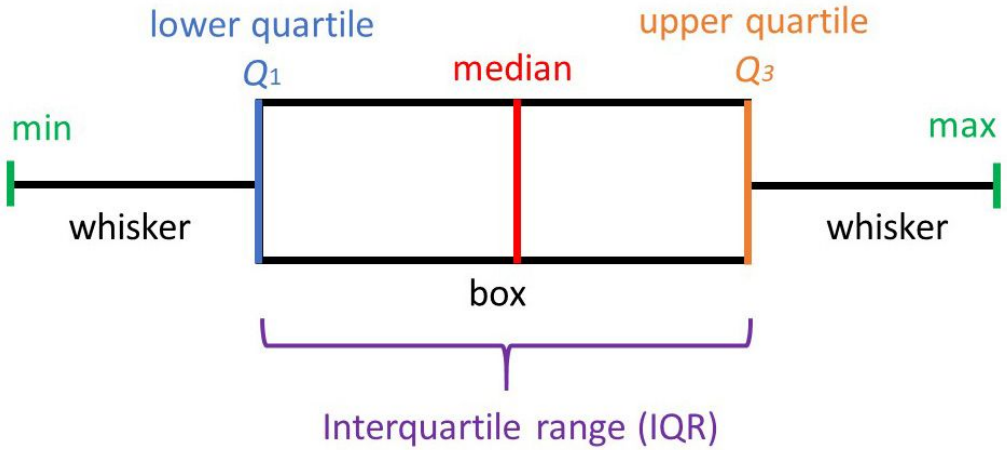
Comparing two histograms

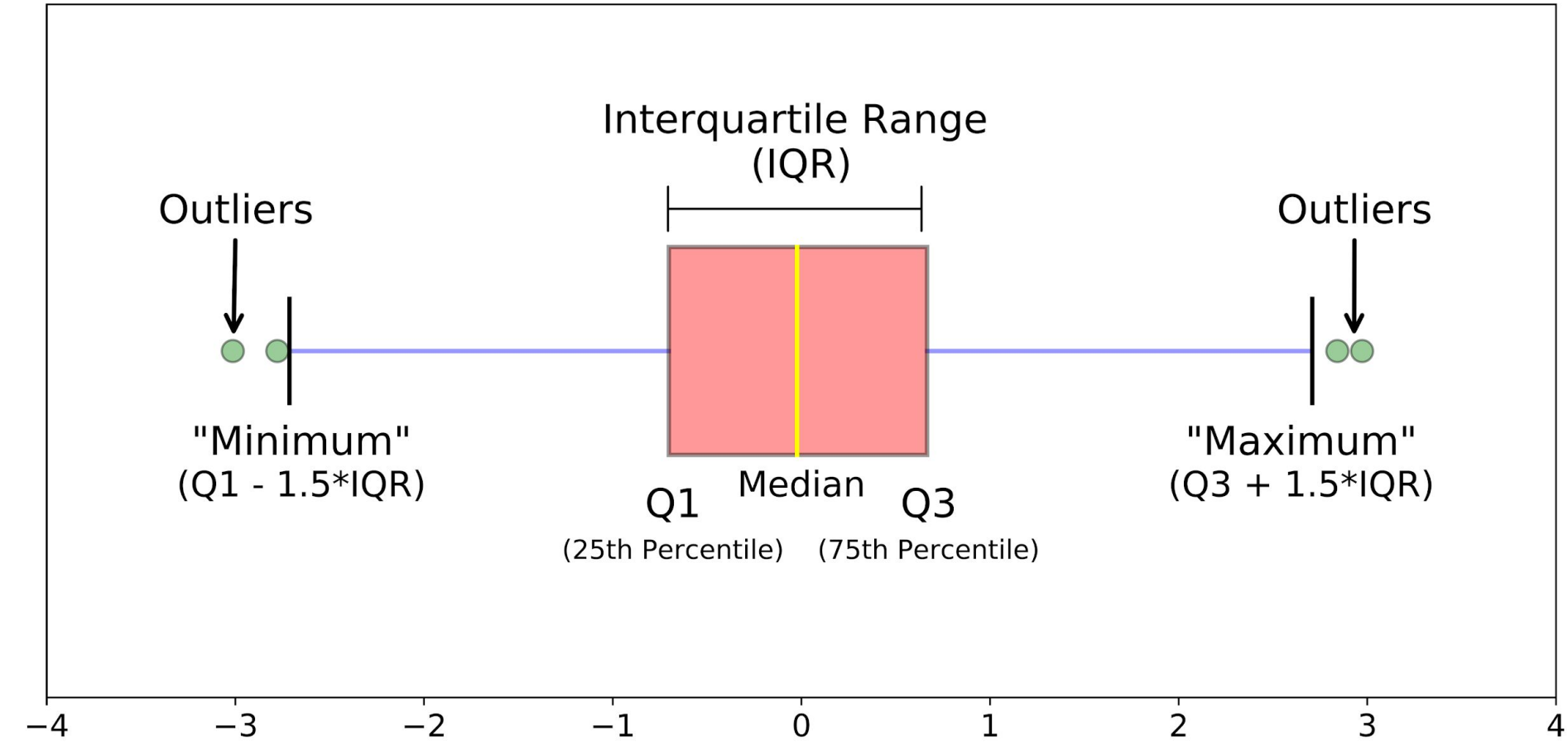


DEMO

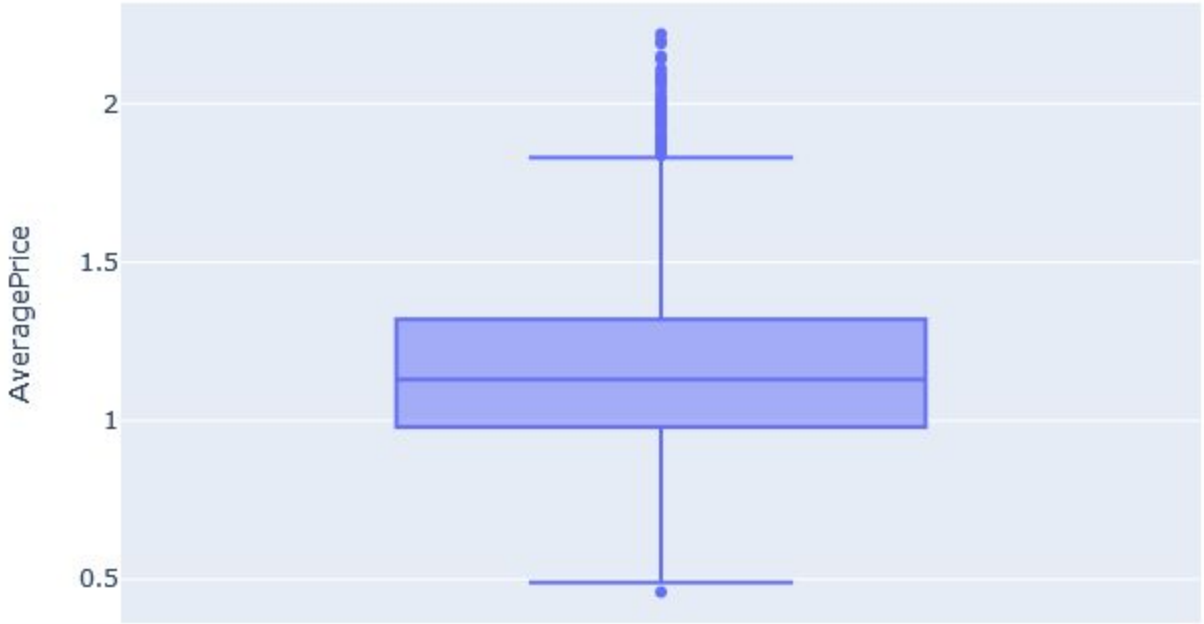
Box Plot

A box plot displays the five-number summary of data set. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.




Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

Box Plot of Conventional Avocado Prices



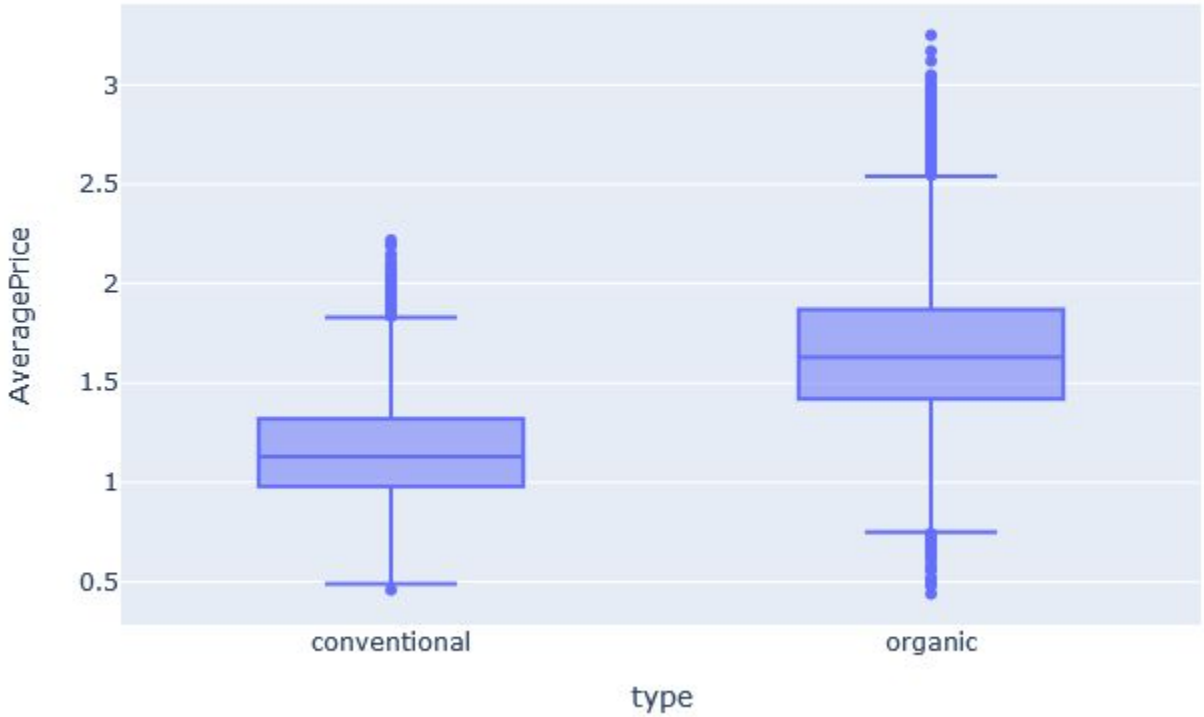
DEMO

Quiz Plot box plot for organic avocado

What is value of upper fence?

Answer: 2.54

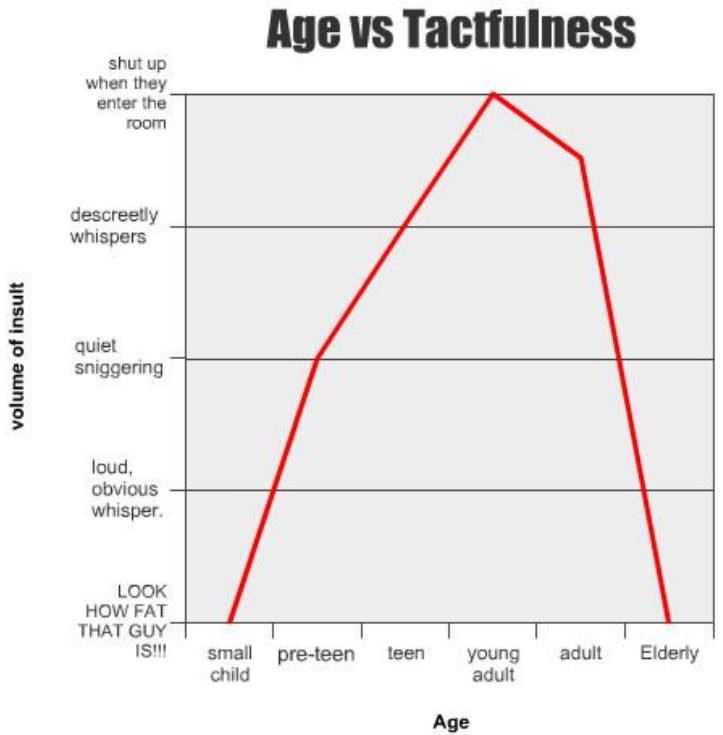
Box Plot for each type of Avocado



DEMO

Line Plot

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.



Average Price of Conventional Avocado Over Time

DEMO

Quiz: make organic plot. What is the date of the highest peak.

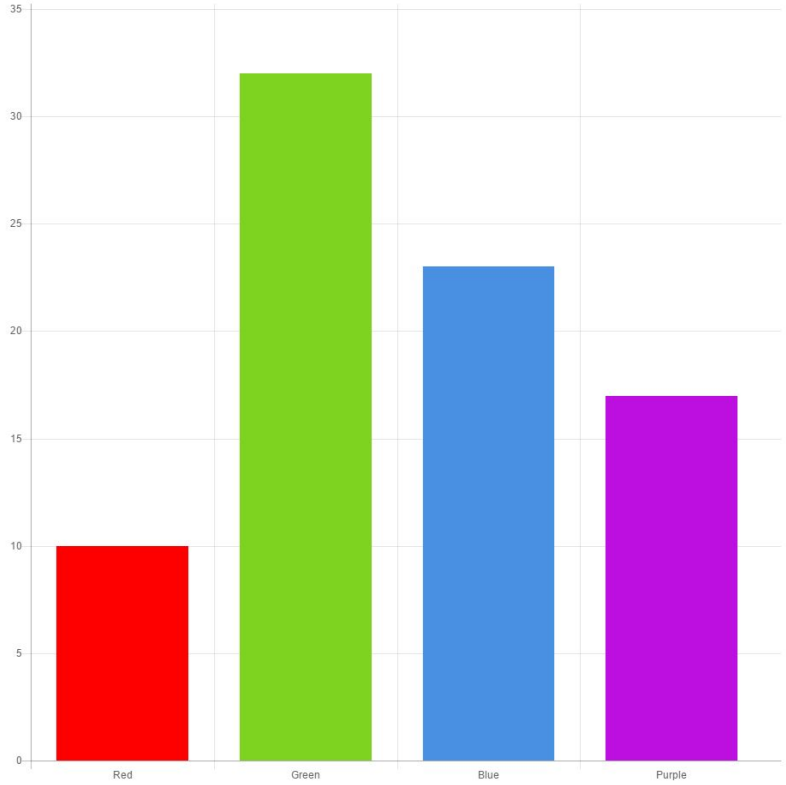
27. Aug 2017

Average Price of Avocado Over Time

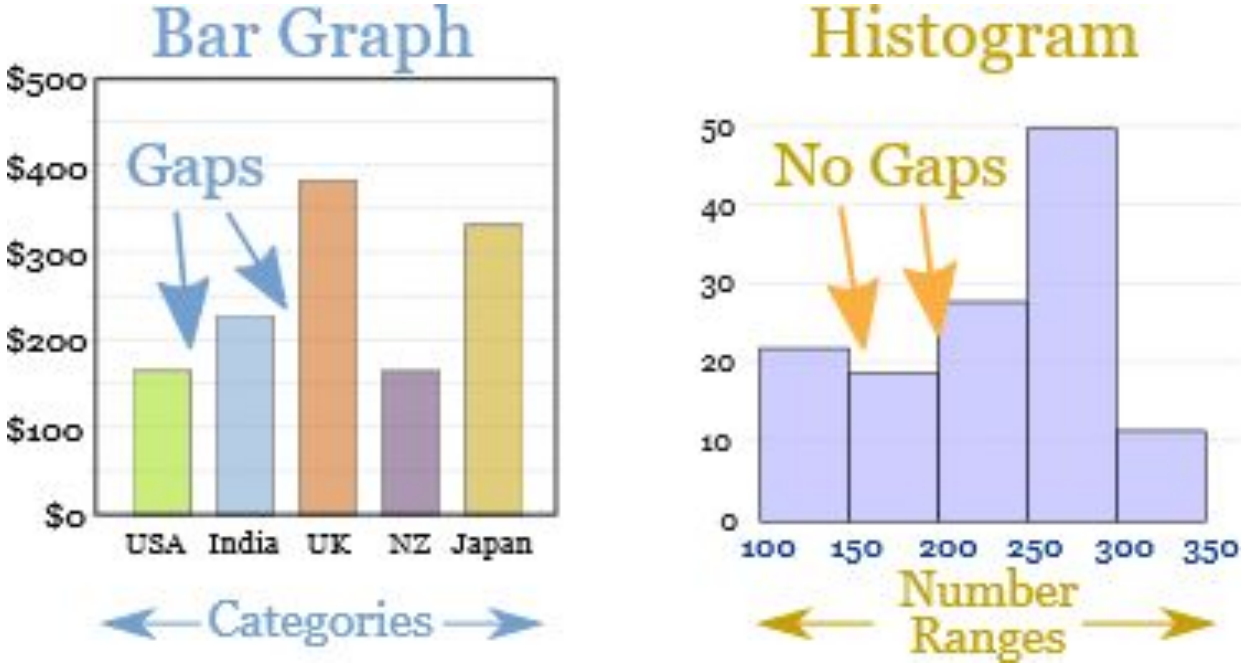
DEMO

Bar plot

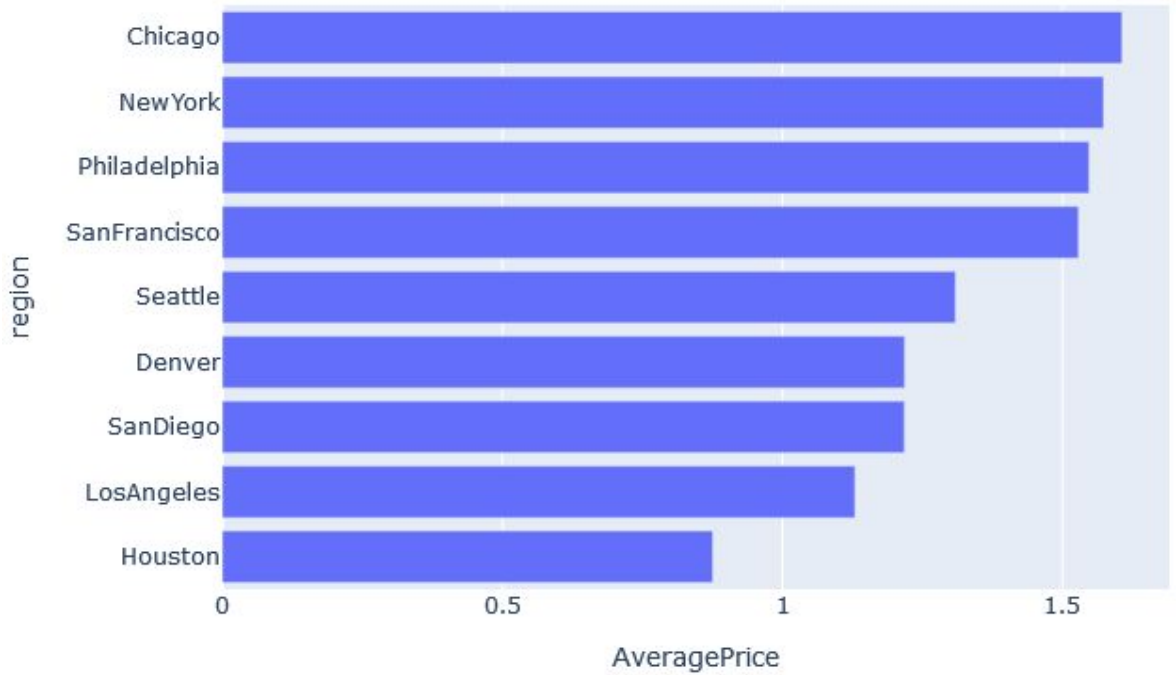
A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.



Bar Plot vs Histogram



Average Price of Conventional Avocado in 2017



DEMO

Quiz df_2017. How many rows?

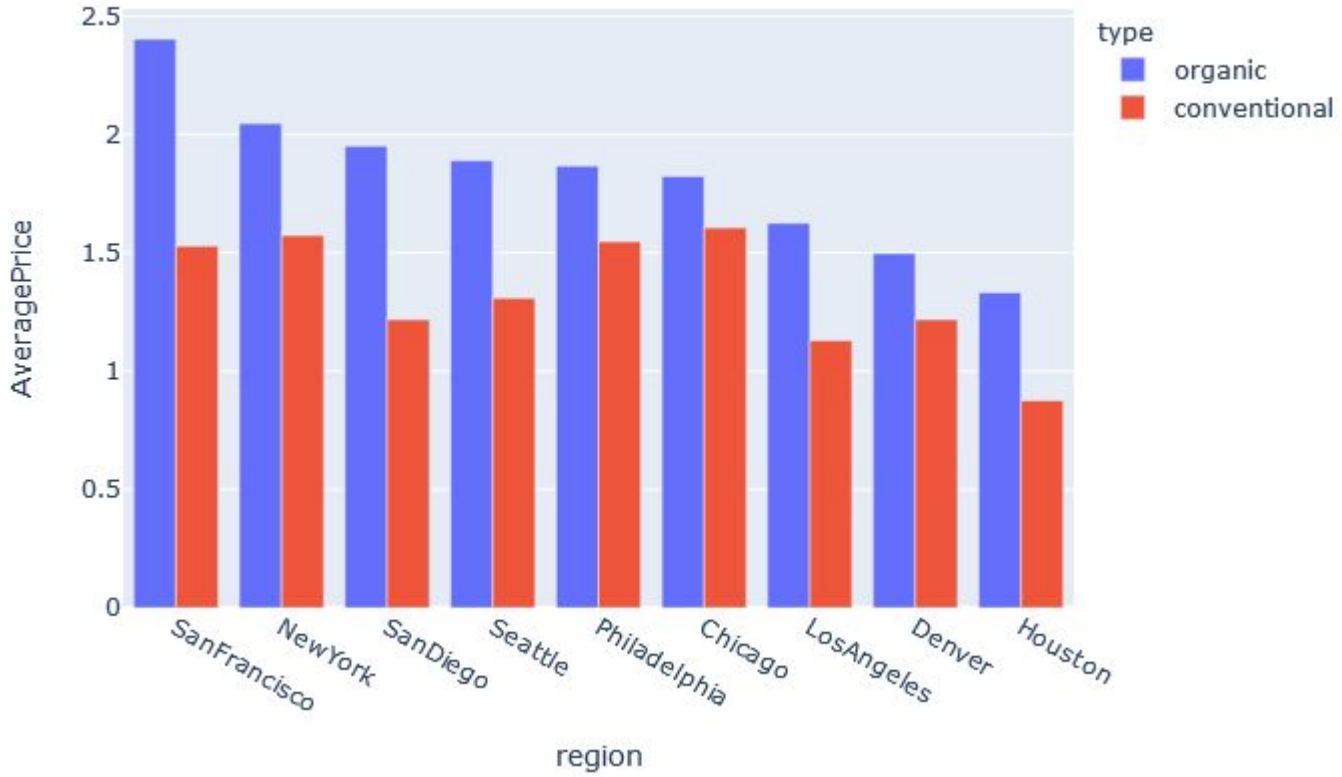
Answer: 5722

Groupby Exercice. **How many rows and columns does "df_2017_region_price" have?**

108,3

DEMO

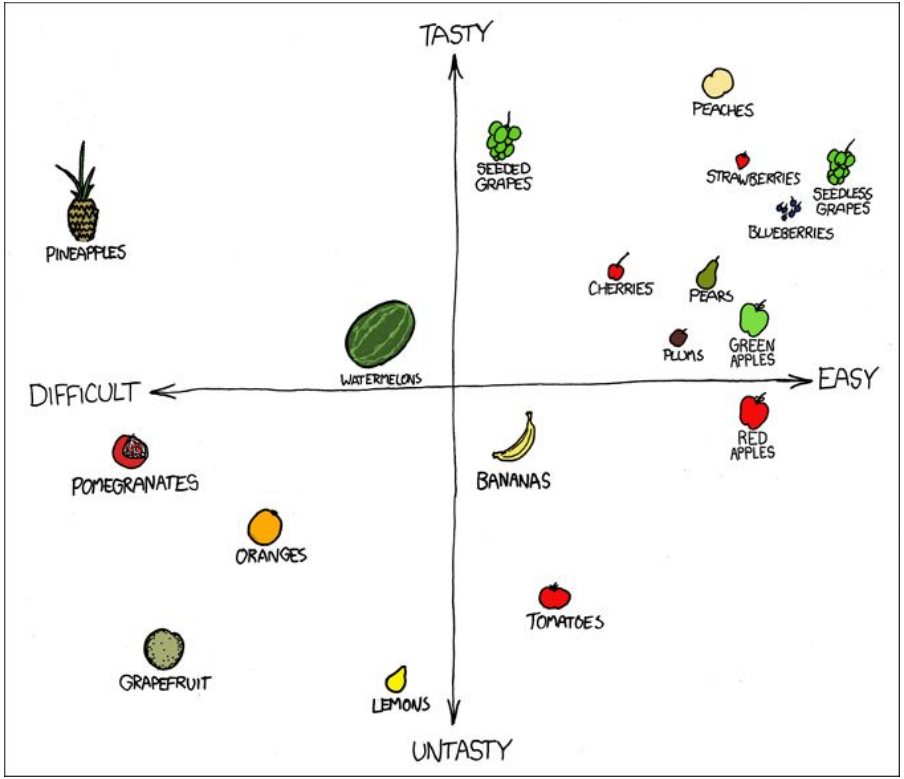
Average Price of Avocado in 2017



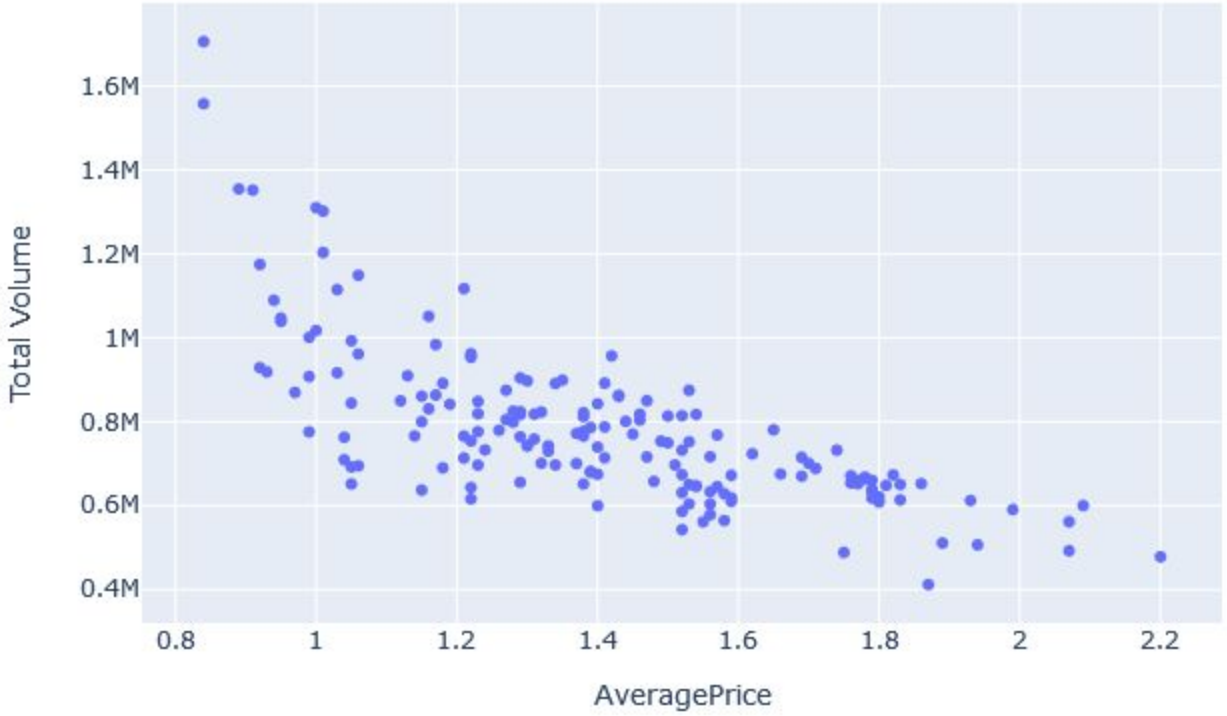
DEMO

Scatter Plot

A scatter plot uses coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.



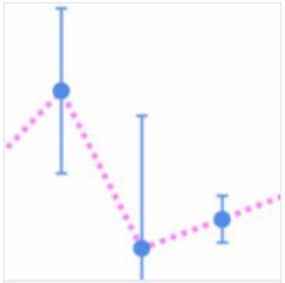
Correlation between Average Price and Total Volume



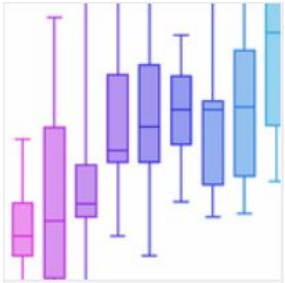
DEMO

Correlation in Houston and San Francisco

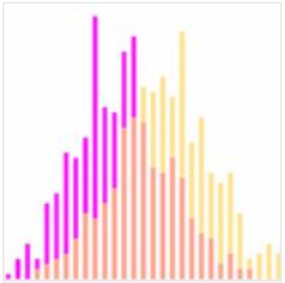
DEMO



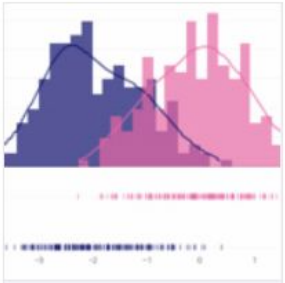
Error Bars



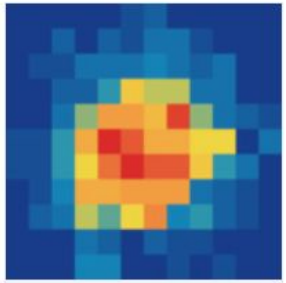
Box Plots



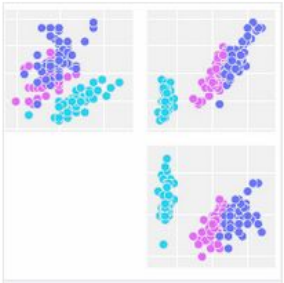
Histograms



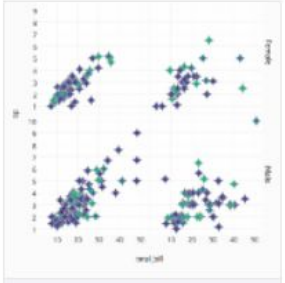
Distplots



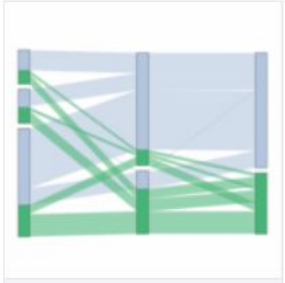
2D Histograms



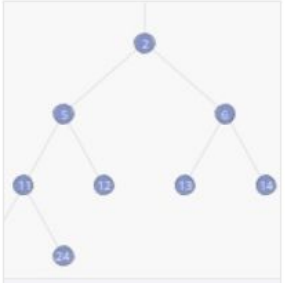
Scatterplot Matrix



Facet and Trellis Plots

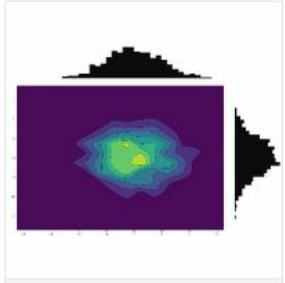


Parallel Categories Diagram

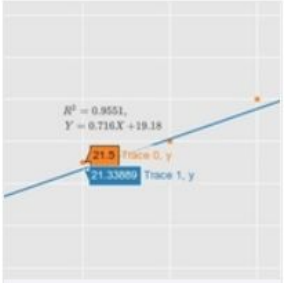


Tree-plots

Violin Plots



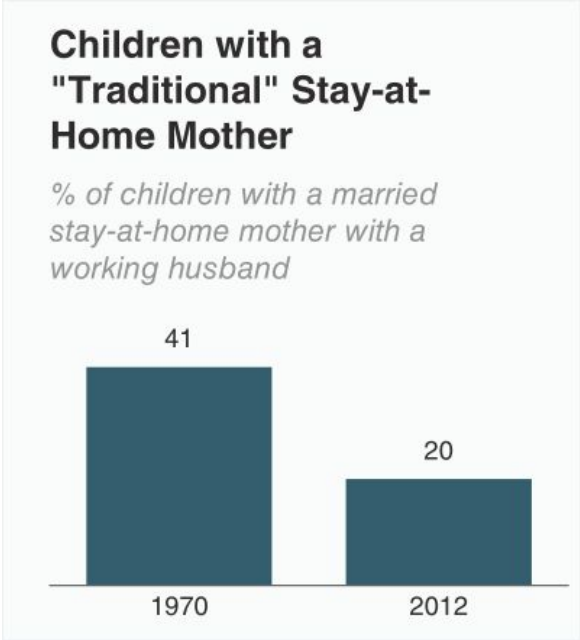
2D Histogram Contour



Linear and Non-Linear Trendlines

Simple text

simple text can be a great way to communicate



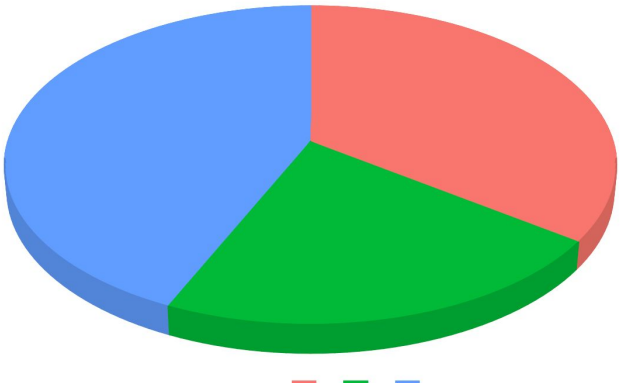
Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

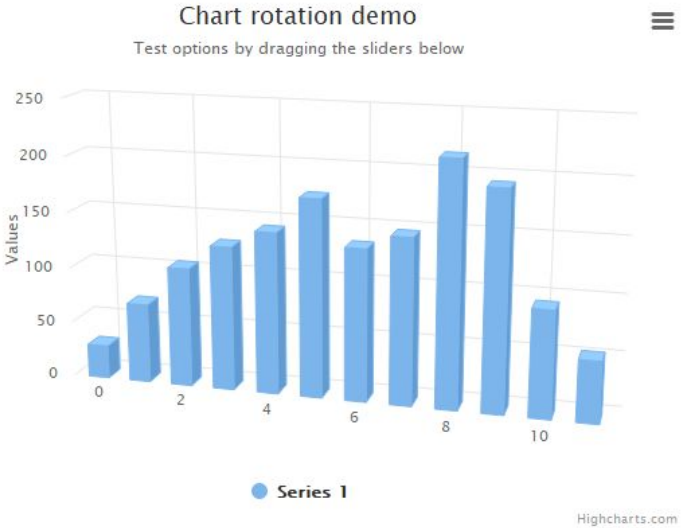
Simple text

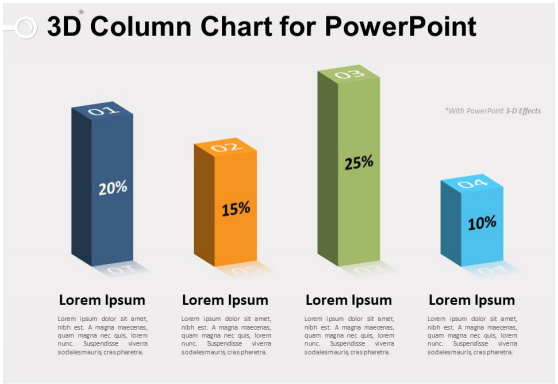
simple text can be a great way to communicate

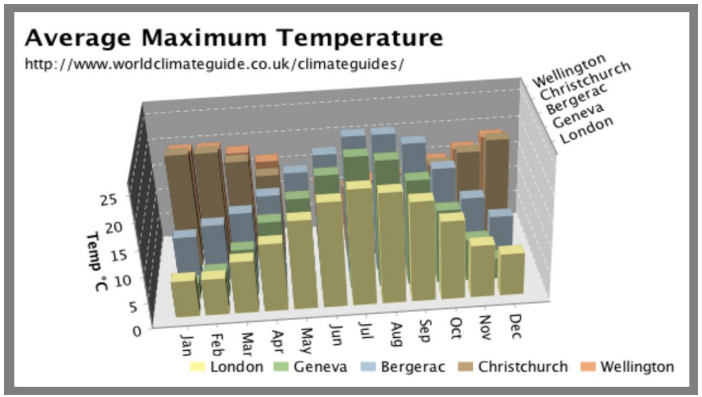
Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Never use 3D







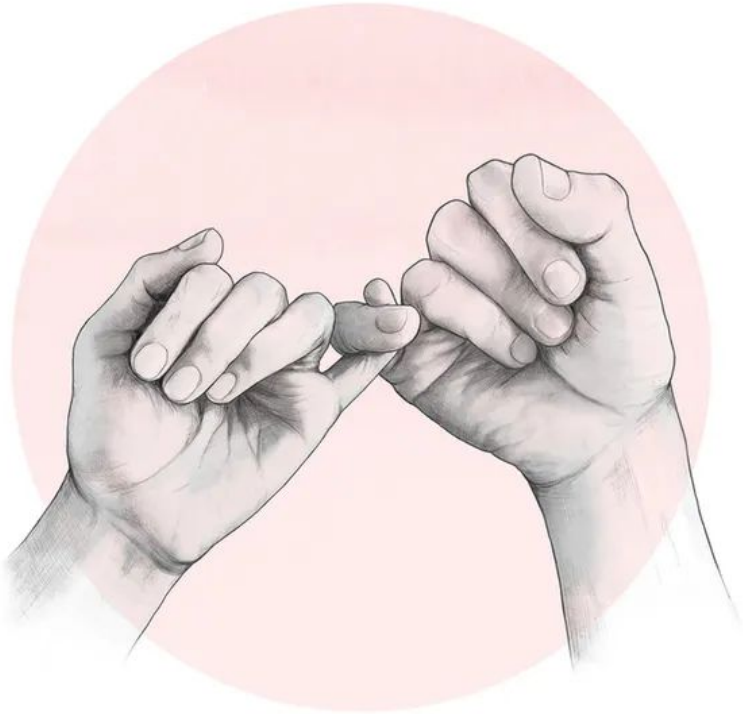


Never use 3D

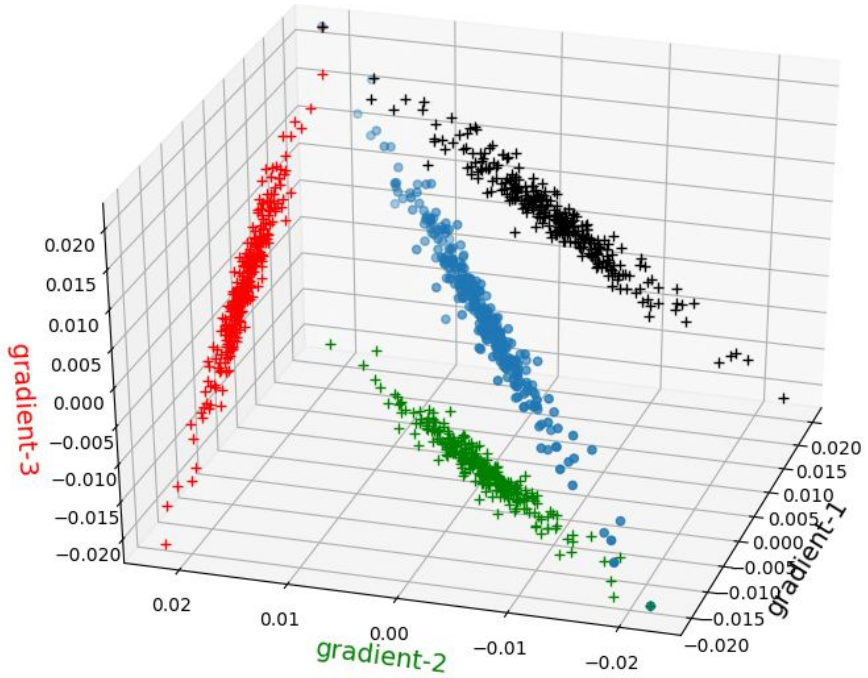
Repeat after me

Repeat after me

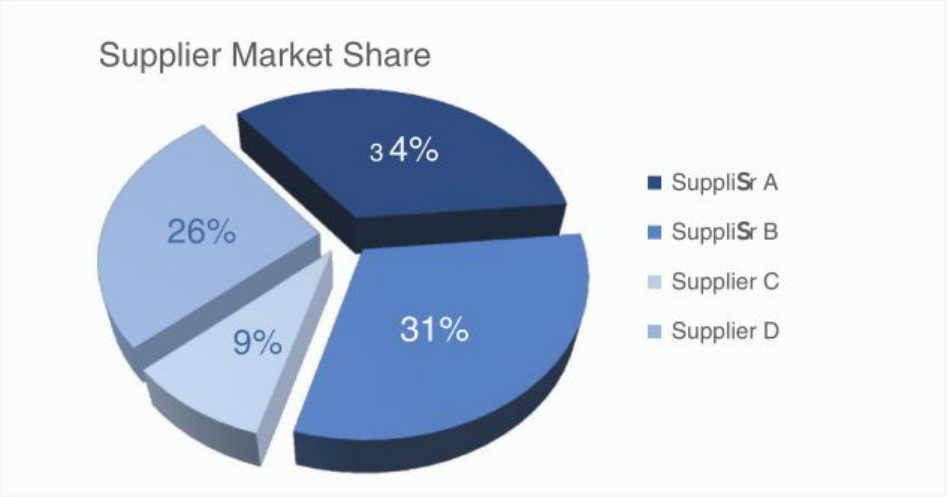
I will **never** use **3D Plots** to impress my managers.



Exception: 3d Scatter Plots



Bar plot as alternative to pie charts.



Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

The secret of good plots:

The secret of good plots:

Remove the clutter.

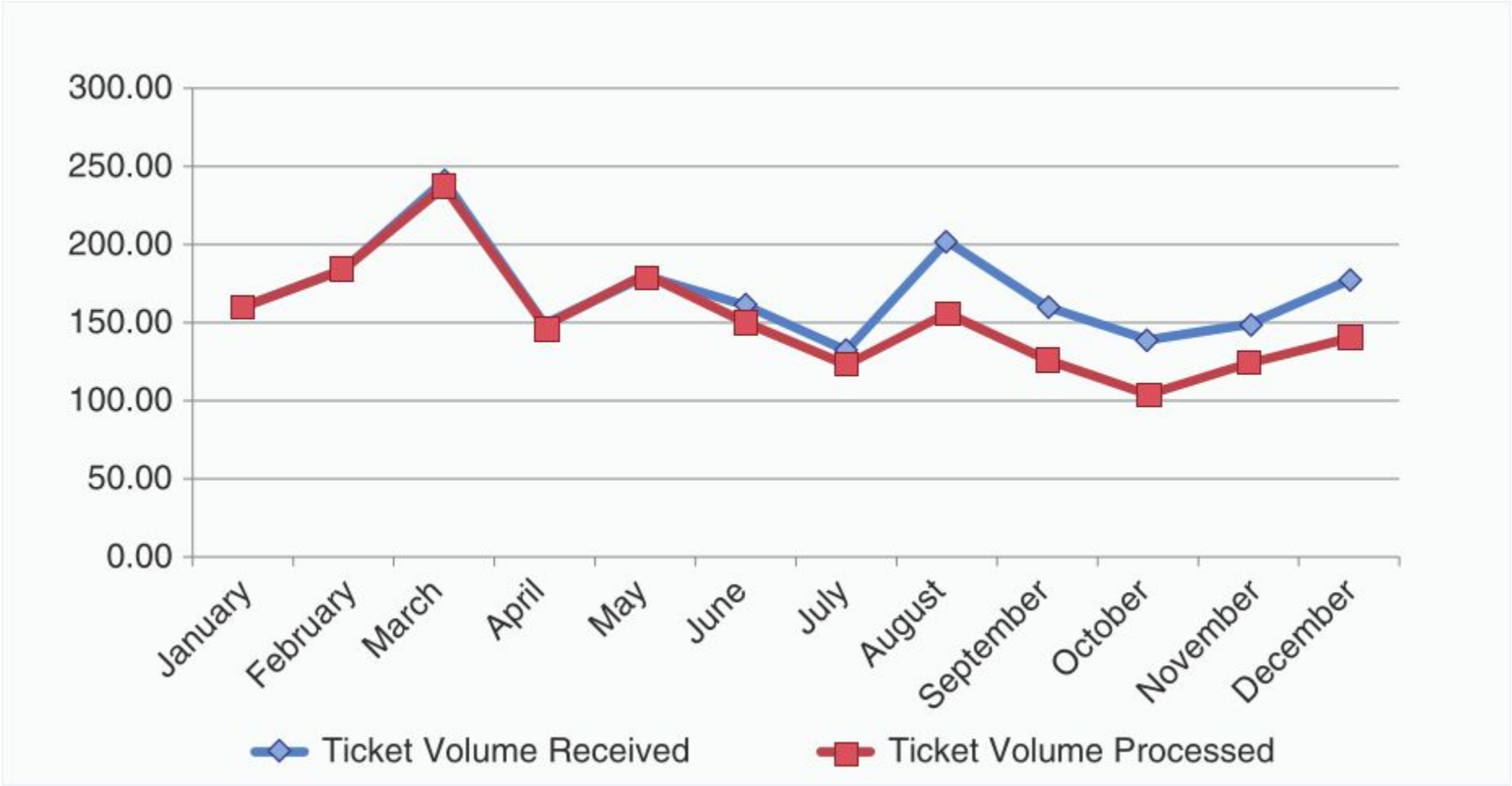
If you don’t remember anything
Remember just this.


“Discard everything that does not spark joy.”
MARIE KONDO
Your tango

Step 1

Remove what has no purpose

Step 1
Remove what has no purpose

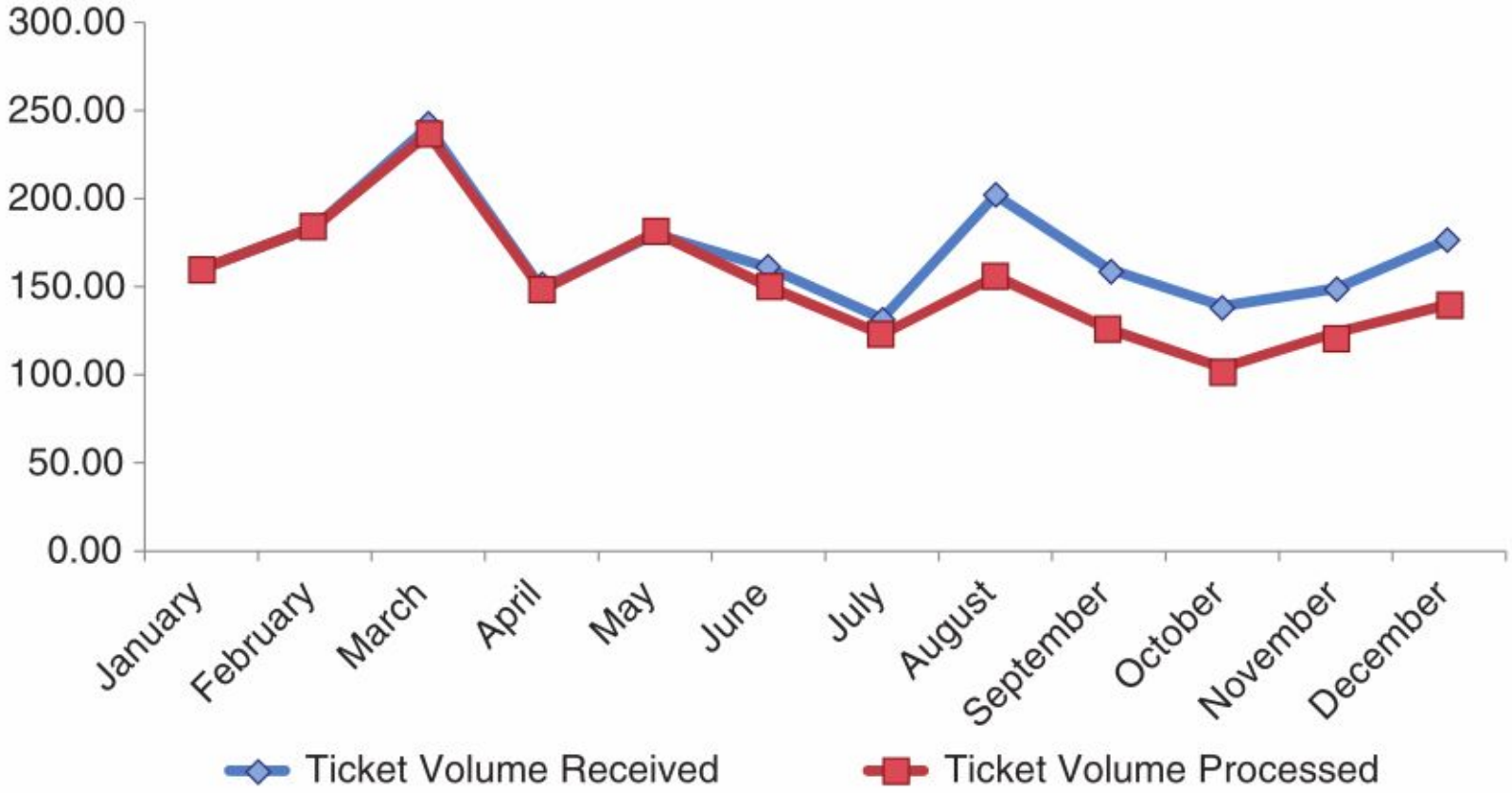


Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

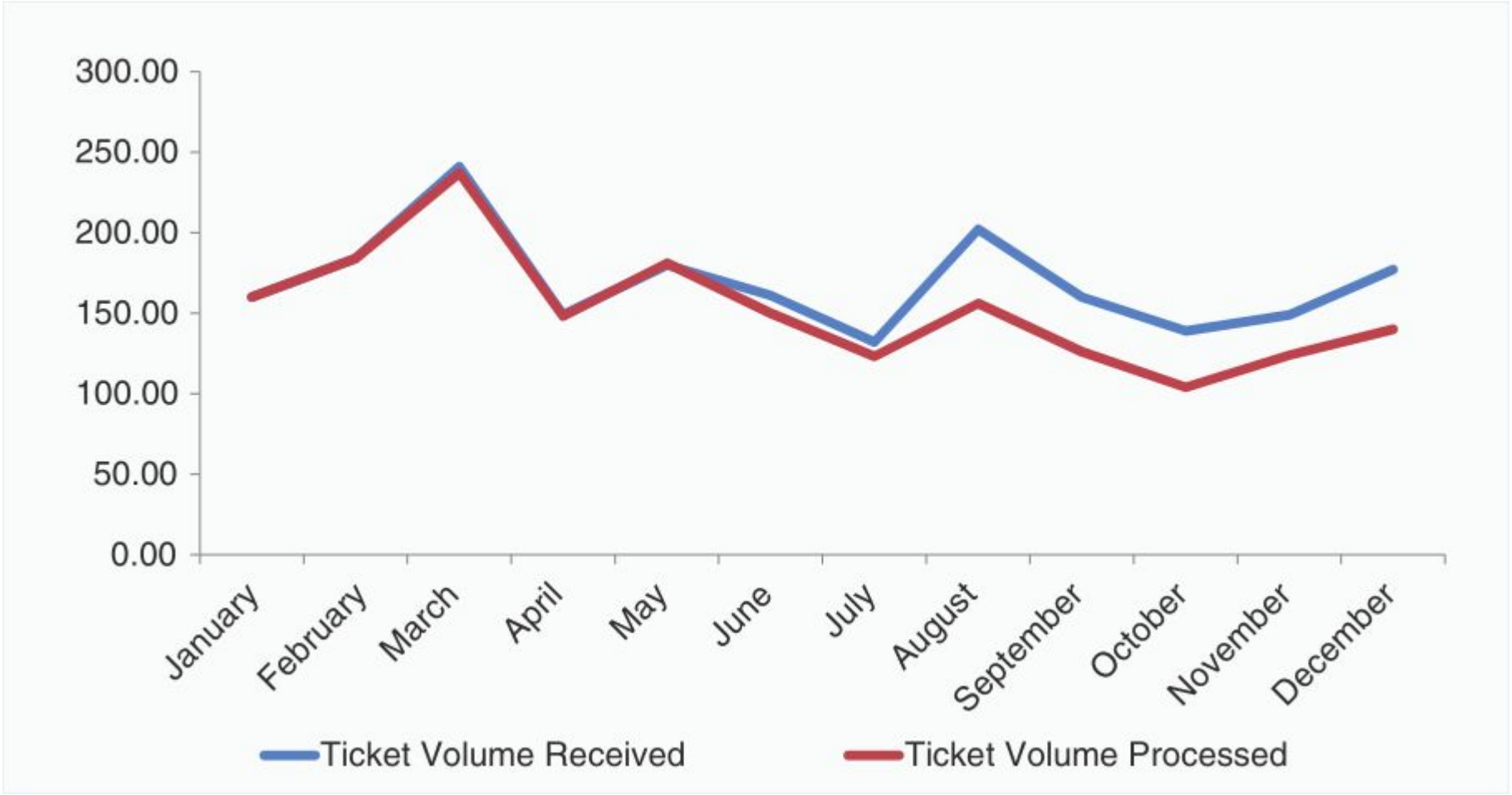
Gridlines?

Do they help the viewer?

No Gridlines = Better Contrast

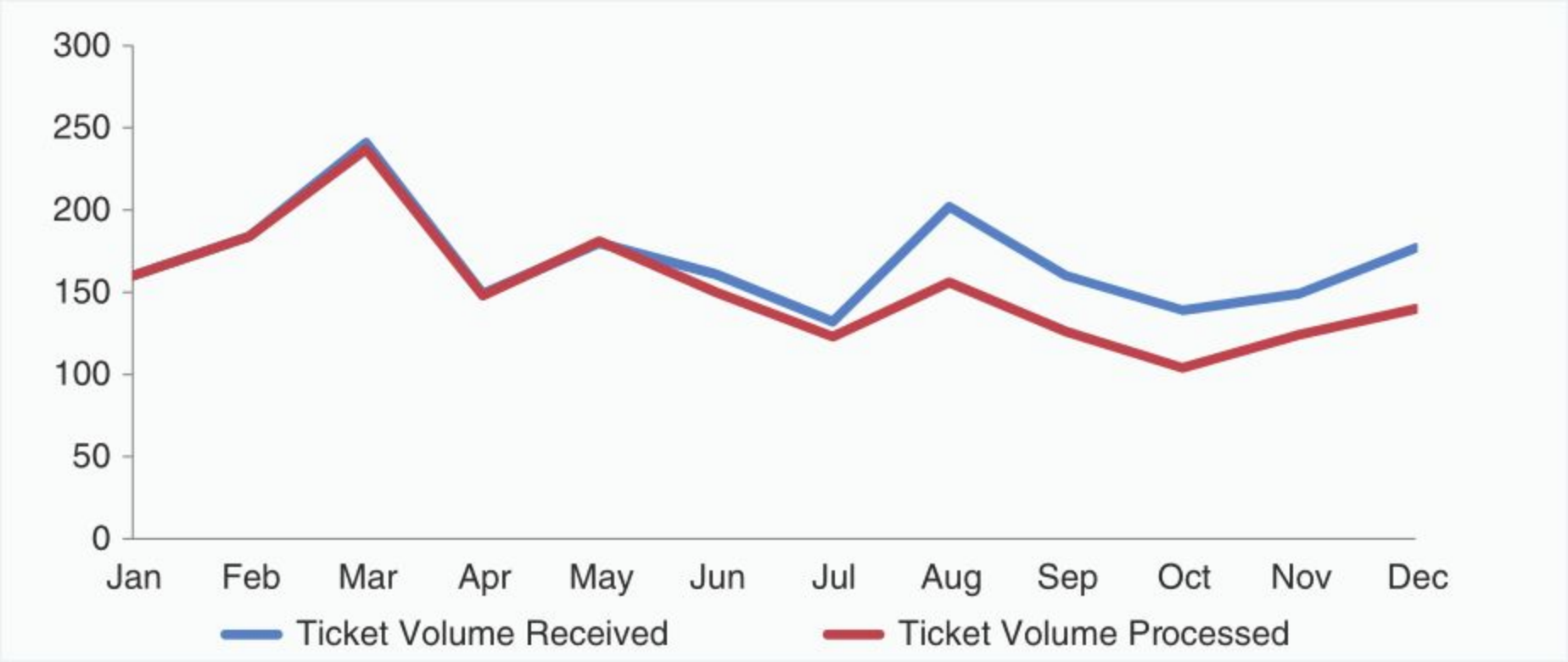


Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.



Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Step 2
Clean your axis.



Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Step 3
The more Obvious The Better.

Step 3

The more **Obvious**
The **Better.**

348307586187236487213 6
478632875632187640732 1
608730164726135876872 1
598237487213498621307 5
732984798321503281479 8
321749025798217349873 2
149827874983217490321 7

How many 0s?

34830758618723648721364786328756321876407321608730164726135876872159823748721349862130753298479832150328147983217490257982173498732149827874983217490321

Orientation

Length

Width

Size

Shape

Curvature

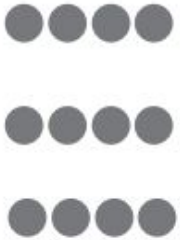
Added Marks

Enclosure

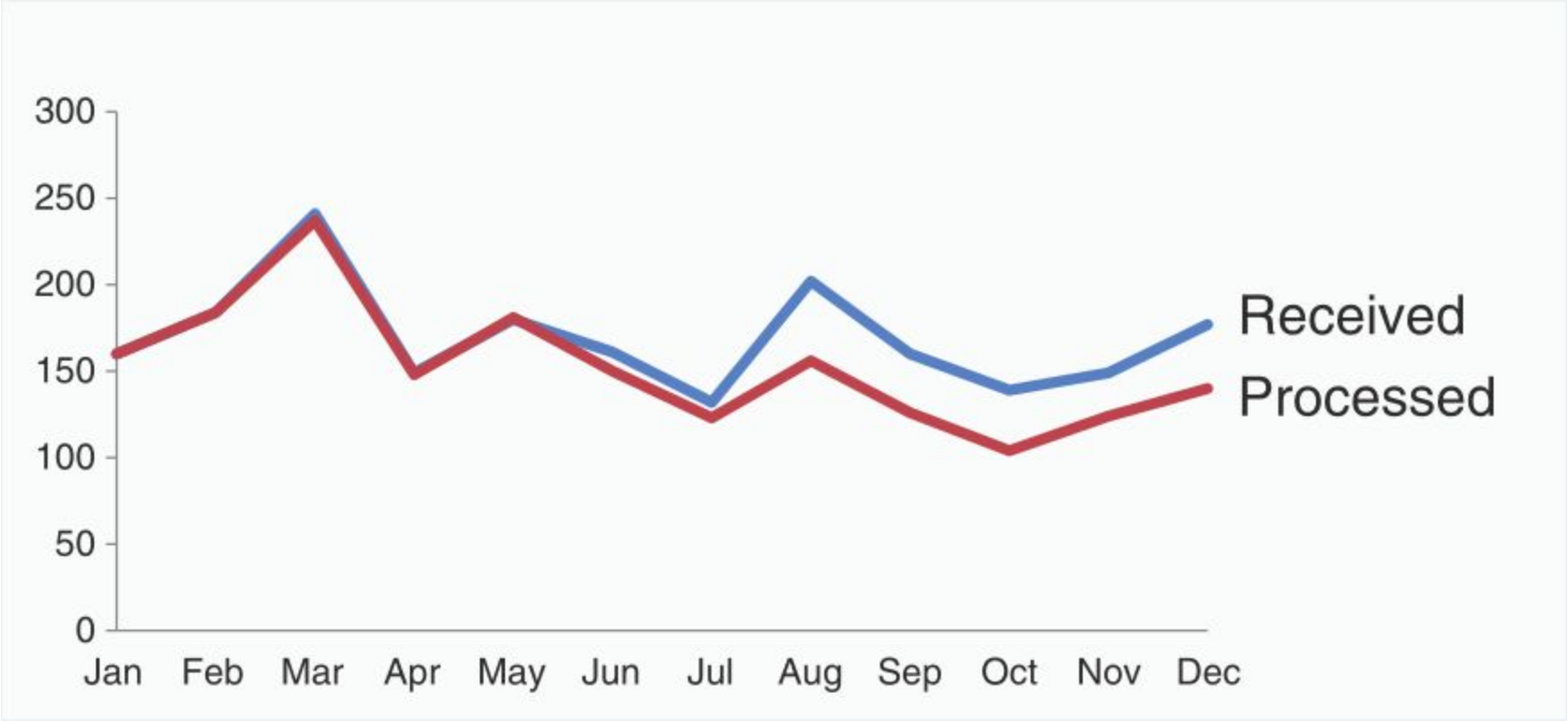
Color Value

Color Hue

Position

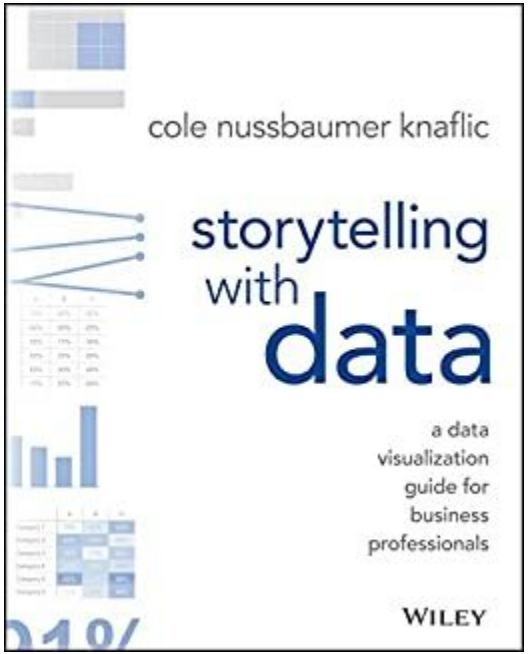


Spatial Grouping



Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Book Recommendation



And no, dont listen the audiobook.