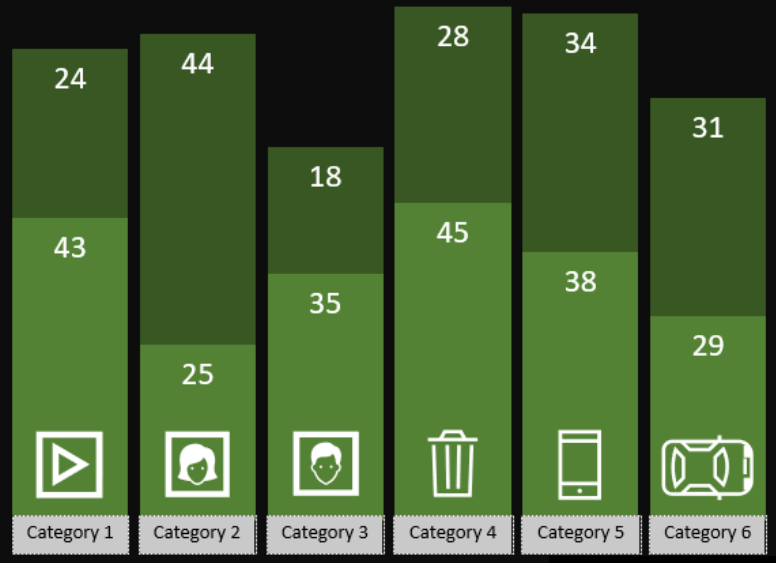
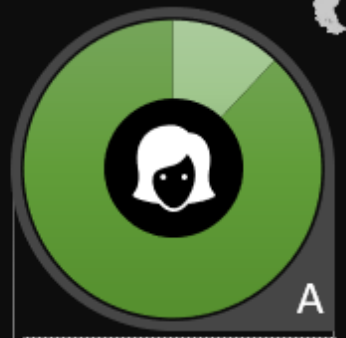
Data Visualization

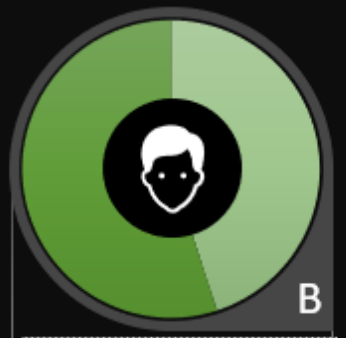
GGE 6505/GGE5405 Introduction to Big Data & Data Science





12%

- Click to add text



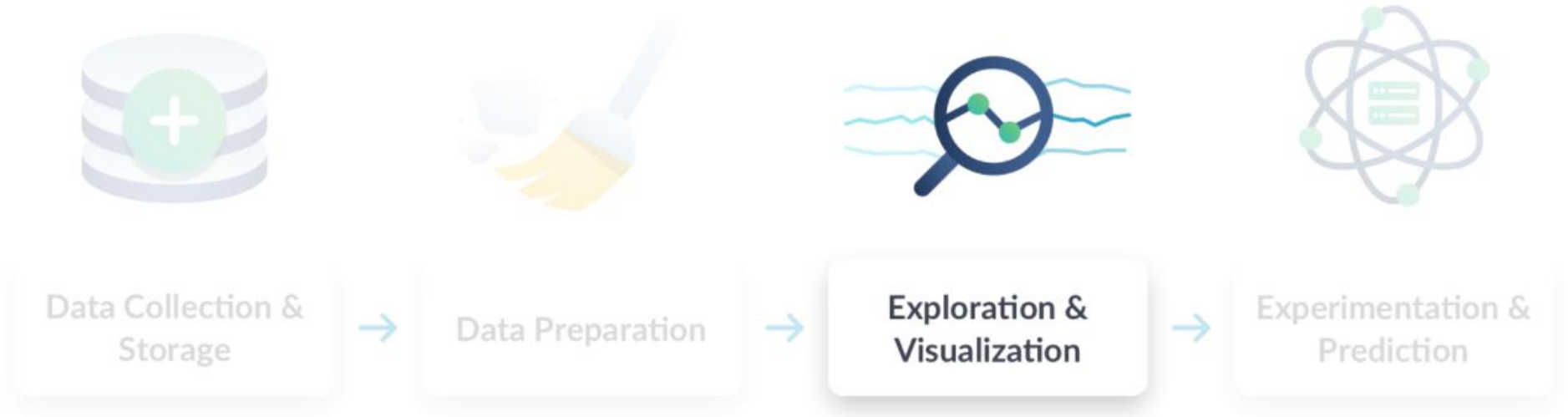
45%

- Click to add text

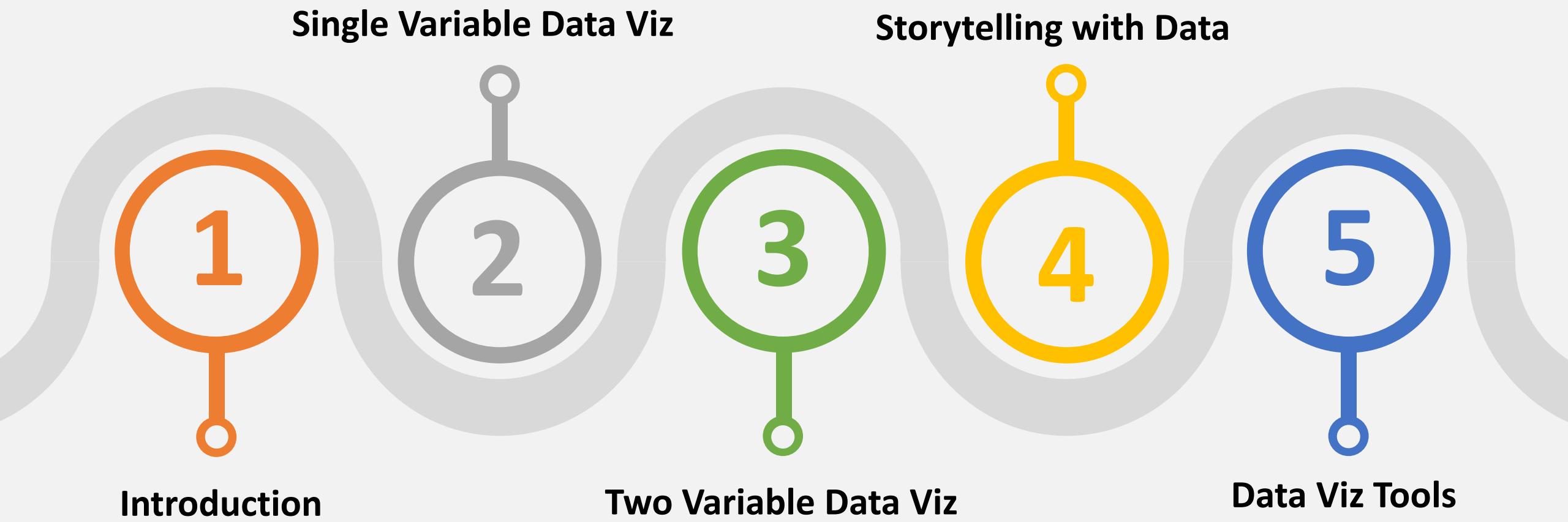
55%

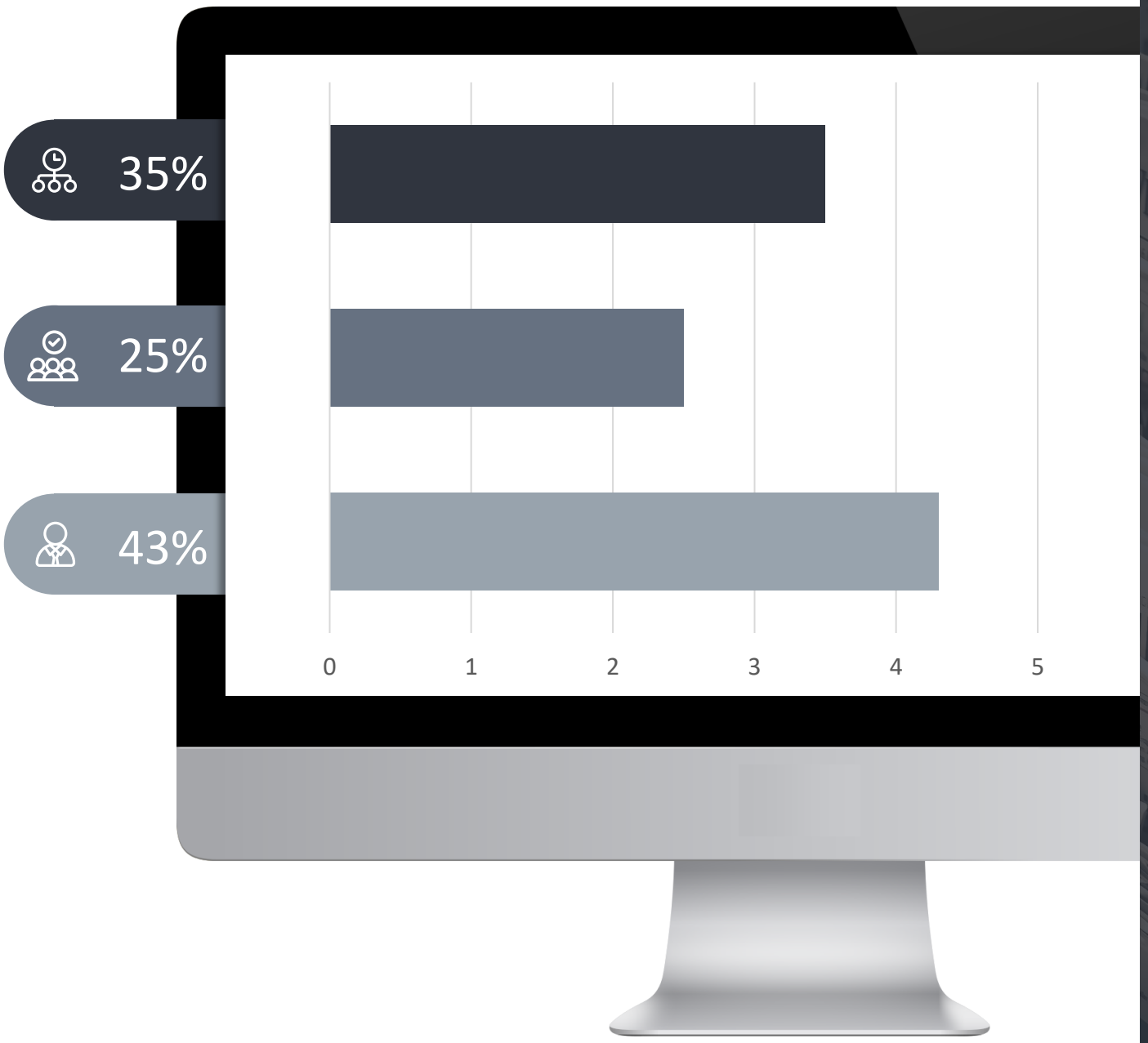
- Click to add text

Data Science Workflow



Outline

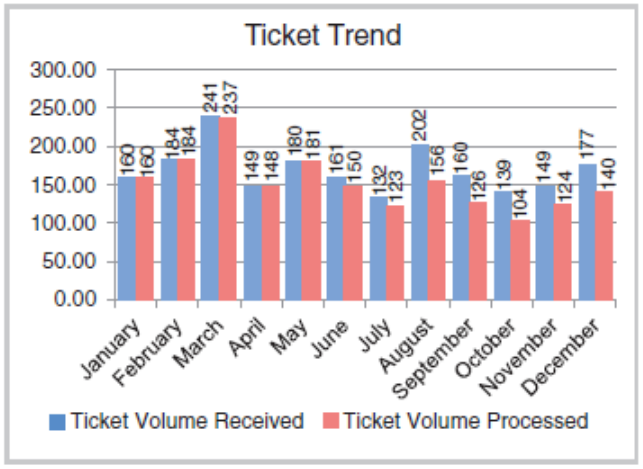

Single Variable Data Viz
Storytelling with Data
1
2
3
4
5
Introduction
Two Variable Data Viz
Data Viz Tools

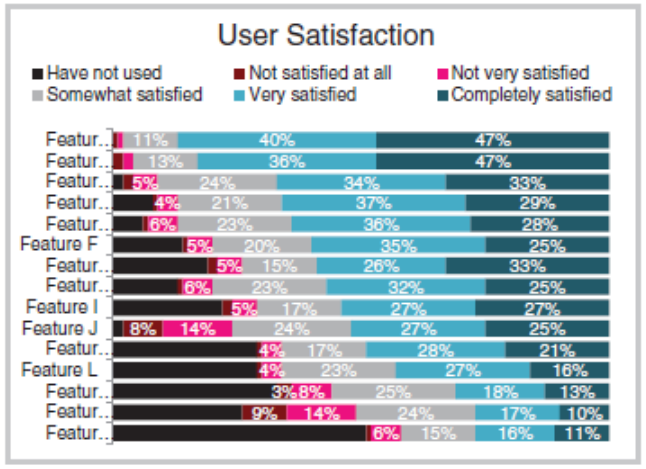


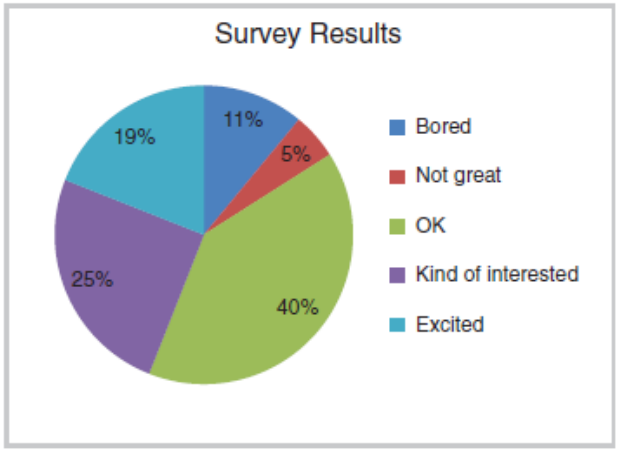
Data visualization (often abbreviated data viz) is an interdisciplinary field that deals with the graphic representation of data. It is a particularly efficient way of communicating when the data is numerous

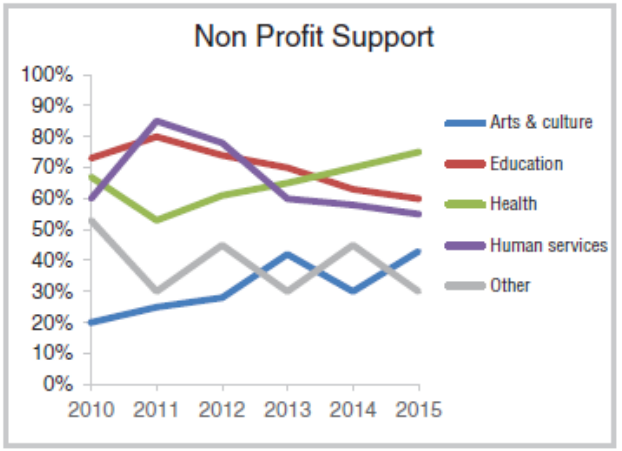
Wikipedia

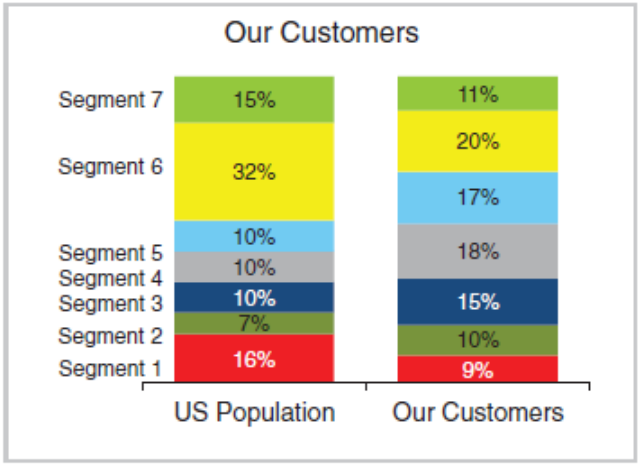
Ineffective Graphs

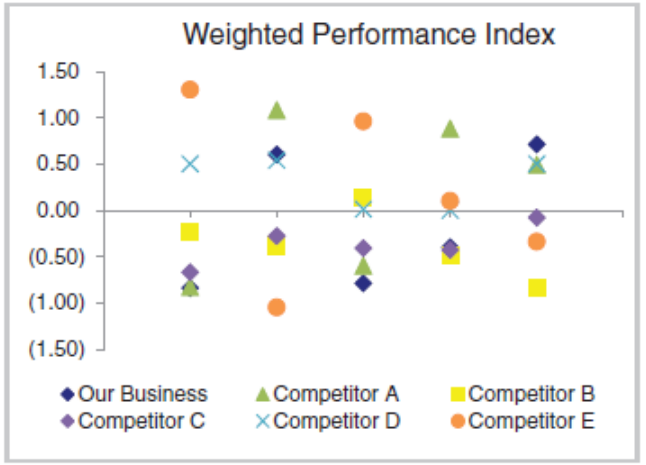












One Variable Data Visualization

- Histogram
- Box Plot

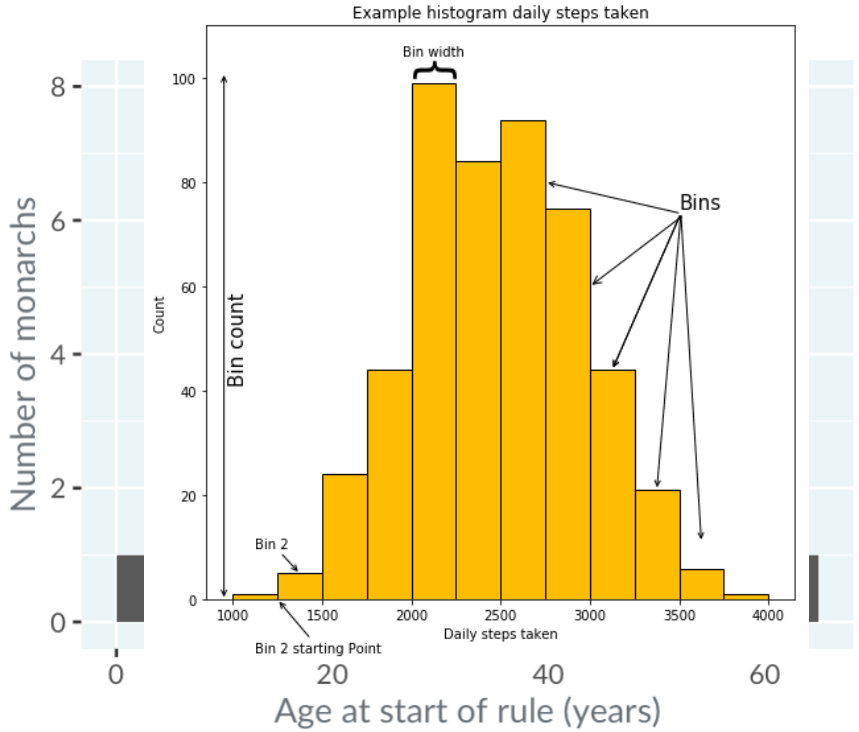
Histogram

- When you have a quantitative variable, that is, an interval or ratio level variable, a histogram is useful.
- A histogram displays numerical data by grouping data into "bins" of equal width.
- You want to show the shape of it’s distribution.

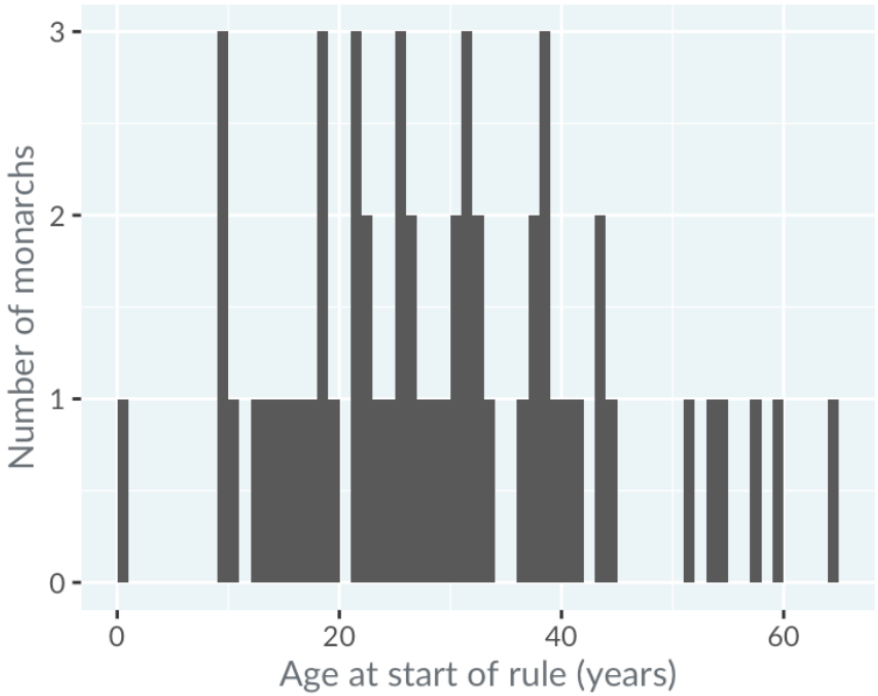
Kings and Queens of England & Britain

official_name	house	birth_date	start_of_rule	age_at_start_of_rule
Elizabeth II	Windsor	1926-04-21	1952-02-06	25.79603
George VI	Windsor	1895-12-14	1936-12-11	40.99110
Edward VIII	Windsor	1894-06-23	1936-01-20	41.57426
...	...	...	...	...
Eadred	Wessex	0923-07-01	0946-05-26	22.90212
Edmund I	Wessex	0921-07-01	0939-10-27	18.32170
Aethelstan	Wessex	0894-07-01	0924-07-01	29.99863

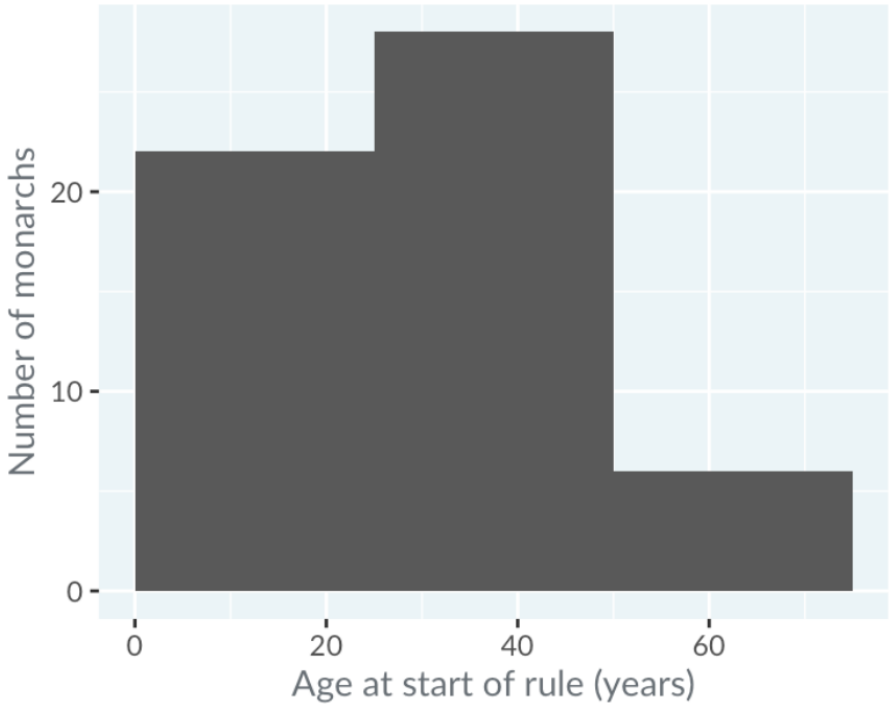
Interpreting histograms | Theory (datacamp.com)



Bin: one year

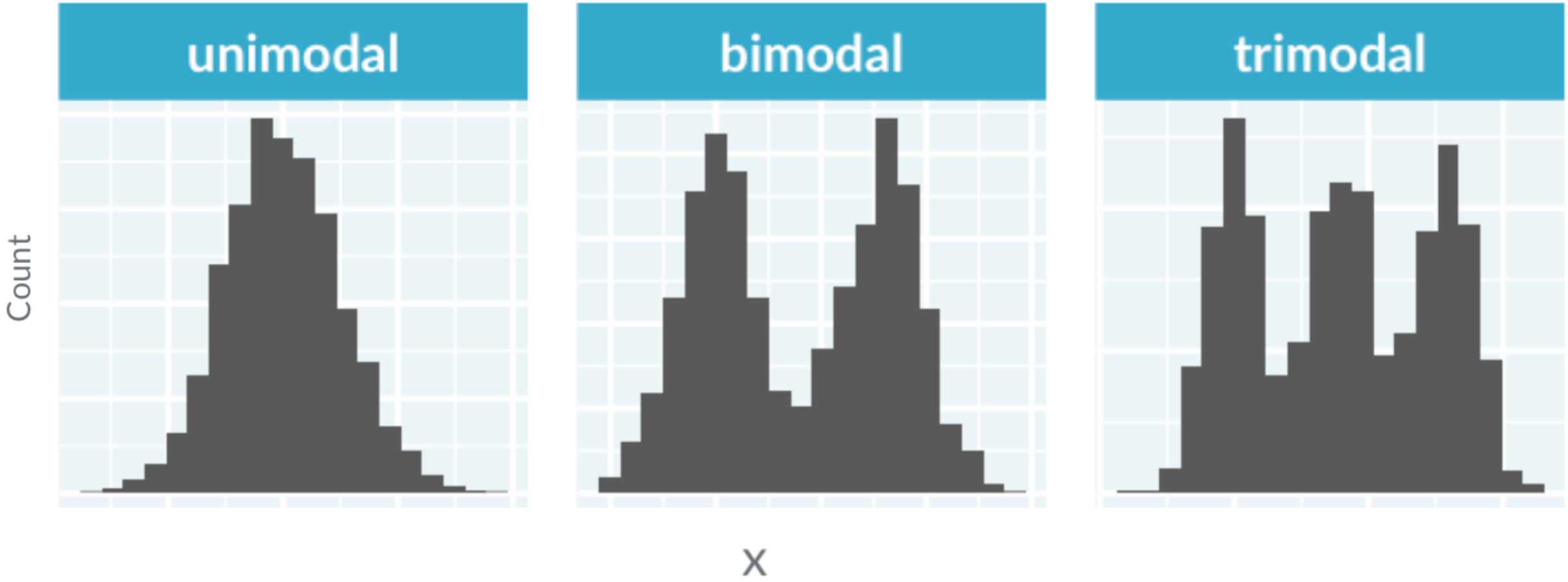


Bin: twenty-five year

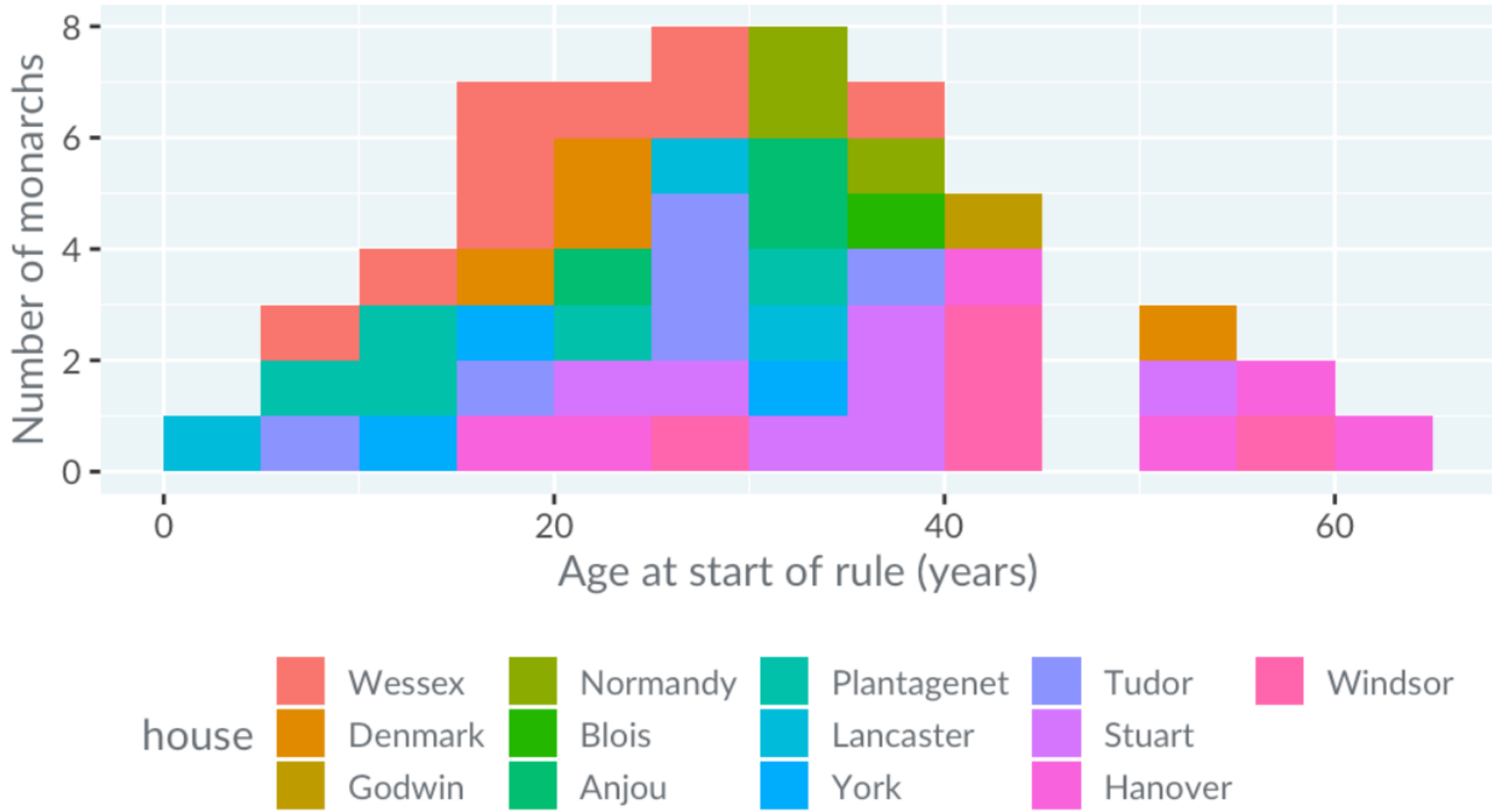


Interpreting histograms | Theory (datacamp.com)

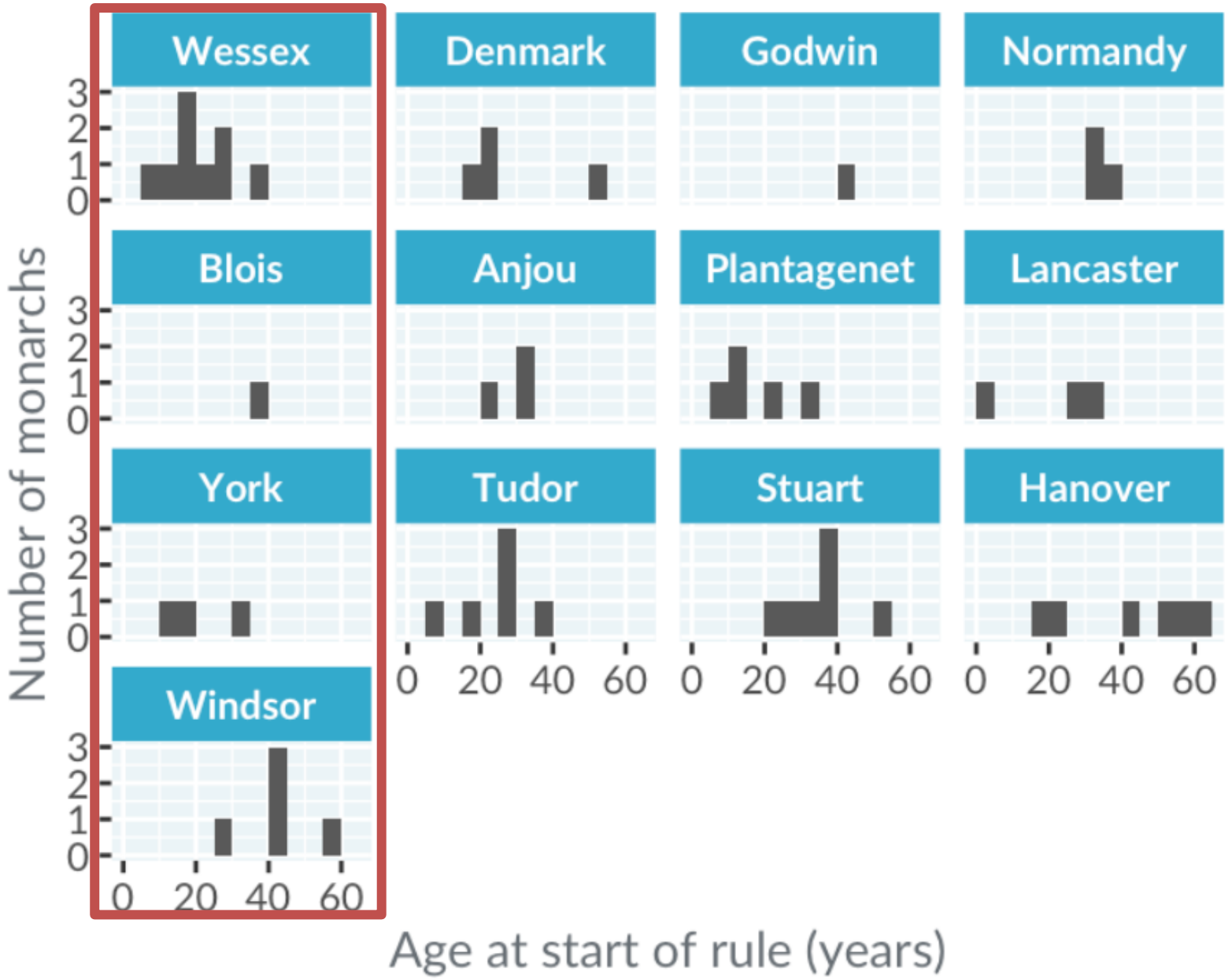
Modality



Interpreting histograms | Theory (datacamp.com)

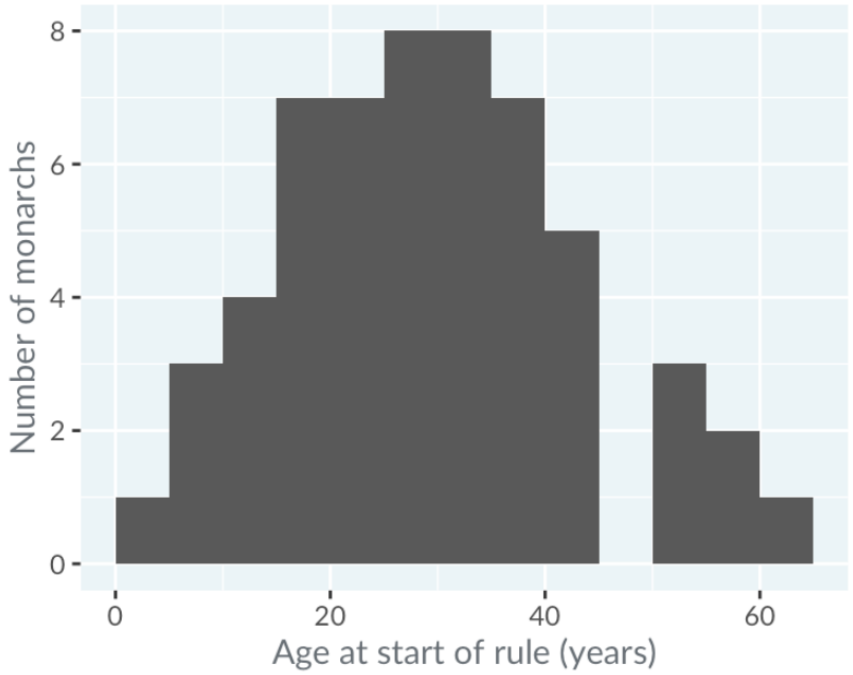


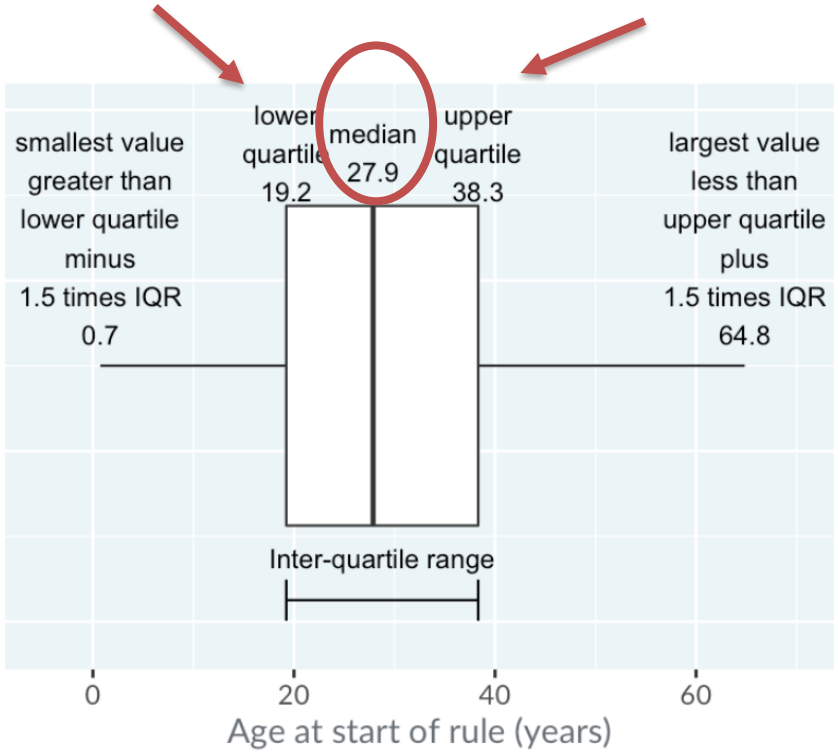
Interpreting histograms | Theory (datacamp.com)


Wessex
Denmark
Godwin
Normandy
Blois
Anjou
Plantagenet
Lancaster
York
Tudor
Stuart
Hanover
Windsor
Number of monarchs
0
1
2
3
0
20
40
60
Age at start of rule (years)

Box Plot

- Boxplot is a way to show the spread and centers of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of center include the mean or average and median (the middle of a data set).



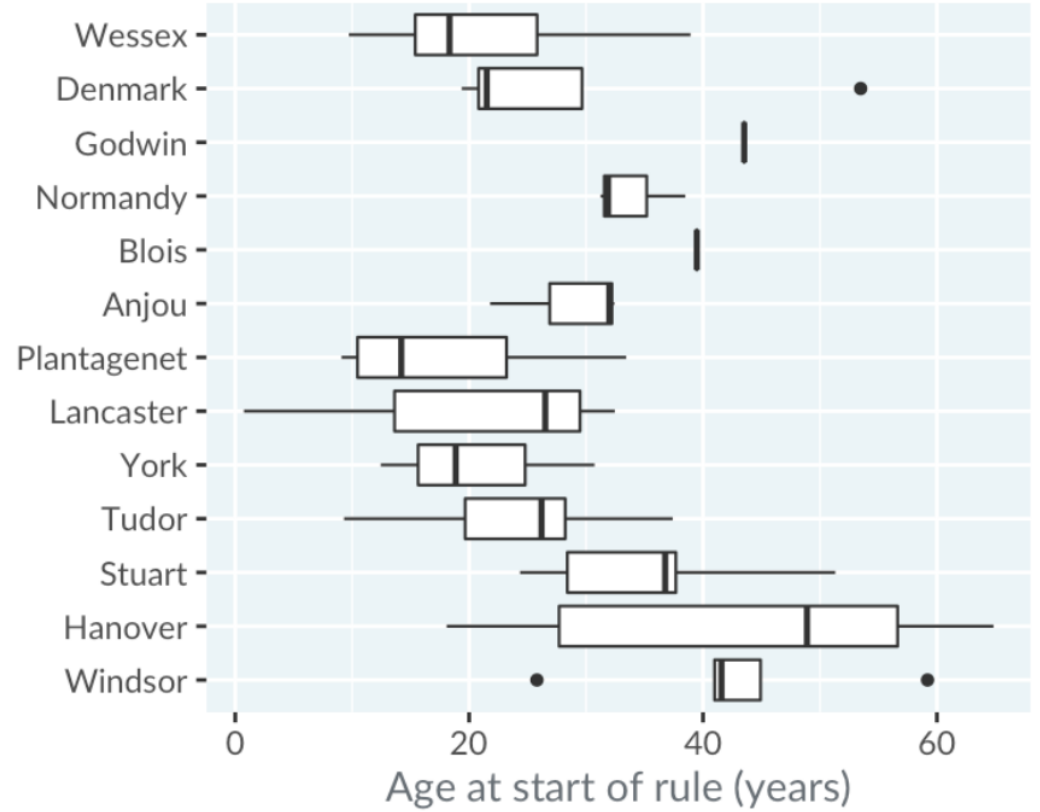


Interpreting histograms | Theory (datacamp.com)

Box Plot

Five pieces of are generally included in the chart:

- The minimum (the smallest number in the data set). The minimum is shown at the far left of the chart, at the end of the left “whisker.”
- First quartile, Q1, is the far left of the box (or the far right of the left whisker).
- The median is shown as a line in the center of the box.
- Third quartile, Q3, shown at the far right of the box (at the far left of the right whisker).
- The maximum (the largest number in the data set), shown at the far right of the box.

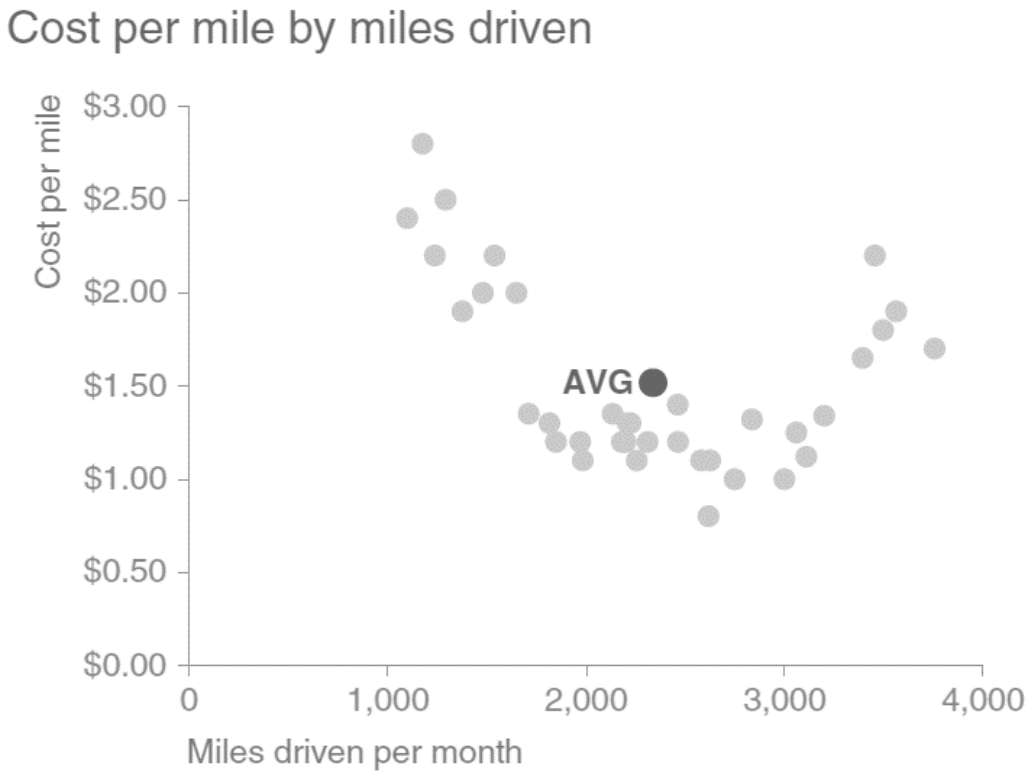


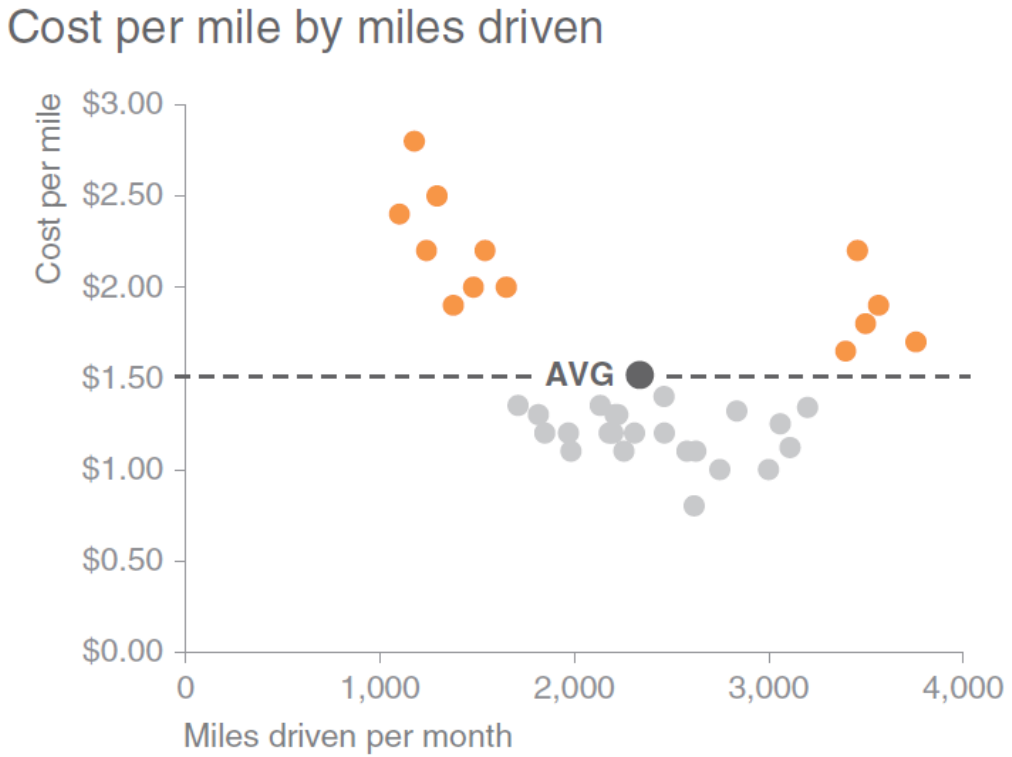
Two Variable Data Visualization

- Points
  - Scatter Plot
- Lines
  - Line Graph
- Bars
  - Bar chart
- Other Types
  - Pie Chart
  - Heatmap

Scatter Plots

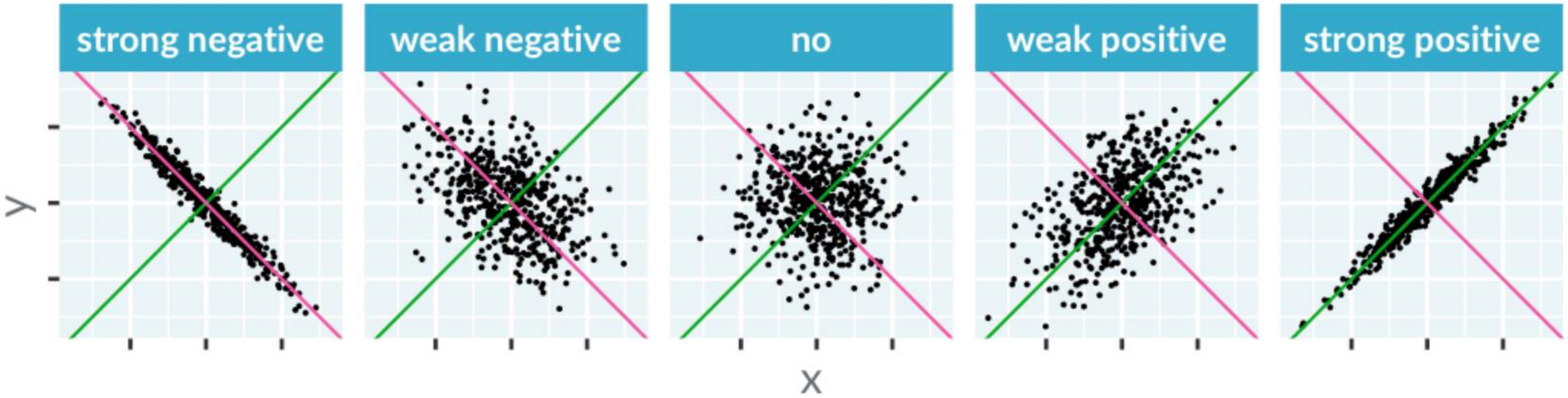
- Showing the relationship between two Variables.





Scatter Plots

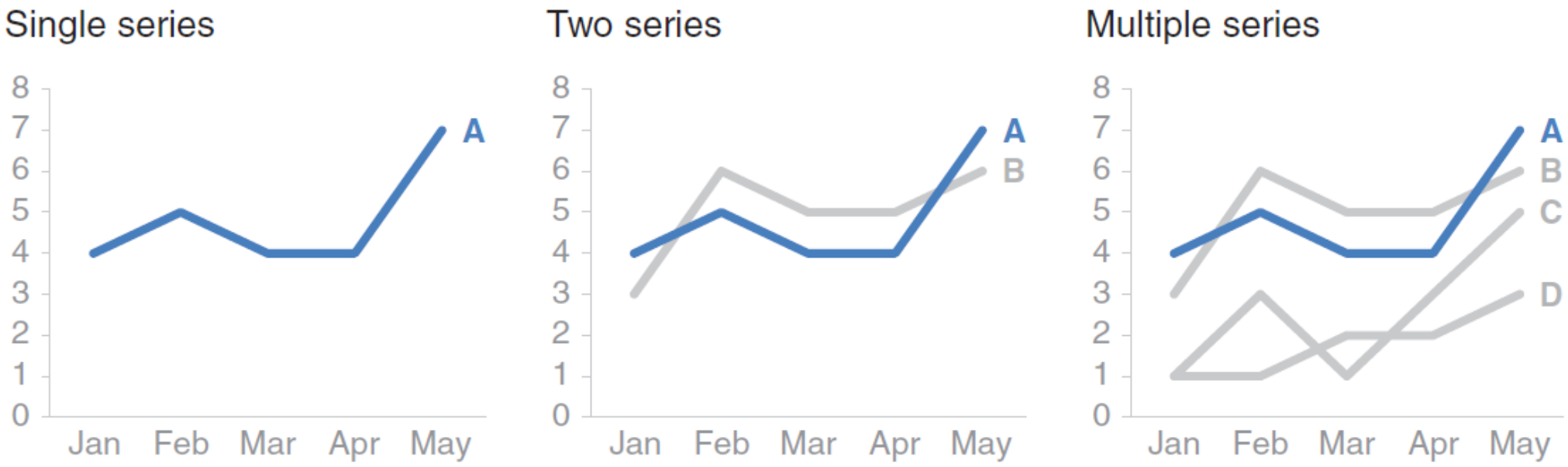
- Identification of correlational relationships are common with scatter plots.

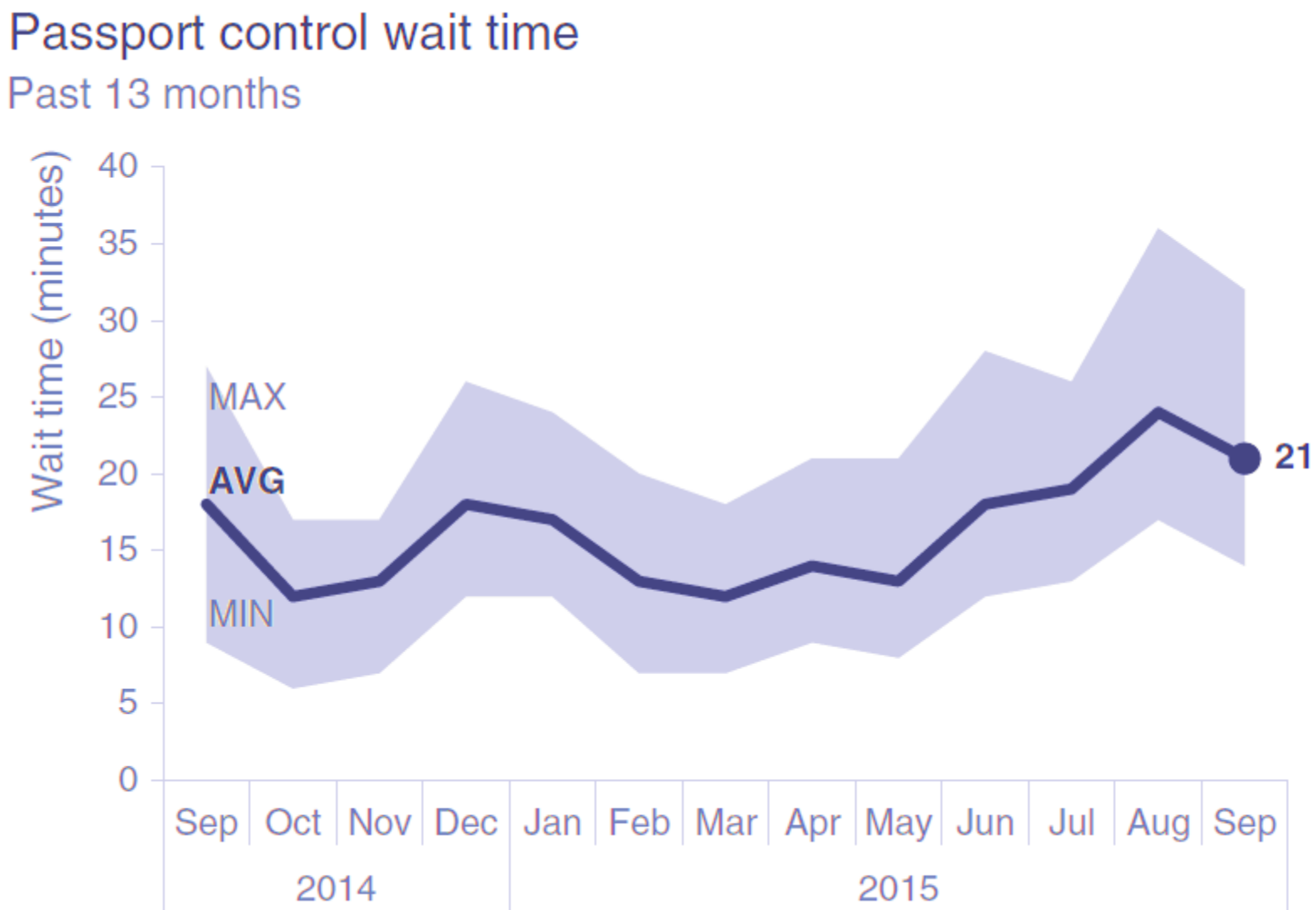


Interpreting histograms | Theory (datacamp.com)

Line Plot

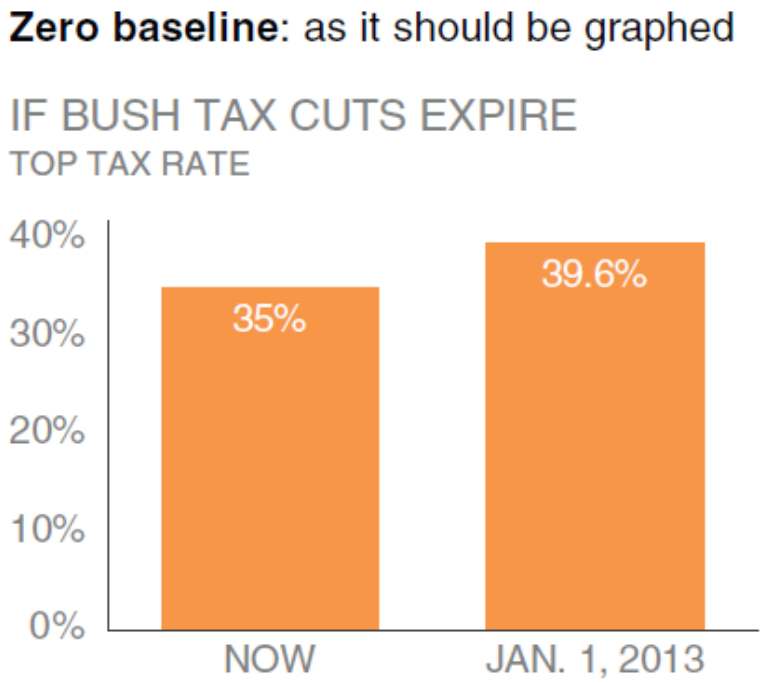
- Line graphs are most commonly used to plot continuous data.
- Consecutive observations are connected somehow.




Passport control wait time
Past 13 months
Wait time (minutes)
40
35
30
25
20
15
10
5
0
MAX
AVG
MIN
21
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
2014
2015

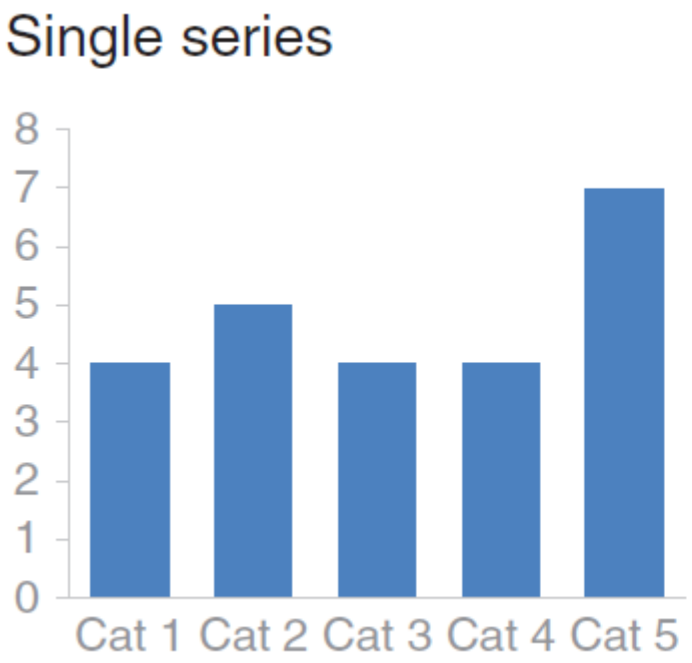
Bar Chart

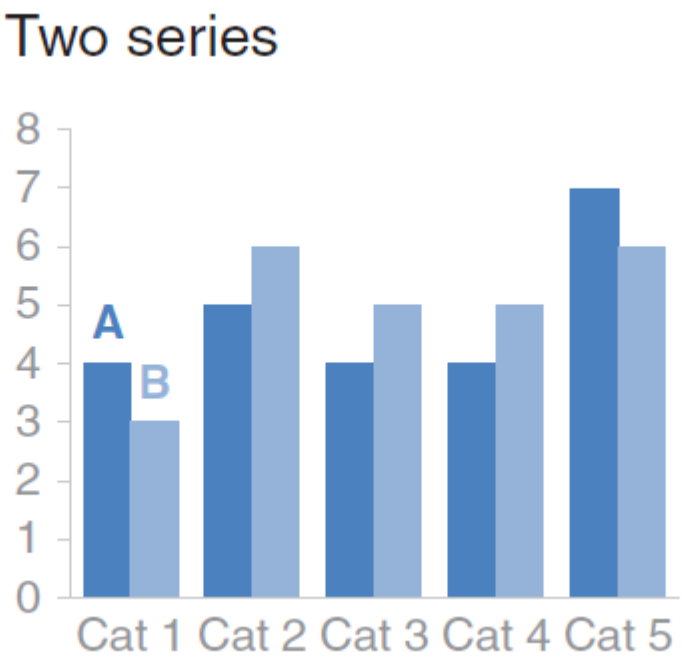
- Bar charts are easy for our eyes to read
- You have categorical variables and wants to counts or percentages for each category

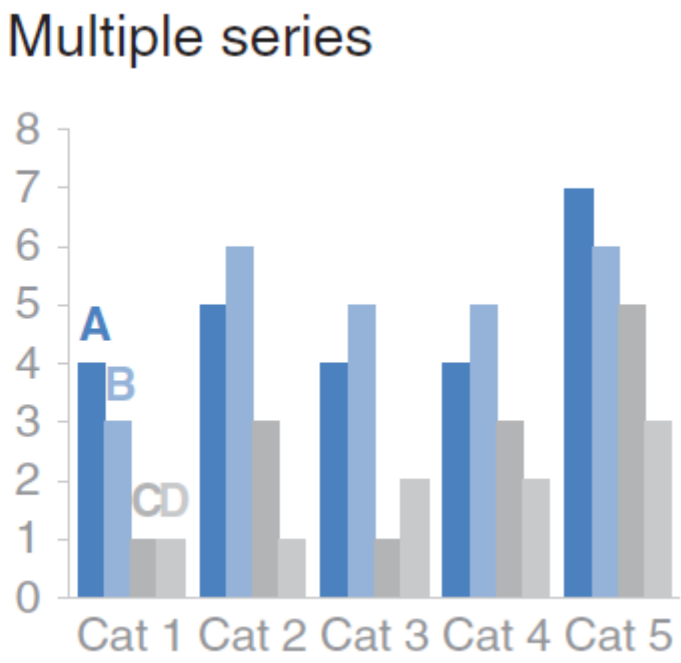


Vertical Bar Chart

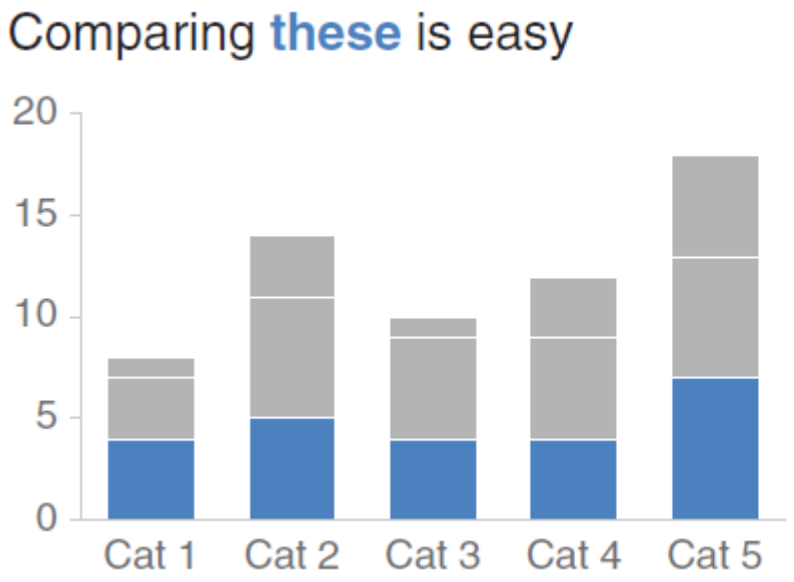
- Vertical bar charts can be single series, two series, or multiple series

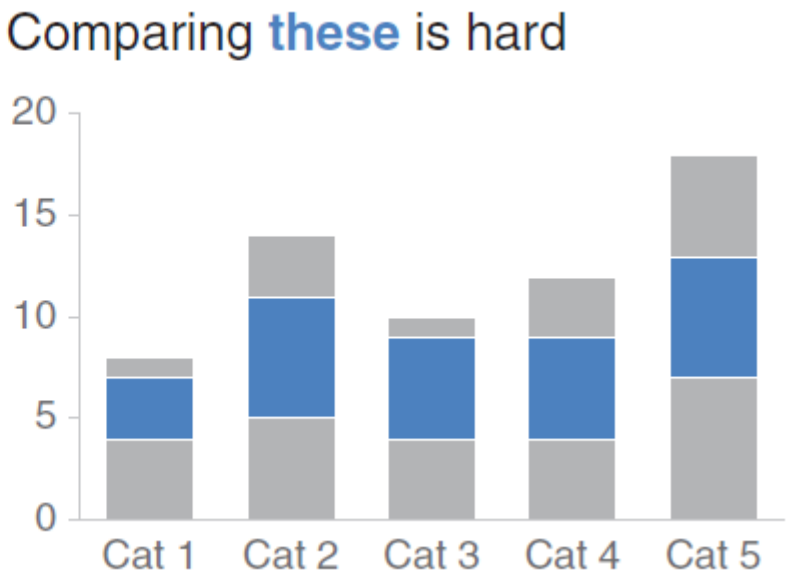






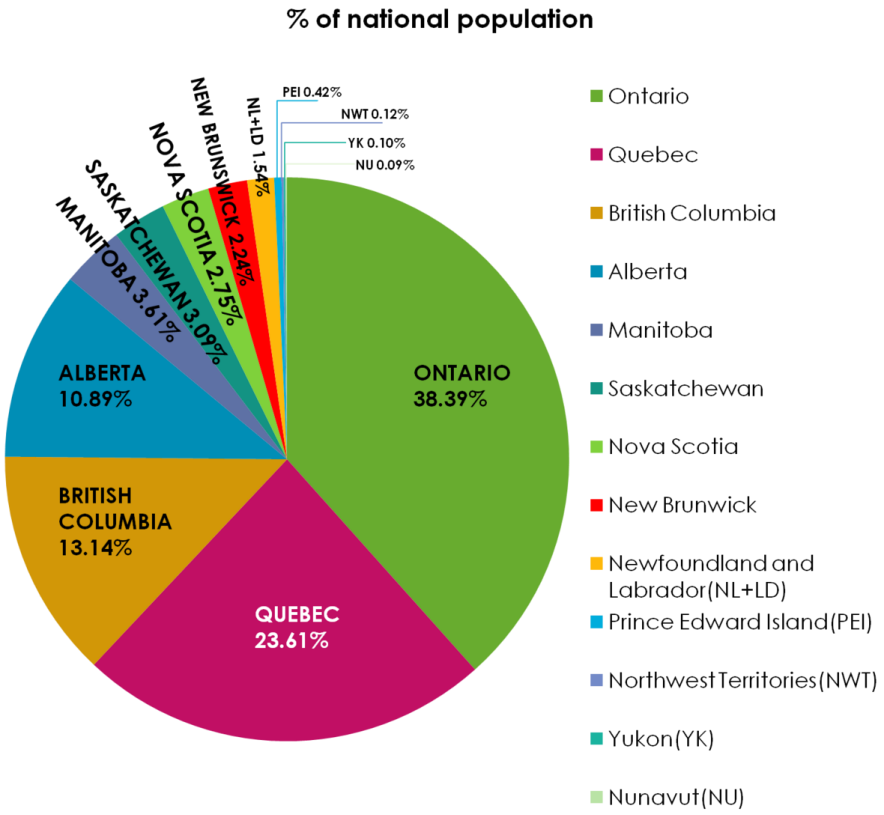
Stacked Vertical Bar Chart





Other types: Pie Chart

- A common way to display categorical variables is with pie charts.
- Unlike bar graphs and line graphs, pie charts do not show changes over time



File:Population of Provinces and Territories of Canada Pie Chart.png - Wikimedia Commons

Other types: Heatmap

- A heatmap is a way to visualize data in tabular format, where in place of (or in addition to) the numbers, you leverage colored cells that convey the relative magnitude of the numbers.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

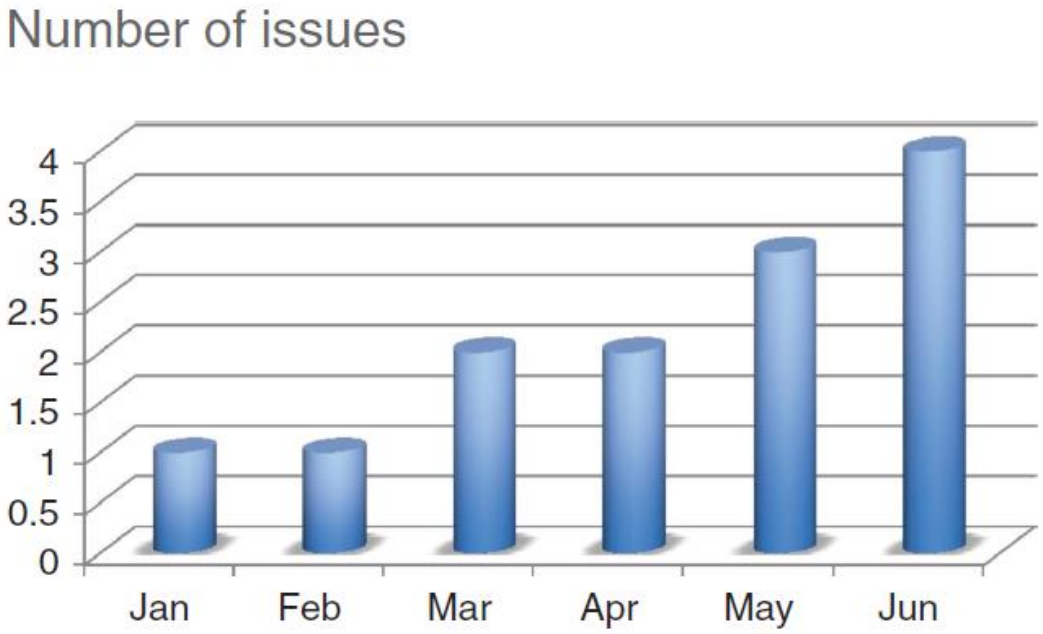
LOW-HIGH

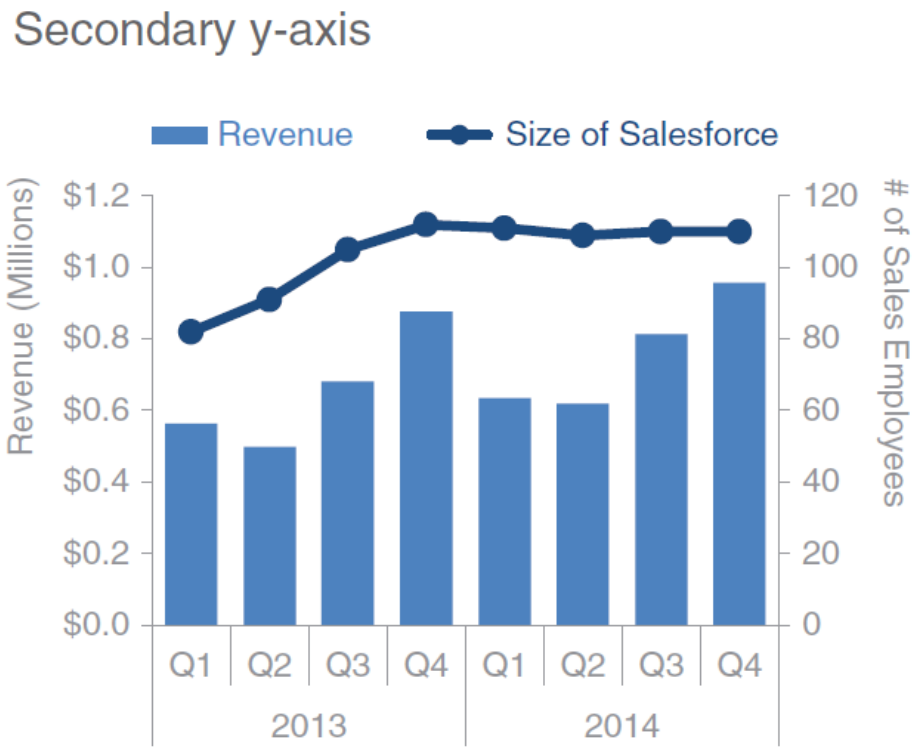
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

To be Avoided!

To be Avoided!

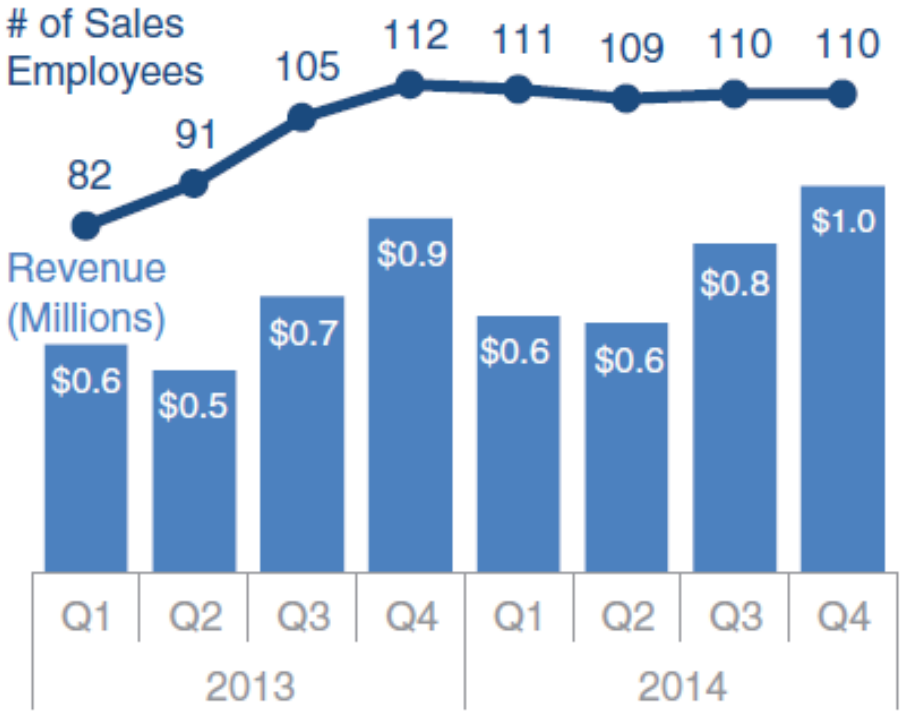
- One of the golden rules of data visualization goes like this: never use 3D.
- Don’t show the second y-axis. Instead, label the data points that belong on this axis directly.



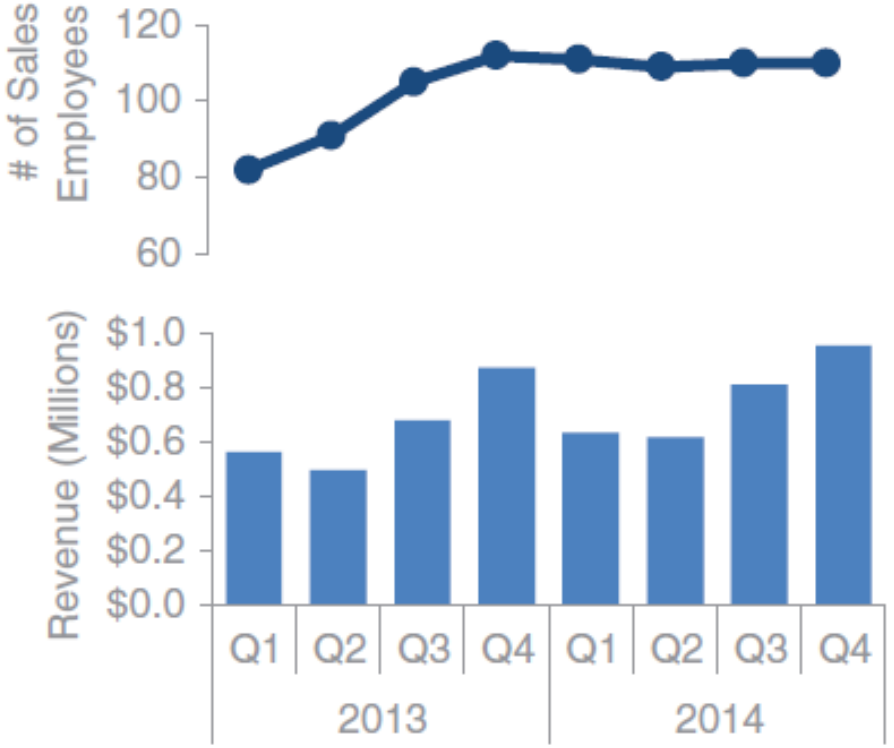


To be Avoided!

Alternative 1: **label directly**



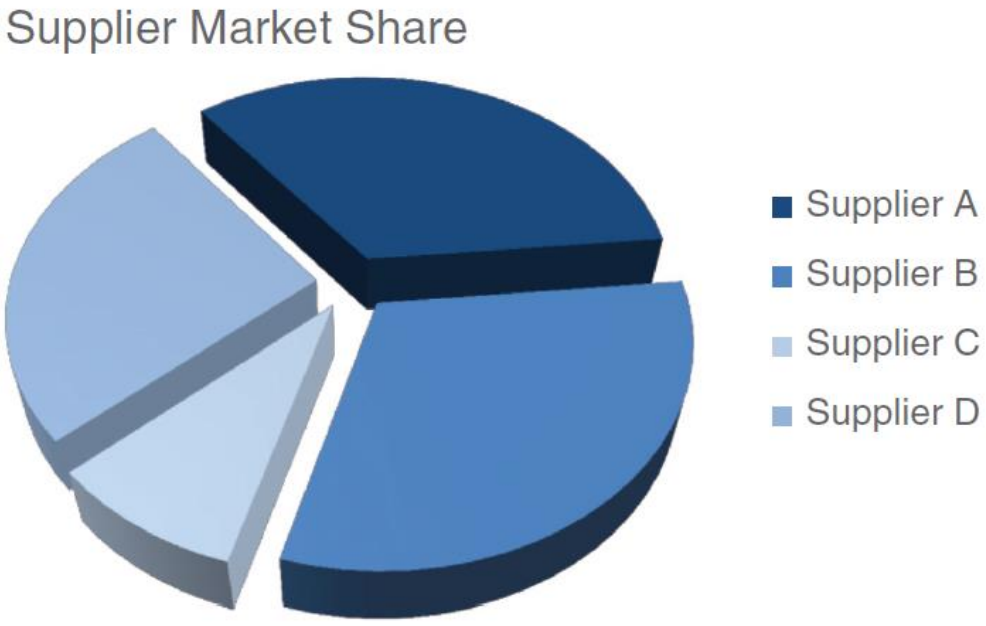
Alternative 2: **pull apart vertically**

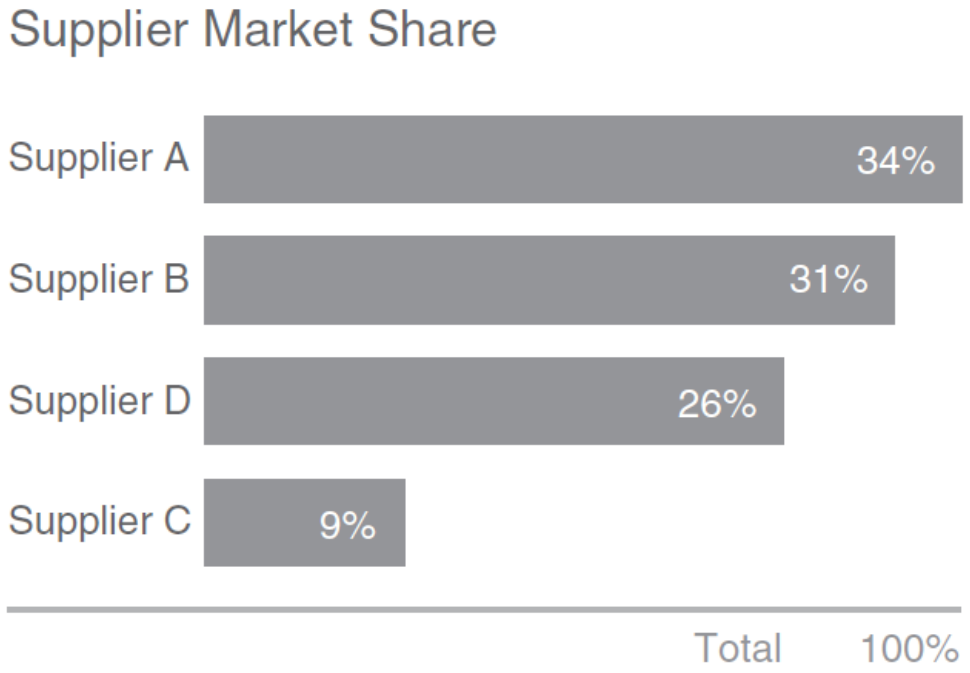


To be Avoided!

- Try to avoid pie charts!

market share across four suppliers: A, B, C, and D. If I asked you to make a simple observation—which supplier is the largest based on this visual—what would you say?

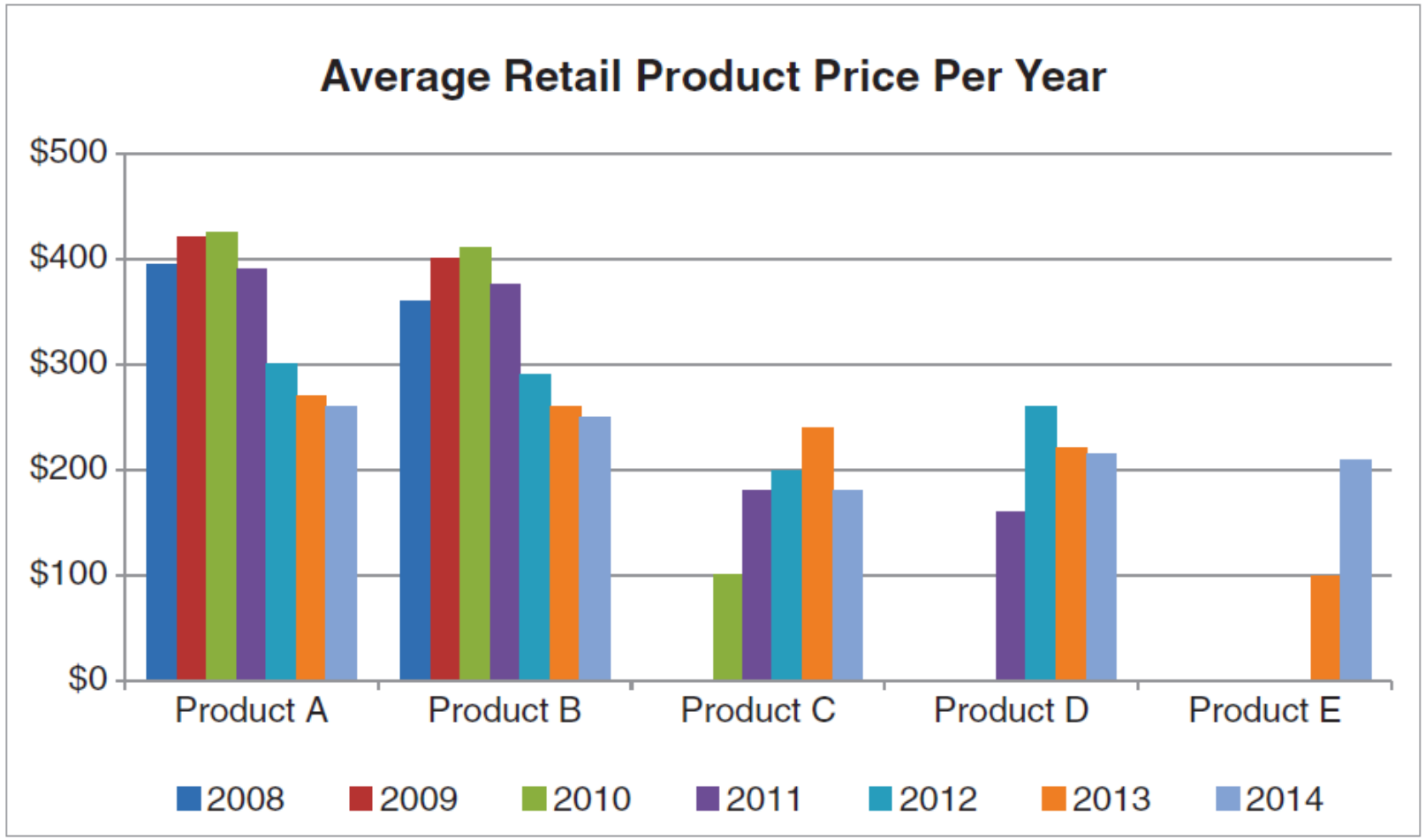




Storytelling with Data

Storytelling with Data

Price has declined for all products on the market since the launch of Product C in 2010



two best selling books that will improve your data visualization skills. — storytelling with data

Storytelling with Data

1. Understand the context

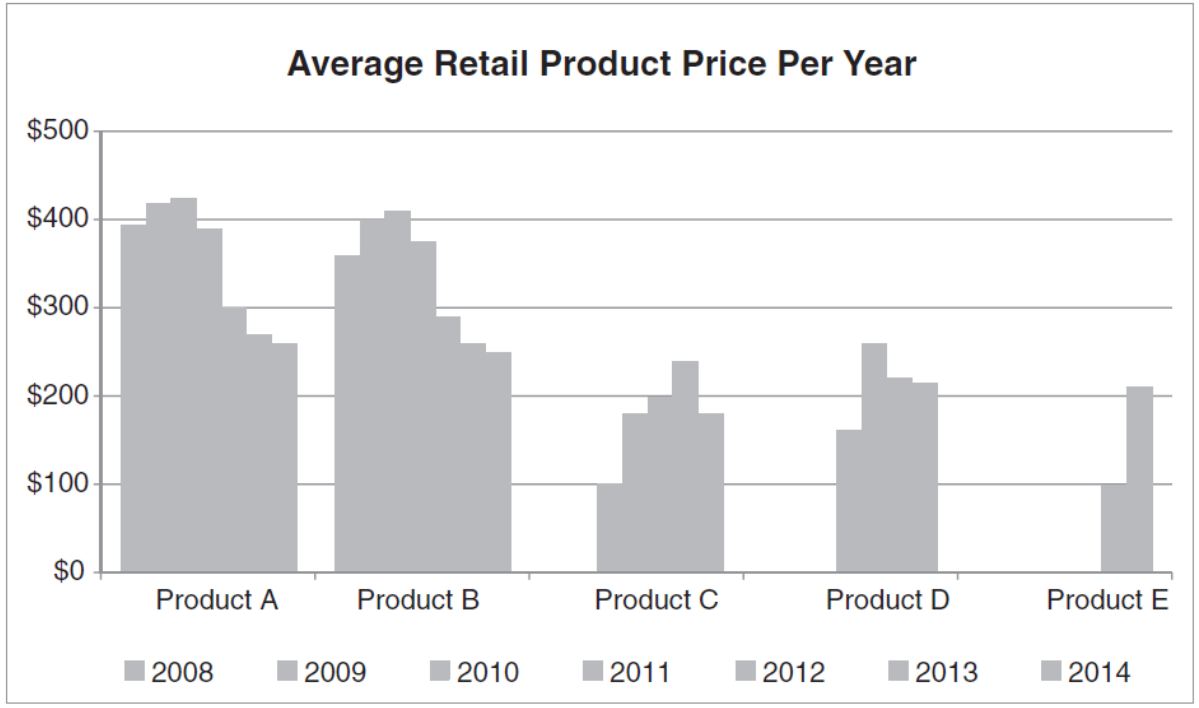
- The first thing to do when faced with a visualization challenge is to make sure you have a robust understanding of the context and what you need to communicate. We must identify a specific audience and what they need to know or do and determine the data we'll use to illustrate our case.

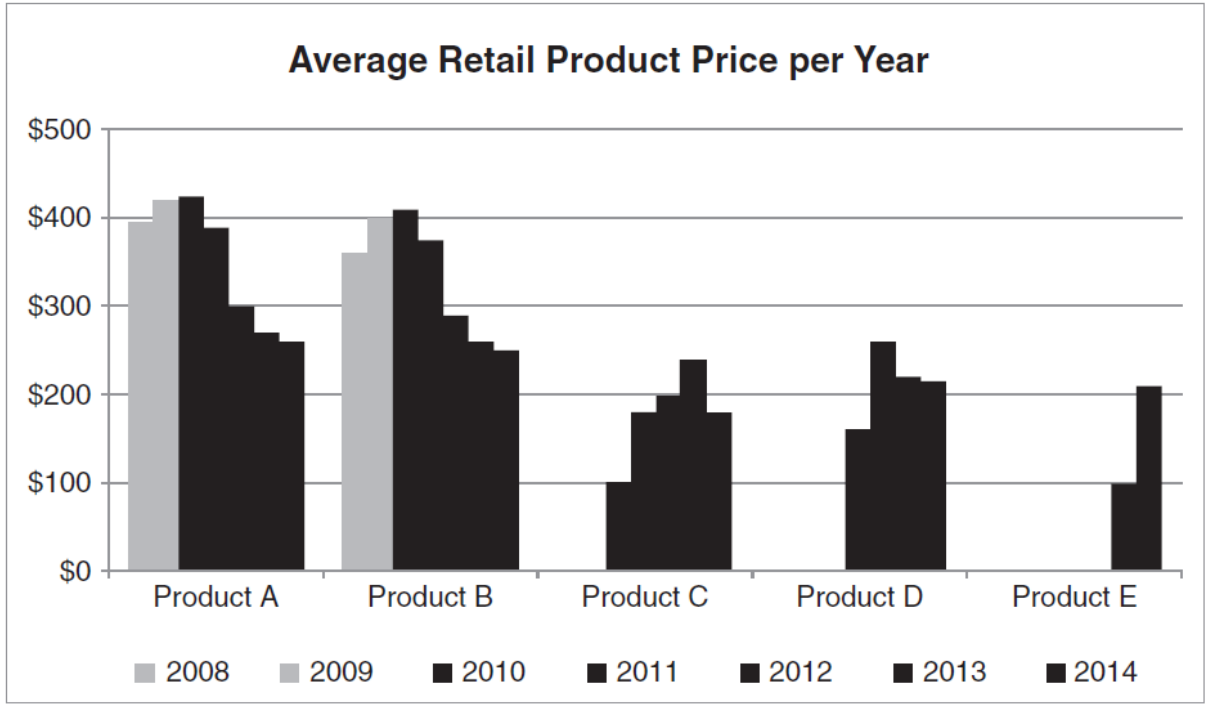
- Case Study: One of the considerations in this decision-making process is **how competitors' retail prices for products in this marketplace have changed over time.**
- **Observation:** Price has declined for all products on the market since the launch of Product C in 2010

Storytelling with Data

2. Choose an appropriate display

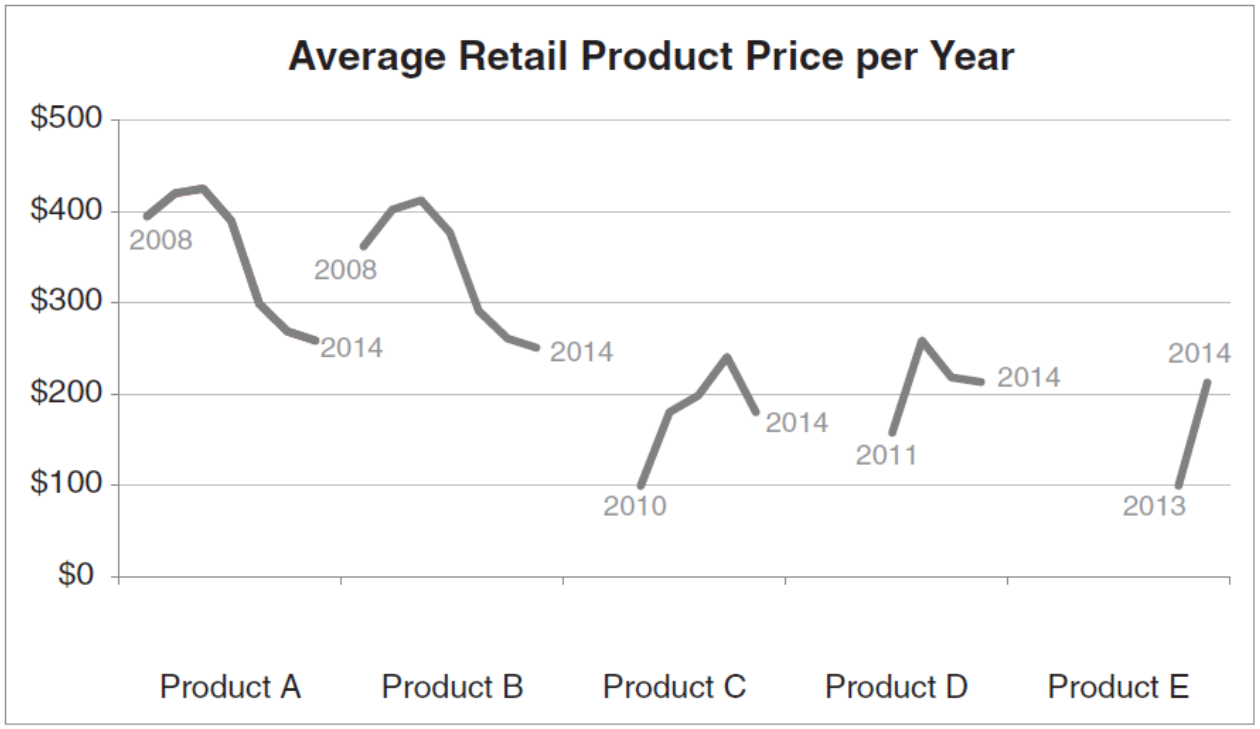
- First, let's remove the visual obstacle of the variance in color
- let's highlight the relevant pieces of data to make it easier to focus our attention there

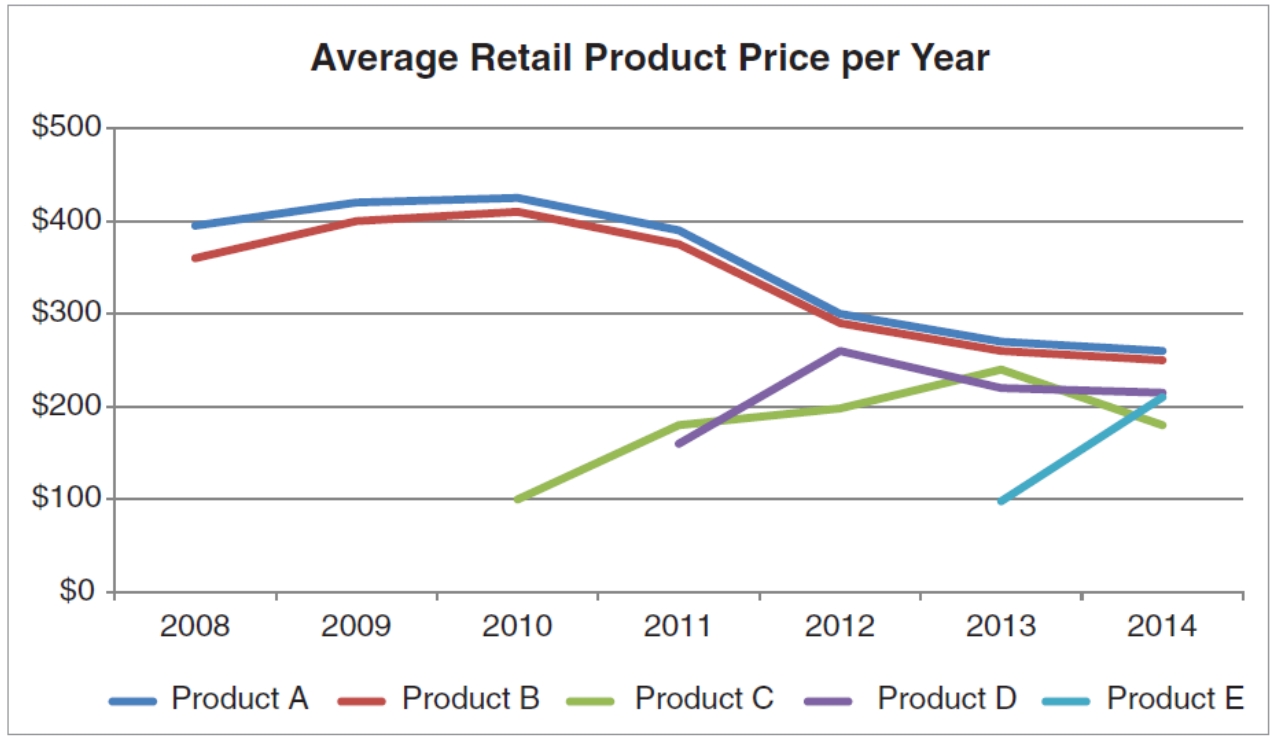




Storytelling with Data

2. Choose an appropriate display





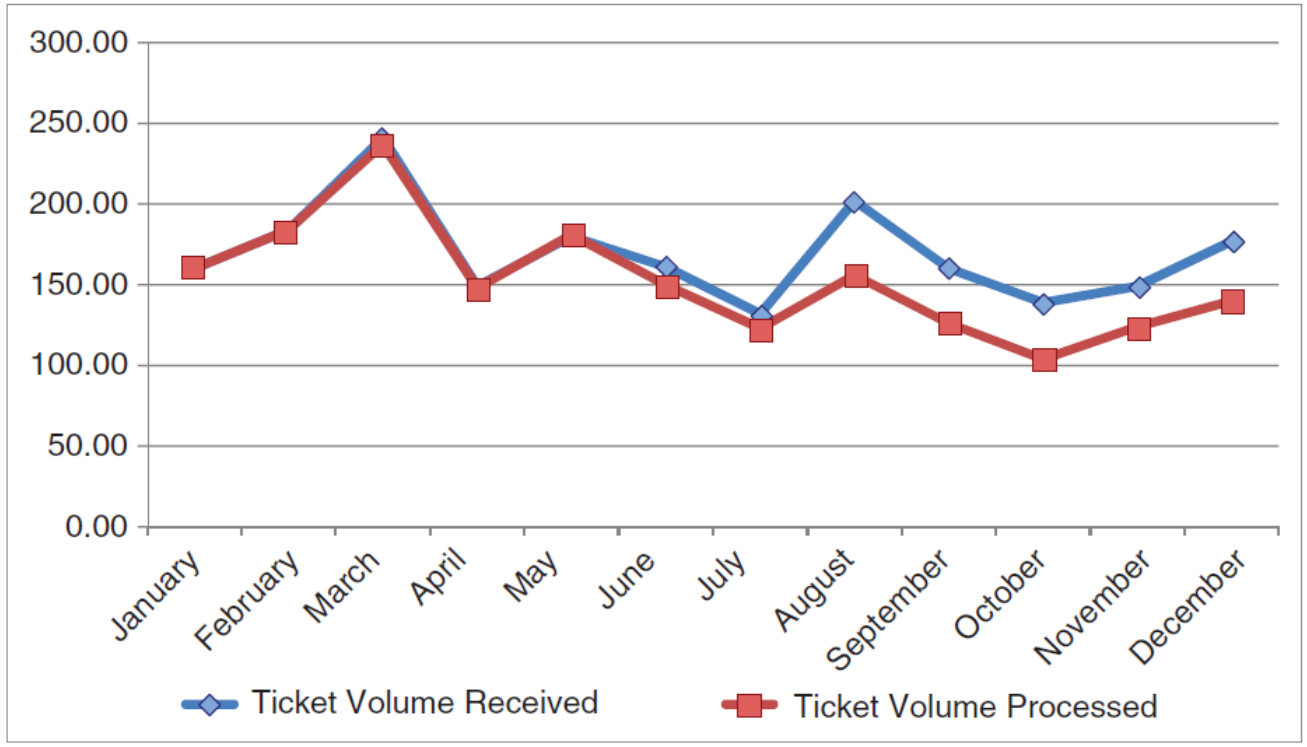
Storytelling with Data

3. Eliminate Clutter

- These are visual elements that take up space but don't increase understanding.
- Clutter can make something feel more complicated than it actually is.

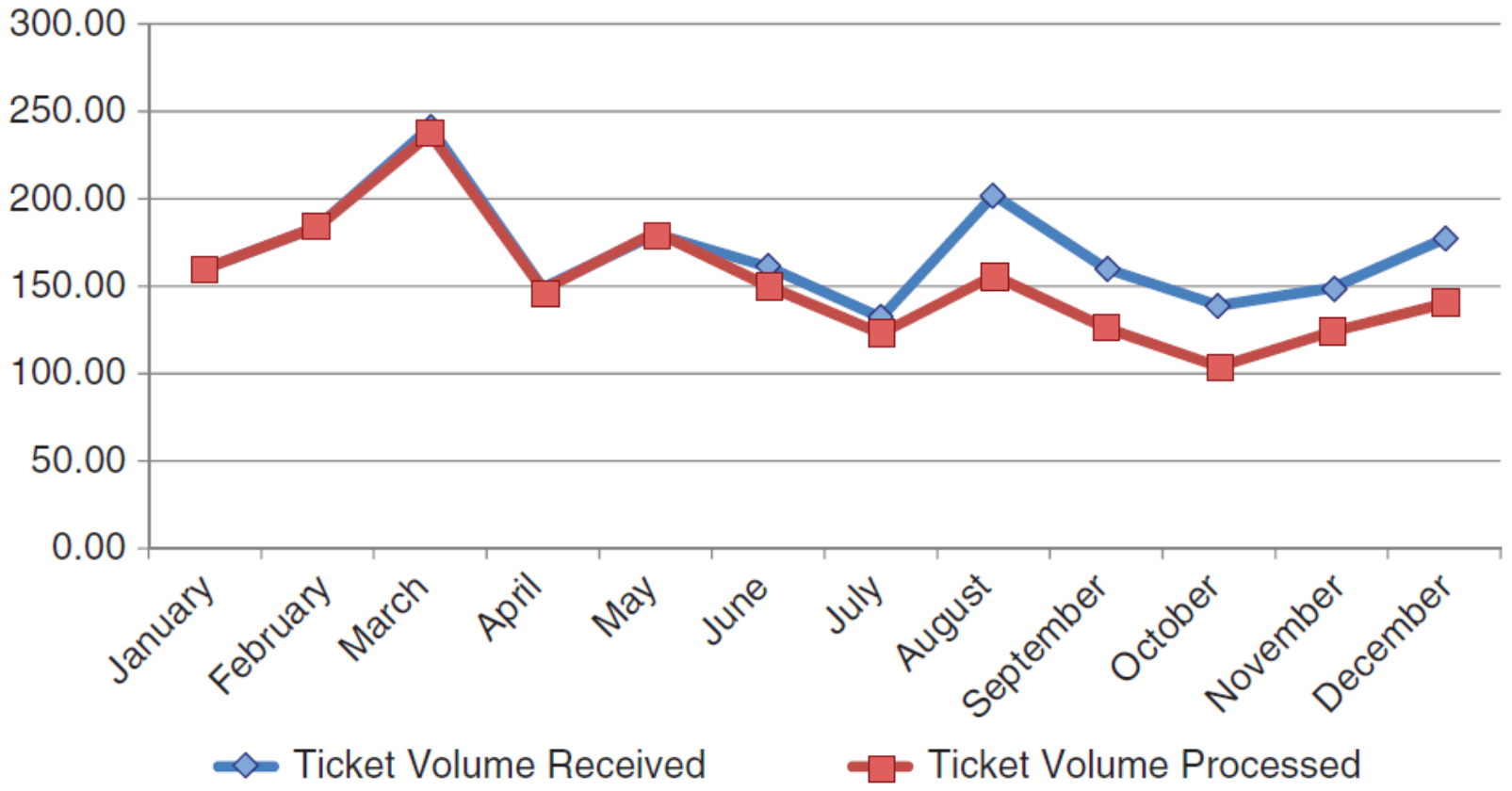
- Case study:

technical issues, from employees



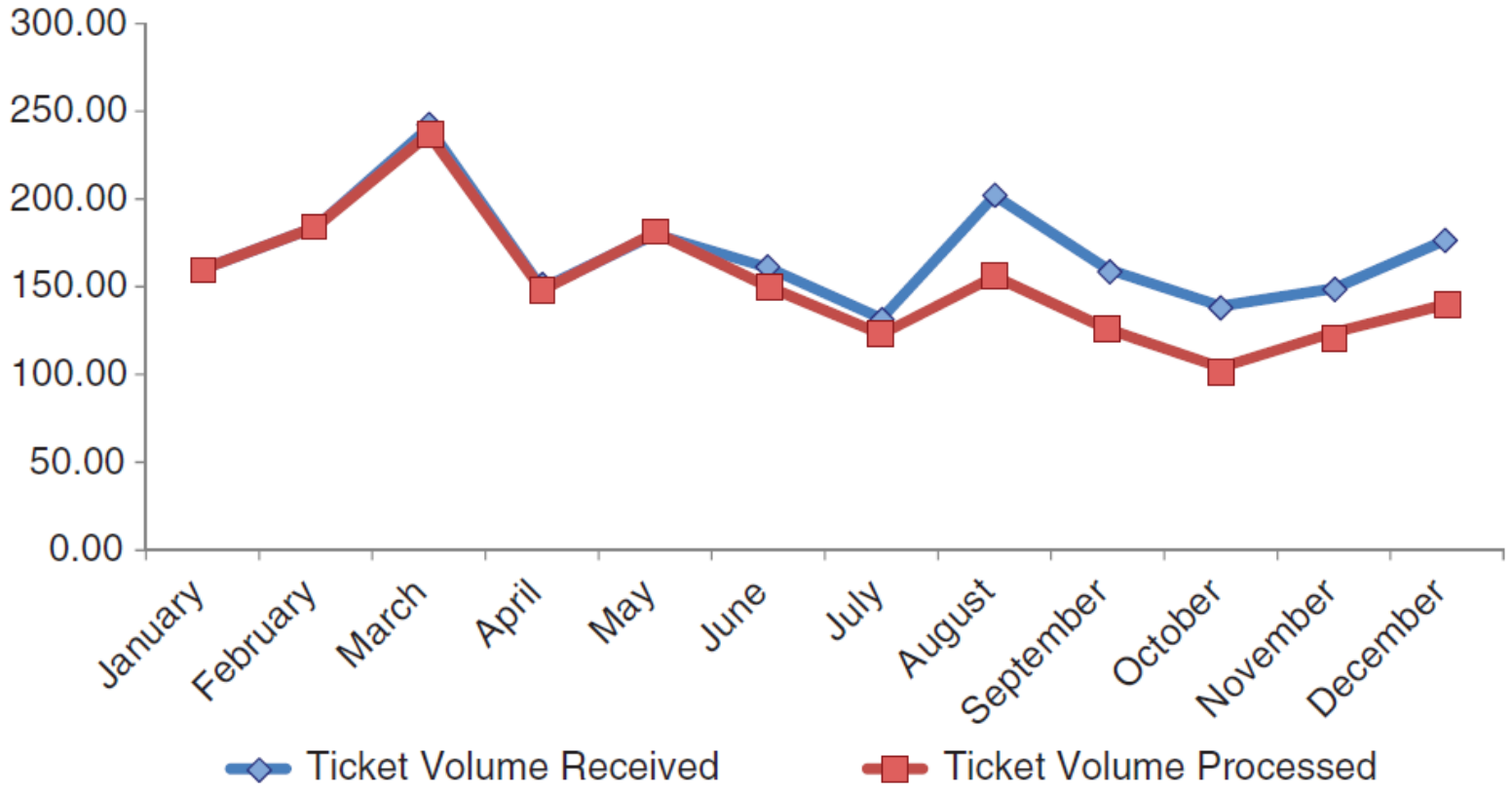
Clutter

- 1. Remove chart border



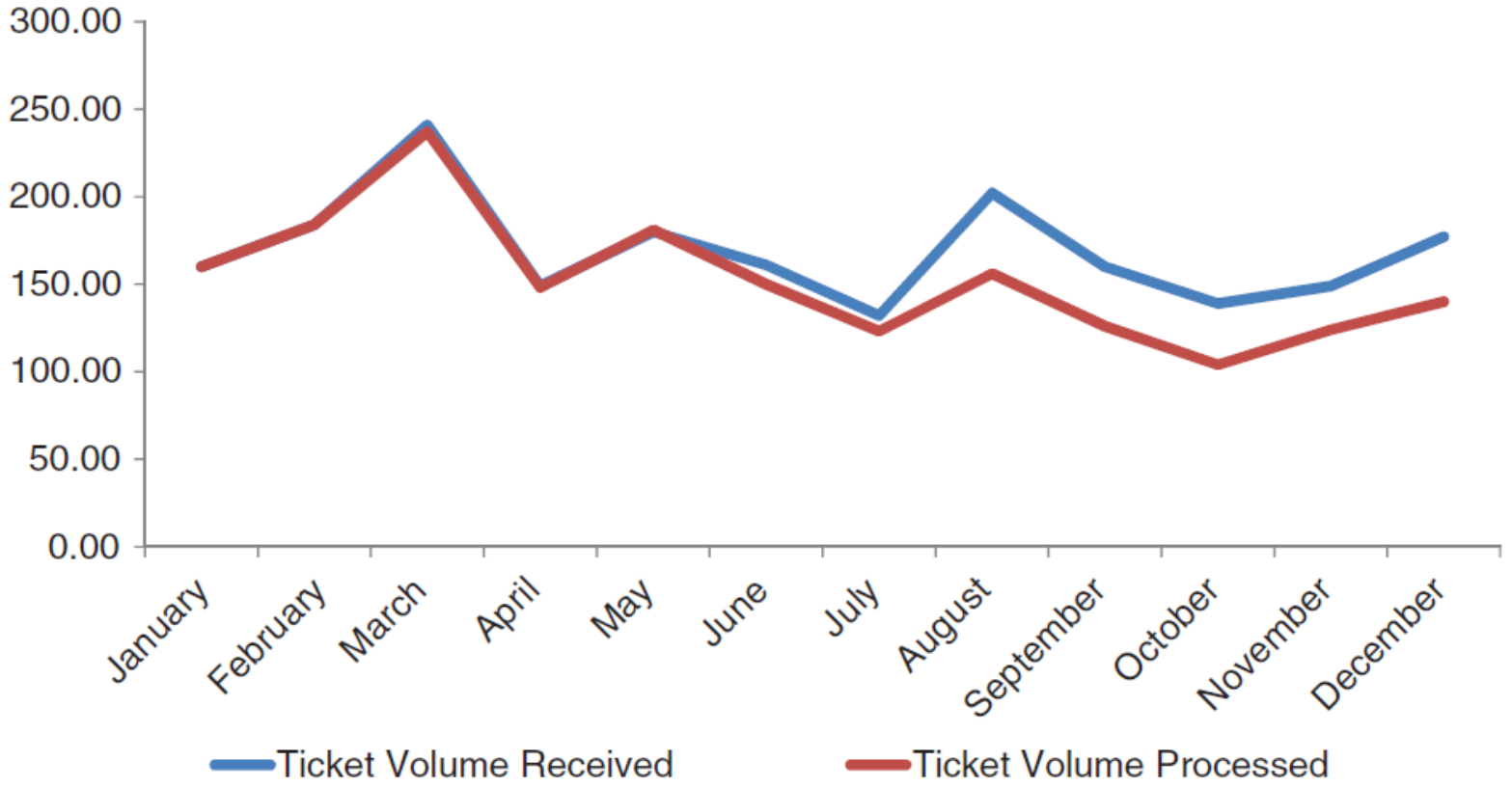
Clutter

- 2. Remove guidelines



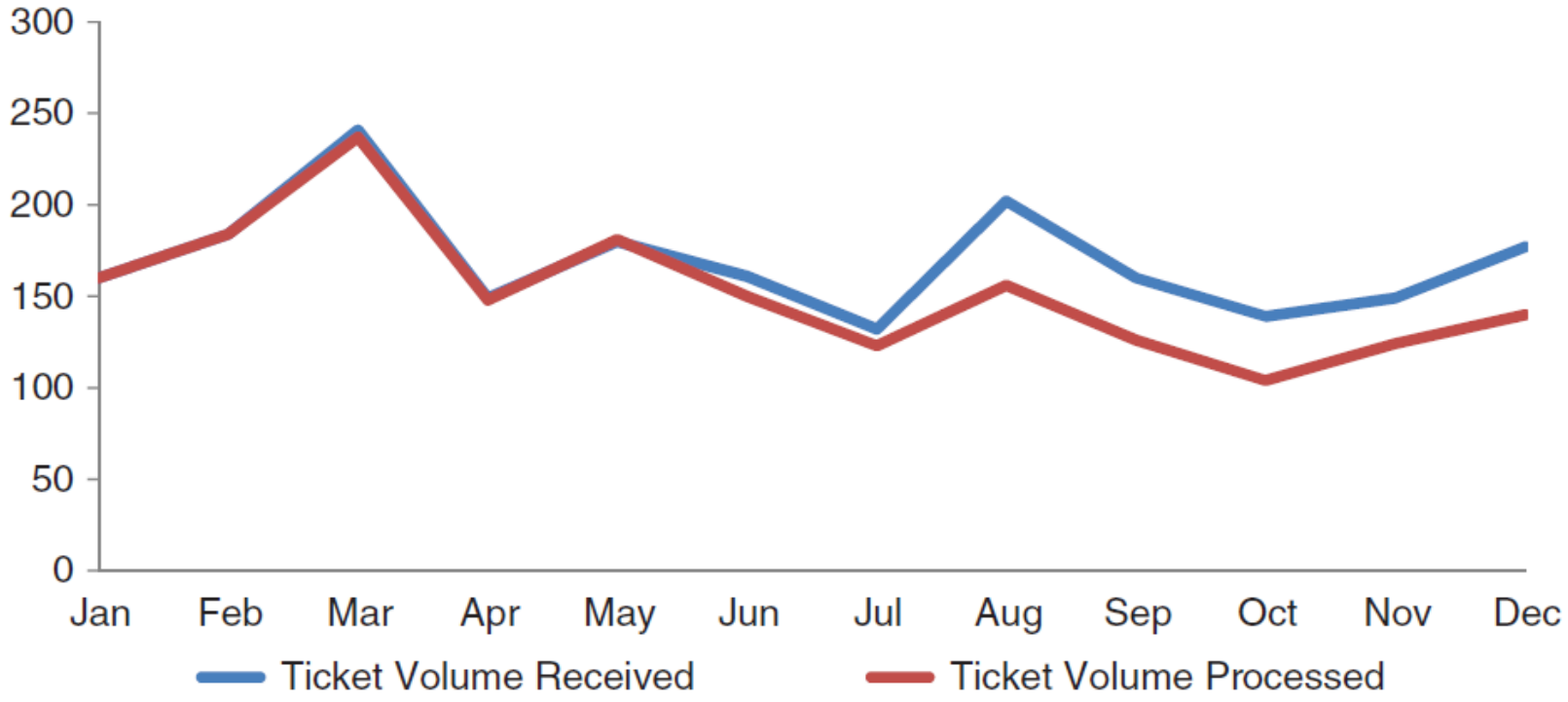
Clutter

- 3. Remove data markers



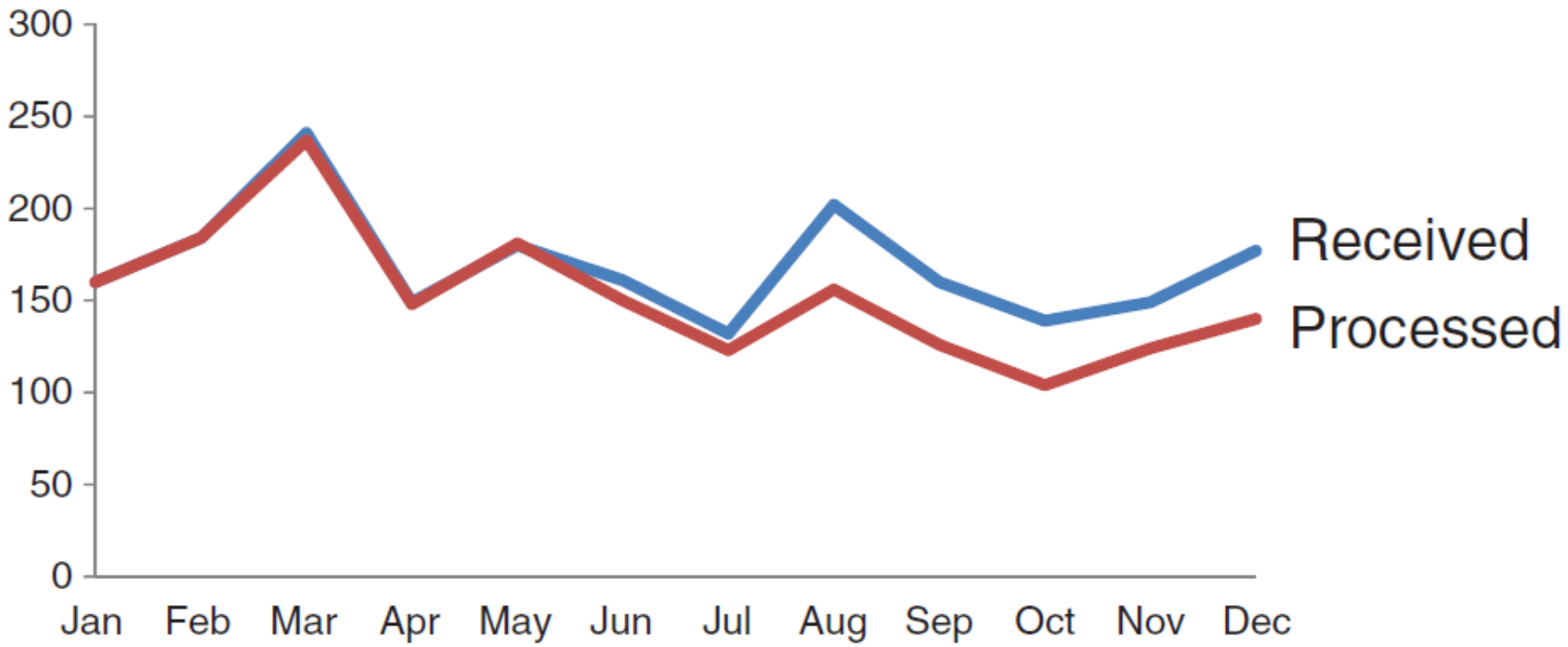
Clutter

- 4. Clean up axis labels



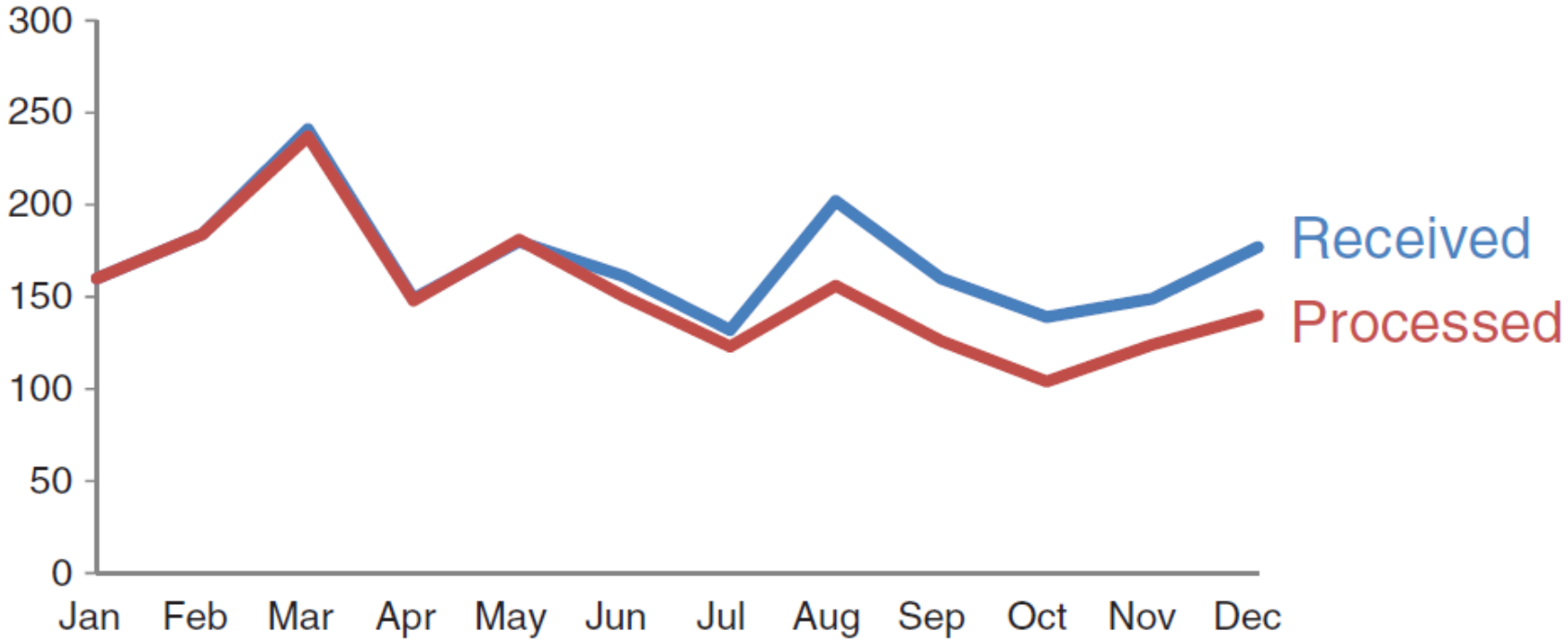
Clutter

- 5. Label data directly



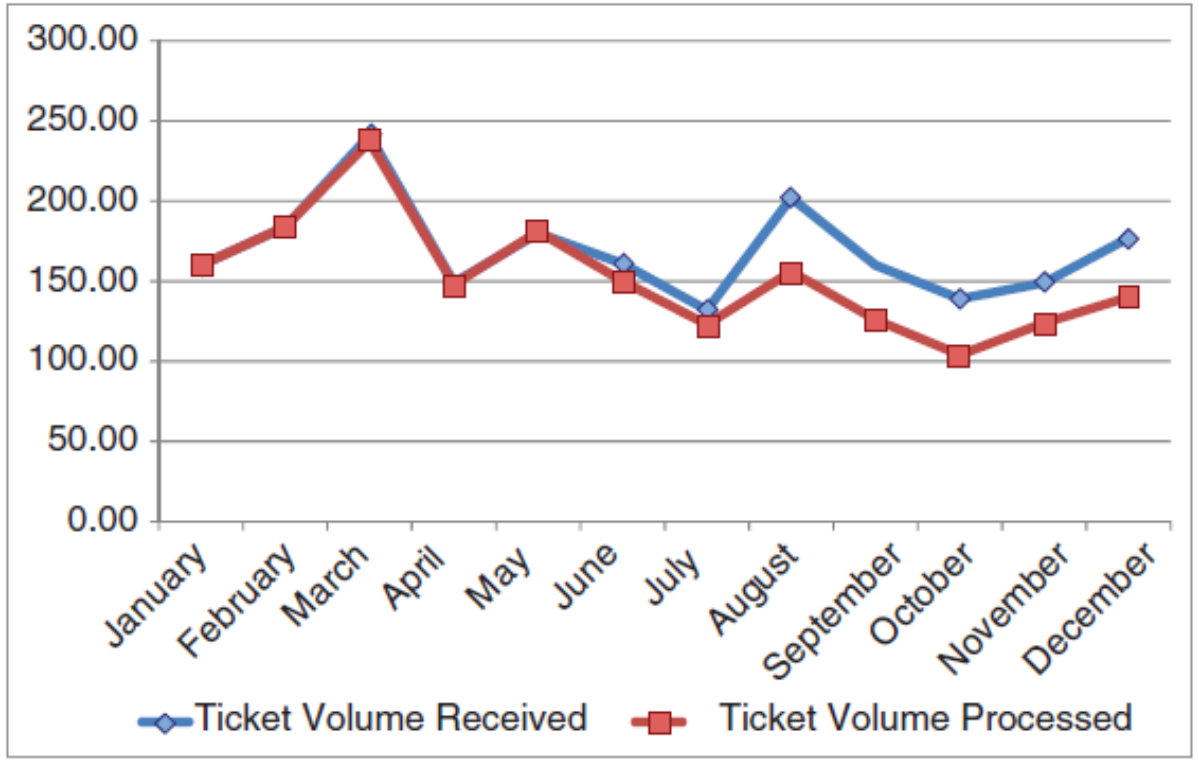
Clutter

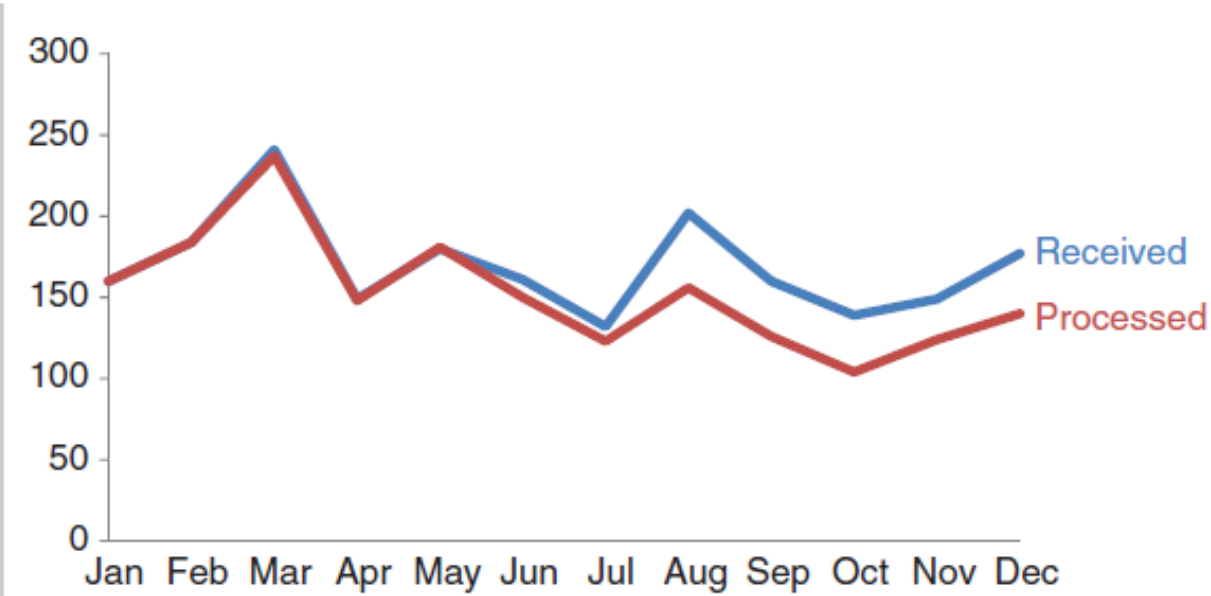
- 6. Leverage consistent color



Clutter

- Before-and-after

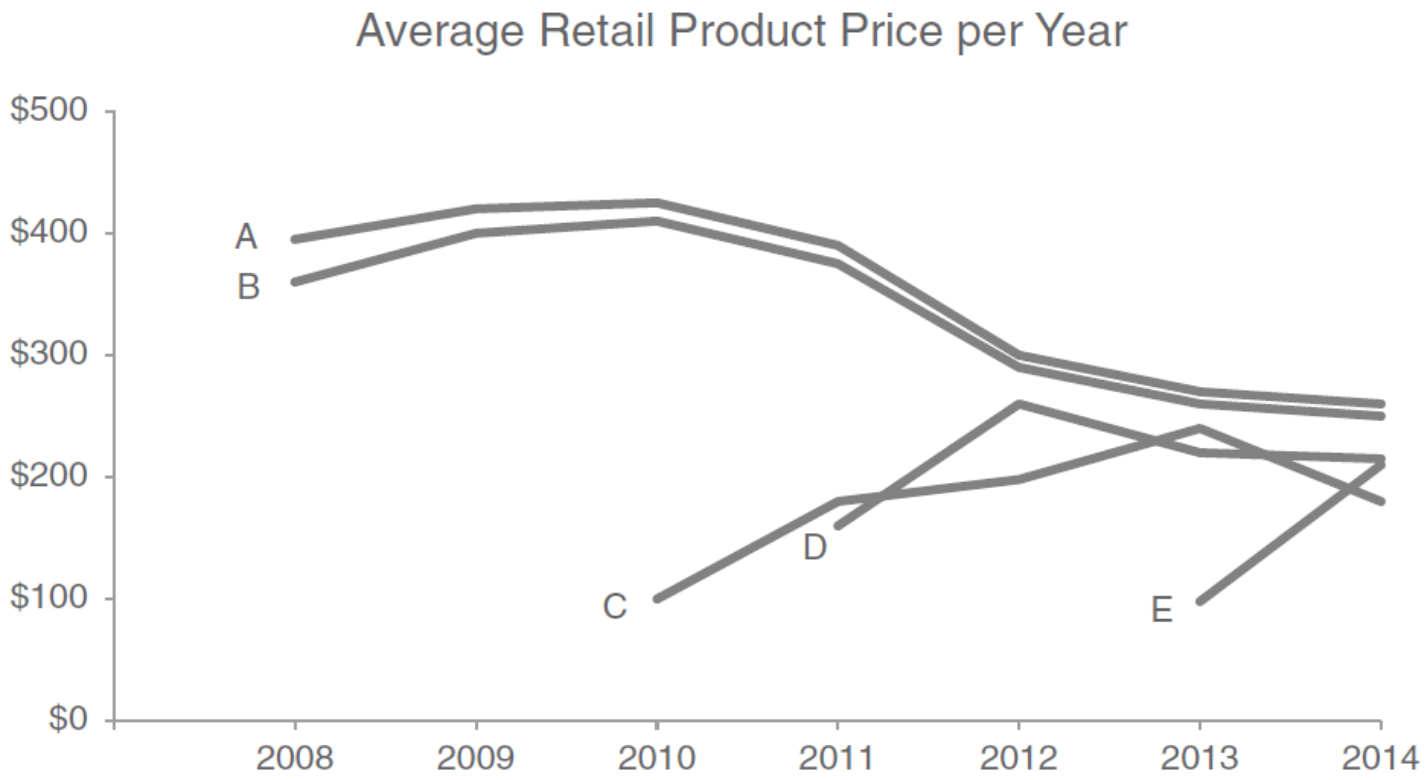




Storytelling with Data

3. Eliminate Clutter

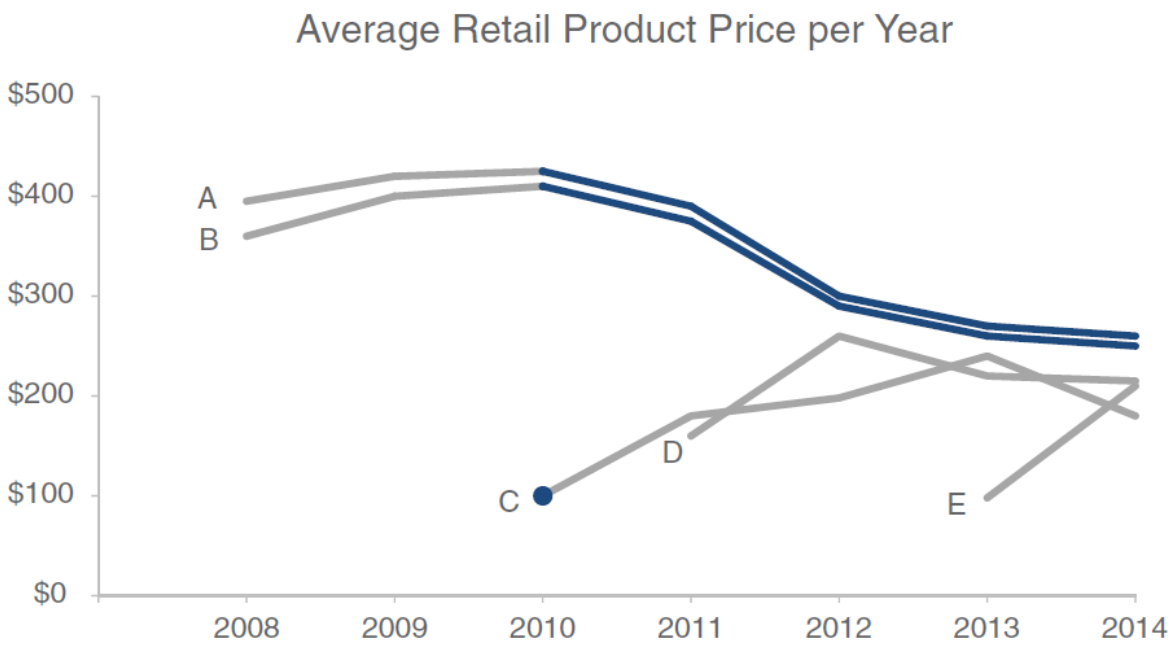
- De-emphasize the chart title
- Remove chart border and gridlines
- Push the x- and y-axis lines and labels
- Remove the variance in colors between the various lines
- Label the lines directly



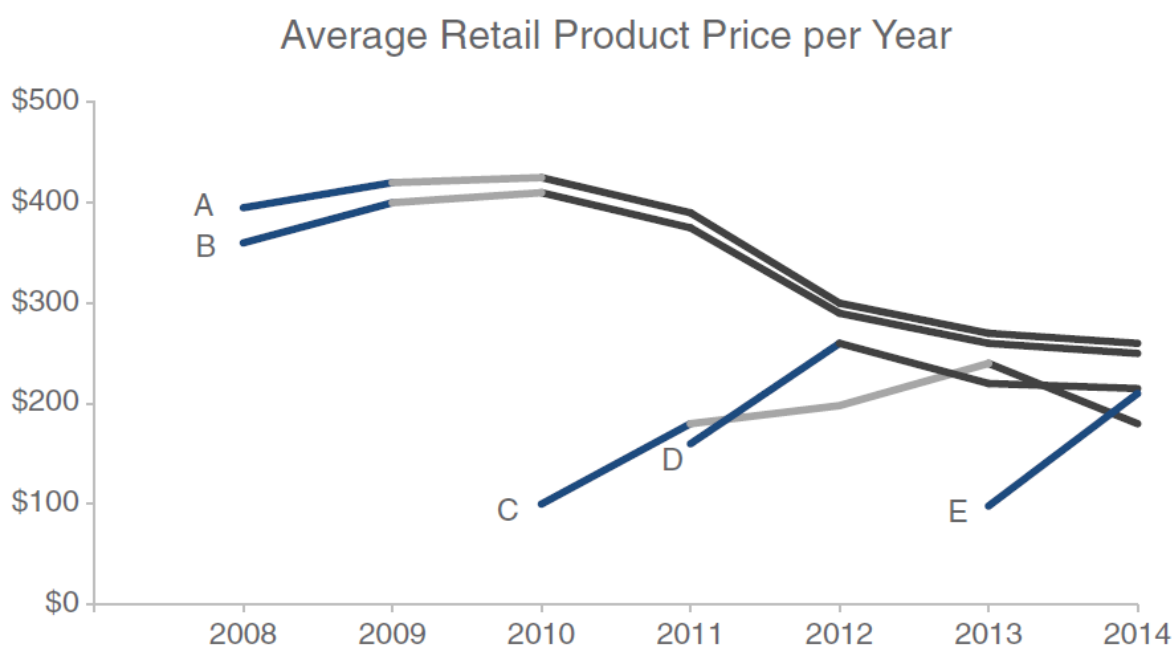
Storytelling with Data

4. Draw attention where you want your audience to focus

"Price has declined for all products on the market since the launch of Product C in 2010."



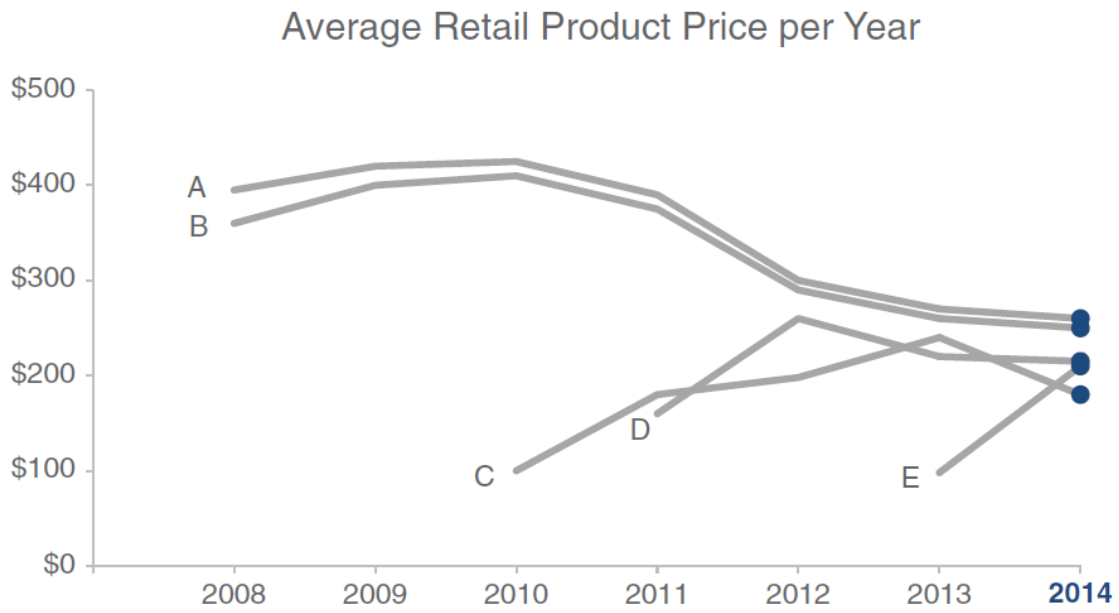
With the launch of a new product in this space, it is typical to see an initial average retail price increase, followed by a decline.



Storytelling with Data

4. Draw attention where you want your audience to focus

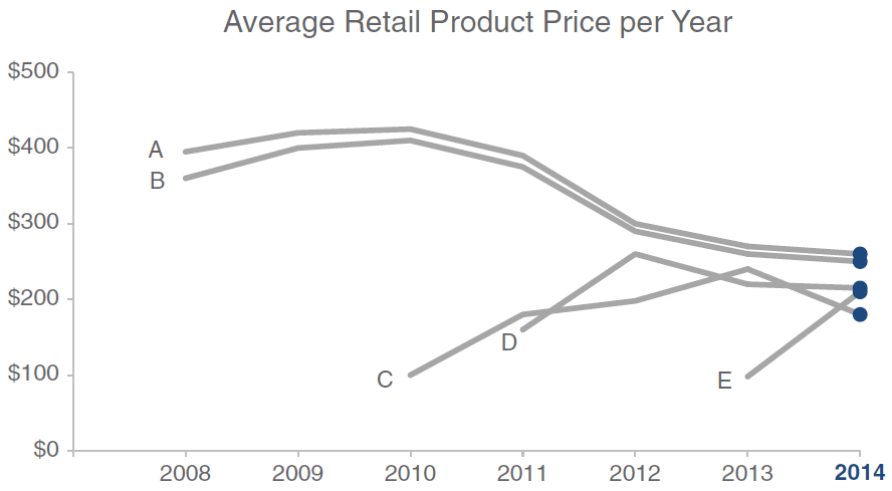
"As of 2014, retail prices have converged across products, with an average retail price of $223, ranging from a low of $180 (Product C) to a high of $260 (Product A)."

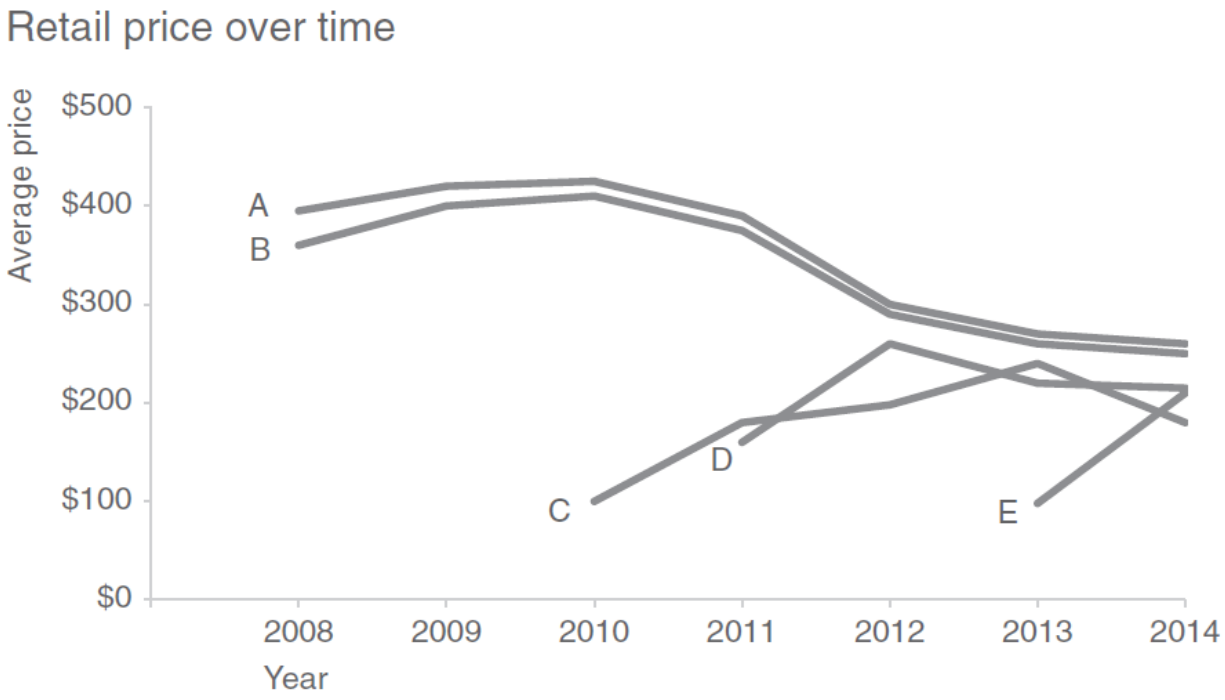


Storytelling with Data

5. Think like a designer

- Make the visual accessible with text.
- Align elements to improve aesthetics.





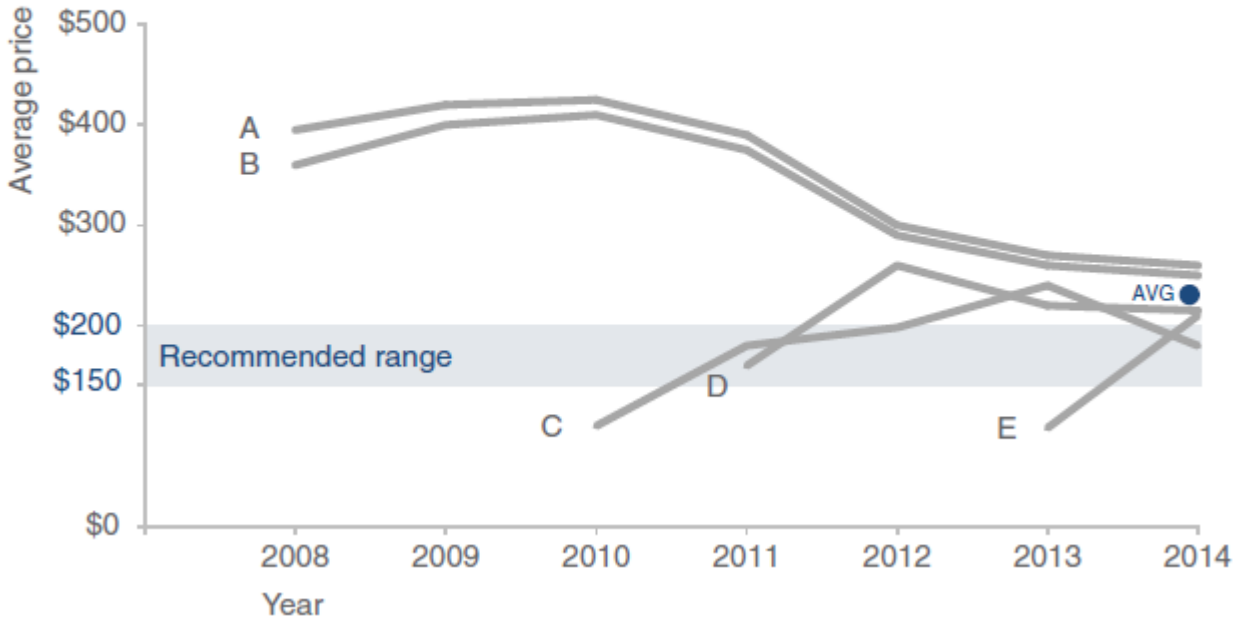
Storytelling with Data

5. Tell a story

Goal:

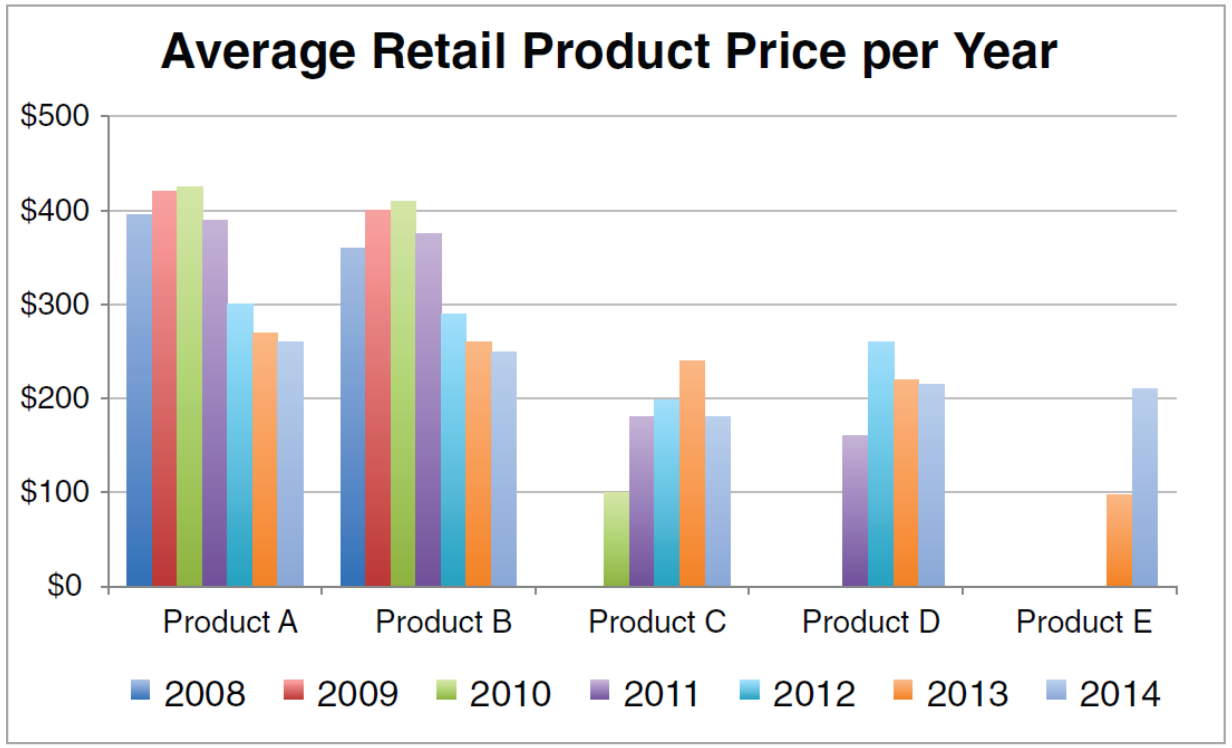
- Understand **how prices have changed over time** in the competitive landscape.
- Use this knowledge to **inform the pricing of our product**

To be competitive, we recommend introducing our product *below the $223 average* price point in the **$150–$200 range**

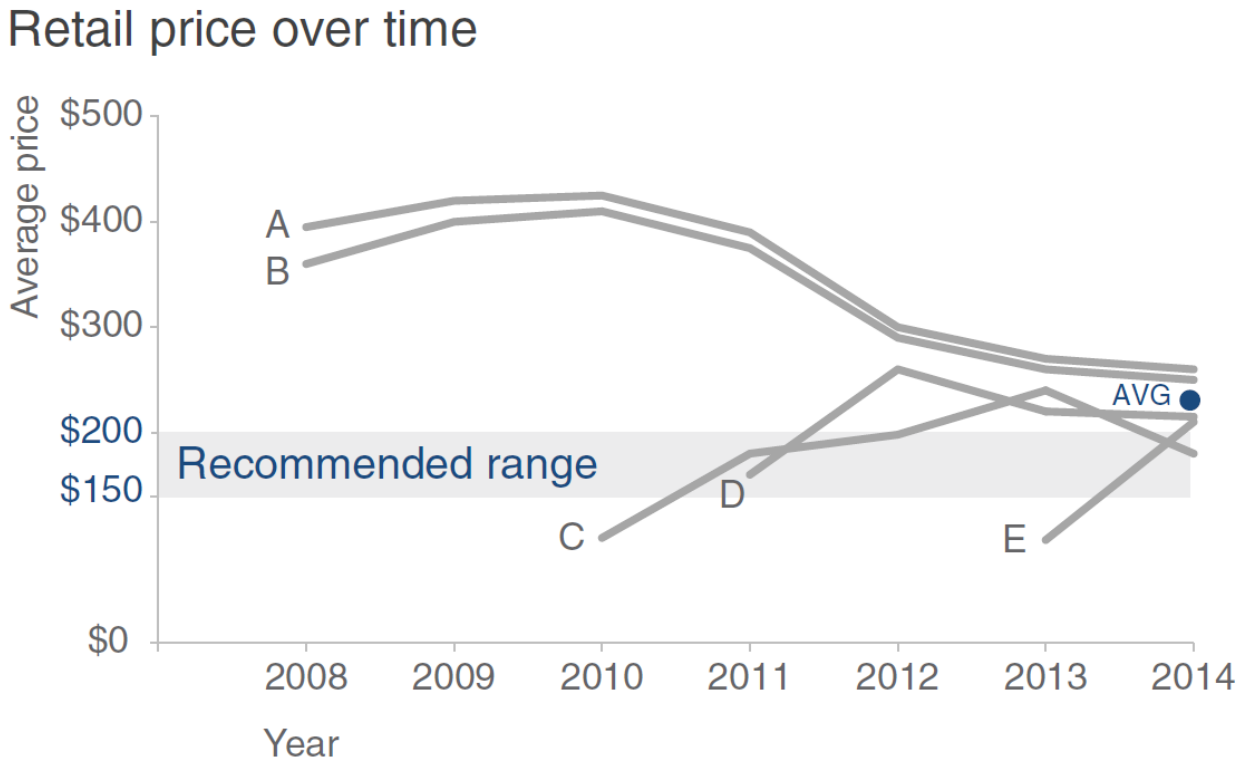


Before-and-after

Price has declined for all products on the market since the launch of Product C in 2010



To be competitive, we recommend introducing our product *below the $223 average* price point in the **$150–$200 range**



Data Visualization Tools

Tableau

Basic version of Tableau data visualization tool is free which can perform regular tasks such as:

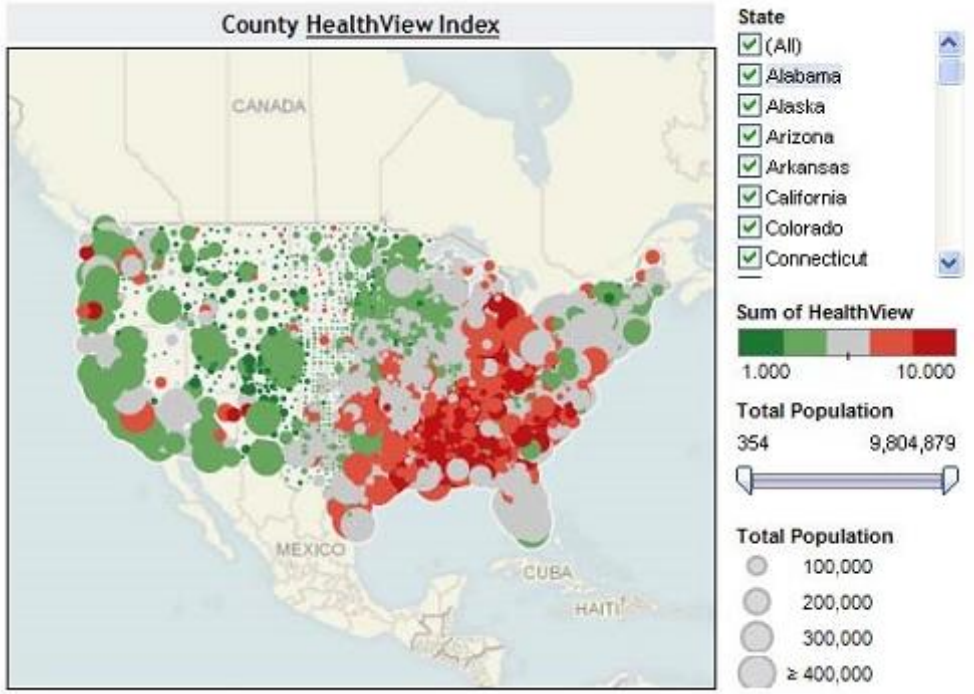


Image Credit :public.tableau.com

1. Data analysis
2. Data monitoring
3. Tracking events
4. Categorizing and sub-categorizing data

Excel

Though excel can accomplish many of the entry level data exploration functions (including heat maps and scatterplots) its limitations are exposed when you want to go beyond the default set of formatting options related to colors, line and styles.

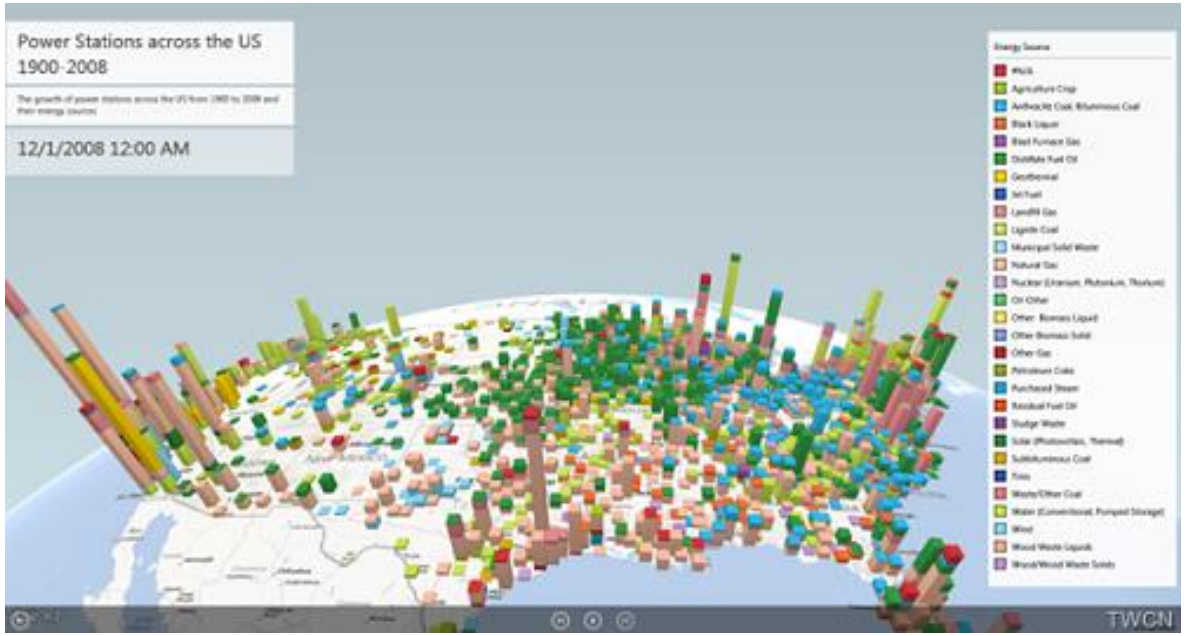
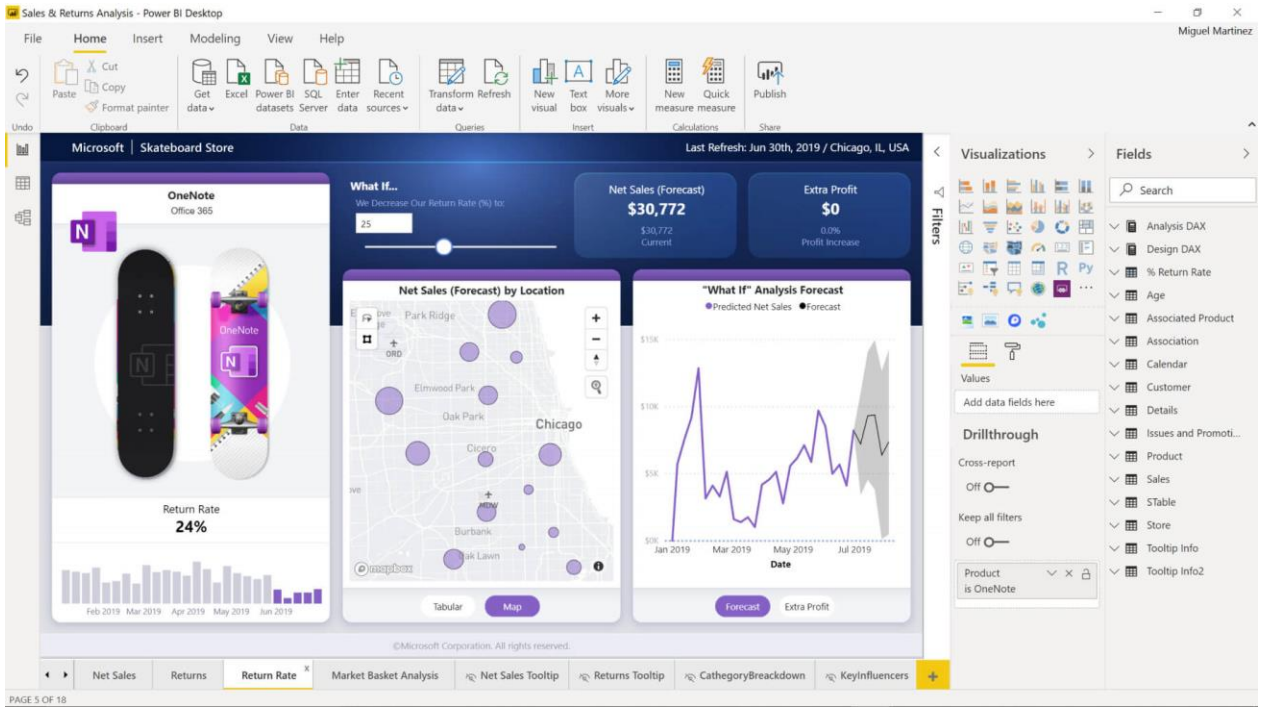


Image Credit : businessintelligence.com

Power BI

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet, or a collection of cloud-based and on-premises hybrid data warehouses.



power bi - Google Search

CartoDB

CartoDB allows easy integration of tabular data with maps. The only downside is that it is free only up to 5 times usage, post which you must pay to use it.

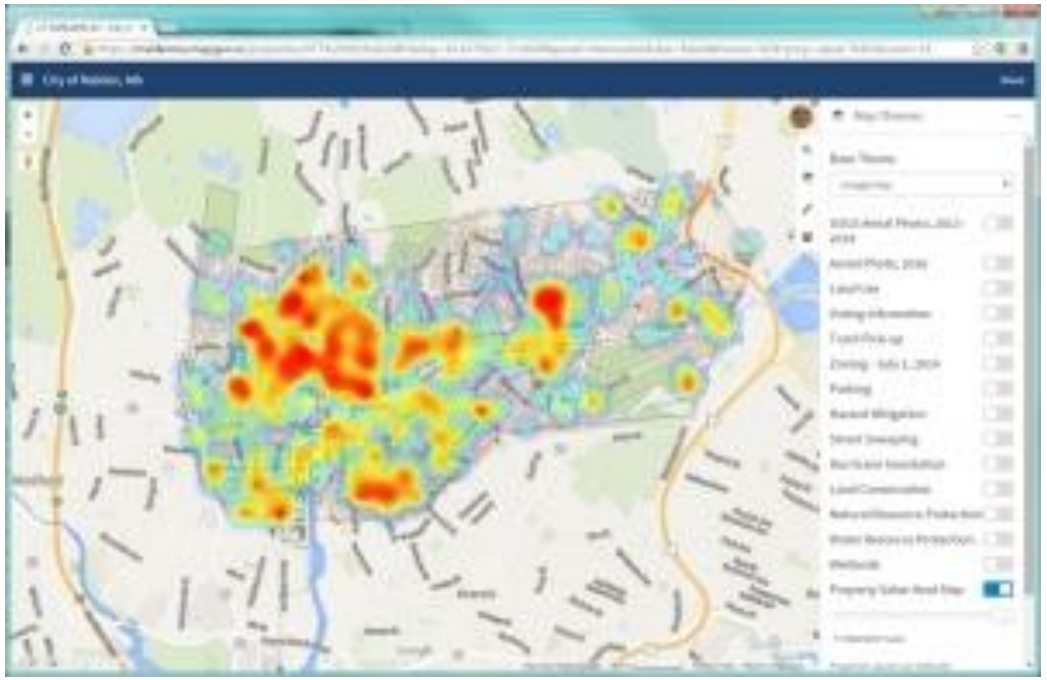
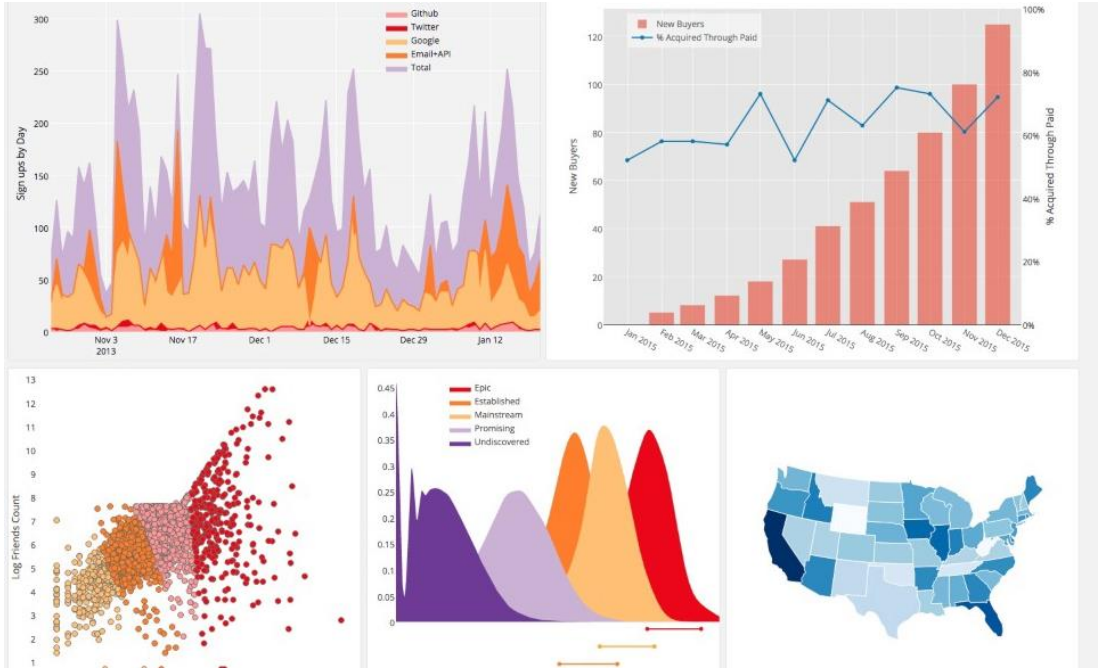


Image Credit : AppGeo.com

Python/matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications

GeoPandas is an open-source project to make working with geospatial data in python easier. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types.



R Studio/Shinny

R Studio: has a separate “diagraphs for R” package that gives the developers an R interface to dygraphs JavaScript charting library which is capable of performing a number of important functions such as time series analysis, graph overlays, etc.

Shiny: It is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

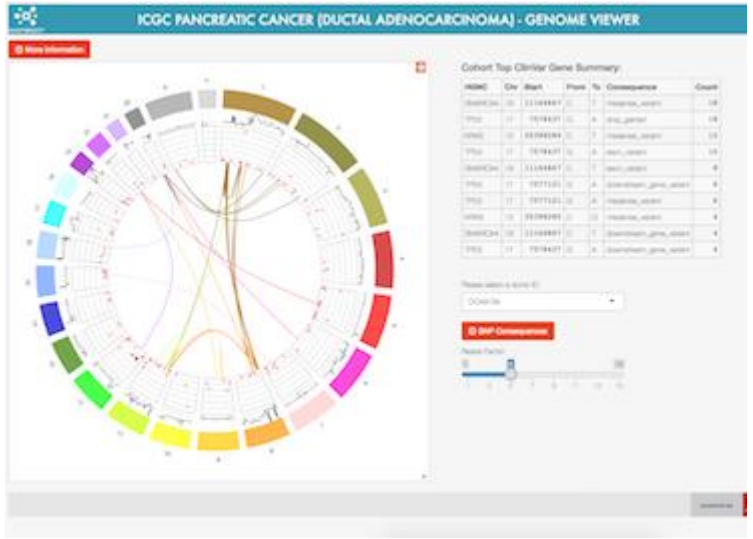


Image Credit : Shiny

Visual.ly

Visually specializes in creating infographics. Just 4 years old, it has succeeded in building up a community of publishers, designers, developers and researchers. Presently they are developing a tool which would allow users to build their own infographic through an automated service.

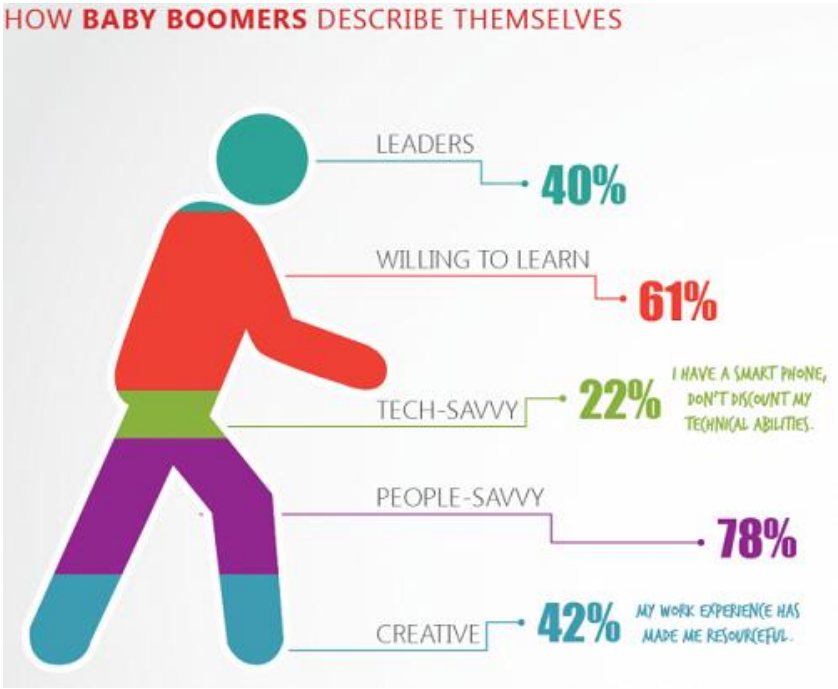
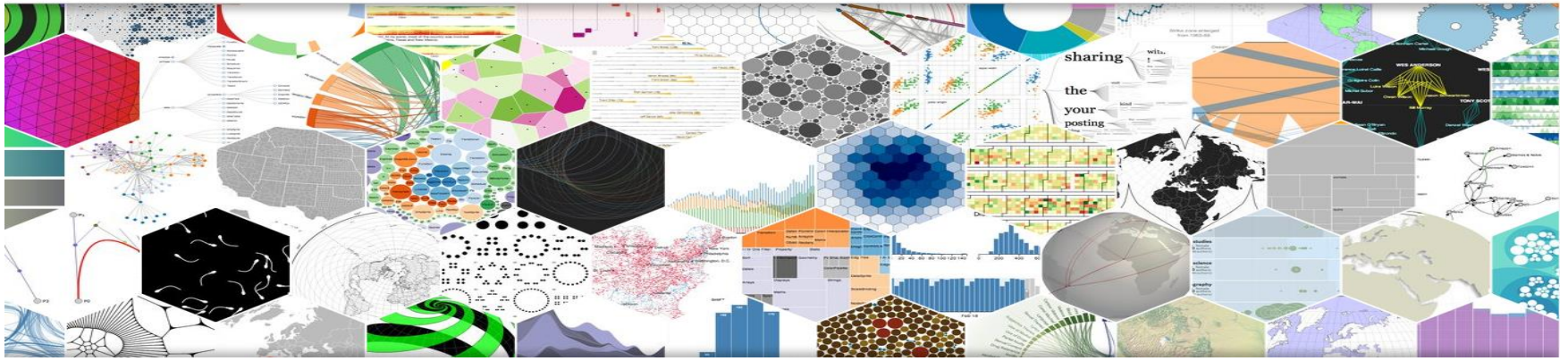


Image Credit: blog.visual.ly

D3.js

Data-Driven Documents



Data Driven Documents gives developers the power to integrate HTML, CSS and SVG into their code to produce highly customized data visualization. Not shipping with standard templates is one of its plus points, as the developers are free to represent data in any manner they want. Being open source, it is always free and there is always an online community to exchange information with. Before you decide on using it, do keep in mind that is has a steep learning curve and there are compatibility issues with older browsers as it is known to work well only with IE9 and above.

Resources

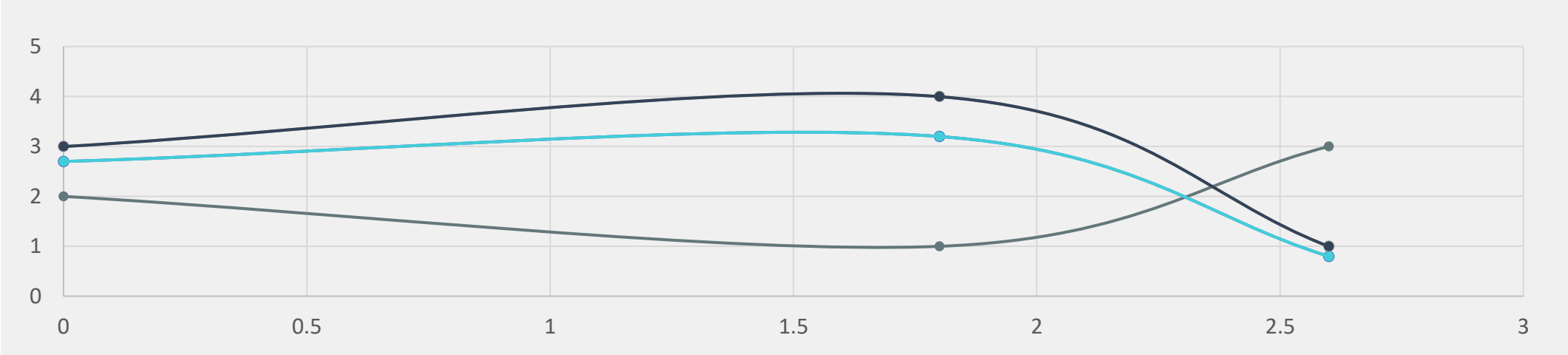
- Datacamp
- Storytelling with data, cole nussbaumer knaflic, Wiley, 2015

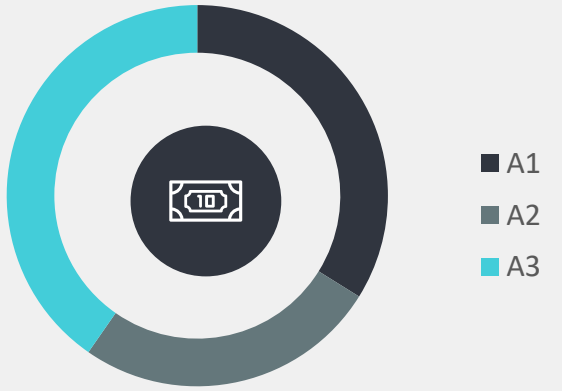

THANK YOU

DATA DRIVEN

Power Point Presentation Template

DATA DRIVEN





Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a

DATA DRIVEN

67%
FINANCING

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

45%
MARKETING

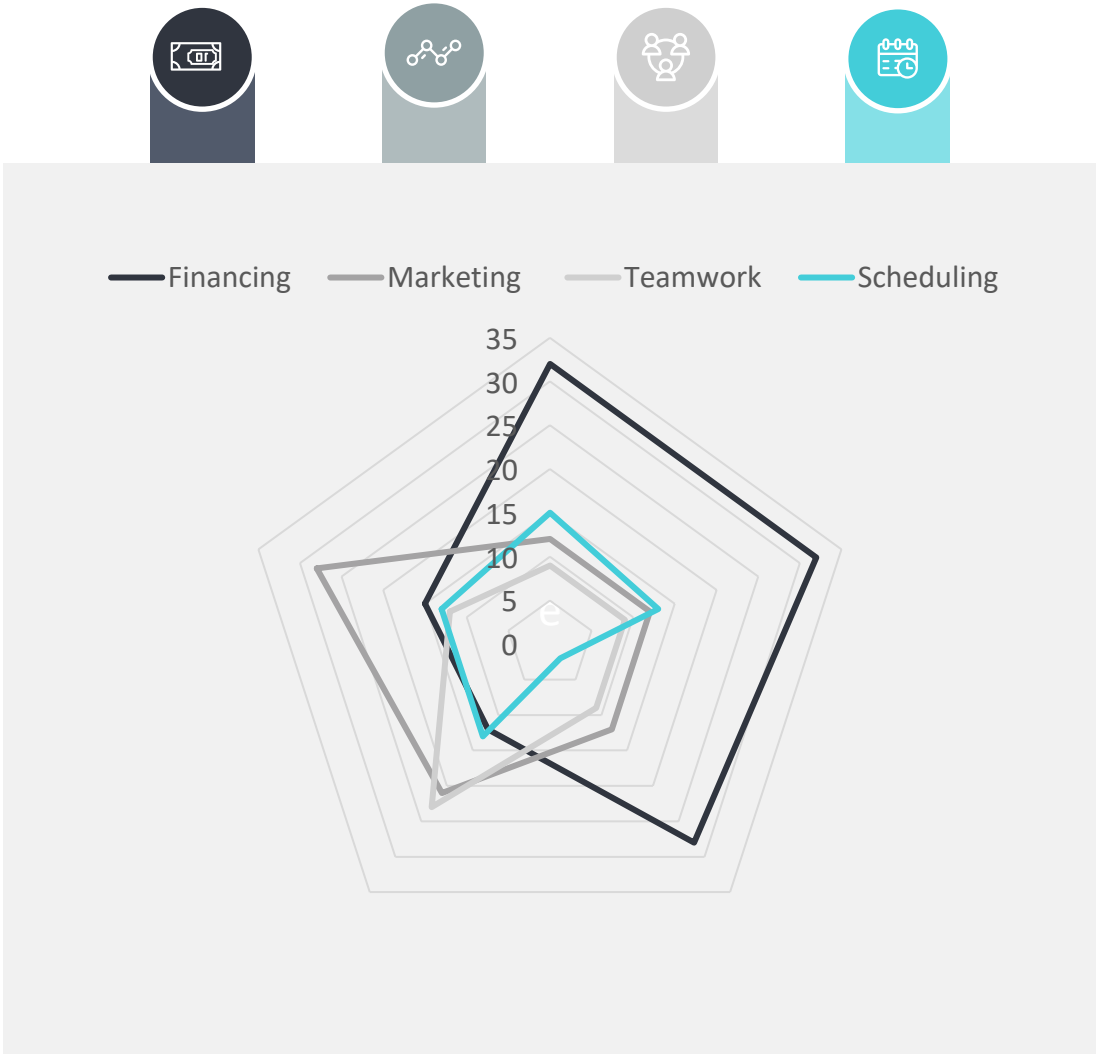
Lorem ipsum dolor sit amet, consectetur adipiscing elit.

31%
TEAMWORK

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

80%
SCHEDULING

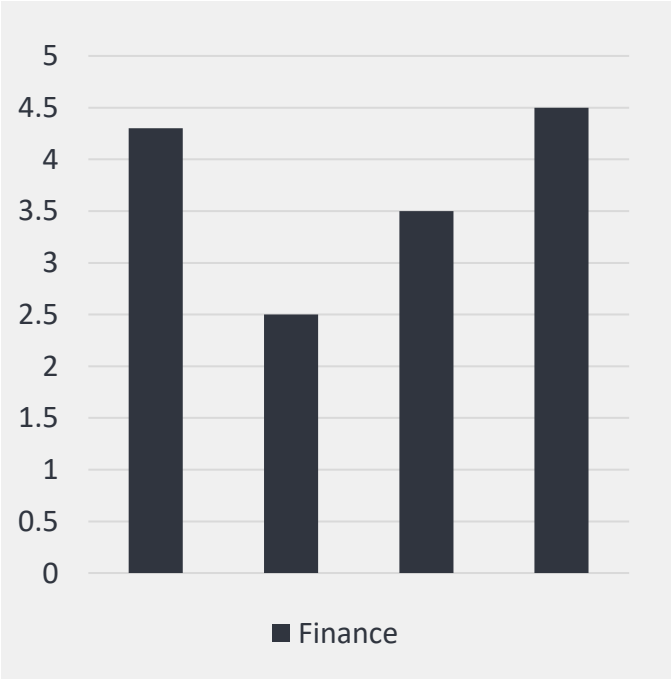
Lorem ipsum dolor sit amet, consectetur adipiscing elit.



Your logo

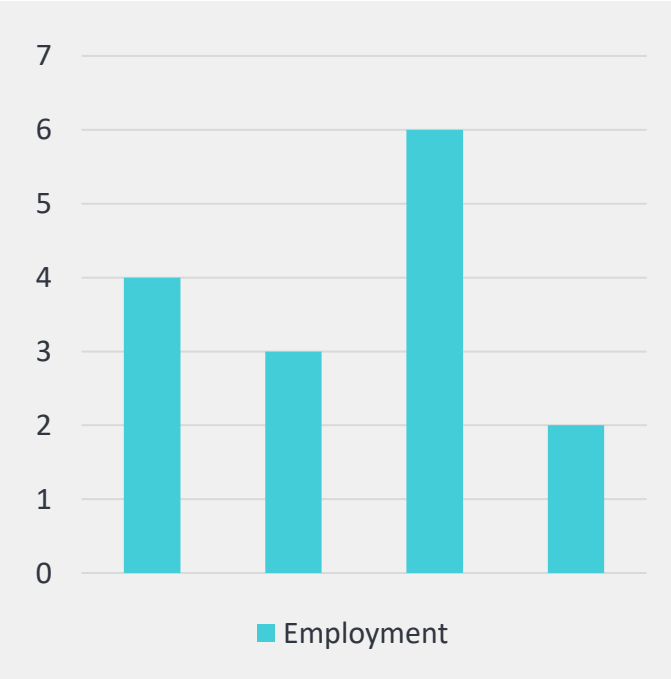
DATA DRIVEN

Lorem ipsum dolor sit amet. 35%



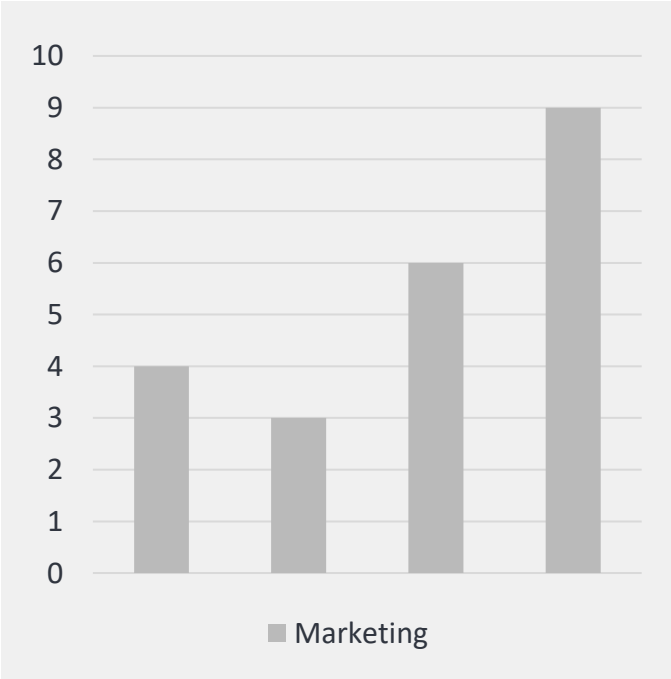
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

Lorem ipsum dolor sit amet. 60%



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

Lorem ipsum dolor sit amet. 73%

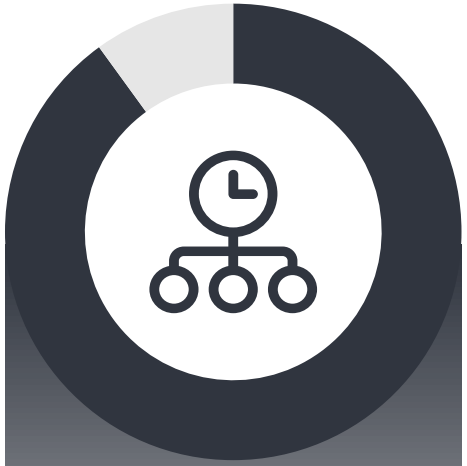


Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

Your logo

DATA DRIVEN

TIME MANAGEMENT



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

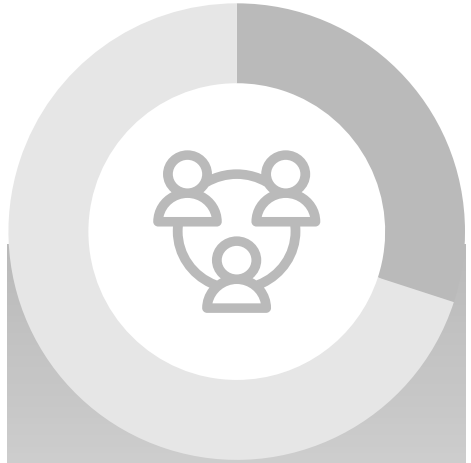
90%

EMPLOYEMENTS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

60%

TEAM CONNECTION



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

30%

DATA DRIVEN

$ 3,577,071

Value Per Capita

Year	Values
2018	- 3.567%
2019	+ 1.007%
2020	+ 2.012%
2021	+ 4,031%

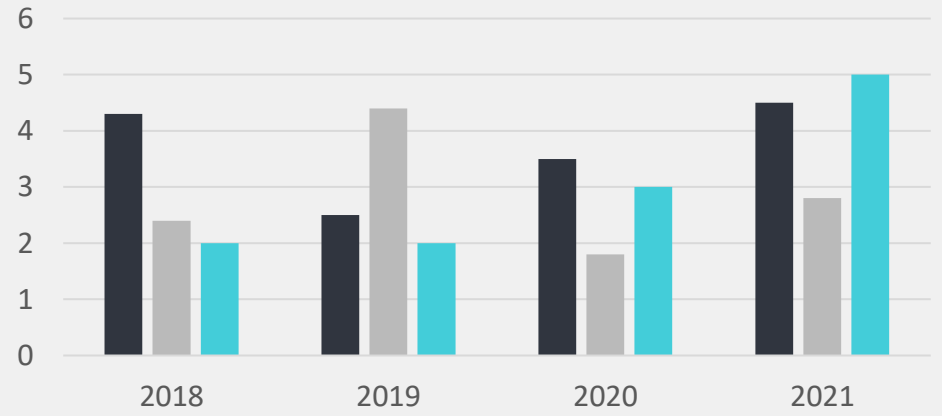
$ 1,672,054

Market Value

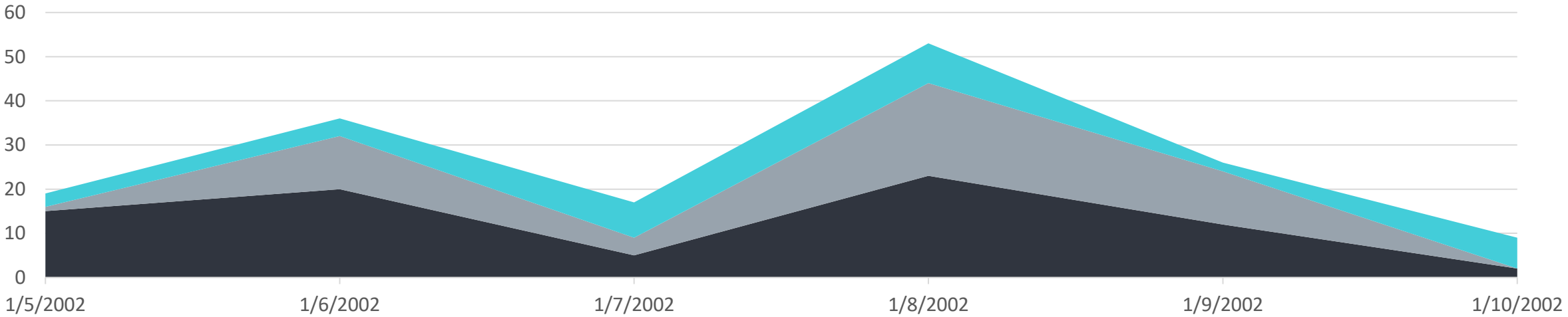
INCOME PERCENTAGE

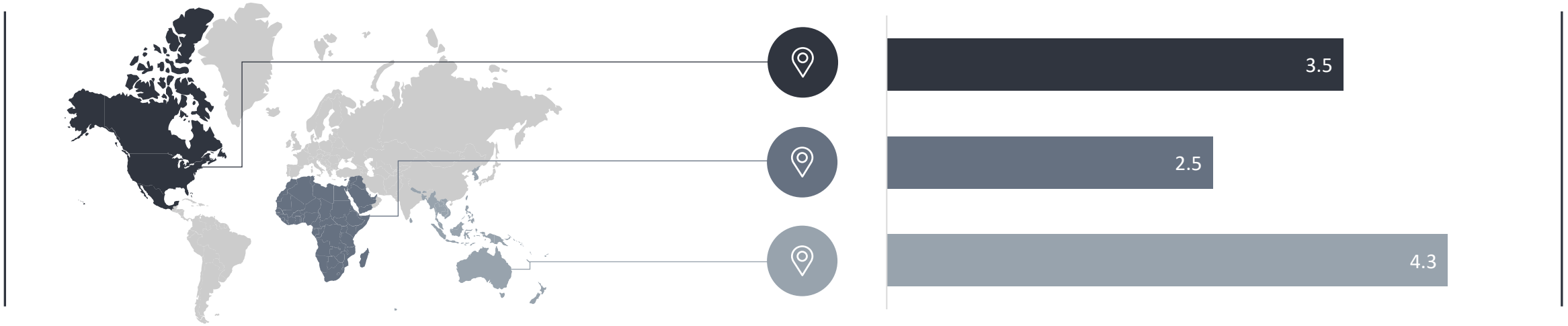
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam vulputate est a convallis placerat. Orci varius

75%



DATA DRIVEN


60
50
40
30
20
10
0
1/5/2002
1/6/2002
1/7/2002
1/8/2002
1/9/2002
1/10/2002


3.5
2.5
4.3

COMPANY TIMELINE

ESTABLISHMENT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

2016

2017

REFINEMENT

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

MARKETING SALES

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

2018

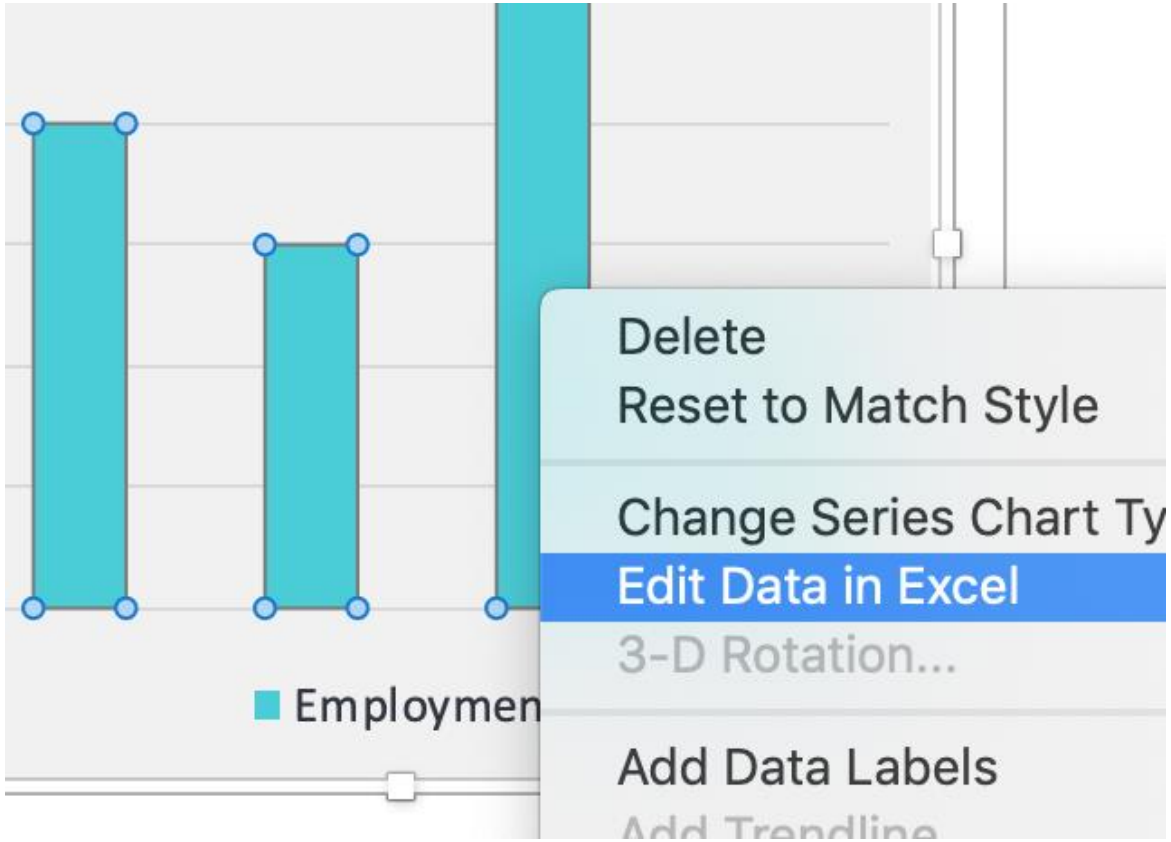
2019

INCREASING INCOME

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam a convallis placerat.

Your logo

GUIDE: EDITING DATA IN THIS TEMPLATE



If you'd like to edit the data in any of the graphs, tables or data-diagrams in this template, just right click on the graph and select *Edit Data in Excel*.