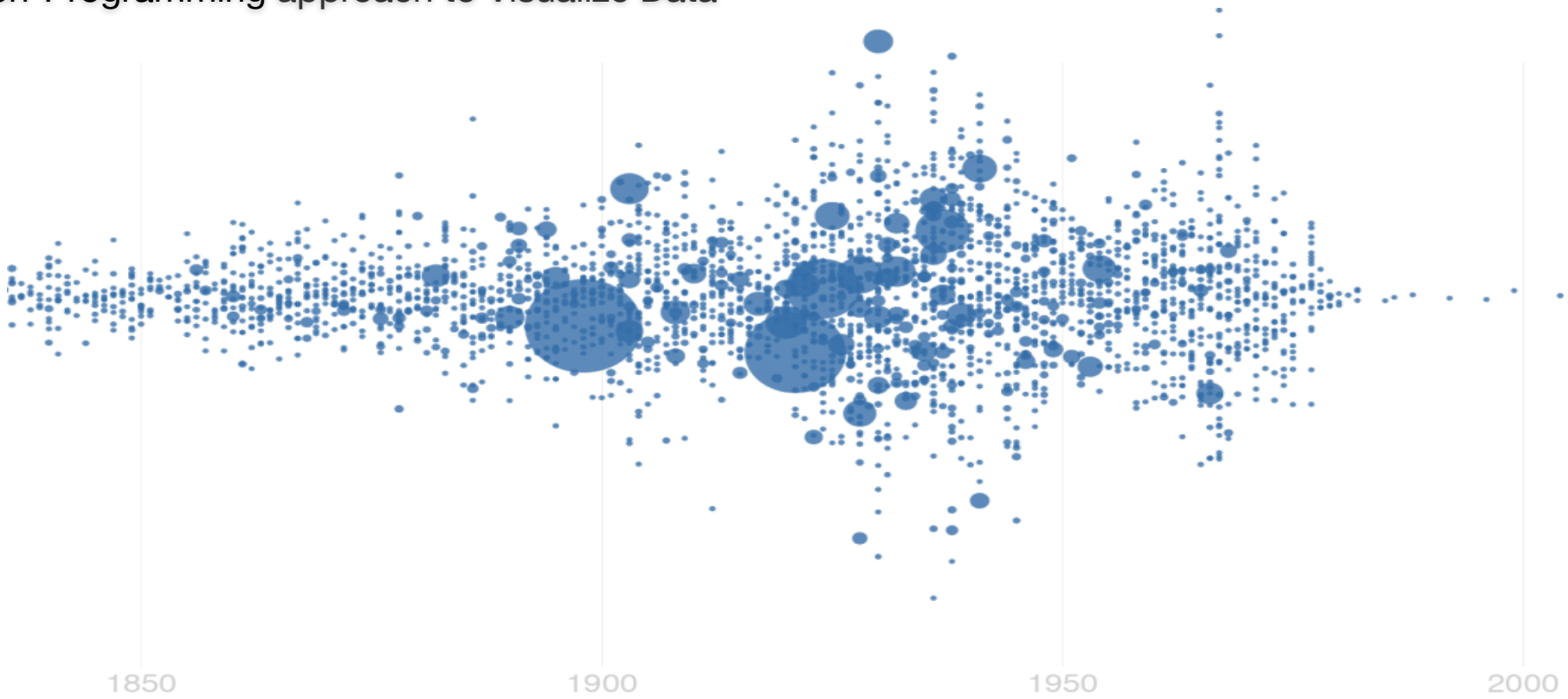
Data Visualization

Non-Programming approach to Visualize Data



Dr. Omer Ayoub
Senior Data Scientist,
House of Mathematical and Statistical Sciences,
King Abdul Aziz Univerrsity, Jeddah, Saudi Arabia

The Abdus Salam International Centre for Theoretical Physics

The World Academy of Sciences
for the advancement of science in developing countries

CODATA

RESEARCH DATA ALLIANCE

Dr. Omer Ayoub

Ph.D in Computer Science (USA)

ICTP Associate

Senior Data Scientist

House of Mathematical Sciences, Consulting Firm

King Abdul Aziz University, Jeddah,

Saudi Arabia

Email: omerayoub@hotmail.com omer@statisticalview.com

Content

1

1. Introduction to Data Visualization
2. What is non-programming approach?
3. How to benefit from this workshop?
4. Data Openness and Open Access policy

2

1. Which type of visual design should I select to present my findings?
2. Chart types and Design best practices

3

1. An idea and discussion about Next sessions
2. Getting yourself ready with the tools to practice

4

1. Questions and Answers Session

5

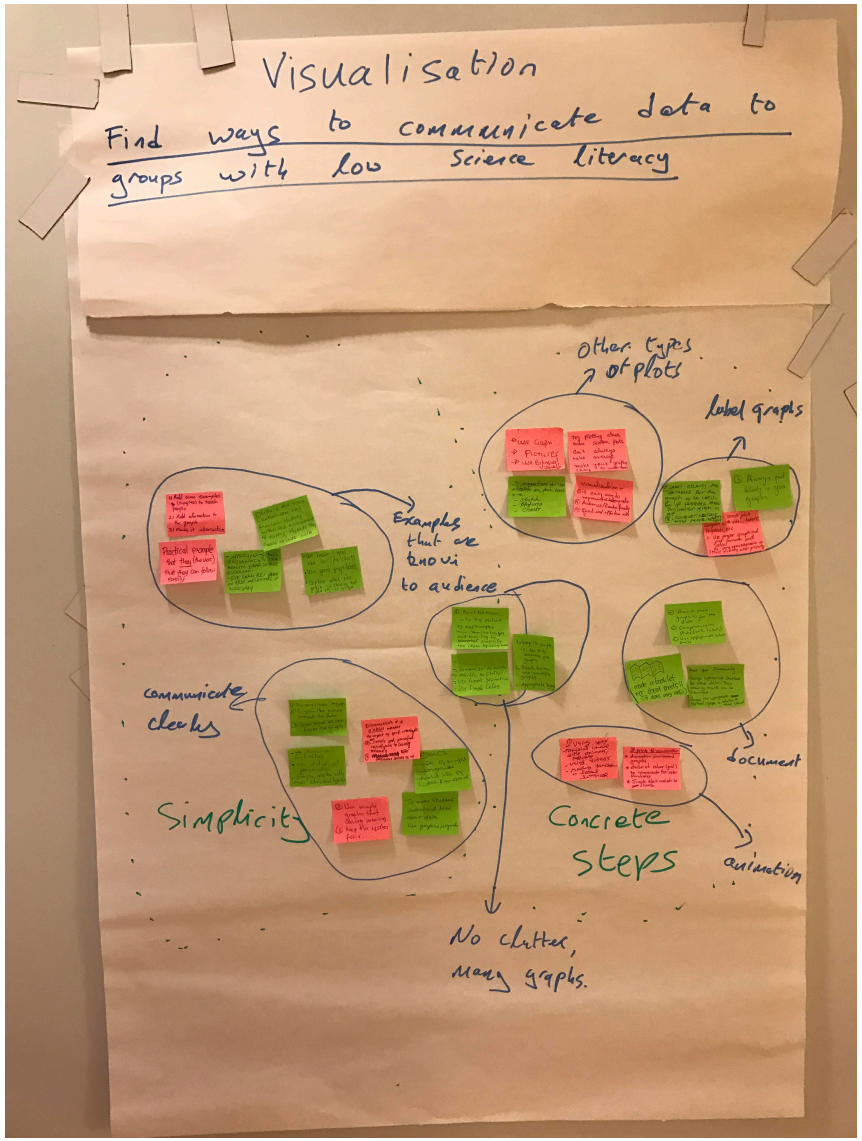
1. Wrap Up

First CoDATA - RDA Summer School Participants in ICTP - 2016

Your contribution to your society …

- Self-assessment questions:
  - How do you plan to contribute to your society in terms of applying the methodologies and practices learnt during this summer school?
  - Any plans to do something for Open data access?
  - Any thoughts on following standardized procedures to overcome the barriers in data sharing?

Feedback and Suggestion



Visualization

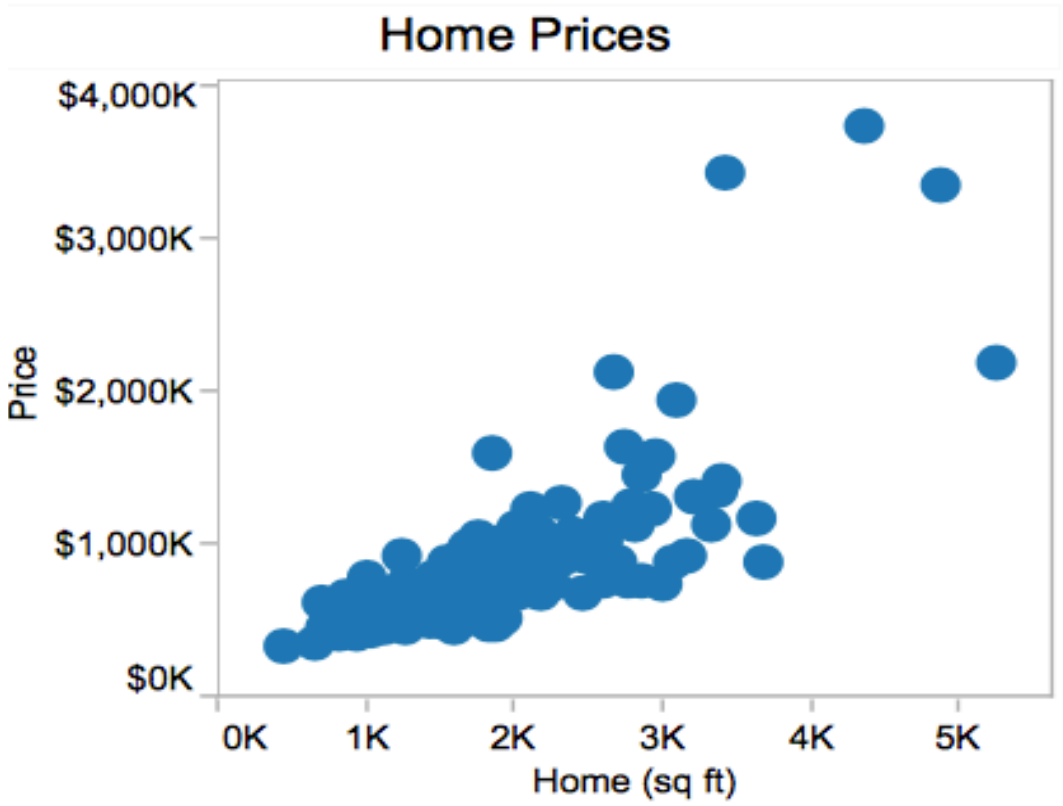
"Numbers have an important story to tell. They rely on you to give them a clear and convincing voice."

– Stephen Few,

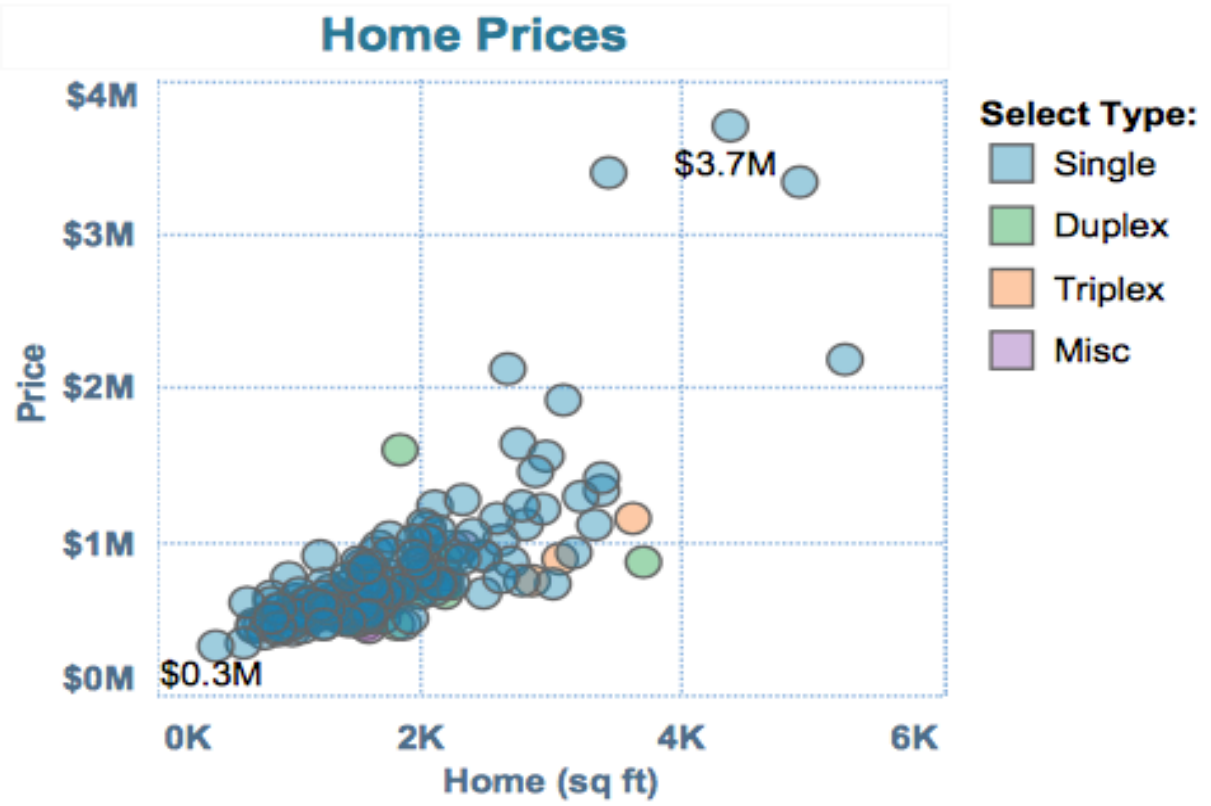
Now You See It: Simple Visualization Techniques for Quantitative Analysis

Visualization

Good visualization



Great visualization



Finding the Story in your Data

- Information can be visualized in a number of ways, each of which can provide a specific insight.
- When you start to work with your data, it's important to identify and understand the story you are trying to tell and the relationship you are looking to show. Knowing this information will help you select the proper visualization to best deliver your message.

TRENDS

Ice Cream sales over time

CORRELATIONS

Ice Cream sales vs. Temperature

OUTLIERS

Ice cream sales in an unusual region

KNOW YOUR DATA

Before understanding visualizations, you must understand the types of data that can be visualized and their relationships to each other. Here are some of the most common you are likely to encounter.

Data Types

QUANTITATIVE
Data that can be counted or measured; all values are numerical.

DISCRETE
Numerical Data that has a finite number of possible values. Example: number of employees in the office

CONTINUOUS
Data that is measured and has a value within a range. Example: Rainfall in a year.

CATEGORICAL
Data that can be stored according to group or category. Example: Types of products sold

Data Relationships

NOMINAL COMPARISON

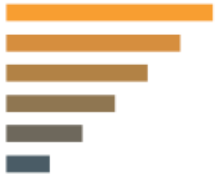
This is a simple comparison of the quantitative values of subcategories. Example: Number of visitors to various websites.

TIME SERIES

This tracks changes in values of a consistent metric over time. Example: Monthly sales etc.

CORRELATION

This is data with two or more variables that may demonstrate a positive or negative correlation to each other. Example: Salaries according to education level.



RANKING

This shows how two or more values compare to each other in relative magnitude. Example: Historic weather patterns ranked from the hottest months to the coldest.

DEVIATION

This examines how data points relate to each other, particularly how far any given data point differs from the mean.
Example: Amusement park tickets sold on a rainy day vs. a regular day.

DISTRIBUTION

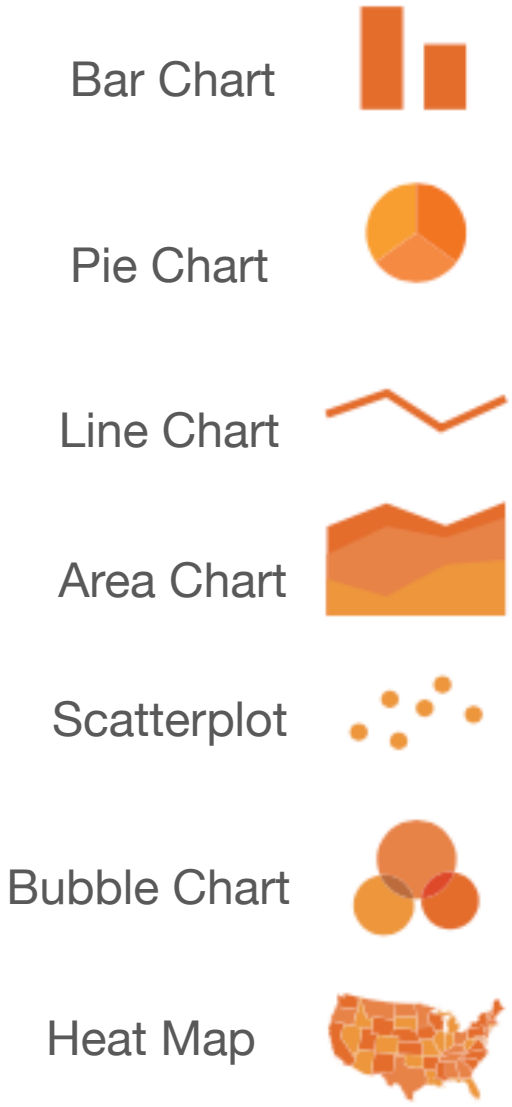
This shows data distribution, often around a central value. Examples: Heights of players in Basketball team

PART-TO-WHOLE RELATIONSHIPS

This shows a subset of data compared to the Larger whole. Example: Percentage of customers purchasing specific products etc.

Chart Types

This section addresses about most common chart types that are usually used for Visualization. Furthermore, we will discuss about the best practices to use these chart types:

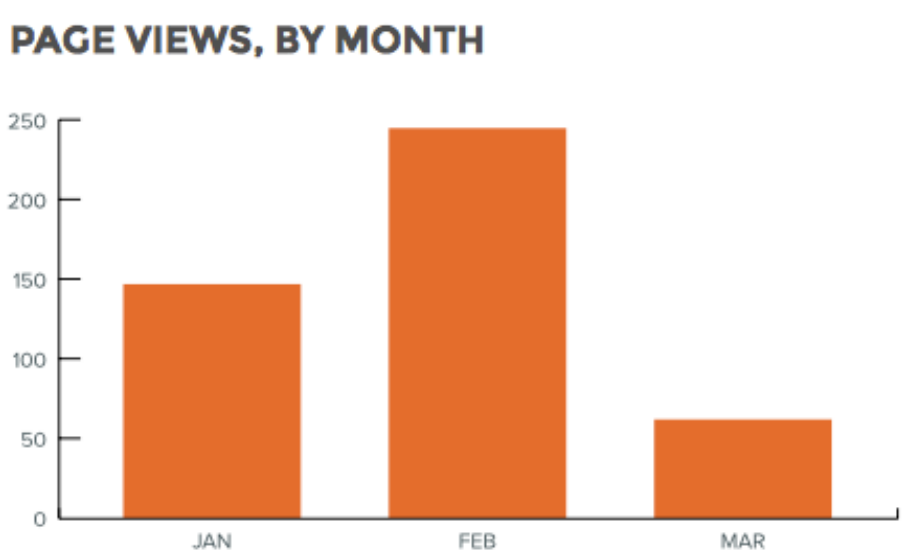


Bar Charts

Bar charts are very versatile. They are best used to show change over time, compare different categories, or compare parts of a whole.

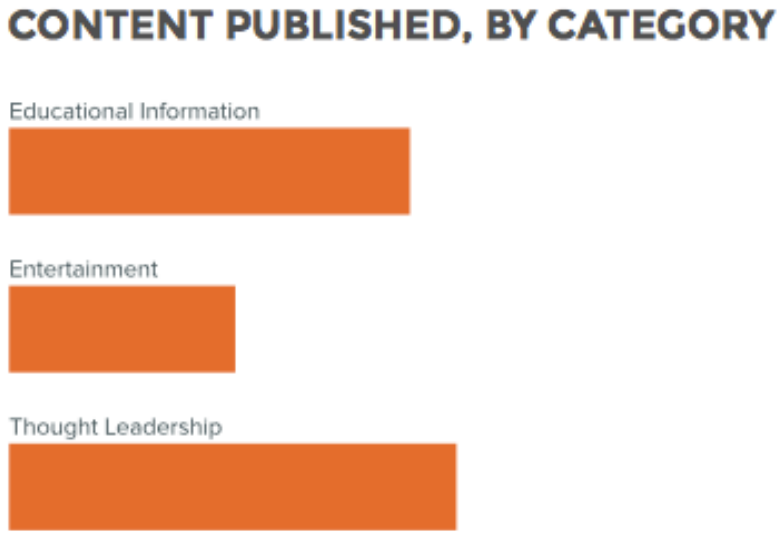
Common Bar chart variations include Stacked, 100% stacked versions. Usually these variations are used to compare multiple part-to-whole relationships. i.e. Monthly online traffic analysis by different sources.

Bar Charts Variations



VERTICAL
(Column Chart)

It is best used for chronological data (time-series should always run left to right) or when visualizing negative values below the axes.



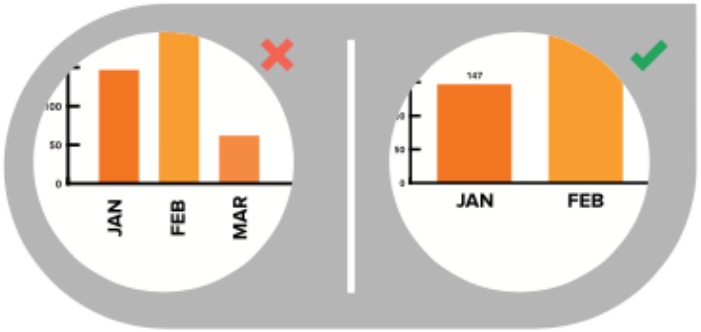
HORIZONTAL

It is best used when data with long categories are to be labelled

Bar Charts

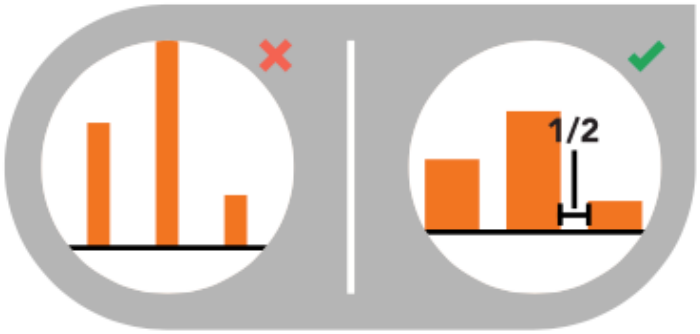
Best Practices

Bar Charts Design Best Practices



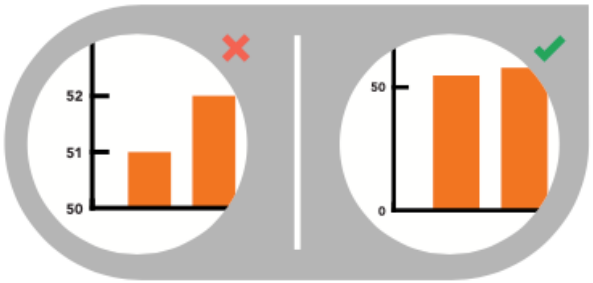
Use Horizontal Labels

Avoid steep diagonal or vertical type, as it can be difficult to read



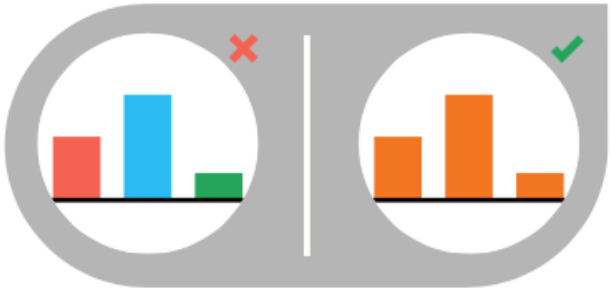
Space Bars Appropriately

Space between the bars should be at least ½ bar width



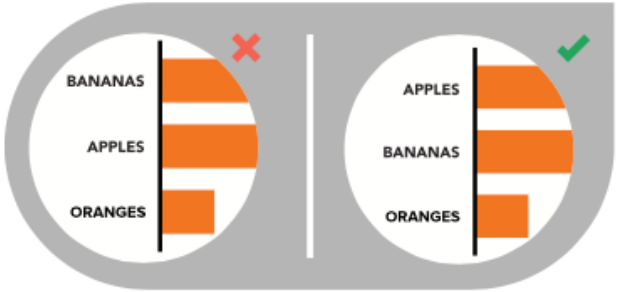
Start the y-axis value at Zero

Starting at a value above zero truncates the bars and doesn't accurately reflect the full value.



Use Consistent Colors

Use one color for bar charts. You may use an accent color to highlight a significant data point.



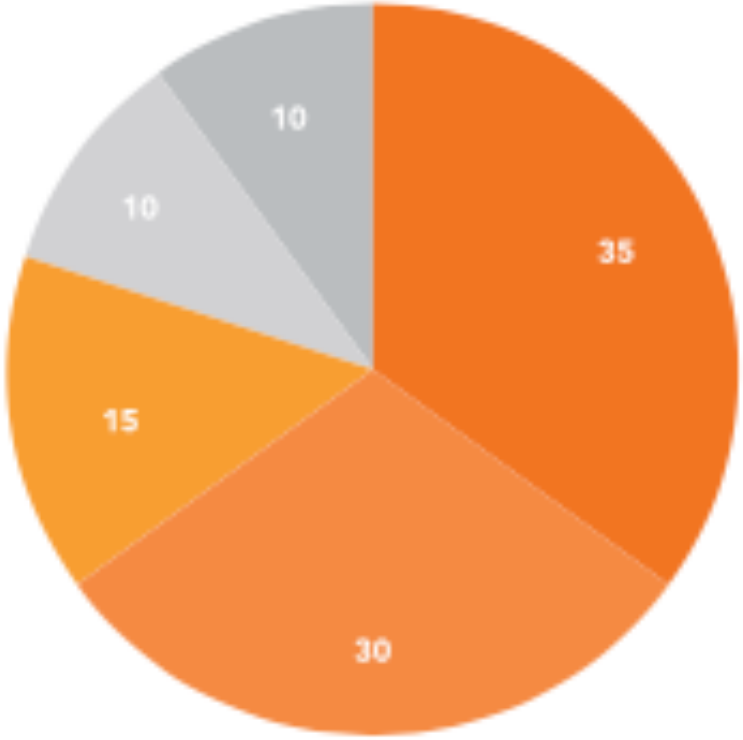
Order Data Appropriately

Order the categories alphabetically, sequentially or by the values.

Pie Chart

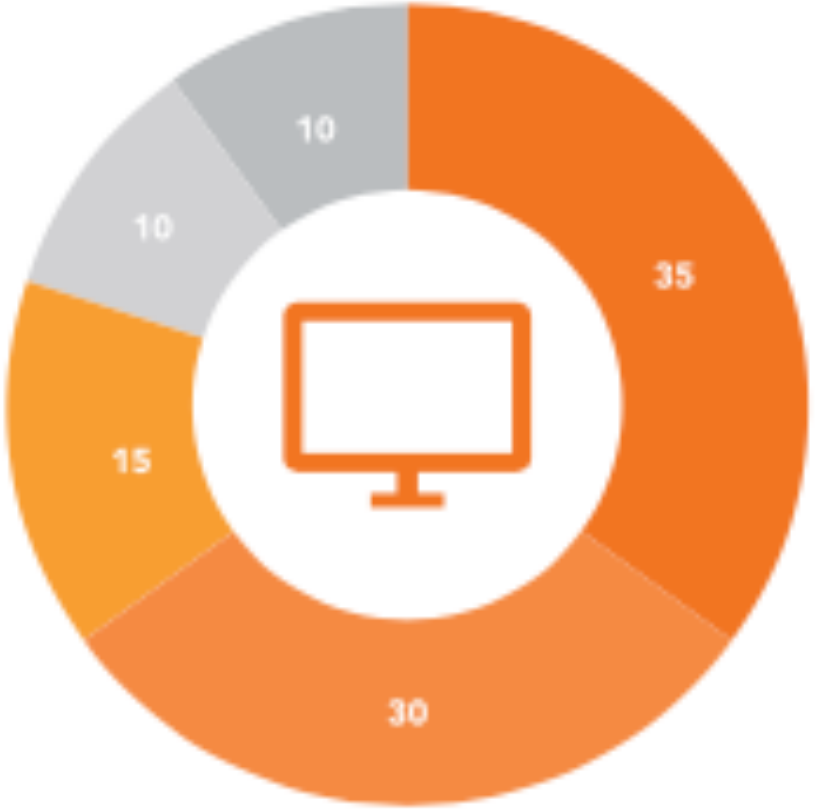
Pie charts are best used for making portion to whole comparisons with discrete or continuous data. They are most impactful with a small data set.

Pie Chart Variations



STANDARD

It is used to show part-to-whole relationships.



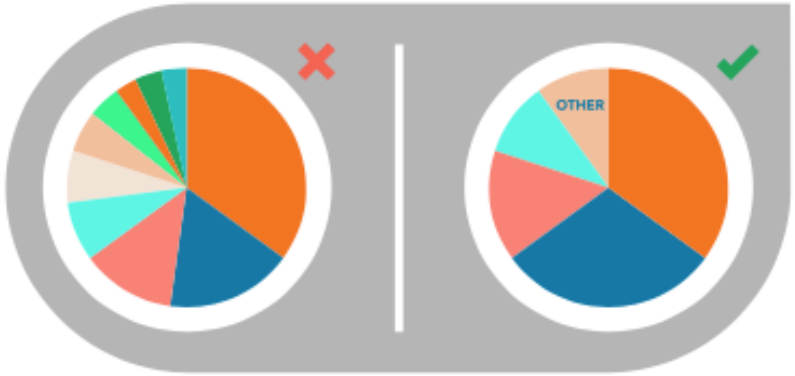
DONUT

A stylistic variation of the original pie chart with an inclusion of a total value or design element in the center.

Pie Charts

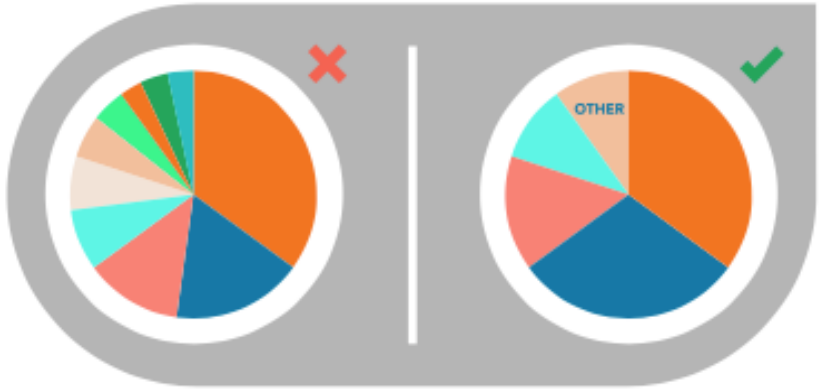
Best Practices

Pie Charts Design Best Practices



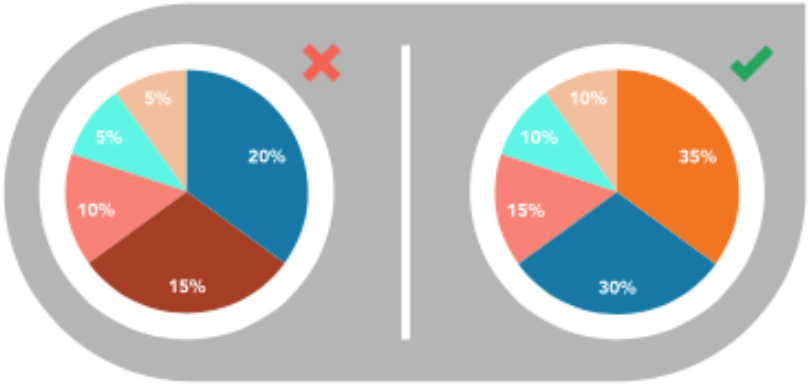
Visualize no more than 5 Categories per Chart

It is difficult to differentiate between the small values; depicting to many slices makes it complex and decreases the visualization impact. If needed, multiple small slices may be categorized as “Miscellaneous” or “Other”



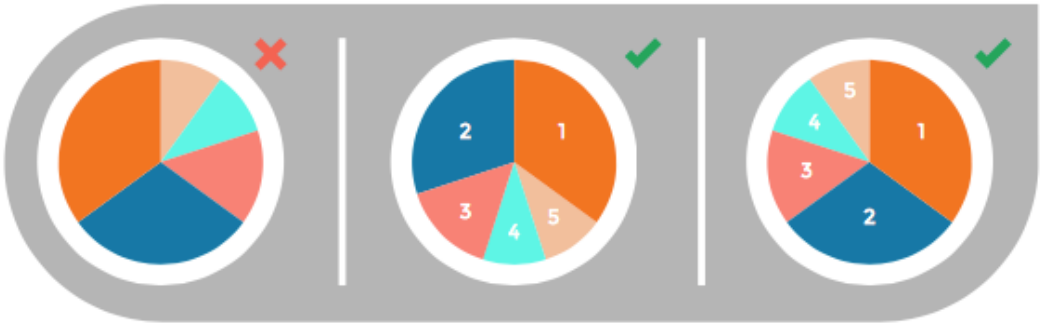
Don’t use Multiple Pie charts for Comparison

Sliced sizes are very complex to compare side by side. Hence, if required; use a stacked bar chart instead.



Total Data Count must be 100%

Make sure that total values sum up to 100% and that pie slices are sized proportionate to their corresponding value



Order the slices Correctly

Option-1: Place the largest section at 12 o’clock going clockwise and second largest at 12 o’clock counterclockwise.

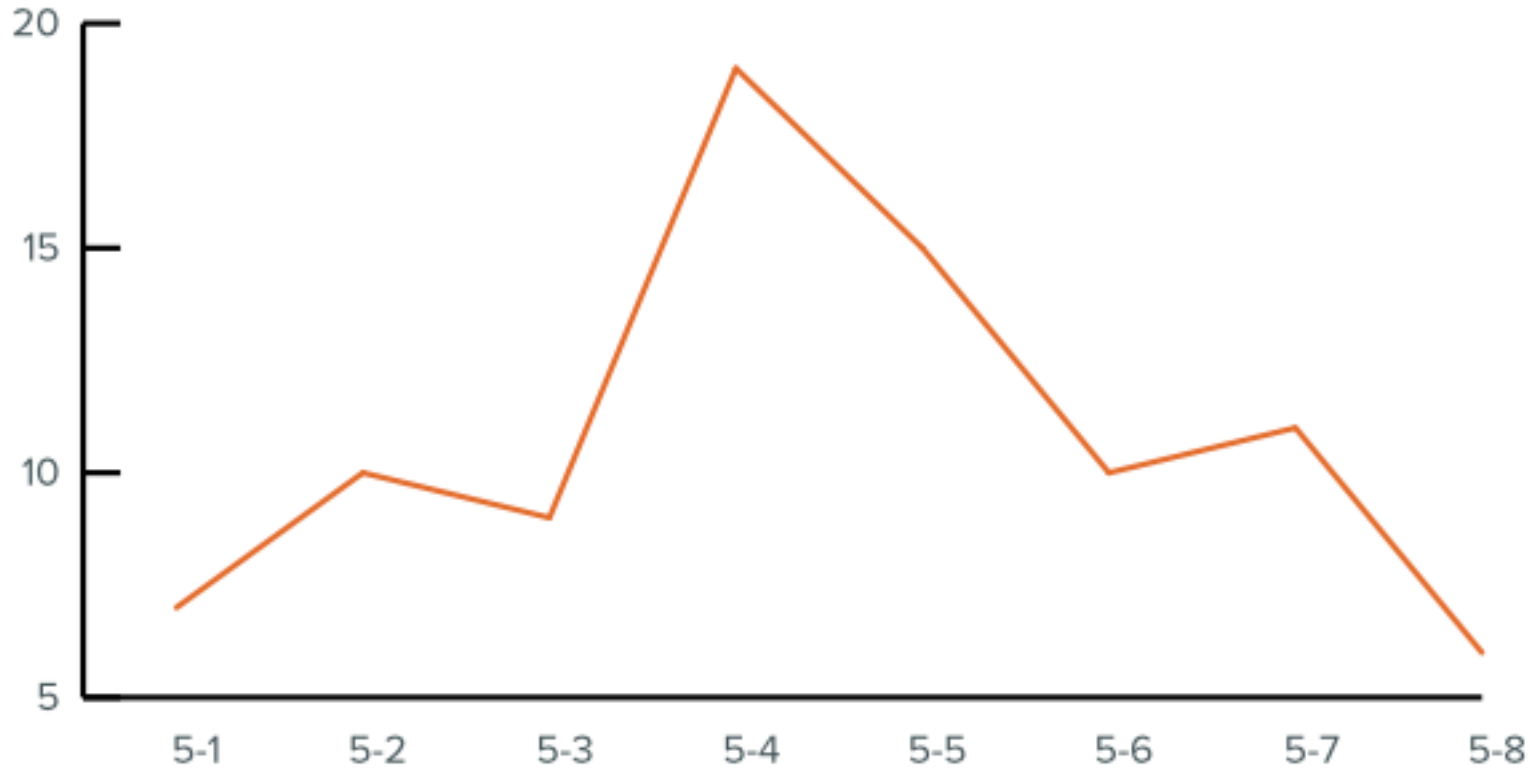
Option-2: Place the largest section at 12 o’clock going clockwise. Place remaining sections in the descending order, going clockwise.

Line Chart

Line charts are used to show time-series relationships with continuous data. They help show trend, acceleration, deceleration, and volatility.

Line Chart Variations

Line chart itself doesn't offer any variations. It may be used to track or identify changing trends in bar chart but it itself doesn't have any variants.

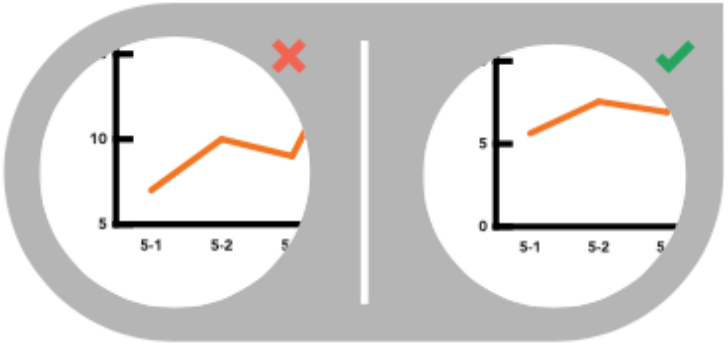


Direct Marketing Views, By Date

Line Charts

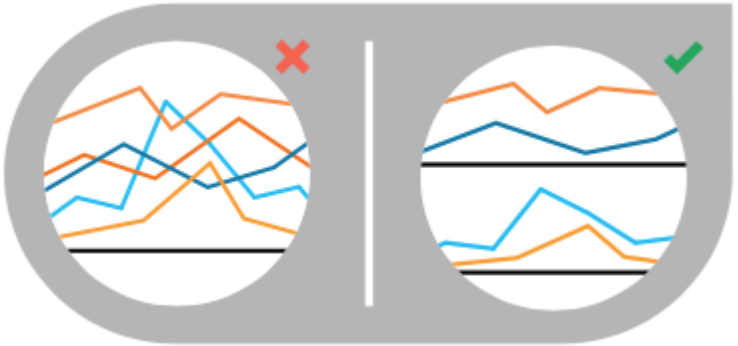
Best Practices

Line Charts Design Best Practices



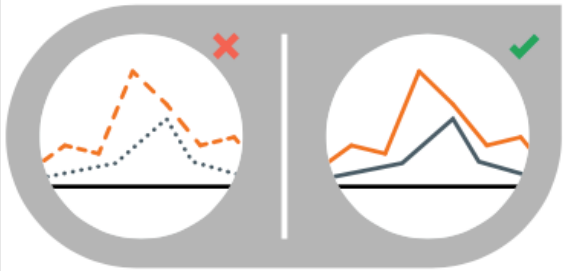
Inclusion of Zero Baseline

Although a Line chart doesn't have to start with a 0 value; it should be included whenever possible.



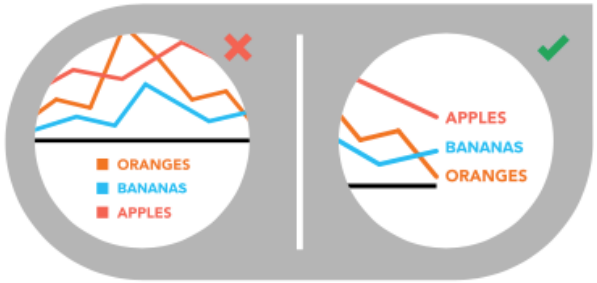
Don't plot more than 4 lines

If you need to display more than 4 lines, break them into separate charts for better comparison



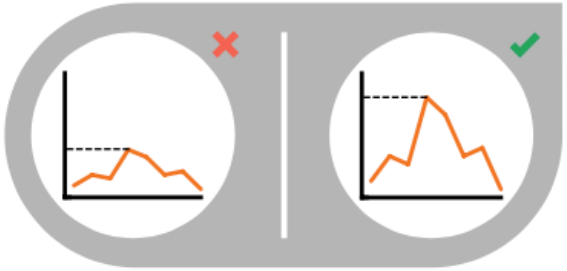
Solid Lines ONLY

Use of dashed and dotted lines can be distracting



Label Directly

This allows readers quickly identify lines.



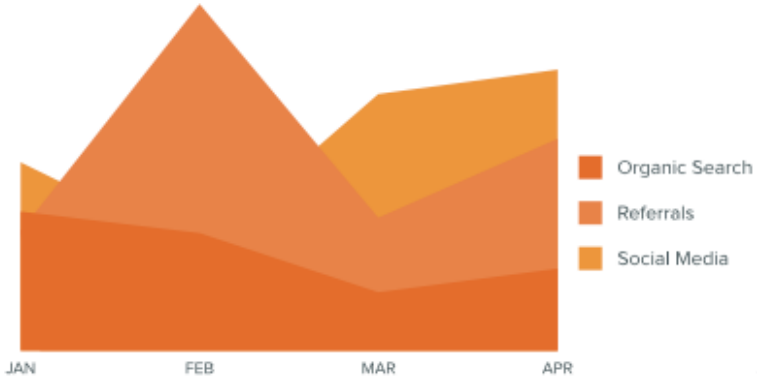
Use the right Height

Plot all lines so that the line chart takes approximately two-thirds of the y-axis's total scale.

Area Chart

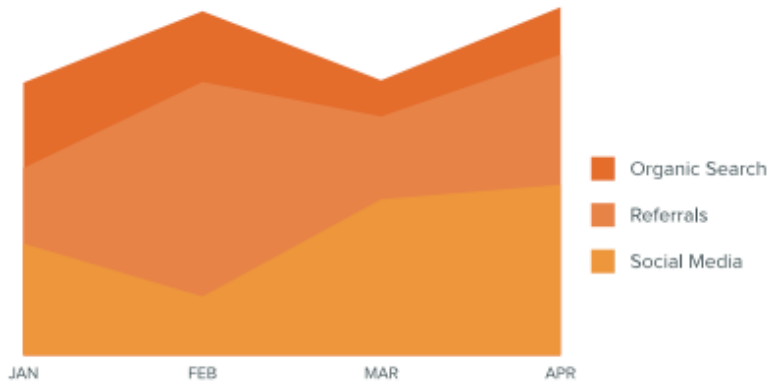
Area charts depict a time-series relationship, but they are different than line charts in that they can represent volume

Area Chart Variations



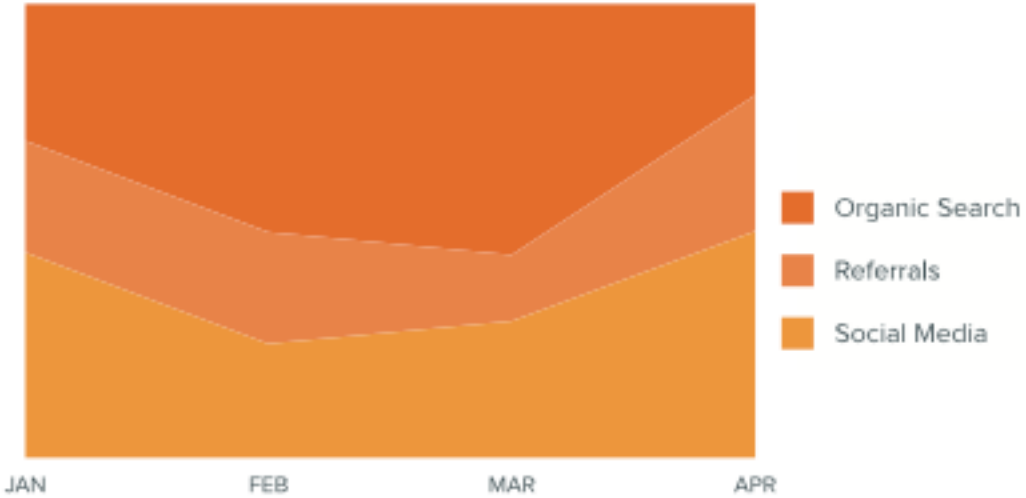
Area Chart

Used to show or compare quantitative progression over time



Stacked Area Chart

Best used to visualize part-to-whole relationship over time, how each category contributes to cumulative total



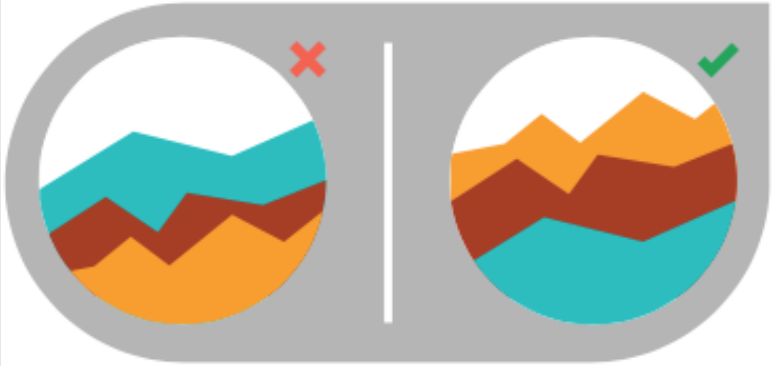
100% Stacked Area Chart

Used to show distribution of categories as part of a whole, where the cumulative total is not important.

Area Chart

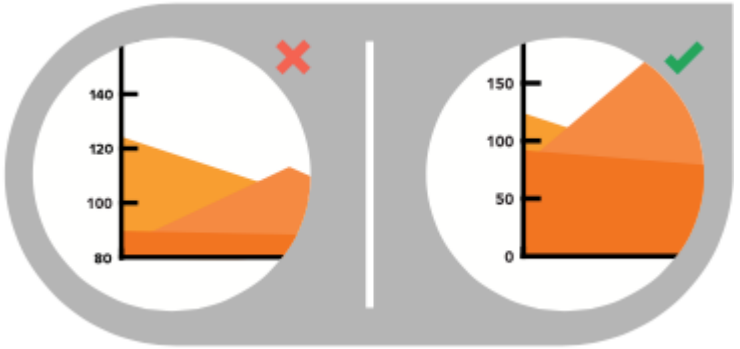
Best Practices

Area Charts Design Best Practices



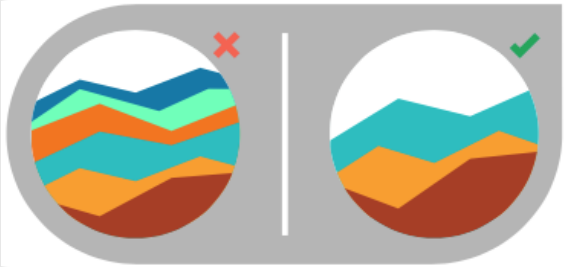
It should be easy to read

In stacked area charts, arrange data to position categories with highly variable data on the top of chart and low variability on the bottom.



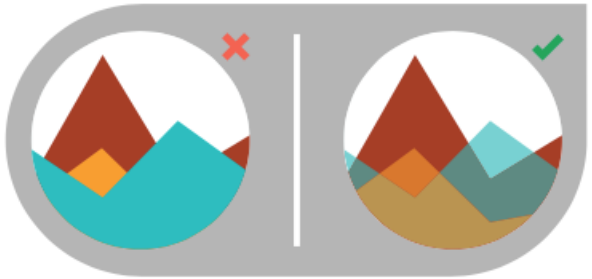
Start y-axis value at 0

Starting above zero truncates the visualization of values.



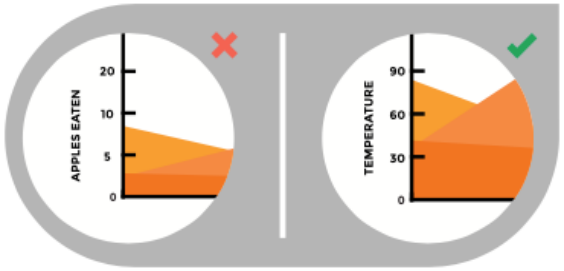
Don’t display more than 4 categories

It will result in a complex cluster visual



Use Transparent Colors

Use of transparency must be ensured for clear visibility



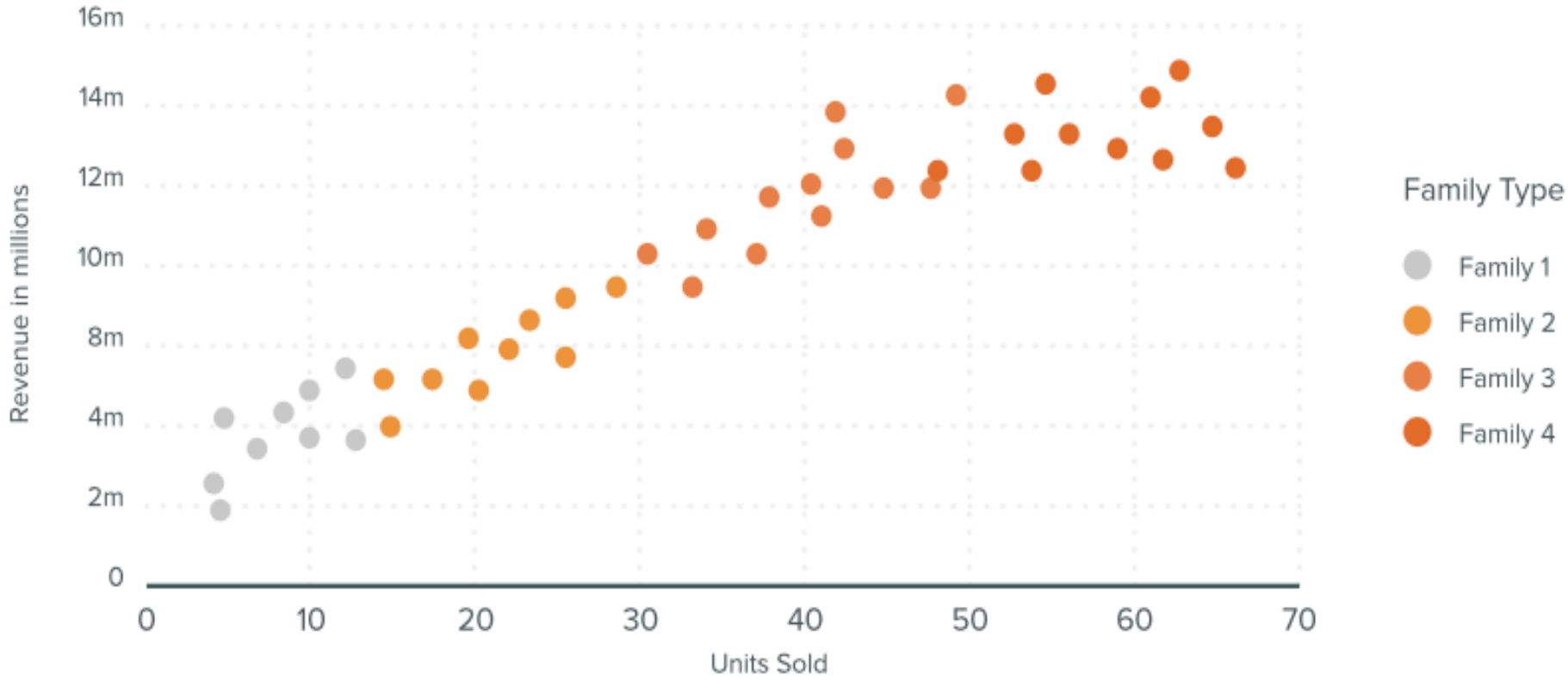
Don’t use for Discrete Data

The connected lines imply intermediate values, which only exist in continuous data

Scatterplot Chart

Scatter plots show the relationship between items based on two sets of variables. They are best used to show correlation in a large amount of data.

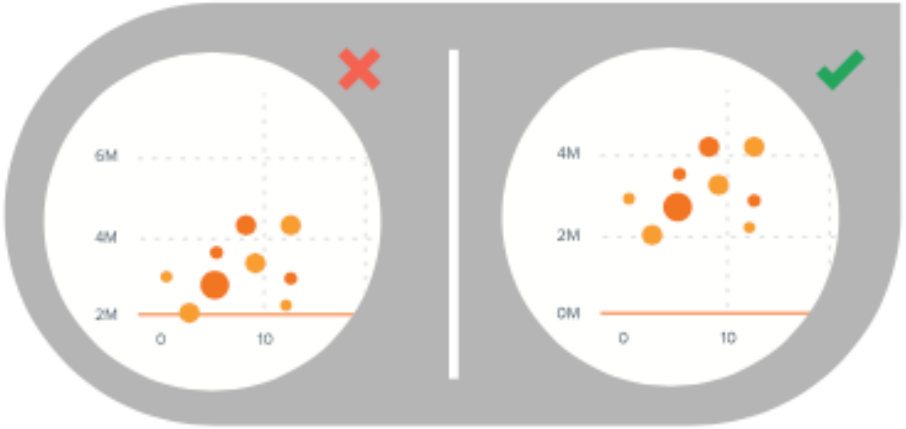
Scatterplot Chart Variations



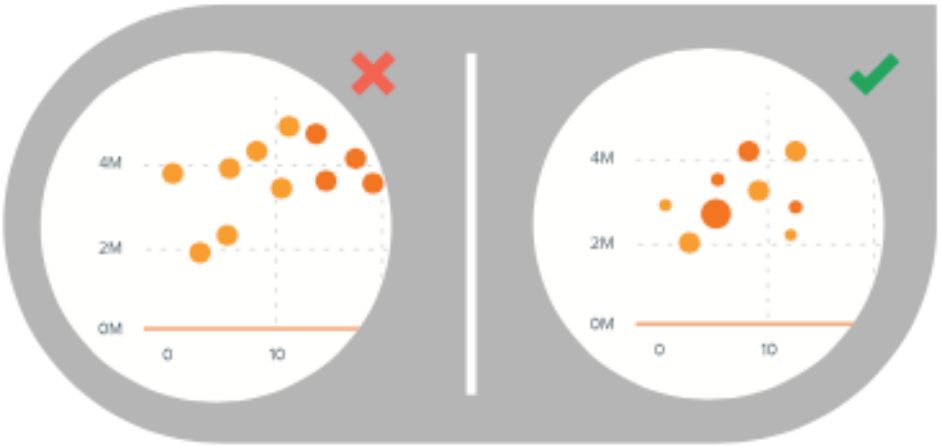
Scatterplot Chart

Best Practices

Scatterplot Charts Design Best Practices



Start with y-axis value at 0

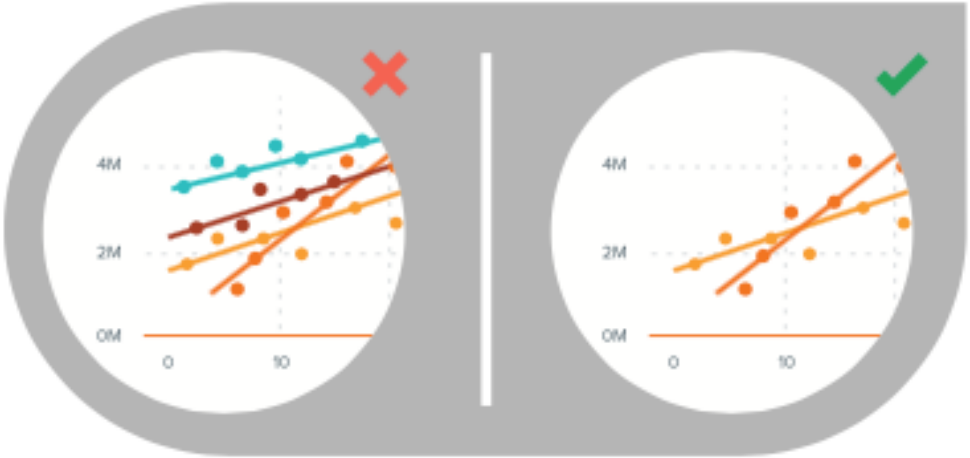


Include more Variables

Use size and dot color to encode additional data variables

Use Trend Lines

These lines help draw correlation between the trending variables



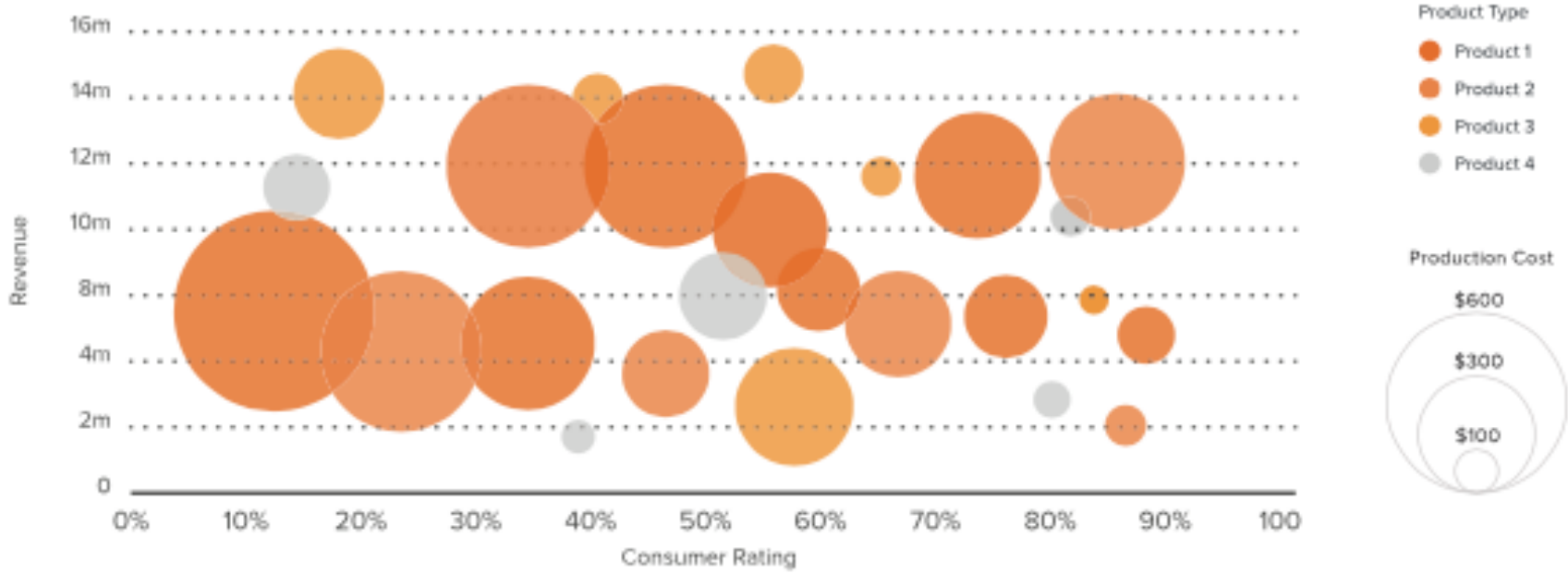
Don't Compare more than 2 Trend Lines

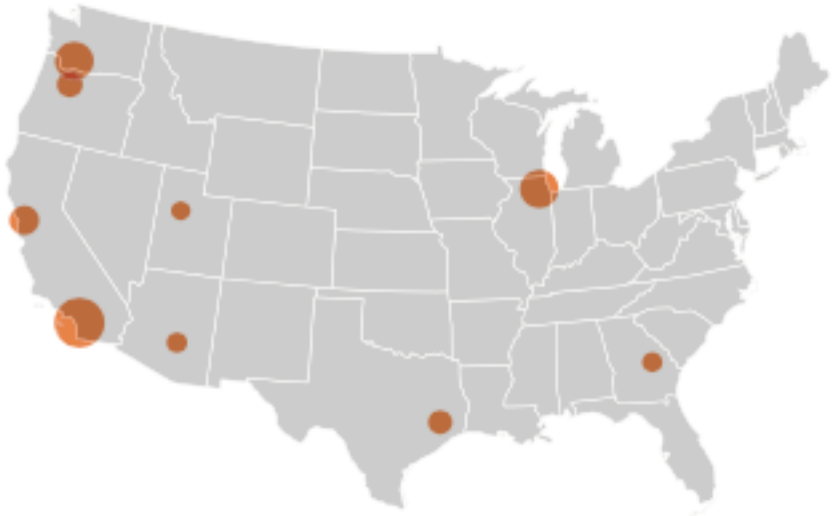
Too many lines make it difficult to interpret

Bubble Chart

Bubble charts are good for displaying nominal comparisons or ranking relationships.

Bubble Chart Variations





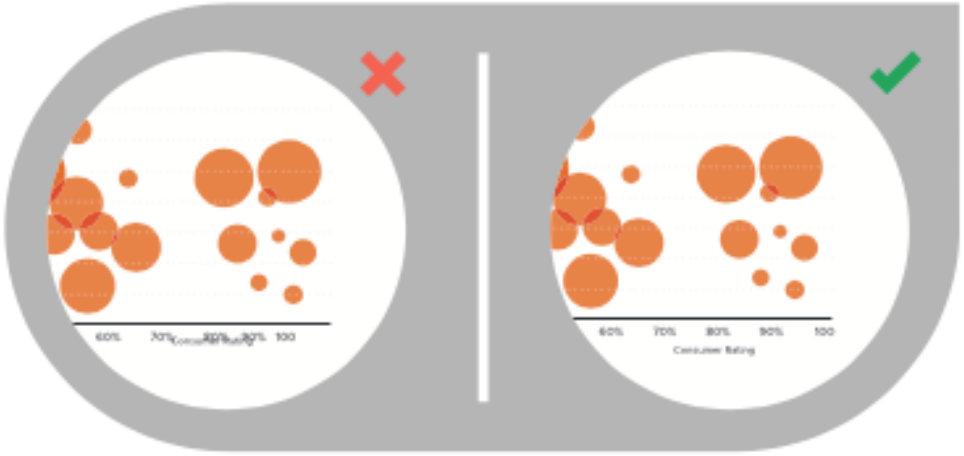
Bubble Plot is a scatterplot with bubbles best used to display an additional variable.

Bubble map is best used to visualize values for specific geographic regions.

Bubble Chart

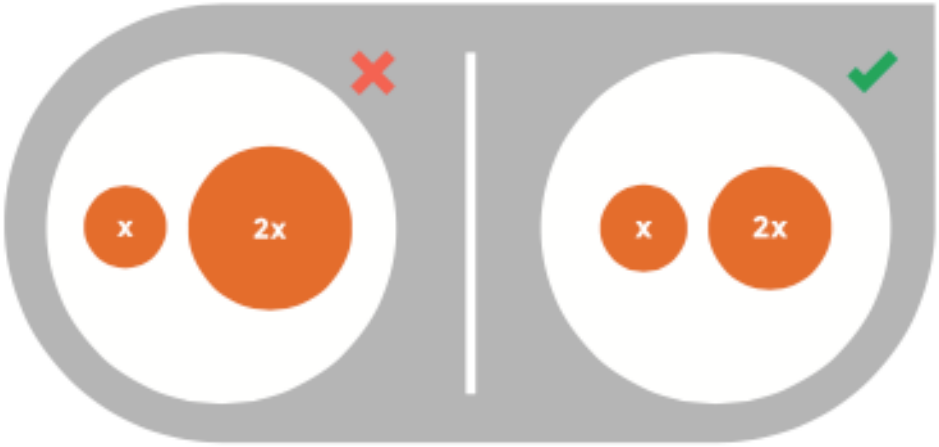
Best Practices

Bubble Chart Design Best Practices



Label Visibility must be ensured

Make sure the labels are visible, easily identifiable and unobstructed



Size the Bubbles Appropriately

Bubbles should be scaled according to the area and not the diameter.

Avoid using Odd shapes

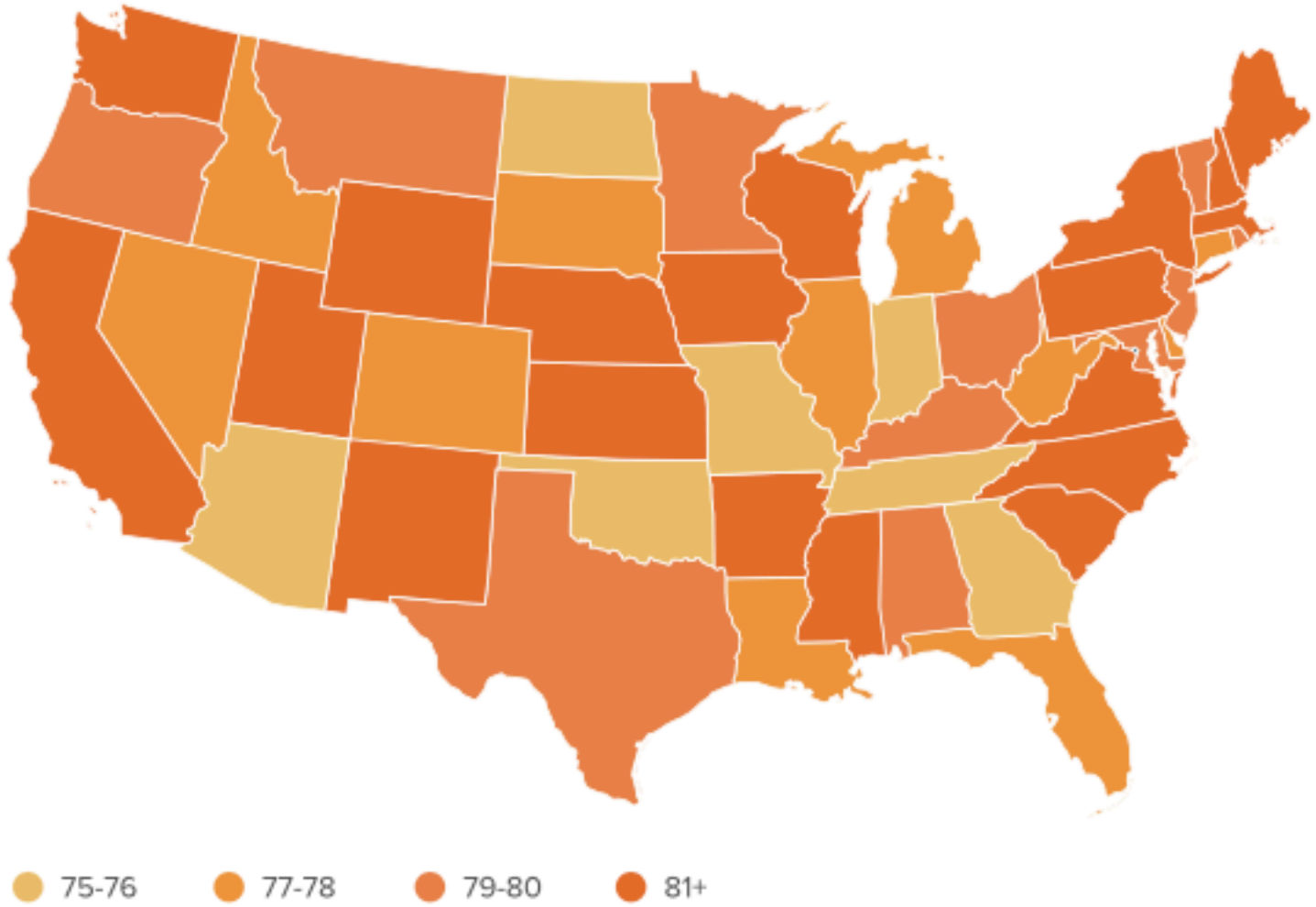
Avoid adding too much details or using shapes that are not entirely circular, this can lead to inaccuracies.

Heat maps

Heat maps are used to display categorical data, using intensity of color to represent values of geographic areas or data tables.

Heat Map Variations

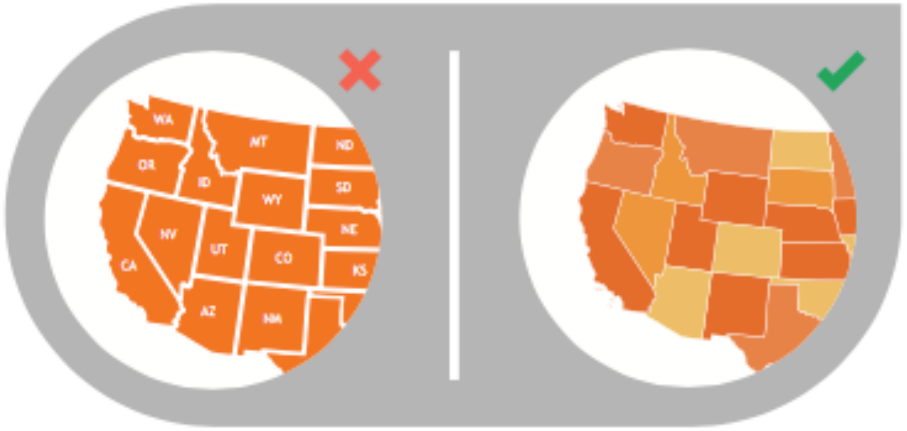
STATES WITH NEW SERVICE CONTRACTS



Heat Map

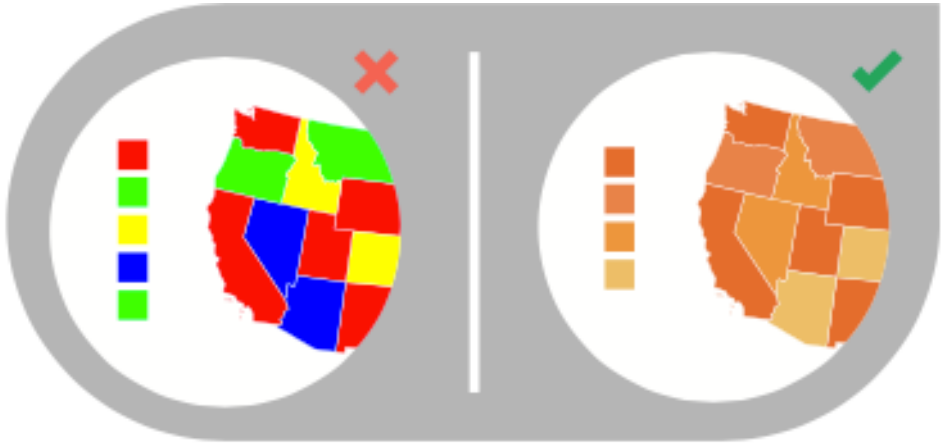
Best Practices

Heat Map Design Best Practices



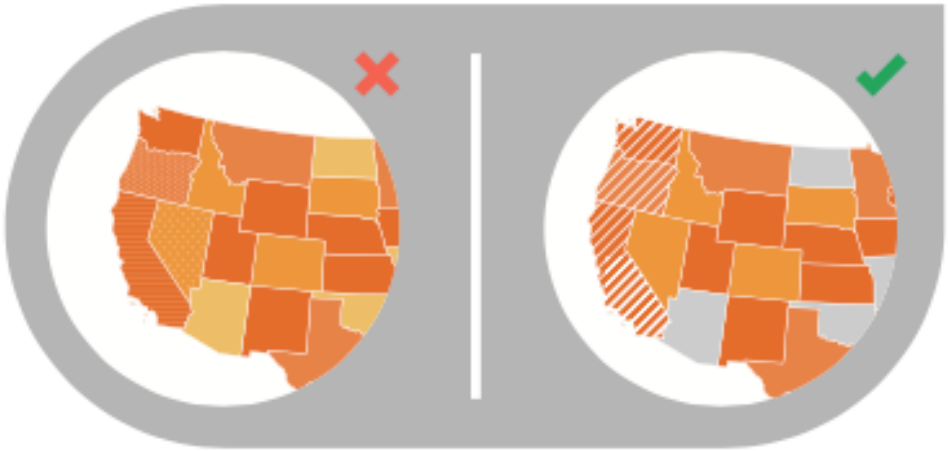
Use a simple Map outline

These lines are meant to frame the data



Appropriate Choice of Colors

Use a single color with varying shades. This will not only make it soothing and appealing visually but also present the results correctly..



Use of Patterns

Use patterns to indicate second variable. But using multiple patterns is overwhelming and distracting

Appropriate Date Ranges

Select 3 to 5 numerical ranges that enable fairly data distribution. Use +/- signs to indicate high and low ranges

Do’s and Don’ts in DATA DESIGN & VISUALIZATION

- Do Use one color to represent each category
- Do order data sets using logical hierarchy
- Do use callouts to highlight important or interesting information
- Do visualize your data in a way that it is easy for readers to compare values
- Do use icons to enhance comprehension and reduce unnecessary labelling

- Don’t use high contrast color combinations such as Red/Green or Blue/Yellow
- Don’t use 3D charts. They can skew perception of the visualization
- Don’t add chart junk. Unnecessary Illustrations, drop shadows or ornamentations distract from the data
- Don’t use more than 6 colors in a single layout
- Don’t use distracting fonts or elements (such as bold, italic or underlined text)

References

- Pham Viao; *Best Practices in Data Visualizations (2014)*, Microstrategy
- Haider Al Seaidy; *Dashboard Design and Data Visualization Best Practices (2016)*, Splunk Conference on Data Science
- Syno et al; *Best Practice Visualization, Dashboard and Key Figures Report (2013)*, Open Data Monitor.
- Hubspot; *How to Design Charts and Graphs*, Data Visualization 101
- Tableu; Visual Analysis Best Practices: Simple Techniques for Making Every Data Visualization Useful and Beautiful