Numeric and Statistical Visualizations Part 1

John Keyser

Taxonomies of Data Visualizations

- There are many ways to categorize and think about visualizations
- Example: Categorical vs Quantitative Variables
  - Variation within each type
  - Variation across time
  - Variation across space
  - Relationships among variables
- We will return to some of these later
  - Not the goal for today's discussion

Chart Types for Numeric and Statistical Data

- We will focus here on standard visualizations for data with a primarily numeric/statistical component
  - Probably the most common type of data
  - Will focus on idea of independent/dependent variables
- Later we will talk about other data, such as relationships and hierarchies, or scientific data
  - Including how to combine with numeric info

UNIVARIATE VISUALIZATIONS

Single Variable

- We often have multiple measurements of a single variable
- Our goal in visualization is usually to understand all the values that variable takes on
  - Usually collective statistics (when individual data points aren't feasible)

Scatter Plot (1D)

- Stripchart
- Dot Plot
- Basically plot points on one axis
- Use is limited, but sometimes insightful

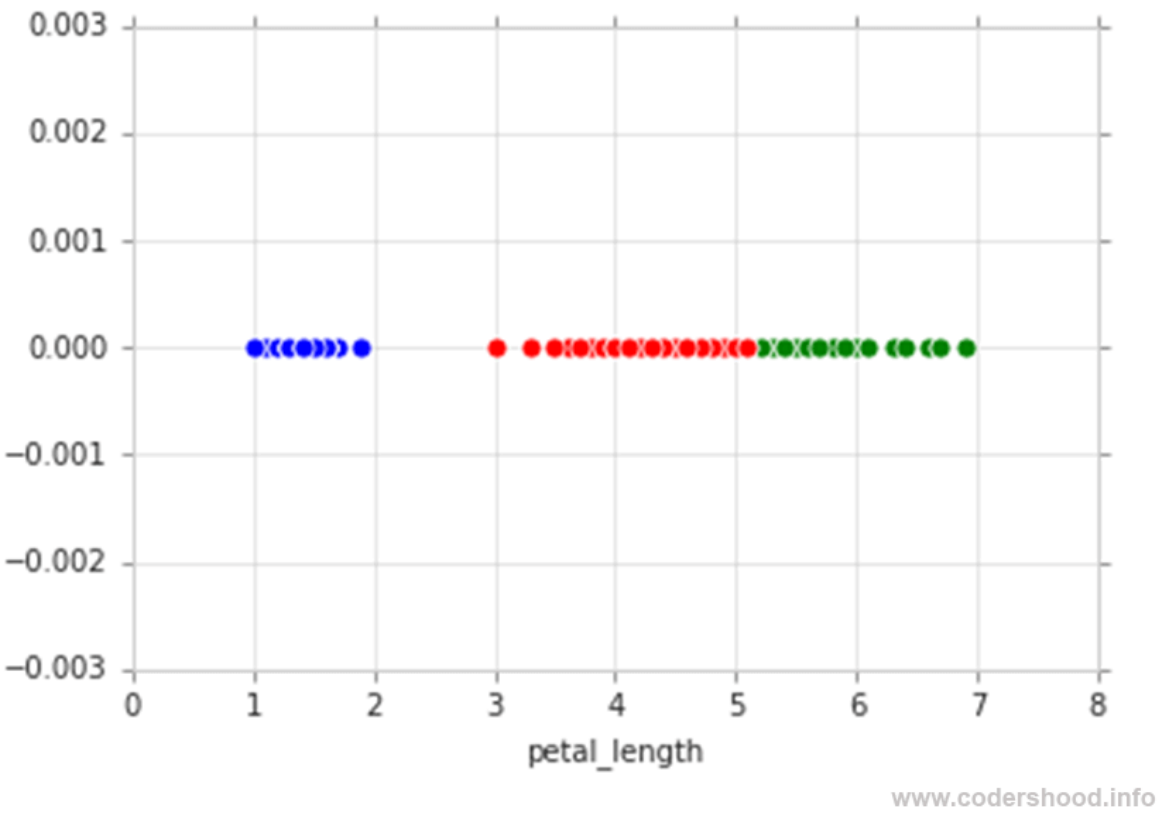


Image from:
https://www.codershood.info/2019/03/25/the-hunger-games-guide-to-exploratory-data-analysis-plotting-in-python/exploratory-data-analysis-plotting-in-python-1d-scatter-plot/

Beeswarm Chart

- If too many points for a dot plot, can use Beeswarm
- Spreads out points along the perpendicular axis

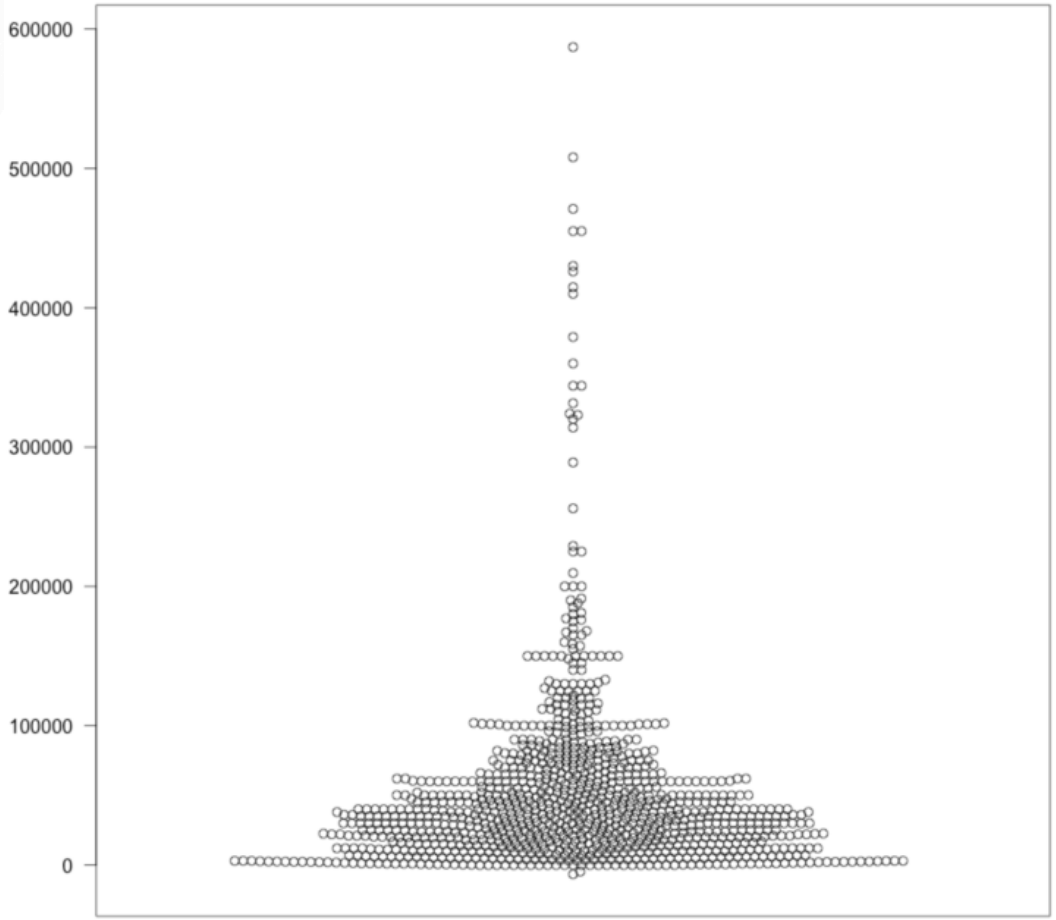


Image from:
https://flowingdata.com/2016/09/08/beeswarm-plot-in-r-to-show-distributions/

Box Plots

- AKA Box-and-Whisker
- Show Median, Quartiles, along with max/min
- Variations: Box/whiskers can show other things, such as Std. Dev. or midpoints of quartiles

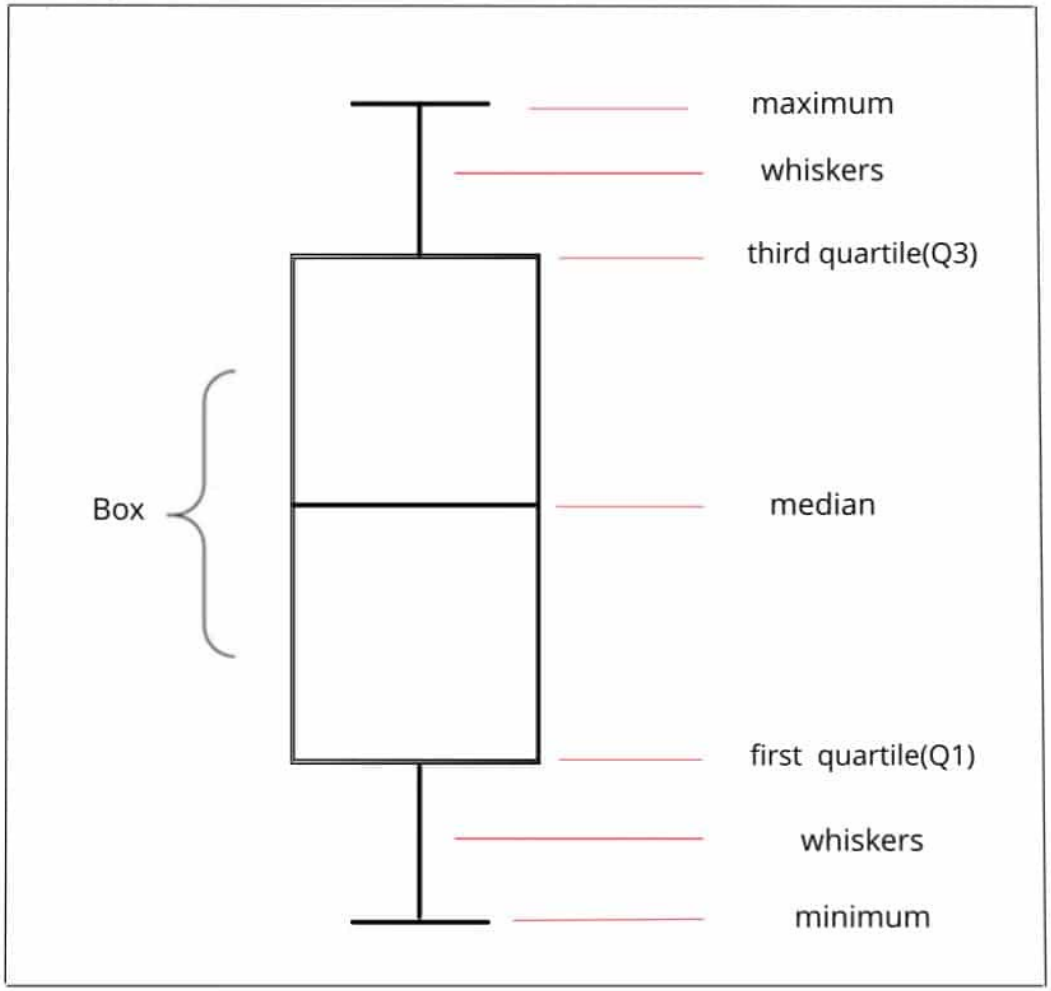


Image from:
https://fahimahmad.netlify.app/images/ComparingDistributionofDatainRbox-plot.jpg

Box Plots With Mean/SD

- Can also show mean and standard deviation on same plot
- Mean sometimes shown with a dot in middle
- Sometimes use box plot with mean at center, boxes at 1 σ
- Many other variations

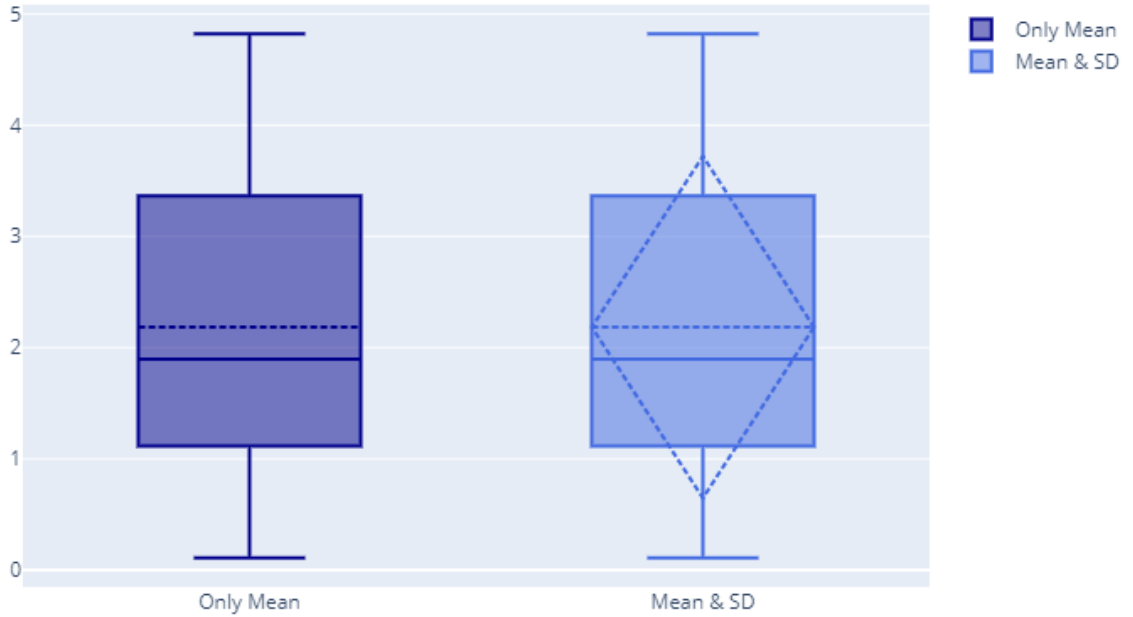


Image from:
https://plotly.com/python/box-plots/

Box Plots With Outliers

- Can show individual outliers on same graph, as points

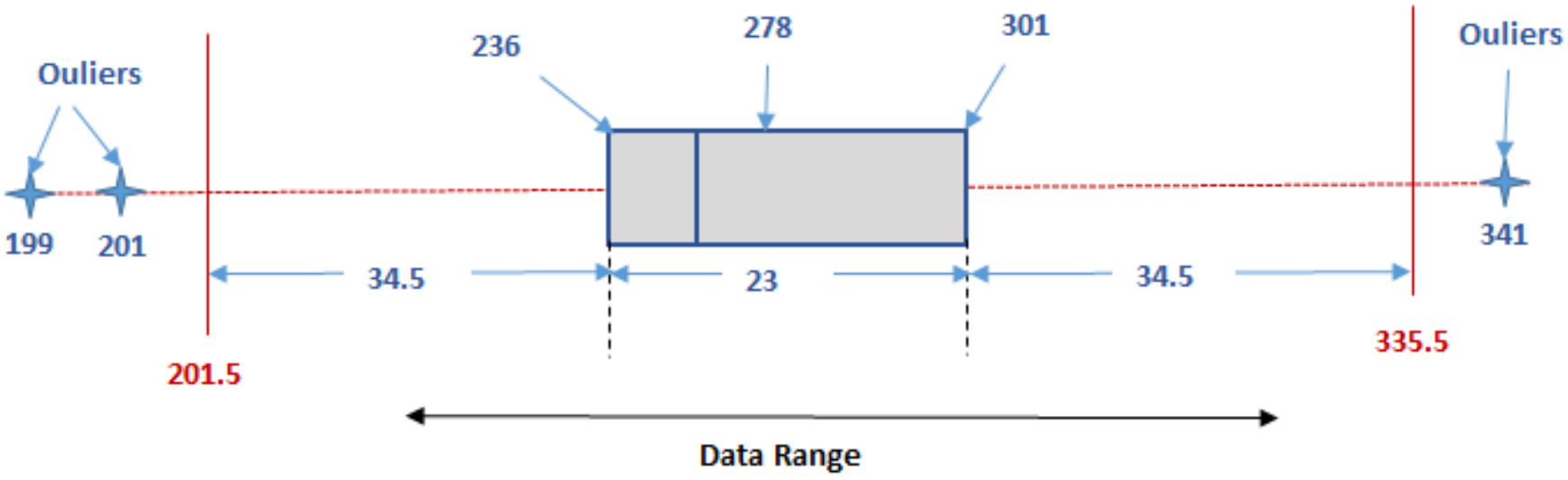


Image from:
https://www.whatissixsigma.net/wp-content/uploads/2015/07/Box-Plot-Diagram-to-identify-Outliers-figure-1.png

Histograms

- Put results into bins, drawn at lengths
- Bin choices can change impression of data

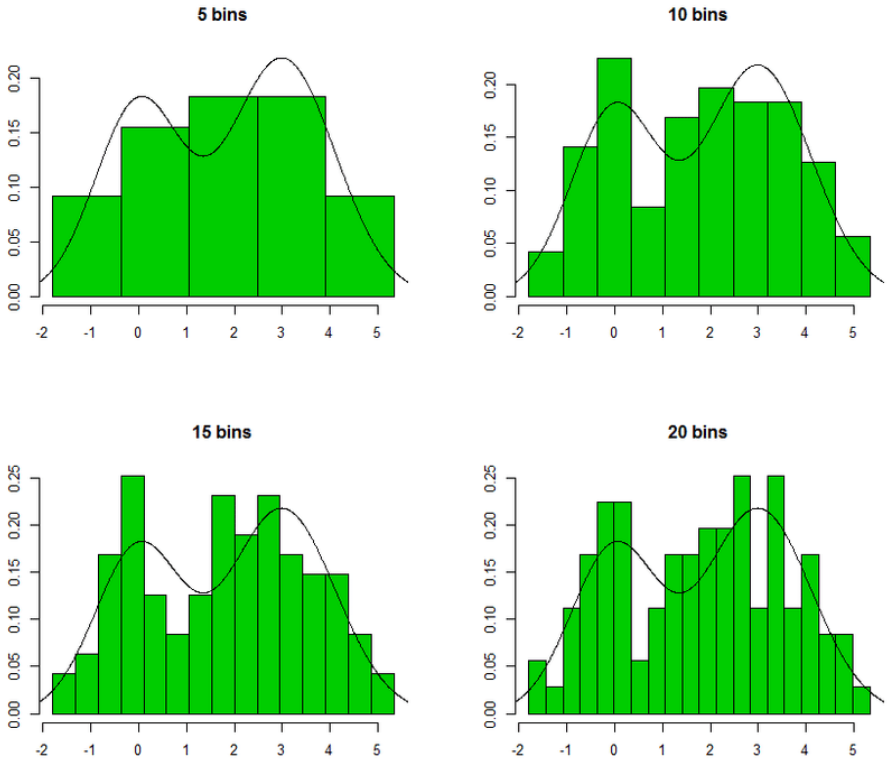


Image from:
https://www.researchgate.net/figure/Influence-of-the-number-of-bins-on-the-histogram-The-number-of-bins-chosen-by-the_fig7_276354826

"Violin" Plots

- Histogram that is mirrored and treated similar to a box plot
- Sometimes combined with stripchart

```
px.violin(tips, y="tip", x="smoker", color="sex", box=True, points="all")
```

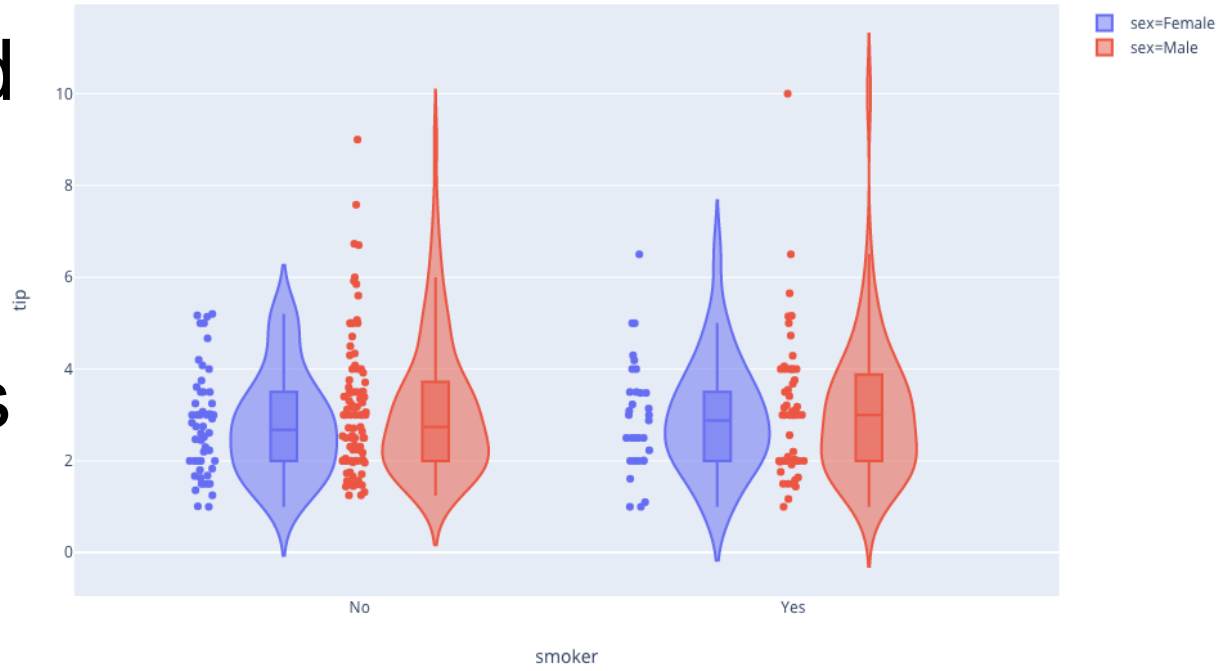


Image from:
https://medium.com/plotly/introducing-plotly-express-808df010143d

Radial (Pie) Charts

- Useful to show relative values as opposed to absolute values
- People have difficulty judging area/angle, though!

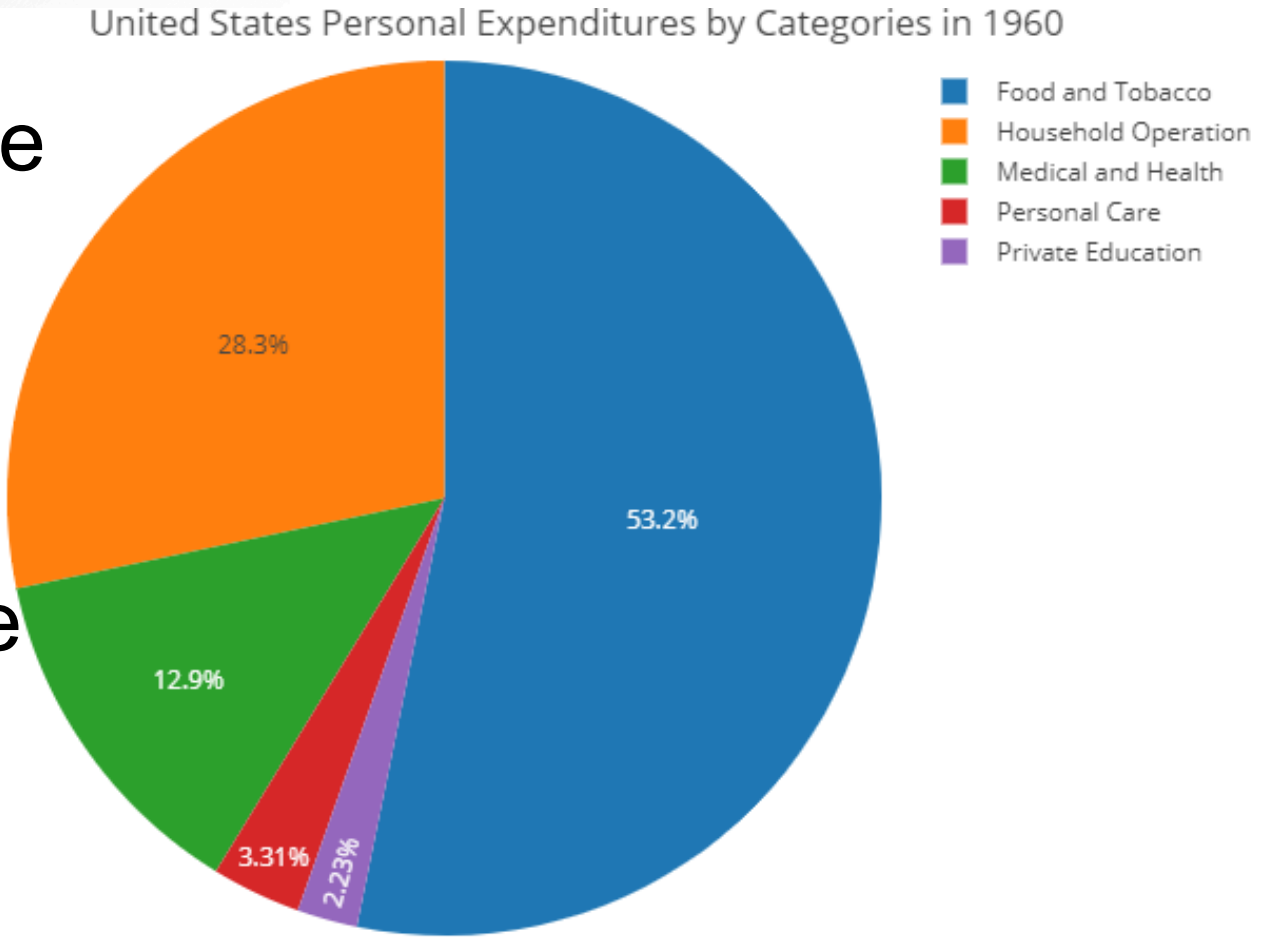


Image from:
https://plotly.com/r/pie-charts/

Donut Charts

- Remove center of pie chart
- Now, comparison is by arc length and surface area, not angle
  - People perceive this better
  - Easier to compare small slices
  - Optimal if inner ring is 40% of outer

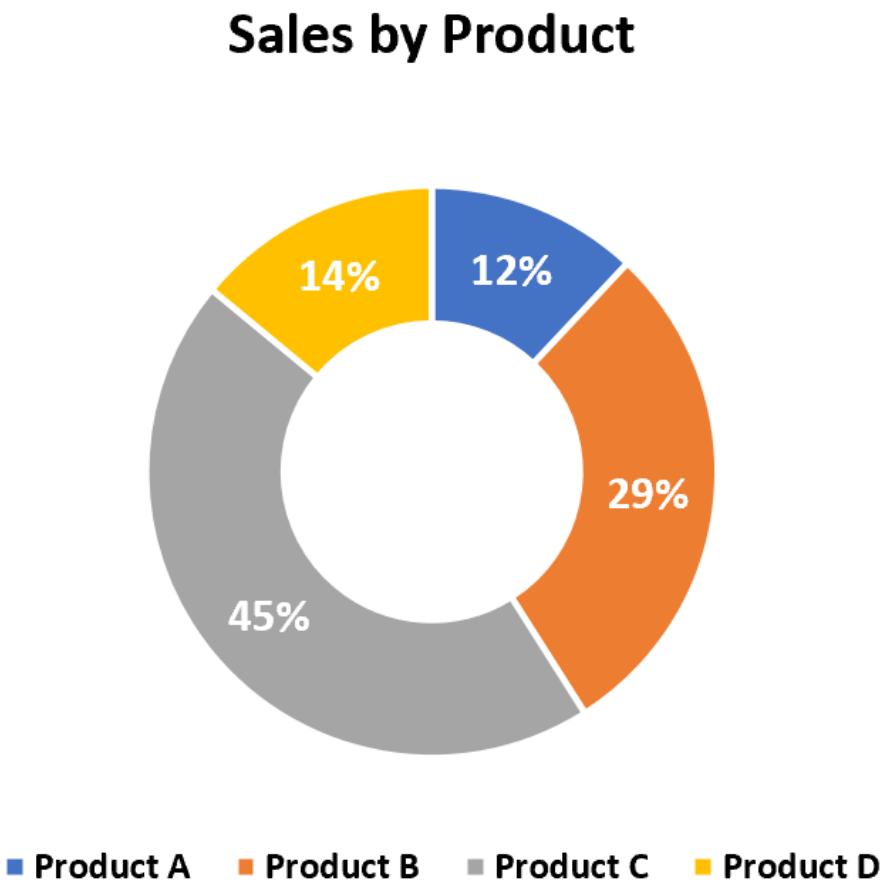


Image from:
https://www.statology.org/double-doughnut-chart-excel/

Using Pie Charts

- Some have argued that pie (and donut) charts should never be used.
- But, can be good to show some things:
  - What is needed for majority
    - Assuming pie slices are in some order, so that a clear straight line is meaningful
  - As an extra axis of information on something already represented as a circle
  - When we want it for aesthetic reasons
    - People like them… and some indication that circles are a naturally preferred shape.

“BASIC” MULTIVARIATE VISUALIZATIONS

Independent and Dependent Variables

- Often there are multiple variables that are involved.
- Independent Variables:
  - The things “controlled” for in an experiment
  - The things that will take on values
- Dependent Variables:
  - The things measured in an experiment
  - The values we want to understand

Appropriate Visualizations

- The number of independent and dependent variables determines a lot about what visualization methods are appropriate
- Also, the number of data points changes the type of visualization you can use
  - Is it possible/useful to show individual data points, or only collective statistics?
- And, the values could be discrete or continuous…

Bar Chart

- One independent, one dependent
- Discrete, limited values for independent
- (Histograms are bar charts where dependent variable is frequency)

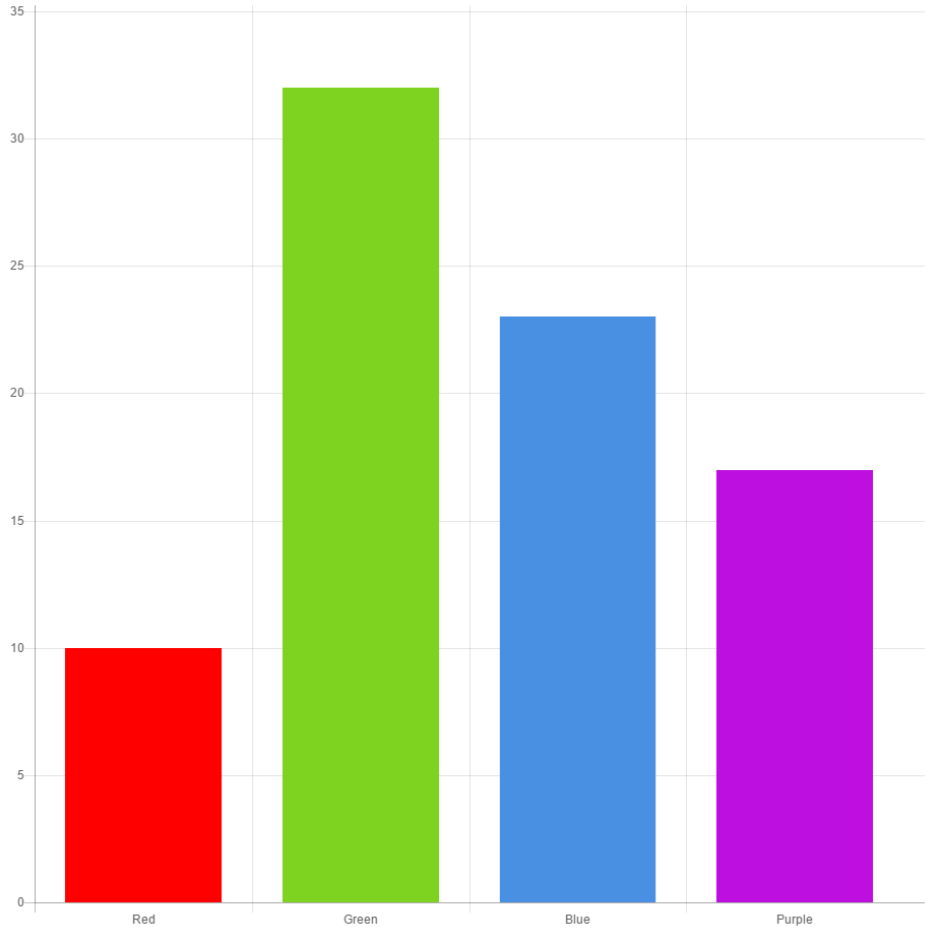
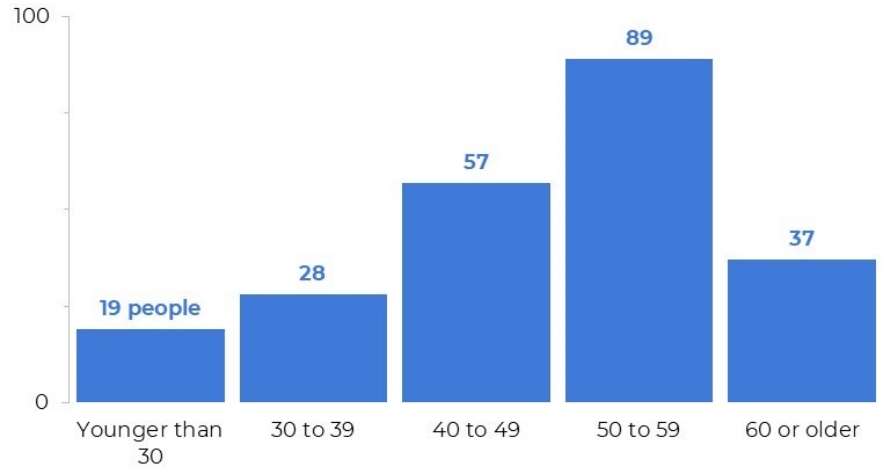


Image from:
https://www.themexpert.com/docs/quix/elements/bar-chart

Bar Chart

- Rules of Thumb:
  - Horizontal axis should be quantitative or ordinal values.
    - Or if just 2-4 categories
  - More than 4 categories: use vertical axis (i.e. horizontal bar chart)



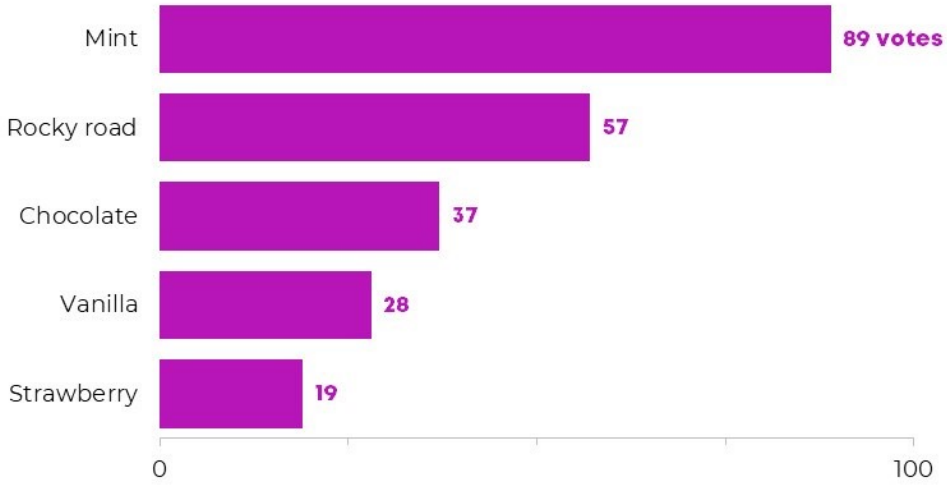


Image from:
https://depictdatastudio.com/when-to-use-horizontal-bar-charts-vs-vertical-column-charts/

Grouped Bar Chart

- Like multi-line graph, adds a second independent variable
- Usually use color to distinguish second independent

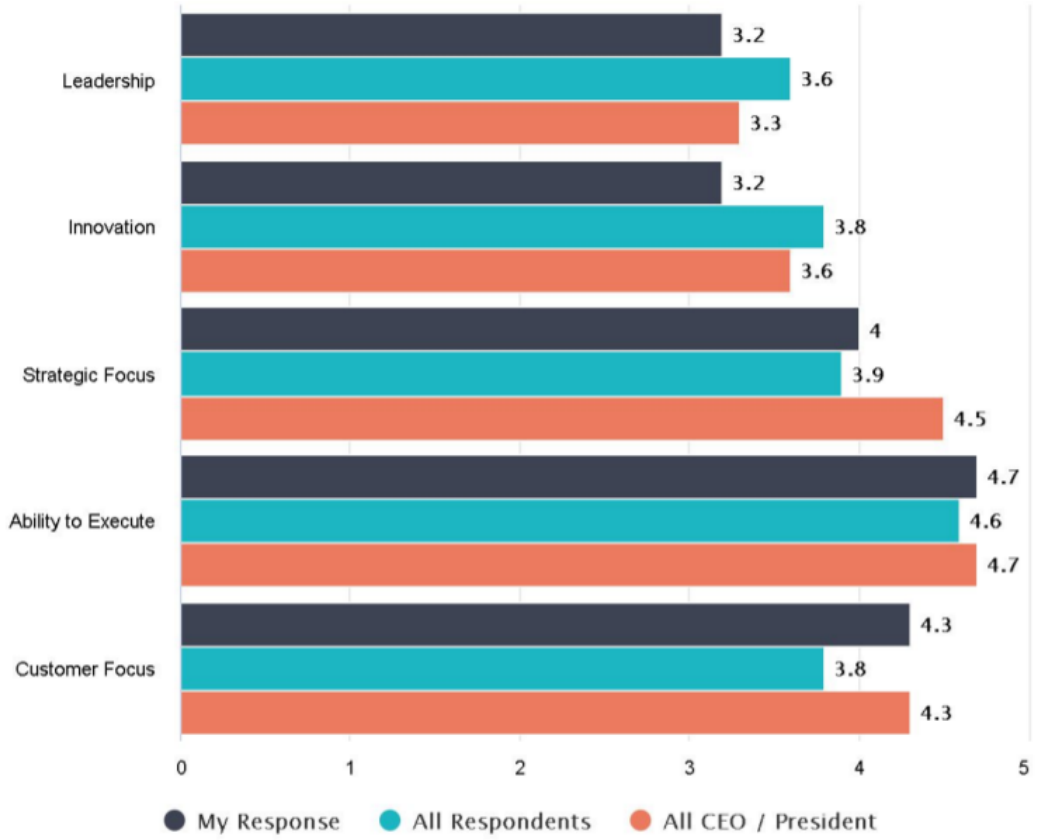


Image from:
https://help.brilliantassessments.com/hc/en-us/articles/360037060794

Poplulation Pyramid

- If one independent variable has two categories, can graph as a population pyramid
- So called since it usually is used for population of men/women (but could be any 2 categories

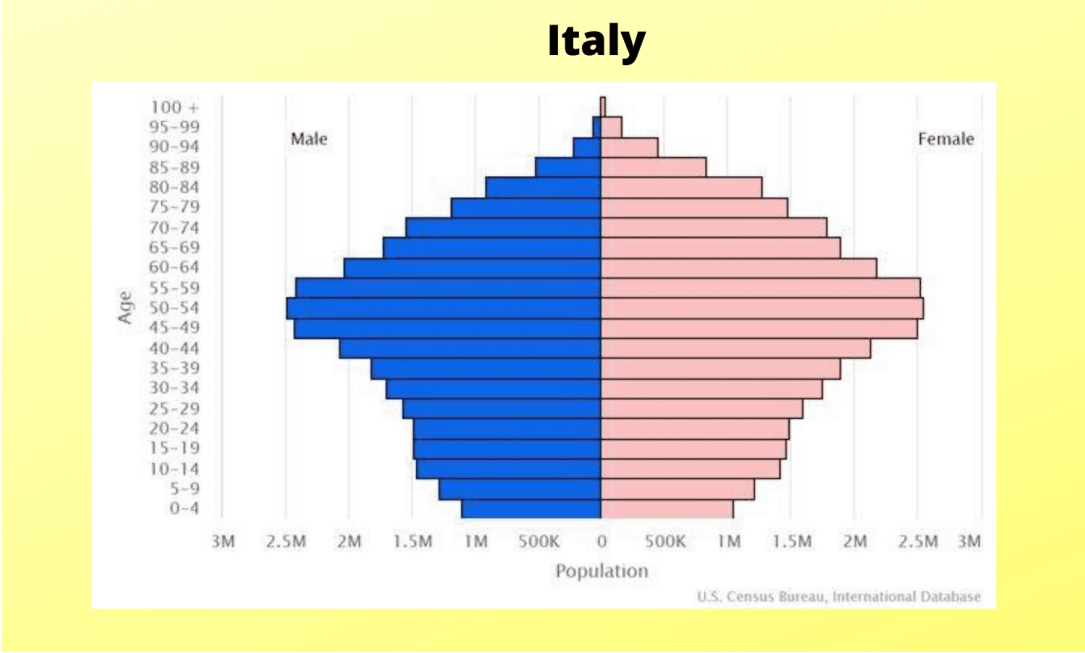


Image from:
https://populationeducation.org/what-is-a-negative-or-top-heavy-population-pyramid/

Stacked Bar Chart

- Adds additional independent variable
  - Like the multi-bar chart, but values are stacked
  - Note: significant issues in interpreting/comparing
- Makes sense when the parts sum to a whole

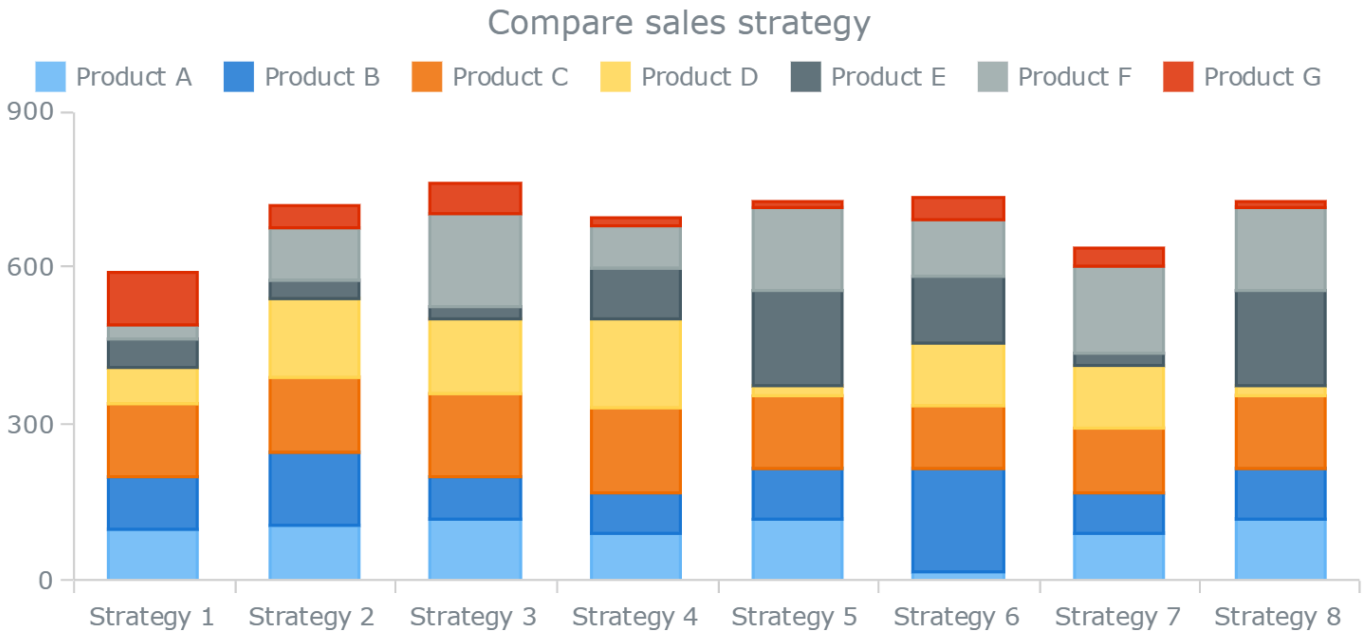


Image from:
https://www.smashingmagazine.com/2017/03/understanding-stacked-bar-charts/

100% Stacked Bar Chart

- Each bar adds up to full amount
  - Shows relative instead of absolute values
- Single one is an alternative to Pie Chart
  - But now can compare multiple
- Easier to compare multiple categories
  - Easiest for those at beginning/end due to common baseline

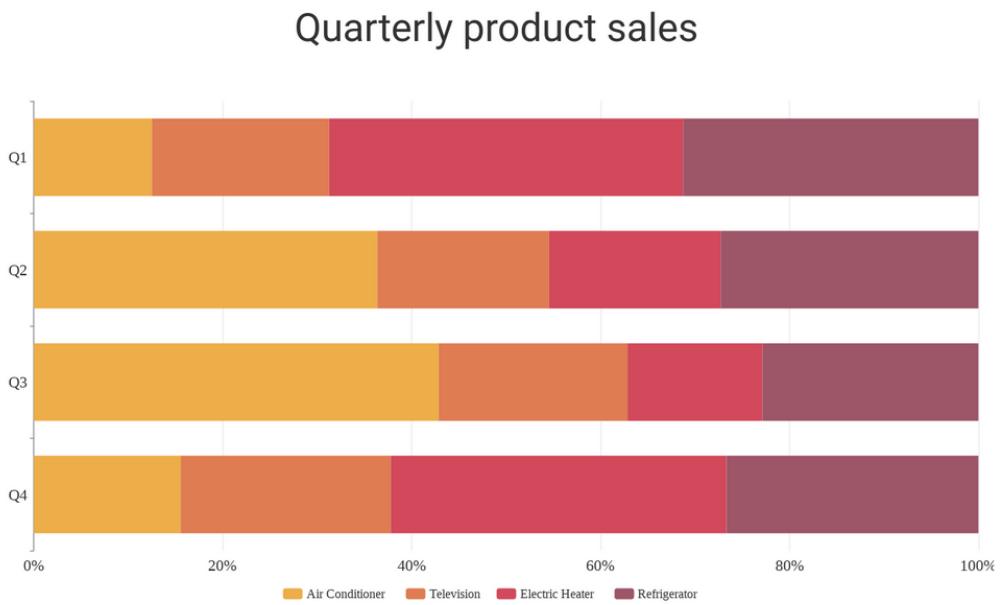


Image from:
https://online.visual-paradigm.com/charts/templates/100-stacked-bar-charts/100%25-stacked-bar-chart/

Line Graphs

- Single independent, single dependent
  - Each on one axis
- Continuous line along independent variable
- Usually assumes independent variables have unique value
  - Otherwise, treat as set of univariate

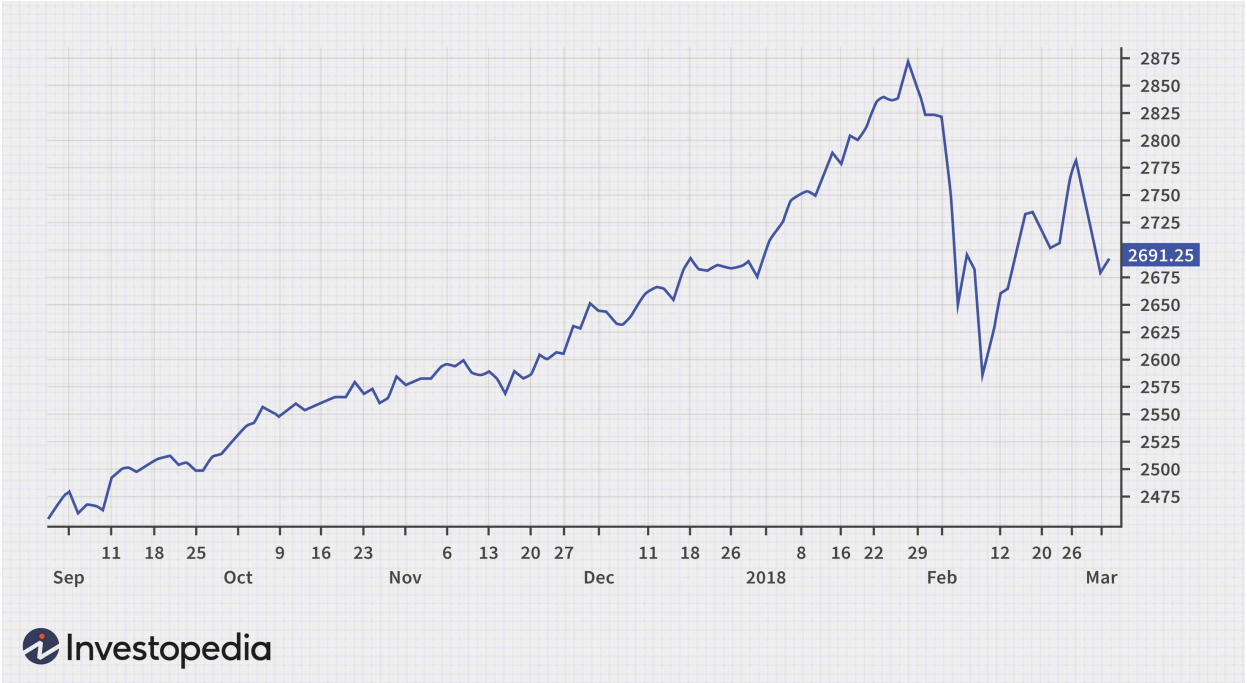


Image from:
https://www.investopedia.com/terms/l/linechart.asp

Multi-Line Graphs (1)

- TWO independent, one dependent
  - One independent on axis (usually x-axis)
  - Other independent is which line (usually categorical)
- Usually use color (or shape of plotted points) to distinguish one variable

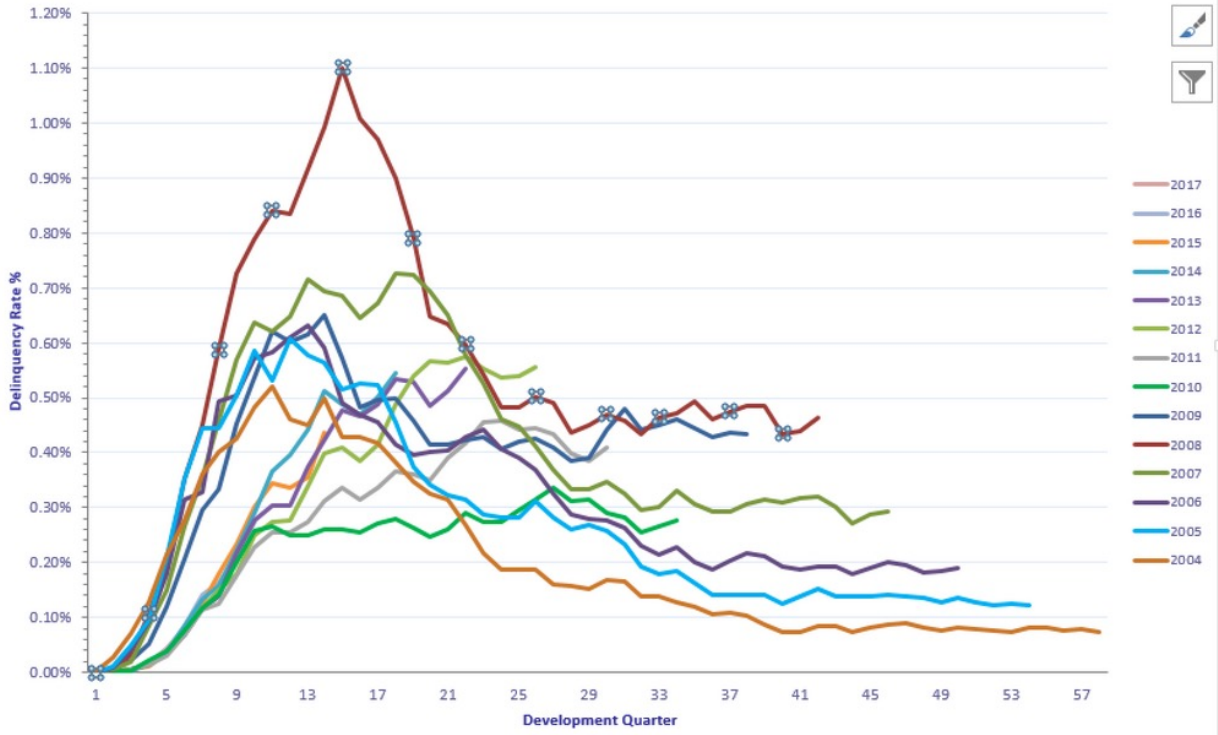


Image from:
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Need-Help-with-Multi-line-charts/td-p/244692

Multi-Line Graphs (2)

- Can also show ONE independent, and two (sometimes more) dependent variables
- Different axis might be needed for each dependent variable

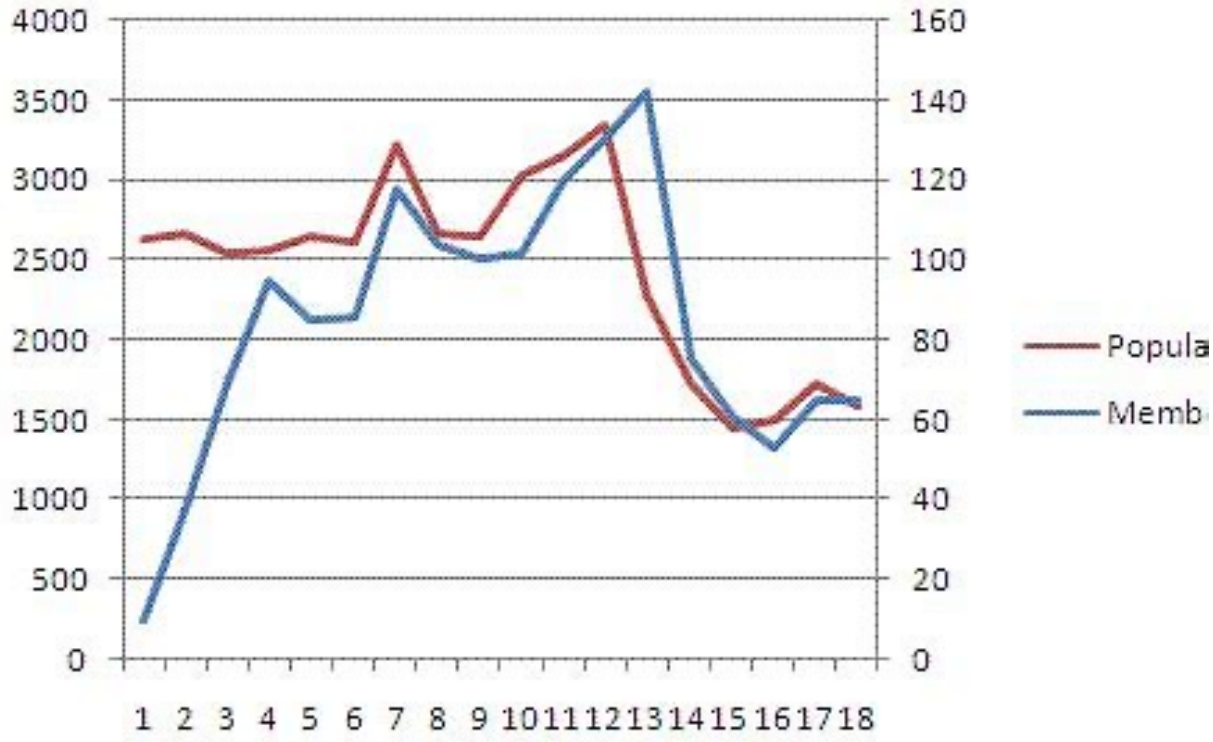


Image from:
https://community.powerbi.com/t5/Desktop/2-Y-axes/td-p/302068

Slopegraph

- Simple line chart
- One independent variable with only two values (horizontal)
- Many dependent variables
- Connecting lines for each variable: slope shows increase or decrease

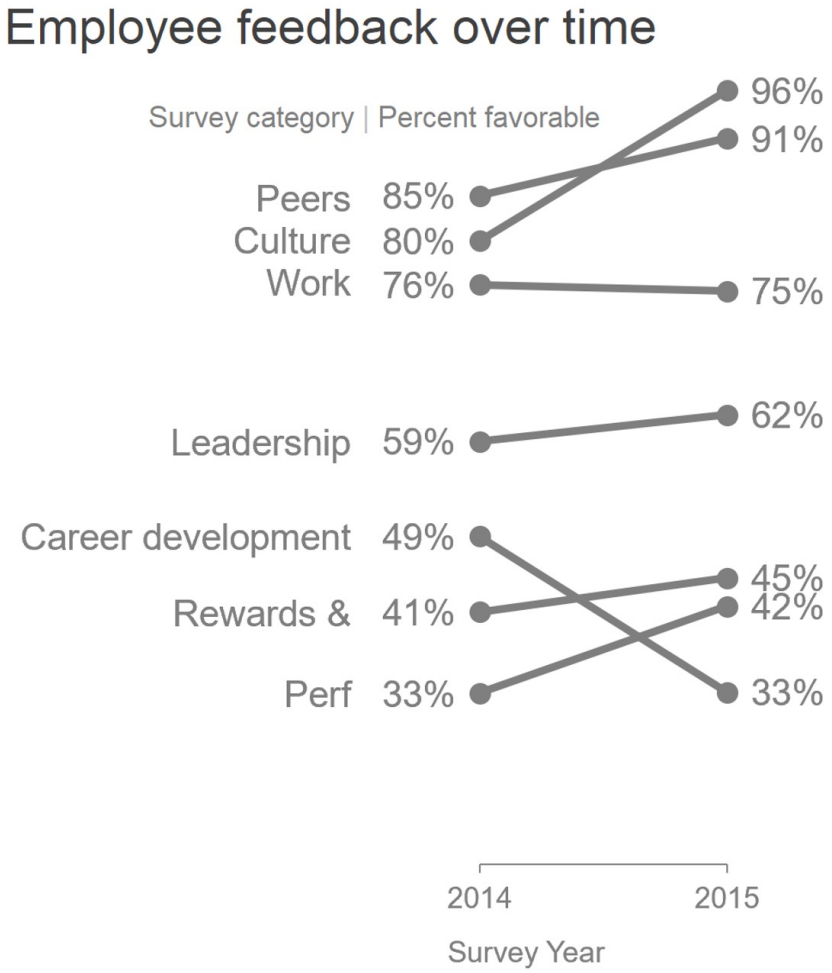


Image from:
https://adrianbaertschi.github.io/2020/06/20/storytelling-with-data-part1.html

Multiple Box Plots

- Independent variable: categories
- Can combine box plots of the dependent variable (distribution) in one graph

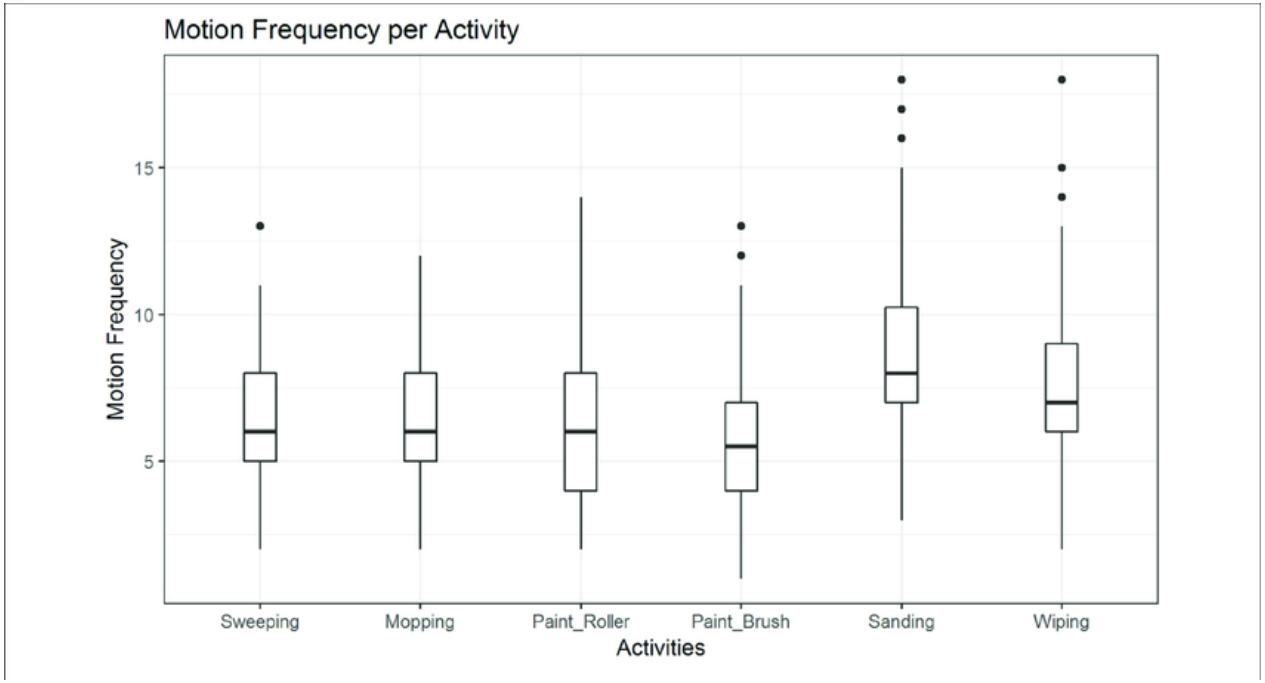


Image from:
https://www.researchgate.net/figure/Distribution-of-motion-frequency-across-activities-A-significant-difference-is-found_fig4_336098202

Simplified: High-Low Charts

- Range Bar Charts
- Graph only max/min values
  - And maybe mean

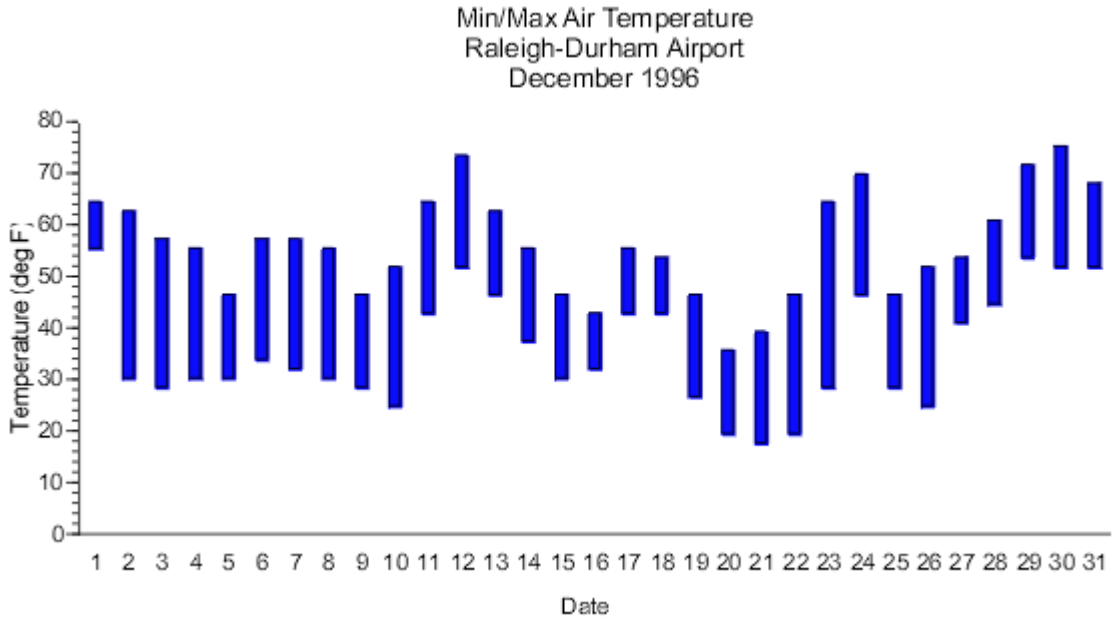


Image from:
https://labwrite.ncsu.edu//res/gh/gh-bargraph.html

Band Charts

- If independent variable is more continuous
- Fill in area between high/low values
- Can plot one category within the range, if helpful

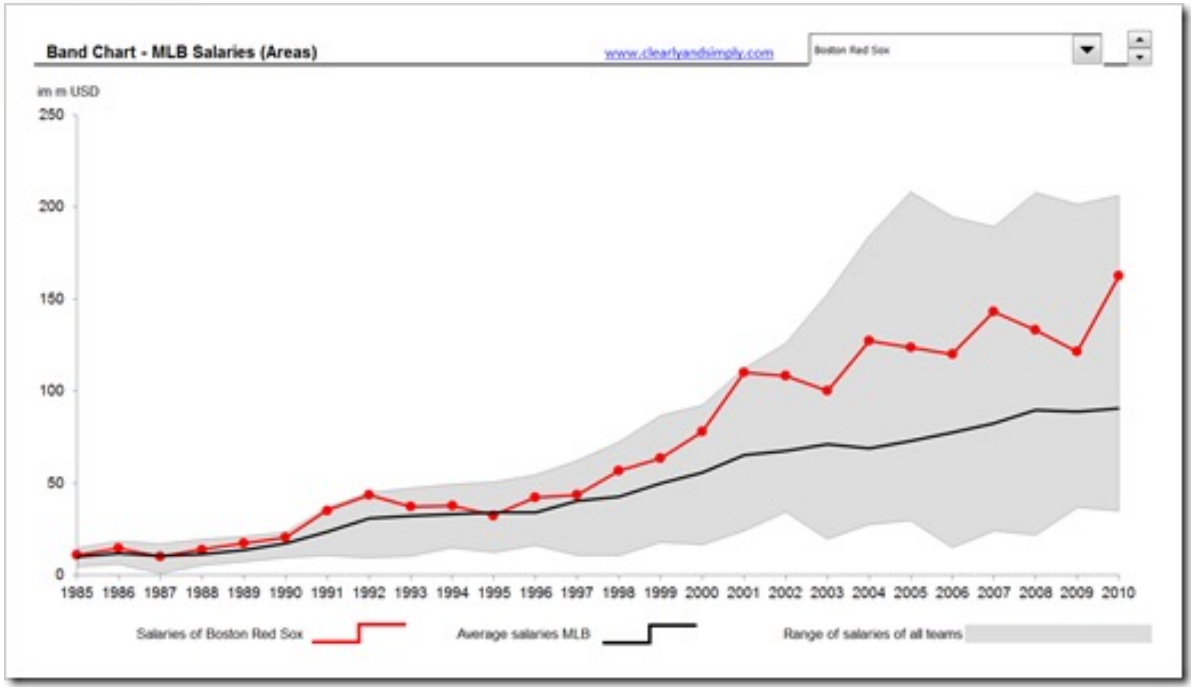


Image from:
https://www.clearlyandsimply.com/clearly_and_simply/2011/04/an-underrated-chart-type-the-band-chart.html

Streamgraph

- 1 continuous independent variable (typically time)
- 1 categorical independent variable
- 1 dependent variable
- Combination of stacked bar chart, area (line) chart
- Shows absolute value, broken up into categories, centered on 0

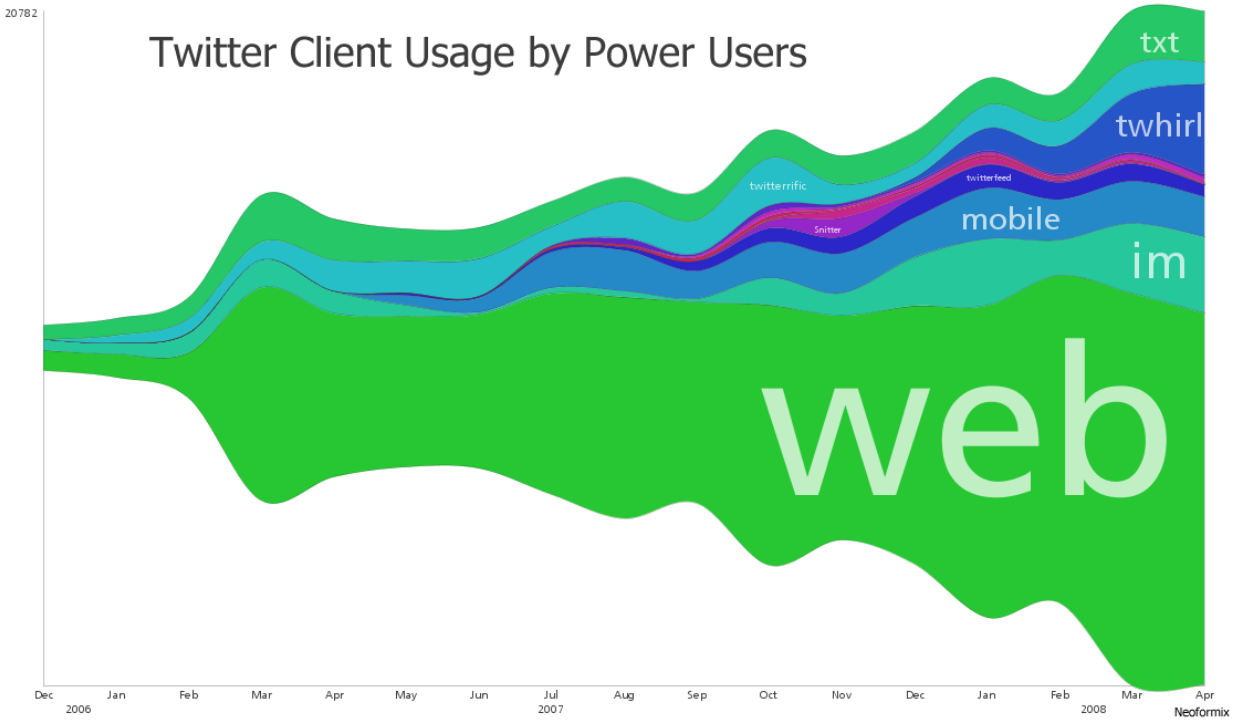


Image from:
http://www.neoformix.com/2008/TwitterClientStream.png

Bump Chart

- 1 independent variable (time)
- 1 dependent variable that shows ranking (only)
- Fixed rows show rank
- Connect with lines – shows progression of relative ranking over time

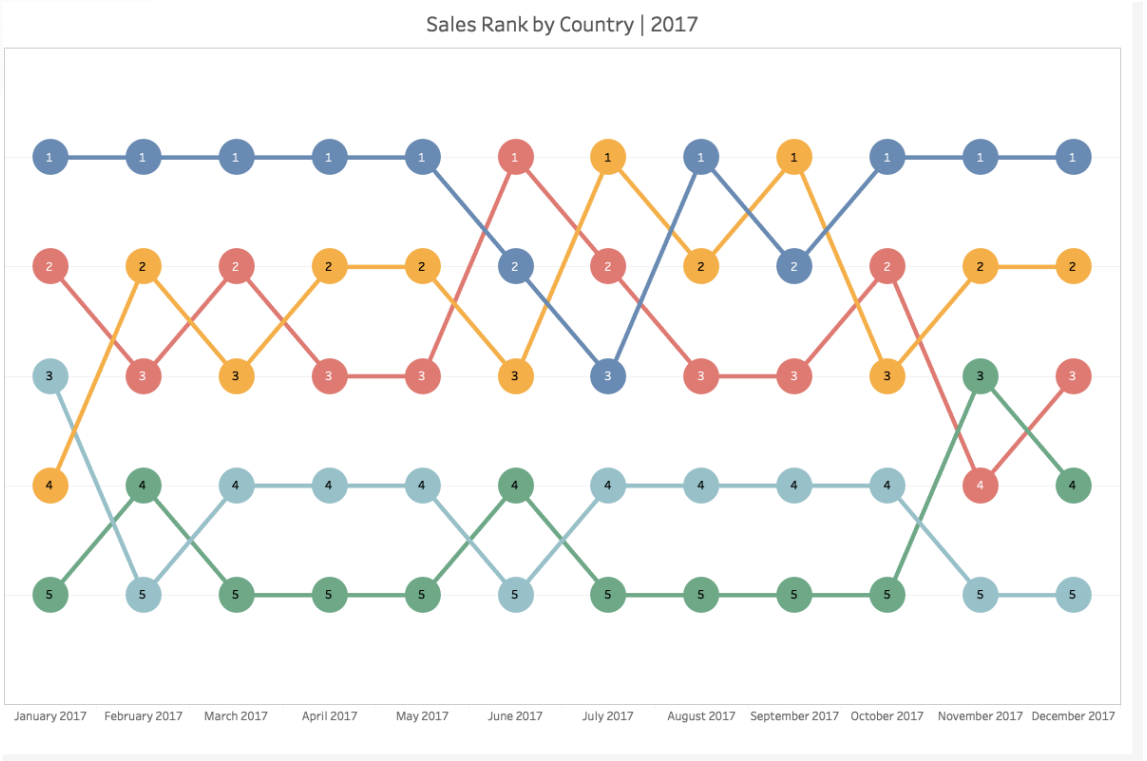


Image from:
https://darren.gosbell.com/2022/12/building-a-bump-chart-in-power-bi-using-deneb/

Icon Representations (Pictogram)

- Generally just an alternative to a bar chart
  - But, does not have to be displayed in a line; icons can be displayed in groups
- Can be less precise, but give greater visual impact than bar charts

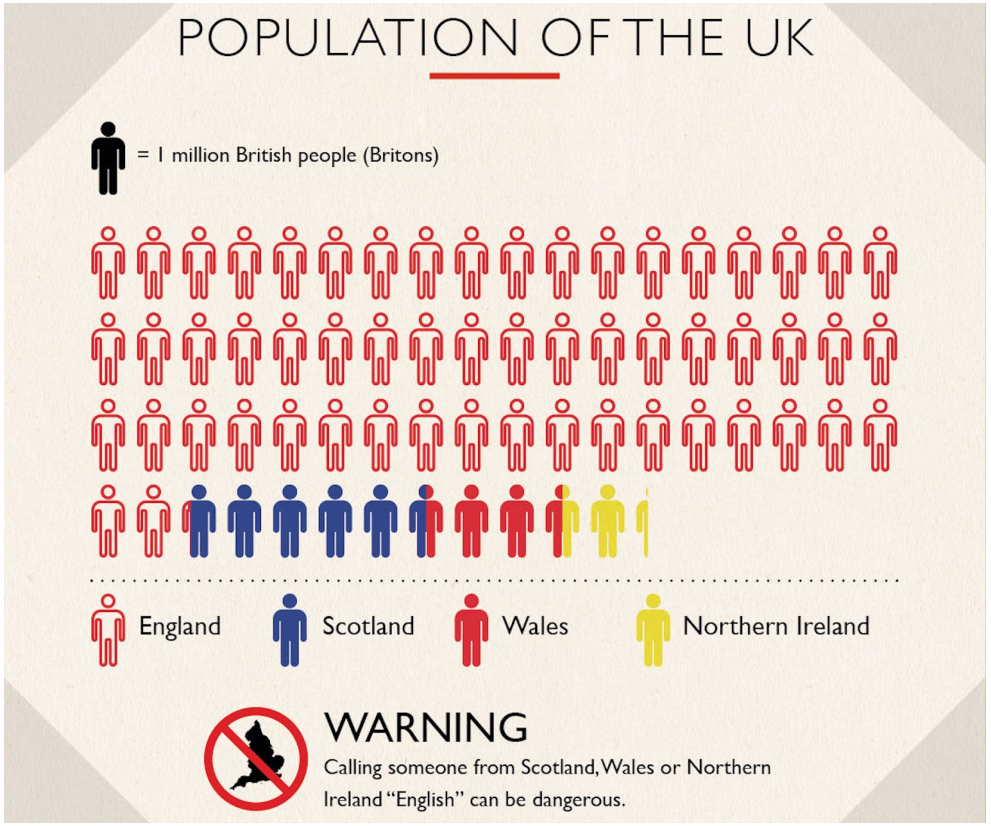


Image from:
https://www.easel.ly/blog/how-to-make-pictograms/

Waffle Chart

- Alternative to a pie chart or stacked bar graph
  - Can be easier to compare amounts
- Can also just show a percentage of the whole
- Area is divided into a grid (typically 10x10)
- Fill in area based on percentage

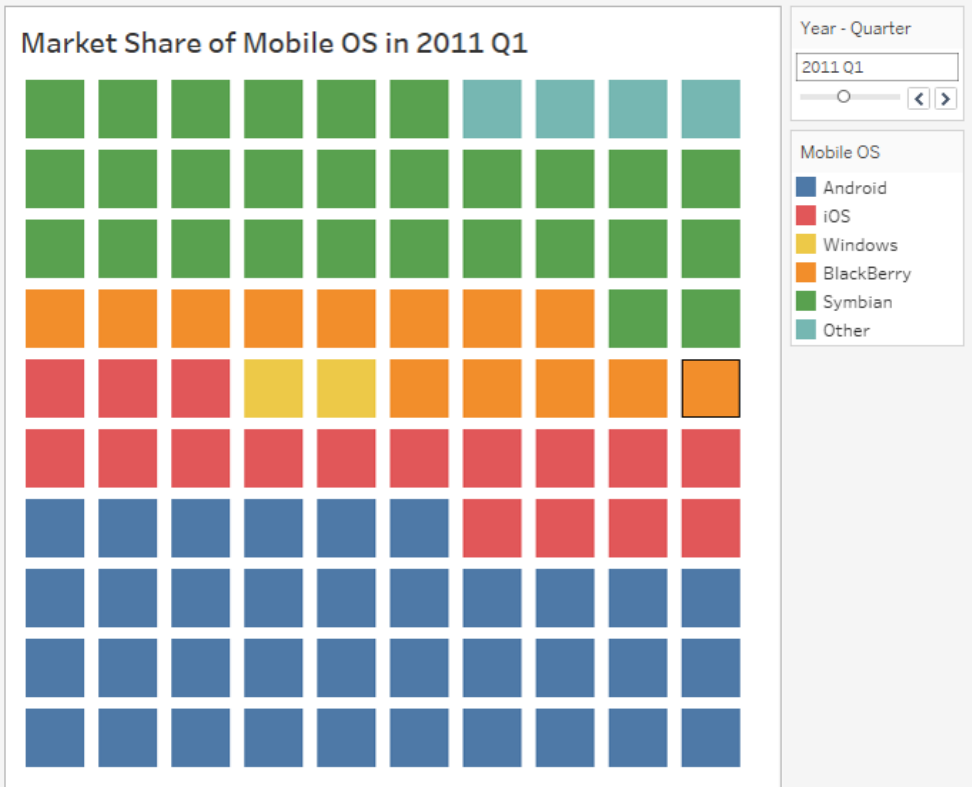


Image from:
https://fhsuguides.fhsu.edu/dataviz/partsof

Scatter Plot

- Single independent, 2 dependent variables
  - Each data point is the independent
  - Plot against 2 different axes (for dependent variables)
  - Option to label points

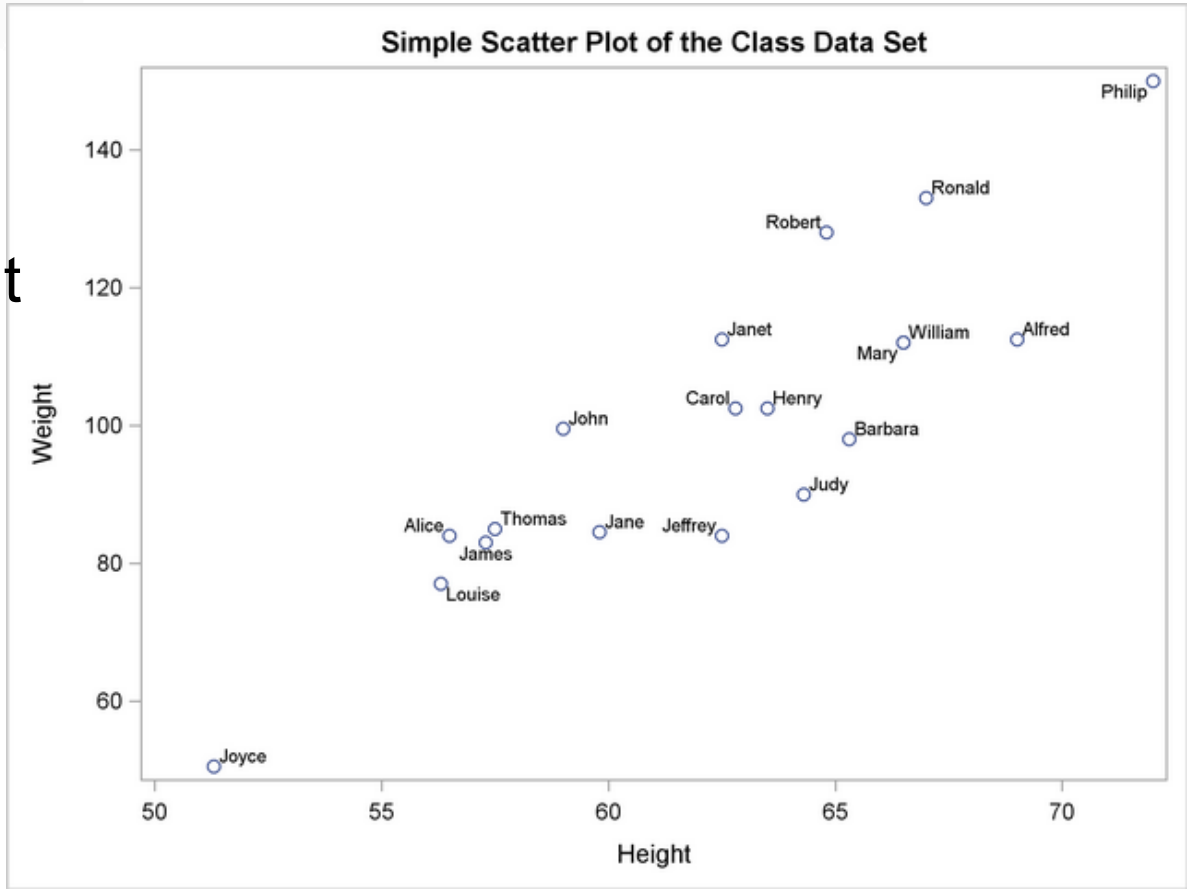
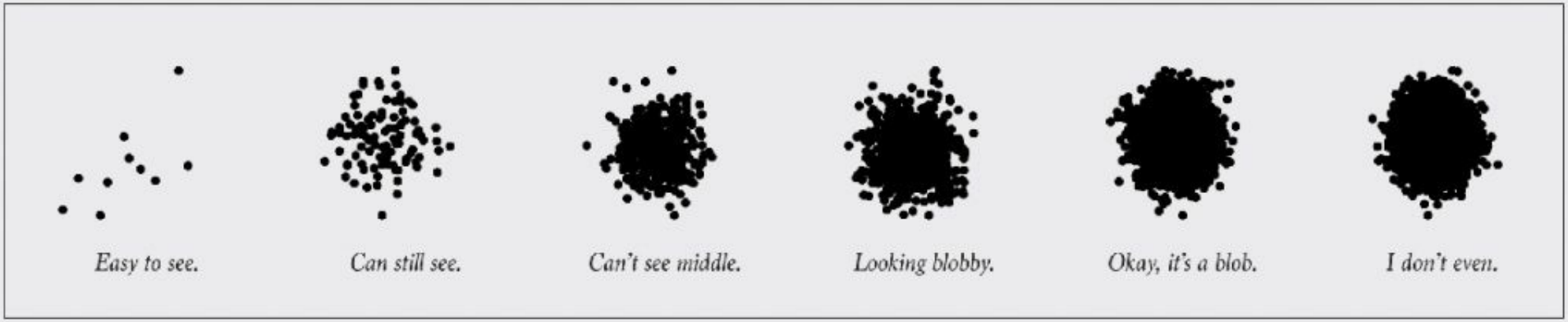


Image from:
https://stackoverflow.com/questions/28665507/how-to-i-create-a-labelled-scatter-plot

Overplotting

- If there are too many points, scatter plot points overwrite
  - Difficult to distinguish; loss of information



- Can cluster points – e.g. only draw point for every x points in radius
  - Will miss single separate points
- Small overplotting can also be avoided by moving points away from each other
  - But this starts introducing false data, and can only be used for small overplotting issues

Image from:
https://visme.co/blog/scatter-plot/

“ADVANCED” MULTIVARIATE VISUALIZATIONS

Heat Map

- 2 Independent, 1 Dependent Variable
  - Discretize the independent variables
  - For each, intensity indicates value of dependent

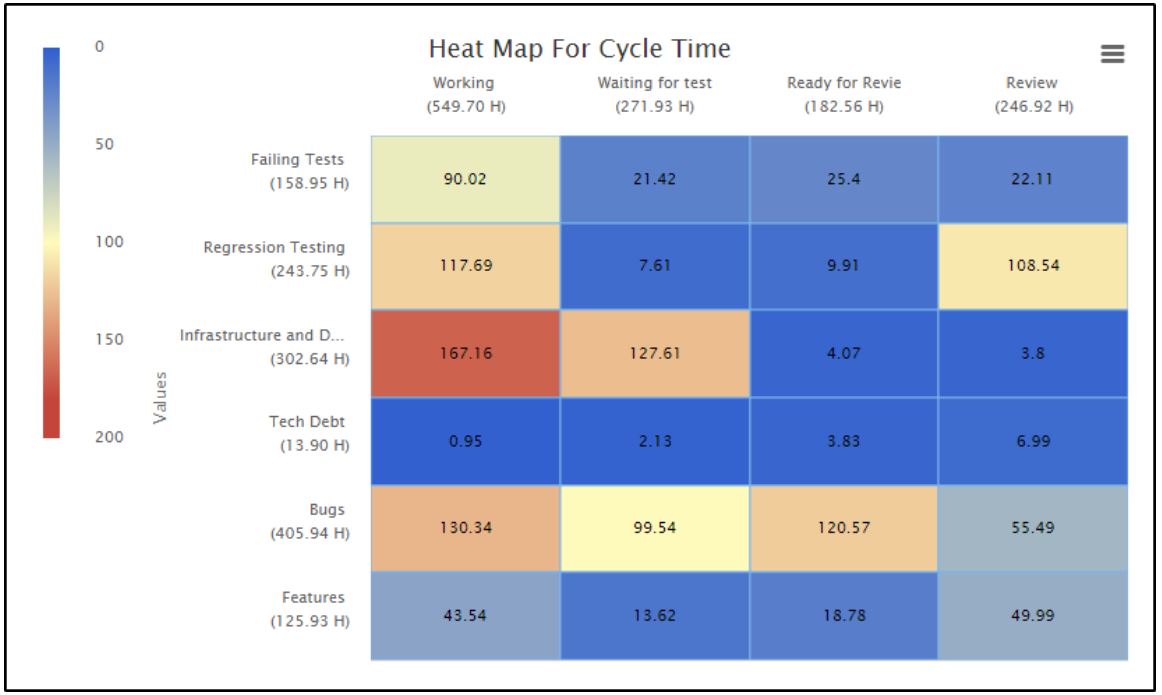


Image from:
https://go2.fullstory.com/heatmap-tools.html

Heat Map as a 2D Histogram

- Can be thought of as a histogram of a scatter plot
  - Independent variables are bins in each of the original “dependent” variables
  - Dependent variable is the total number of samples in that area
- Can be used this way when too many points would lead to overplotting
- This is an example of a “2D Histogram”

Heat Map as a 2D Histogram

- Cumulative clicks on areas of a website

Image from:
https://sitetuners.com/blog/do-heat-maps-really-work/

Bubble Charts

- Enhancing Scatter Plots by drawing points with different size/color can give an additional variable or 2 (dependent or independent)

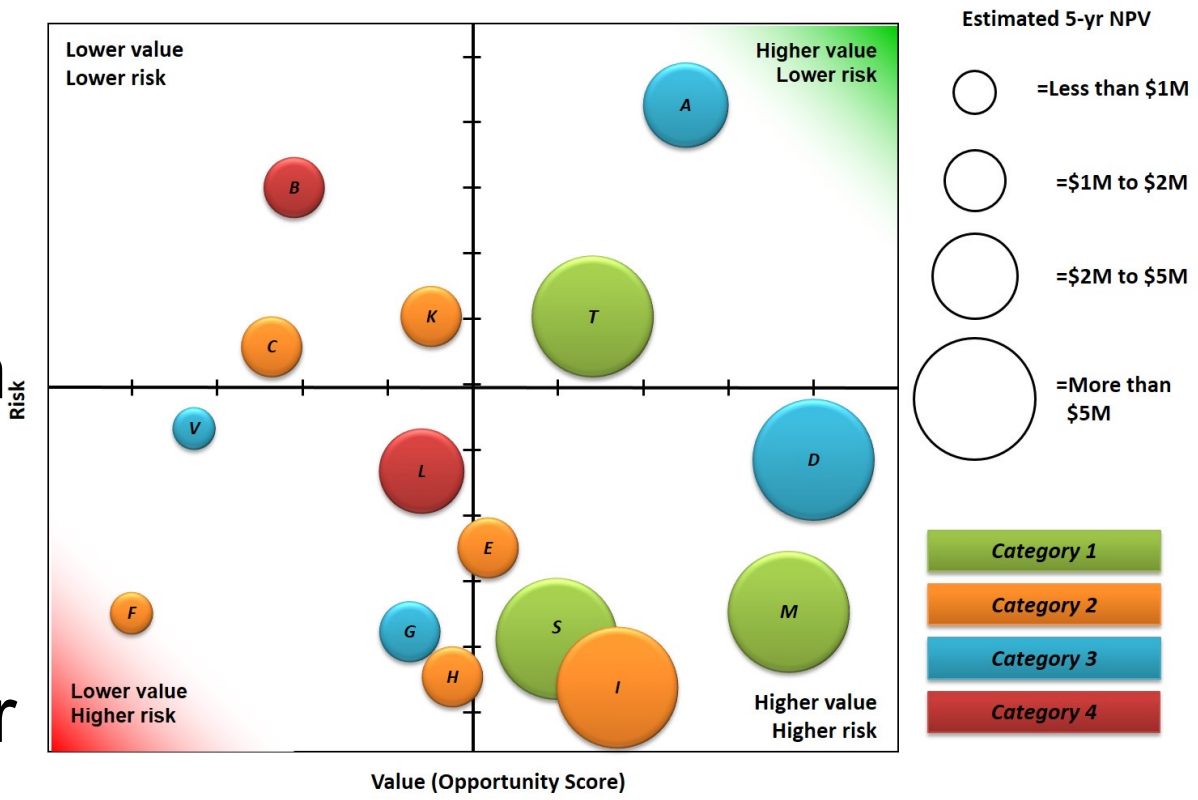


Image from:
https://www.planview.com/products-solutions/solutions/project-portfolio-management/

Scatter Plot Matrix

- To deal with many additional dependent variables, can make matrix of Scatter Plots
- Each plot in the matrix is combination of one scatter plot variable vs. another
  - N variables -> NxN matrix of plots
- Diagonal is usually a histogram
  - Since can't plot variable vs. itself
  - And, overall chart is a reflection over diagonal, so sometimes only half is used
- Helps identify pairwise relationships

Scatter Plot Matrix

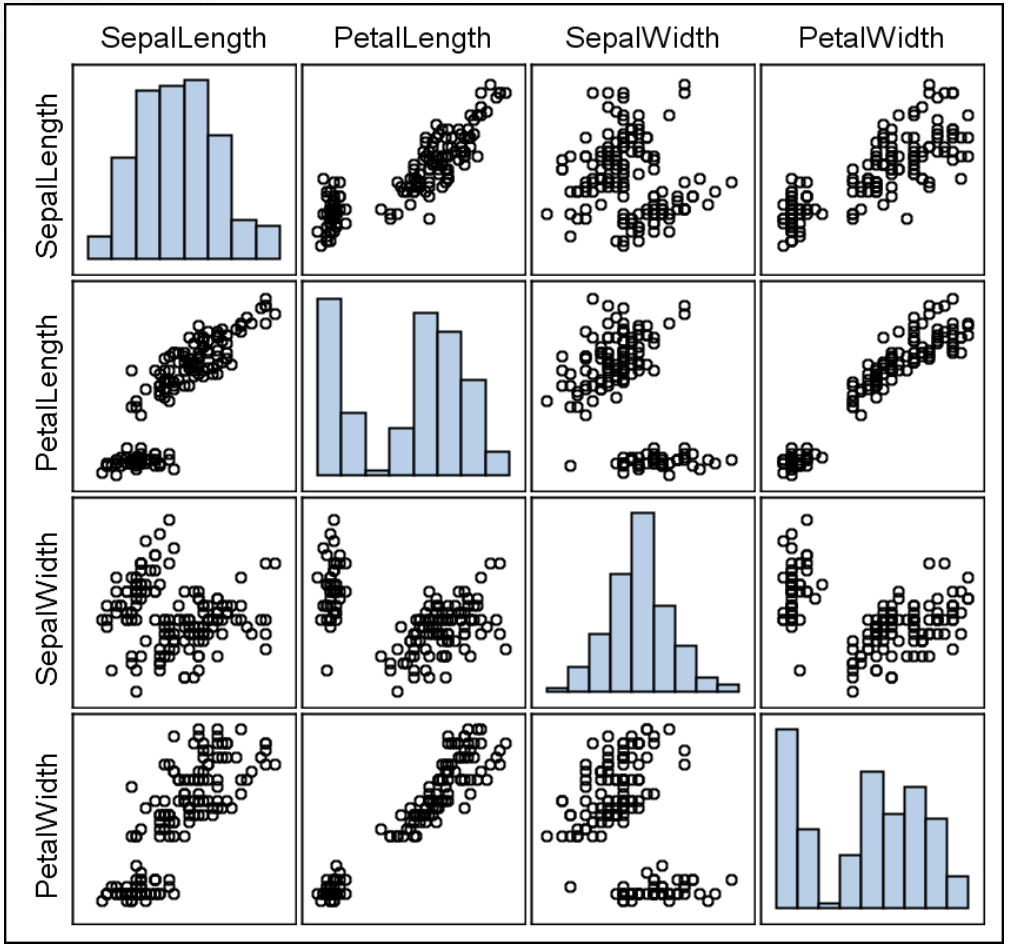


Image from:
https://blogs.sas.com/content/graphicallyspeaking/2012/10/07/scatter-plot-matrix-with-a-twist/

Scatter Plot Matrix

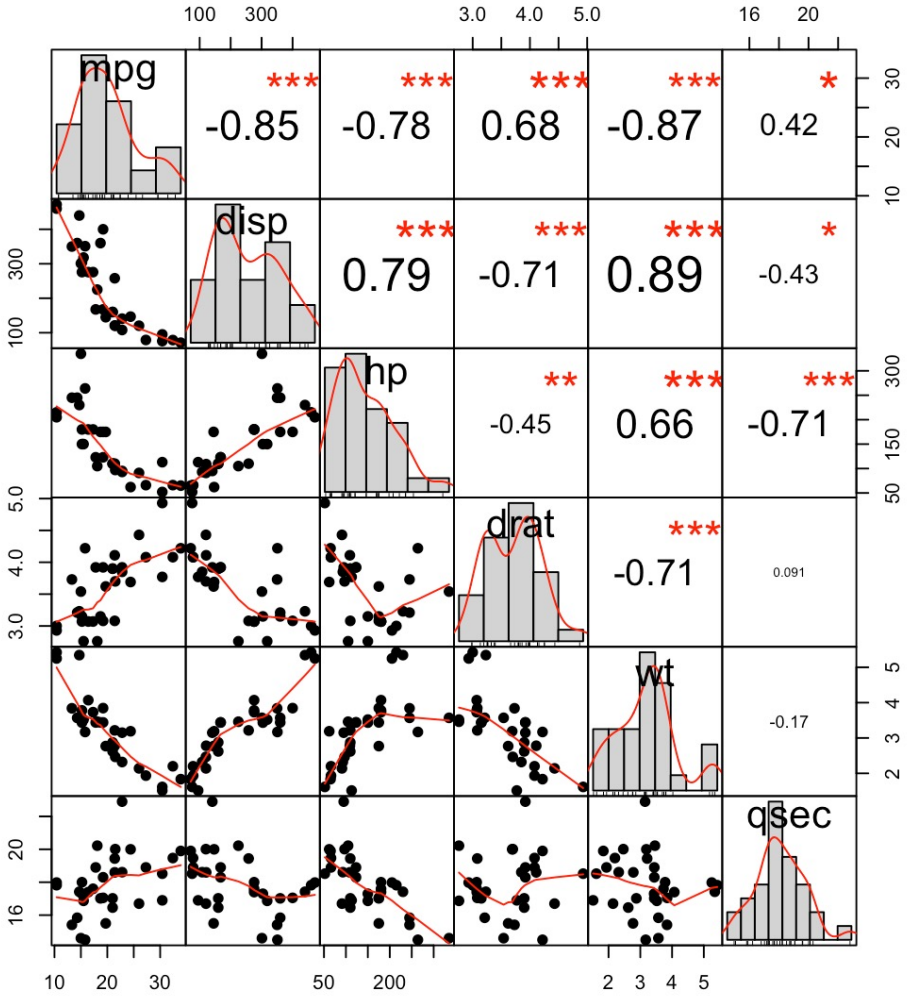


Image from:
http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

Matrix of Charts

- To deal with additional independent variables, can make matrix of charts
  - Arbitrary chart types
- Rows can identify variation in one variable, columns another
  - For more than 2, can group hierarchically
- Works for a limited set of independent values (otherwise grid is too large)

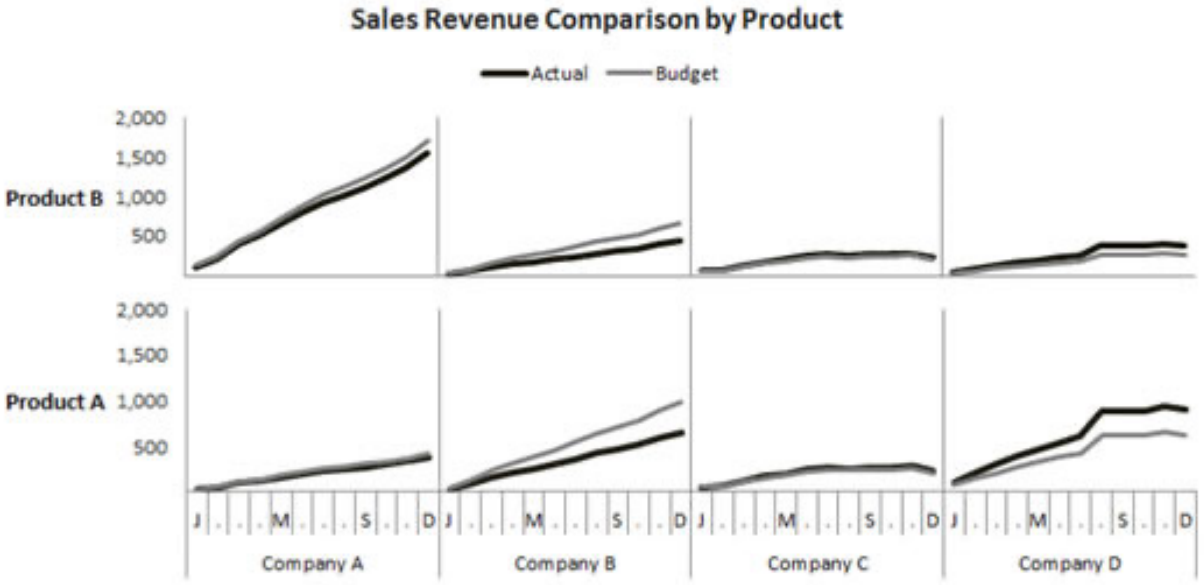


Image from:
https://www.lucidchart.com/pages/examples/orgchart_software

Parallel Coordinates

- One independent variable, many other variables
  - Can be independent or dependent
- Will create one axis for each variable
  - Axis scale or ordering can be chosen per-variable
  - Ordering of the axes can be chosen
- Each instance (value of the ind. variable) is one polyline, through each axis.

Parallel Coordinates

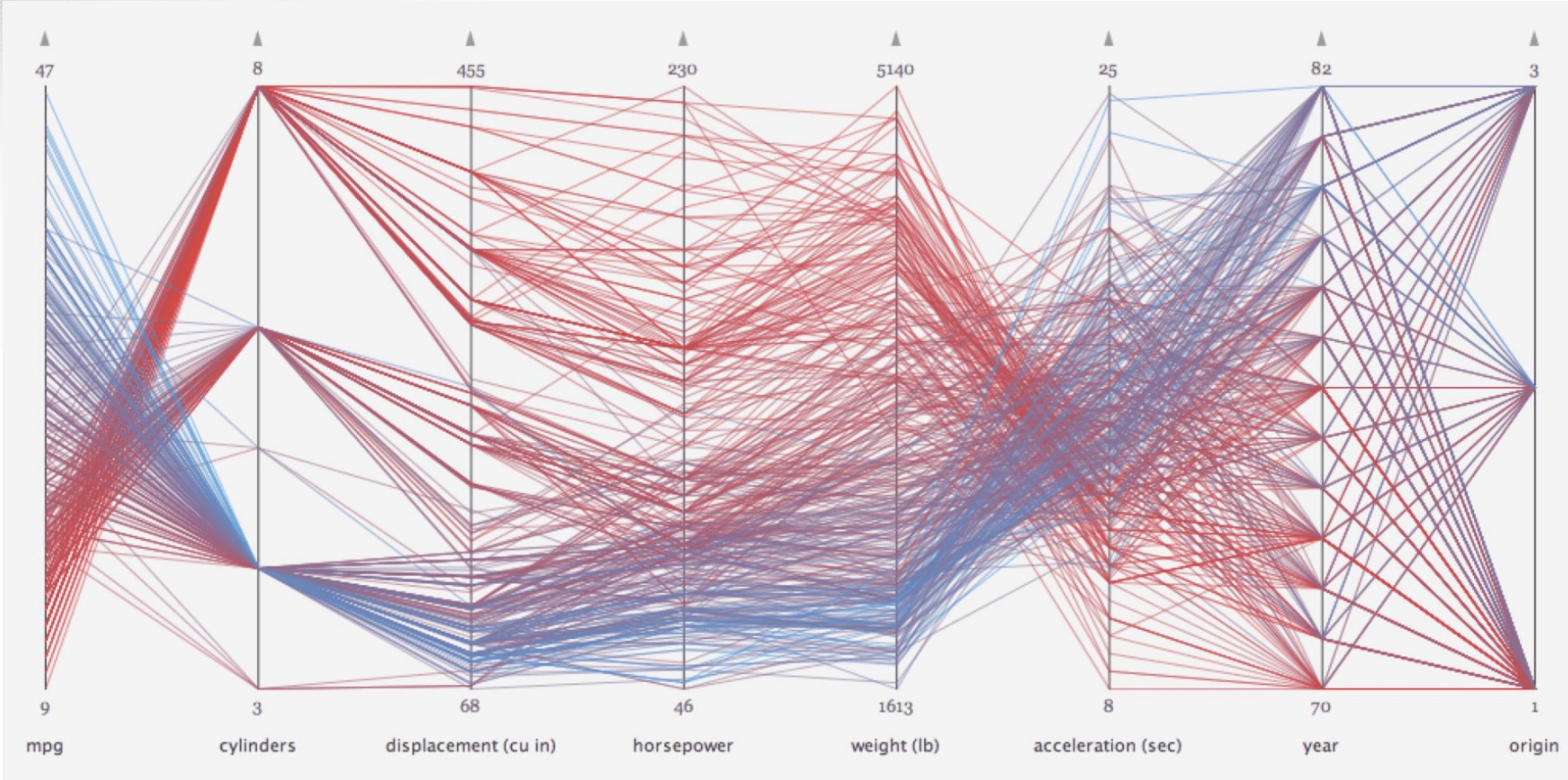


Image from:
https://datavizproject.com/data-type/parallel-coordinates/

Understanding Parallel Coordinates

- Can infer relationships from line patterns
  - Parallel lines: positive correlation
  - X shape: negative correlation
  - Random: difficult to tell

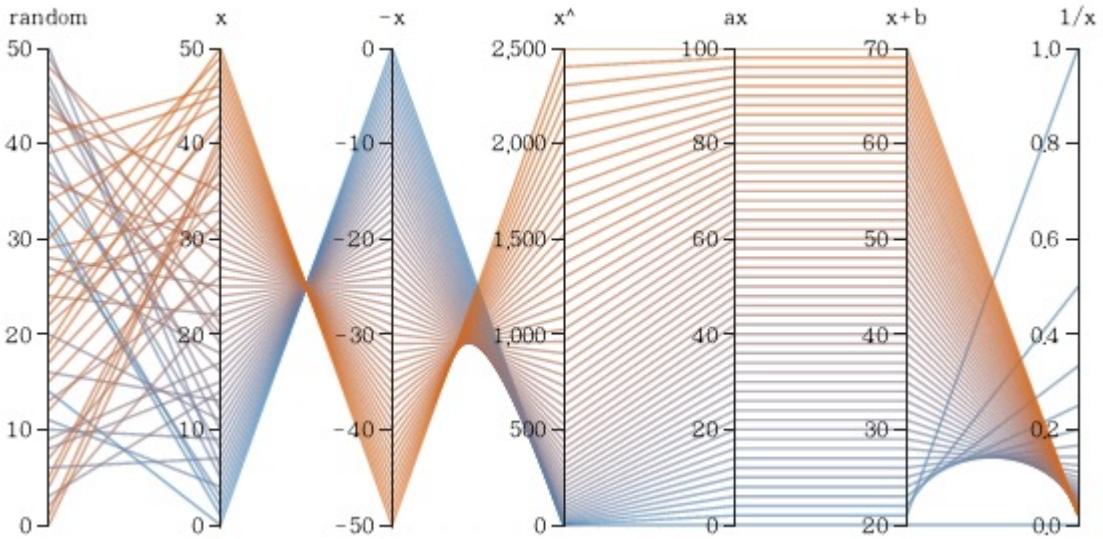


Image from:
https://en.wikipedia.org/wiki/Parallel_coordinates

Designing Parallel Coordinates

- The order of axes makes a big difference
- The scale of each axis makes a difference
- Color helps understand flow
- Interactive charts tend to allow much more examination of the data

Image from:
https://en.wikipedia.org/wiki/Parallel_coordinates

Understanding Parallel Coordinates

- ```
  px.parallel_coordinates(iris, color="species_id", color_continuous_scale=["red", "green", "blue"])
  ```

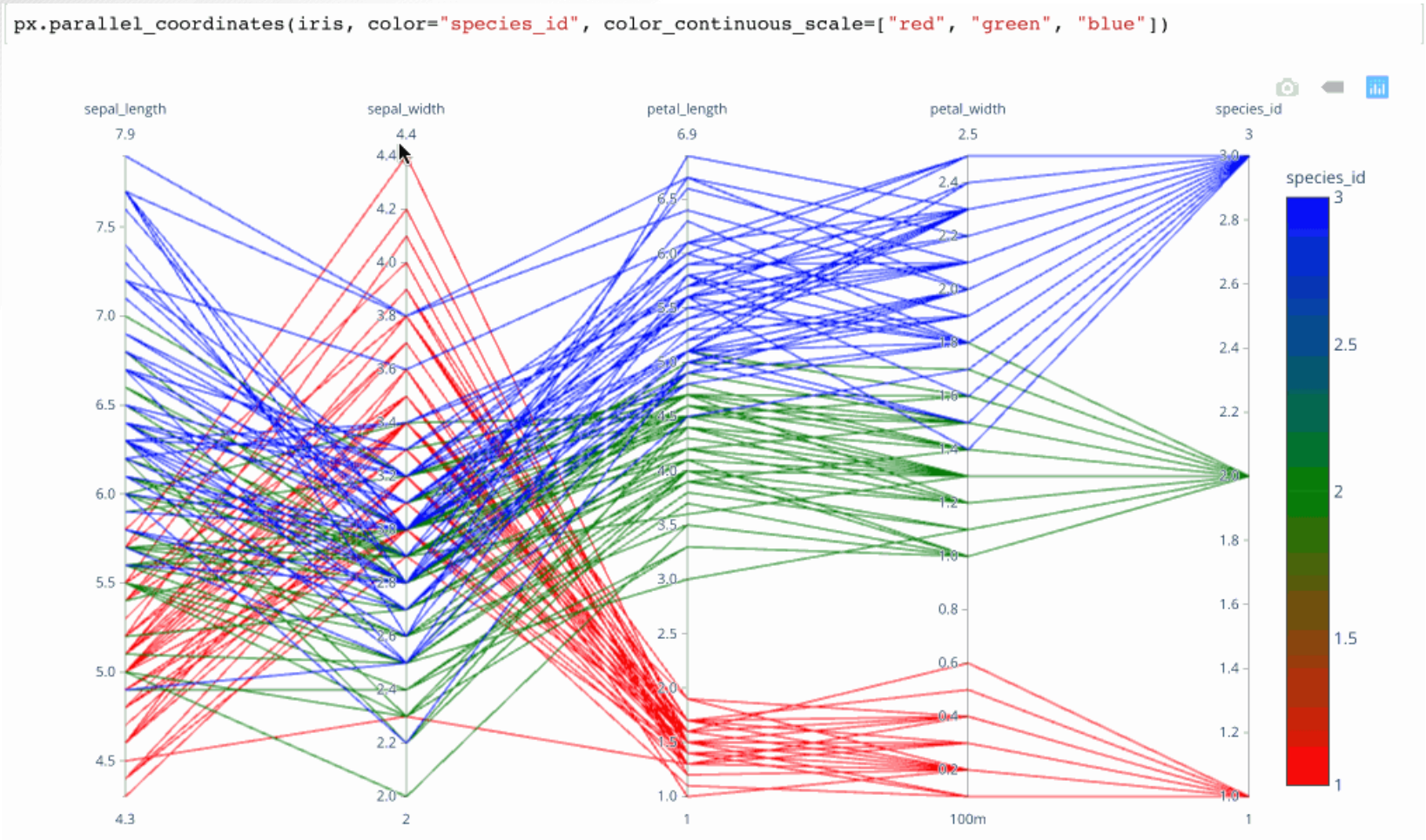


Image from:
https://medium.com/plotly/introducing-plotly-express-808df010143d
  ```

Star/Radar Charts

- Plotting multivariate data, usually for a single entity or small number of entities
- Arrange the axes for each variable radially from the center
- Each entity forms a polygon connecting the values on each axis

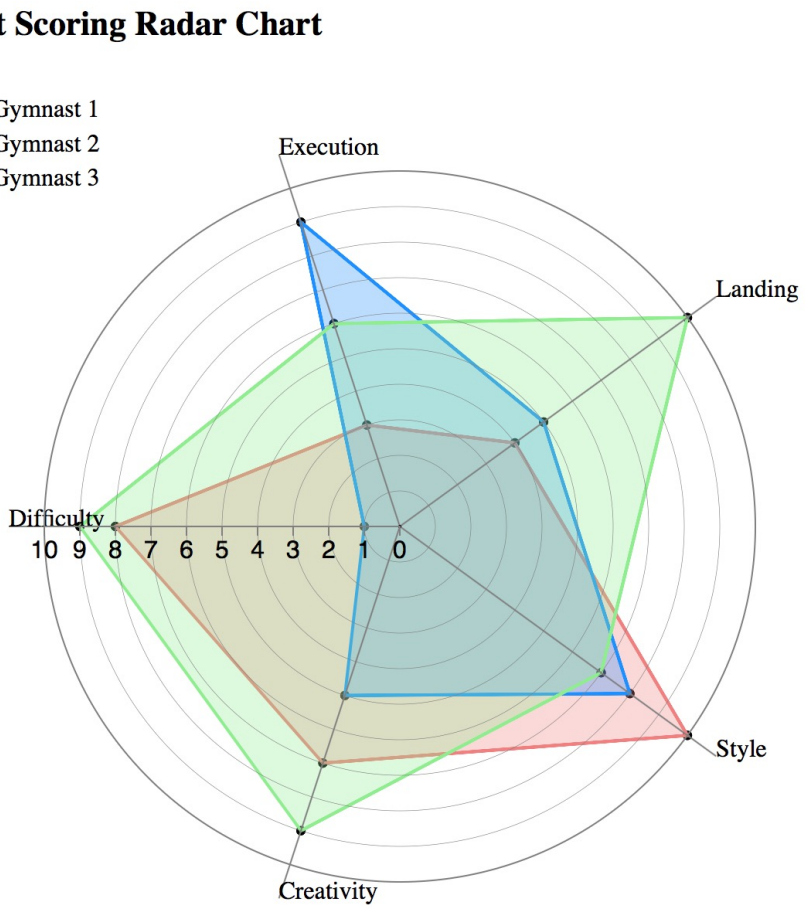


Image from:
https://www.cs.middlebury.edu/~candrews/showcase/infovis_techniques_s16/radar_chart/

Star/Radar Charts

- Like Parallel Coordinates, but limited number of entities, arrangement of axes is different
- Axis order, scaling can still matter, though
- Careful: can be helpful or deceiving
  - e.g. each axis contributes equally to visual “weight” of the polygon

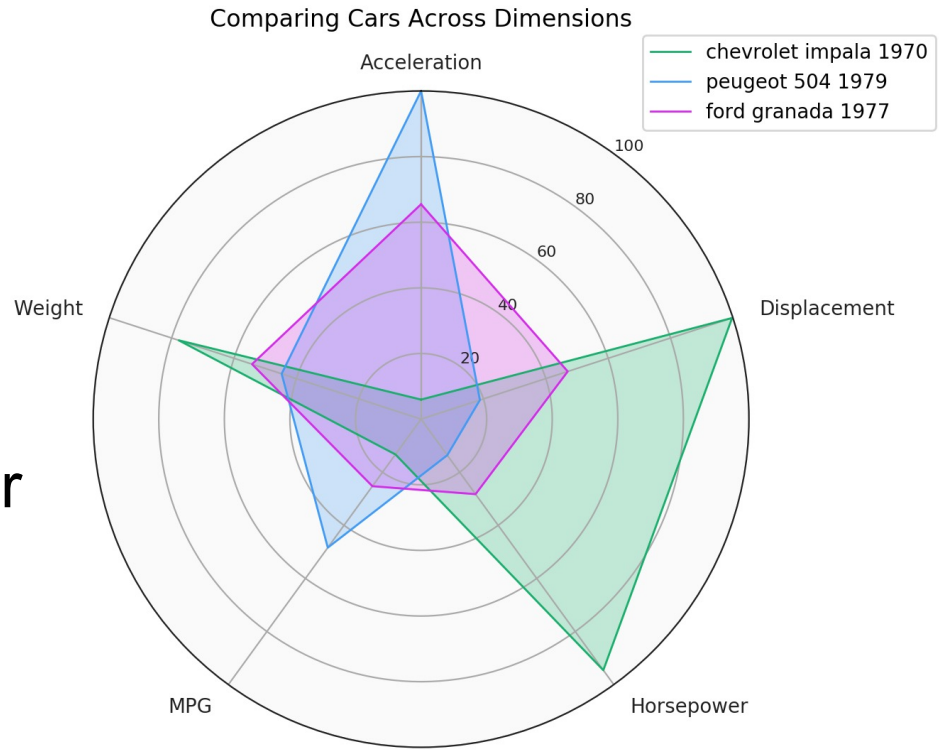


Image from:
https://www.pythoncharts.com/matplotlib/radar-charts/

Using Color, Shape, Time, etc.

- Viewing additional variables is a challenge
  - So, people have come up with ways to try to display more at once
  - This often leads to difficulty in perception, understanding
- We will return to some of these later.

Using Color, Shape, Time, etc.

- Color can reflect one variable
  - Independent or Dependent
- Shape can reflect one variable
  - e.g. in scatter plot, to determine category
- Size can reflect one variable
  - e.g. in Bubble Plot
- Texture (image/pattern overlay)
  - For categorization
- Time variation adds one dimension
  - But is complicated to use effectively (other than for representing time itself)