Data Visualization

Why Visualize Data?

Why Visualize Data?

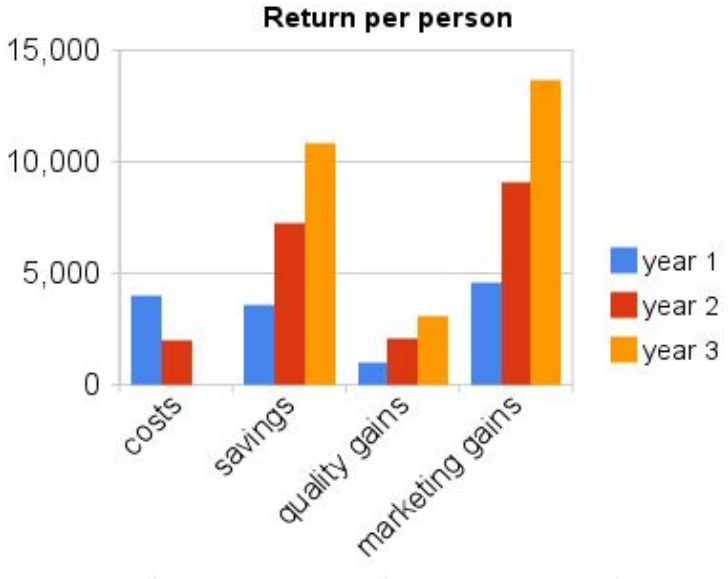
- Abstraction
- Analysis
- Context
- Insight
- Communication
- Outlier Detection

Abstraction

- Computers are good with details - humans are not
- Recall: Abstraction - turn details into ideas
- I.e. ignore individuals and reason about populations
- Abstractions are pragmatic - pick an abstraction to fit a purpose

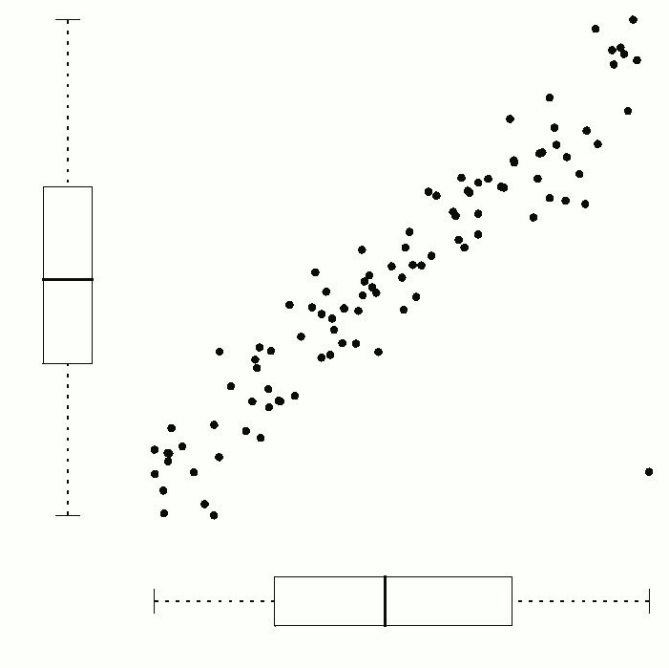
Analysis

- Even after processing and querying data, it can be difficult to interpret
- Visualize data to extract meaning
- When information is abstracted, analysis becomes more obvious



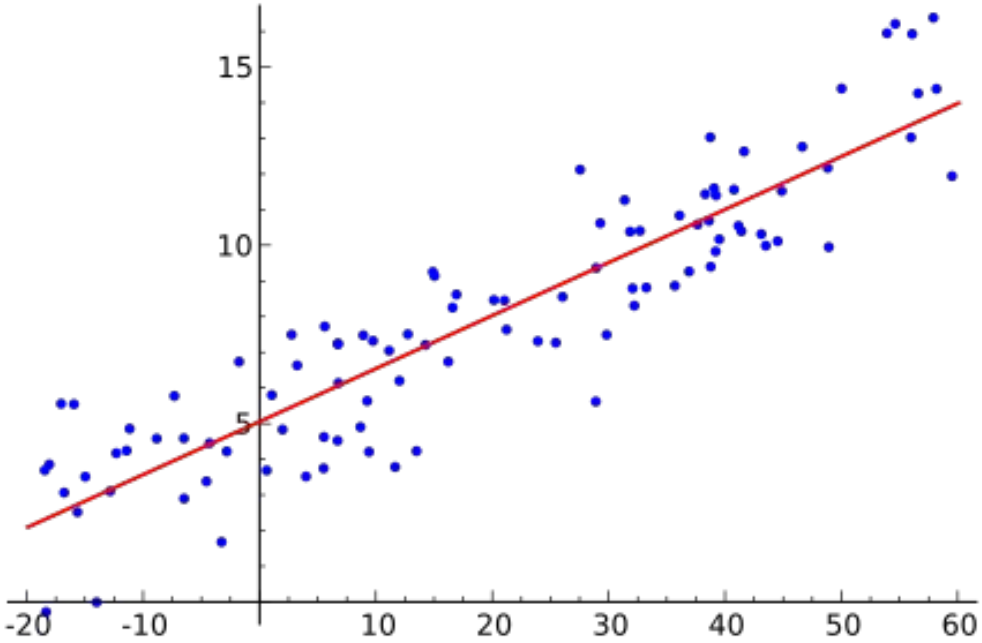
Context

- A single value on its own ignores the bigger picture
- A key aspect of analysis is **comparison**



Insight

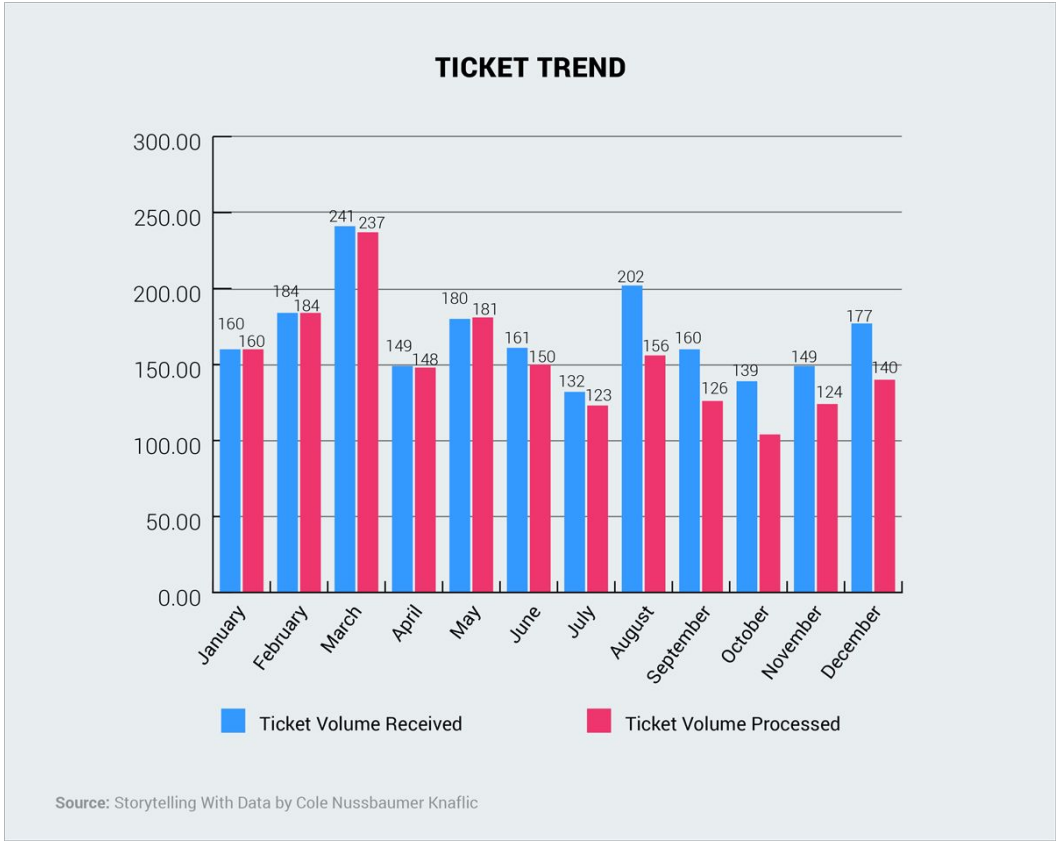
- Notice relationships between data points that would not be obvious when just looking at numbers
- Get new ideas about the data



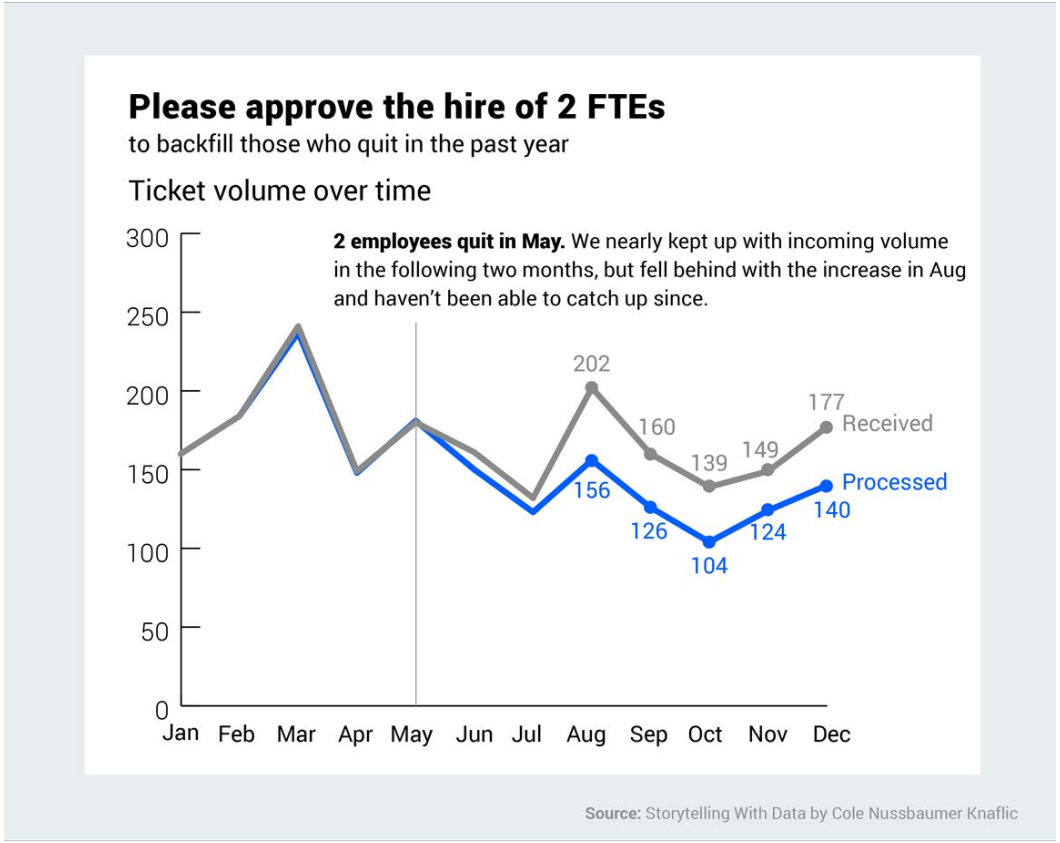
Communication

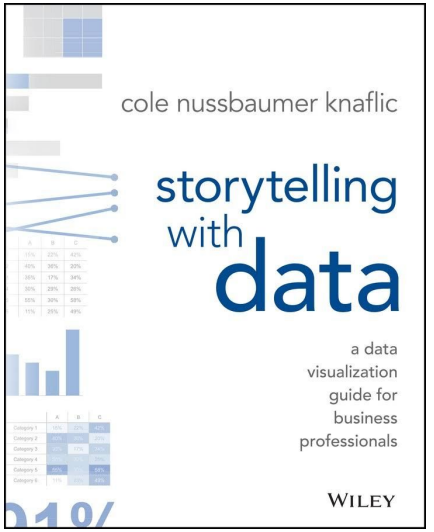
- Insight is useless without action
- Action involves communicating findings to others
- Even if “The numbers speak for themselves”, how they are presented will affect how others interpret them
- Visuals are a much more efficient way to share information than numbers

Communication



Communication





Storytelling With Data
by Cole Nussbaumer Knaflic

Worth watching (long): https://www.youtube.com/watch?v=8EMW7io4rSI
Worth watching (short): https://www.youtube.com/watch?v=6xsvGYIxJok

Outlier Detection

- Visualizations make outliers really obvious
- Outlier detection is useful for many applications:
  - Detecting errors
  - Cleaning data
  - Understanding where an analysis/prediction fails
  - Monitoring

Visualization Mediums

Dashboards

- Persistent
- Automated (Real time-ish)
- Interactive
- Purposes:
  - Monitoring
  - Decision making
  - Enabling others

Reports

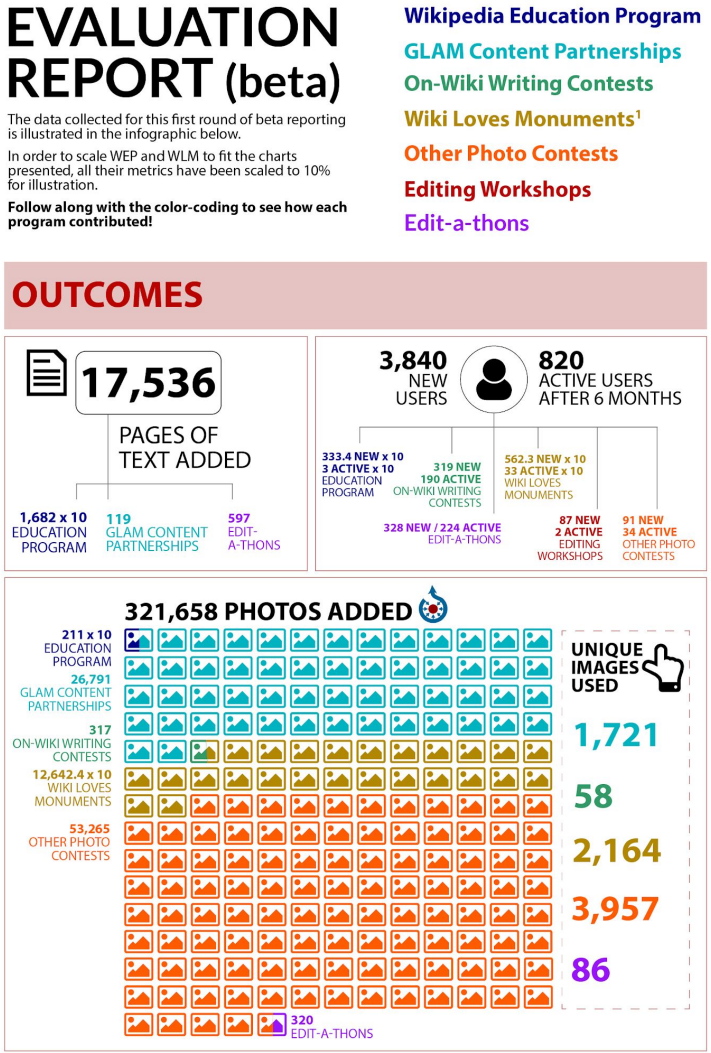
- Can be automated or manually produced
- Often produced with a set frequency (but not necessarily)
- Mix of text and visualizations
- Visualizations are key to quickly informing report consumers
- Not interactive

Insights

- Visualizations can be used to highlight or emphasize a particular finding
- A visualization with the purpose of pointing out a particular idea is sometimes called an insight
- Insights can be included in reports or dashboards
- Commonly used in meetings as a discussion topic
- When building a business case (around a decision) insights are invaluable

Infographics

- Combines text and visualizations
- Often done by a graphic designer
- Meant for public consumption or marketing



https://commons.wikimedia.org/wiki/File:CROP_Evaluation_Report_(beta)_Infographic.png

Types of graphs

Product Type	Region	Store Type	Sales
Shoes	West	Kiosk	145
Shirts	West	Kiosk	85
Pants	West	Kiosk	46
Shoes	West	Outlet	241
Shirts	West	Outlet	143
Pants	West	Outlet	89
Shoes	East	Kiosk	154
Shirts	East	Kiosk	101
Pants	East	Kiosk	32
Shoes	East	Outlet	216
Shirts	East	Outlet	205
Pants	East	Outlet	67

Bar Plot

- https://python-graph-gallery.com/barplot/
- Shows relationship between numerical and categorical variable
- Great for comparing aggregations from a “group-by” operation

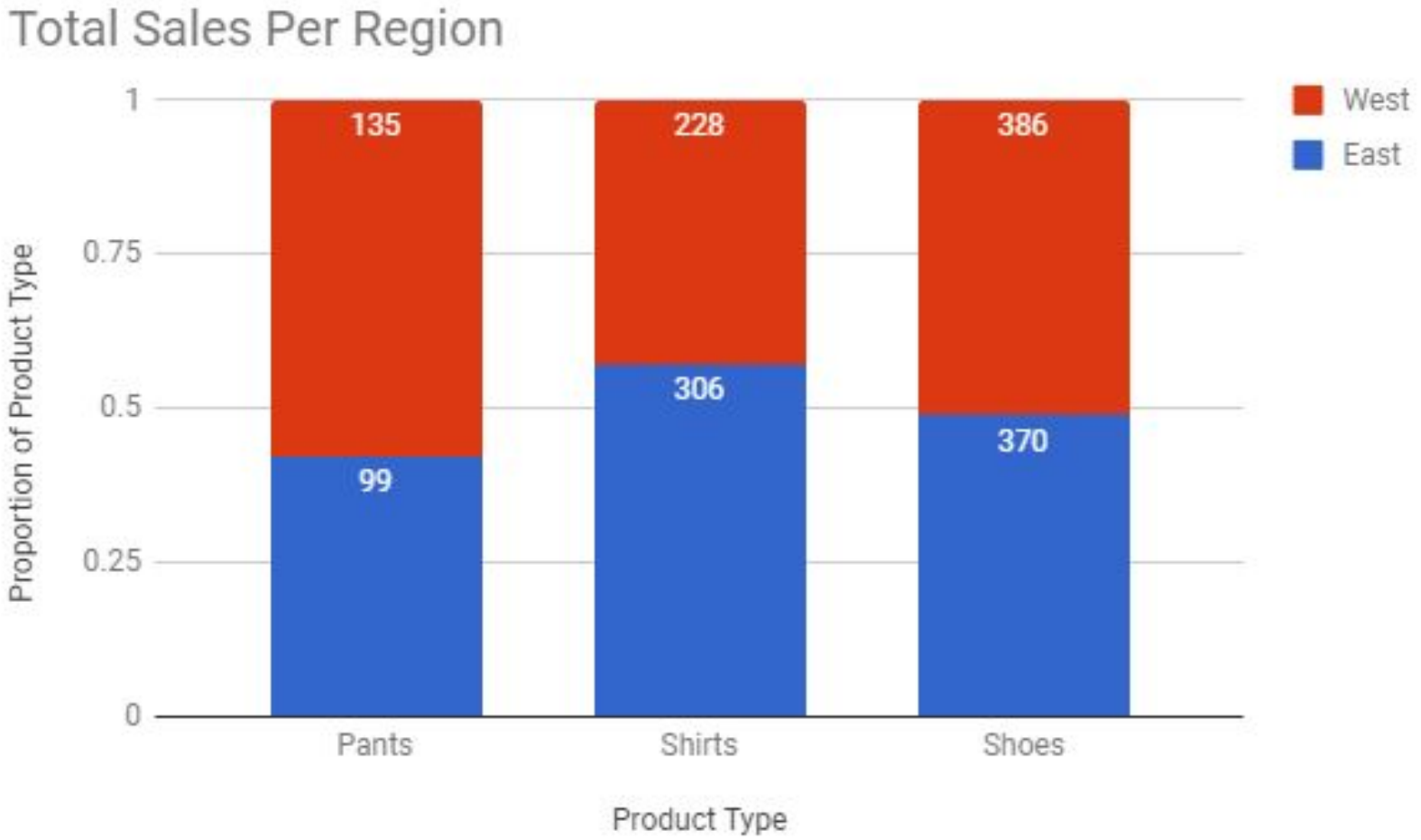

Total Sales
800
600
400
200
0
Sales Count
756
534
234
Shoes
Shirts
Pants
Product Type
Sales

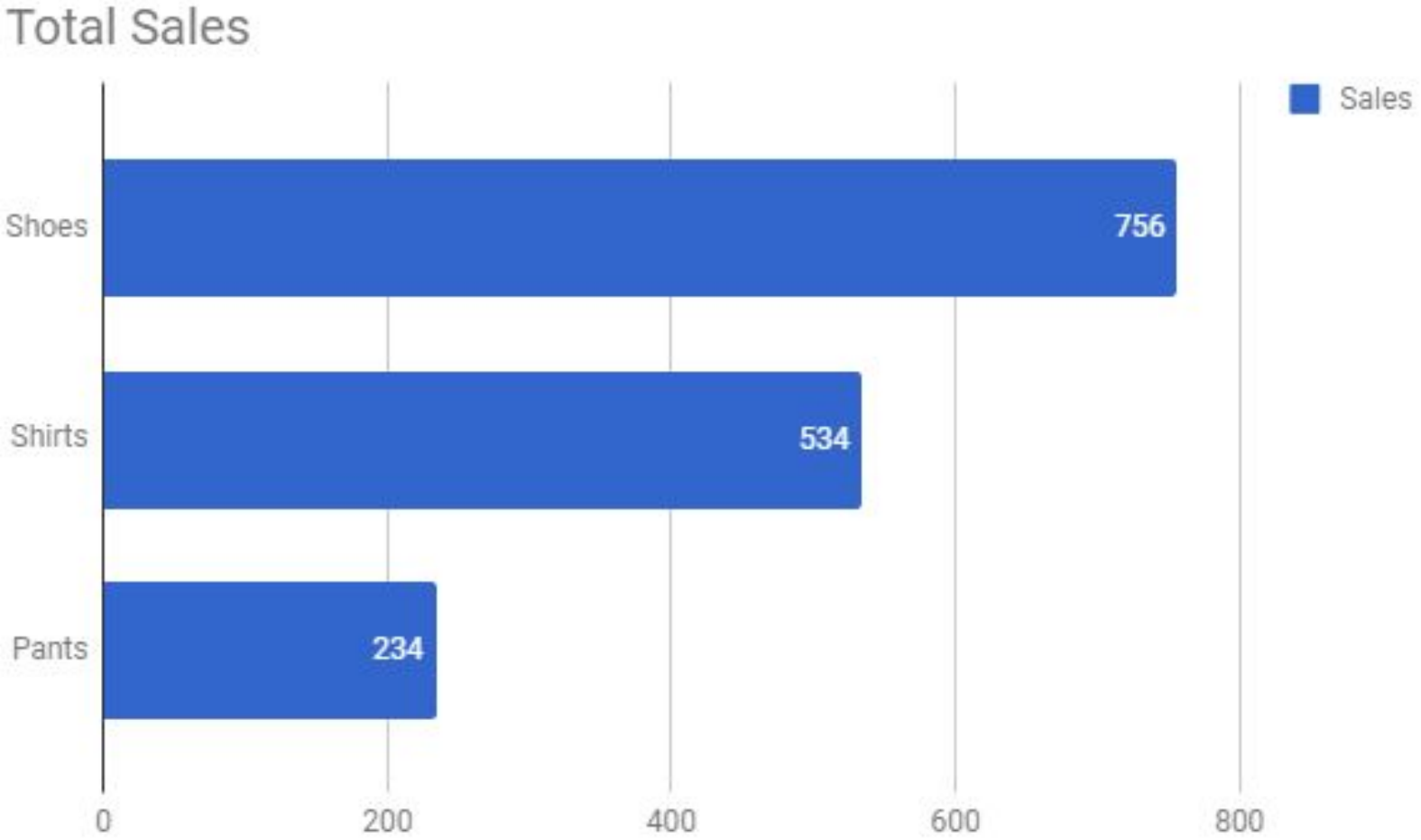

Total Sales Per Region
400
300
200
100
0
East
West
99
135
306
228
370
386
Pants
Shirts
Shoes
Product Type

Stacked Bar Plot

- https://python-graph-gallery.com/stacked-barplot/
- For visualizing sub-groups
- Stacking vs. Grouping Bars
  - Grouping:
    - For comparisons across **all** sub-groups
  - Stacking:
    - Helps with comparing totals for a group
    - Helps with comparing proportion of a sub-group
- Normalize the bars if proportion is all you care about

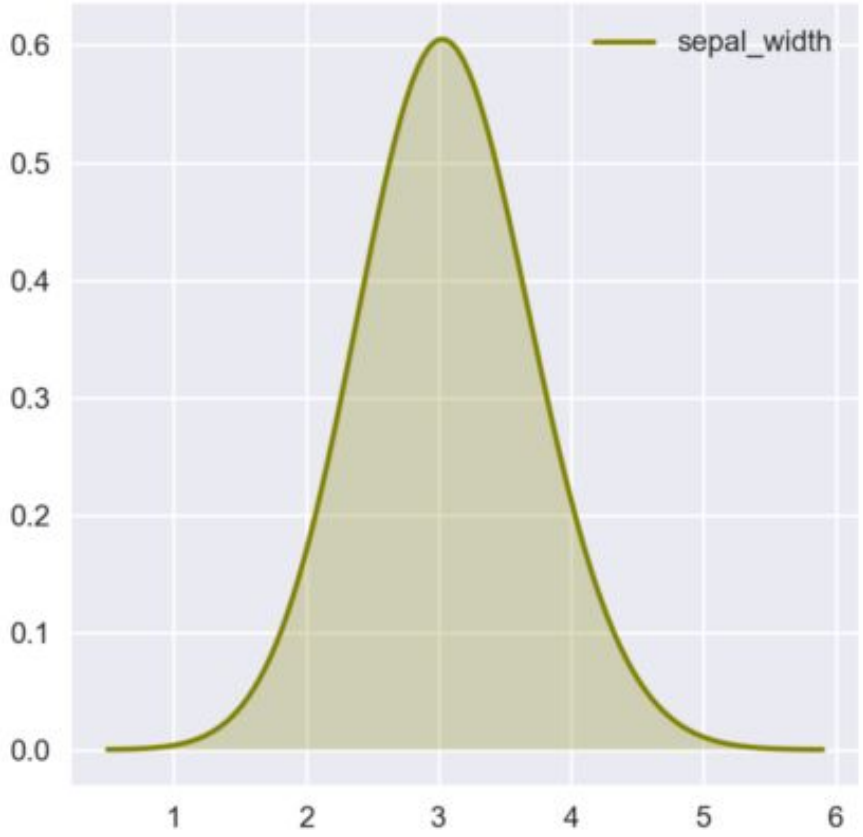

Total Sales Per Region
800
600
400
200
0
West
East
386
228
370
306
135
99
Pants
Shirts
Shoes
Product Type


Total Sales Per Region
1
0.75
0.5
0.25
0
Proportion of Product Type
135
228
386
306
99
370
West
East
Pants
Shirts
Shoes
Product Type

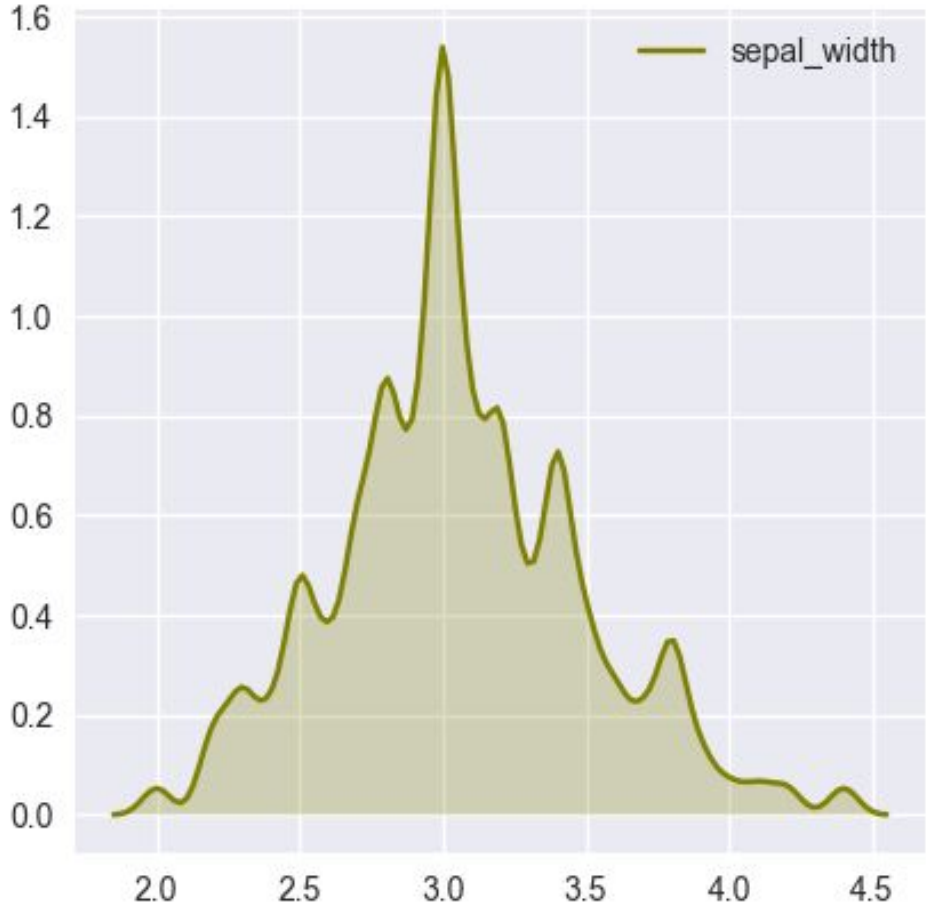

Total Sales
Sales
Shoes
Shirts
Pants
756
534
234
0
200
400
600
800

Density Plot

- https://python-graph-gallery.com/density-plot/
- For showing the **distribution** of a numerical variable
- Bandwidth parameter - determines smoothness of plot
  - Higher values = smoother curve, close approximation of underlying distribution
  - Lower values = more fluctuations, curve fits tighter to the data



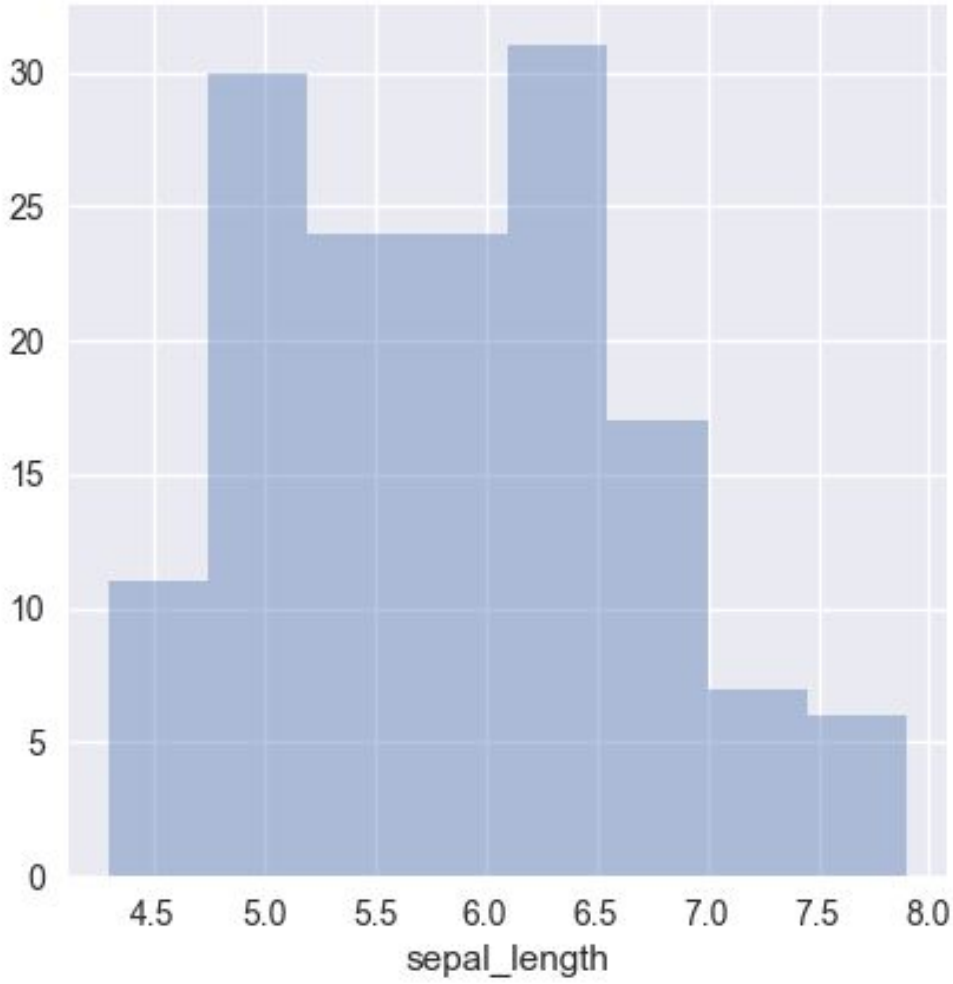
https://python-graph-gallery.com/73-control-bandwidth-of-seaborn-density-plot/

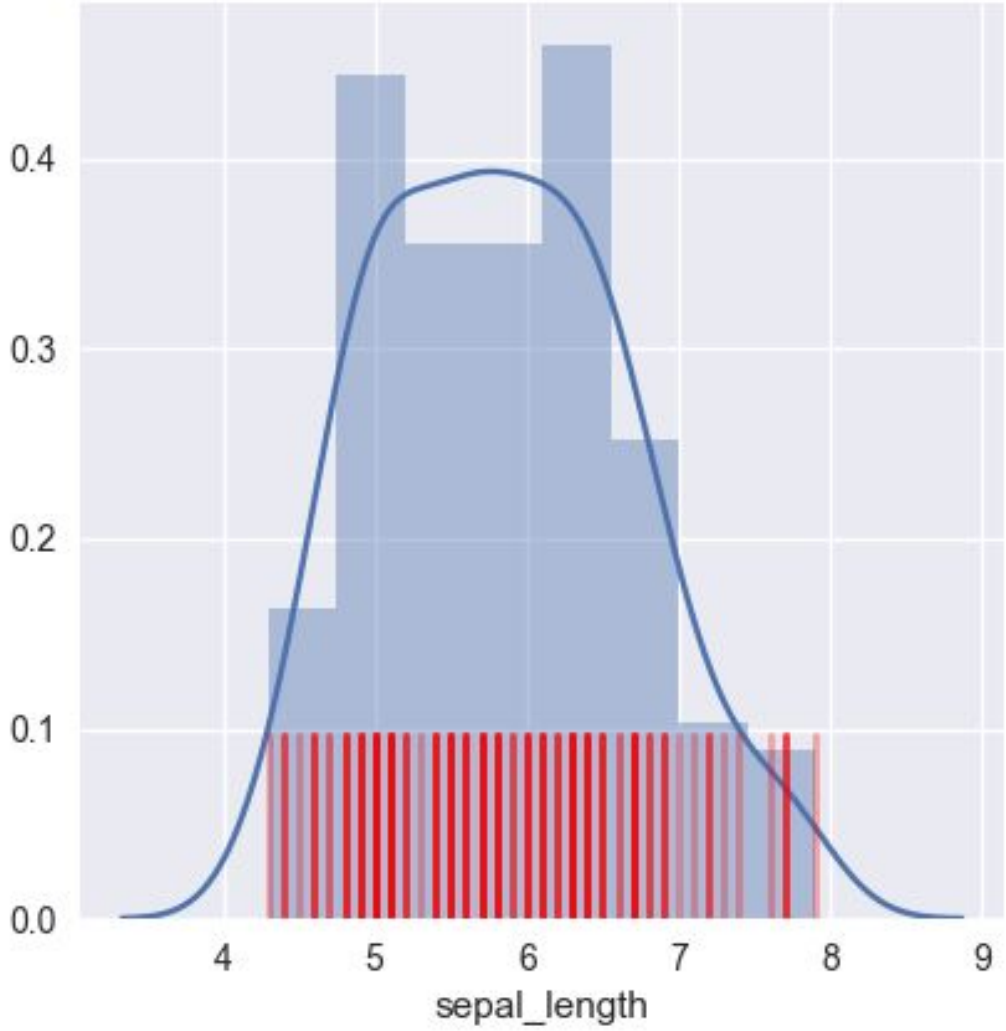


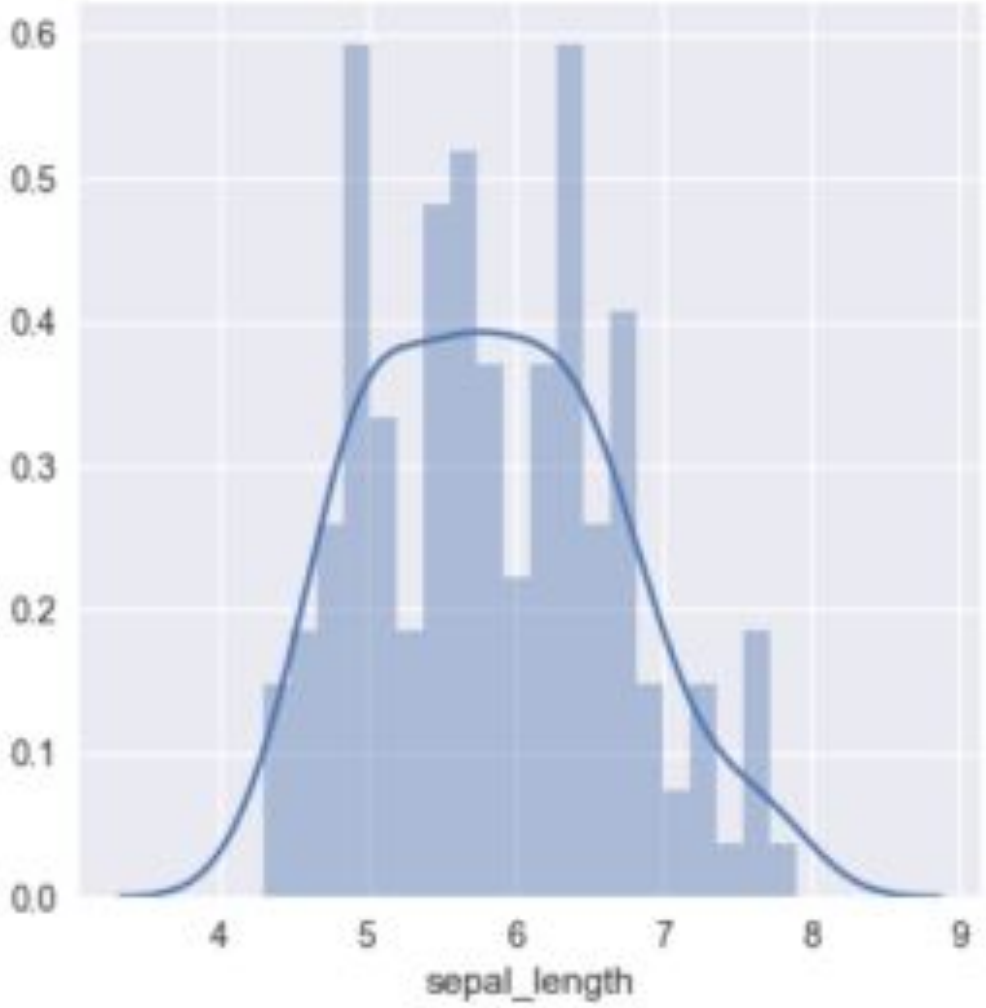
https://python-graph-gallery.com/73-control-bandwidth-of-seaborn-density-plot/

Histogram

- https://python-graph-gallery.com/histogram/
- For showing the shape of a numerical variable
- Similar to Density plot
- Histograms visualize the **actual** shape of the variable and not the underlying distribution
- Number of bins is an important parameter
- Outliers can really decrease the effectiveness of this visualization

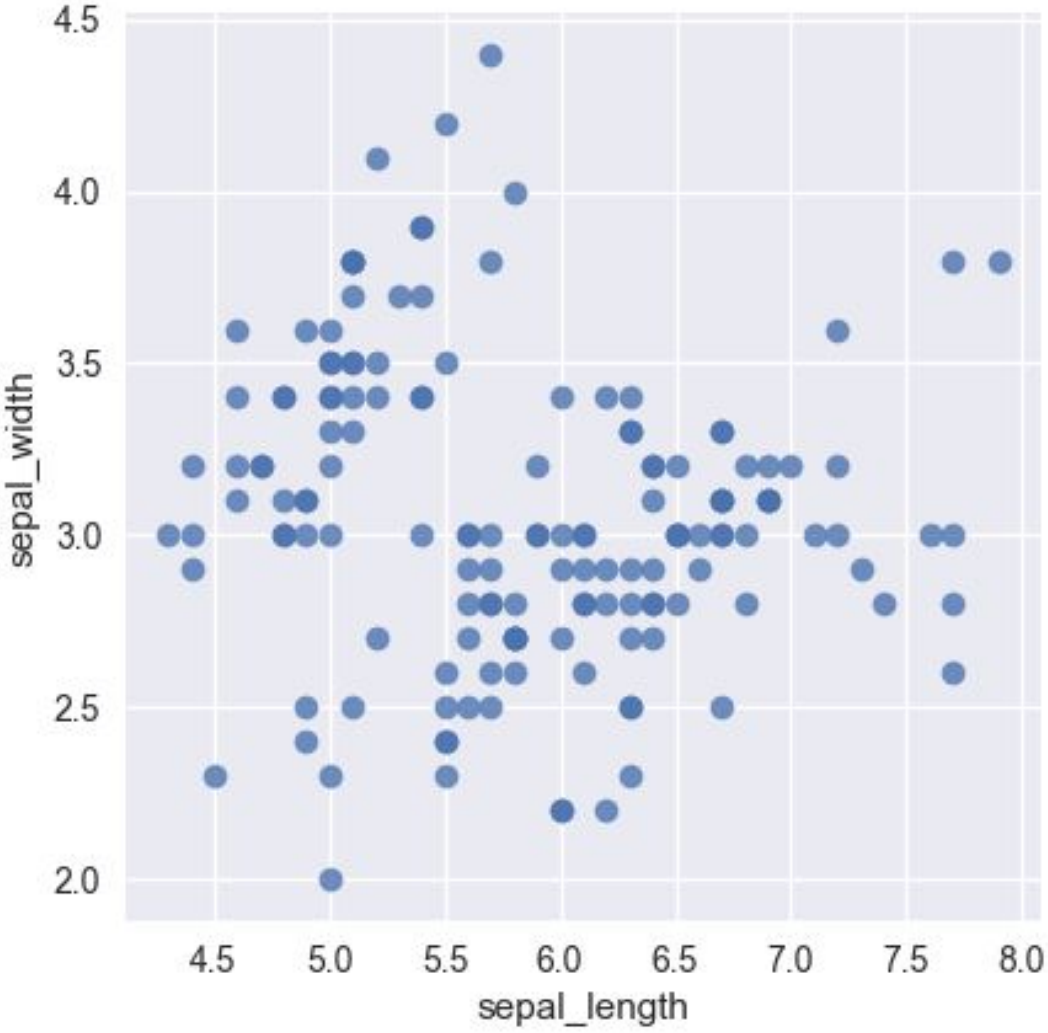

30
25
20
15
10
5
0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sepal_length

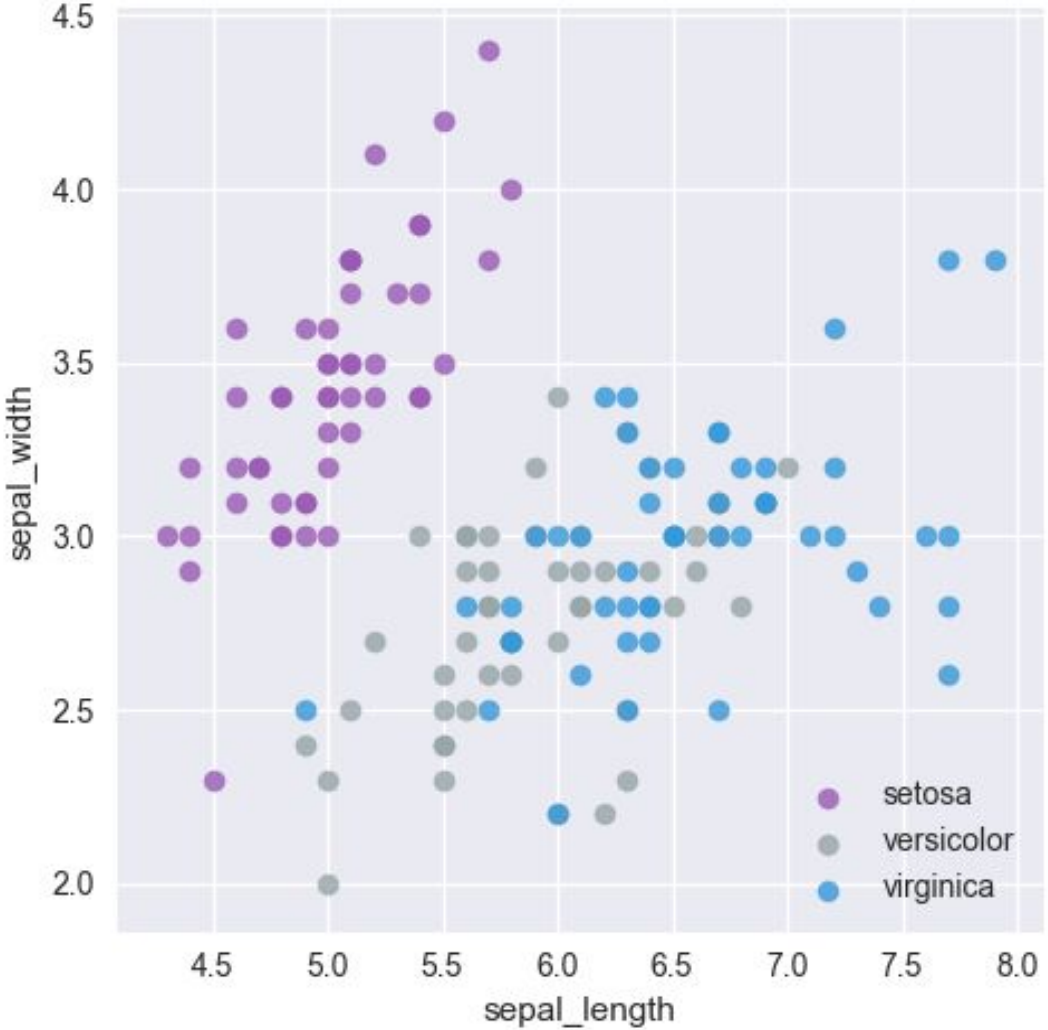

0.4
0.3
0.2
0.1
0.0
4
5
6
7
8
9
sepal_length

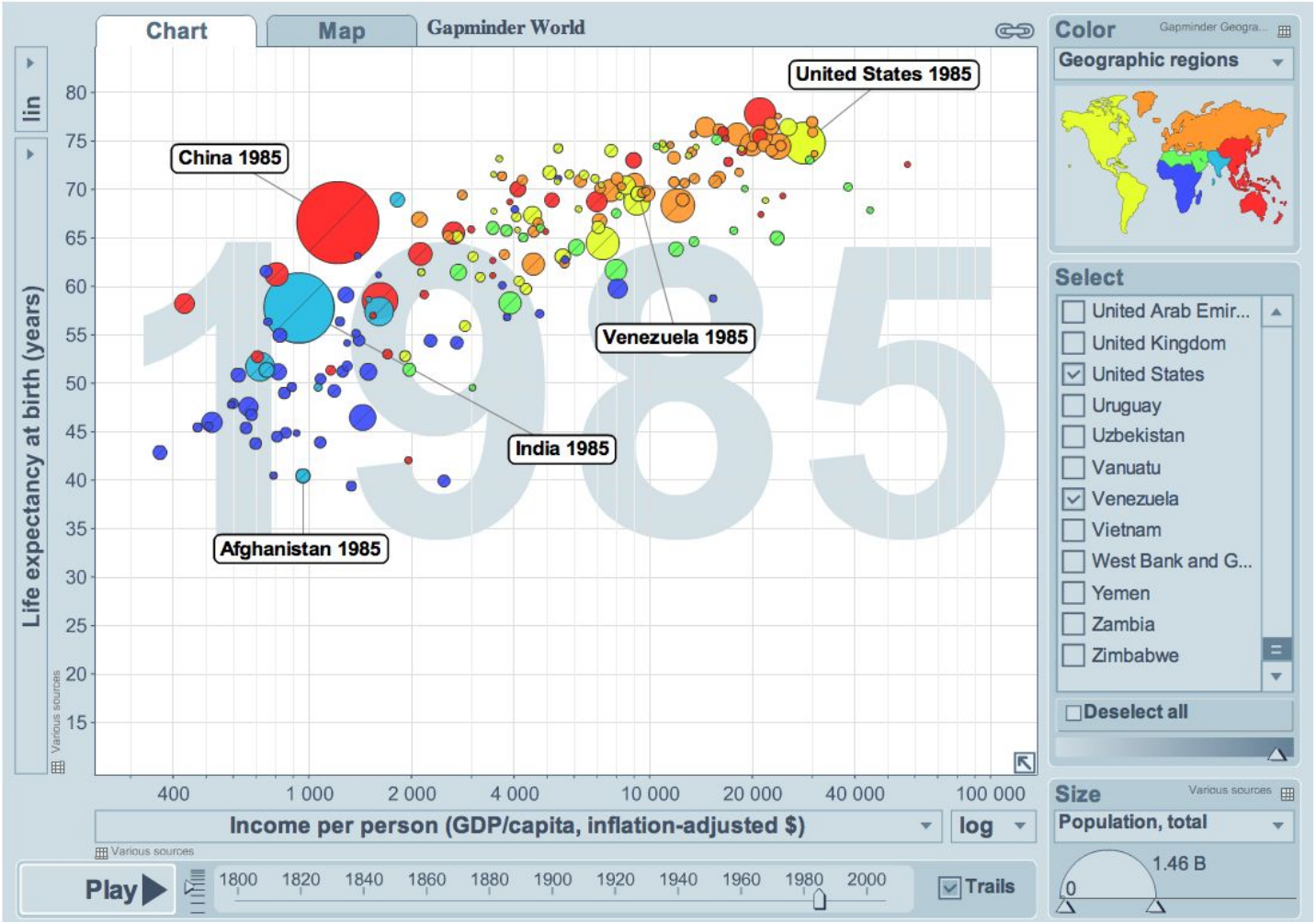

0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
5
6
7
8
9
sepal_length

Scatterplot

- https://python-graph-gallery.com/scatter-plot/
- Visualize the **relationship** between two variables
- Usually two numerical variables
- A third categorical variable can be visualized by adding colour
  - Or numerical with color scale
- An additional numerical variable can be represented by the points **size** (called a bubble plot)
  - It is also possible to add fifth by using animation over time
- Great for visualizing a **clustering** that can be projected onto two or three dimensions


4.5
4.0
3.5
3.0
2.5
2.0
sepal_width
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sepal_length

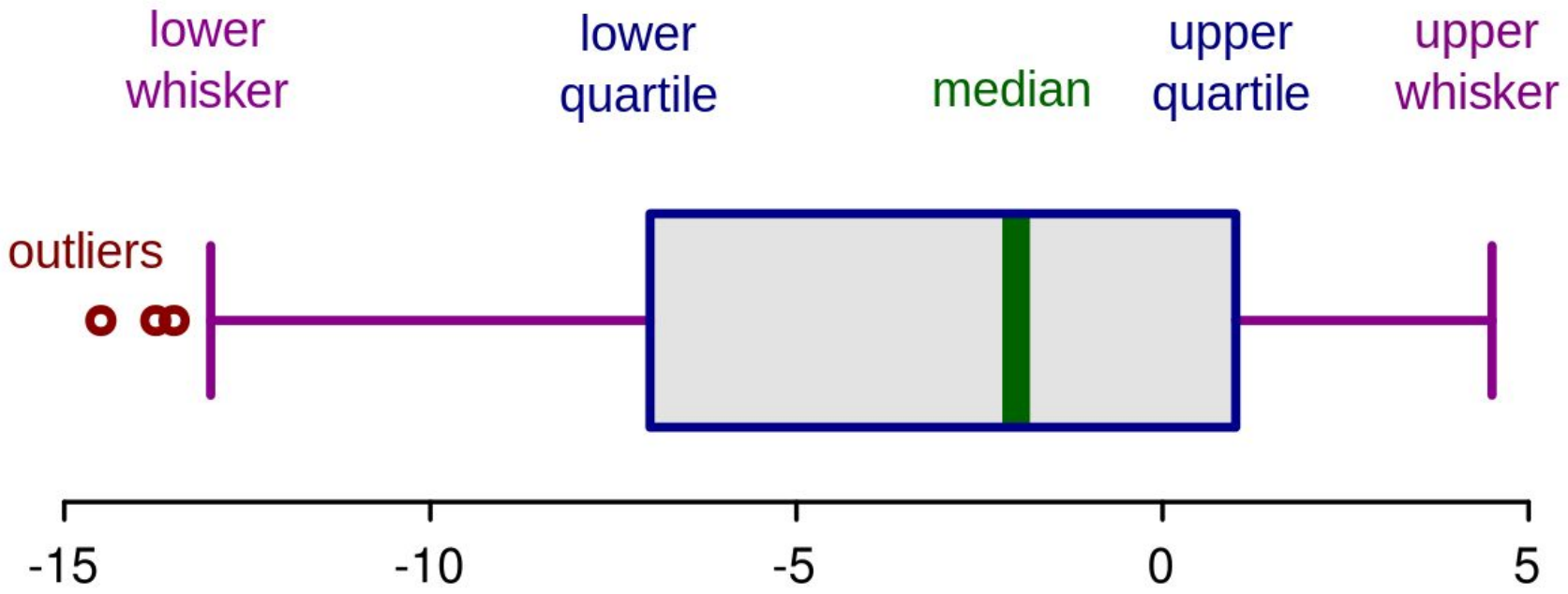

4.5
4.0
3.5
3.0
2.5
2.0
sepal_width
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sepal_length
setosa
versicolor
virginica



https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

Box Plots

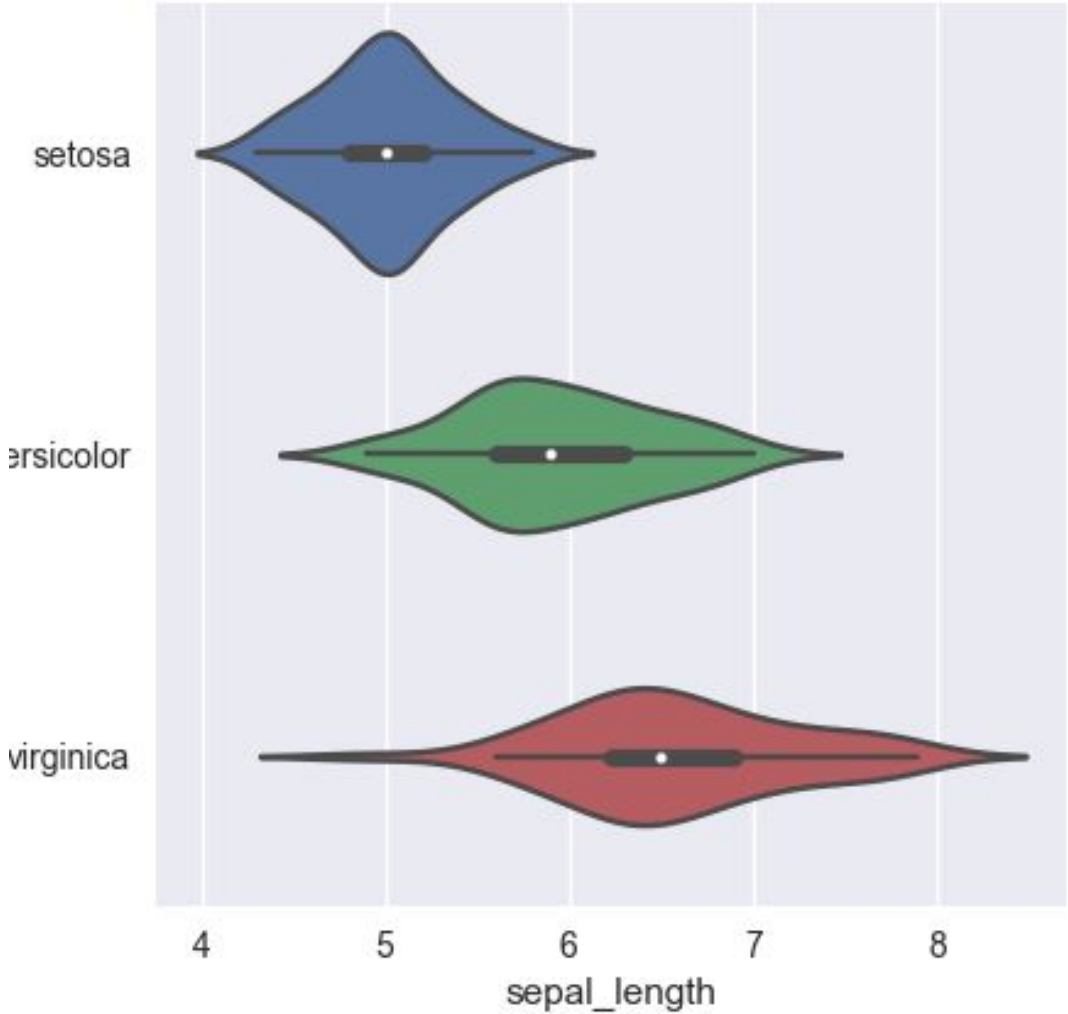
- http://python-graph-gallery.com/boxplot/
- Summarizes a numerical variable
- Can compare several numerical variables or one numerical variables split by a categorical variable
- Good for visualizing obvious outliers
- Hides the underlying distribution

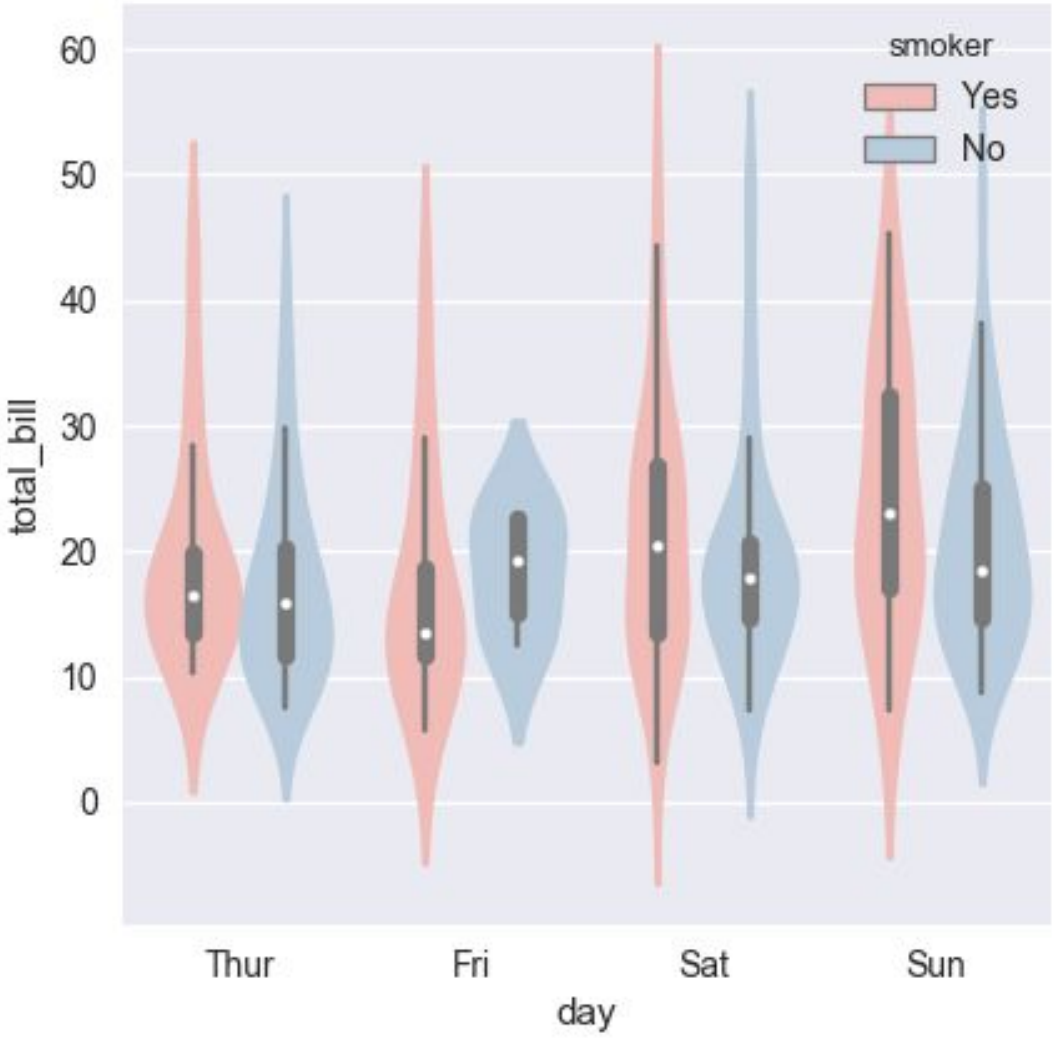


https://upload.wikimedia.org/wikipedia/commons/d/d1/Elements_of_a_boxplot_en.svg

Violin Plot

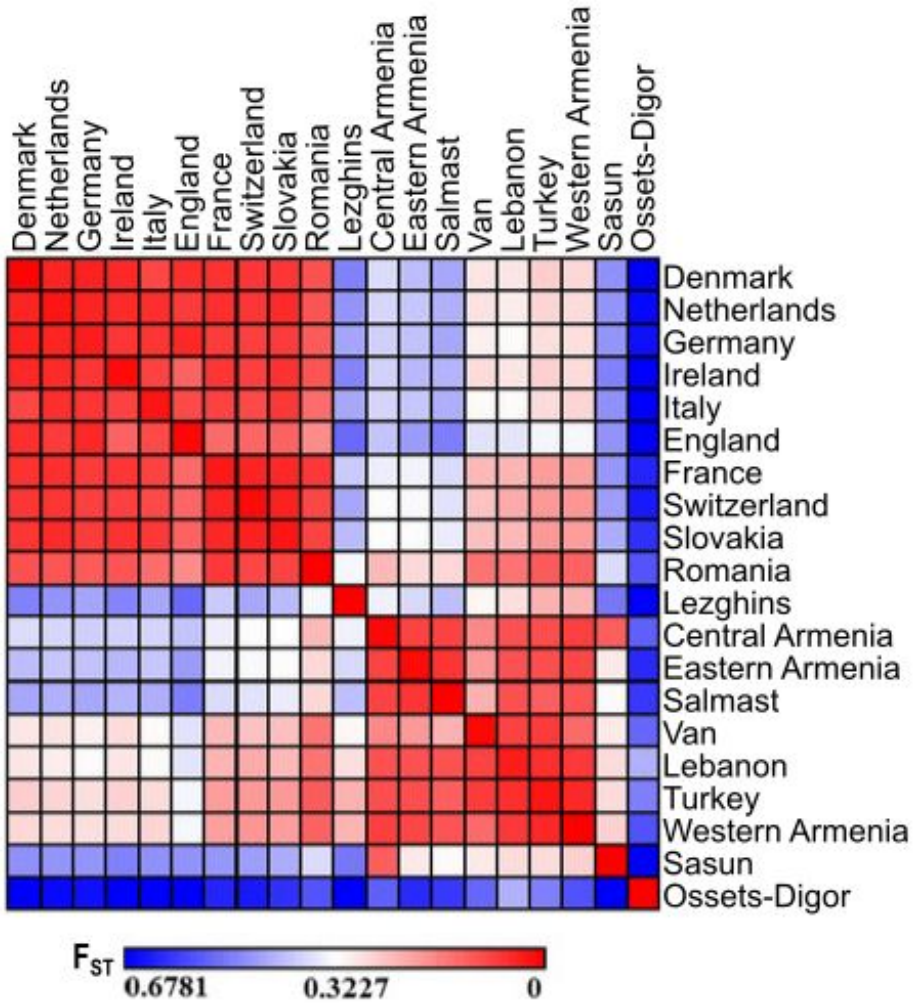
- https://python-graph-gallery.com/violin-plot/
- Similar to boxplot but it shows the underlying distribution
- Can display a single numerical variables, several categorical variables, or a single numerical variable split by group
- Good for comparing numerical variables while maintaining the underlying distribution
- Can be used in place of a density plot if variables overlap each other a lot


setosa
ersicolor
virginica
4
5
6
7
8
sepal_length

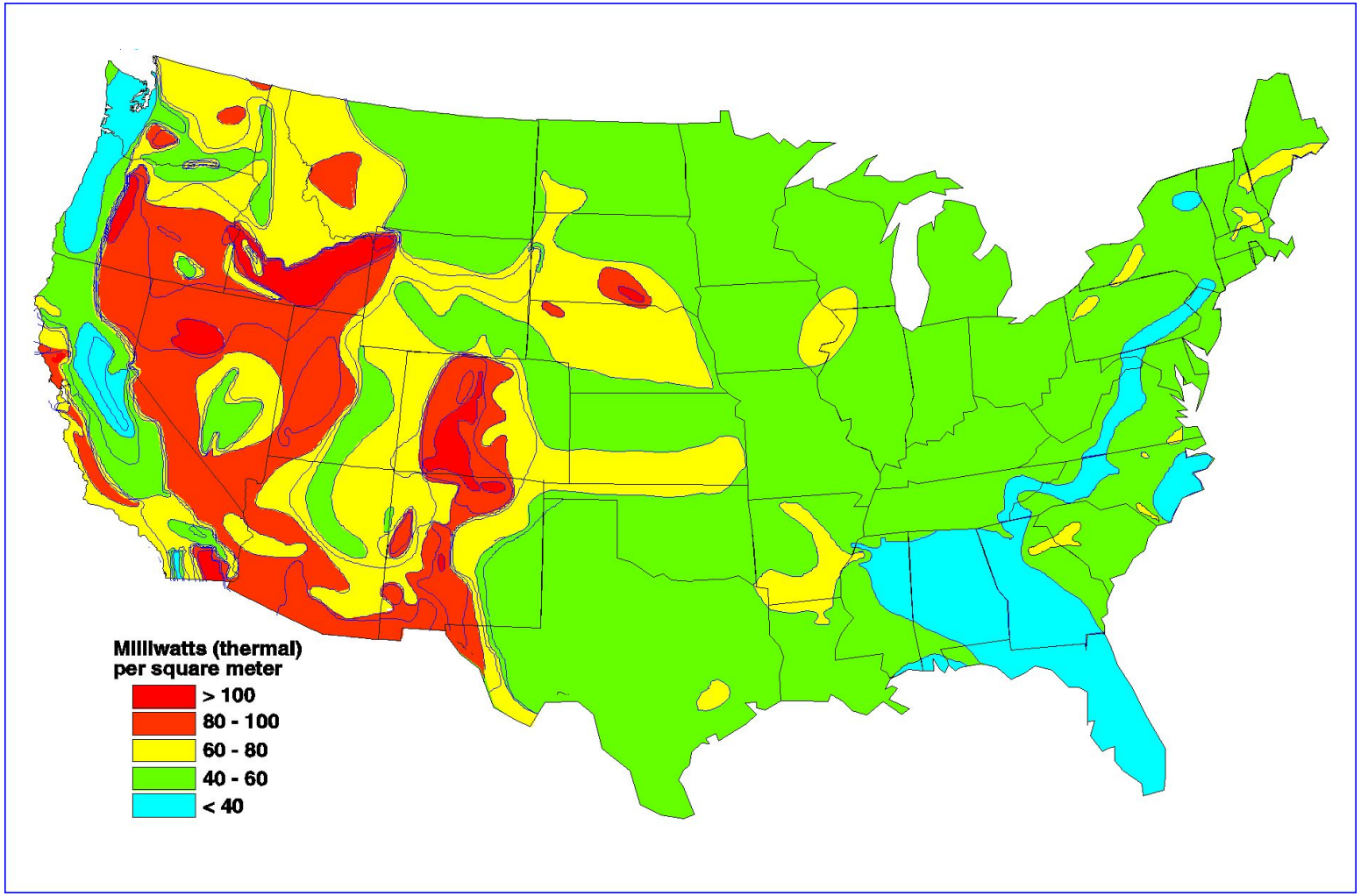

smoker
Yes
No
60
50
40
30
20
10
0
total_bill
Thur
Fri
Sat
Sun
day

Heatmap

- https://python-graph-gallery.com/heatmap/
- Graphical representation of a matrix
- Cells are real numbers
- Number values are mapped to a colour gradient
- Generalized - can map numbers to a gradient in any shape (e.g. a literal map)



https://upload.wikimedia.org/wikipedia/commons/e/ec/Heatmap_of_pairwise_Fst_genetic_distances_calculated_for_the_Y-DNA_haplogroup_R1b1a2.png


Milliwatts (thermal)
per square meter
> 100
80 - 100
60 - 80
40 - 60
< 40

Graphing Best Practices

Best practices

- Simplify
- Avoid 3D, animations etc.
- Use a colour pallette
  - Consider color-blind people
- Think about the point you want to make
- Label your axes
- Start values at 0 (if it makes sense)
- https://www.darkhorseanalytics.com/blog/data-looks-better-naked
- http://blog.visme.co/data-storytelling-tips/
- https://gramener.github.io/visual-vocabulary-vega/