CS410: Principles and Techniques of Data Science

Module 8: Data Visualization

https://drive.google.com/drive/folders/1yvBwrnug2J1INAxv4cfS42wCzVBYLZCi?usp=sharing

Introduction

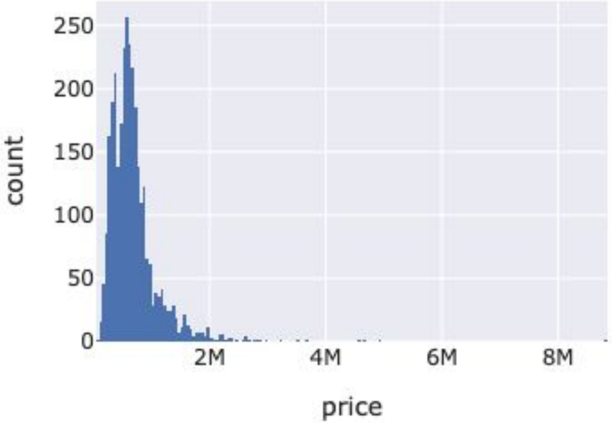
Data scientists create data visualizations in order to understand our data and explain our analyses to other people. Every plot has a message. And it’s our job to use plots to communicate this message as clearly as possible.

Software packages for visualization change all the time, so any code we use can quickly get out-of-date.

Choosing Scale to Reveal Structure

Below is the San Francisco house prices example and take a look at the following histogram of prices.

```
px.histogram(sfh, x='price', width=350, height=250)
```



The plot displays all the data, but most of the data are crammed into the left side of the plot, which makes it hard to understand house prices.

We can’t clearly see important features about the data like whether there are multiple modes or skewness.

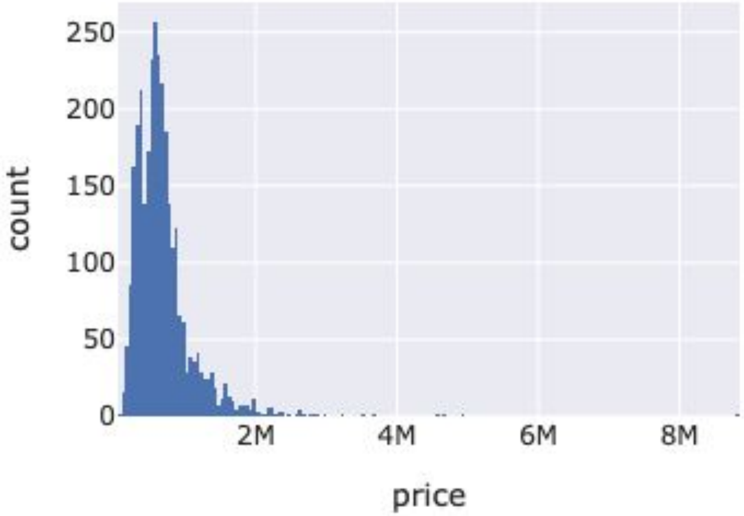
Solution?

Filling the Data Region

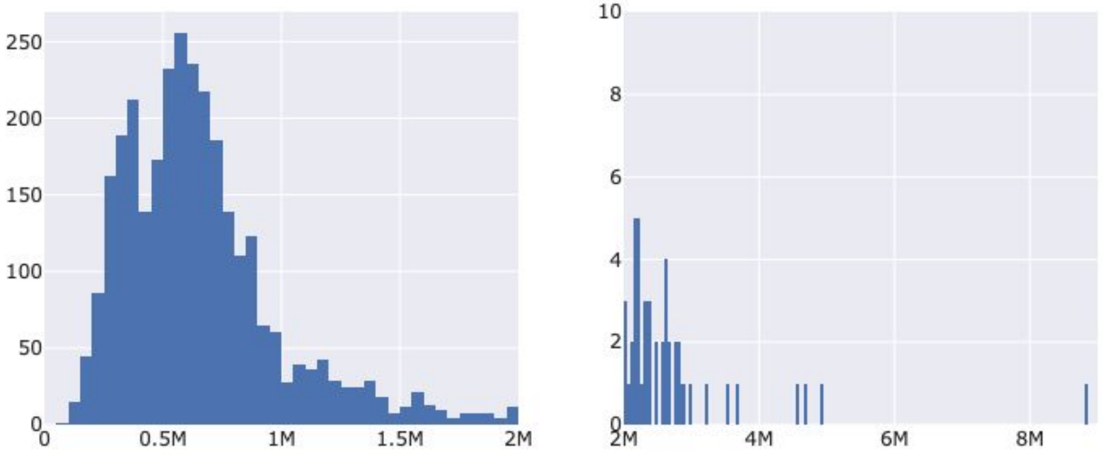
This issue can happen when there are a few unusually large observations.

In order to get a better view of the main portion of the data we can drop these observations from the plot by adjusting the x- or y-axis limits, or by removing outlier values from the data before plotting.

In either case, we must mention this exclusion in the caption or on the plot itself.



Filling the Data Region



```
fig.update_xaxes(range=[0, 2e6], row=1, col=1)

fig.update_xaxes(range=[2e6, 9e6], row=1, col=2)

fig.update_yaxes(range=[0, 10], row=1, col=2)
```

We addressed the issue by making two plots, one for the bulk of the data and one for the tail.

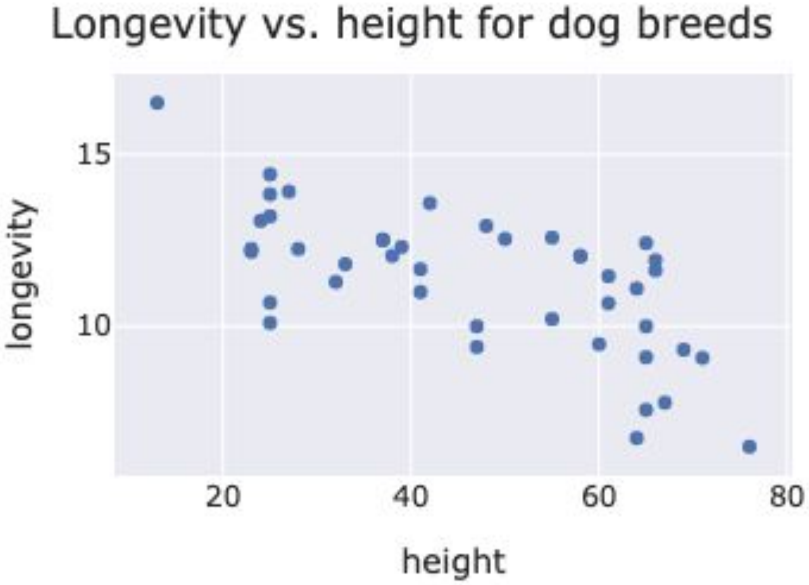
Notice that the x-axis in the right plot includes 0, but the left plot begins its x-axis at $2,000,000.

Including Zero

We often don’t need to include 0 on an axis, especially if including it makes it difficult to fill the data region.

Scatter plot shows the average longevity plotted against average height for dog breeds.

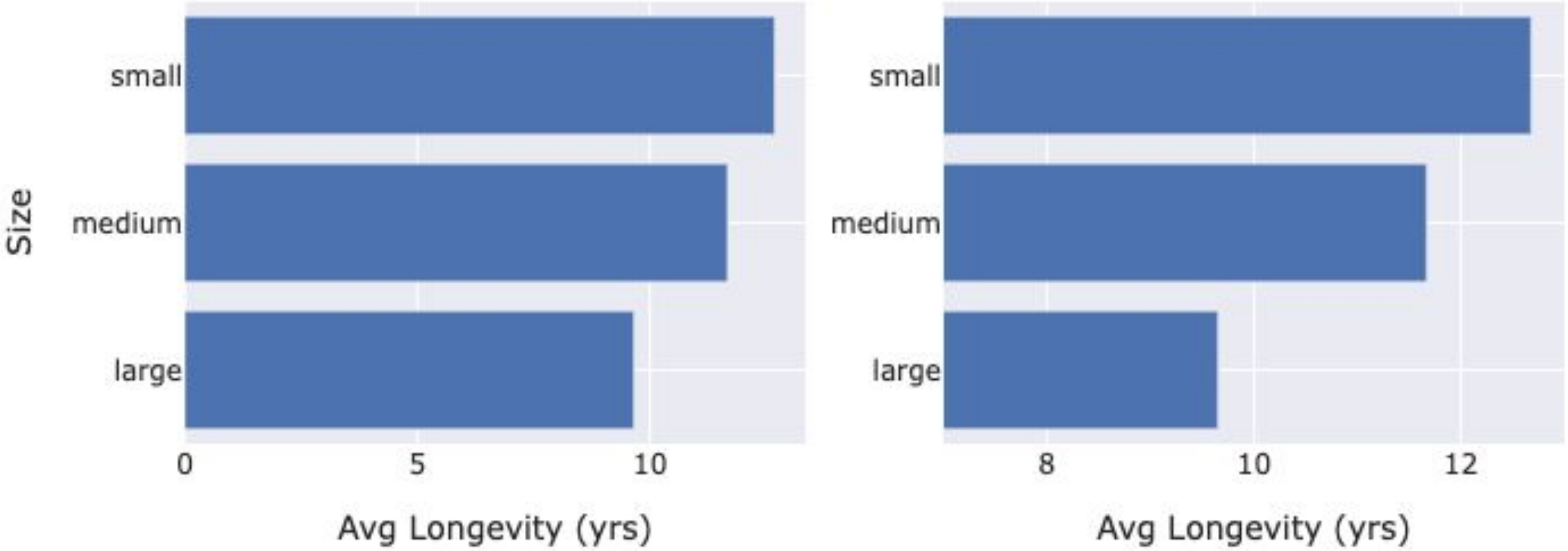
The x-axis of the plot starts at 10 cm since all dogs are at least that tall, and, similarly, the y-axis begins at 5 years.



Can you think of any cases where 0 has to be included?

Including Zero

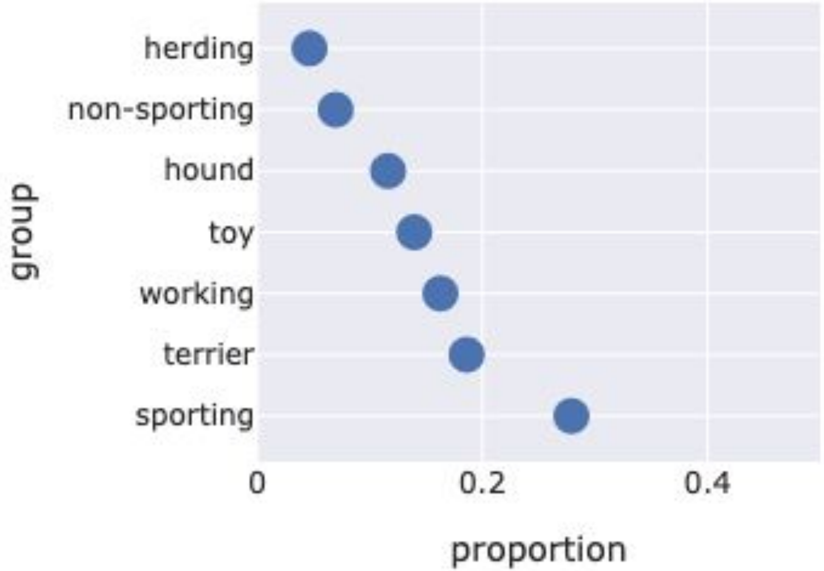
For bar charts, including 0 is important so the heights of the bars directly relate to the data values.



The left plot includes 0, but the right plot doesn't. It's easy to incorrectly conclude from the right plot that medium-sized dogs live twice as long as large-sized dogs.

Including Zero

We also typically want to include zero when working with proportions, since proportions range from 0 to 1.

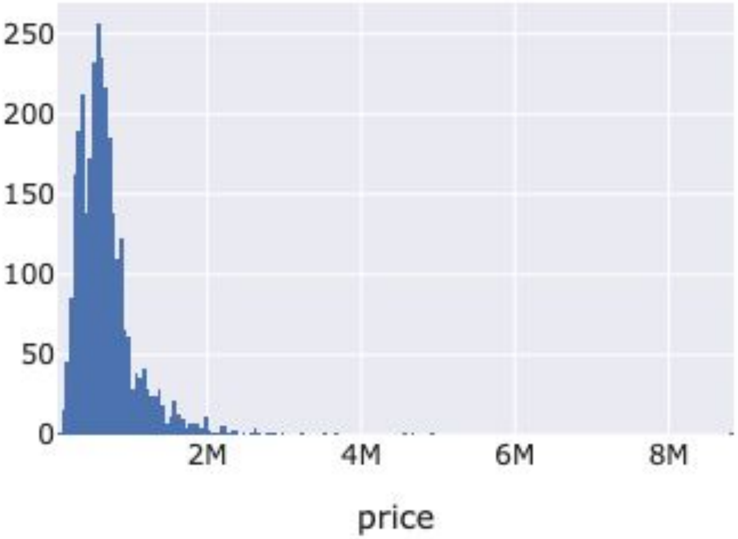


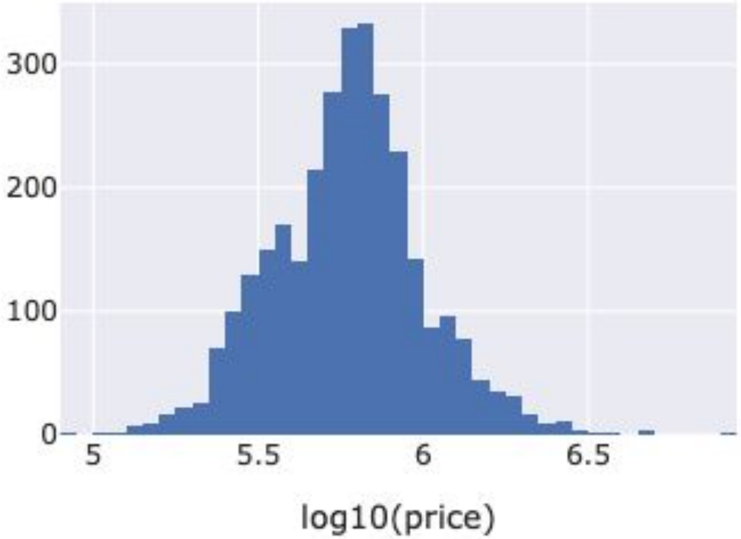
Proportion of dogs of each dog group in our dataset.

By including 0, it is easier to accurately compare the relative size of groups.

Revealing Shape Through Transformations

Another common way to adjust scale is to transform the data or the plot axes. And, when the transformation produces a symmetric distribution, the symmetry carries with it useful properties in later modeling steps.



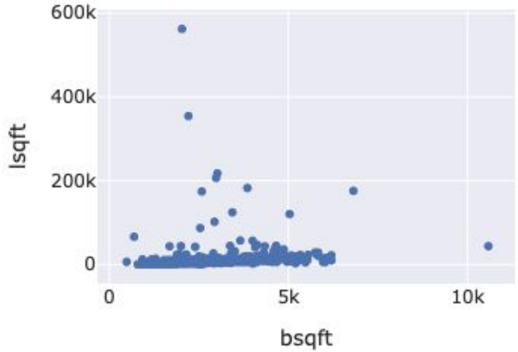


Log-transform is especially useful. The left histogram is the original SF house prices data.

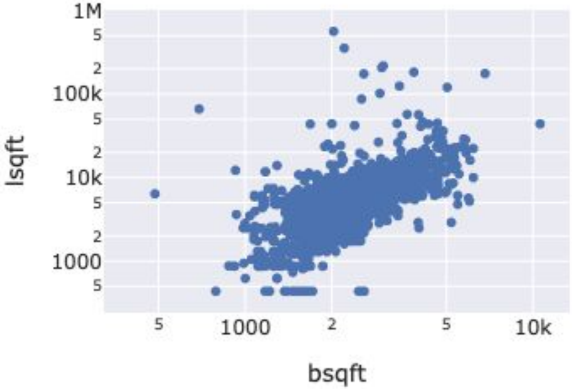
On the right, we’ve taken the log (base 10) of the prices before plotting.

Revealing Shape Through Transformations

The log transform can also reveal shape in scatter plots. The building size on the x-axis and the lot size on the y-axis. It’s hard to see the shape in this plot since many of the points are crammed.



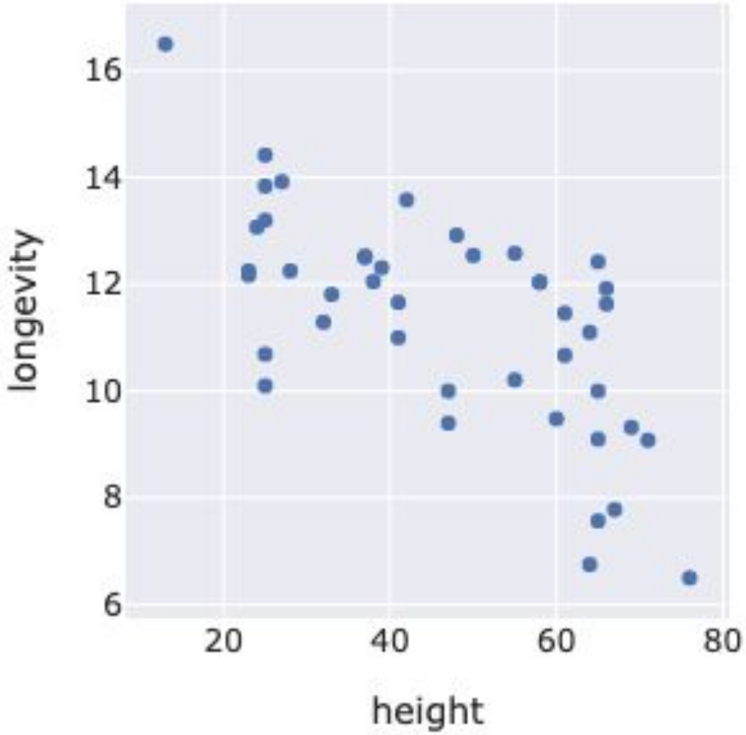
Using a log scale for both x- and y-axes, the shape is much easier to see. we can see that the lot size increases roughly linearly with building size (on the log scale)



Banking to Decipher Relationships

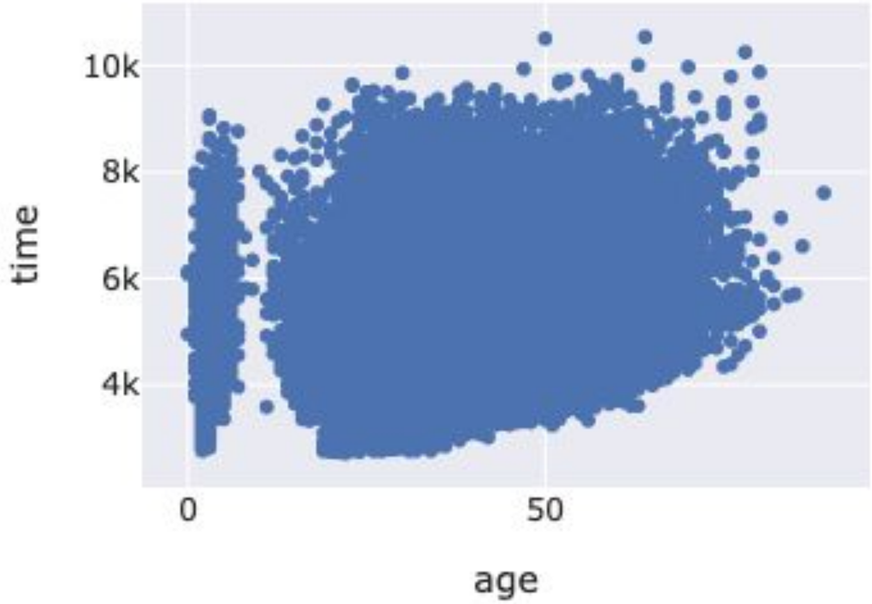
We can change the aspect ratio of the plot, by adjusting the length and width of the rectangular plot. That is, we can stretch or shrink a plot without changing its axes limits. This adjustment is called “banking”.

```
px.scatter(dogs, x='height', y='longevity',
                 width=300, height=300)
```



Smoothing and Aggregating Data

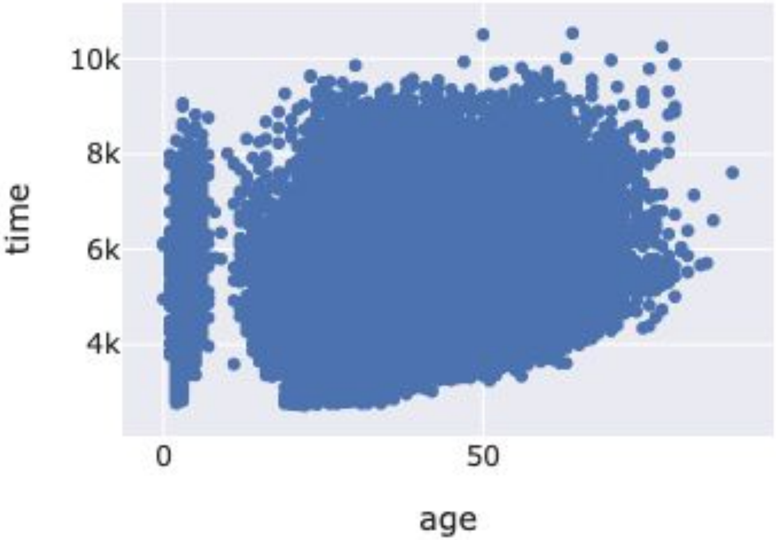
Overplotting



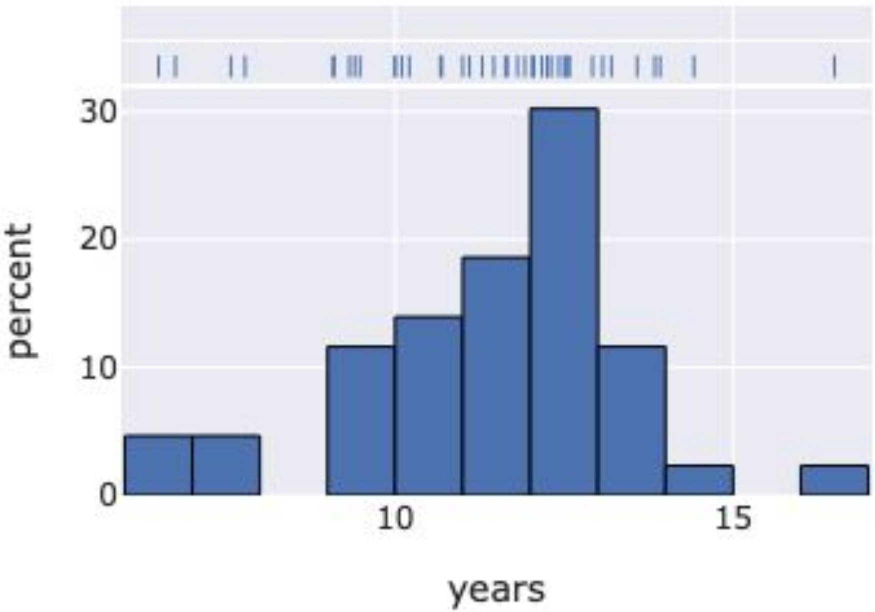
How to avoid overplotting?

Smoothing and Aggregating Data

Overplotting



Histogram



When not to smooth?

Smoothing and aggregating can help us see important features and relationships, but when we have only a handful of observations, smoothing techniques can give misleading representations of the data.

With just a few observations, we prefer rug plots over histograms, and we use scatter plots rather than smooth curves.

Facilitating Meaningful Comparisons

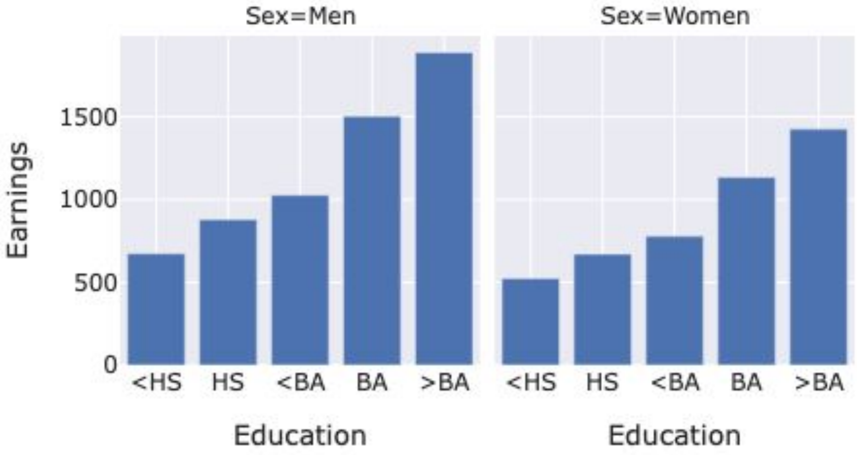
The same data can be visualized many different ways.

How can we decide between two plots of the same data?

Emphasize the Important Difference

Whenever we make a plot that compares groups, we should ask: does the plot emphasize the important difference?

The US Bureau of Labor Statistics publishes data on income. The 2020 median full-time-equivalent weekly earnings for people over 25 and given below. We've split people into groups by education level and sex.



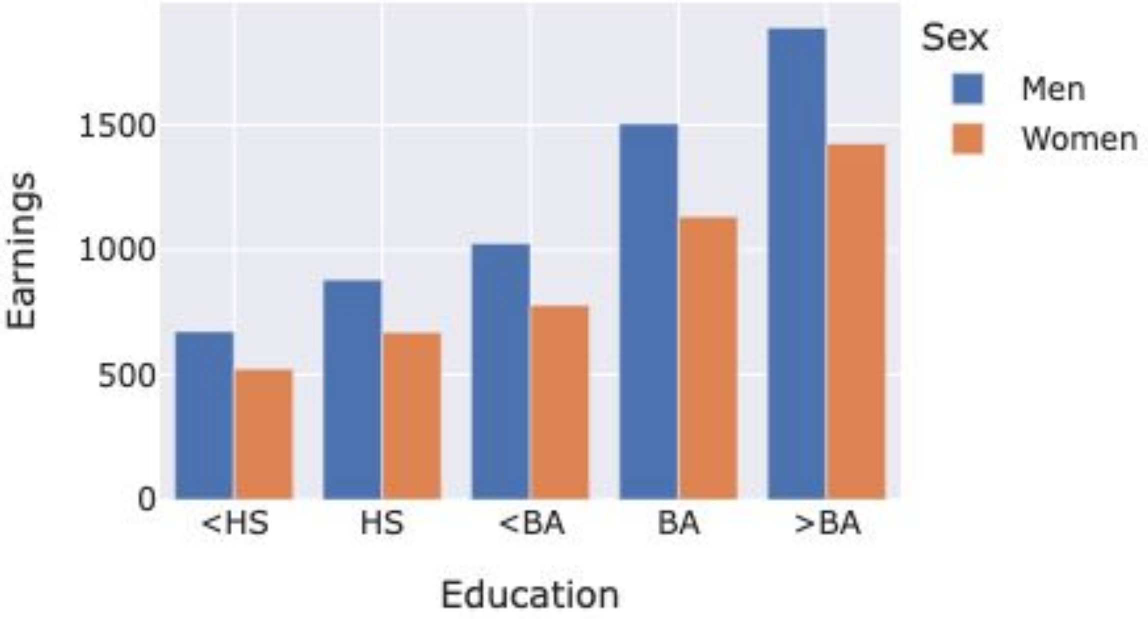
These bar plots show that earnings increase with more education.

Any other plots for the same data?

Emphasize the Important Difference

A more interesting comparison is between men and women of the same education level.

```
px.bar(earn, x='educ', y='income', color='gender',
             barmode='group',
             labels=labels,
             width=450, height=250)
```



This plot is much better, we can more easily compare the earnings of men and women for each level of education.

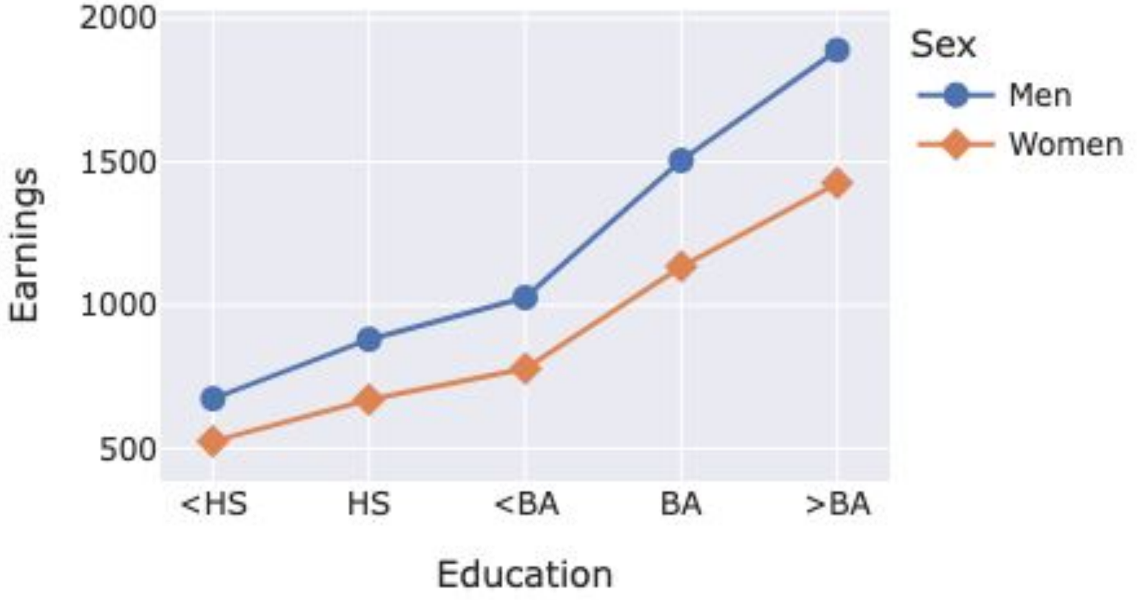
Can we create a better plot?

Emphasize the Important Difference

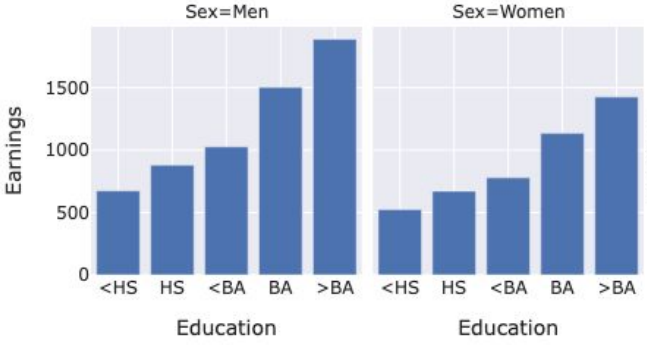
We can make this difference even more clear using vertical alignment. Instead of bars, we use dots so that we can align the dots vertically for each education level.

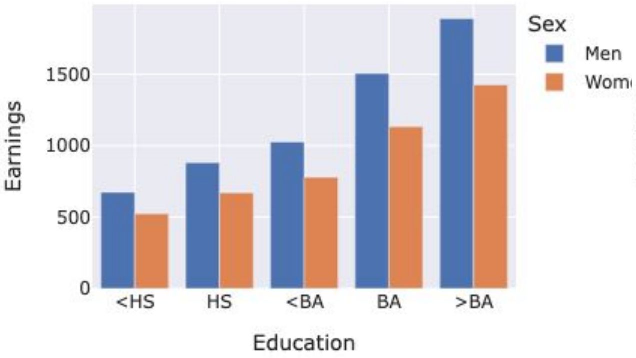
```
px.line(earn, x='educ', y='income', color='gender', symbol='gender',
              labels=labels,
              markers=True,
              width=450, height=250)
```

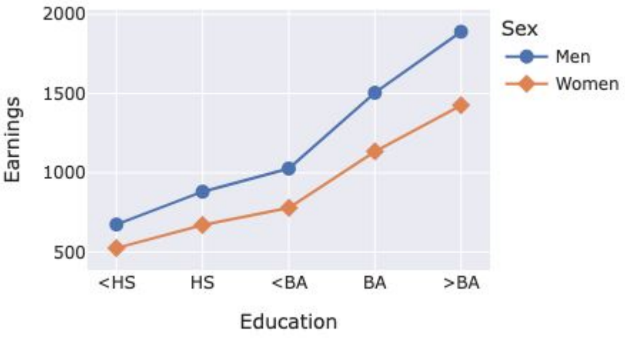
It’s visually clear that in 2020, the median weekly earnings for men was higher than the earnings for women at each education level, and that this gap grows at higher education levels.



Emphasize the Important Difference







Three plots that all plot the same data, but they are not the same in how readily you can see the message in the plot. We prefer the last one because it aligns the income differences vertically, making them easier to compare.

We ordered the education categories from the least to greatest number of years of education. This ordering makes sense because education level is ordinal.

What happens when we are comparing against a categorical variable that does not have a natural ordering?

Ordering Groups

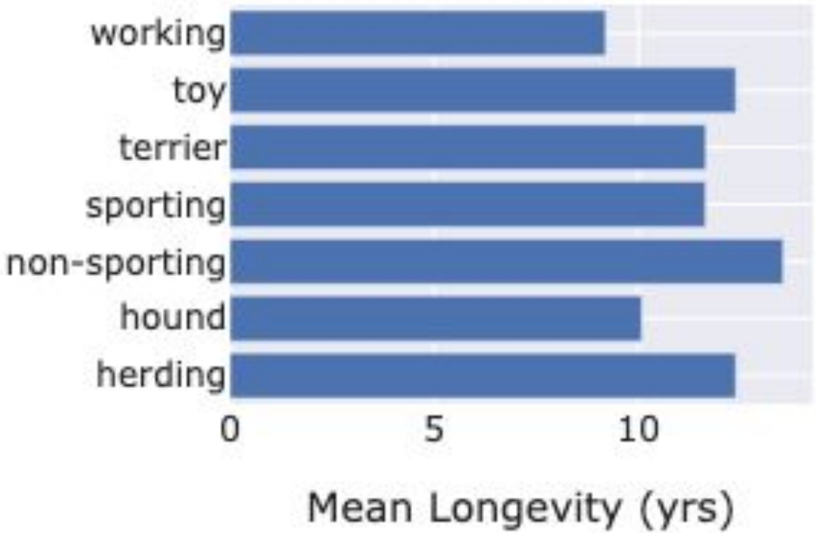
For nominal features we adjust the ordering to facilitate comparisons, but we leave ordinal data in their natural order.

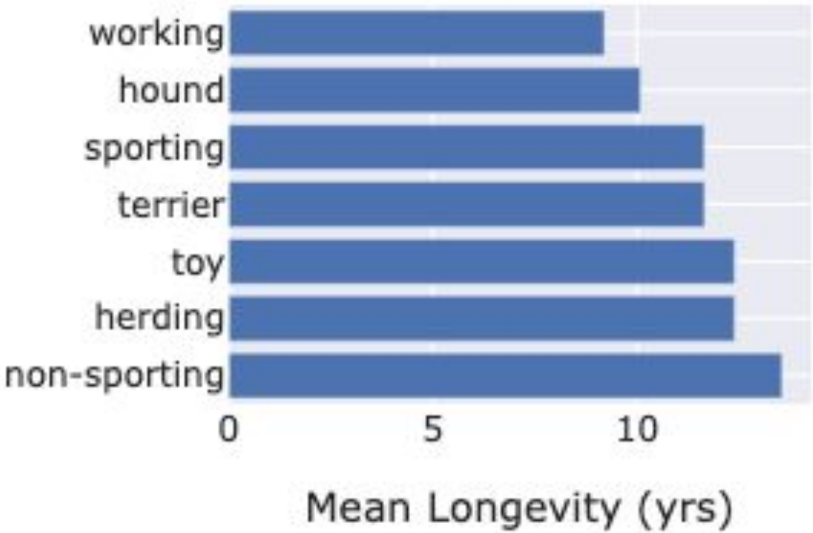
In general, for bar plots, it’s good practice to order the bars according to their height.

For box plots, we typically order them according to the medians.

Ordering Groups

The two bar plots below each compare the mean lifespan for groups of dog breeds. We prefer the plot on the right since it has ordered bars, which make it easier to compare longevity across groups.

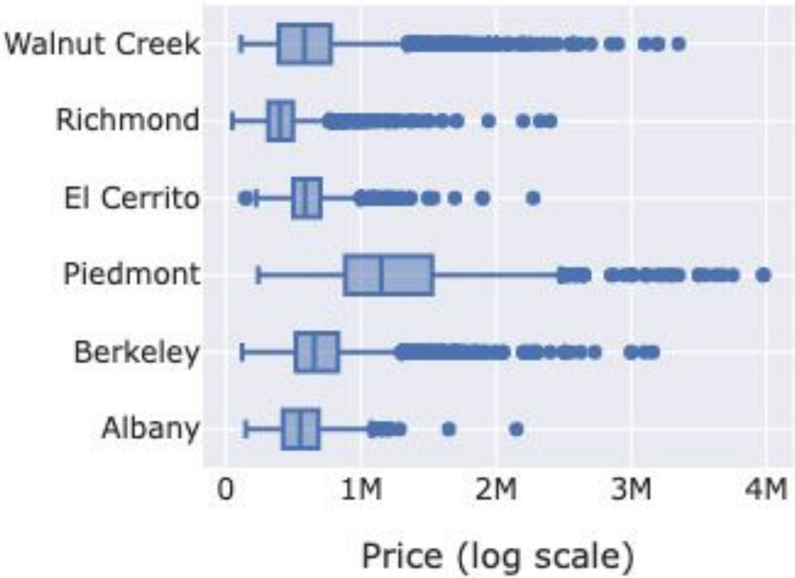


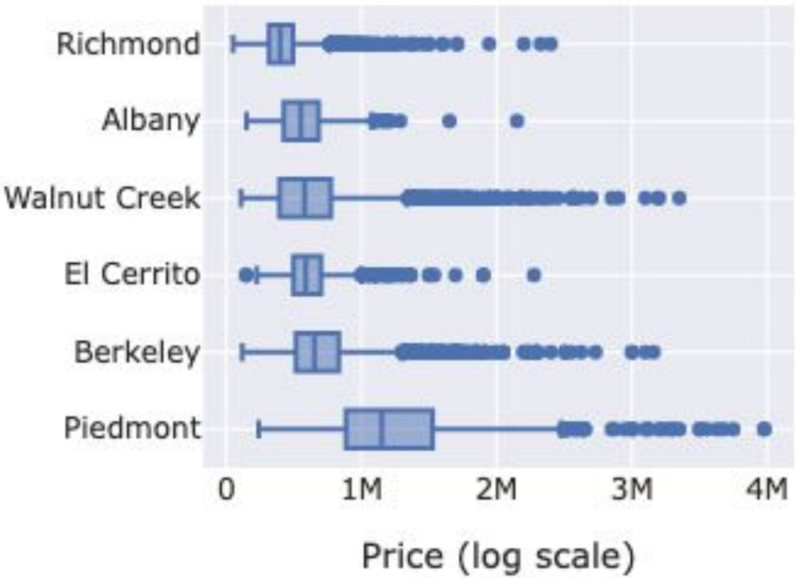


Ordering Groups

Two sets of box plots below each compare the distribution of price for houses in different cities in the San Francisco East Bay.

We prefer the plot on the right since it has ordered boxes, according to the median price for each city. Again, this ordering makes it The lower quartile and median price in Albany and Walnut Creek are roughly the same but the prices in Walnut Creek have greater right skew.





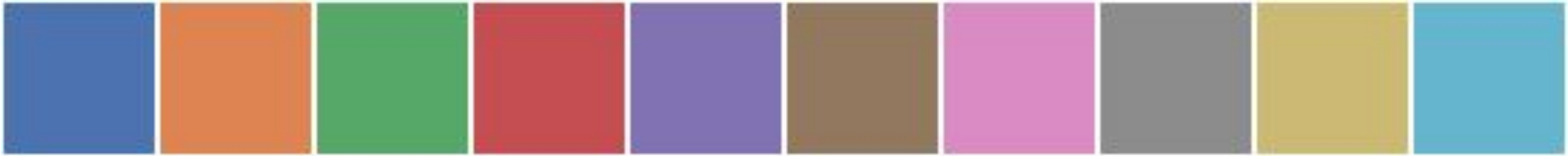
Selecting a color palette

Choosing colors also plays an important role in data visualization.

We want to avoid overly bright or dark colors so that we don’t strain readers’ eyes.

We should also avoid color palettes that might be difficult for color-blind people - 7-10% of people are red-green color-blind.

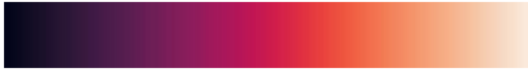
For categorical data, we want to use a color palette that can clearly distinguish between categories.



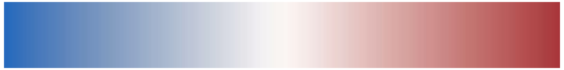
Selecting a color palette

For numeric data, we want to use:

Sequential color palette that emphasizes one side of the spectrum more than the other



Diverging color palette that equally emphasizes both ends of the spectrum and deemphasizes the middle.



We choose a sequential palette when we want to emphasize either low or high values, like cancer rates. We choose a diverging palette when we want to emphasize both extremes, like for two-party election results.

Selecting a color palette

It’s important to choose a perceptually uniform color palette. The term “perceptually uniform” means that when a data value is doubled, the color in the visualization looks twice as bright to the human eye.

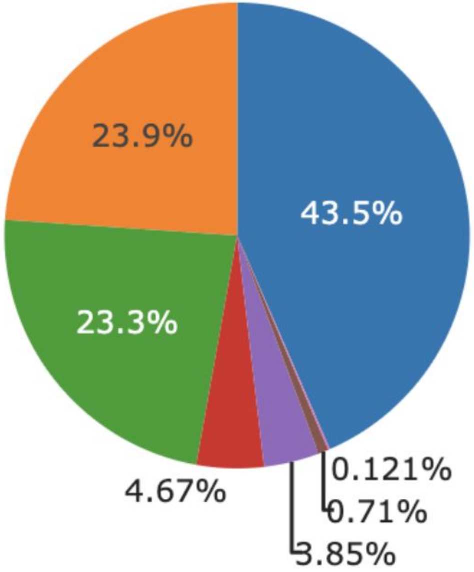
Guidelines for comparisons in plots

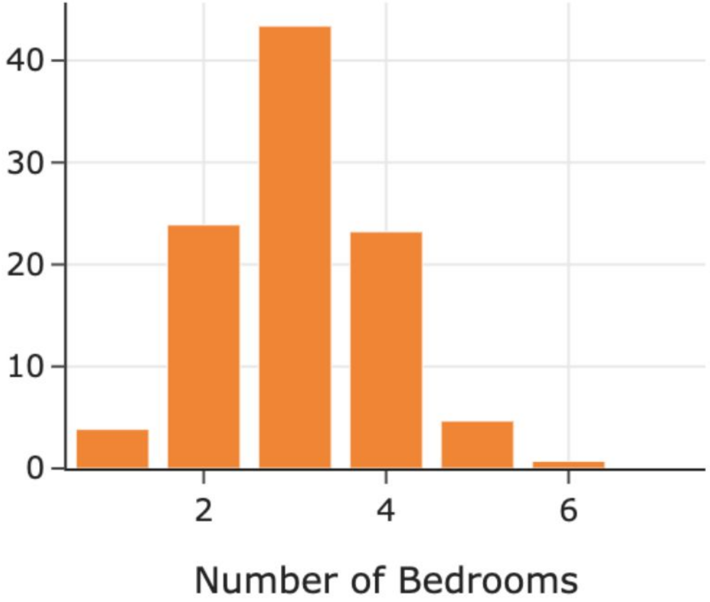
People understand visualizations better when the visualization uses positions and lengths rather than angles, areas, and volume and color.

Pie Charts vs Bar Charts

Pie chart showing the proportion of houses sold in San Francisco that are 1, 2, 3, …, 8+ bedrooms.

It’s hard to judge the angles in the pie, and the annotations with the actual percentages are needed. We also lose the natural ordering of the number of bedrooms. The bar chart doesn’t suffer from these issues.





Adding Context

- It’s important to add additional context to the plot. We can do this with reference markers, color, and labels.
- Aim to give enough context in our plots so that they can stand alone - a reader should be able to get the gist of each plot without needing additional explanation elsewhere.
- Every element of a statistical graph should have a purpose. Superfluous text or plot features, often referred to as “chartjunk”, should be eliminated.

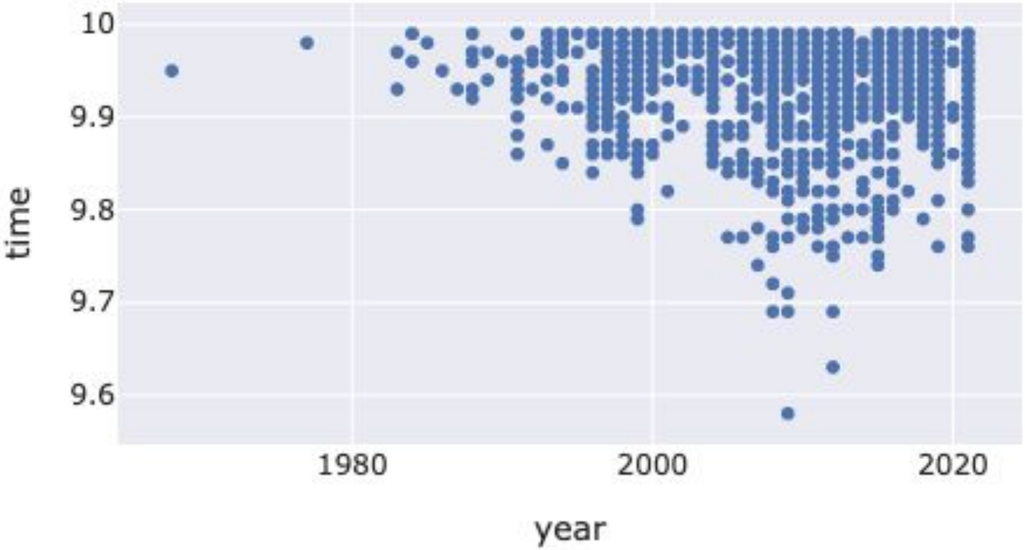
Adding Context

- *Reference Markers*: Reference points and lines can provide benchmarks, historical values, and other external information to compare our data against and help interpret the results. We might also add a vertical line on a time-series plot to mark a special event, like a natural disaster.
- *Labels:* It is good practice to consistently use informative labels on tick marks and axes. Axis labels often benefit from including units of measurement. Our graphs should contain titles and legends when needed.
- *Captions*: Captions point out the important features of the plot and their implications. It's okay for the caption to repeat information found in the text. Readers often skim a publication and focus on section headings and visualizations so plot captions should be self-contained.

Example: 100m Sprint Times

The race times in the men's 100-meter sprint since 1968. These data include only races that were electronically timed and held outdoors in normal wind conditions, and the times included are only for those runners' that came in under 10 seconds. The plot is a basic scatter plot showing race time against year.

Beginning with this plot, we augment it to create a plot featured in an article about the Olympic 100-meter sprint.



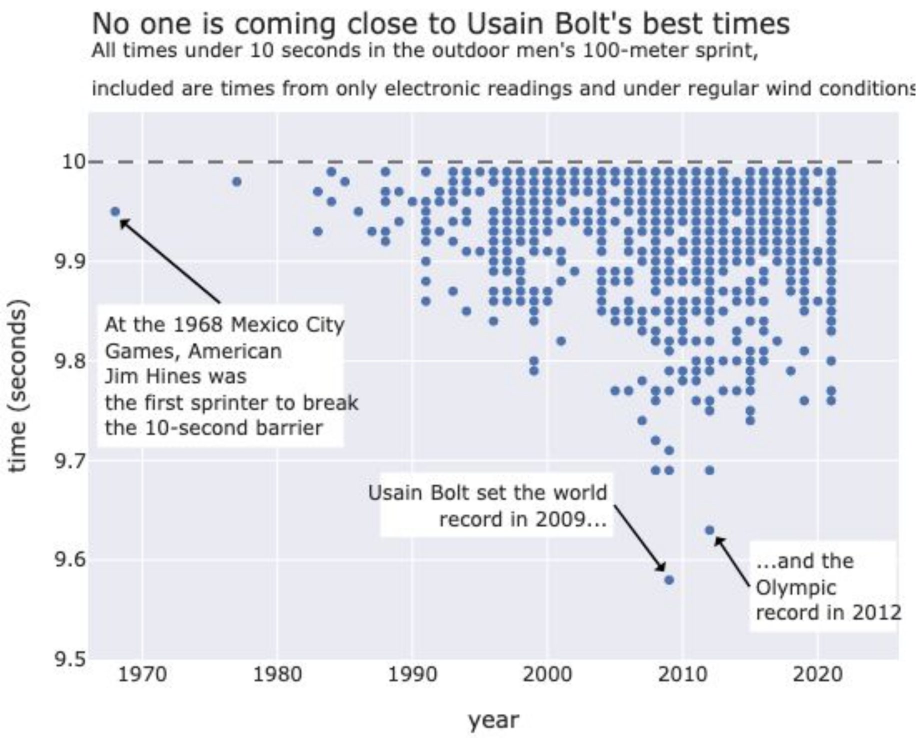
Example: 100m Sprint Times

When we want to prepare a plot for other people, it’s important to first think about the plot’s takeaway message. The main story is that Usain Bolt’s world record time of 9.58 seconds set in 2009 is still untouched. The second-best race time also belongs to Bolt.

We provide context to this plot by adding:

- a title that directly states the main takeaway,
- units of measurement in the y-axis label,
- annotations on key points in the scatter plot, including the two best race times that belong to Usain Bolt, and
- a caption describing the data.

In addition, we add a horizontal reference line at 10 seconds to clarify that only times below 10 seconds are plotted, and we use a special symbol for the world record time to draw the reader’s attention to this crucial point.



No one is coming close to Usain Bolt's best times

All times under 10 seconds in the outdoor men's 100-meter sprint,

included are times from only electronic readings and under regular wind conditions

At the 1968 Mexico City Games, American Jim Hines was the first sprinter to break the 10-second barrier

Usain Bolt set the world record in 2009...

...and the Olympic record in 2012

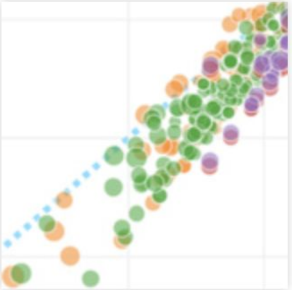
Creating plots using plotly

The plotly package has several advantages over other plotting libraries.

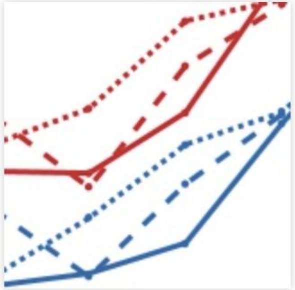
- It creates interactive plots rather than static images. When you create a plot in plotly, you can pan and zoom to see parts of the plot that are too small to see normally. You can also hover over plot elements, like the symbols in a scatter plot, to see the raw data values.
- Iit has a simple API for creating basic plots, which helps when you're doing exploratory analysis and want to quickly create many plots.

Plotly basic charts

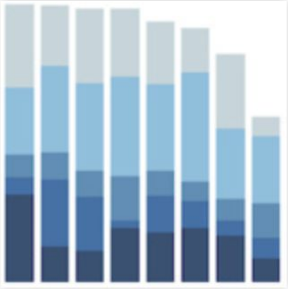
https://plotly.com/python/basic-charts/



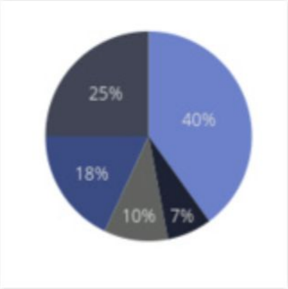
Scatter Plots



Line Charts



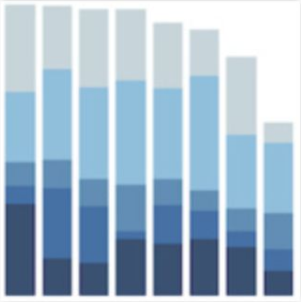
Bar Charts



Pie Charts

Eve
Seth
Enos
10
42% of parent
15% of entry
15% of root

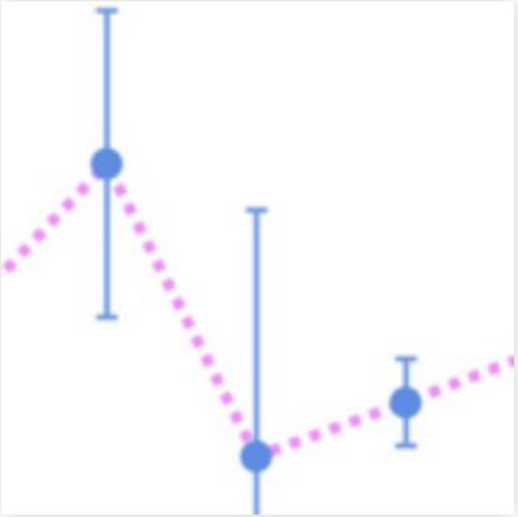
Treemap Charts



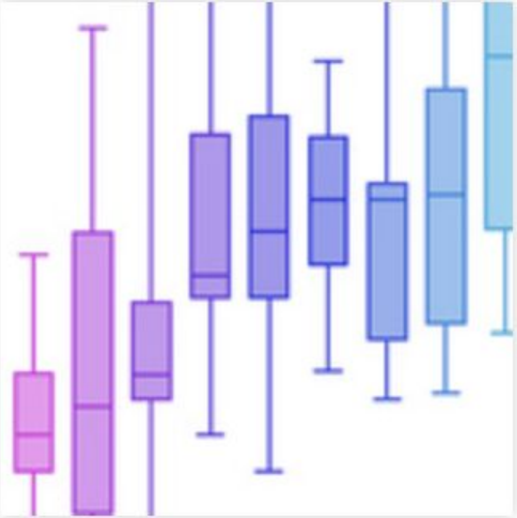
Categorical Axes

Plotly statistical charts

https://plotly.com/python/statistical-charts/



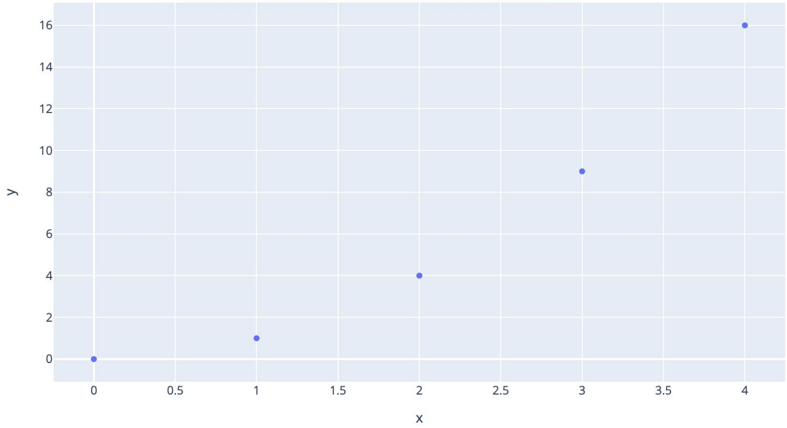
Error Bars

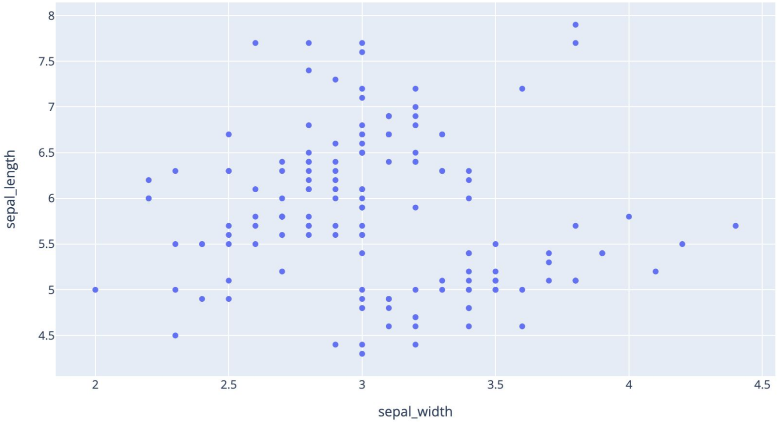


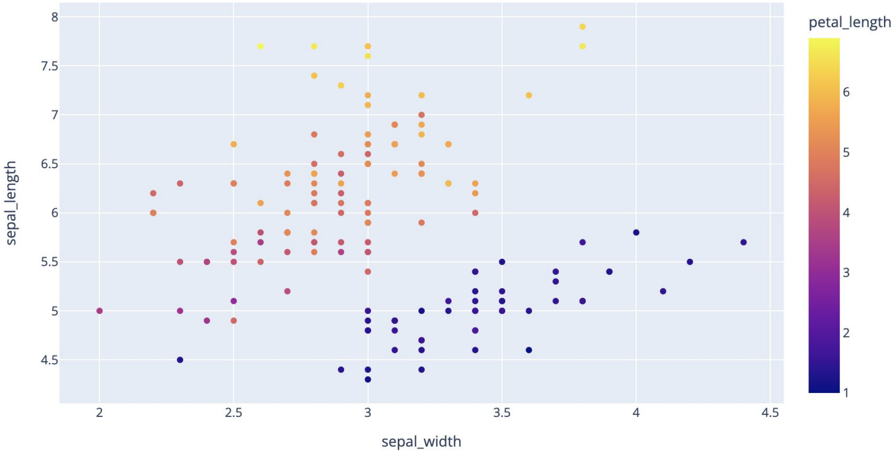
Box Plots

Histograms

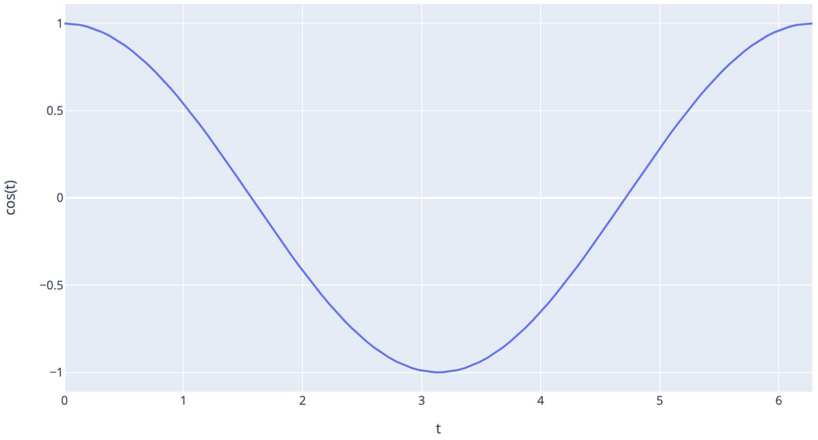
Scatter plots



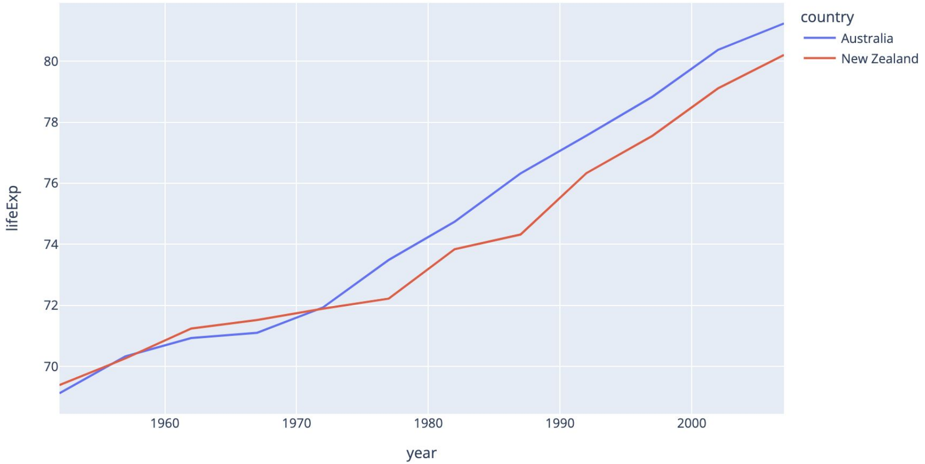




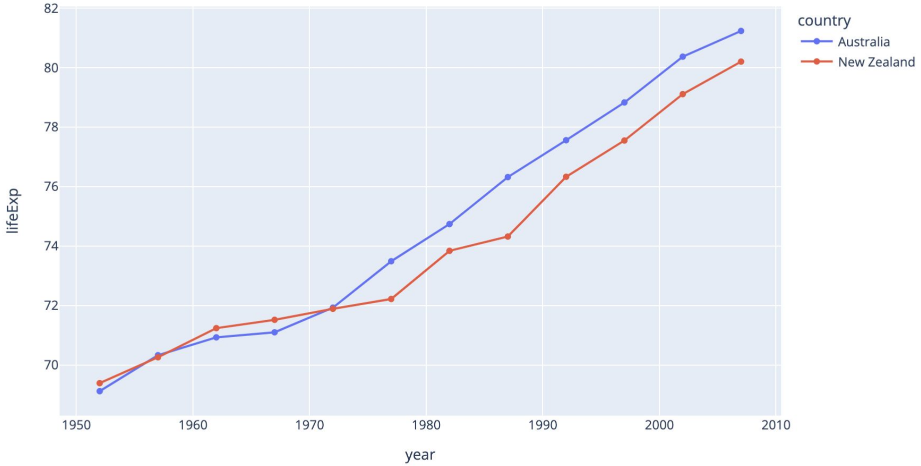
Line plots



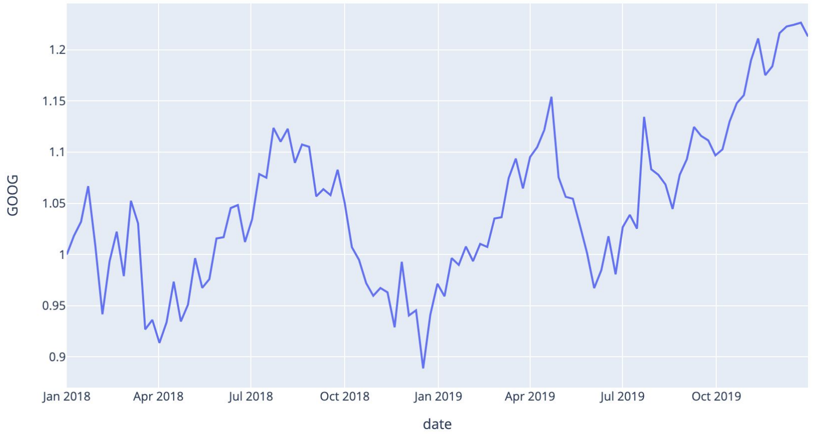
1
0.5
0
−0.5
−1
cos(t)
0
1
2
3
4
5
6
t



country
Australia
New Zealand
80
78
76
74
72
70
lifeExp
1960
1970
1980
1990
2000
year



country
Australia
New Zealand
82
80
78
76
74
72
70
lifeExp
1950
1960
1970
1980
1990
2000
2010
year

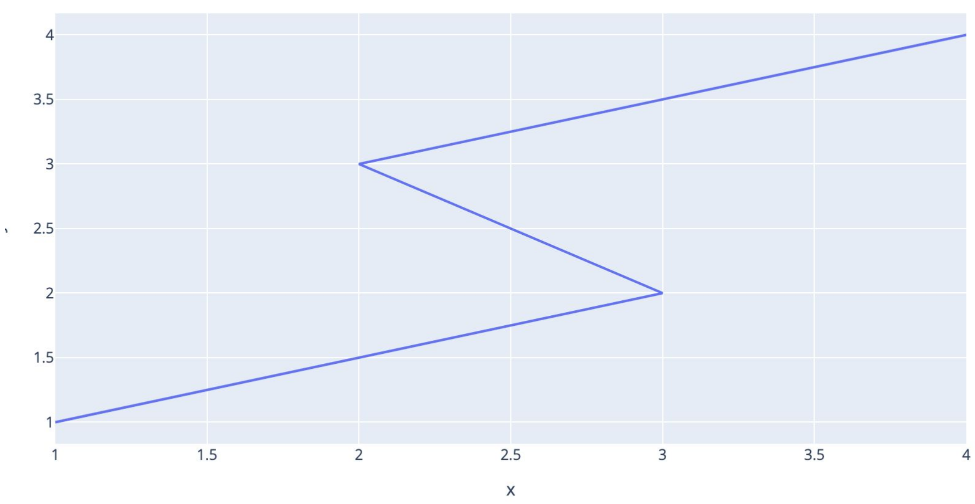


1.2
1.15
1.1
1.05
1
0.95
0.9
GOOG
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
Apr 2019
Jul 2019
Oct 2019
date

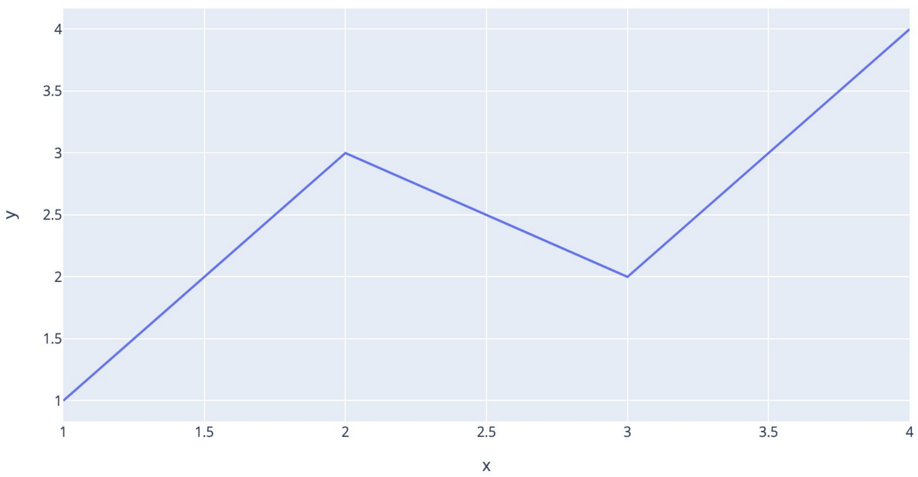
Data order in scatter and line charts

Plotly line charts are plotted and connected with lines in the order they are provided, with no automatic reordering.

It is required to explicitly sort data before passing it to Plotly to avoid lines moving "backwards" across the chart.

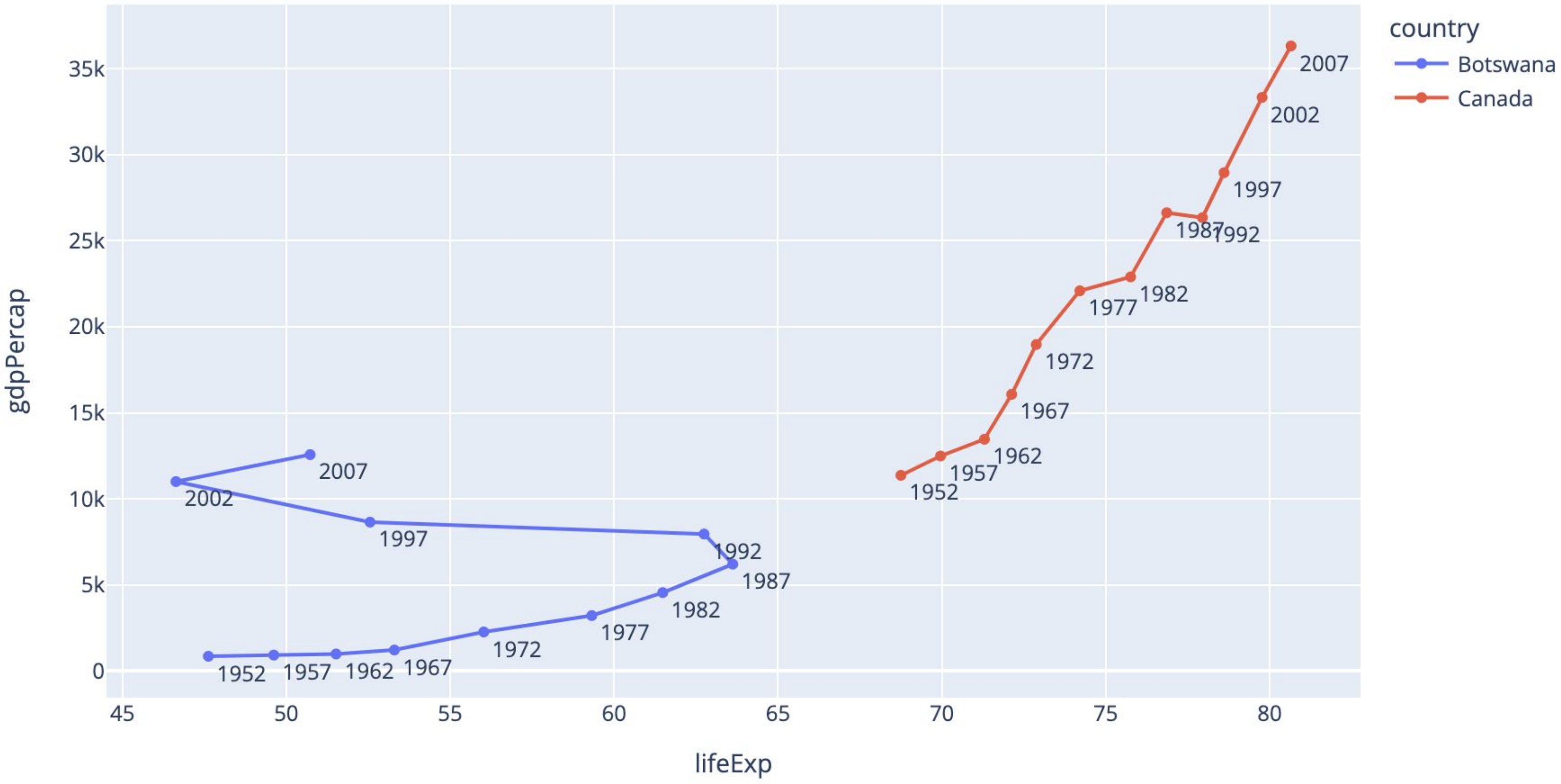


Unsorted Input

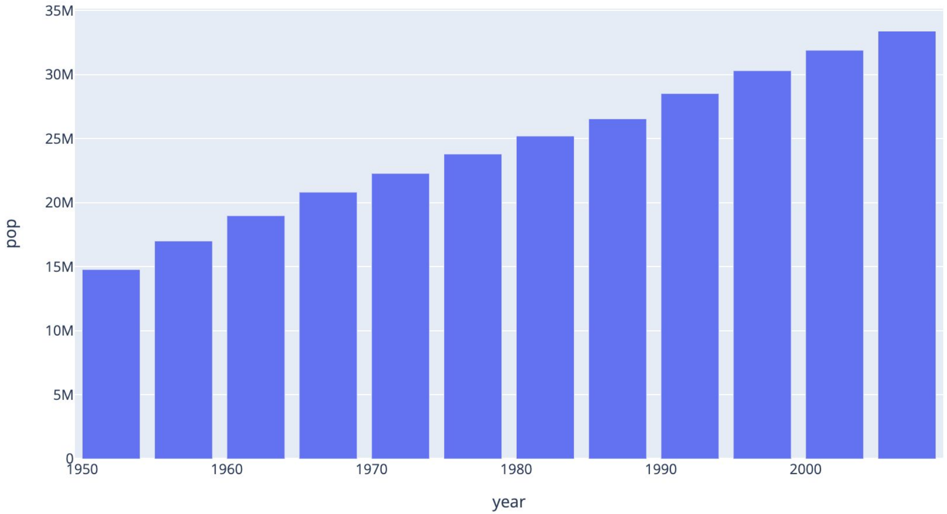


Sorted Input

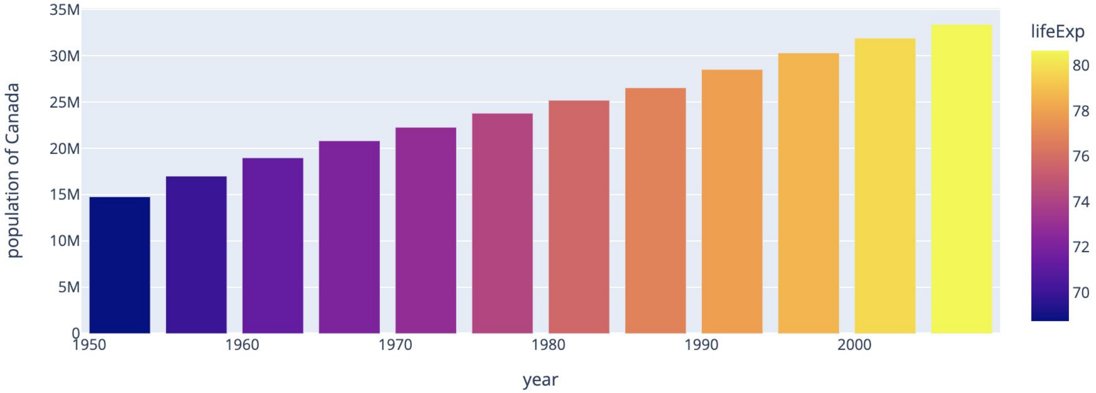
Update traces

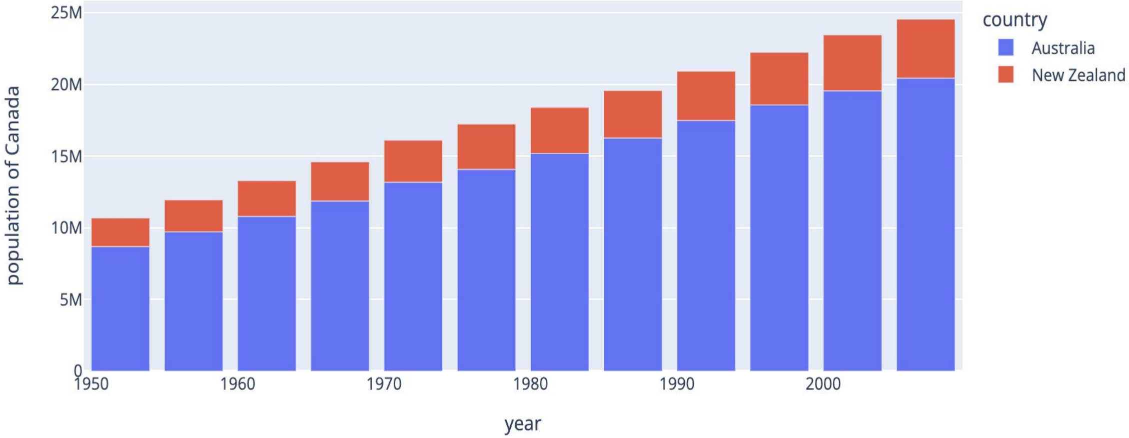


Bar charts

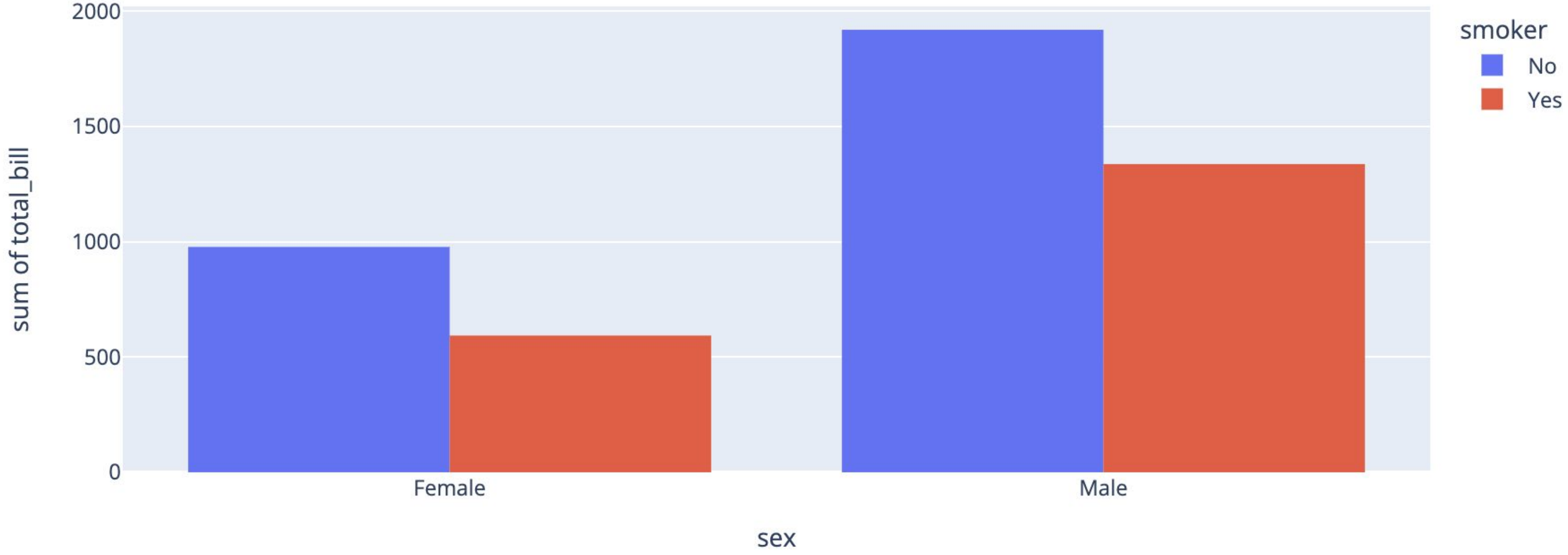

35M
30M
25M
20M
15M
10M
5M
0
pop
1950
1960
1970
1980
1990
2000
year

Colored bars



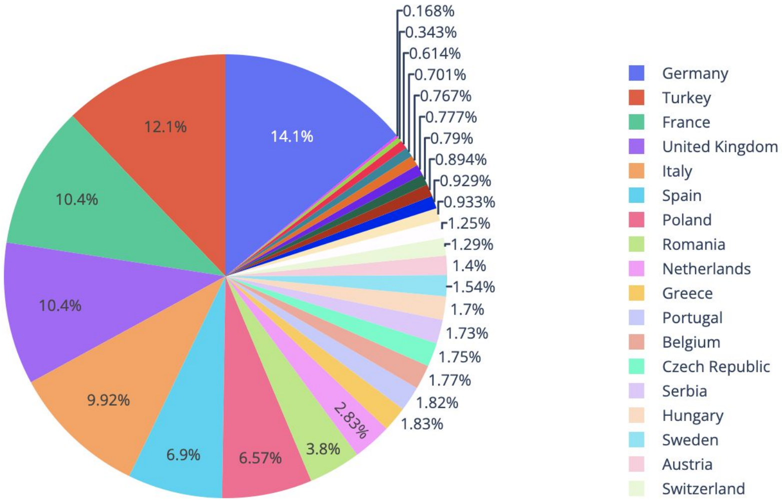


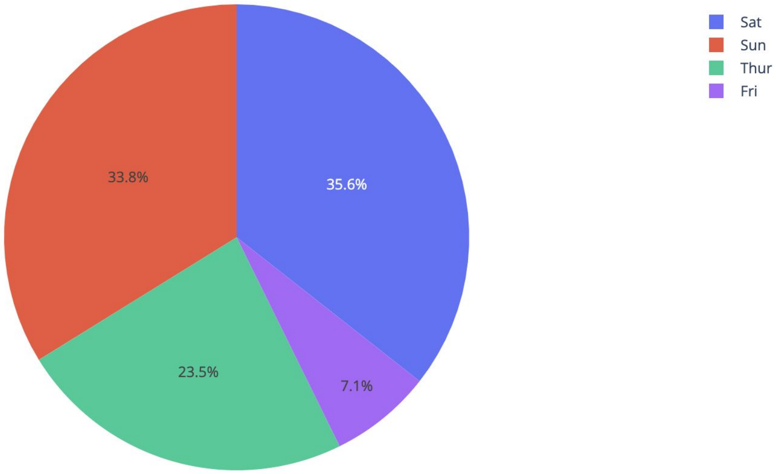
Grouping bar plots



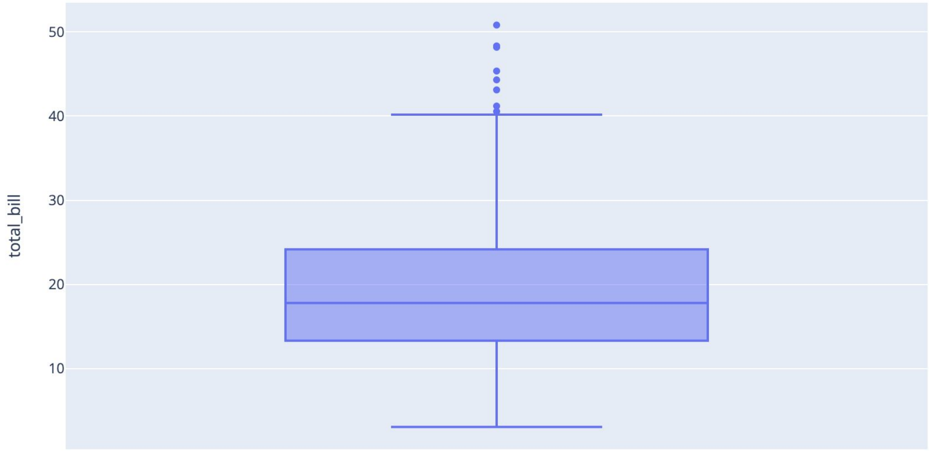
Pie charts

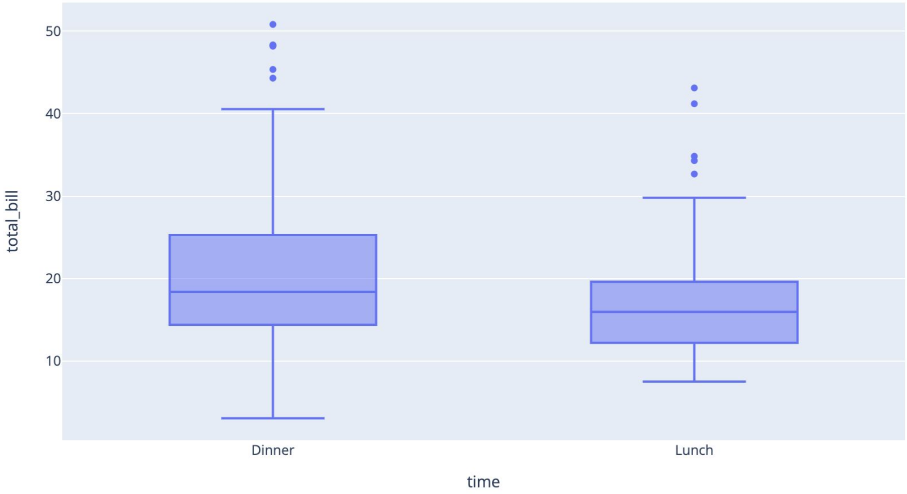
Population of European continent



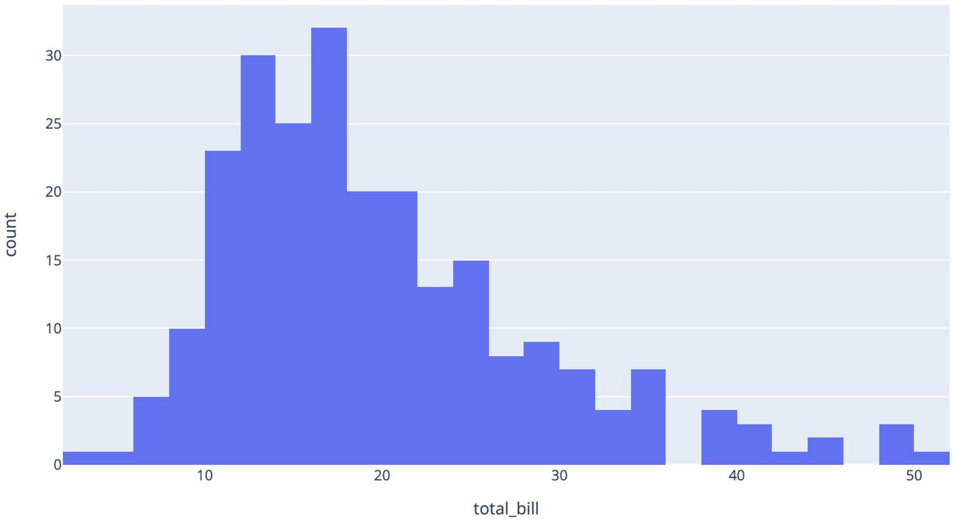


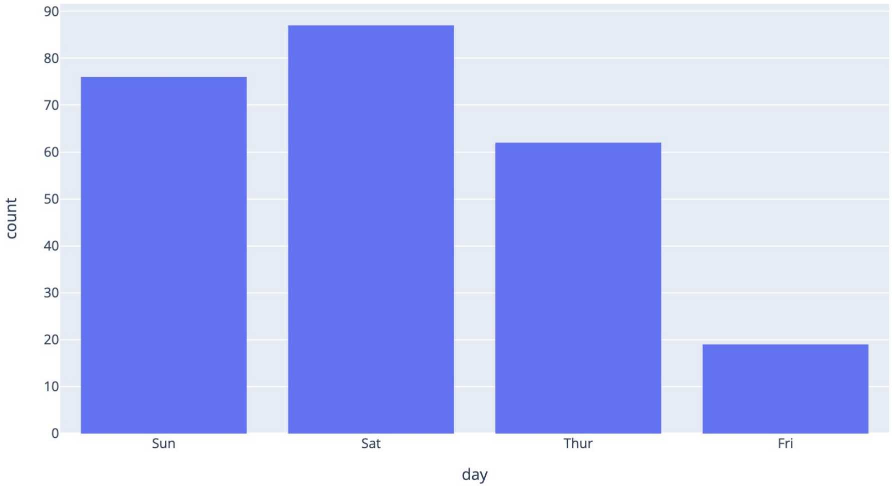
Box plots


total_bill
50
40
30
20
10

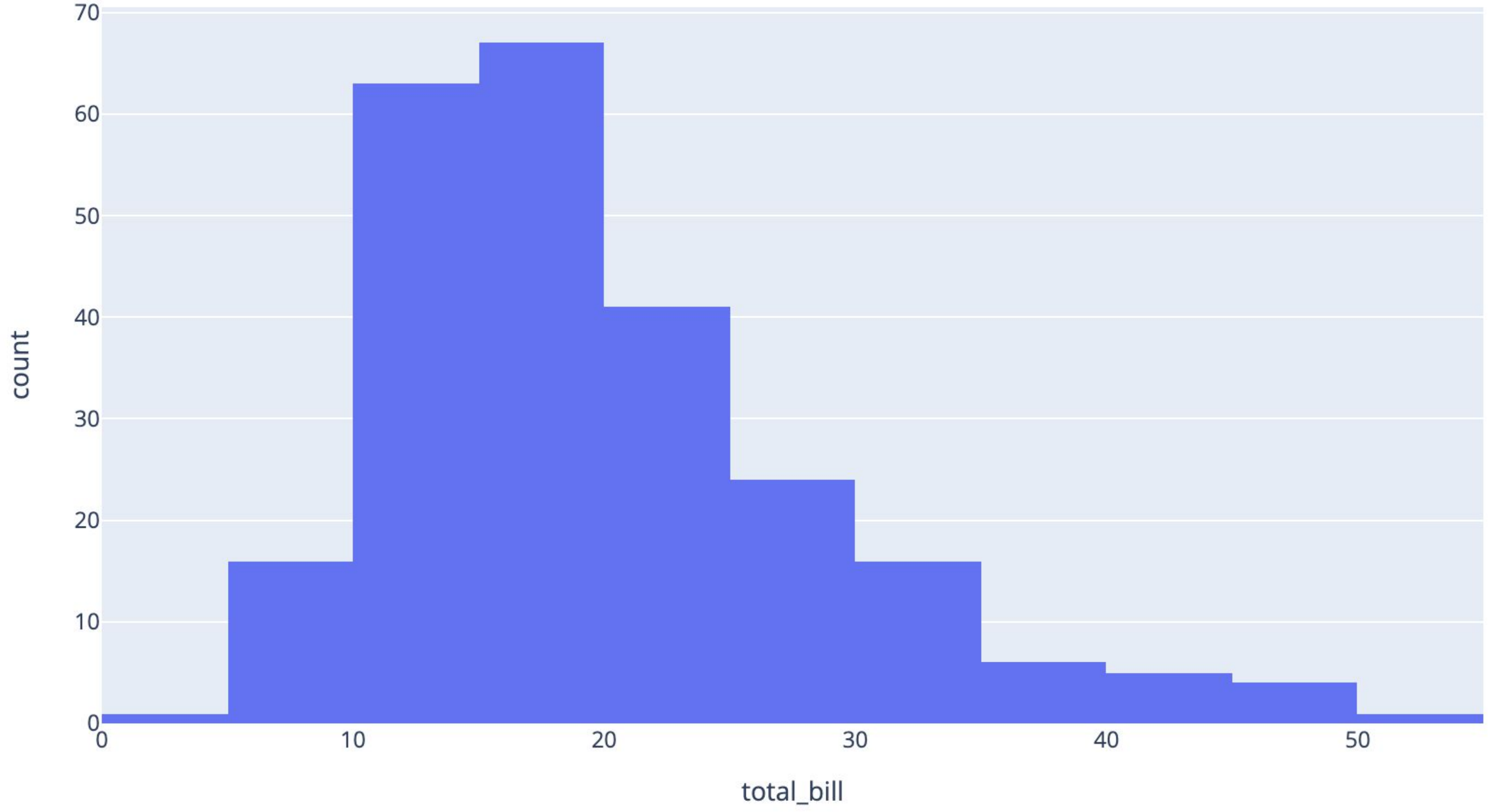

total_bill
50
40
30
20
10
Dinner
Lunch
time

Histogram plot





Histogram plot: choosing the number of bins



Exercise

In addition to the plots shown in the previous slides, generate one more plot of your choice using any of the data frames you worked so far.

Summary

- It’s important to choose plots that emphasize the important differences.
- Recommendations for ordering categories in plots - order box plots by medians and bar plots by height (nominal features).
- If they are ordinal than it’s typically best to keep that ordering in the plot.
- We provided guidelines for selecting color palettes according to the type of data plotted.

THANK YOU!