Data Analysis for Business Analytics

CIS621

Week 3 - Descriptive Statistics

Lecture Goals

1. Understand ways to present frequency of data
2. Understand measures to central tendency
3. Understand measures of data dispersion

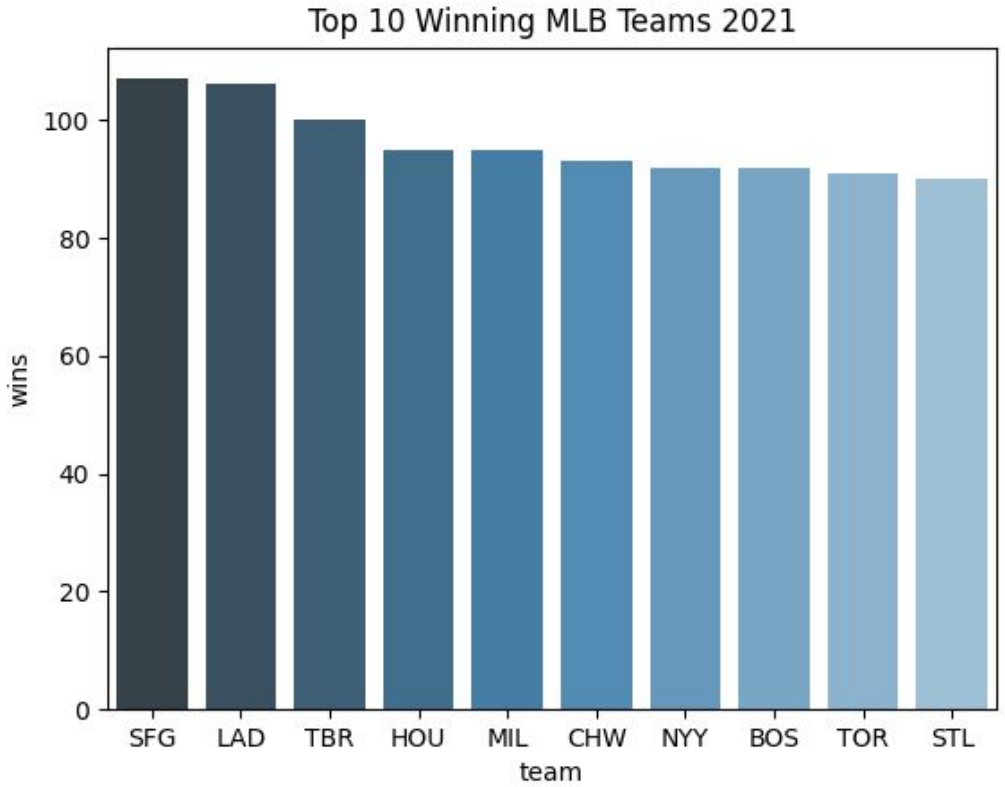
The World Without Descriptive Statistics

sales
10
12
9
25
27
24
23
17
18
18
20
21
30
34
35
35
37
40
35
31

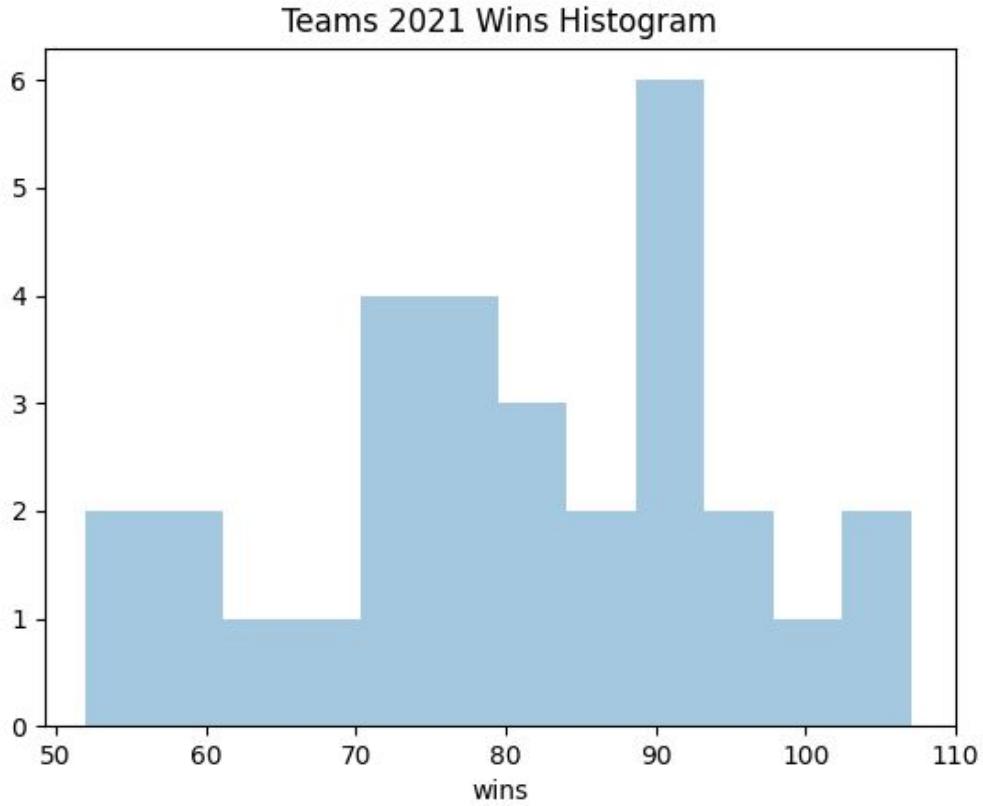
Frequency

- Counting!
- Though simple, sometimes showing counts is exactly what we need.
- For example: it could be useful to know monthly subscriber numbers in aggregate (though other ways to look at the data could be a useful or necessary complement).

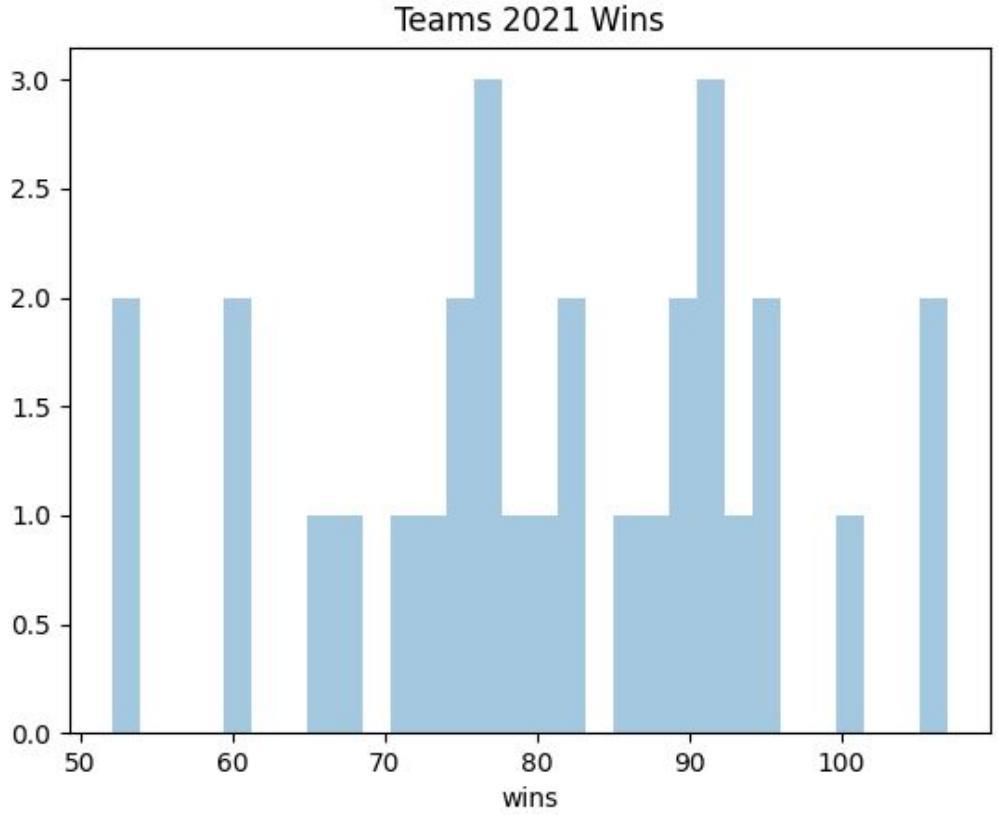
Bar Chart


Top 10 Winning MLB Teams 2021
wins
0
20
40
60
80
100
SFG
LAD
TBR
HOU
MIL
CHW
NYY
BOS
TOR
STL
team

Histogram



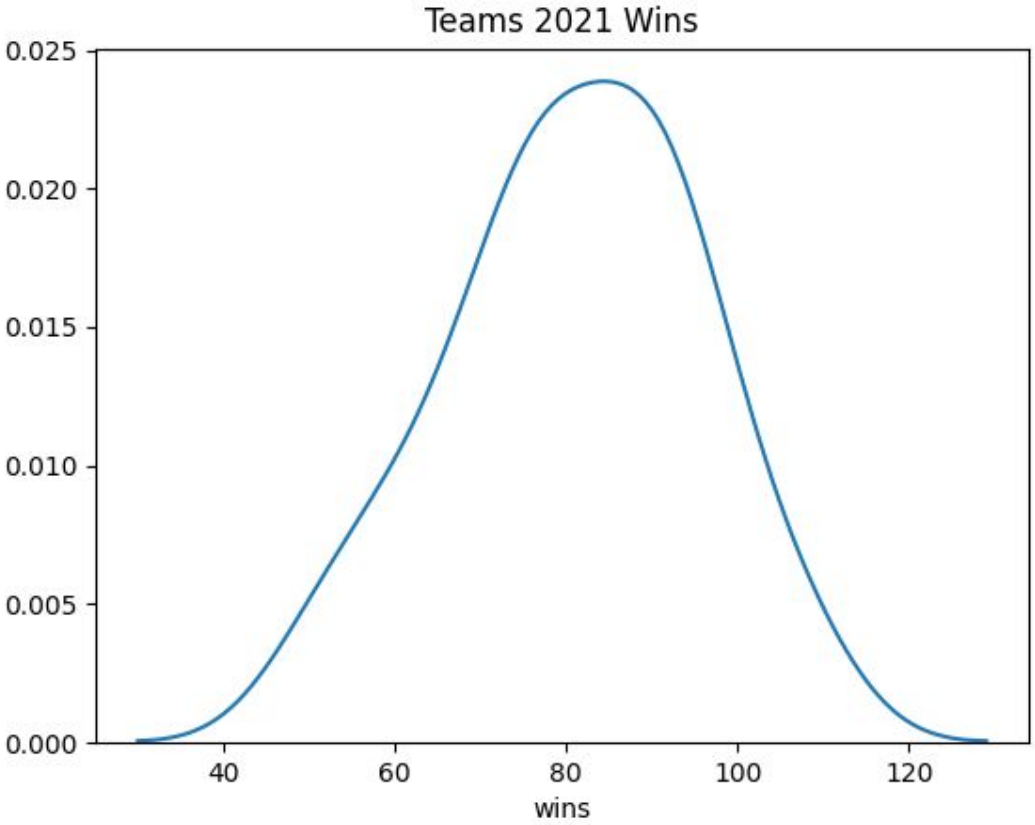
Histogram - More Bins



How Many Bins to Choose?

- Choosing the number of bins is part art and trial-and-error.
- You can either base it on the 1) number of bins you want or 2) the width of the bins (e.g., [10, 20, 30, 50]).
- These decisions should be based on the data at hand and what looks most useful.

An Alternate: Density Plots

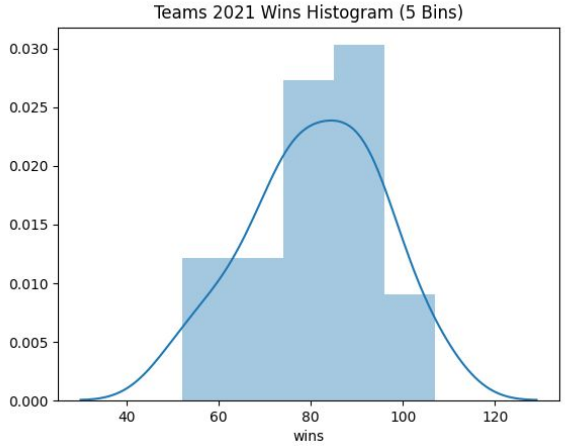


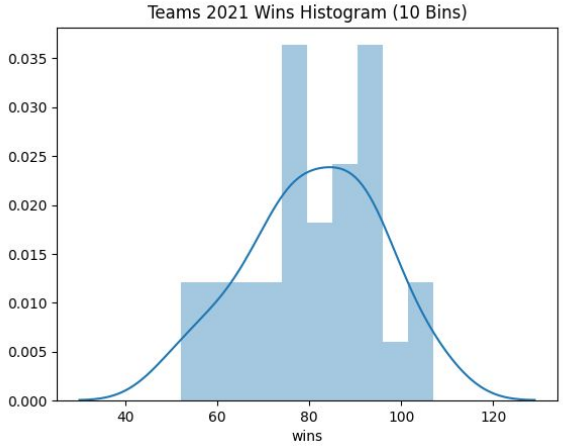
Density plots help answer a question like the following:

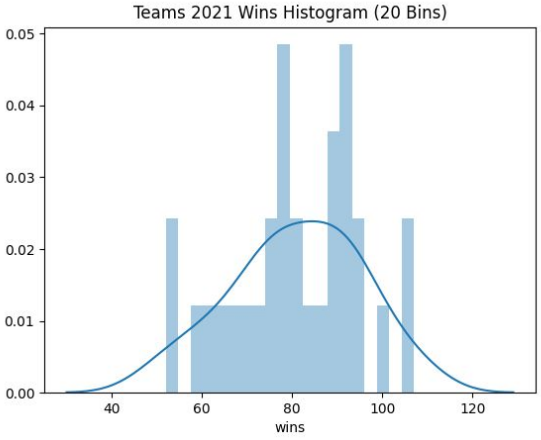
If we were to pick a random team, what is the likelihood they have, say, 60 wins?

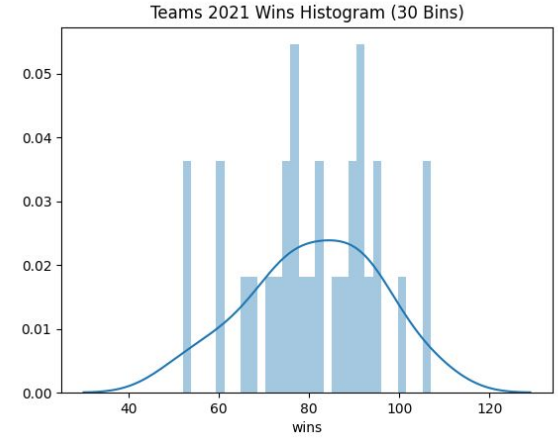
To note, the values are on the y-axis and difficult to interpret and should generally be removed when sharing with a business audience.

Histogram vs. Density Plot





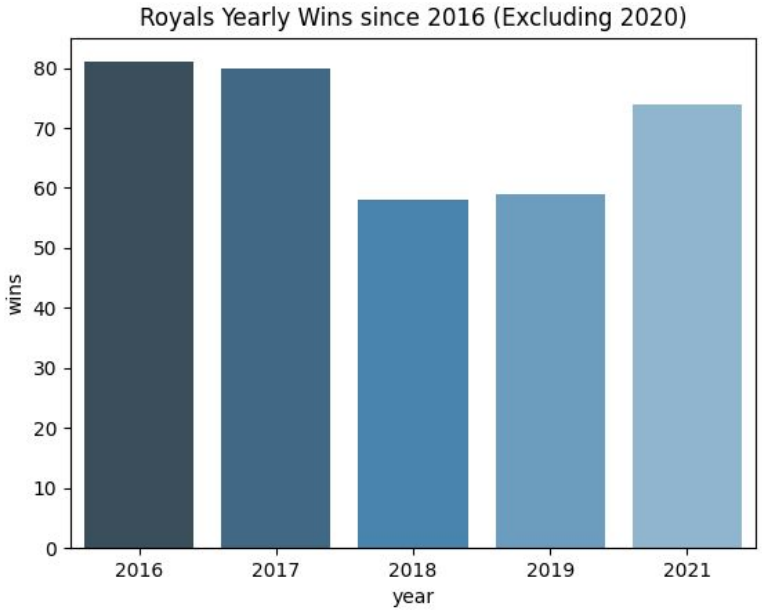


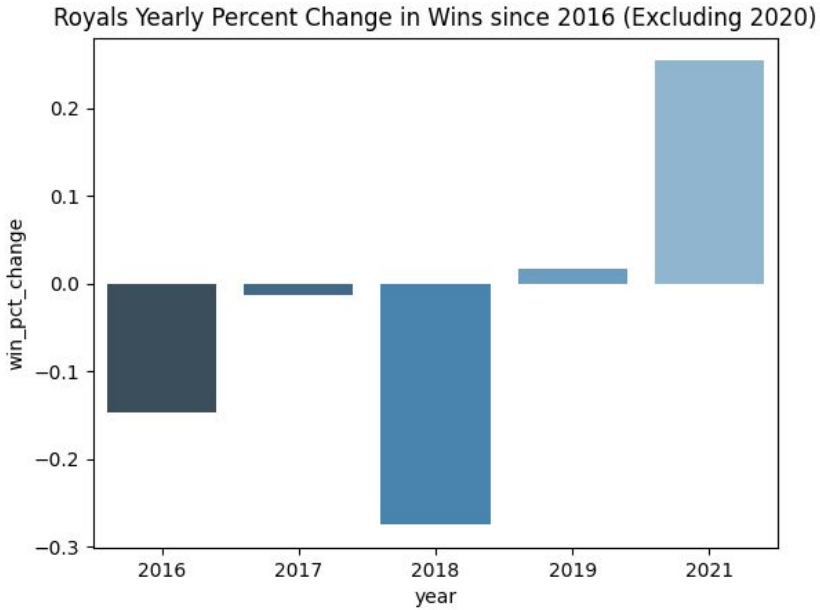


Histogram vs. Density Plot: When to Use Each?

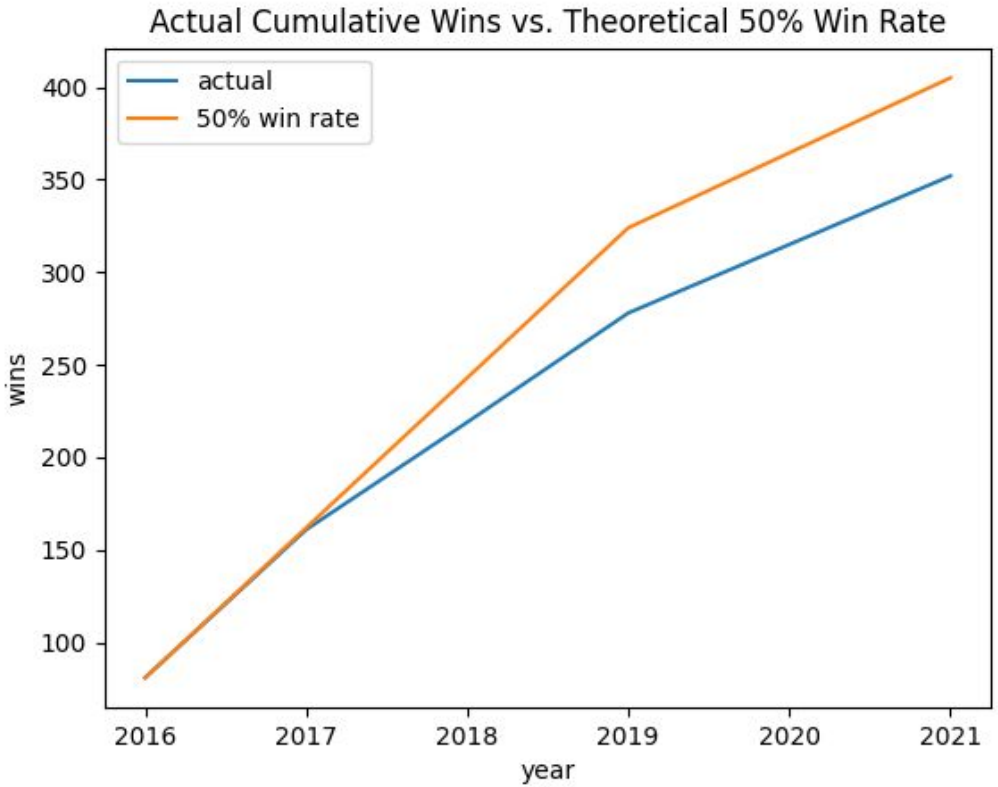
- Use histograms as the default since they are more widely understood.
- If you have a compelling reason to convey a *general pattern*, use a density plot.
  - This might be the case when you're data is particularly noisy.
  - This could also be the case when you know your data is a small sample, and adding generalization would tell a more complete story.
  - Keep in mind that the density plot will extend the range of your data.
  - Also recall that a density chart can be largely - though certainly only partly - replicated by histograms with a comparatively small number of bins.
- Generally avoid presenting a combination histogram-density plot to business stakeholders since these can be busy.

Pivoting Frequency: Percent Change



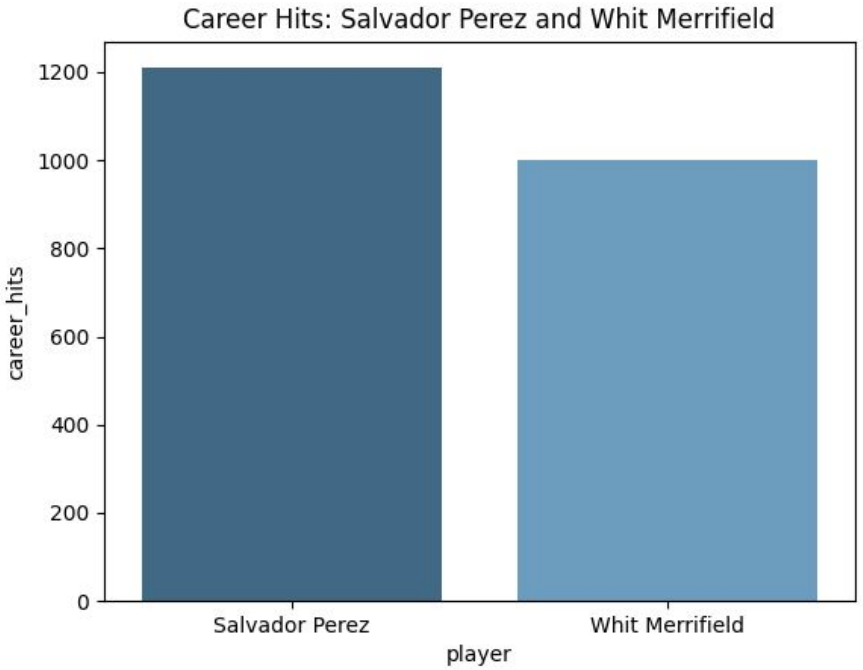


Pivoting Frequency: Cumulative Growth



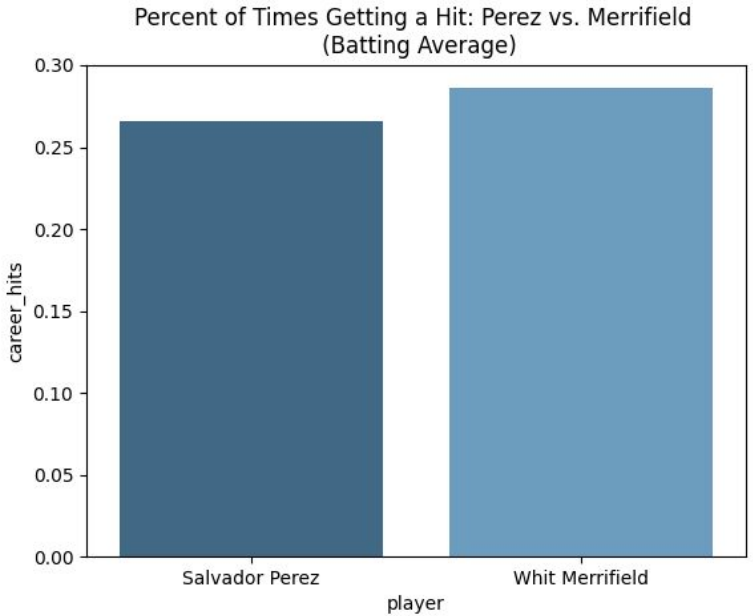
Be Careful when Comparing Counts!

What's might be wrong with the below chart?



Counts vs. Percentages

Oftentimes, we want to use percentages to compare multiple entities to ensure the values are comparable.



Measures of Central Tendency

- Mean: the average
- Median: the midpoint
- Mode: the most frequent value

Class Discussion

In what situations would we want to use the mean? The median? The mode?

Mean, Median, Mode

- Mean: use as default choice
  - If the mean and median are close, favor reporting the mean if you want to only report one measure
- Median: use when outliers are present
  - Sharing both the mean and the median is often a good idea
  - When the mean and median are noticeably different, the median is likely the better measure of central tendency since this case indicates the presence of outliers
- Mode: use when values are in a very limited range or certain values are of high interest

Points for Sharing Central Tendency with a Business Audience

- Use the mean *if appropriate*, which is won't always be. Everyone understands the mean.
- Depending on the the audience, you may have to explain the median. If so, do so in a simple way: "The median is the mid-way point: where half of data is above the point and half of data is below the point."
- Instead of "mode", use "most frequent value". If you need to refer to it multiple times, explain the terms are interchangeable.
- Sharing both the mean and the median - and training your audience to look at both - can often be a wise choice.

Common Measures of Dispersion

- Percentiles
- Range
- Interquartile Range
- Standard Deviation
- Skew
- Kurtosis

Percentiles

- Best demonstrated as an example: the 10th percentile represents the point at which 10% of observations are at or lower than.
- Generally, the percentiles used are those below, which supplies four buckets of values (this version is aptly called quartiles).

min	10
25%	30
50%	42
75%	68
max	100

Range

- The range is simply the maximum value - the minimum value.
- In the example below, the range is 90. You might also say something like “the values range from 10 - 100”.

min	10
25%	30
50%	42
75%	68
max	100

Interquartile Range (IQR)

- The difference between the 75th percentile value and the 25th percentile value.
- Unlike the range, the IQR controls for outliers.

min	10
25%	30
50%	42
75%	68
max	100

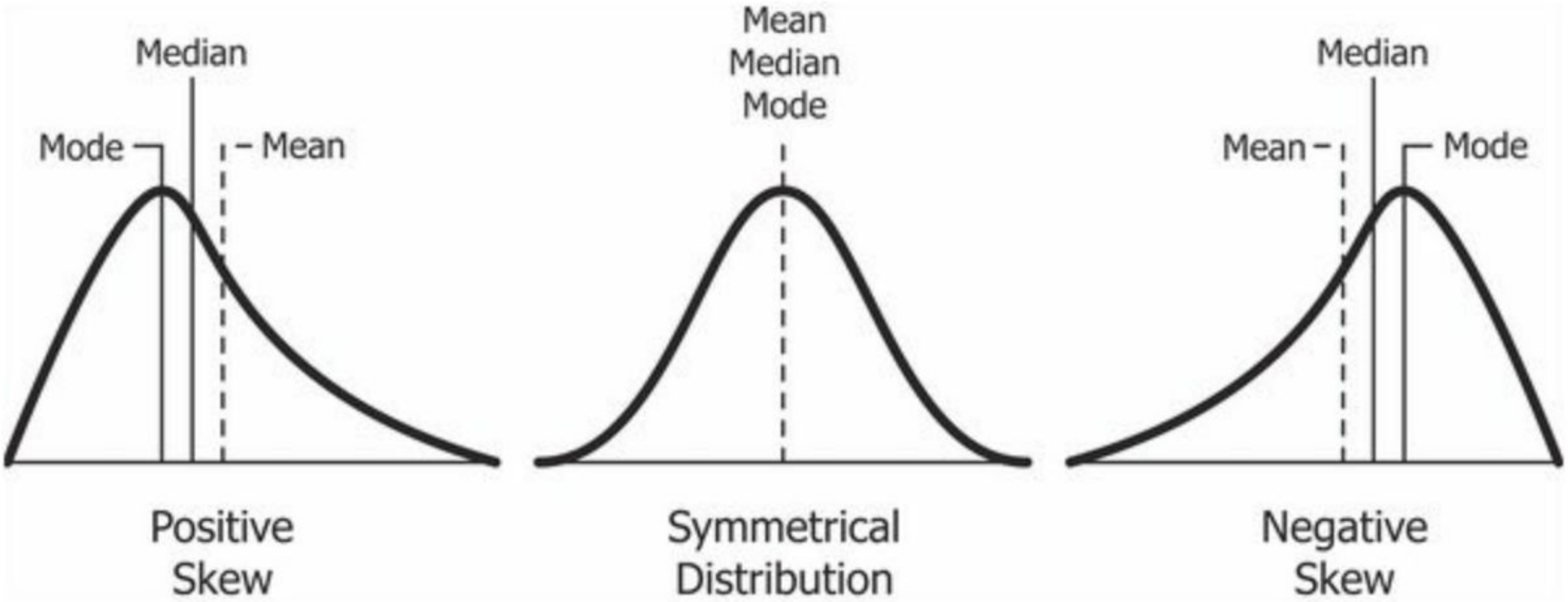
Standard Deviation

- Tells us the standard amount values deviate from the mean.
- Spreadsheet demo!

Skew

- A measure of asymmetry compared to a normal distribution.

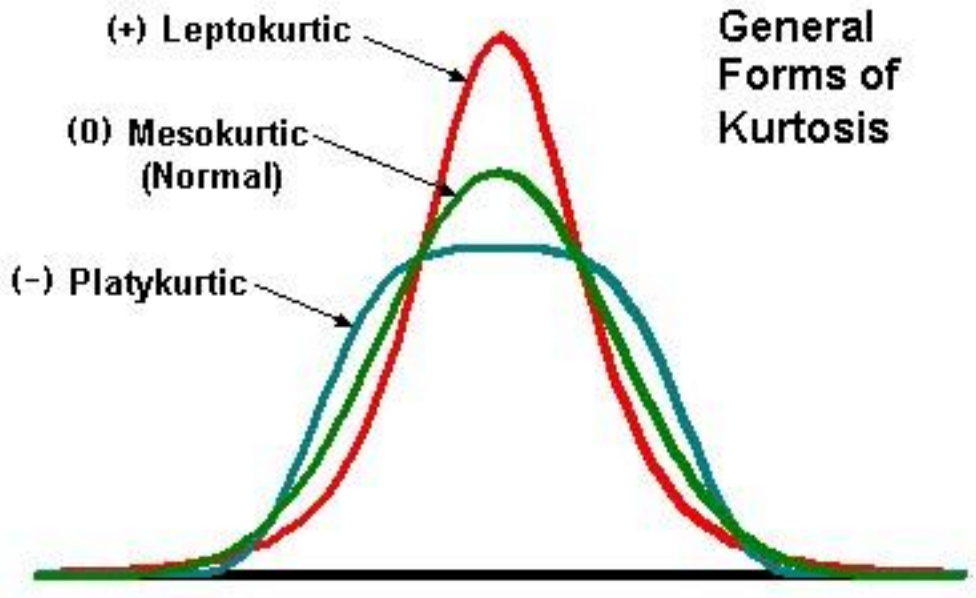
  Image Credit - Wikipedia



Kurtosis

- A measure of the “fatness” of tails.

Image Credit - [IT Feature]



How to Share Dispersion with a Business Audience

- Measures like skew and kurtosis are generally not suited for a non-technical audience.
- If your report would benefit from something more statistical, percentiles - probably specifically quartiles - would be the best place to start. Remember to provide concise explanations as needed.
- Depending on the audience, you may be able to get away with sharing the standard deviation, though some repeated education may be requisite.
- Remember: Not all analysis needs to be reported! Business stakeholders often don’t want to need to see all the backend math and stats!

How to Share Dispersion with a Business Audience

- The best way to communicate dispersion is in business terms.
- Example: “5% of our clients buy 40% of products”

Some Examples of Dispersion for Multivariate Data

- TSNE
- Euclidean Distance
- Nearest Neighbors
- Clustering
- Edit Distance

Demo: Exploring multivariate data dispersion in Python

Classroom Activity

- Find a data visualization online that shows 1) central tendency, 2) dispersion, or both in an effective way.