ChartMogul

The Ultimate **Data Literacy** Cheat Sheet

REMEMBER!

The five characteristics of good data:

Credible
Actionable
Unbiased
Statistically relevant
Easy to interpret

Averages: Which one should you use?

An average, also referred to as a "Measure of central tendency", is a value that attempts to identify the central position within a set of data. Mean, Median and Mode are types of average.

MEAN Does your data have a continuous distribution that's relatively symmetrical? Use **MEAN** (often referred to as just 'average').

MEDIAN Does your data contain significant outliers? Use **MEDIAN** - it's less influenced by this.

MODE represents the most common value in a dataset. If you're dealing with Nominal data (non-numeric categories like "industry vertical"), **MODE** is the only appropriate average to use.

CHARTING TIPS

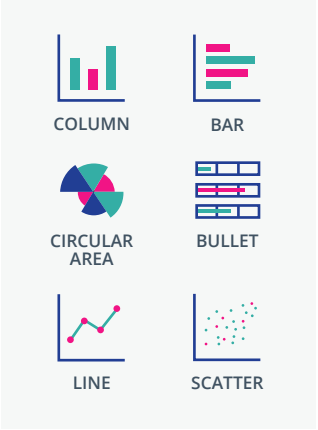
Weighted average

A weighted average is a type of **MEAN**, where some values in the data set are given more influence than others. Each value to be averaged is given a **weight**, representing the importance of that value in the average.

Weighted averages are important when you are dealing with frequencies or distributions, or when working with data that's unequal in some way.

Charts: Which one should you use?

Comparing multiple values



Displaying the composition of a whole

Showing distribution of values

Analyzing trends

Showing the relationship between sets

Common cognitive biases

CONFIRMATION BIAS
The tendency to search for results that confirm your preconceptions.

OBSERVATION BIAS
The tendency to see what we expect (or want) to see in results.

FUNDING BIAS
The tendency to support the interests of a study's financial sponsor or business.

SELECTION BIAS
The tendency to select data for analysis that is not properly random.

SAMPLING BIAS
(A type of Selection bias) The tendency to collect a sample of data in such a way that some members of the population are less likely to be included than others.

How to question data

SOURCE
Do you know where the data came from?

SCALES
Are the scales of each axis clear and effective?

FILTERS
Have any specific filters been applied to the data set?

TIMEFRAME
What is the date range for the presented data?

GAPS
Are there obvious omissions to the data set?

EXCESS
Is there anything presented that's not relevant?

UNIT (S)
Is it clear what the data in the chart represent?

LABELS
Is the data clearly titled and labelled, in a descriptive way?

ACTIONABLE
Can the insights presented be used in an actionable way?

TREND
Is it trending upwards, downwards or flat?

PATTERNS
Are there cyclic patterns (e.g. seasonality) in the data?

DIMENSIONS
Is the data segmented into clear, meaningful dimensions, e.g. "Pricing plan"?

CHARTING TIPS

Should I truncate the Y axis?

Truncating the Y axis (i.e. not starting at zero) is controversial in the world of data visualization. It can be misinterpreted and has been used to mislead consumers in a number of cases. In general, it's not a good idea unless you can clearly show that the axis doesn't start at zero.

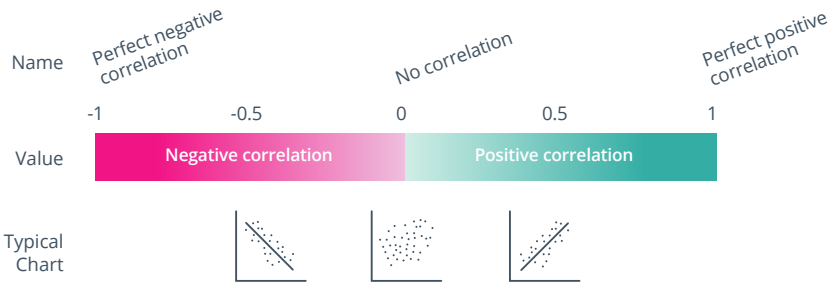
• It can be acceptable to truncate when you're displaying data that would never feasibly reach zero anyway, e.g. global average temperature.

• If you need to emphasise small changes and trends in data, consider showing a chart of the change rather than absolute numbers.

NEVER truncate the Y axis when:
1. Using a bar or column chart (Bar charts should start at zero).
2. The information is likely to be mis-interpreted.
3. It doesn't help in some way with understanding the chart.

Data correlation

Correlation is a statistical relationship between two data sets. Correlation can have a numeric value on a scale from -1 to 1.
A **POSITIVE** correlation is present when both values increase together, whereas in a **NEGATIVE** correlation, one value decreases as the other increases.



CHARTING TIPS

Correlation is not causation!

A strong correlation between two data sets does not necessarily mean that one thing causes the other (causation). There could be other reasons the data has a strong correlation.

Glossary

QUALITATIVE DATA is descriptive - it describes something, e.g. Reason for customer cancellation.

QUANTITATIVE DATA is always numerical (involves numbers), e.g. Revenue lost from customer cancellations.

DISCRETE DATA can only take certain values (like whole numbers), e.g. Number of customers churned.

CONTINUOUS DATA can take any value, within a given range, e.g. Customer churn rate.

CATEGORICAL DATA can be sorted, according to defined groups or categories, e.g. Industry vertical.

STATISTICAL SIGNIFICANCE is when the observed outcome of an experiment is unlikely to have occurred due to chance. This is an important factor when running multi-variant (A/B) tests on your product or website.

Determining statistical significance can be complex. We recommend using a free tool such as AB Testguide: https://abtestguide.com/calc

More resources

The Data Viz Checklist by Stephanie Evergreen: http://stephanieevergreen.com/wp-content/uploads/2016/10/DataVizChecklist_May2016.pdf

Graph Choice Chart by Tuva Labs: https://tuvalabs.com/static/documents/Graph_Choice_Chart.pdf