Data Visualization

Table of Contents

[Lesson 1: Why visualize data? 1](#_Toc501878330)

[Anscombe’s quartet 1](#_Toc501878331)

[Datatypes review 1](#_Toc501878332)

[Type of plots 3](#_Toc501878333)

[Univariate plots 3](#_Toc501878334)

[Scatter Plots 4](#_Toc501878335)

[Line Plots 4](#_Toc501878336)

[Lesson 1 Summary 5](#_Toc501878337)

[Lesson 2 Design 6](#_Toc501878338)

[Chart Junk 7](#_Toc501878339)

[Lie Factor 7](#_Toc501878340)

[Colors 8](#_Toc501878341)

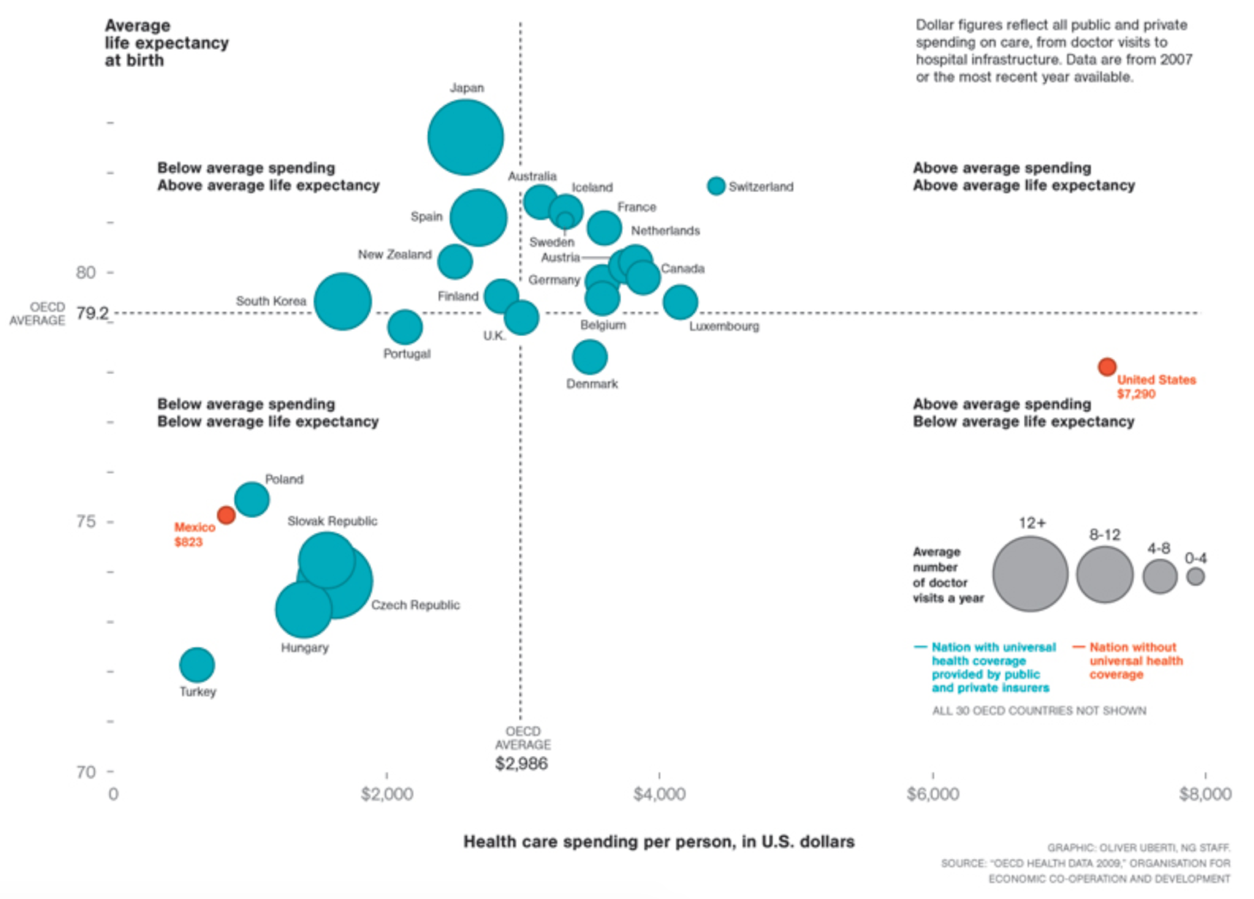
[Groups vs Sets in Tableau 8](#_Toc501878342)

# Lesson 1: Why visualize data?

## Anscombe’s quartet

It is a good example showing multiple datsets with same summary stats but which are actually very different. This tells us how important it is to visualize data along with summary statistics analysis.

# Datatypes review



In the plot above,

1. Doctor visits per year is a ordinal variable ( i.e. ordered categorical).
2. Spending per person is continuous
3. Has universal care (categorical)
4. Life expectancy (continuous)

# Type of plots

## Univariate plots

For **quantitative** data, if we are just looking at one column worth of data, we have four common visuals:

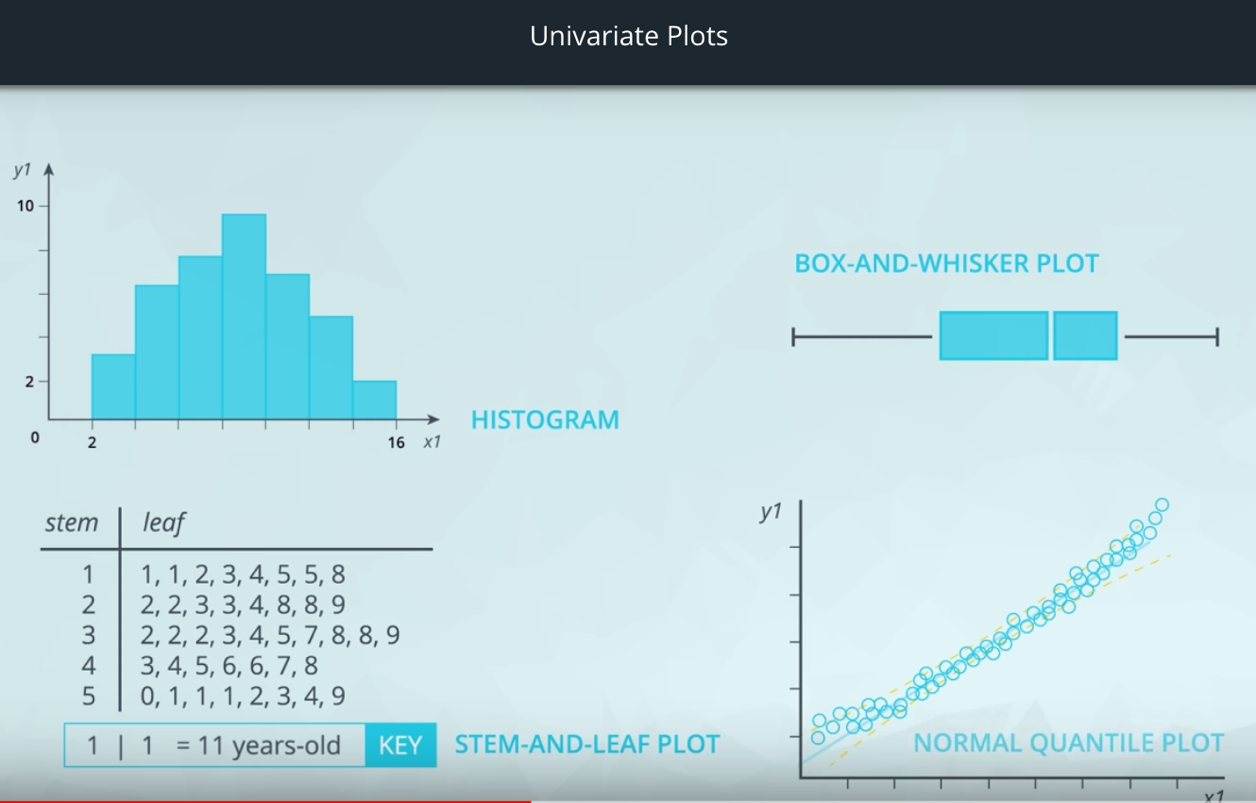
1. Histogram
2. Normal Quantile Plot
3. Stem and Leaf Plot
4. Box and Whisker Plot

In most cases, you will want to use a **histogram**. Histogram: Both x and y are quantitative (continuous or discrete).

For **categorical** data, if we are looking at just one variable (column), we have three common visuals:

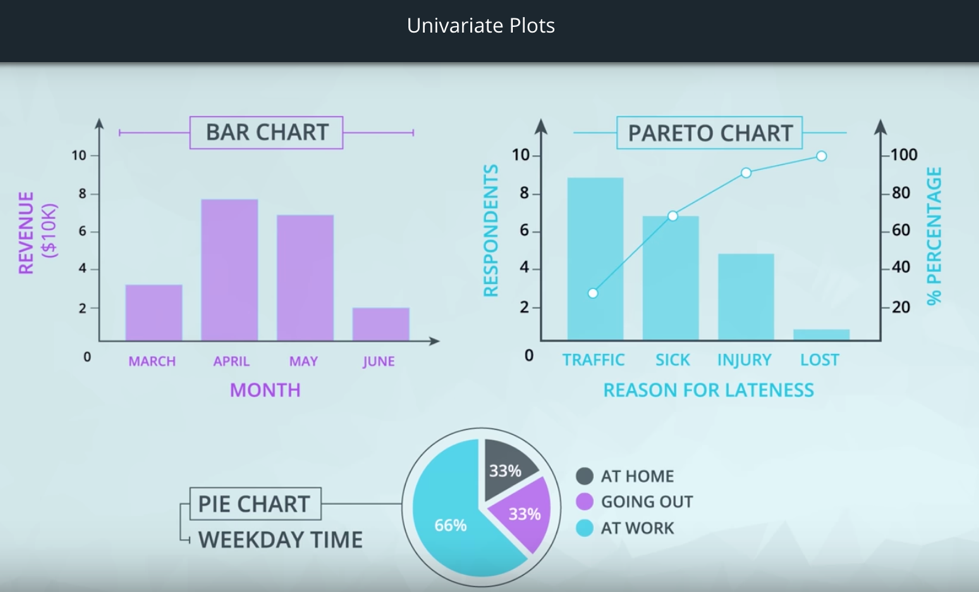
1. Bar Chart
2. Pie Chart
3. Pareto Chart

In most cases, you will want to use a **bar chart**.



Categorical Data: Bar chart,

Pareto chart (most frequent bar at beginning). Example: type of fault at factory.



## Scatter Plots

Scatter plots are a common visual for comparing two **quantitative** variables.

A common summary statistic that relates to a scatter plot is the **correlation coefficient** commonly denoted by **r**.

Though there are a [**few different ways**](http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/) to measure correlation between two variables, the most common way is with [**Pearson's correlation coefficient**](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient).

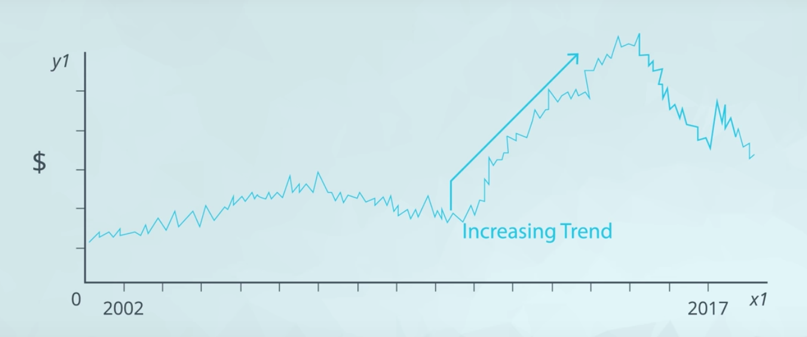
**Pearson's correlation coefficient** provides the:

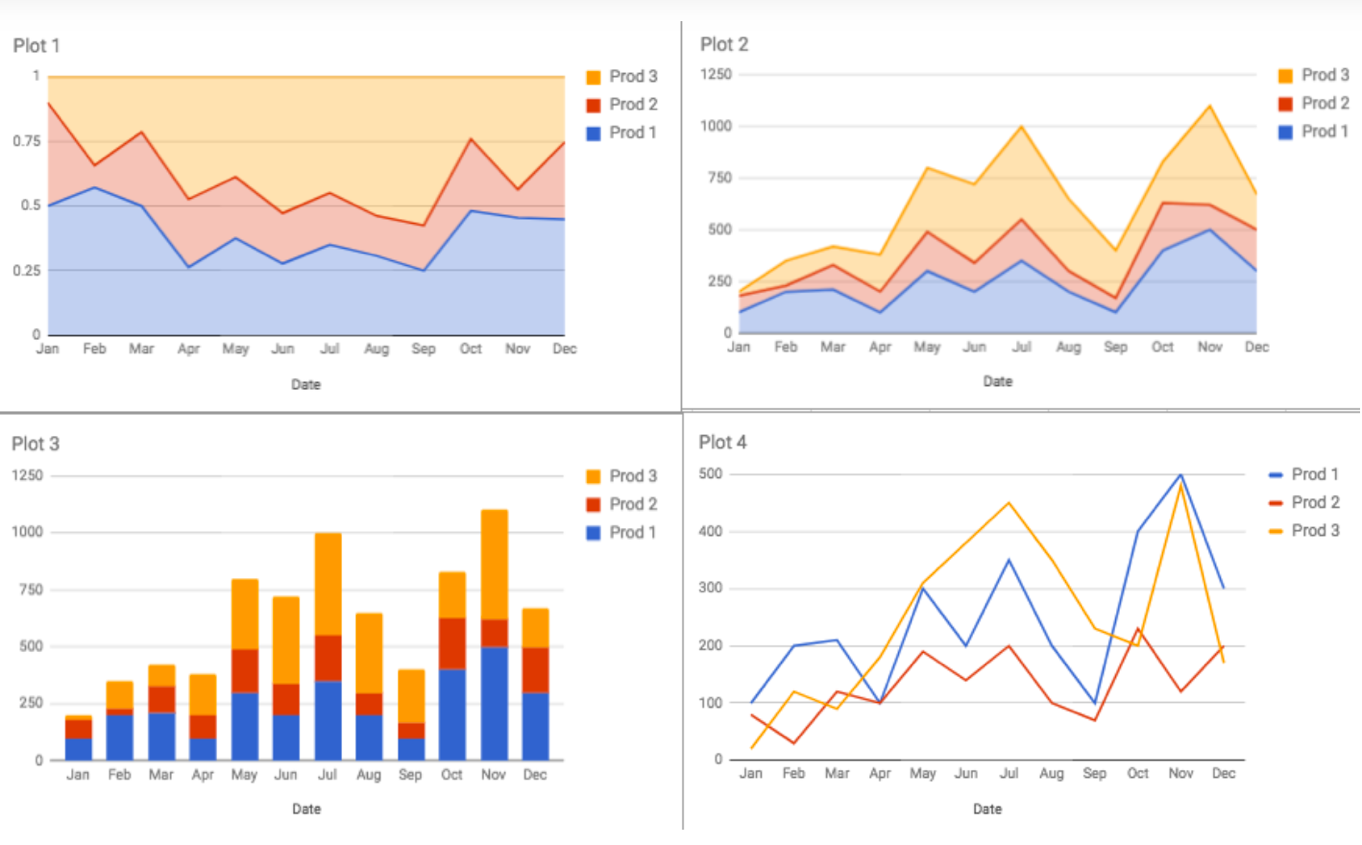
1. Strength
2. Direction of a **linear relationship**.

[**Spearman's Correlation Coefficient**](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient) does not measure linear relationships specifically, and it might be more appropriate for certain cases of associating two variables.

## Line Plots

Line plots are a common plot for viewing data over time. These plots allow us to quickly identify overall trends, seasonal occurrences, peaks, and valleys in the data. You will commonly see these used in looking at stock prices over time, but really tracking anything over time can be easily viewed using these plots.





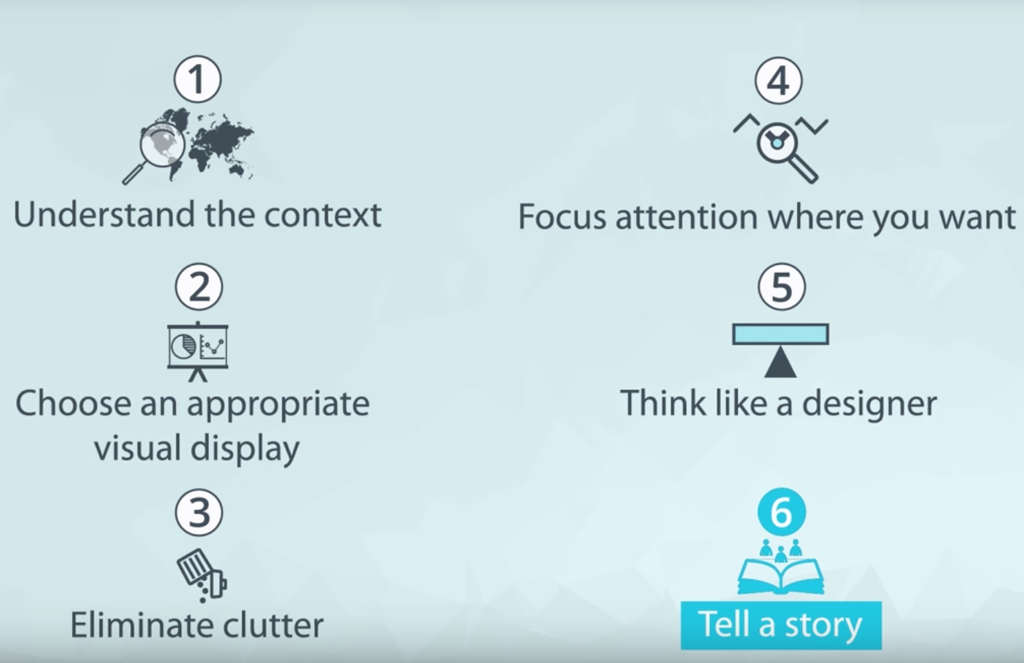
Specific plots for specific questions:

1. Product 3 had fastest growth from Jan to July.
2. Product 1 had more than 50% sales in February.
3. Total sales exceeded 1000 units in November

# Lesson 1 Summary

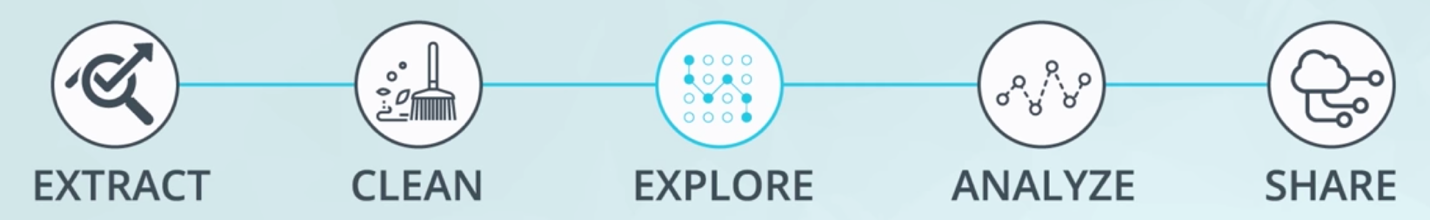
1. You motivated the need for data visualization by showing that summary statistics don't tell the full story. You saw datasets where the summary statistics were the same, but the actual data were very different!
2. You did a review of data types. In general there are quantitative and categorical variables. Quantitative variables can be either discrete or continuous, while categorical variables are either ordinal or nominal.
3. You looked at univariate plots. In most cases a **histogram** should be used for quantitative data, while a **bar chart**should be used for categorical data. There are some cases where you might use one of the other plots.
4. You then looked at bivariate plots, where you were comparing two variables to one another. **Scatter plots** are the most common way to visualize two quantitative variables, while a **line chart** is common for data that you are watching over time. If you are comparing two categorical variables, the best choice is probably a **side-by-side bar chart**.
5. You learned about **correlation coefficients**, which provide the strength and direction of linear relationships. You learned a rule of thumb for determining the whether the relationship between two quantitative variables is **strong**, **moderate**, or **weak**.
6. You then looked at cases where we had more than two variables. You learned that using these plots effectively is about building the plot that helps you see the insight that answers the question you have.
7. You gained some insight into visual encodings and data dashboards, which will be a part of the next lessons!

# Lesson 2 Design



1. **Understand the context** - this means knowing your audience and conveying a clear message about what you want your audience to know or do with the information you are providing.
2. **Choose an appropriate visual display** - this was covered in the last lesson. Check out the recap page [**here**](https://classroom.udacity.com/nanodegrees/nd002/parts/84946f6a-429e-434a-b7be-a98f15d96913/modules/32c2a780-2f58-4556-b494-ddf0198e754c/lessons/3170e950-5743-4dbb-875f-6da4a3ed30f1/concepts/link) if you need a quick refresher.
3. **Eliminate clutter** - you should only provide information to the user that helps convey your message.
4. **Focus attention where you want it** - build visualizations that pull attention to the message you want to highlight.
5. **Think like a designer** - you will learn a number of design principles in this lesson to assist as you start to put together your own data visualizations.
6. **Tell a story** - your visualizations should give the audience a story. The most powerful data visualizations move people to take action.





## Chart Junk

Examples of chart junk you saw in this video include:

1. Heavy grid lines
2. Unnecessary text
3. Pictures surrounding the visual
4. Shading or 3d components
5. Ornamented chart axes

## Lie Factor



The graphical effect is (146-27)/27 = 4.407, and the data effect is (39.6-35)/35 = 0.131, giving a lie factor of 33.54. Any lie factor different than 1 suggests that a visual is distorting the data.

When the factor is greater than 1, we are making an effect larger than it actually is and factors less than 1 are hiding the magnitude of an effect.

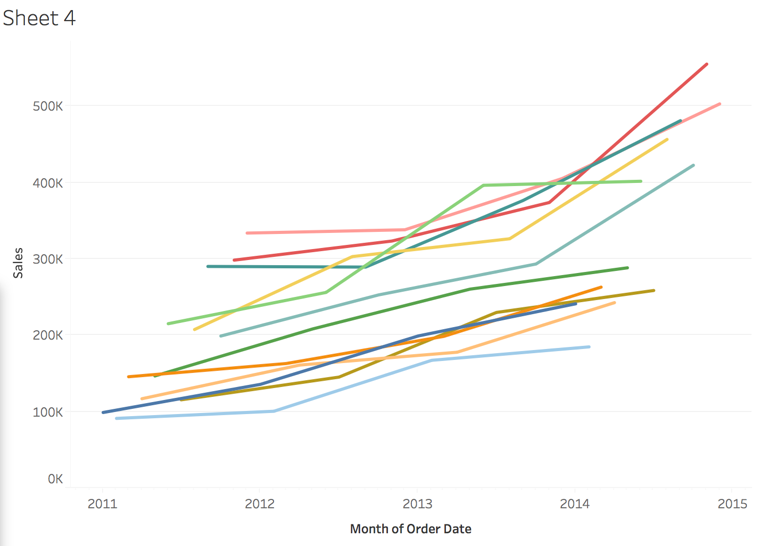
## Colors

No high intensity rainbow colors

Don’t use colors if unnecessary

Use blue and orange hues to make visuals color-blind friendly.

Continuous



## Groups vs Sets in Tableau