Week 2: Graphical EDA

STA238: Probability, Statistics, and Data Analysis II

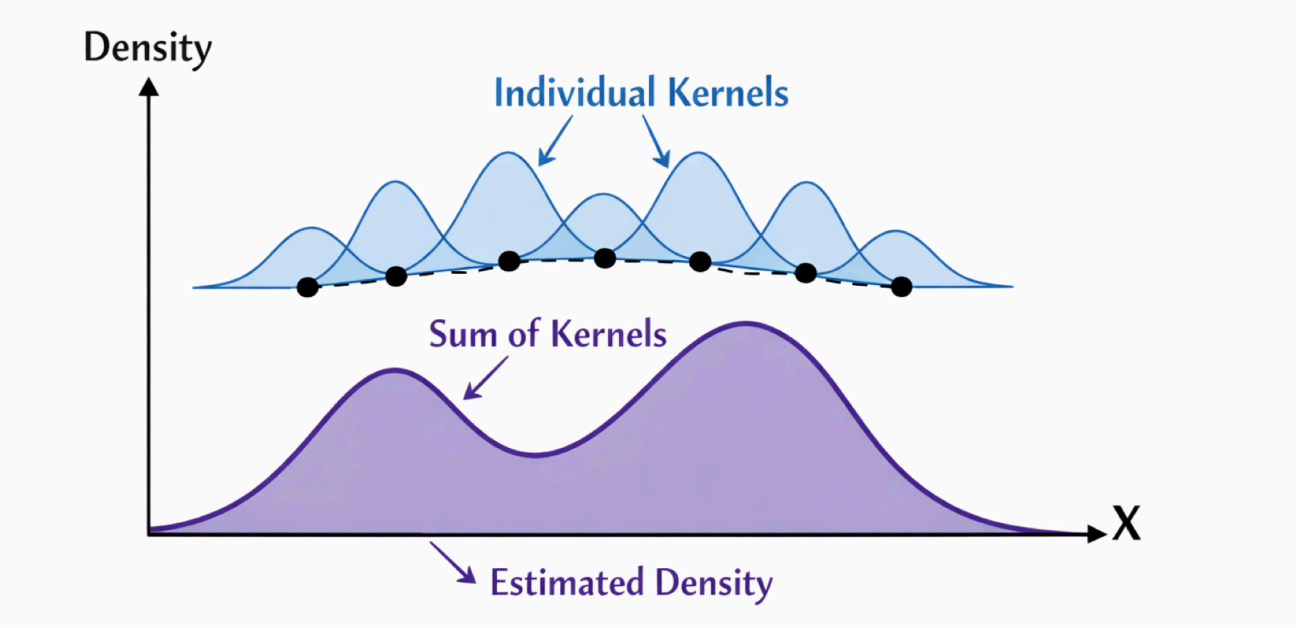
Luis Sierra Muntané

2026-01-15

KDE Review

How should we construct a density if all we have are samples?

Use each observation to give us local information on what the density should look like:



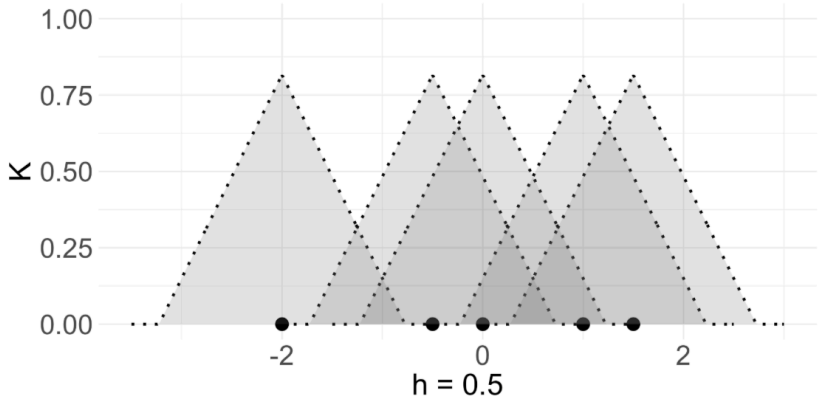
Constructing a KDE

1. Choose a valid kernel K for your data.
2. Choose a bandwidth h to scale K:

$$t \mapsto \frac{1}{h} K\left(\frac{t}{h}\right).$$

3. To each data point x_i, associate a kernel to be centered at it:

$$t \mapsto \frac{1}{h} K\left(\frac{t - x_i}{h}\right)$$



Constructing a KDE

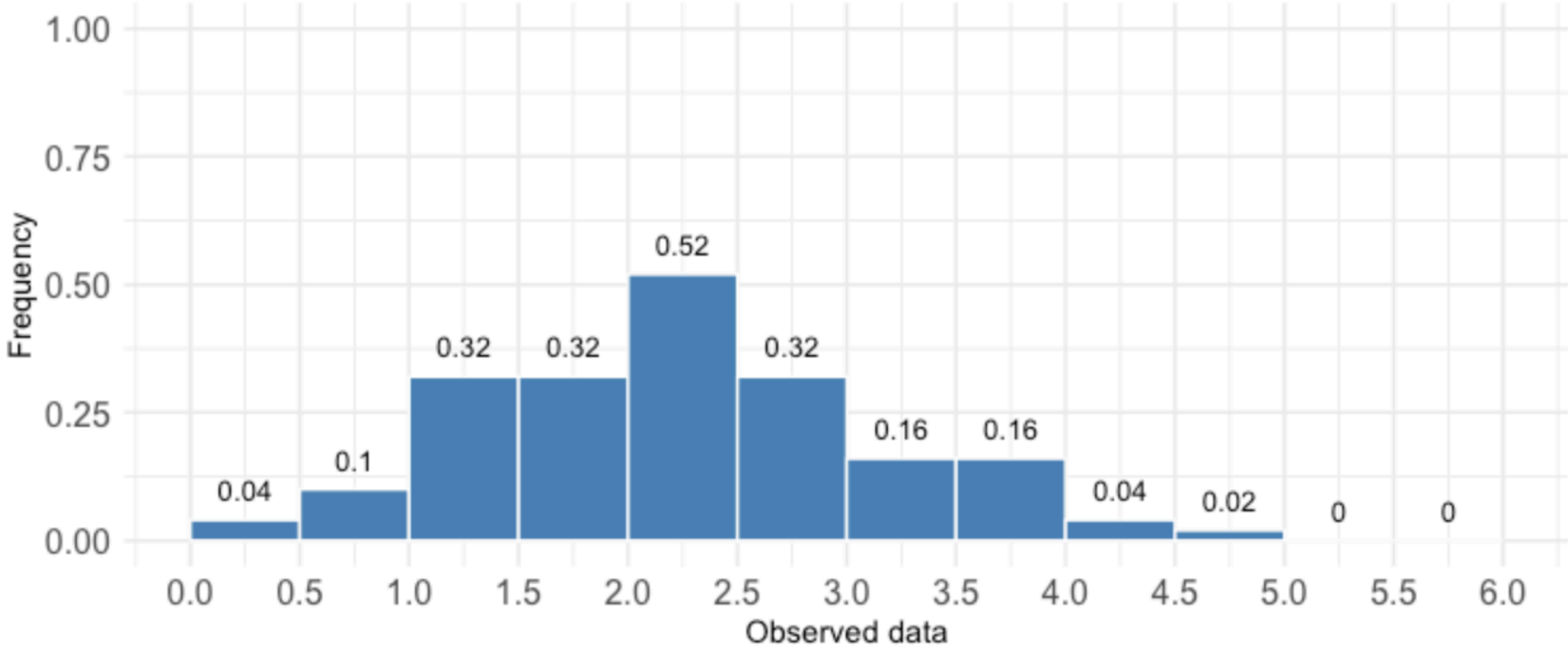
Combine them by averaging their value. We obtain a function of x that estimates the true density.

$$f_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{k}\frac{1}{h}K\left(\frac{x - x_i}{h}\right).$$

Take a look at the R examples posted on Quercus.

Example

Approximating numerical summaries from a histogram by visual inspection (e.g., percentiles)



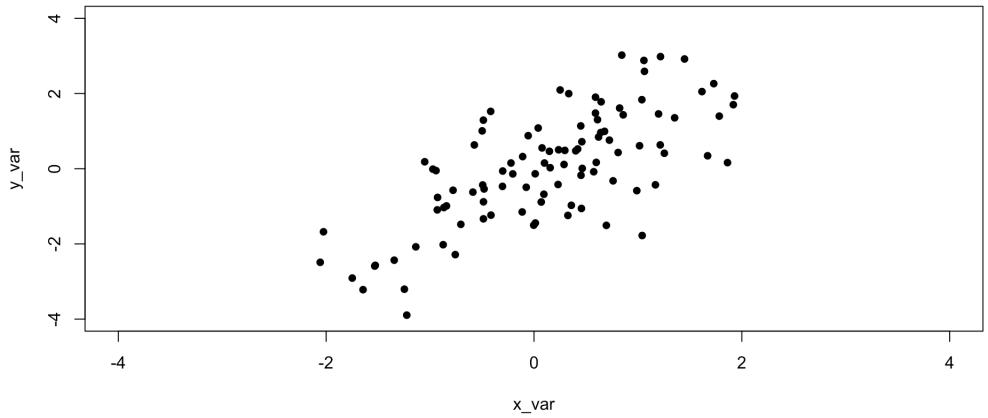
Scatterplots

Visualization for bivariate data

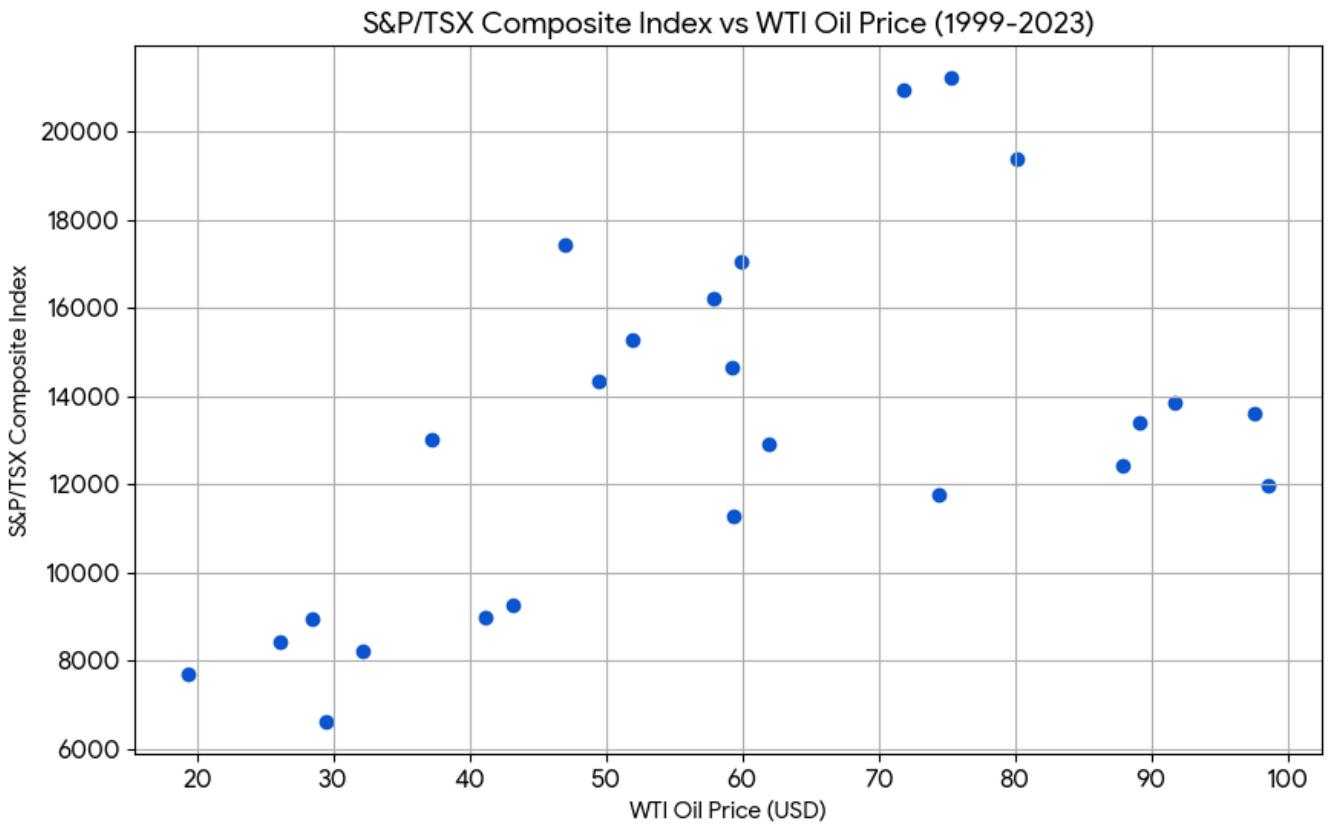
Constructing a scatterplot in R:

Suppose we have two different sets of data that are *paired* $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, and we wish to compare them. A natural way is to plot them at the same time:

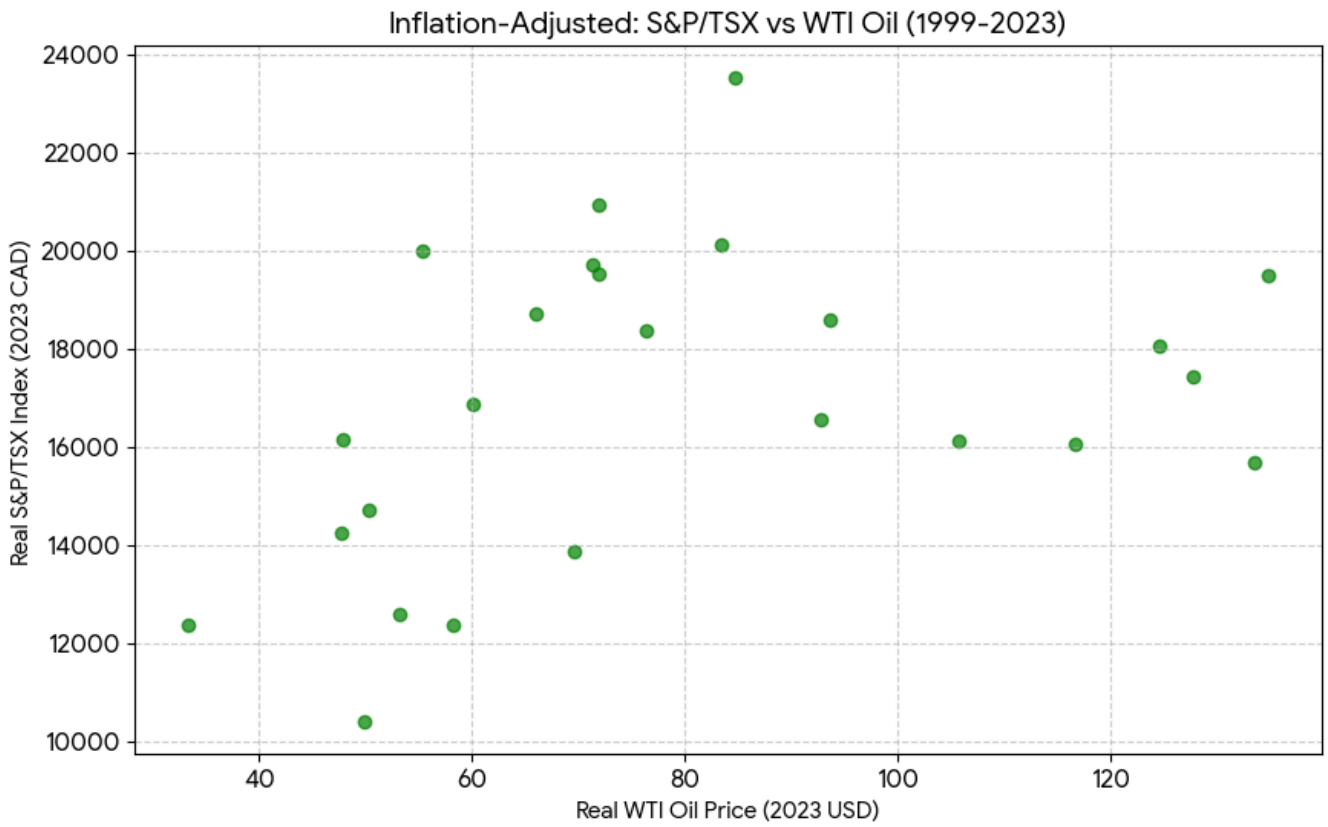
```r
set.seed(238)
x_var <- rnorm(100, mean = 0, sd = 1)
y_var <- x_var + rnorm(100, mean = 0, sd = 1)

# Create the scatterplot
plot(x_var, y_var, xlim = c(-4,4), ylim = c(-4,4), pch = 16)
```



Scatterplot example I



Scatterplot example II



Sample covariance

Our previous numerical summary statistics were for one-dimensional data. For data points $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ we now introduce a two-dimensional statistic, the sample covariance:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

> If we let $Y_i = X_i$, then we have the sample covariance of a dataset with itself, and $S_{XX} = S_X^2$ so it is equal to the sample variance!

Pearson's correlation

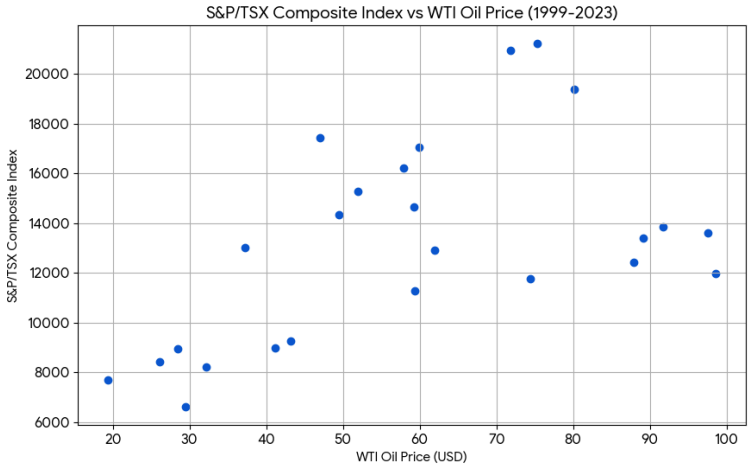
The issue with the covariance is that it does not take into account the scale of the two datasets, which may be very different. In the Oil vs TSX stocks example: - The price of oil had an approximate range of $\approx [20, 100]$ - The stock index had an approximate range of $\approx [6000, 26000]$

Pearson's correlation r_{XY} aims to fix this by normalizing the covariance:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}.$$

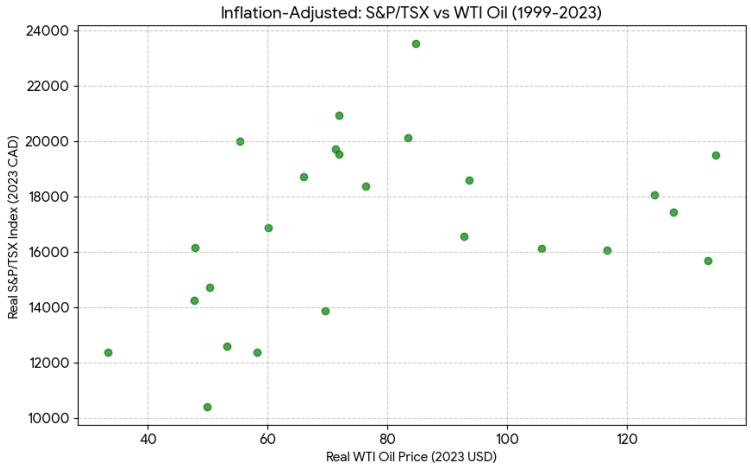
This will tell us the "strength of association" between the two datasets. It is always the case that $-1 \leq r_{XY} \leq 1$.

Scatterplots and correlation



Scatterplot showing Oil prices against Toronto Stock Index

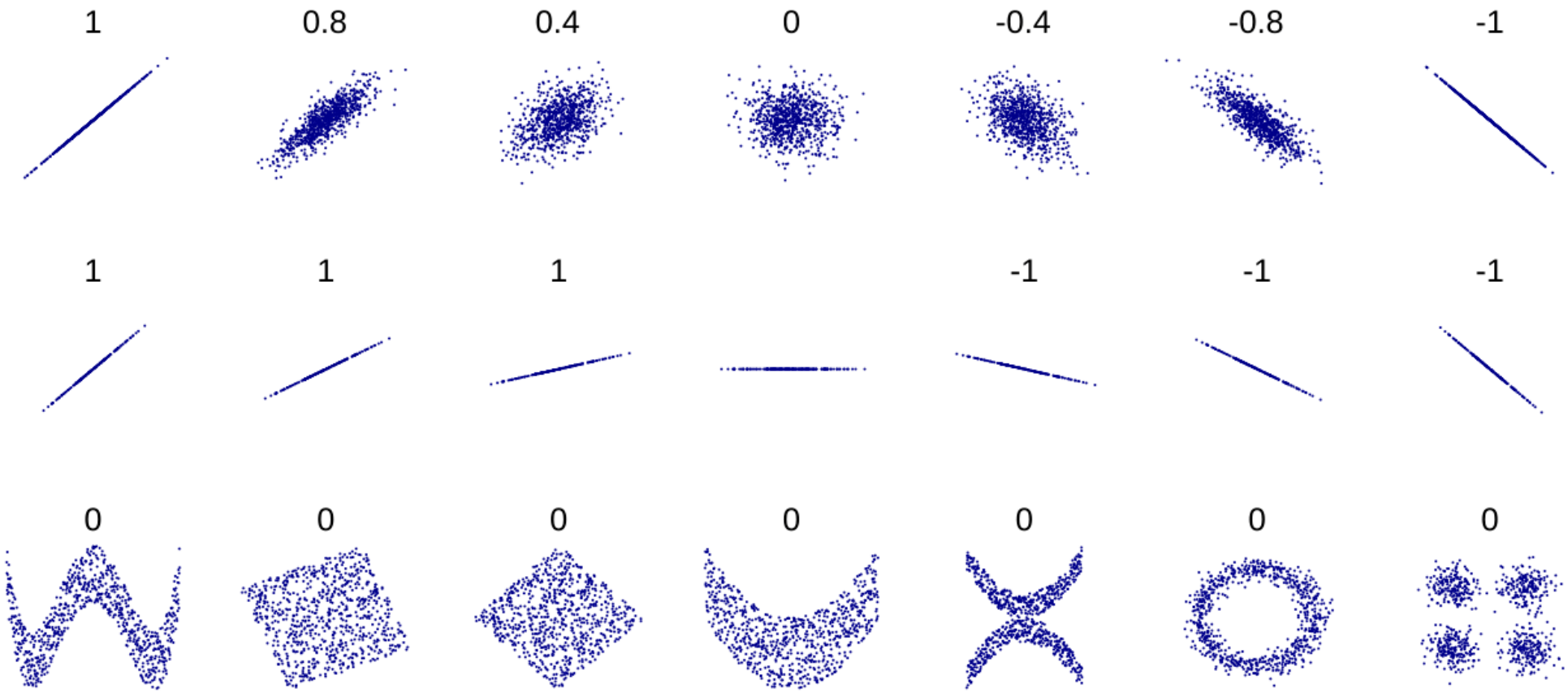
- Covariance: $S_{xy} = 50,403.15$
- Correlation: $r_{xy} = 0.524$



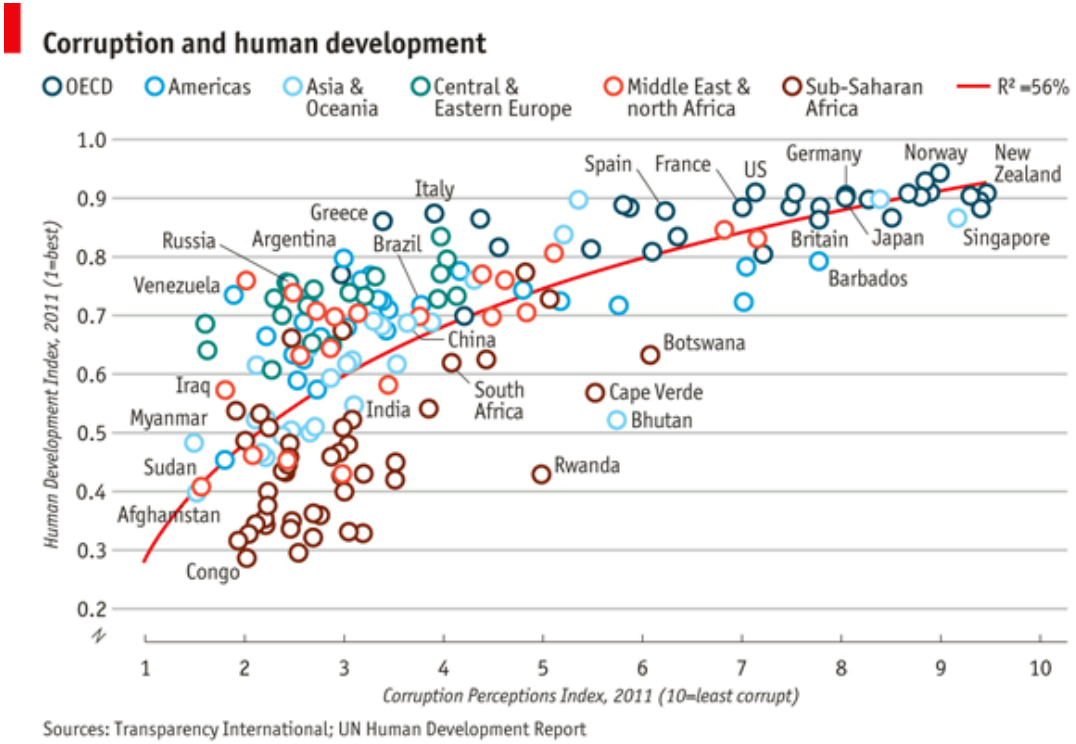
Scatterplot showing inflation adjusted Oil prices against Toronto Stock Index

- Covariance: $S_{xy} = 33,650$
- Correlation: $r_{xy} = 0.36$

Correlation Example

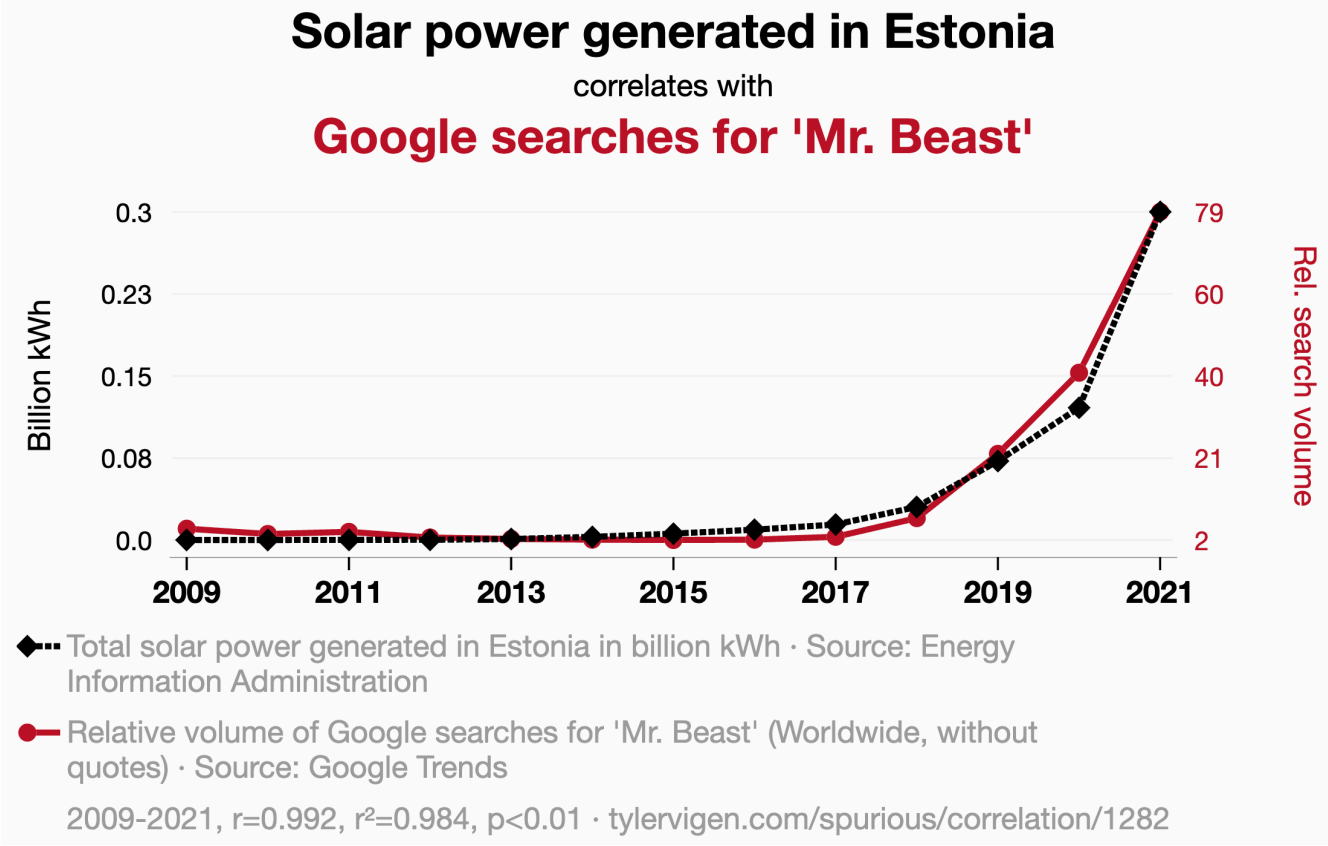


Correlation Example



Another example of a scatterplot showing a corruption index against the Human Development Index in 2011 from this article in The Economist.

Correlation and Causation



Spurious correlation from https://www.tylervigen.com/spurious-correlations

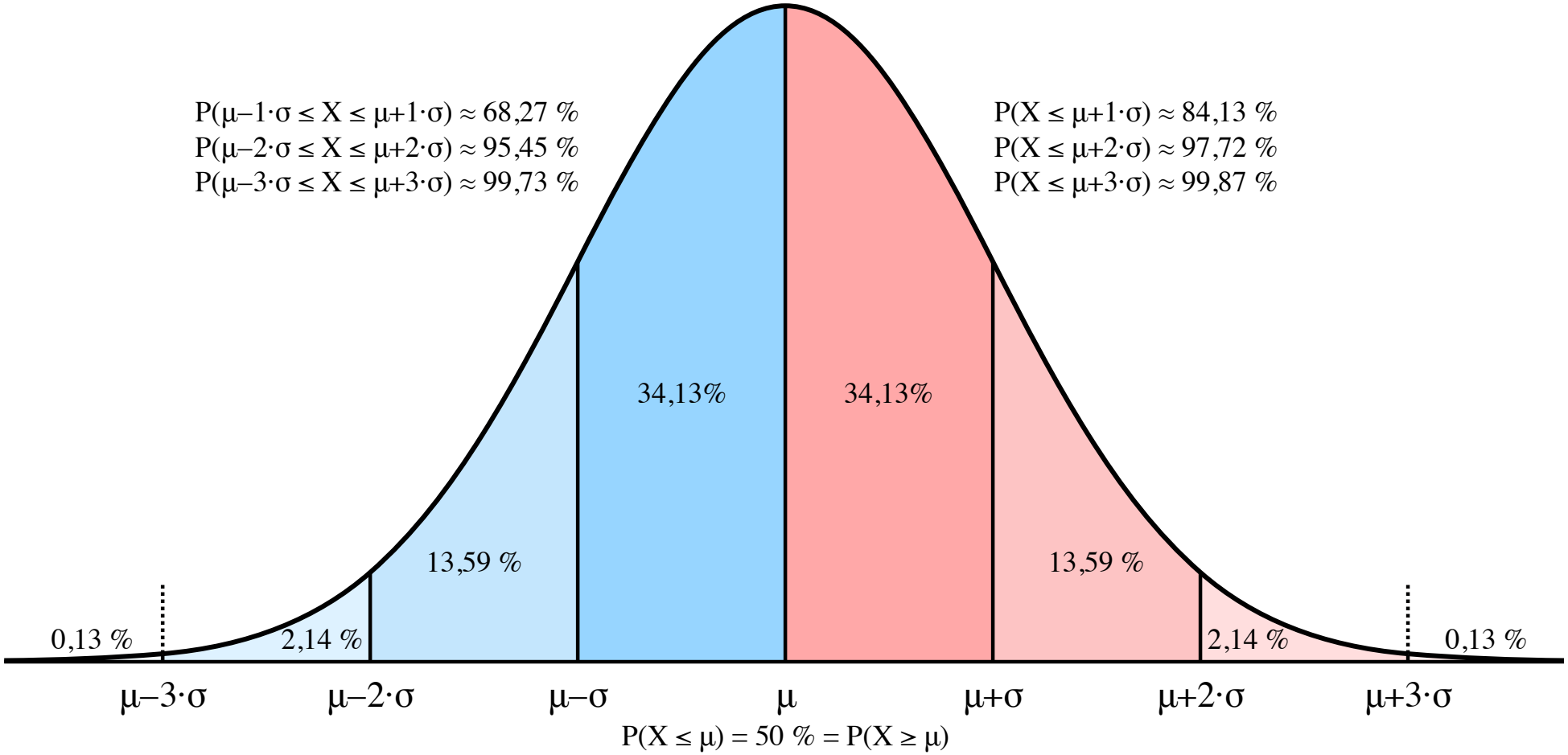
Summary

- Graphical summaries are a tool for visually estimating distributional features.
- Choosing the right tool can support your analytical goals.
- Even for a single tool, important decisions may have to be made by the user for the information to be displayed accurately.
- Correlation is an important measure of association
- A strong correlation (close to ± 1) does not exclude the possibility that there may not be a causal relationship between two variables

Next up: QQ Plots

A Quantile-Quantile plot is used to determine if a dataset follows a specific theoretical probability distribution (commonly the Normal distribution) or if two different datasets share the same distribution.

It works by plotting the quantiles of the data against the quantiles of the theoretical distribution.



QQ Plots Example