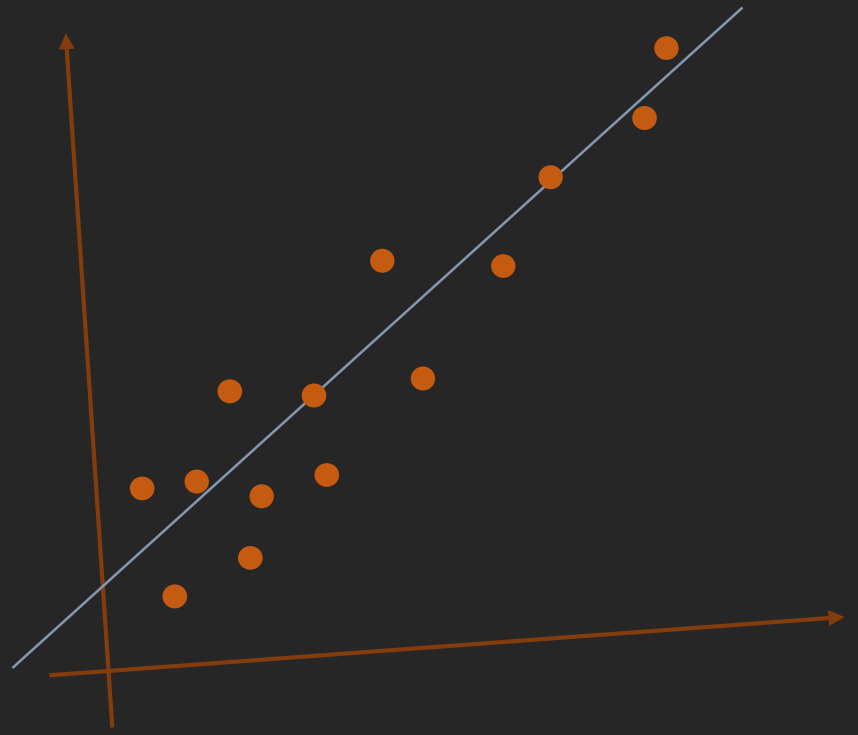
Applied Regression Analysis

STAT 4043/ STAT 5543



Correlation Analysis

Pratyaydipta Rudra

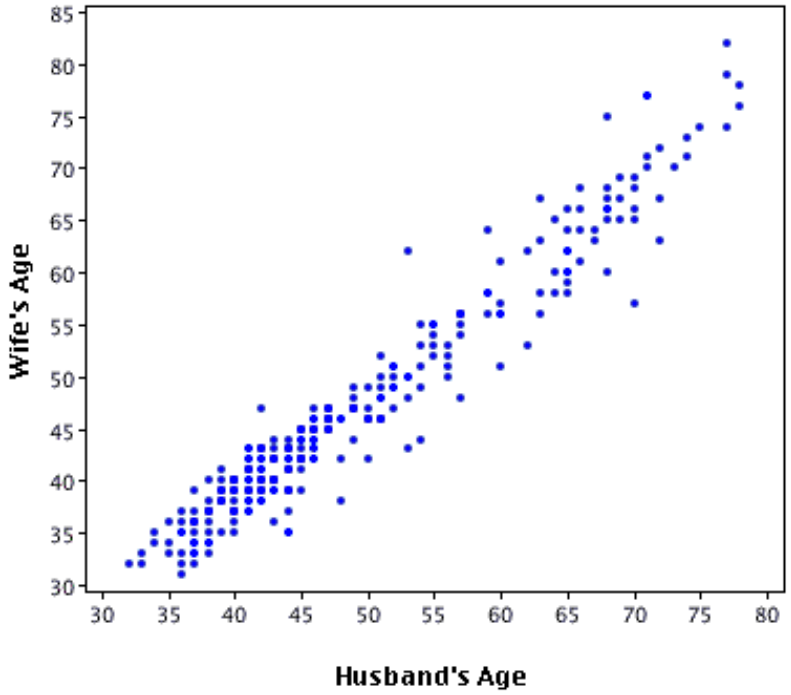
Correlation Analysis

Correlation vs Regression

- Correlation:
  - How linear is the relationship of two variables? (descriptive)
  - Is there statistically significant (linear) association between the two variables?

- Simple Linear Regression (SLR):
  - How good is a linear model to explain my data?
  - $y = \beta_0 + \beta_1 x + \epsilon$ (β_0 = intercept, β_1 = slope)
  - Does a predictor (independent, x) variable cause the outcome (dependent, y) variable to change?
  - Can use multiple regression techniques to study effects of other variables.
  - Prediction.

Covariance



Scatter plot

- Do the two variables vary together?
  - Do the variables increase or decrease together?
  - Does one decrease when the other increases?

Source: http://onlinestatbook.com/

Formula

- $\bar{X} = \frac{\sum X_i}{n}$: Sample mean of x
- $\bar{Y} = \frac{\sum Y_i}{n}$: Sample mean of y
- $Var(X) = \frac{\sum (X_i - \bar{X})^2}{n-1}$: Sample variance of x
- $Var(Y) = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$: Sample variance of y

- $Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$: Sample covariance of x and y

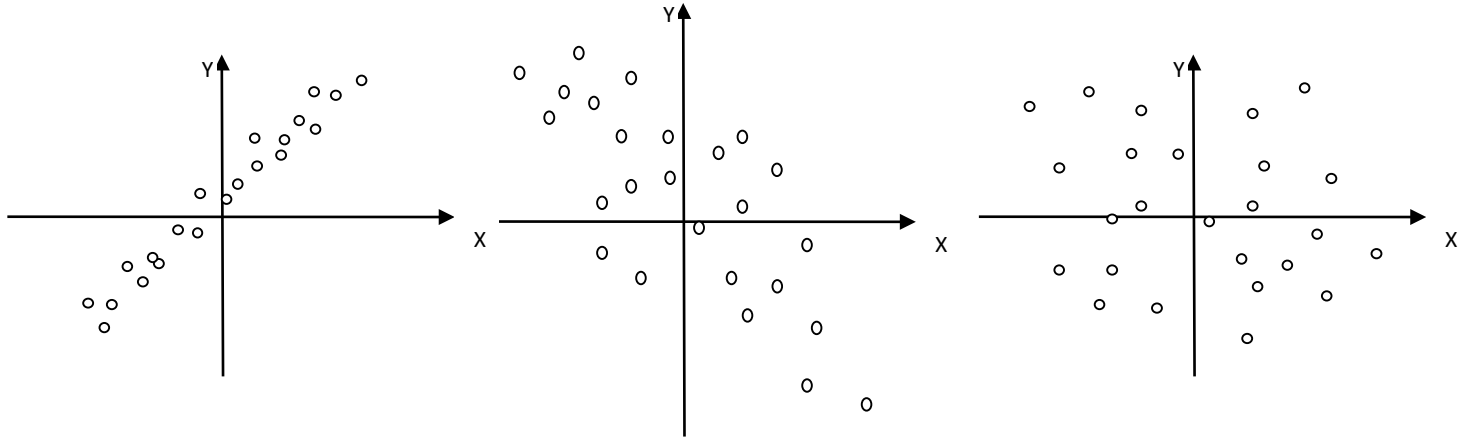
Population versions of the same things

- Population mean of X: $E(X)$
- Population variance of X: $Var(X) = E\big(X - E(X)\big)^2$
- Population covariance of X and Y:

$$\mathrm{Cov}(\mathrm{X}, \mathrm{Y}) = E[(X - E(X))(Y - E(Y))]$$

Correlation

- Correlation measures the magnitude of linear association.
- The tighter the data points are around a line, the higher the correlation.



high positive correlation

weaker negative correlation

no correlation

- When X and Y have positive covariance , they are also positively correlated.
- When X and Y have negative covariance, they are also negatively correlated.

Measure of correlation

- Measure of correlation (scaled version of covariance):

$$r = \frac{Cov(X,Y)}{SD(X)SD(Y)}$$

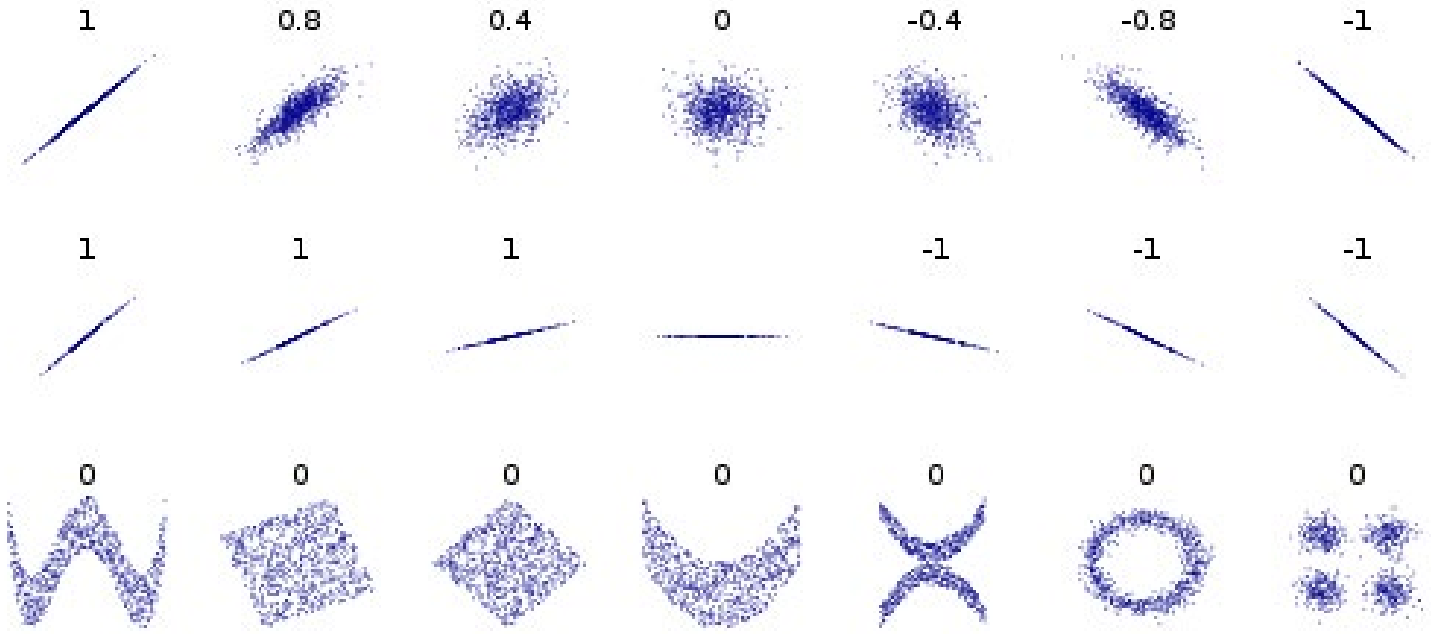
r is called **(Pearson) Correlation Coefficient**.
This is the sample version.

The population version ρ can be obtained similarly by using the population covariance and standard deviations.

Properties of correlation coefficient

- It is unitless.
- $-1 \leq r \leq 1.$
  - r=0: No correlation, r=1 or -1, perfect correlation.
- Symmetric. $r(x, y)$ and $r(y, x)$ are the same.
- The magnitude doesn't change for linear transformation, but the sign might change.
  - $u = a + bx, v = c + dy$
  - $r(u, v) = r(x, y) \times sign(bd)$
- Example: $r(x, y) = 0.45, u = 2 - 3x, v = 4y, r(u, v) =?$

Zero correlation does not imply no association



Source: wikipedia

Hypothesis testing involving correlation coefficient

Hypothesis testing involving population correlation coefficient

- Question: Is there a significant association between height and weight?
- $H_0: \rho = 0$ vs $H_1: \rho \neq 0$.
- This is done using a t-test.
  - What is "a t-test" by the way?
  - Note that we must assume x and y follow a **bivariate normal distribution** – We will see what that is.
- The relevant R functions for correlation analysis:

  `cor()` and `cor.test()`

Bivariate normal distribution

- Suppose we have two random variables following normal distributions and they are not independent, they are jointly distributed.
- Such distribution is a bivariate normal distribution. It has a mean vector and a variance covariance matrix.
- Example: The random vector $(X_1, X_2) \sim N_2\left((0, 0.5), \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix}\right)$

$$E(X_1) = 0, E(X_2) = 0.5, Var(X_1) = 1, Var(X_2) = 4,$$
$$Cov(X_1, X_2) = 1.5$$

Bivariate normal distribution

- The mean vector and the variance covariance matrix uniquely define the joint distribution of X_1 and X_2.
- We can also obtain the correlation matrix from the variance covariance matrix.

$$\begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}$$

- Note: The variance covariance and correlation matrix, both must be symmetric.
- We have a multivariate normal distribution for any number of jointly distributed random variables. Bivariate normal is a special case.

Some things to be careful about

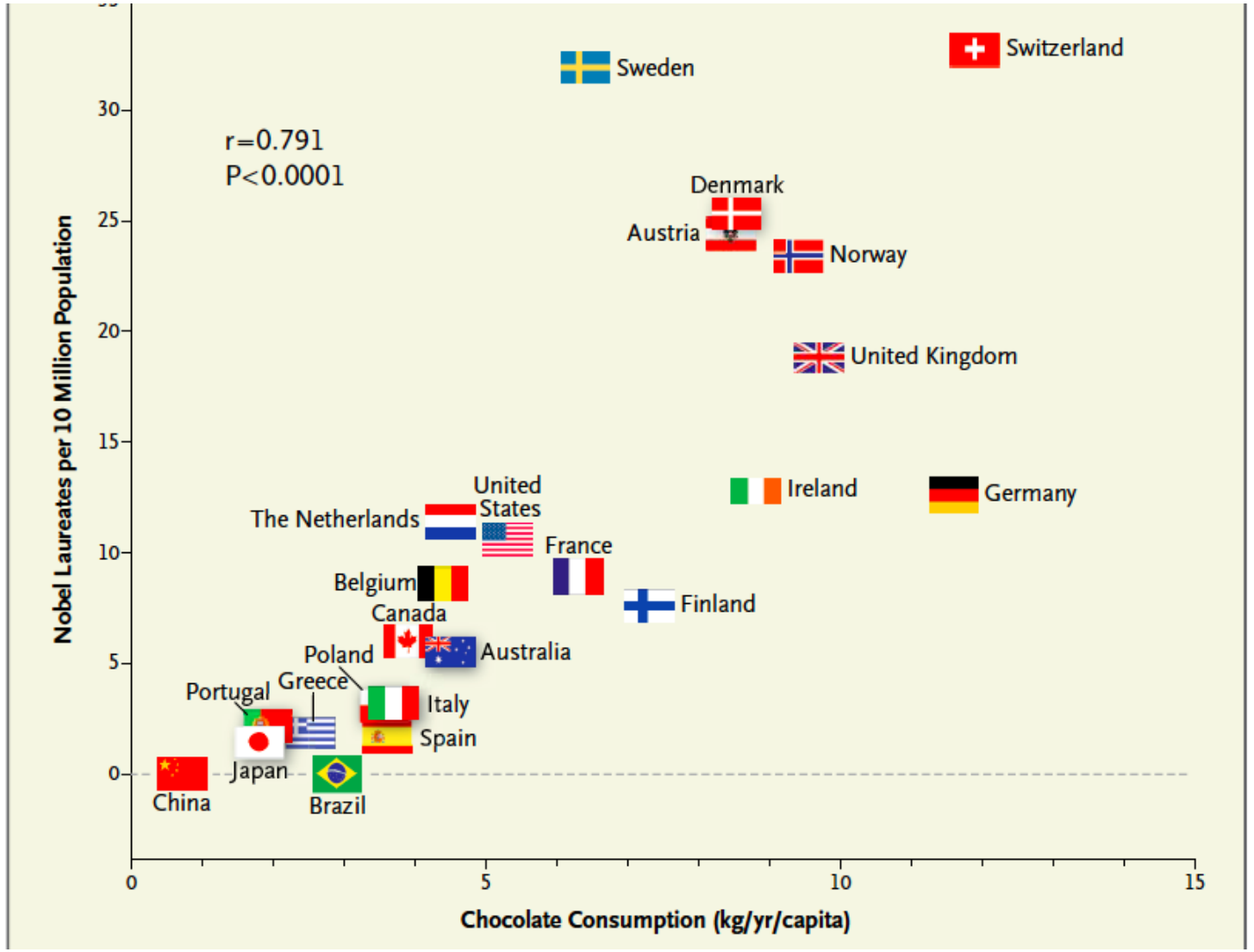
Correlation and independence

- Independence occurs when knowing the value of one variable does not provide any extra information about the value of the other variable.
- Independent implies zero correlation (population correlation).
- But uncorrelated does not imply independent. There might be other non-linear relationships.
- What to do for non-linear relationships? Use other measures. More on that later.

What may be going on when 2 variables are correlated?

- Direct cause and effect
  - overexpression of gene A causes cancer.
- Both cause and effect
  - Coffee consumption causes nervousness and nervous people drink more coffee.
- Relationship caused by a third variable
  - Drinking alcohol and lung cancer. Both are related to cigarette smoking.
  - Smoking is a *confounder* in the relationship of alcohol to lung cancer.
  - For this, we will learn a technique called 'partial correlation' later.
- Coincidental relationship (spurious correlation)

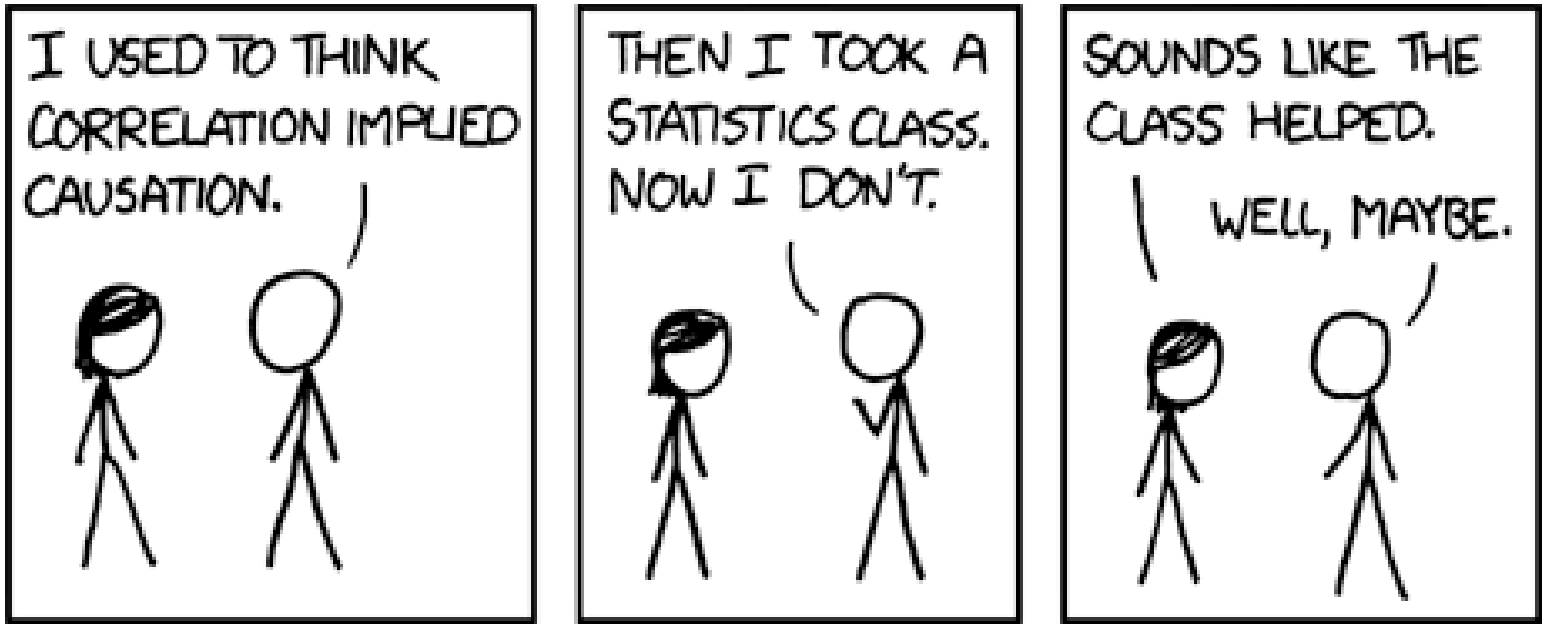
Correlation does not imply causation



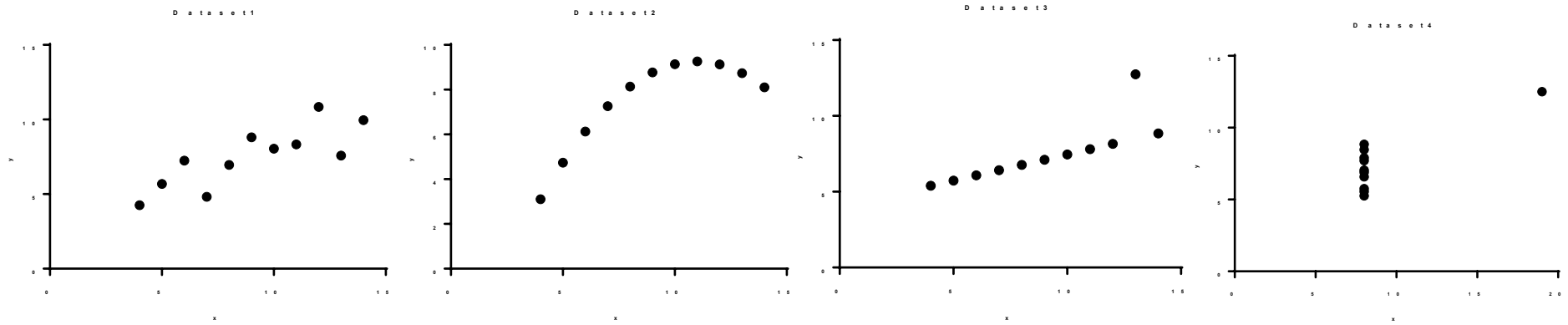
Spurious correlation

Messerli FH N Engl J Med. 2012 Oct 18;367(16):1562-4.

Correlation does not imply causation



Importance of visualization



Correlation coefficient measured from all the four data sets are equal (0.816)

Anscombe's quartet

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. **27** (1): 17–21.

Summary

- What is the difference between correlation and regression?
- Which of correlation and regression is related to the slope of the linear pattern in the data?
- What are two measures of correlation that we learned today?
- What does zero correlation imply?
- What does a high correlation imply? Can you say one variable causes the other?
- Importance of visualization.