


**DATA SCIENCE**  
PPGIa/PUCPR

Prof. Jean Paul Barddal



1

**CORRELATIONS**

2

### Variables that are related

- Let's analyze changes in a variable and how it impacts other variables
- Studying correlations allows us to analyze two or more variables together and to quantify whether:
  - The relation is direct or inverse;
  - The relation is strong or weak.

3

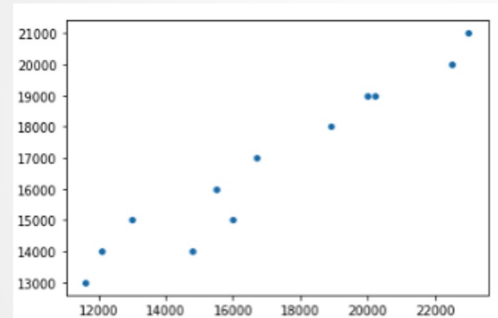
### Example

- The table depicts monthly juice production rates and its cost

Month	Production (l)	Total cost (R\$)
Jan	20200	19000
Feb	16700	17000
Mar	14800	14000
Apr	16000	15000
May	12100	14000
Jun	13000	15000
Jul	11600	13000
Aug	15500	16000
Sep	18900	18000
Oct	20000	19000
Nov	22500	20000
Dec	23000	21000

4

## Scatterplots



- In a cartesian system, these plots allow us to see the correlation between variables

5

## Pearson correlation coefficient

- Quantifies the correlation between a pair of numeric variables
- The computation of the pearson correlation coefficient  $r$  for variables  $X$  and  $Y$  is as follows:

$$r(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{cov(X, Y)}{\sqrt{var(X) \times var(Y)}}$$

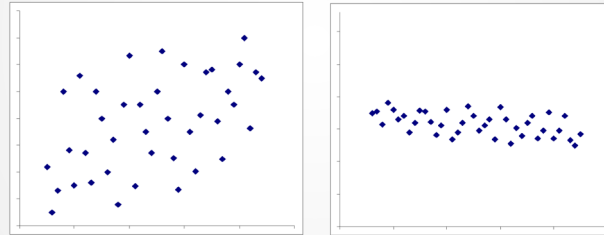
- $r$  lies between  $[-1; +1]$

6



## Graphical analysis

### NULL OR ABSENT CORRELATION



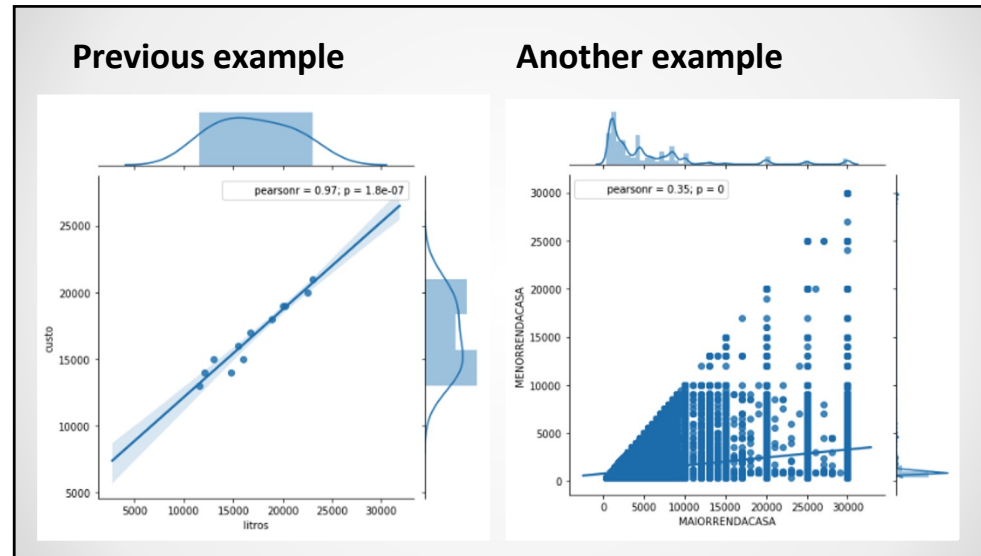
9

## (not so) formal definition

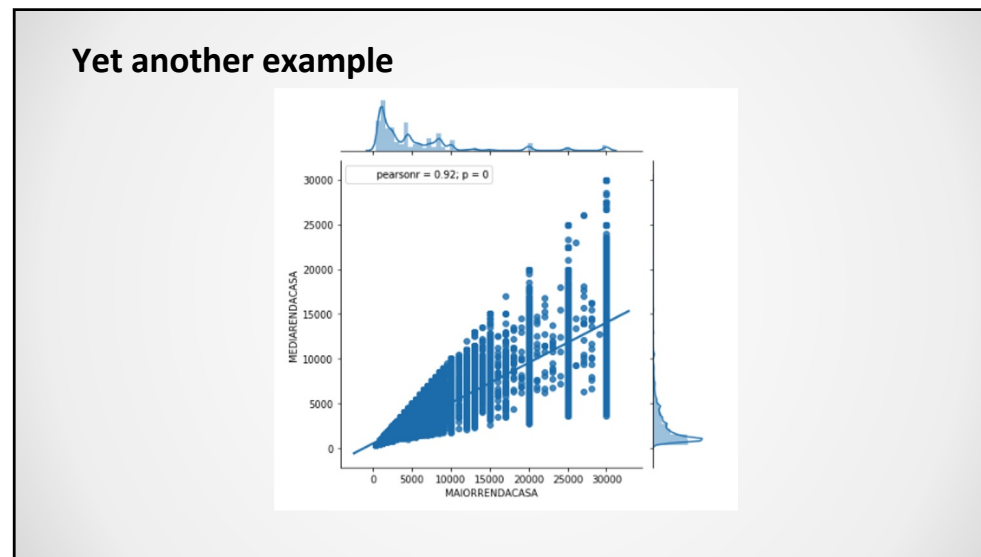
$ r $ (modulus)	Interpretation
$ r  < 0.4$	Weak correlation
$0.4 \leq  r  < 0.7$	Mild correlation
$0.7 \leq  r $	Strong correlation

There is no consensus on these thresholds.  
Each area may assume different values.

10



11



12

### **Pearson correlation**

- Apply this to numeric data
- It also assumes that data:
  - Follow a gaussian distribution
  - That the correlation between variables is linear

13

**ACTIVITY**

14

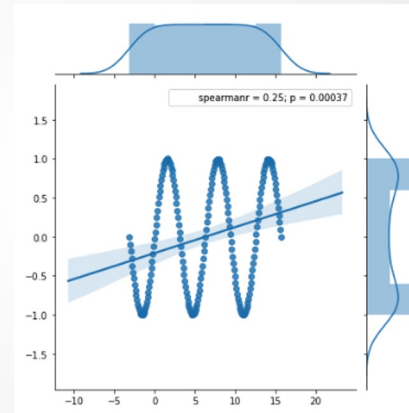
### Activity

- Create two pandas data frames, each with two variables x and y
- In the first data frame, make sure that y is given by  $n * x$ , where n is a value you can choose
- In the second data frame, make sure that  **$y = \sin(x)$**
- Compute the correlation between x and y for both dataframes

15

### Example

Here, feature x and y are correlated, but not linearly



16



## SPEARMAN CORRELATION

17

### Spearman correlation

- Should be used when variables are **ordinal**
- Spearman's  $\rho$ 
  - Data is sorted from the largest to the smallest value per variable
  - $\rho$  is computed the same way as Pearson's  $r$ , yet, assuming the position of each data point
- Or simply use **scipy.stats.spearmanr**

18

## CORRELATION BETWEEN CATEGORICAL VARIABLES

19

### Cross-tabulating

- A nice way to check whether two categorical variables are correlated is to cross-tabulate them
- Using `pd.crosstab`, we can check the behavior of two variables
- For a more visual approach, we can use a heatmap to check the correlation

20

Example

- Let's take a quick look at this approach using the titanic data

21

**CORRELATION AND CAUSATION**

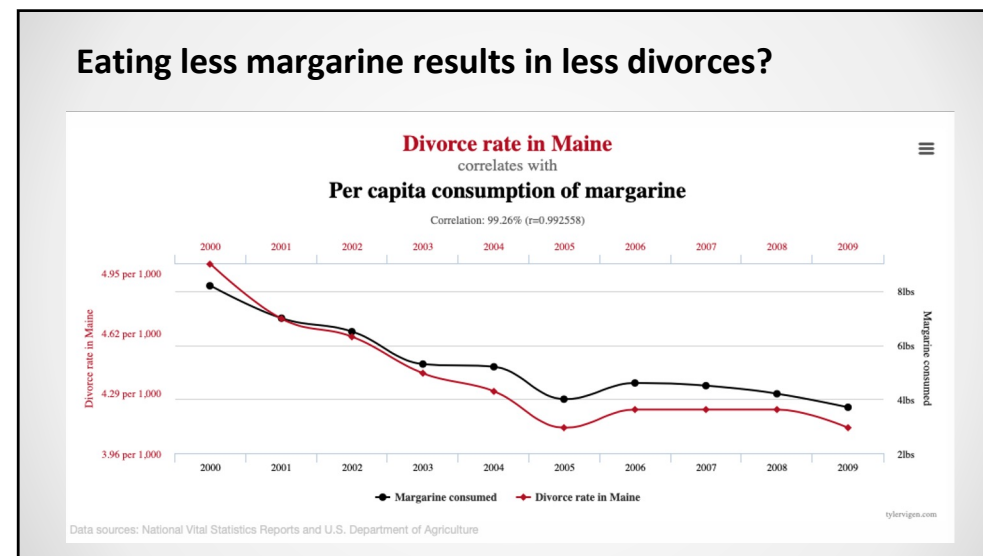
25

## Correlation and causation

- Sometimes we will find correlations and we will assume that a variable A is causing variable B to have a certain behavior
- That **may** be true, but not necessarily

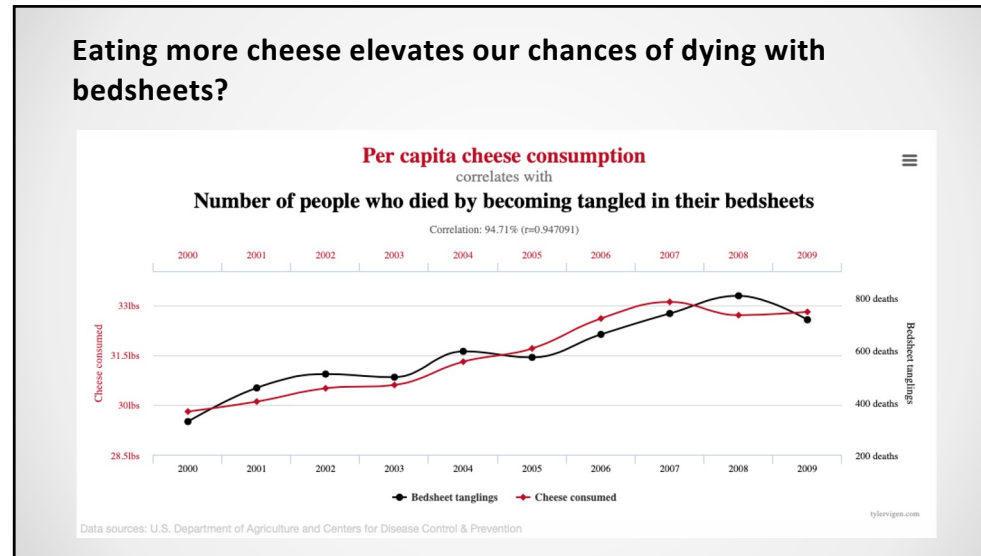
26

## Eating less margarine results in less divorces?



27

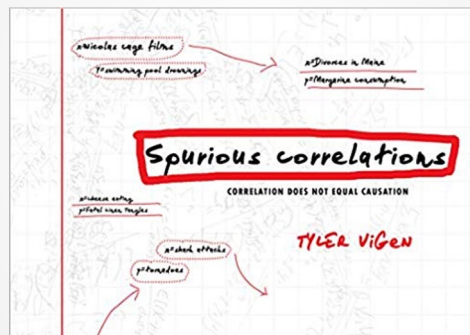
## Eating more cheese elevates our chances of dying with bedsheets?



28

## Spurious Correlations

- <https://www.tylervigen.com/spurious-correlations>



29

## ACTIVITY

35

### Activity

- You have until the **end of the day** to find a dataset that you are interested on working with
- The dataset must be in tabular shape and it must:
- Have at least 10 features (do not pick a large dataset if you are starting on data science projects)
- At least 100 instances
- Avoid datasets with text data (unless you have a background to work with it, stick with categorical/numerical features)
- You must send the dataset to [jean.barddal@ppgia.pucpr.br](mailto:jean.barddal@ppgia.pucpr.br)

36