


DATA SCIENCE

PPGIa/PUCPR

Prof. Jean Paul Barddal



1

CORRELATIONS

2

Variables that are related to one another

- Correlation analysis is a way to analyze the relationship between two or more variables. The goal is to quantify and assess whether:
 - The relation is direct or inverse;
 - Strong or weak.
- **Important:** correlation does **NOT** imply causation.

3

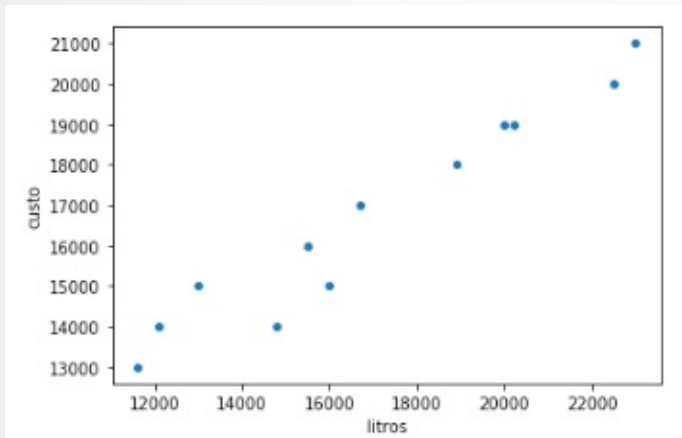
Example

How can we determine whether production is correlated with the cost?

Month	Production (I)	Total cost (R\$)
Jan	20200	19000
Fev	16700	17000
Mar	14800	14000
Abr	16000	15000
Mai	12100	14000
Jun	13000	15000
Jul	11600	13000
Ago	15500	16000
Set	18900	18000
Out	20000	19000
Nov	22500	20000
Dez	23000	21000

4

Scatterplot



- Points in a cartesian system, which allow a quick overview of the data and potential correlation

5

Pearson correlation

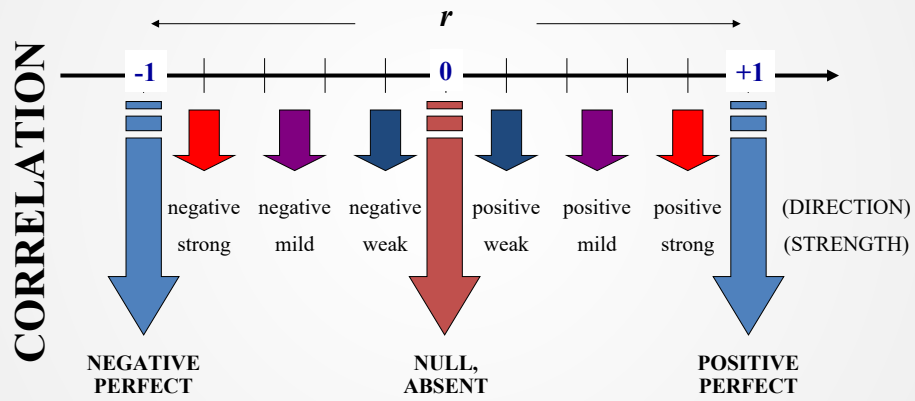
- Quantifies the correlation amongst variables
- r between variables X and Y is given by

$$r = \frac{\text{Cov}(X,Y)}{\sqrt{S_x^2 \cdot S_y^2}}$$

- r is bounded in $[-1, +1]$

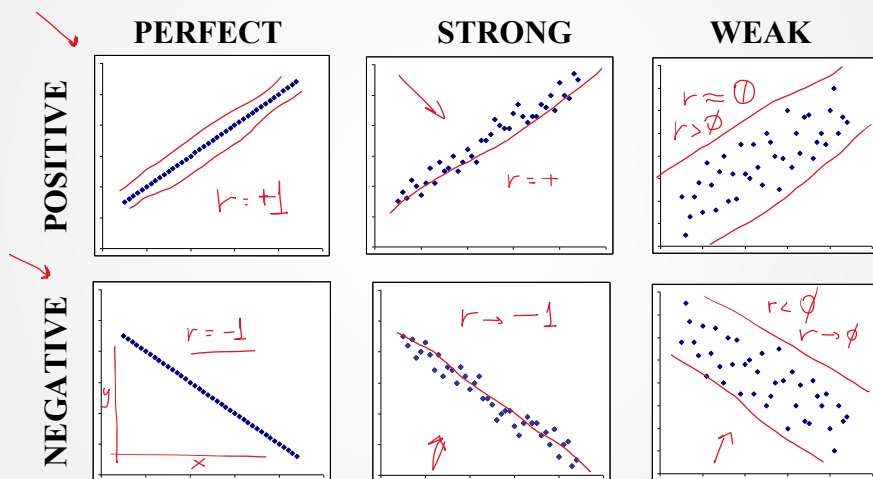
6

Understanding Pearson's coefficient



7

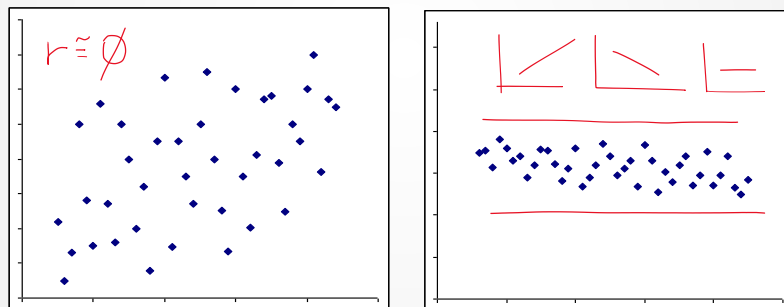
Visual analysis



8

Visual analysis

NULL OR ABSENT CORRELATION



9

(not so) formal definition

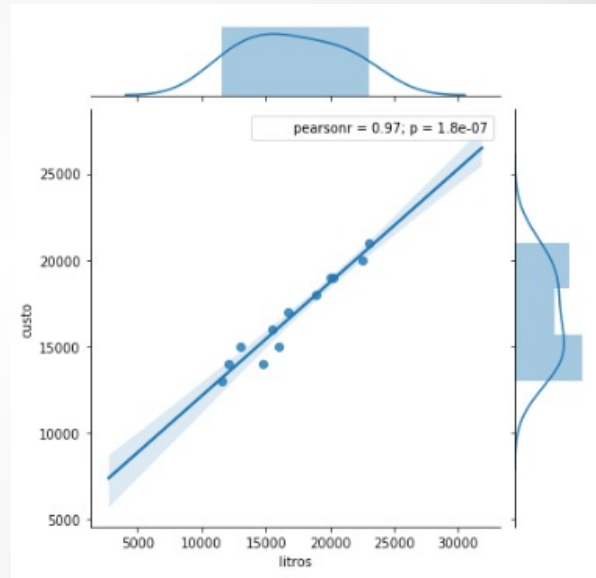
$ r $ (modulus)	Interpretation
$ r < 0.4$	Weak correlation
$0.4 \leq r < 0.7$	Mild correlation
$0.7 \leq r $	Strong correlation

Note that there is no consensus on these threshold.
Different areas assume different values

10

Previous example

Let's code the previous example of production amount and costs



11

SPEARMAN'S CORRELATION

12

Spearman's correlation

- Should be used when
 - We have ordinal data
- Spearman's ρ (rhô)
 - Elements are sorted from the most positive to the most negative
 - ρ is computed using Pearson's equation, taking into account the ranking of each observation

13

**(COR-)RELATION BETWEEN
CATEGORICAL VARIABLES**

14

Cross-tabulating

- A simple approach to check the behavior of categorical variables is to use cross-tables (contingency tables)
- Using **pd.crosstab** and **heatmaps** allow us to quickly identify interesting behavior in data

15

Example

- Let's analyze an example using the titanic dataset

16

CORRELATION AND CAUSATION

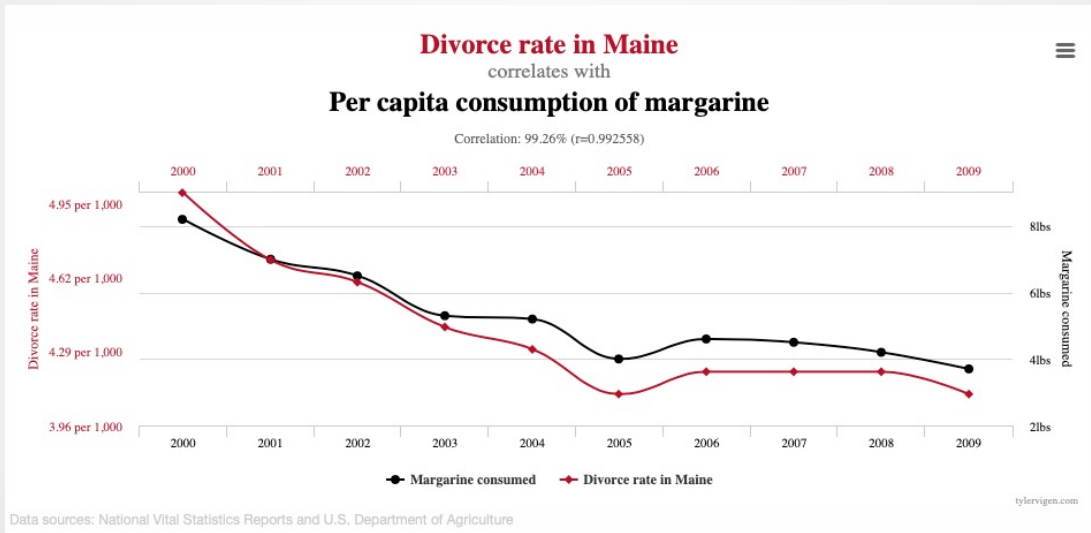
21

Correlation and causation

- Several times, we will observe a correlation and assume that one variable is leading the other
- This **may** be true, but not necessarily

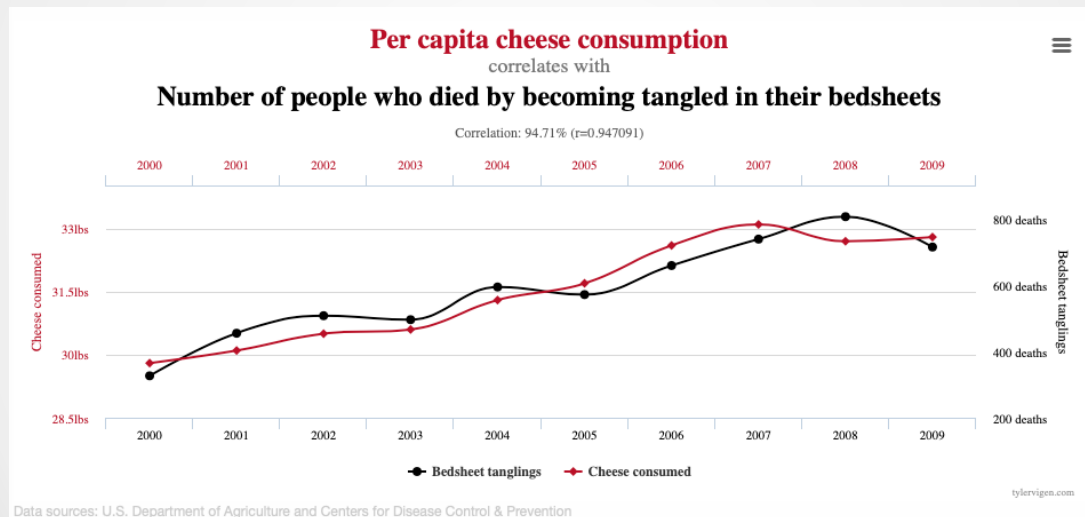
22

Eating margarine is bad for your marriage?



23

Eating more cheese increases our chances in dying tangled in our bedsheets?



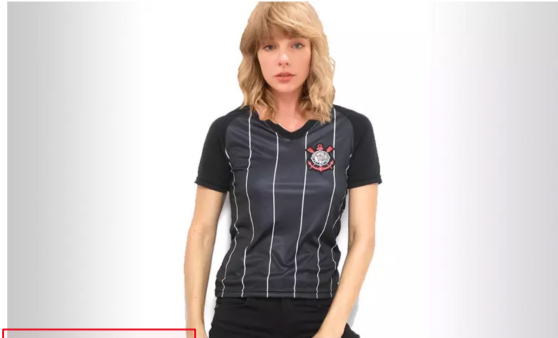
24

Taylor Swift and Corinthians

Taylor Swift é corintiana? Time nunca perdeu quando cantora lança álbum

Anna Satie, da CNN em São Paulo
11 de dezembro de 2020 às 15:33 | Atualizado 11 de dezembro de 2020 às 15:43

Ouvir



Será que Taylor Swift é corintiana?
Foto: CNN Brasil/Reprodução/Instagram

O lançamento surpresa de Evermore, nono álbum de Taylor Swift, nesta sexta-feira (11) foi motivo de alegria para os torcedores do Corinthians.

O porquê? O time nunca perdeu uma partida imediatamente antes ou depois da cantora apresentar um novo CD. Por um terço das vezes, o alvinegro paulista empatou, mas a maior parte das partidas terminou em vitória corintiana.

Compartilhar

MAIS DA CNN BRASIL

Após embate com Pazueto, Doris escala técnicos para falar...
O que os episódios do 'O que eu faço?'...
Quem já teve Covid-19 precisará tomar vacina?
Vacina da Covid-19 é contraindicada para menores de 18 e...

MAIS DA CNN BRASIL

Planalto pressionou Pazueto a mudar estratégia da vacina em reunião na segunda

25

Facebook, Cholesterol, and Justin Bieber

Evidence That Facebook Cancelled Out the Cholesterol-Lowering Effects of Justin Bieber

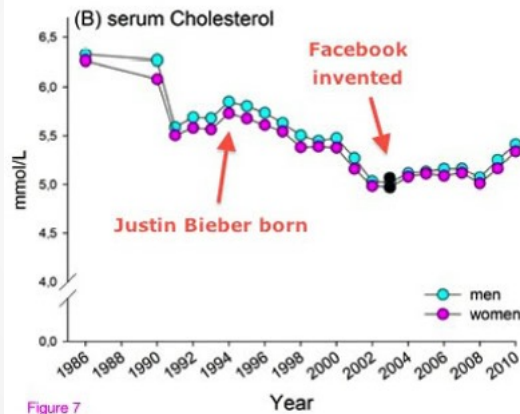


Figure 7

26

More examples: Spurious Correlations

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

