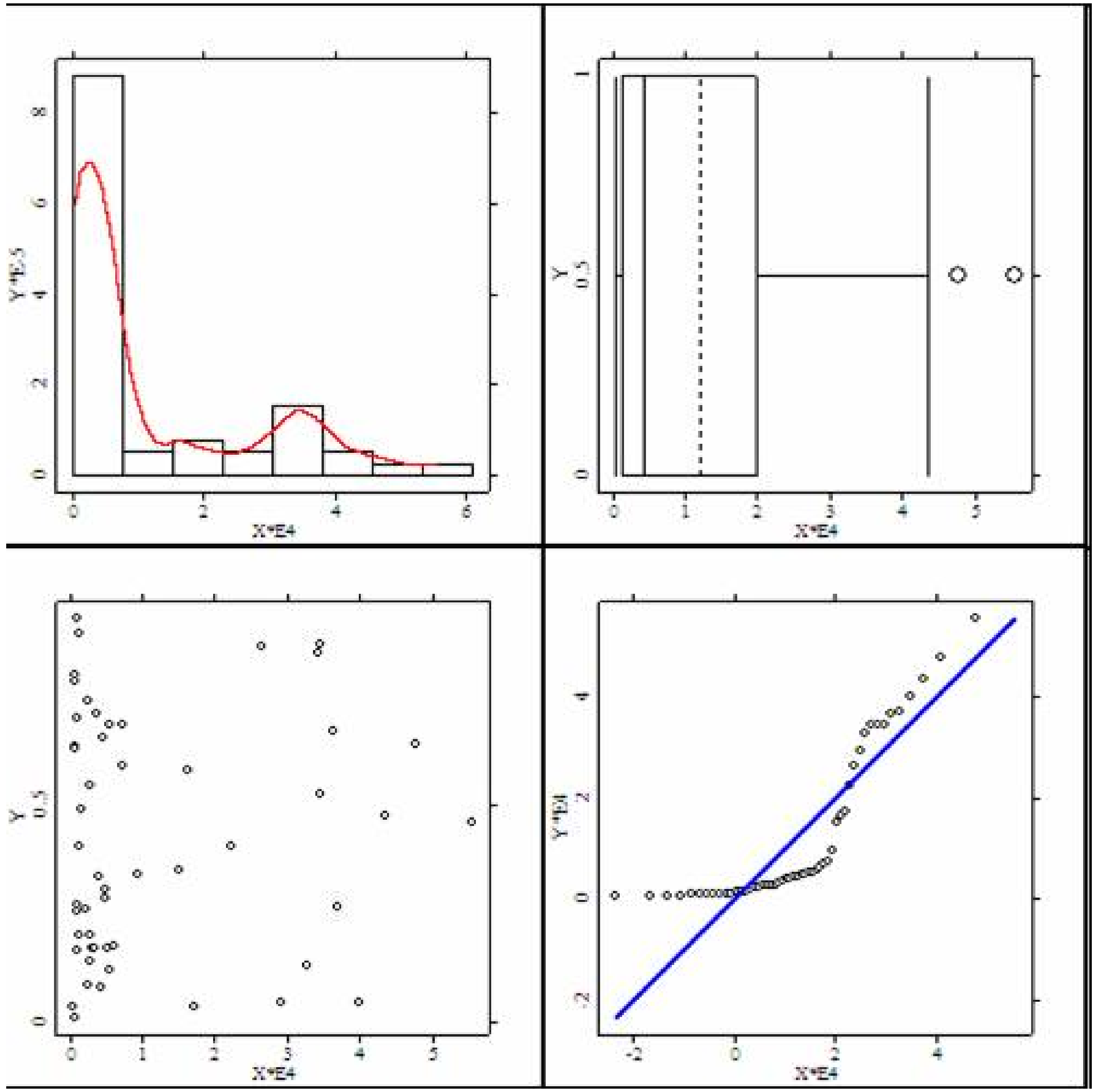
Descriptive statistics

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

What are descriptive statistics?

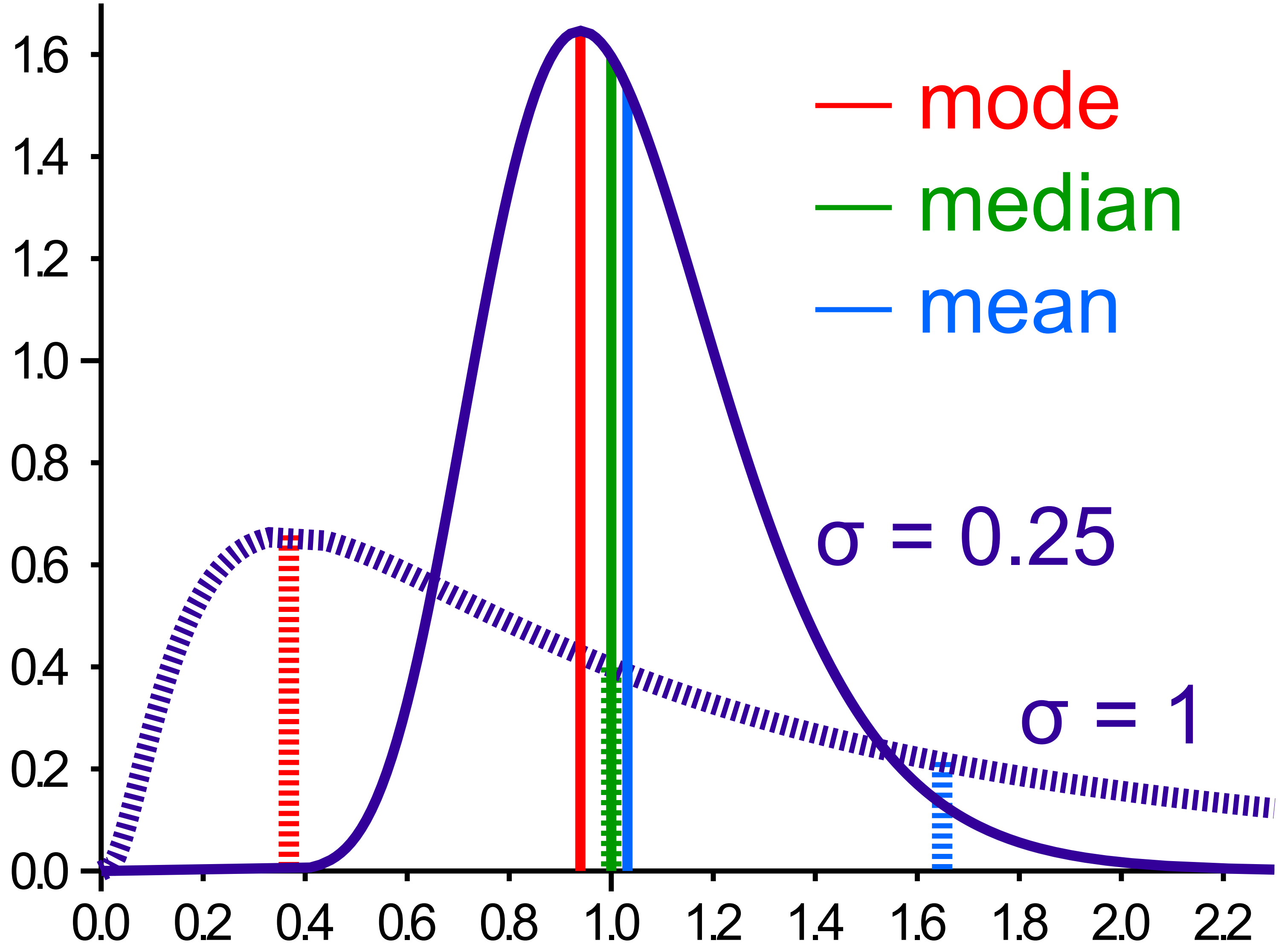


[1] Wikimedia

Measures of centrality

- Mean
- Median
- Mode

Measures of centrality



[1] Wikimedia

Measures of variability

- Variance
- Standard deviation
- Range

Measures of variability

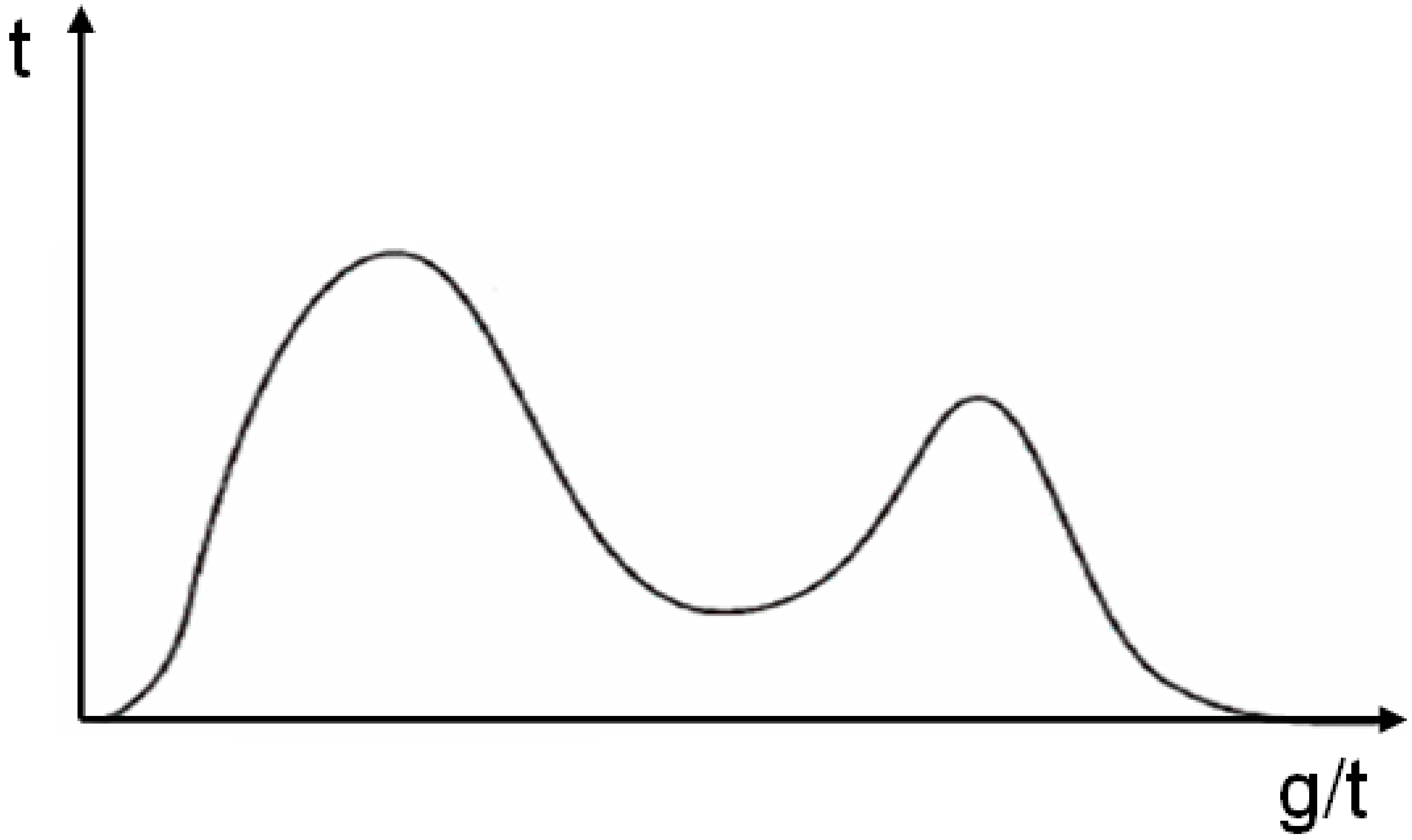
Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Standard Deviation

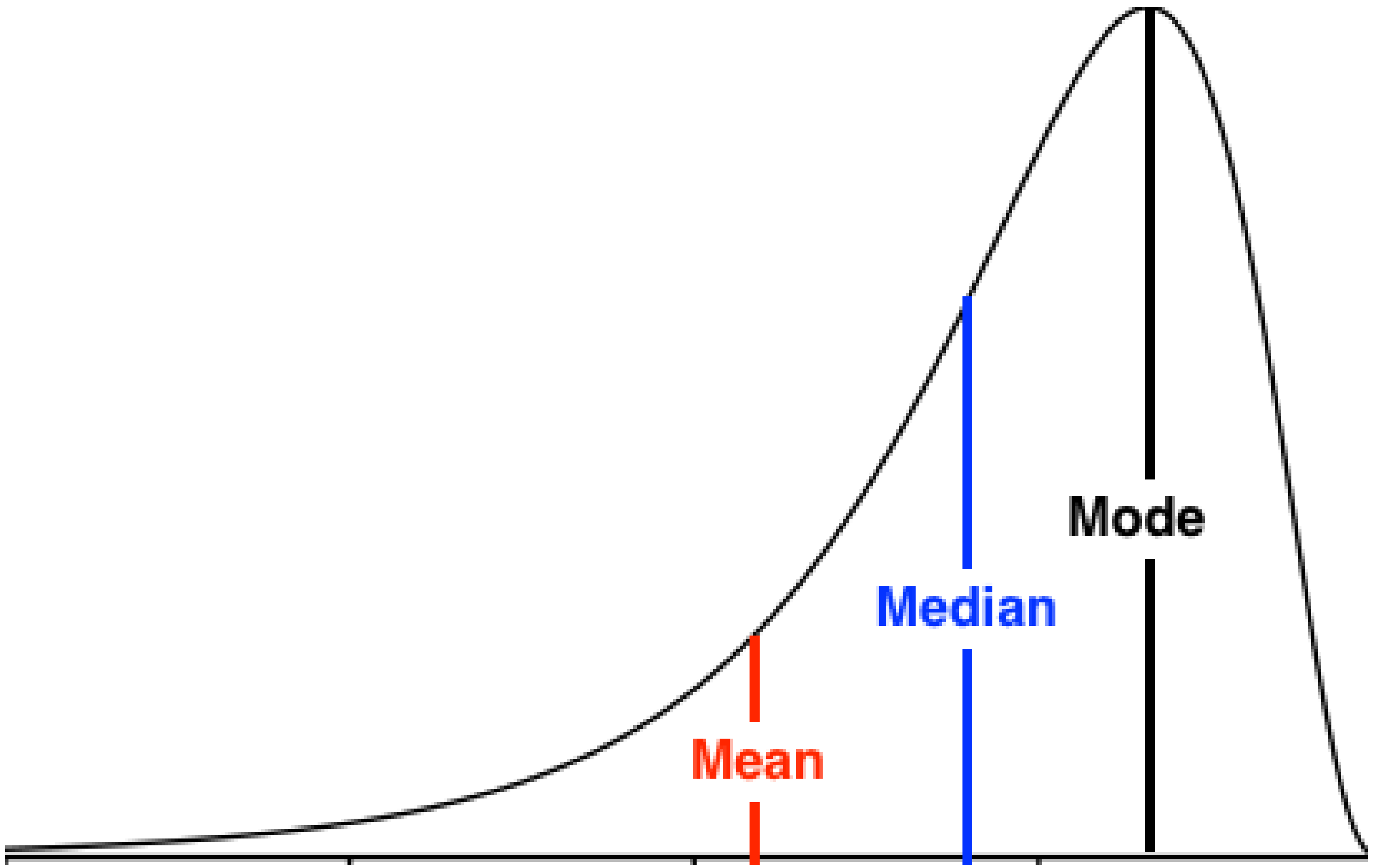
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Modality



[1] Wikimedia

Skewness



[1] Wikimedia

Summary

- Defining descriptive statistics
- Mean, median, and mode
- Standard deviation and variance
- Modality and skewness

Let's prepare for the interview!

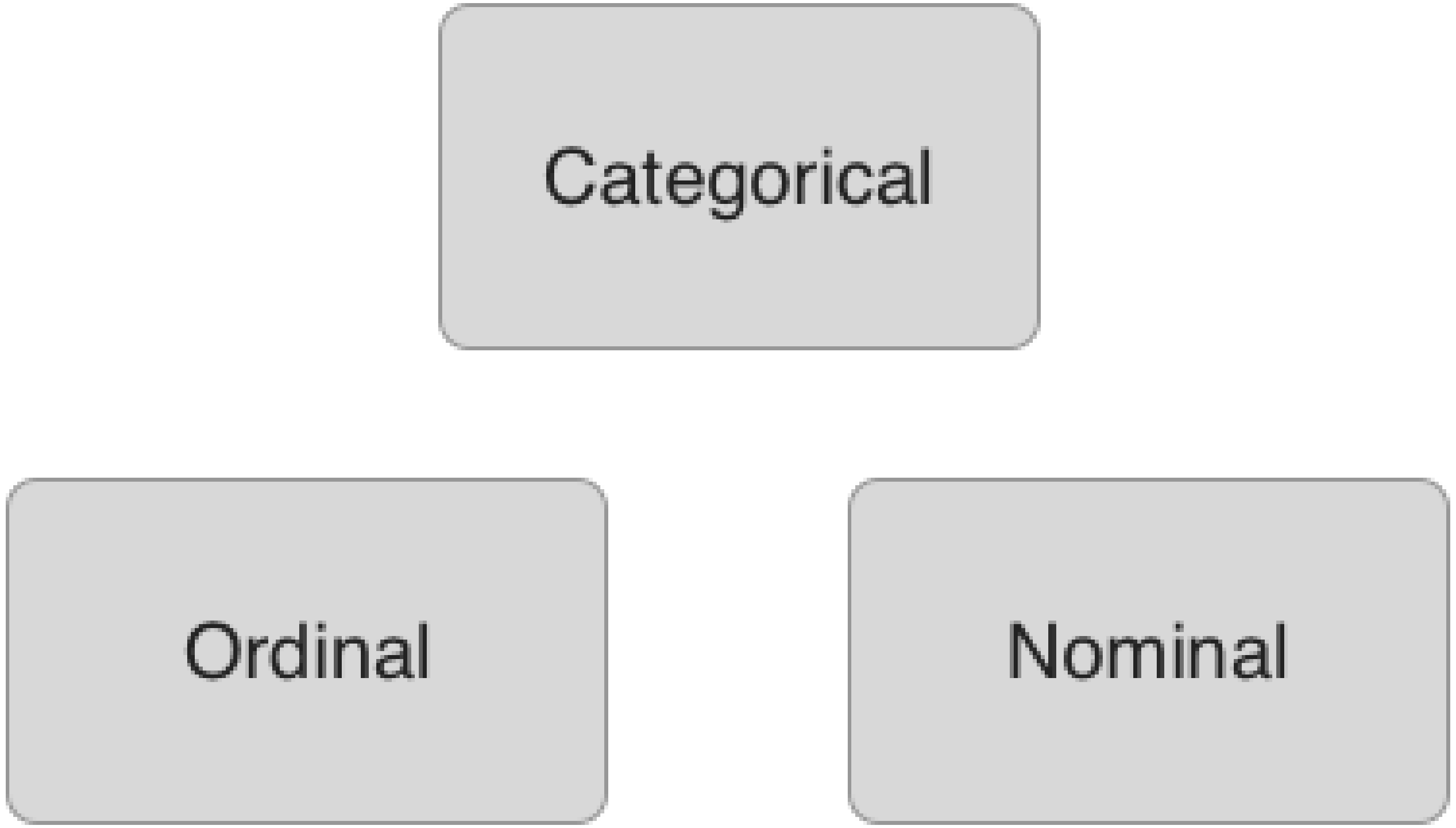
PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Categorical data

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

Types of variables

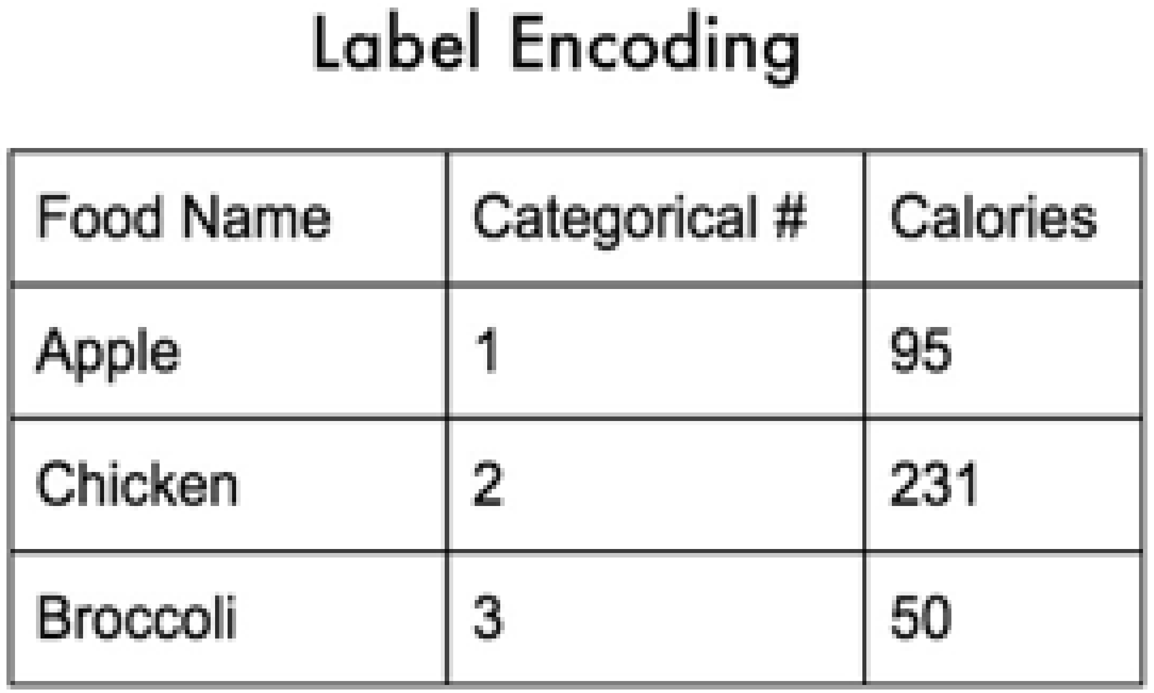


Categorical

Ordinal

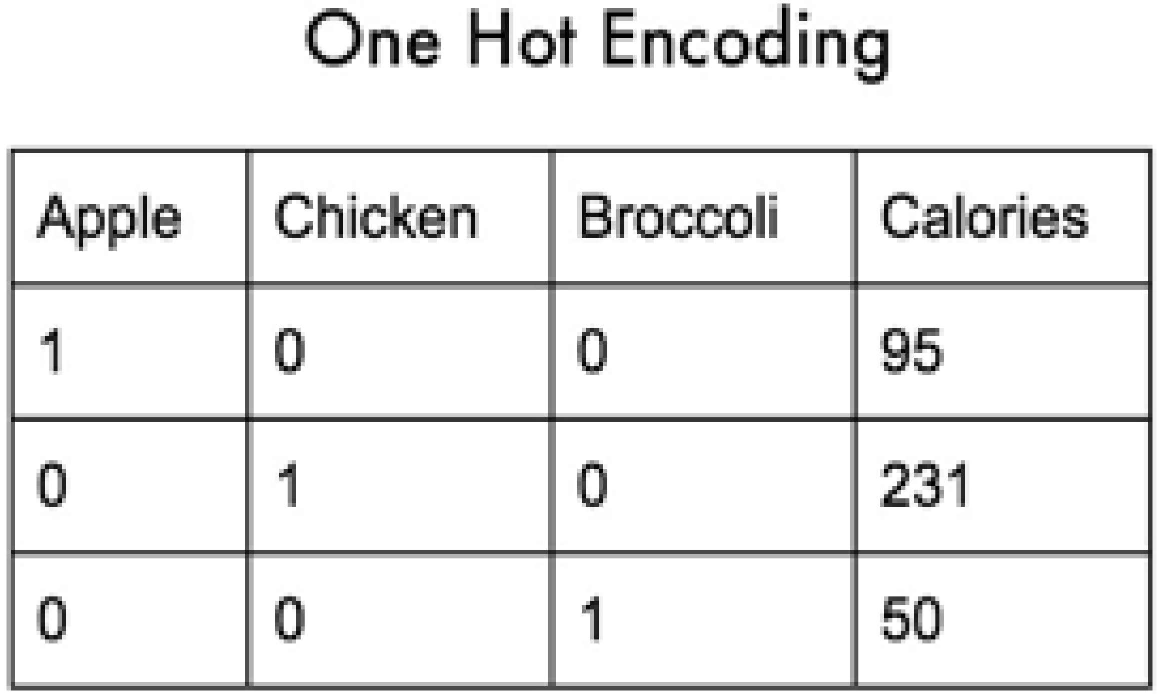
Nominal

Encoding categorical data



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

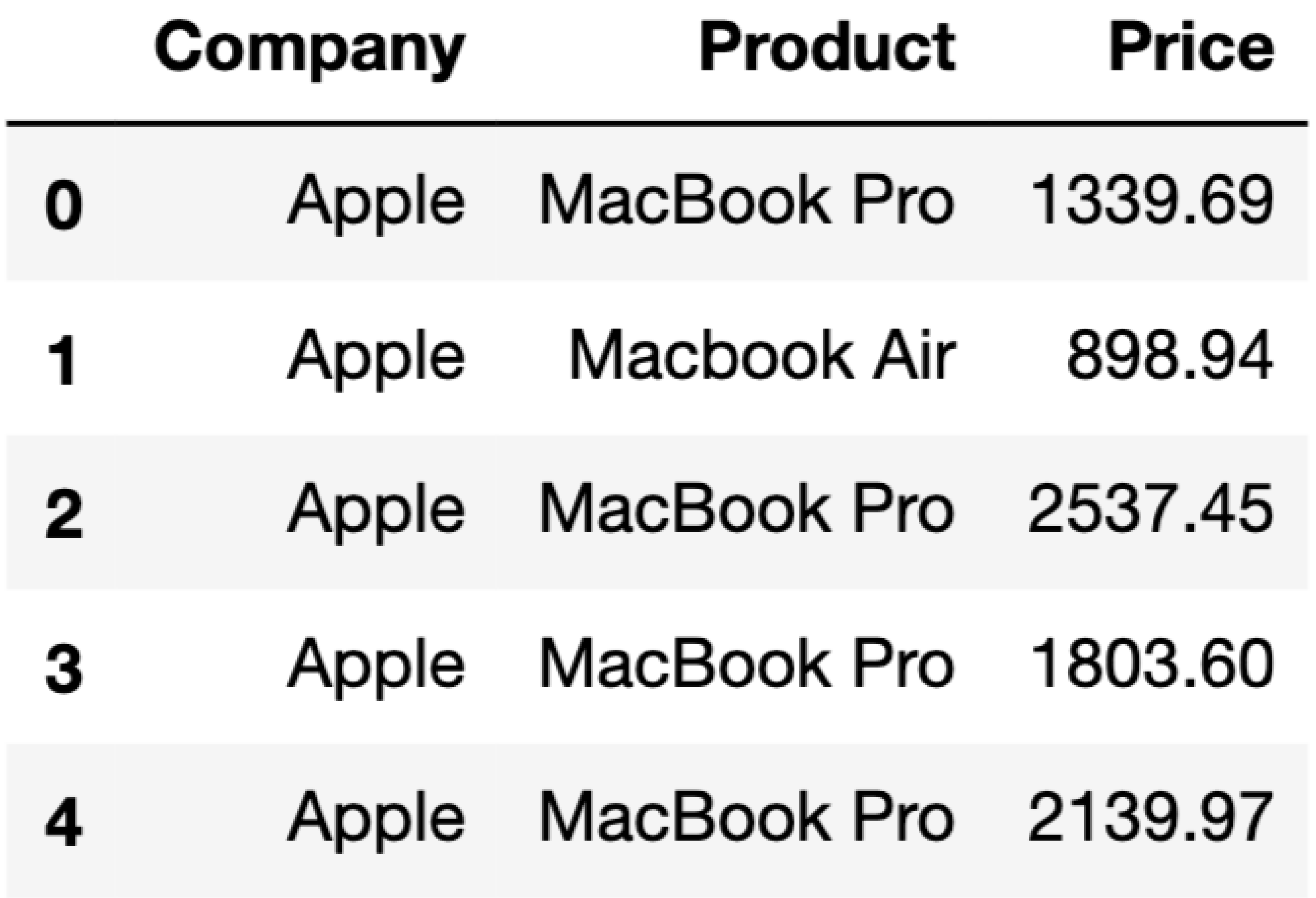


One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

[1] What is One Hot Encoding and How to Do It

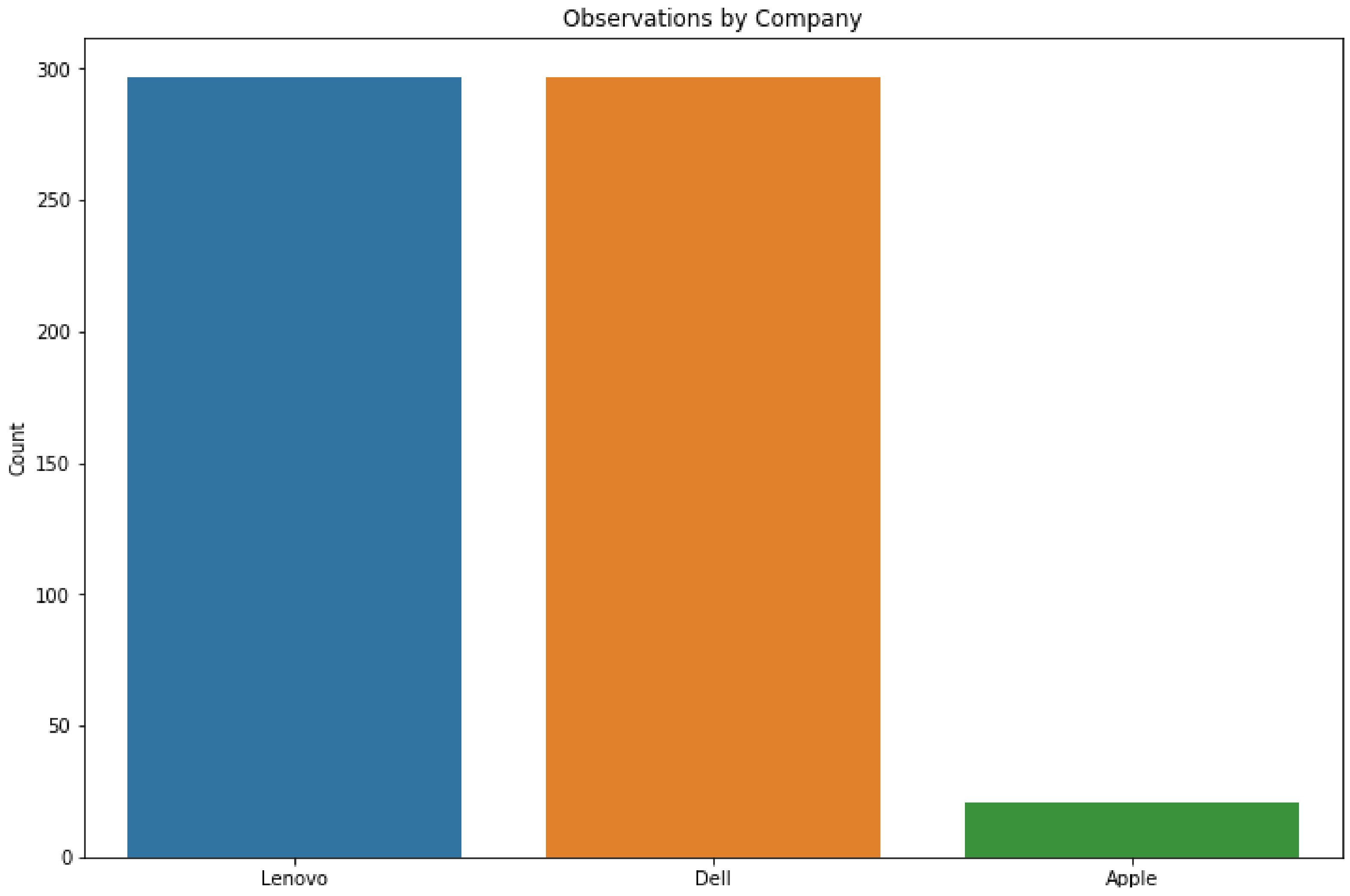
Example: laptop models



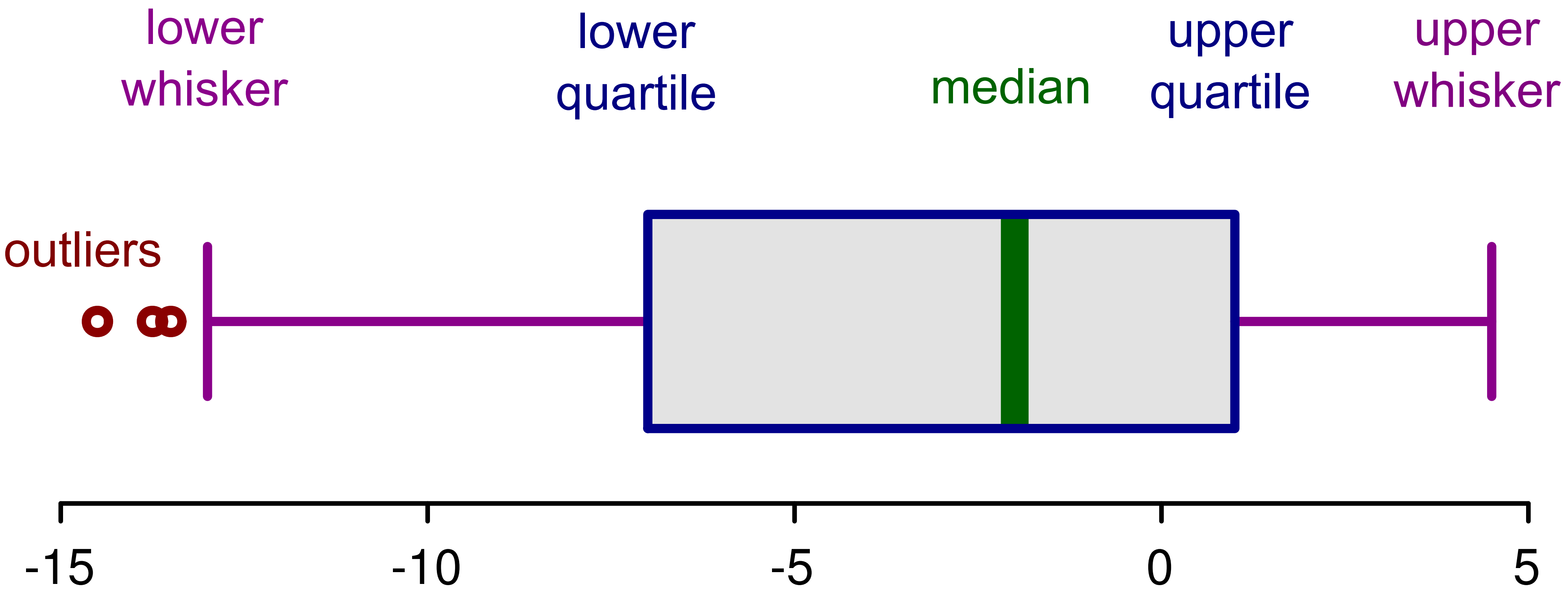
	Company	Product	Price
0	Apple	MacBook Pro	1339.69
1	Apple	Macbook Air	898.94
2	Apple	MacBook Pro	2537.45
3	Apple	MacBook Pro	1803.60
4	Apple	MacBook Pro	2139.97

Example: laptop models

```
company_count = df['Company'].value_counts()
sns.barplot(company_count.index, company_count.values)
```



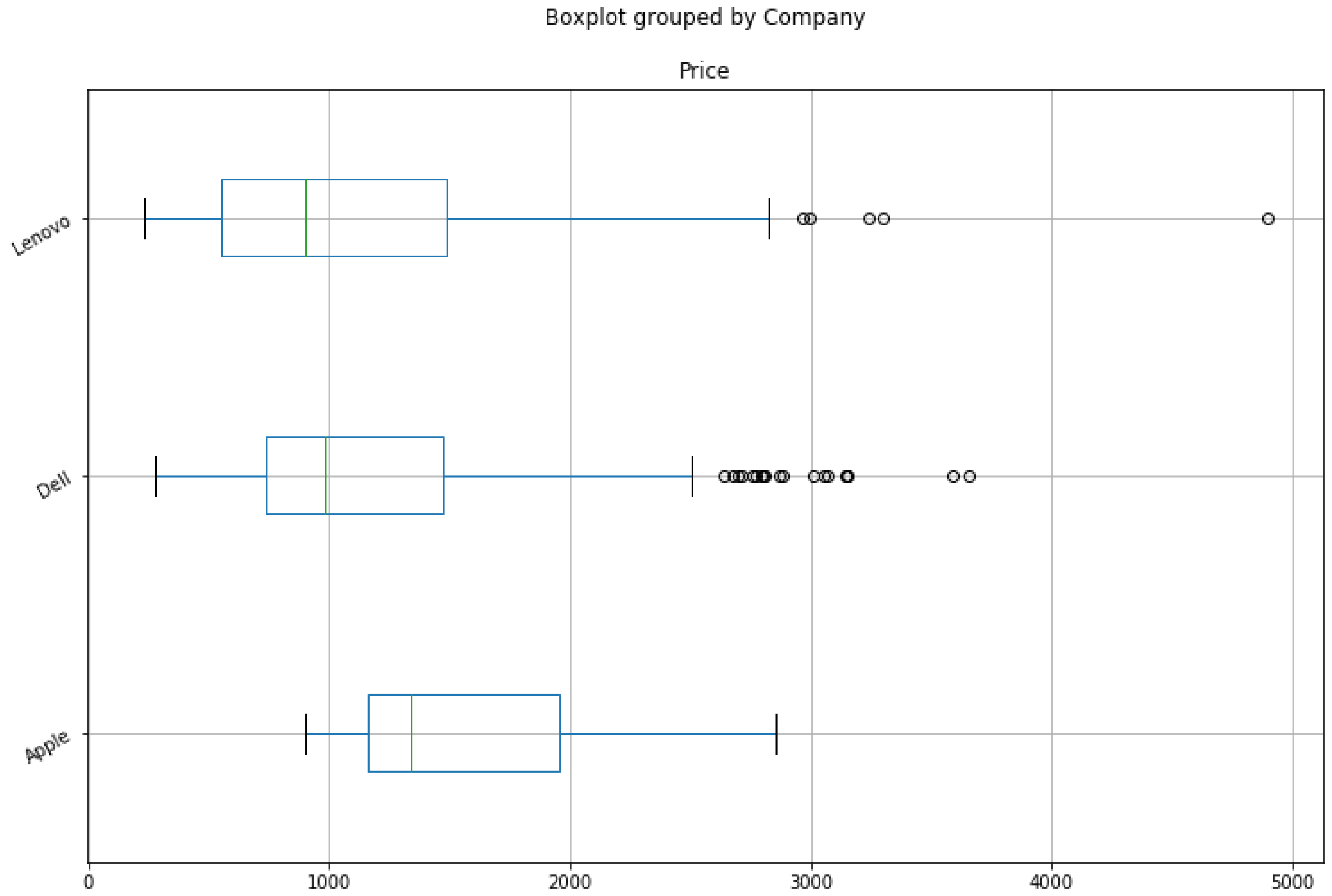
Box plots



[1] Wikimedia

Example: laptop models

```
df.boxplot('Price', 'Company', rot = 30, figsize=(12,8), vert=False)
```



Summary

- Types of variables
- Encoding techniques
- Sample exploratory data analysis

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

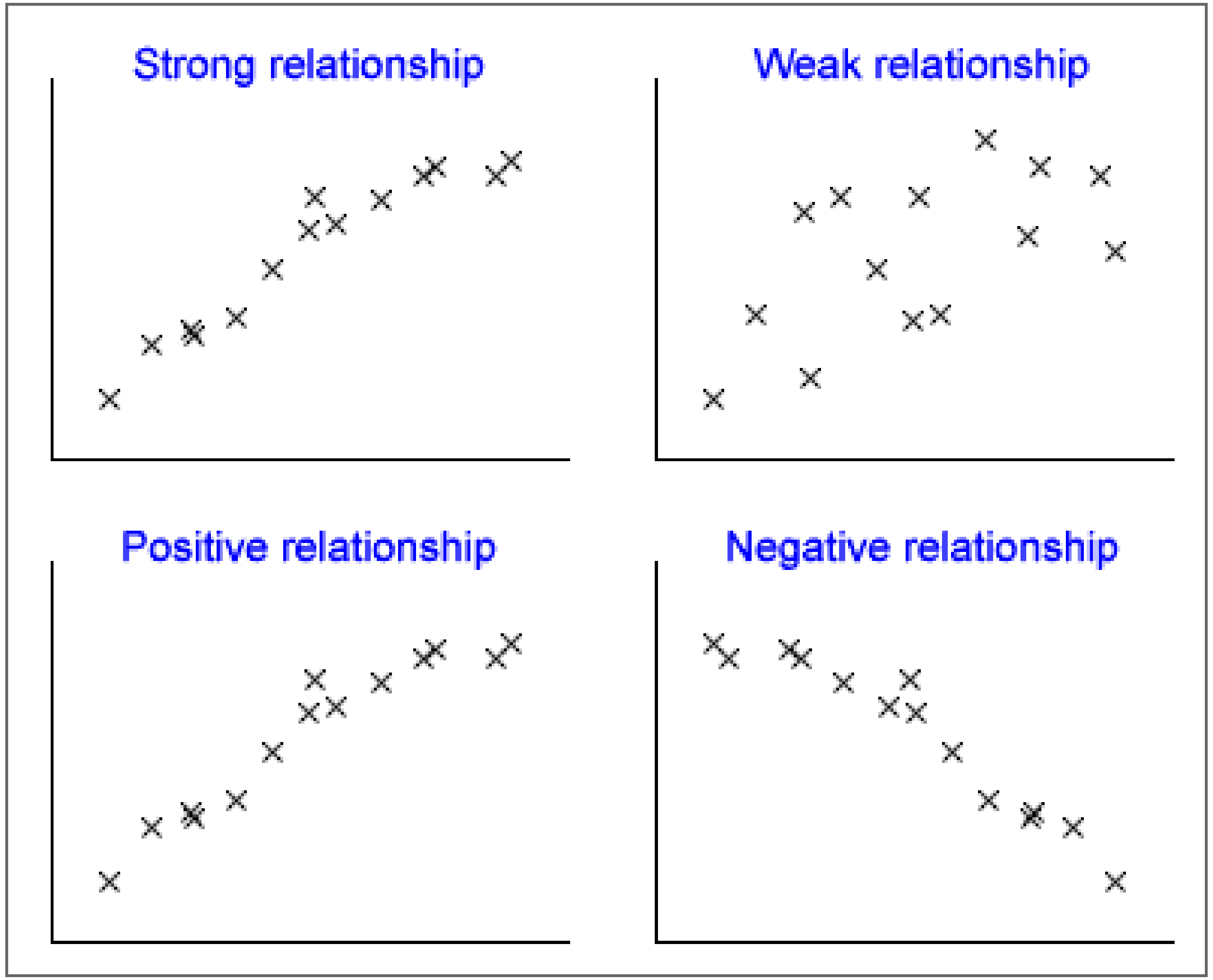
Two or more variables

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey

Data Scientist, Squarespace

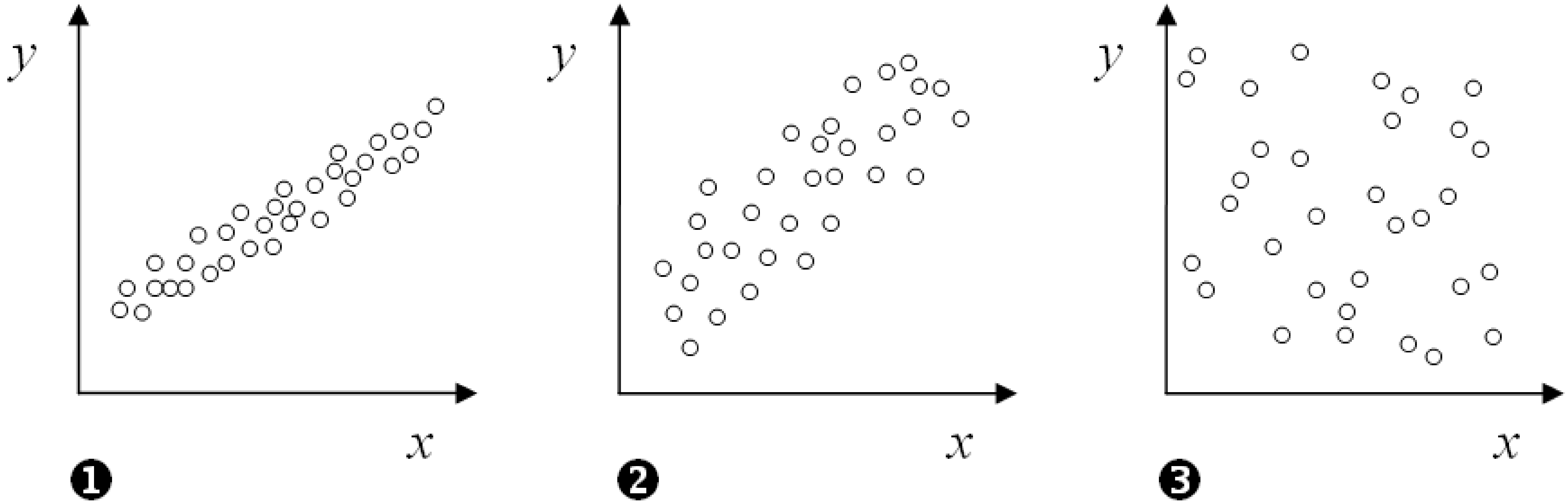
Types of relationships



[1] Wikimedia

What is correlation?

- Statistical relationship between variables
- Stronger correlation = more information



[1] Wikimedia

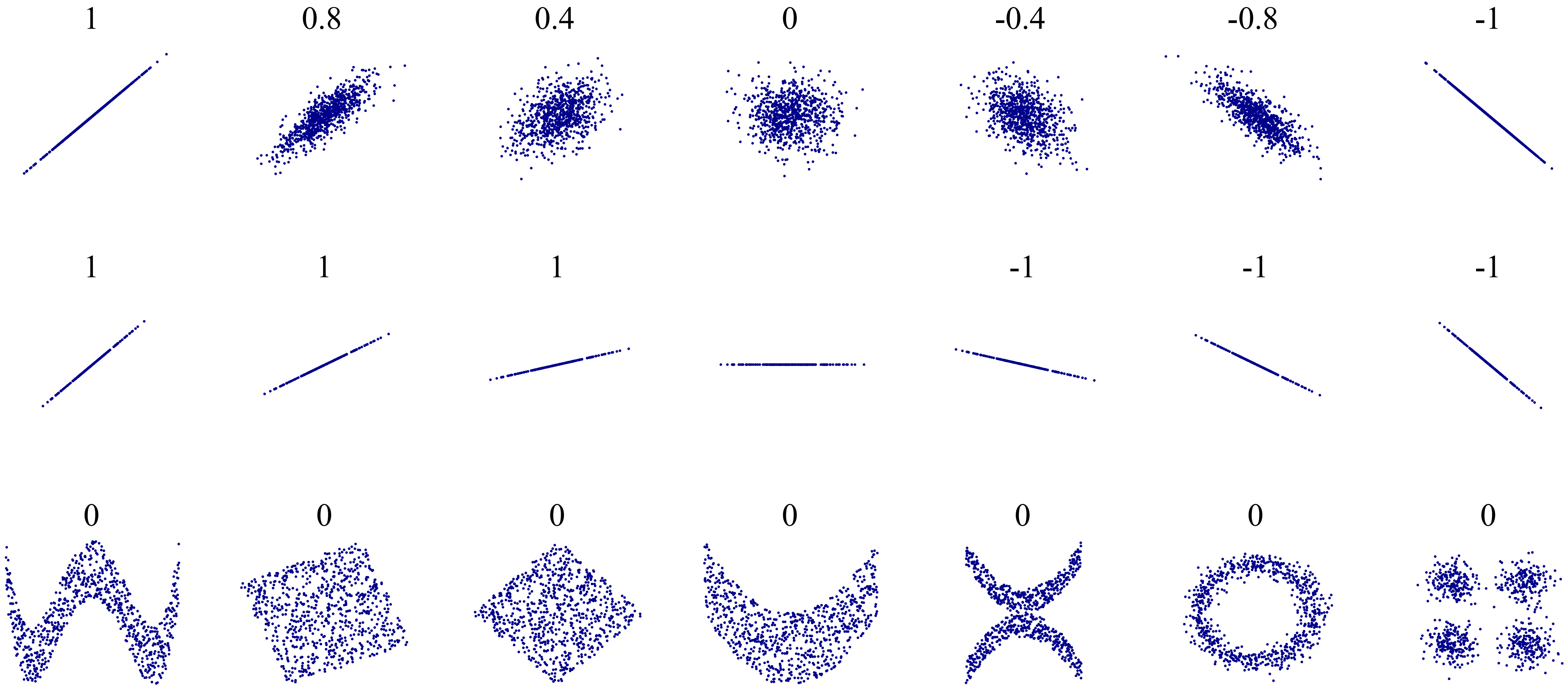
Covariance

$$\mathrm{Cov}_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)}$$

Pearson's correlation

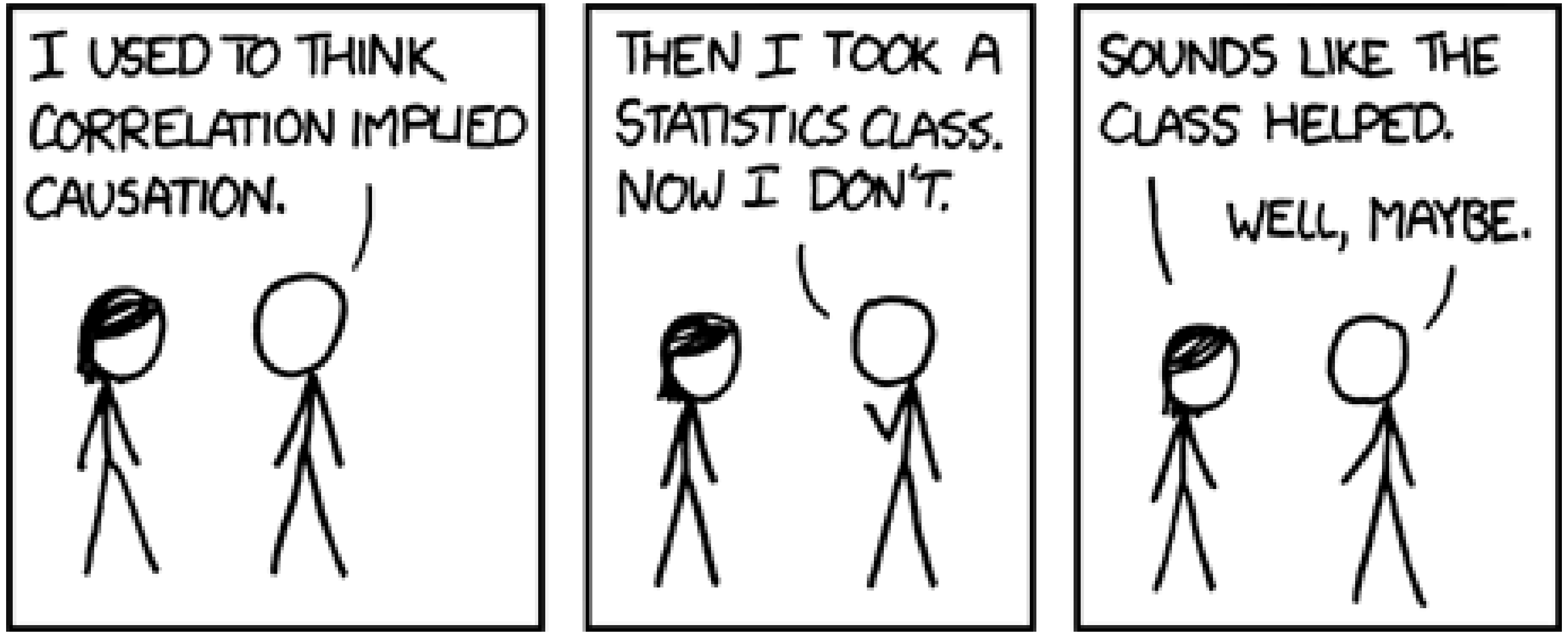
$$r = \frac{\mathrm{Cov}(x, y)}{S_x \cdot S_y}$$

Pearson's correlation



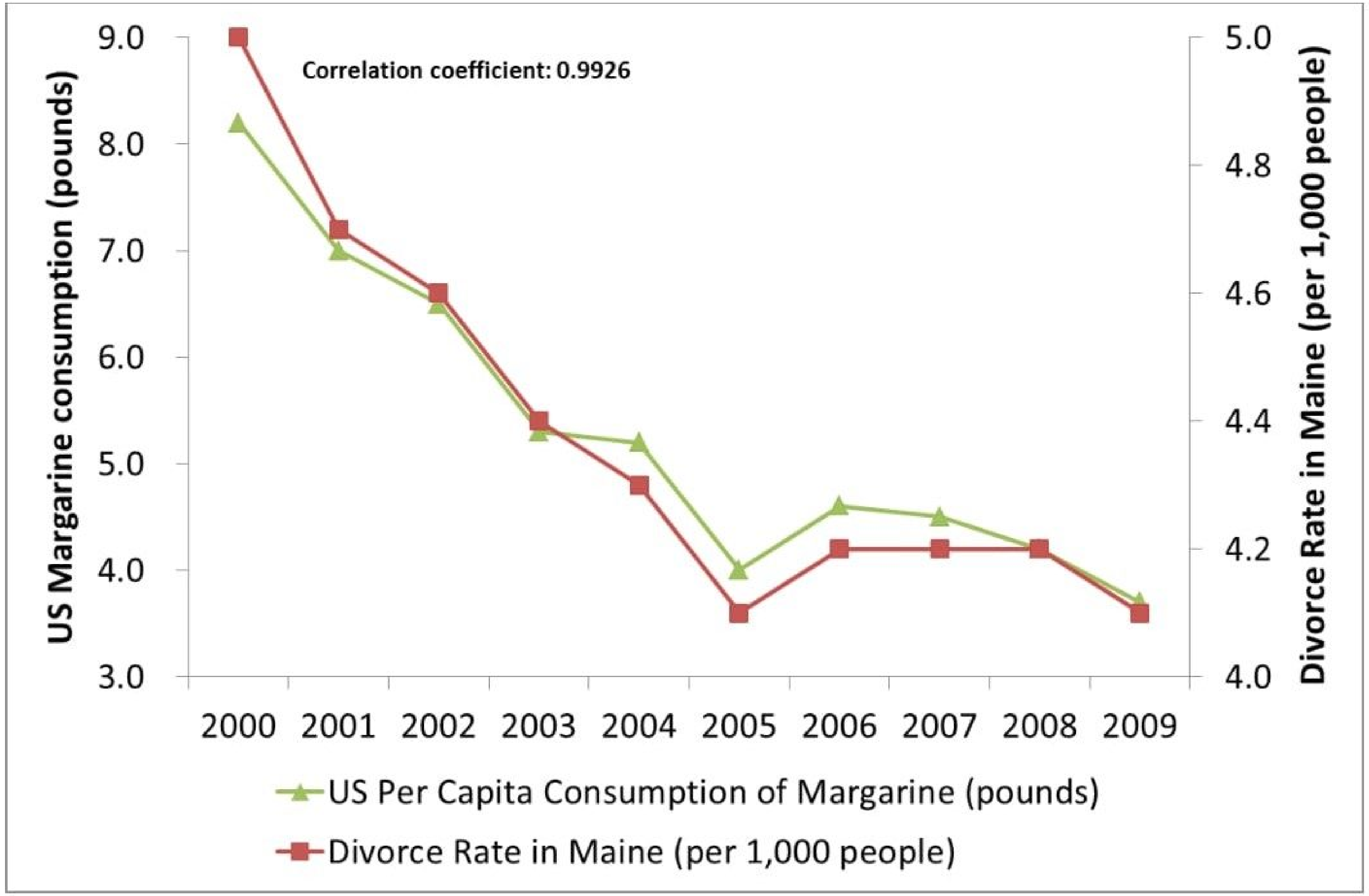
[1] Wikimedia

Correlation vs. causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

[1] xkcd

Correlation vs. causation



[1] Correlation does not mean Causation

Summary

- Types of relationships
- Review of correlation
- Covariance
- Pearson's correlation
- Correlation vs. causation

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON