Statistics (II)

SLIDES BY:

JIANNAN WANG

https://www.cs.sfu.ca/~jnwang/

Outline

Correlation Analysis

- Big Picture
- How to do correlation analysis
- Causal analysis

Hypothesis Testing

- Big Picture
- A/B Testing

Outline

Correlation Analysis

- Big Picture
- How to do correlation analysis

Hypothesis Testing

- Big Picture
- A/B Testing

Correlation Analysis

Correlation

- It is a measure of relationship between two variables

Why is correlation analysis useful?

- For understanding data better
- For making predictions better

Case Study: How to do correlation analysis

Height and weight are correlated

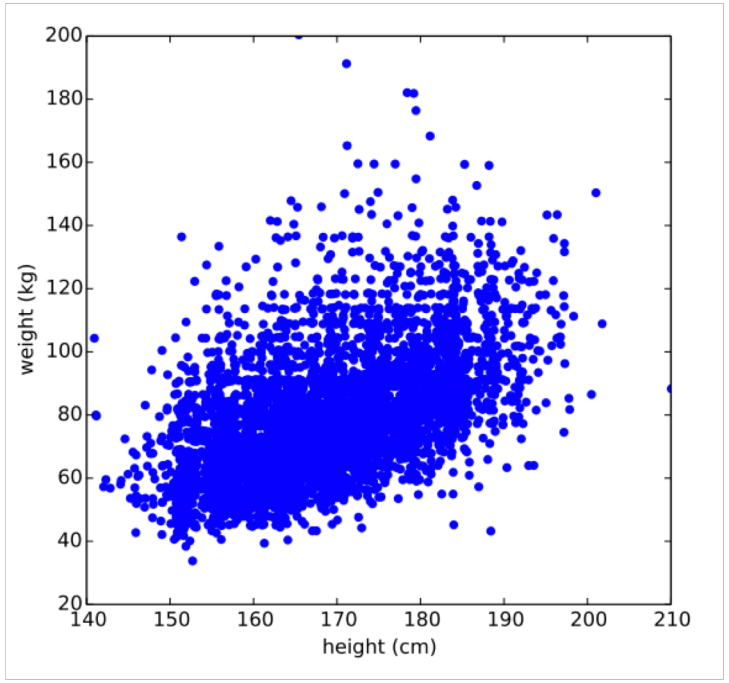
	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0

Source: *Think Stats -- Exploratory Data Analysis in Python*

Idea 1. Visualization

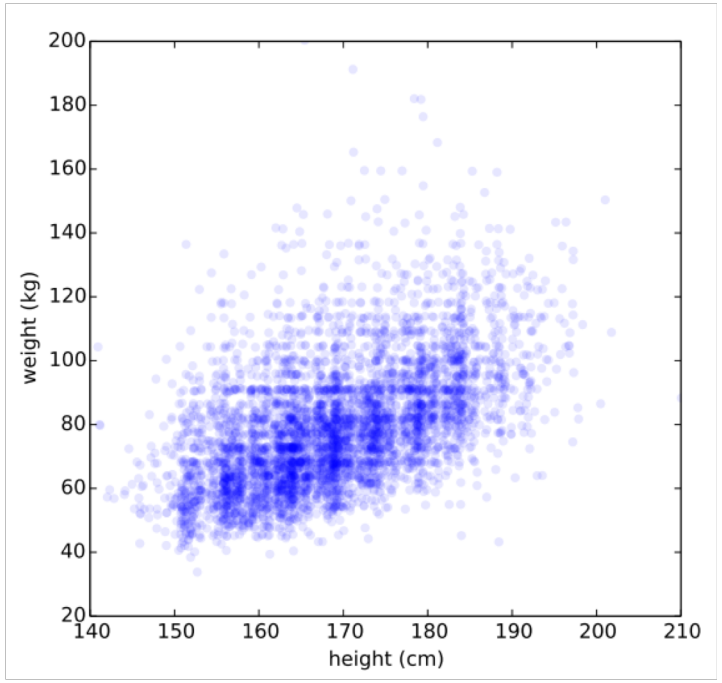
Scatter Plot

	height	weight	age	male
1				
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



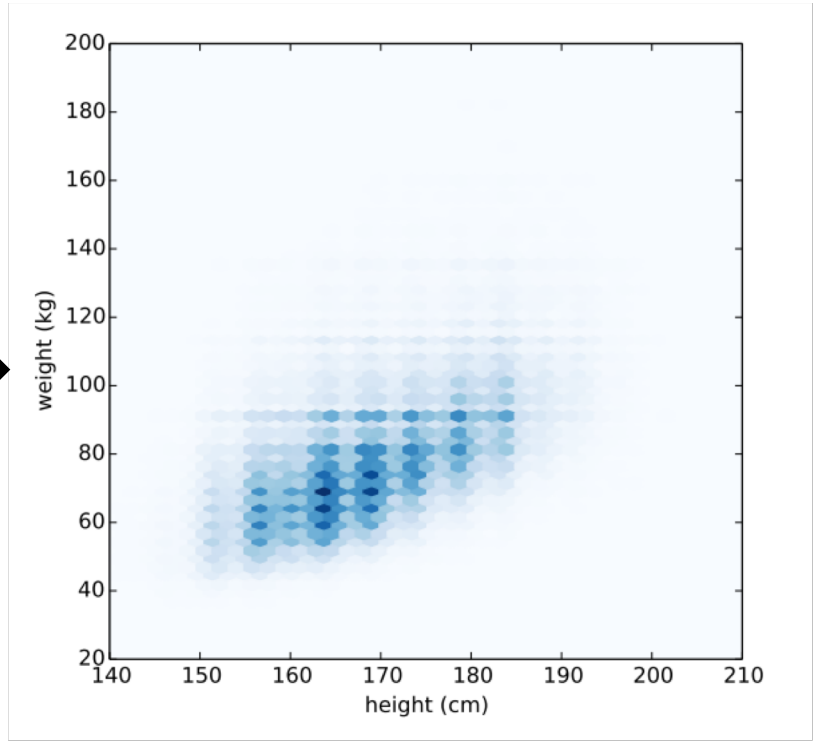
Scatter Plot (with transparency)

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



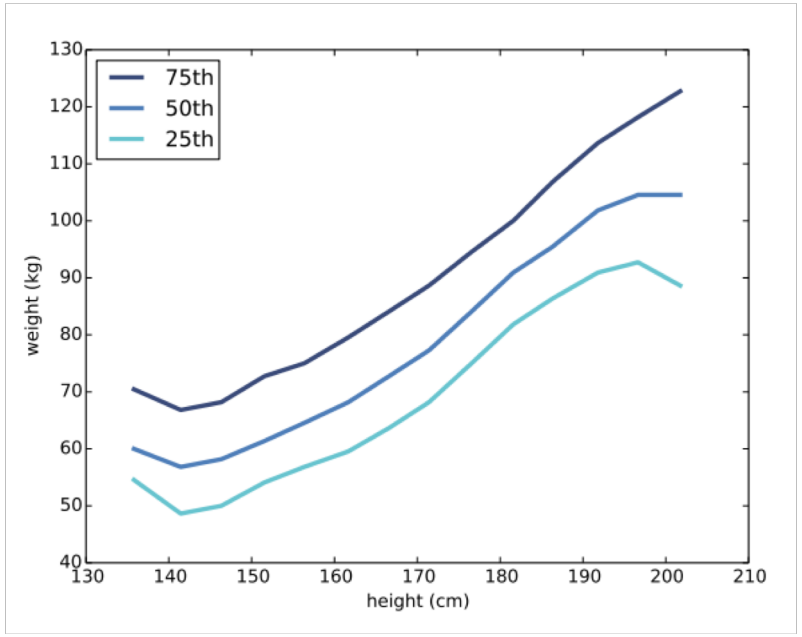
Hexbin Plot

	height	weight	age	male
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



Characterizing relationships

	height	weight	age	male
1				
2	151.765	47.8256065	63	1
3	139.7	36.4858065	63	0
4	136.525	31.864838	65	0
5	156.845	53.0419145	41	1
6	145.415	41.276872	51	0
7	163.83	62.992589	35	1
8	149.225	38.2434755	32	0



Idea 2. Correlation Coefficient

Covariance

Covariance is a measure of the **tendency** of two variables to vary together.

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

$$\operatorname{cov}(X,Y) = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]$$

Hard to interpret

113 kilogram-centimeters

Pearson's correlation

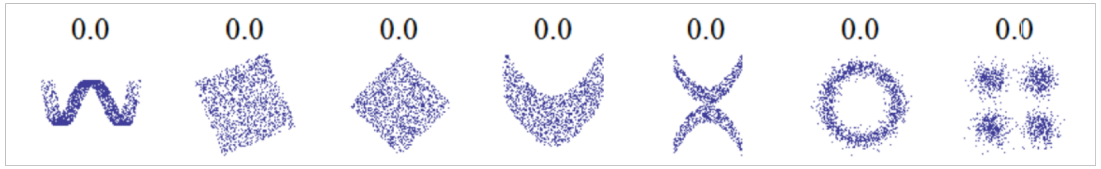
Pearson's correlation is a measure of the **linear relationship** between two variables

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Easy to Interpret

- [-1, 0) → Negative Correlated
- (0,+1] → Positive Correlated
- -1 or +1 → Perfectly Correlated

What about non-linear relationship?



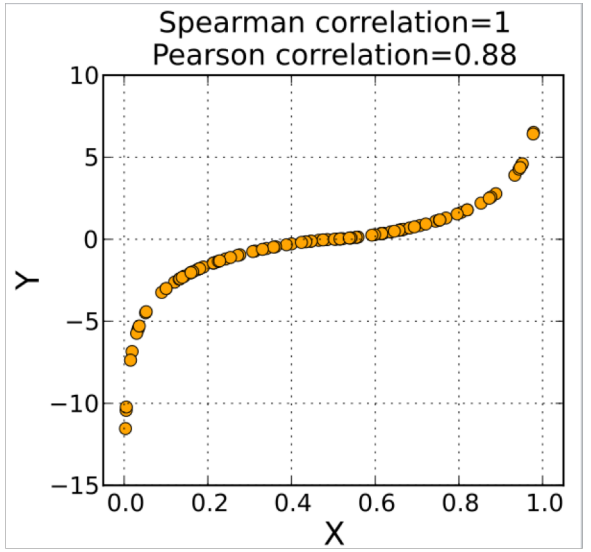
Spearman's rank correlation

Spearman's rank correlation is a measure of **monotonic relationship** between two variables

$$r_s = \rho_{\mathrm{r}_X, \mathrm{r}_Y} = \frac{\mathrm{cov}(\mathrm{r}_X, \mathrm{r}_Y)}{\sigma_{\mathrm{r}_X} \sigma_{\mathrm{r}_Y}}$$

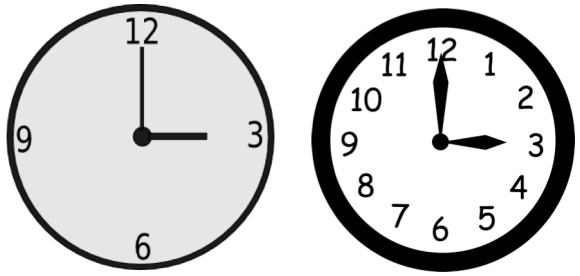
Advantages

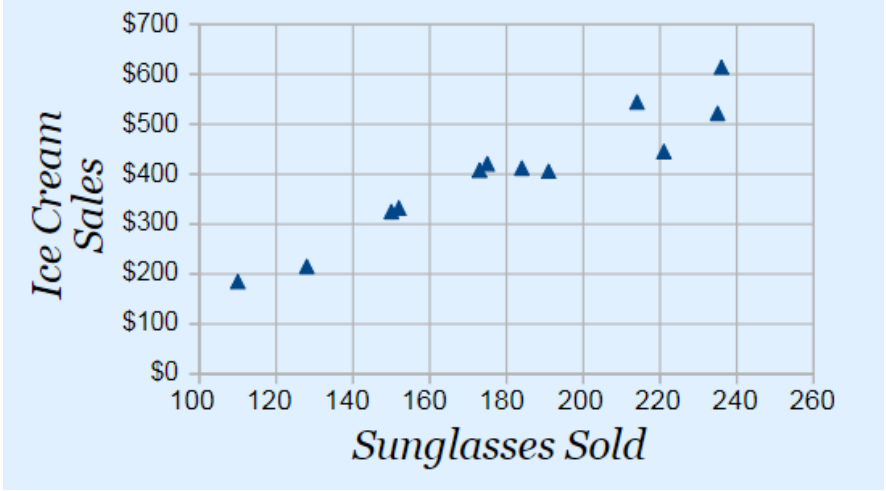
- Mitigate the effect of outliers
- Mitigate the effect of skewed distributions



Causal Analysis

Correlation ≠ Causation





Causal Analysis

If A and B are correlated, then

1. A causes B,
2. B causes A, or
3. C causes both A and B

How to tell which case it is?

1. Use time (If A comes before B, then A can cause B but not the other way around)
2. Use Randomness (randomized controlled trial)

Outline

Correlation Analysis

- Big Picture
- How to do correlation analysis
- Causal analysis

Hypothesis Testing

- Big Picture
- A/B Testing

Why Hypothesis Testing?

We want to make a claim from our data

But, data is just a sample

How to prove our claim in this situation?

Using Hypothesis Testing

Example

- Claim: A data scientist earns more money than a data engineer
- Data: A sample of 50 data scientists and 50 data engineers
- Result: 100K vs. 70k

Can we use this result to prove that our claim is correct?

Hypothesis Testing

Equivalent Terms

- Hypothesis == Claim
- Hypothesis Testing == Claim Proving

Key Idea

- Prove by contradiction

Analogy

- How to prove: There is no smallest rational number greater than zero.
- Hint: a rational number is any number that can be expressed as the fraction a/b of two integers

Alternative and Null Hypotheses

Alternative Hypothesis (H_a)

- This is the claim that you want to prove it's correct

Null Hypothesis (H_0)

- The opposite side of H_a

Possible Outcomes

- Reject H_0 (a contradiction is found) → Accept H_a
- Fail to reject H_0 (no contradiction is found)

Example

Alternative Hypothesis (H_a)

- A data scientist earns more money than a data engineer

NULL Hypothesis (H_0)

- A data scientist earns less (or equal) money than a data engineer

If H_0 is true, what's the probability of seeing:

- ~~Data Scientist (100 K) vs. Data Engineer (70 K)~~
- Salary(Data Scientist) – Salary(Data Engineer) >= 30 K

This is called P-value

Make a decision based on p-value

We hope that

- p-value is as low as possible so that we can reject H_0 (i.e., accept H_a)

Level of Significance (e.g., $\alpha = 0.01$)

- How low do we want p-value to be?

Level of Confidence (e.g., $c = 1 - \alpha = 99\%$)

- How confident are we in our decision?

P-Hacking (Cheating on a P-Value)

Common Mistakes

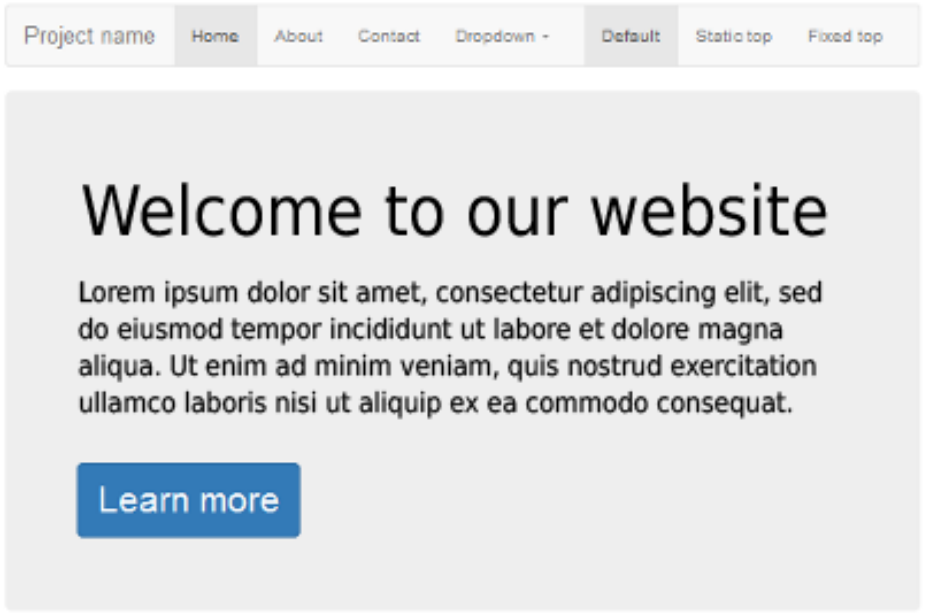
1. Collect data until the hypothesis testing is passed
2. Keep doing analysis on the same data until you find something significant

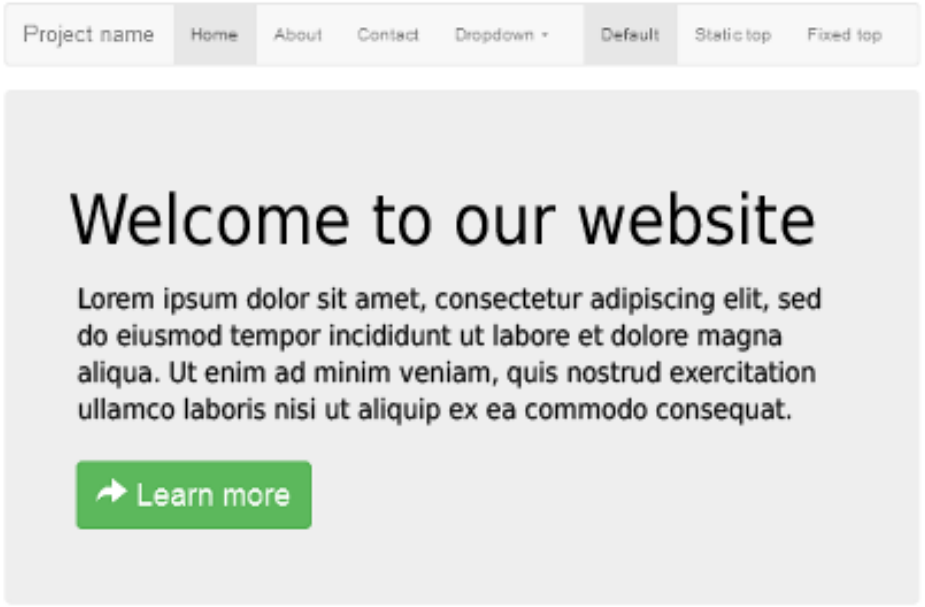
Solution

- You should know what you're looking for (H_0 and H_a) before you start
- Decrease the level of significance (e.g., $\alpha/2$ for two hypothesis tests on the same data)

A/B Testing

What UI is better?





Surprising A/B Tests

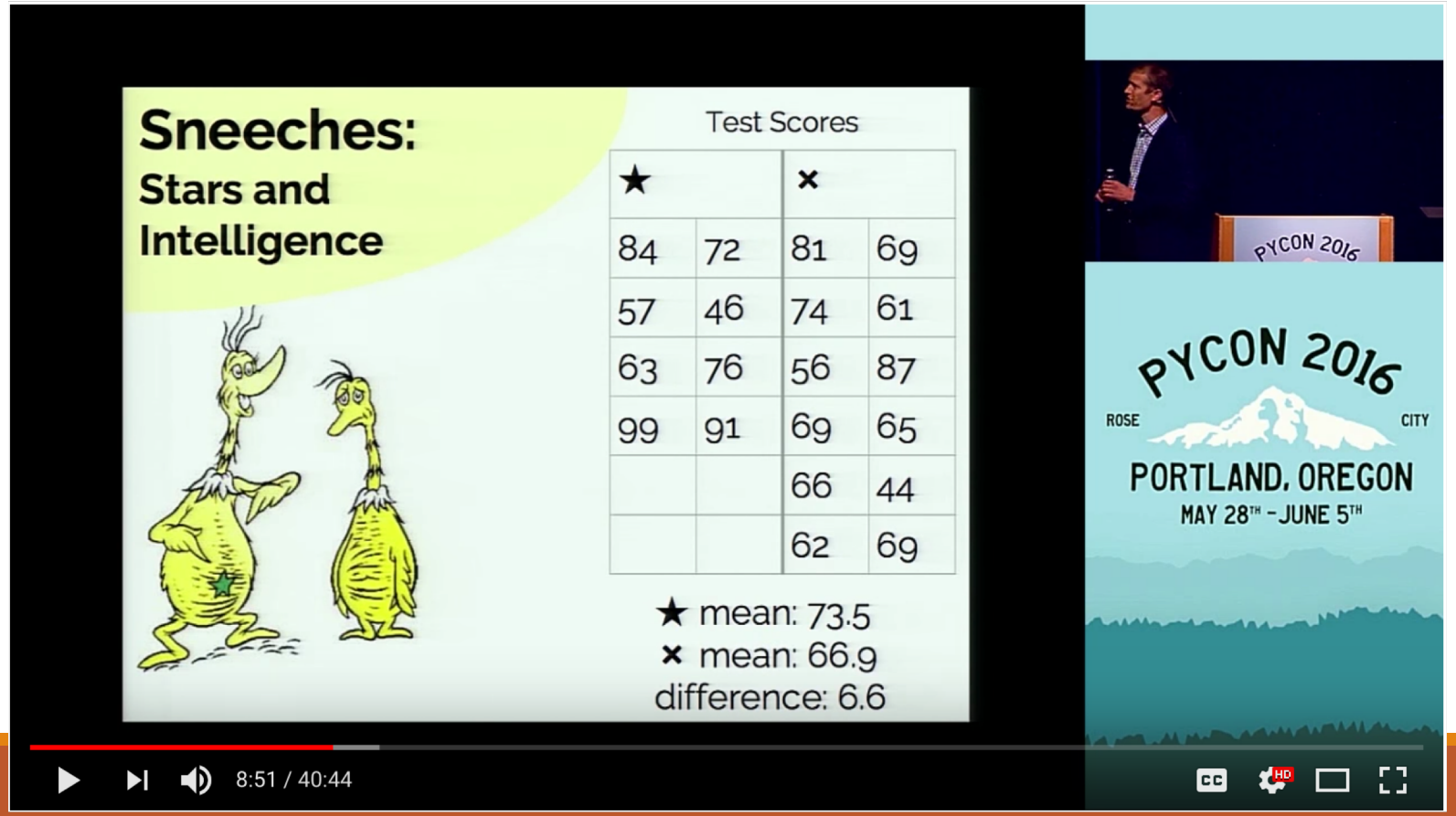
A. Get $10 off the first purchase. Book online now!

B. Get an additional $10 off. Book online now.

https://www.wordstream.com/blog/ws/2012/09/25/a-b-testing

Permutation Test

https://youtu.be/Iq9DzN6mvYA?t=8m9s



Conclusion

Correlation Analysis

- Using visualizations (scatter plot, hexbin plot)
- Using correlation coefficients (Pearson, Spearman's rank)

Hypothesis Testing

- Null Hypothesis (H_0) and Alternative Hypothesis (H_a)
- P-value and P-hacking
- A/B Testing