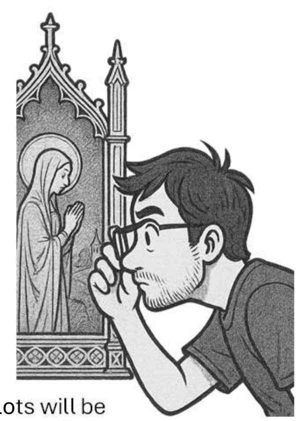
A medieval archive room. Three students - Anna, Markus, and Elena - sit at a large wooden table, dusty documents spread before them.

Anna: "Look at this! These contracts from 15th-century altar painters are fascinating. They specify payments based on painting size, the number of figures, and even how many assistants the master employed."

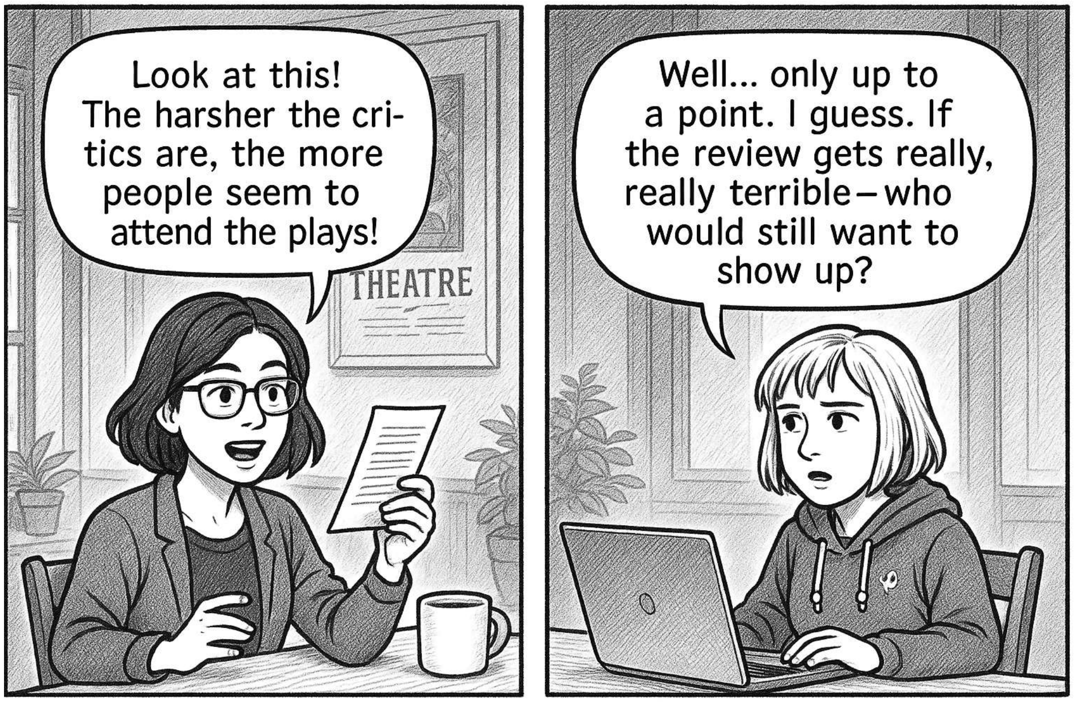
Markus: "We could collect these numbers and see if there's any connection between them. For example, does a larger painting automatically mean more figures?"

Elena: "And how does that affect the honorarium? Let’s build a dataset and explore!"

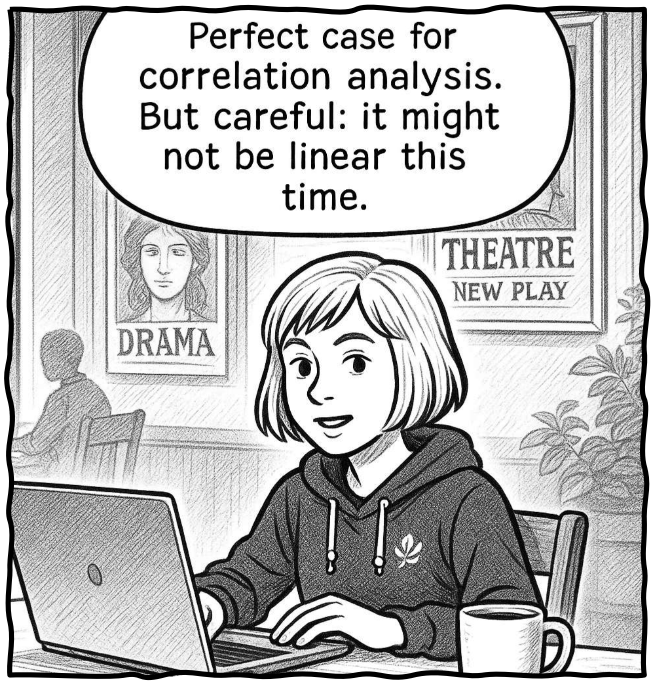
- Run the program *renaissance-retables.py*.
- What are the dimensions of the data being visualized? Name them: painting size, honorarium, etc.
- How do you interpret the figures?
- The fee the painters receive seems to be higher if there are more figures in the painting. Are there any other interesting correlations? How would you explain them?
- Look at the code: does it confirm your explanations?



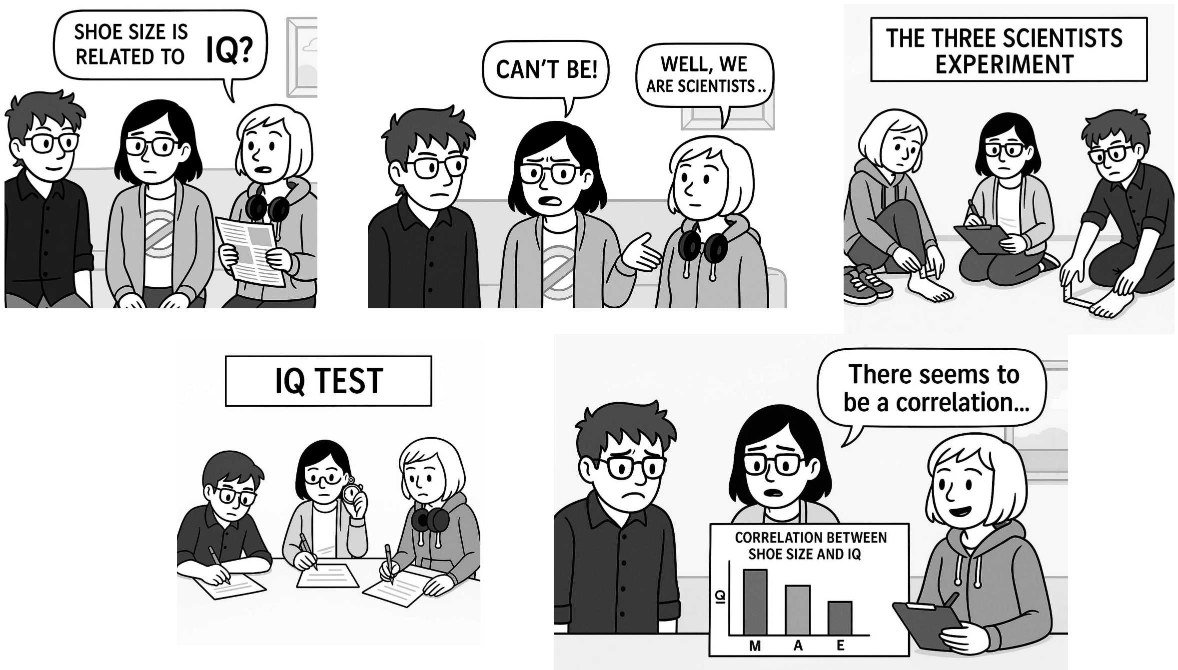
- Run the program *renaissance-retables-2.py*. Several plots will be displayed.
- What is the regression line that is plotted? How is it calculated in the code? What does it suggest?
- Look at the last plot (the overview). What are the curves that can be seen from the top left to the bottom right?
- Focus on “city size.” What do you discover?
- What is the Pearson-coefficient?



1. **Run the program critics.py**
   - Execute the code and observe the two plots and the printed correlation coefficients.
2. **Interpret the scatter plot**
   - Describe the shape of the data (inverted U-curve).
   - What does this shape tell you about the relationship between Kritikschärfe and Publikumsinteresse?
3. **Compare correlation coefficients**
   - Write down the Pearson and Spearman coefficients.
   - Why is Pearson low despite a visible relationship?
   - Why is Spearman higher?
4. **Check the linear regression plot**
   - Why is a linear regression not suitable here?
5. **Look at the code**
   - Analyze how Publikumsinteresse is calculated.
   - Explain why this leads to a non-linear, but still correlated, relationship.



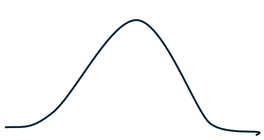
1. **Run the program experiencedCritics.py**
   - Execute the code and observe the plots and printed correlation coefficients.
2. **Interpret the scatter plot**
   - Describe the relationship between Erfahrung (Jahre) and Schreibzeit (Stunden).
   - Is it linear or non-linear?
3. **Compare correlation coefficients**
   - Write down Pearson and Spearman values.
   - Why might Pearson underestimate the strength of the relationship?
4. **Evaluate the regression plots**
   - Compare the linear and logarithmic regression fits.
   - Why does the logarithmic model fit the data better?
5. **Analyze the code**
   - Examine how Schreibzeit is calculated.
   - Explain why the dependency follows a hyperbolic/logarithmic trend.



- Which results could be expected from the experiments the three friends did?
- What correlation would you expect to find, between IQ and shoe size?
- If someone presented you with the results of this study (the three students comparing shoe sizes and IQ), what objections would you have?
- What could be done to get to a more reliable result?
- In case you still do not believe there is really a correlation – how would you try to falsify the hypothesis? What would be your *null-hypothesis*?

The BMI:

1. Body lengths in a population are distributed in a gaussian distribution.
2. The body weight in a population is also distributed in a gaussian distribution.
3. On average tall people are heavier than smaller people (positive correlation).
4. The BMI is defined as $\frac{kg}{m^2}$.The BMI is distributed in a gaussian distribution.



Look at the (sketch of the) curve of the BMI:

- Where would be someone placed, who is extremely thin?
- Would that person be tall or small?
- Would that person be rather light or heavy?
- How likely would it be in a gaussian distribution for someone to be extremely thin or extremely heavy?

When testing for a correlation between two variables, we usually test the **null hypothesis:H_0**:
There is no correlation between the two variables (i.e. Pearson's r=0).

- The p-value tells us how likely it is to observe a correlation as extreme as the one we found — assuming the null hypothesis is true.
- The **p-value** is often misunderstood. It does not tell, how likely it is that a correlation happened just by chance.

  - Run correlationFinder.py.
  - Check some of the results it provides in folder plots.
  - Can you explain the correlations?
  - Do you agree with these two statements?

    A small p-value (e.g., less than 0.05) means: *The correlation is probably real – not random.*
    A large p-value means: *There's no strong evidence for a real connection*

  - Run correctedCorrelations.py
  - It uses a correction mechanism (Benjamini-Hochberg). What has changed?