Data Science Survival Skills

Homework 5

Description of the Homework

In this homework you will work with the wine quality dataset. In the first step, you need to apply a dimensionality reduction to the features to visualise them in 2D. After that, you will have to come up with a hypothesis and apply statistical tests to analyse it.

Homework 5: Tasks 1/2

We provide you a csv file containing the **wine quality dataset**:

- Load the dataset using pandas.
- Create **three** nice looking plots:
  - Use the whole dataset except the ‘quality’ column
  - Apply **PCA, t-SNE and UMAP** to all remaining features (reduce dimensionality to two dimensions)
  - Plot the features for each of the three methods ➜ use the quality column to color code the recorded points (use a suitable colormap)

    ➜ **Slide: Screenshots of the three plots**

Homework 5: Task 2/2

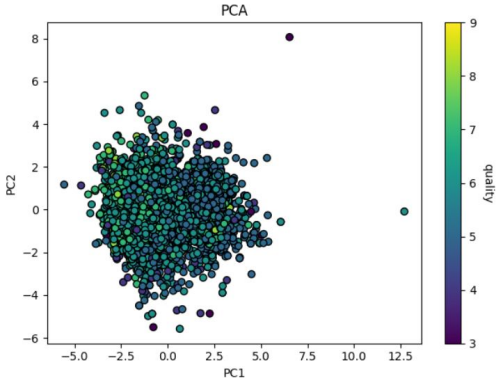
- Think about a hypothesis regarding wine features and wine quality. Apply a statistical test of your choice to investigate if your chosen feature contributes significantly to the wine quality.

➜ **Slide: Your hypothesis.**

➜ **Slide: State and explain why you chose a specific test.**

➜ **Slide: What are your conclusions? Report a p-value.**

Example solution



\+ **Two more plots**

H0: Residual sugar does not contribute significantly to the quality of a wine.

Why: Test ... is used because ...

➔ p-value=X.XXX

➔ Conclusion: ....

Homework: Requirements

You must complete **all** homework assignments (**unless otherwise specified**) following these guidelines:

- **One** slide/page.
- **PDF** file format only.
- It has to contain your **name, student** (**matriculation**) **number** and **IdM** in the down-left corner.
- Font: **Arial,** Font-size: **> 10 Pt.**
- Answer **all** the questions and solve all the tasks requested.
- Be careful with **plagiarism**. Repeated solutions will not be accepted!