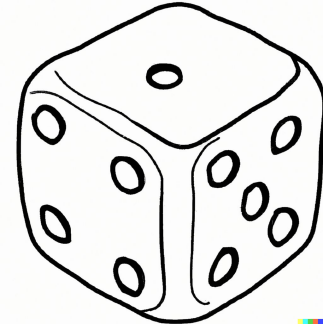
School of Computer Science
University of St.Gallen

Data Science: Exercise 2

Bernhard Bermeitinger, Thomas Huber
03.10.2023

Task 0 - Spurious Correlations



- Provide your own example of a spurious correlation that could be found when analyzing (any) data.

  Explain why it is spurious.

  - *2 minutes*

Task 1.1 - Data Mining

- Split into four groups, one for each of the following data mining techniques:
  - association rule mining
  - clustering
  - classification
  - regression
- Find (or construct) an example of that technique being applied in a real-world scenario.
  - *5 minutes*

Task 1.2 - Data Mining

- Split off into new groups
- All new groups should have (at least) one expert of the four techniques
- Briefly present your technique and example in the group
  - *2 minutes per technique*

Task 2.0 - Notebook Setup

- Set up Google Colab or Jupyter Notebook (or whatever else you prefer for solving the coding exercises)
- Open / load the notebook
  - *~5 minutes*

Task 2.1 - Dataset Loading

- Complete Tasks 1 and 2 in the notebook
- Download the dataset from the paper: https://osf.io/fv8c3
- Load it into a *pandas.DataFrame*
- Clean the DataFrame, by dropping all rows with NaN values
  - *10 minutes*

Task 2.2 - Simple Statistics

- Complete Task 3 in the notebook
- Calculate the **mean**, **median**, **min** and **max** values for all numeric columns
  - *10 minutes*

Task 2.3 - Average Cards

- Complete Task 4 in the notebook
- Calculate the average number of yellow and red cards per game for each player. Then print out the 5 players with the highest average number of cards per game.
  - *10 minutes*

Task 2.4 - Average Per Country

- Task 5 in the notebook
- Calculate the average number of yellow and red cards per game for each country.
    - *5 minutes*

Task 2.5 - Correlation

- Task 6 in the notebook
- For each of the variables, find the variables that have the highest correlation with it.
- Then, form **groups of three** and pick out some correlations and explain why you think they are interesting and what might be the cause of them.
- Present your correlations and what you think causes them.
    - *10 minutes*

Task 2.6 - Simple Analysis

- Task 7 in the notebook
- Create a **boxplot** of the **average rating** grouped by the **average skin color (using the annotator's ratings)**.
- Explain how to read a boxplot.
- Is the boxplot surprising?
    - *5 minutes*

Additional Tasks

Task 3.1 - Descriptive Task

- What four methods to summarize the main aspects of the data in a descriptive way have you learned in the lecture?
- Form four groups, one for each method.
- In your group, using your method, find a way to summarize the data from the “Many Analysts, One Dataset” paper
  - *5 minutes*

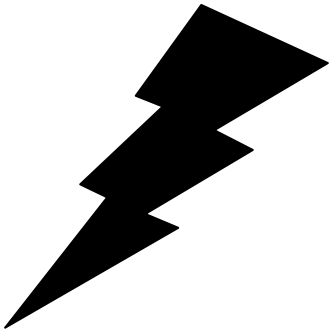
Task 3.2 - Explorative Task

- What four methods to find patterns, relationships, anomalies or trends in the data have you seen in the lecture?
- Form four groups, one for each method.
- In your group, explain the purpose of performing this method.
  - *5 minutes*

Task 3.3 - Predictive Task

- What four methods to make predictions about future data points have you seen in the lecture?
- Form four groups, one for each method.
- In your group, explain on what variable(s) you could apply this method in the dataset, or, if none are applicable, what additional data you would need.
    - *5 minutes*

Task 4 - From Data To Knowledge



- Starting from **Data**, what are the steps you need to take to reach **Knowledge**?
- Give an example of how to perform this step for the soccer player dataset.