School of Computer Science
University of St.Gallen

Data Science: Exercise 3

Bernhard Bermeitinger, Thomas Huber
03.10.2023

Task 2.4 - Average Per Country (*Exc 2*)

- Task 5 in the notebook from *exercise 2*
- Calculate the average number of yellow and red cards per game for each country.
  - *5 minutes*

Task 2.5 - Correlation (*Exc 2*)

- Task 6 in the notebook from *exercise 2*
- 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 (*Exc 2*)

- Task 7 in the notebook from *exercise 2*
- 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*

Task 1.1 - Plotting Player Weight vs. Height

- Create a **Scatter Plot** of the player

  weight (x-axis) vs. **height** (y-axis)

Task 1.2 - Data Manipulation

- Create a new column with the name **NameLength**, that contains the length of the player's name for each row

Task 1.3 - Scatter Plot

- Create a **Scatter Plot** of player **weight** vs. **name length**.

- Do you see a correlation between the two? Why or why not?
- What makes this plot different from the one of weight vs. height?

Task 1.4 - Linear Regression

- Create a **Linear Regression** model that predicts the player's **height** based on the player's **weight**.
  Use [scikit-learn LinearRegression] for this

According to your model:

- What is the height of a player who weighs 80 kg?
- What is the height of a player who weighs 100 kg?

Task 1.5 - Scatter Plot with Regression Line

- Create a **Scatter Plot** of player **weight** vs. **height**.
- Draw the **regression** line into the scatter plot.

- Bonus: Repeat this for **height** vs. **name length**.

Task 2.1 - SQL Query with SQLAlchemy

- Write a query that returns the player's **weight**, **height**, and **position** using SQLAlchemy.

Task 3.1 - SQL Query with SQLAlchemy

- To enrich our data we will collect information about the countries. For this we will use an API.

Task 3.2 - Data Cleaning

- The '**name**' column contains dictionaries. This makes it annoying for us to work with.
- Simplify the column by **replacing** all entries in it with the value in 'common' in that dictionary.

Task 3.3 - Joining DataFrames

- **Combine** the two DataFrames on the **leagueCountry** column.
- For the DataFrame with the countries, you only need the **name** and **fifa** columns.

Task 4.1 - Joining crowdstorming data and country data with SQL

- Select all columns from the **crowdstorming** table, and only the **fifa** column from the **countries** table.
- Then join the two tables on the **leagueCountry** column of the **crowdstorming** table and the **name** column of the **countries** table.

Task 5 - Calculating the mean

- Calculate the **mean height** and **weight** of each player in the database, **using** SQLAlchemy.
- Repeat this, but on the **DataFrame**. Are the results the same?

Task 6 - Calculating the mean per position

- Calculate the **mean height** and **weight** of each player **per position** in the database, using SQLAlchemy.
- Repeat this, but on the **DataFrame**. Are the results the same?

Task 7 - Calculating the mean per position and league

- Calculate the **mean height** and **weight** of each player **per position** and **per league** in the database, **using** SQLAlchemy.
- Repeat this, but on the **DataFrame**. Are the results the same?