Work on project. Stage 4/5: The statistics

Project: Data Analysis for Hospitals

The statistics

Hard 1 hour **220** users solved this stage. Latest completion was **about 3 hours ago**.

Description

2 / 2 Prerequisites

- ✓ Summarizing categorical columns 2
- ✓ Reshaping and Pivot Tables 2

You have cleared your dataset of empty rows and values. Some values have also been corrected, and now we can start a comprehensive study of our data. In this stage, we will find the main statistical characteristics of our data, consider data distributions, and so on.

Answer the following questions and output the answers in the specified format.

1. Which hospital has the highest number of patients?
2. What share of the patients in the general hospital suffers from stomach-related issues? Round the result to the third decimal place.
3. What share of the patients in the sports hospital suffers from dislocation-related issues? Round the result to the third decimal place.
4. What is the difference in the median ages of the patients in the general and sports hospitals?
5. After data processing at the previous stages, the `blood_test` column has three values: `t` = a blood test was taken, `f` = a blood test wasn't taken, and `0` = there is no information. In which hospital the blood test was taken the most often (there is the biggest number of `t` in the `blood_test` column among all the hospitals)? How many blood tests were taken?

Hint

Objectives

Steps 1-8 are the same as steps 2-9 in the third stage. It's not necessary here to set the maximum number of columns to display. The fourth stage requires completing the following steps:

1. Read the CSV files with datasets.
2. Change the column names. The column names of the sports and prenatal tables must match the column names of the general table.
3. Merge the data frames into one. Use the `ignore_index = True` parameter and the following order: `general`, `prenatal`, `sports`.
4. Delete the `Unnamed: 0` column.
5. Delete all the empty rows.
6. Correct all the gender column values to `f` and `m` respectively.
7. Replace the `NaN` values in the gender column of the prenatal hospital with `f`.
8. Replace the `NaN` values in the `bmi`, `diagnosis`, `blood_test`, `ecg`, `ultrasound`, `mri`, `xray`, `children`, `months` columns with zeros.
9. Answer the 1-5 questions using the `pandas` library methods. Output the answers on the separate lines in the format given in the Example section.

If you have corrupted CSV files, please download them and unzip in your working directory.

Example

The input is 3 CSV files, `test/general.csv`, `test/prenatal.csv`, and `test/sports.csv`.

The output: the following answers are given for reference only, the actual answers might be different.

```
The answer to the 1st question is Brighton
The answer to the 2nd question is 0.645
The answer to the 3rd question is 0.873
The answer to the 4th question is 35
The answer to the 5th question is Oxford, 178 blood tests
```

Report a typo