Unit 1 homework instructions

DKU Stats 101 Fall 2025 Session 2

2025-10-24

Table of contents

Scoring guide **2**

Content . . . 2

Technical . . . 2

Questions **3**

Question 1: Displaying and describing the data (25 points) . . . 3

1a. Investigating offensive shots . . . 3

1b. Investigating goals . . . 4

1d. Thinking about your results . . . 4

Question 2: Comparing groups (25 points) . . . 4

2a. Compare `home_foul_advantage` by the variable `covid_match` . . . 5

2b. Compare `home_yellow_advantage` by the variable `covid_match` . . . 5

2c. Thinking about your results . . . 5

Question 3: Considering deviations (25 points) . . . 5

3a. Selecting your data . . . 5

3b. Finding the average . . . 6

3c. Normalizing the data . . . 6

3d. Thinking about your results . . . 6

Question 4: Your own investigation (25 points) . . . 6

4a. Selecting your own question . . . 6

4b. In summary . . . 7

Scoring guide

Content

- Getting the right answer is only a small part of the grade
- Good quality interpretation of your results is the name of the game
- If you see something that looks unusual in your data (outlier, some unusual distribution type) - investigate it!
- When explaining your results, say something interesting about them. Did it match your expectations? Why or why not?
- Brief explanations that simply repeat what I can visually see myself will not receive a good score
- On the other hand, filling the homework with pages of not very interesting description is not valuable either. The goal isn't to write the most words, but find the most interesting things in the data.
- You do not need to be an expert in football to get a good grade on this assignment, but I will expect you to look up basic information, such as "what is a shot in football" and "what kind of games were Covid games"? and so on to help you understand and set expectations your data.
- The information requested in the question prompts are only a starting point, if you find other interesting information along the way, please report that. You don't need to look at the data forever but if there is obviously something else interesting in the data you should report it.
- You must have up to Question 1 completed for the homework check on October 30th

Technical

- Make sure your graphs are produced using `ggplot()`, are well labeled, and are easy to read.
- Make sure your tables are produced with the `kable()` function from the `knitr` package, are well labeled, and are easy to read. You can make your tables prettier with the `kableExtra` package.
- Make sure you do not have anything rendered in your PDF file besides your results and, when asked for by a question, your code. That means no warnings, messages, or other output should appear in your final rendered PDF file.
- Make sure to accurately mark each page a question answer appears on when submitting on GradeScope.

Questions

Question 1: Displaying and describing the data (25 points)

For this investigation, we are going to examine the distribution of goals.

1a. Investigating offensive shots

Using the Think-Show-Tell framework from the textbook, investigate the distribution of `home_shots` and `away_shots`.

Note: I recommend you use the internet to look up how a shot is defined in football

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

1b. Investigating goals

Using the Think-Show-Tell framework from the textbook (example on page 71), investigate the distribution of `home_goals` and `away_goals`.

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

1d. Thinking about your results

Consider the results of 1b. and 1c. together. What can we understand about these offensive statistics?

Question 2: Comparing groups (25 points)

One popular theory about why teams have a home field advantage is that the intensity of the crowd influences the officials to be more favorable to the home team (see this article). Let's see if our dataset supports this hypothesis.

First, create a variable called `covid_match` where the variable is `TRUE` if the match was conducted in an empty stadium during Covid-19 and `FALSE` if it was not. Next, create a variable called `home_foul_advantage` that subtracts `home_fouls` from `away_fouls` and another variable called `home_yellow_advantage` that subtracts `home_yellow` from `away_yellow`.

2a. Compare `home_foul_advantage` by the variable `covid_match`

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2b. Compare `home_yellow_advantage` by the variable `covid_match`

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2c. Thinking about your results

Consider the results of 2b. and 2c. together. What can we learn about hypothesis for this question? What conclusion would you draw? And what other information do you think you would need to be more confident in your conclusion?

Question 3: Considering deviations (25 points)

3a. Selecting your data

Pick any team and create two subsets using the `filter` verb of just matches featuring that team, one subset for home, the other for away.

3b. Finding the average

Make a table of the averages of goals (home and away, separate columns), shots (home and away, separate columns), possession % (home and away, separate columns), and pass % (home and away, separate columns). Show your code using the `#| echo: true` code block option.

3c. Normalizing the data

Add a row to your table; find how many z units each of the averages are away from the overall averages in the dataset. Show your code using the `#| echo: true` code block option.

3d. Thinking about your results

Interpret your results - what do the z scores indicate about the offensive capabilities of the team? How does it vary by home vs. away? What other kind of data would you like to have to answer this question?

Question 4: Your own investigation (25 points)

4a. Selecting your own question

Similar to the previous questions, think of your own question that you would like to ask of the data, ideally one that goes deeper into one of the questions considered above. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means.

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

4b. In summary

Sum up everything that you have learned from questions 1-4. Do not simply repeat/rephrase your previous results but try to say something larger that synthesizes the results together to draw a more meaningful general conclusion.