Unit 2 homework instructions

DKU Stats 101 Spring 2025 Session 3

AUTHOR
Anonymous

PUBLISHED
February 4, 2025

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 art for a good score, but I will expect you to look up basic information, such as "what does art typically sell for?" and "what is a standard size for paintings"? 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.

Technical

- Make sure your graphs are produced using `ggplot()`, are well labeled, and are easy to read.
- Make sure your tables (including regression 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 HTML 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 HTML file.
- Convert your HTML file to PDF using the Microsoft Print to PDF option in the Print menu (PC) or the PDF button option from the Print menu (Mac)
- Make sure to accurately mark each page a question answer appears on when submitting on GradeScope.
- Delete the Scoring Guide section of the instructions before final rendering and submission.

Shanghai used house price analysis

Introduction

Question 1: Describing your data (10 points)

1a. Where is this data from?

For this dataset, describe the data according to the five Ws & *how* defined in the textbook Chapter 1.2. What are some possible problems with the *who* and *what* of the dataset?

The dataset you are using for this assignment is a subset of the original dataset that can be found [here](#).

1b. What are the variable types?

For the following variables, please make a table.

One column should be the variable name, the second should be the variable type as defined in the textbook Chapter 1.3, and the third the units of the variable (if applicable). Note that you can find the units in the `sh.house.raw.sample.hw2.csv` file, though that dataset has different cases than the one you are given for your homework.

- `id`
- 标题链接
- 居室数
- 总面积
- 居民楼总层数
- 小区均价
- 物业费用

- 建造年份
- 楼层分布
- 区
- 近地铁
- 价格

Question 2: Association (20 points)

2a. Investigating average community price vs. green area: association

Using the Think-Show-Tell framework from the textbook (example on page 213), please examine the relationship *in association terms* between average price of a community and the percent of green area a community has. How strongly are they associated?

Note: for this question and all other Think sections in the homework, you do not need to report the W's of the data (you have already completed this in Q1)

Think

Show

Tell

2b. Investigating average community price vs. year built

Using the Think-Show-Tell framework from the textbook, please examine the relationship *in association terms* between the average price for community and the year built. How strongly are they associated?

Think

Show

Tell

2c. Thinking about your results

Consider the results of 2a. and 2b. together. What can we understand about average community price in Shanghai? Why do you think you observe these results?

Question 3: Simple regression (20 points)

3a. Investigating price vs. area

Using the Think-Show-Tell framework from the textbook, please examine how the management fee of an apartment is related to the price of the apartment.

Think

Show

Tell

3b. Checking model fit

Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think again

Revising conclusions

3c. Investigating price vs. apartment floor

Similar to 3a. and 3b., fully analyze the relationship between price and apartment floor.

Think

Show

Tell

Think again

Revising conclusions

3d. Thinking about your results

What can we learn about the determinants of apartment prices in Shanghai from these two investigations? Do the results surprise you? What lurking variables do you think could be at work here, if any?

Complete up to here for Homework Check - due January 26th at 11:59 pm.
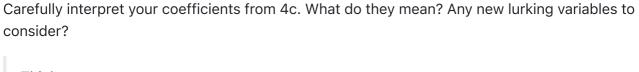
Question 4: Multiple regression (30 points)

4a. Investigating area + floor and price

Using the Think-Show-Tell framework from the textbook, please examine how area and floor of an apartment are related to price. Make use of all the tools described in the textbook to assess model fit in the `Think again` section - if it is necessary to revise your model, do it in the `Think again` section. Then state any updated conclusions in the `Revising conclusions` section.

Think

Show

Tell

Think again

Revising conclusions

4b. Interpreting coefficients of 4a. model

Carefully interpret your coefficients from 4a. What do they mean? Are there any lurking variables here?

Think

Show

Tell

4c. Add the variable near a subway

Now add the variable near a subway to your model and analyze the relationship similar to what you did in 4a.

Think

Show

Tell

Think again

Revising conclusions

4d. Reinterpret your coefficients

Carefully interpret your coefficients from 4c. What do they mean? Any new lurking variables to consider?

Think

Show

Tell

4e. Thinking about your results

Consider the results of 4a.-4d. together. What can we learn about used housing prices in Shanghai? How did your conclusions change from 3d.? Why do you think they changed?

Question 5: Your own investigation (20 points)

5a. Selecting your own question

Develop your own model of used housing price in Shanghai. Use the Think–Show–Tell procedure to conduct your investigation. Think deeply about what your result means and interpret your coefficients carefully.

Think

Show

Tell

Think again

Revising conclusions

5b. In summary

Sum up everything that you have learned in this investigation. 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.