STAT 542 – Homework 5 Example Solutions

Problem 1 in the Chapter 8 Exercises (in Section 8.12 of the book).

Answers may vary, but one possible comment is that creating more than one account is misrepresenting the scientist’s identity and defeating the purpose of the website’s business model which allows a limited amount of free access but may want users to pay a premium for unlimited access.

Problem 2 in the Chapter 8 Exercises (in Section 8.12 of the book).

Answers may vary, but the benefit of such research may be overshadowed by the ethical and privacy issues of using real people’s personal information to predict something very sensitive.

Problem 5 in the Chapter 8 Exercises (in Section 8.12 of the book).

This is an instance of multiple testing, where numerous exploratory inferences are being made in the hopes of finding something significant. Ideally, some type of adjustment for the multiple tests should be done, to try to prevent “false discoveries”.

Problem 8 in the Chapter 8 Exercises (in Section 8.12 of the book).

Posting screen name, location, and demographic information runs the risk of identifying a person even if name and email address were removed. There is a strong need for privacy when a sensitive variable such as “preference for relationships” is part of the data set.

Problem 9 in the Chapter 8 Exercises (in Section 8.12 of the book).

Answers may vary, but the result could guide consumers of this research toward harmful stereotypes about employment roles for men and women. Also, algorithms have been known to reinforce and accentuate biases that exist in society, so we should be cautious about unthinking use of machine learning algorithms in guiding policy decisions.

Problem 3 in the Chapter 14 Exercises (in Section 14.8 of the book).

The look of the plot may vary slightly, but here is one plot that captures the requested information:



Problem 6 in the Chapter 14 Exercises (in Section 14.8 of the book).

From this plot shown below, we can see a few things: First, there is a positive association between lot size and price, although it is not completely linear. For smaller lots, the price rises steeply as lot size increases. For larger plots, the price rises much less steeply (the slope flattens out) as lot size increases. In terms of the marginal distributions of each variable: The distribution of lot size is skewed to the right, with a few large lots way out in the right tail. The distribution of house prices is also skewed to the right, with most of the prices being smaller numbers, but a long tail of higher prices.



Problem 7 in the Chapter 14 Exercises (in Section 14.8 of the book).

See website for the solution code. The eventual plot should look something like:

