Lab-08

Answer Key

March 17, 2022

Preface

The goal of this assignment is to help you gain more familiarity with using **ggplot** to visualize relationships. In this lab we provide less scaffolding and more open-ended questions. As always, please come to office hours and reach out to your teaching staff if you have any questions.

FYI: in this assignment we added `\newpage` before some questions to keep question statements, code blocks, and output together.

Data

We will work with data on Airbnb listings from Inside Airbnb.[1] For this lab, we will use numeric reviews data contained in `review_summary.csv`. Each row corresponds to a listing `id`, and variables summarize all the reviews for that listing. For context, here is the reviews page for the first listing in the data. In the top left corner you can see that Airbnb reviews include an overall rating (`review_scores_rating`) and several sub-ratings for specific things (e.g., cleanliness, stored in the column `review_scores_cleanliness`). Start by importing these data and assigning them to a name.

[1] Inside Airbnb is a mission driven activist project with the objective to: *Provide data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals.*

Part 1: Exploring the Data

1. We'll start by making a simple visualization of a relationship in the reviews data. Make a scatter plot of `review_scores_rating` on the y-axis and `review_scores_accuracy` on the x-axis. What can you learn from this visualization about the individual ratings and their relationship?

2. To help see patterns more clearly, modify your plot from 1 by setting `alpha = 0.05` for points and adding a smoothed conditional mean using `geom_smooth()`.

3. Use `GGally::ggpairs` to make a scatter plot matrix to visualize pairwise relationships for the review scores for accuracy, cleanliness, and location. Which of these pairwise combinations is most correlated? Which is least correlated?

Part 2: Extracing Insights from the Data

In this part of the lab, we will use the reviews data to learn about consumer preferences. Put yourself in the shoes of an entrepreneur who wants to build an Airbnb hosting business. Before making any investments, you need to develop a strategy for your business. You know that reviews are important, and so you want to find ways to boost your ratings (so you can then increase your occupancy rates and/or prices in order to stack loot and join the FIRE movement). But your capital and time are limited so you want to figure out how to allocate them. In order to maximize your **overall** ratings, should you prioritize accuracy, cleanliness, checkin, communication, location, or value?[2]

4. Estimate a univariate model for `review_scores_rating` as a function of `review_scores_accuracy`. Use `summary()` to print the results. How do you interpret the relationship between these variables in practical terms?

```
# estimate univariate model


# summarize univariate model
```

[2]For now let's keep things simple and put aside the fact that your costs would vary by location and that value ratings will depend on price, which you are also going to be choosing...

5. Estimate a multivariate regression with `review_scores_rating` as the dependent variable and all the other `review_scores_*` variables as independent variables. Make a coefficient plot with point estimates and confidence intervals. Omit the intercept. Order the coefficients in the plot according to the magnitude of the estimate. Which sub-rating has the largest coefficient? Which has the smallest?

```
# estimate multivariate model


# tidy the estimates


# make a coefficient plot
```

6. Make a plot that visualizes marginal effects from both regression models, focusing on the relationship between `review_scores_rating` and `review_scores_accuracy`. For the multivariate visualization, hold all the other sub-ratings at their means. You can choose how to present them. For example: you could overlay them on one plot with a common y-axis, use faceting to make two separate plots, use the approach from the temperature/snow example in class to create a single plot with two subplots, or make two separate plots. When making your choice, think about what you want to convey and how you can do that accurately and effectively.

7. Use your results from 5 to revisit the qualitative question from 4: How do you interpret the relationship between `review_scores_rating` and `review_scores_accuracy` using the multivariate regression results? How does this affect your understanding of the univariate regression results?

8. What do the multivariate regression results tell you about the preferences of reviewers? How could you use them to inform your business strategy?