SOC6707 Intermediate Data Analysis, Winter 2022

Assignment 2

Due date: 28 February, 11:59pm

Details

There are **100 points** in total.

You will need to submit both your answers to the questions and accompanying R code. You should submit:

- your R Markdown file; and
- the knitted PDF resulting from your R Markdown file.

Please submit both files via Quercus.

Remember to:

- Label the answers to each question
- Label any graphs clearly with suitable axis labels and titles
- Comment your code so that it is easy to understand

Question 1 (50 points)

This question relates to the Airbnb dataset. This contains variables describing Airbnb listings in Toronto as of 7 December 2019.

a)

Create a histogram of price by room type, with all histograms shown on the same chart but colored in different colors. Interpret the graph descriptively.

Note: for readability, I suggest:

- changing the y-axis scale to be density, not frequency; and using `position = dodge` so that the bars are shown next to each other (e.g. `geom_histogram(aes(y = ..density..), position = 'dodge')`)
- changing the x-axis so it displays on the log scale, i.e. `scale_x_log10()`

b)

Create a boxplot of price by whether or not the host is a superhost. Interpret the graph descriptively.

c)

Calculate the correlation of price and overall rating (`review_scores_rating`) separately by room type. Interpret your results.

d)

i) Run a simple linear regression of price versus overall rating. Interpret the coefficient and significance on `review_scores_rating`.

ii) Run a simple linear regression of log(price) versus log(overall rating). Interpret the coefficient and significance of `log(review_scores_rating)`.

e)

Run a multiple linear regression of log(price) with covariates `room_type`, `log(review_scores_rating)` and `host_is_superhost`. Interpret the coefficients and significance

f)

Compute a correlation coefficient between the model residuals from e) and `log(review_scores_rating)`. Interpret the results.

Question 2 (50 points)

This question relates to the `income_payments` data set. These data refer to a set of 1000 fathers who are divorced and required to pay child support payments. The `income` variable refers to the father's income, the `payment` variable refers to the amount of child support payments paid monthly.

The fathers were asked to respond to a survey, and the `surveyed` variable is TRUE if they responded to the survey and FALSE otherwise. Note that this dataset is unique in that we observe income and payments for all fathers – in most cases, we would only observe information on income and payments for those fathers who are surveyed.

In this question we will be investigating the relationship father's income and child support payments, and how this differs for the whole sample compared to just those fathers who were surveyed.

a)

Read in the dataset and create two new columns, `log_income` and `log_payments`, that is the log transform of the `income` and `payment` variables.

b)

Create the following graphs:

i) A histogram of income
ii) A histogram of log income
iii) A scatterplot of income (x axis) and payment (y axis)
iv) A scatterplot of log income (x axis) and log payment (y axis)

Interpret the graphs descriptively in a few sentences. Do you think there's a relationships (i.e. a positive or negative correlation) between father's income and child support payments? Why or why not?

c)

Fit the following simple linear regression model:

$$Y_i = \alpha + \beta X_i$$

where Y_i is `log_payments` and X_i is `log_income`. Interpret the sign and significance of the slope coefficient.

d)

Create the following histograms i) `log_income` by whether or not fathers were surveyed or not, with both histograms shown on the same chart but colored in different colors ii)`log_payments` by whether or not fathers were surveyed or not, with both histograms shown on the same chart but colored in different colors

Interpret both graphs descriptively.

e)

Create a scatterplot of log income (x axis) and log payment (y axis) for just those fathers who were surveyed. Interpret the graph descriptively.

f)

Fit the same model as in c) above, but this time just use data from those fathers who were surveyed. Interpret the sign and significance of the slope coefficient.

g)

Comment briefly on what conclusions you make from this analysis, keeping in mind that usually we would only have data (and therefore be making inferences) based on surveyed fathers. Are all fathers equally likely to respond to the survey? Why or why not? How does survey response affect the conclusions we make from our analysis?