SOC6707 Intermediate Data Analysis, Winter 2021

Assignment 1

Due date: 5 February, 5pm

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 (30 points)

This question relates to the GSS dataset. We will be looking at how age at the time of first birth varies by education and current age. Note there are a few different education variables in the GSS dataset but for this question, we will be focusing on the binary `has_bachelor_or_higher` variable.

a)

Report the following descriptive statistics:

i) What proportion of respondents have a non-missing observation for their age at the time of the birth of their first child?
ii) What proportion of respondents have a non-missing observation for their highest level of education?
iii) For those respondents who have a non-missing education value:

- What is the number of respondents by education group (at least a Bachelor's degree, less than a Bachelor's degree) that have a non-missing observation for age at first birth?
- What is the proportion of respondents by education group (at least a Bachelor's degree, less than a Bachelor's degree) that have a non-missing observation for age at first birth?

Comment briefly on your calculations.

b)

For parts b) and c), we will be looking at the subset of respondents who have an education level reported, so you can filter out those respondents who have missing values of education.

Plot histograms of age at first birth by education level (at least a Bachelor's degree, less than a Bachelor's degree), with both histograms shown on the same chart but colored in different colors. Use `geom_histogram(position = "dodge")` so that the histograms are plotted next to each other. Interpret your chart.

c)

i) Calculate the correlation between age and age at first birth. Interpret your finding.
ii) Create a variable called `age_group` which groups the continuous `age` variable into 5-year age groups (hint: there is code to do this in Lab 2b).
iii) Calculate the mean age of first birth by age group and education level (`has_bachelor_or_higher`)
iv) Create a line chart of the results from part iii), plotting mean age of first birth (y axis) versus age group (x axis), with a separate line (and different color) for education level. Comment on your chart. Does the pattern over age agree with your findings from part i)? Why or why not?

Question 2 (20 points)

This question relates to the country indicators dataset.

Choose two different countries and describe, with the aid of at least two graphs, the country's fertility rate (TFR) and child mortality rate, including the levels, trends over time, and the relationship between the two quantities. Note that the two variables are

- `tfr` = total fertility rate, which is the average number births per woman in that particular country and year
- `child_mort` = under-five child mortality rate, which is the number of deaths to children aged 5 or less per 1,000 live births.

Question 3 (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.