SOC6302 Winter 2023

Assignment 3

Due date: Friday 31 March, 11:59pm

Details

- There is one assignment question worth **50 points** in total.
- In addition, there are details on what and how to submit your research proposal and EDA for the project. These are due at the same time as Assignment 3.

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

- your Quarto file; and
- the knitted PDF resulting from your Quarto 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)

Plot the the model residuals from e) (y-axis) versus `log(review_scores_rating)` (x-axis). Interpret the graph descriptively.

Research project

The goal of the research project is to carry out a regression analysis on a research question of interest using one of the following datasets (which are available on Quercus):

- `gss`: the Canadian General Social Survey, 2017
- `cchs`: the Canadian Community Health Survey, 2017-2018
- `census`: data from the 2011 Canadian Census

Also on Quercus are files that give further information on the variables in the datasets.

You will need to hand in a short **research proposal** and your **exploratory data analysis**.

Notes:

- Please submit research proposal and EDA as a separate document to Assignment 1 (there will be a separate part of Quercus).
- These should be submitted as a Quarto document; please also submit the resulting pdf.

Research proposal

Please describe

- your research question(s) of interest, and why they are of interest
- any hypotheses you may have
- the dataset you plan to use
- your main dependent variable of interest
- your main independent variables of interest

This should be no longer than 1 page.

Exploratory Data Analysis

Carry out some exploratory data analysis, given your research question of interest. This should be no longer than 3 pages, including graphs. Here are a few things to keep in mind:

- **General characteristics of dataset**: for example, how many observations, how were the data collected (is the dataset representative of the population of interest?)
- **Missing data**: If your dataset does not have any missing observations, then fine to just say this (don't need to do EDA graphs or discuss). If you have missing observations, summarize what is missing, and give a brief discussion about whether or not you think missingness may be a problem (e.g. is there more likely to be missing data for some groups compared to others?)
- **Summary statistics of variables of interest**: for example, you could present a table of summary statistics of main variables, including things like number of observations, mean/median/sd (if a continuous variable), proportions by group, etc…
- **Graphs showing both univariate and bivariate patterns**: Remember back to EDA lecture about appropriate graphs to show patterns in different types of variables. We are likely to be interested in both univariate patterns (e.g., the distribution of continuous variables, proportions for categorical outcomes…) and bivariate patterns (e.g. scatterplots, proportions/boxplots by group).
- **Pick one to three key graphs**: You could keep going ad infinitum. But good EDA reports will just pick a few key graphs to show key relationships/patterns, and have a good discussion of these. If you put in 20,000 graphs things become hard to distill and understand. I know you have probably done 20,000 graphs to pick the 1-3 graphs that go into the report.

- **Discuss what you see**: Good reports will have a good discussion about patterns in the graphs and what they potentially mean. In most cases this is more than just one sentence. If patterns (or absence of patterns) are surprising, you can note that down.