Experimental Design for Data Science

INF2178 Technical Assignment 2

Due: March 9th at 11:59pm on GitHub (**in the Assignment 2 folder**)
Note: You **MUST** name your files in the format FIRSTNAME-LASTNAME-A2.ipynb and FIRSTNAME-LASTNAME-A2.pdf (example: VICTORIA-CHUI-A2.pdf). You will be penalized if you do not.

In this assignment, you will be exploring a dataset and performing analysis using **one-way AND two-way ANOVAs**. We are looking for knowledge of quantitative analysis through the use of ANOVAs in Python, and a theoretical understanding of their use through the write-up. You will need to **submit both a .ipynb** Jupyter notebook (with comments throughout) **and a .pdf** (with a short write-up/narrative). To submit these on GitHub, please make a pull request at the appropriate location, following the instructions in the GitHub reference video on Quercus.

Data

It can be difficult to find licensed or unlicensed child care for children in Ontario due to high fees and low availability of spaces for children. Toronto Children's Services found that 75% of families cannot afford child care, to which the provincial government pledged 100,000 new child care spaces from 2016 to 2026[1].

In this assignment, you will examine Toronto licensed child care centres by using a dataset (titled INF2178_A2_data.xlsx) that collects information on the operation and capacity of these centres for multiple age groups (updated February 2024). The first tab of the data file provides a comprehensive list of different centres, demographic information and classifications, and locations. This tab also provides each centres daily capacity information. The second tab of the dataset details what each of the features mean.

Instructions

1. Examine the dataset. Not every column is going to be useful for you in this assignment; however, you are welcome to utilize every column if you are interested.
2. **Python**: Perform quantitative analysis using **one-way AND two-way ANOVAs**. Utilizing your knowledge of ANOVAs and the code from class, perform quantitative analysis as you see fit to examine the differences in the continuous variables, based on the categorical ones of your choosing.
    a. Feel free to calculate overall centre capacity numbers to create another continuous variable.
    b. Create a narrative, tell us something about the data. What story could you tell from this preliminary analysis, or what further analysis would you need to do to explore the research question(s) you have in mind?
    c. In your code, you need to create interaction plots, test the assumptions for running a one-way/two-way ANOVA, and complete post-hoc tests to earn the maximum points.

3. **PDF**: Write a short narrative (no longer than 6 pages IN TOTAL - including figures and tables) explaining your process and what you learned from the data.
    a. Include headings and sub sectioning as needed.
    b. Your figures/tables should be professional. Tip: **You do not want to provide screenshots of your code/output/data frames** in your write up and you do not need to mention the various functions or Python libraries you used to conduct your analyses. We will see this from your code.
    c. Your write up needs to include research questions, interaction plots, results of testing the assumptions for running a one-way and two-way ANOVA, ANOVA results, and post-hoc test results for maximum points.

Marks Breakdown

This assignment is worth 20% of your final grade. The grading will be broken down as follows:

% of Assignment	**Item**
40%	Functionality of code and use of appropriate code (we run your code)
10%	Code comments
40%	Narrative of findings (**research questions**, results, discussion…)
10%	Successful submission to GitHub and naming of files

IMPORTANT Feedback from A1 to Improve A2

Check your assignment 2 write-up for these common errors **before submission**

1. P-value cannot be 0, instead should be stated as < 0.001.
2. What were your research questions and did your analysis answer them?
3. Figure text (legends/axis labels) should be in large enough font and nonoverlapping. Use figure and table numbering.
4. Dataframes, summary statistics, and t-stat/p-value should be in properly formatted tables instead of screenshots. Round your p-value/t-stat to 2 or 3 significant figures.
5. Instead of using verbatim column names, use a meaningful phrase to represent what the variable means.
6. Put your name on your write-up pdf and code notebook!!!!
7. Use the dataset that was uploaded to Quercus when running any code analysis. I.e. if we provide an excel, run an excel file. **Do not create your own csv file**. We grade your code by loading the dataset we provide and running all your code, so if we cannot run your code then you will lose marks. A good step to take before submission is to restart your kernel and run all cells to check you have not missed defining variable names or packages etc.