MIE 1624 Introduction to Data Science and Analytics – Fall 2022

Assignment 1
Due Date: 11:59pm, Oct 16, 2022
Submit via Quercus

Background:

For this assignment, you are responsible for answering the below questions based on the dataset provided. You will then need to submit a 2-page report in which you present the results of your analysis. In your report, you should use visual forms to present your results. How you decide to present your results (i.e. with tables/plots/etc.) is up to you, but your choice should make the results of your analysis clear and obvious. In your report, you will need to explain what you have used to arrive at the answer to the research question and why it was appropriate for the data/question. You must interpret your final results in the context of the dataset for your problem.

Dataset:

Kaggle has hosted an open data scientist competition in 2021 titled "**Kaggle ML & DS Survey Challenge.**" The purpose of this challenge was to "*tell a data story about a subset of the data science community represented in this survey, through a combination of both narrative text and data exploration.*" More information on the competition, data, and prizes can be found on:
https://www.kaggle.com/c/kaggle-survey-2021/data

The dataset provided (**kaggle_survey_2021_responses.csv**) contains the survey results provided by Kaggle. The survey results from 25973 participants are shown in 369 columns, representing survey questions. Not all questions are answered by each participant, and responses contain various data types.

In the dataset for Assignment 1, column Q25 "*What is your current yearly compensation (approximate $USD)?"* contains a numerical target variable. Rows with null salaries have been dropped. (Please refer to **clean_kaggle_data.csv**). **You should work with the clean dataset for this assignment.**

Questions:

The objective of this assignment is to explore the survey data to understand **(1)** the nature of women's representation in Data Science and Machine Learning and **(2)** the effects of education on income level. The following tasks should be completed:

1. **[3pts]** Perform exploratory data analysis to analyze the survey dataset and to summarize its main characteristics. Present 3 graphical figures that represent different trends in the data. For your explanatory data analysis, you can consider Country, Age, Education, Professional Experience, and Salary.

2. **[4pts]** Estimating the difference between average salary (Q25) of men vs. women (Q2).
    a. **[0.5pts]** Compute and report descriptive statistics for each group (remove missing data, if necessary).
    b. **[0.5pts]** If suitable, perform a two-sample t-test with a 0.05 threshold. Explain your rationale.
    c. **[1.5pts]** Bootstrap your data for comparing the mean of salary (Q25) for the two groups. Note that the number of instances you sample from each group should be relative to its size. Use 1000 replications. Plot two bootstrapped distributions (for men and women) and the distribution of the difference in means.
    d. **[0.5pts]** If suitable, perform a two-sample t-test with a 0.05 threshold on the bootstrapped data. Explain your rationale.
    e. **[1pts]** Comment on your findings.
3. **[5pts]** Select "highest level of formal education" (Q4) from the dataset and repeat steps **a** to **e**, this time use analysis of variance (ANOVA) instead of t-test for hypothesis testing to compare the means of salary for three groups (Bachelor's degree, Master's degree, and Doctoral degree) [**0.75pts** for **a; 0.5 pts** for **b; 2pts** for **c; 0.75 pts** for **d; 1pt** for **e**].

Submission:

1) Produce a 2-page report explaining your response to each question for the given data set and detailing the analysis you performed. When writing the report, make sure to explain for each step, what you are doing, why it is important, and the pros and cons of that approach.

2) Produce an IPython Notebook detailing the analysis you performed to answer the questions for the given data set.

What to submit:

1. Submit via Quercus a Jupyter (IPython) notebook containing your implementation and motivation for all the steps of the analysis with the following naming convention:

    lastname_studentnumber_assignment1.ipynb

    Make sure that you **comment** on your code appropriately and describe each step in sufficient detail. Respect the above convention when naming your file, making sure that all letters are lowercase, and underscores are used as shown. **A program that cannot be evaluated because it varies from specifications will receive zero marks.**

2. Submit a report in PDF including the findings from your analysis. Use the following naming conventions **lastname_studentnumber_assignment1.pdf**.

Late submissions will receive a standard penalty:

- up to one hour late - no penalty
- one day late - 15% penalty
- two days late - 30% penalty
- three days late - 45% penalty
- more than three days late - 0 mark

Tools:

- **Software:**
  - **Python Version 3.X** is required for this assignment. All libraries are allowed but here is a list of the major libraries you might consider: Numpy, Scipy, Sklearn, Matplotlib, Pandas. **Specify `!pip install <library_name>` commands in the first cell of your notebook for all the 3rd party libraries you use.**
  - No other tool or software besides Python and its component libraries can be used to touch the data files. For instance, using Microsoft Excel to clean the data is NOT allowed.
  - Read the required data file from the same directory as your notebook: for example, pd.read_csv("clean_kaggle_data.csv").

- **Required data files:**
  - **clean_kaggle_data.csv**: survey responses with yearly compensation.
  - The data file cannot be altered by any means. The Jupyter notebook will be run using a local version of this data file. Do not save anything to file within the notebook and read it back.

Other requirements:

1. A large portion of marks is allocated to analysis and justification. Full marks will not be given for code alone.

2. Output must be shown and readable in the notebook. The only files that can be read into the notebook are the files posted in the assignment without modification. All work must be done within the notebook.

3. The notebook should be presentable, do not show large amounts of raw output.

4. Ensure the code runs in full before submitting. Just before you submit, rerun the entire notebook (navigate to Kernel -> Restart Kernel and Run all Cells). Ensure that there are no errors.