Introduction to data analysis - Spring 2023

Mini-Project

Submission guidelines:

- Submission deadline: 08/08/23 at 23:55.
    - The submission box in Moodle will close 48 hours after this deadline. To avoid penalties for late submissions (as stated in the syllabus), submit by this deadline.
- Submission in pairs only (unless special permission is given by the head TA)
- Submission must include (at least) two files (**not a single zip**):
    - One jupyter notebook with your answers to parts 1, 2, 3, and 4 with both markdown and code cells. Markdown cells should contain brief explanations of your analyses. No need to elaborate in this file - you will do that in the PDF - but it should be clear enough that we know which questions the code relates to and what it does. Code cells must enable complete reproduction of all your results. Your code should be clearly documented.
    - One PDF of exported content of your jupyter notebook. The PDF has to contain your outputs, similar to homework submission.
    - One PDF file with your answers to part 5.
    - You should merge the exported PDF and the PDF for part 5. You may use online free tools to do that, like:

      https://tools.pdf24.org/en/merge-pdf

    - As listed in the syllabus, if you use generative AI tools, you must also submit a (third) docx file with details on your usage. Refer to the syllabus for details.
- File names must be in the following format: final_project_ID1_ID2.pdf, final_project_ID1_ID2.ipynb .
- You must adhere to the syllabus when you prepare and submit your work. Any deviation from the syllabus and/or guidelines herein will result in point deduction.
- You can write in either Hebrew or English, but it's better to use the language you are more comfortable with.

Instructions

For the course mini-project, you will work with a dataset of your choice (from a set of possible datasets) to answer questions you are curious about.

IMPORTANT NOTES, please read carefully:

- In all of the following analyses, you will likely need to make some choices regarding what variables to include, whether to do some pre-processing (e.g., addressing missing values, generating new variables), what techniques to use in the analysis, etc. Clearly state each decision you made, explain why you made it and what might have been alternative choices.
- You can earn up to 15 bonus points for your project (can reach a maximum of 115 points) if you do a particularly thoughtful analysis, involving either an additional (unprovided, but somewhat complimentary) dataset, or analysis of a complex data type (e.g., text). Note, simply using complex data or an additional data source does not guarantee a bonus. If you ask an interesting question and think of an original way to address it, that will get you the extra points.

Part 1: Choose a dataset

Choose one of the following datasets:

- Adult Income dataset: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset
- US Estimated Crimes dataset: https://www.kaggle.com/datasets/tunguz/us-estimated-crimes
- Diabetes Prediction dataset: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
- US Married Couples in 1976 dataset: https://www.kaggle.com/datasets/utkarshx27/labor-supply-data
- World Air Quality Index dataset: https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates
- Mobile Phones dataset: https://www.kaggle.com/datasets/artempozdniakov/ukrainian-market-mobile-phones-data
- Students Exam Scores dataset (Only use the version in the file named 'Expanded_data_with_more_features.csv', not the version in the other file): https://www.kaggle.com/datasets/desalegngeb/students-exam-scores?select=Expanded_data_with_more_features.csv

Once you have chosen your dataset:

1) State which dataset you chose.
2) Provide a brief (2-4 sentences) description of the dataset. What is this dataset about?
3) List the features in the dataset and their types.
4) List the number of records in the dataset.

Part 2: Exploratory data analysis

In this part, you will do an initial exploration of the dataset you chose. This part should serve the next parts. That is, you should look at variables that can influence your analyses for parts (3) and (4). Of course, you can (and probably should) also explore further, and/or use this as a way to motivate questions for parts (3) and (4). You should explain why you are exploring the particular variables you chose.

1) Show plots illustrating the distribution of at least 5 variables in your dataset. Comment on anything interesting you observe.
2) Show plots illustrating bivariate relationships for at least 2 pairs of variables. Explain what you observe (e.g., positive/negative correlation, no correlation, etc.).

Part 3: Estimation and hypothesis testing

In this part, you will formally test a hypothesis using your data.

1) What is the question you want to explore? Why is it interesting to you?
2) Clearly state your null hypothesis and alternative hypothesis.
3) Run a test and report the results in a comprehensive way.

Part 4: Prediction/clustering

In this part, you will see how well you can address a classification problem or a clustering problem using your data. Choose one of the following two options.

Option 1: classification

1) What do you want to try to classify? Why? What is a potential application of an algorithm that classifies your target variable?
2) Clearly state what is the target variable (class) you are trying to predict, which variables (features) you are using to predict the class, and why you chose these variables..
3) Use kNN for the classification task and report the results.

Option 2: clustering

1) Why do you want to form clusters of the data? What is a potential application of the output of your clustering?
2) Clearly state which variables (features) you are using for clustering, and why you chose these variables.
3) Use K-means for the clustering task and report the results.

Part 5: Communication and reflection

Write a report summarizing your work. Your report should include the following sections:

- QUESTIONS: What are the questions you wanted to explore? Why are they interesting to you?
- DATASET: Describe the dataset you use; Explain why it is appropriate for answering these questions.
- ANALYSIS & FINDINGS: What analyses did you conduct to answer your questions? What did you find? (support with plots, but no code here). **This part should summarize everything you've done in parts 2-4. A person reading this should be able to understand the questions you asked, the analysis you've done, and the results, without looking at the jupyter notebook.**
- LIMITATIONS: What are some limitations of your analyses and potential biases of the data you used? How might these biases affect your findings?
- FUTURE DIRECTIONS: What new questions came up following your exploration of this data? Describe at least one question that could not be answered using your data alone, and specify what additional data you would collect to address it.