DIT863 Project

2024-11-20 to 2025-01-18
Maximum points: 40
Please contact Yinan (yinan@chalmers.se) if you have any questions

Deliverables

- A detailed report describing the results of the following tasks.
- Runnable Python code that reproduces the results presented in the report. Make sure the code is well-documented and easy to follow, as a lack of clarity may result in a deduction of points.
- Any datasets you have used.

Remarks:

- Be prepared to defend and explain your solutions. You may be asked to present your results and reasoning during a follow-up review if we have any questions. Points will be deducted if you cannot adequately justify your approach or the steps taken in your solutions.
- The submission should be individual. However, pairs are allowed if Python skills are insufficient. In such cases, Python code may be shared, but each individual must submit their own report. In the report, please clearly explain each person's contribution.

General description

This project involves the following steps:

Step 1. Identify datasets:

- Find datasets that interest you and align with your exploration goals. Use your creativity and curiosity to select datasets that inspire insightful analyses.
- Explain why the selected datasets are particularly interesting to you. **Max points: 1.**
- If you cross reference multiple datasets, please explain your reasoning. **Max points: 2.**

Step 2. Visualize data and compute descriptive statistics:

- Analyze your selected datasets by creating visualizations and computing descriptive statistics. This step provides an initial understanding of the data and its characteristics. **Max points: 5.**
- Clearly explain the reasons for choosing these analyses. **Max points: 3.**
- Explain what you observe from data. **Max points: 3.**

Step 3. Formulate hypotheses:

- Based on your observations from Step 2, propose **three hypotheses** driven by your interests, domain knowledge, and insights from the data. Make sure that the hypotheses are logically connected and contribute to a larger, cohesive question. **Max points: 5.**
- For example, "The average income correlates with education level across different regions". Avoid common-sense hypotheses that lack novelty, e.g., "More rain increases umbrella sales".

Step 4. Validate or invalidate your hypotheses with statistical analysis:

- Formulate your hypotheses as statistical hypotheses. Perform statistical analyses to test these hypotheses. Illustrate the statistical significance and provide a clear, logical explanation of your findings. **Max points: 8.**

Step 5. Extend with machine learning predictions:

- Define a task that involves machine learning models. Clearly motivate why the task is relevant and interesting. **Max points: 2.**
- Break down the task and phrase it as Research Questions (RQs). **Max points: 1.**
- Develop at least three machine learning models to make predictions relevant to the RQs, compare their performance statistically, and explain the trade-offs and your preferred model. Highlight statistical significance in your analyses. **Max points: 5.**

Step 6: Final conclusion and potential pitfalls:

- Present a comprehensive conclusion that synthesizes the entire project. **Max points: 2.**
- Provide in-depth analysis of potential pitfalls and their impact on the study. **Max points: 3.**

Suggested data sources

- World Bank: https://datacatalog.worldbank.org/search?q=&sort=last_updated_date%20desc
- City of Rochester: https://data.cityofrochester.gov/search?tags=demolition
- Any other similar data sources of your choice
- Alternatively, you may choose a simpler dataset (e.g. from scikit-learn) and focus on creating a standard report. Make sure you follow the steps correctly. You will get a passing grade.

Scoring criteria

Total points: 40

1. Problem justification**:** Clearly justify your research questions and hypotheses.
2. Novelty and depth**:** Propose hypotheses that demonstrate originality based on your unique interests and data exploration. Avoid overly simplistic or obvious hypotheses.
    - The depth and complexity of the hypotheses will influence the grades you get.
3. Dataset cross-referencing:
    - It is recommended to cross-reference multiple datasets, ideally at least three, to enhance the depth of analysis.
    - While it is acceptable to use a single dataset, higher grades require integrating insights from at least two datasets. For example, comparing economic data with education data to uncover insights.

Note: You may be asked to present and defend your report if we have additional questions. This does not require additional preparation. Simply be ready to explain and discuss your work. Please do not use generative AI models to complete the work on your behalf and ensure the results are your own.

Good luck and have fun!