

Homework 1

Data Analysis and Machine Learning with Python

Problem description:

Please use two datasets and answer their corresponding questions. Each question requires an explanation of your thoughts and actions, along with relevant evidence or charts if possible. Additionally, not all questions have standard answers.

Dataset 1: housing_data.csv

Field Descriptions:

'Area': House area,

'No. of Rooms': Number of rooms,

'No. of Bathrooms': Number of bathrooms,

'Location': Location of the house (Suburb, City Center, or Rural),

'Miles (dist. between school and house)': Distance between the house and the school,

'Rent Price per Month': Monthly rent price,

'Sell Price': House selling price

Q1. What steps will you take upon receiving this dataset before commencing data analysis?

Q2. If you are to inquire about Q1 from ChatGPT or Bing, what responses will you receive? Do you find them reasonable? If not, how will you rectify it?

Q3. If you are restricted to renting a house, which one or ones will you select, and why?

Q4. Assuming you have enough funds to purchase a house, will you opt to continue renting or proceed with a purchase? If renting, which one will you choose? If buying, which one will you select? Why?

Q5. Are there any properties with rent or selling prices that seem unusually high or low? Why?

Dataset 2: family_data.csv

Field Descriptions:

'Family': Family ID,

'Member': Member (Adult or Child),

'Income': Annual income,

'Spend': Annual spending

Q1. Which family boasts the highest annual income, and which has the lowest? How do you ascertain this?

Q2. Which families do not possess adequate annual income to cover all members' spending? What is the maximum shortfall? How do you determine this?

Q3. Are there any single-parent families, where only one Adult is present? Are there any childless families? How do you discern this?

Q4. Do you suspect any errors within this dataset? Examples may include negative figures, missing or duplicate data, etc. Why?

Q5. Can ChatGPT or Bing assist with the aforementioned four questions? If so, to what extent? How do you issue commands to the AI tool? If not, why not?

Submission details:

please submit all the required files in a folder, named as "<studentID>_hw1", then zip it as a .zip file, named as "<studentID>_hw1.zip", kindly check the example below.

- required:
 - 1. your report, named as "<studentID>_report.pdf" (example: b12345678_report.pdf)
 - 2. your source code:
 - if you are using .ipynb, please submit as two files, named as "Housing_Main.ipynb" and "Family_Main.ipynb"
 - if you are using .py, please name your two main codes as "Housing_Main.py" and "Family_Main.py", all the other source code can be named as you pleased.
- optional:
 - 1. a README file, specifying how to execute your code, as well as your program's dependencies.
- not required:
 - 1. the datasets. Please assume the datasets are located under the same directory as your main code, **DO NOT** include the dataset in your submission. (20% score will be deducted on any violation)

example submission:

```
/b12345678_hw1.zip
---/b12345678_hw1
    ---/b12345678_report.pdf
```

---/Housing_Main.ipynb
---/Family_Main.ipynb
---/ ... <all other source code>
---/README.md (optional)

Homework policies:

- **deadline:** before class on 7th week
- We encourage you to discuss or ask questions at **Homework 1 Discussion**, but you can still email TAs and professor.
- You can discuss the homework with your classmates, but you **CANNOT plagiarize** your classmate's codes and reports. If we find **two files are exactly the same**, the **score** for these two files are **0**.
- The result in your report must be reproducible with your source code.