Practical Assignment

General Information

Description: Your assignment comprises of two parts. In the first part, you are going to preprocess and analyze the market historical data set of real estate valuation that were collected from Sindian Dist., New Taipei City. In the second part, you are supposed to prepare the Machine Learning model that can estimate the house price based on the selected input data.

Tools: The assignment shall be prepared in Python, in Jupyter Notebook. Students may choose any libraries they want, but it is recommended to use the libraries that we covered during classes (Pandas, PyPlot, Sci-Kit Learn)

Solution: Solutions should be provided in a form of Jupyter Notebook. Each solution should contain:

- Names and student ID numbers
- Python code that solves the tasks
- Results of the code (tables, charts, accuracy metrics)
- Answers for the questions **in English** (Task 7)

Dataset Description

The data comprises of three files:

Transaction.csv – this file contains information about the real transactions. The columns in this file are as follows:

- AgentId – Foreign key that points out the agent from 'agents.csv'. This column does not contain real data – agents and agencies are the artificial data, prepared for the purpose of this exam.
- X1 – the transaction date in decimal date format
- X2 – the house age (unit: year)
- X3 – the distance to the nearest MRT station (unit: meter)
- X4 – the number of convenience stores in the living circle on foot (integer)
- X5 – the geographic coordinate, latitude. (unit: degree)
- X6 – the geographic coordinate, longitude. (unit: degree)
- Y – house price of unit area (10,000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared). This column will be your target value (output) for the second part of exam

Agents.csv – this file contains the data about agents that were responsible for the transactions. The columns in this file are as follows:

- AgentId – Primary key, unique id of every agent.
- AgencyId – Foreign key that identifies agency from 'agency.csv'
- First Name – first name of the agent
- Last Name – last name of the agent

Agency.csv – this file contains the data about agencies. The columns in this file are as follows:

- AgencyId – Primary key, unique id of each agency.
- Name – Name of the Agency

TO DO

Part 1 - data analysis

Task 1 – 5 points

Prepare a table that presents how many transactions were conducted by each agency. The results should be sorted by the number of transactions - descending. Each row of your table shall contain at least a number of transactions and the name of the agency.

Task 2 – 5 points

Prepare a table that presents the mean house price of the unit area for each agent (mean value of column Y for each agent). The results shall be sorted by the mean price - descending.

Task 3 –5 points

Prepare a table that presents the mean house price of the unit area for each agency and for each year. The results shall be sorted by the year (ascending) and then by the name of the agency.
Hint: You may want to add a new column with year only, to solve this task.

Task 4 – 5 points

Prepare a chart that presents the results of task 3.

Part 2 – Machine learning

For this part, your goal is to prepare, train and test the Machine Learning model that can estimate the house price of unit area.

Task 5 – 10 points

Select the features and labels (X and y) from your dataset. Make sure that your selection is reasonable. Split the data into two datasets (training and testing).

Task 6 – 10 points

Prepare the Machine Learning model, then train and test this model. For this task, you should provide your code in Python as well as the accuracy of the model. As an accuracy, you should report ***Mean Absolute Error*** and ***Mean Absolute Percentage Error***. You may also report other metrics (R2, SMAPE, MSE) but this is not necessary.

Note: You do not have to achieve the best possible results, but your results should be reasonably good.

Task 7 – 10 points

Analyze your results and answer the following questions:

- Is your model overfitting? (Yes, No, cannot say), explain your answer.
- Is your model underfitting? (Yes, No, cannot say), explain your answer.

For this task, you should present the answers to the questions above.