CPE 232 Data Models (GRADING CRITERIA)

Midterm Exam (2nd Portion: Coding)

Please read the instructions below carefully.

1. This portion of the midterm exam is open-book.
2. Students are allowed to work individually only. Communication of any forms (e.g. chat, discord, messenger, etc.) is NOT allowed. Those who are involved in plagiarism, of any kind, will receive zero for the midterm.
3. Students have until 15.30 to complete this coding portion.
4. Use Jupyter Notebook to write and run your code, along with giving an explanation if asked.
5. For your submission, upload the notebook file (.ipynb) and its corresponding PDF on LEB2. It is recommended that you start uploading your submission 10 minutes before 15.30 to avoid any technical error. *Submissions that are marked late will not be accepted.*

Tasks

Use this dataset to perform the tasks below.

1. [8 points] Display information of the data: *size, shape, number of dimensions*, and *overview information*.
   - ➔ 2 pts for each display. All or nothing.

2. Display statistics (min, max, average, S.D.) of the following:
   - 2.1. [3 points] All attributes
     - ➔ 3 pts. All or nothing.
   - 2.2. [6 points] Selected attributes: `Price, Landsize, Propertycount`
     - ➔ 2 pts for each attribute. All or nothing.
   - 2.3. [6 points] Selected attributes with a specific condition:
     - ■ `Landsize` < 500
       - ➔ 3 pts. All or nothing.
     - ■ `Bedroom2` =2 and `Bathroom` =1 and `Car` =1
       - ➔ 3 pts. All or nothing.

     Remarks for #2.3: Whether you performed the above conditions "combined" or "separately", both are acceptable.

3. Inspect if there are any missing values; and If there are, perform the following cases:
   - 3.1. [5 points] Use *the original data*, remove *rows* that contain missing values. Display data shape and overview information after the removal.
     - ➔ 3 pts for removing rows

- ➔ 2 pts for displaying post-removal shape and info

3.2. [12 points] Use *the original data*, replace missing values *with zeros*. For those columns that contain missing values, compare the average value of the before versus after replacement. Based on this result, briefly discuss if this method should be used, and why.
- ➔ 3 pts for replacement with zeros
- ➔ 5 pts for comparison
- ➔ 4 pts for discussion

3.3. [12 points] Use *the original data*, apply data imputation for *numeric* columns (use average), and remove rows that contain missing values for non-numeric column(s). Compare the average value of the before versus after replacement for those imputed numeric columns. Based on this result, briefly discuss if this method should be used, and why.
- ➔ 3 pts for mean imputation
- ➔ 2 pts for removing rows of non-numeric column
- ➔ 4 pts for comparison
- ➔ 3 pts for discussion

[Note: Use the data from 3.3 for task #4 and onwards]

4. [3 points] Format the datetime of the attribute `Date` to YYYY/MM/DD
- ➔ 3 pts. All or nothing

5. [15 points] Create a pie chart to demonstrate *unique* values of the attribute `Method`. In your visualization, also display chart title, percentage of distribution, and a legend.
- ➔ 5 pts for obtaining unique values of `Method`
- ➔ 10 pts for pie chart (-1 each for missing title, % distribution, legend)

6. [15 points] Create two bar charts to present these attributes: `Rooms` and `CouncilArea` individually. For each visualization, also display the chart title, a legend both axes, and data labels. Make sure that data labels do not overlap one another.
- ➔ 5 pts for `Rooms` bar chart
- ➔ 5 pts for `CouncilArea` bar chart
- ➔ 5 pts for chart info (-1 each for missing title, legend, label)
- ➔ -2 if labels overlap

7. [15 points] Group the data by `Regionname` and `Type`, then display the sum of these attributes: `Price`, `Bedroom2`, `Bathroom`, `Car`, and `Landsize`. The expected output looks like the following snapshot.
- ➔ 3 pts for using group-by
- ➔ 2 pts for computing sum (after grouped)
- ➔ 10 pts if output matches the snapshot (i.e. column headers)
  - ◆ If not matched, give 5 pts for the attempt

		Price	Bedroom2	Bathroom	Car	Landsize
Regionname	**Type**					
Eastern Metropolitan	h					
	t					
	u					
Eastern Victoria	h					
	u					
Northern Metropolitan	h					
	t					
	u					
Northern Victoria	h					
South-Eastern Metropolitan	h					
	t					
	u					
Southern Metropolitan	h					
	t					
	u					
Western Metropolitan	h					
	t					
	u					
Western Victoria	h					