CITS 2401
Computer Analysis and Visualisation

Project 2

Data Analysis and Visualisation of Weather Prediction

Worth: 15% of the unit
Submission: (1) your code and (2) your data analysis and visualisation report on the quiz server.
Deadline: **18th October 2024, 11:59 PM**
Late submissions: late submissions attract a 5% raw penalty per day up to 7 days (i.e., 25th October 2024, 11:59 PM). After that, the mark will be 0 (zero). Also, any plagiarised work will be marked zero.

1. Outline

In this project, we will continue from our Project 1 where we implemented a weather prediction detection system. But instead of implementing the features (which we completed in Project 1), we will now focus on data analysis and visualisation skills to better present what our datasets contain. For this project, you will be given a dataset(**`weather_data.csv`**) that contains weather record with timestamp. Your task is to perform the following steps (more details in the tasks section):

- Data analysis
- Data visualisation
- Write data analysis and visualisation report

Note 1: This is an individual project, so please refrain from sharing your code or files with others. However, you can have high-level discussions about the syntax of the formula or the use of modules with other examples. Please note that if it is discovered that you have submitted work that is not your own, you may face penalties. It is also important to keep in mind that ChatGPT and other similar tools are limited in their ability to generate outputs, and it is easy to detect if you use their outputs without understanding the underlying principles. The main goal of this project is to demonstrate your understanding of programming principles and how they can be applied in practical contexts.

Note 2: you do not necessarily have to complete project 1 to do this project, as it is more about data analysis and visualisation of the datasets you are given.

2. Tasks

You program must define the function `main()` with the following syntax. All other functions will get called inside the main function:

```
def main(fileName):
```

The input arguments for this function are:

- `fileName`: The name of the CSV file (as string) containing the record of weather for different timestamps. The first row the CSV file will contain the headings of the columns. A sample CSV file "weather_data.csv" is provided with project sheet on Moodle.

To begin, you need to define a `main(filename)` function that will read the dataset, call all other functions and return following results of Task 1, Task 2 and Task 3(i).

```
var, medianResult, corr, pca, task2i, task2ii, task3i = main(fileName)
```

CITS 2401
Computer Analysis and Visualisation

Task 1: Data Analysis using NumPy **Mark: 15**

Answer the following five NumPy related tasks for data analysis. These will require use of NumPy functions and methods, matrix manipulations, vectorized computations, NumPy statistics, NumPy where function, etc. To complete this task, write a function called `task1(data)`, where `data` contains all records from the dataset which we retrieved in main function. The `task1(data)` function **returns four values** from the following four questions (i - iv). Return all results rounded to two decimal points.

Input:

```
var, medianResult, corr, pca[0:5] = task1(data)
```

output:

```
0.03 [7.79, 23.56] -0.14 [[-1.37][-1.41][-1.2 ][-1.66][-1.53]]
```

i. `var`: Filter weather record based on `Mostly Cloudy` weather summary and calculate the variance of Humidity. Use sample variance formula for calculation.

Output:

```
print(var) = 0.03
```

ii. `medianResult`: Filter data based on Rain (Yes) and then calculate the median of `Temperature` that are in the lower 25th (inclusive) and upper 75th (inclusive) percentile.

Output:

```
print(medianResult) = [7.79, 23.56]
```

iii. `corr`: Filter weather record where summary is `Mostly Cloudy` and `Temperature` is greater than `Apparent Temperature`. Then, using filtered data calculate dot product between `Wind Speed` and `Wind Bearing` and then perform correlation between the resultant vector and `Visibility` column.

Output:

```
print(corr) = -0.14
```

iv. `pca`: Create a `N x 5` matrix where `N` is number of rows in the dataset and `5` is the number of columns, we will call these features (`Summary`, `Rain`, `Temperature`, `Apparent Temperature`, and `Humidity`) (before that you need to convert all string values to numerical values. You can assume there will always be 6 `Summary` and use the following values – Partly Cloudy: 1, Mostly Cloudy: 2, Overcast: 3, Foggy: 4, Clear: 5, and Breezy and Mostly Cloudy: 6 and two `Rain` status Yes: 1, No: 2). Calculate principal component analysis (PCA) to reduce the dimensionality of data to `N X 1`.

The algorithm for PCA is:

a) Standardize the data along all the features (subtract mean and divide by standard deviation over the feature dimension).
b) Calculate the covariance matrix for the features

c) Perform eigen decomposition on the covariance matrix to get eigenvectors (principal components) and eigenvalues
d) Sort the eigenvectors based on their eigenvalues from highest to lowest
e) Select top k eigenvectors (k=1)
f) Transform the data using the selected eigenvectors (dot product of eigenvectors and Standardized data in step a)

Output:

```
print(pca.shape) = (999, 1)

print(pca[0:5]) = [[-1.37][-1.41][-1.2 ][-1.66][-1.53]]
```

Task 2: Data Analysis using Pandas **Mark: 5**

Answer the following 2 Pandas related tasks for data analysis. To complete this task, write a function called `task2(data)`, where `data` contains all records from the dataset. The function **returns a list from the task 2(i) and value from task 2 (ii)** from the following questions. Return all results rounded to two decimal points.

Input:

```
task2i, task2ii = task2(data)
```

output:

```
[9.47, 8.29, 9.22, 8.77, 13.77, …] False
```

i. Using pandas, filter the weather record that have `Temperature` lower than the mean value of `Apparent Temperature` and `Rain` is Yes.

Output:

```
print(task2i[:5]) = [9.47, 8.29, 9.22, 8.77, 13.77]
```

ii. Using pandas find whether there are any `NaN` values in the dataset or not. If there are any `NaN`, replace them with zeros. Return a Boolean True if there were any `NaN` or False otherwise.

Output:

```
print(task2ii) = False
```

Task 3: Data Analysis using SymPy **Mark: 5**

Answer the following two `SymPy` related tasks for data analysis. These will require use of `SymPy` functions. To complete the task 3(i), write a function called `task3(data)`, where `data` contains all records from the dataset. The function **returns a 2-dimension list** containing values from the following question. Use another function (`multivariate()`) to complete task3(ii).

i. Filter the weather record based on Rain category Yes and No; then calculate the derivatives of `Temperature` for both rain category using the following formula.

```
expr = 2.5x**3 + 3*x**2 + 3.5*x + 5
```

Input:

```
task3i = task3(data)
```

output:

```
print(task3i[0][0], task3i[1][0])= 733.26 716.08
```

ii. Write a function `multivariate()` that solves following multivariate problems and **prints the result**.
Multivariate problems: Given the following objective function, solve it for multiple variables (Y and Z) using `sympy`.

$$X = Y^{0.25} + Z^{0.34}$$

Output: []

Note: Return all results rounded to two decimal points. However, do not round during any calculation.

Task 4: Data Visualisation using matplotlib **Mark: 10**

You must demonstrate following 5 matplotlib related skills for data visualisation. To complete this task, write a function called `task4(data, pca)`, where `data` contains all records from the dataset and `pca` values you calculated in Task 1. The function should display the following plots. The presentation of the visualisations (e.g., customising labels, points etc.) will determine your fluency in data visualisation skills.

i. Plot `Temperature` values using bar plot for weather record where `Wind Speed` is below 10 and `Visibility` is above 9.
ii. Extract weather data where `Rain` is `No` for different `Summary` separately and display their total `Temperature` using a pie plot. Number of different `Summary` will vary.

iii. Extract data from all weather record for different `Summary` separately and display the `Humidity` values using a boxplot. Number of different `Summary` will vary.
iv. Create a `N X 2` matrix where `N` is number of weather records in the dataset, 2 is the index for `Temperature` and `Apparent Temperature` values. Plot the matrix using scatter plot.
v. Create a colour code using the values(`pca`) in the matrix created in Task 1(iv). Plot `Wind Speed` values where values are over 8 as a bar plot using the corresponding colour codes. [You can use the rounded value of `pca` to create the colours].

Task 5: Write a summary report **Mark: 10**

Write a summary report for your data analysis and visualisation. The report presents your findings and recommendations using above findings and articulate your understanding of the datasets. You should clearly explain methodology (how), and results (what). The report should also address how the data analysis might extract some inherent patterns in the dataset. Use appropriate style to present your explanation in a story like format that will enable the reader to understand how the above data analysis might help better understand weather prediction. Add your suggestions. Report should be within **800-1000** words excluding code.

Note: Ensure to include your name and student ID in the report!

CITS 2401
Computer Analysis and Visualisation

3. Submission

Submission items (1) and (2) – Code and Report

Submit your whole code (tasks 1 to 5) in the quiz answer box by the due date (18 October 2024 11:59 pm, drop dead due date 25 October 2024 11:59 pm with 5% raw penalty per day), containing all functions, objects etc., as well as attaching the python file containing all the code you wrote for this project. You should name the file as [student id]_P2.py. For example, if your student ID is 12345678, then your file name is 12345678_P2.py. Similarly, attach your report as a PDF format ONLY on the quiz server.

Final remarks: make sure you have the module docstring for your project code, indicating your name and your student ID number along with the description of the project.

Failure to follow these instructions will be regarded as ***NO SUBMISSION*** *(i.e., you will receive 0 for this project)*

4. Rubrics

	Criteria	Highly Satisfactory (D, HD)	Satisfactory (P, CR)	Unsatisfactory (N)
Data Analysis (25 marks)	• Understand the use of Python for data analysis. • Demonstrate the ability to write and execute Python code for data analysis.	**Demonstrated the ability to analyse data using Python fluently:** • Correct use of five or more NumPy related skills for data analysis as appropriate in articulate ways. • Meaningful data analysis results.	**Demonstrated the ability to analyse data using Python:** • Correct use of four or five NumPy related skills for data analysis. • Mostly meaningful data analysis results.	**Failed to demonstrate the ability to use Python for data analysis:** • Less than two correct uses of NumPy related skills for data analysis. • The data analysis results are not intuitive/hard to understand.
Visualization (10 marks)	• Understand the use of Python visualisation tools. • Demonstrate the ability to visualise data in Python.	**Demonstrated the ability to use Python for visualisation fluently:** • Correct use of matplotlib for visualisations with advanced visualisation outputs. • Visually appealing format for various data analysis results.	**Demonstrated the ability to use Python for visualisation:** • Correct use of matplotlib for visualisations. • Appropriate visualisation format for various data analysis results.	**Failed to demonstrate the ability to use Python functions:** • Incorrect use of matplotlib for visualisations. • Inappropriate visualisation format for different data.
Report (10 marks)	• Understand the concept of data analysis and visualization. • Demonstrate the ability in writing.	**Demonstrated the ability to write a report:** • Advanced level of referencing to the coding and theoretical concept. • Detailed suggestion to further data analogy.	**Demonstrated the ability to write a report:** • Referenced coding and theoretical concept. • Provided some suggestion to further data analogy.	**Failed to demonstrated the ability to write a report:** • Scant referencing to the coding and theoretical concept. • Zero or minimal suggestion to further data analogy.
Coding Style (5 marks)	• Code is written in accordance with the style guideline. • Code is written legibly and is of high standard.	**Demonstrated the ability to present Python code comprehensively:** • The coding style conforms to the style guideline with attention to details.	**Demonstrated the ability to present Python code:** • The coding style conforms to the style guideline.	**Failed to demonstrate the ability to present Python code:** • The coding style does not conform to the style guideline.

This project is worth a total of 50 marks. Please note that your submission will be manually checked before we release the final marks.