MEHARRY
SCHOOL OF APPLIED
COMPUTATIONAL SCIENCES

MSBD566 – Predictive Modeling and Analytics

Assignment 1

Instructions

1. Answer all problems listed below.
2. You can use any layout to provide the answer, but the problem numbers and answers must be arranged clearly and properly.
3. Total points are 100 points.
4. Submission **must be following the steps below and submitted through Blackboard**. Any submission through other platforms (email, chat, etc.) will not be accepted unless there is an issue on Blackboard before the deadline.
5. Homework must be submitted by **9/18/2025 11:59 pm CT**. Late submissions will be accepted up to 2 days with penalty of 5 points per day. No submission will be accepted after 8/20/2025 11:59 pm CT without prior approval.
6. AI Policy: AI can be used in terms of understanding terms and asking for examples but not directly to answer the questions.

Problems

Part 1 – Pandas

1. Create a Python Jupyter notebook (`.ipynb`) to answer this problem. Name it as **`MSBD566_<LastName>_<FirstName>_Assignment1.ipynb`**.
2. In the header, please include:
   - Name
   - Course number
   - Date
   - Honor statement:
3. Copy this problem into the notebook and answer each question using Python. All work must be present in the notebook. Print the answers nicely after calculations, plots, etc. are made.

   The Air Quality Index (AQI) is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health effects you may experience within a few hours or days after breathing polluted air. The AQI is reported according to the Environmental Protection Agency's scale. The Health Department obtains the pollen forecast from Pollen.com. Pollen forecasts are based on a variety of environmental and seasonal factors, including past and current pollen counts over the past 24 - 72 hours and the weather conditions. The pollen forecasts estimate how much pollen an allergy sufferer is likely to be exposed to in the future. The pollen forecast is currently reported on a scale of 0 to 12 as follows:

Pollen Count	**Pollen Level**
0.0 to 2.4	Low

2.5 to 4.8	Low – Medium
4.9 to 7.2	Medium
7.3 to 9.6	Medium – High
9.7 to 12.0	High

Source:
1. https://data.nashville.gov/datasets/Nashville::air-quality-and-pollen-count/about
2. https://www.nashville.gov/departments/health/environmental-health/air-pollution-control/daily-aqi-and-pollen-count

Question 1: Explore the data by plotting AQI and Category across the years in two separate plots. Make sure they have proper labels and titles. Use a datetime format for the dates.

a. How many times has the AQI been recorded above 120?
b. When did the Air Quality Index Category become unhealthy?

Question 2: Which pollen type occurs the most? Hint: You can use a scatterplot if needed.

Question 3: [Open-ended] Based on this data, do you think Nashville is a city that is comfortable to live (based on the air quality and pollen only)? Why?

Question 4: [A vectorization problem]. Compare the manual approach versus a vectorized approach to find the mean AQI for each Air Quality Category (**`'Category'`** in the table). For the vectorized method, you can use **`groupby()`** and **`mean()`** method in pandas datatype (ex: **`data.groupby('ColumnName')`** and **`dataList.mean()`**). For the manual method, a regular for-loop can be used. Calculate the time difference between the two approaches and evaluate.

Note: Another way to record the time is using the time module. An example is shown below:

```
# import time module
import time
time_start = time.time() # record the starting time

# put all your calculations that you want to time here.

time_end = time.time() # record the end time
time_taken = time_end - time_start
print(f"Time taken: {time_taken} seconds")
```

<u>Part 2 – GitHub practice</u>

1. Create a GitHub repository on your account called MSBD566. Make the repository public (you can make it private after the assignment has been graded).
2. In your repository, you should have:
   a. A folder called Assignment 1, and in the folder, add:
      i. Your Python notebook created in Part 1: **`MSBD566_<LastName>_<FirstName>_Assignment1.ipynb`**
      ii. The data file provided to run Part 1: **`Air_Quality_and_Pollen_Count.csv`**
   b. A **`README.md`** containing a short description of your repo. <u>Add the last edited date</u> on the last line of the README, something like this:

3. If you created your repository on your local machine, you can zip the whole folder for submission. If you have it only on the cloud (i.e., the GitHub website), you will need to clone or download it and upload the zip file to the Blackboard.
4. Submit your repository link AND your downloaded repository where you find this document.

Note #1: The reason we upload a copy in the submission is for future assessment as we have to have a record in the Blackboard. But it's a good thing that we also practice making our own local copy.

Note #2: If you need help on markdown syntaxes (pretty much like what you did in the Python notebook), you can refer a documentation here: https://docs.github.com/en/get-started/writing-on-github/getting-started-with-writing-and-formatting-on-github