Big Data Analytics and Security

CSEC 5311/CETE 4392

Spring 2025

TEXAS A&M UNIVERSITY SAN ANTONIO

Assignment- COVID-19 Data Analysis (US-Specific)

If you use any AI tools (e.g., ChatGPT) to answer any of the questions, you MUST cite and add a section "My thoughts on AI-generated answers" at the end of that questions.

Total Marks: 100

Objective: This assignment will test your ability to load and prepare US-specific COVID-19 data, visualize trends, and critically analyze findings. Use the file **time_series_covid19_confirmed_US.csv** from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series for the analysis.

Question 1: Load and Prepare Data (30 Marks)

Write a Python program to:

1. Load the dataset using **Pandas**.
2. Prepare the data by:
    - Aggregating the data at the state level (e.g., sum up cases for each state across all counties).
    - Transposing the date columns into a single column named Date and adding a column State for the state names.
3. Provide a table (or DataFrame) preview of the first 10 rows of the prepared dataset.

***Hint*: Use groupby() to aggregate at the state level and melt() to reshape the dataset.**

Submission: Include a screenshot of your preview along with the code.

Question 2: Visualization of Trends (40 Marks)

Using the prepared data from Question 1:

1. Identify the 5 states with the highest cumulative confirmed cases.
2. Plot the cumulative confirmed cases for these 5 states over time on a line chart.

3. Annotate the graph with:
    - Titles
    - Axis labels
    - A legend indicating the states.

Submission: Include a screenshot of your plot along with the code.

Question 3: Critical Analysis (30 Marks)

Based on your visualization and data exploration:

1. Identify one state with a distinct trend (e.g., an early sharp rise, a plateau, or a second wave).
2. Research and describe a real-world event or factor that could explain this trend (e.g., state-specific policies, population density, vaccination rates).
3. Write a **short essay (250–300 words)** explaining:
    - The observed trend.
    - The real-world events or policies contributing to it.
    - Lessons for public health or policymaking based on this analysis.

***Hint*: Go beyond surface-level information. Use credible sources to identify potential explanations for trends in specific states.**

Grading Criteria

- **Question 1 (30 Marks)**:
    - Data loaded correctly (10 Marks)
    - State-level aggregation and date conversion performed (15 Marks)
    - Clear table preview (5 Marks)
- **Question 2 (40 Marks)**:
    - Correct calculation of top 5 states (10 Marks)
    - Visualization quality (e.g., clear graph, proper labels, legend) (20 Marks)
    - Code clarity and organization (10 Marks)

- **Question 3 (30 Marks):**
    - Identification of trends or outliers (10 Marks)
    - Depth of analysis and connection to real-world events (15 Marks)
    - Writing quality (grammar, structure, originality) (5 Marks)

Submission Guidelines:

Submit a detailed report containing your code (Q1 and Q2), preview screenshots (Q1), plot screenshots (Q2), and discussion (Q3) in .PDF format. Also, attach your code file as a separate document.