CS 122 Assignment 5

Due: Nov 1 , 11:59 PM
Total Points: 100

Overview

You are an analyst working for a national transportation agency.
Your task is to analyze and compare traffic-accident trends across four U.S. states to uncover regional patterns, evaluate accident severity, and provide insights for policy and safety planning.

You will use Python (Pandas + Matplotlib/Seaborn) to perform data cleaning, aggregation, and visualization on a large real-world dataset.
Each task builds on your analytical and visualization skills.

Dataset Description

File: accident_100k.csv (~100 000 rows × 21 columns)

A portion of the dataset (≈ Aug 2017 – Apr 2019) has missing or irregular records.
You may keep the gap visible, interpolate counts, or highlight the missing period in your visualization.
Explain your chosen approach briefly in your report.

Use the following four states for all analyses:
California (CA), Florida (FL), Texas (TX), and New York (NY).

Task 1 – Time-Series Analysis per State (20 pts)

Goal: Show daily accident counts per state.

1. Convert Weather_Timestamp → Date using pd.to_datetime().
2. Group by ['State','Date'] and count accidents.
3. Plot one multi-line figure (four lines = four states).
4. Label axes clearly and include legend.

Analytical Focus:

- Identify peak periods or holidays with the most accidents.
- Compare temporal patterns across CA, FL, TX, and NY.

Task 2 – Heatmap of Accident Density by Day-of-Week and State (20 pts)

Goal: Visualize accident *density* (proportion within each state) across weekdays.

1. Extract the day name → DayOfWeek.
2. Compute normalized frequency per state (groupby('State')['DayOfWeek'].value_counts(normalize=True)).
3. Create one heatmap (X = State, Y = DayOfWeek).

Analytical Focus:

- Which states show the strongest weekday/weekend contrast?
- Discuss any unusual daily patterns.

Task 3 – Accident Severity vs Weather Category (20 pts)

Goal: Examine how severity varies under different weather types.

1. Filter to ['Fair','Mostly Cloudy','Cloudy','Clear'].
2. Create **one combined figure** containing **four subplots (2×2 layout)**—one per state.
   - Each subplot shows a **boxplot** of Severity vs Weather_Condition.
3. Optionally, include small bars showing **mean severity** per weather category on the same or a second row.
4. Title each subplot (e.g., *Accident Severity in CA*).

Analytical Focus:

- Which weather type has the highest median/mean severity?
- Which state shows the widest spread?
- Comment on outliers or flat distributions.

Submit only one combined 2×2 figure for Task 3. ***Do not submit separate plots.***

Task 4 – Histogram of Accident Severity (20 pts)

Goal: Visualize the frequency distribution of accident severity across all states.

1. Create **one combined figure** with **four subplots (2×2 layout)**—one per state.
2. Use the same bin range (1–4) and identical axis limits.
3. Label axes and include subplot titles.

Analytical Focus:

- Describe the skew (minor vs severe accidents).
- Note any spikes or gaps between states.

Submit only the combined 2×2 histogram figure. ***No separate images.***

Task 5 – Open-Ended Exploration (20 pts)

Goal: Design your own mini-analysis using at least **two variables**.

1. Pose a clear question (e.g., *Does lower visibility lead to higher severity?*).
2. Choose an appropriate visualization.
3. Include your question, plot, and short conclusion (3–4 sentences).

Analytical Focus: Your reasoning and plot choice must be logical and data-driven.

Expected Figures (5 Total)

Task	Figure	Expected File
1	Multi-line Time Series (4 States Combined)	task1_timeseries.png
2	Heatmap (Day-of-Week × State)	task2_heatmap.png
3	Box/Bar Plot (2×2 Subplots Combined)	task3_weather.png
4	Histogram (2×2 Subplots Combined)	task4_histograms.png
5	Open-Ended Exploration (1 Figure)	task5_exploration.png

Total = 5 combined images. All plots must be labeled and saved as .png in the plots/ folder.

Submission Requirements

Submit a zip named:

HW5_<YourStudentID>.zip

src/

hw5_solution.py ← all tasks 1–5 in this one script

plots/ ← five .png files

report.pdf ← one page per task with figure + insight

Grading Rubric (100 pts)

Category	Description	Pts
Data Cleaning & Preprocessing	Logical handling of missing data; correct state filtering	10
Task 1	Correct plot + interpretation	20
Task 2	Normalized heatmap + analysis	20
Task 3	Combined 2×2 box/bar figure + insight	20
Task 4	Combined 2×2 histogram figure + discussion	20
Task 5	Creative question (≥ 2 variables) + sound reasoning	10