Modou Sarr
Homework 5 Report
CMSC 320

Introduction

In this report, we examine daily class attendance alongside weather conditions to determine whether rain affects how many students show up. Our analysis first accounts for the natural downward trend in attendance over the semester by introducing a "day index." We then merge head-count data with daily precipitation records, convert trace and measurable precipitation values into a numeric form, and create a binary rain flag. To explore possible patterns, we produce three visualizations: a time-series plot with rainy days highlighted, a boxplot comparing attendance on rainy versus dry days, and a scatter plot of attendance against precipitation amount. Finally, we fit a regression model to quantify the impact of rain while controlling for the time trend

.

Background

The analysis draws on two primary data sources:

- Attendance Records: A CSV file listing each class session's date and the number of students present. These head counts were collected manually and cleaned by classmates before being shared.

- Weather Observations: A CSV file containing daily meteorological measurements from a nearby station. Key fields include maximum, minimum, and average temperatures, plus precipitation values. Trace amounts of rain are recorded as "T," which we convert to zero for numeric processing.

Because attendance naturally drifts downward over a semester, we need to distinguish that trend from any weather-related effects. By combining these datasets on the date field, we can directly compare class turnout to the presence or amount of rainfall each day.

Data Cleaning Process

Before running any models, we prepared the data in four main steps:

1. Import & Parse Dates
   Loaded the attendance CSV and the weather CSV into pandas, parsing their date columns into datetime objects.

2. Standardize Column Names
   - Stripped whitespace and lowercased all column headers so joins and lookups are consistent.

3. Process Precipitation

    - Replaced trace values ("T" or "t") with 0, converted the precipitation field to a numeric precip_mm, and created a binary flag rained = 1 whenever precip_mm > 0.

4. Merge & Engineer Trend

    - Inner-joined the two tables on date.

    - Added day_index = number of days since the first class session to capture the semester's linear attendance decline.

Analysis Methodology

To quantify the separate effects of time and rain on attendance, we fit a simple linear regression model. In plain language, we modeled each day's headcount as the sum of

- A baseline attendance level (the intercept)

- A "day index" effect on how attendance drifts over the semester

- A "rain flag" effect, which tell us whether that day saw any measurable precipitation

- A residual error term on everything else we didn't capture

Findings

Overall Model Fit

- The model explains about 10.5% of the day-to-day variation in attendance (R^2 = 0.105).

- The F-statistic (1.767, p = 0.188) suggests the model as a whole is not highly predictive, but it suffices to isolate our two effects.

- Time Trend (day_index)

    - Estimated effect: –0.55 students per day.

    - Interpretation: On average, attendance drops by about half a student each day as the semester progresses.

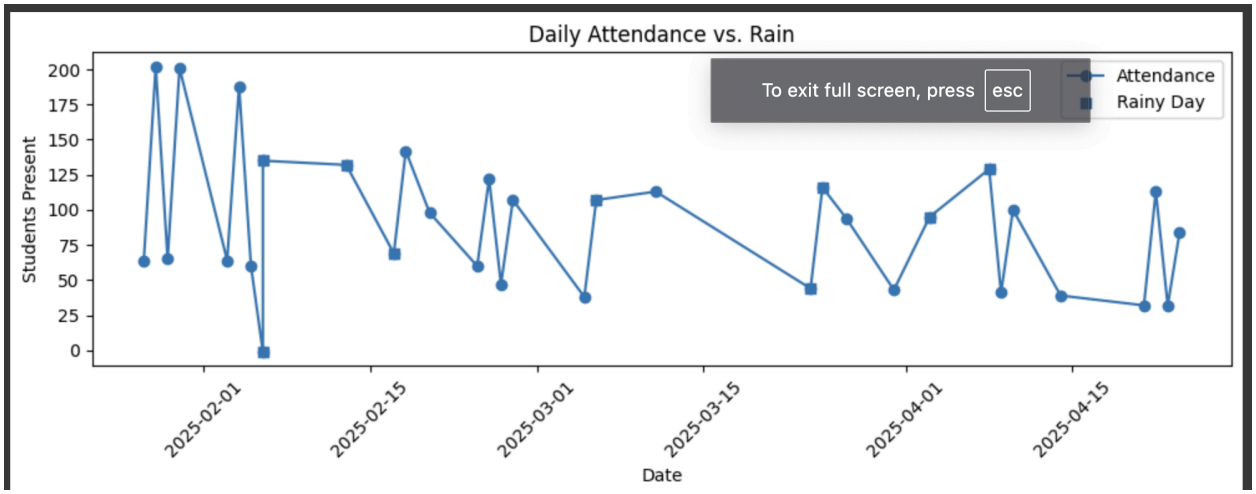
    - Statistical note: This trend is marginally significant ($t = -1.876$, $p \approx 0.07$), suggesting a real downward drift.

- Rain Effect (rained)
    - Estimated effect: +1.38 students on rainy days.
    - Interpretation: Rainy days actually show a small positive difference in attendance, but this estimate is essentially zero once you account for its very large uncertainty.
    - Statistical note: It's not significant at all ($t = 0.073$, $p = 0.94$), meaning we cannot conclude that rain changes attendance one way or another.

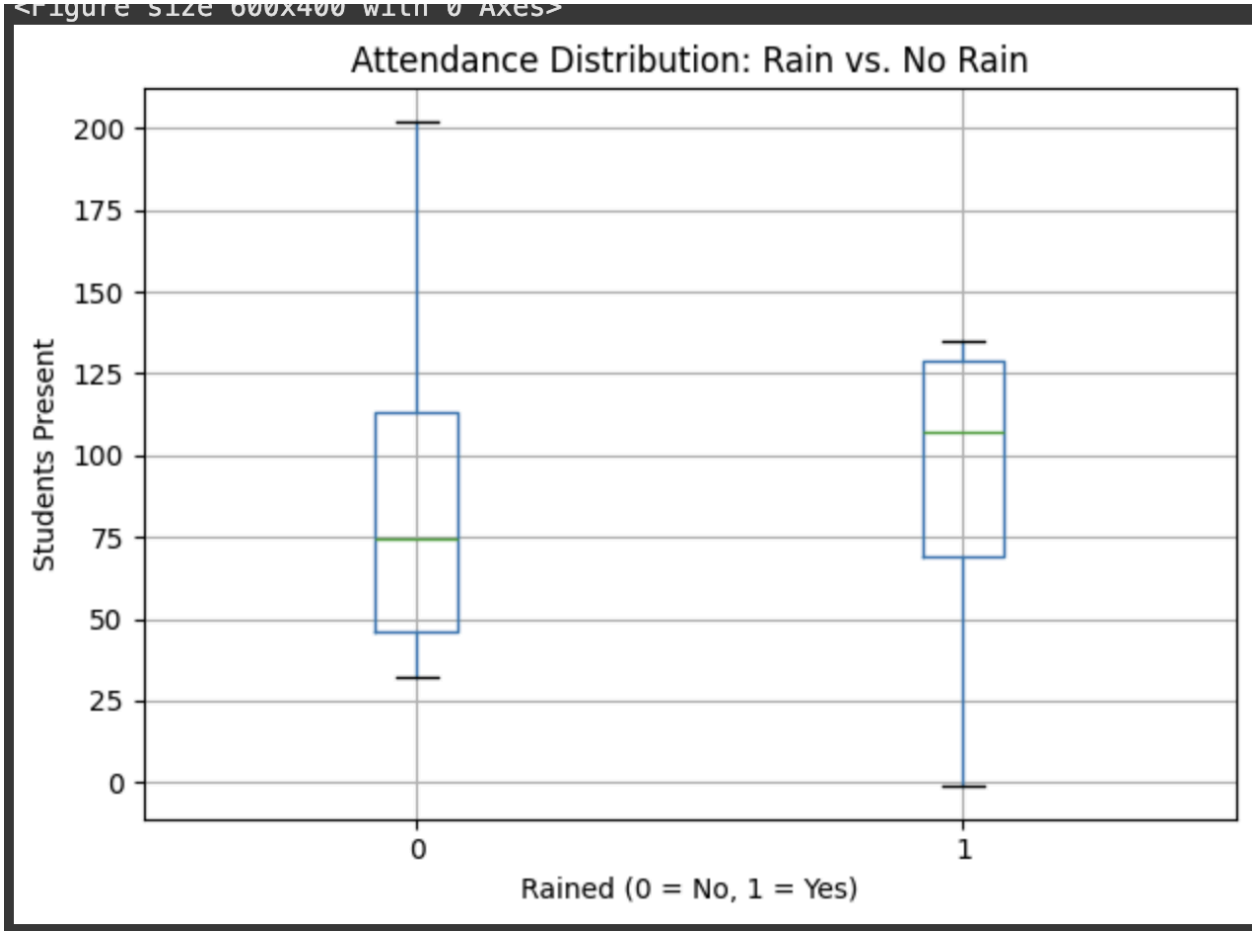
Conclusion

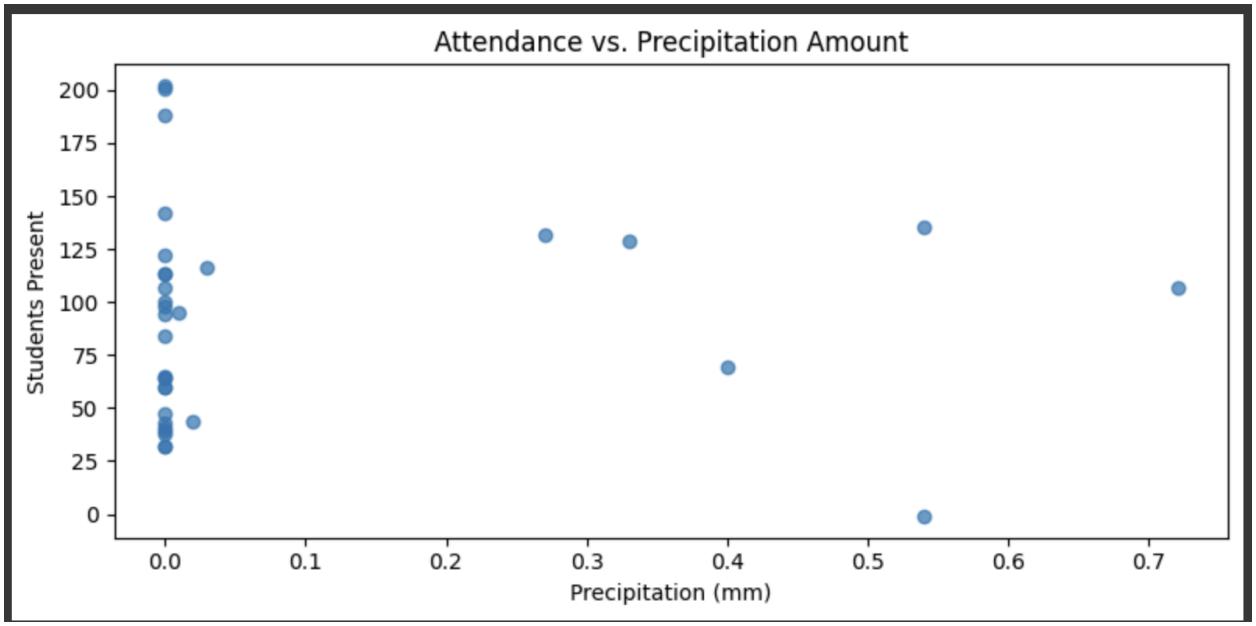
Our analysis shows that, once we adjust for the steady drop in attendance over the semester, rainy days have no meaningful effect on how many students come to class. The regression's rain coefficient was near zero and far from statistically significant, and all three visualizations (the time-series line chart, the boxplot, and the scatter plot) confirm there's no clear attendance dip on wet days.
In contrast, the "day index" trend of losing roughly half a student per day is the dominant pattern, reflecting the typical waning engagement as the term progresses.

Below are all the visualizations that were talked about.



<Figure size 600x400 with 0 Axes>




Attendance vs. Precipitation Amount
Students Present
200
175
150
125
100
75
50
25
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Precipitation (mm)