Environmental Exposure Analysis for LA Schools: My Workflow & Assumptions

How I Worked on This Project

When I first got this assignment, I was like "wow, that's a lot of data to collect!" But I broke it down into steps that made sense:

Step 1: Finding the Schools

First, I needed to find all the schools in LA. I used the California Department of Education website to scrape info about the schools. I wasn't sure at first how to define "Los Angeles" exactly, so I created a geographic box (a polygon) that I think covers the main LA area. Maybe I missed some schools on the edges, but I had to draw the line somewhere!

Some schools didnt have coordinates, which was annoying. I had to drop those from my analysis because without knowing where they are, I can't match them to the closest pollution monitors.

Step 2: Getting Pollution Data

This part was tricky! AirNow's website has tons of data files, one for each hour of every day. The assignment asked for 4 years of data (2020-2024), but I ended up only collecting data for 2024 because of time constraints and my laptop limitations. Even getting one year's worth of data took almost 900 minutes! because the website is kinda slow sometimes.

My biggest assumption here was that the nearest monitoring station to a school is representative of the air quality at that school. We know that's not completely true - pollution can vary even within neighborhoods! But it's the best we can do without having monitors at every school.

Step 3: Weather Data Collection

Similar to the pollution data, I grabbed weather info from Weather Underground for 2024 only. Again, I assumed that weather stations represent conditions across their zip codes. I know thats not perfect, but I dont think weather varies as much as pollution within small areas.
Some days had missing data, which I filled by taking averages from nearby days. Not ideal, but better than having holes in the dataset!

Step 4 & 5: Analyzing Everything Together

This is where things got interesting! I had to decide what "school hours" actually means. I went with:
- Elementary schools: 8 AM - 2 PM
- Middle and high schools: 8 AM - 3 PM

This is another big assumption - I know some schools start earlier or end later, and some have different schedules on different days. But I needed something consistent to analyze.

I also had to think about school holidays and weekends. Should I include summer vacation? I decided to focus just on official school days in 2024, because that's when students are actually exposed to these conditions.

Difficulties I Faced

The biggest challenge was definitely the scope of the project. The assignment asked for 4 years of data (2020-2024), but this was impossible for me to complete with my available time and equipment. My laptop just couldn't handle processing that much data, and the scraping would have taken weeks to complete! I made the decision to focus on just 2024 data to make the project manageable, hoping that one year would still provide meaningful insights.

The web scraping part was super challenging! Sometimes the websites would block me if I requested too many pages too quickly. I had to add random delays between requests and even use different browser headers to make it look like a real person browsing.

Another big problem was the format of data files from AirNow. Even with just one year of data, there were thousands of hourly files to download and process. My poor laptop was struggling so much with even this reduced dataset that I had to figure out ways to process it in chunks.

I also lost like a whole week of work when my computer crashed and I hadn't backed up some of my processed data files. Had to redo a bunch of calculations which was a real pain.

Figuring out the closest monitoring station to each school was harder than I thought it would be. I had to learn some geospatial calculations that wasn't covered in any of my classes before.

What I Learned

I learned so much about air quality during this project! I didnt realize how much pollution levels change throughout the day. Morning rush hour really does affect the air quality when kids are going to school.

I also learned that there's a BIG difference in exposure depending on where schools are located. Schools near major highways had much higher average PM2.5 and PM10 levels than schools in residential areas away from traffic.

Final Thoughts

If I could redo this project, I would definitely try to get more computing power or cloud resources so I could analyze the full 4 years of data as originally intended. I think having multiple years would show interesting trends and patterns that my single-year analysis can't reveal.

I would also try to incorporate more info about school schedules and maybe even indoor air quality. Most kids spend most of their school day inside, and buildings can filter out some pollution.

Also, I think a cool next step would be to look at health data for these schools - like asthma rates or absences due to illness - and see if there's any correlation with the environmental data.

Anyways, I hope this analysis helps highlight how environmental factors affect our students everyday! Maybe it can even lead to some positive changes for schools in high-pollution areas.