Project Initial plan

Authors: Zijia Tang, zt81
Peilin He, ph155
Xingyu Bai, xb30

1. Why did you choose the method you used?

We chose statistical analysis and data visualization methods because they allow us to identify correlations, trends, and geographic patterns in air pollution and asthma rates.

- **Data Cleaning**: We may need to clean and merge data from different sources effectively.
- **Exploratory Data Analysis (EDA)**: To detect missing values, outliers, and general trends in air pollution levels.
- **Statistical Inference**: We will use regression analysis and correlation tests to determine the strength of the relationship between air pollution and asthma cases.
- **Visualization (Geospatial Mapping & Time Series Analysis)**: Visualizing the levels of different pollutants over time may help us address the problem.

2. What patterns do you see in what you find interesting?

Some initial patterns we expect to see:

- Seasonal Variations: Certain pollutants (e.g., PM2.5, ozone) tend to peak during the summer months and decline to their lowest levels in winter.

3. What research topics or questions did your group generate from this brainstorming?

Through our discussions, we generated several researchable questions:

- How has air pollution changed over time in major U.S. cities, and what aspects relate to this change?
- Does it correlate with population density and urbanization?
- How have major environmental policies (e.g., Clean Air Act amendments) affected air pollution levels in different U.S. states?
- How do wildfires affect air pollution levels in different U.S. regions?

4. Which of these ideas can you see your group potentially pursuing?

We see ourselves pursuing the following questions:

- How has air pollution changed over time in major U.S. cities, and what aspects relate to this change?
- What are the potential reason(s), and what correlates with the change?

5. Do you feel like more brainstorming is needed before you find a topic?

- At this stage, we feel confident in moving forward with one of the identified research questions. However, we may need to further brainstorm data sources that will help us address the question effectively.

6. If you used an LLM (Large Language Model), how satisfied were you with its answers? Why?

We used an LLM to help refine our research questions and brainstorm possible patterns in the data.

Through talking with ChatGPT, we got three different ideas: 1. How do socioeconomic factors (e.g., parental education, family size) correlate with student academic performance in secondary education? 2. How do socioeconomic factors affect standardized test scores across different school districts? 3. How has air pollution changed over time in major U.S. cities, and what aspects relate to this change?

Then, we discussed these three topics together and analyzed how we could find data to address these questions. We discovered a dataset from the EPA that contains a substantial amount of data, allowing us to fully develop our final research question. As a result, we chose the third option.

However, we still need to validate the research direction by checking data availability and ensuring that our chosen statistical methods align with our dataset.

Part 2: Collaboration Plan

1. How will you divide responsibilities? Will some students be responsible for certain portions of the project, or will you be more integrated and decide on responsibilities on a weekly basis?

We will divide our research question into three subsections, with each team member responsible for one part. Responsibilities will be re-evaluated weekly, and we will review each other's work to ensure a balanced workload and make adjustments as needed.

2. About how much time do you expect every group member to spend on the project each week, on average? It is okay if this number is higher toward the last couple of weeks of the semester.

On average, every group member would spend 3 hours on the project.

3. When and how will you meet?

We will hold a regular Zoom meeting every Friday at 7:00 PM to review each member's progress.

4. What platform(s) will you use to communicate between meetings?

We will use WeChat for communications.

5. Where will you track who is doing what tasks and when those tasks will be done?

We have a Google document to share our ideas and write a skeleton of our whole project.

6. Where will you store data, code, writing, etc., so that all group members have easy access to shared materials?

We will store our code and writing on a GitHub repository.

7. Is your group willing to publicly share your project, for example, as part of a portfolio of work?

We will have a public Github repository. All of us would be collaborators of that GitHub repository. We will revisit on April 1 to confirm our initial plan.

Appendix

Github link: https://github.com/TzJ2006/CS216-Final-Project

GPT conversation:

Question:

Now please help me think of some ideas about the initial mind map of my class final project, to come up with a good research question for the 216 Final Project, try testing your potential questions against the following criteria:

Is your research question clear? Can the question be answered thoroughly through this project? Is your research question concise? Is your research question complex? Is your research question arguable? Can your question be tested? Is your research question analytical? Does your research question result in a description of the problem or an analysis of the problem?

Note that our group members are interested in the following aspects: Education, Health & Society, Politics & Social Trends, and Environmental Factors.

GPT's answer:

Education

- **Socioeconomic Factors and Standardized Test Scores:**
    - **National Center for Education Statistics (NCES):** Provides data on various aspects of education in the U.S., including assessments and school demographics.
    - **U.S. Census Bureau:** Offers socioeconomic data at the county and school district levels.
- **School Funding and Student Performance:**
    - **Common Core of Data (CCD):** Contains detailed information on public elementary and secondary schools, including financial data.
    - **School District Finance Survey (F-33):** Provides data on revenues and expenditures of school districts.
- **Access to Free School Meals and Academic Performance:**
    - **Food and Nutrition Service (FNS):** Part of the USDA, offers data on school meal programs.
    - **NCES:** Includes data on student participation in free or reduced-price lunch programs.
- **Homework Hours and Student Performance:**
    - **High School Longitudinal Study (HSLS):** Tracks a cohort of students and includes data on study habits and academic outcomes.

Health & Society

- **Air Pollution and Asthma Rates:**
    - **Environmental Protection Agency (EPA):** Provides data on air quality indices across various regions.
    - **Centers for Disease Control and Prevention (CDC):** Offers data on asthma prevalence and hospitalizations.
- **Income Level and Health Insurance Coverage:**
    - **U.S. Census Bureau's American Community Survey (ACS):** Includes data on income levels and health insurance status.

- **Obesity Rates and Access to Fresh Food:**
    - **USDA Food Environment Atlas:** Provides data on food access indicators and health outcomes.
    - **CDC's Behavioral Risk Factor Surveillance System (BRFSS):** Offers data on obesity rates at the state and metropolitan levels.
- **Sleep Duration and Well-being Among College Students:**
    - **National College Health Assessment (NCHA):** Collects data on various health behaviors, including sleep patterns, among college students.

Environmental Factors

- **Climate Change and Agricultural Productivity:**
    - **USDA National Agricultural Statistics Service (NASS):** Provides data on crop yields and agricultural practices.
    - **National Oceanic and Atmospheric Administration (NOAA):** Offers climate data relevant to agriculture.
- **Renewable Energy Adoption and Economic Growth:**
    - **International Energy Agency (IEA):** Provides data on renewable energy adoption.
    - **World Bank:** Offers data on economic indicators.
- **Natural Disasters and Migration Patterns:**
    - **Internal Displacement Monitoring Centre (IDMC):** Provides data on displacement due to natural disasters.
- **Public Transportation Access and Carbon Footprints:**
    - **American Public Transportation Association (APTA):** Offers data on public transit usage.
    - **Environmental Protection Agency (EPA):** Provides data on emissions and environmental impact.

Question:

Thank you for helping us brainstorm. We are particularly interested in the environment part, can you help us think of a more detailed question?

GPT's answer:

Here are some research questions that align well with your requirements:

Environmental Factors

- What is the correlation between renewable energy adoption and economic growth?
- How has air pollution changed over time in major U.S. cities, and what aspects relate to this change?
- How does climate change impact agricultural yields in different regions?
- Does public transportation availability reduce carbon footprints in urban areas?
- What role do extreme weather events play in driving migration patterns?

Possible Data Sources:

- U.S. Department of Energy renewable energy reports
- USDA crop yield data
- EPA climate change impact assessments

Potential Analytical Methods & Modules:

- Geospatial Analysis (Mapping climate change effects)
- Machine Learning (Forecasting energy adoption rates)
- Statistical Inference (Examining causation vs. correlation)

Question:

Thank you, we want to explore this question: How has air pollution changed over time in major U.S. cities, and what aspects relate to this change? Can you help us find some potential data websites?

GPT's answer:

To explore how air pollution has changed over time in major U.S. cities and identify factors contributing to these changes, you can utilize the following data sources:

1. **EPA Air Quality Data:**
    - **Description:** The Environmental Protection Agency (EPA) provides comprehensive air quality data, including historical trends for various pollutants across U.S. cities and counties.
    - **Access:** Visit the EPA's "Air Quality - Cities and Counties" page to download spreadsheets detailing air quality status and trends for Core-Based Statistical Areas (CBSAs).
      epa.gov
2. **American Lung Association's State of the Air Reports:**
    - **Description:** These reports offer insights into pollution levels, trends over time, and health-related statistics for major U.S. cities.
    - **Access:** Explore city rankings and detailed charts on the American Lung Association's website.
      lung.org
3. **Bureau of Transportation Statistics (BTS) - Air Pollution Trends:**
    - **Description:** The BTS provides data on air quality trends in selected Metropolitan Statistical Areas, focusing on pollutants like PM2.5 and ozone.
    - **Access:** Access the data and accompanying analyses on the BTS website.
      bts.gov
4. **NASA's U.S. Air Quality Trends:**
    - **Description:** NASA offers visualizations and data on the changes in air pollutant concentrations over time, highlighting significant improvements and remaining challenges.
    - **Access:** Explore the trends and download relevant data from NASA's air quality page.
      airquality.gsfc.nasa.gov
5. **Our World in Data - Air Pollution:**
    - **Description:** This platform provides global and country-specific data on air pollution, including historical trends and health impacts.
    - **Access:** Navigate to the air pollution section on Our World in Data to find interactive charts and downloadable datasets.
      ourworldindata.org

By leveraging these resources, you can analyze temporal changes in air pollution across major U.S. cities and investigate factors such as policy interventions, industrial activities, and transportation trends that may influence these patterns.