**Air Quality Project Proposal**

* **Basic Info**. The project title, your names, e-mail addresses, UIDs, a link to the project repository.

Project title: COVID-19 & Air Quality / Group Name: The Wranglers

Link: <https://github.com/Jonathan-Barton/AirQualityProject>

Members:

|  |  |  |
| --- | --- | --- |
| Name | E-mail Address | UID |
| Adriana Reyes Miranda | adriana.reyesmiranda@utah.edu | u6014167 |
| Teigen Judd | teigen.judd@utah.edu | u0955471 |
| Jonathan Barton | jonathan.barton@utah.edu | u1329000 |
| Yi-Jin Chen | yi-jin.chen@utah.edu | u1302428 |

* **Background and Motivation.**
  + This project will give us great insight into how data wrangling projects are executed. The topic at hand, COVID-19 and air quality is of extra relevance to the state of Utah, which has reoccurring air quality issues. This data could be used to predict severe COVID case numbers by monitoring AQI.
* **Project Objectives.**
  + The primary question is trying to understand the effect of air quality on COVID-19 in the US. We want to learn how to clean and format data to accomplish the data analysis. The data could be used to predict COVID severity based on air quality.
* **Data.**
  + We are obtaining the COVID-19 data from the Johns Hopkins Center from Systems Science and Engineering repository: [https://github.com/CSSEGISandDat](https://github.com/CSSEGISandData)a. This data is presented in a visual dashboard format that draws from multiple sources which include: WHO, CDC, European Centre for Disease Prevention and Control (ECDC), and several other institutions. Data is gathered for each US state from sources often provided by the state, city, or county. Data is also collected from several independent national sources. The data is licensed under Creative Commons for public use. At the request of the data providers, credit is given to: “Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1” as well as: “ Badr, H. S., B. F. Zaitchik, G. H. Kerr, J. M. Colston, P. Hinson, Y. Chen, N. H. Nguyen, M. Kosek, H. Du, E. Dong, M. Marshall, K. Nixon, and L. M. Gardner, 2021: Unified COVID-19 Dataset.” The data is available for public viewing in both desktop and mobile formats. Two data points (COVID cases and COVID deaths) are tracked at various levels of organization.
  + The air quality data is coming from the EPA: <https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html>. The EPA keeps measurements of various air pollutants at thousands of locations across the United States. This data is publicly available at EPA.gov.
* **Data Processing.**
  + **COVID data:**
  + Epidemiological data on **positivity, mortality,** and **total cases** at the regional scale (we could use zip codes maybe?)
    - Positivity = Total number of people tested positive for COVID
    - Mortality = the number of people dead because of COVID
    - Total cases = total number of COVID cases, which is the sum of dead people, tested positive and recovered
  + **Air quality data**:
  + Air Quality Index (AQI) data
    - 5 years of AQI data based on the main pollutants
    - Organize this data by geographical regions that correlate with the COVID data.
* We will do data cleanup using the python packages NumPy and pandas:
  + We will remove any unnecessary columns of data not relevant to our analysis.
  + We will clean the columns important for our analysis:
    - Find solutions to possible incomplete data.
    - Important columns will include those that tell us about the epidemiological data and the AQI data.
    - Columns that will help us look for correlation will be associated with helping us plot data based on geographical regions.
* **Design.**
* We will use python to do linear regression and analyze the data:
  + We will look for any correlation between our COVID data and our AQI data based on region
  + We will use python to display the data as a scatterplot.
* An alternative prototype could be to use a bar chart to plot the COVID and AQI data based on designated regions.
* **Must-Have Features.**
* Particulates from EPA dataset
* Mortality outcomes
* Positivity rates : 7-day moving average
* Age comparison
* Source and components of air pollution
* Relations between COVID-19 and air quality --- correlation coefficients, etc
* Spatiotemporal clustering: geographical pattern and COVID-19 cases and air quality
* **Optional Features.**
  + Other possible factors of correlation (i.e. different climate conditions and air pollution, socio-economic conditions)
* **Project Schedule**

From the timeline guide:

**February 12th**

Submit Project Proposal

**February 12th**

Project Overview and Data Description Presentation

**February 15th-19th**

Setup Call with Professor and TA to get feedback on project proposal

**March 12th**

Project Milestone Submission via Github

**March 17th**

Peer Feedback Session- PeerFeedback Guidelines

**March 26th**

Submit updated Process book with feedback incorporated if any

**April 14th**

Data Wrangling Initial Work Presentation

**April 28th**

Project Presentation

Submission of Presentation and any remaining documentation.

**May 5th**

Final Project Submission

|  |  |
| --- | --- |
| Week (Due Date) | Milestone |
| Week 1 (Feb 26) | Collect and Merge data |
| Week 2 (March 5) | Merge/Clean Data |
| Week 3 (March 12) | Submit Data |
| Week 4 (March 26) | Update dataset if need |
| Week 5 (April 2) | Charts / Integrated presentation |
| Week 6 (April 9) | Testing / Debugging |