Mithil Patel

Bellevue University

DSC540-T302 Data Preparation (2231 – 1)

Prof. Catie Williams

September 25, 2022

Milestone 1

Q1) Project Subject Area

* The purpose of our project is to conduct analytical studies investigating potential health impacts of air pollutants on individuals and the environment within countries and hope to guide public health actions.

Q2) 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)

* Excel: displays the total number of air pollution related death along with type of respiratory disease by each country.
  + <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/ambient-air-pollution-attributable-deaths>
* Website: contains a table showing the concentration of harmful particles present in the atmosphere from 2010 to 2016 by each country
  + <https://www.who.int/data/gho/data/themes/air-pollution/modelled-exposure-of-pm-air-pollution-exposure>
* API: request dataset for the air quality index ranging from 1 to 5 for each country.
  + <https://openweathermap.org/api/air-pollution>

Q3) The relationships between them, or the relationship you will make

between them

* All three dataset are related by the list of countries. A column ‘country’ will be used to combine three dataset that will help gain valuable insights. Additionally, index column from API and number of deaths from excel data will form relation to show how air pollution relates to number of deaths by country. Excel data set and website data will show weather high concentrations of particulate matter leads to higher number of deaths.
* Excel: contains a column with list of country names while the other three columns contain statistics of the number of deaths by gender.
* Website: Contains number of concentrations of particulate matter in urban and rural over the years area for each country.
* API: Contains air quality index by country along with other polluting gases concentration

Q4) What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

* Dataset is fundamental for a data scientist to tackle issues and provide confidence in a given resolution. The three datasets mentioned above will be used to perform an analytical study to evaluate the health effects of air pollution. In order to formulate our data to gain insight, numerous data wrangling techniques will be applied to our datasets for each milestone. For the excel dataset retrieved from the World Health Organization (WHO) containing the number of deaths corresponding to respiratory disease due to air pollution in each country, data filtering is required to only extract the total number of deaths for all respiratory diseases instead of by each. The excel file will be valuable considering that ambient air pollution ranks as one of the top ten risk factors and the particles involved in these emissions have the greatest effect on health. Another dataset/table from the WHO’s website involving the concentration of harmful particles present in the atmosphere from 2010 to 2016 requires data transformation. The columns will be filtered to only show the oldest and the most recent years from the data set to determine how significant the changes were. Additionally, I shall find a way to extract data from the table into a pandas data frame for readability and build a data science model. The last dataset from the OpenWeatherMap API returns the AQI (air quality index) by passing in latitude, longitude, and the start and end date as the parameters for a request. An algorithm is necessary to receive the AQI by country from API and then convert JSON output into a pandas data frame. However, since latitude and longitude are required as a parameter, an additional dataset containing longitude, latitude, and country name will be used for automating API requests ( <https://www.kaggle.com/datasets/paultimothymooney/latitude-and-longitude-for-every-country-and-state>). Another issue will be to convert regular dates (i.e., UTC) into UTI to suffice start and end date parameter input for an API request.

Q5) 250 Words describing how you plan to tackle the project, what the data means, the ethical implications of your project scenario/topic, and what challenges you might face.

* People, especially in mainstream media, have long argued for the human contribution to the rapid climate change and its overall impact on the environment. A recent news article highlights how a population’s proximity to nuclear power plants can change and alter human DNA and have a significant risk of cancer. With the power of data science, I decided to discover the truth through data to dispel the myth revolving around climate change and its impact on our society, especially on human health. To search for the truth, I will employ various datasets from different data sources such as API, excel sheets, etc. For instance, one of the data sets will address the risk of air pollution by country and related deaths due to pollutant molecules in the air. With such heavy potential implications, I am anticipating that the data sources used will bring more awareness to society. By performing transformations to the datasets, I wish to build a model that not only leads to awareness but sheds light on the controversial topic and helps to prevent further problems in the future. One challenge I may encounter while working on this project is that data may lead to biased interpretations. For instance, even with less severe respiratory diseases, under-developed countries with less advanced health care systems are more likely to die in comparison with developed countries. One assumption made for the project is that since air pollution increases every year, we should expect the number of deaths to rise because more people’s health is being compromised. However, the overall quality of health care is likely to improve in each country resulting in more lives saved. Hence, we should also factor in the advancement of the health care system in our analysis.