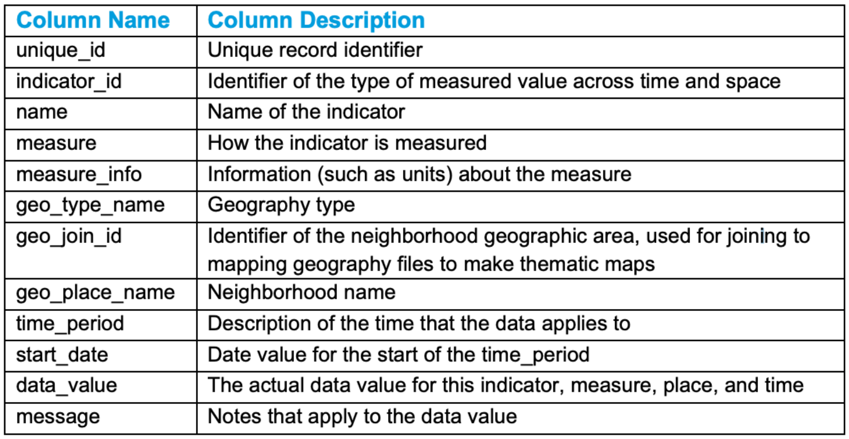
CS210 Final Project Dataset Selection Summary

4/2/2024

Group:
Dylan Legg (dl1073)
Aaditya Rayadurgam (ar1873)
Bhavya Patel (bsp75)
Maxwell Chuo (mhc88)

Dataset Description and Source

- Source: https://catalog.data.gov/dataset/air-quality
  - Dataset is managed by NYC DOHMH – Environmental Health, the Bureau of Environmental Surveillance and Policy
  - Dataset provides data on different levels of air pollutants collected from various geographical locations around NYC. These indicators provide a perspective across time and NYC geographies to better characterize air quality and health in NYC.
- Why you chose this dataset
  - Dataset includes practical data that affects millions of New Yorker's health every day.
  - It is local to us, and as frequenters of the city it is interesting to see what geographical locations in the city may be the most detrimental to human health as compared other locations.
  - In addition, we would like to see whether or not some of NYC's pollution reduction action actually has an effect.
- Dataset statistics - number of rows, columns, missing values, etc.
  - Dataset has 12 columns, 1629 rows, 0 missing values.



Column Name	Column Description
unique_id	Unique record identifier
indicator_id	Identifier of the type of measured value across time and space
name	Name of the indicator
measure	How the indicator is measured
measure_info	Information (such as units) about the measure
geo_type_name	Geography type
geo_join_id	Identifier of the neighborhood geographic area, used for joining to mapping geography files to make thematic maps
geo_place_name	Neighborhood name
time_period	Description of the time that the data applies to
start_date	Date value for the start of the time_period
data_value	The actual data value for this indicator, measure, place, and time
message	Notes that apply to the data value

How we will use the data:

Visual Display:

1. Trends over time for different pollutants.
2. Comparison of air quality before, during, and after specific pollution reduction actions.

Hypothesis we plan to test:

1. Are there higher concentrations of pollutants in areas with high traffic density or are zoned for industry?
2. Have pollution policies in NYC had an effectiveness in reducing air pollutants in certain neighborhoods?
3. Possible seasonal variations in pollutant levels due to changes in heating use, traffic patterns, or weather conditions?

Summary Statistics:

1. Mean and median levels of each pollutant by geographical area and over time.
2. Percent change in pollution levels year-over-year or before and after policy implementations.
3. Correlation coefficients between pollutant levels and potential factors such as traffic data, industrial activity, and population density.
4. Air Quality Index (AQI) calculations to assess health implications.