Air Quality Data - Documentation

**Data for Ozone** - Ozone (44201) Year 2022 - <https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual>

**Data for PM2.5** - PM2.5 FRM/FEM Mass (88101) – Year 2022 - Daily data 8 Hour Run Average <https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual>

**Data for PM10** - PM10 Mass (81102) – Year 2022 daily data - Daily data 8 Hour Run Average <https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual>  
  
The column Arithmetic Mean was taken for in all the three datasets to computing.

**Process -**

1. Geolocation Mapping: The geographical locations of the sensor sites were established based on their longitude and latitude coordinates. These locations were then mapped for individual particles (PM2.5, PM10, and Ozone).
2. Census Tract Integration: Census tracts were added to the map as a separate layer to provide additional geographic context.
3. Spatial Join: A spatial join operation was performed to associate the census tracts with the sensor sites located within their boundaries. The corresponding particle values from the sensor sites were assigned to these census tracts.
4. Data Aggregation: The above process was executed separately for each particle type (PM2.5, PM10, and Ozone). The resulting tables were then merged to create a comprehensive dataset that includes particle values for each census tract.

**Process for handling missing values using weighted K-Nearest Neighbors Imputation**

1. Centroid Calculation: The centroid of each polygon was calculated using the centroid property of the geometric objects. The x and y coordinates of these centroids were extracted and stored in separate columns in the DataFrame. These coordinates were used as the basis for identifying nearest neighbors in the KNN imputation.
2. Weighted KNN Imputation: The missing values in the target column were filled in using the K-nearest neighbors algorithm. For each entry that lacked a target value, the algorithm identified the K closest neighbors by calculating the Euclidean distance between their centroid coordinates. These neighboring entries then contributed their target values to estimate a replacement for the missing value. This estimate was produced by taking a weighted average of the neighbors' target values. Weights were assigned as the inverse of the squared distance between each neighbor, which ensured that neighbors closer to the target entry had a larger impact on the final imputed value.