Case Study #1: Predicting Annual Air Pollution – Table of features

Variable	Details
id	Monitor number – the county number is indicated before the decimal – the monitor number is indicated after the decimal **Example**: 1073.0023 is Jefferson county (1073) and .0023 one of 8 monitors
fips	Federal information processing standard number for the county where the monitor is located – 5 digit id code for counties (zero is often the first value and sometimes is not shown) – the first 2 numbers indicate the state – the last three numbers indicate the county **Example**: Alabama's state code is 01 because it is first alphabetically (note: Alaska and Hawaii are not included because they are not part of the contiguous US)
Lat	Latitude of the monitor in degrees
Lon	Longitude of the monitor in degrees
state	State where the monitor is located
county	County where the monitor is located

Variable	Details
city	City where the monitor is located
CMAQ	Estimated values of air pollution from a computational model called **Community Multiscale Air Quality (CMAQ)** – A monitoring system that simulates the physics of the atmosphere using chemistry and weather data to predict the air pollution – ***Does not use any of the $PM_{2.5}$ gravimetric monitoring data.*** (There is a version that does use the gravimetric monitoring data, but not this one!) – Data from the EPA
zcta	Zip Code Tabulation Area where the monitor is located – Postal Zip codes are converted into "generalized areal representations" that are non-overlapping – Data from the 2010 Census
zcta_area	Land area of the zip code area in meters squared – Data from the 2010 Census
zcta_pop	Population in the zip code area – Data from the 2010 Census
imp_a500	Impervious surface measure – Within a circle with a radius of 500 meters around the monitor – Impervious surface are roads, concrete, parking lots, buildings – This is a measure of development

Variable	Details
imp_a1000	Impervious surface measure – Within a circle with a radius of 1000 meters around the monitor
imp_a5000	Impervious surface measure – Within a circle with a radius of 5000 meters around the monitor
imp_a10000	Impervious surface measure – Within a circle with a radius of 10000 meters around the monitor
imp_a15000	Impervious surface measure – Within a circle with a radius of 15000 meters around the monitor
county_area	Land area of the county of the monitor in meters squared
county_pop	Population of the county of the monitor
Log_dist_to_prisec	Log (Natural log) distance to a primary or secondary road from the monitor – Highway or major road
log_pri_length_5000	Count of primary road length in meters in a circle with a radius of 5000 meters around the monitor (Natural log) – Highways only

Variable	Details
log_pri_length_10000	Count of primary road length in meters in a circle with a radius of 10000 meters around the monitor (Natural log) – Highways only
log_pri_length_15000	Count of primary road length in meters in a circle with a radius of 15000 meters around the monitor (Natural log) – Highways only
log_pri_length_25000	Count of primary road length in meters in a circle with a radius of 25000 meters around the monitor (Natural log) – Highways only
log_prisec_length_500	Count of primary and secondary road length in meters in a circle with a radius of 500 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_1000	Count of primary and secondary road length in meters in a circle with a radius of 1000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_5000	Count of primary and secondary road length in meters in a circle with a radius of 5000 meters around the monitor (Natural log) – Highway and secondary roads

Variable	Details
log_prisec_length_10000	Count of primary and secondary road length in meters in a circle with a radius of 10000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_15000	Count of primary and secondary road length in meters in a circle with a radius of 15000 meters around the monitor (Natural log) – Highway and secondary roads
log_prisec_length_25000	Count of primary and secondary road length in meters in a circle with a radius of 25000 meters around the monitor (Natural log) – Highway and secondary roads
log_nei_2008_pm25_sum_10000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 10000 meters of distance around the monitor (Natural log)
log_nei_2008_pm25_sum_15000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 15000 meters of distance around the monitor (Natural log)
log_nei_2008_pm25_sum_25000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 25000 meters of distance around the monitor (Natural log)

Variable	Details
log_nei_2008_pm10_sum_10000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 10000 meters of distance around the monitor (Natural log)
log_nei_2008_pm10_sum_15000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 15000 meters of distance around the monitor (Natural log)
log_nei_2008_pm10_sum_25000	Tons of emissions from major sources data base (annual data) sum of all sources within a circle with a radius of 25000 meters of distance around the monitor (Natural log)
popdens_county	Population density (number of people per kilometer squared area of the county)
popdens_zcta	Population density (number of people per kilometer squared area of zcta)
nohs	Percentage of people in zcta area where the monitor is that **do not have a high school degree** – Data from the Census
somehs	Percentage of people in zcta area where the monitor whose highest formal educational attainment was **some high school education** – Data from the Census

Variable	Details
hs	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing a **high school degree** – Data from the Census
somecollege	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing **some college education** – Data from the Census
associate	Percentage of people in zcta area where the monitor whose highest formal educational attainment was completing an **associate degree** – Data from the Census
bachelor	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **bachelor's degree** – Data from the Census
grad	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **graduate degree** – Data from the Census
pov	Percentage of people in zcta area where the monitor is that lived in **poverty** in 2008 – Data from the Census

Variable	Details
hs_orless	Percentage of people in zcta area where the monitor whose highest formal educational attainment was a **high school degree or less** (sum of nohs, somehs, and hs)
urc2013	2013 Urban-rural classification of the county where the monitor is located – 6 category variable - 1 is totally urban 6 is completely rural – Data from the National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm){target="_blank"}
urc2006	2006 Urban-rural classification of the county where the monitor is located – 6 category variable - 1 is totally urban 6 is completely rural – Data from the National Center for Health Statistics
aod	Aerosol Optical Depth measurement from a NASA satellite – based on the diffraction of a laser – used as a proxy of particulate pollution – unit-less - higher value indicates more pollution – Data from NASA