Group 7: Tanmay Kapse, Galav Sharma, Pratyush Joshi.

Github link: https://github.com/galav12/CSC442-Project

Dataset 1: US Pollution data from 2000-2016. This dataset has been collected by the Environmental Protection Agency (EPA). 4 different pollutants, organized by location and specific site numbers in the US from 2000-2016 are all recorded.
https://www.kaggle.com/datasets/mexwell/us-air-pollution

Dataset 2: Heart Disease and Stroke Prevention Data. This is one of the datasets from the National Cardiovascular Disease Surveillance System. It is a comprehensive list of indicators that contribute to the public health burden of cardiovascular diseases. The dataset is organized by location, and includes type of disease and risk factors. Can be stratified by age, sex and other factors.
https://www.kaggle.com/datasets/mazharkarimi/heart-disease-and-stroke-prevention

Cleaning and Wrangling:
Our datasets were a little tricky, in the fact that they were very messy, and could not be merged easily on one variable. We initially were confused whether the pollution dataset was based on specific sites or on state location like the heart disease dataset is. We had a discussion with Dr.Mallavarapu, and she suggested that we turn some of the row values into column headers, and clean each file separately before merging. Merging the unclean datasets would have created a mess to even begin cleaning. One of our group members attempted the cleaning and merging, and we have a good dataset to begin EDA on now. We still have to take a look at the outliers, and see how they affect the dataset and any analysis output. After looking at the outliers, and seeing their impact, we decided to remove outliers for the pollution dataset, and then merge with the heart disease dataset. We could not get “outliers” from the heart disease dataset as each category for age range, race, and gender have percentages of adults that suffered from a certain type of cardiovascular disease. Hence, we have one percentage value, and an outlier is not feasible. For the pollution dataset, we created boxplots to see the spread, and decided to remove the top 5% and bottom 5% of values for each of the pollutant variables per state. This greatly helped the cleanliness of the boxplots. We chose to keep the middle 90% as we are unsure of all the known outliers in the data in the pollution dataset to be a calculation error or a “true” outlier, or an actual data point where the pollution levels just happened to be that crazy.

Data Merging:
Once the datasets were cleaned, merging them was not that difficult to accomplish. We merged using a left join on the Location Description and Year for the heart disease dataset, and using a right join on the State and Year variables for the US pollution dataset. Our combined dataset has

21,997 rows and 22 columns. With the merged dataset, some of our values that were categorical variables in the rows become column headers, with only numerical values under them.

Group Contributions:
Galav: Found Datasets, Cleaned and Merged Datasets.
Tanmay; Found Datasets, Created the report, assisted in EDA.
Pratyush: Found Datasets, assisted in report creation and EDA.