Analysis Pipeline

Description of your dataset

We will be downloading the dataset directly from UCI Machine Learning Website. Our dataset is Beijing air quality: Beijing Multi-Site Air-Quality Data Data Set. It was marked as a regression analysis task. This data set includes hourly air pollutants, air-quality data, and meteorological data monitored at different Beijing monitoring sites.

Explore and visualize your data

We will explore this dataset by linear regression about different columns. After exploration and linear regression, we will visualize our data by plotting out our interesting variables and our linear regression line.

We will use the matplotlib function in Python to plot graphs including a Scatter plot with a linear regression line. After plotting the scatter plot with the data, we will use a bar chart to illustrate the difference between the different regression models.

Research question

What variables or features affect the amount of air pollution in Beijing? Would we be able to predict this amount by collecting various weather data? How accurate the prediction would be?

Your planned analysis

We are using linear regression to test out the major variables that affect the amount of air pollution in Beijing. We will also use cross-validation to train our models and find out if this validates by applying the model to the test data. The input would be the weather columns such as the time of the year, temperature, and the output would be the amount of PM2.5 or PM10.

Your planned report

Since we have many potential features for the predictive task, we will report the model with features where we found it to be the most accurate predictor after cross validation, and the error metrics. Therefore, the report should include the process of data cleaning, data imputation, plotting/visualizations, at least 2 models (as the task required), the error metrics.