Final Project: Regression, Inference, and Classification

Vaishavi Kagita

Part 1:

Where did I find the data? Kaggle!

What organization collected the data originally? Environmental Protection Administration(EPA) in Taiwan

Part 2:

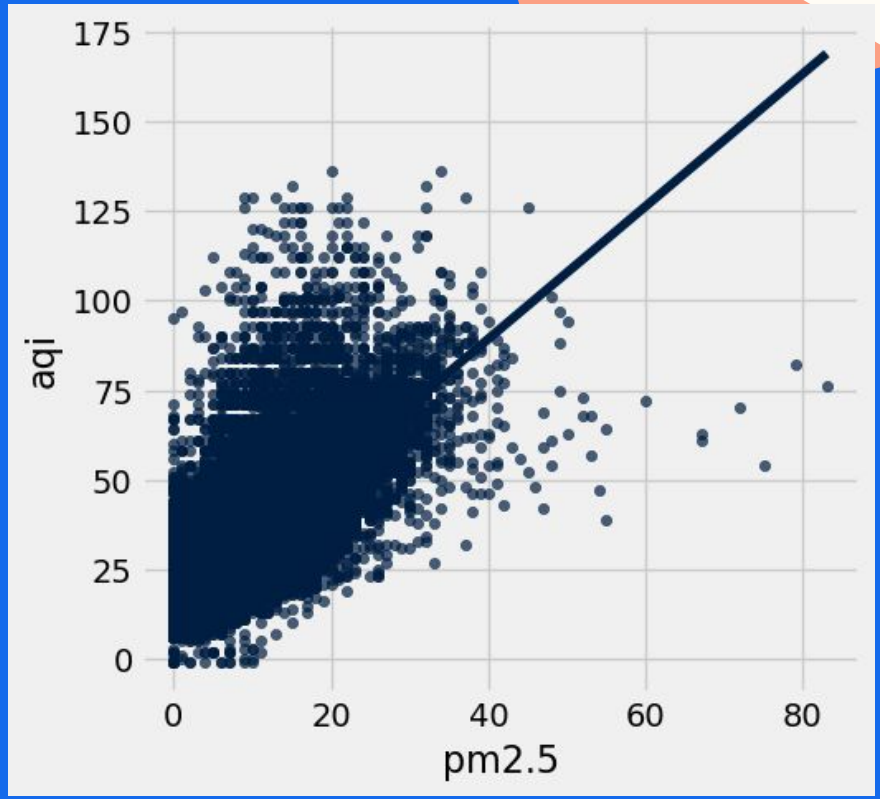
Which pairs of numerical variables did I look at? Pm2.5 and AQI

What was the confidence interval for correlation between these two variables? 95% CI: [0.7469, 0.7528]

What value of x was your input for prediction? x = 50

What was your 99% prediction interval? [107.32, 108.54]

Were there any potential biases or issues you saw in your analysis? Yes some.



Part 3:

My features? Included PM2.5, PM10, and SO2

How I will build my classifier.

- I will use the k-Nearest Neighbors (k-NN) method for classification.
- The dataset will be shuffled and split into 75% training data and 25% testing data.
- The classifier will calculate Euclidean distances between the features of the test row and all rows in the training set.
- The labels of the nearest k rows (starting with k=3) will be used for majority voting to predict the label.
- The accuracy of the classifier will be tested on the test set, and adjustments will be made if necessary.

Part 4:

What did I learn?

- PM2.5 has a strong linear relationship with AQI, making it a reliable predictor for air quality.
- The data may have biases, such as not being a simple random sample, which could affect generalizability.
- Big Question my classifier will explore → Can pollutant levels (PM2.5, PM10, SO2) reliably categorize air quality as good or bad?
- This project taught me the importance of splitting data for training and testing and ensuring it is representative.
- I gained a deeper understanding of regression, classification, and working with real-world datasets.