PREDICTING POLLUTION LEVELS

SNEHA VERMA
JORANIA F. ALVES

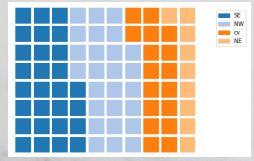
EXPLORATORY DATA ANALYSIS

Data Shape:
43,824 instances
13 features.

Data Description:

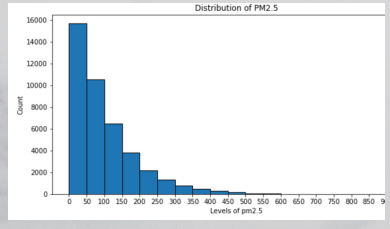
- There are 9 variables. One is categorical.
- There are no duplicate values.
- There are missing values in the target variable, PM 2.5.

Variable Distributions:



The categorical variable, wind direction, is imbalanced because there are not equal occurrences of each value.

We did not under- or over-sample because this dataset represents a specific location and wind direction can be a characteristic of a specific location which would create bias in the dataset.



The target variable is right-skewed showing a common range of PM 2.5 levels of between 0 and 50 micrograms per cubic meter of air.

DATA CLEANING

Missing Values:

- We removed the missing values from the target variable.
- **New Data Shape:** 41, 757 instances; 9 features.

Outliers

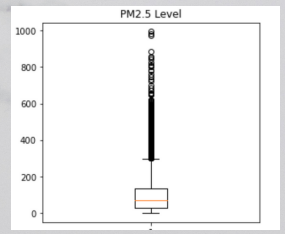
- Removed outliers across all numerical variables.
- Found out that the differences in the outliers and non-outliers for the cumulated hours of snow and rain are not big.
- Decided to remove outliers based on all numerical features except cumSnow and cumRain.
- Due to the arbitrary nature of the outliers, we decided to create two models:

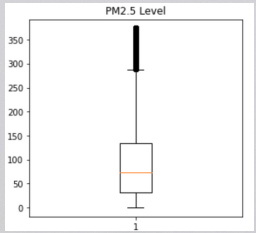
1) With Outliers

Data Shape: 41,757 observations, 12 features (one-hot encoded values of wind direction).

2) Without Outliers

Data Shape: 39,789 observations, 12 features.





DATA ANALYSIS

- We split each of the datasets (with and without outliers) into x and y parts.
- Then, we split each of the x and y variables into training and testing sets.

Results:

Linear Regression

Data with Outliers	Data without Outliers
R^2 = 26%	R^2 = 26%
Bias = 5860.1	Bias = 4545.4
Variance = 1.996	Variance = 1.407

Polynomial Regression

Data with Outliers	Data without Outliers
R^2 = 49%	R^2 = 46%
Bias = 4099.9	Bias = 3345.0
Variance = 32.689	Variance = 40.506

Conclusion:

Based on the R-squared values, the polynomial regression with outliers is the model that should be used to predict pollution levels. However, the bias and variance values indicate that the polynomial regression without outliers is a good model because it has better accuracy than the other models; however, all models are underfitting.