INFO411: Data Mining and Knowledge Discovery

Instructions:
This task is a real-world data mining problem. You are required to prepare a set of presentation slides that must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion.
Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Air pollution prediction in the United States

Background: The US records daily ozone, SO_2, CO and NO_2 levels in several counties of every state. The data set for this task contains the yearly summary data for these readings, and associated meteorological data such as air quality index (AQI) and particulate matter (PM) index. The data are available from https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual.

Download the AQI by County annual summary data for 2020. The "Days with AQI" column indicates the number of days in the year that the AQI index was recorded in the county and the following six columns indicate how many of those days had which index level.

Requirements:
1. Explore the relationships between air pollution (this could be what you judge to be "bad" AQI days, or high median/high max AQI, or another criterion of your own definition), the meteorological variables and the states.
2. Present relevant visualisations of the data, which help to illustrate the relationships, trends and differences found in the previous items.
3. Develop models to predict the number of days of PM > 2.5 concentrations using the rest of the meteorological data. Two of these that you develop should be the standard linear model and the random forest.
4. Provide the performance evaluation of any fitted models, including details of cross-validation or splitting into training, validation and/or testing sets.
5. Present your interpretations and conclusions.