For my final project, I have chosen three datasets regarding information on fuel-powered motor vehicles. Links to the three sets can be found below. All three datasets are related by vehicle model.

For my flat file, I have chosen a CSV file containing the engine specifications and emissions data of several models of motor vehicles. Each row is one vehicle model, and each column is one specification such as engine size, fuel economy, CO2 emissions, etc. To complete Milestone 2, I will conduct the following steps:

1. I will remove all hybrid electric vehicles from the set. They are considered outliers due to their infrequency and their high difference in specifications compared to other cars.
2. I will remove all categorical columns, as they are irrelevant to my analysis.
3. I will also remove the combined fuel consumption column that is in MPG. There is already a column for combined fuel consumption in L/100km, making the second column redundant.
4. I will find the average numerical values of duplicate values to aggregate them.
5. For car models with different trim names, I will drop the trim names. This allows for easier frame merging. Each model’s trim is already indicated by the engine specifications.

For my website, I have chosen one that contains a table of the number of each vehicle model sold during 2019. Each row is one vehicle model. To complete Milestone 3, I will conduct the following:

1. I will import the HTML data and get it into a more readable format such as a data frame.
2. The site contains other tables; I will remove the other tables first.
3. I will separate the make and model names into two columns just like the CSV.
4. The website aggregates all car models of BMW and Volvo (e.g., the site aggregates the sales of three Volvo models into the ‘Volvo 60-series’). I will split each aggregated sale number for these two among their models.
5. To make matching to the file easier, I will only keep the columns for model, sales, and change.

For my API, I have chosen one that contains the vehicle dimensions and weight of each vehicle model. To complete Milestone 4, I will conduct the following:

1. I will convert the returned data into JSON format for ease of access.
2. I will then change the JSON labels from letters to actual names for readability.
3. I will organize the retrieved data into a data frame.
4. I will remove trim names to allow for merging with the other data sources.

For Milestone 5, I will join the separate data tables using Python and build visualizations in Matplotlib such as the following:

1. Sales vs CO2 emissions
2. Sales vs fuel consumption
3. Fuel consumption vs dimensions
4. Dimensions vs sales
5. CO2 emissions vs dimensions

Note that all these steps are tentative. More steps may arise as I complete the project.

Links:

CSV: https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles

Website: https://www.goodcarbadcar.net/2019-us-vehicle-sales-figures-by-model/

API: https://vpic.nhtsa.dot.gov/api/