Pipeline.

Data Generation model

The main purpose for our pipeline is versatility. At the start of the project, it was unclear to us what dataset we would be using. Three options had been presented to us for the data acquisition:

1. Pull from a rest API.
2. Pull from an SQLite database.
3. Import a .csv

At the start of the project there was some confusion amongst our team whether we ought to build for a static or a dynamic dataset. In an early stage of the project, we made the design choice to start building a pipeline that could import from the three sources. Because we were unsure if the supplied data would be static or dynamic, we made the choice to automate the generation of different datasets prior to the EDA conclusions what dataset to use for the final version.

The data models that made it to the next stage have an enumerated name structure df1 to df5.

For use in the next stage of the project the data model generated its output in .csv and stores it in an SQLite.db

Data analyses determined that df4 would be the model we’d be using for the final regression model. The other models are currently commented out to reduce system load and clutter.