Daniel Ardoin

Data Analytics Bootcamp

April 1, 2020

ETL Project: Gas Prices

This project draws data from two separate sources. The first being an API from the US Energy Information Administration website. A total of four API’s from their site are used to comprise a large portion of the projects data. One for viewing the child series options, the second for average annual gas prices in California, the third for average monthly gas prices in California, the fourth for average weekly gas prices in California. The secondary data source finished the data for this project is a .csv file from data.world. This source provides average weekly gas prices for the entire US.

Extract

Extracting the API data was straightforward. A jupyter notebook is used to bring in the API’s and decoded the json into something usable. The one hiccup in extracting the data was moving through the various key-value pairs. After navigating to a key, a list was being returned so further code was needed to move from the list to the dictionary contained inside. Beyond that, the extraction process went smoothly.

Extracting the .csv file was also straightforward. The same jupyter notebook via Pandas is used to bring in the .csv. No problems were run into during the extraction of the .csv.

Transform

Transforming the API wasn’t too difficult. The dates used in the API’s weren’t quite as user friendly as I would have wanted and I made the decision to change that. The native format was “yyyymmdd” but I wanted to the year, month, and day to be in their own columns. I used a for loop to achieve this by iterating through the list elements and using the datetime module to parse the date. After that, I just converted that new list with the new columns with the parsed dates into a pandas dataframe.

Transforming the .csv file was done in the exact same way as the API’s. The only thing I might change is to turn my iterations into a function rather than just copying for each data element. My last step in transforming the data was to take my average weekly California gas prices and my average weekly US gas prices and merge their respective dataframes into a single dataframe based on the date.

Load

Loading the data was the most foreign aspect of this project by far as I wanted to use MongoDB but had missed the class covering it. My desire to learn about it is what lead me to the decision to use it. Luckily it was more straightforward than I was expecting. Beyond basic syntactical errors as I was learning, I got through loading the data fairly quickly. I’d say it was the fastest part of the project. The Database I made contains a total five collections: average annual California gas prices, average monthly California gas prices, average weekly California gas prices, average weekly US gas prices, merged average weekly California and US gas prices.

Works Cited

“U.S. Energy Information Administration – EIA – Child Series Options.” *U.S. Energy Information Administration (EIA) - Qb*, U.S. Energy Information Administration, 30 Mar. 2020, [www.eia.gov/opendata/qb.php?category=240839](http://www.eia.gov/opendata/qb.php?category=240839).

“U.S. Energy Information Administration - EIA – Annual California Gas Prices.” *U.S. Energy Information Administration (EIA) - Qb*, U.S. Energy Information Administration, 30 Mar. 2020, [www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM\_EPM0\_PTE\_SCA\_DPG.A](http://www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM_EPM0_PTE_SCA_DPG.A).

“U.S. Energy Information Administration - EIA - Monthly California Gas Prices.” *U.S. Energy Information Administration (EIA) - Qb*, U.S. Energy Information Administration, 30 Mar. 2020, [www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM\_EPM0\_PTE\_SCA\_DPG.M](http://www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM_EPM0_PTE_SCA_DPG.M).

“U.S. Energy Information Administration - EIA - Weekly California Gas Prices.” *U.S. Energy Information Administration (EIA) - Qb*, U.S. Energy Information Administration, 30 Mar. 2020, [www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM\_EPM0\_PTE\_SCA\_DPG.W](http://www.eia.gov/opendata/qb.php?category=240839&sdid=PET.EMM_EPM0_PTE_SCA_DPG.W).

“U.S. Retail Gasoline Prices - Dataset by Eia-Wind.” *Data.world*, 24 Feb. 2020, <data.world/eia-wind/us-retail-gasoline-prices>.