Capstone Project 1 Data Wrangling

My data sets are respectively from the California Independent System Operator (CA ISO) and the Midcontinent ISO (MISO) for their respective electrical grids. Each of these data sets is an enormous labyrinth with myriad associated challenges (some quite unexpected, such as the configuration of the CAISO system around the antiquated Internet Explorer browser). Most parties access this data indirectly via an expensive subscription service, such as SNL Energy. For this reason, I suspect that the data has not been much studied.

The Good News;

1. First, there is a lot of data from CAISO, including 5-minute power prices for more than 2,000 locations (or "Nodes"), and hourly data for Wind and Solar generation. Moreover, there is essentially a futures market for the electricity prices (the so-called "Day-Ahead-Market", or DAM).

2. Second, billions of dollars are traded daily based on this information, so there are only rarely missing or obviously erroneous data.

3. Third, CAISO has its own API (however, accessing this has so far been challenging, and it takes 7-15 minutes to receive the results of a query with data output of only 0.5 MegaBytes. This remains a work in progress (with perhaps 40 hours invested in understanding this technicality only).

However, the data is largely configured to meet the needs of historical large users, who tend to be large companies with strong IT departments who have been working with these data sets for years.

Data Wrangling Challenges

To cite but a single example, one item of curiosity is the occasional occurrence of negative power prices, and whether this presents a valuable opportunity to charge Electric Vehicles cheaply in an intermittent and opportunistic matter. An academic paper on this subject by a Berkeley Professor mentioned the importance of hydro-energy as a driver of the negative prices. However, the hydro data cited in the esteemed Professor's analysis was annual – an extremely low level of granularity that might be deemed anecdotal at best. It seems that there is in fact hourly hydro data available, but not in an accessible CSV format. Specifically, the data is included manually in a daily text file that allows computation of the so-called California "Duck Curve", which is displayed in a daily HTML report that is co-published daily. While my limited Python skills were fortunately sufficient to write a function that successfully extracted the hydro data from a sample Daily Report that I had downloaded as a csv file (def text_to_csv( )), my function did not work so well with files downloaded via a simple web scraper using the Requests package from the urllib library. Sometimes the columns are separated by empty columns, and sometimes dashes are manually entered next to a number to indicate a continuation. Fortunately, with an approach of reading the Response file as text, and then cleaning it up with a series of Regular expressions, the files read via the urllib library now can be processed (and readily concatenated to make an annual time series, one hopes). Many thanks here to my esteemed Mentor, Jeff Hevrin, who provided the critical breakthrough with his elegant and succinct code.

An Iterative Cycle of Data gathering, Processing, EDA, Interviews,

One of my favorite videos in this Springboard course raises the Data Science issue of "*What is the question?"*. This I revisit daily as new information is gathered through informational interviews, data collection, data wrangling, and exploratory data analysis.

Initially the question(s) might have been framed loosely as:

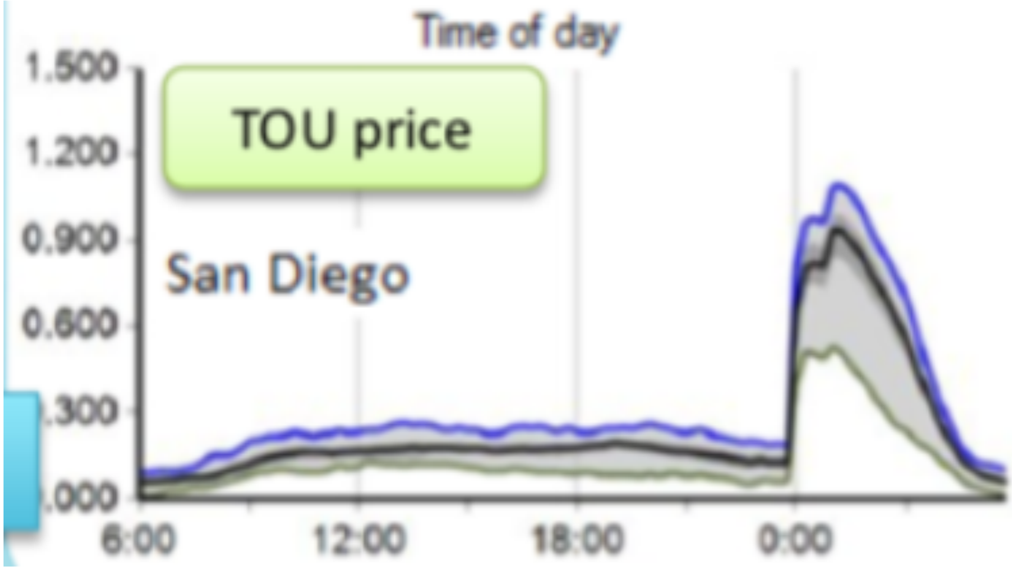
1. If intermittent and opportunistic EV charging in Northern California during the night (from midnight to 6am) was successful in hitting the 24 5-minute periods with the lowest prices, what might the savings be at

wholesale prices relative to a supposed naïve strategy of charging for two hours starting midnight? Preliminary Answer: c.25%

2. If a powerful Level Two charger was purchased so the 12-lowest price 5-minute periods were accessed, how much would the savings increase? Answer: to c.29%.

The underlying idea was, *"Could more Level Two EV chargers be beneficially used to "mop up" the excess supply that caused the very low or negative prices, and the electrical grid thereby be stabilized, with a more harmonious integration of Renewables and fossil fuel plants being achieved?"*

However, concurrent informational interviews have surfaced a surprisingly pressing infrastructure problem, known as "*The Last Mile*". The point is that local transformers can very easily become overheated by geographic and temporal clustering of EV charging. A great irony of this is that the proposed time-of-use solution seems to cause charging by timer at exactly the onset of the designated off-peak period, and this may overwhelm the transformers. So charging worth less than one hundred dollars can quickly overheat the nearby $7,000 transformer (supposedly shortening its life by a factor of 10,000x). See TOU chart below. Who would have guessed!



A more sophisticated and artful strategy is therefore required, which ideally incentivizes EV chargers to cede control to a third party, so that the entire system can be optimized with respect to multiple constraints (some much better understood by the general public than others).

Critical in this context is that it was quickly realized from EDA of the CAISO data that the nighttime prices were first the most stable during the day, and secondly they were NOT the lowest despite being off-peak. The lowest prices in CAISO occur around 10am, when solar energy is strong but the land has not warmed up enough from overnight to have much of an air conditioning load.

Accordingly, it would be helpful to look at an area such the Midcontinent ISO (MISO) where wind is dominant and nighttime prices would likely be more volatile, as well as being the lowest in the 24-hour day. Further justifying this analysis is that my wide reading had surfaced that a large transmission interconnect line will be opened between MISO and CAISO, so the two markets will be connected in the future.

Incidentally, the hourly MISO price data can be downloaded conveniently as 9 MB files. However, the hourly prices for each day are arrayed horizontally. Fortunately the judicious use of stacking and Time Stamp data got around this problem, but it did take me a few hours of thinking and experimentation. It is easy for me personally to now see how 80-90% of the Project time can be devoted to artfully collecting and processing data (and figuring out what the right question is, now that data has been sampled).

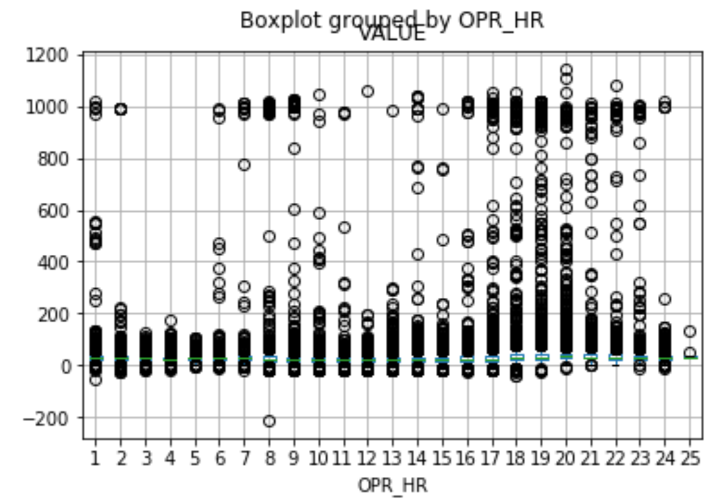
The foregoing is a highly abbreviated narrative of the data wrangling issues encountered. Most assuredly the battle continues. The goal is a comprehensive, Economic, Infrastructure, and Data Science analysis that works for all parties. A portfolio of ideas is being developed in this context, and may be marketed to interested parties (including VW and their national network of stations).

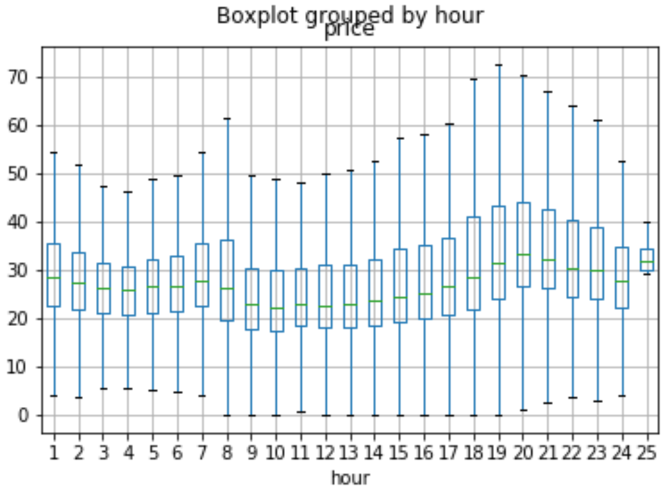
The current idea is to access wind in sufficient volume that the lowest power prices occur at night, and not in a manner that can be easily predicted by an EV owner charging at home. The charger then has an incentive to cede control of his timing to allow a sophisticated third part to hit he lowest prices, and in the process the dangerous local transformer load constraint can then be artfully navigated by an appropriate algorithm framed with multiple objectives. This has the potential to give rise to a rather paradoxical result. A major disadvantage of Wind Energy has been that it is unpredictable and occurs mostly at night. But this is exactly what is needed for home EV chargers to cede control of their timing (and hence also thereby limit inadvertent damage to the vulnerable local transformers).

Outliers

In the context of analyzing thousands of 5-minute prices over the course of a year, there were in fact so many outliers that the underlying economics were almost completely obscured, and the scales were also rendered meaningless. By removing the outliers for plotting purposes only, this challenge was successfully managed.

In the chart on the left below, the interquartile range is obscured. However, in the chart on the right the medians and the interquartile ranges are clearly discernible, and one can readily see that the lowest prices occur not during the night, but rather around 10am.


Boxplot grouped by OPR_HR
VALUE
OPR_HR


Boxplot grouped by hour
price
hour