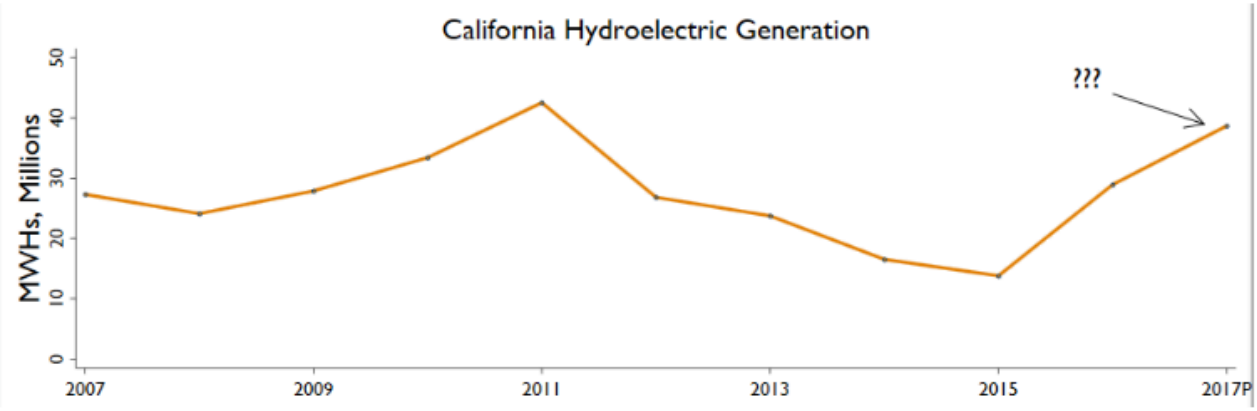
Capstone Project 1 Milestone Report

I. Problem Statement:

This Project seeks to address a fundamental outstanding question regarding the future growth in electric vehicles (EV's). This is an important issue as transportation-related CO2 emissions constitute more than 25% of US GHG emissions. Moreover, to the extent that EV's can be charged with power derived from Renewables, such as Solar and Wind, then a material reduction in GHG emissions could be achieved, and this would also provide an attractive model for other countries to follow.

Potentially, EV charging could greatly support the stability of the electrical grid. For example, as generation from Renewable forms such as Solar and Wind Energy has burgeoned, very low prices (even negative occasionally) have been experienced. While this may at first seem attractive, as a practical matter such low or negative prices typically require subsidies to incentivize other power plants to remain available when their power may be needed in the future.

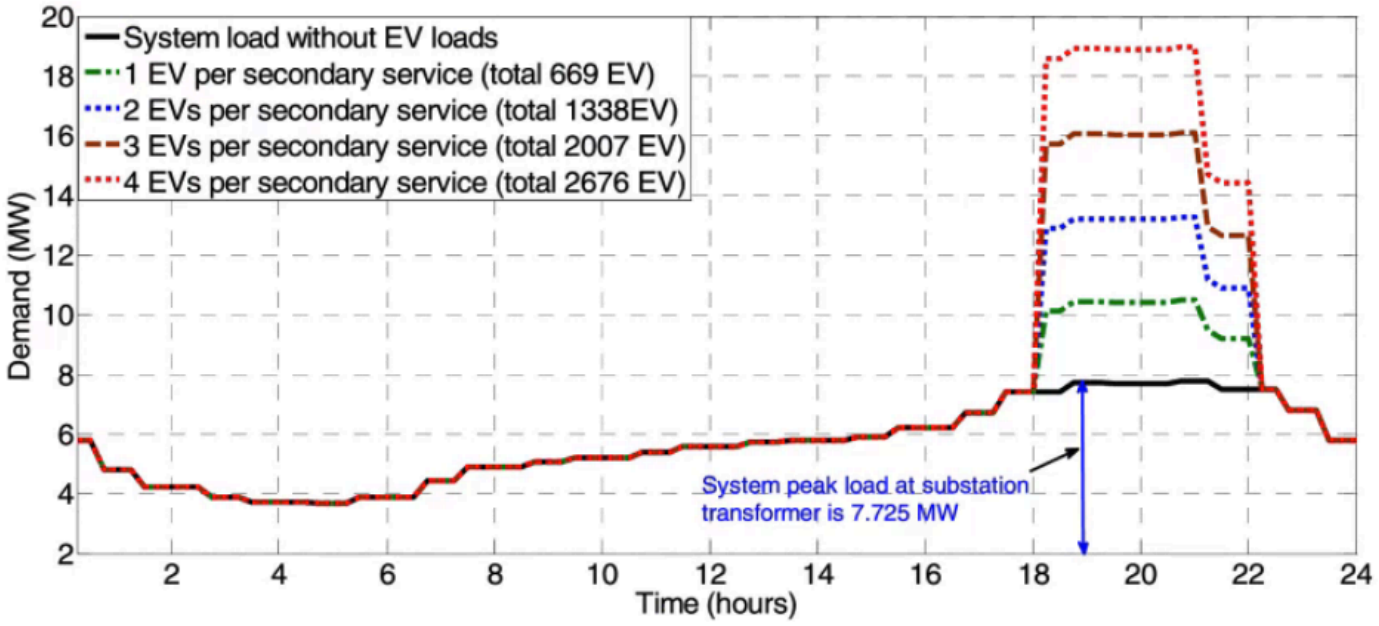
While there has been some statistical analysis of the negative power prices in California, it has been quite limited, and not especially granular. For example, Lucas Davis (a Professor at the Energy Institute at Haas Berkeley) published an analysis in which hydro power was considered to be a critical variable. However, the hydro data (below) was at the annual level, and moreover there was no breakout of the small-stream hydro which he believed to be particularly important (because its level is driven by the natural flow of water, without curtailments).



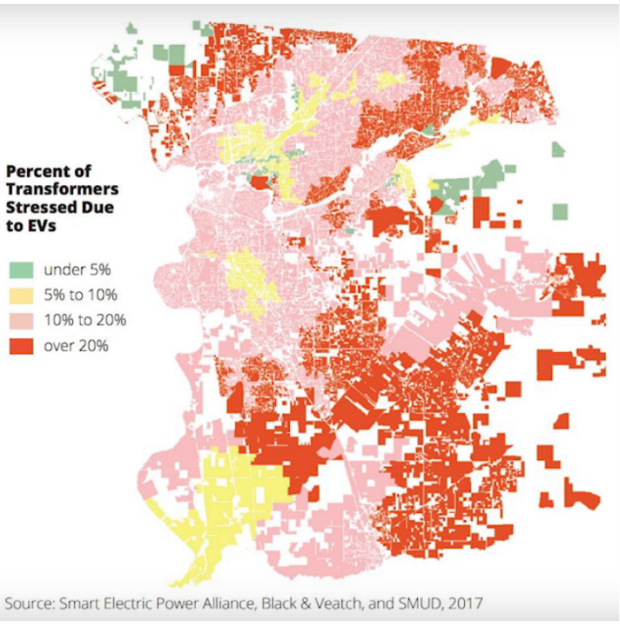
Note: This figure was constructed by Lucas Davis (UC Berkeley) using annual California hydroelectric generation from EIA. The prediction for 2017 uses actual generation Jan-May and Jun-Dec generation from 2016.

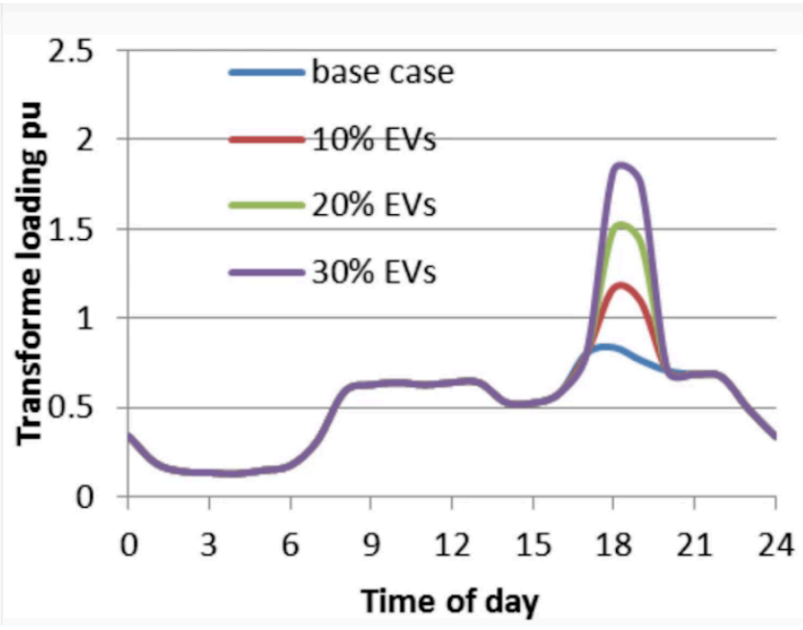
This analysis utilizes an apparently overlooked source to assess the importance of hydro power, namely that hourly hydro generation for small and large hydro is provided as back-up for daily reports by the California Independent System Operator (or CA ISO). While this data is not available as a continuous time series, such a time series can be resourcefully compiled by querying and concatenating the subject daily data (although it turns out that he code for this is non-trivial for some unexpected reasons). It is hoped that this increase in granularity (from annual to hourly small hydro data) will yield new insights as to instances of low or even negative prices, and that superior predictions will enable a sustainable approach to EV charging.

One very attractive opportunity is that EV charging could be managed to instantaneously mop up brief periods of oversupply, and thereby to stabilize the integration of Renewables into the electrical grid. However, a thorny problem in this context is the so-called "*Last Mile*" problem. Specifically, the hourly electrical load of charging an EV is so high that it typically exceeds the Peak load for an entire household.



The challenge here is that if there is so-called "clustering" of charging (either geographical and/or temporal) then the local transformers can quickly become overheated, and their life shortened by a factor as large as 10,0000-fold. These transformers are local, and physically appear as cylinders suspended by telegraph poles for local power lines. Typically the local transformers do not have sensors, so the utility has no feedback as to their current status and whether they are being overloaded. However, physical inspections have suggested that in certain areas more than 20% of the transformers are already stressed due to EV charging (see map below), and this percentage is likely to increase sharply as EV's are increasingly adopted.





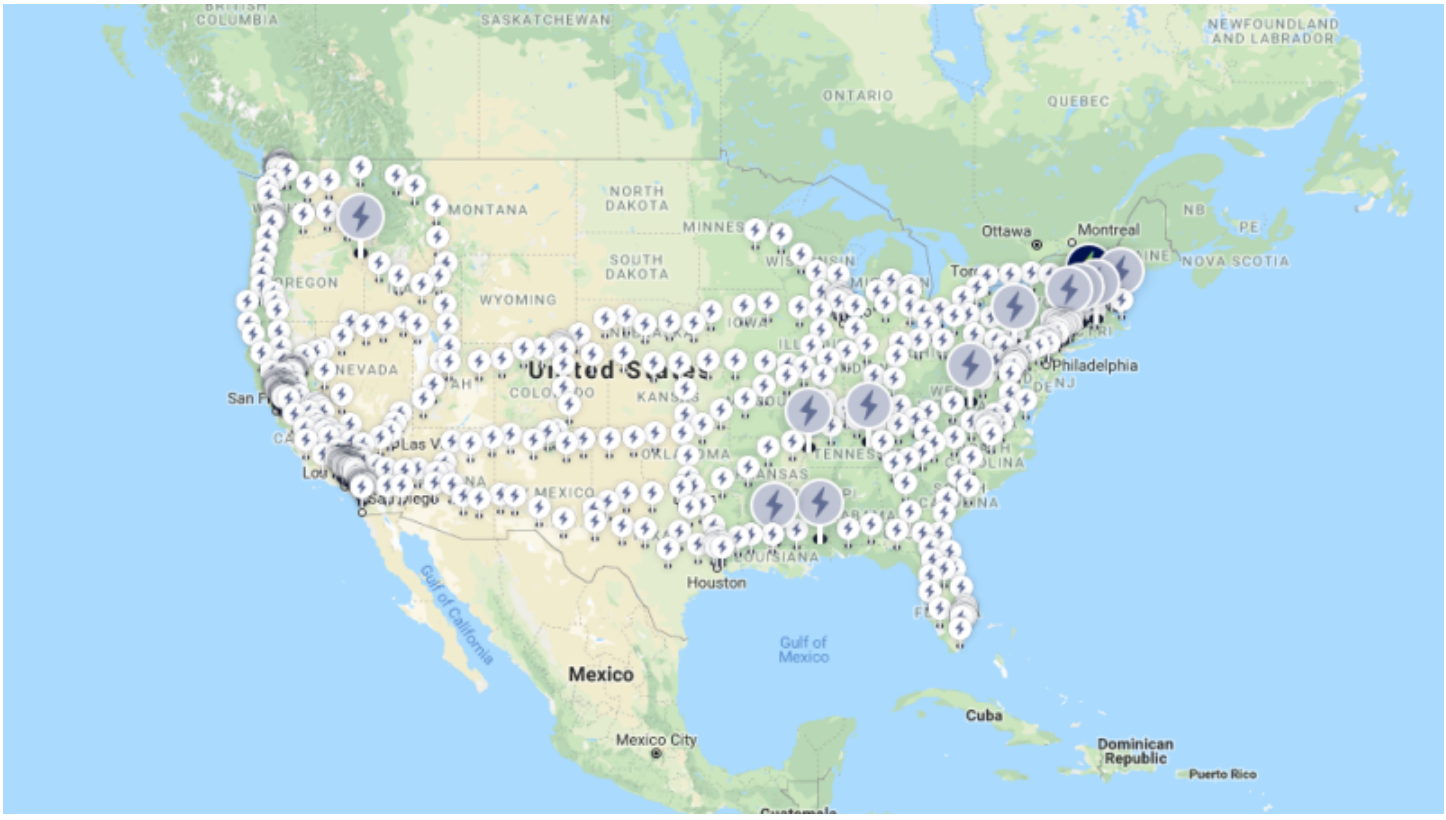
<u>Potential Approaches</u>

The use of intermediate batteries: first, one may note that although the "*Last Mile*" problem is perceived as being new, its fundamental nature was recognized at least as early as 2011, when Ron Prosser began writing about it and developed a business plan for a company later named "Green Charging" . A pilot project was later initiated with ConEd. The approach adopted there was to reduce stress on the local transformer system by using large batteries to charge the EV's, and also to use solar PV energy to charge the battery.

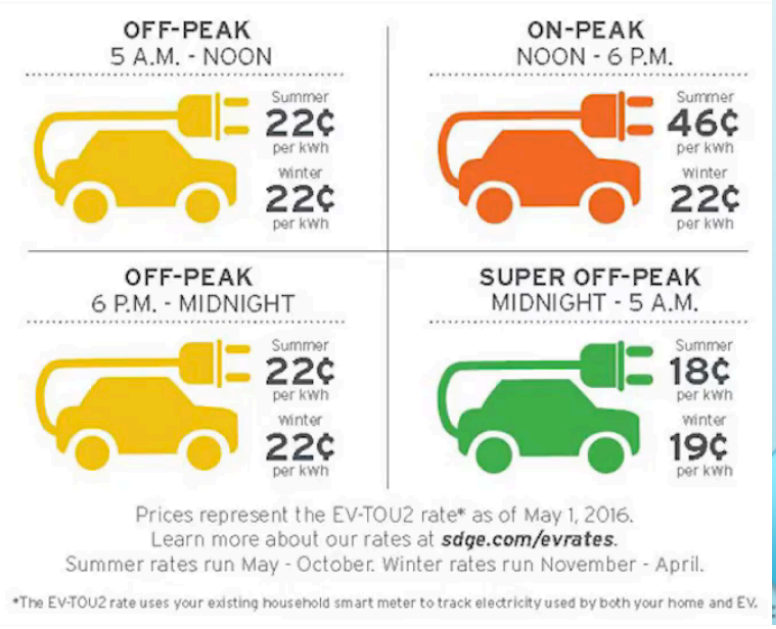
In fact, the "Electrify America" national network of EV charging systems being implemented by Volkswagen uses exactly this approach of using large (here mostly 350 kv) batteries, with the variation that the batteries can be inexpensively sourced from EV's whose batteries have lost more than 80% of their original capacity. The approach is succinctly described below:

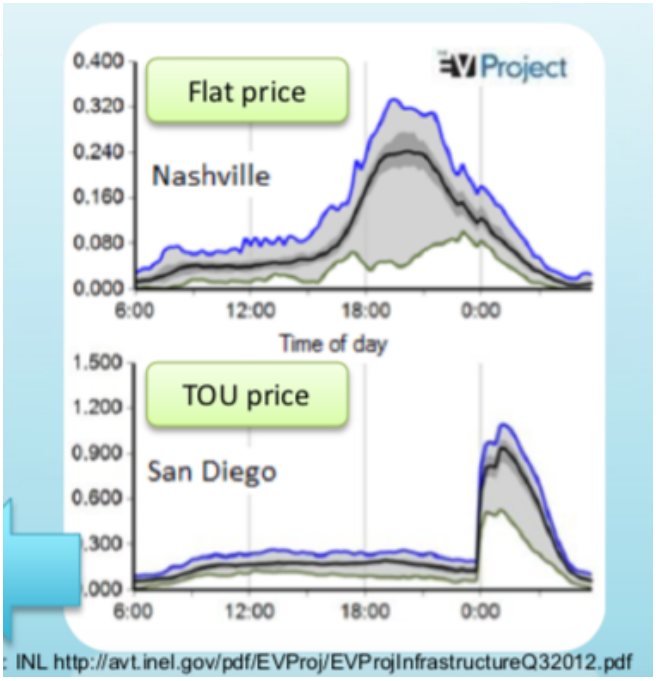
"Electrify America, the Volkswagen subsidiary that's tasked with building up America's electric vehicle charging infrastructure ahead of VW's plan to introduce a new lineup of electric cars, has made a deal to use Tesla Powerpack batteries in over 100 of its charging stations. The company will include 210 kW Tesla Powerpack batteries with 350 kWh of energy capacity, which will help mitigate Electrify America's charging station power demand during peak charging periods, according to the company's announcement. By supplementing with the energy stored in the batteries during busier charging periods, the company can avoid putting too much demand on the grid and ideally dodge any charges associated with that. The design is modular so that more energy capacity can be added in the future. The current setup will rollout to stations across the country throughout 2019".

What is interesting to this author about the commentary above is that it recognizes that charges may be imputed to the culprit for "*putting too much demand on the grid*". If this approach were to be consistently applied to home EV charging, a cluster of EV owners who innocently each decided to charge at the onset of the "Off-Peak" period



Tariff-Based Approaches: the current approach adopted by large utilities such as PG&E and San Diego Gas & Electric, seems primarily to seek to incentivize EV chargers to not charge during Peak hours through the use of a so-called "*Time-of-Use*" tariff system. An example from San Diego Gas & Electric is provided below.





However, as we have already seen above, the "Hump" from EV charging can easily be a <u>multiple</u> of the Peak demand. Moreover, experience has shown that there is typically highly clustered charging at the <u>*exact*</u> onset of the Off-Peak time period, as shown in the chart below comparing the "Flat Price" experience for Nashville with the "Time-of-Use" (TOU) experience in San Diego (chart supplied by San Diego Gas & Electric).

A Centralized Approach: the key issue seems to be to spread out the EV charging load over time, and ideally in a centrally-coordinated non-random manner. The central challenge is that any random process will inevitably have statistically-likely peaks, and the cost of such peaks may be quickly prohibitive as the lives of local transformers may be shortened as much as 10,000-fold by overheating.

<u>Relationship with Autonomous Vehicles</u>

To date, many of the world's largest car manufacturers and tech companies (including Apple, Google, and Tesla) have made enormous financial and intellectual capital investments in the development of autonomous vehicles. Notable in this context is that most autonomous vehicles are widely anticipated to be electric (see for example, "*Why most self-driving cars will be electric*" which is available at: https://www.usatoday.com/story/money/cars/2016/09/19/why-most-self-driving-cars-electric/90614734/ .

However, it is becoming increasingly apparent that the "Achilles Heel" of electric vehicles is the "Last Mile" problem, which by comparison has received very much less attention. This author believes that EV charging and the Last Mile problem are worthy of much greater consideration, and radically new approaches. Indeed, one may suspect that as a practical matter the implementation of autonomous vehicles may be constrained by the electrical grid, as evidenced by the last Mile problem. These issues have underpinned the motivation of this particular Data Science-based initiative.

II. Data Wrangling Issues

The subject 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 publicly, other than by a narrow community of business interests.

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 8 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 and are highly familiar with it and its idiosyncrasies.

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).

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 designated as "Off-peak". In fact, the lowest prices in CAISO occur around 10am, when solar energy is strong but the land mass has not yet warmed up enough from overnight cooling to have much of an air conditioning load.

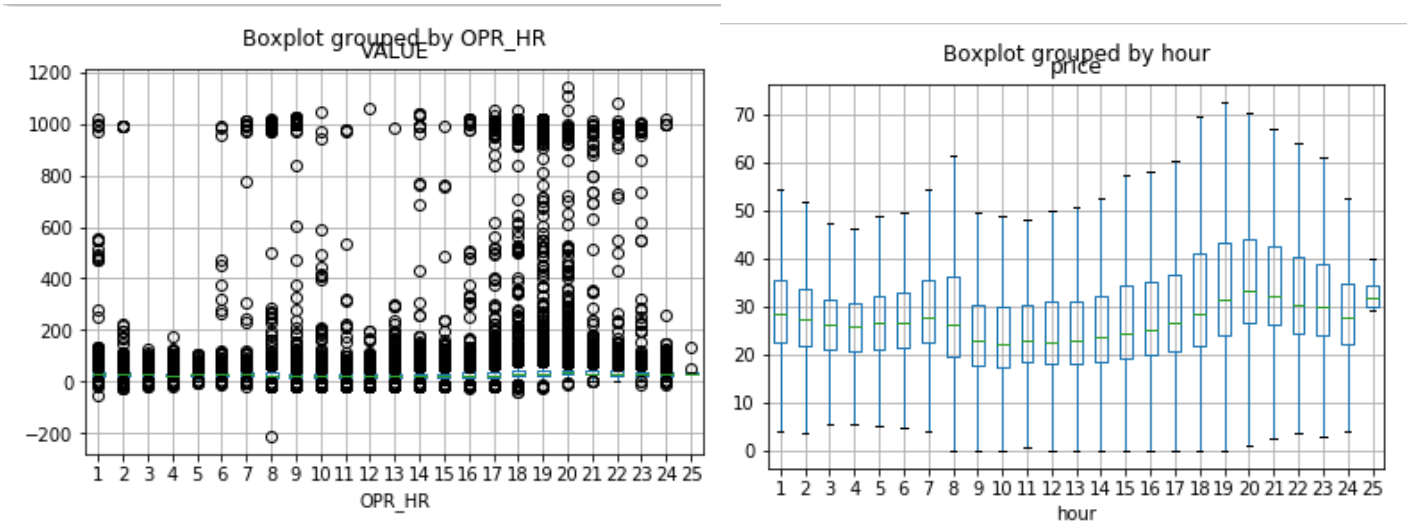
Accordingly, it would be helpful to look at an area such the Midcontinent ISO (MISO) where wind is dominant and nighttime prices would likely be more volatile, as well as being the lowest in the 24-hour day. Further justifying this analysis is that my wide reading had surfaced that a large transmission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



From data Story Notebook

Background

While electric vehicles are attractive from the standpoint of their low carbon footprint, there is concern that EV charging could potentially destabilize the electrical grid in the future. The argument is that if EV owners simply opt to charge their vehicle when they arrive home from work in the evening, then peak demand could be very sharply increased. Since it has been suggested that EV charging could potentially increase a household's total electricity consumption by as much as 50%, the increase in peak demand could be much higher.

Peak power demand is often met with simple cycle gas turbines which can be started up within minutes and are relatively inexpensive to purchase and expeditious to install. However, when their electrical output is ramping up and down sharply, their CO2 emissions per kwh can be similar to those of a coal-fired power plant. Accordingly,absent an economic form of power storage, the benefit of carbon-free Renewable energy may be much reduced. While battery storage has been much discussed, the cost been calculated as 25 cents per kwh, which far exceeds the generation cost which is often below 4 cents per kwh for Renewables. For this reason, 98% of the storage implemented to date has been in the form of pumped hydro, with batteries accounting for less than 2%.

An opposing view is that EV charging could actually help stabilize the electrical grid because it can increase power demand at night when it is commonly lowest. In particular, output from wind turbines is often lowest in the "wee hours" (eg from midnight till 4am) , and sometimes power prices are even actually negative then as a result. While negative prices might at first seem a good thing, as a practical matter they invite lobbying for all manner of subsidies so the generator can remain profitable and justify continuing to supply the grid with power.The type of backroom deals that result are usually far from optimal.

A Data Science Perspective

In general, most EV owners as a matter of actual practice tend to charge their vehicles for about two hours each night, and in one continuous stretch. However, output from wind turbines can vary greatly within the space of just five minutes. This raises the question of whether significant costs savings could be achieved by charging intermittently at times when the power price is lowest (or even negative). For example, a Decision Rule could be adopted to charge for any Five-Minute period if the price was below 2 cents per kwh at the onset of that period. So

now, instead of a continuous stretch of two hours (or 24 Five-Minute intervals), we could have 24 Five-Minute intervals chosen within a time from 11pm to 6am when the price happened to be below 2 cents/kwh. Of course, it is not practical for the EV owner to be up all night checking each Five-Minute price. Rather the owner would subscribe to an App that would do this on his behalf. Specifically, the EV owner would cede control of the charging to the App, and in return receive a hopefully significantly lower charging cost.

This of course is the Data Science Question. Is it reasonable to suppose that the distribution of prices is sufficiently dispersed that the opportunity exists to significantly lower the cost through selective, opportunistic charging in the manner outlined above. For example, if nighttime prices did not really vary that much from interval to interval, then the benefit of selective charging would be insufficient to justify the effort and the cost involved of setting up the whole system.

While there are many possible metrics, a simple one ("Metric One") was someone arbitrarily chosen just to start exploring the data and making a crude preliminary assessment. If this was encouraging, then further more comprehensive analysis would appear to be justified.

Metric One: what percentage savings are achieved for the wholesale power price at a major grid node by charging at the lowest-price (non-consecutive) 24 periods rather than simply by charging from midnight to 2am? The latter approach has been chosen as a benchmark, as it could easily and inexpensively be implemented with a timer.

Northern California was the subject of the initial analysis because of (a) the state's focus on Renewable forms of Energy, (b) the interest in electric vehicle ownership, and (c) the availability of power prices at 5 minute intervals (versus the hourly prices available for other grids). The dominant utility is Pacific Gas & Electric ("PGE"), and so the major PGE clearing Node was chosen.

[Metric Two:] A variation on the above approach is when a so-called Level Two charger is used, which is more expensive than a standard EV charger but charges at twice the rate. Accordinly the above analysis could be repeated, but this time with the prices of the 12 lowest intervals, rather than the 24 lowest.

Data Collection and Processing

Process: time series data (specifically in Five Minute intervals) for the PGE Node was downloaded in the form of twelve monthly csv files from the California Independent System Operator ("CAISO"), which is based in Sacramento, the state capital. Each file took about seven minutes to download on the weekend, but during weekdays many of the attempted downloads failed. It seems that most interested parties obtain the data from a third-party provider, for which subscriptions are reportedly (according to a leading academic) very expensive. After requisite data cleaning, these monthly files were then merged to make a continuous time series.

Commentary:

An Initial Surprise: Plotting this time series revealed quite a surprise. While one can indeed discern low and negative prices, the "shocker" is the recurrent price spikes of around $1.00 per kwh (apparently the statutory maximum), compared to the average price of closer to just 3 cents/kwh. While there has been much discussion of the negative prices (which occur in c.3% of the intervals), I have personally never heard of the prodigious price spikes. To investigate this matter further, separate plots were done for the four months with the highest incidence of such spikes, namely February through May. The spikes seem to be spaced out and non-consecutive, further adding to the mystery. An email dialogue via LinkedIn with a leading Energy Data Analytics company in Northern California confirmed the veracity of the spikes, but no explanation was provided other than their supposedly random nature and that they may be "congestion related". While these spikes are indeed rare (counting them reveals a frequency of only 1%), they are so large that they still make a large contribution to the mean Node price. This phenomenon, although not the subject of the initial inquiry, clearly needs further investigation.

A Personal Hypothesis: from my somewhat extensive background reading and historical domain knowledge, March is the month with the steepest hourly ramping of fossil fuel production. My guess is that this occurs in March because at the onset of peak demand in the evening there are still relatively few fossil fuel plants running, as the seasonal Air Conditioning demand has not started yet and most of the power is coming from Renewables. So the fossil fuel plants are "cold" and slow to start up (just like a car that has not been driven for a while). One theory to test is whether the rare spikes are positively correlated with a faster ramping rate. One simple chart that might be informative is a scatter-

plot of the percentage of the load that is supplied by Renewables in a given week versus the frequency of price spikes in that week.

Quantification of Metric One: the crudest initial exploration is to look at the tail of the distribution.

Each day have to first sort to find the lowest-price 24 periods, then compute the average for these periods (again, for each average).

Conceptual Algorithm

To derive a floor for the daily charging costs associated with the App:

1. Take the concatenated DF and carve out only the hours 1 to 6.
2. Currently this DF is sorted by 'day', then 'hour', then by 'interval'.
3. We want it sorted by 'day','hour', then by 'price' (for the 6 x 12 = 72 intervals)
4. For a two hour charging period of 24 non-consecutive intervals, we clip off the first 24 off the 72 intervals for each day, and then compute their mean.
5. This derives the mean of the 24 lowest-price intervals in the subject six-hour period for each day

Footnote

Personally, the ("publicly undiscussed") price spikes are a much more interesting topic than the negative power prices. Why? Because the negative prices can be made to go away simply with Curtailments. While this does not solve the broader issue of over-abundant supply of Renewable energy (primarily from Solar PV), it does appear as a statistical solution. Secondly, since the Solar and Wind generators often receive valuable tax credits worth more than 2 cents per kwh, for these generators the net price is still positive. Moreover, while the phenomena of negative prices has attracted a great deal of discussion, we can easily calculate that they effect the overall average price only modestly. In stark contrast, if the huge Price spikes should increase in frequency from 1% to 4%, average power prices could more than double! My economic intuition therfore suggests that the most valuable problem to understand and to solve is that of the still mysterious and largely undiscussed price spikes, which might ultimately threaten the viability of the Renewables paradigm.

A Second Surprise:

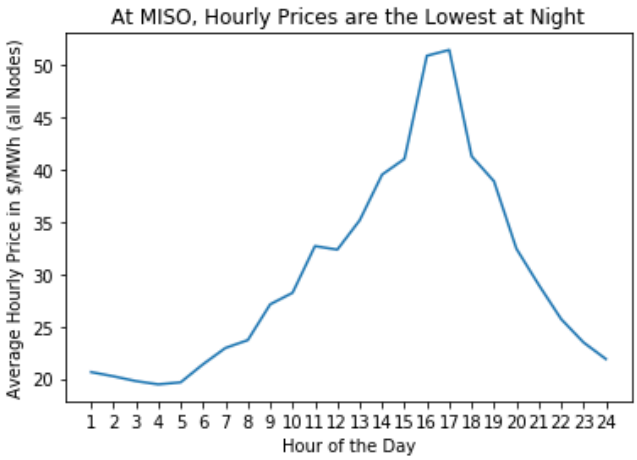
The wholesale prices for the year were grouped by their respective hour of the day, and then averaged. The surprise was that the lowest prices did NOT occur in the middle of the night as PG&E's published materials had suggested, but rather in the middle of the day - in the hours surrounding solar noon. In short, in recent years PV solar generation (including from rooftop solar panels) has become so abundant that it may at times even exceed the total power demand (known as the Load). Another reported factor is hydroelectric power. Apparently much of this is from flowing streams rather than dams, and for these streams the power can not be managed in the way that it can be with dams.

Preliminary Conclusion

Interestingly the above chart provides a strong hint as to our likely conclusion. Specifically, it shows the interquartile ranges as well as the medians. So we can see that the median price is around 28 dollars per MWh for hours 1 & 2, corresponding to the first two hours (so from midnight). But we can see that for the hours Three, Four, and Five, about 25 percent of the observations are below $20/MWh, and the mean of these observations may be around $15/MWh.

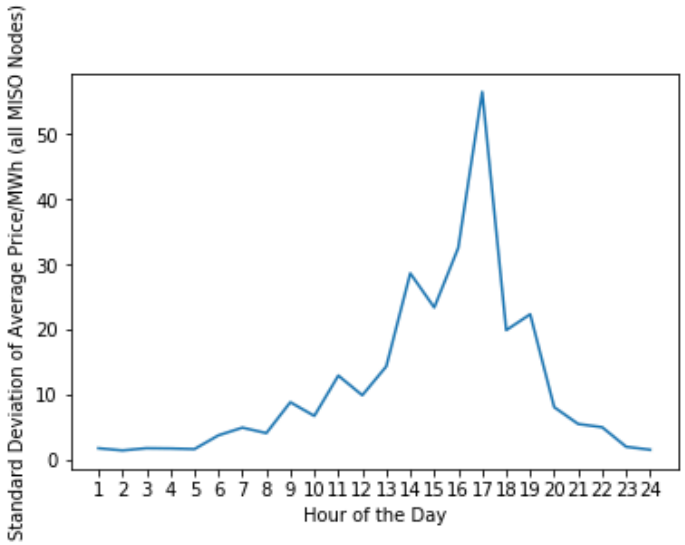
So a first guess is that we might realise an almost 50 percent reduction in the wholesale cost (the 15 dollar mean for the observations below the interquartile range, versus the above cited median for the first two hours of $28/MWh. This is encouraging, and suggests that the additional effort to investigate matters in a much more rigorous manner may not be wasted.

MISO Price Commentary



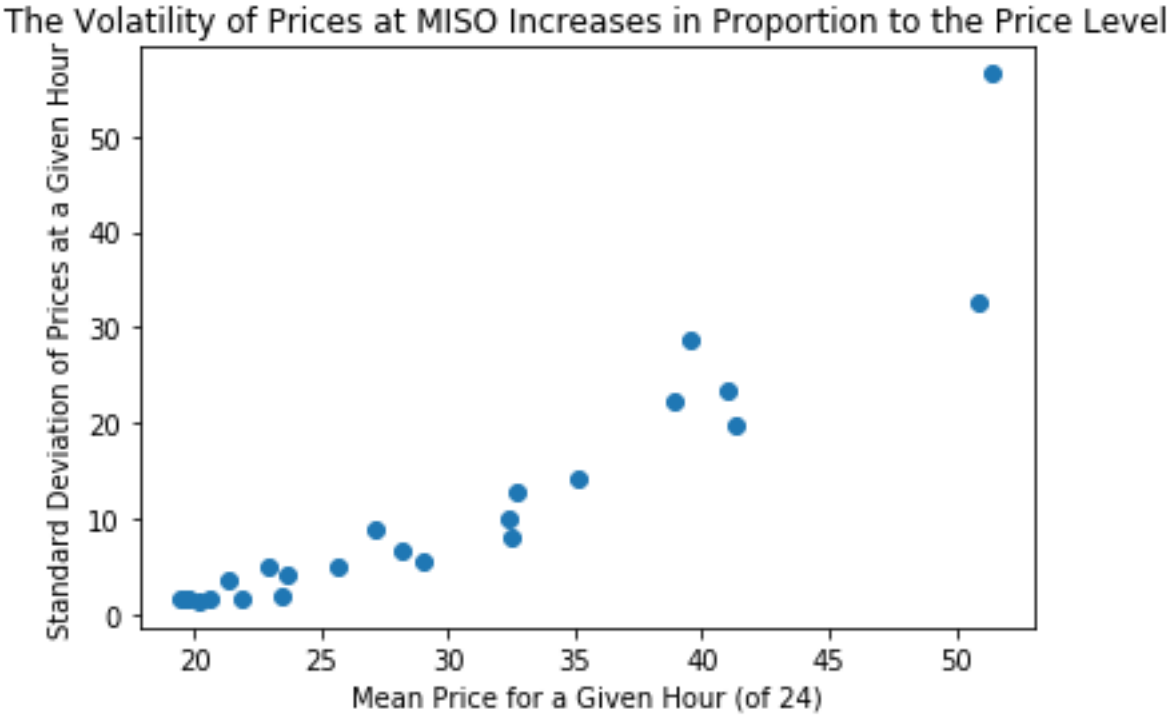
At MISO the pattern of hourly prices is quite different from at CAISO

Specifically, the lowest prices occur at night (hours Three & Four), and at around $20/MWh are lower than the $30/MWh level at night for CAISO. At CAISO the lowest prices are mid-morning as the sun rises in the sky, while at MISO mid-morning prices are rising steadily.



At MISO, the standard deviation of prices at night is even lower than at CAISO

But at CAISO and MISO the night prices are the least volatile, with the volatility broadly increasing in proprtion to prices.



Capstone Project 1 Milestone Report Students typically spend 10 - 20 Hours

Think of a milestone report as an interim report that you may be asked to share with your client to keep them updated on your findings. It's also an opportunity for you to take stock of how far you've come, what you've found, and practice your data storytelling skills. This is similar to an early draft of the final capstone project 1 report.

The milestone report compiles all the reports that you've been writing throughout the course. Hopefully, you've been keeping your findings organized and documenting in a systematic manner. You should not need to do any new data analysis for this report.

Steps:

1. Write your capstone project 1 milestone report (Google Doc, 5-6 pages) and include the following:
2. Problem statement: Why it's a useful question to answer and for whom (*get this from your proposal*)
3. Description of the dataset, how you obtained, cleaned, and wrangled it *(get this from your data wrangling report)*
4. Initial findings from exploratory analysis (*get this from your data story and inferential statistics reports)*
    1. Summary of findings
    2. Visuals and statistics to support findings
5. Update your presentation slides.

6. Update your GitHub repository with the capstone project 1 code, milestone report, document, and slides .
7. Use the link below to share your report with your mentor for feedback, and update as needed.
8. Convert to .pdf and add to your repository. Share with your peer community.

NB This report is not simply a cut and paste job of prior work. During the last two months numerous Informational Interviews have been conducted, and much research has subsequently been undertaken. Accordingly, the issues are continually being re-framed.