NYSERDA PropTech Challenge Submission

Prepared by: Michael Sweeney, BEMP, LEED AP - AKF Engineers
March 26, 2021

Introduction and Workflow Summary:

The answers and discussions below were developed alongside a data science workbook built in Google Colab. The purpose of the workbook is to document the Python machine learning workflow and to provide additional insight and context within an interactive environment. It can be viewed alongside this document and is accessible below:

https://colab.research.google.com/drive/1C4NSwZxvYLvGp3VsIeFFCzQbiN6ccCjB#scrollTo=YIDjhIZJ3NLy

A link to the 2-minute submission video can be found here:

https://vimeo.com/529105976

Input data was analyzed, parsed and visualized in the Colab notebook using Pandas, Plotly and Matplotlib. Using SciKit-Learn, a separate Random Forest Regressor model was created for each submeter as well as for an aggregate meter representing the sum of all tenant submeters. Input variables for each model included temperature, hour of day, daily building entrants, and whole-building electric consumption. Once each model was fit and evaluated for accuracy, it was then used to predict 8/31/20 consumption using inputs for that day.

The Colab notebook features more granular analyses and a deeper dive into some of the dataset's intricacies.

NYSERDA PropTech Problem Statement Answers

1. **What is your forecasted consumption across all 18 tenant usage meters for the 24 hours of 8/31/20 in 15 minute intervals (1728 predictions)?**

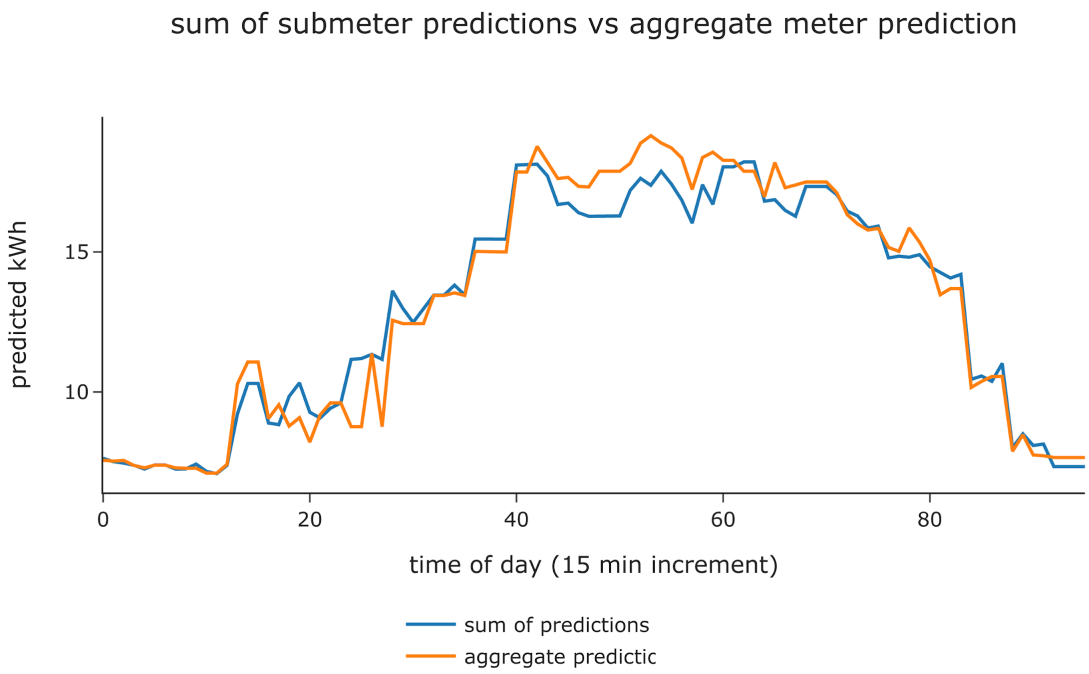
   Daily Consumption on August 8/31/2020:

   - Sum of tenant meter predictions: 1,248 kWh/day
   - Tenant aggregate meter prediction*: 1,262 kWh/day

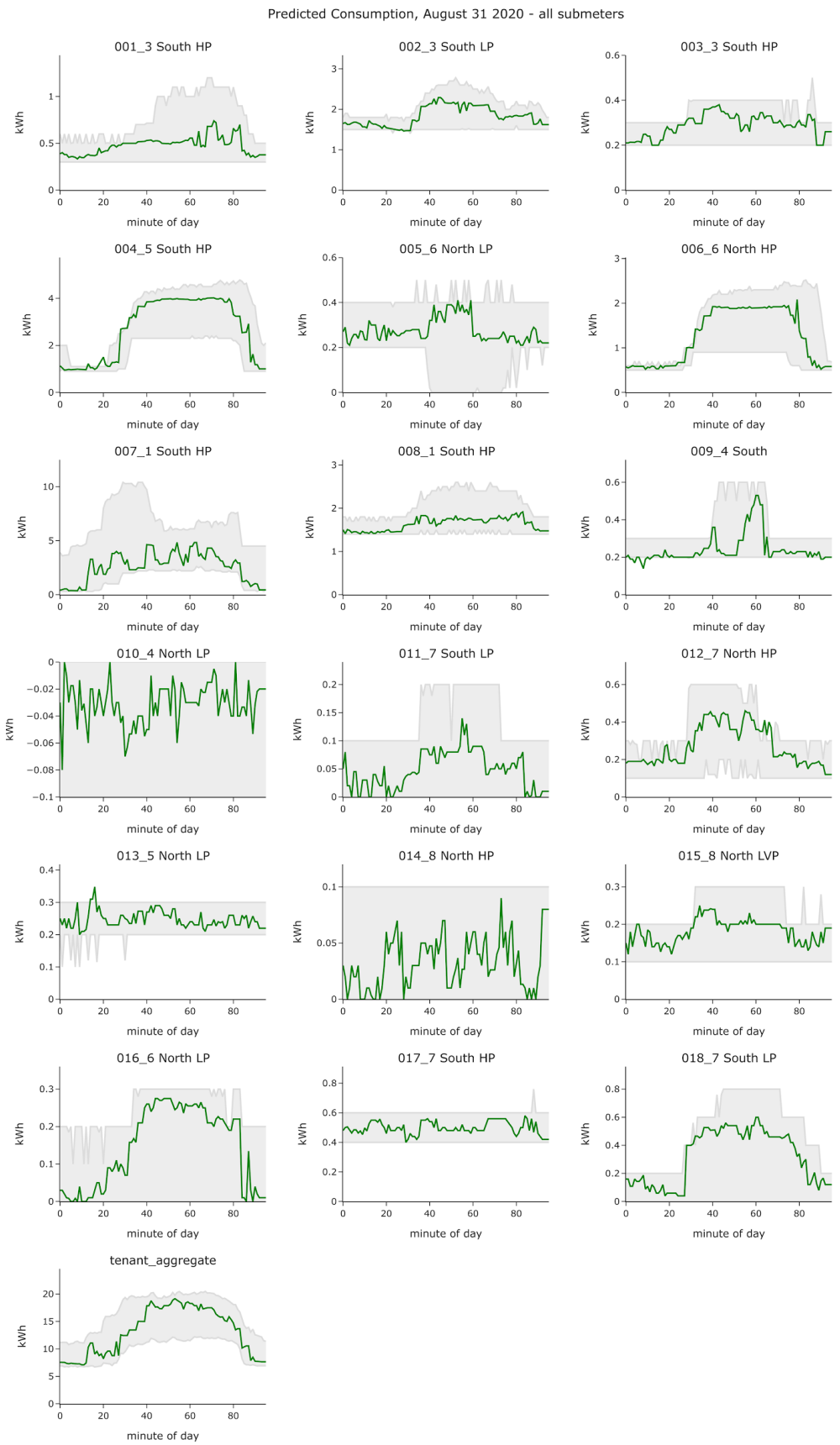
   *Although the main requirement of the submission was to provide 15 minute interval predictions for each submeter for August 2020 2021 (1728 predictions), we decided for the sake of context and a more high-level analysis that it would be helpful to create an

aggregated meter consisting of sum of all submeters, and use it as an input for a separate predictive model. Whenever the term ‘aggregate’ or ‘tenant aggregate’ is used throughout this report, it should be understood to mean the sum of tenant submeters being used as its own input and model, as opposed to individual submetered data. However, the “primary” prediction as it relates to this challenge is the sum of tenant meter predictions and the individual submeter predictions from which it is derived.

The file included in this submission, “predictions.csv”, shows predictions for all 18 tenant usage meters as well as the tenant aggregate prediction. The following graph shows the prediction outputs for the sum of all tenant submeter predictions vs the single aggregate meter prediction. They align fairly well in terms of scale, but, as expected, they vary a bit due to the fact that they were derived from two independent pipelines.



The figures on the following page show predictions for each meter and tenant aggregate (green line), alongside the 5th and 95th percentiles of observed post-pandemic data (gray band). The predictions appear to be reasonable in scale when compared to the previously-observed values.


Predicted Consumption, August 31 2020 - all submeters
001_3 South HP
002_3 South LP
003_3 South HP
004_5 South HP
005_6 North LP
006_6 North HP
007_1 South HP
008_1 South HP
009_4 South
010_4 North LP
011_7 South LP
012_7 North HP
013_5 North LP
014_8 North HP
015_8 North LVP
016_6 North LP
017_7 South HP
018_7 South LP
tenant_aggregate
kWh
minute of day

2. **How correlated are building-wide occupancy and tenant consumption?**

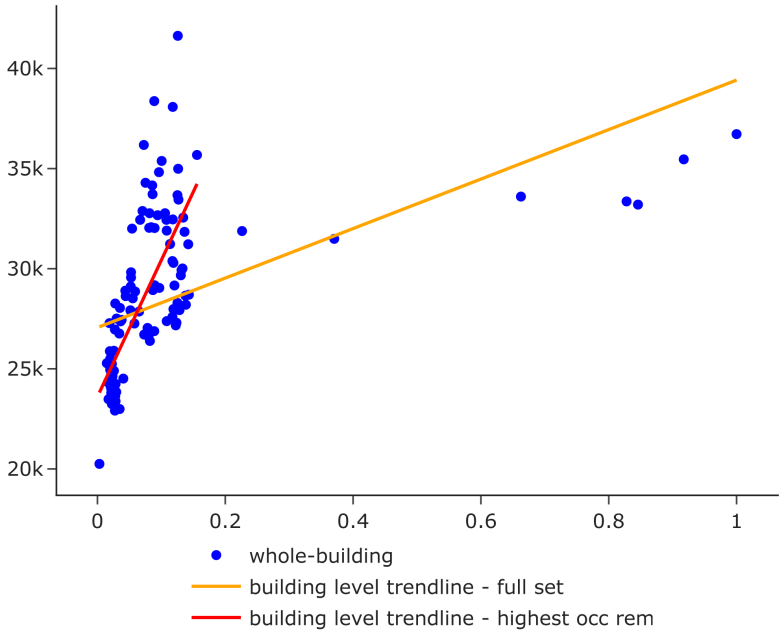
The correlation between building-wide occupancy and tenant consumption varies substantially. At the high end of correlation, meters “006_6 North HP”, “015_8 North LVP”, and “014_8 North HP”, have Pearson coefficients (a linear measure of correlation between two datasets) around 0.45, meaning their readings are moderately correlated with occupancy. Some meters, like “008_1 South HP”, are actually negatively correlated with occupancy (-0.39). However, negatively correlated meters tend to be relatively small in magnitude and are thus more sensitive to signal noise. The tenant aggregate meter has a correlation of 0.11. The table below show shows Pearson coefficients for each submeter:

006_6 North HP	0.469668
015_8 North LVP	0.467148
014_8 North HP	0.446689
013_5 North LP	0.423619
018_7 South LP	0.422683
004_5 South HP	0.417728
011_7 South LP	0.325256
002_3 South LP	0.309240
017_7 South HP	0.281697
009_4 South	0.276448
016_6 North LP	0.254589
submeter_total	0.109636
005_6 North LP	0.104485
012_7 North HP	-0.008201
003_3 South HP	-0.106884
001_3 South HP	-0.188034
007_1 South HP	-0.307786
010_4 North LP	-0.330160
008_1 South HP	-0.394561

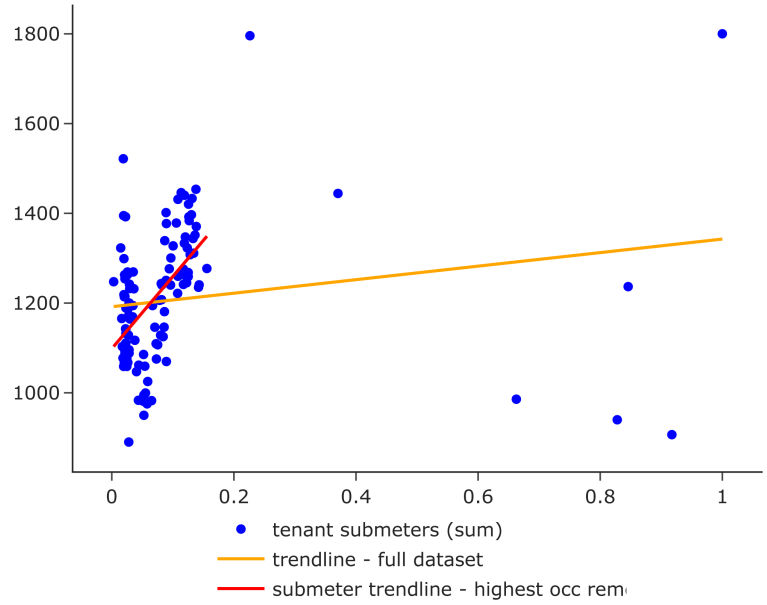
Because many tenant submeters are insensitive to building occupancy, the graphs on the following page show peak occupancy fraction (%) compared to tenant aggregate and whole-building electricity consumption.

OLS (ordinary least squares) trendlines have been shown for the full datasets (orange) and for the dataset with occupancies above ~20% removed (red). The latter shows more localized effects during post-pandemic occupancy.

Peak occupancy fraction vs. daily kWh
whole building



Peak occupancy fraction vs. daily kWh
tenant submeter



3. **What is the mean absolute error for your model?**

Sum of submeter predictions:

1. Mean Absolute Error (MAE): 1.018
2. Mean Absolute Percentage Error (MAPE): 0.059

Tenant Aggregate:

1. Mean Absolute Error (MAE): 0.734
2. Mean Absolute Percentage Error (MAPE): 0.060

Because mean absolute error (“MAE”) is scale-dependent and because multiple models were created for this submission, it may be inappropriate to use MAE to compare submeters of different scales. For this reason, the mean absolute percentage error (“MAPE”) was included, which is scale normalized. The tenant aggregate, for example, has a 0.73 MAE but a 0.06 MAPE, suggesting it may be among the stronger fits in the set despite its high MAE.

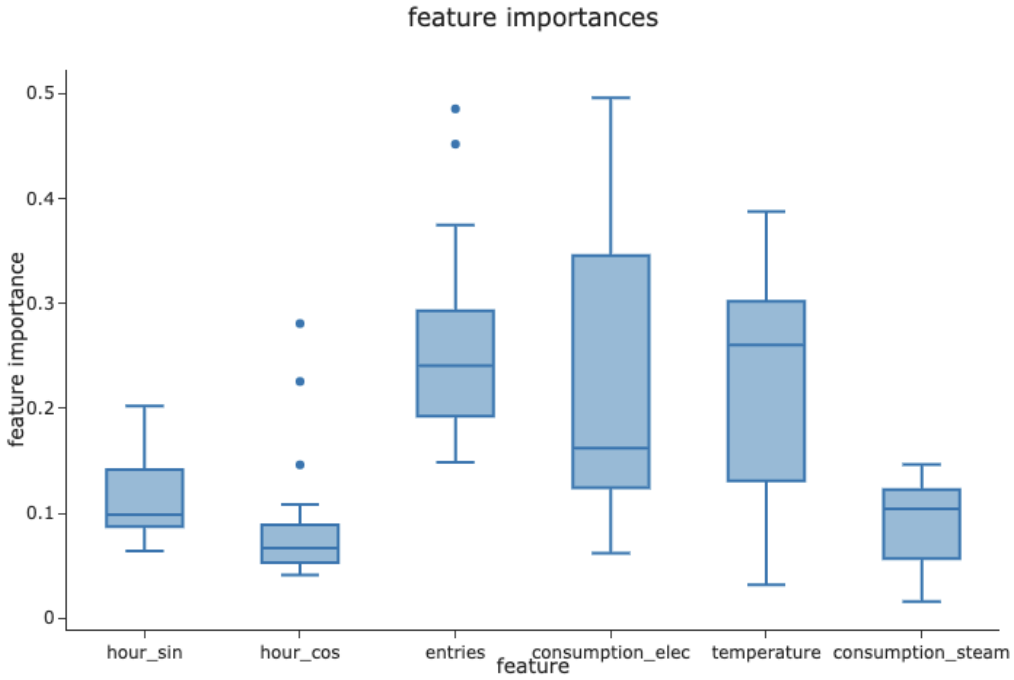
The below table shows mean absolute error for each model, as well as several other accuracy metrics and the proportion that each submeter contributes to the total.

0	train score	test score	mean_abs_error	mean_abs_percentage_error	pct_total_consumption	rmse
004_5 South HP	0.991980	0.956325	0.125138	0.069209	0.175737	0.226967
006_6 North HP	0.983920	0.925458	0.082628	0.092570	0.079526	0.140652
tenant_aggregate	0.980636	0.900387	0.733867	0.060286	1.000000	1.118181
001_3 South HP	0.957675	0.786444	0.056280	0.129274	0.036053	0.080198
007_1 South HP	0.959140	0.785067	0.571152	0.234495	0.272211	0.942559
002_3 South LP	0.944566	0.748024	0.084987	0.049263	0.135665	0.113374
016_6 North LP	0.947801	0.719431	0.033979	0.351013	0.006733	0.052026
018_7 South LP	0.946879	0.718295	0.088003	0.419495	0.013507	0.116081
008_1 South HP	0.943504	0.664677	0.095866	0.056616	0.132996	0.144217
012_7 North HP	0.892066	0.471599	0.051776	0.290767	0.017362	0.073925
003_3 South HP	0.882791	0.420575	0.033826	0.133269	0.019626	0.050819
009_4 South	0.858246	0.359951	0.040346	0.166540	0.018237	0.062015
013_5 North LP	0.824208	0.169863	0.037208	0.167133	0.020388	0.051448
005_6 North LP	0.825544	0.113600	0.072134	0.283235	0.020426	0.093571
011_7 South LP	0.819314	0.063352	0.034467	0.571848	0.002354	0.046860
015_8 North LVP	0.798712	-0.004828	0.050527	0.358547	0.012952	0.058921
014_8 North HP	0.746203	-0.274873	0.034870	0.708890	0.001837	0.047565
017_7 South HP	0.748112	-0.277581	0.089169	0.190365	0.037006	0.112706
010_4 North LP	0.697035	-0.493422	0.050114	0.731655	-0.002615	0.056960

4. **What feature(s)/predictor(s) were most important in determining energy efficiency?**

   Whole-building electric consumption, daily building entries, and outdoor temperature were most important in determining energy consumption.

   The below box plot shows the distribution of feature importances for each Random Forest Regressor model. Temperature and daily entries are roughly tied for highest mean importance, and total building electric consumption was the most widely-varied. Individual rankings for ‘hour_sin’ and ‘hour_cos’ are likely underweighted because the two input variables are representative of a single cyclical encoding for hour of day. Combined, these features may be significantly more important than when disaggregated.



The table below shows individual feature importances for each meter/model:

metername	hour_sin	hour_cos	entries	consumption_elec	temperature	consumption_steam
001_3 South HP	0.130208	0.055032	0.451754	0.108082	0.182065	0.072859
002_3 South LP	0.139888	0.280780	0.349193	0.061833	0.126573	0.041733
003_3 South HP	0.064003	0.090712	0.485372	0.110398	0.181954	0.067560
004_5 South HP	0.202218	0.070682	0.183648	0.495882	0.031665	0.015905
005_6 North LP	0.102872	0.083745	0.239993	0.151417	0.302430	0.119544
006_6 North HP	0.172361	0.048829	0.243551	0.451033	0.066923	0.017302
007_1 South HP	0.093408	0.225586	0.157449	0.102684	0.289116	0.131757
008_1 South HP	0.155932	0.047901	0.374611	0.183282	0.174911	0.063363
009_4 South	0.086946	0.108487	0.278444	0.121406	0.300405	0.104313
010_4 North LP	0.086876	0.066146	0.148641	0.164210	0.387516	0.146611
011_7 South LP	0.088322	0.076654	0.219580	0.162104	0.329910	0.123430
012_7 North HP	0.081557	0.081937	0.240779	0.223211	0.260315	0.112200
013_5 North LP	0.075631	0.064479	0.297920	0.133172	0.296553	0.132245
014_8 North HP	0.098570	0.065643	0.152182	0.209473	0.332041	0.142091
015_8 North LVP	0.089404	0.146141	0.240116	0.134384	0.283020	0.106936
016_6 North LP	0.142148	0.041279	0.223070	0.393062	0.143992	0.056450
017_7 South HP	0.089532	0.066850	0.248424	0.155462	0.328216	0.111516
018_7 South LP	0.120205	0.052283	0.269438	0.386350	0.113681	0.058042
tenant_aggregate	0.164812	0.046975	0.161355	0.484411	0.106288	0.036159

5. **What is the most energy-efficient occupancy level as a percentage of max occupancy provided (i.e., occupancy on 2/10/20)?**

   Generally speaking, maximum occupancy (2/10/20) is the most energy-efficient occupancy level for the whole building in the period for which tenant occupancy was provided. Reductions in occupancy are not sufficiently offset by proportionate reductions in consumption because the building will still remain “on” to some extent (plug loads, emergency lights, space conditioning, etc) in areas with limited tenant occupancy.

   Energy-efficient occupancy level (kwh/person/day), was evaluated for both whole-building metered kwh and tenant aggregate metered kwh.

   - Whole-building peak: February 10th, 2020. 1900 daily entries (100% occupancy), 19.32 kWh/person/day.
   - Tenant aggregate peak: March 9th, 2020. 1743 daily entries (91.7% occupancy), 0.52 kWh/person/day

   Because this analysis only includes building entries and not actual occupants in the submetered space, the tenant aggregate efficiency may be inaccurate in the event that occupancy in the tenant space diverges significantly from whole-building occupancy.

   The below table shows the top efficiency for the most efficient ten days, sorted by whole-building meter.

date_time	daily entries	frac max occ	whole building kwh/p/dy	submeter total kwh/p/dy
2020-02-10	1900.0	1.000000	19.326316	0.947421
2020-03-09	1743.0	0.917368	20.344234	0.520252
2020-03-10	1607.0	0.845789	20.659614	0.769571
2020-03-11	1573.0	0.827895	21.207883	0.597584
2020-03-12	1259.0	0.662632	26.687847	0.783002
2020-03-13	704.0	0.370526	44.730114	2.051562
2020-03-16	430.0	0.226316	74.139535	4.176279
2020-08-17	272.0	0.143158	105.514706	4.561029
2020-07-29	263.0	0.138421	107.224335	5.211787
2020-07-28	262.0	0.137895	109.351145	5.548092

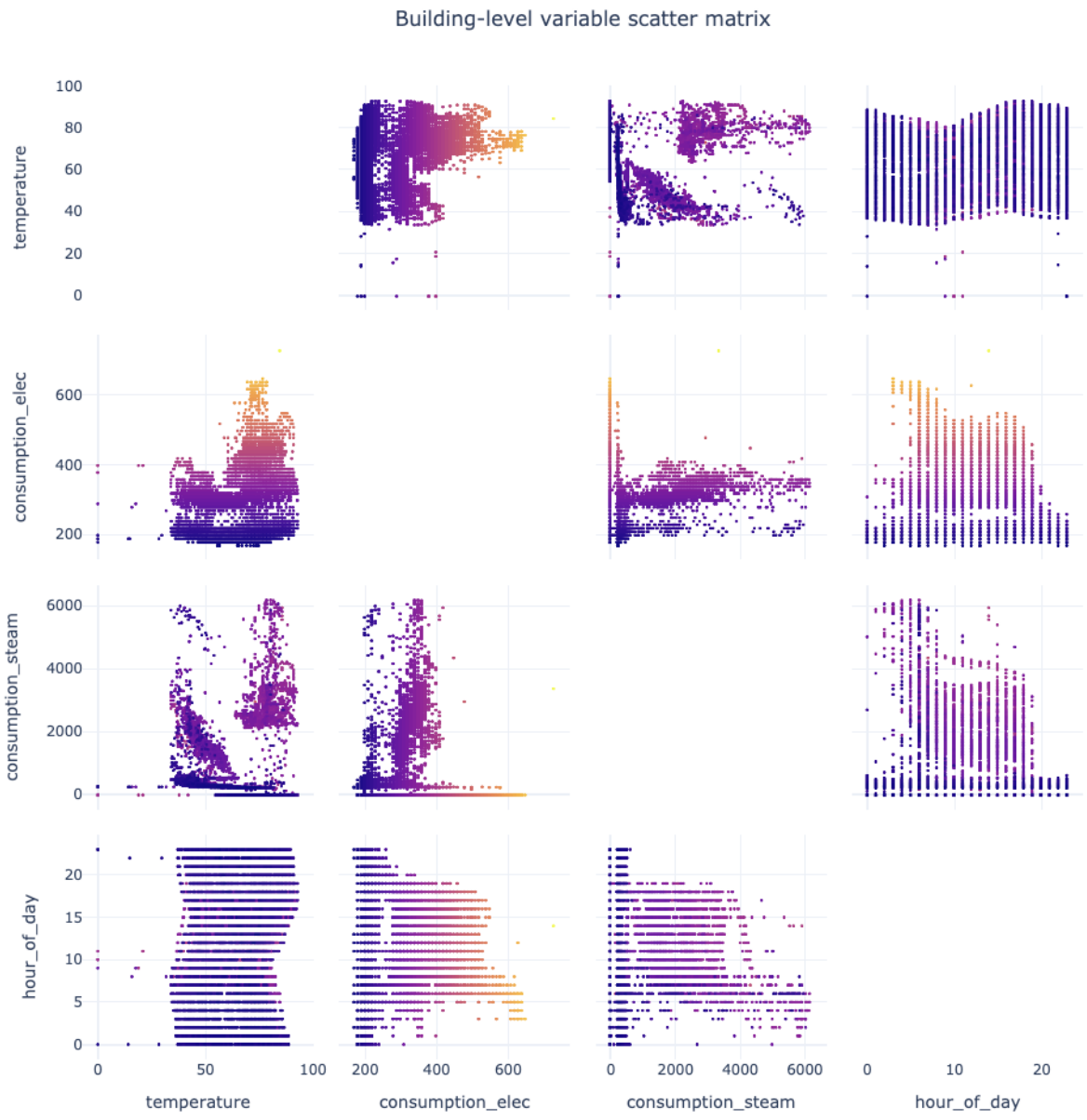
6. **What else, if anything, can be concluded from your model?**

One of the first tasks in the analysis was to evaluate and analyze the input data. A scatter matrix was created for sub-hourly input data using only periods for which daily occupancy data was available. Scatter matrices can be useful to get an overall impression of inter-related parameters and for understanding which inputs might be useful in developing the predictive model. The scatter matrix is shown on the following page. The color dimension encodes building electric consumption (yellow is on the high end of observed values, purple is on the lower end).

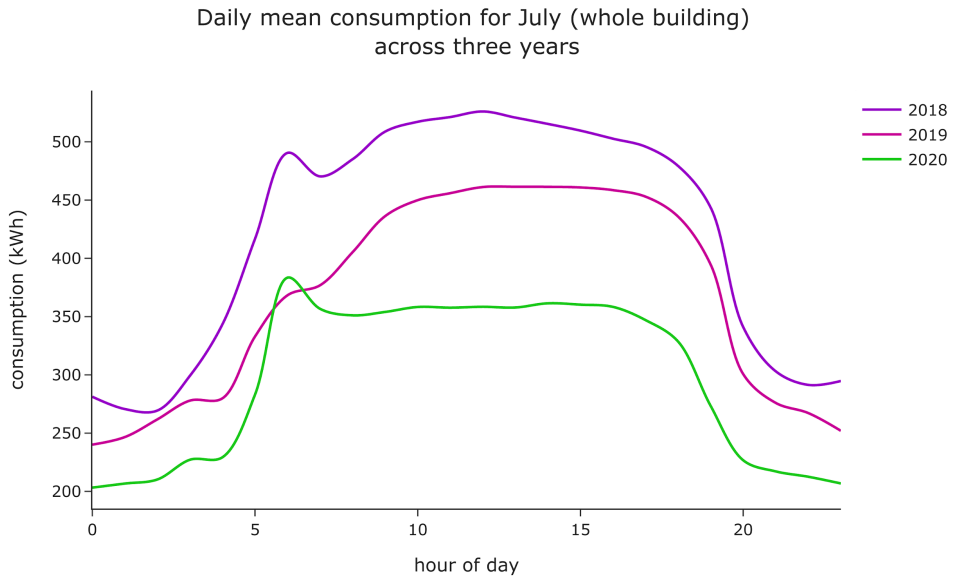
Although included in the machine learning models, tenant occupancy data was excluded from the scatter matrix because it was only available on a per-day basis. Humidity weather data was also excluded in order to simplify the scatter matrix and based on the assumption that drybulb temperature was a close-enough proxy to view ambient impacts on tenant energy consumption.

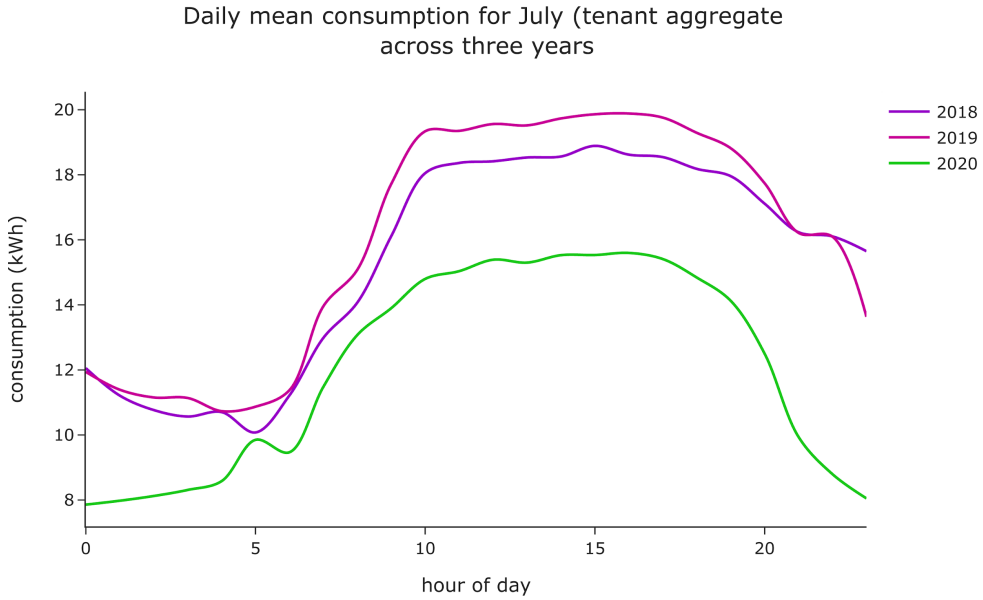
A few key trends jump out that lead to the following assumptions:

1. There is a clear correlation between steam and outdoor temperature, with steam consumption increasing noticeably above 70 degrees and below 55 degrees, but remaining fairly low in-between. This suggests that in addition to steam heating, the building has either a steam turbine or absorption chiller to meet at least part of the building's cooling loads.
2. Electric consumption also increases at outdoor air temperatures above 70 degrees. In addition to the steam chillers mentioned above, there is likely electric cooling equipment in the building (either electric chillers or DX units).
3. There is a pronounced spike in electricity consumption between the hours of 3 and 6 am, indicating a potential morning warm-up routine for the building’s air handlers.


Building-level variable scatter matrix
temperature
consumption_elec
consumption_steam
hour_of_day
0
20
40
60
80
100
200
400
600
2000
4000
6000
0
5
10
15
20
0
50
100
200
400
600
0
2000
4000
6000
0
10
20

An interesting higher-level question is the extent to which the overall building electric profile has been reduced due to the pandemic. The below graph shows daily mean consumption for each weekday in July for the years 2018, 2019, and 2020–for both the tenant aggregate and the building electric meter. Consumption and peak dropped measurably in both instances. Of potential interest is that the whole-building electricity profile has a spike around 5AM in 2018 and 2020 but not 2019. It's possible that 2019 did not have a morning warm-up routine whereas 2018 and 2020 did, or that 2019 nights in July were cooler and required less morning warm-up—or for another unknown reason.





7. **What other information, if any, would you need to better your model?**

   The following are items could help in training and/or understanding the models:

   - Actual tenant space occupancy levels rather than whole-building entries to more-tightly couple occupancy and electricity consumption in tenant spaces.
   - Hourly rather than daily occupancy information could potentially be very helpful.
   - More information regarding the submeters and the loads they serve. For example, meters serving supplemental AC units could have outdoor air temperature inputs weighted more heavily than meters serving server rooms / IDF closets, which might be less sensitive to ambient conditions.
   - More information regarding the meter errors and/or granularity of metered observations. It appears that the loads serving some submeters are too small to be captured in detail and oscillate between fairly round values. This could either be a function of meter accuracy or of decimal rounding during reporting.
   - Weekend occupancy data (even if zero), and more pre-pandemic / fully occupied building occupant data would help set lower and upper boundaries for the models' sensitivity to occupancy.