SUMMER RESEARCH REPORT

Rahim Shamsy

Student Number: 250734673

First of all, I would like to thank you – Dr. Capretz, Dr. El-Yamany, and Dr. Grolinger – for the continuous support and guidance throughout this research period. I am truly grateful for the opportunity you have given me, and the trust you put in me to pursue research in a field I was not well versed with. I have grown in not only my skills, but in my thinking capacity and my interpersonal skills. It has made a very sizable impact in my career, and I am sure I will look back upon it in the future, as one of the most transformational experiences.

ABSTRACT

The purpose of this report is to outline my findings from this summer of Research. It defines the objectives for research that evolved as a result of greater transparency throughout the summer. From correspondences with Building Management staff, it was discovered that the lack of human resources with adequate expertise has limited the Smart-Building efforts that have been undertaken. The Information Management of the buildings is being used for reactionary purposes, at best, and the true value of Data Science's application to energy consumption data is not being harnessed. This had its impacts on the nature and quality of the energy consumption data that was provided to us – with unexplained values and very high-level data. Nonetheless, data exploration and cleaning was performed with the aim to conduct Time Series Analysis for long term consumption forecasting. The poor quality of the data, coupled with the lack of CMLP-building-specific data and information on factors causing changes in energy consumption, made the data exploration and preparation phase challenging. Features used in modelling were generated based on general contextual knowledge. Seasonal Exogenous Autoregressive Integrated Moving Averages (ARIMA) models, which made use of the time series information and the generated factors/features (known as exogenous variables), were trained and diagnosed. These models were not able to explain the randomness in the data, which is an essential requirement for accurate predictions. This was primarily attributed to the lack of relevant building data (other than energy consumption), among other reasons. Nonetheless, since introduction of Exogenous Variables improved the model's ability to explain randomness in the data, moving forward, generation or collection of more building specific data (such as occupancy and equipment usage schedules) would further enhance the model. A well planned out metering system with easily accessible knowledge of the physical network of appliances and electrical circuits will prove beneficial to creating a conducive Data Science pipeline. More specific data collection would aid in providing insights tailored to specific buildings. These include, but are not limited to, occupancy in physical regions measured by various meters, schedule of electrical equipment usage, active electrical equipment inventories, and maintenance information. Finally, the university's facility-management's complexities require the creation of end-to-end solutions, which would minimize human intervention.

SUMMER RESEARCH OBJECTIVES

The democratization of data analysis algorithms and computer processing power give a greater ability to organize and make use of scores of data generated by different equipment used in a building. Although buildings' electricity systems have evolved over several years, to incorporate several state-of-the art equipment, an understanding of intricate electricity systems is exclusive to the electrical trades. Although this is a significant barrier to understanding the underlying principles and foundations of a building's equipment and distribution of electricity consumption, a marriage between Data Science applications and integrated Electrical Systems will lead to optimization of energy consumption patterns in response to several aspects such as occupancy.

The objective for summer research was to explore the electrical systems and procedures in place for the Claudette MacKay-Lassonde Pavilion (CMLP) Engineering Building. In doing so, key insights were to be provided for making the new ThreeC+ Engineering Building more energy efficient. The open-ended nature of this task allowed for flexibility and aided in pursuing the most ideal path towards transferrable insights. The following explains the evolution of objectives for summer research:

1. Explore the procedures currently upheld by Facilities Management personnel, and how conducive they are to supporting energy efficient procedures. This would be useful in enabling the research group to make more informed decisions.

2. Perform Time Series analysis on energy consumption data collected over several years for different electrical circuits within CMLP; visualize and manipulate data to generate key insights that can contribute to the research group's knowledge of energy consumption patterns.

3. Use information found in time series exploration to create Exogenous and Non-Exogenous Seasonal Autoregressive Integrated Moving Averages (ARIMA), so as to forecast long-term energy consumption for the building. This step would be useful in explaining how insights translate to a more accurate model. Although the research group has carried out research on energy consumption predictions with London Hydro, only short term energy predictions were made. Since energy predications in the short term horizon are more accurate, finding a means to predict long term energy consumption patterns with different 'regressors' would benefit decision makers.

RESEARCH PROCESS AND FINDINGS

Institutional Findings

In order to accomplish the objectives, which were set out for the summer, research and learning had to be performed on two fronts: institutional and analytical. An understanding the building management procedures at Western University was developed, which defines the constraints within which new solutions to existing problems can be deployed for the new ThreeC+ Building. By exploring and analyzing the data, key insights were obtained pertaining to energy usage patterns of CMLP building, which will be similar to that of the new building.

Over 3 meetings with Facilities Management personnel, several aspects of the building-management procedures and resources were learnt. A great number of sensors make a variety of measurements across campus; however, only a handful of these measurements are stored permanently. A vast majority of the sensors collect measurements that are only available for live visualization through a Web Platform. Therefore, most of the data that is measured is not actively monitored for optimization, but only for troubleshooting when issues arise. It is also evident that the University has not extracted maximum value from the Building Information System in place, as it is used for reactionary rather than for precautionary purposes and planned activities. Because the electricity metering for the building was not planned at the time according to needs of the future, it remains to be used for functional purposes at best. The system of meters and sensors, and the type of data collected are not tailored to needs of energy-efficient buildings, rather investments have been made on the former for trivial purposes such as decisions pertaining to whether or not an electricity 'line' has enough capacity to support an entire new building. Nonetheless, this sentiment was echoed by Facilities Management personnel – much cannot be accomplished with the resource constraints on the human resources.

In spite of these persistent problems that have limited the university's accomplishments pertinent to building efficiency and energy conservation, there exists several opportunities to fill human resource gaps by automating several tasks. These tasks include monitoring abnormal electricity consumption, deploying field engineers to issue-sites, and performing predictive maintenance among others.

Analytical Findings

Tools Used

The summer research period, during which I was tasked to accomplish the aforementioned objectives, consisted of steep learning curve early on, given my background in different design aspects of a building. The month of May was primarily spent on learning various tools that would be necessary for Data Science-related work, which would ensue for the remainder of the summer. These include learning about the aspects of the Python programming language and environment (through the usage of browser-based Jupyter Notebooks IDE) pertaining to Data

Science work. This step proved to be an important one, since the tools were essential to manipulate, edit, clean, and explore the electricity consumption data. These tools are as follows: Python Numpy for Matlab-like multi-dimensional array creation and manipulation, used for improvement of algorithms' performance on the Array Data Structure; Python Pandas package for arrangement, easy manipulation and exploration of multivariate data, represented by Series and DataFrames (such as those in R programming language); Python Matplotlib package for creation of graph objects for variety visualizations of different aspects of data in DataFrames; Python 'seaborn' package, built on top of Matplotlib, for greater ease of creating more sophisticated graphing plots.

Description of Data and Data Exploration

Equipped with a basic understanding of data manipulation and exploration tools offered through the several Python packages, data that was collected for the CMLP building was explored to better understand underlying effects attributed to time. Data collected for CMLP spanned over 5 years – from 2012 to 2017 – with a granularity of 5-minute intervals, i.e. measurements were made and data points were stored every 5 minutes. The fields in which measurements were made include steam and ethanol flow rates for the heating systems, water flow rates through several pipelines within the building, instantaneous electricity consumption data (in kW units), and cumulative electricity consumption data (in kWh units) for various circuits within the building. Every 5 minutes for each data field, the timestamp value and the measured value are stored. While a total of greater than 100 000 data fields exist throughout campus (collecting different types of data pertaining to buildings), there are only about 50 such data fields for CMLP that are actually stored in the cloud, many of which do not involve electricity data. Nevertheless, the data fields that explain electricity consumption over the years were categorized into differently numbered meters, representing separate electricity circuits.

There were several barriers to Facilities Management personnel's active involvement pertaining to building energy efficiency, the most prominent being lack of human resources. As a result, the data collected over the past 5 years has not been detailed and diverse enough to train Data Science models with the specificity necessary to get key insights. More specifically, each data field consists of a time-stamp and an energy consumption value. Neither is the physical region of CMLP represented by this data known, nor is it known which appliances lie in the circuits within which the meters are placed. Although exploring how this information can be determined as is would be important, it significantly reduced the information available and limited what could have been achieved through analysis.

Given the inherent limitations to the data made available, the most suitable course of action for exploration of the data was to perform Time Series Analysis. Furthermore, because of the group's focus on electricity-related research, only energy consumption data fields for CMLP were used in analysis. Through Time Series Analysis, the energy consumption data fields were explored through visual representation and numerical summaries to get a better understanding of the quality of data. Upon this analysis, the data was cleaned and prepared for Time Series Forecasting through Seasonal Autoregressive Integrated Moving Averages (ARIMA) models.

Considered among a group of several Supervised Machine Learning algorithms, ARIMA models are strengthened by the use of regressor-variables other than electricity consumption (called exogenous variables), the effect of which were also observed. Nonetheless, having to generate these from scratch limited the specificity to which they could be generated, and the type of data pertaining to the CMLP building that could be generated. For instance, while data for occupancy at the modular level, performance of different equipment of the building, and other variables that affect these would have been useful, they could not have been generated.

Upon preliminary exploration of cumulative energy consumption (in kWh units) data for 6 different meters, it was found that there were four distinct groups of kWh values for each meter. Determining this made it relatively easy to extract only the values that made logical sense to use in Time Series Analysis. One group of values, which was deemed to be regular, included a variety of kWh values for time stamps at 5-min intervals. Nonetheless, this type of data did not include several timestamps, which would later be imputed through interpolation. The second such distinct group of data included repeated timestamps with repeating extremely high values. These were deemed to be irregular, either caused by meter reading errors or database errors, and were subsequently removed. Lastly, two other distinct groups of data included values that seemed normal at first, but were for time-values between the 5-min intervals – the irregularity of this was proof for its removal as well. The following give examples for 3 irregular data groups within all electricity consumption data fields that were removed. The data provided is for Meter 2 in the building:

	seq	t	ts	m2kWh_tot
3968	999595916	9	2012-06-13 14:08:55.180	-205.0
7680	999599628	9	2012-06-26 11:11:14.240	-99.0

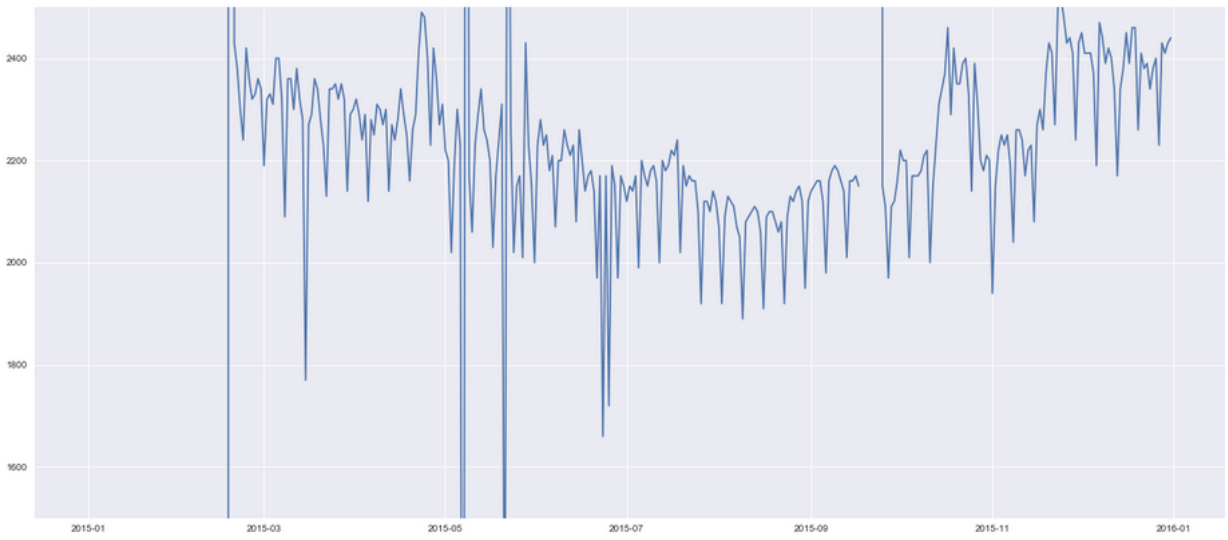
	seq	t	ts	m2kWh_tot
3355	999595303	0	2012-06-11 11:19:06.000	32770.0
3356	999595304	0	2012-06-11 11:19:06.000	2.0
4163	999596111	0	2012-06-14 06:19:11.200	32770.0
4164	999596112	0	2012-06-14 06:19:11.200	2.0

	seq	t	ts	m2kWh_tot
3278	999595226	8	2012-06-11 04:50:00.100	16974085.0
3279	999595227	8	2012-06-11 04:50:00.100	16974085.0
3280	999595228	8	2012-06-11 04:50:00.100	16974085.0
3281	999595229	8	2012-06-11 04:50:00.100	16974085.0
3282	999595230	8	2012-06-11 04:50:00.100	16974085.0

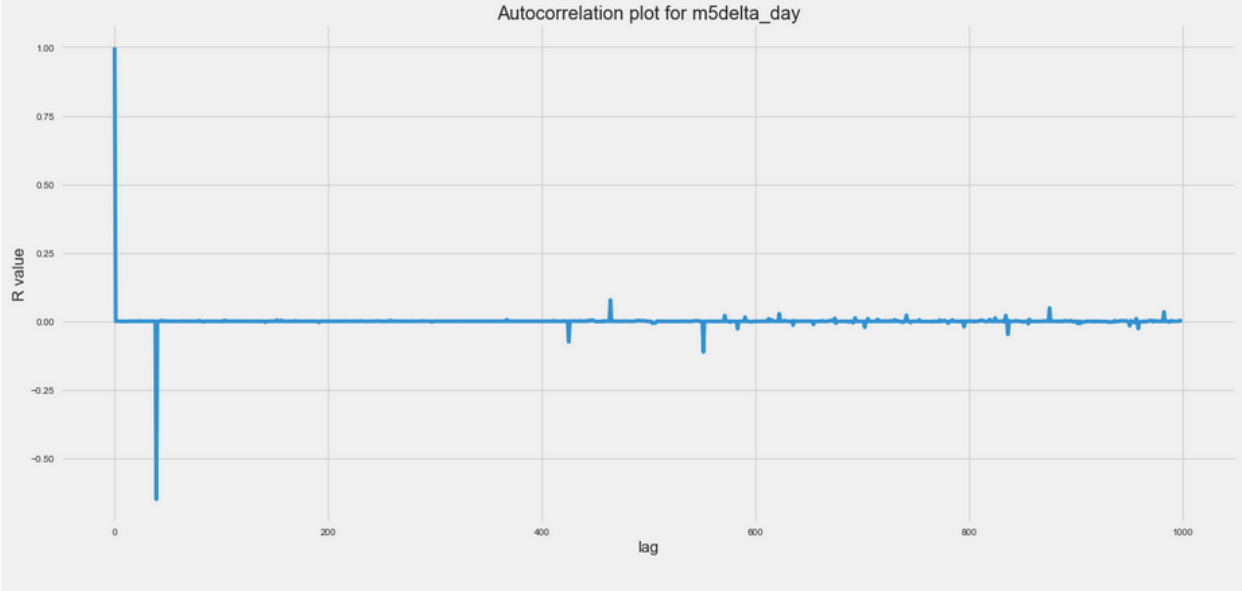
Data Cleaning and Preparation, and further Exploration

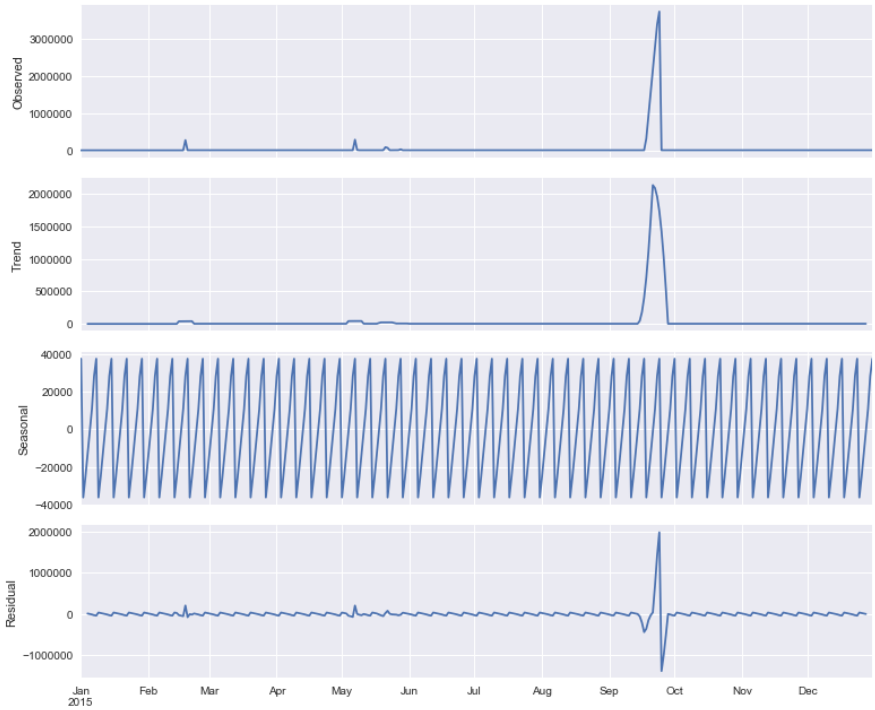
Although data of each of the 6 meters (representing 6 different circuits in the CMLP building) recorded the same type of data (kWh), the number of reliable data points were more numerous for Meter 5 (m5) than that of any other meter. For this reason, only m5 data was used, given that the same programming code for other data fields could be reproduced due to similarities. From now onwards, the analysis presented is for m5. Although the small granularity of the m5 data was conducive to creating more accurate forecasting models, it was difficult to visualize individual movements of data since of more than three hundred thousand timestamp recordings for m5 alone. At the same time, it was realized that cumulative electricity consumption did not yield significant visual insights since a small granularity limited low-level visualizations with ease. Only an upward trend was noticeable. Two operations were performed

to solve this: differential data was calculated and used (using change in cumulative kWh as the data, known as the 'delta'), and the data was resampled to daily granularity (every 24 hours) by summing up delta values for every 24 hours. Visualization of the data following these changes led to the following key insights about change in energy consumption in m5:
Seasonality of change in energy consumption (delta values) was observed within a week (every seven days) with peaks at midweek, and troughs at weekends. Values were also observed to peak (on average) towards the beginning of the year, and to reach minimum average values in the middle of the year – signifying lesser use in the summer due to fewer students, and greater energy consumption in the winter due to low temperatures. Although the reasoning is obvious, it confirms the validity of the data to a certain extent. The following visualization shows the weakly seasonal cycles of delta values (kWh) for year 2015 for m5 (delta vs date graph):



Autocorrelation analysis was performed on the data to determine how values within the same time series (m5) were correlated to each other, i.e. how one value in the past affects, another in the future. In this analysis, the number of 'lags' represents how far apart any two data points are; i.e. how many consecutive timestamps apart. For example, the value of autocorrelation (similar to R-value in linear regression – how well a trend fits data) at lag 39 below represents the correlation between any value in the time series and the 39th value after it. In the case below, the autocorrelation value is approximately -0.65, which is quite strong. This would mean that every 6 weeks, the data is heavily correlated. The autocorrelation plot can be seen below.

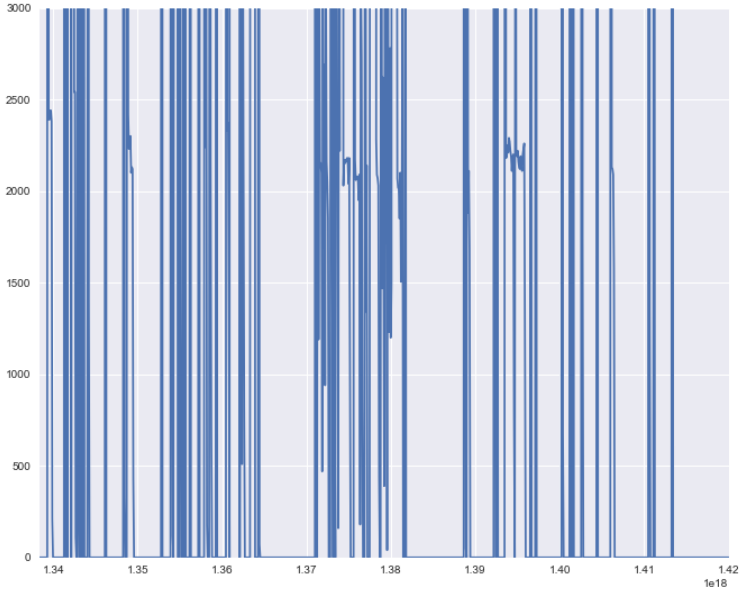




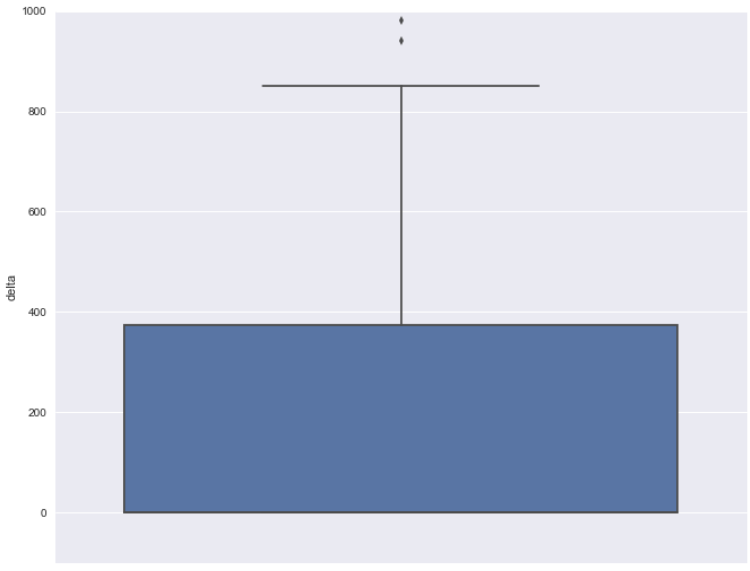
In addition to visualizing the data points themselves, and doing an Autocorrelation analysis, the time series was decomposed into its distinctive parts to further deconstruct signs in the previous visualization. The decompositions created for data within m5 for year 2015 is shown to the left, as an example. These parts include the trend (signifying overall increases/decreases in average values), the seasonality and the

residuals (significant outliers), and are separated using the seasonal decomposition method of 'statsmodels.tsa' package in Python. From this decomposition, it can be seen that weekly seasonality is prominent for the year 2015 for the m5 time series, like inferred in the time series plot shown above. Furthermore, there is no visible trend in general during all years; however, this is attributed to differencing that was performed on the cumulative data to obtain delta values. The visible seasonality and lack of trend is similar for all other years.

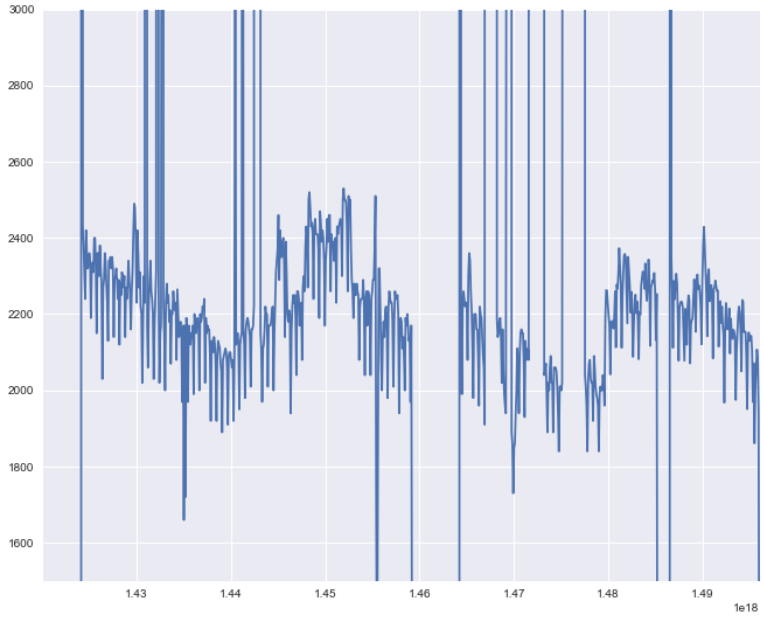
To further explore the high variation in certain regions of the data visualised above, data was separated into two periods – before 2015, and after (and including) 2015. Visualizing these separately and seeking summary statistics through box plots, the following was observed for the two periods (respectively):



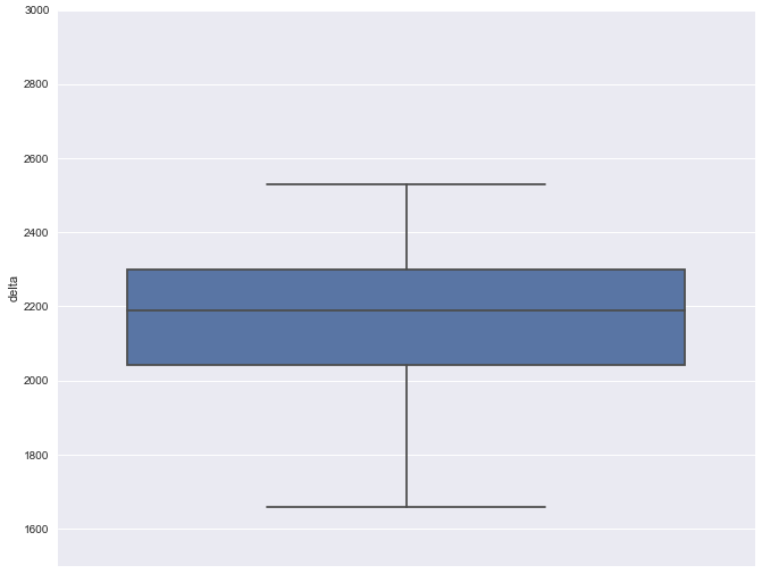
2012-2014 Daily Delta Data:



2012-2014 Daily Delta Boxplot



2015-2017 Daily Delta Data



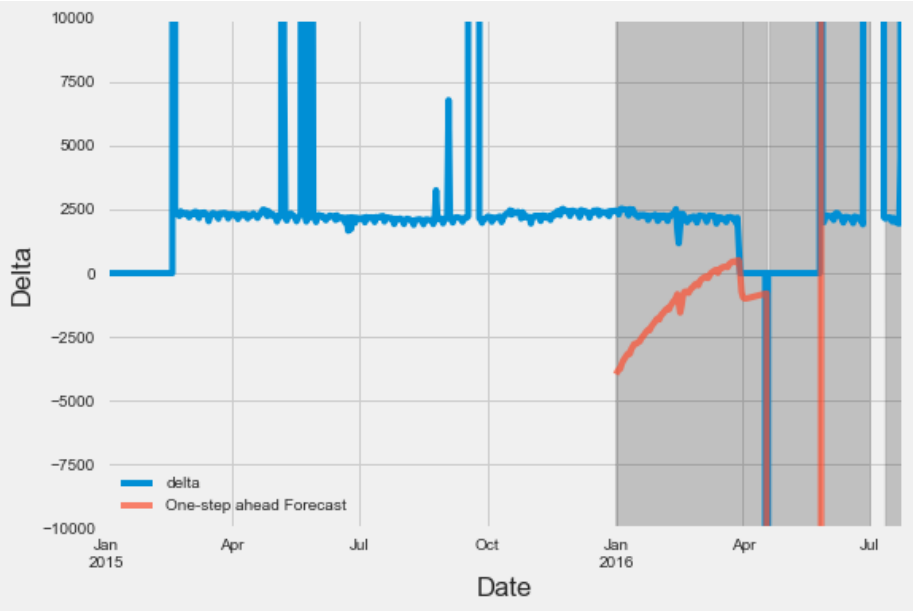
2015-2017 Daily Delta Boxplot

Through these visualizations, it was observed that the data before 2015 was excessively erratic, and not reliable, since the median delta value was that of 0 kWh (an inconceivable number since it is not possible for daily consumption to be zero for many days). The data values for 2015 onwards showed more regularity with visible seasonality in line with our expectations for yearly and weekly changes. Furthermore, the second boxplot reveals a more reliable median daily consumption of approximately 2200 kWh, with minimum values dropping as low as 1600. These figures were more regular, and as a result, the data was truncated to include values from 2015 onwards only. This was the final (cleaned) data that was to be used for Time Series Forecasting, using ARIMA model training.

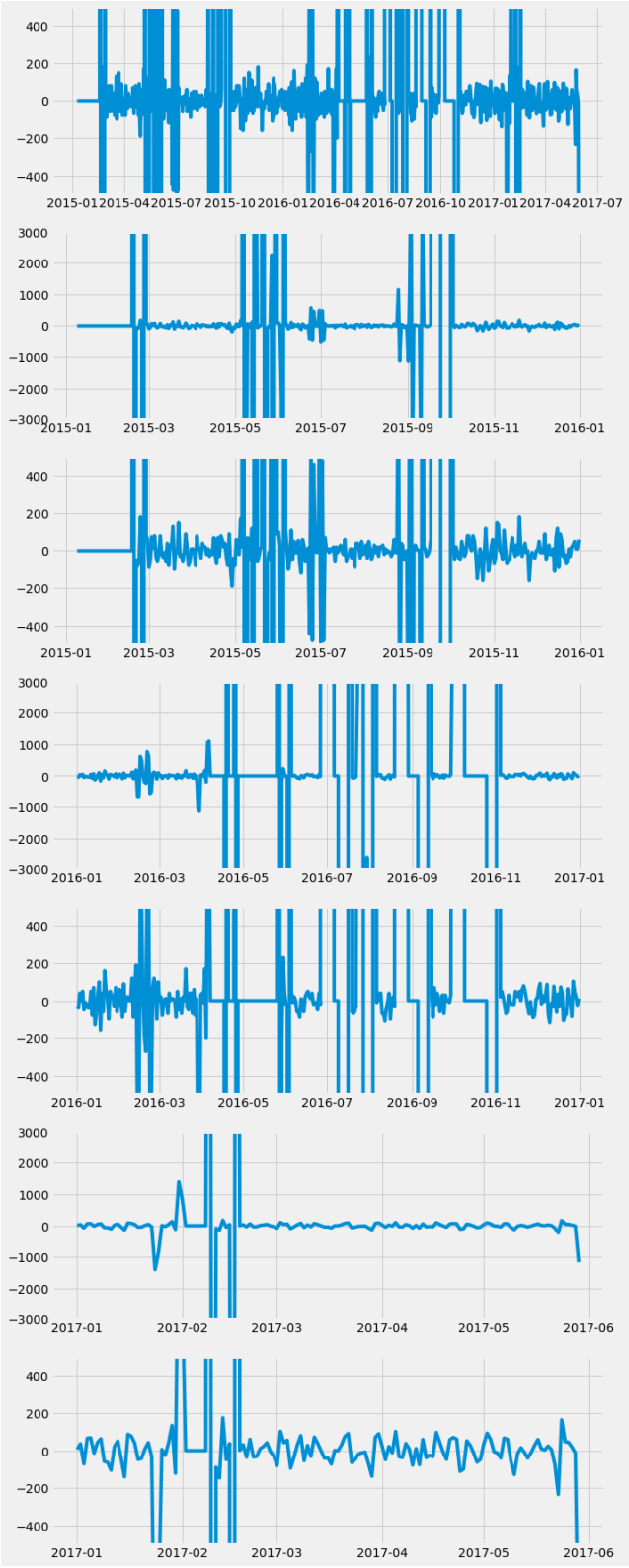
ARIMA Modelling

ARIMA Modelling combines Autoregressive (AR) models, which use correlations among values of the same time series to predict future values, and Moving Averages (MA) models which uses the correlation among residuals, which is the difference between observed values and predictions of those observed values. Additionally, it employs differencing of the values (denoted by the 'I' in the name) so as to make the data stationary. Stationary data is a prerequisite for ARIMA modelling. Stationary data represents data that does not exhibit seasonality or trends, but rather appears as white noise – the information of the data that is not explained simply by a trend or seasonality. ARIMA modelling attempts to create relationships and explanations of the residual white noise, so that predictions can be accurately made. By using the autoregressive relationships, seasonality and differencing performed, the model could only go so far as to make predictions. Predicting the residuals would significantly improve predictions.

The first step in preparing the m5 daily delta consumption data was to make it stationary. Two possible methods would serve this purpose. A first order difference is performed when data is transformed to the difference between consecutive data points (reducing the size of the data by one observation consequently). In other words, a certain data point's value is subtracted from a consecutive data point's values – the difference being the transformed data point. A seasonal difference is performed when a difference is computed between a certain data point and another data point in a consecutive seasonal cycle that appears in the same point in the seasonal cycle. Therefore, the number of data points between the two data points used to compose a seasonal difference would be one less than the length (in data points) of one cycle – 6 days in this case. It is deprecated to difference a data set more than twice. Therefore, a combination of the two types of differences were used on the dataset (which was non-stationary initially). Not stationarizing the data set at all would yield poor forecasts, as showed in a rough model's predictions:



The combination of differencing that yielded the most stationary data set was one seasonal difference, followed by one first order difference. The ensuing dataset was visualized as follows:

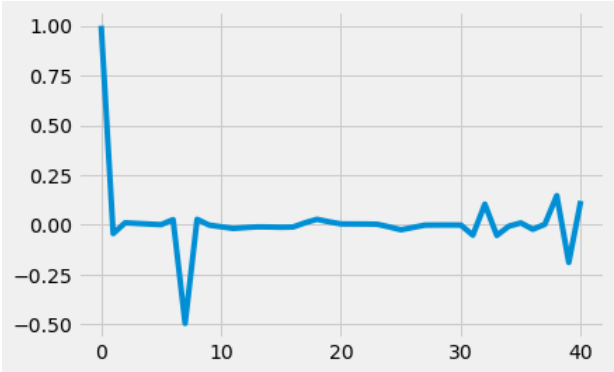


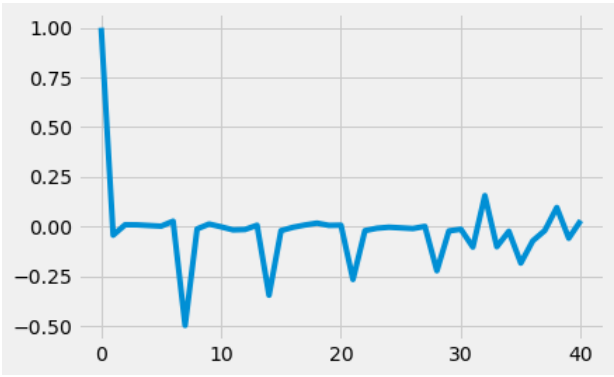
From the figure, which shows different aspects of the data set, it can be seen that the values, other than extreme ones, show non-seasonal data points and have a consistent average value of 0. For the non-extreme values, the data looks like white noise, which contributes to the signs that it is stationary. Other combinations of differencing didn't eliminate seasonality completely.

At this point, the data was separated into training, and test sets in a 3:1 ratio. This was done so that the model could be tested on 'new' data that it had not seen before in training/modelling.

Having prepared the data, the next step in ARIMA modelling was to determine its parameters. These parameters were that of the Autoregressive (AR) and Moving Averages (MA) models' equations, put together to form the ARIMA equation. Each of the AR and MA equations could include up to an infinite number of parameters, with circumstantial dependencies that will be discussed below. The method that was used to determine these parameters was the Maximum Likelihood Estimate (MLE), by trying different values of the parameters, and calculating the proportion of observations that were accurately predicted within a confidence interval. The parameters that would yield the highest proportion (hence, the probability) would be used as the parameters for the ARIMA equation. Within the same Python statespace.tsa package, there were methods to fit a model using the MLE method. Nonetheless, the number of AR and MA parameters had to be determined.

The number of parameters in the ARIMA model (attributed to AR and MA terms) was determined using two methods simultaneously – graphical and through the use of a test statistic. The provisional number of AR and MA parameters were determined using the graphical method, by plotting the Autocorrelation and Partial Autocorrelation Functions (ACF and PACF). The following was visualized (first the ACF, followed by the PACF):



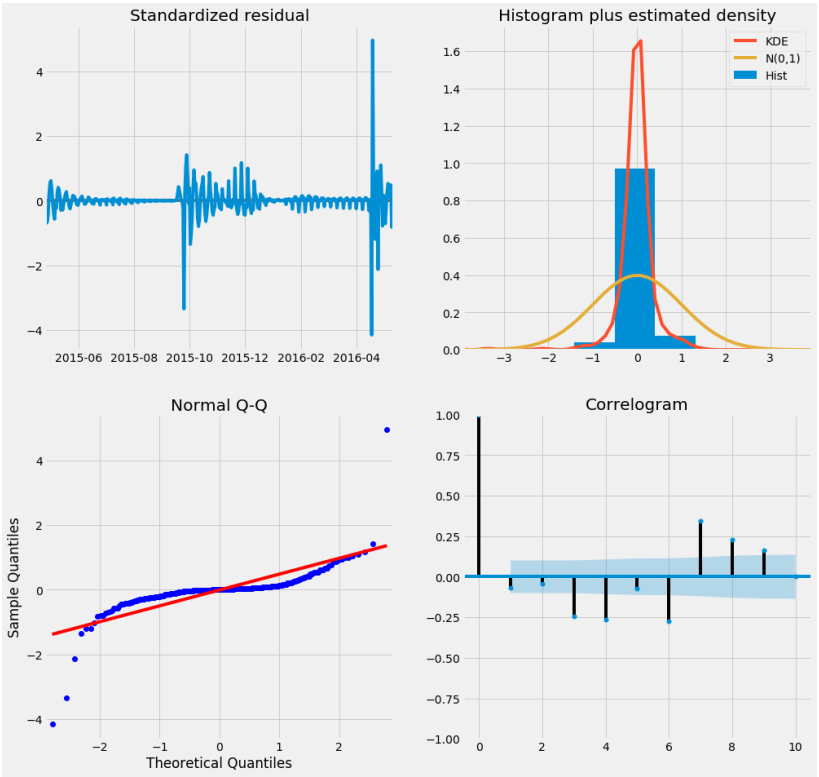


Rules[1] for determining the number of AR and MA parameters (each for seasonal and non-seasonal portions of the data set) dictated that there should be 7 MA terms from the ACF plot, since there is a sharp cut-off of ACF values after lag 7. The PACF plot did not come to use because the ACF value at the first lag (lag-1) was negative. The number of AR terms would have been determined using the PACF plot otherwise. Knowing that these ACF and PACF plots are classical MA model plots, the number of AR terms was set to 0. The ARIMA model-fitting function provided by statespace.tsa included the parameters as follows for seasonal data (before stationarizing): ARIMA(p,d,q)x(P,D,Q,S), where p = number of non-seasonal AR terms, d = number of non-seasonal differences (i.e. first order), q = number of non-seasonal MA terms, P = number of seasonal AR terms, D = number of seasonal differences, Q = number of seasonal MA terms, and S = number of lags between two points on a seasonal cycle. Immediately, p = 0, d = 1, q = 7, P = 0, D = 1, and S = 7 can be set aside. Therefore, ARIMA(0,1,7)x(0,1,Q,7). Next, the test statistic Aikeke Information Criterion (AIC) was used to determine the ideal values for Q, while also testing different values of q to ensure the ideal value is selected. The model yielding the lowest AIC test statistic would be used to model the time series. It was confirmed that the value of q (the number of non-seasonal MA terms) yielding the lowest AIC value was in fact 7. At the same time, the number of seasonal MA terms that yielded the lowest AIC value was that of 13. Therefore, the Seasonal ARIMA model that was to be fit, with its corresponding number of terms was ARIMA(0,1,7)x(0,1,13,7).

Upon training the model on training data provided, the model was diagnosed using reports and graphs.

[1] http://people.duke.edu/~rnau/arimrule.htm

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-1.0219	0.067	-15.301	0.000	-1.153	-0.891
ma.L2	0.3563	0.169	2.106	0.035	0.025	0.688
ma.L3	0.2650	0.185	1.435	0.151	-0.097	0.627
ma.L4	0.0037	0.157	0.024	0.981	-0.305	0.312
ma.L5	-0.0968	0.182	-0.530	0.596	-0.454	0.261
ma.L6	0.2544	0.167	1.526	0.127	-0.072	0.581
ma.L7	-0.5388	0.181	-2.969	0.003	-0.894	-0.183
ma.S.L7	-2.0730	0.135	-15.313	0.000	-2.338	-1.808
ma.S.L14	1.2385	0.372	3.333	0.001	0.510	1.967
ma.S.L21	-0.1556	0.640	-0.243	0.808	-1.411	1.100
ma.S.L28	0.0239	0.743	0.032	0.974	-1.433	1.481
ma.S.L35	-0.0566	1.032	-0.055	0.956	-2.080	1.966
ma.S.L42	0.0709	1.607	0.044	0.965	-3.079	3.221
ma.S.L49	0.0404	2.182	0.019	0.985	-4.236	4.317
ma.S.L56	-0.0096	2.062	-0.005	0.996	-4.051	4.032
ma.S.L63	-0.0502	1.425	-0.035	0.972	-2.842	2.742
ma.S.L70	-0.0287	0.993	-0.029	0.977	-1.975	1.917
ma.S.L77	-0.0201	0.988	-0.020	0.984	-1.957	1.917
ma.S.L84	0.0089	1.062	0.008	0.993	-2.073	2.091
ma.S.L91	0.0460	0.519	0.089	0.929	-0.972	1.064
sigma2	1.352e+12	1.05e-12	1.29e+24	0.000	1.35e+12	1.35e+12

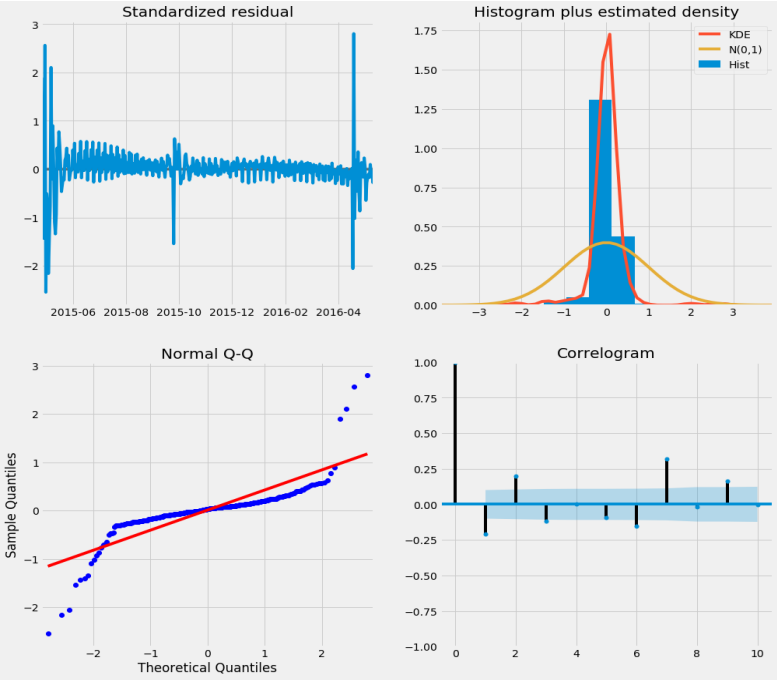


The first figure on the left shows a summary table, the rows of which represent each parameter (equation coefficient) of the ARIMA model. Each parameter's value is given, as well as other information under other columns. The 'P > abs (z)' shows the significance of each feature weight/coefficient of the ARIMA model parameters. In other words, the column shows the test statistic P-value of the feature weights to quantify, put simply, the probability/confidence with which the coefficients capture accurate predictions for all data points in the data set. Ideally, the value in that column should be less than a confidence level selected – one that I thought should be around 10% for a provisional model. Unfortunately, several of the parameters have P-values much greater than that significance level, suggesting a poor model. A solution to this problem would be eliminating those parameters completely, but that would severely under fit the data, overgeneralize relationships and making subsequent predictions of such a model would have even higher error rates.

The second figure above shows graphical summaries of the model's performance. The top right graph shows that the distribution of residuals (the prediction errors). This distribution is ideal, since the residuals (denoted by the blue-shaded histogram) are normally distributed. Recall that the data set for ARIMA modelling is stationarized so that we could extract as much information from the random white noise in the data. Similarly, residuals of predictions must not only be

normally distributed, but must not have any correlations. Viewing a plot of the residuals (top left plot) and the Autocorrelation Function of the residuals (bottom right plot), it can be seen that there is information in the residuals (seasonality) that is present. The autocorrelations for lags beyond the blue-shaded region in the bottom right plot explain the observable seasonality in the top left plot. The presence of seasonality in residuals of predictions is deprecated, and this contributes to the weaknesses of the model trained. The model has a clear need to be made stronger. Given that there is seasonality in the residuals that is not captured by the model, the introduction of additional regressors could solve the problem. For this reason, and Exogenous ARIMA model was trained.

	coef	std err	z	P>|z|	[0.025	0.975]
x1	2986.9339	nan	nan	nan	nan	nan
x2	-2875.9444	2.29e+04	-0.126	0.900	-4.77e+04	4.19e+04
x3	2.076e-05	1.94e+04	1.07e-09	1.000	-3.8e+04	3.8e+04
x4	9.361e+04	nan	nan	nan	nan	nan
x5	2.478e+04	4200.602	5.900	0.000	1.66e+04	3.3e+04
x6	1961.2281	1.44e+04	0.136	0.892	-2.63e+04	3.02e+04
x7	-6160.8999	2.08e+04	-0.296	0.767	-4.69e+04	3.46e+04
const	-5.563e-11	304.169	-1.83e-13	1.000	-596.160	596.160
x8	6169.1876	1.64e+04	0.376	0.707	-2.6e+04	3.83e+04
x9	-6681.2289	1.87e+04	-0.357	0.721	-4.33e+04	3e+04
x10	-7427.0448	1.48e+04	-0.500	0.617	-3.65e+04	2.17e+04
x11	-1.244e+04	6390.419	-1.946	0.052	-2.5e+04	87.836
ma.L1	-0.7821	0.234	-3.344	0.001	-1.240	-0.324
ma.L2	-0.0234	0.485	-0.048	0.962	-0.974	0.928
ma.L3	0.4324	0.611	0.708	0.479	-0.764	1.629
ma.L4	0.1989	0.702	0.283	0.777	-1.177	1.575
ma.L5	0.0989	0.785	0.126	0.900	-1.439	1.637
ma.L6	0.1802	0.472	0.381	0.703	-0.746	1.106
ma.L7	-0.5981	0.712	-0.840	0.401	-1.994	0.798
ma.S.L7	-2.9020	0.493	-5.885	0.000	-3.869	-1.936
ma.S.L14	1.7214	1.429	1.205	0.228	-1.079	4.522
ma.S.L21	0.2897	2.491	0.116	0.907	-4.592	5.171
ma.S.L28	0.2591	7.457	0.035	0.972	-14.356	14.874
ma.S.L35	-0.0887	11.071	-0.008	0.994	-21.788	21.610
ma.S.L42	0.0200	11.156	0.002	0.999	-21.846	21.886
ma.S.L49	-0.1257	12.645	-0.010	0.992	-24.910	24.658
ma.S.L56	-0.1815	15.263	-0.012	0.991	-30.096	29.733
ma.S.L63	-0.0434	16.230	-0.003	0.998	-31.854	31.767
ma.S.L70	-0.1035	15.357	-0.007	0.995	-30.202	29.995
ma.S.L77	-0.0530	13.772	-0.004	0.997	-27.045	26.939
ma.S.L84	0.1128	11.145	0.010	0.992	-21.732	21.957
ma.S.L91	0.0789	4.803	0.016	0.987	-9.334	9.492
sigma2	1.337e+12	nan	nan	nan	nan	nan



Exogenous ARIMA

Exogenous variables could be introduced in the training of the model through the same statespace.tsa package in Python, with the SARIMAX method, used for Seasonal Exogenous ARIMA (SARIMAX model). For a provisional model, the following exogenous variables were passed to the modelling function: the month of the year, the day of month, the day of week, and the semester (fall/winter or summer). Additionally, the following weather-related daily data was obtained[2] and passed as exogenous variables: maximum and minimum temperature, maximum wind speed, total precipitation, total rainfall, total snow, and the total snow accumulated on ground for each day. Overall, the introduction of these variables improved the model, as can be signified by the figures to the left. Other than a

[2] https://www.weatherstats.ca/

few exceptions, it can be seen that the significance of each parameter is improved by the introduction of exogenous variables; however, there still exist several parameters that we cannot be sufficiently confident about. This is the case for a few exogenous variables as well. The most significant impact on model improvement comes from the following variables: maximum temperature, and accumulated snow on ground. Furthermore, by the graphical diagnostic above, it can be seen that the randomness of the residuals has increased, and seasonality reduced in general (as indicated by the Correlogram). In spite of the improvement, there still exist significant autocorrelation of residuals at lags 1, 2, 3, and 7., which signifies that there is information in the data that has not yet been utilized completely.

From the process that was undertaken, the following are suggestions for further improvements of the model, as evidenced by the diagnostics

1. Using contextual knowledge of general building energy consumption data, incorrect measurements within the data sets (as opposed to anomalies due to surges in usage) could be removed to make the data more regular and representative of the true consumption of the CMLP. This removal of extreme values will ensure that ARIMA model parameters are trained on legitimate values that truly represent consumption, which will yield better predictions with lower error.
2. The generation of more building specific data fields such as occupancy and special events that can be used in more explicit supervised learning problems. These new data fields would contribute to more reliable exogenous variables which, when passed to the ARIMA training function, would be more effective in capturing signals in the residuals, which has not been done effectively with the generated exogenous variables thus far. Seeing the differences between the non-Exogenous and Exogenous models' diagnoses, it can be seen clearly that the introduction of relevant data as exogenous variables can significantly improve model performance. Nonetheless, out of the 11 exogenous variables introduced, only 2 significantly affected the model. Therefore, care should be given to obtaining specific variables that improve model performance.
3. In visualizations of data shown above, it is evident that although the weekly seasonality is prominent, what is also prominent is several observations that are essentially zero, or too extreme to be true. Finding a way to not only generate new exogenous variables, but also impute these anomalous data observations with more regular observations would be beneficial. Perhaps the first step should be using time series modelling to model the otherwise obscene numbers, and then use Seasonal Exogenous ARIMA modelling to use the regularized data set to make more accurate forecasts.