Using Weather Patterns to Predict a City's Energy Consumption and Renewable Energy Production

Angelene Arito – Xiao Deng – Steven Jordan – Huy Tran – Justin Winfield
DePaul University | DSC 672 - Predictive Analytics Capstone | November 20, 2019

Executive Summary

For this research paper, we are tasked by the fictitious Power City, USA to help it increase its reliability on renewable energy and reduce its reliance on fossil fuels. We were provided several time-based datasets that include: weather-related features, the electricity produced at their existing wind farms and solar arrays, and the electricity consumed by eight different sectors of the city. We developed models to predict the renewable energy production and consumption in a scenario year, and provided recommendations so that they can grow their renewable energy production – especially when taking into account consumption growth related to an expected 22% annual increase of electric vehicle (EV) adoption.

We split and transformed the provided data into ten different datasets – two for production (wind farm, solar array), and eight for consumption in the city's key sectors (food service, grocery stores, healthcare, K-12 schools, lodging, offices, residential, and stand-alone retail). Consumption data was subset by sector because the energy consumption of different sectors exhibit very different patterns (e.g. K-12 schools are closed on weekdays, holidays, and summer/winter breaks). Each dataset included sixty-five weather-related or time-related (which we engineered) features. Initial feature selection was conducted for each dataset using a Lasso wrapper method.

Using ten different algorithms/approaches, ten models were created for each of the ten datasets, resulting in one hundred models for evaluation. The algorithms used are: time series regression, artificial neural networks, multi-linear regression, decision tree regression, decision tree regression with bagging, gradient boosting, AdaBoost, XGBoost, random forest, and support vector regression. The models' parameters were tuned and evaluated using the root mean squared error (RMSE) score and explained variance.

While gradient boosting and AdaBoost did yield some of the best models, random forest yielded the best models for the majority (six of ten) of datasets, was in the top four performing models for the remaining (four of ten) datasets, and yielded the best average RMSE and explained variance scores across the ten datasets. Since we were not informed of Power City's technical competency, we have opted to provide predictions based solely on the random forest algorithm – as it is a more parsimonious solution.

We recommend that Power City scales up their wind farm and solar array energy production annually by 30% and 20%, respectively. This recommended growth will enable Power City to compensate for, or even exceed, the energy consumption's increase due to the adoption of electric vehicles (EVs) for 86% of the year. It is still possible Power City may need to use non-renewable resources to compensate for the EVs on days with extreme weather conditions, but the recommended level of growth makes up for it on days with more "predictable" weather. In fact, if Power City can store all excess energy and the weather is as predicted, this growth level will produce an extra 16.8 full days of energy, which can be saved for an

emergency, or redistributed throughout the year to even further reduce Power City's dependence on fossil fuels.

Abstract

Sustainable living and harnessing renewable resources are becoming increasingly integral in preserving the environment and reducing carbon footprint to ensure long-term habitation on the planet. With the goal of decreasing reliance on fossil fuels for energy production, renewable resources such as wind and sunlight are being leveraged to transition away from carbon-dense sources. The biggest challenge in turning to these renewable resources is their unpredictability; weather patterns are highly volatile and are often hard to forecast with great accuracy. Using data from the fictitious Power City, the goal of this research is to construct predictive models for electrical energy production from wind and solar sources, and energy consumption across eight city sectors that can effectively forecast at an hourly level. We also forecasted production and consumption values for a diverse set of days over the course of a single scenario year. Ultimately, these models can be used to assist in planning the demand-supply of renewable energy for a single city. Multiple machine learning methods (time series, neural networks, boosting, decision trees, random forest, and support vector machines) were explored and their performances were compared across ten datasets: wind and solar energy production, and consumption for eight sectors. Time series analysis considered the data chronologically, whereas all other methods considered the data in a randomized fashion. The random forest models performed the best, on average, and were used to forecast the production and consumption values for the specified dates. The random forest models were used to predict values for the full scenario year and provide recommendations on how Power City should grow their wind and solar energy sources to account for expected growth in electric vehicle adoption.

1. Introduction

Renewable energy is essential for a more sustainable existence. One of the greatest challenges when utilizing renewable sources like wind energy and solar energy is their unpredictability, as production capacity is largely rooted in weather patterns. Understanding how both production and consumption can be predicted will allow for future planning of production needs and consumption demand. The goal of this project was to study the fictitious Power City's electrical energy production and consumption – specifically: wind energy production at the wind farm, solar energy production at the solar array, and energy consumption of eight different sectors of the city – then devise a plan for the future. Specifically, Power City estimates that there will be an annual 22% increase in electric vehicle (EV) adoption, and we need to provide them insight on how they can account for the EVs' increased demand on the grid. Time series analysis was an obvious initial step to model the chronological data; while it did not return the most accurate models, it helped in understanding relationships between features and the overall timeline of production and consumption. In addition, neural networks, boosting methods, random forests, support vector machines (SVM), and multi-linear regression were tested and compared to one another in order to determine which was the best to model wind production, solar production, and Power City's consumption. The best models were determined for forecasting, and recommendations were made in regards to the city's future energy supply and demand.

2. Related Work

In reviewing work related to machine learning in predicting renewable energy consumption and production, some studies tend to be more focused on exploring one aspect of renewable energy – such

as consumption in residential areas or wind power generation – while others focus more on the overall picture of production and consumption. The research paper *Machine Learning Techniques for Supporting Renewable Energy Generation and Integration: A Survey* (Perera et al., 2014) looks at wind, solar, and hydro resources; two out of these three energy sources were also a focus in this project, with parallel findings. Regarding wind energy, wind speed is cited to be the most significant factor in prediction given it is directly related to power output, with a note that wind turbines cap out at certain speeds to maintain safe operation (Perera et al., 2014). The paper *Different Models for Forecasting Wind Power Generation: Case Study* (Alencar et al., 2017) references Betz's Law with the following statement – "only 59.3% of the energy contained in the air flow can theoretically be extracted by a wind turbine," and further handles this limitation by introducing a coefficient to account for the physical cap. Certain environmental conditions are also mentioned as noteworthy, such as humidity and temperature, as they affect the density of air, which ultimately affects power output (Perera et al., 2014).

Both papers (Perera et al., 2014 and Alencar et al., 2017) mention different levels of forecasting for both short and long term, the former explaining the importance of capitalizing on energy when it is available in the short-term to be stored for future use. Alencar et al. (2017) granularizes the intervals into four buckets: ultra-short-term (minutes to an hour ahead), short-term (one to several hours ahead), medium term (several hours to one week ahead), and long-term (one week to a year or more ahead). We did not apply these buckets; rather, we aggregated the data by day and determined a number of lags to use in time series analysis - determined individually for each dataset.

From a modeling standpoint, both papers discuss multiple machine learning techniques for modeling wind. Perera et al., (2014) brings together other studies' results to determine an optimal model for wind, including time series analysis, neural networks, and SVM, the last of which obtained the best performance with the lowest root mean squared error (RMSE) score. Regression trees and random forests are also cited as being widely used when creating models for wind power prediction. In an attempt to deal with some of the unpredictable factors involved in wind power, Alencar et al., (2017) attempts to outline a hybrid approach to modeling by leveraging an ARIMA model and two neural networks. For each of the aforementioned forecasting intervals, variables of humidity, pressure, temperature, and direction are first fed through ARIMA models, and then through the first and second neural networks to achieve the final predictions. When compared to each algorithm alone, the hybrid model performed best when compared using three performance metrics: mean average error (MAE), RMSE, and mean average percentage error (MAPE) (Alencar et al., 2017).

Shifting to solar renewable energy, the paper *A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting* (Li et al., 2015) focuses on energy production from a solar photovoltaic (PV) system located in Florida, exploring predictions at different increments of time – fifteen minutes, one hour, and one day ahead of time – employing what they cite as the two most commonly applied methods: artificial neural networks (ANN) and support vector regression (SVR). It is cited that a potential shortcoming of other research on the topic is using the information of the total power plant, rather than breaking things down on a more micro-level; the solution to this is what is coined as a "hierarchical view" to forecasting, where machine learning is used at the inverter level (different machine learning models for each inverter) and then summed to give a picture for the entire plant (Li et al., 2015). Mean bias error (MBE), MAE, RMSE, relative MBE, MPE, and relative RMSE are all used as performance metrics – the first three being cited as having limitations due to the size of their reported errors not being obvious as to whether they are large or small (Li et al., 2015). The hierarchical technique is found to outperform traditional models in one step ahead forecasts, regardless of the algorithm (ANN vs. SVR), while twenty-four hours ahead shows the hierarchical and traditional

approaches to be fairly even in performance (Li et al., 2015). The hierarchical approach also outperforms the traditional approach looking at the micro-level as well, as it is better able to account for which smaller generation units have the biggest impact on the overall forecast (Li et al., 2015).

Unlike wind and solar energy production, energy consumption involves greater complexity beyond just weather conditions, such as facility type and size, day of week, etc. In *Machine learning approaches for estimating commercial building energy consumption* (Robinson et al., 2017), energy consumption in commercial buildings is analyzed, citing that building energy consumption accounted for 40% of total energy consumption in the U.S. in 2015, according to the Energy Information Administration. Robinson et al. (2017) considers both what is referred to as a "common" feature set – those that are more commonly available features for greater generalization, and an "extended" feature set – features that are less common but are important in creating a more accurate model. Common features include information such as building square footage, building type (office, warehouse, etc.), number of days above and below a certain threshold (for heating and cooling); extended features include number of hours of operation, natural gas used, and number of employees (Robinson et al., 2017). In terms of modeling, similar to predictions for wind and solar energy production, multiple different machine learning algorithms are used and then compared against one another based on performance metrics including MAE, median absolute error, and explained variance (R^2). Of the thirteen algorithms tested within the scope of Sci-kit learn, XGBoost performed the best with both the common feature set and the extended feature set (Robinson et al., 2017). Having started with "representative" building data, the results are noted to be most applicable to an entire metropolitan area, rather than one specific building (Robinson et al., 2017).

The paper *Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach* (Zhang et al., 2018) looks at energy consumption from a residential standpoint and cites that the residential sector accounts for one-fifth of total energy demand in the U.S., according to the U.S. Energy Information Administration. It is noted that the two biggest limitations on different modeling approaches for predicting consumption are 1) heavy reliance on micro-level data, which can be difficult to obtain and/or be time-consuming and expensive to collect and 2) difficulty in generalizing the results of building energy consumption to a greater region (Zhang et al., 2018). In order to address these issues, Zhang et al. (2018) uses a few different data sources that are readily available in all major metropolitan areas in the U.S. as inputs with outputs being focused on synthesized household energy consumption for a region. These datasets are statistically matched on the variables they share and attributed with energy consumption to create a synthetic population of households that represents an entire region; features include aspects of the housing units themselves as well as socio-economic and demographic attributes around the residents (Zhang et al., 2018). Several algorithms are then used for modeling, using MAE, median absolute error, R^2, and MAPE to compare performance. Models are tested for each electricity BTU, natural gas BTU and other BTU with elastic net regression, lasso regression, ridge regression, linear regression, random forest, and boosting methods being among those considered. In terms of electricity BTU and natural gas BTU prediction, the elastic net regression outperforms all other models across most or all metrics; when it comes to other BTU predictions, random forest performs the best (Zhang et al., 2018).

The common thread between each study, be it about wind energy production, solar energy production, or energy consumption in a certain sector, is that there is no one "go-to" machine learning method that accounts for everything. It is common to see multiple machine learning methods employed and then compared using similar performance metrics - RMSE, R^2, MAE, MAPE, etc. It appears to also be a theme for ensemble algorithms to outperform simpler models, particularly in some of the more complex

scenarios around production and consumption. The work done in this project follows many similar approaches regarding multiple machine learning techniques evaluated against one another, but also takes a step to look at the relationship between production and consumption, and how that relationship drives a plan for renewable energy conversion forward.

3. Methodology

3.1 The Data

The data used for this project originated from fourteen files regarding energy production and consumption for Power City. At a high level, the raw files include:

- Weather conditions, calendar day features, and electricity production at the city's wind farm, where the turbines are located. Approximately two years' worth of data.
- Weather conditions, calendar day features, and electricity production at the city's solar array, where the solar panels are located. Approximately four and a half years' worth of data.
- Weather conditions, calendar day features, and electricity consumption among eight sectors within the Power City. One year's worth of data.
- Weather conditions and calendar day features for a single scenario year, which is to be used for making predictions using final models.

All production and consumption files include hourly data – twenty-four records per day. Weather conditions and calendar specifics between all files include:

- **Weather Condition Features:** solar elevation, cloud cover, dew point, humidity, precipitable water, temperature, visibility, wind speed, and pressure
- **Calendar Day Features:** day of the week, holiday, school day

Table A1.1 in the appendix presents a detailed view of all input files, including the number of records in each as well as the number of features and their respective types.

3.2 Exploratory Analysis & Pre-processing

In order to get the files into a more comprehensible and cohesive format, some preliminary features were added and others were normalized in order to merge the files together. Initially separated day, month, hour columns were used to generate a single DateTime column. Only two files had missing data: the scenario year weather data and the solar array weather data. Missing values were largely filled in using the median value of the two weeks preceding and following the missing data point (median was preferred over mean due to the count of zeros in the data). The only exception was filling in values for Precipitation and Pressure values in the solar array weather set, since many consecutive values were missing; for those values, the average of previous years was used to complete the data.

Full details on the preliminary preprocessing steps taken can be found in Table A1.1 in the appendix. At the end of the initial preprocessing, four consolidated files were created, all based on hourly data: wind farm production, solar array production, Power City consumption, and Power City scenario. Table 1 summarizes each concatenated dataset and its features.

Once the files were merged, each was aggregated (using mean) by day to simplify the problem and smooth some of the noise coming from the weather data. The consumption data was subset into eight

Wind Farm Production	Solar Array Production	Power City Consumption	Power City Scenario
Date	Date_Time	City	City
Location	Location	Year	Year
Hour	Year	Date_Time	Date_Time
Year	Month	Month	Month
Month	Day	Day	Day
Day	Hour	Hour	Hour
Date_Time	Electricity_KW_HR	Day_of_week	Day_of_week
Electricity_KW_HR	Solar_Elevation	Weekdays	Weekdays
Wind_Speed	Cloud_Cover_Fraction	Holiday	Holiday
Wind_Speed_Bin	Dew_Point	School_Day	School_Day
	Humidity_Fraction	Sector	Consumption_type
	Precipitation	Electricity_KW_SQFT	Electricity_KW_SQFT
	Pressure	Solar_Elevation	Cloud_Cover_Fraction
	Temperature	Cloud_Cover_Fraction	Dew_Point
	Visibility	Dew_Point	Humidity_Fraction
	Wind_Speed	Humidity_Fraction	Precipitation
	Wind_Speed_Bin	Precipitable_Water	Pressure
		Temperature	Temperature
		Visibility	Visibility
			Wind_Speed
			Wind_Speed_Bin

Table 1. Overview of features in consolidated datasets.

separate files, each representing a single sector of usage (residential, K-12 schools, etc.). Feature engineering was used to create additional calendar variables (such as 'Weekend' and 'Season'), and dummy variables were extracted from categorical variables ('Month', 'Day', 'Day of week'). All files except for the scenario file were split into 80/20 train/test sets. For all machine learning techniques other than time series analysis, where chronological data is necessary, the 80/20 split was randomized with the date time stamp removed. As a final step, Lasso feature selection was run on all 10 training datasets to determine the most optimal features going into modeling. A summary of the feature selection results is included in Table A1.2 in the appendix.

Exploratory analysis of the data was conducted on each primary input dataset: wind, solar, and consumption. In observing the wind data, a nearly perfect linear relationship can be seen between wind speed and electricity production, which is no surprise; the greater the speed of the wind, the higher the production of electricity (see Figure A1.1 in the Appendix). It is noteworthy, however, to point out the level-off effect happening with production at higher wind speeds - at a certain point no matter how high the wind speed is, production is the same. This is referenced in both Perera et al. (2014) and Alencar et al. (2017), which explain that wind turbines have a maximum level of operation in order to remain safe. On the other end of the spectrum, there are also some data points that show no electricity production even when wind speeds are above zero; it is possible this can be attributed to non-functional turbines or periods of maintenance. Given the data points were few and likely represented a more realistic representation of the wind farm in question, they were left in the dataset.

In order to observe the behavior of wind production over the course of a single day, every hour in a day was aggregated by mean across the entire dataset for wind (roughly three years of data). This aggregation showed the most active points of wind energy production (which in effect also shows when wind speeds are highest) being at night, mostly between the hours of 10:00pm to 5:00am; there is also a

sharp drop-off in the late morning, around 10:00am (see Figure A1.2 in the Appendix). Given that in pre-processing the decision was made to aggregate the data by day rather than keeping it hourly, this story of wind behavior on a given day was essential in being able to take daily predictions back down to an hourly level after final predictions were determined.

In observing the solar production data, the overall behavior of production was plotted over time to get a sense of the seasonal effects, since the solar data came with the longest span of time (over four years). As seen in Figure A1.3 in the Appendix, solar production tends to have higher production output in the beginning and end of the year with a slight drop in the middle of the year. When observed over the course of a single day, and aggregated by hour in the same manner as wind production, we see a normal curve (Figure A1.4 in the Appendix); the highest solar energy production is in the middle of the day, with no production overnight (as expected). Solar elevation (the angle the sun is in the sky) was also observed to see if there was any relationship between that and the amount of solar energy produced - analysis showed only mild correlation between the two.

Sector	Size
Residential	84,832,407 sq ft
Office	9,999,731 sq ft
K-12	2,426,480 sq ft
Stand-alone retail	971,962 sq ft
Food Service	967,558 sq ft
Health Care	965,374 sq ft
Lodging	931,129 sq ft
Grocery	539,981 sq ft

Table 2. Size of each sector in square feet from largest to smallest

Exploring consumption wound up being most beneficial when the eight sectors provided in the dataset were separated: food service, grocery, health care, K-12 schools, lodging, office, residential, and stand-alone retail. Table 2 shows the sizes in square feet of each sector in the dataset.

The consumption datasets have the fewest data points - only one year's worth - but the seasonal patterns of consumption over that time still surfaced in a digestible way. See Figures A1.5 - A1.12 in the Appendix.

- **Food Service:** One of the most volatile consumption patterns. Relatively low consumption with small standard deviations in the winter months with very large spikes in consumption over the summer months
- **Grocery:** Similar to food service in that consumption is lower in the winter months and higher in the summer months. Overall, volatility is low, and there appears to be weekly seasonality
- **Health Care:** Relatively high consumption overall. Peak consumption appears to happen in summer months but extends further into spring and autumn months as well. Weekly seasonality observed
- **K-12 Schools**: Consumption appears consistent all year apart from a dip in the summer months when school isn't in session. Weekly seasonality observed
- **Lodging:** High volatility throughout the year. Large spikes in consumption at the beginning and the middle of the year. No clear seasonality at any level
- **Office:** Very steady consumption throughout most of the year, though spikes in usage appear to be the opposite of some of the other sectors in that elevation in consumption happens in winter months rather than summer months. Weekly seasonality observed
- **Residential:** Similar shape and volatility to food service consumption. Relatively low consumption with small standard deviations in the winter months and large spikes in the summer months

- **Stand-Alone Retail:** Steady consumption throughout most of the year with some spikes in consumption in the summer months. Weekly seasonality observed

3.3 Modeling – Time Series

Prior to the construction of the time series models in Python, initial analysis was conducted in RStudio v1.1.456. The independent variables (electricity produced or consumed) from the ten datasets and their corresponding time stamps were separated into their own data frames (removing all exogenous variables). This testing was done in order to identify the starting autoregression, moving average, seasonality, or other differencing parameters to be used in the SARIMAX() function afterward in Python.

Each independent variable was then tested for normality. Because the majority had non-normal distributions, SARIMA analysis was performed on both the log-transformed and non-modified versions of the independent variables. ACF, PACF, and EACF plots were created for each data frame in order to identify potential contenders for time series parameters. Additionally, the auto.arima() function (from the forecast package) was conducted to validate the parameter selections and/or identify new potential parameter contenders using the lowest Bayesian Information Criterion (BIC) scores for evaluation.

	Log	p	d	q	Seasonality	p	d	q	BIC
Solar	Yes	0	0	2	n/a	0	0	0	39574.24
Wind	Yes	1	0	0	n/a	0	0	0	1675.59
Food Service	Yes	1	0	0	n/a	0	0	0	-1521.90
Grocery	Yes	2	0	0	7	2	0	0	-909.15
Healthcare	Yes	1	0	0	7	2	0	0	-808.87
K-12	Yes	1	0	0	7	1	0	0	-116.65
Lodging	Yes	2	0	2	n/a	0	0	0	-1207.85
Office	Yes	2	0	1	7	2	0	0	67.02
Residential	Yes	3	0	1	n/a	0	0	0	-999.98
Retail	Yes	2	0	0	7	2	0	0	-498.16

Table 3. The best-performing time series parameters in the initial time series analysis in RStudio

Using the ACF, PACF, and EACF plots, potential seasonal parameters were identified, and also validated using the auto.arima() function. The best overall parameters (Table 3), which passed the Ljung-Box test (null hypothesis is accepted and the residuals are likely due to white noise) were recorded and subsequently tuned further in Python using the SARIMAX() function.

	Log	p	d	q	Seasonality	p	d	q	RMSE	Exp. Var.
Solar	Yes	1	0	2	n/a	0	0	0	4607	0.85
Wind	Yes	1	0	0	n/a	0	0	0	19642	0.69
Food Service	No	1	1	0	7	2	0	1	261	0.00
Grocery	No	2	1	2	n/a	0	0	0	79	0.81
Healthcare	No	1	0	0	7	1	1	0	330	0.80
K-12	No	2	0	0	n/a	0	0	0	311	0.89
Lodging	No	2	1	2	4	1	0	0	62	0.10
Office	No	5	1	1	n/a	0	0	0	3423	0.77
Residential	Yes	1	0	1	n/a	0	0	0	7412	0.00
Retail	Yes	3	1	0	7	2	0	0	124	0.72

Table 4. The best-performing time series parameters after tuning SARIMAX() in Python

In Python, models were constructed using the SARIMAX() function from the statsmodels library. The SARIMAX() function takes as input ARIMA and seasonal parameters, in addition to an array of exogenous features, and can be fit to predict values on sequential time periods. The best performing parameters from the R analysis were used to construct initial models using SARIMAX, and the exogenous variables from the datasets were reintroduced (weather conditions, calendar day features, etc.). The models were then tested against a test set using the RMSE and explained variance scores for evaluation. The parameters of each model were tweaked until the highest performing model was produced – as seen in Table 4.

Despite our best efforts, we were unable to produce models with non-zero explained variance for the food service and residential datasets (and we produced a model with only a 10% explained variance for the lodging dataset). We believe this is due to the fact that the consumption datasets only included one year of data to train, and consumption is especially volatile in the summer (with GARCH-like effects).

Because the summer seasonality is unable to be captured in such a limited dataset, that volatility could not be incorporated into the predictions.

3.4 Modeling – Multilinear Regression

A multilinear regression algorithm was explored for model creation, using the LinearRegression function in Sci-Kit Learn. RMSE and R-Square were used for evaluating the model performance. Overall, these models performed decently well but much worse than other algorithms which might indicate that nonlinear models could fit these datasets better; so further tuning of these models were halted.

3.5 Modeling – Artificial Neural Networks

Initially, artificial neural networks were trained using the three consolidated datasets: wind production, solar production, and consumption. After reviewing and taking into consideration the different consumption patterns of eight sectors, ANNs were then trained on the ten datasets. Recurrent neural networks (RNN) and long-short term memory models (LSTM) on Keras were trained and tuned. An additional pre-processing step to prepare the training data frame was applied: turning time series data into supervised data frame. The output data frame has *Mx(N+1)-C* columns, where M is the number of variables from the original dataset, N is the number of time steps (or previous values) as input for the model to learn from in order to make a prediction of a future step, and C is the number of predictor variables. More specifically, in this project, RNN and LSTM use a twenty-four-time step, or a day's worth of data, to learn and ultimately make predictions. One important aspect of using RNN and LSTM in Keras is that they require the time series data to be in chronological order. As a result, wind, solar, and consumption sets were split 80/20 for train/validation sets without randomization of rows. The final step applied to training and validation data was to reshape the data frame to have dimensions in the following order: number of samples, time-step size, number of features. Training RNN and LSTM models using Keras can be time-consuming due to multiple factors, for example: higher numbers of trainable parameters, or high numbers of training epochs. In an effort to see whether a model is a good fit, the initial trained models were trained with only fifty epochs. Model building and training started out with a simple model of [one input layer, one hidden layer, and one output layer]. The training process continued with adding layers and tuning the parameters, such as the number of nodes in each layer, activation function, loss function, optimizer, and performance metrics.

The model trained with the training set is validated with validation set, by looking at the output of the losses after each training epoch. The lower value of loss (approaching zero) and the faster the loss decreases, the better and faster a neural network model learns. Ideally, the sign of a potential good model is a continuously small gap (low level of oscillation) between the training's losses and validation's losses. While LSTM model showed that the losses were decreasing, there was a gap between training's losses and validation's losses, and the decrease was slow. RNN was less successful, as the losses shown oscillating pattern. Further effort of training didn't return any ANN with higher potential. This led the project to make a decision of moving towards using ANNs run by Sci-kit learn. Sci-kit learn neural network function helped make the training process much simpler: data split with randomization of rows, no requirement to convert time series data into supervised data frame, smaller number of parameters to be tuned. Interestingly, the ANNs' performances showed improvement and potential, but did not show as much success when evaluated against other algorithms experimented in this project.

3.6 Modeling – Adaptive Boosting, Gradient Boosting & XGBoost

Boosting is a supervised machine learning ensemble method that takes a family of algorithms to convert weak learners into strong learners. The goal of a boosting algorithm is to correct the errors that were left from the previous predictor and reduce bias and variance in the model. Three boosting algorithms were used in this project.

Adaptive Boosting, or AdaBoost, is an algorithm where all observations are given equal weights and gets updated each time a model runs and calculate the errors. Higher weights are given to each data point for getting and predicted value incorrectly. Gradient boosting is a similar boosting algorithm to AdaBoost where its approach is to convert weak learners into strong learners. However, instead of creating weights for the next model, the errors calculated in the previous model will be now the target variable in the next model in order to minimize error. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the gradient boosting framework. Different from gradient boosting, XGBoost used a more regularized model formalization to control over-fitting, which gives it better performance.

For implementation, the AdaBoostRegressor and GradientBoostingRegressor from the Sci-kit learn package and XGBRegressor from xgboost package were used on the production and consumption training datasets. After running the model using all the default parameters, a grid search was applied to find the best parameters for each of the datasets. Lastly, using the best parameters found in the grid search approach, there was manual tuning of the hyperparameters to better see if the performance of the model can be improved further. The boosting models immediately resulted in high performance scores (see Results section), so were tuned further until the best RMSE and R^2 scores were achieved.

3.7 Modeling – Decision Trees & Random Forest

Working with Sci-kit learn helped reduce the complexity of building and training machine learning algorithms significantly, as well as reducing computational time. In the case of decision trees and random forests, the training and fitting of models for wind production, solar production, and consumption were performed on the 10 datasets, where consumption is subset into eight sectors. The process is similar to that with ANNs by Sci-kit learn, where DecisionTreeRegressor and RandomForestRegressor functions are called, and parameters are tuned by Grid search and manually. For decision tree, the following parameters were used: criterion (friedman mse), splitter (best), min samples split (2), min samples leaf (1). Random tree models were also utilized to find features importance in order to reduce features for training ANNs. For random forest, the following parameters were used: n estimators (410), max features (0.7 - considering 70% of all features at splitting), min samples split (3). Decision tree models have high RMSE scores across ten datasets, especially for solar production with the second highest RMSE of 24,703. On the other hand, random forests have some of the best scores for RSME across eight consumption sectors (K-12 schools, lodging, office, residential), as well as solar production.

From building and training random forest models, there was a consistent pattern of good performances based on RMSE and R-squared scores. As a result, random forest was chosen as the main algorithm for all ten datasets, including wind production, solar production, and the eight consumption sectors. As one potential random forest model for each of these datasets was identified, it was further tuned and tested using test sets. Models for wind and solar production were more straightforward, as they are two separate models. However, this is not the case for Consumption sectors, as the project's objective is to build a model for Power City's consumption as a whole. The best random forest model identified for each consumption sector had different parameters' values compared to the other sectors' best random forest models. In order to provide a more parsimonious solution, all eight sectors' random forest models were compared and manually tested to find the best combination of parameters for a final model that can maintain or improve predictive performance for each sector's consumption. In order to evaluate the final random forests model, the predicted value for each calendar day (average of electricity consumption of the day) was calculated by summing up the predicted values of all eight sectors'

consumption on that day. These predicted values were generated by calling the predict() method on the eight consumption sectors' test datasets. The last step was to use this model and make predictions using scenario dataset. These predictions are used as a tool to demonstrate the functionality of random forest model for Power City's daily average energy consumption.

3.8 Modeling – Support Vector Machines

Perera et al. (2014) built and trained a least-square SVM using multiple kernel functions, and their results show SVM outperformed other solar production models, with up to 27% higher accuracy. In this project, SVMs were far less successful; performance-wise they ranked at the bottom compared to the other algorithms tested. Similar to many other models in this project, SVM models were built and trained using Sci-kit functionality. SVM tends to be used for classification problems, thus for this project's regression problem, SVR by Sci-kit (SVM for regression model) was applied to training sets for wind production, solar production, and the eight consumption sectors. The two parameters used and tuned for SVR models were: kernel for algorithm (*linear*), gamma - coefficient of the kernel (*auto*). In terms of performance, SVR models have markedly higher RMSE scores, as well as lower explained variance scores. As a result, identifying the cause of, and potentially improving, SVR's weak performance is a related future work this project can embark on.

4. Results

4.1 Model Evaluation

Considering multiple models were used to create a predictive model for energy production and energy consumption, common performance metrics were established. As a regression problem, the most common metric to use to evaluate the performance is the RMSE value. This metric indicates if the model can be good at predicting the observed data. Also, another performance metric to consider is the R^2 value. R^2 is a statistical measure of how close the data is to the regression line. This also tells how the target variable can be explained by the predictors from the model. Table 5 shows the RMSE and R^2 evaluation of the highest performing model constructed by every algorithm on all ten datasets. The highest performing models overall for each data set are highlighted in yellow.

	Production				Consumption															
	Wind		Solar		Food Service		Grocery		Healthcare		K12		Lodging		Office		Residential		Retail	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Time Series	4607	0.85	19642	0.69	261	0.00	79	0.81	330	0.80	311	0.89	62	0.10	3423	0.77	7412	0.00	124	0.72
Neural Network	2064	0.96	18405	0.72	231	0.76	60	0.97	155	0.93	273	0.93	61	0.83	2257	0.87	2180	0.95	82	0.92
Multi-linear Regression	2802	0.92	18033	0.63	244	0.70	76	0.95	158	0.92	307	0.92	66	0.80	2403	0.83	4620	0.77	94	0.87
Gradient Boosting	1701	0.97	17082	0.76	66	0.98	**42**	**0.99**	**99**	**0.97**	168	0.97	41	0.93	1868	0.91	1278	0.98	41	0.98
AdaBoosting	**1550**	**0.98**	18511	0.72	81	0.97	75	0.95	159	0.92	159	0.98	41	0.93	2169	0.88	1644	0.97	**36**	**0.98**
XGBoost	2047	0.96	17264	0.66	63	0.98	50	0.98	122	0.95	207	0.96	45	0.91	2087	0.88	1385	0.98	51	0.97
Decision Tree	2299	0.95	24703	0.50	85	0.97	86	0.94	168	0.92	128	0.99	43	0.92	1942	0.90	1631	0.97	92	0.90
DT + Bagging	1596	0.98	17499	0.75	73	0.98	71	0.96	173	0.91	285	0.93	43	0.92	1560	0.94	1705	0.97	48	0.97
Random Forest	1558	0.98	**16812**	**0.77**	**66**	**0.98**	59	0.97	160	0.92	**158**	**0.98**	**37**	**0.94**	**1426**	**0.95**	**1170**	**0.99**	44	0.98
SVM	9766	0.29	31136	0.20	273	0.67	158	0.81	343	0.69	1126	0.12	66	0.80	6519	0.08	7015	0.51	242	0.37

Table 5. A model evaluation matrix showing the RMSE and R-squared value for energy production between wind and solar and energy consumption for each of the eight sectors. The best performing models for each dataset are highlighted in yellow.

4.2 Predictive Model for Energy Production

In order to find the best model, we need to find the model with the lowest RMSE with a corresponding high R^2 value. When comparing the different values for the best predictive model for solar energy production we found a random forest model was the best model. The model had the lowest RMSE value (16,812) and the highest R^2 value (0.77) compared to the other models used. The parameters used to create the model required a minimum of six samples before splitting the tree, two-hundred trees, and used the Mean Absolute Error as a criterion to determine the quality of the split. When looking at the model's feature importance, we found cloud coverage (55.2%), humidity (9.4%), solar elevation (6.8%), dew point (4.7%), and temperature (4.4%) as the top five features from model.

A similar approach was performed to determine the best predictive model for wind energy production. Both AdaBoost and random forest were found to be the best models where they had the same r-squared value (0.98) and a nearly identical RMSE value (1,550 for AdaBoost and 1,558 for random forest). After much consideration which model to choose, we decided to use the random forest model since the model for the solar energy production was also a random forest model. The parameters used to create the model required seven samples before splitting the tree and two-hundred trees. The feature importance for this model was only Wind Speed (99%) as the other predictors were dummy variables from feature engineering mentioned previously.

4.3 Energy Production Predictions

Once the best model for wind and solar energy production was finalized, the next step is to create predictions for the energy production from each energy source for a given day specified by Power City. Table 4 shows the energy production for the 6 days per hour and per day. As an example, for October 13th, the model predicted that both solar and wind will generate about 41,291 kW/hr and 3,441 kw/hr of electricity respectively or 990,976 kW and 210,666 kW of electricity per day.

	Solar Energy		Wind Energy	
Date	Electricity Production (kW/hr)	Daily Electricity Production (kW)	Electricity Production (kW/hr)	Daily Electricity Production (kW)
March 15th	59,364	1,424,748	1,173	28,169
June 26th	47,331	1,135,937	1,176	28,224
July 3rd	78,336	1,880,053	469	11,261
October 13th	41,291	990,976	3,441	82,598
November 19th	35,449	850,788	8,777	210,666
December 25th	15,487	371,694	8,534	204,825

Table 6. Predicted electricity production for solar and wind energies for a specific day

4.4 Predictive Model for Energy Consumption

After the initial models across ten datasets were built and evaluated, Table 5 summarizes the performances of ten algorithms for each dataset with RMSE and Explained variance scores. Based on this table, the best RMSE score for each dataset was used as an indicator for the suitable algorithm for the respective dataset. In regards to the eight consumption sectors, it appears that for five out of eight sectors, random forests algorithm outperformed other algorithms. Subsequently, one final random forest built, tuned, and fitted with all eight sectors' training datasets. With the exception of office and restaurant sectors, RMSE scores of this random forest model for all other six sectors (Table 7) remained reasonably stable compared to their respective initial best scores (Table 5), indicating that the random forest model is a good fit. The parameters of the final random forest for energy consumption are as follows: max features (0.7), min samples leaf (1), min samples split (3), n estimators (410). The final

RMSE for consumption as a whole, calculated by generating predicted values from eight consumption sector's test datasets as well as using actual consumption values from these test datasets, was 2,963. Additionally, the explained variance score was 0.96, also signaling the effectiveness of the model for Power City's energy consumption. The top five most important features for each of the sectors are shown in Table 9.

Food Service		Grocery		Healthcare		K12		Lodging		Office		Residential		Retail	
RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
68.25	0.98	58.24	0.97	161.57	0.92	210.92	0.96	39.25	0.93	1817.77	0.92	1323.45	0.98	51.12	0.97

Table 7. RMSE scores of final random forest model for eight consumption sectors

Food		Grocery		Healthcare		K-12	
Temperature	74.5%	Dew Point	51.7%	Temperature	51.6%	Weekend	61.8%
Dew Point	19.3%	Temperature	29.7%	Weekend	16.9%	Saturday	6.9%
Solar Elevation	2.6%	Sunday	8.5%	Dew Point	15.5%	Temperature	5.6%
Summer	0.9%	Weekend	3.4%	School Day	3.4%	Solar Elevation	5.3%
Humidity	0.7%	Solar Elevation	1.9%	Solar Elevation	3.3%	School Day	4.9%

Lodging		Office		Residential		Retail	
Temperature	69.1%	Weekend	45.6%	Temperature	71.9%	Sunday	48.5%
Dew Point	23.6%	Temperature	18.3%	Dew Point	20.9%	Temperature	26.1%
Solar Elevation	2.4%	Sunday	12.6%	Solar Elevation	2.4%	Dew Point	7.3%
Summer	2.0%	Dew Point	8.5%	Summer	1.6%	Weekend	7.2%
Cloud Cover	0.7%	Holiday	4.1%	Humidity	0.7%	Holiday	5.1%

Table 9. Top 5 most important features of each consumption sector

4.5 Energy Consumption Predictions

As the final predictive model for energy consumption was finalized, consumption predictions for various calendar days were made by applying random forest model to the scenario dataset. The predictions can be reviewed in Table 8 below. The middle column consists of predicted average energy consumption (kW/hr) of six different days in the scenario year, whereas the last column shows the total energy consumption (kW) of these six respective days, which were calculated by multiplying the values in the middle column by 24 hours. The high consumption values appear to be in June (summer), and the last quarter of the year October - December (fall/winter).

Date	Electricity Consumption (kW/hr)	Electricity Consumption (kW)
March 15th	83,803	2,011,266
June 26th	129,677	3,112,248
July 3rd	97,559	2,341,416
October 13th	102,660	2,463,840
November 19th	103,033	2,472,792
December 25th	103,164	2,475,936

Table 8. Predicted energy consumption for various calendar days

5. Conclusion

The highest performing models were determined using RMSE and explained variance scores for evaluation, and are summarized in Table 5. While gradient boosting and AdaBoost did yield some of the best models, random forest yielded the best model for the majority (six of ten) of datasets, was in the top four performing models for the remaining (four of ten) datasets, and yielded the best average RMSE and explained variance across the ten datasets. For these reasons, and because we are uncertain of the client's technical capabilities, we have opted to provide Power City predictions based solely on the

random forest algorithm – as it is a more parsimonious solution. For additional parsimony, we developed a single random forest set of parameters that were applied to the eight consumption datasets for a single effective solution.

As the client has recognized – a 22% increase in electric vehicle (EV) adoption by Power City residents will result in an increase of daily energy consumption by 6-7%. The energy for this consumption increase will have to be produced somewhere; if there is not an increase in renewable energy production, it will be produced by coal, natural gas, or other non-renewable sources – thus eliminating the positive environmental impact of the EVs.

There is not a straightforward solution because wind energy production peaks in the winter months, and is lowest during the summer. Inversely, solar energy production peaks in the summer months, and is lowest in the winter. Furthermore, strange and unexpected weather phenomena can occur at any time (e.g. a summer could be extra cloudy). If Power City were to invest in a sole renewable energy production source (either wind or solar), it risks not producing enough energy to compensate for the EVs during one of the seasons, or when the weather is especially non-conducive to energy production. Furthermore, there is a risk of renewable energy overproduction (particularly solar energy production during the summer), in which energy would be unused and lost, if Power City does not have adequate energy storage systems.

For these reasons we have the following recommendations to Power City:

1. Scale up their wind farm energy production annually by 30%
2. Scale up their solar park energy production annually by 20%
3. Increase their energy storage technology

This recommended growth in solar and wind energy production will enable Power City to compensate for or exceed the energy consumption increase due to electric vehicles on 86% days of the year. It is still possible Power City may need to use non-renewable resources on days with extreme weather conditions, but the recommended level of growth more than makes up for it on days of more "predictable" weather. In fact, if Power City can store all excess energy and the weather is as predicted, this growth level will produce an extra 16.8 full days of energy, which can be saved for an emergency, or redistributed throughout the year to even further reduce Power City's dependence on fossil fuels. We do not know their current capabilities of energy storage, but if invested in, it can enable the city to store any excess for the literal rainy days (when solar energy production is minimal). Because the city produces 50% – 1600% more solar energy than wind energy (depending on the day), we recommend that they scale up their wind energy production at a greater rate than their solar energy production to better safeguard their production on unexpected cloudy days. If Power City cannot grow at that scale, but has enough energy storage to account for any excess production, we recommend at minimum that they grow their wind and solar energy by 12% annually. That rate should cover the net increased energy consumption on an annual basis, and assumes perfectly predicted weather patterns (has only about 10-14 hours of excess energy at the end of the year).

In addition to this growth, we recommend that the city invest in more efficient energy technology like LED light bulbs and ENERGY Star-Certified appliances, in addition to launching a city-wide initiative to reduce consumption. These last two suggestions can greatly reduce energy consumption overall, reducing the risk of under-production of energy as Power City becomes increasingly dependent on wind and solar energy.

6. Future Work

We recommend obtaining more data and updating the models annually because of the limited Consumption dataset. With more than one year of Consumption data, we can more easily account for volatility and annual seasonality effects. Additionally, one could update the model to account for the effects of climate change – and use a time series model to predict the weather conditions of future years (which this research does not do).

We also recommend exploring more models using SVMs, since it was indicated in some of the previous work (Perera et al., 2014) that it could produce higher performing models than what we were able to achieve.

7. References

[1] Kasun S. Perera, Zeyar Aung, Wei Lee Woon. *Machine Learning Techniques for Supporting Renewable Energy Generation and Integration: A Survey.* 2014.

[2] David Barbosa de Alencar, Carolina de Mattos Affonso, Roberto Célio Limão de Oliveira, Jorge Laureano Moya Rodríguez, Jandecy Cabral Leite, José Carlos Reston Filho. *Different Models for Forecasting Wind Power Generation: Case Study.* 2017.

[3] Caleb Robinson, Bistra Dilkinaa, Jeffrey Hubbs, Wenwen Zhang, Subhrajit Guhathakurta, Marilyn A. Brown, Ram M. Pendyala. *Machine learning approaches for estimating commercial building energy consumption*. 2017.

[4] Wenwen Zhang, Caleb Robinson, Subhrajit Guhathakurta, Venu M. Garikapati, Bistra Dilkina, Marilyn A. Brown, Ram M. Pendyala. *Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach*. 2018.

[5] Zhaoxuan Li, SM Mahbobur Rahman, Rolando Vega, Bing Dong. *A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting*. 2015.

8. Appendix

Table A1.1: Detailed Dataset Summary

File Name	Attributes	Records	Missing Data	Pre-Processing Applied
calendar_days_consumption	7 (3 interval, 3 nominal, 1 Boolean)	365	None - apart from the blank observations in HolidayName, which is deliberate	● HolidayName was converted to a Boolean (1 if it's a holiday) ● The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder

				since no year is provided (year will not be used in the analysis
calendar_days_scenario	7 (3 interval, 3 nominal, 1 Boolean)	366	None - apart from the blank observations in HolidayName, which is deliberate	• HolidayName was converted to a Boolean (1 if it's a holiday) • The date variables were concatenated into a single column with date format, using the year 1992 as a placeholder since no year is provided (year will not be used in the analysis
car_charging	6 (3 interval, 2 nominal, 1 ratio)	8760	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis
powercity_ consumption	6 (3 interval, 2 nominal, 1 ratio)	70080	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis
powercity_ population	8 (2 nominal, 6 ratio)	44	None	• None
powercity_ solarangle_ consumption	6 (3 interval, 2 nominal, 1 ratio)	8783	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date

				format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis
powercity_ weather_ consumption	11 (5 interval, 2 nominal, 4 ratio)	8760	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis
powercity_ weather_ scenario	13 (5 interval, 2 nominal, 6 ratio)	8784	1079 total values: Cloud_Cover_Fraction: 2 Dew_Point: 26 Humidity_Fraction: 26 Precipitation: 369 Pressure: 625 Temperature: 25 Visibility: 2 Wind_Speed: 4	• The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis • Missing variables were replaced with the median value, of the 14 existing values preceding, and the 14 values succeeding (to account for seasonal changes). Data points at the end only used the median of the 14 values preceding. • Because the wind_speed equipment cannot measure winds under 1.5 and were set at 0, a new wind speed column created with the wind speed values binned for every 1.5 meters per second
Sector_Use_ Matrix	6 (1 nominal, 5 ratio)	8	None	• None

solararray_ production	3 (1 interval, 1 nominal, 1 ratio)	18704	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis
solararray_ solarangle	6 (5 interval, 1 nominal)	50800	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format
solararray_ weather	13 (6 interval, 1 nominal, 6 ratio)	41322	29139 total values: Cloud_Cover_Fraction: 191 Dew_Point: 270 Humidity_Fraction: 270 Precipitation: 12590 Pressure: 15342 Temperature: 150 Visibility: 142 Wind_Speed: 184	• The date variables were concatenated into a single column with date format, using the year 1991 as a placeholder since no year is provided (year will not be used in the analysis • Missing variables were replaced with the median value, of the 14 existing values preceding, and the 14 values succeeding (to account for seasonal changes). Data points at the end only used the median of the 14 values preceding. However, because the majority of the precipitation and pressure values in 2013 and 2014 were missing, those were replaced with the average values of those corresponding calendar days in 2010-2012 • Because the wind_speed equipment

				cannot measure winds under 1.5 and were set at 0, a new wind speed column created with the wind speed values binned for every 1.5 meters per second
windfarm_ production	3 (2 interval, 1 ratio)	15835	None	• The Hour variable was converted from (1-24) to (0-23) to align with other spreadsheets • The date variables were concatenated into a single column with date format
windfarm_ windspeed	4 (2 interval, 1 nominal, 1 ratio)	15390	None	• The date variables were concatenated into a single column with date format • The wind speed was binned as per the other files

Table A1.2: Final Features

Wind	Solar	Consumption - Stand-alone Retail	Consumption - Residential	Consumption - Office
Electricity_KW_HR_AVG	Electricity_KW_HR_AVG	Electricity_KW_AVG	Electricity_KW_AVG	Electricity_KW_AVG
Wind_Speed_AVG	Solar_Elevation_AVG	Weekend	Weekend	Weekend
Month_1	Cloud_Cover_Fraction_AVG	Holiday	Holiday	Holiday
Month_2	Dew_Point_AVG	School_Day	School_Day	School_Day
Month_3	Humidity_Fraction_AVG	Solar_Elevation_AVG	Solar_Elevation_AVG	Solar_Elevation_AVG
Month_4	Precipitation_AVG	Cloud_Cover_Fraction_AVG	Cloud_Cover_Fraction_AVG	Dew_Point_AVG
Month_5	Pressure_AVG	Dew_Point_AVG	Dew_Point_AVG	Humidity_Fraction_AVG
Month_6	Temperature_AVG	Temperature_AVG	Humidity_Fraction_AVG	Temperature_AVG
Month_7	Visibility_AVG	Visibility_AVG	Temperature_AVG	Visibility_AVG
Month_8	Wind_Speed_AVG	Month_1	Visibility_AVG	Month_1
Month_9	Month_1	Month_2	Month_1	Month_2
Month_10	Month_2	Month_4	Month_2	Month_3
Month_11	Month_3	Month_5	Month_3	Month_4
Month_12	Month_4	Month_7	Month_4	Month_5
Day_1	Month_5	Month_8	Month_5	Month_6
Day_2	Month_6	Month_9	Month_6	Month_7
Day_3	Month_7	Month_10	Month_7	Month_8
Day_4	Month_8	Month_12	Month_8	Month_9
Day_5	Month_9	Day_3	Month_9	Month_10
Day_6	Month_10	Day_4	Month_10	Month_11
Day_7	Month_11	Day_6	Month_11	Month_12
Day_8	Month_12	Day_7	Month_12	Day_1
Day_9	Day_1	Day_10	Day_1	Day_2
Day_10	Day_2	Day_11	Day_2	Day_3
Day_11	Day_3	Day_14	Day_3	Day_4
Day_13	Day_4	Day_15	Day_4	Day_5
Day_14	Day_5	Day_16	Day_5	Day_6
Day_15	Day_6	Day_17	Day_6	Day_7
Day_17	Day_7	Day_18	Day_7	Day_8
Day_18	Day_8	Day_23	Day_8	Day_9
Day_19	Day_9	Day_24	Day_9	Day_10
Day_20	Day_11	Day_25	Day_10	Day_12
Day_21	Day_12	Day_26	Day_11	Day_13
Day_22	Day_13	Day_27	Day_12	Day_14
Day_23	Day_14	Day_29	Day_14	Day_15
Day_24	Day_15	Day_31	Day_15	Day_16
Day_25	Day_16	Day_of_week_1	Day_16	Day_17
Day_26	Day_17	Day_of_week_3	Day_17	Day_18
Day_28	Day_18	Day_of_week_5	Day_18	Day_19
Day_29	Day_19	Day_of_week_6	Day_19	Day_20
Day_30	Day_20	Season_Winter	Day_20	Day_21
Day_31	Day_21		Day_21	Day_22
Day_of_week_1	Day_22		Day_22	Day_23
Day_of_week_2	Day_23		Day_23	Day_24
Day_of_week_3	Day_24		Day_24	Day_25
Day_of_week_5	Day_25		Day_25	Day_26
Day_of_week_6	Day_27		Day_26	Day_27
Day_of_week_7	Day_28		Day_27	Day_29
Season_Spring	Day_29		Day_28	Day_30
Season_Summer	Day_30		Day_29	Day_31
Season_Winter	Day_of_week_1		Day_30	Day_of_week_1
	Day_of_week_2		Day_31	Day_of_week_2
	Day_of_week_3		Day_of_week_2	Day_of_week_3
	Day_of_week_4		Day_of_week_3	Day_of_week_4
	Day_of_week_6		Day_of_week_4	Day_of_week_5
	Day_of_week_7		Day_of_week_5	Season_Autumn
	Season_Autumn		Day_of_week_6	Season_Spring
	Season_Spring		Day_of_week_7	Season_Summer
	Season_Summer		Season_Autumn	Season_Winter
	Season_Winter		Season_Spring	
			Season_Summer	
			Season_Winter	

Consumption - Lodging	Consumption - K-12 Schools	Consumption - Health Care	Consumption - Grocery	Consumption - Food Service
Electricity_KW_AVG	Electricity_KW_AVG	Electricity_KW_AVG	Electricity_KW_AVG	Electricity_KW_AVG
Weekend	Weekend	Weekend	Weekend	Weekend
School_Day	Holiday	Holiday	Holiday	School_Day
Solar_Elevation_AVG	School_Day	School_Day	School_Day	Solar_Elevation_AVG
Cloud_Cover_Fraction_AVG	Solar_Elevation_AVG	Solar_Elevation_AVG	Solar_Elevation_AVG	Cloud_Cover_Fraction_AVG
Dew _Point_AVG	Cloud_Cover_Fraction_AVG	Cloud_Cover_Fraction_AVG	Cloud_Cover_Fraction_AVG	Dew _Point_AVG
Temperature_AVG	Dew _Point_AVG	Dew _Point_AVG	Dew _Point_AVG	Humidity_Fraction_AVG
Visibility_AVG	Humidity_Fraction_AVG	Temperature_AVG	Temperature_AVG	Temperature_AVG
Month_1	Temperature_AVG	Visibility_AVG	Visibility_AVG	Visibility_AVG
Month_3	Visibility_AVG	Month_1	Month_1	Month_1
Month_4	Month_1	Month_2	Month_3	Month_2
Month_5	Month_2	Month_3	Month_4	Month_3
Month_7	Month_5	Month_4	Month_5	Month_4
Month_8	Month_6	Month_5	Month_6	Month_5
Month_10	Month_7	Month_7	Month_7	Month_7
Month_11	Month_8	Month_9	Month_8	Month_8
Day_4	Month_10	Month_10	Month_10	Month_9
Day_6	Month_11	Month_12	Month_11	Month_10
Day_7	Month_12	Day_1	Day_3	Month_12
Day_9	Day_1	Day_3	Day_6	Day_2
Day_10	Day_2	Day_4	Day_7	Day_3
Day_14	Day_3	Day_6	Day_9	Day_4
Day_15	Day_4	Day_7	Day_10	Day_6
Day_16	Day_5	Day_9	Day_11	Day_7
Day_22	Day_6	Day_10	Day_14	Day_9
Day_23	Day_7	Day_11	Day_15	Day_10
Day_24	Day_10	Day_13	Day_16	Day_12
Day_29	Day_12	Day_15	Day_17	Day_14
Day_30	Day_13	Day_16	Day_20	Day_15
Day_31	Day_14	Day_17	Day_23	Day_16
Day_of_week_5	Day_15	Day_18	Day_24	Day_17
Season_Autumn	Day_16	Day_20	Day_29	Day_18
Season_Spring	Day_17	Day_22	Day_30	Day_19
Season_Summer	Day_18	Day_23	Day_of_week_1	Day_21
	Day_19	Day_24	Day_of_week_2	Day_23
	Day_20	Day_25	Day_of_week_3	Day_24
	Day_21	Day_26	Day_of_week_4	Day_25
	Day_22	Day_27	Day_of_week_5	Day_26
	Day_23	Day_29	Day_of_week_6	Day_27
	Day_24	Day_31	Season_Autumn	Day_28
	Day_25	Day_of_week_1	Season_Spring	Day_29
	Day_29	Day_of_week_2	Season_Summer	Day_30
	Day_31	Day_of_week_3	Season_Winter	Day_31
	Day_of_week_2	Day_of_week_4		Day_of_week_2
	Day_of_week_3	Day_of_week_5		Day_of_week_3
	Day_of_week_4	Day_of_week_6		Day_of_week_4
	Day_of_week_5	Season_Spring		Day_of_week_5
	Day_of_week_6	Season_Winter		Day_of_week_6
	Day_of_week_7			Day_of_week_7
	Season_Autumn			Season_Summer
	Season_Spring			Season_Winter
	Season_Summer			
	Season_Winter			

Figure A1.1: Electricity Production by Wind Speed

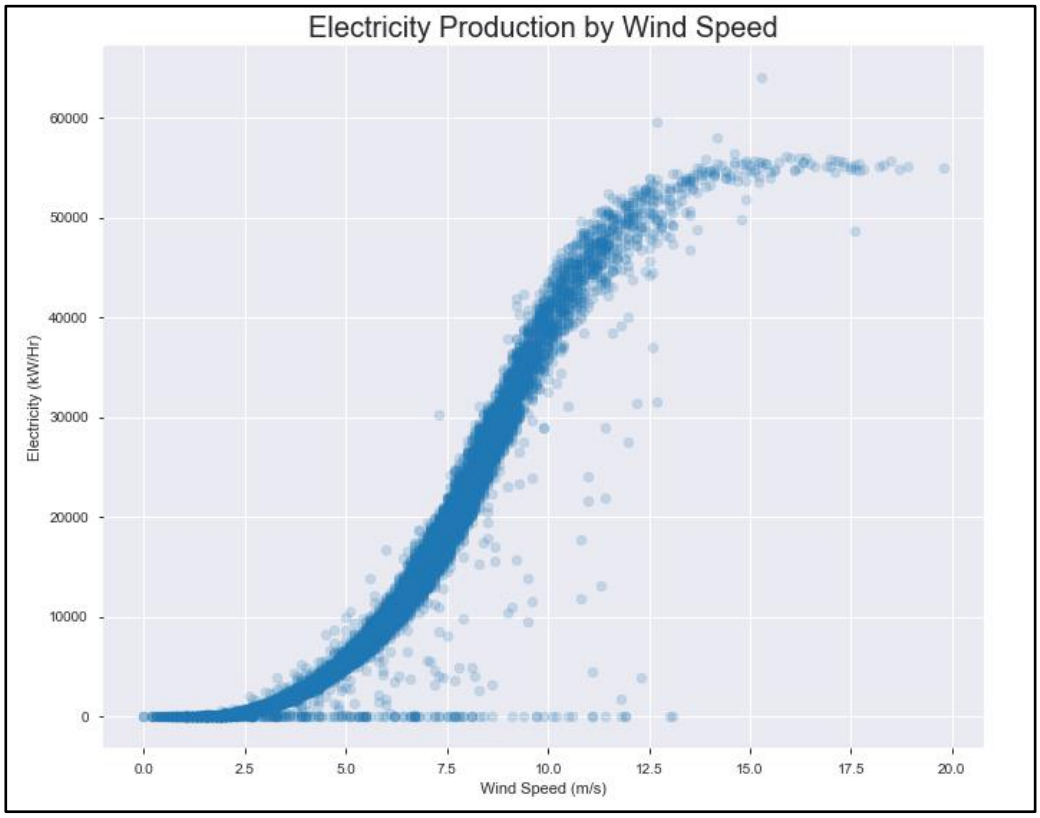


Figure A1.2: Wind Energy Production Over 1 Day

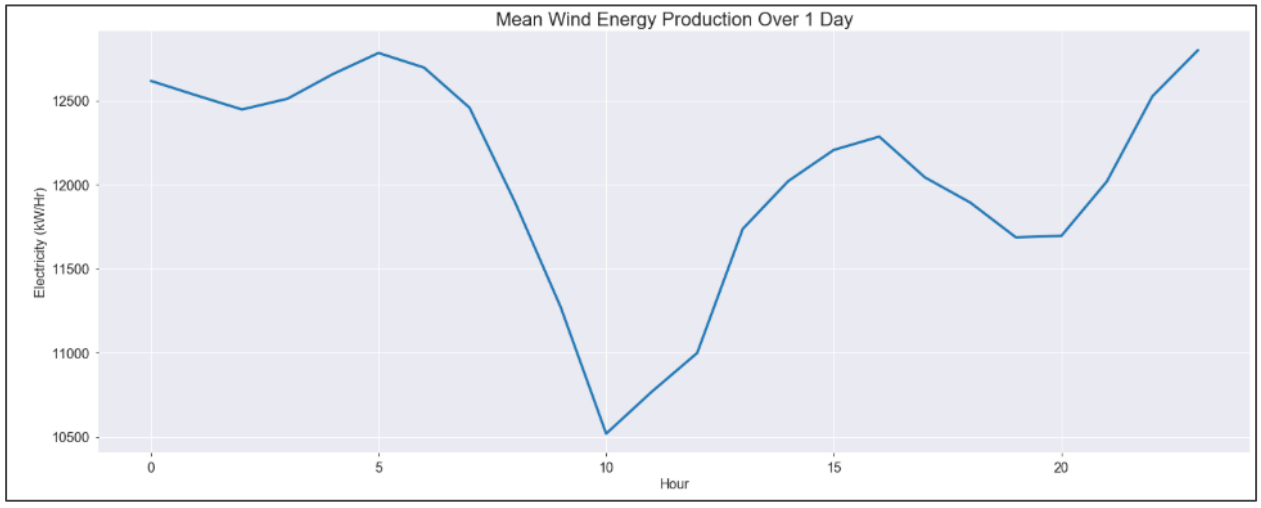


Figure A1.3: Electricity Production by Solar Energy

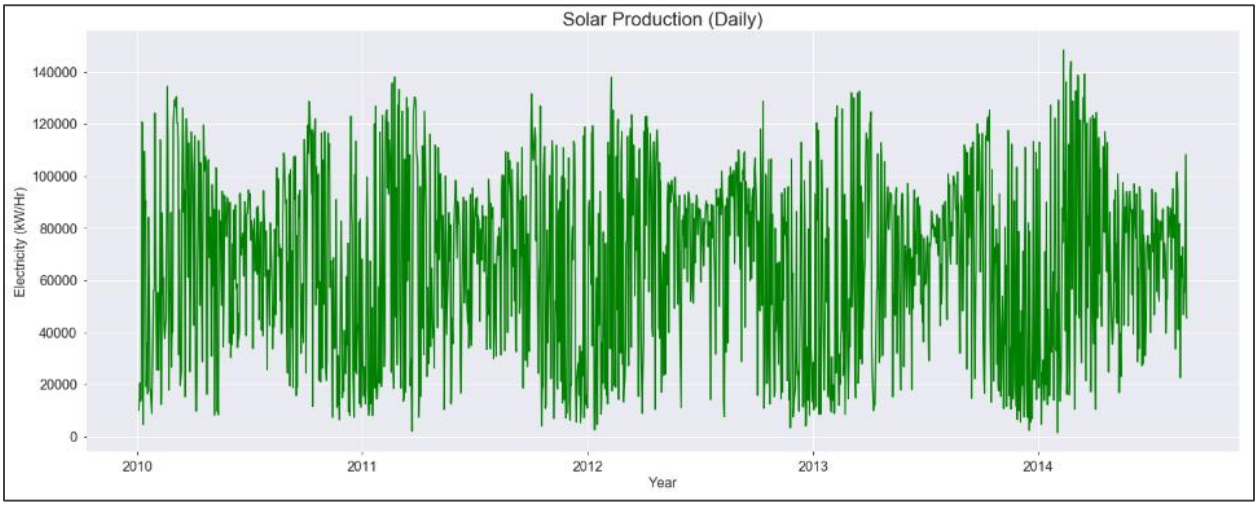
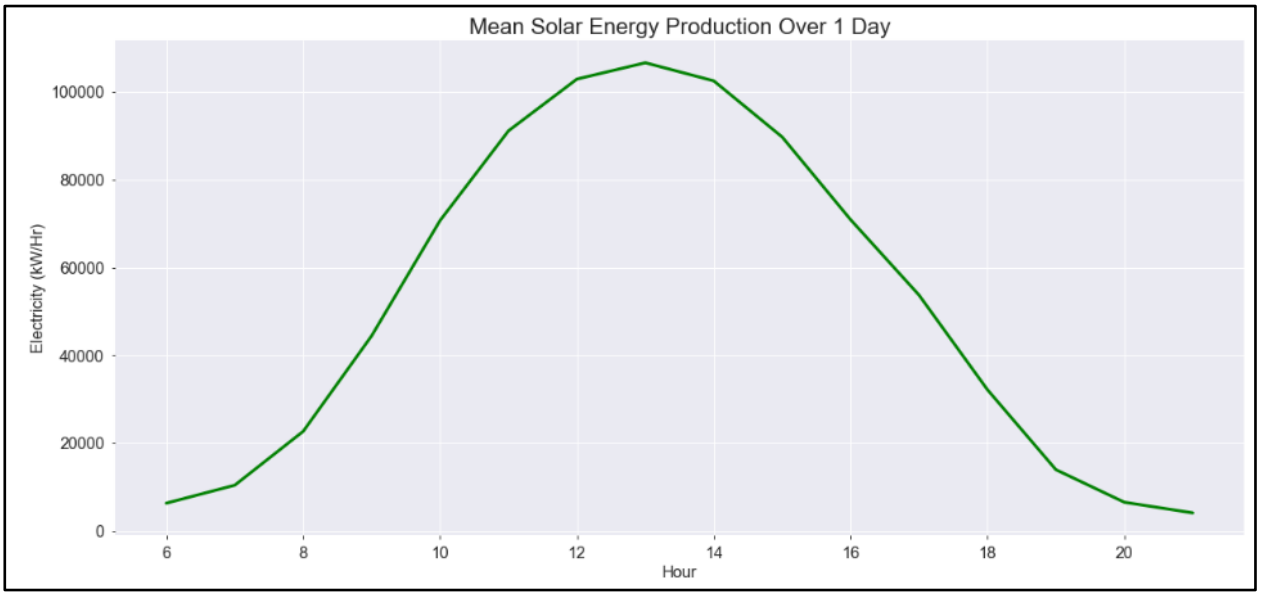


Figure A1.4: Solar Energy Production Over 1 Day



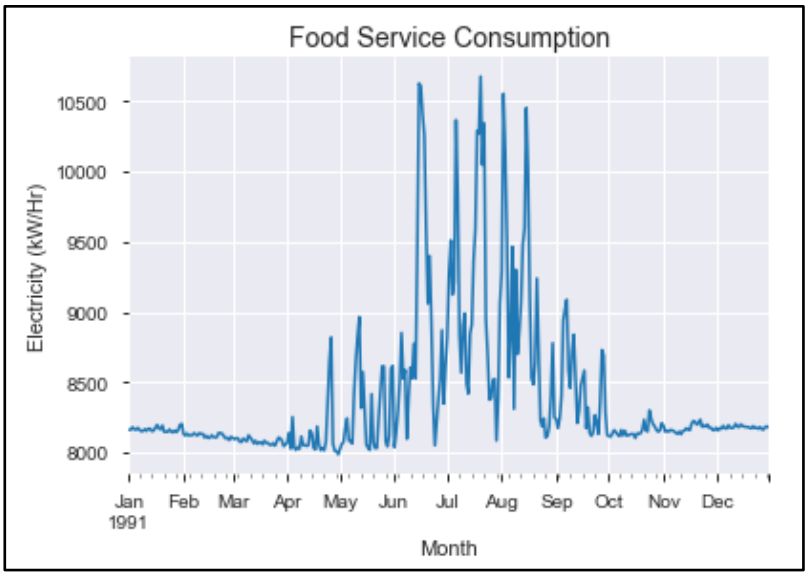


Figure A1.5: Food Service Energy Consumption

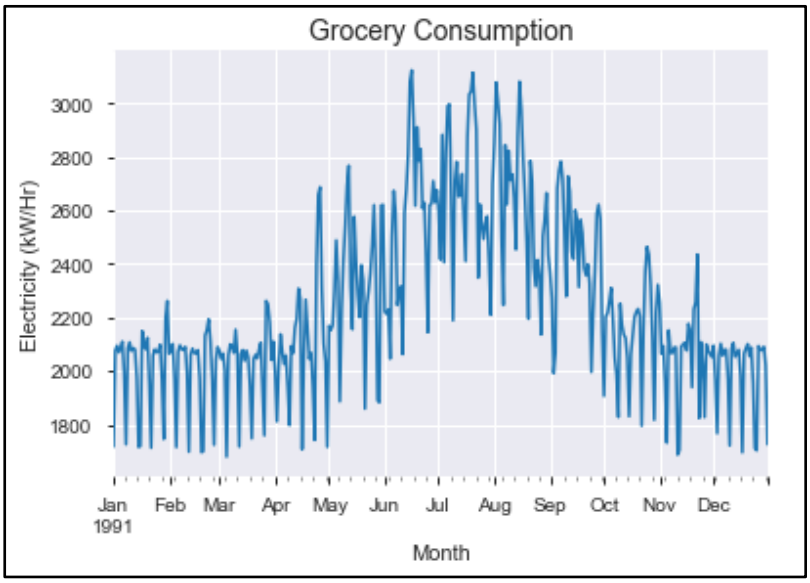


Figure A1.6: Grocery Energy Consumption

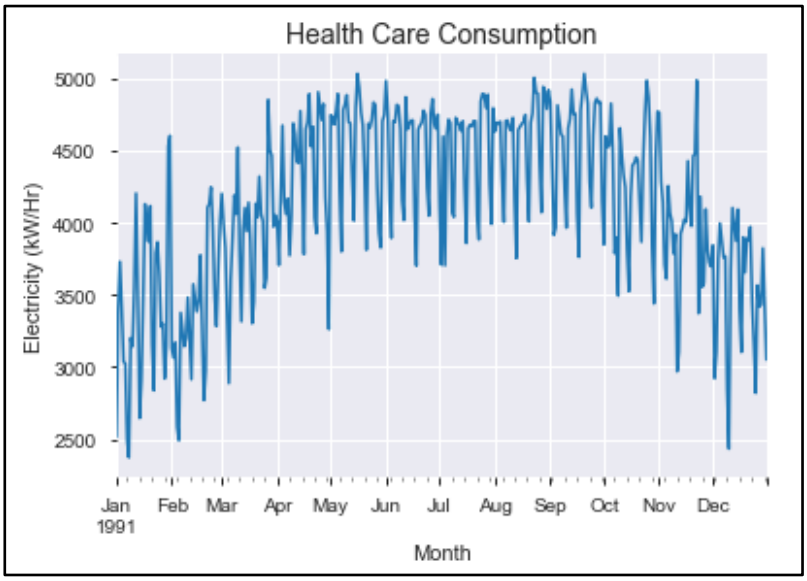


Figure A1.7: Health Care Energy Consumption

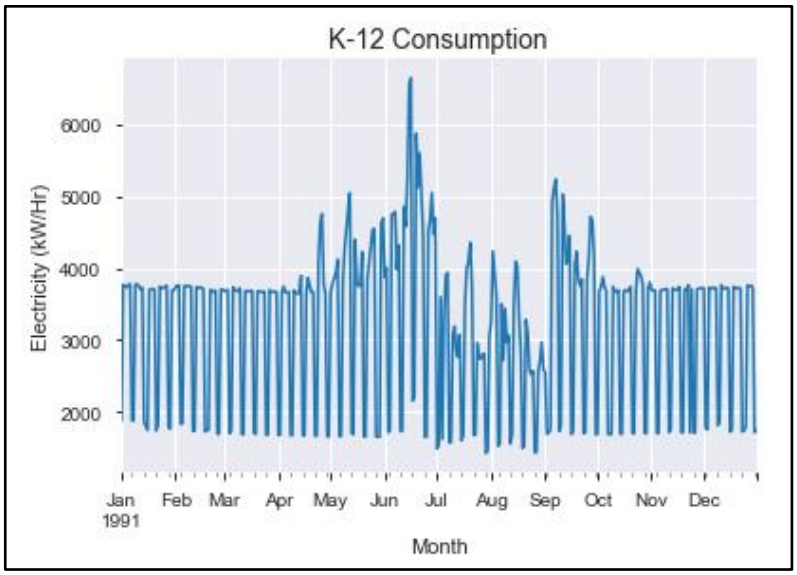


Figure A1.8: Education K-12 Energy Consumption

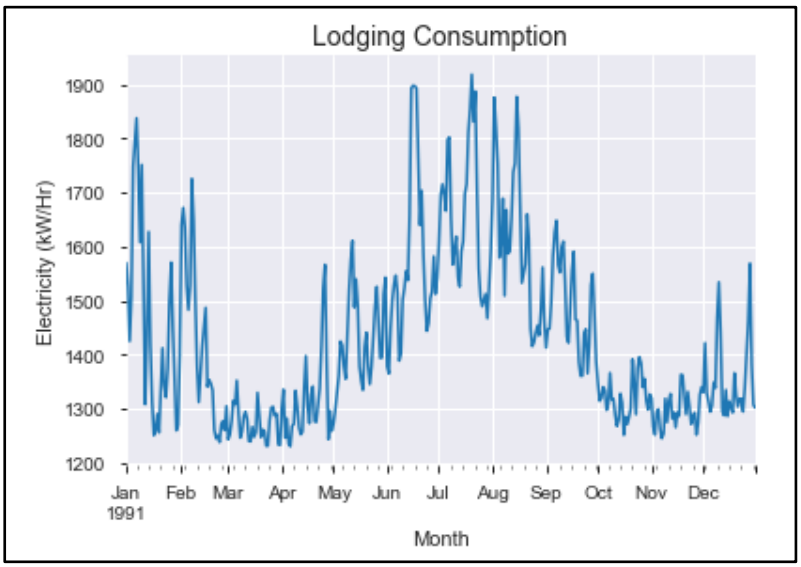


Figure A1.9: Lodging Energy Consumption

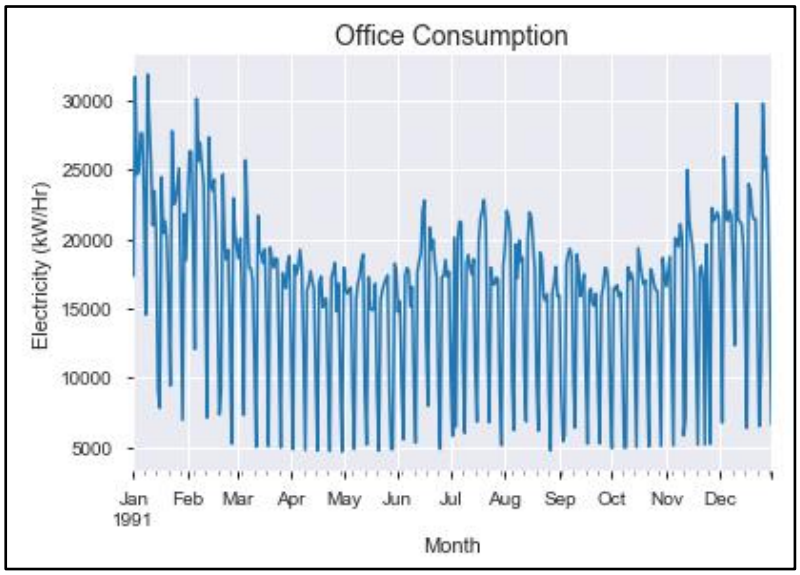


Figure A1.10: Office Energy Consumption

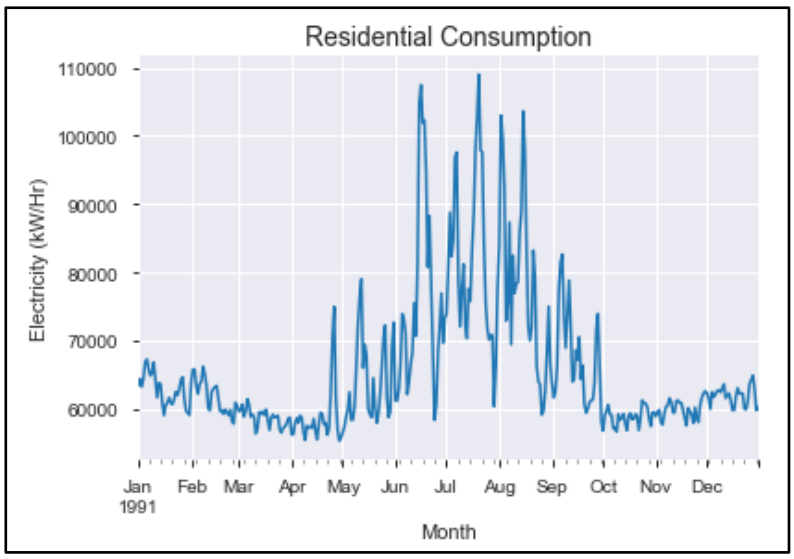


Figure A1.11: Residential Energy Consumption

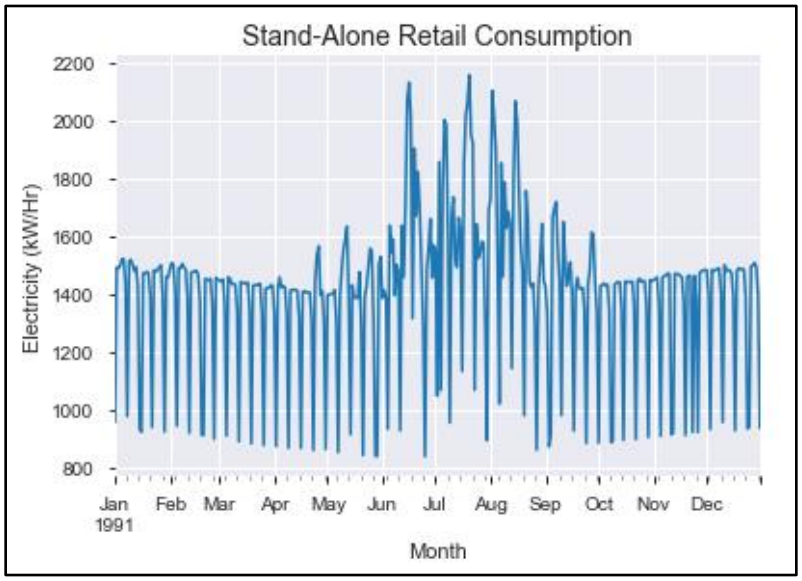


Figure A1.12: Stand-Alone Energy Consumption