Energy Consumption

in the

Steel Industry

Mason Cocroft

Prof. Robert Coyne

DSP 557

27 February, 2022

**Table of Contents**

1. **Introduction (3 – 8)**
   1. Problem Overview (3)
   2. Outcomes of Interest (3)
   3. Background – Personal Interest (4)
   4. Background – Data Sources (4 – 5)
   5. Data Features & Dimensions (5 – 8)
2. **Methodologies (8 – 28)**
   1. Data Preprocessing (8)
   2. Descriptive Analysis (8 – 10)
   3. Data Exploration & Visualization (10 – 26)
   4. Methods, Implementation, & Measures of Success (27 – 28)
3. **Outcomes & Analysis (28 – 34)**
   1. First Evaluation: Linear Regression (28 – 29)
   2. First Evaluation: Ridge Regression (29)
   3. First Evaluation: Lasso Regression (30)
   4. Model Selection, Further Evaluation, & Visualization (30 – 34)
   5. Limitations (34 – 35)
4. **Conclusion (36 – 38)**
   1. Outcome Summary (36 – 37)
   2. Future Research Directions (37 – 38)
   3. Reflection (38)
5. **Resources (39 – 40)**

**Introduction**

**Problem Overview**

The steel industry requires an exuberant amount of energy. Liquifying steel and iron for molding products, moving those massive, dense products, cutting them, and shipping them long distances understandably requires a significant amount of energy. Because of this, it is vital for industry members to maximize efficiency regarding their energy consumption. Accurately predicting energy consumption allows for a more accurate energy cost prediction, and in turn a more efficient product pricing strategy to counter act any changes in the energy consumption patterns. Reducing or mitigating the overhead energy cost is vital for improving any company’s baseline profits.

**Outcomes of Interest**

This primary aim for this project is to determine if a regression model can accurately predict energy consumption measured in kilowatt-hours (kWh) for a single firm in the steel industry, and, by extension, could pricing data be used to accurately predict the cost of energy consumption for said company. There are multiple secondary outcomes to be investigated over the course of the project: Does plotting energy consumption in the steel industry reveal any cyclical patterns? If so, at what time level do these patterns exist (day, month, season, year)? Does adding additional features to better define the time of day and date for each data sample improve model accuracy? Do added time and date related features have a significant impact on prediction accuracy? Can energy consumption be accurately predicted with a model fed only data pertaining to the time and date? How does a time-and-date-only model compare to other models?

**Background – Personal Interest**

I recently started a new job with a steel firm located out of Horsham, Pennsylvania. The company is a steel wholesaler with five locations throughout the United States, and one additional location in Canada. While the company does not operate its own mills, it frequently does business with major domestic and foreign mills. Through my position I have been able to do business with people from every stage of a steel products life cycle, from the smelting and coil rolling stage through the various final product stages (building fabrication, steel servicing centers, produces of metal furniture, etc.). I have also been able to see a steel warehouse in person to witness the massive of the products and get just a glimpse at the total amount of power necessary to get steel products from the mill to an end stage customer. Through this project I hope to show engagement with my new position, and I hope that any significant takeaways can be used by my company to increase knowledge on their energy consumption patterns, and/or to update and improve their product pricing strategy by understanding energy consumption trends.

**Background – Data Sources**

The data for this project is sourced from Kaggle (Csafrit), titled ‘Steel Industry Energy Consumption’. It is originally sourced from the UCI Machine Learning Repository, and was uploaded in accompaniment to three research papers by V.E. Sathishkumar, et al (2020). These research papers assess the predictability of energy consumption using Random Forest models, as well as how the industry can improve its energy consumption to become ‘greener’.

The data itself is provided by the Korean Electric Power Corporation, and is from an anonymous South Korean steel plant that produces the largest steel products, coils and plates. The data for a South Korean firm is used because no data from an American firm was freely distributed or publicly accessible. The difference in steel production between the two countries is relatively similar, with the U.S. being the world’s third largest producer and South Korea being the sixth largest as of 2021 (World Steel Association), so applying any findings of this project to American data is not unreasonable on those grounds. The largest difference between the two countries’ steel industry is in the transportation of the steel, where American steel firms depend heavily on long haul trucking, while South Korean firms do not operate on the within the same scope as South Korea is comparable in size (Worlddata.info) to the state of Virginia (State Symbols USA), rather than to the whole of the United States.

**Data Features & Dimensions**

The data set contains 35,040 samples. Each sample represents a 15-minute interval during the 2018 calendar year. The original dataset contains 10 features:

**Table 1: Features**

|  |  |
| --- | --- |
| Feature Name | As Appears in the Data |
| Energy Usage in kWh | Usage\_kWh |
| Lagging Current Reactive Power in kVarh | LaggingCP\_Reactive |
| Leading Current Reactive Power in kVarh | LeadingCP\_Reactive |
| CO2 Emissions | CO2 |
| Lagging Current Power Factor | LaggingCP\_Factor |
| Leading Current Power Factor | LeadingCP\_Factor |
| Number of Seconds Since Midnight | NSM |
| Week Status | Week\_Status |
| Day of the Week | Day\_of\_week |
| Type of Load | Load\_Type |

Lagging and leading current reactive power are measured in reactive energy (kVarh), which is a measure of how an electrical system is affected by the distribution of energy that is either a lagging remnant of previous use or is leading up to the next use. Lagging and leading current power factor are measured as a percentage. CO2 emissions are measured by parts per million. Each of these features are continuous. Number of seconds go midnight is simply the number of seconds that have passed since midnight, and is a continuous integer variable. Week status, day of the week, and type of load are categorical variables. They are converted to a numeric scale for calculation purposes. Week status is either 1 for a weekday or 0 for a weekend. Day of the week starts as 1 on Monday and finishes as 7 on Sunday. Load type is 0 for a light load, 1 for a medium load, and 2 for a large load. The data source does not provide insight into how a load is determined to fit into one of these categories. Four additional numerated categorical variables were addended to this dataset:

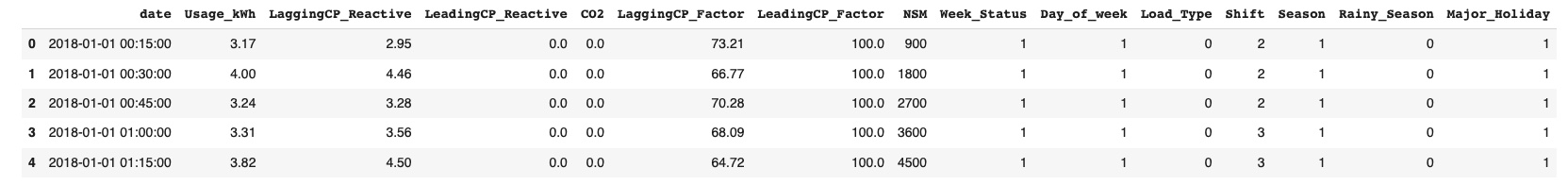
**Table 2: Addended Features**

|  |  |
| --- | --- |
| Feature Name | As Appears in Data |
| Shift | Shift |
| Season | Season |
| Rainy Season | Rainy\_Season |
| Major Holiday | Major\_Holiday |

Shift represents what work shift was on at the time the sample was taken, and has values of 1 for first shift (9am-5pm), 2 for second shift (5pm-1am), or third/graveyard shift (1am-9am). Season is for season of the year, 1 for winter, 2 for spring, 3 for summer, 4 for fall. Rainy season is a binary variable for whether the sample is in the rainy season or not. The rainy season is a period of nearly constant precipitation, and is also equivalent to the summer period, which is the months of July and August (Aspeli, 2020). Lastly, a binary variable for whether or not a sample is on a day that is a major holiday. Major holidays are defined as any holiday that Koreans typically have off from work. There were 11 holidays that met this criterion according to a Korean state operated tourism website (VisitKorea):

**Table 3: Holidays**

|  |  |
| --- | --- |
| Holiday | Affected Date |
| New Year’s Day | January 1st |
| Korean New Year Celebrations | February 15th, 16th, and 17th |
| Korean Independence Day | March 1st |
| Children’s Day  (Celebration of children) | May 5th |
| Buddha’s Birthday | May 22nd |
| Liberation Day  (Victory over Japan day) | August 15th |
| Chuseok Celebrations  (Autumn harvest festival) | September 24th, 25th, and 26th |
| National Foundation Day  (Celebration of the first Korean state) | October 3rd |
| Hangul Day  (Celebration of the Korean alphabet) | October 9th |
| Christmas | December 25th |

Here is a glimpse at the dataset, including the original and addended features:

**Methodologies**

**Methodologies**

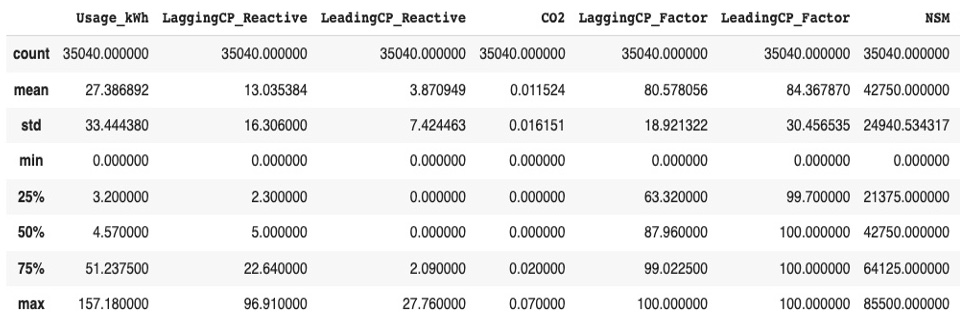
**Data Preprocessing**

The original data set included zero missing values. The original csv file was uploaded to Microsoft Excel for manipulation. The categorical variables in the original data set were numerated from their original text-based format. The four addended variables were entered in numerated form to begin with. After adding the additional features, there remained zero missing values. The Excel file was uploaded to Google Collab for Python work. The file was saved as ‘data’, and descriptive analysis was performed on this dataset as it was all encompassing. For work on the regression models, ‘data’ was split into three datasets: the addended data set ‘addended’ (which was a copy of ‘data’), the original data set ‘original’ (containing only the variables found in the original Kaggle dataset), and a dataset made up of only the features pertaining to time and date, ‘date\_only’.

**Descriptive Analysis**

Table 4 displays various descriptive statistics for each of the continuous variables of the ‘data’ dataset, containing all the original variables and the addended ones:

**Table 4: Descriptive Statistics – Continuous Variables**

****

At first glance, Usage\_kWh appears to show the greatest range of outcomes, while CO2 and the current power variables show generally little range. This becomes explicitly clear when looking at the standard deviation and interquartile range values, rather than just the minimum and maximum values. NSM is a repetitive cycle of seconds since midnight, with each day’s worth of samples following the same cycle of seconds from 0 to 85,500 since midnight, so that variable is not well defined by these descriptive statistics.

The categorical features are explored in table 5:

**Table 5: Descriptive Statistics – Categorical Variables**

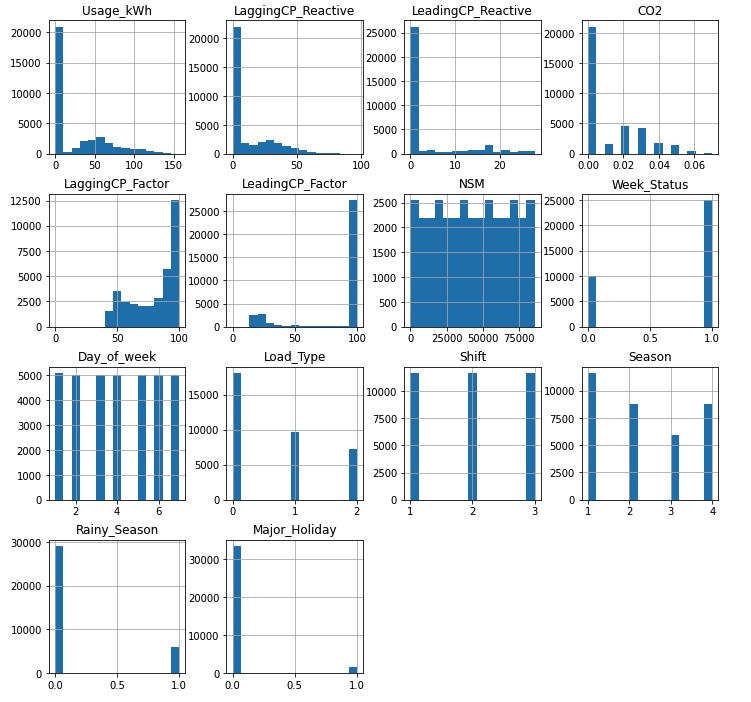
****

Minimum, maximum, and the interquartile range measurements are not very informative due to the features being categorical, but the standard deviation and mean do give some hint as to how those features are distributed.

**Data Exploration & Visualization**

To further analyze the distribution of the various features, a grid of plots, Figure 1, was created to display the values for each variable on the x-axes, and the sample counts on the y-axes.

**Figure 1: Grid of Distribution Plots**



The categorical variables are easily identifiable at a quick glance due to the few bins needed to contain all samples’ values. The days of the week are nearly evenly divided amongst the seven days, only slightly off due to the weekday alignment of the start (Monday) and end date (Saturday) of 2018. Week status is distributed as approximately 2/7ths of the samples are during the weekend and 5/7ths are during weekdays, again just off due to the starting and ending date of the calendar year. Shift is evenly distributed amongst three values as every day of the year has the same evenly distributed shift periods. Load type reveals that the majority of the loads worked during the taken samples were minimum loads, medium loads being approximately half the number of samples as minimum loads, and maximum loads being approximately ¾ the loads as medium loads. For the purpose of this project, t5he South Korean seasons were not considered to be evenly distributed 3-month periods, but rather were aligned with different temperate periods. Winter is the coldest months, spring is the warmer months before the rainy season, summer is the rainy season, and fall is the remaining months leading back to winter. The Season feature mirrors that: 4 months for winter, 3 months for spring, 2 months for summer, and 3 months for fall. As mentioned, rainy season mirrors summer, so approximately 1/6th of the values equal 1 for rainy season, and 5/6ths equal 0 for not rainy reason. Lastly for the categorical variables, major holidays make up only 15 of the 365 days of the year, so the distribution of 1s for holiday and 0s for not matches that ratio.

The continuous variables are less neatly distributed by nature. Usage\_kWh, LaggingCP\_Reactive, LeadingCP\_Reactive, and CO2 have a significant percentage of their variables at or around zero. If the values closest to zero are ignored for Usage\_kWh, LaggingCP\_Reactive, and CO2 would appear approximately normally distributed. That would still not be the case for LeadingCP\_Reactive. These features have so many values at zero because zero is the floor value, energy consumption cannot be below zero. On the other hand, LagginCP\_Factor and LeadingCP\_Factor have a significant portion of their values at or around 100, but would still not be approximately evenly distributed if those values were ignored. This is partly due to 100 being the ceiling value as these features are measured as a percentage. NSM is displayed as a ‘castle wall’ shaped block due to the high range of seconds on the x-axis and the exactly repeating cycle of seconds in a day during the year the samples were recorded over. The distribution of these variables at or around zero, or at and around 100, are further investigated in Figures 2 through 6 below. These figures zoom in on the x-axis to only show the values close to zero (or 100), and how they are distributed. CO2 is not further investigated as a quick glance at the data reveals that the vast majority of the samples are simply equal to zero, not values close to zero.

**Figure 2: Magnified Plot – Usage\_kWh**

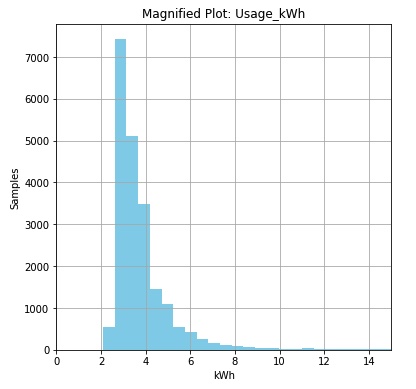
****

Figure 2 reveals that the majority of samples are actually valued between 2 and 4 kWh, and very few are valued at zero. The descriptive statistics did reveal there is at least one sample valued at zero, but a bin with 1 out of 35,040 samples would be very hard to see with the human eye, as is the case here.

**Figure 3: Magnified Plot – LaggingCP\_Reactive**

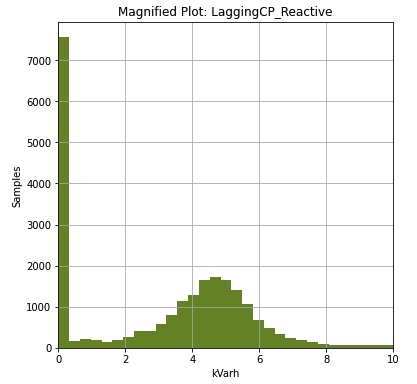
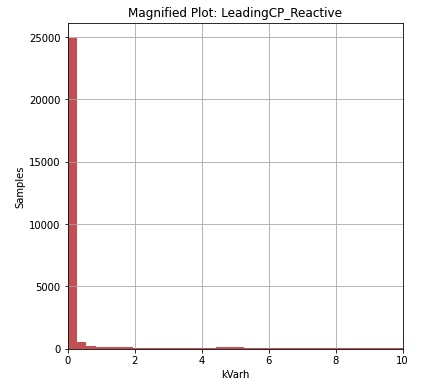
****

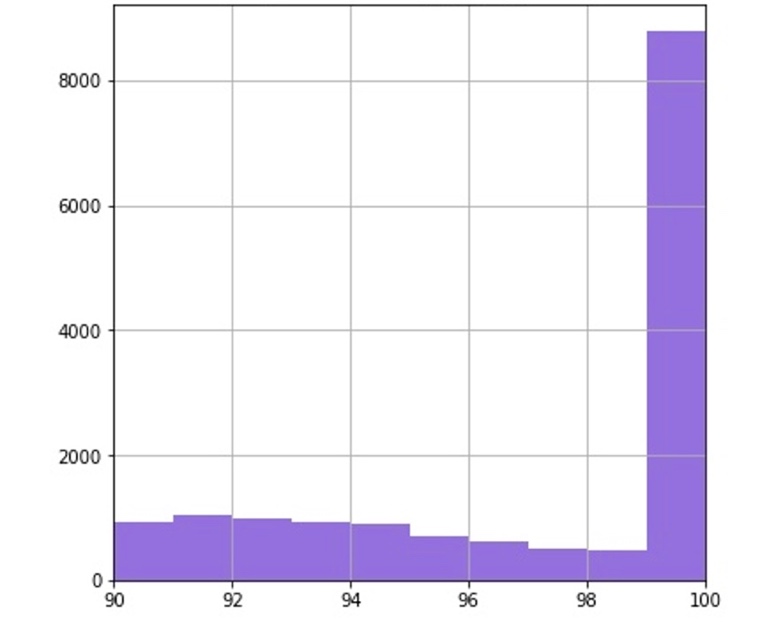
Figure 3 reveals that LaggingCP\_Reactive has a significant portion of samples valued at zero, but the samples outside of these are approximately evenly distributed.

**Figure 4: Magnified Plot – LeadingCP\_Reactive**

****

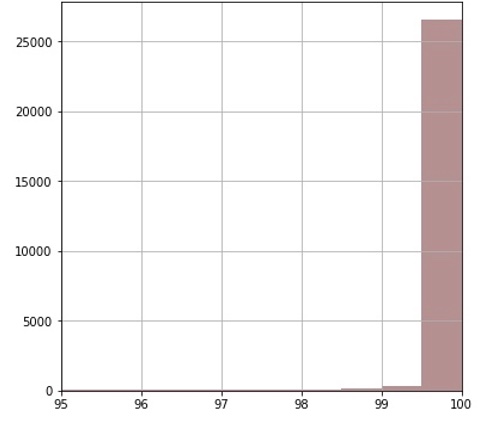
It is not the same case for LeadingCP\_Reactive, as Figure 4 shows that about 2/3rds of the samples are simply equal to zero.

**Figure 5: Magnified Plot – LaggingCP\_Factor**



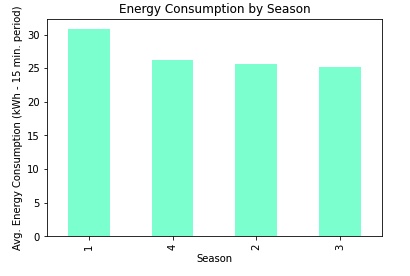
LaggingCP\_Factor has many instances of samples being equal to 100, and viewing the entire plot in the Figure 1 grid does not reveal any sign of an approximately normal distribution The same can be said for LeadingCP\_Factor in Figure 6 below.

**Figure 6: Magnified Plot – LeadingCP\_Factor**



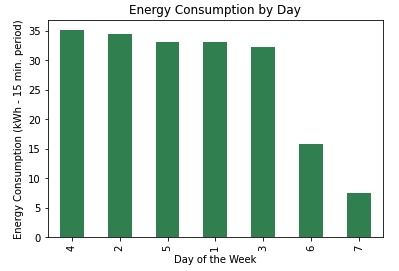
The categorical features Season, Day of the Week, Major Holiday, Shift, and Load Type were each individually analyzed for how energy usage was distributed for each features’ subcategories. These features’ relationship with energy usage is visually explored in Figures 8 to 11. Each of these plots are organized by category the category with the highest average energy usage on the left, to the category with the lowest average energy usage on the right.

**Figure 8**



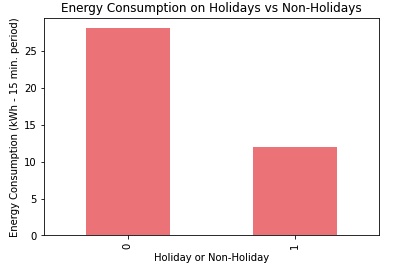
For Season, winter consumes the most energy, followed by fall, spring, and lastly summer. This appears to imply a correlation between seasonal temperature and energy consumption, though outside factors may also be the cause of this phenomena (for example: an unidentified cyclical nature to steel consumption over the course of a year).

**Figure 9**



As one might assume, the weekdays consume significantly more energy in any given fifteen-minute interval than the weekends. Similar to the United States, Saturday and Sunday (6 and 7 on the x-axis, respectively) are traditionally days off from work. Steel is considered a 24/7 industry, but Saturday and Sunday are traditionally days for maintenance, inspection, or catching up on any similar tasks. Saturday consumes nearly twice as much energy as Sunday, which is likely due to any work that was not completed on Friday being finished on Saturday morning, with the rest of the day being used for other miscellaneous tasks. Sunday would only be used for miscellaneous tasks, further reducing its average energy consumption by not having any ‘catch-up’ work to complete.

**Figure 10**



Holidays that are traditionally taken off from work in South Korea also consume less energy than non-holidays, similar to how weekends consumed less than weekdays in Figure 9. Because weekends are not removed from the data to measure this feature, less energy is consumed on average on non-holidays than on the weekdays displayed in Figure 9. Comparing holidays to the weekends, holidays consume about 12 kWh per 15-minute interval, less than Saturdays (15 kWh), but more than Sundays (7 kWh).

**Figure 11**

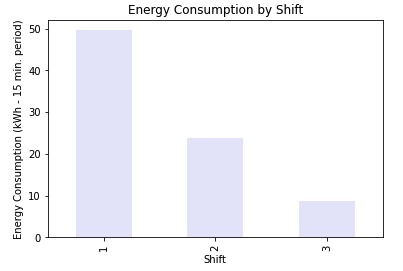
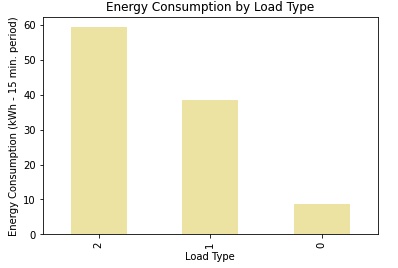


Figure 11 shows that energy consumption follows a pattern similar to labor standards in the steel industry. First shift consumes the most power, and first shift is when the bulk of material moving, smelting, transport loading, and other high energy consuming tasks are completed. Second shift consumes the next most, and second shift is when any of the previously mentioned tasks are completely followed by the start of maintenance and cleanup procedures. Third shift consumes the least as third shift is used to complete any maintenance tasks and cleanup procedures, and prepares any machinery or work spaces for use the upcoming morning.

**Figure 12**

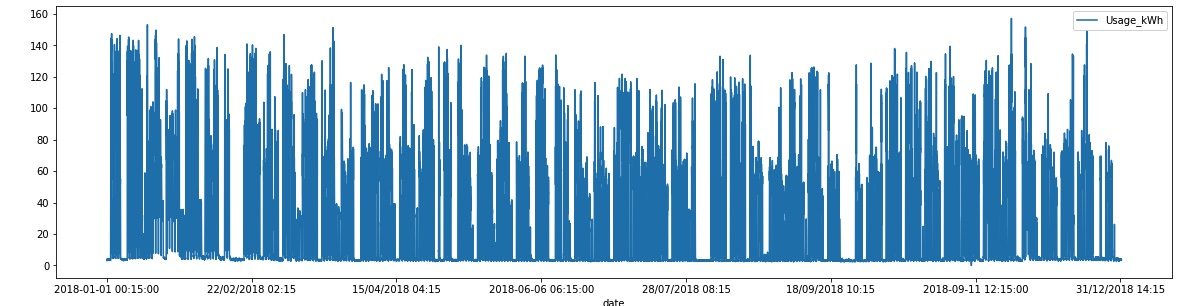


The consumption of energy by load type also appears to follow conventional thinking: the larger the load of steel, the more energy is consumed. As previously mentioned, a ‘load’ was not clarified by the poster of the data set. Minimum loads, marked as zero, also include the times when no load was being handled, which skews the data. There was no informed way to separate minimum loads from no loads.

The energy usage data was also visualized over time for various periods of time to see if there was a pronounced, cyclical nature to the consumption of energy. It was visualized by year, random month in each season, and second day in each selected month.

It was first visualized by year:

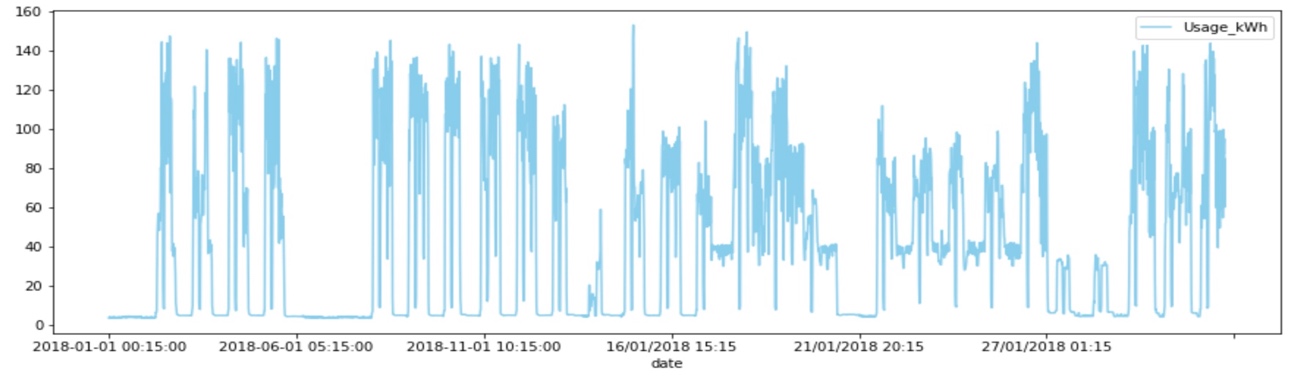
**Figure 13: Usage over the Course of a Year**



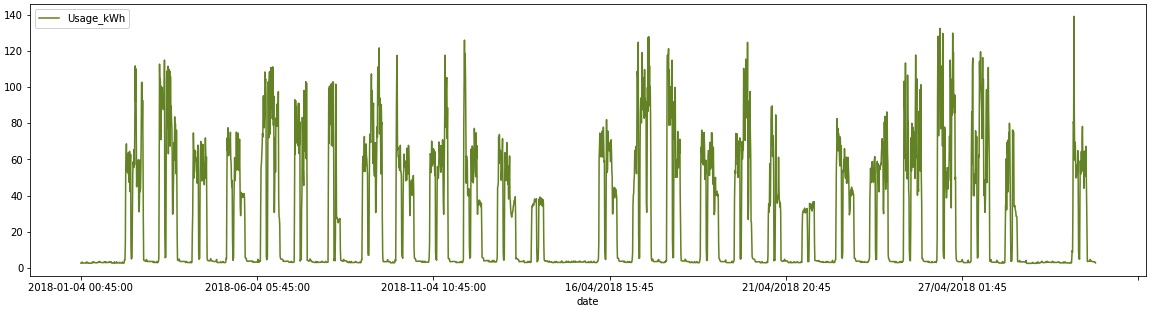
Most of the gaps in the data are explained by the cycle of weekdays and weekends over the course of a year. Some of the larger gaps can be explained by the major holidays, specifically the gap in February, which aligns with the three-day holiday for Korean New Year, and the gap in September, which aligns with the three-day Chuseok celebrations. The other larger gaps do not line up as cleanly with the marked major holidays, and therefore have no explanation known to this study. Paying attention to the peaks of the energy usage over the course of the year, this diagram seems to corroborate with Figure 8 that winter has the highest energy consumption and summer has the lowest.

Next, one month was singled out for each season: January for winter (Figure 14), April for Spring (Figure 15), July for summer (Figure 16), and October for fall (Figure 17):

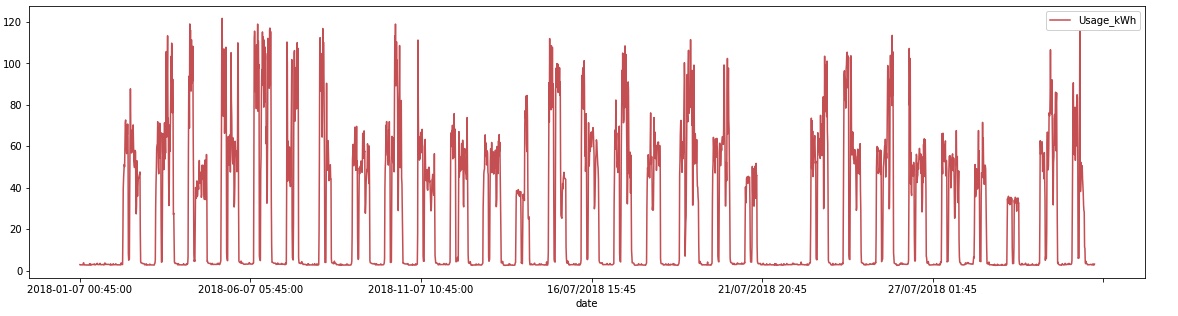
**Figure 14: Energy Usage in January**

****

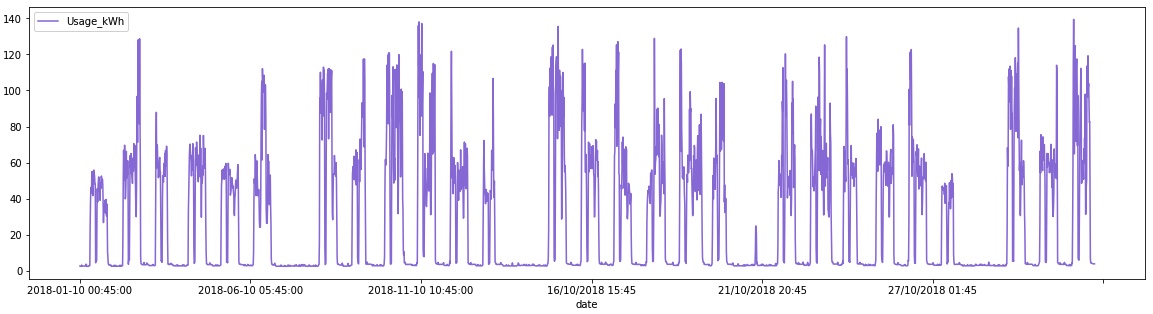
**Figure 15: Energy Usage in April**

****

**Figure 16: Energy Usage in July**

****

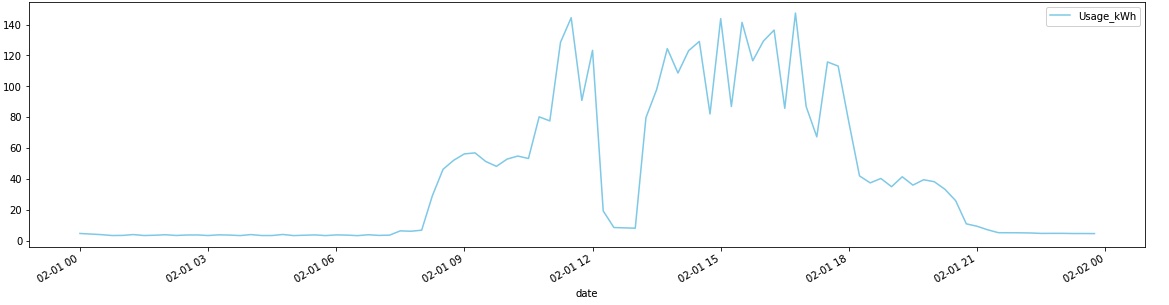
**Figure 17: Energy Usage in October**

****

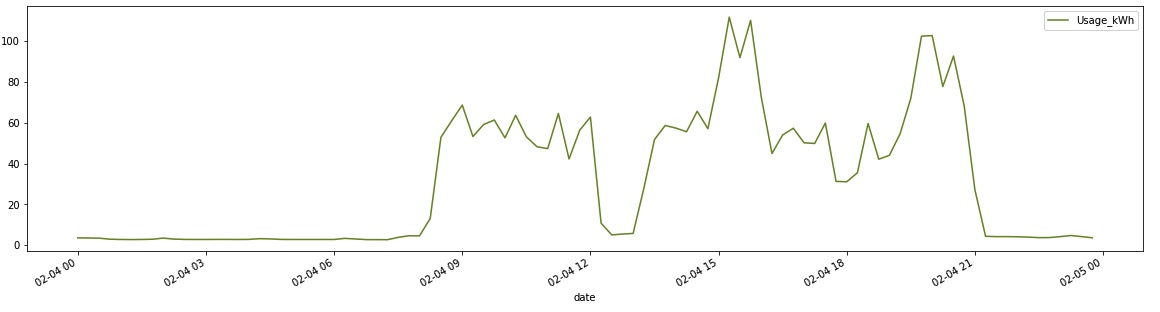
The many smaller gaps for each of these diagrams align with the day/night cycle, with significantly less energy being used at night, which agrees with the energy distribution by shift in Figure 11. Most of the larger gaps align with single day holidays in these Figures, when less energy was used over the course of an entire day, like New Year’s Day on January 1st. Additionally, the y-axis scales for each month similarly to how energy usage fluctuates per season. Winter has a peak energy expenditure around 160 kWh, spring is around 140 kWh, summer is around 120 kWh, and fall is around 140 kWh. The energy peaks bottoming out in summer and ticking back up again in fall also appear to support the hypothesis that energy consumption is connected to seasonal temperature. This may also indicate that the rainy season prohibits the Korean steel industry from operating at max capacity.

Lastly, the energy consumption during the second day of each selected month was visualized:

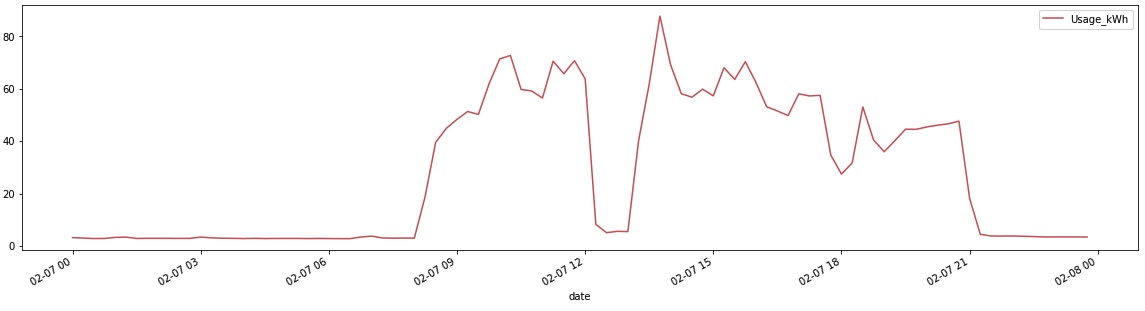
**Figure 18: Energy Usage on January 2nd**

****

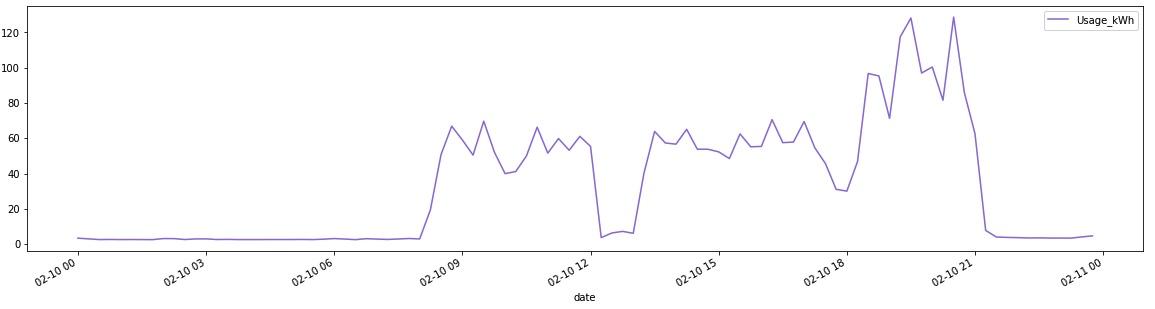
**Figure 19: Energy Usage on April 2nd**

****

**Figure 20: Energy Usage on July 2nd**

****

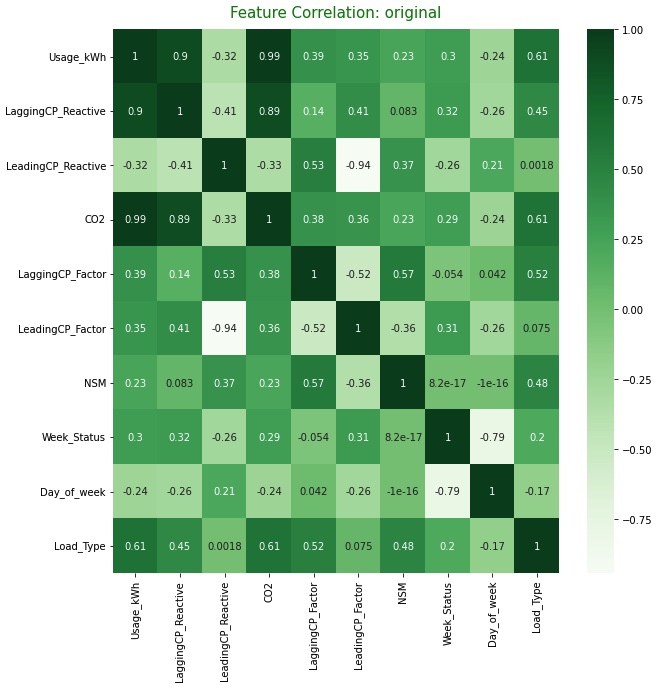
**Figure 21: Energy Usage on October 2nd**

****

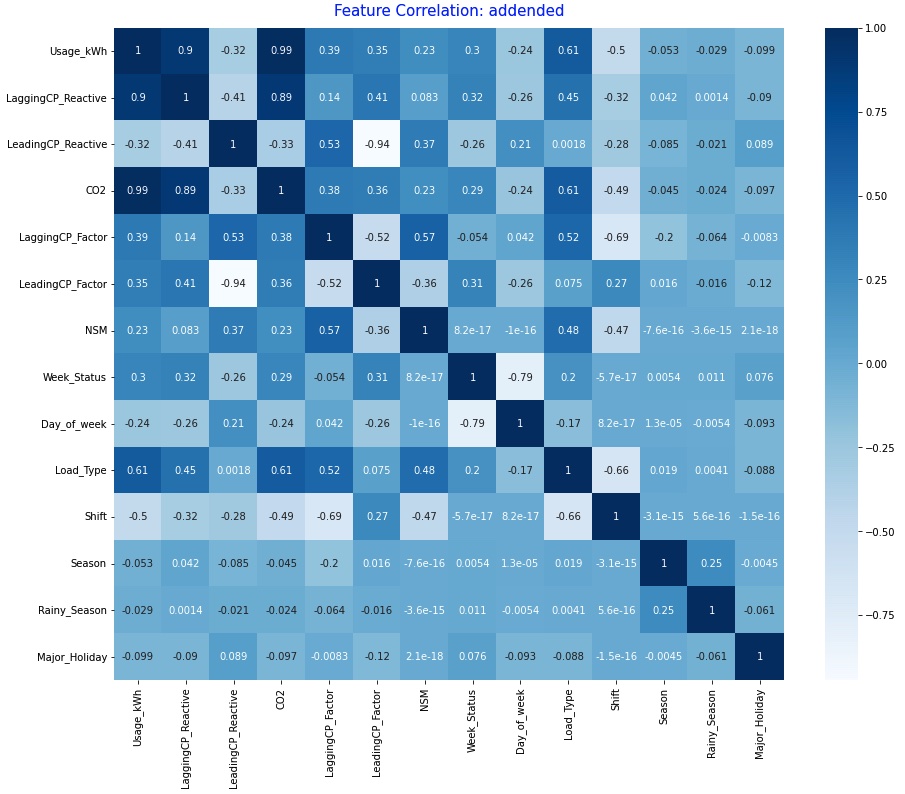
Each of the selected days follows an energy usage pattern that concurs with energy usage per shift in Figure 11. The bulk of the energy used each day is consumed between approximately 8:30am and 9pm, matching first shift consuming the bulk of the energy, second shift consuming the next most, and third shift consuming the least. There is variation in when each day has the most energy used, which is likely explained by the variation in each day’s work load and the order in which that work is completed. Each day also has a noticeable dip from 12pm to 1pm, the traditional lunch hour, which interrupts the high energy usage period. There is a less pronounced dip each day around 6pm, which would align with second shift’s break. Energy consumption is low enough throughout third shift that their break is not visually detectable in these graphs. The y-axis scale for each of these days follow a similar pattern to the months’ scale changes. The winter day (Jan. 2nd) has the highest peak energy, while the summer day (Jul. 2nd) has the lowest. Spring (Apr. 2nd) and fall (Oct. 2nd) have peaks in between.

Feature correlation was also investigated. The data was first copied into three separate datasets: addended which includes all original and added features, original which includes only features from the original Kaggle dataset, and date\_only which included only the features pertaining to time and date. This was done at this point in case changing the selected features had any effect on feature correlations. A correlation matrix was visualized for each of the data sets:

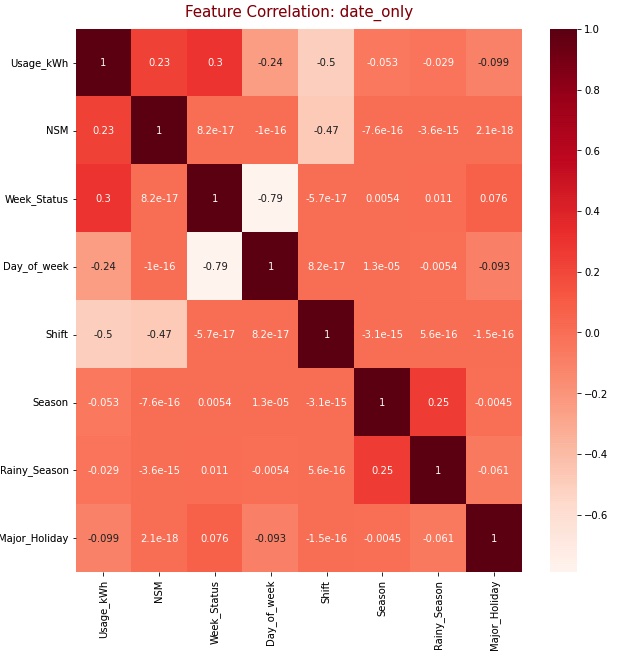
**Figure 22: Feature Correlation - Original**

****

**Figure 23: Feature Correlation – Addended**

****

**Figure 24: Feature Correlation – Date Only**

****

The three matrices do mirror each other fairly well, no significant changes to feature correlation are made when adding or subtracting variables. The most significant correlations are Usage\_kWh, LaggingCP\_Reactive, CO2, and Load\_Type. The strongest correlation is between CO2 and Usage\_kWh (0.99). Of the addended features, the strongest correlation is between Shift and LeadingCP\_Factor (0.27). Of the variables remaining in the date\_only dataset, the strongest correlation is between Week\_Status and Usage\_kWh (0.30). There is no significant change in correlation when adding and subtracting variables.

**Methods, Implementation, & Measures of Success**

This project is specifically investigating whether a regression model can be used to accurately predict energy consumption. To do so requires the use of a few Python packages. In set up, pandas and numpy are used for data uploading, preprocessing, and exploration. Matplotlib.pyplot is used for figure creation, with the seaborn package being used to help visualize correlation. The data will be split into training and test sets, done so with the help of the train\_test\_split function from sklearn.model\_selection. Three different regression models will be tested for each of the datasets: linear, ridge, and lasso. The different models will be tested to see which is most appropriate for this data. These model functions are each pulled from sklearn.linear\_model module.

The models will be evaluated first using a simple accuracy score and comparing model run times measured in milliseconds (ms). The score evaluates the model’s accuracy as a decimal percentage. The best performing model in terms of the simple accuracy score will then be evaluated further using mean absolute error, mean squared error, and root mean squared error. These measurements evaluate the size of the errors the models make. The models will be run on each of the three datasets at each stage of evaluation. The best model will also be visualized for the three datasets. The visualizations will show the actual energy consumption on the x-axis, and the predicted energy consumption on the y-axis. To perform these evaluations, metrics from the sklearn module will be used alongside functions from the time and math packages. The best performing model will be used to evaluate the primary outcome of whether or not a regression model can accurately predict energy consumption. The second outcome of the ability to predict energy usage using only date and time related features will be answered by comparing the performance of the models on the date\_only dataset compared to performance of the original and addended datasets. The other secondary outcomes found answers in the data exploration and visualization phase. The papers by Sathiskumar, et al. focus on Random Forest models used to predict Usage\_kWh, so the regression outcomes will also be loosely compared to the performance of the models from their work.

**Outcomes & Analysis**

**First Evaluation: Linear Regression**

* Addended Data Set
  + CPU Times user 13.8 ms, sys: 11.1 ms, total: 24.9ms
  + Wall time: 22.5 ms
  + **Score: 0.98207**
* Original Data Set
  + CPU times: user 13.8 ms, sys: 11.1 ms, total: 24.9 ms
  + Wall time: 17.4 ms
  + **Score: 0.98199**
* Date Only Data Set
  + CPU times: user 16.2 ms, sys: 26.1 ms, total: 42.3 ms
  + Wall time: 26.9 ms
  + **Score: 0.34642**

The linear regression model performed strongly on the addended and original data sets, but saw a significant decline in accuracy for the date\_only dataset with accuracy decline to the 0.3s from the high 0.9s. The addended data set saw marginal improvement in accuracy of 0.00008 over the original data set. Despite the date\_only dataset taking nearly twice as long to run, each of these models ran in well under a second which is impressive considering there were 28,032 samples in the training set.

**First Evaluation: Ridge Regression**

* Addended Data Set
  + CPU times: user 16 ms, sys: 29.3 ms, total: 45.3 ms
  + Wall time: 32.2 ms
  + **Score: 0.96969**
* Original Data Set
  + CPU times: user 15.3 ms, sys: 16.1 ms, total: 31.4 ms
  + Wall time: 17.4 ms
  + **Score: 0.96955**
* Date Only Data Set
  + CPU times: user 9.79 ms, sys: 3.95 ms, total: 13.7 ms
  + Wall time: 11.6 ms
  + **Score: 0.34643**

The ridge regression models also performed well, but were outperformed by the linear regression models. Like linear regression, addended was the most accurate with original just behind with the scoring lowering by only 0.00014. Again mirroring linear regression, the date\_only dataset performed significantly worse, once more dropping to the 0.3s from the high 0.9s, but actually saw an improvement in accuracy by 0.00001. Each model ran in under a second once again, though this time date\_only took the least amount of time to run and the addended set took the most amount of time to run.

**First Evaluation: Lasso Regression**

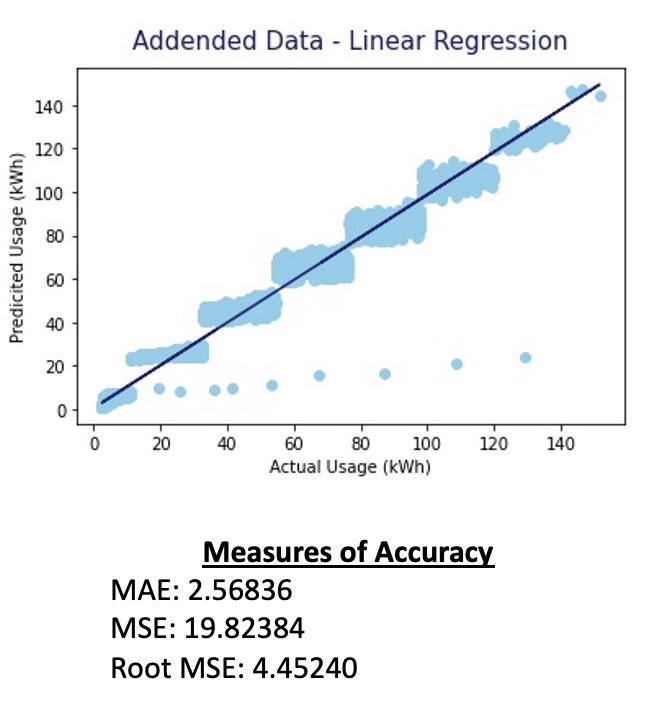
* Addended Data Set
  + CPU times: user 54.9 ms, sys: 59.2 ms, total: 114 ms
  + Wall time: 62 ms
  + **Score: 0.91364**
* Original Data Set
  + CPU times: user 51.6 ms, sys: 35.9 ms, total: 87.5 ms
  + Wall time: 57.8 ms
  + **Score: 0.91165**
* Date Only Data Set
  + CPU times: user 16.2 ms, sys: 8 ms, total: 24.2 ms
  + Wall time: 18.3 ms
  + **Score: 0.34212**

The lasso regression models saw a relatively significant decrease in accuracy from the addended and original datasets. It performed approximately 0.5 worse on both datasets compared to ridge regression, and 0.7 worse on both datasets compared to linear regression. Once again addended slightly outperformed original, this time by a slightly higher margin of 0.00199. Date\_only was in the 0.3s again but did not see the same change in performance compared to the previous models that addended and original saw, only falling ~0.004. Each dataset again ran in well under a second, but saw increased run times compared to the previous methods.

**Model Selection, Further Evaluation, & Visualization**

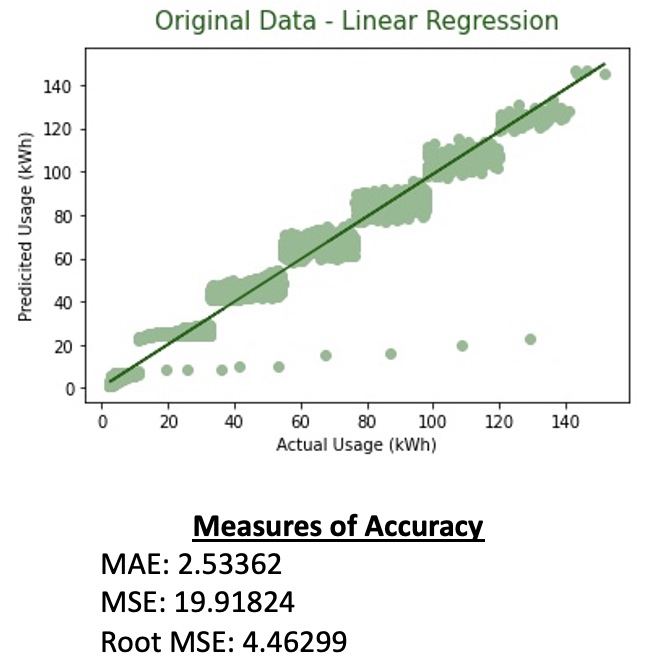
The linear model appears to be the best regression model to use of the three tested for this project. It performed the best on both the addended and original data set, and was very comparable in performance on the date\_only data set compared to the ridge regression model. The linear models also ran the quickest of the three model types, though all three ran in under a second for each data set. Moving forward with the linear model, mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) will now be used to further evaluate performance on each of the datasets, and each the model for each data set will be visualized:

**Figure 25: Further Evaluation – Addended Data**

****

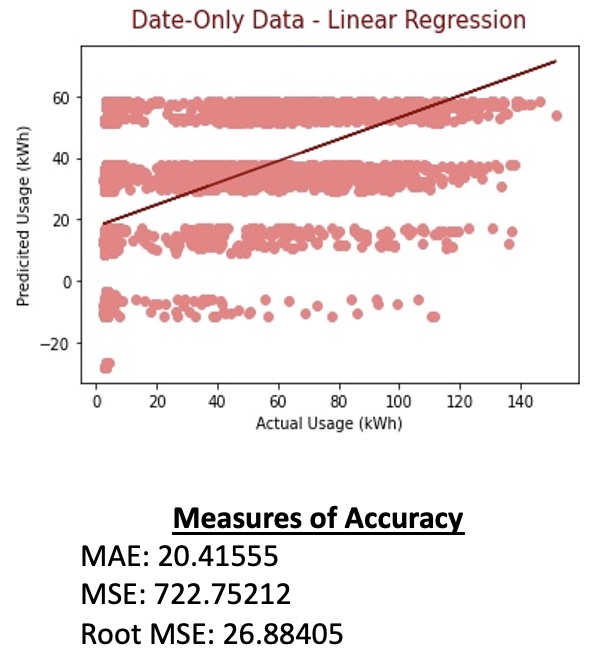
The addended model’s strong accuracy score seems to be reflected by both the graph and additional accuracy measurements. By MAE measures, the model is on average only 2.5 kWh off on its predictions compared to the actual result. MSE and RMSE are slightly less confident in the model’s ability, with RMSE saying that predictions are an average of 4.45 kWh off from the actual. The graph shows a step-ladder like structure, though the trend line is an approximately 45-degree angle starting at zero. The thickest ‘step’ is approximately 10 kWh wide (approximately equally distributed above and below the trend line) for prediction and 20 kWh long for actual. The thinnest step is approximately 5 kWh wide (again approximately evenly distributed) for prediction and 20 kWh long for actual. The step-ladder like structure is likely due to unaccounted for variables. The alpha of the points was set to 0.1, so it is evident that the steps are densely packed and are not formed by a loose collection of points. There are nine distinct outlier points located toward the bottom of the graph, where the actual energy usage is much less than the actual usage. There are no outliers where the predicted energy usage is much higher than the actual.

**Figure 26: Further Evaluation – Original Data**



Just as it did for the simple evaluation of the three regression types, the original dataset model performed nearly identically to the addended model. The steps are similarly characterized to the steps for the addended visualization. In terms of MAE, it actually performed slightly better by 0.03474, while it performed slightly worse for the MSE (-0.0944) and RMSE (-0.01059). The same stepladder shape is present in this plot, and the same outliers are present where the prediction was much higher than the actual. The alpha was again set to 0.1 so the step structures are densely packed.

**Figure 27: Further Evaluation – date\_only Data**



The date\_only data set performs significantly worse, as it did with the simple accuracy score, for all the additional measurements and in the visualization. Interestingly, the steps turned into stripes, and there was a number of predictions made below zero despite no actual values below zero. The stripes are approximately equal in width, being about 10 kWh thick based on the y-axis variance. The MAE, MSE, and RMSE are only significantly worse compared to the addended and original data set results. The maximum kWh of the predictions is also significantly lower than the other models.

**Limitations**

Some limitations are extensions of the data itself. It is not very robust, and there are likely many significant features not included in the data. Additional non-categorical features like tonnage moved, loads handled, and cost of energy would significantly improve the character of the data. Likewise, the data not including any financial features means the extended primary outcome of predicting energy costs is impossible, though prediction accuracy would likely be high due to high prediction accuracy for usage. The overabundance of zero or near zero values for the various continuous variables likely plays a part in the step ladder like structure of the graphs. The continuous variables also not being normally distributed likely also played a part. Normalizing the continuous variables may fix or improve this issue. The structure of the plots could also be due to the cyclical nature of energy consumption.

Many outside factors not accounted for in this study have a significant impact on the steel industry and on the usefulness of this study. The difference between the American and Korean labor markets may cause unreconcilable differences between American and Korean data. Korean culture is not familiar to me innately, so the major holidays, shift lengths, or general work schedule may not be accurately portrayed in the addended features. The state of the steel market itself would have a significant impact on the data, and the Korean steel market saw two significant outside events in the year this data was taken, 2018. Firstly, the Trump administration enacted significant tariffs on steel imports, and later struck a deal with the South Korean government negating the tariffs but reducing Korean steel exports to the United States by 30% of the average of the previous three years (Jin, 2018). At the same time, exports of products primarily made from steel (cars, nails, etc.) dramatically increased as they were not affected by the tariffs (Swanson, 2020), so it is difficult to tell would affect this would have had on the Korean steel market. Secondly, there was a push to dramatically alter the Korean labor market, with the South Korean governments enacting laws to shorten the work week from 68 to 52 hours (Wamsley, 2018). The enforcement of these laws is difficult to track, but a labor-intensive industry like the steel industry likely would have been significantly impacted by such legislation.

The data being for only one year also limits the study. The impact of the described outside market forces is not readily detectable without additional years’ information. Likewise, a yearly energy usage cycle is not detectable without more than one year’s worth of information. This is further complicated by this data predating the COVID-19 pandemic. Disruptions to the global market during the pandemic may make the data after 2019 unintelligible when compared to this data. While there were 35,040 samples in the dataset, measuring by 15-minute intervals is likely not as effective for this type of study as comparing samples equal to days over the course of a decade would be. Likewise, the random forest models of Sathiskumar’s papers (2020) more accurately predicted this data so regression models themselves may just not be the ideal tool to use on this data.

**Conclusion**

**Outcome Summary**

The primary outcome of whether or not regression models can be used to predict energy consumption in the steel market appears to be affirmatively answered. The tested regression methods each produce a simple accuracy score above 0.9 for the original and addended data. The date\_only data will be addressed further later. Of the tested regression methods, linear regression performed the best with a simple accuracy score over 0.98. The MAE, MSE, and RMSE likewise confirmed that the model was accurate, though they were not necessarily as flattering as the simple accuracy score was. The plots of the linear regression showed some unique characteristics which raised further questions about the model’s performance. Overall, it appears safe to say that linear regression is capable of accurately predicting energy consumption, though it is likely not the best tool for the job. This appears to be confirmed by Sathiskumar’s work with random forest models.

The secondary outcomes each had a much clearer answer than the primary outcome. Solely using features pertaining to date and time cannot produce a model that accurately predicts energy consumption. Each regression model produced an accuracy score in the 0.3s for the date\_only dataset, and the plot for the linear regression model further highlighted the inability of model to predict when using date and time features alone, predicting a number of samples to have used a negative amount of energy. Adding additional features to further define the date and time via the addended data set did slightly improve model accuracy. However, the improvement was miniscule and was more likely due to an increase in random chance of correct prediction from adding variables than it is to being true improvement. None of the added features (Major Holiday, Shift, Season, or Rainy Season) had a significant individual or combined impact on the model’s accuracy. Lastly, the plotting of energy consumption over various time periods did appear to reveal a cycle nature to energy consumption, at least at the daily, weekly, and monthly levels. Without additional years’ data it is impossible to see if seasonal or yearly patterns are in apparent. There was an apparent day and night cycle, as well as a weekly cycle of high weekday energy use followed by a decrease on weekends. The major holiday feature did help to explain various gaps in energy consumption when looking at the plot of the entire year’s consumption, and when looking at different months’ consumption.

**Future Research Directions**

There are many directions to take to further investigate this data. First and foremost, obtaining data from an American steel firm would be the ideal next step. Likewise, obtaining multiple years’ worth of data would allow for a more in-depth investigation into the patterns and predictability of energy use in the steel industry. Data from a wider time period would help to define seasonal or yearly energy consumption cycles. Adding financial information, particularly the average cost of energy for a time period (ideally daily if investigating multiple years), would also provide invaluable information for predicting overhead costs and would greatly assist companies in drafting strategic pricing plans to negate any predictable changes in energy consumption. If sticking with the data available here, it would be worthwhile to spend more time investigating events that may have impacted the steel industry. These events could be significant labor disputes, major weather events, or other large outside factors. Knowing more about any events like these would help to explain any gaps in energy consumption that could not be explained by the major holiday feature. A greater understanding or Korean culture, both the general culture and specifically the work culture, would go a long way in understanding and characterizing this data.

**Reflection**

I would like to lead by saying I enjoyed working on this project, and it was a unique and valuable experience working on it. That being said, I believe working alone was my main hinderance over the course of this class. All other problems sort of branched out of not having fellow group members to siphon work off to or to assign work to based on their strengths and weaknesses with. My strengths are writing, presenting, and researching content, so it would have been nice to have someone to complement those strengths by having a greater technical knowledge or more coding experience. Having a groupmate with a background in a STEM field (math, statistics, computer science, etc.), or just someone that this material came more naturally to, likely would have improved this project’s results and decreased the overall workload which would reduce class related stress by extension. Mainly due to time crunches (admittedly, often self-imposed), I ran into problems where I was unsure of what to do and simply ran out of time to find the best solution possible and had to settle for something that was okay but lacking something. Working with a group may have helped to mitigate this. Overall I would say I am content with the results of my work, but I still see plenty of room for improvement.

**Resources**

Advisor, International Market. “Doing Business in South Korea.” *South Korean Culture - Doing Business in South Korea*, Doing Business Guide, http://www.southkorea.doingbusinessguide.co.uk/the-guide/south-korean-culture/.

Aspen, Mira Dahl. “Embassy of the Republic of Korea to Norway.” *The Phenomenon of Rainy Season in Korea, How Koreans Enjoy It to Its Fullest, And The 2020 Rainy Season Situation 상세보기|Citizen JournalistsEmbassy of the Republic of Korea to Norway*, 14 Sept. 2020, https://overseas.mofa.go.kr/no-en/brd/m\_21237/view.do?seq=87.

Csafrit. “Steel Industry Energy Consumption.” *Kaggle*, 21 Dec. 2021, https://www.kaggle.com/csafrit2/steel-industry-energy-consumption.

“Fundamentals of Smart Metering - Kwh and Kvarh Meters.” *EIT*, Engineering Institute of Technology, https://www.eit.edu.au/resources/fundamentals-of-smart-metering-kwh-and-kvarh-meters/.

Jin, Hyunjoo, and Joyce Lee. “U.S., South Korea Revise Trade Deal, Korean Steel Faces Quota.” *Reuters*, Thomson Reuters, 26 Mar. 2018, https://www.reuters.com/article/us-southkorea-trade-usa/u-s-south-korea-revise-trade-deal-korean-steel-faces-quota-idUSKBN1H206V.

“May 2021 Crude Steel Production.” *Worldsteel.org*, World Steel Association, 16 Jan. 2022, https://worldsteel.org/media-centre/press-releases/2021/may-2021-crude-steel-production/.

“Public Holidays.” *VisitKorea*, https://english.visitkorea.or.kr/enu/TRV/TV\_ENG\_1\_1.jsp.

Sathishkumar, VE, et al. “Industry Energy Consumption Prediction Using Data Mining Techniques.” *Gvpress*, International Journal of Energy, Information and Communications, 1 Nov. 2020, https://gvpress.com/journals/IJEIC/vol11\_no1/2.pdf.

“Size of States.” *State Symbols USA*, State Symbols USA, <https://statesymbolsusa.org/symbol-official-item/national-us/uncategorized/states-size>.

“South Korea: Country Data and Statistics.” *Worlddata.info*, World Data.info, https://www.worlddata.info/asia/south-korea/index.php.

Swanson, Ana, and Peter Eavis. “Trump Expands Steel Tariffs, Saying They Are Short of AIM.” *The New York Times*, The New York Times, 27 Jan. 2020, https://www.nytimes.com/2020/01/27/business/economy/trump-steel-tariffs.html.

Wamsley, Laurel. “South Korea Shortens 'Inhumanely Long' Workweek.” *NPR*, NPR, 1 Mar. 2018, <https://www.npr.org/sections/thetwo-way/2018/03/01/589895641/south-korea-shortens-inhumanely-long-work-week>.