ETM
ENGINEERING & TECHNOLOGY MANAGEMENT

Load Forecasting for Northwest Natural Gas

Course Title:	Decision Support Systems: Data Warehousing
Course Number:	ETM 538
Instructor:	Daniel Sagalowicz, Mike Freiling
Term:	Winter
Year:	2017

Author(s):
Andrew Dang
Justin Moore
Niguel Morfin
Mohammed Al-Hadi
Maoloud Dabab

ETM Office Use Only
Report Number
Type: Student Team Project Report
Note:

Table of Contents

Contents

Introduction

NW Natural is a publicly traded utility headquartered in Portland, Oregon. The company is a primary distributor of natural gas and serves residential, industrial, and commercial consumers in the Pacific Northwest. This project aims at improving gas load forecasting. Load forecasting is the crucial first step for any planning study. This is an exercise of applying different methods and models on past data such as weather and load consumption data to predict load behavior in the short- and long-term.

Company Background

NW Natural Gas buys natural gas from suppliers in the Western U.S. and Canada and distributes it to residential, commercial, and industrial customers throughout the service territory in Oregon and southwest Washington. NW Natural Gas serves about 720,000 customers in Oregon and southwest Washington. NW Natural builds, maintains, and operates the local natural gas distribution system – that is, the pipes and related equipment that transport natural gas to homes and businesses. In recent years, NW Natural's growth rate has exceeded the national average for local distribution companies. This growth is due to strong customer preference for natural gas for space heating and water heating and the relative cost-efficiency of natural gas.

Problem Statement

NW Natural buys gas from sources in Canada and the rocky mountain regions of the United States. They then distribute the gas to local consumers through their network of interstate, city, and local pipes. NW Natural also has two Liquefied Natural Gas facilities and one underground storage unit. Gas in these facilities is more costly than gas from Canada or the rocky mountain region. Therefore, load forecasting is crucial in helping the company plan for the short-term and long-term efficiently.

Load forecasting is a predictive analysis using past data to predict future load. These factors include time factors such as hours of the day (day/night), day of the week (week day/weekend), or season (spring/summer/fall/winter). Other factors that affect accurate load forecasting include weather conditions such as temperature, humidity, pressure, and wind. Additionally, there is a downward trend in residential consumption of natural gas in the past few years. This may be due to better home insulation, higher efficiency equipment, and/or better technology that help consumers reduce their gas consumption.

Currently, NW Natural does most of their load forecasting manually using spreadsheets. This process is time consuming, error prone and poor in quality. This process does not have any business intelligence analysis behind it. There are many improvement opportunities in this area of the business; including improving forecasting accuracy. Better processes may also help to strengthen the data quality and reporting capability.

Data Acquisition

The data was made available by teammate Andrew since he works at NW Natural and was asked by his manager to take this class to solve this particular problem. The data includes the load consumption from NW Natural from 1/1/2009 to 8/31/2015. Weather related data are downloaded from publicly available weather data such as Weather Underground (www.wunderground.com) and the National Weather Service (www.weather.gov).

Model Formulation

In general, there are three types of forecasting techniques. Extrapolation is a time series method which uses historical data as the basis for estimating future outcomes. The best trend curve is obtained by using regression analysis, then the best estimate may then be obtained by using the equation of the best trend curve. Correlation is an econometric forecasting method in which one would identify the underlying factors that might influence the variable that is being forecast. The outcome of this method depends heavily on the good judgment and experience to make the forecasting method effective. The third technique is using a hybrid method which combines extrapolation and correlation.

Typically, weather has the greatest impact on gas consumption. Primarily, this includes temperature, humidity, and wind. Of those factors, temperature generally has the greatest impact on natural gas load variation. However, temperature and load may not be related linearly. It is further complicated by the influence of humidity, wind speed, and other factors such as pressure and precipitation.

For our own analysis, we tested several of the methods learned in class, and multivariate regression. For our method me finally settled on the regression method as the best predictor, and as R1 as the second best.

R1 Method

Mechanically, the R1 rule built on the sum load as a load level then the several variables: month, weekday, average daily temperature, daily snowfall, daily precipitation, average daily wind speed, and pressure were considered in order to find the min error. When we chose the attribute which gave us the lowest average error, we found that the month was the best predictor Appendix E shows the snapshots for all steps, and the below table gives us the error, about 33%.

Count of Load Level	Column Labels											
Row Labels	(A) very low	(B) low	(C) meduim	(D) high	#N/A	(blank)	Grand Total	Max	Sum	Sum without max	Error	total Error
(A) very low	334	2					336	334	336	2	0.005952	0.336419753
(B) low	88	5					93	88	93	5	0.053763	
(C) meduim	62	85	5				152	85	152	67	0.440789	
(D) high	18	130	138	105	1		392	138	391	253	0.647059	
Grand Total	**502**	**222**	**143**	**105**	**1**		**973**					

Using month as an attribute gave us a rule for each month. We decided to use a very liberal estimation, using the max load as the point estimate. We then took the difference of the actual load against our estimated load and that gave us an error, the estimate minus the actual. By averaging all of the errors, we got a mean of around 33%. The variance of errors was also very high, we saw errors ranging from around 4% to up to near 50%. We split the data into a training set and a test set, each set had 50% of the data.

As a second step, we decided to do R2 to see how the second attribute will affect. The average temperature was the second attribute, and we ended up with a 23% error rate. It makes sense that the error rate would be reduced, but predicting the temperature far into the future is difficult. We decided to use a high point estimate because it's more important for NW Natural to overestimate how much the load will be than underestimate. Overestimation is a bit more inventory, but underestimation means that there may not be enough gas to keep up with demand.

Bayesian Model

The Bayesian model was developed on a training data (1/1/2009-12/31/2012) by calculating the probability for load levels of each attribute (Precipitation, Speed Level, Temperature Level, Snowfall, Weekday, and Pressure Level) and then by selecting the highest probability of each attribute, a combination which predicts one of several load level ranges (A-D).

Multivariate Regression Model

The regression model to predict total daily load was developed using R. The data, which spans from 1/1/2009-8/31/2015, were split into a training set (1/1/2009-12/31/2012) and a test set (1/1/2013-8/31/2015) that were consistent with the split used in the R1 and Bayesian analysis. Using multivariate regression, a model was developed that considers the impact of the month, weekday, average daily temperature, daily snowfall, daily precipitation, average daily wind speed, and pressure. Next, the test set was loaded into R and the "predict" function was used to use to apply the training model to the test set to see if the model would work with new data. To determine the effectiveness of the model, the average error rate was calculated by comparing the prediction data with the actual daily load in the test set. The R code that was used for this analysis can be found in Appendix A. After comparing the regression methodology to R1 and Bayesian, the analysis was rerun with a random training and test split (70% training and 30% split from 1/1/2009-8/31/2015) using the caTools library in R to determine if time had any significant impact on the original predictions. The R code used for this analysis can be found in Appendix B.

Research Analysis

R1 Method

We built a simple R1 rule based on the data we climate data we were able to get, and used that to give a prediction about what the sum of the load is to be expected. We found that our error rates were the lowest when we used the month as a predictor. So we took the month, and assigned an estimated load value based on that. We took the maximum load values for each individual month and had the month assign that value as the prediction. We found that taking the maximum gave us higher error rates than taking the average would have (taking the average gave us an average error of about 28% instead of 33%), but it also never was a good predictor of extremely cold weather, and it underpredicted the amount of gas that would be needed in the winter. After discussing with Andrew, who had the most industry knowledge, we decided to estimate the maximum loads instead of average, which gave us higher error rates, but also never underpredicted the amount of gas needed. Andrew said that it is a greater sin to under-prepare for the winter and over prepare for the summer than it is to have better forecasts. Knowing which is preferable, we decided to trade precision for security against running out of inventory.

Using month as an attribute gave us a rule for each month. We decided to use a very liberal estimation, using the max load as the point estimate. We then took the difference of the actual load against our estimated load and that gave us an error, the estimate minus the actual. By averaging all of the errors, we got a mean of around 33%. The variance of errors was also very high, we saw errors ranging from around 4% to up to near 50%. We split the data into a training set and a test set, each set had 50% of the data.

In terms of using the all row data, from 2009 to 2015, to see if it will effect in the finding, the table below shows the both error rates. therefore, we concluded that there is no significant change in the results, which improved in the regression model.

	Percipitation	Speed Level	Temp Level	Snow Fall	Pressure	Month	Weekday
Error rates 2009-2012	0.4606	0.457	0.4213	0.4797	0.4789	0.2838	0.4825
Error rates for all Years	0.4745	0.4564	0.3784	0.4823	0.4774	0.2814	0.484

Bayesian Model

The analysis of the bayesian model has introduced a predictive model which is specifically for training data. And the results was showing that a combination of highest probability of all attributes and shows a prediction of each load level.

		Attributes						
		Precipitation	Speed Level	Temp Level	Pressure	Month	Weekday	Snow Fall
Load Level	A	Very Low (0.8328)	Low (0.5570)	Medium (0.6100)	Medium (0.4880)	7 (0.1644)	Sat (0.1511)	M (0.7506)
	B	Very Low (0.5695)	Medium (0.3430)	Low (0.9935)	High (0.4012)	3 (0.2233)	Sat (0.1618)	M (0.9482)
	C	Very Low (0.5580)	High (0.3258)	Low (0.9962)	Medium (0.4082)	2 (0.2434)	Thu (0.1685)	M (0.8988)
	D	Very Low (0.7480)	Low (0.5570)	Low (0.7251)	High (0.6030)	12 (0.4045)	Wed (0.1679)	M (0.8625)

We found the Bayesian method complicated and also not very good for predicting. It seemed to work better for specific conditions, but not all that useful for general situations (ie. an entire month). We believe this would not be useful in decision making over long periods of time, and so may not be the best choice as decision support. The only advantage of using the Bayesian model over other methods that if we have the time attribute and that was not available in the data.

However, that does not mean that the Bayesian method is useless. It may give interesting probabilistic information which may be useful for other problems that NW Natural is dealing with. We will recommend in the future research section that the method be developed further to see if there is actually useful information to be gained from it. However, we think it may not be the best fit for this particular application.

Multivariate Regression Model

The multivariate regression model built in R considers the impact of several variables: month, weekday, average daily temperature, daily snowfall, daily precipitation, average daily wind speed, and pressure. Using the "lm()" function with the training set (see Appendix A for the full R code), the model below was developed, yielding the coefficient values included in Appendix B.

```
sumFit <- lm(Sum.of.Load ~
Mo+Week.Day+Tavg+SnowFall+PrecipTotal+AvgSpeed+StnPressure, data=dataReduced)
```

None of the numeric variables (Tavg, SnowFall, PrecipTotal, StnPressure, and AvgSpeed) were strongly correlated with each other (if the correlation approaches -1 or +1), which was calculated in R to yield the results below:

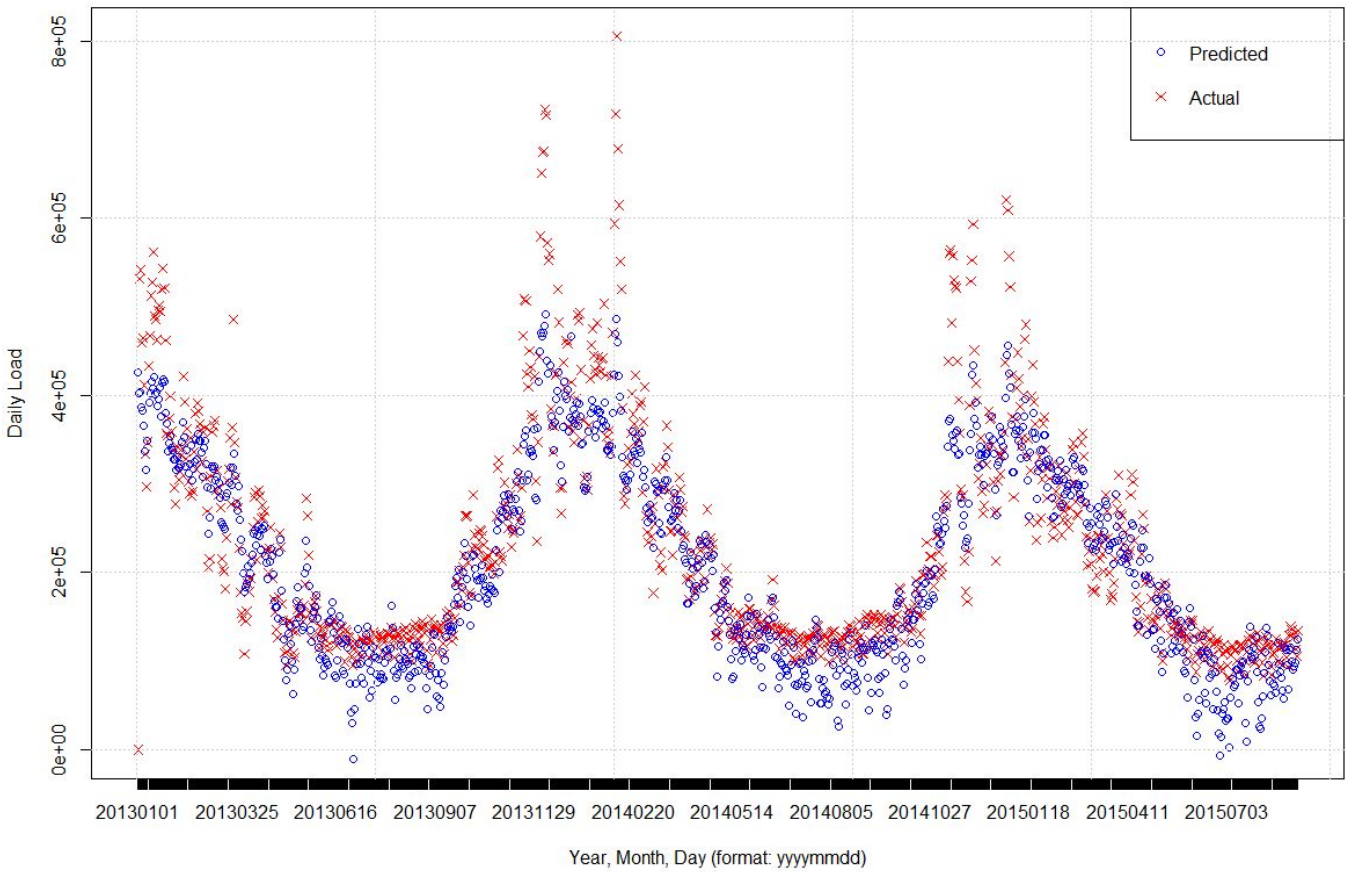
cor(dataReduced[,6:10])

	Tavg	SnowFall	PrecipTotal	StnPressure	AvgSpeed
Tavg	1.0000000000	-0.025217336	-0.06984538	0.06157219	-0.163928802
SnowFall	-0.02521734	1.0000000000	0.13850505	-0.07318501	-0.000727013
PrecipTotal	-0.06984538	0.1385050537	1.0000000000	-0.4744505	0.2133501579
StnPressure	0.06157219	-0.073185007	-0.4744505	1.0000000000	-0.132106949
AvgSpeed	-0.1639288	-0.000727013	0.21335016	-0.13210695	1.0000000000

The coefficient table in Appendix B indicates that the model has an R^2 value of 0.85, and the variables for month (specifically, March-November), day of the week (Saturday and Sunday), average daily temperature, and pressure are all significant, with $p <= 0.001$. The two factor variables, month and day of the week, are compared relative to January and Friday, so January and Friday don't show up in the coefficient table. Next, the "predict()" function was used to compare the training model results to the test set daily load using the following command (see Appendix A for the full R code):

prediction <- predict(sumFit, type="response", newdata=testCleaned)

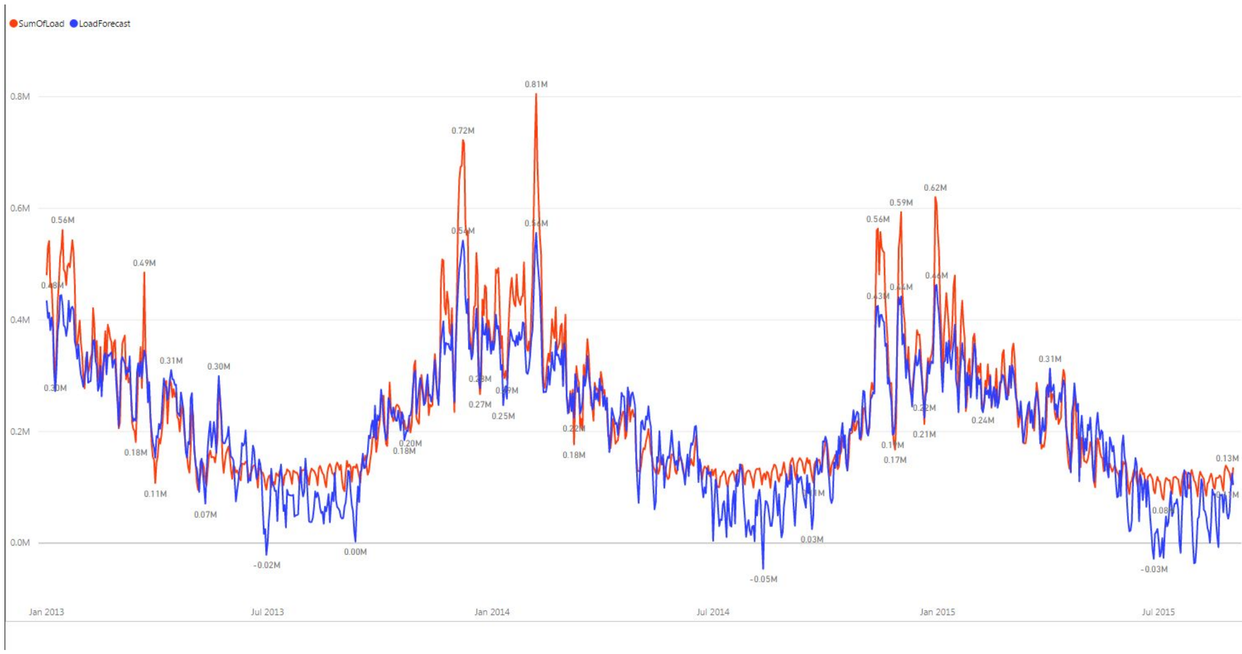
The following plot shows the comparison between the prediction estimates (blue) and the actual daily load values from the test set (red):



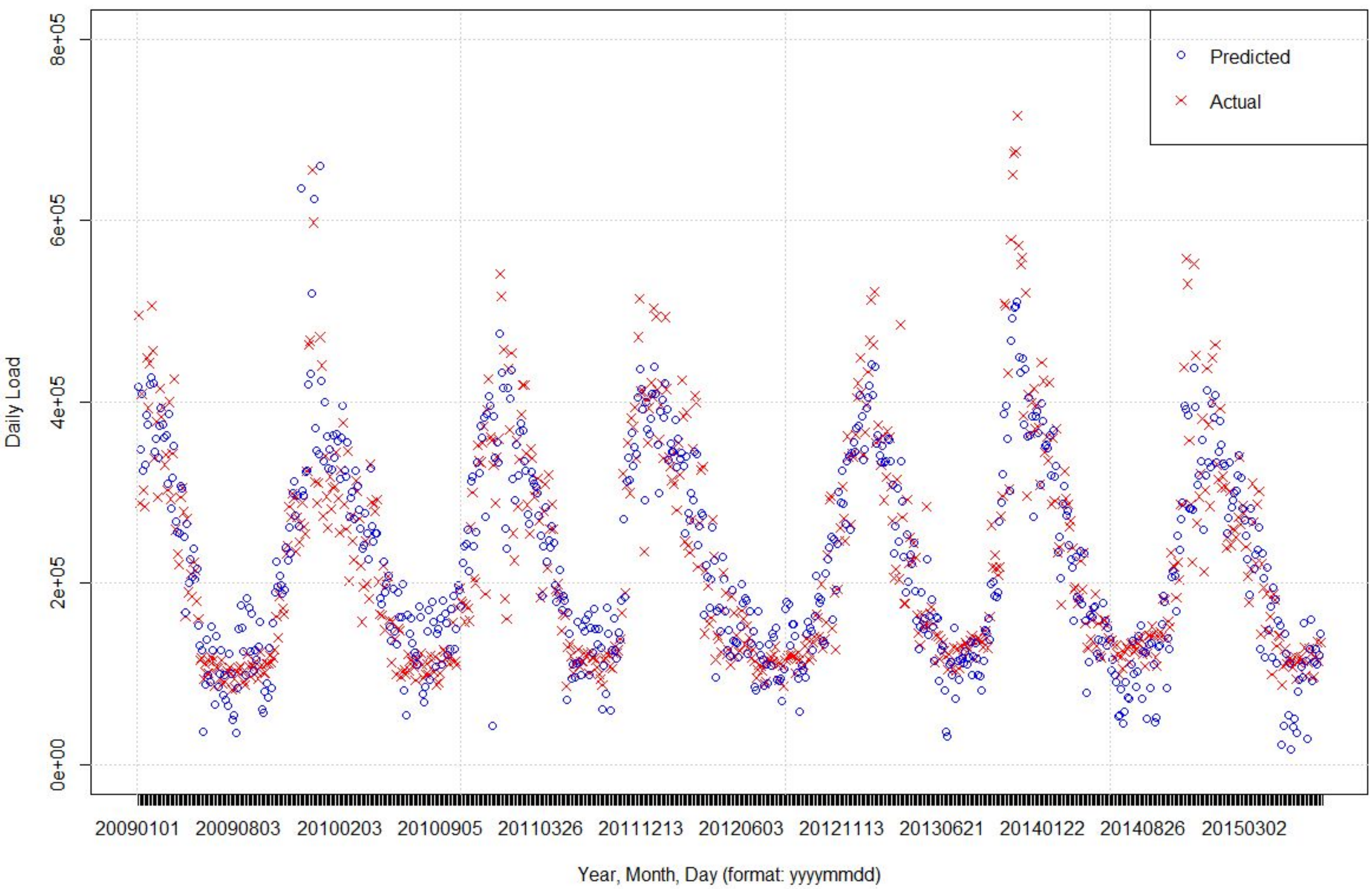
Finally, the last step was to calculate the average error rate of the model to compare its effectiveness to the R1 model. This was done using the following equation (included in Appendix A), which calculated an average error rate of 17.16%. Note that the first record was excluded because gas load data was missing for that day:

```
mean(abs((prediction[2:973]-testCleaned$Sum.of.Load[2:973])/testCleaned$Sum.of.Load[2:973]))
```

We ran our regression analysis independently using both R and Microsoft Power BI for comparison. Below is the output from Microsoft Power BI::



Since we had used a non-random training and test set to compare the R1, Bayesian, and regression analyses, we decided to rerun the regression analysis using a random training and test split (70% training and 30% test from 1/1/2009-8/31/2015) since it had the lowest error rate to see if time had a significant impact on the output (for instance, if customers are using less gas on average over time). Using the R package "caTools" to create a random training and test split (code in Appendix F), the model yielded the coefficients in Appendix G (R^2=0.8449) and had an error rate of 17.47% (compared to 17.16% for the original model). Although the error rate is similar, the significant coefficients changed slightly, with Wednesday, average daily wind speed, and average daily precipitation being significant variables in addition to those that were significant in the previous model. The following plot shows the comparison between the prediction estimates (blue) and the actual daily load values from the test set (red):



Discussion

The R1 does not give a bad predictor. Against the test data, it does a decent job of predicting the load rates, only giving an error rate of about 33%. This is not a bad heuristic to go on, and it is easy to understand, for example, in January we forecast a load of about 571,000. This is both easy to communicate and easy for decision makers to use as a general rule of thumb, but there is a lot of variance in the decisions, which is not ideal.

While the simplicity of R1 is very nice to have, regression analysis is not too complex as to be prohibitive to people in business management. In addition, doing regression cut the error rate in half when compared with R1. We tried performing a Bayesian analysis, but we were unable to calculate a total error rate for the entire model, although it worked well for predicting individual cases. In addition, the Bayesian analysis is more complex and more difficult for most business users to communicate and understand (much less make decisions on), compared to regression, which is pretty well known in the industry and most people can understand, even if they don’t always know how the model is created. Since this is a Decision Support Systems class, we decided it was important to balance predictive value and simplicity, hence regression.

It is very interesting to learn that we independently conducted two different regression analyses on two different platform. The results of these two studies came out about the same. Please see the two charts above for this demonstration.

Our model has predicted the gas usage right on par with the actual load on the test data set with an error rate of about 17%, and so we are pretty satisfied. There are some areas that the gaps are little higher than normal. This is just the matter of fine-tuning the model to narrow down the gaps.

Future Research

We did not conduct any studies on seasonality. One of the suggested future research would be conducting a study on the affecting of gas load on different seasons (summer vs. fall vs. winter vs. spring). We do see the trends over the year but we did not do any study on this area.

Another suggested future research is modeling gas consumption on an hourly basis during the day for weekdays and weekends. NW Natural could use this information to help meet the demand from consumers while minimize the cost of purchasing the gas. This will help NW Natural plan better for their customers' gas consumption. Due to the limited timing of this project, we could not acquire any weather data on the hourly granularity yet.

The model currently looks at the gas load as a whole and does not break it down into different segments. This brings an opportunity to model the gas load at the different segments of the customers such as residential, industrial, or commercial. It will also provide more details on interruptible customer and uninterruptible customer where NW Natural could interrupt service for some customers to meet the demand for other customers if needed.

One last future research recommended for NW Natural is to study the gas theft. This is an on-going issue for the company that they want to get some insights into it. By studying the loads and consumptions at gate stations and at individual service locations, NW Natural could prevent gas theft.

Appendix

Appendix A: R code used for the regression analysis.

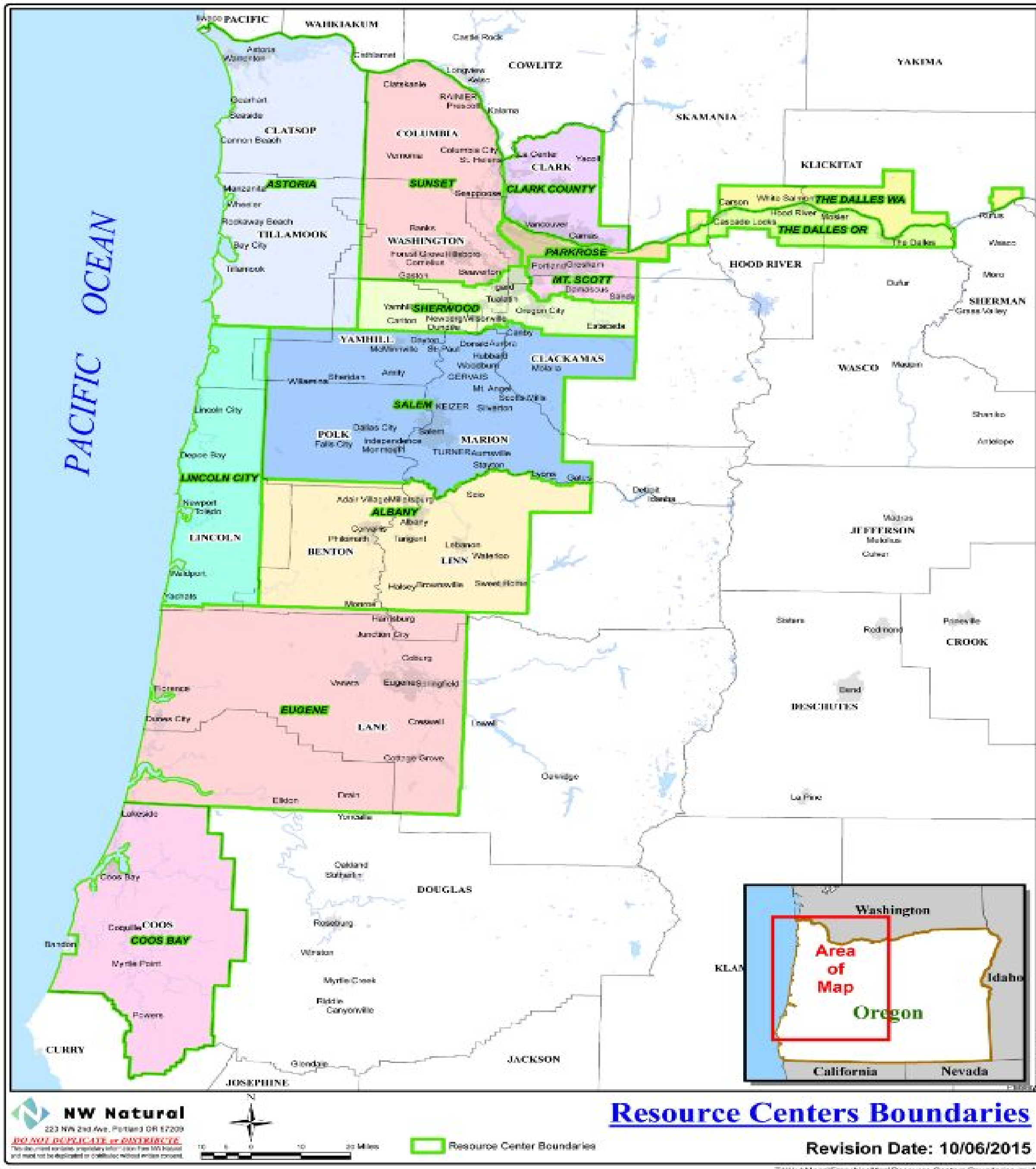
```
#Import data, display first six rows, and show variable definitions
#In Excel, removed #N/A values and "M" values (appear to be "Misssing") and replaced with
blank
dataset <- read.csv("./ETM538_jm_trainingset.csv")
head(dataset)
str(dataset)

#Remove columns WBAN, Year, Day, Date
dataReduced <- subset(dataset, select = -c(1,3,5,6))
str(dataReduced)

#Convert Tavg, SnowFall, PrecipTotal, StnPressure, and AvgSpeed to numeric; convert Mo to
factor
#(otherwise lm() function will check significance of each value/factor of those variables, keep
Week.Day as factor
#since each state in day of week could be significant)
dataReduced$Tavg <- as.numeric(as.character(dataReduced$Tavg))
dataReduced$StnPressure <- as.numeric(as.character(dataReduced$StnPressure))
dataReduced$AvgSpeed <- as.numeric(as.character(dataReduced$AvgSpeed))
dataReduced$Mo <- as.factor(dataReduced$Mo)

#Since SnowFall and PrecipTotal have a factor "  T" and "M", presumably for "Trace", which is
typically <0.1 inches.
#These values need to be replaced by 0 to convert to numeric
dataReduced$SnowFall[dataReduced$SnowFall == "  T"] <- "0"
dataReduced$SnowFall[dataReduced$PrecipTotal == "  T"] <- "0"
dataReduced$SnowFall <- as.numeric(as.character(dataReduced$SnowFall))
dataReduced$PrecipTotal <- as.numeric(as.character(dataReduced$PrecipTotal))
dataReduced[is.na(dataReduced)] <- 0
str(dataReduced)

#Initial linear model using Sum.of.Load
sumFit <- lm(Sum.of.Load ~
Mo+Week.Day+Tavg+SnowFall+PrecipTotal+AvgSpeed+StnPressure, data=dataReduced)
summary(sumFit)
SSE_sum <- sum(sumFit$residuals^2)
RMSE_sum <- sqrt(SSE_sum/nrow(dataReduced))
SSE_sum
RMSE_sum

#Residual plots
```

```
par(mfrow=c(2,2))
plot(sumFit)

#Import test set
testset <- read.csv("./ETM538_jm_testset.csv")

#Clean test set
testCleaned <- subset(testset, select = -c(1,3,5,6))
testCleaned$Tmax <- as.numeric(as.character(testCleaned$Tmax))
testCleaned$Tmin <- as.numeric(as.character(testCleaned$Tmin))
testCleaned$Tavg <- as.numeric(as.character(testCleaned$Tavg))
testCleaned$StnPressure <- as.numeric(as.character(testCleaned$StnPressure))
testCleaned$AvgSpeed <- as.numeric(as.character(testCleaned$AvgSpeed))
testCleaned$Mo <- as.factor(testCleaned$Mo)
testCleaned$SnowFall[testCleaned$SnowFall == "  T"] <- "0"
testCleaned$SnowFall[testCleaned$PrecipTotal == "  T"] <- "0"
testCleaned$SnowFall <- as.numeric(as.character(testCleaned$SnowFall))
testCleaned$PrecipTotal <- as.numeric(as.character(testCleaned$PrecipTotal))
testCleaned[is.na(testCleaned)] <- 0
str(testCleaned)

#Predict using the Sum.of.Load model
prediction <- predict(sumFit, type="response", newdata=testCleaned)
predictConf <- predict(sumFit, newdata=testCleaned, interval='confidence')

#Calculate R^2 of prediction and RMSE
SSE <- sum((prediction - testCleaned$Sum.of.Load)^2)
SST <- sum((mean(dataReduced$Sum.of.Load) - testCleaned$Sum.of.Load)^2)
R2 <- 1 - SSE/SST
RMSE <- sqrt(SSE/nrow(testCleaned))
R2
RMSE

#Plot results
par(mfrow=c(1,1))
plot(testCleaned$Sum.of.Load, col="red", xlab="Year, Month, Day (format: yyyymmdd)",
                                        ylab="Daily Load", xaxt="n")
points(prediction, col="blue", xlab="Year, Month, Day (format: yyyymmdd)", ylab="Daily Load")
axis(1, at=1:973, labels=testCleaned$YearMonthDay)
legend(x="topright", c("Predicted","Actual"), col=c("blue","red"), pch=1)

#Calculate average error rate; excluded row 1 because test case was missing
mean(abs((prediction[2:973]-testCleaned$Sum.of.Load[2:973])/testCleaned$Sum.of.Load[2:973
]))
```

Appendix B: Coefficient values for the multivariate regression model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  459262.2    33695.6  13.630  < 2e-16 ***
Mo2           -9158.5     6047.9  -1.514 0.130159
Mo3          -33419.7     5977.1  -5.591 2.69e-08 ***
Mo4          -78567.2     6247.6 -12.576  < 2e-16 ***
Mo5         -103827.3     6649.7 -15.614  < 2e-16 ***
Mo6         -106367.4     7284.7 -14.602  < 2e-16 ***
Mo7          -88128.1     7989.7 -11.030  < 2e-16 ***
Mo8          -77629.3     8194.7  -9.473  < 2e-16 ***
Mo9          -92526.4     7747.8 -11.942  < 2e-16 ***
Mo10         -90165.7     6521.3 -13.826  < 2e-16 ***
Mo11         -35232.6     6012.8  -5.860 5.74e-09 ***
Mo12           9313.3     5938.9   1.568 0.117056
Week.DayMon   -1392.7     4554.3  -0.306 0.759807
Week.DaySat  -17584.5     4549.5  -3.865 0.000116 ***
Week.DaySun  -26408.9     4548.3  -5.806 7.84e-09 ***
Week.DayThu    2115.4     4561.0   0.464 0.642863
Week.DayTue    2251.2     4557.4   0.494 0.621399
Week.DayWed    5762.9     4555.2   1.265 0.206036
Tavg          -6056.7      198.2 -30.553  < 2e-16 ***
SnowFall     174457.2   116759.5   1.494 0.135354
PrecipTotal    1146.2     1829.0   0.627 0.530962
AvgSpeed        369.4      366.7   1.007 0.313906
StnPressure    5126.7     1087.9   4.712 2.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46430 on 1438 degrees of freedom
Multiple R-squared:  0.8458,	Adjusted R-squared:  0.8435
F-statistic: 358.6 on 22 and 1438 DF,  p-value: < 2.2e-16
```

Appendix C: Northwest Natural Service Areas



Appendix D: Initial Set of Use Cases

System modeling load profiles

- Factors:
  - Customer equipment
  - Time of day
  - Heating degree days
  - Wind
  - Day of the week
  - Synergy tool
- Objective:
  - This could be used for system design to provide better system reinforcement in areas with high peak load.

Peak Day / Peak Hour modeling in Integrated Resource Planning

- Objective: more accurate estimation of peak loads

Residential Load Studies

- Objective: identity driver for lower usage per customer:
  - Better or more efficient equipment?
  - Heat pump?

Daily supply planning

- Objective: Increased accuracy in daily planning (purchase gas, supply gas)
  - Gas control
  - Nomination or allocation gas
  - Historically, this is driven by SMEs with past experience (need to be data driven)

Gas theft

- Objective: identify and prevent gas theft.

End of month usage estimates

- Objective: Increase accuracy in monthly usage from customer.

Appendix E: Data and Pivot tables of R1:

WBAN	Year Month Day	Year	Mo	Day	Date	Week Day	Tmax	Tmin	Tavg	Tem Level	SnowFall	Precip Total	Precip Level	Stn Pressure	Pressure Level	Avg Speed	Speed Level	Sum of Load	Average of Load	Min of Load	Max of Load	Load Level
24229	20120101	2012	1	1	1/1/2012	Sun	51	34	43	Cold	0	0	Very Low	30.06	High	13.4	High					
24229	20120102	2012	1	2	1/2/2012	Mon	54	36	45	Cold	M	T		29.99	Meduim	10.6	High	338053.976	14085.58232	7605.63661	20619.0513	(C) meduim
24229	20120103	2012	1	3	1/3/2012	Tue	50	43	47	Cold	M	T		30.17	High	9.6	Meduim	279347.836	11639.49316	7144.74828	17667.2373	(B) low
24229	20120104	2012	1	4	1/4/2012	Wed	54	42	48	Cold	M	0.16	Low	30.1	High	9.2	Meduim	306936.698	12789.02909	7530.83561	19922.8983	(C) meduim
24229	20120105	2012	1	5	1/5/2012	Thu	51	34	43	Cold	M	0.02		30.29	High	4.7	Low	313385.194	13057.7164	7348.74261	18314.0936	(C) meduim
24229	20120106	2012	1	6	1/6/2012	Fri	39	31	35	Very Cold	M	0.12	Low	30.23	High	3.8	Low	454613.587	18942.23277	12581.2643	26134.9026	(D) high
24229	20120107	2012	1	7	1/7/2012	Sat	47	34	41	Cold	M	T		30.33	High	3.9	Low	355147.607	14797.81694	11736.9213	19803.9026	(C) meduim
24229	20120108	2012	1	8	1/8/2012	Sun	45	34	40	Cold	0	0	Very Low	30.25	High	1.6	Very Low	393306.729	16387.78038	12373.3823	23094.7353	(C) meduim
24229	20120109	2012	1	9	1/9/2012	Mon	43	35	39	Cold	M	0.3	Low	30.18	High	1.5	Very Low	421044.893	17543.5372	10690.4186	25870.1956	(D) high
24229	20120110	2012	1	10	1/10/2012	Tue	45	29	37	Cold	M	0.02		30.33	High	1.7	Very Low	405403.163	16891.79844	10364.9326	24348.9983	(D) high
24229	20120111	2012	1	11	1/11/2012	Wed	46	29	38	Cold	M	0	Very Low	30.27	High	13.4	High	474486.808	19770.28368	14298.0433	29400.5716	(D) high
24229	20120112	2012	1	12	1/12/2012	Thu	44	27	36	Cold	0	0	Very Low	30.2	High	4.8	Low	488206.619	20341.94245	14413.6943	31353.1766	(D) high
24229	20120113	2012	1	13	1/13/2012	Fri	43	26	35	Very Cold	0	0	Very Low	30.11	High	4	Low	473666.27	19736.09459	14316.2126	30145.6809	(D) high
24229	20120114	2012	1	14	1/14/2012	Sat	45	33	39	Cold	M	0.13	Low	29.96	Meduim	7.5	Meduim	426943.355	17789.30645	13641.9979	22958.8519	(D) high
24229	20120115	2012	1	15	1/15/2012	Sun	38	30	34	Very Cold	M	T		29.95	Meduim	6.5	Meduim	446517.015	18604.87562	13321.3443	23835.2413	(D) high
24229	20120116	2012	1	16	1/16/2012	Mon	38	29	34	Very Cold	M	0.08		30.05	High	11	High	503293.785	20970.57437	14102.7076	26305.3226	(D) high
24229	20120117	2012	1	17	1/17/2012	Tue	40	32	36	Cold	M	0.59	Low	29.86	Meduim	9.9	Meduim	494254.11	20593.92123	13825.4276	27456.5616	(D) high
24229	20120118	2012	1	18	1/18/2012	Wed	53	32	43	Cold	M	0.99	Low	29.6	Meduim	10.9	High	413933.847	17247.24361	8729.89995	23586.1043	(D) high
24229	20120119	2012	1	19	1/19/2012	Thu	53	38	46	Cold	M	1.91	Meduim	29.51	Meduim	9	Meduim	358134.247	14922.26029	8021.40527	19213.8483	(C) meduim
24229	20120120	2012	1	20	1/20/2012	Fri	41	36	39	Cold	M	0.6	Low	29.37	Low	16.7	High	396275.496	16511.47901	11478.9383	21355.3869	(C) meduim

The Training Data

Or high	Load Level
50000	(A) very low
200000	(B) low
300000	(C) meduim
400000	(D) high

Or high	Temp Level
75	Hot
60	Warm
35	Cold
0	Very Cold

Or high	Speed Level
10.1	Hot
6.1	Meduim
3.1	Low
0	Very Low

Or high	Precip Level
2	Hot
1	Meduim
0.1	Low
0	Very Low

Or high	Pressure Level
30	Hot
29.5	Meduim
0	Low

Month	Av Load
1	366992.158
2	339995.55
3	296374.554
4	221581.788
5	160416.819
6	125167.168
7	105740.392
8	107870.495
9	116002.547
10	182917.398
11	294515.588
12	389336.184

Month	Max Load
1	571312.285
2	573512.43
3	439543.442
4	356649.95
5	274152.236
6	180556.176
7	130202.178
8	134532.018
9	163746.214
10	321013.904
11	616657.953
12	656113.487

The References for Vlookup

Max	Sum	Sum without max	Error	total Error
46	120	74	0.616667	0.283849
65	113	48	0.424779	
69	124	55	0.443548	
65	120	55	0.458333	
106	124	18	0.145161	
120	120	0	0	
124	124	0	0	
124	124	0	0	
120	120	0	0	
89	124	35	0.282258	
57	120	63	0.525	
57	122	65	0.532787	

The Pivot Table for Month

Max	Sum	Sum without max	Error	total Error
2	2	0	0	0.460648
123	384	261	0.679688	
7	20	13	0.65	
334	458	124	0.270742	

The Pivot Table for Precip

Max	Sum	Sum without max	Error	total Error
173	478	305	0.638075	0.475895
12	29	17	0.586207	
576	945	369	0.390476	

The Pivot Table for Pressure

Max	Sum	Sum without max	Error	total Error
188	244	56	0.229508	0.4797251
1	1	0	0	
1	1	0	0	
2	2	0	0	
565	1207	642	0.531897	

The Pivot Table for Snow

Max	Sum	Sum without max	Error	total Error
86	239	153	0.640167	0.450724
420	605	185	0.305785	
232	443	211	0.476298	
59	164	105	0.640244	

The Pivot Table for Speed

Max	Sum	Sum without max	Error	total Error
305	915	610	0.666667	0.421271
48	48	0	0	
46	46	0	0	
439	439	0	0	

The Pivot Table for Temperature

Max	Sum	Sum without max	Error	total Error
120	207	87	0.42029	0.482474227
107	208	101	0.485577	
107	208	101	0.485577	
102	208	106	0.509615	
96	208	112	0.538462	
107	208	101	0.485577	
114	208	94	0.451923	

The Pivot Table for Weekday

BAN	Year Month Day	Year	Mo	Day	Date	Week Day	Tmax	Tmin	Tavg	Snow Fall	Precip Total	Stn Pressure	Avg Speed	Sum of Load	Average of Load	Min of Load	Max of Load	Load Level	Estimation Load	New Load Level
4229	20150801	2015	8	1	8/1/2015	Sat	98	62	80	0	0	29.78	7.6	98192.27	4091.34455	2839.9889	5143.6369	(A) very low	107870.4952	(A) very low
4229	20150802	2015	8	2	8/2/2015	Sun	80	67	74	0	T	29.78	3.4	83571.08	3482.12831	1376.5843	4832.3299	(A) very low	107870.4952	(A) very low
4229	20150803	2015	8	3	8/3/2015	Mon	84	64	74	0	T	29.81	5.4	115898.5	4829.10237	2352.9509	6595.1329	(A) very low	107870.4952	(A) very low
4229	20150804	2015	8	4	8/4/2015	Tue	83	59	71	0	0	29.89	8.6	128081.6	5336.73517	3804.4763	6872.2423	(A) very low	107870.4952	(A) very low
4229	20150805	2015	8	5	8/5/2015	Wed	78	56	67	0	0	29.98	7.5	122567	5106.96037	3816.6596	6495.6946	(A) very low	107870.4952	(A) very low
4229	20150806	2015	8	6	8/6/2015	Thu	79	56	68	0	0	29.95	7.1	118822.7	4950.94608	3366.7173	6667.1766	(A) very low	107870.4952	(A) very low
4229	20150807	2015	8	7	8/7/2015	Fri	87	59	73	0	0	29.79	7.8	114901.2	4787.55107	3454.5539	6470.9466	(A) very low	107870.4952	(A) very low
4229	20150808	2015	8	8	8/8/2015	Sat	82	63	73	0	0	29.8	5.8	95491.33	3978.80544	2378.4266	5513.2756	(A) very low	107870.4952	(A) very low
4229	20150809	2015	8	9	8/9/2015	Sun	86	61	74	0	0	29.83	5.7	84888.9	3537.03768	1849.0079	4714.4143	(A) very low	107870.4952	(A) very low
4229	20150810	2015	8	10	8/10/2015	Mon	87	65	76	0	0	29.79	4.6	106982.9	4457.61949	2141.8573	6209.9906	(A) very low	107870.4952	(A) very low
4229	20150811	2015	8	11	8/11/2015	Tue	91	63	77	0	0	29.77	5.8	109839.4	4576.643	2886.2329	6071.3436	(A) very low	107870.4952	(A) very low
4229	20150812	2015	8	12	8/12/2015	Wed	91	66	79	0	0	29.84	5.3	118669.2	4944.55164	3005.0806	6571.4253	(A) very low	107870.4952	(A) very low
4229	20150813	2015	8	13	8/13/2015	Thu	86	64	75	0	0	29.86	5.7	124444.7	5185.19704	3588.0339	6290.9816	(A) very low	107870.4952	(A) very low
4229	20150814	2015	8	14	8/14/2015	Fri	74	64	69	0	0.12	29.98	8	115631.1	4817.96128	3318.9893	6475.9649	(A) very low	107870.4952	(A) very low
4229	20150815	2015	8	15	8/15/2015	Sat	77	63	70	0	0	30.1	7.3	100386.3	4182.76158	3047.0129	5353.9533	(A) very low	107870.4952	(A) very low
4229	20150816	2015	8	16	8/16/2015	Sun	83	56	70	0	0	30.03	8.2	97339.93	4055.83042	2962.5686	5475.8993	(A) very low	107870.4952	(A) very low

Test Data Using the Average

Max	Sum	Sum without max	Error	total Error
451	491	40	0.081466	0.286008
137	243	106	0.436214	
106	238	132	0.554622	

The Error of the results for Average

WBAN	Year Month Day	Year	Mo	Day	Date	Week Day	Tmax	Tmin	Tavg	Snow Fall	Precip Total	Stn Pressure	Avg Speed	Sum of Load	Average of Load	Min of Load	Max of Load	Load Level	Estimation Load	New Load Level
24229	20150801	2015	8	1	8/1/2015	Sat	98	62	80	0	0	29.78	7.6	98192.27	4091.34455	2839.9889	5143.6369	(A) very low	134532.0184	(A) very low
24229	20150802	2015	8	2	8/2/2015	Sun	80	67	74	0	T	29.78	3.4	83571.08	3482.12831	1376.5843	4832.3299	(A) very low	134532.0184	(A) very low
24229	20150803	2015	8	3	8/3/2015	Mon	84	64	74	0	T	29.81	5.4	115898.5	4829.10237	2352.9509	6595.1329	(A) very low	134532.0184	(A) very low
24229	20150804	2015	8	4	8/4/2015	Tue	83	59	71	0	0	29.89	8.6	128081.6	5336.73517	3804.4763	6872.2423	(A) very low	134532.0184	(A) very low
24229	20150805	2015	8	5	8/5/2015	Wed	78	56	67	0	0	29.98	7.5	122567	5106.96037	3816.6596	6495.6946	(A) very low	134532.0184	(A) very low
24229	20150806	2015	8	6	8/6/2015	Thu	79	56	68	0	0	29.95	7.1	118822.7	4950.94608	3366.7173	6667.1766	(A) very low	134532.0184	(A) very low
24229	20150807	2015	8	7	8/7/2015	Fri	87	59	73	0	0	29.79	7.8	114901.2	4787.55107	3454.5539	6470.9466	(A) very low	134532.0184	(A) very low
24229	20150808	2015	8	8	8/8/2015	Sat	82	63	73	0	0	29.8	5.8	95491.33	3978.80544	2378.4266	5513.2756	(A) very low	134532.0184	(A) very low
24229	20150809	2015	8	9	8/9/2015	Sun	86	61	74	0	0	29.83	5.7	84888.9	3537.03768	1849.0079	4714.4143	(A) very low	134532.0184	(A) very low
24229	20150810	2015	8	10	8/10/2015	Mon	87	65	76	0	0	29.79	4.6	106982.9	4457.61949	2141.8573	6209.9906	(A) very low	134532.0184	(A) very low
24229	20150811	2015	8	11	8/11/2015	Tue	91	63	77	0	0	29.77	5.8	109839.4	4576.643	2886.2329	6071.3436	(A) very low	134532.0184	(A) very low
24229	20150812	2015	8	12	8/12/2015	Wed	91	66	79	0	0	29.84	5.3	118669.2	4944.55164	3005.0806	6571.4253	(A) very low	134532.0184	(A) very low
24229	20150813	2015	8	13	8/13/2015	Thu	86	64	75	0	0	29.86	5.7	124444.7	5185.19704	3588.0339	6290.9816	(A) very low	134532.0184	(A) very low
24229	20150814	2015	8	14	8/14/2015	Fri	74	64	69	0	0.12	29.98	8	115631.1	4817.96128	3318.9893	6475.9649	(A) very low	134532.0184	(A) very low
24229	20150815	2015	8	15	8/15/2015	Sat	77	63	70	0	0	30.1	7.3	100386.3	4182.76158	3047.0129	5353.9533	(A) very low	134532.0184	(A) very low
24229	20150816	2015	8	16	8/16/2015	Sun	83	56	70	0	0	30.03	8.2	97339.93	4055.83042	2962.5686	5475.8993	(A) very low	134532.0184	(A) very low
24229	20150817	2015	8	17	8/17/2015	Mon	89	60	75	0	0	29.91	6.9	116648.8	4860.36666	3081.9493	6606.8236	(A) very low	134532.0184	(A) very low
24229	20150818	2015	8	18	8/18/2015	Tue	96	60	78	0	0	29.8	6	116798.3	4866.59482	3151.6109	6817.3496	(A) very low	134532.0184	(A) very low

Test Data Using Max

Max	Sum	Sum without max	Error	total Error
334	336	2	0.005952	0.336419753
88	93	5	0.053763	
85	152	67	0.440789	
138	391	253	0.647059	

The Error of the results for Average

Appendix F: R code used for regression analysis with random training/test split.

```
#Import data, display first six rows, and show variable definitions
#In Excel, removed #N/A values and "M" values (appear to be "Misssing") and replaced with
blank
dataset <- read.csv("./ETM538_jm_data.csv")
library(caTools)
set.seed(18274)
head(dataset)
str(dataset)

#Remove columns WBAN, Year, Day, Date
dataReduced <- subset(dataset, select = -c(1,3,5,6))
str(dataReduced)

#Convert Tavg, SnowFall, PrecipTotal, StnPressure, and AvgSpeed to numeric; convert Mo to
factor
#(otherwise lm() function will check significance of each value/factor of those variables, keep
Week.Day as factor
#since each state in day of week could be significant)
dataReduced$Tavg <- as.numeric(as.character(dataReduced$Tavg))
dataReduced$StnPressure <- as.numeric(as.character(dataReduced$StnPressure))
dataReduced$AvgSpeed <- as.numeric(as.character(dataReduced$AvgSpeed))
dataReduced$Mo <- as.factor(dataReduced$Mo)

#Since SnowFall and PrecipTotal have a factor "  T" and "M", presumably for "Trace", which is
typically <0.1 inches.
```

```
#These values need to be replaced by 0 to convert to numeric
dataReduced$SnowFall[dataReduced$SnowFall == "  T"] <- "0"
dataReduced$SnowFall[dataReduced$PrecipTotal == "  T"] <- "0"
dataReduced$SnowFall <- as.numeric(as.character(dataReduced$SnowFall))
dataReduced$PrecipTotal <- as.numeric(as.character(dataReduced$PrecipTotal))
dataReduced[is.na(dataReduced)] <- 0
str(dataReduced)

#Split into a random training and test set using sample.split() from caTools (70%/30%)
split <- sample.split(dataReduced$Sum.of.Load, SplitRatio = 0.7)
training <- subset(dataReduced, split==TRUE)
test <- subset(dataReduced, split==FALSE)
nrow(training)
nrow(test)

#Initial linear model using Sum.of.Load
sumFit <- lm(Sum.of.Load ~
Mo+Week.Day+Tavg+SnowFall+PrecipTotal+AvgSpeed+StnPressure, data=training)
summary(sumFit)
SSE_sum <- sum(sumFit$residuals^2)
RMSE_sum <- sqrt(SSE_sum/nrow(training))
SSE_sum
RMSE_sum

#Residual plots
par(mfrow=c(2,2))
plot(sumFit)

#Predict using the Sum.of.Load model
prediction <- predict(sumFit, type="response", newdata=test)
predictConf <- predict(sumFit, newdata=test, interval='confidence')

#Calculate R^2 of prediction and RMSE
SSE <- sum((prediction - test$Sum.of.Load)^2)
SST <- sum((mean(test$Sum.of.Load) - test$Sum.of.Load)^2)
R2 <- 1 - SSE/SST
RMSE <- sqrt(SSE/nrow(test))
R2
RMSE
```

```
#Plot results; excluded row 1 because test case was missing
par(mfrow=c(1,1))
plot(test$Sum.of.Load[2:731], col="red", xlab="Year, Month, Day (format: yyyymmdd)",
                    ylab="Daily Load", xaxt="n", pch=4, ylim=c(0,800000))
points(prediction[2:731], col="blue", xlab="Year, Month, Day (format: yyyymmdd)",
                                        ylab="Daily Load")
axis(1, at=1:731, labels=test$YearMonthDay)
legend(x="topright", c("Predicted","Actual"), col=c("blue","red"), pch=c(1,4))
grid()

#Calculate average error rate; excluded row 1 because test case was missing
mean(abs((prediction[c(2:731)]-test$Sum.of.Load[c(2:731)])/test$Sum.of.Load[c(2:731)]))
```

Appendix G: Coefficient for the multivariate regression model with random training/test split.

```
Coefficients: (1 not defined because of singularities)
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 276490.6    46977.8   5.886 4.78e-09 ***
Mo2         -8602.4     5591.8  -1.538   0.1241
Mo3        -45012.0     5603.2  -8.033 1.77e-15 ***
Mo4        -82495.7     5875.9 -14.040  < 2e-16 ***
Mo5        -97703.2     6383.6 -15.305  < 2e-16 ***
Mo6        -90031.4     6984.3 -12.890  < 2e-16 ***
Mo7        -66307.0     7768.2  -8.536  < 2e-16 ***
Mo8        -51835.9     8061.4  -6.430 1.66e-10 ***
Mo9        -74967.5     7563.0  -9.912  < 2e-16 ***
Mo10       -81747.0     6274.0 -13.029  < 2e-16 ***
Mo11       -28174.1     5801.7  -4.856 1.31e-06 ***
Mo12        -3749.1     5759.5  -0.651   0.5152
Week.DayMon    890.1     4356.3   0.204   0.8381
Week.DaySat -11604.2     4370.3  -2.655   0.0080 **
Week.DaySun -19367.8     4333.6  -4.469 8.37e-06 ***
Week.DayThu   7903.7     4330.6   1.825   0.0682 .
Week.DayTue   3760.6     4384.5   0.858   0.3912
Week.DayWed   9336.7     4360.0   2.141   0.0324 *
Tavg        -6954.8      192.1 -36.201  < 2e-16 ***
SnowFall         NA         NA      NA       NA
PrecipTotal -28860.6     5538.3  -5.211 2.11e-07 ***
AvgSpeed      1834.7      365.1   5.026 5.54e-07 ***
```

StnPressure 12642.8   1568.3   8.061 1.42e-15 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47440 on 1681 degrees of freedom
Multiple R-squared: 0.8468,   Adjusted R-squared: 0.8449
F-statistic: 442.5 on 21 and 1681 DF, p-value: < 2.2e-16

Appendix H: Bayesian Model Probabilities Data

Precipitation:

Label	Percipitaion Level	Load Level	Count	Total	Probability
Very Low|A	Very Low	A	628	754	0.832891247
Very Low|B	Very Low	B	176	309	0.569579288
Very Low|C	Very Low	C	149	267	0.558052434
Very Low|D	Very Low	D	98	131	0.748091603

Wind Speed:

Label	Speed Level	Load Level	Count	Total	Probability
High|C	High	C	87	267	0.325842697
High|D	High	D	33	131	0.251908397
Medium|B	Medium	B	106	309	0.343042071
Medium|D	Medium	D	33	131	0.251908397
Low|A	Low	A	420	754	0.557029178
Low|D	Low	D	33	131	0.251908397

Temperature:

Label	Temp Level	Load Level	Count	Total	Probability
Medium|A	Medium	A	460	754	0.610079576
Low|B	Low	B	307	309	0.993527508
Low|C	Low	C	266	267	0.996254682
Low|D	Low	D	94	131	0.72519084

Pressure:

Label	Pressure Level	Load Level	Count	Total	Probability
High|B	High	B	124	309	0.4012945
High|D	High	D	79	131	0.6030534
Medium|A	Medium	A	577	754	0.4880637
Medium|C	Medium	C	151	267	0.4082397

Month:

Label	Month	Load Level	Count	Total	Probability
7|A	7	A	124	754	0.1644562
8|A	8	A	124	754	0.1644562
3|B	3	B	69	309	0.223301
2|C	2	C	65	267	0.2434457
12|D	12	D	53	131	0.4045802

Weekday:

Label	Weekday	Load Level	Count	Total	Probability
Sat|A	Sat	A	114	754	0.1511936
Sat|B	Sat	B	50	309	0.1618123
Thu|C	Thu	C	45	267	0.1685393
Wed|D	Wed	D	22	131	0.1679389

Snow Fall:

Label	Snow Fall Level	Load Level	Count	Total	Probability
M|A	M	A	0	754	0.75066313
M|B	M	B	293	309	0.948220065
M|C	M	C	240	267	0.898876404
M|D	M	D	113	131	0.86259542

Load Levels:

Load Level	Load Level Count	Total	Probability
A	754	1461	0.5160849
B	309	1461	0.211499
C	267	1461	0.1827515
D	131	1461	0.0896646

Predictive Model:

	Percipitation	Speed Level	Temp Level	Snow Fall	Pressure	Month	Weekday	Load Level	Product (x 10^{-5})	Likelihood (%)
Observation:	Very Low	High	High	M	High	12	Sun			
A	0.832891247	0.05437666	0.0729443	0.7506631	0.229443	0.006631	0.160477	0.516084873	31.24969679	100.0%
B	0.569579288	0.26860841	0	0.9482201	0.401294	0.029126	0.126214	0.211498973	0	0.0%
C	0.558052434	0.3258427	0	0.8988764	0.389513	0.213483	0.131086	0.18275154	0	0.0%
D	0.748091603	0.2519084	0	0.8625954	0.603053	0.40458	0.10687	0.089664613	0	0.0%
TOTAL									31.24969679	100.0%