Predicting Daily Peak Load of ISO New England Region

Team We R
Final Report

Group Members
Debora Maia Silva - 0029989458
Gagandeep Singh Khanuja - 0029971620
Yash Kothari - 0030144260

June 5, 2018

Here's a link to Overleaf document.

1 Abstract

A methodology that relates climate parameters to Peak Demand at county level scales has been applied to six states in the Independent System Operator (ISO) New England region to predict daily peak demand. These models allow for detailed analyses of electricity demand and its vulnerabilities to climate change at county level scale. The results include the comparison between various models and indicate the best model for predicting the daily peak demand significantly.

2 Introduction

Energy. Like water, an essential item for humans, and society loses money, prestige and life quality without it. Most people do not know the complex system behind the fact that their air conditioning is running on the hottest day of the year, and one of the objectives of this study is to make sure they will still be on even on that devastating summer afternoon.

Peak loads or peak demand is the highest demand in energy in a given period of time, usually daily. The electric infrastructure system has to be prepared for this load, even if it only happens in a certain day of the year. This study do not intend to discuss the resilience of the electric system, but it will be working on public available data to show what this load could be based on climate values.

3 Objective

The main objective of this project is to predict the peak load of the region under the New England ISO area by identifying the major parameters that affect the demand of energy consumption.

4 Project Details

The peak load in this project refers to the highest demand of electrical power daily. The power demand depends on various factors like - demography, economy, weather, season, day of the week and more. This project focus on weather inputs, public available from the NOAA website[1].

New England is a northeastern region of the United States comprising the states of Maine, Vermont, New Hampshire, Massachusetts, Connecticut and Rhode Island. The data is provided by ISO New England Inc.- an independent, non-profit Regional Transmission Organization[2]. The dataset contains hourly data of peak demand from 2011 to 2017.

Peak demand may exceed the maximum supply of power, resulting in power outages and load shedding. The peak demand helps decide the size of transformers, generators, transmission lines and circuit breakers, being an important factor when designing the systems and performing risk analysis in resilience. Considering our currently technology in energy generation, power plants based on fossil fuels are faster to get into the grid, and they are usually the ones used to cover peak demand - and at a high price[3]. This is not an environmental friendly actions, but it is what most electric systems can offer currently[4]. The use of the results in this study is to prevent the worst case scenario for any society nowadays: outages.

4.1 Future of this work

It is expected that North America will hold the largest share of the global energy in the coming years due to the presence of large enterprises and technical experts.[5] This will lead to an ever-increasing demand for energy. At the same time the concern for environment-friendly means is driving investment of millions in energy efficient measures by businesses around the globe. Global warming has redefined the perception of the utility and energy industry. With the help of smart grid systems, it is possible to collect the operational and consumption data and further using advanced analytics tools and techniques can help the energy industry to run efficiently. For example, BuildingIQ, a San Francisco-based energy analytics company, has helped save USD 700,000 in the third quarter of 2014.[5]

5 Literature Review

Maximilian Auffhammera, Patrick Baylisc and Catherine H. Hausman in their paper focus on the cost implications due to highest load observed in a period. They analyze multiyear data from 166 load balancing authorities in USA to find relationship between peak demand and average demand with temperature. They find that peak demand is affected by temperature more than average demand. Moreover, the impacts on peak load vary substantially across space, driven by differences in the distribution of heating and cooling degree days as well as differences in heating and cooling technologies.Estimating the response function of average and peak loads to weather, they estimate a set of time series models, one for each load zone and rely on inter day variation in total load or peak load as a function of daily weather to identify the regression coeffs used in simulations.
However, they don't consider the effects due to population density and other climatic factors like station pressure, humidity or wind speed affecting the energy consumption which we intend to consider while estimating the peak load.[6]

David J. Sailor and J. Ricardo Munoz in their paper assess the sensitivity of electricity consumption to climate at regional scales across eight of the most energy intensive states in USA. Two sets of variable types were used – primitive variables such as temperature, relative humidity, and wind speed, and derived variables including cooling degree days, heating degree days, and enthalpy latent days are used, and their advantages and disadvantages are discussed. Linear least-squares regression statistical method was used in this study assuming a linear relationship between the dependent variables (electricity consumption) and the climatic independent variables. The least-squares method gives good predictions of the regression coefficients if the two Gauss-Markov (G-M) conditions meet. Further, two approaches for relating energy consumption to climate have been demonstrated. The primitive variable approach uses data that are directly available from meteorological stations and the degree-day approach involves some significant transformations of the raw data. The degree-day approach is best of the two.[7]
However, the primitive variable approach contains anomalies that result in poor electricity yield results. Also, in such models there can be only one coefficient for dependence of electricity consumption on temperature and such problem is only partially solved by dividing the data into two seasons. Also, the effect of other climatic factors other than temperature haven't been well included in the model and the effect of population on the energy consumption hasn't been considered.

David J. Sailor in his paper in the year 2000 relates the residential and commercial electricity loads to climate while further evaluating state level sensitivities and vulnerabilities. Model sensitivities to climate changes were investigated to estimate electricity demand based on climatic changes. The previously developed methodology was applied to eight energy-intensive states in

diverse geographical locations resulting in monthly aggregated statewide predictive models for electricity consumption. Monthly statewide aggregated variables were combined in the simplest physical significance model structure possible. The electricity consumption models all depend upon temperatures, humidity and wind speed, with temperature being the most important variable. The results show a wide range of electricity demand impacts, with Washington state experiencing decreased loads associated with climate warming, but the other 7 experience a significant increase in annual per capita residential and commercial electricity consumption.[8]
However, the results assume only climatic factor change and do not contain socio-economic data as an influence on energy consumption which we intend to study. Further, it should be noted that the long-term base values for degree day calculations may also shift due to populations acclimating to the warmer climates. Such analysis is beyond the scope of this paper.

A.C. Menezes, A. Cripps, R.A. Buswell, J. Wright, D. Bouchlaghem in their paper build two models to estimate small power consumption in office buildings along with typical power demand profiles. The first model depends entirely on random sampling of monitored data, and the second depends on a bottom-up approach that establishes power demand and energy use. The testing for both the models is done by a blind validation demonstration of a good correlation between metered data and monthly predictions of energy consumption. The prediction ranges for power demand profiles are observed from metered data with minor exceptions. These methods help in improving the prediction of the operational performance of small power equipment in offices.[9]
However, the focus here is to improve the metered data collection in predicting the power consumption and not the factors affecting the discrepancy in the data. They do not consider the climatic effects or the socio-economic effects on the power consumption.

Eric Fox in the AEIC Load Research Conference presents about using load research data to develop long-term peak demand forecasts. Peak demand has been growing faster than energy due to growth in air-conditioning, closed spaces and increased lighting. The problem with the standard approach of building a monthly, yearly or seasonal forecast can be solved by capturing the impact of end-use saturation and efficiency trends and use class level and end-use forecasts to drive peak demand. For this, two approaches were used wherein the first method estimated customer hourly load profiles from data, combined it with monthly class energy forecast that reflected changing end-use energy composition and then aggregated monthly class profiles to scale to actual system load to load profile forecast. The second approach explicitly captures end-use load and peak-day conditions and requirements to create variables and use it to build monthly base-use load variable. The load build-up model performs slightly better than other.[10]
However, good load data is required to support both modelling approaches and the model will lose it's accuracy when used on data having various missing values.

6 Variable Description

The response variable is the daily Peak Demand in GWh [11] and there are 22 parameters for each county in New England ISO that will help us predict the response variable.

6.1 Summary Statistics of all Variables

No.	Variable Name	Description	Units	Source
1	COUNTY	County in the study region		iso-ne.com
2	Date	Date on which the load reading was taken		iso-ne.com
3	Peak.Demand	Peak load	GWh	iso-ne.com
4	STATE.x	State belonging to study region		iso-ne.com
5	DryBulb.max	Maximum temperature of air measured by a thermometer freely exposed to the air	oC	ncdc.noaa.gov
6	DryBulb.min	Minimum temperature of air measured by a thermometer freely exposed to the air	oC	ncdc.noaa.gov
7	DryBulb.avg	Average temperature of air measured by a thermometer freely exposed to the air	oC	ncdc.noaa.gov
8	WetBulb.max	Maximum measure of moisture present in the air.	oC	ncdc.noaa.gov
9	WetBulb.min	Minimum measure of moisture present in the air.	oC	ncdc.noaa.gov
10	WetBulb.avg	Average measure of moisture present in the air.	oC	ncdc.noaa.gov
11	DewPoint.max	Maximum temperature to which the air would have to cool to reach saturation	oC	ncdc.noaa.gov
12	DewPoint.min	Minimum temperature to which the air would have to cool to reach saturation	oC	ncdc.noaa.gov
13	DewPoint.avg	Average temperature to which the air would have to cool to reach saturation	oC	ncdc.noaa.gov
14	RelHumi.max	Maximum ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water	%	ncdc.noaa.gov
15	RelHumi.min	Minimum ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water	%	ncdc.noaa.gov
16	RelHumi.avg	Average ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water	%	ncdc.noaa.gov
17	WindSpeed.max	Max Wind speed	MPH	ncdc.noaa.gov
18	Windspeed.min	Min Wind speed	MPH	ncdc.noaa.gov
19	WindSpeed.avg	Average Wind speed	MPH	ncdc.noaa.gov
20	StPress.max	Maximum pressure felt at that spot, without being adjusted for altitude	Hg	ncdc.noaa.gov
21	StPress.min	Mnimum pressure felt at that spot, without being adjusted for altitude	Hg	ncdc.noaa.gov
22	Stpress.avg	Average pressure felt at that spot, without being adjusted for altitude	Hg	ncdc.noaa.gov
23	Population	No. of people living in the county		www.census.gov
24	Area.sqm	Area of the county	M2	www.census.gov
25	PopDensity	Population density in the county	Unit/m2	www.census.gov

Table 1: Variable description

7 Methodology

Using climate data such as temperature, wind speed and air pressure, the models can be developed with those features with our target variable as the peak load.

7.1 Data Preprocessing

1. Data from official sources
2. Extracting relevant information
3. Merging the peak load data with climate and socioeconomic data [11]
4. Identifying the daily peak load from the hourly data collected
5. Merging these parameters with the peak load data according to the date

7.2 Problems faced during Data Pre-processing

7.2.1 Missing Counties

The dataset of 14 counties, so we estimated the climate data values based on their distance to the closest county with the available values. Since our weather data was from airports, we looked for the closest airport (next station) to it, and copied that climate data to the missing county.

7.2.2 Population data

Population for some years was not available. However, it is known that population does not change much each year. Hence, we assume the same population as 2016 for those missing years.

7.3 Model Building

1. Splitting the data into 85-percent training and 15-percent testing sets.
2. General linear model (GLM) with all the variables left after data cleaning is run on the training set to find the significant parameters.

3. Forward Stepwise regression is carried out on the training dataset for variable selection. This builds a regression model from a set of candidate predictor variables by entering predictors based on p values, in a stepwise manner until there is no variable left to enter any more. The stepwise selection process adds all variables one at a time and at each step keeps the ones that contribute significantly to the model. This gives the following variables as significant: WetBulb.max, WindSpeed.max, RelHumi.max, DewPoint.avg, WetBulb.avg, RelHumi.avg, DryBulb.avg, DewPoint.min, Population, PopDensity, AREA.sqm., WetBulb.min, StPress.min, StPress.avg.

4. Using these 14 variables MLR is run on the training set. We see that RMSE values decreases a lot compared to previous model but is still high. The residuals were checked, and they didn't show any pattern and were normally distributed.

5. Boxcox transformation gives lambda value = -1.09. But instead of reducing, the RMSE is drastically increased. Hence, we don't transform the variable.

6. Ridge regression is an extension for linear regression. It's basically a regularized linear regression model that solves the issue of overfitting. A super important fact we need to notice about ridge regression is that it enforces the coefficients to be lower, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the trained model. Hence, ridge regression was carried out on these 14 variables and it didn't give any reason to leave out any predictors.

7. Lasso method overcomes the disadvantage of Ridge regression by not only punishing high values of the coefficients but setting them to zero if they are not relevant. Therefore, we end up with the following variables: DryBulb.avg, WetBulb.max, DewPoint.avg, RelHumi.max, WindSpeed.max, StPress.avg, Area and PopDensity having the major effect on the response.

8. Since, the RMSE values were still high due to over-fitting we try K-fold Cross validation 3 models GLM, Ridge and Lasso using the 8 variables selected after Lasso.

9. The variables given by Lasso are then used in the GAM Model which helps fit various predictors by adding new variables each time while keeping the rest of them fixed. GAM uses backfitting and pre-defined smoothing splines approach. The model with the lowest AIC (233785.5) is the best model. GAM gives a lower RMSE (325). So, this is the best model uptil now. Though GAM allows non-linear fit to predict the response it has a propensity to overfit and because of this limitation the model loses predictability when the smoothed variables have values outside of the range of training dataset. Essentially, you are sacrificing predictability outside of your data range for precision within your data range.

10. RPART regression trees builds classification or regression models of a very general structure using a two-stage procedure; the resulting models can be represented as binary trees. Typically, you will want to select a tree size that minimizes the cross-validated error. The RMSE for rpart comes out to be 381.794 which is very high compared to others. This is because the decision tree changes when the dataset is perturbed a bit. This reduces the robustness of the classification algorithm to noise and isn't able to generalize well to future observed data. This can undercut confidence in the tree and hurt the ability to learn from it.

11. Random forest is then used as it is one of the most accurate learning algorithms available. It produces a highly accurate classifier and runs efficiently on large databases. It reduces overfitting by averaging several trees and reduces variance by using multiple trees which avoids the inclusion of a classifier that doesn't perform well between training and testing data. The significant variables given by Random Forest are: DryBulb.avg, WetBulb.avg, DewPoint.avg, RelHumi.min, WindSpeed.avg, StPressure.min, Population, PopDensity, AREA.sqm. The RMSE value for random forest is 273.51 which is the lowest and hence this shows that it learns from the training data and helps predict on the test data very well.

12. SVM model is also fit to our data; the parameters are specified in the SI file.

13. BartMachine followed the same modeling as ANN and SVM, using the function in its own package[12] to set the best parameters (detailed in SI) before fitting the cross validated models.

Now, we start different sections for each state

8 Connecticut

8.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

8.1.1 Density Plots

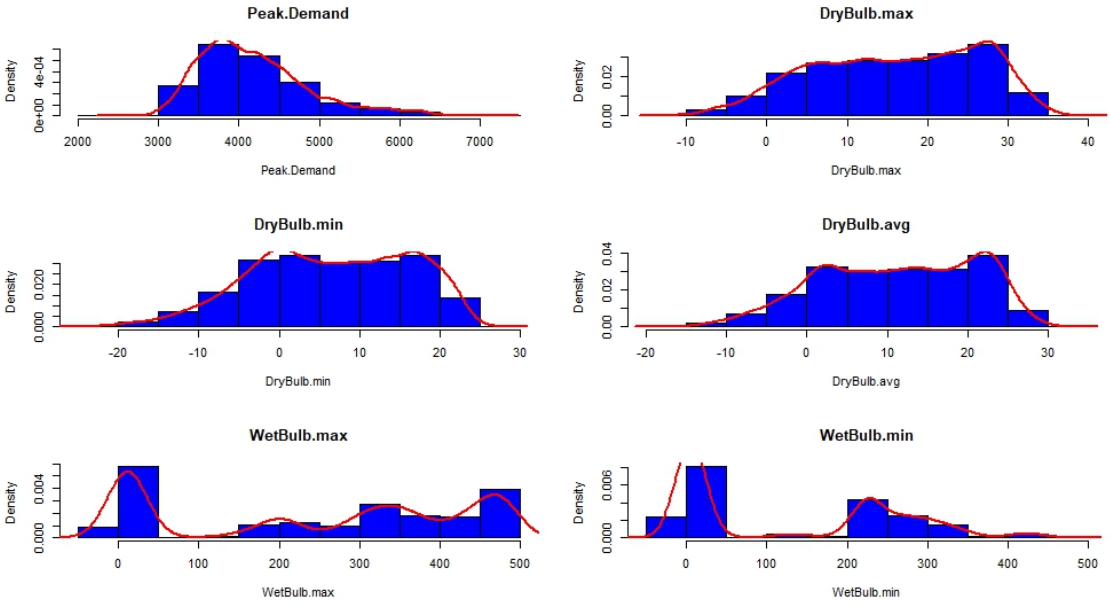

Figure 1: Distribution of variables

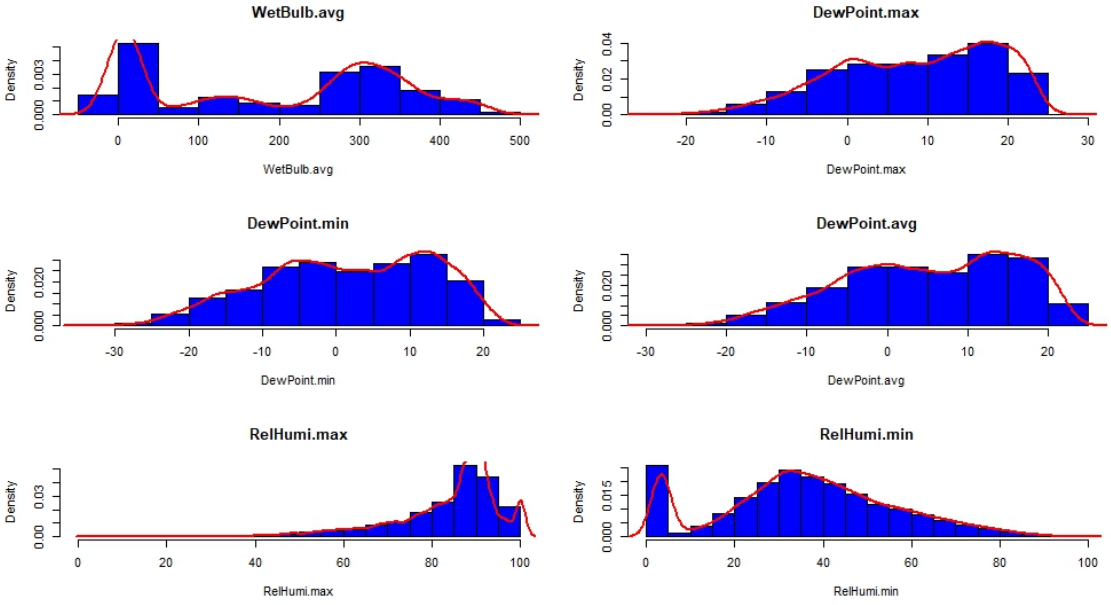

Figure 2: Distribution of variables

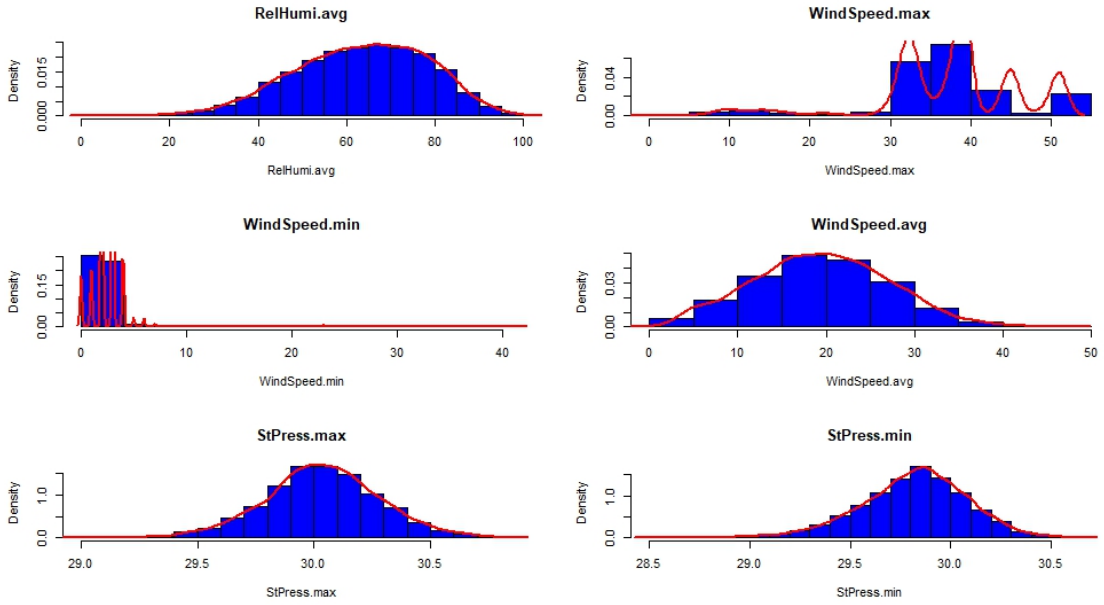

Figure 3: Distribution of variables

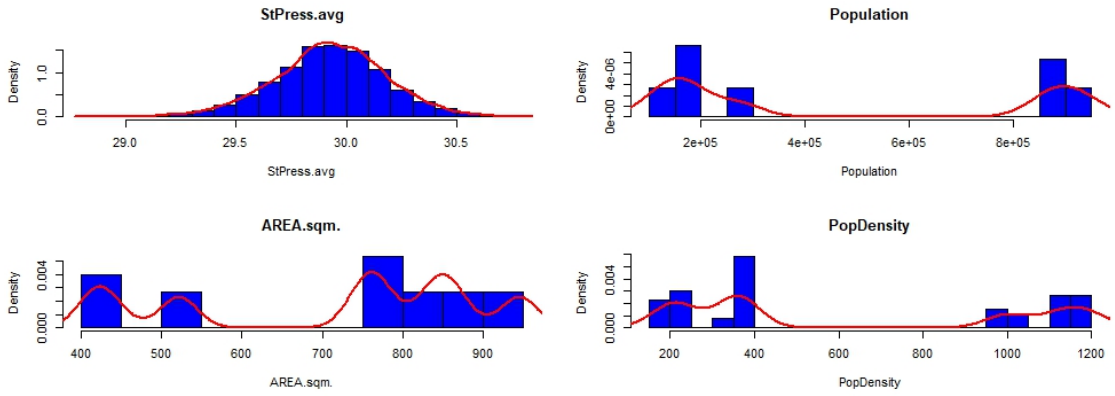

Figure 4: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

8.1.2 Scatter Plots between Peak Demand and other variables

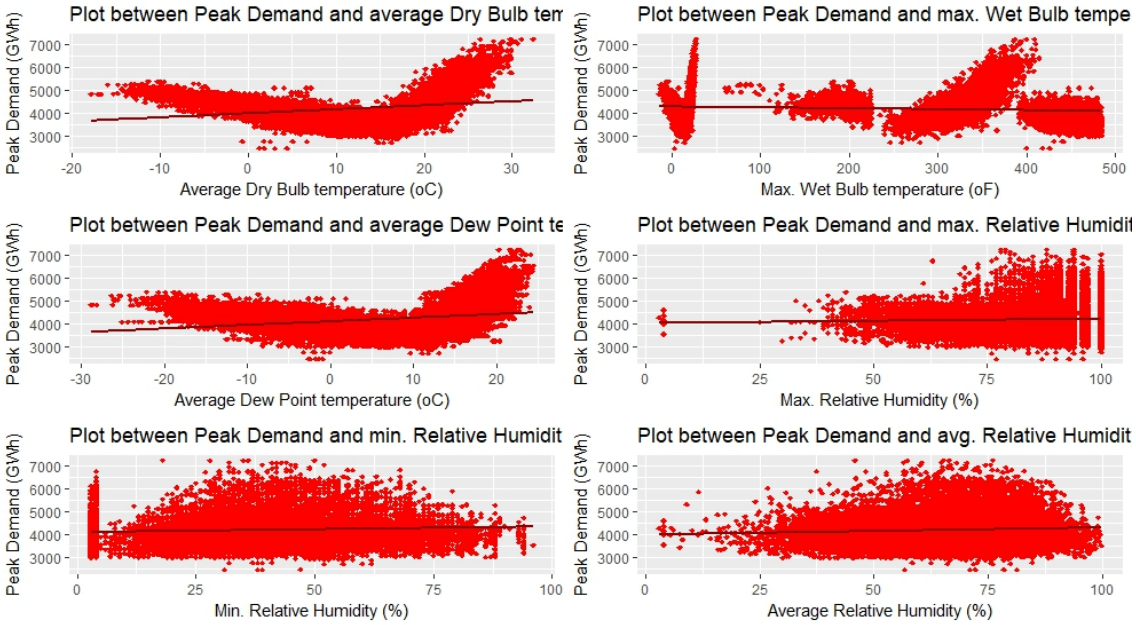

Figure 5: Scatter Plots between Peak Demand and other variables

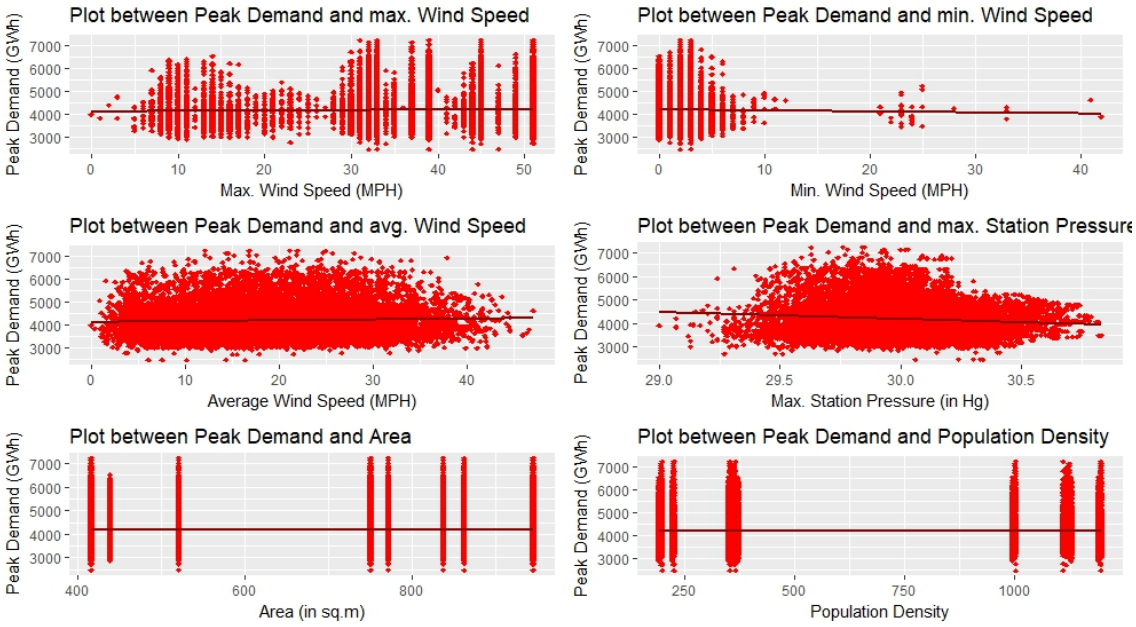

Figure 6: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 300-350oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

8.1.3 Violin Plots

Our target variable, Peak.Demand, was plotted in violin plots against our important variables - the ones we selected for the models.

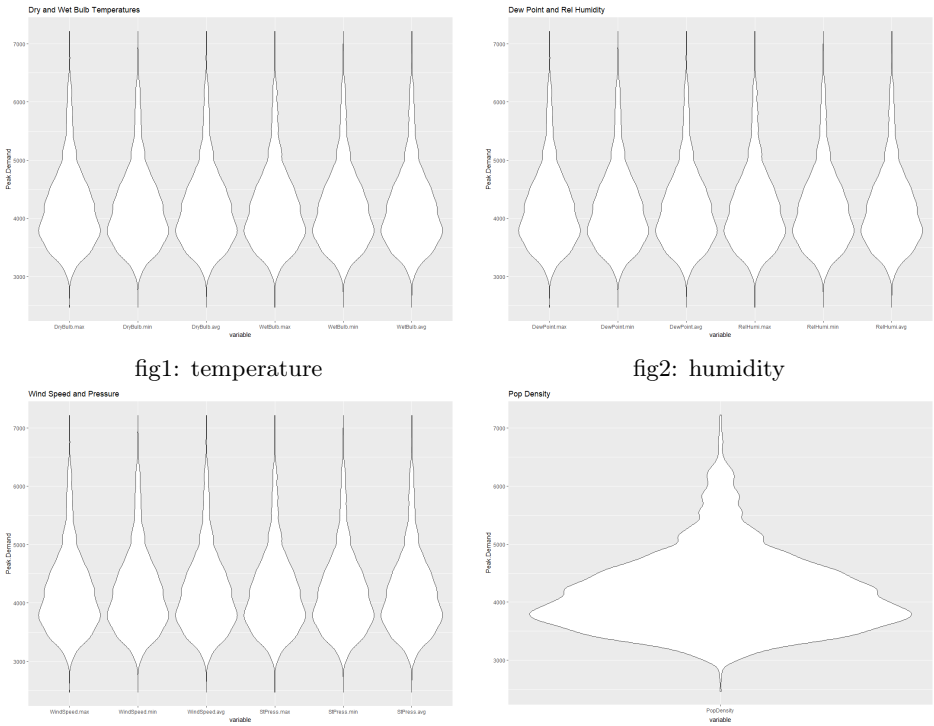

fig1: temperature

fig2: humidity

fig3: test3

fig4: test4

Figure 7: Violin Plots

We do not see variation between max, min or average values when plotting against Peak.Demand.

8.1.4 Correlation plot

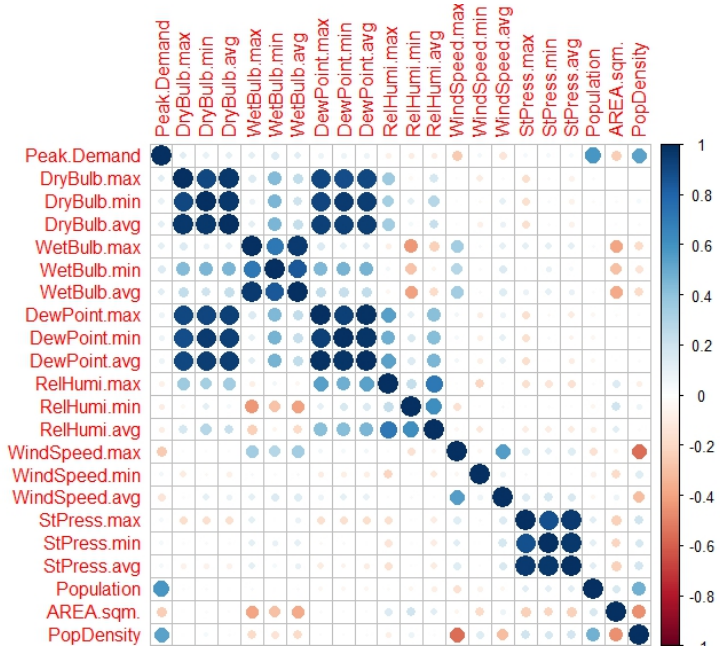

Figure 8: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.
Hence, transformation and other techniques will be required while building the predictive model.

8.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 fits.

	RMSE.train Mean	**RMSE.test Mean**
MLR	660.2638	660.0548
Ridge Model	661.367	661.2023
Lasso Model	722.8522	660.2714
GAM	951.4673	323.9911
Decision Trees	286.7145	377.3713
Random Forest	126.9767	273.5183
SVM	190.4676	308.8617
BartMachine	234.2435	305.7965

SVM and BartMachine both show signs of overfit models, with test errors higher than train. More details of this results are found in the Supporting Information document, like Q-Q and Actual vs Predicted plots.

8.3 Best Model- Random Forest

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = 273.5183

b) It performs better than decision trees because it reduces overfitting by averaging several trees.

c) Also, it reduces variance by using multiple trees which avoids the inclusion of a classifier that doesn't perform well between training and testing data.

d) Independent training of each base classifier on a training set sampled with replacement from the original training set. As the number of trees increase the error decreases. This technique is known as bagging, or bootstrap aggregation. In Random Forest, further randomness is introduced by identifying the best split feature from a random subset of available features.

e) The LOOCV Lasso and MLR do not consider the interaction between variables that is taken into consideration by Random Forest.

f) The difference in RMSE values between LOOCV Lasso and Random Forest isn't much but Random Forest is one of the most accurate learning algorithms available and hence have a great predictive ability compared to other models even though it's complexity is high, and interpretability low compared to others.

g) Though it is complex than GAM, Random Forest has better predictive capability compared to GAM. GAM has a propensity to overfit and because of this limitation the model loses predictability when the smoothed variables have values outside of the range of training dataset. Essentially, you aren't sacrificing predictability outside of your data range for precision within your data range when using Random Forest

8.3.1 Variable Selection

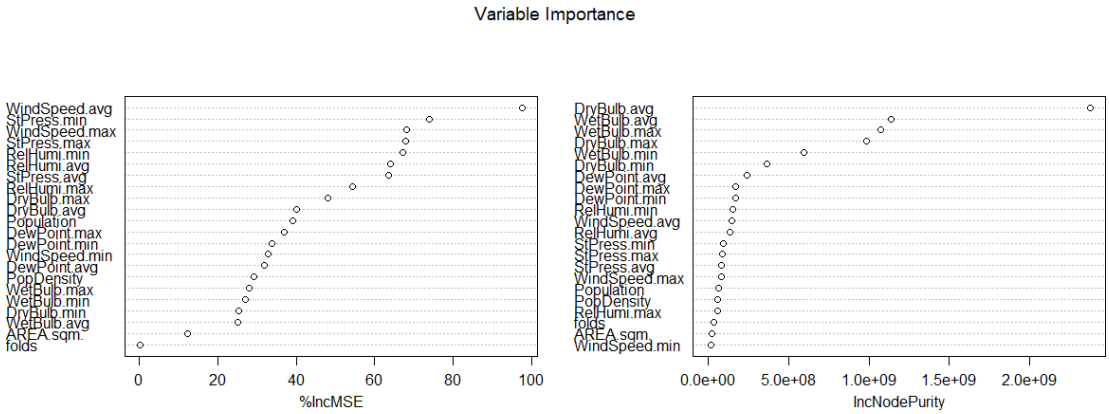

Figure 9: Variable selection Random Forest

The predictors based on Node Purity that have been selected here are DryBulb.avg, WetBulb.avg, DewPoint.avg, RelHumi.avg, StPressure.min, WindSpeed.avg Population, PopDensity, Area

8.3.2 Boxplots of Models

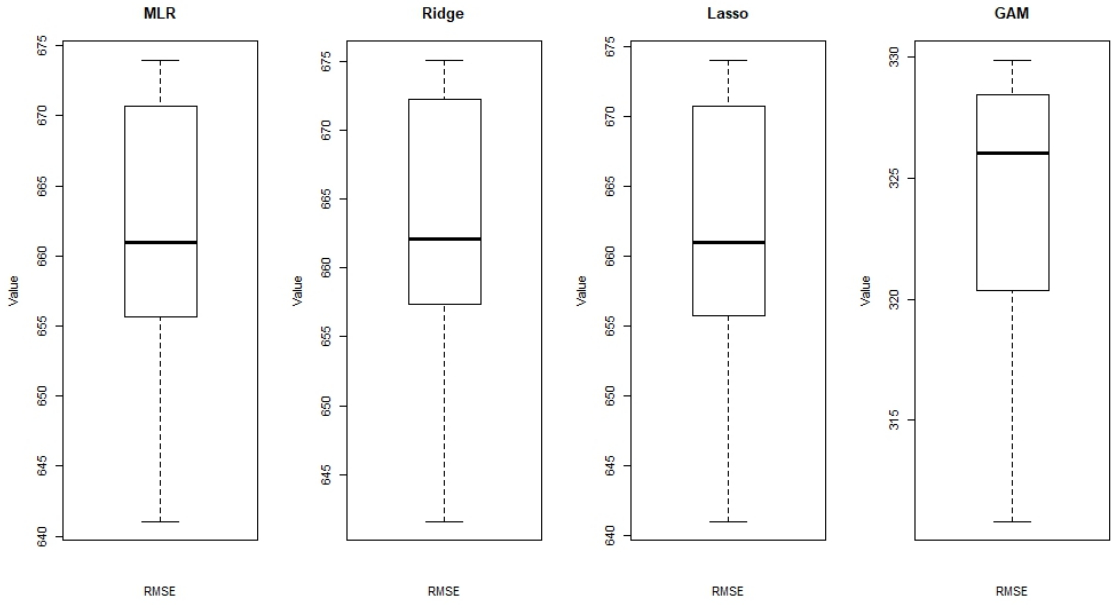

Figure 10: Boxplot of Models(RMSE) - MLR, Ridge, Lasso and GAM

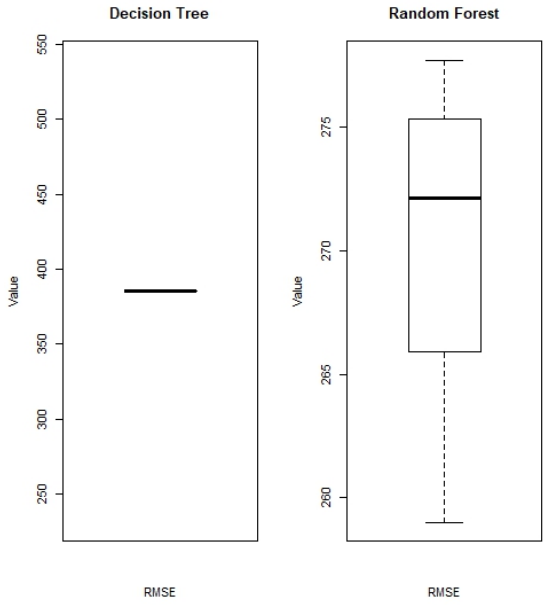

Figure 11: Boxplot of Models(RMSE) - Decision Trees,Random Forest

The boxplots show the distribution of the 10 fold out of sample RMSE's for different models. It can be seen that Random Forrest has a concentrated distribution with the lowest mean RMSE value.

9 Rhode Island

9.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

9.1.1 Density Plots

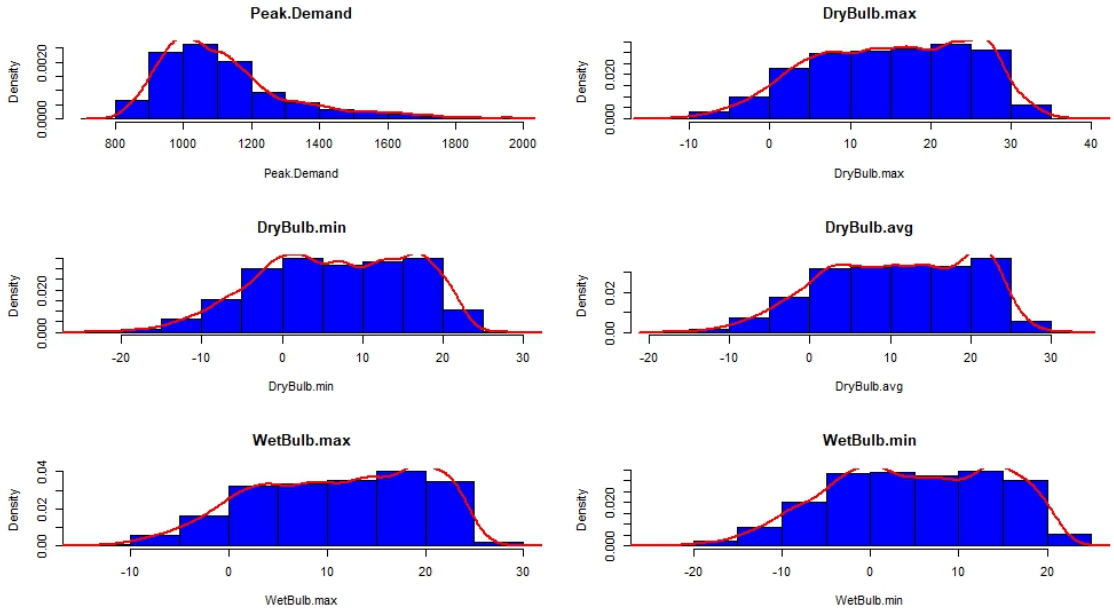

Figure 12: Distribution of variables

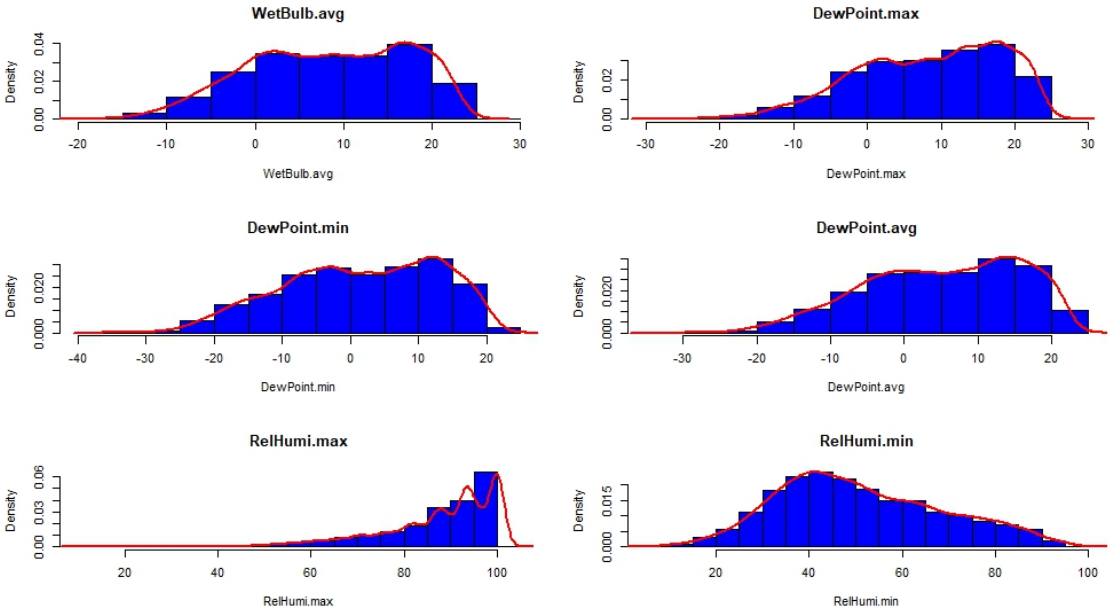

Figure 13: Distribution of variables

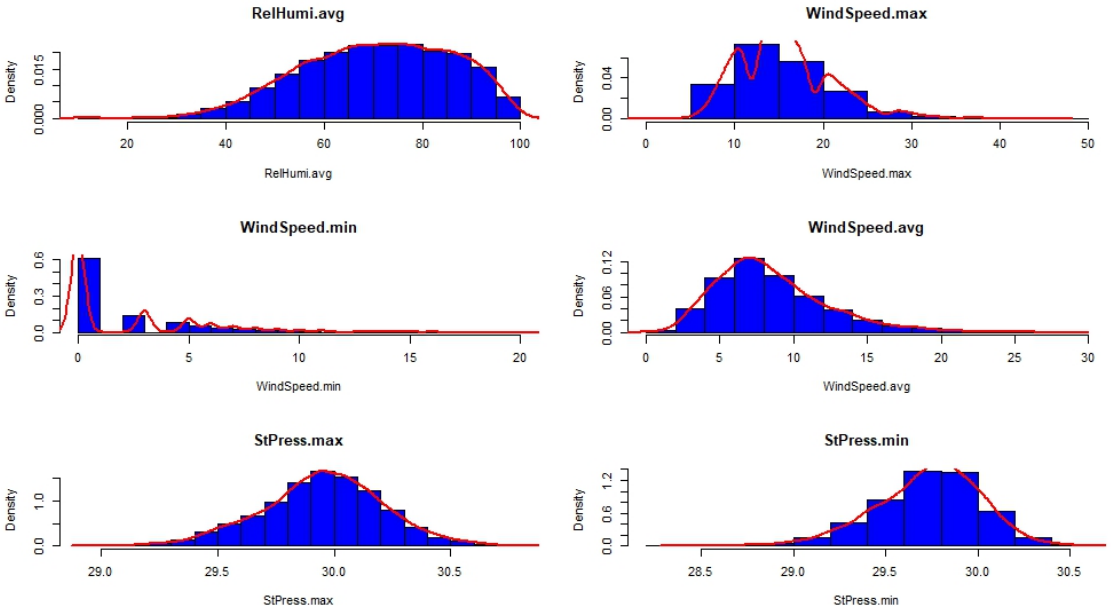

Figure 14: Distribution of variables

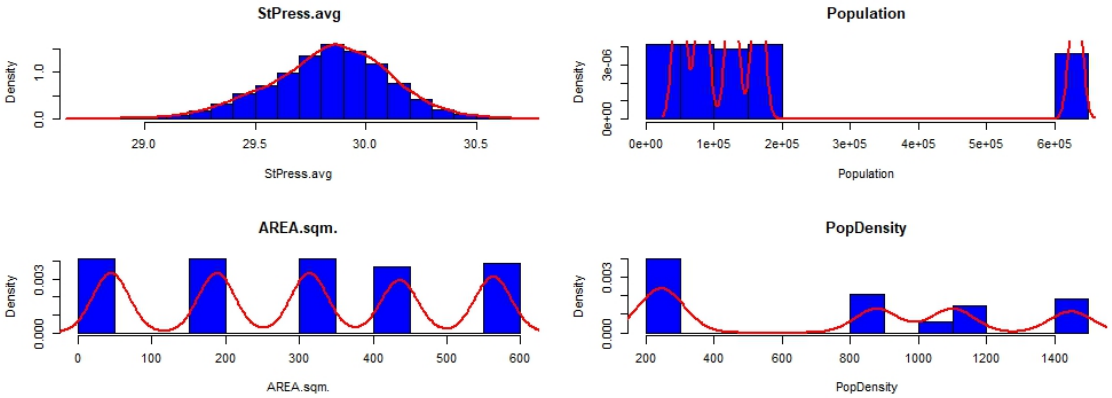

Figure 15: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

9.1.2 Scatter Plots between Peak Demand and other variables

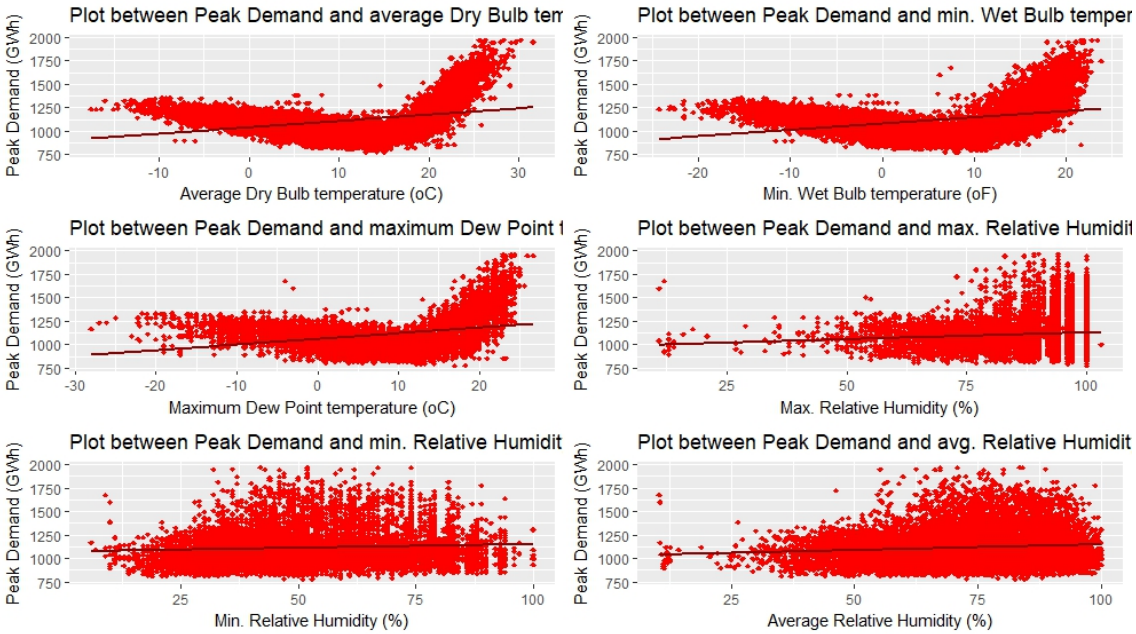

Figure 16: Scatter Plots between Peak Demand and other variables

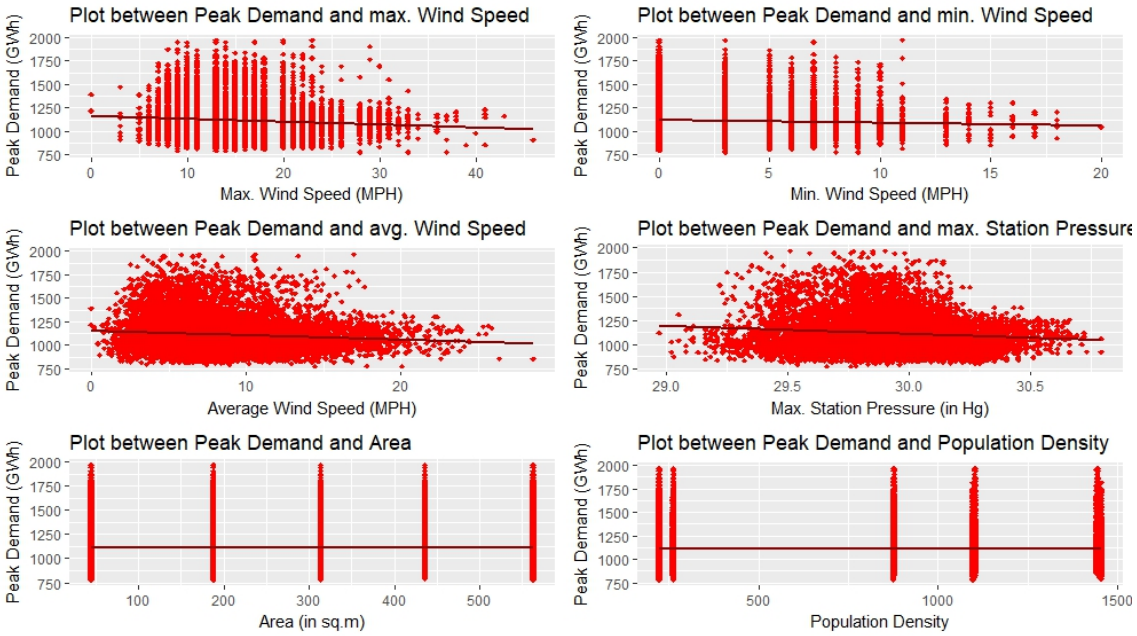

Figure 17: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 10-20oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

9.1.3 Violin Plots

Our target variable, Peak.Demand, was plotted in violin plots against our important variables - the ones we selected for the models.

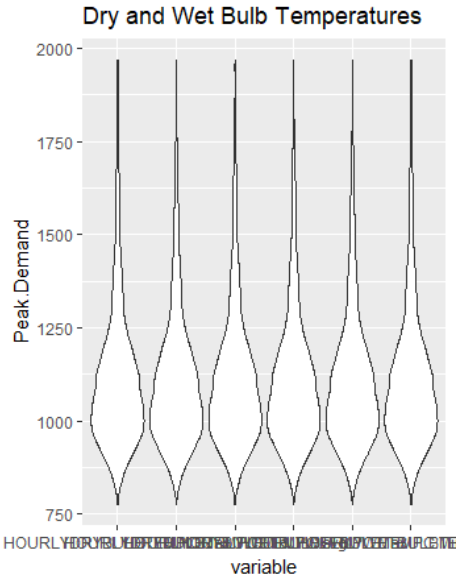

fig1: temperature

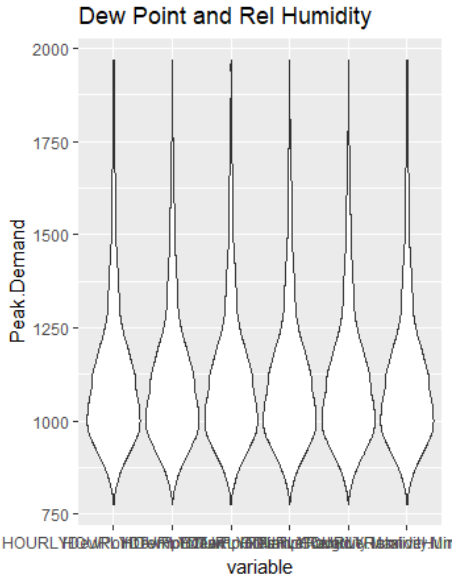

fig2: humidity

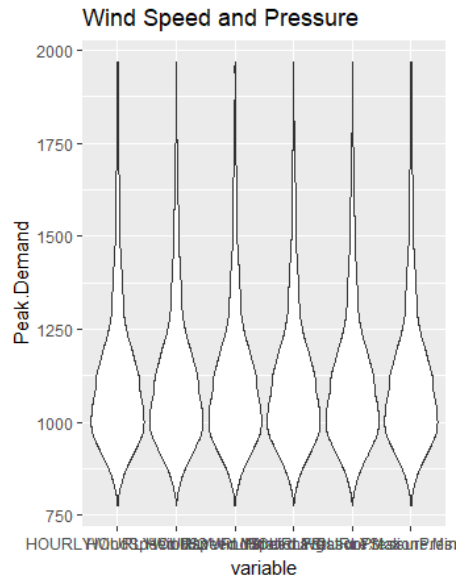

fig3: windspeed

fig4: PopulationDensity

Figure 18: Violin Plots Rhode Island

We see the values are predominant in the same range of results. The Population density is smaller than CT, but still higher than VT, which are results we expected from the characteristics of those states.

9.1.4 Correlation plot

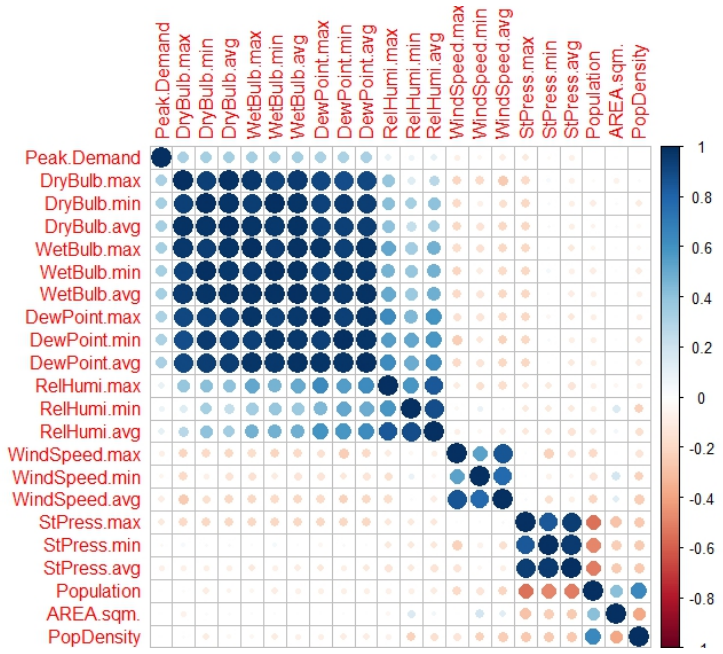

Figure 19: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.
Hence, transformation and other techniques will be required while building the predictive model.

9.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 models fit.

	RMSE.train Mean	RMSE.test Mean
MLR	171.9378	172.0485
Ridge Model	172.8364	172.9274
Lasso Model	198.5958	172.0508
GAM	253.2276	84.4418
Decision Trees	165.165	181.256
Random Forest	133.472	151.0774
MARS Unpruned	80.98665	87.1866
MARS Pruned	80.9929	87.2281
SVM	45.84439	70.37895
BartMachine	76.97309	82.98487

We see that Rhode Island did better than Connecticut, and since it is a considerably smaller dataset, we assume that was the reason for a better accuracy in predicting our target variable.

9.3 Best Model- MARS

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = value

b) MARS makes no assumptions about the underlying functional relationships between dependent and independent variables. In general, the splines are connected smoothly together, and these piecewise curves (polynomials), also known as basis functions (BFs), result in a flexible model that can handle both linear and nonlinear behavior.

c) MARS generates BFs by stepwise searching overall possible univariate candidate knots and across interactions among all variables. An adaptive regression algorithm is adopted for automatically selecting the knot locations. The MARS algorithm involves a forward phase and a backward phase. The forward phase places candidate knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares residual error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and overfitted model. The backward phase involves deleting the redundant BFs that made the least contributions

d) MARS can be considered to be more computationally efficient than other models, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding and pruning. In addition, the developed MARS models are easier to be interpreted. Furthermore, since MARS explicitly defines the knots for each design input variables, the model enables engineers to have an insight and understanding of where significant changes in the data may occur.

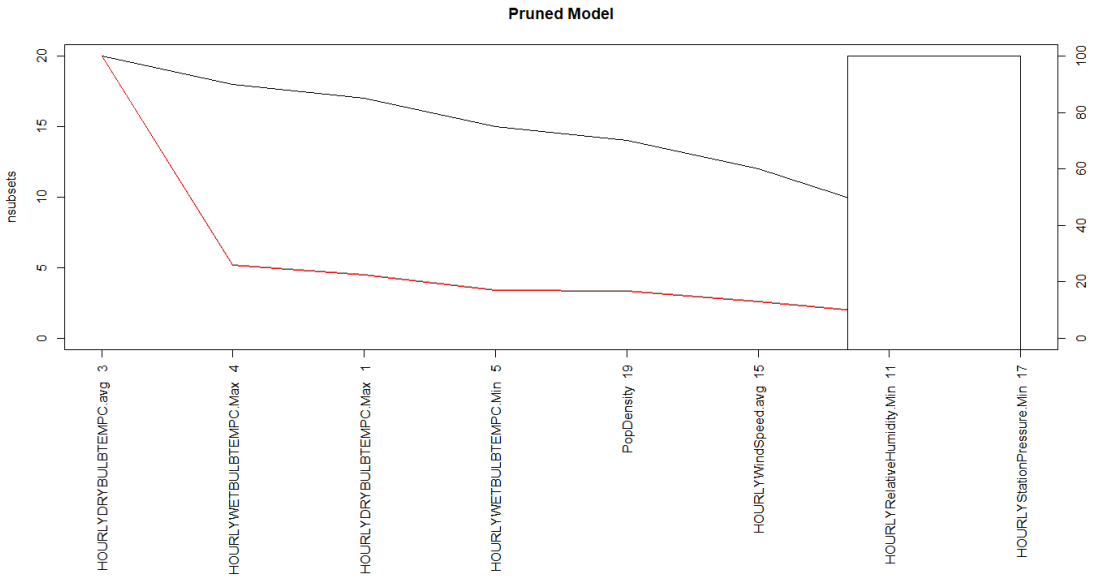

Figure 20: Variable Selection Pruned

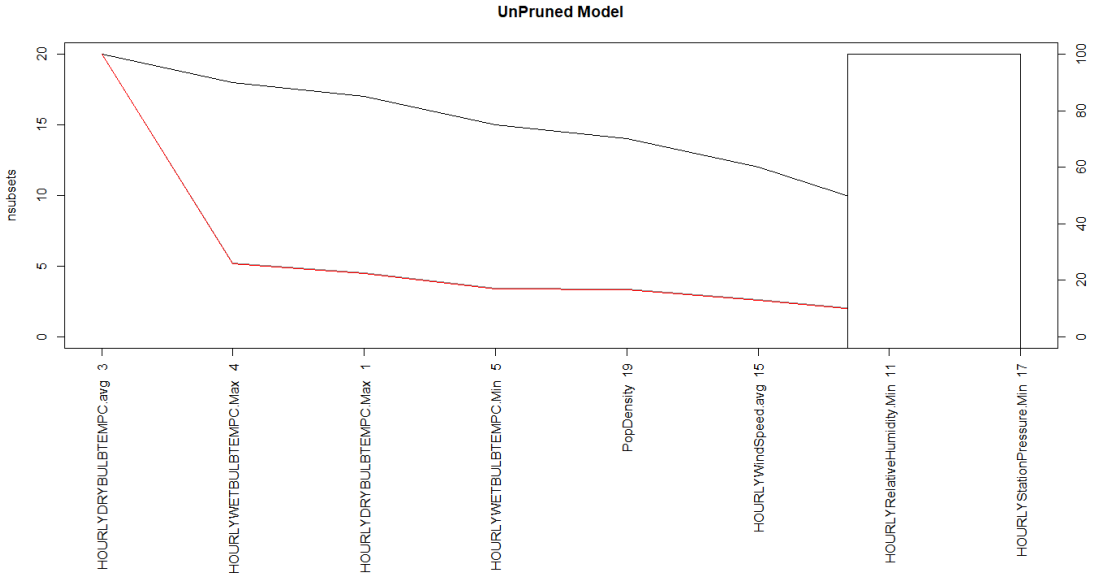

Figure 21: Variable Selection Pruned

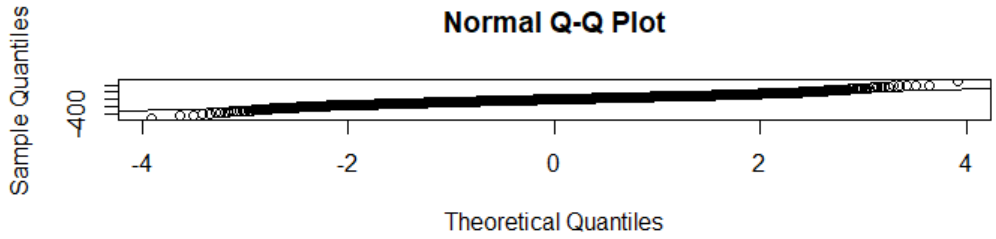

Figure 22: QQ plot for Unpruned MARS

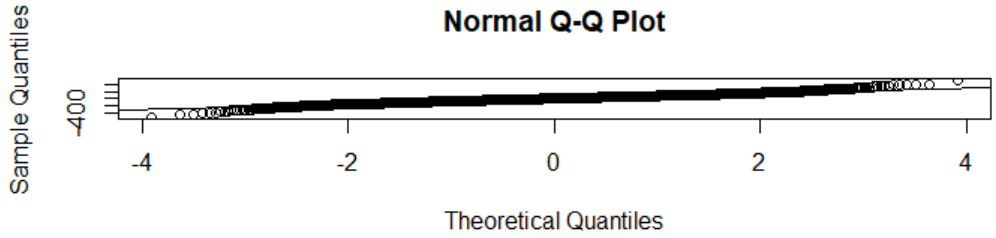

Figure 23: QQ plot for Pruned MARS

10 Massachusetts

10.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

10.1.1 Density Plots

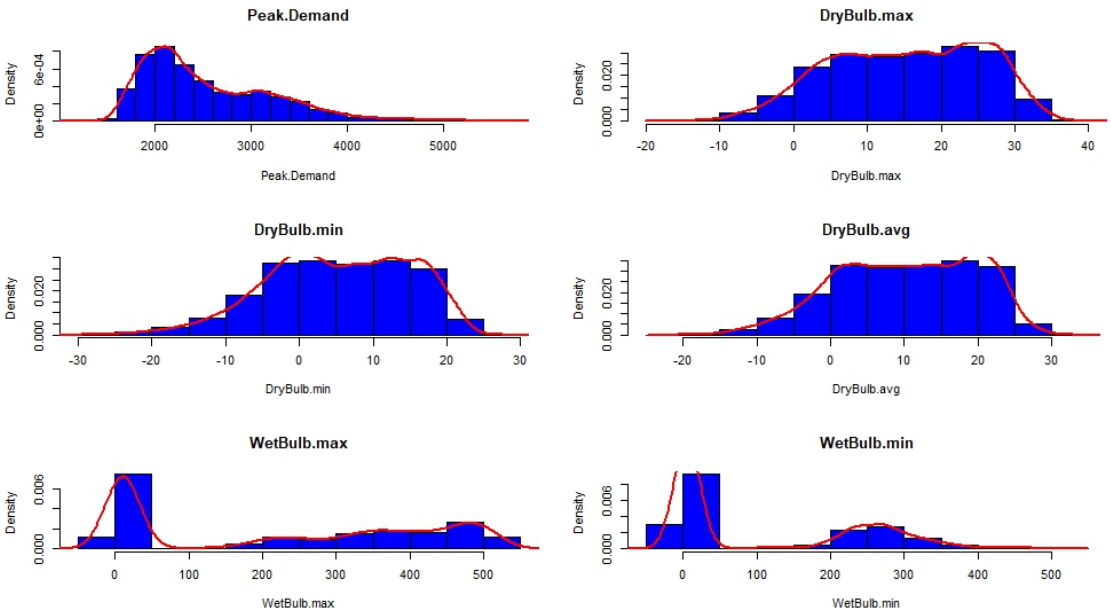

Figure 24: Distribution of variables

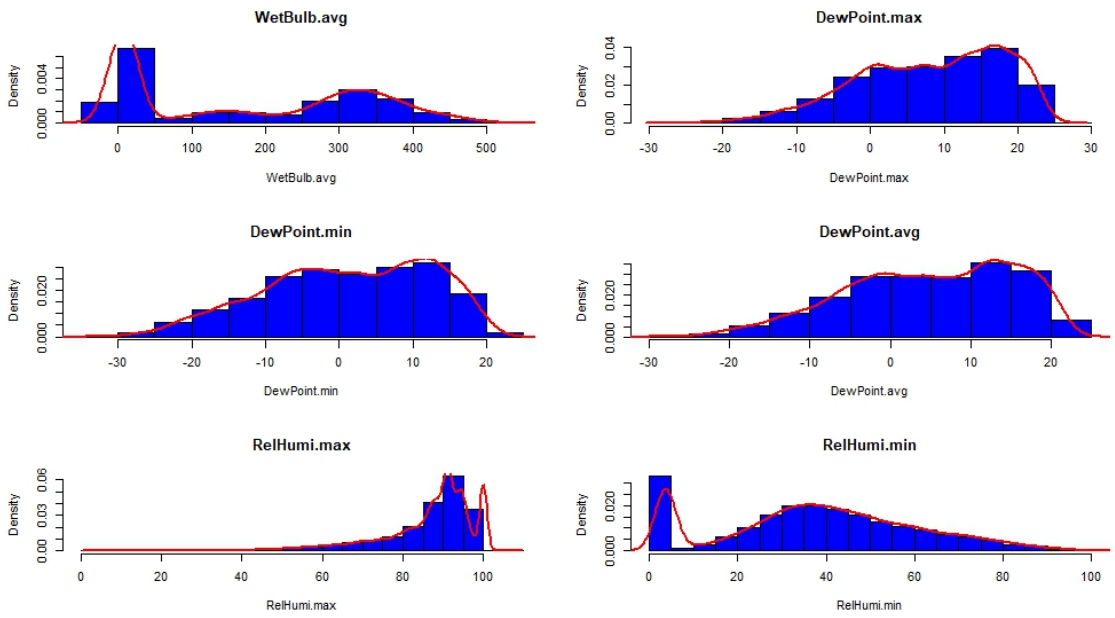

Figure 25: Distribution of variables

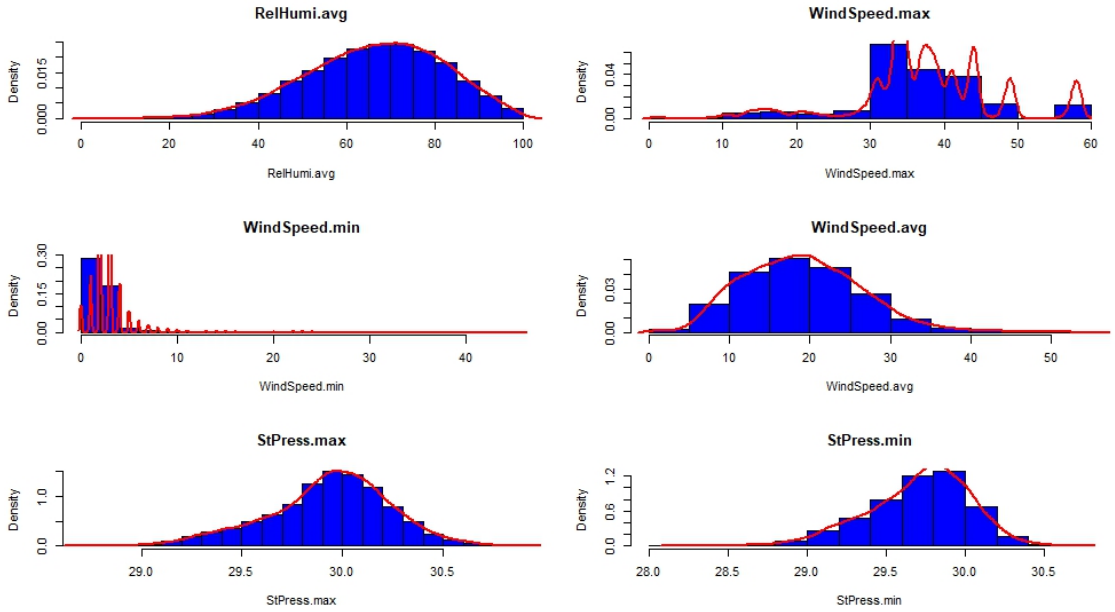

Figure 26: Distribution of variables

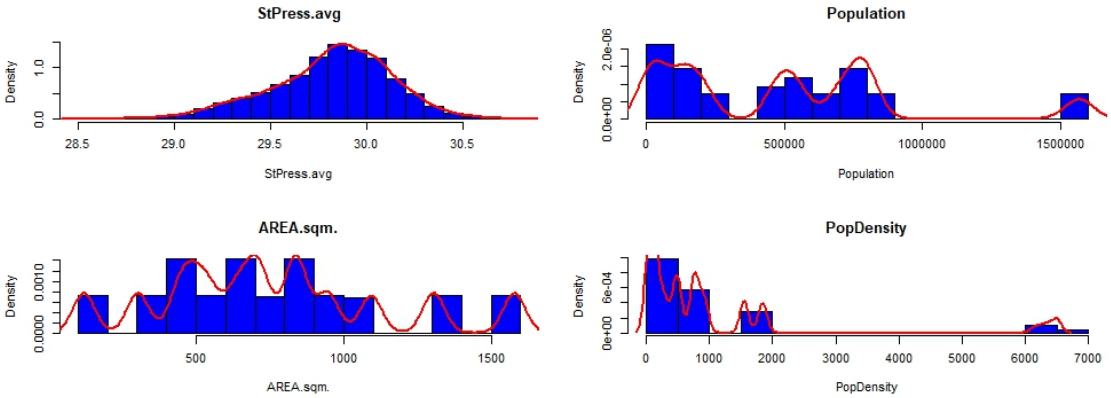

Figure 27: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

10.1.2 Scatter Plots between Peak Demand and other variables

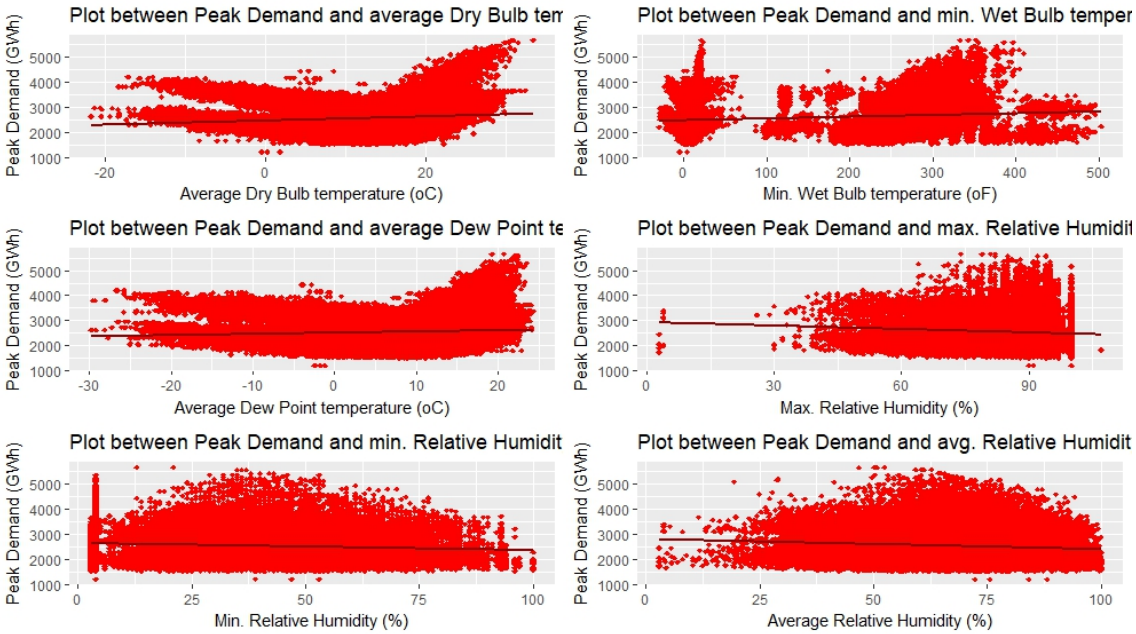

Figure 28: Scatter Plots between Peak Demand and other variables

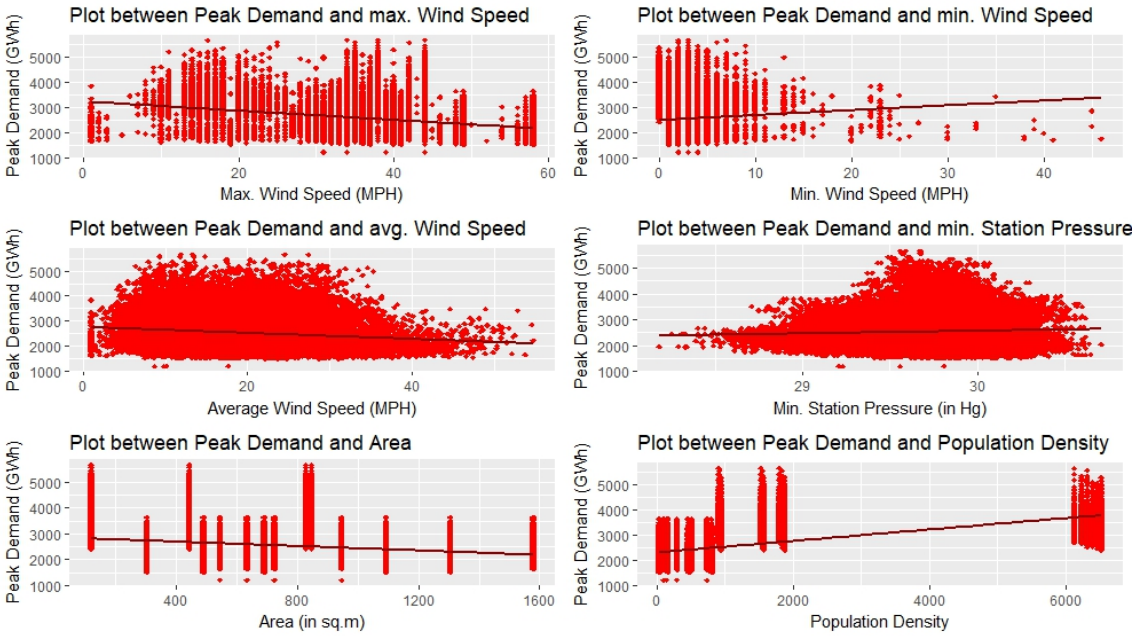

Figure 29: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 10-20oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

10.1.3 Violin Plots

Our target variable, Peak.Demand, was plotted in violin plots against our important variables - the ones we selected for the models.

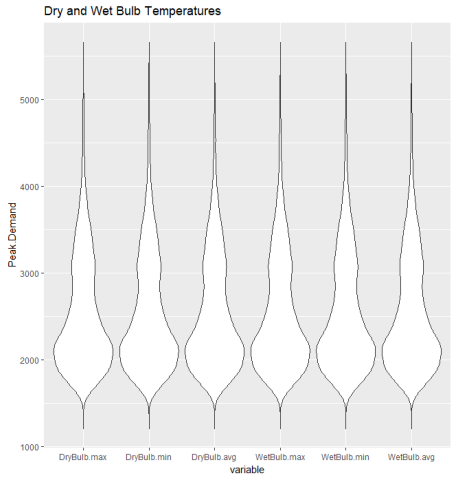

fig1: temperature

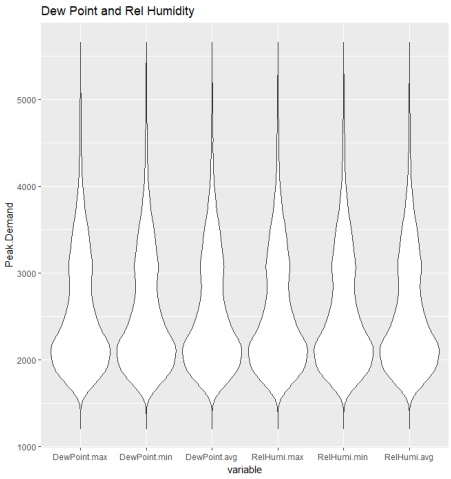

fig2: humidity

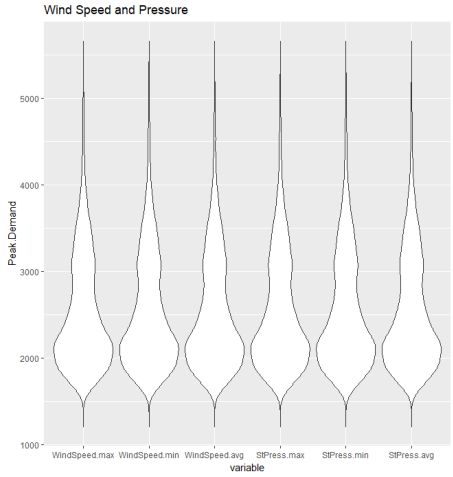

fig3: windspeed

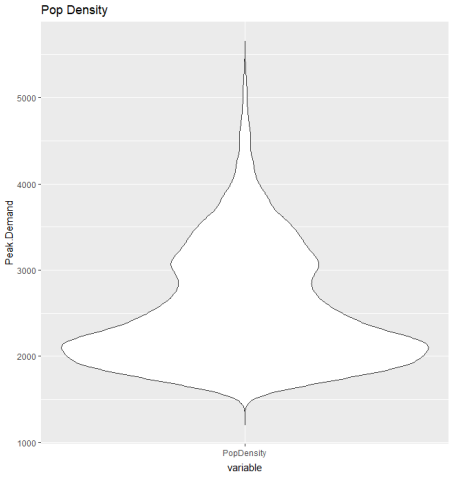

fig4: PopulationDensity

Figure 30: Violin Plots Massachusetts

We see the values are predominant in the same range of results. The Population density is the highest among all other states, which are results we expected from the characteristics of this states.

10.1.4 Correlation plot

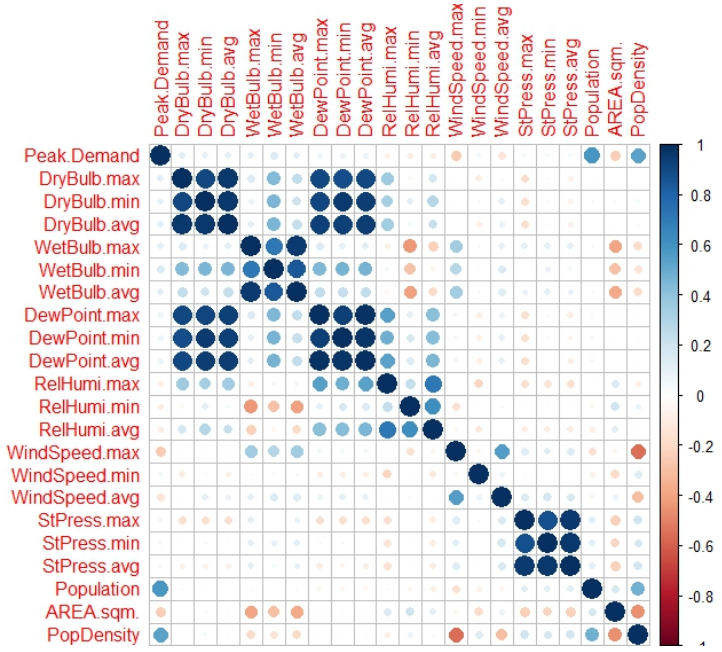

Figure 31: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.
Hence, transformation and other techniques will be required while building the predictive model.

10.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 models fit.

	RMSE.train Mean	RMSE.test Mean
MLR	530.5549	530.5894
Ridge Model	533.0431	533.0644
Lasso Model	757.737	530.605
GAM	820.7037	284.5181
Decision Trees	280.83235	292.2337
Random Forest	81.67636	223.6686
MARS Unpruned	21.7618	23.97228
MARS Pruned	21.7618	23.97228
SVM	243.0005	170.6444
BartMachine	380.7098	239.3022

Since Massachussets is a state with many counties, so it has a larger dataset than Rhode Island, we see the RMSE back up to another order of magnitude, like Connecticut.

10.3 Best Model- MARS

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = value

b) MARS makes no assumptions about the underlying functional relationships between dependent and independent variables. In general, the splines are connected smoothly together, and these piecewise curves (polynomials), also known as basis functions (BFs), result in a flexible model that can handle both linear and nonlinear behavior.

c) MARS generates BFs by stepwise searching overall possible univariate candidate knots and across interactions among all variables. An adaptive regression algorithm is adopted for automatically selecting the knot locations. The MARS algorithm involves a forward phase and a backward phase. The forward phase places candidate knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares residual error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and overfitted model. The backward phase involves deleting the redundant BFs that made the least contributions

d) MARS can be considered to be more computationally efficient than other models, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding and pruning. In addition, the developed MARS models are easier to be interpreted. Furthermore, since MARS explicitly defines the knots for each design input variables, the model enables engineers to have an insight and understanding of where significant changes in the data may occur.

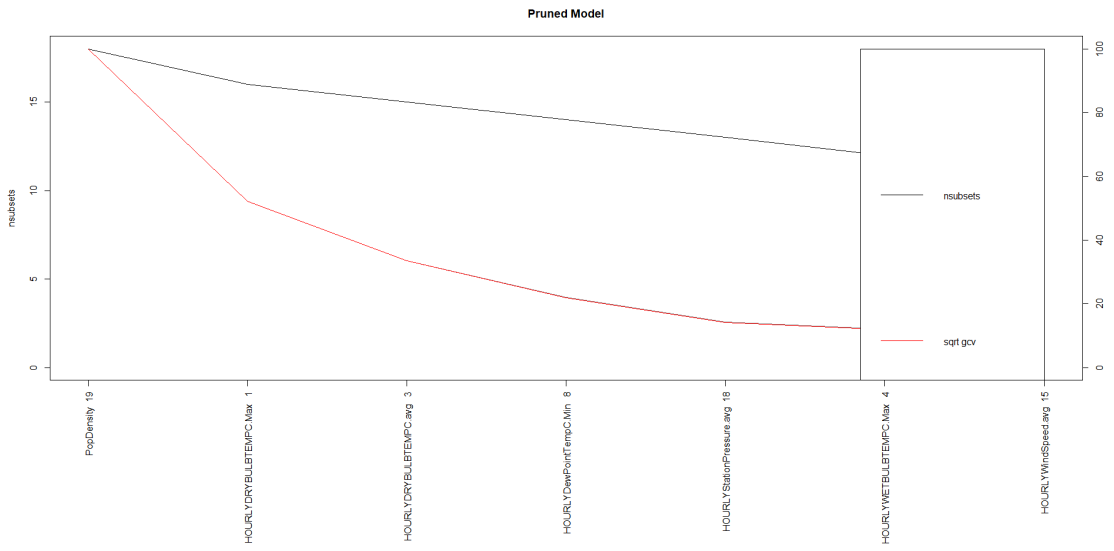

Figure 32: Variable Selection Pruned

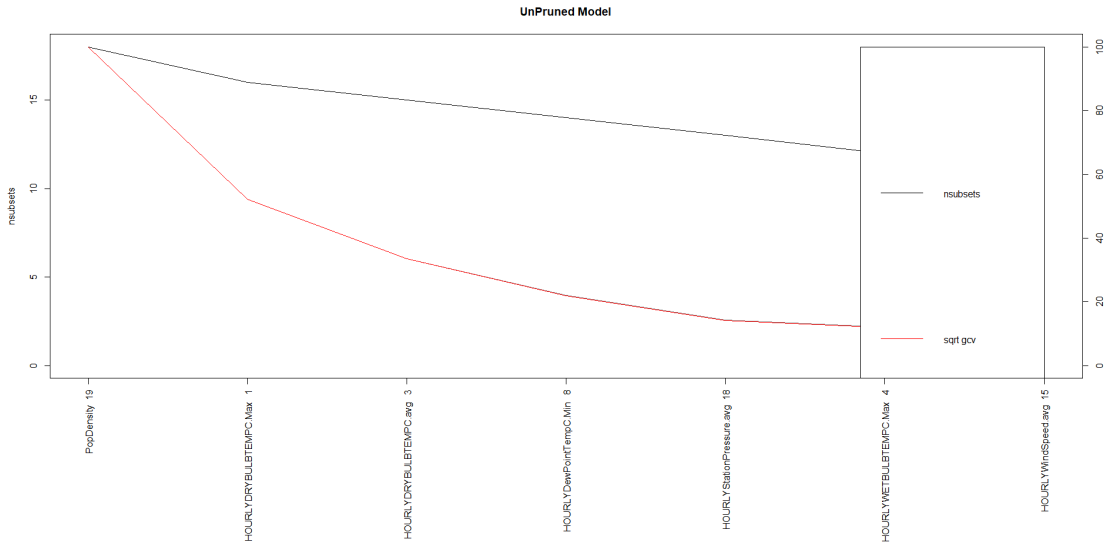

Figure 33: Variable Selection Unpruned

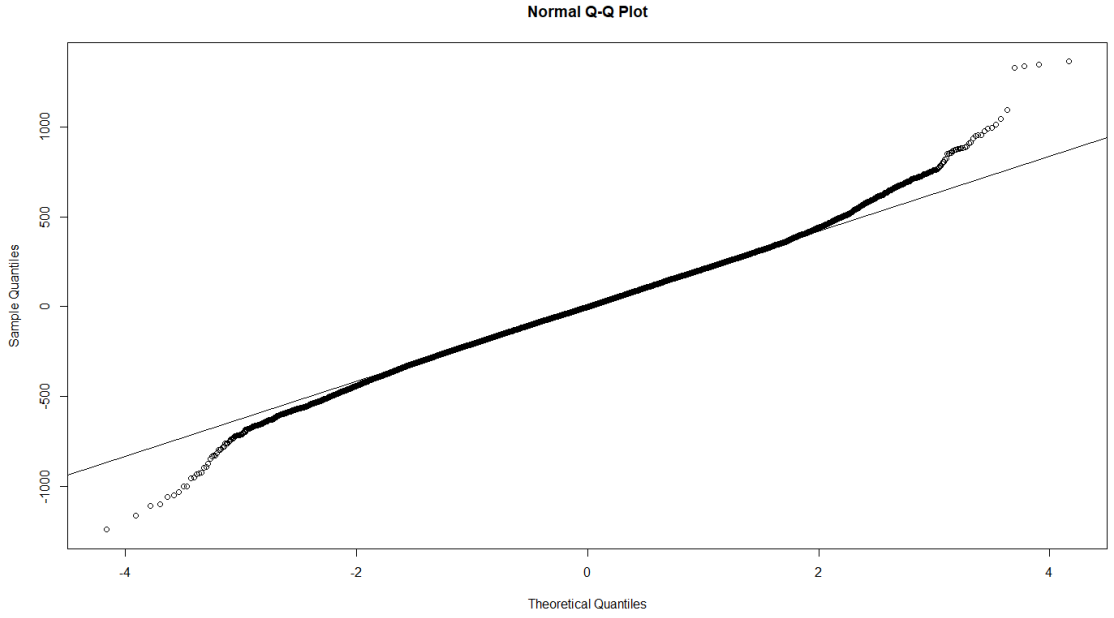

Figure 34: Normal QQ Plot for Pruned

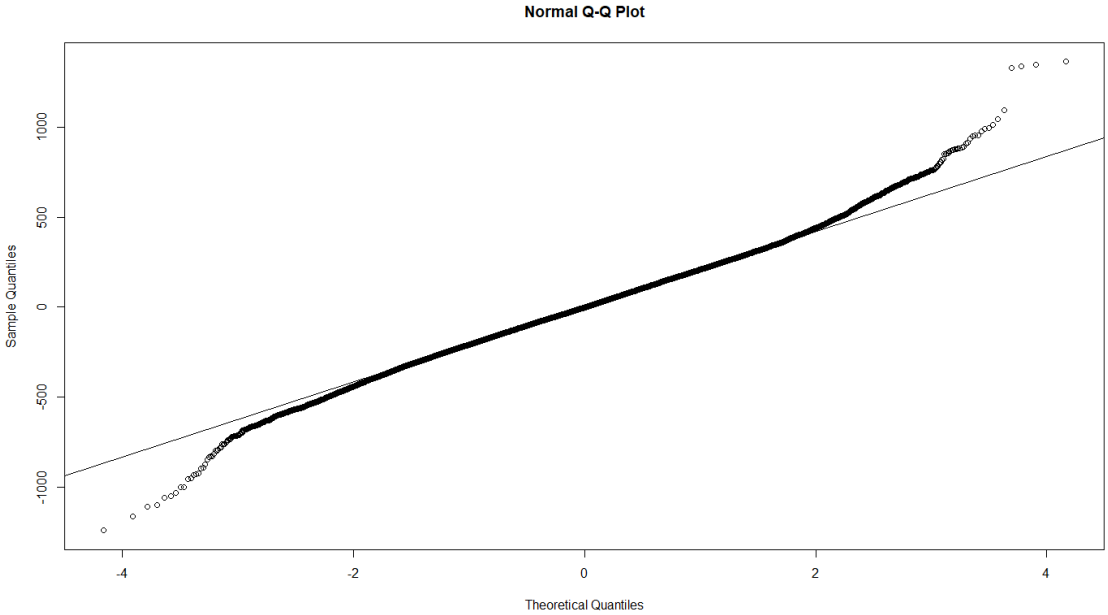

Figure 35: Normal QQ plot for Unpruned

11 Vermont

11.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

11.1.1 Density Plots

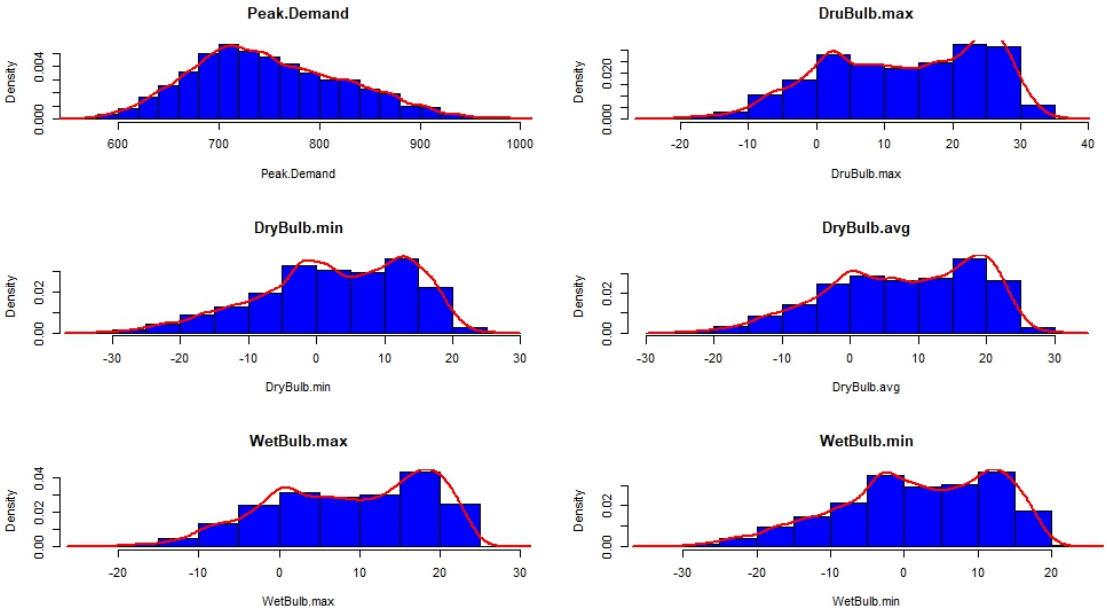

Figure 36: Distribution of variables

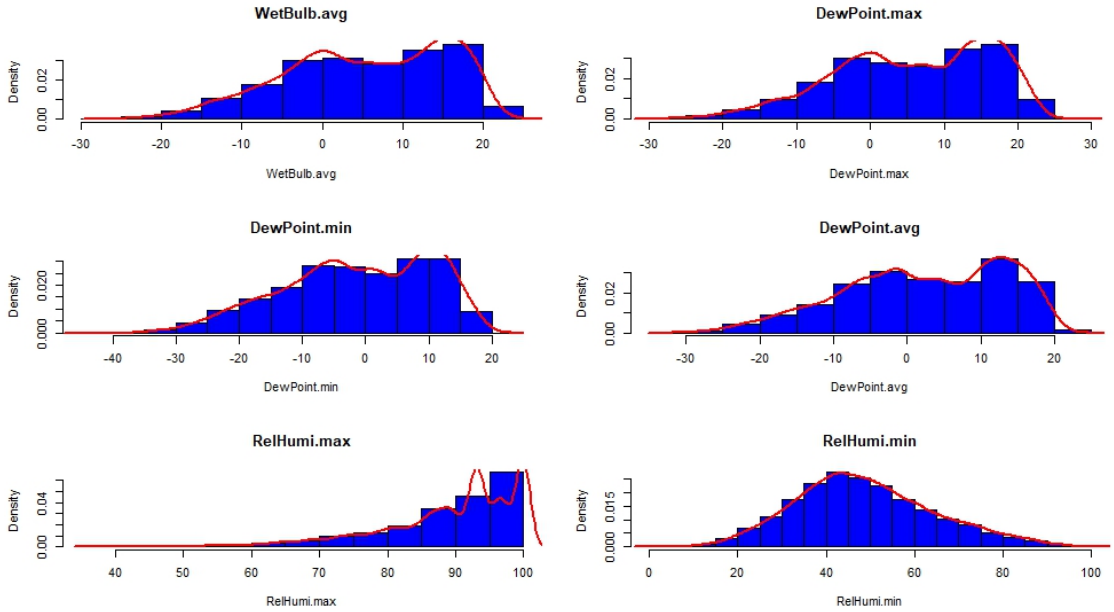

Figure 37: Distribution of variables

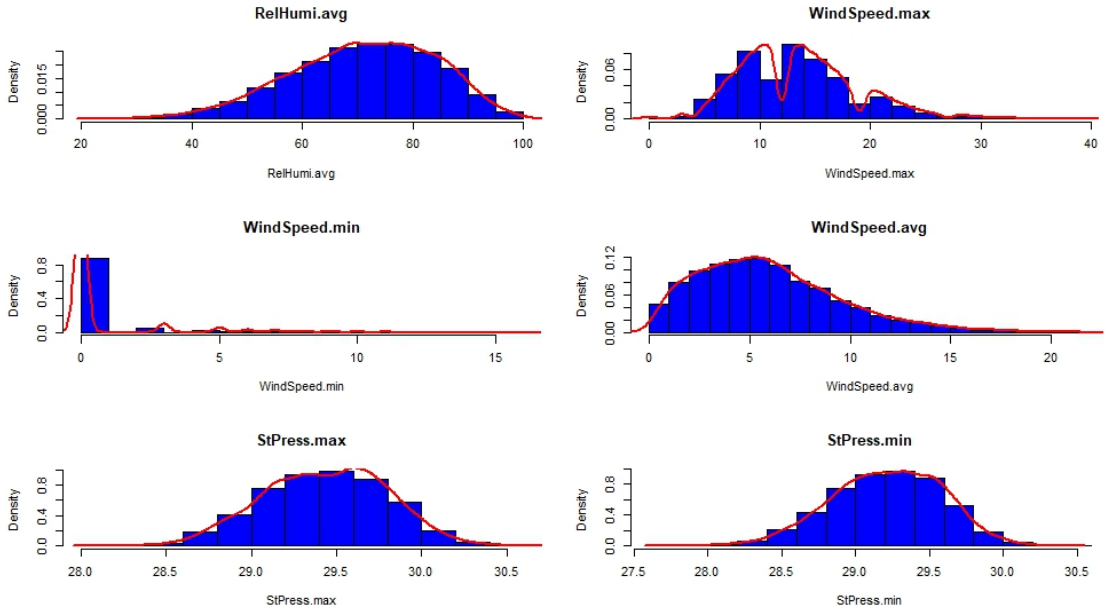

Figure 38: Distribution of variables

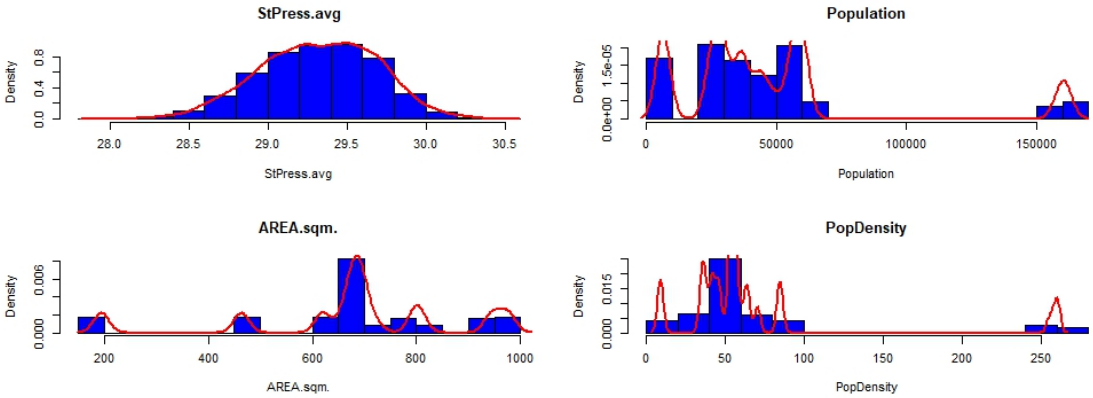

Figure 39: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

11.1.2 Scatter Plots between Peak Demand and other variables

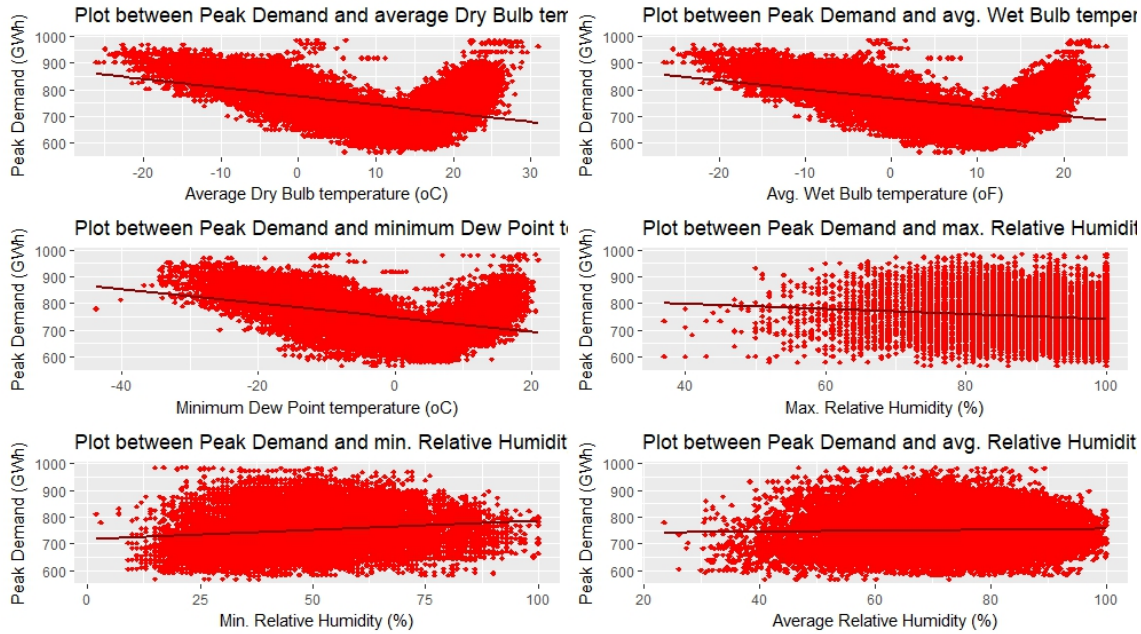

Figure 40: Scatter Plots between Peak Demand and other variables

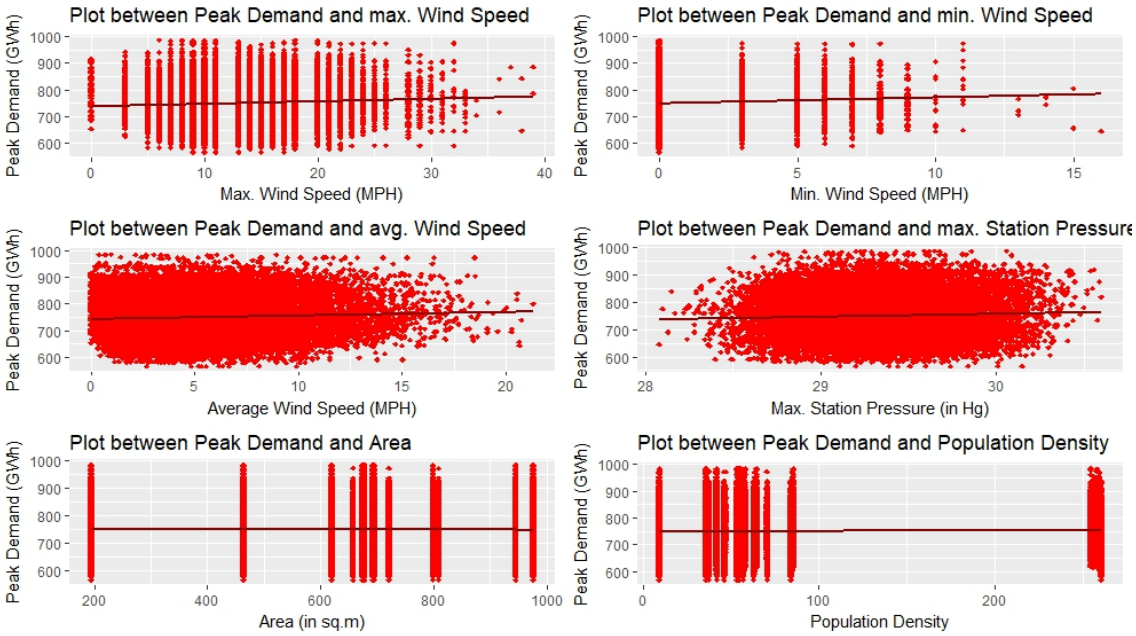

Figure 41: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 10-20oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

11.1.3 Violin Plots

Our target variable, Peak.Demand, was plotted in violin plots against our important variables - the ones we selected for the models.

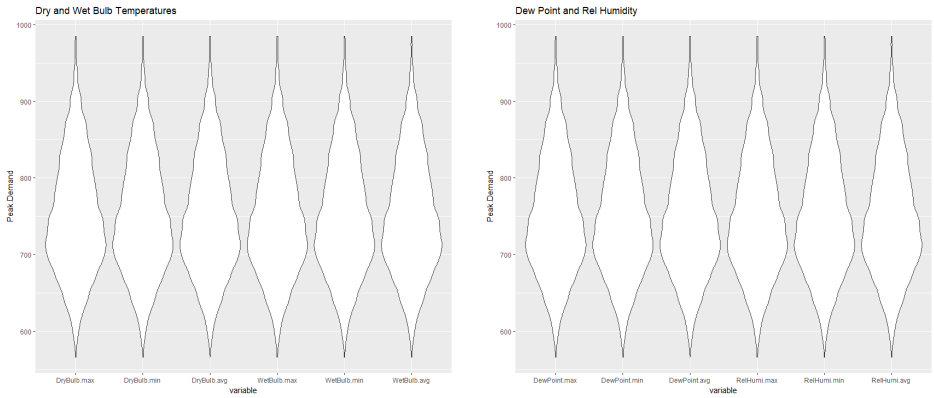

fig1: temperature

fig2: humidity

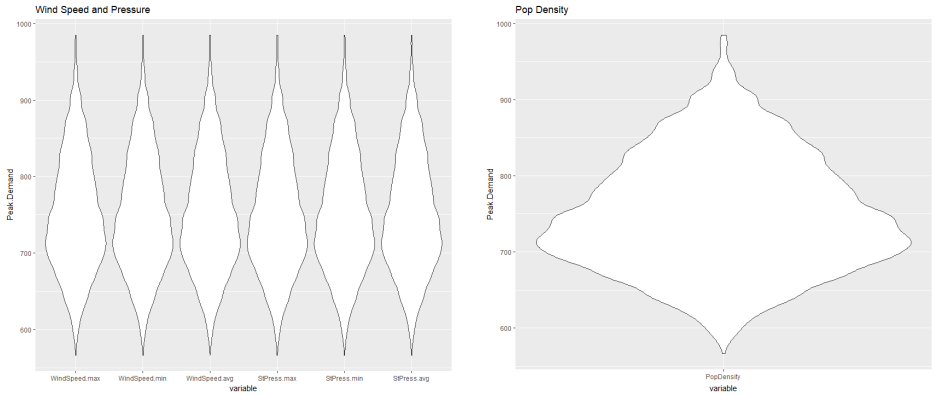

fig3: windspeed

fig4: PopulationDensity

Figure 42: Violin Plots Vermont

We see VT has a distribution of the climate values not as concentrated in an specific range as the other states. The population density is also smaller than the other states.

11.1.4 Correlation plot

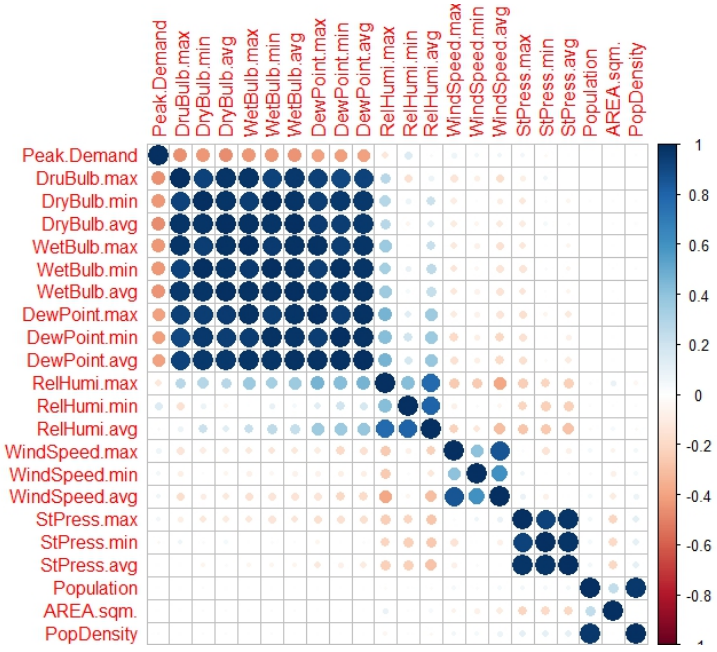

Figure 43: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.
Hence, transformation and other techniques will be required while building the predictive model.

11.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 models fit.

	RMSE.train Mean	**RMSE.test Mean**
MLR	660.2638	63.30829
Ridge Model	661.367	66.04832
Lasso Model	722.8522	63.31468
GAM	951.4673	44.56402
Decision Trees	47.8617	48.1811
Random Forest	13.3521	29.6096
MARS Unpruned	4.3072	4.2548
MARS Pruned	4.3081	4.2546
SVM	27.78869	36.1615
BartMachine	40.42754	44.76609

Vermont is also a dataset with less input data, and we got better results, like in Rhode Island.

11.3 Best Model- MARS

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = value

b) MARS makes no assumptions about the underlying functional relationships between dependent and independent variables. In general, the splines are connected smoothly together, and these piecewise curves (polynomials), also known as basis functions (BFs), result in a flexible model that can handle both linear and nonlinear behavior.

c) MARS generates BFs by stepwise searching overall possible univariate candidate knots and across interactions among all variables. An adaptive regression algorithm is adopted for automatically selecting the knot locations. The MARS algorithm involves a forward phase and a backward phase. The forward phase places candidate knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares residual error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and overfitted model. The backward phase involves deleting the redundant BFs that made the least contributions

d) MARS can be considered to be more computationally efficient than other models, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding and pruning. In addition, the developed MARS models are easier to be interpreted. Furthermore, since MARS explicitly defines the knots for each design input variables, the model enables engineers to have an insight and understanding of where significant changes in the data may occur.

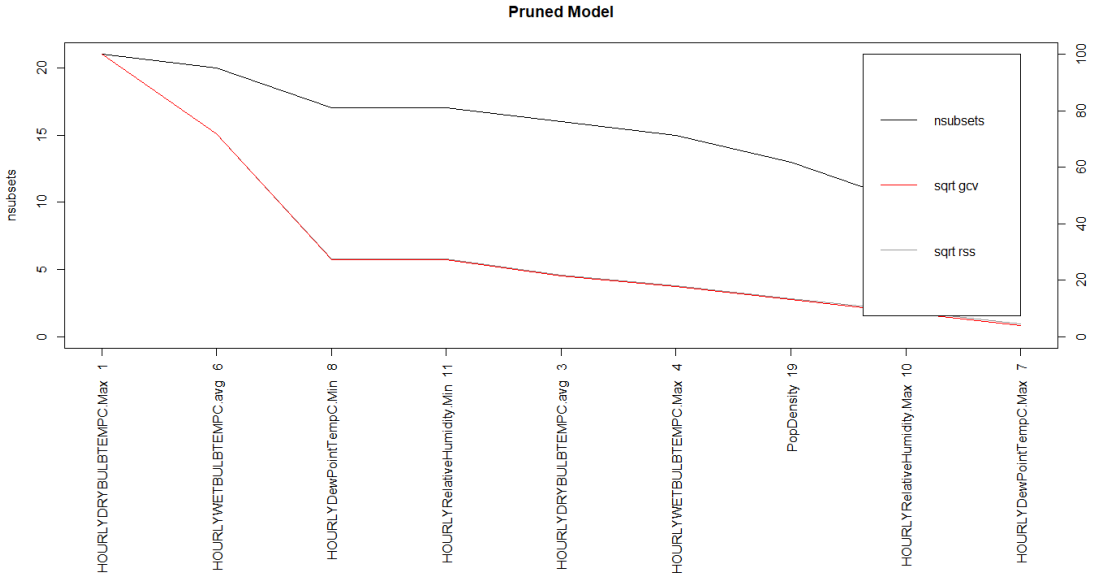

Figure 44: Variable Selection Pruned

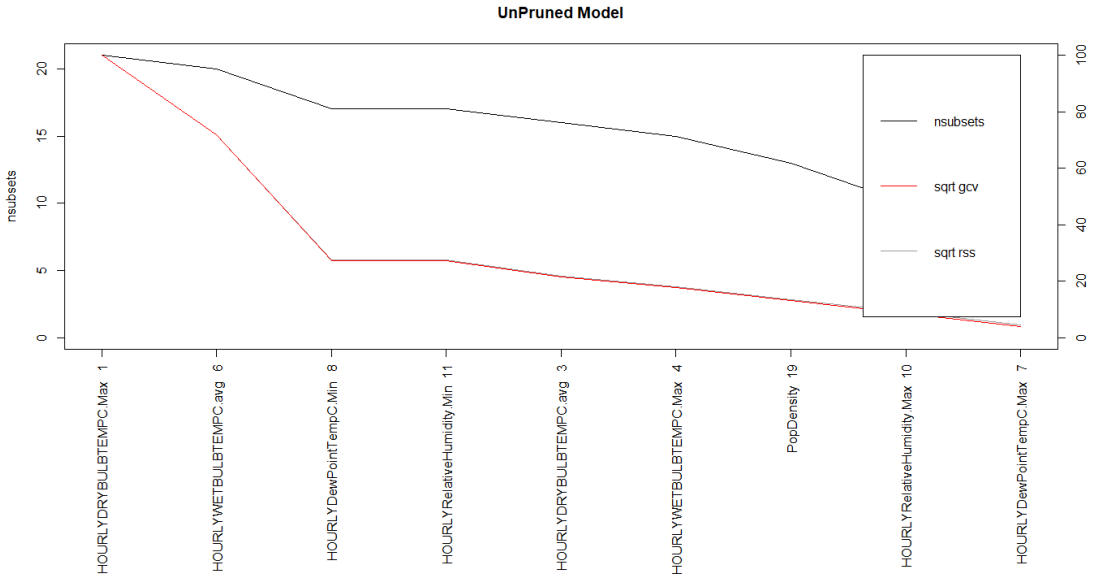

Figure 45: Variable Selection Unpruned

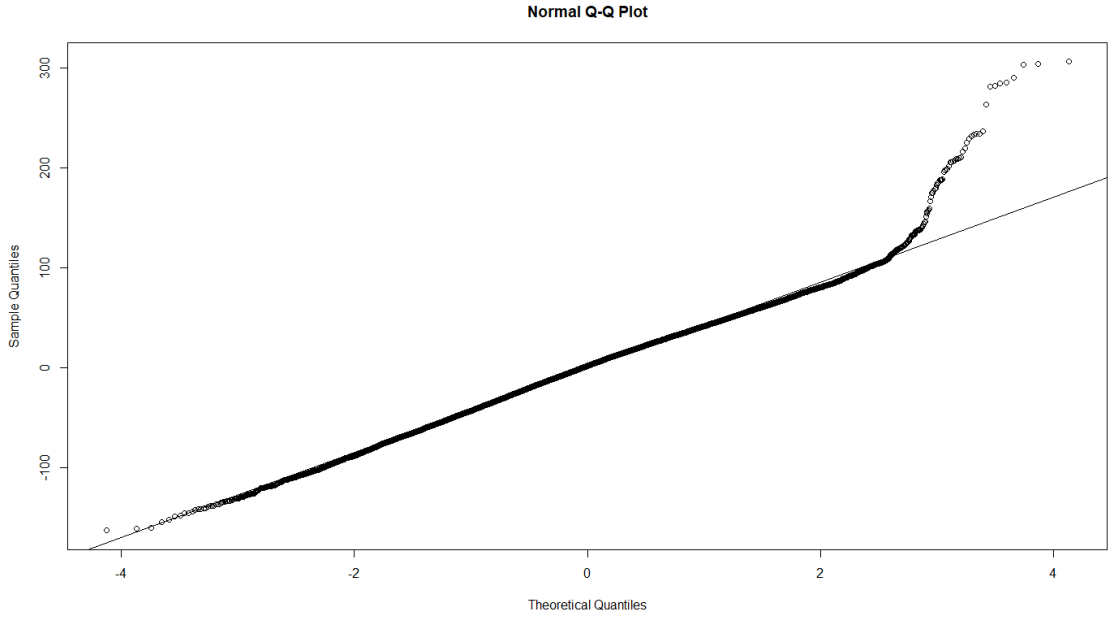

Figure 46: Normal QQ Plot for Pruned

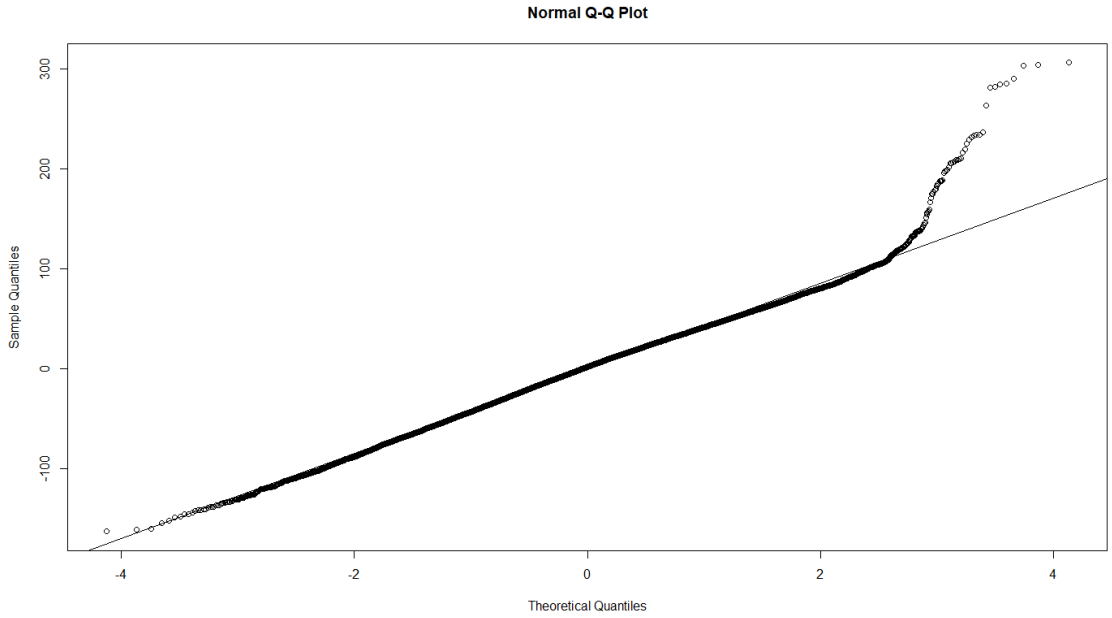

Figure 47: Normal QQ plot for Unpruned

12 New Hampshire

12.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

12.1.1 Density Plots

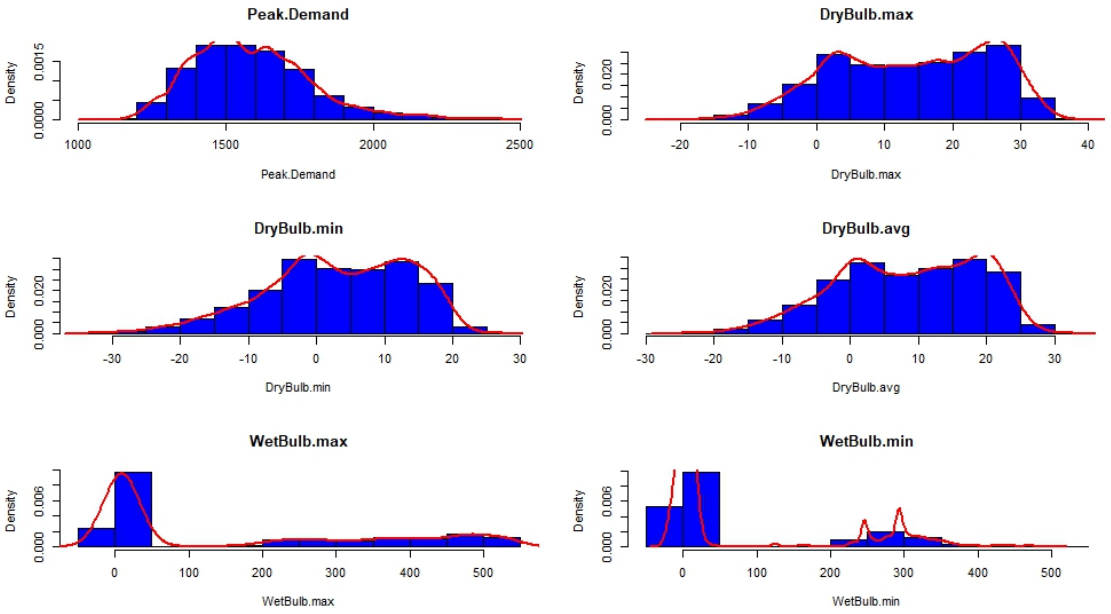

Figure 48: Distribution of variables

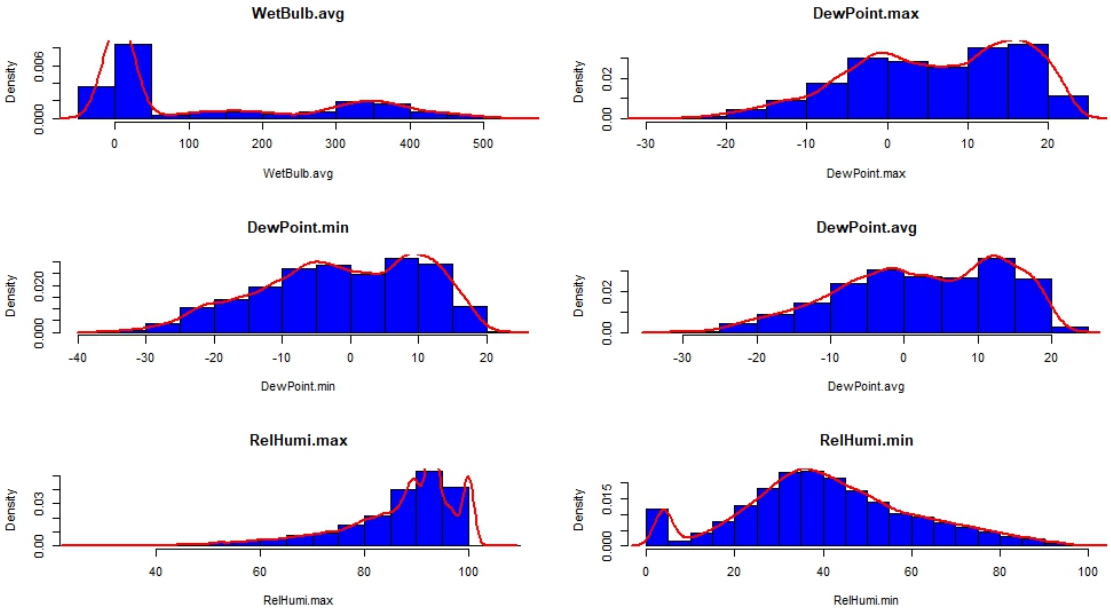

Figure 49: Distribution of variables

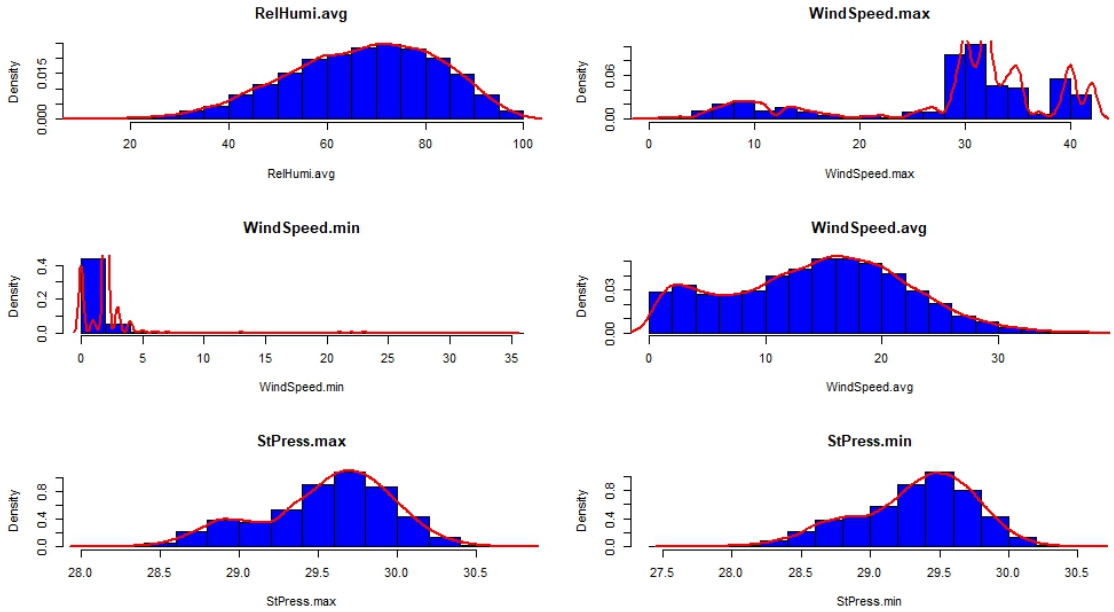

Figure 50: Distribution of variables

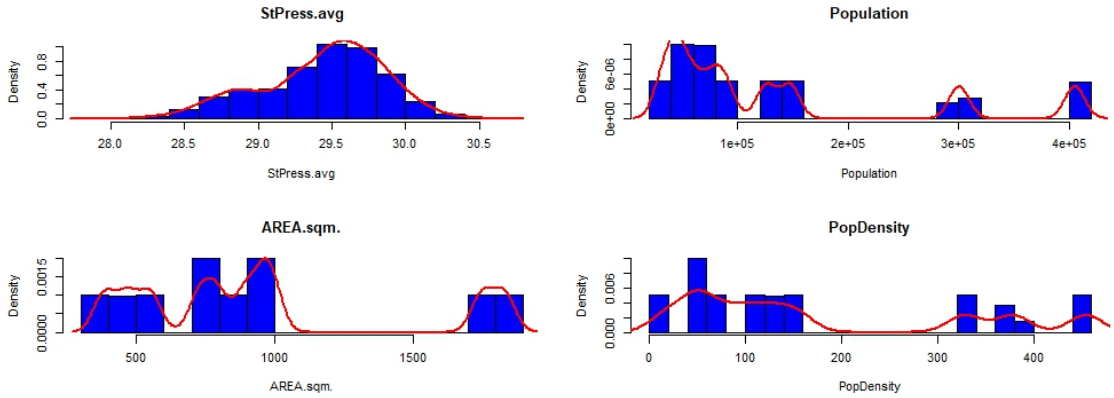

Figure 51: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

12.1.2 Scatter Plots between Peak Demand and other variables

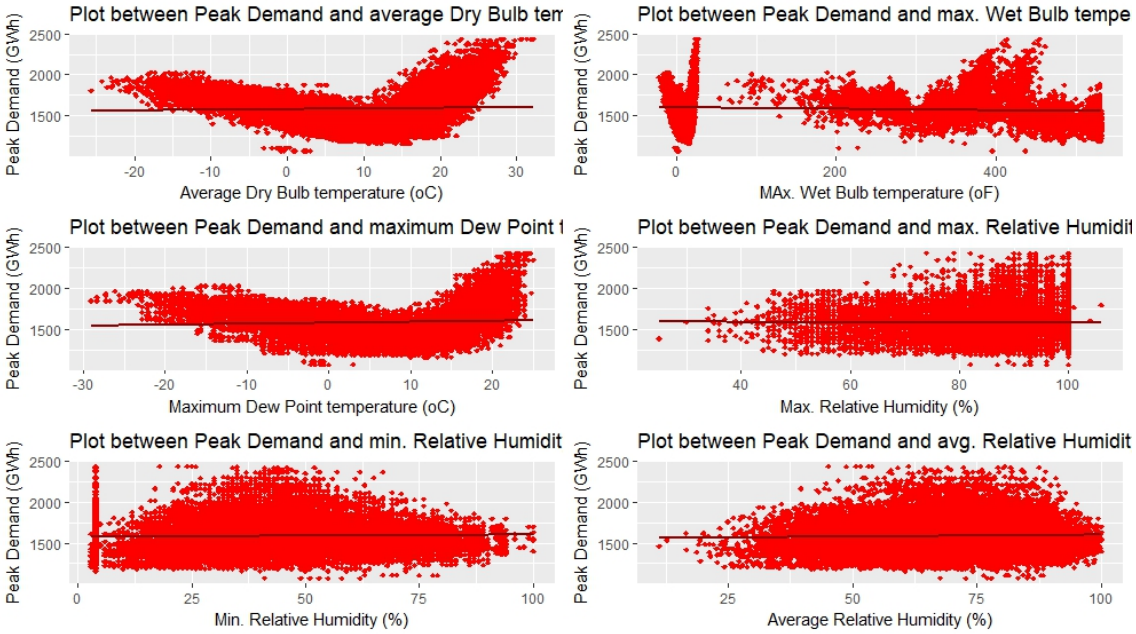

Figure 52: Scatter Plots between Peak Demand and other variables

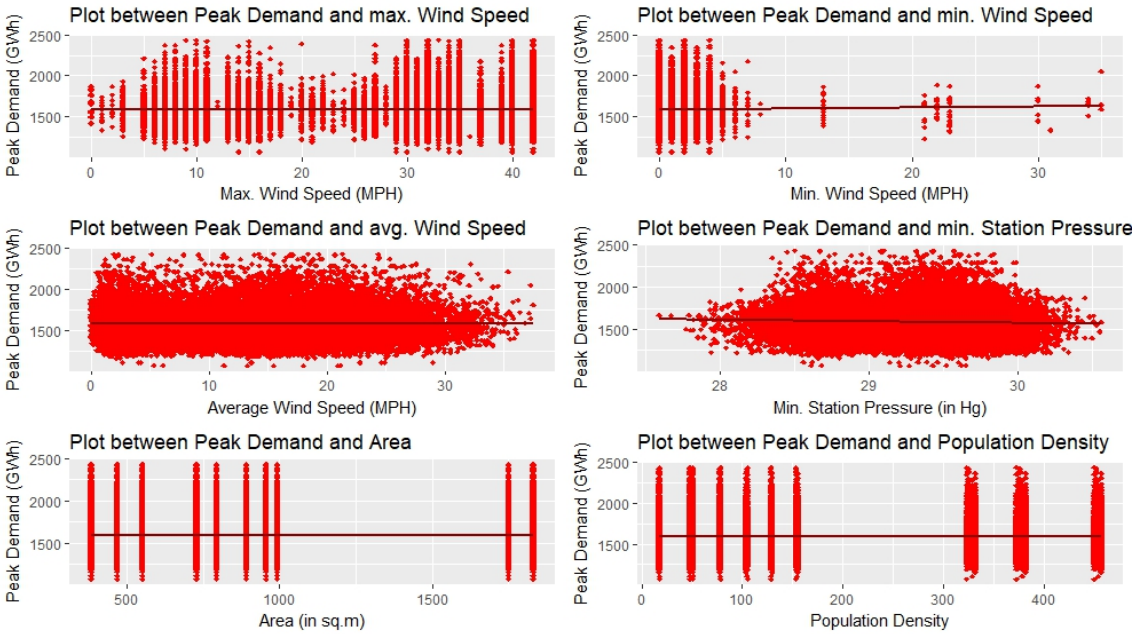

Figure 53: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 10-20oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

12.1.3 Violin Plots

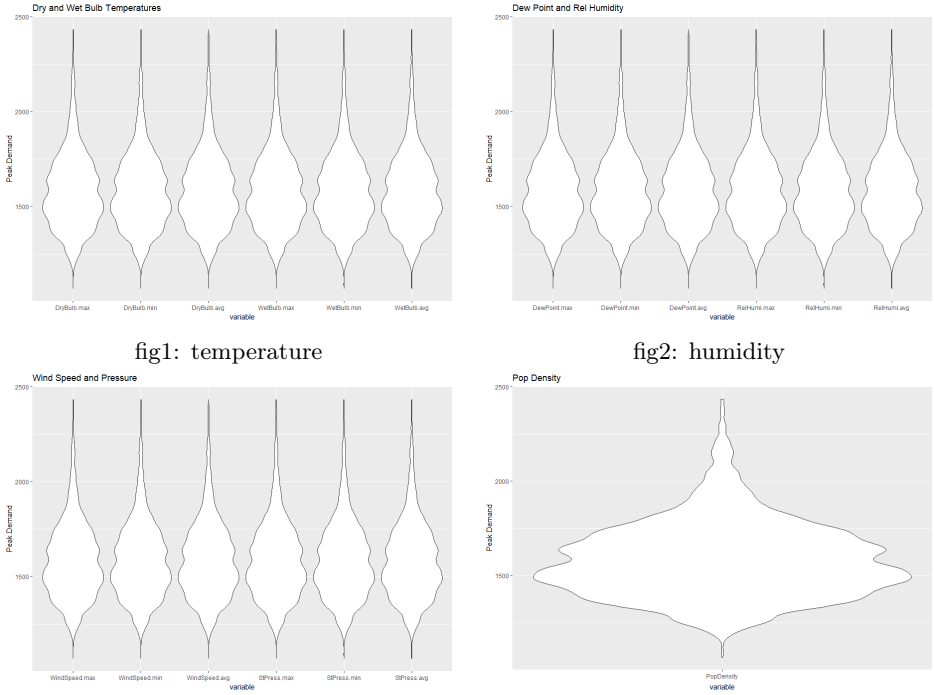

fig1: temperature

fig2: humidity

fig3: windspeed

fig4: PopulationDensity

Figure 54: Violin Plots New Hampshire

We see NH has concentrated climate values in an specific range as the other states. The population density is also smaller than the other states.

12.1.4 Correlation plot

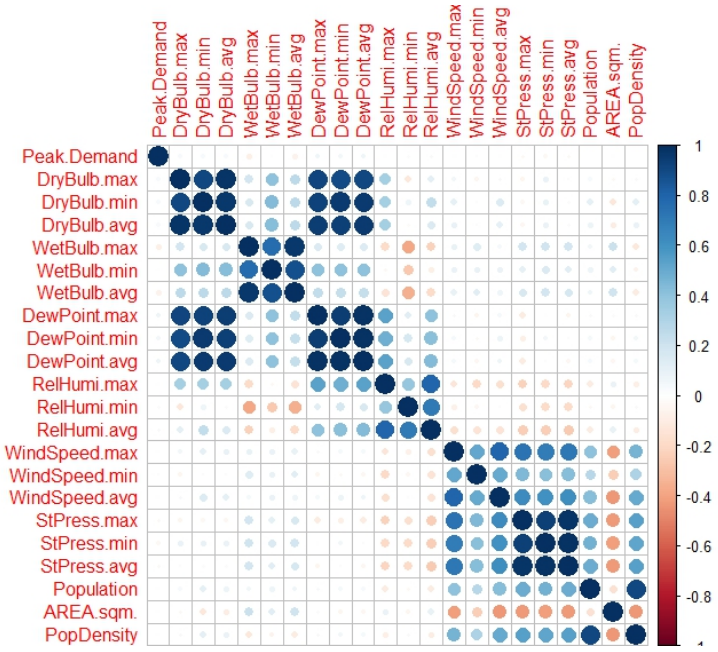

Figure 55: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.

Hence, transformation and other techniques will be required while building the predictive model.

12.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 models fit.

	RMSE.train Mean	**RMSE.test Mean**
MLR	194.2134	194.2582
Ridge Model	198.7856	198.8303
Lasso Model	203.8772	194.2595
GAM	255.151	116.2854
Decision Trees	121.6158	122.9325
Random Forest	37.8412	81.9748
MARS Unpruned	10.8187	10.5532
MARS Pruned	10.8187	10.5532
SVM	71.04172	110.167
BartMachine	97.53213	109.4414

New Hampshire is also a dataset with less input data, and we got better results, like in Vermont and Rhode Island.

12.3 Best Model- MARS

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = value

b) MARS makes no assumptions about the underlying functional relationships between dependent and independent variables. In general, the splines are connected smoothly together, and these piecewise curves (polynomials), also known as basis functions (BFs), result in a flexible model that can handle both linear and nonlinear behavior.

c) MARS generates BFs by stepwise searching overall possible univariate candidate knots and across interactions among all variables. An adaptive regression algorithm is adopted for automatically selecting the knot locations. The MARS algorithm involves a forward phase and a backward phase. The forward phase places candidate knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares residual error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and overfitted model. The backward phase involves deleting the redundant BFs that made the least contributions

d) MARS can be considered to be more computationally efficient than other models, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding and pruning. In addition, the developed MARS models are easier to be interpreted. Furthermore, since MARS explicitly defines the knots for each design input variables, the model enables engineers to have an insight and understanding of where significant changes in the data may occur.

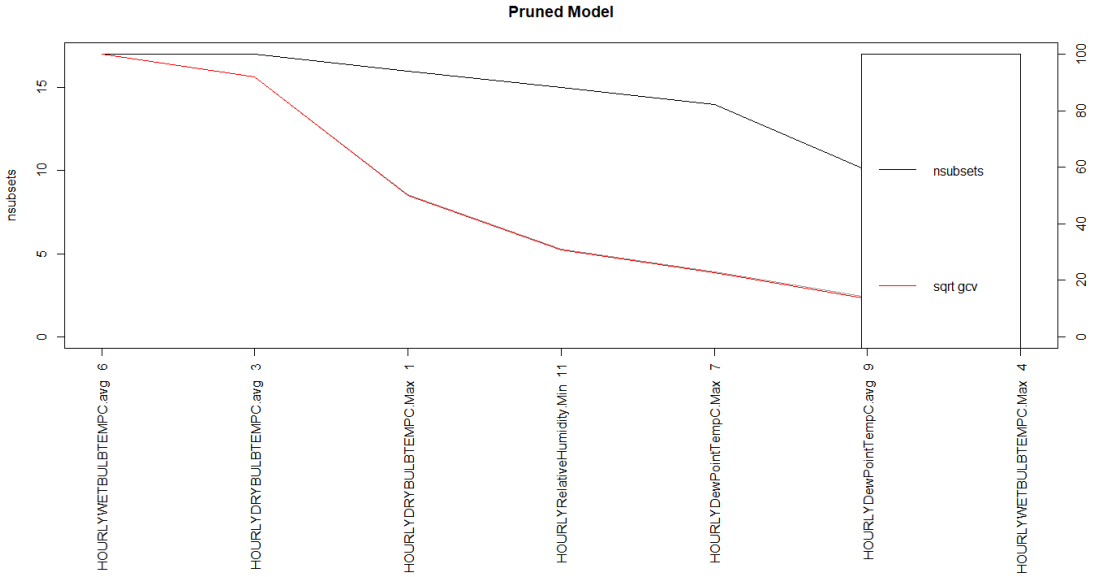

Figure 56: Variable Selection Pruned

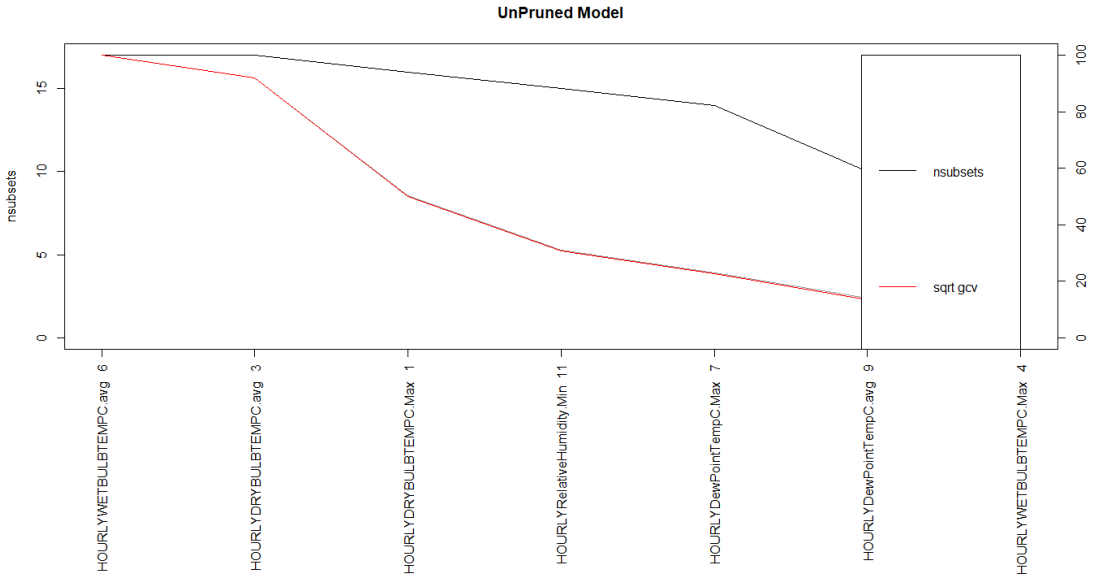

Figure 57: Variable Selection Unpruned

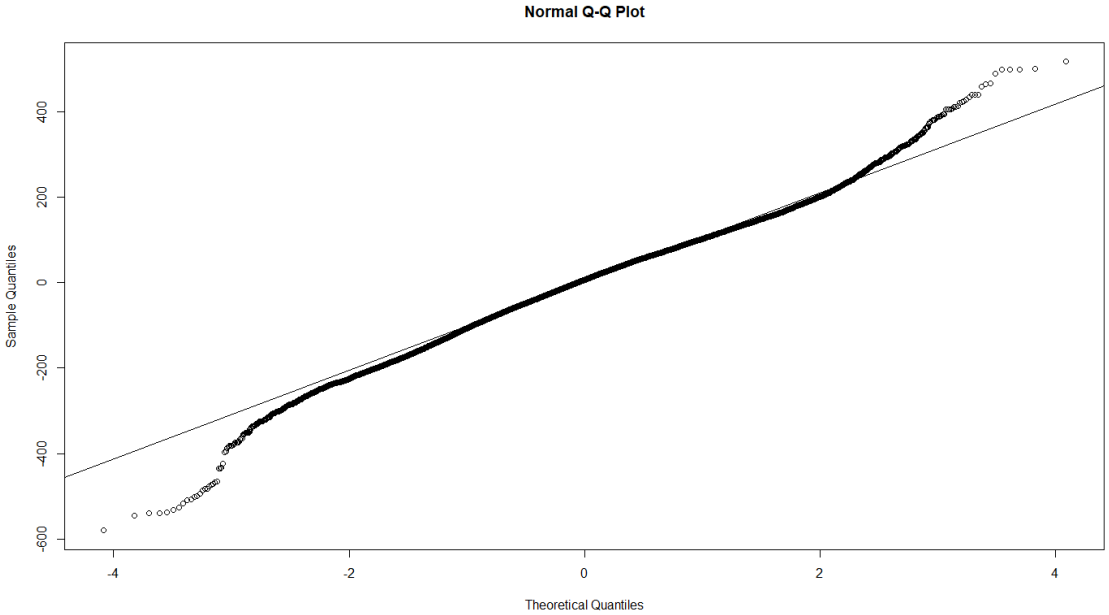

Figure 58: Normal QQ Plot for Pruned

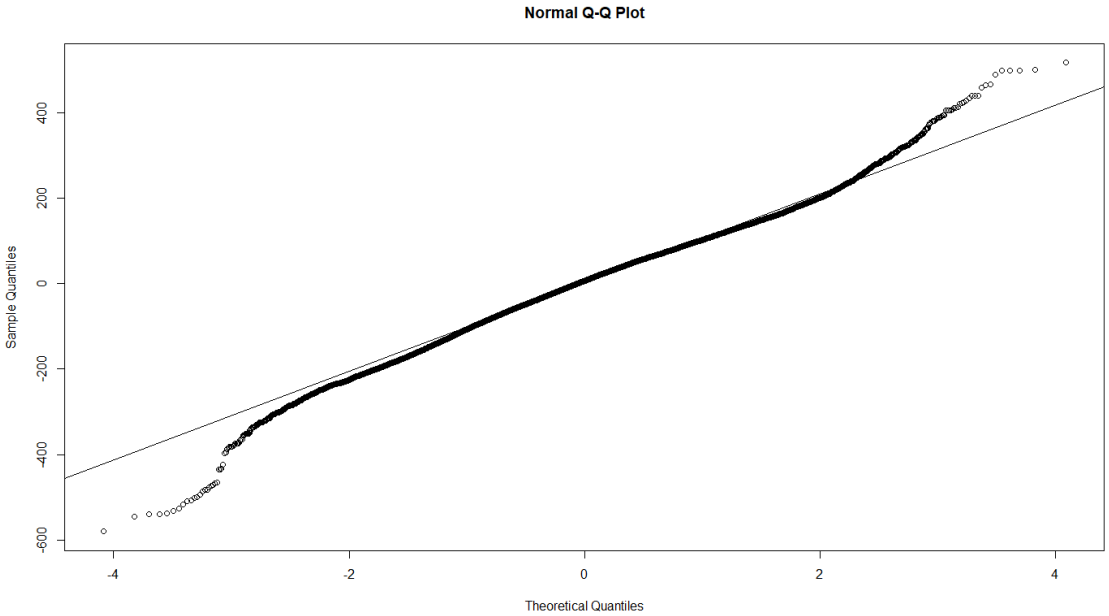

Figure 59: Normal QQ plot for Unpruned

13 Maine

13.1 Exploratory Data Analysis

We study the Density, Scatter and Violin Plots to better understand the distribution of the variables

13.1.1 Density Plots

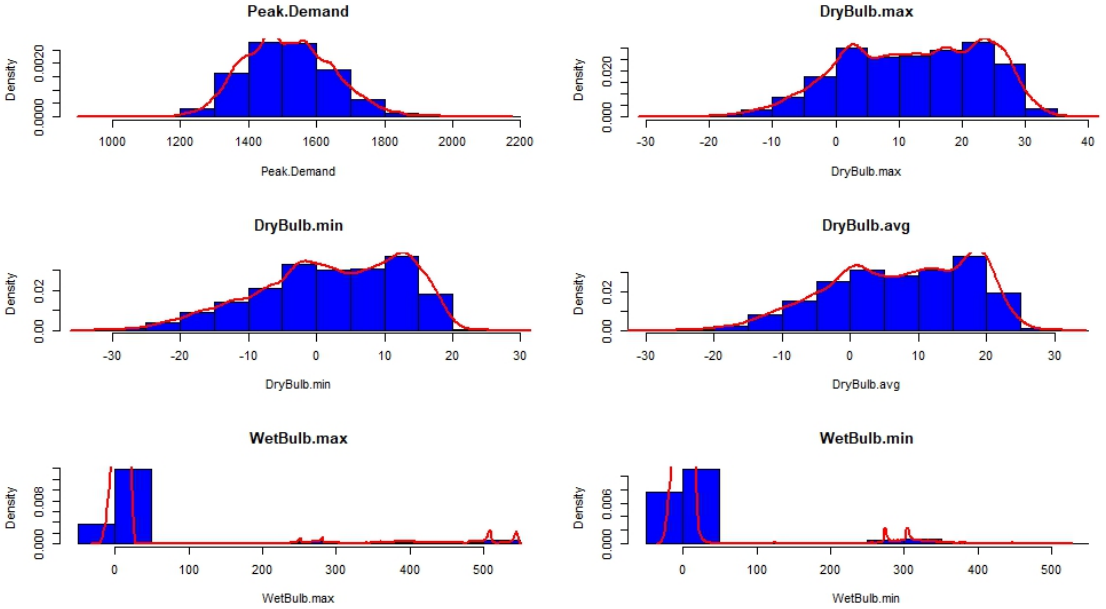

Figure 60: Distribution of variables

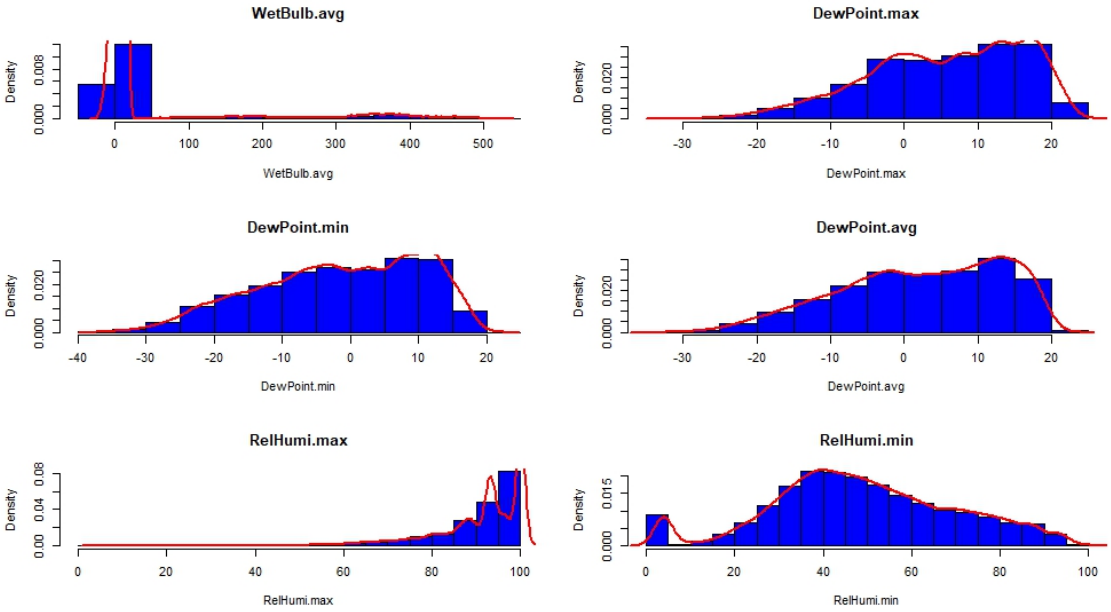

Figure 61: Distribution of variables

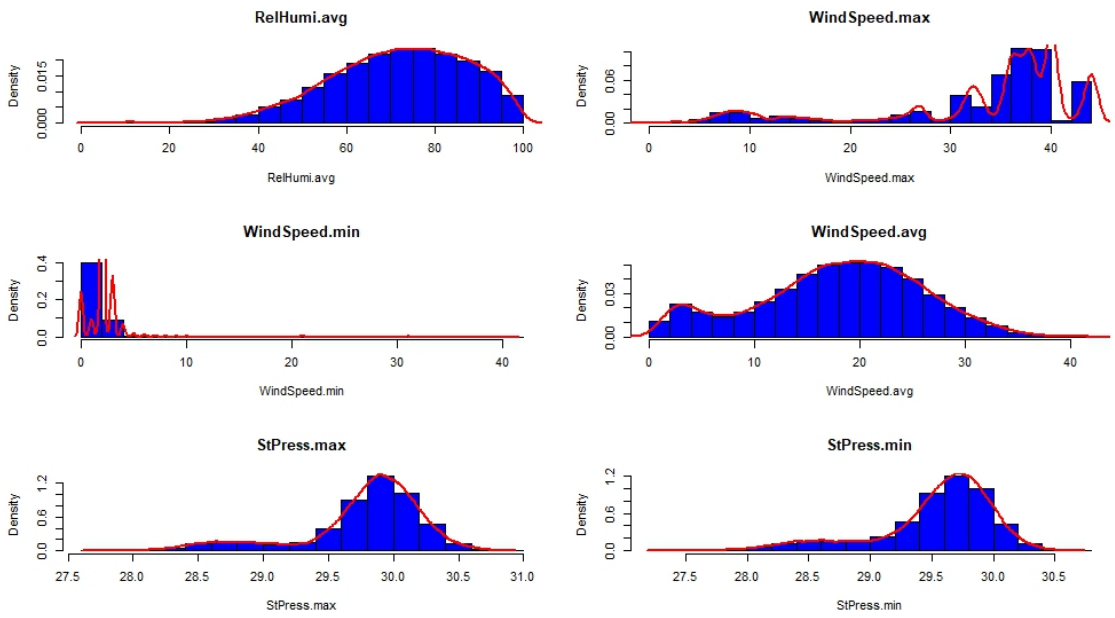

Figure 62: Distribution of variables

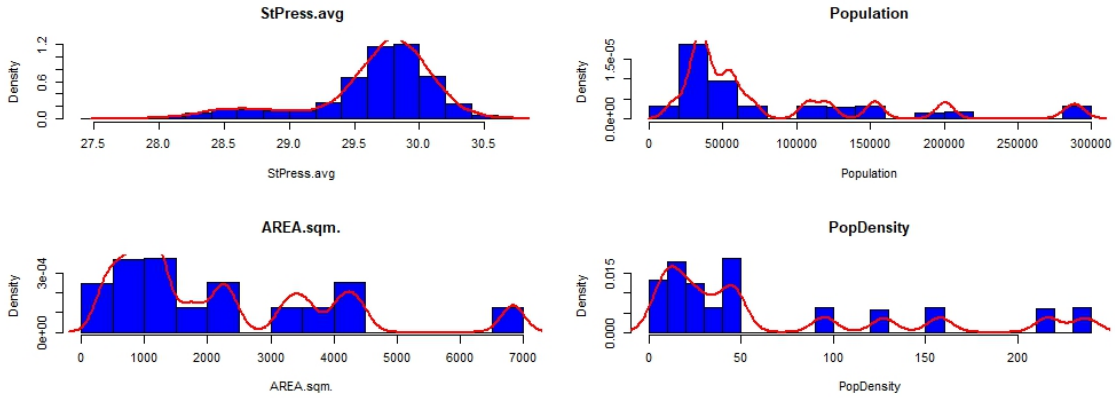

Figure 63: Distribution of variables

The above plots show the concentration of the values of dependent variables

Peak Demand, Dry Bulb temperature, Wet Bulb temperature, Dew Point, Relative Humidity, Station Pressure have a concentrated distribution whereas others have a wide range.
The histogram bins help us visualize the data better and show the frequency distribution of the data.

13.1.2 Scatter Plots between Peak Demand and other variables

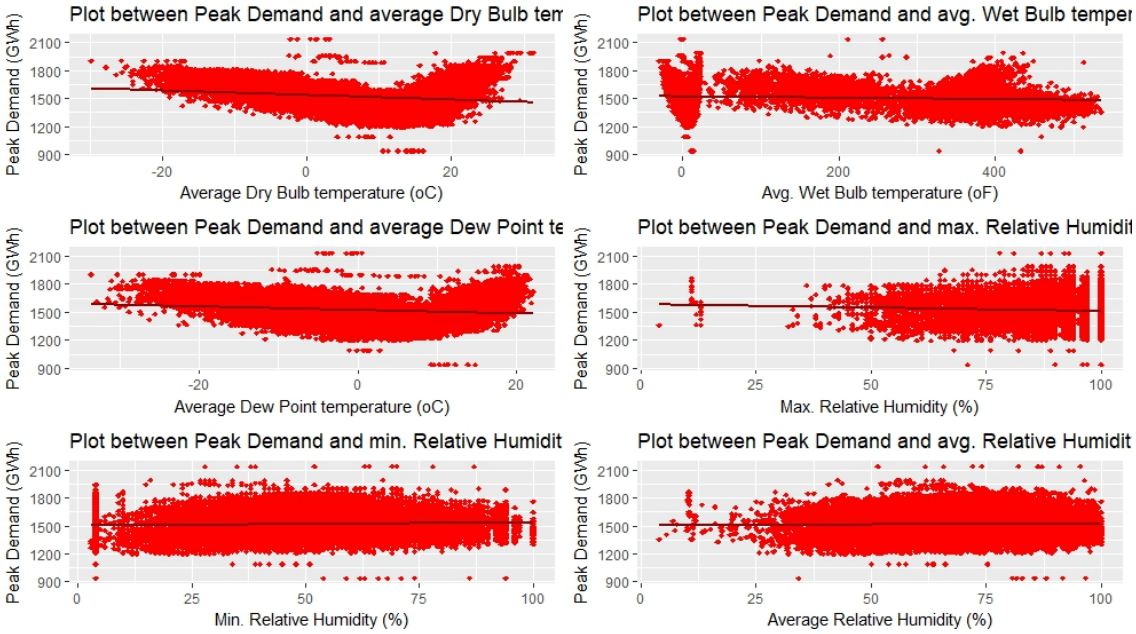

Figure 64: Scatter Plots between Peak Demand and other variables

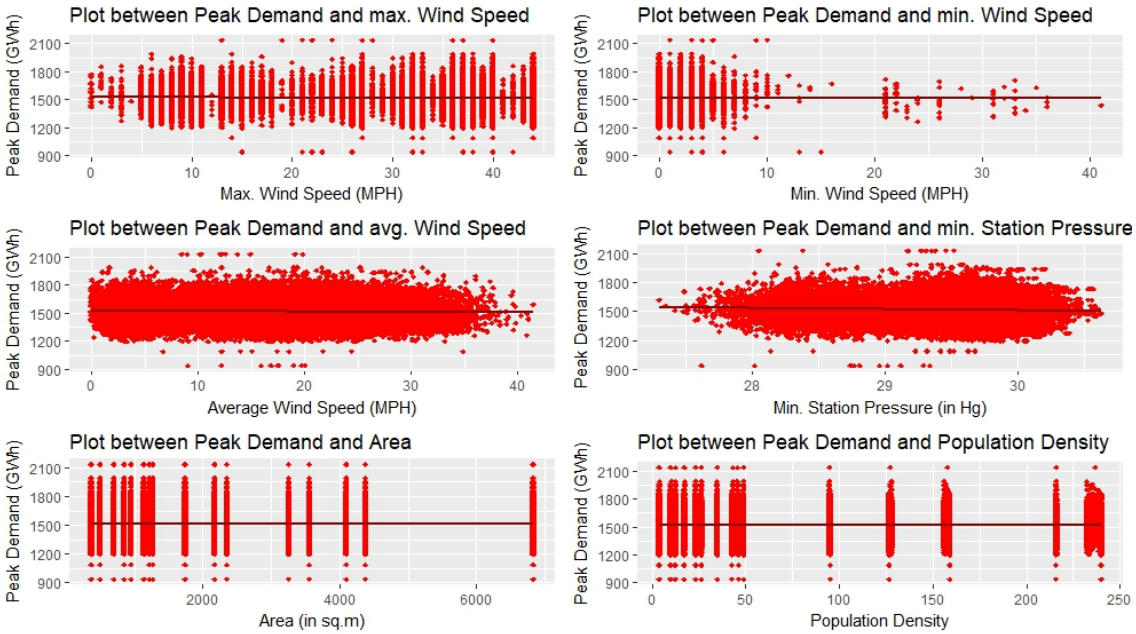

Figure 65: Scatter Plots between Peak Demand and other variables

As seen from above plots we see that energy consumption increases when dry bulb temperature is low (below 0oC) or high (above 20oC).
The same can be said about the relationship between Dew Point temperature and Peak Demand.

There is a sudden increase in energy consumption when the Wet Bulb temperature is in the range of 10-20oF

The relationship between peak demand and Relative Humidity, Wind Speed and Station Pressure is quite vague and requires further analysis

13.1.3 Violin Plots

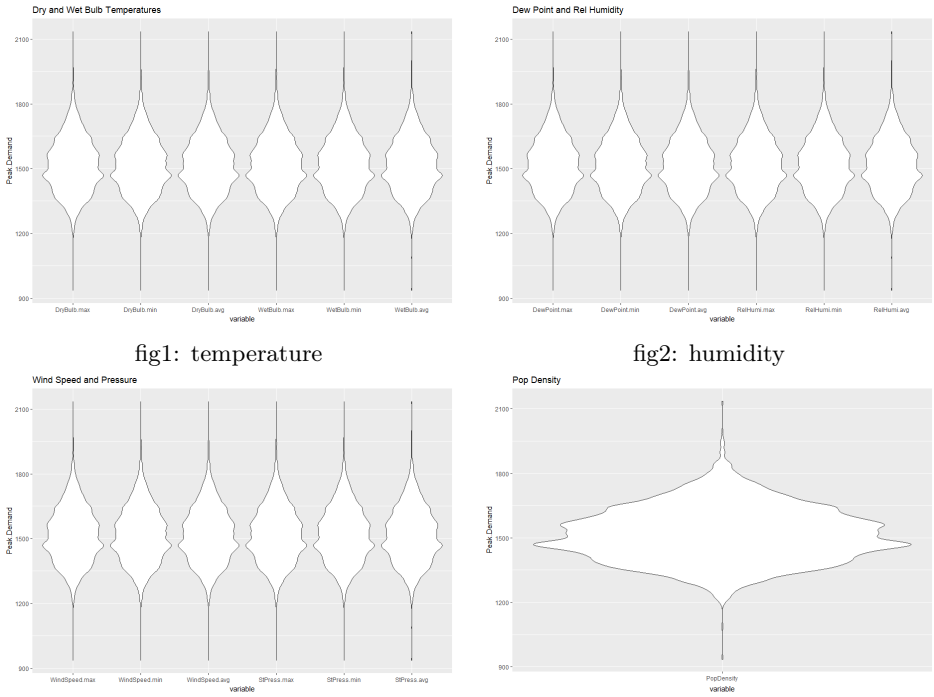

fig1: temperature

fig2: humidity

fig3: windspeed

fig4: PopulationDensity

Figure 66: Violin Plots Maine

We see ME has a concentrated distribution of the climate values in a specific range as the other states. The population density is wider than the other states.

13.1.4 Correlation plot

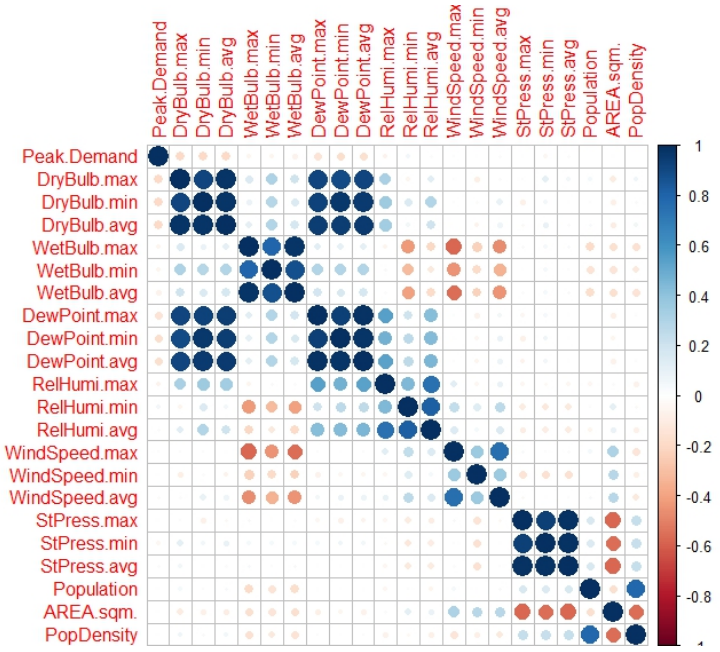

Figure 67: Correlation plot

As seen from above, there isn't much correlation between Peak Demand vs Relative Humidity, Wind Speed and Station Pressure. Hence, it requires further analysis.

Hence, transformation and other techniques will be required while building the predictive model.

13.2 Results

Results are from cross validation fitting with 10 folds, and we show the mean RMSE of the 10 models fit.

	RMSE.train Mean	**RMSE.test Mean**
MLR	122.6973	122.7127
Ridge Model	122.9657	122.9789
Lasso Model	130.423	122.7131
GAM	159.934	82.8929
Decision Trees	88.6997	89.0109
Random Forest	29.0483	72.5581
MARS Unpruned	4.3072	8.4585
MARS Pruned	8.0118	8.4578
SVM	58.46071	81.87843
BartMachine	74.08006	86.52305

As seen from the table above, MARS is the best model for the state of ME

13.3 Best Model- MARS

The main criteria of comparison between different models was the RMSE values. The RMSE value shows how a model performs on the test data set after learning from the training dataset. The lower the RMSE the higher is the predictive power of the model.

The following were the justification points for model selection:
a) It gave the lowest RMSE = value

b) MARS makes no assumptions about the underlying functional relationships between dependent and independent variables. In general, the splines are connected smoothly together, and these piecewise curves (polynomials), also known as basis functions (BFs), result in a flexible model that can handle both linear and nonlinear behavior.

c) MARS generates BFs by stepwise searching overall possible univariate candidate knots and across interactions among all variables. An adaptive regression algorithm is adopted for automatically selecting the knot locations. The MARS algorithm involves a forward phase and a backward phase. The forward phase places candidate knots at random positions within the range of each predictor variable to define a pair of BFs. At each step, the model adapts the knot and its corresponding pair of BFs to give the maximum reduction in sum-of-squares residual error. This process of adding BFs continues until the maximum number is reached, which usually results in a very complicated and overfitted model. The backward phase involves deleting the redundant BFs that made the least contributions

d) MARS can be considered to be more computationally efficient than other models, as the MARS algorithm builds flexible models using simpler linear regression and data-driven stepwise searching, adding and pruning. In addition, the developed MARS models are easier to be interpreted. Furthermore, since MARS explicitly defines the knots for each design input variables, the model enables engineers to have an insight and understanding of where significant changes in the data may occur.

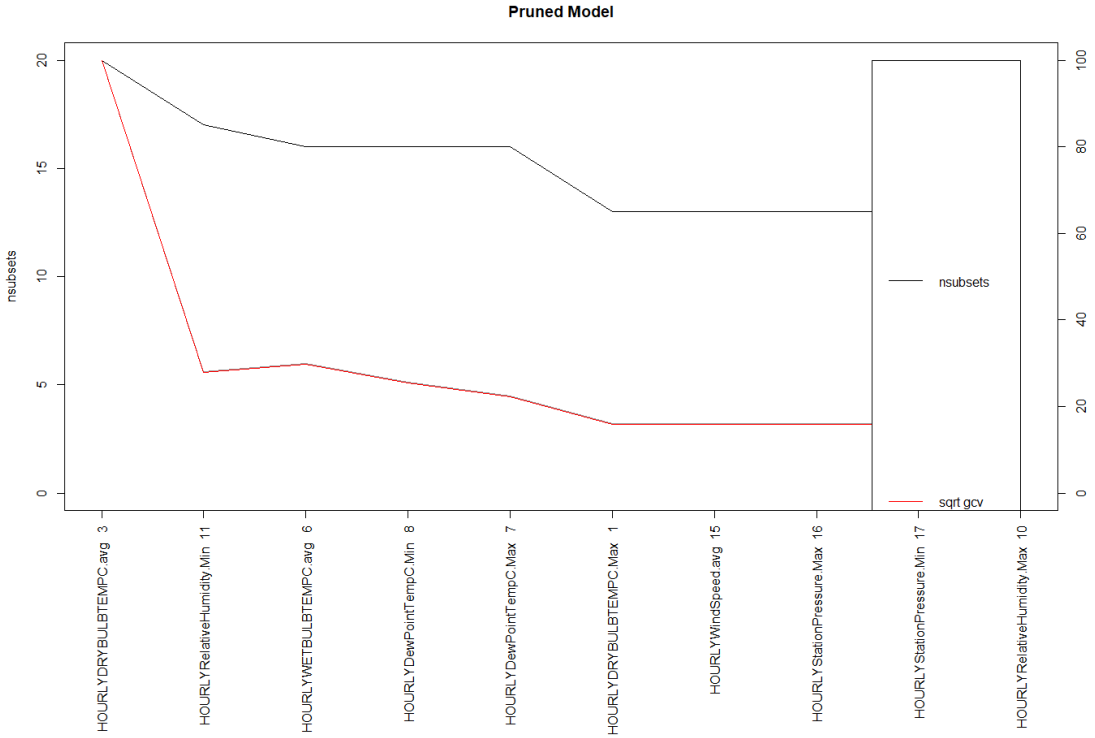

Figure 68: Variable Selection Pruned

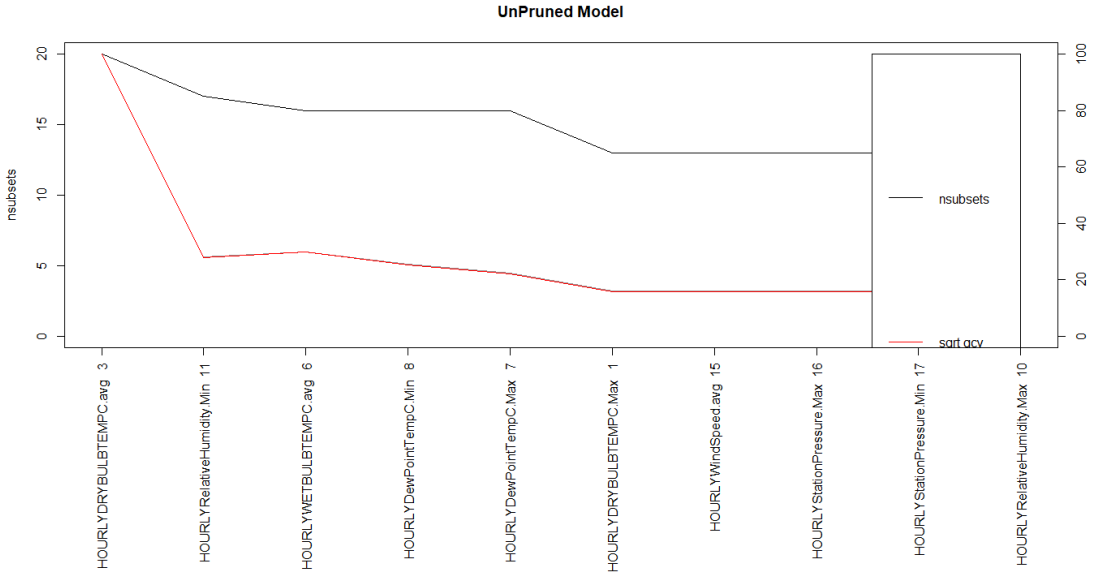

Figure 69: Variable Selection Unpruned

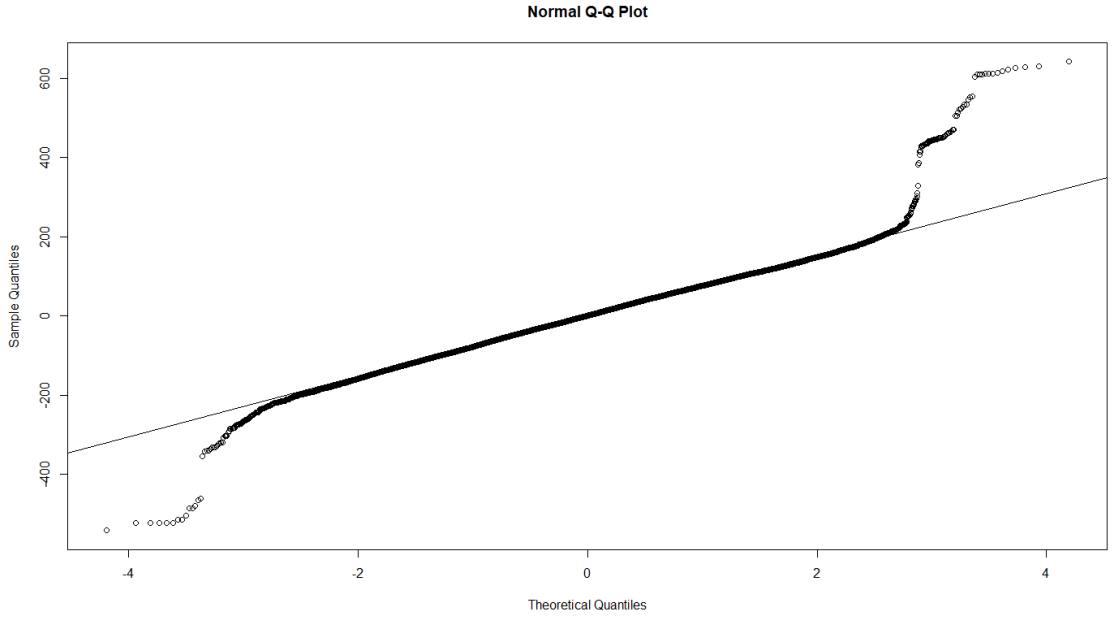

Figure 70: Normal QQ Plot for Pruned

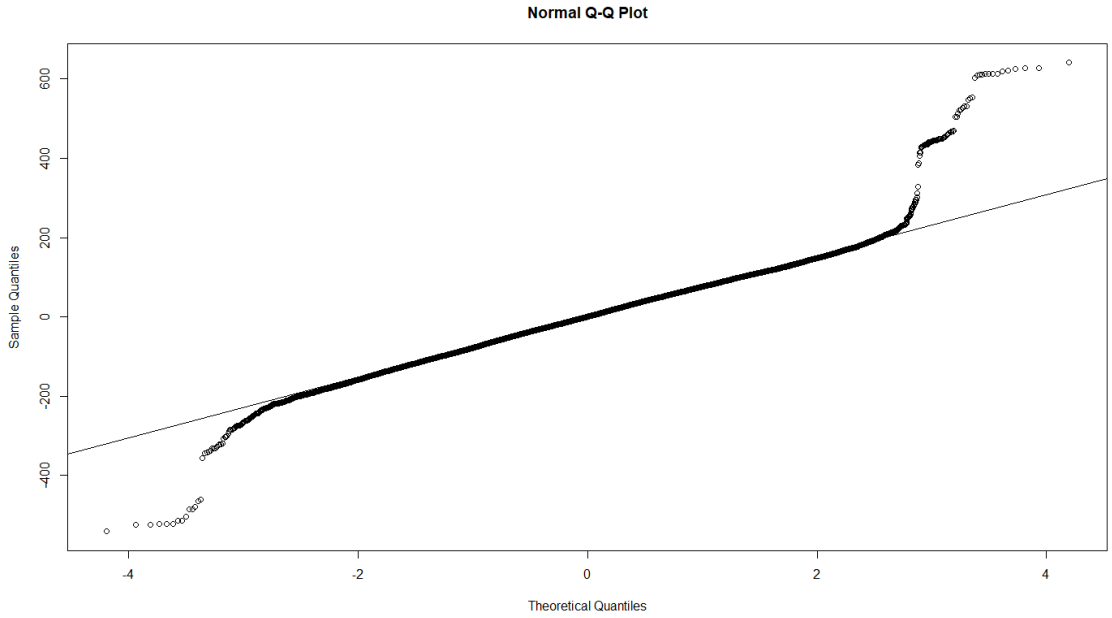

Figure 71: Normal QQ plot for Unpruned

14 Conclusion

10 models are run for all the six states and the best model is selected based on the parameter - Out of Sample RMSE. The lower the RMSE value the better is the prediction of the Peak Demand. A better prediction for the daily peak demand is needed so that the states are ready with the infrastructure in times of sudden surge in electricity consumption.
Most of the models also help us to select which parameters affect the response more than others. The parameters affecting each state's peak demand is explained below.

As seen, it's mostly the climatic parameters and the population density that affects the electricity demand. This is correct intuitively as well and the same is expalined by our models.

14.1 Conneticut

As seen from the RMSE table, Random Forest gives the best prediction for Peak Demand.

The predictors based on Node Purity that have been selected here are DryBulb.avg, WetBulb.avg, DewPoint.avg, RelHumi.avg, StPressure.min, WindSpeed.avg Population, PopDensity, Area

The QQ plot shows that the residuals behave normally.

14.2 Rhode Island

As seen from the RMSE table, MARS Pruned gives the best prediction for Peak Demand.

The variable importance plots help to select the variables which is DryBulb.avg,DePoint.min, StPress.min, WetBulb.max, WindSpeed.avg,RelHumi.min and PopDensity.

The QQ plot shows that the residuals behave normally.

14.3 Massachusetts

As seen from the RMSE table, MARS gives the best prediction for Peak Demand.

The variable importance plots help to select the variables which is DryBulb.max,DePoint.min, StPress.avg, WetBulb.max, WindSpeed.avg,RelHumi.avg and PopDensity.
The QQ plot shows that the residuals behave normally.

14.4 Vermont

As seen from the RMSE table, MARS Pruned gives the best prediction for Peak Demand.

The variable importance plots help to select the variables which is DryBulb.max,DewPoint.min, StPress.avg, WetBulb.max, WindSpeed.avg,RelHumi.min and PopDensity.

The QQ plot shows that the residuals behave normally.

14.5 New Hampshire

As seen from the RMSE table, MARS gives the best prediction for Peak Demand.

The variable importance plots help to select the variables which is DryBulb.avg,DewPoint.max, StPress.avg, WetBulb.max, WindSpeed.avg,RelHumi.min and PopDensity.

The QQ plot shows that the residuals behave normally.

14.6 Maine

As seen from the RMSE table, MARS Pruned gives the best prediction for Peak Demand.

The variable importance plots help to select the variables which is DryBulb.avg, DewPoint.min, StPress.max, WetBulb.avg, WindSpeed.avg, RelHumi.min and PopDensity.

The QQ plot shows that the residuals behave normally.

References

[1] Local Climatological Data (LCD) | Data Tools | Climate Data Online (CDO) | National Climatic Data Center (NCDC).

[2] ISO New England - Energy, Load, and Demand Reports.

[3] Jonathan Koomey and Richard E. Brown. The role of building technologies in reducing and controlling peak electricity demand. Technical Report LBNL-49947, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), September 2002.

[4] P. Regulagadda, I. Dincer, and G. F. Naterer. Exergy analysis of a thermal power plant with measured boiler and turbine losses. *Applied Thermal Engineering*, 30(8):970–976, June 2010.

[5] Energy and Utility Analytics Market worth 3.41 Billion USD by 2021.

[6] Maximilian Auffhammer, Patrick Baylis, and Catherine H. Hausman. Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 114(8):1886–1891, February 2017.

[7] David J. Sailor and J. Ricardo Muñoz. Sensitivity of electricity and natural gas consumption to climate in the U.S.A.—Methodology and results for eight states. *Energy*, 22(10):987–998, October 1997.

[8] David J Sailor. Relating residential and commercial sector electricity loads to climate—evaluating state level sensitivities and vulnerabilities. *Energy*, 26(7):645–657, July 2001.

[9] Anna Carolina Kossmann de Menezes, Andrew Cripps, Richard A. Buswell, Jonathan A. Wright, and Dino Bouchlaghem. Estimating the energy consumption and power demand of small power equipment in office buildings. 2014.

[10] Eric Fox. UEsleicntricg/ GaLs o/ Waatder Research Data to Develop Long-Term Peak Demand Forecasts. page 33.

[11] Global Utility and Energy Analytics Market - Expected to be Worth USD2.9 Billion by 2020 - Research and Markets, November 2017.

[12] Adam Kapelner and Justin Bleich (R package). bartMachine: Bayesian Additive Regression Trees, April 2018.