FORECASTING ENERGY CONSUMPTION FOR SPACE CONDITIONING IN US USING MACHINE LEARNING

Author
Abrar Ul Farhan Mohammed
0030180435
Graduate student

Supervisor
Dr. Roshanak Nateghi
Assistant Professor
School of Industrial Engineering

PURDUE
UNIVERSITY

DATA:

For this assignment data used was obtained from United States Energy Information Administration (US E.I.A) which has data of individual buildings and related factors. Total related variables reported are '1119' and total dataset has record of 6720 buildings.

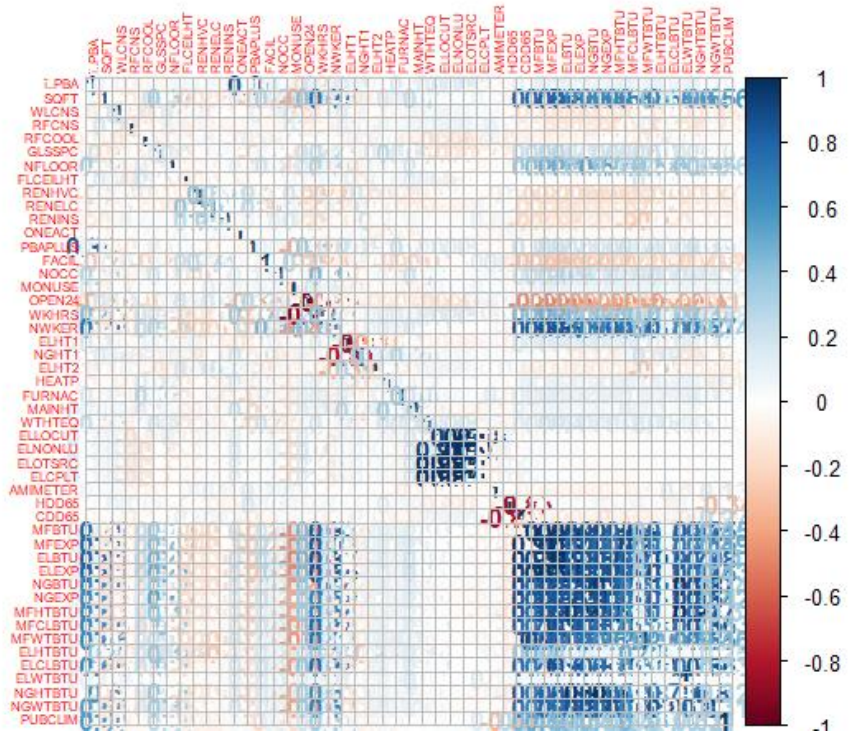
RESPONSE VARIABLE:

Total energy used for space conditioning (i.e. heating, cooling and water heating in the East South-Central Region)

Total energy=Electricity + Natural Gas + Fuel Oils + District energy source + wood + coal + Solar + other energy sources

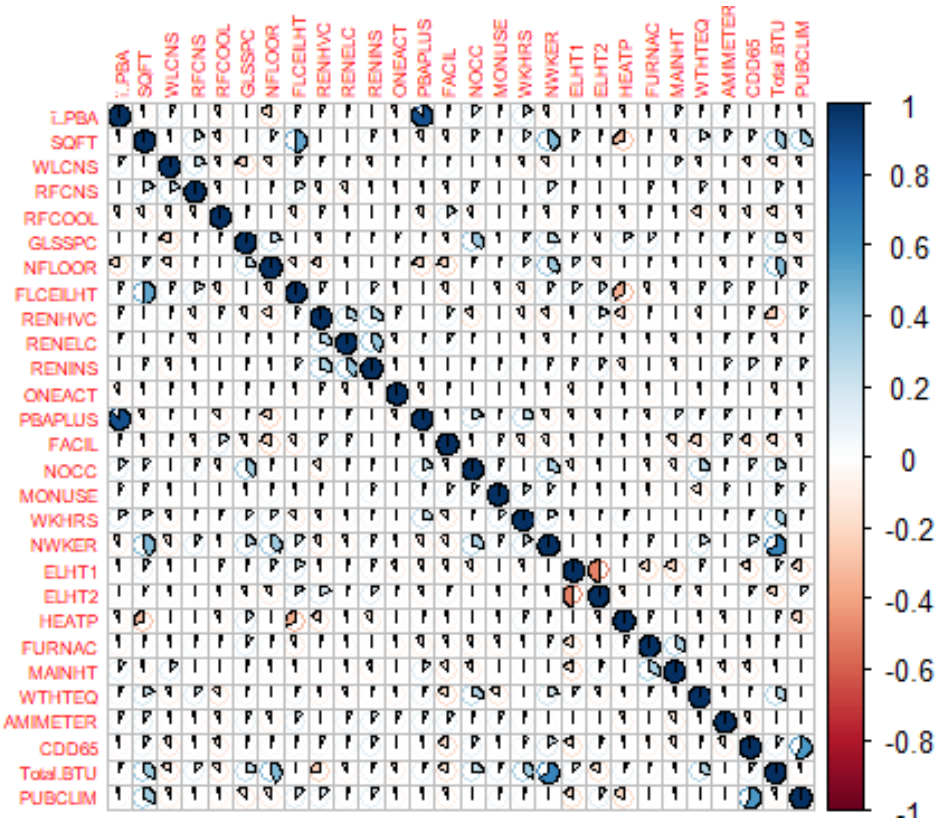
EXPLORATORY ANALYSIS/VARIABLE SELECTION:

Variables were selected based on there importance and effect they might have on total energy consumption. On total energy consumption total energy, energy from individual energy resources have been added to create a single response variable. The energy variable taken were in BTU units (British Thermal Units) to maintain uniformity between different energy variables. As the given geography of response variable was East South-Central Region, data was filtered out to contain the required buildings data.



This CORR-PLOT shows variables which were selected from various variables

After this data was cleaned using correlation plot as benchmark to remove redundant variables which might have an adverse effect on model performance. Correlation matrix was taken after data cleaning and is shown below.



Correlation matrix of variables taken after cleaning data

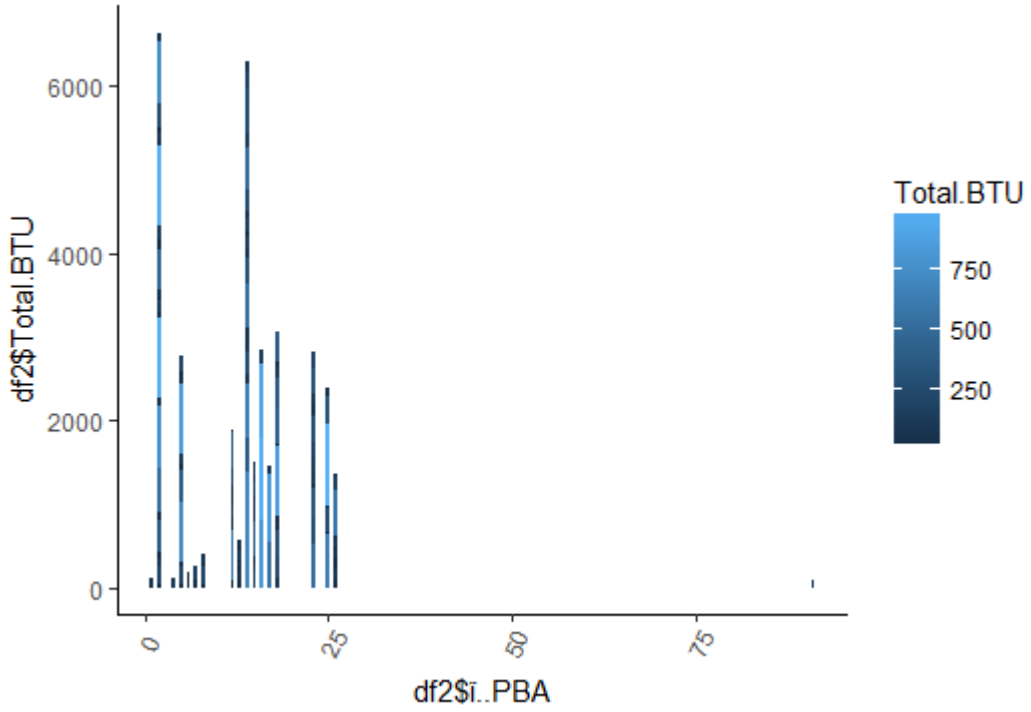
Initial list of variables had 71 variables with most of them being redundant and related to other parameters. Variables taken for building models are given below.

Serial No.	Variable code	Variable explanation
1	ï..PBA	Principal building activity
2	SQFT	Square footage
3	WLCNS	Wall construction material
4	RFCNS	Roof construction material
5	RFCOOL	Cool roof materials
6	GLSSPC	Percent exterior glass
7	NFLOOR	Number of floors
8	FLCEILHT	Floor to ceiling height
9	RENHVC	HVAC equipment upgrade
10	RENELC	Electrical upgrade
11	RENINS	Insulation upgrade
12	ONEACT	One activity in building
13	PBAPLUS	Specific building activity
14	FACIL	On a multiple building complex
15	NOCC	Number of businesses
16	MONUSE	Months in use
17	WKHRS	Total hours open per week
18	NWKER	Number of employees
19	ELHT1	Electricity used for main heating
20	ELHT2	Electricity used for secondary heating

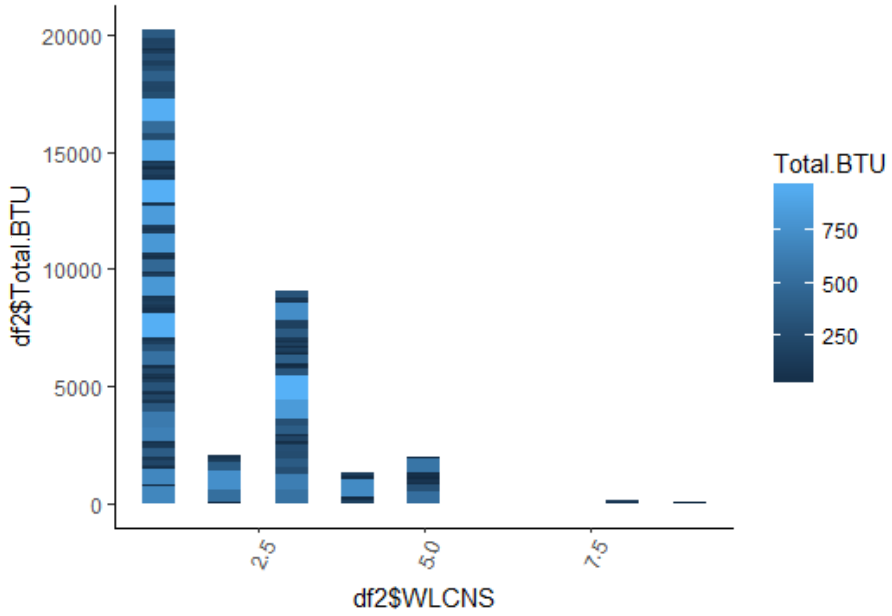
21	HEATP	Percent heated
22	FURNAC	Furnaces that heat air directly
23	MAINHT	Main heating equipment
24	WTHTEQ	Water heating equipment
25	AMIMETER	AMI Smart metering
26	CDD65	Cooling degree days
27	Total.BTU	Total BTU= summation of individual BTUs used in space conditioning
28	PUBCLIM	Building America climate region

Final Database was built containing the above independent variables and total energy as dependent variable. The dependent variable was scaled down by dividing it by 10^4 to get uniformity over complete data. Few variables and there relation with total energy consumption has been explained below.

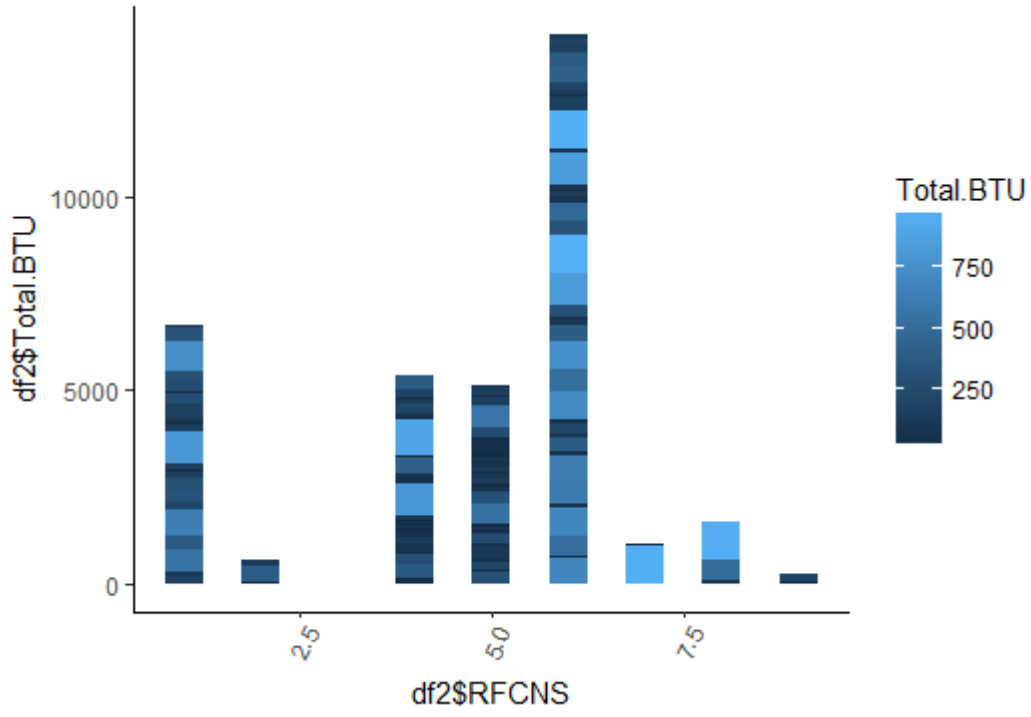
- Principal Building activity:
  This variable had 21 categories and gives the purpose/activity the building is being used for. The below given plot shows energy usage with respect to different categories. It can be observed highest energy is used by office buildings.



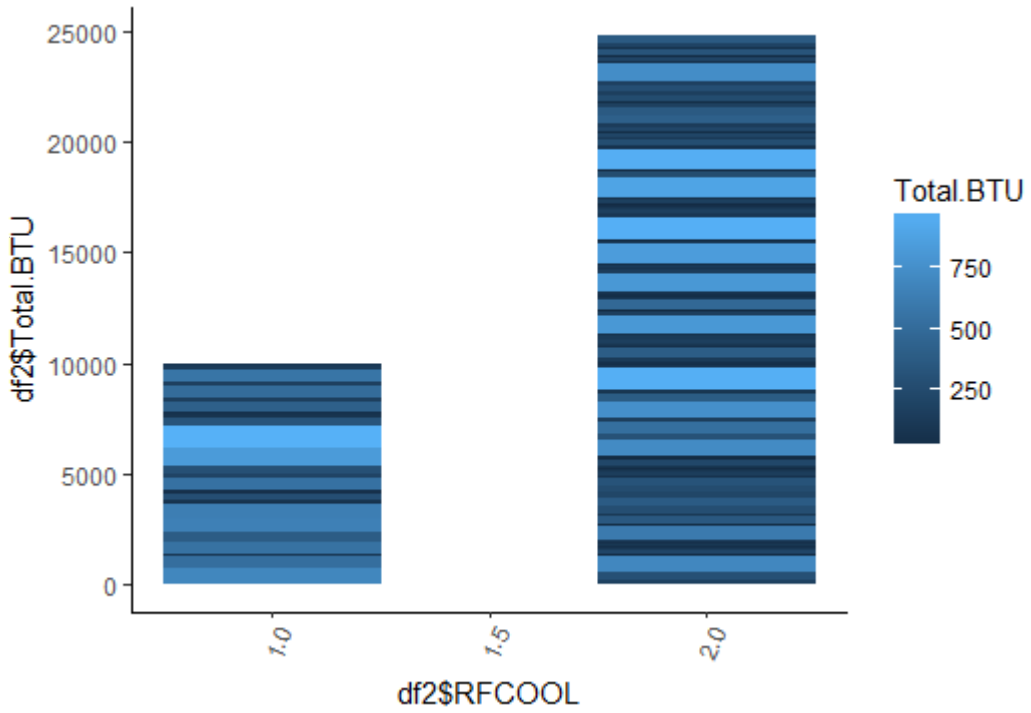
- Wall Construction Materials:
  This Variable took under consideration type of material used in construction of house. And it can be observed that houses constructed by brick and constructed were consuming more energy.



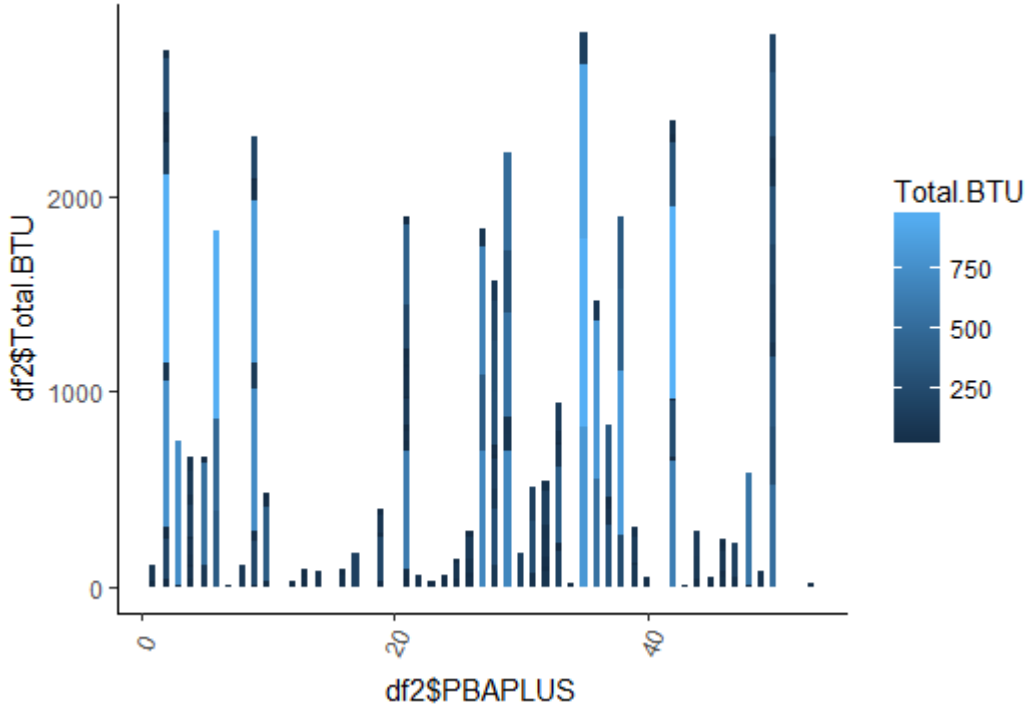
- Roof Construction Material:
  This Variable took under consideration type of material used in construction of roof of house. And it can be observed that roofs constructed by synthetic, plastic or rubber sheeting are consuming most energy.



- Cool Roof Materials:
  This Variable took under consideration if cool materials were used in the construction of roof of house. And it can be observed that roofs without cool roof materials consume more energy.

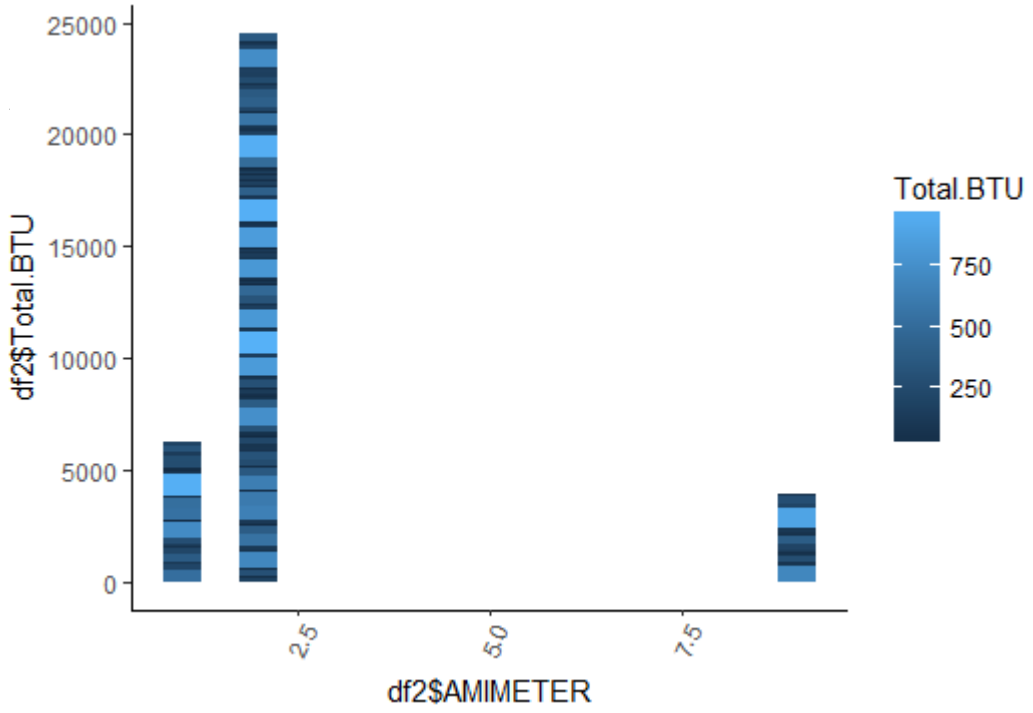


- More Specific Building Activity:
  This variable takes in count all types of building activity possibly the buildings are being used for. It is an extension of PBA with more in depth understanding and specifically pointing at activities which are consuming energy. It can be seen administrative/professional offices, Distribution/shipping center, High schools, Hospital/inpatient health, Retail stores and Strip shopping mall are most consumers of energy.

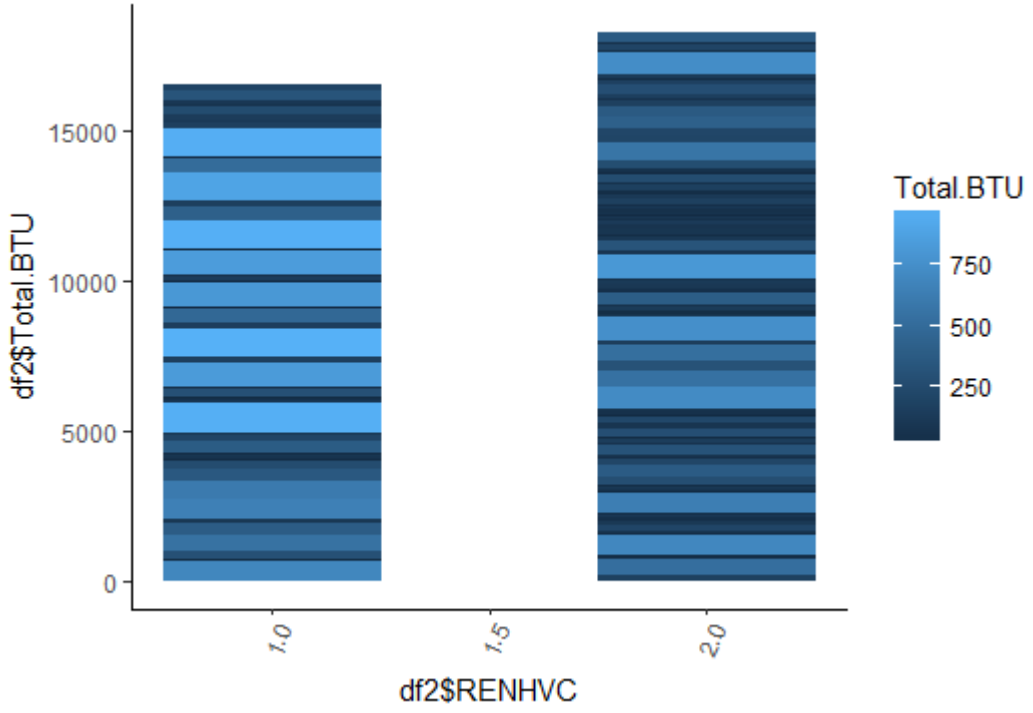


- AMI Smart Metering:

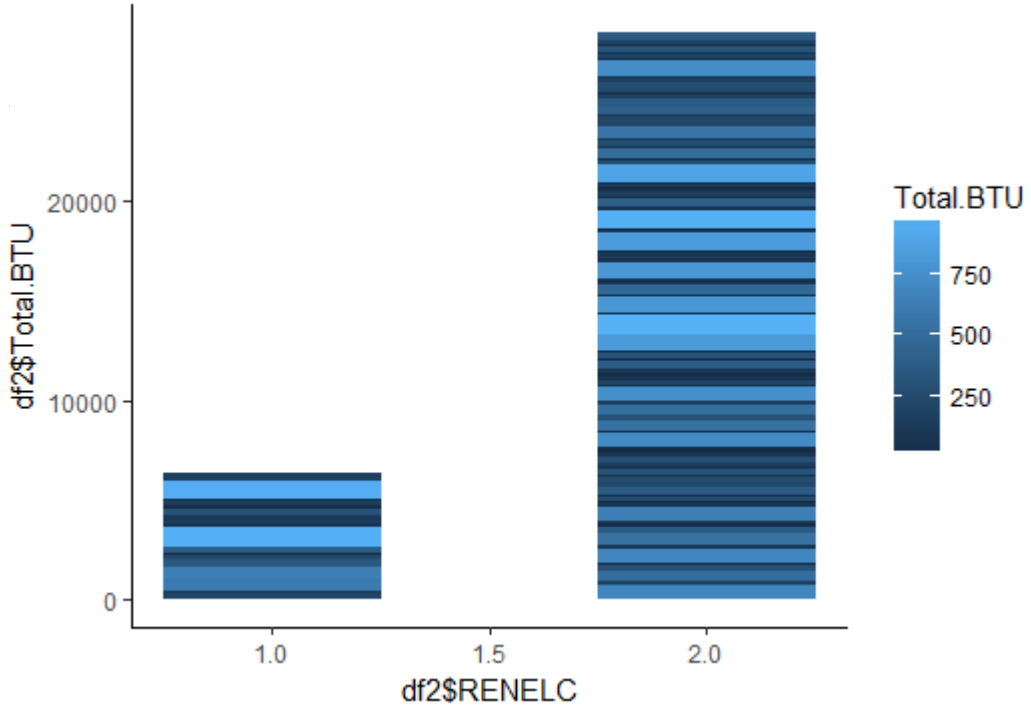
This variable takes into consideration houses under observation have AMI smart metering installed, which regulates total energy consumption. It can be observed that the houses without AMI smart metering were consuming large amounts of energy.



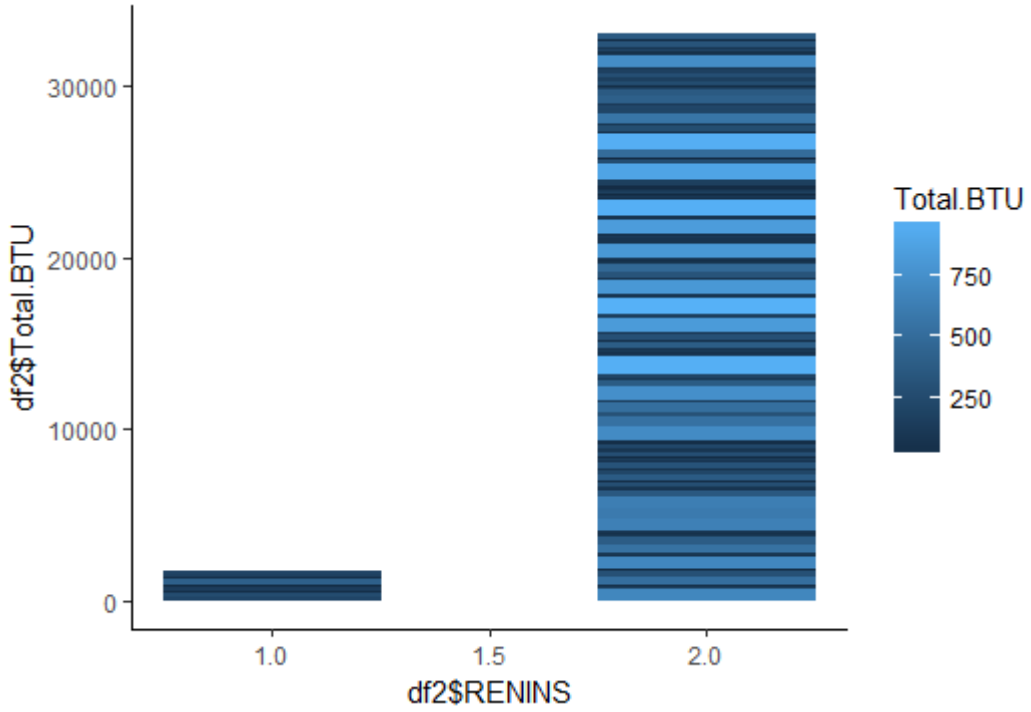
- HVAC Equipment Upgrade:
  This variable takes under consideration the upgrade of Heating, Ventilation and Cooling(HVAC) equipment. It can be observed that the HVAC upgraded buildings had low energy consumption.



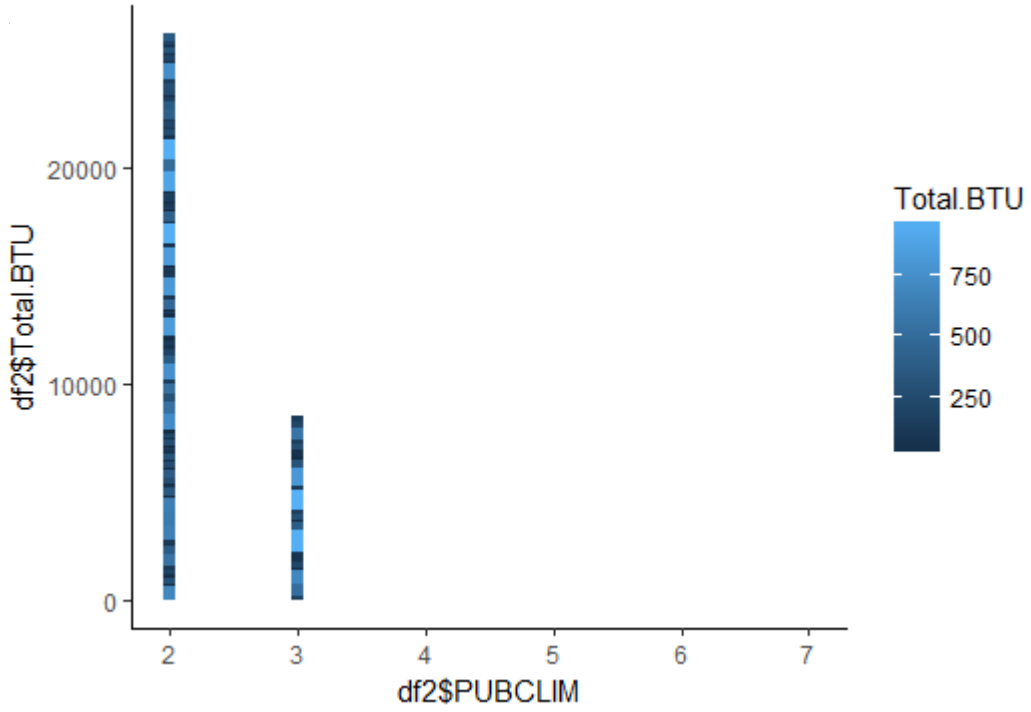
- Electrical Upgrade:
  This variable takes under consideration the upgrade of electrical wiring etc. It can be observed that the upgraded buildings had low energy consumption.



- Insulation Upgrade:
  This variable takes under consideration the upgrade of electrical equipment. It can be observed that the upgraded buildings had low energy consumption.



- Building America Climate Region:
  This variable took into consideration different climates buildings are situated in. But data might be skewed as the response variable is related to east south-central region, which might not have much variation in climates. It can be observed buildings situated in mixed-humid climate.



MODELS COMPARISON AND INTERPRETATION:

Data obtained after exploratory analysis was used to build different models. Models used for fitting data were linear, Generalized Linear Models(GLM), Generalized Additive Models(GAM), Random Forest(RF), Support Vector Machines(SVM), Bayesian Additive Regression Trees(BART), Multivariate Adaptive Regression Splines(MARS) and Classification and Regression Trees(CART). Neural Networks weren't used as the dataset was small and wouldn't train neural networks in the best way possible and interpretation of neural networks is difficult.

The below given table shows Root Mean Square values(In-Sample), RMSE(Out-of-sample) of various fitted models.

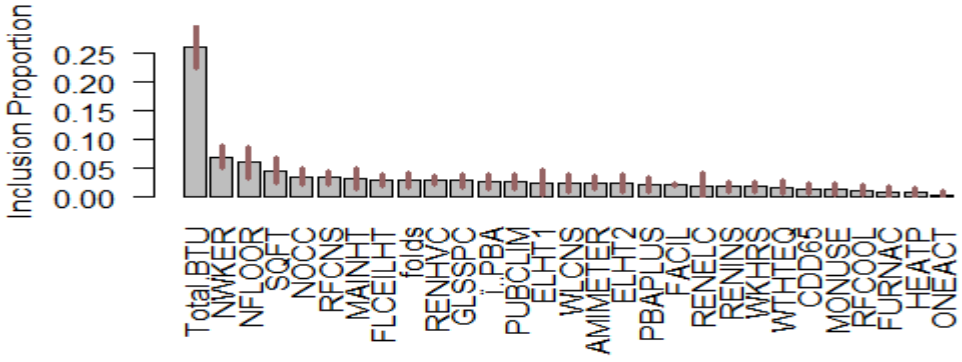
linearmodelis	linearmodelos	lassomodelis	lassomodelos	rdigemodelis	ridgemodelos	GAMmodelis	GAMmodelos	cartmodelis	cartmodelos
130.7185	128.0702	219.97	182.1969	219.97	182.1969	130.9187	125.3178	104.71599	92.07443
126.8846	179.5315	208.9107	295.6474	208.9107	295.6474	126.9188	177.6438	104.00377	114.81126
131.8524	106.8453	222.4792	132.0884	222.4792	132.0884	131.986	104.628	102.55391	107.12448
130.8231	116.4835	219.2951	167.1493	219.2951	167.1493	130.8307	116.6512	105.89628	119.86236
126.675	156.2274	218.7285	181.6019	218.7285	181.6019	126.6829	156.322	96.20329	119.48776
131.9253	113.731	222.282	138.3316	222.282	138.3316	131.9461	113.835	103.31494	117.17999
123.4099	173.0602	208.1551	266.3858	208.1551	266.3858	123.4754	172.6998	95.54381	171.73814
112.7401	322.7786	209.9882	246.4198	209.9882	246.4198	112.7414	322.8647	95.4696	135.67096
122.0886	358.6788	201.4387	349.6147	201.4387	349.6147	122.1448	358.4916	99.16379	147.74935
127.3745	153.0141	216.567	194.216	216.567	194.216	127.3765	153.161	102.37045	157.09008
126.4492	180.84206	214.78145	215.36518	214.78145	215.36518	126.50213	180.16149	100.923583	128.278881

rfmodelis	rfmodelos	marsprunedmodelis	marsprunedmodelos	marsunprunedmodelis	marsunprunedmodelos	SVMmodelis	SVMmodelos	bartmodelis	bartmodelos
58.78344	102.29057	103.59873	108.8554	90.06996	73.82237	21.72189	184.0027	3.116243	32.500715
57.51775	148.00123	93.287	101.6333	88.34096	80.16014	20.677	294.6169	2.76369	8.540608
59.02802	83.22659	87.10021	124.9303	87.76352	88.99219	22.01644	135.1642	2.164814	49.166914
58.36834	109.73378	92.82516	112.56	87.38763	92.51217	21.7216	167.9544	2.76004	33.594095
55.58738	134.30677	90.27981	139.2419	86.13458	100.85036	21.62639	184.8142	3.239064	24.759967
58.68321	89.38578	95.23299	117.0254	88.60139	81.74468	21.9808	139.9407	2.445557	45.009552
53.36577	160.72572	89.65521	153.8868	85.34471	107.46389	20.58738	264.4285	3.002363	40.21297
57.24139	115.81547	95.58924	184.3026	89.21633	78.12604	20.84082	244.4473	2.245774	53.593635
54.03108	190.67727	89.23261	168.807	88.63846	77.10114	19.93513	346.7453	2.304571	89.325982
55.84663	132.23094	100.85746	126.56	87.40601	93.25974	21.40289	195.9205	2.256668	21.184411
56.845301	126.639412	93.765842	133.78027	87.890355	87.403272	21.251034	215.80347	2.6298784	39.7888849

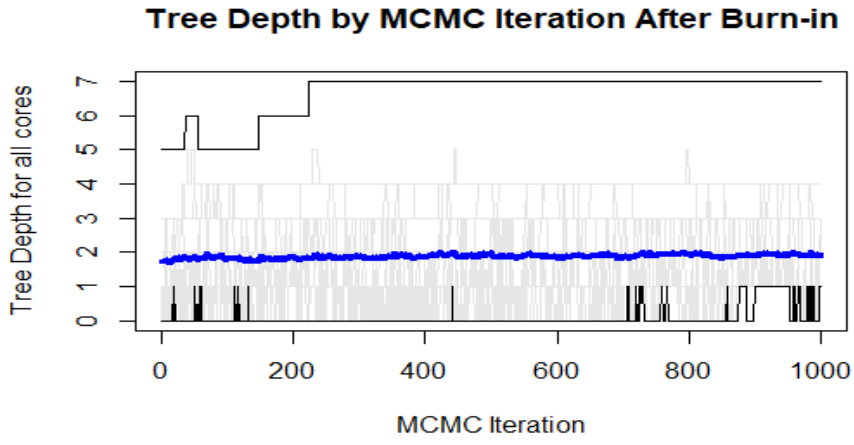
It can be observed best RMSE values both in-sample and Out-of-Sample was for BART.

Adjusted R^2 value for BART was 0.96 which is very good.

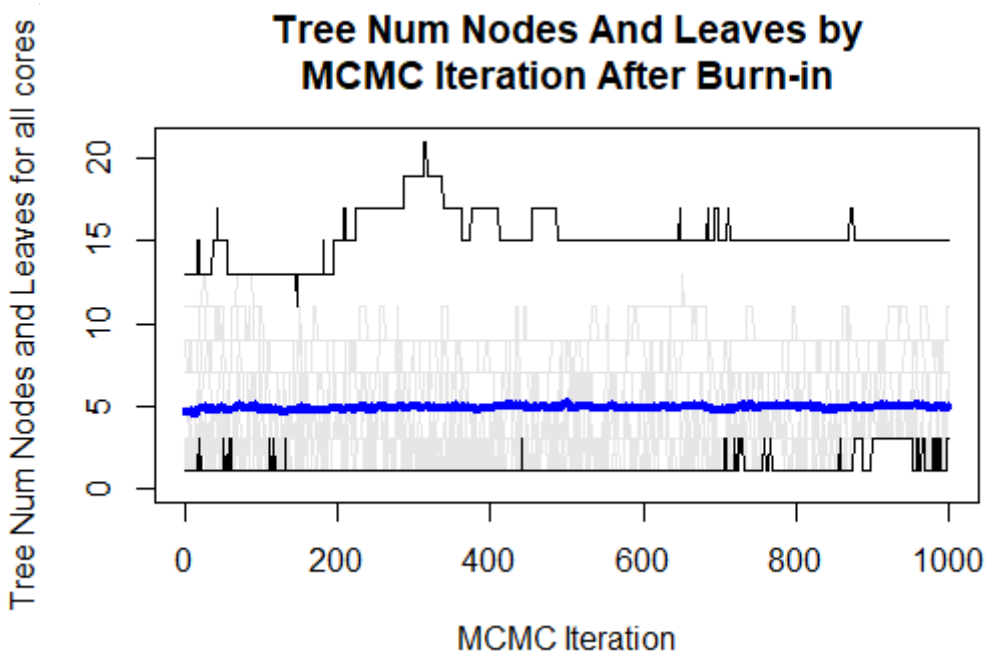
This model was selected based on both adjusted R^2 and RMSE values which were significantly lower compared to other models. R^2 values for other were found but weren't significant, as there RMSE were very high compared to BART.



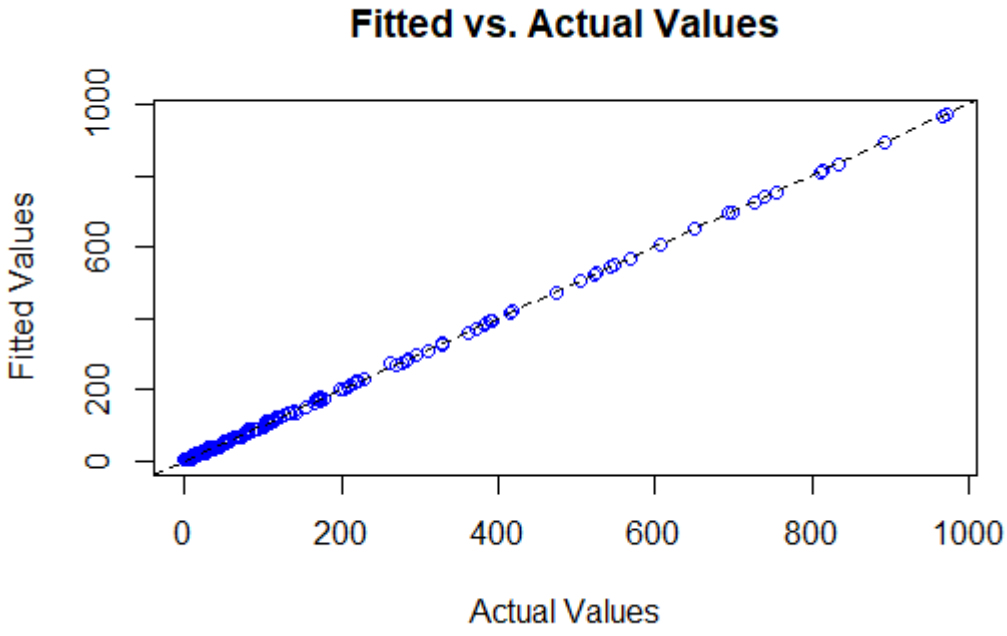
The above graph gives variable importance of different variables included in BART model.



The above given plot gives average tree depth across each tree in the BART model by Gibbs sample number (Gibbs sampling assumes we can compute conditional distributions of one variable conditioned on all of the other variables and sample exactly from these distributions).

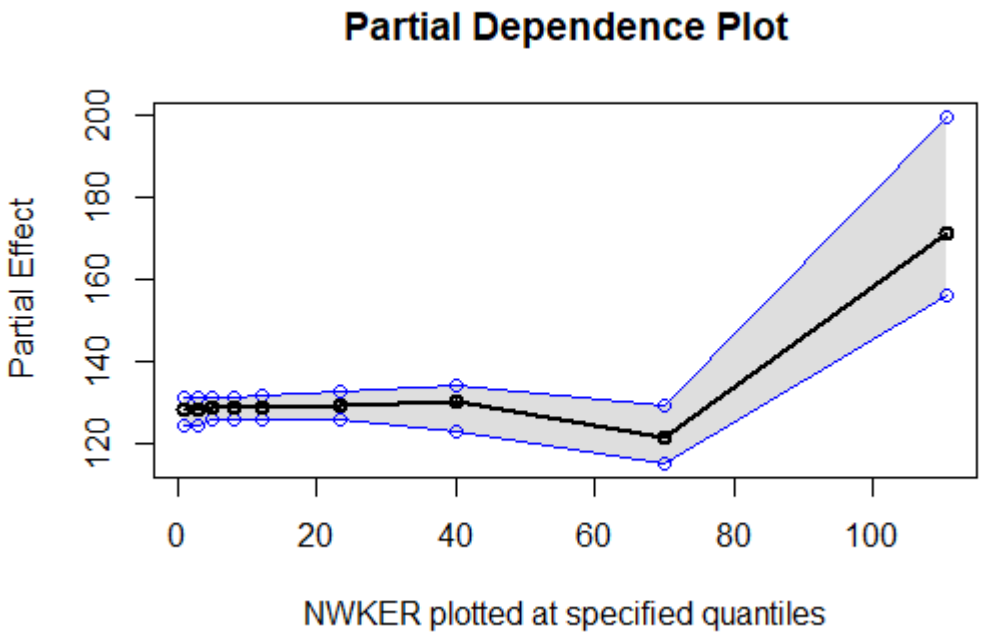


The above given plot gives average number of nodes across each tree in the BART model by Gibbs sample number.

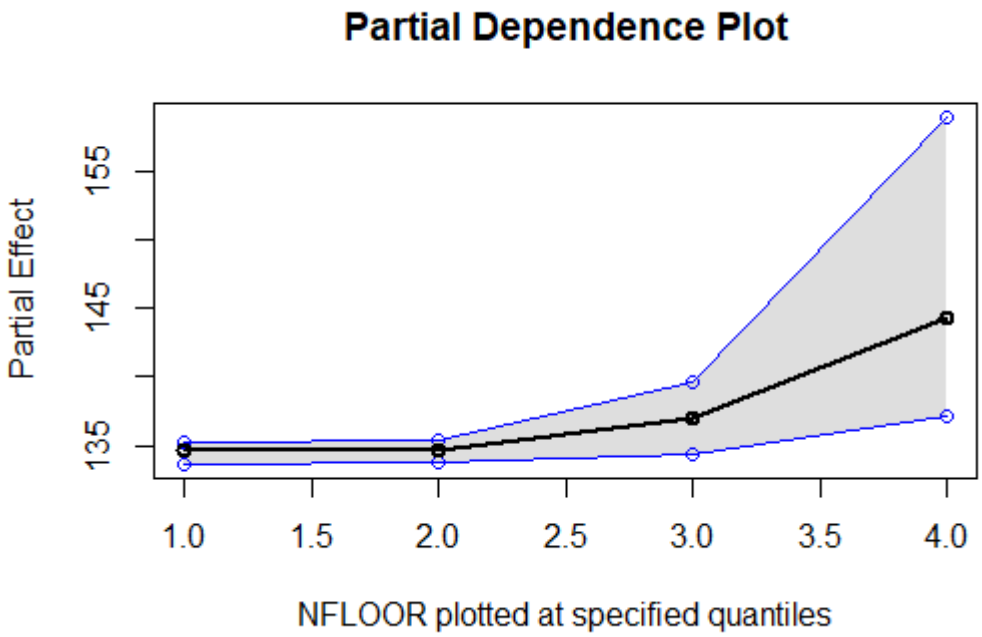


The above graph gives actual vs fitted values in BART model.

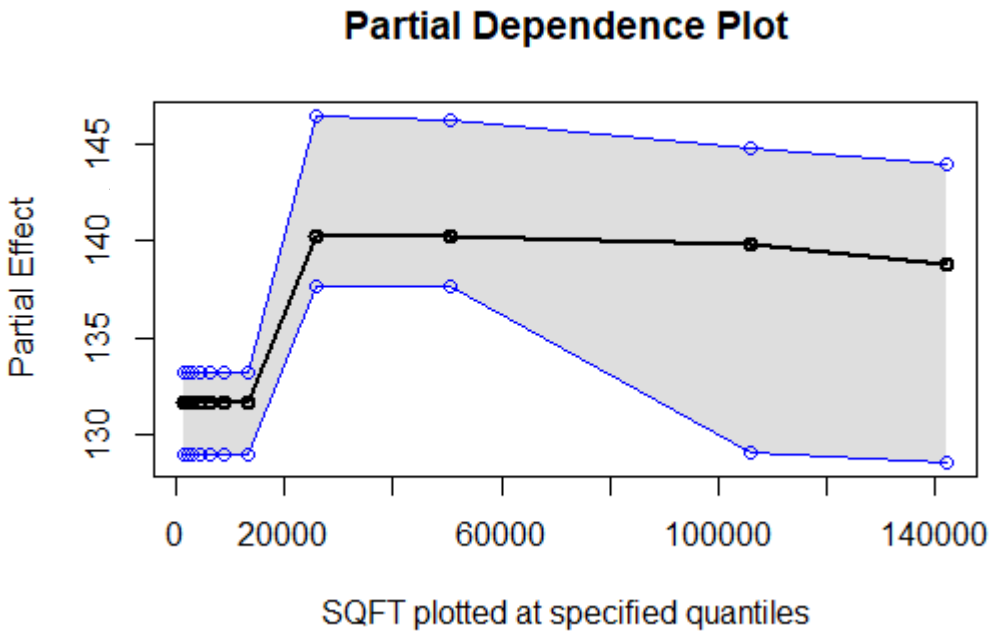
PARTIAL DEPENDENCY PLOTS:



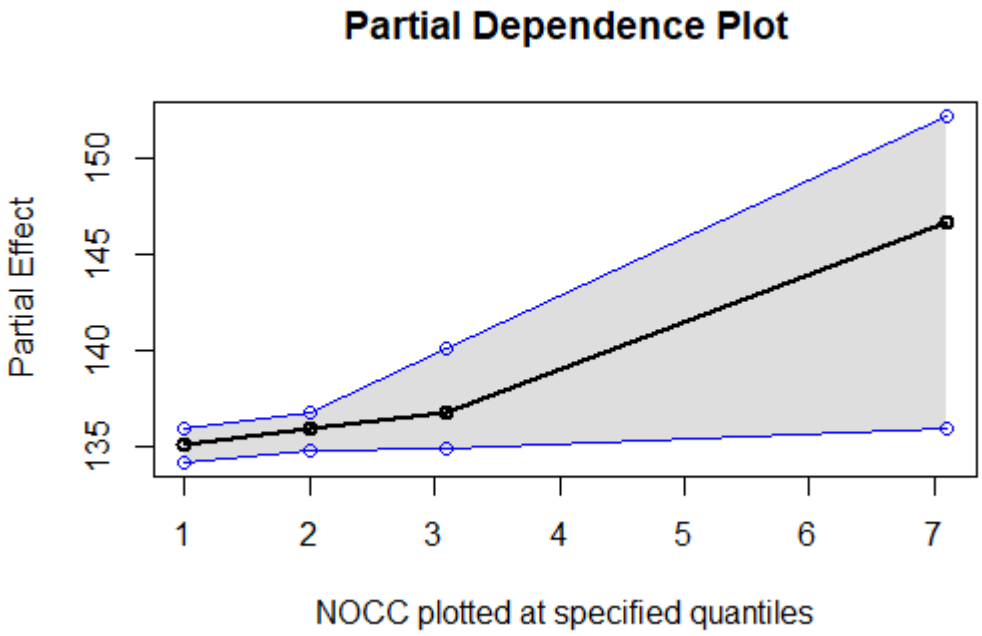
This above plot gives partial dependency of number of employees on total energy consumption and clearly explains as number of employees increases consumption increases.



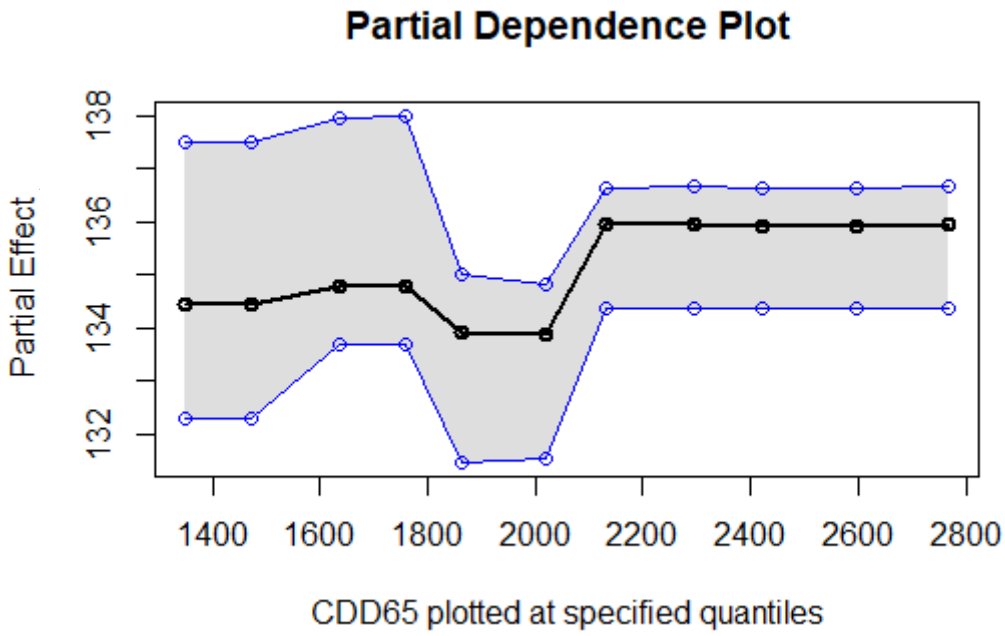
This above plot gives partial dependency of number of floor of buildings on total energy consumption and explains as number of floors increases consumption increases.



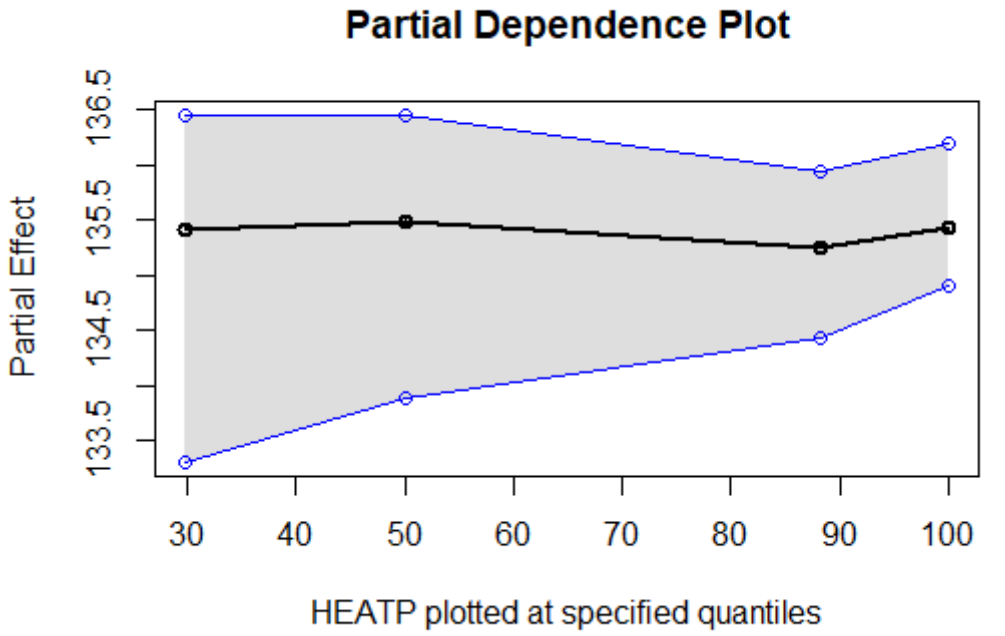
This above plot gives partial dependency of square feet area of buildings on total energy consumption and it is observed to a certain extent the consumption remains same and increases rapidly and wouldn't change much after certain area.



This above plot gives partial dependency of number of businesses in a building on total energy consumption and it is observed energy consumption increases as number of businesses increases.



The above plot shows partial dependency of cooling degree days on total energy consumption. The graph is U type which explains, to maintain the building temperature energy Is highly consumed during cold/hot weather and decreases during optimal temperature.



The above graph gives the partial dependency of percent heated to total energy consumption. The plot shows dependency isn't varying much that might be due to both heating and cooling energy consumption which would balance out total energy consumption.

The other partial dependency plots weren't shown in this report as they are categorical values and wouldn't be wise to show partial dependency on the response variable.