Modeling Energy Consumption to Predict Future Demand

Final Project Report

Submitted by: Chrish David Douglas,
Kulveen Kaur,
Tejaswini Vibhute,
Yashaswi Pandey,
Yerramorrusu Harshitha Reddy

Submitted to: Professor Erik Anderson

Table of Contents

Abstract

The rising temperatures due to global warming has resulted in the steady increase in electricity consumption around the globe, this caused concern to eSC, an energy company based in South Carolina as they want to reduce electricity consumption if next summer if ‘extra-hot’, so as to avoid blackouts. We analyzed the data from the month of July, as eSC thinks that it is the month with the highest energy usage. After thoroughly examining the data, we cleaned it and transformed it to build a model using the features having the most effect on the total energy consumption, the result from the model predicted the energy consumption with a high accuracy and we recommended them to advocate for better insulation to reduce the total energy consumption.

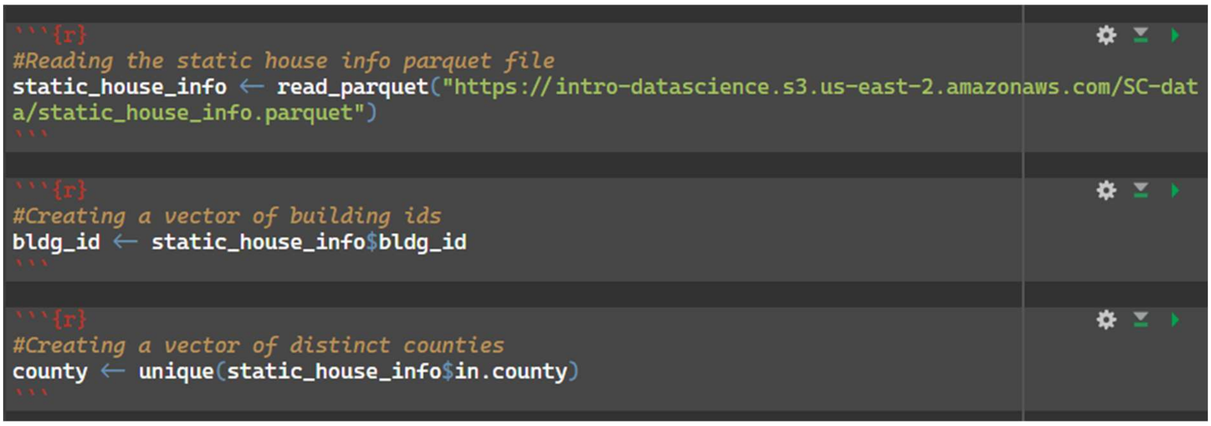
Introduction

The aim of the project was to provide a data driven approach to solving the issue of rising electricity consumption due to global warming. The data provided contained information about the electricity consumption of houses in various counties, the details of the houses and the weather information for the counties, the analysis of the data gave a positive correlation between the rise in temperature and higher electricity consumption. This caused concern to eSC, an energy company, prompting them to find ways to decrease electricity consumption if next summer is ‘extra-hot.’ The project uses the data to analyze the electricity consumption in various counties to build a model that accurately predicts the electricity consumption for next year and provides the company with recommendations for steps to take to reduce demand on their electricity grid.

Data Pipelining

The project started with creating a data pipeline to get all the required data from all the sources and create a csv file with only the data needed for the project.

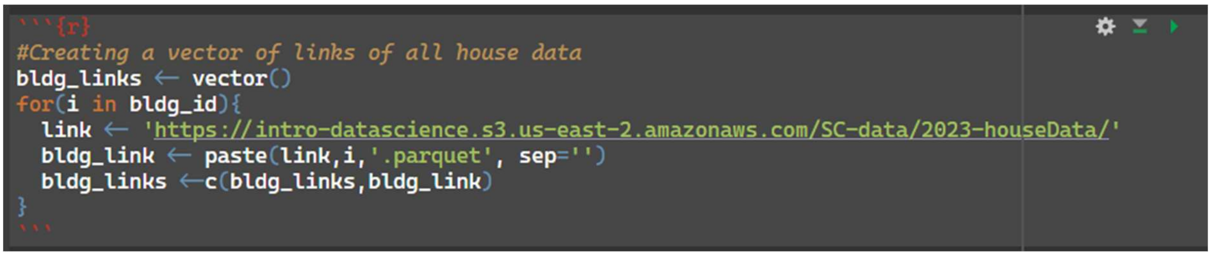
The first step was to read in the Static House Info file and extract the unique building and county ids present. Then we created a vector containing all the links for the energy data of each house, iterating over the vector we read each house's energy data into a temporary dataframe, then extracted the data for the month of July, added the total for all energy usage and added the building id before   binding the temporary dataframe to our main dataframe.



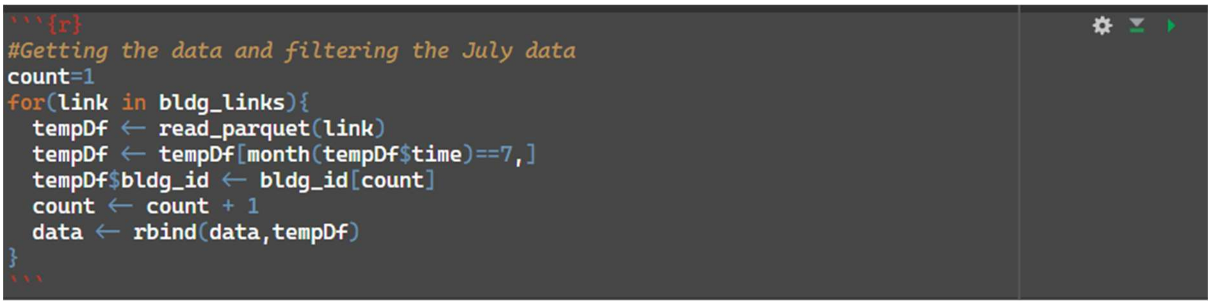
~~~
```{r}
#Reading the static house info parquet file
static_house_info ← read_parquet("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-dat
a/static_house_info.parquet")
```
~~~

~~~
```{r}
#Creating a vector of building ids
bldg_id ← static_house_info$bldg_id
```
~~~

~~~
```{r}
#Creating a vector of distinct counties
county ← unique(static_house_info$in.county)
```
~~~

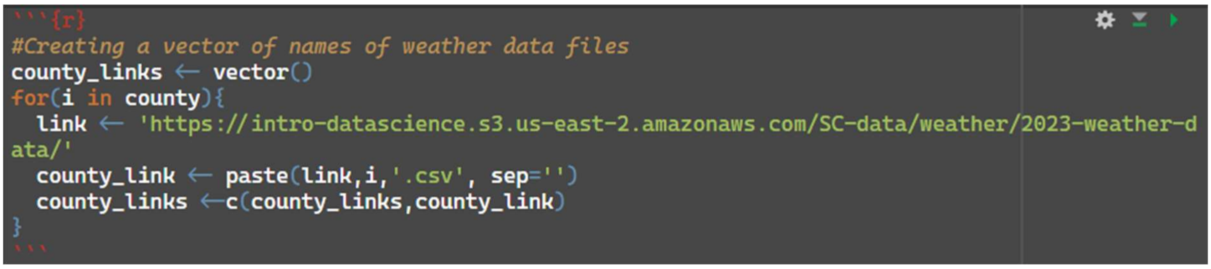


~~~
```{r}
#Creating a vector of links of all house data
bldg_links ← vector()
for(i in bldg_id){
  link ← 'https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/'
  bldg_link ← paste(link,i,'.parquet', sep='')
  bldg_links ←c(bldg_links,bldg_link)
}
```
~~~

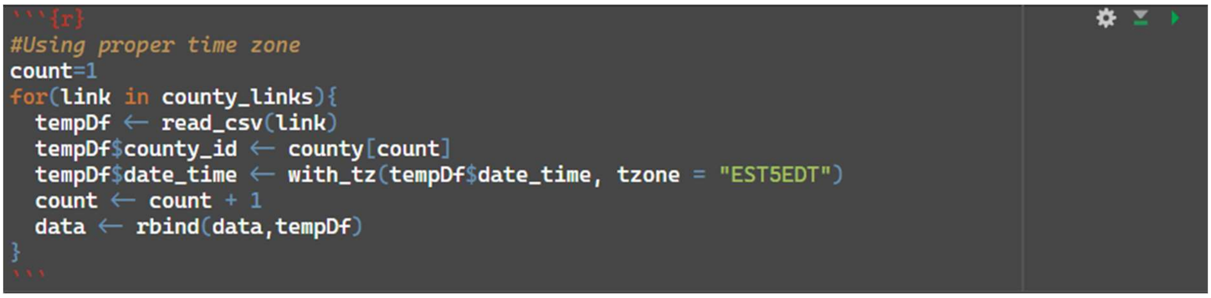


~~~
```{r}
#Getting the data and filtering the July data
count=1
for(link in bldg_links){
  tempDf ← read_parquet(link)
  tempDf ← tempDf[month(tempDf$time)==7,]
  tempDf$bldg_id ← bldg_id[count]
  count ← count + 1
  data ← rbind(data,tempDf)
}
```
~~~

The second step was creating a vector with all the links to weather data for each county, iterating over this vector we read all the weather data one by one into a temporary dataframe, then converted the time zone from UTC to EDT before extracting the data for the month of July and adding the county id, then we bind this with our main dataframe for weather data.



~~~
```{r}
#Creating a vector of names of weather data files
county_links <- vector()
for(i in county){
  link <- 'https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/'
  county_link <- paste(link,i,'.csv', sep='')
  county_links <-c(county_links,county_link)
}
```
~~~



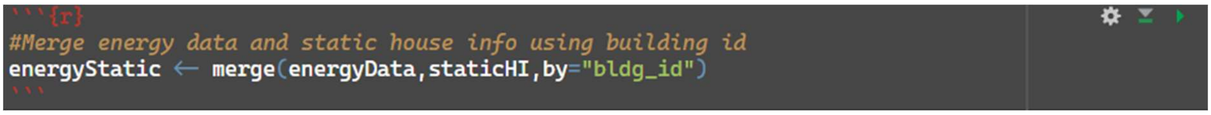
~~~
```{r}
#Using proper time zone
count=1
for(link in county_links){
  tempDf <- read_csv(link)
  tempDf$county_id <- county[count]
  tempDf$date_time <- with_tz(tempDf$date_time, tzone = "EST5EDT")
  count <- count + 1
  data <- rbind(data,tempDf)
}
```
~~~

The main issue we faced in this step of the project was formulating a plan to get the data we need as quickly as possible, then we later had to comeback and correct it as we had not checked the time zones earlier and were having trouble merging our data.

At the end of this phase of the project we had 4 csv files containing all the data we need to complete the project

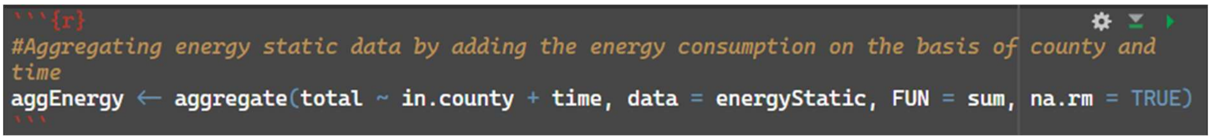
Data Preparation

In this phase of the project, we started with merging the house data and the static house information to get the details of the house along with its energy consumption so we could extract the features that predict the energy consumption the most.



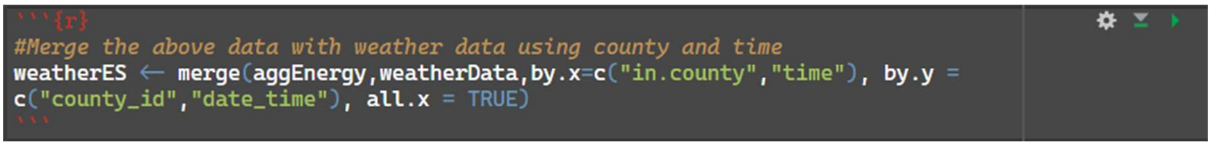
~~~
```{r}
#Merge energy data and static house info using building id
energyStatic <- merge(energyData,staticHI,by="bldg_id")
```
~~~

The merged data was then aggregated according to the county and the time, the total consumption was summed to get the total consumption for a county for every hour.



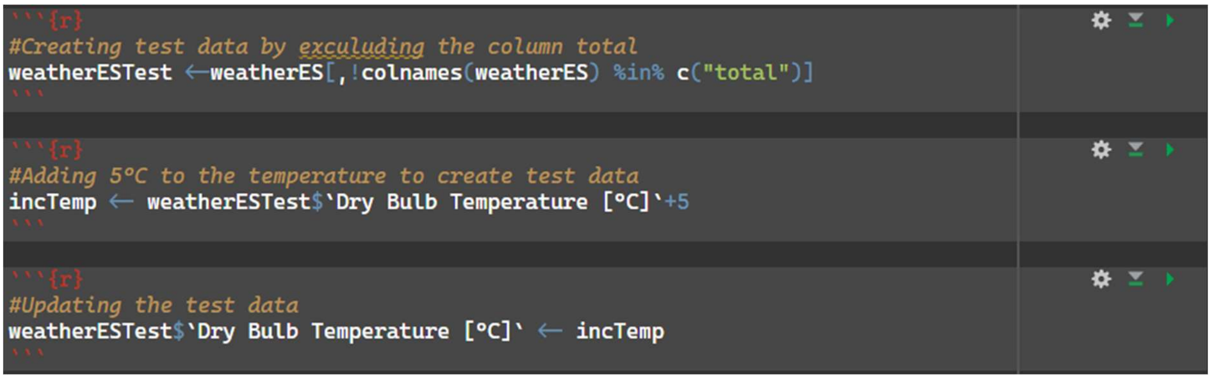
~~~
```{r}
#Aggregating energy static data by adding the energy consumption on the basis of county and
time
aggEnergy <- aggregate(total ~ in.county + time, data = energyStatic, FUN = sum, na.rm = TRUE)
```
~~~

We merged this aggregated data with the weather data for every county for every hour to get the weather information, county, and the total consumption, as we have to predict the change in total consumption if temperatures rise.



~~~
```{r}
#Merge the above data with weather data using county and time
weatherES <- merge(aggEnergy,weatherData,by.x=c("in.county","time"), by.y =
c("county_id","date_time"), all.x = TRUE)
```
~~~

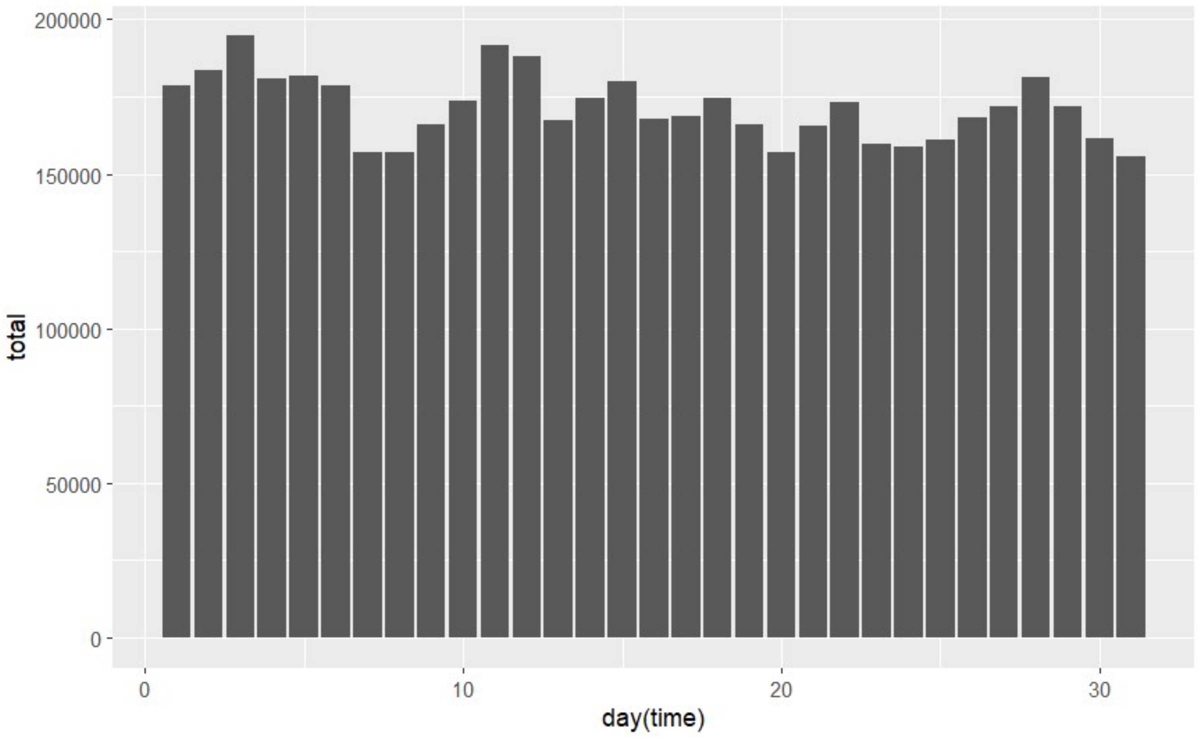
We dropped the total consumption and added 5 to the temperature to get the dataframe that we will use to predict the energy usage for every hour needed for next year.



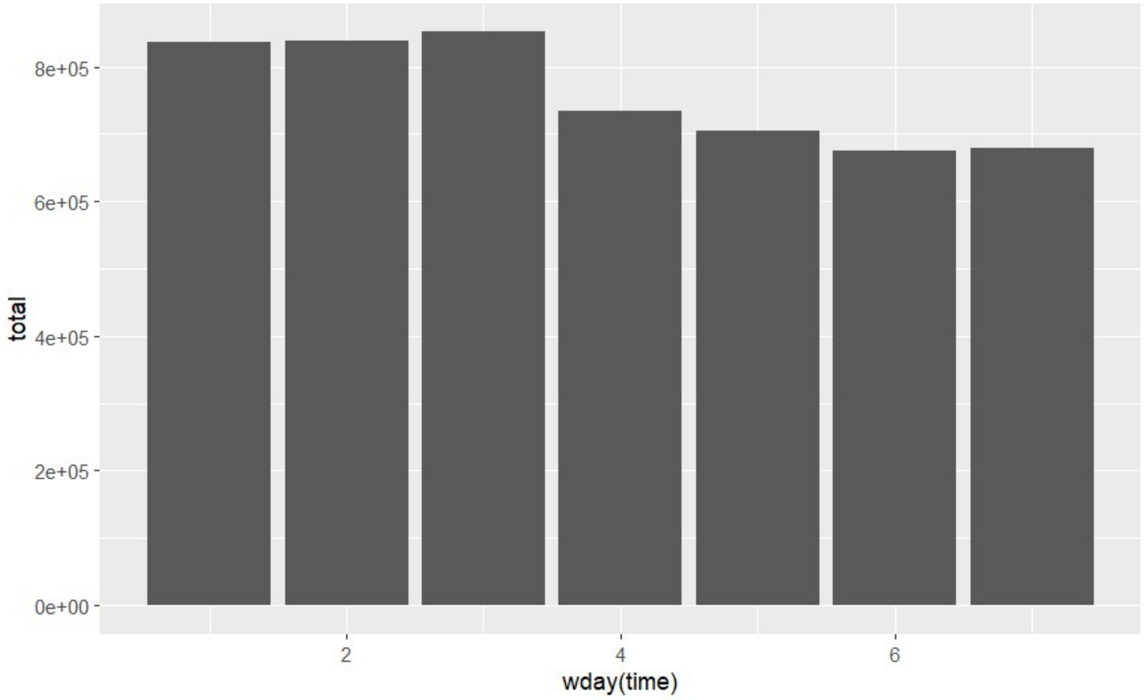
~~~
```{r}
#Creating test data by exculuding the column total
weatherESTest <-weatherES[,!colnames(weatherES) %in% c("total")]
```

```{r}
#Adding 5°C to the temperature to create test data
incTemp <- weatherESTest$`Dry Bulb Temperature [°C]`+5
```

```{r}
#Updating the test data
weatherESTest$`Dry Bulb Temperature [°C]` <- incTemp
```
~~~

Data Analysis

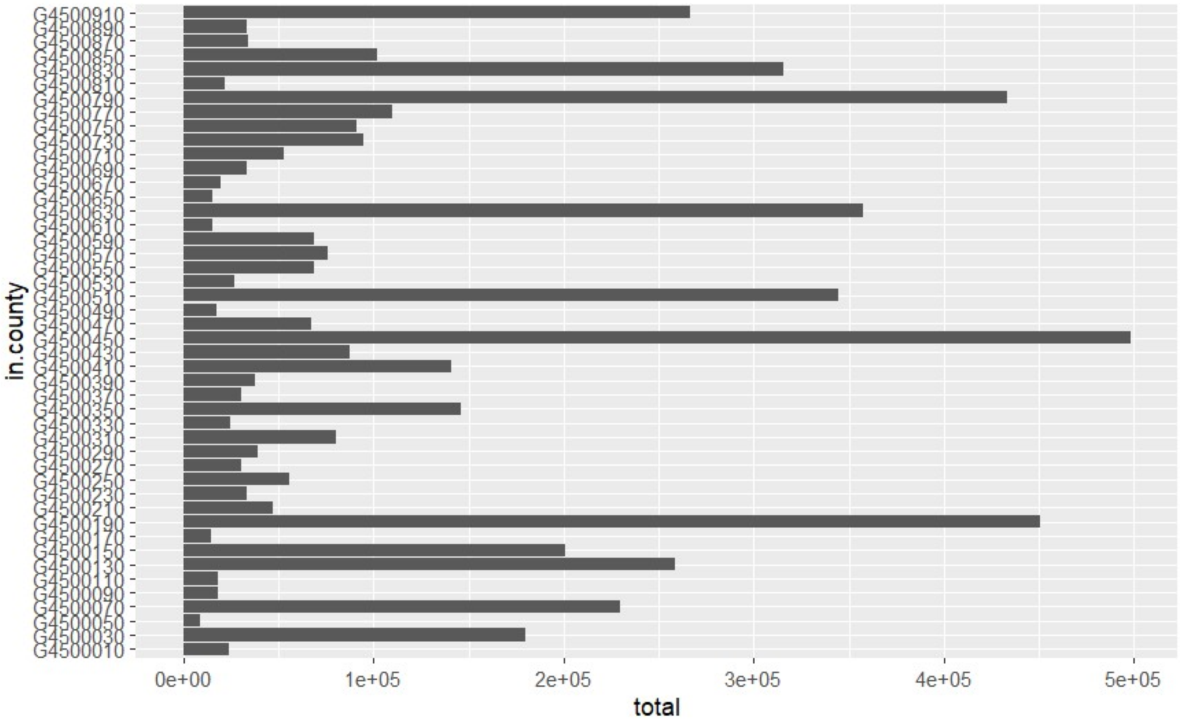
The data we prepared was then analyzed to get an idea of how the variables are related to improve our knowledge of the data so we can prepare a model that will accurately predict the energy consumption. The first thing that was analyzed was the relationship between the days of the week and the total energy demand on eSC's electrical grid to get an idea of the days in July when electricity demand was highest and lowest. We plotted this data and found out that the energy demand was highest at the start of the month and slowly decreased as we got further into July.



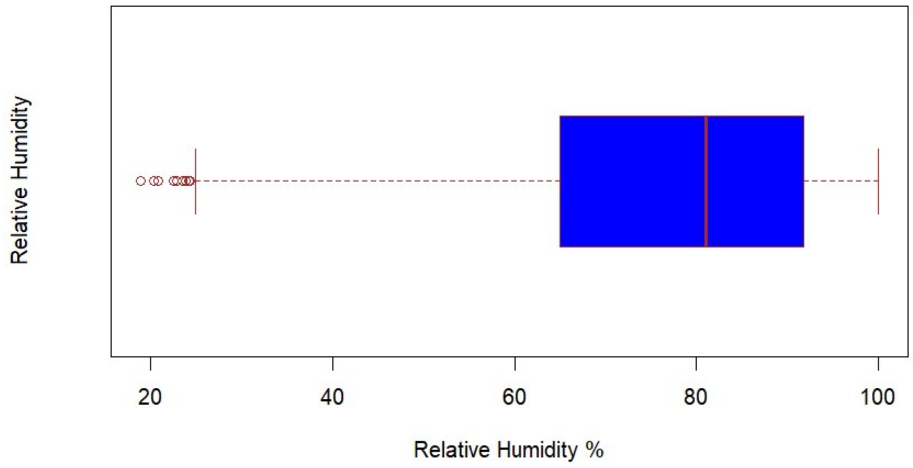
The other thing that stood out was that there were clear patterns of highs and lows and on inspection showed us that the demand was highest on Monday through Wednesday before slowly decreasing as the week ended.

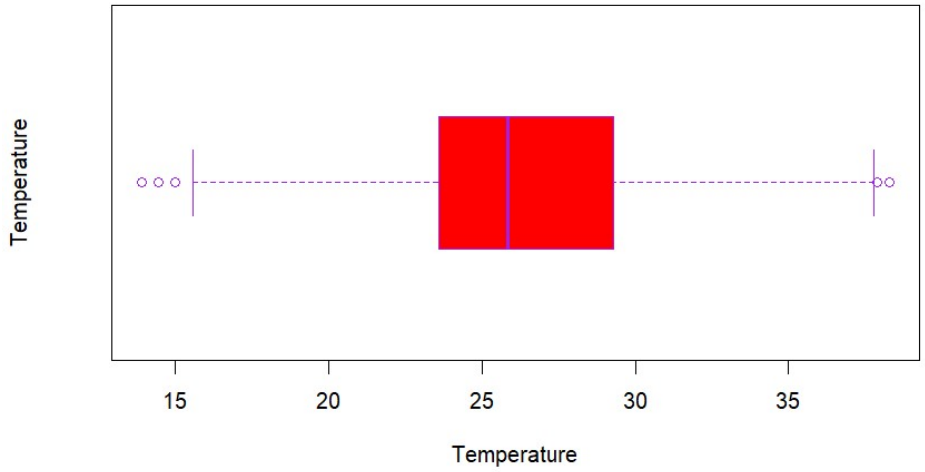


We then decided to have a look at energy consumption in different counties throughout the month of July, the graph showed us clearly that most of the energy consumption was concentrated in only a few counties.

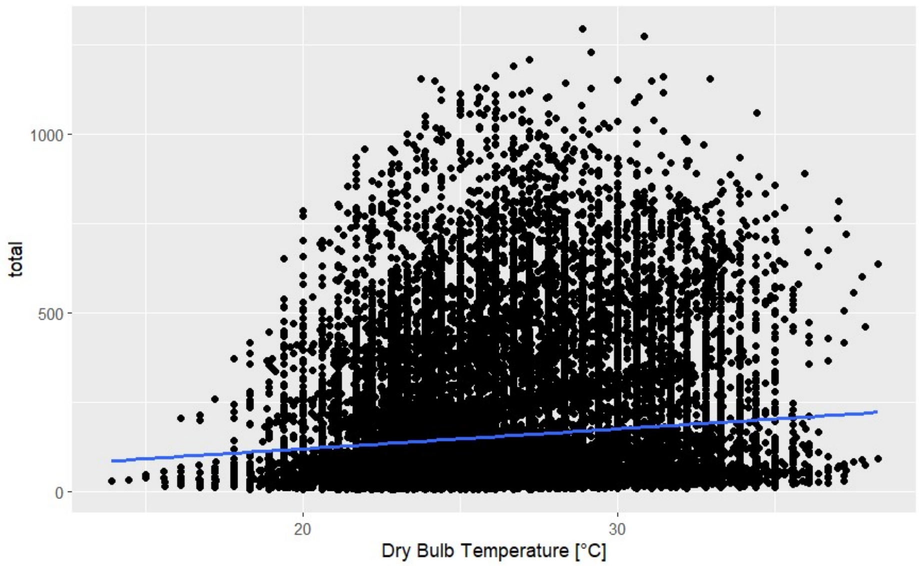


The analysis of the temperature and relative humidity for the month of July showed that while most of the data points were relatively close together there were some significant outliers.



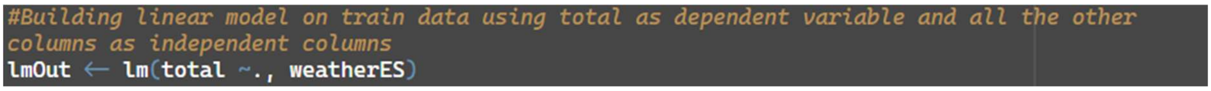


As eSC is concerned with energy consumption when temperature rises, we decided to plot the relationship between the two to see if anything stood out, this confirmed a linear relationship between temperature and energy consumption.


total
1000
500
0
20
30
Dry Bulb Temperature [°C]

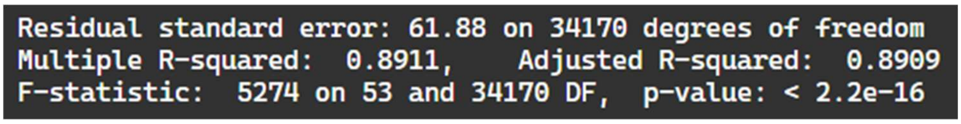
Data Modeling

We decided to start by running a linear model as we had just confirmed a linear relationship between temperature and energy consumption.



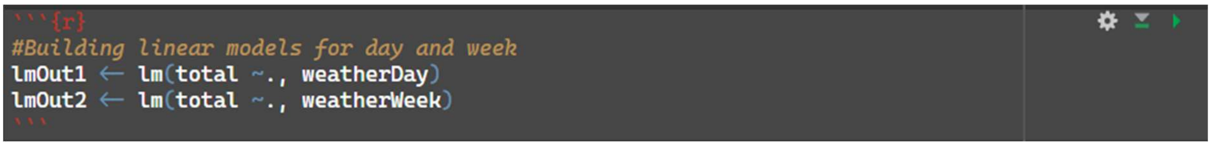
```
#Building linear model on train data using total as dependent variable and all the other
columns as independent columns
lmOut <- lm(total ~., weatherES)
```

The summary showed us that the model had an adjusted R^2 of 0.8909 and p-value of <2.2e-16, this meant that the model was explained 89% of change in the total energy consumption using the features we had chosen.



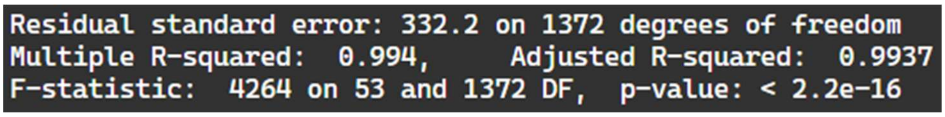
```
Residual standard error: 61.88 on 34170 degrees of freedom
Multiple R-squared:  0.8911,	Adjusted R-squared:  0.8909
F-statistic:  5274 on 53 and 34170 DF,  p-value: < 2.2e-16
```

After this we decided to model the relationships between the day of the month and day of the week and the total energy consumption as our analysis had shown that there were certain patterns in the relationship.



~~~
```{r}
#Building linear models for day and week
lmOut1 <- lm(total ~., weatherDay)
lmOut2 <- lm(total ~., weatherWeek)
```
~~~

The linear model between day and total had an adjusted R^2 of 0.9937 and p-value of <2.2e-16, but the residual standard error of 332.2 was extremely high.



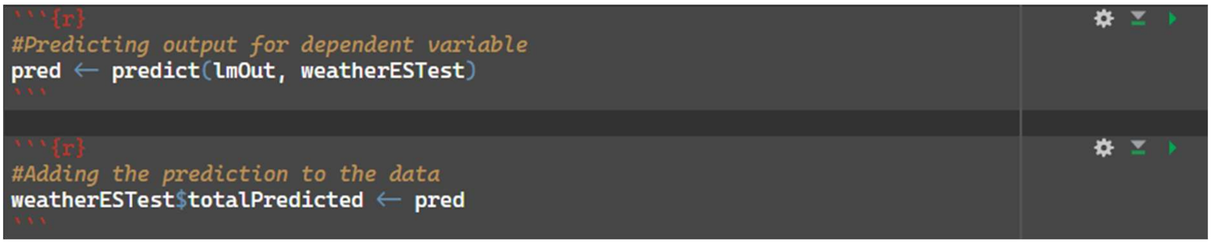
```
Residual standard error: 332.2 on 1372 degrees of freedom
Multiple R-squared:  0.994,	Adjusted R-squared:  0.9937
F-statistic:  4264 on 53 and 1372 DF,  p-value: < 2.2e-16
```

Similarly, the model between day of the week and total returned an adjusted R^2 of 0.9854 and p-value of <2.2e-16, but had an even higher residual standard error at 2260.

We also built a couple of support vector machines and a neural network but had below par performance. In the end we chose our first model as it was the best predictor out of all our models.

Prediction

Having already created the data needed during the preparation phase, we ran a prediction on it using our best model and added the resulting predictions to a new dataframe to get the predicted energy consumption for every hour of July for all the counties if there was an increase in temperature by 5°.



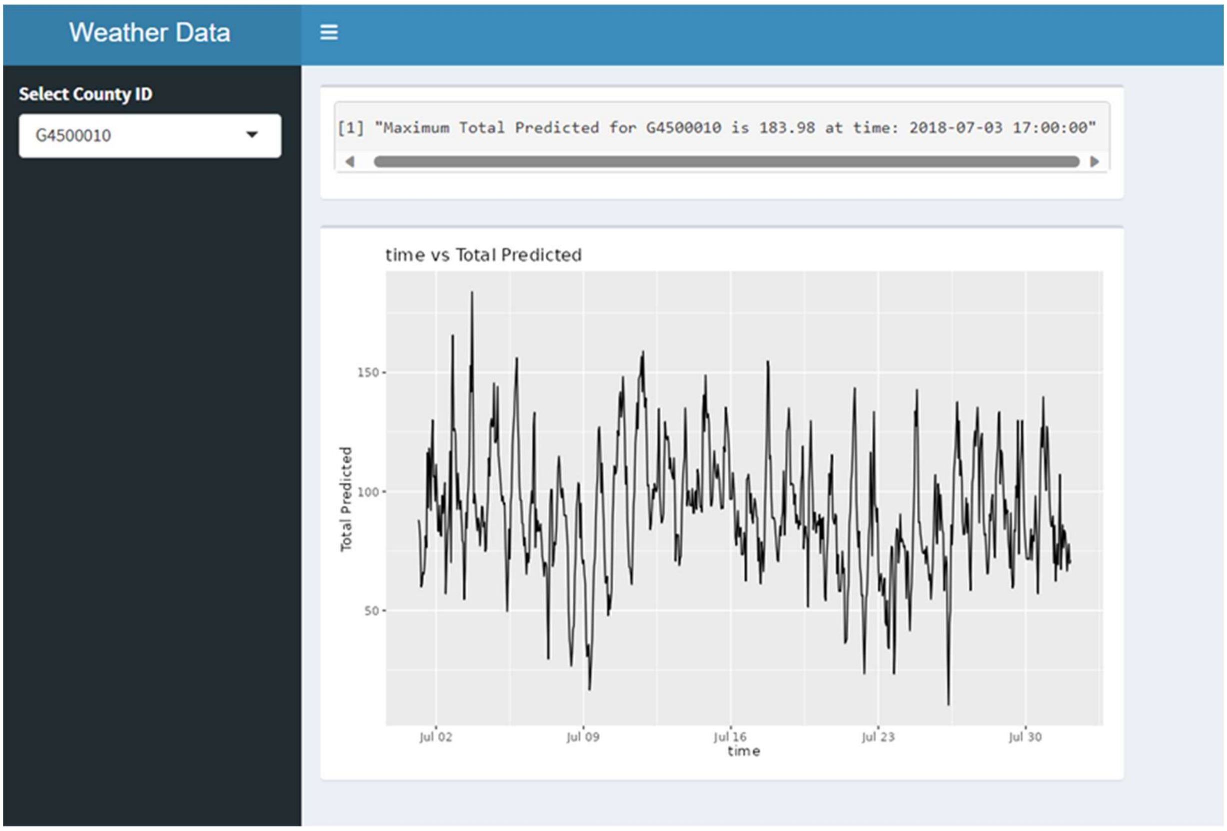
~~~
```{r}
#Predicting output for dependent variable
pred <- predict(lmOut, weatherESTest)
```
~~~

~~~
```{r}
#Adding the prediction to the data
weatherESTest$totalPredicted <- pred
```
~~~

We saved the resulting dataframe as a csv to use it in our shiny app.

Shiny App

A webapp using R Shiny was created so that eSC could interact with the data and to visualize the demand of energy for all the counties. It also displays when in that county will energy demand peak so that the company can be prepared to supply enough energy to that county. The app also plots the demand of energy overtime in that county so that the demand can be met easily.



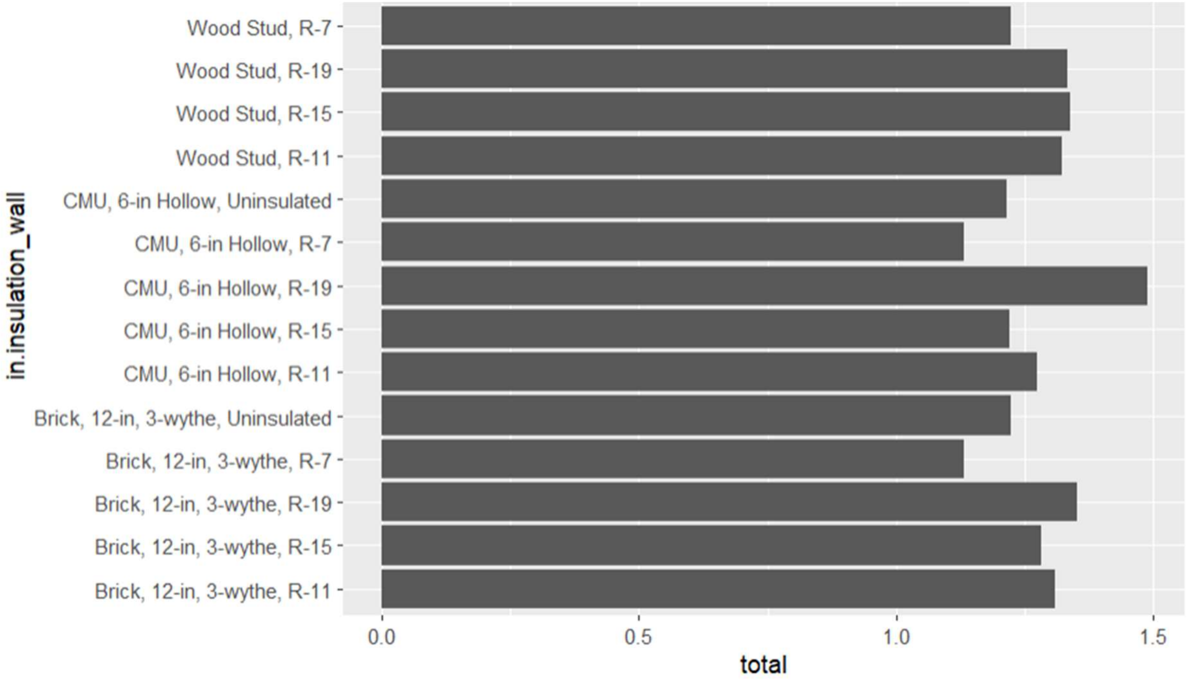
```
[1] "Maximum Total Predicted for G4500010 is 183.98 at time: 2018-07-03 17:00:00"
```

Weather Data Dashboard (shinyapps.io)

Recommendations

eSC required us to provide them with recommendations to reduce the energy demand as creating new power plants is not an option for them. We analyzed the data focusing our attention on the houses and how energy usage changed from one house to another, carefully considering which features could be easily changed by the customers of eSC, this was done because features like number of bedrooms which were having a quite significant impact on the electricity usage would be something that remains constant as changing the number of bedrooms is impractical.

The feature that stood out to us was insulation and how different types of insulations were affecting energy usage, we plotted the average energy consumption for each insulation type to see if there were any differences.



From the above graph we can see that there are certain types of insulation that perform better.

The final recommendation we would make to eSC would be to mandate for better standardization in insulations by lobbying and incentivizing builders so that the energy consumption can be reduced. eSC should also encourage customers to get solar panels at home to offset the energy demands.

Conclusion

Reducing peak energy consumption is a challenge being dealt with worldwide as global warming continues to increase and while it is important for energy companies to come up with ways to reduce the demand it is extremely important to accurately predict the demand so they can provide uninterrupted energy to everyone. Improving the insulation is only going to be the first step towards the goal of reducing demand, accurate housing information, up to date knowledge on various breakthroughs in other industries and opinions of subject matter experts will play an important role in meeting these goals.

Appendix A

Work distribution: -

1. Data pipelining and preparation: - Yashaswi Pandey
2. Analysis and Visualizations: - Kulveen Kaur and Chrish David Douglas
3. Modelling and Predictions: - Tejaswini Vibhute
4. R Shiny application: - Yerramorrusu Harshitha Reddy

Appendix B

Link to the shiny application: -

https://nyvzt1-yashaswi-pandey.shinyapps.io/finalProject-Group2/