IST 687 – Introduction to Data Science

Group 2

Submitted by:

Advait Narvekar - amnarvek@syr.edu

Amol Borkar - aborkar@syr.edu

Harshil - harathod@syr.edu

Jill Karia – jkaria@syr.edu

Rishikesh Thakker - rrthakke@syr.edu

TABLE OF CONTENTS

Project Description

In this project, we saw the concerns of an energy company (eSC) being addressed regarding the potential impact of global warming on electricity demand. eSC is specifically worried about the upcoming summer, particularly the month of July, which is historically the peak period for energy consumption. The primary goal is to prevent blackouts by understanding and managing increased energy demand during exceptionally hot periods.

To get some background about Esc, it provides electricity to residential properties in South Carolina and a small part of North Carolina. Instead of investing in new power plants to meet potential demand surges, eSC aims to identify key drivers of energy usage and develop strategies to encourage customers to save energy. The overall objective is to reduce energy consumption during peak periods, specifically in July, to avoid the need for significant infrastructure expansion.

eSC's approach involved analyzing energy consumption patterns, understanding factors influencing usage, and formulating strategies to incentivize customers to adopt energy-efficient practices. By doing so, eSC aims to ensure a reliable power supply during high-demand periods, contribute to environmental sustainability, and explore cost-effective alternatives to expanding energy production facilities.

This project is crucial for eSC to proactively manage the potential strain on their electrical grid, ensuring uninterrupted service to customers while aligning with environmental goals. It emphasizes a strategic focus on customer behavior, energy conservation, and sustainable practices to create a more resilient and environmentally responsible energy infrastructure.

Project Scope

The primary objective of this project is to address the concerns of the energy company (eSC) regarding potential challenges associated with global warming and increased energy demand. The focus is on proactive management of energy consumption, optimization of resources, and the development of sustainable practices.

Tasks:

1. Data Preparation: Read and merge static house data, energy usage data and weather data. In this way create a unified dataset for analysis.

2. Exploratory Data Analysis (EDA): Explore basic statistics and distributions and visualize energy consumption patterns, correlations, and trends. Gain insights into relationships between energy usage, house attributes, and weather conditions.

3. Energy Usage Prediction Model: Split data into training and testing sets and then experiment with various regression models for predicting energy usage. Select the best-performing model based on evaluation metrics.

4. Model Accuracy: Evaluate the model using appropriate metrics. Provide a clear explanation of the model's accuracy. Present insights into factors influencing energy usage predictions.

5. July Weather Dataset Modification: Adjust the temperature data for July to be 5 degrees warmer and ensure consistency with the original format and structure.

6. Evaluate Peak Future Energy Demand: Use the best model to predict future energy demand in July considering potential changes in customer behavior. Provide a model-driven estimation of peak energy demand.

7. Show Future Peak Energy Demand: Break down future peak energy demand by geographic regions and explore other relevant dimensions or attributes impacting demand.

8. Shiny Application Development: Develop a Shiny application for interactive data exploration. Incorporate features to understand model predictions and future energy needs.

9. Identification of Peak Energy Demand Reduction Approach: Identify potential approaches to reduce peak energy demand. Consider factors like demand response, energy-efficient technologies, etc.

10. Modeling Impact of Peak Demand Reduction Approach: Implement a data-driven approach to model the impact of the identified approach and explain the impact using relevant data and insights.

Project deliverables

1. The code of your analysis
2. A presentation (to the CEO of the power company)
3. A URL for the shiny app
4. A document explaining the work done (will be reviewed by your 'technical manager') → This document must also explain which team member did which task
5. You need to provide an update every two weeks (one per group, not per person) For each update (including for the final submission), provide:
    a. Work done by each person (since the last update)
    b. Work planned to be done by each person (by the next update)
    c. Key issues / challenges → this should be the basis for your deliverable in item
    d. the document explaining the work done (including for example, models that were bad, but still evaluated)

Data Description

We had been provided with 4 datasets for this project. The datasets that we were given were:

1. Static House Dataset

Description:

The "Static House Data" is a dataset containing basic information for a random sample of single-family houses served by eSC. It includes details such as the building ID, which is used to access energy data, and other unchanging attributes like the size of the house. The dataset is designed to provide a snapshot of static house characteristics that do not vary over time.

Size:

Approximately 5,000 houses are included in the dataset.

Format:

The dataset is stored in 'parquet' format, a storage-optimized CSV file.

Link:

https://intro-datascience.s3.us-east2.amazonaws.com/SCdata/static_house_info.parquet

2. Energy Usage Data

Description:

The "Energy Usage Data" dataset provides detailed hour-by-hour energy consumption information for a collection of single-family houses served by eSC. Each house has its own dataset file, containing calibrated and validated energy usage data with 1-hour load profiles. This data covers various sources like air conditioning systems and dryers, offering insights into the specific energy consumption patterns of each house.

Size:

There are approximately 5,000 houses represented in the dataset, with each house having its own individual file.

Format:

The energy usage data is stored in 'parquet' format, a storage-optimized CSV file. Each file is uniquely named after the building ID, which serves as an identifier for the corresponding house. For example, the URL for 'building_id' 102063 is:

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet.

Link:

All files are centrally located in one folder on Amazon AWS. For instance, the directory contains data for various houses, such as: https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData.

3. Meta Data

Description:

The "Meta Data" file serves as a data description document, providing clear and human-readable explanations for the fields used across different housing data files. It acts as a guide to understanding attributes present in both the static house data and the energy usage data.

Link:

The data description file is available at https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/data_dictionary.csv.

Format:

This CSV file contains comprehensive descriptions of attributes found in the dataset, aiding users in interpreting and utilizing the data effectively.

4. Weather Data

Description:

The "Hour-by-hour Weather Information" dataset comprises time series weather data collected for each county, stored based on a unique county code. The weather data is organized into separate files, with one file dedicated to each geographic area, identified by its county code. The county code for each house can be cross-referenced using the 'in.county' column in the static house dataset.

Size:

The dataset covers approximately 50 counties, with each county having its own individual file.

Format:

The weather data is stored in a simple CSV format. Each file is named after the county code, facilitating easy identification and retrieval. For example, the URL for the weather

data of county 'G4500010' is https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weatherdata/G4500010.csv

Link:

All weather files are centrally located in one folder on Amazon AWS, accessible through the directory https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weatherdata/

Business Questions

1. Identifying Counties with High Energy Consumption (G4500710 and G4500630): These countries represent a significant market opportunity for energy companies to reach the perfect target audience for their energy-saving campaigns and smart gadgets.

2. Mid-July Peak Energy Demand: Businesses can implement seasonal energy management strategies to meet peak demand in mid-July. Offering special promotions or services to encourage energy-efficient practices, such as setting thermostats to optimal temperatures during hot summer months, could be part of this.

3. Hourly Energy Consumption Peak (16:00 - 17:00): Companies that specialize in energy-efficient cooling appliances and systems can promote their products during this time period in collaboration with eSC. Innovative cooling technologies that reduce energy consumption can be marketed to these countries' households and businesses.

4. Inefficient Windows and Roofing Materials: Manufacturers of energy-efficient windows and roofing materials can profit from the 73% of the population that does not have adequate insulation. Businesses can run educational campaigns to raise awareness about the benefits of energy-efficient materials and offer incentives to encourage their use.

5. From Incandescent Bulbs to LEDs: Company should promote more usage of LED lamps to reduce the energy consumption.

6. Thermostat Preferences: HVAC (Heating, Ventilation, and Air Conditioning) companies can create smart thermostats that optimize cooling and heating based on outside weather conditions. They can market these devices to homes and businesses with the goal of lowering energy consumption while maintaining comfort in collaboration with the enrgy provider as they know the perfect target audience for the same.

7. The Effects of Global Warming and Energy Efficiency Campaigns: Businesses can position themselves as sustainability leaders by organizing energy efficiency campaigns. Educational workshops, energy-saving tips, and product demonstrations can all be part of these campaigns. Companies can also invest in R&D to create innovative, eco-friendly products that meet future energy efficiency requirements.

8. Decision-Making Based on Data: Companies can invest in data analytics tools to gain insights into these countries' energy consumption patterns and preferences. This information can be used to inform product development and marketing strategies, ensuring that businesses meet the unique needs of the local population.

Data Cleaning and Merging

We were provided with four different datasets that we could use out of which the two datasets; namely static house data set and the metadata were easy to download using R.

The other two datasets, namely the weather data and. energy consumption data were much trickier to collect because one file contained only information about one county in case of weather data or one building in case of energy consumption data. Download all the data by writing a custom script in which the used building ids are mentioned in the static data house and download each file from the. energy consumption data set and then merged all these files into one huge data asset.

Once this process was completed, we modified the code a little to do the same thing for the weather data and download the data for each county. This is how we collect the totality of the data available to us.

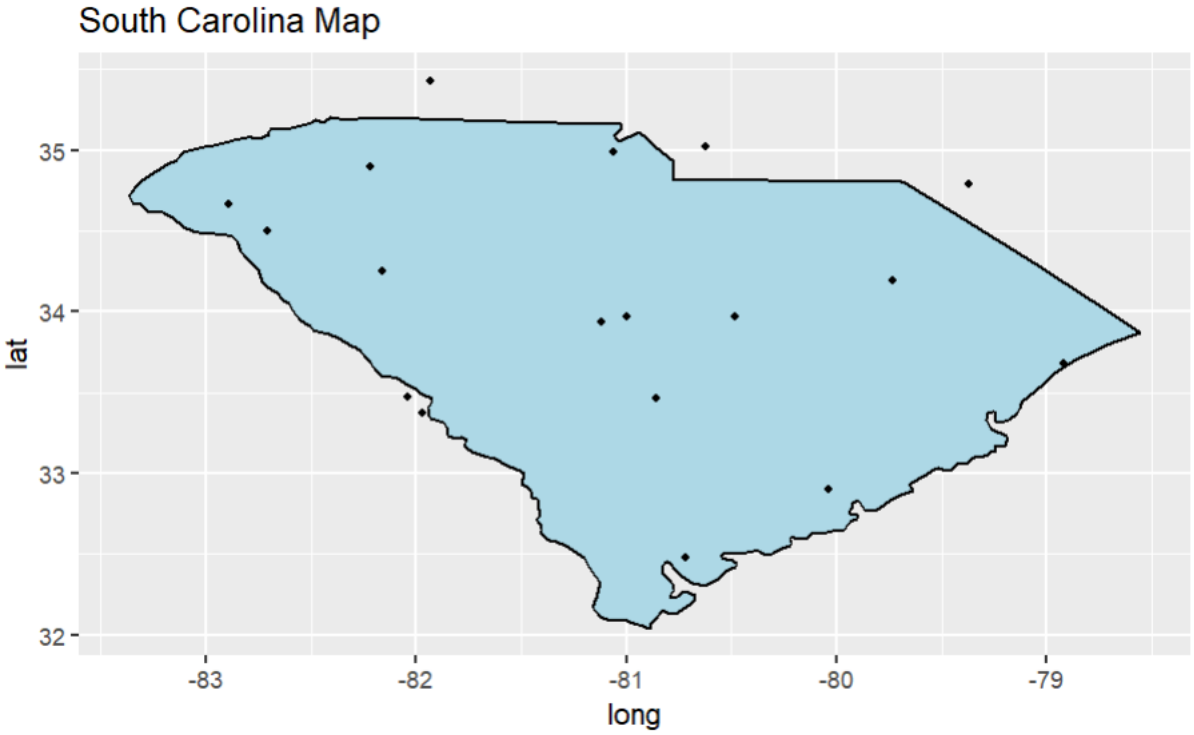
Once we had all the data in the next part was merging all the data sets into just 1 big data file. We did this using the building id column. Once the data was merged, we looked at the data descriptions in the metadata file to determine which columns would be useful to us in our analysis. We narrowed down around 23 columns, and then we tested these columns by creating linear models to see what accuracy we could achieve. The higher the accuracy, better the model.

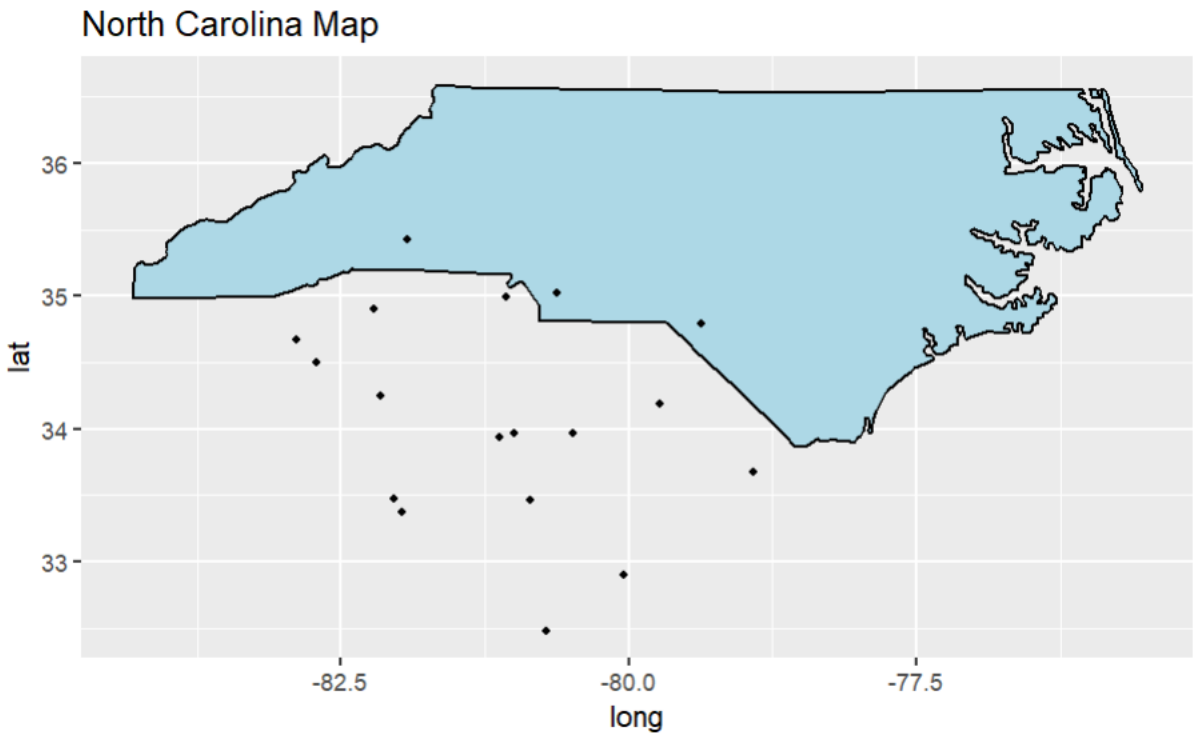
This process further enhanced our understanding of which columns can work as good predictors for the predictive modelling stage.

Exploratory Data Analysis

To perform exploratory data analysis, we are getting started by the metadata file and looking at the column descriptions mentioned. Then we used a bunch of R functions like str() and summary() to analyze various columns and draw some basic insights about the data. After this we tried to plot some graphs to visualize the trend for the energy consumption over the day.

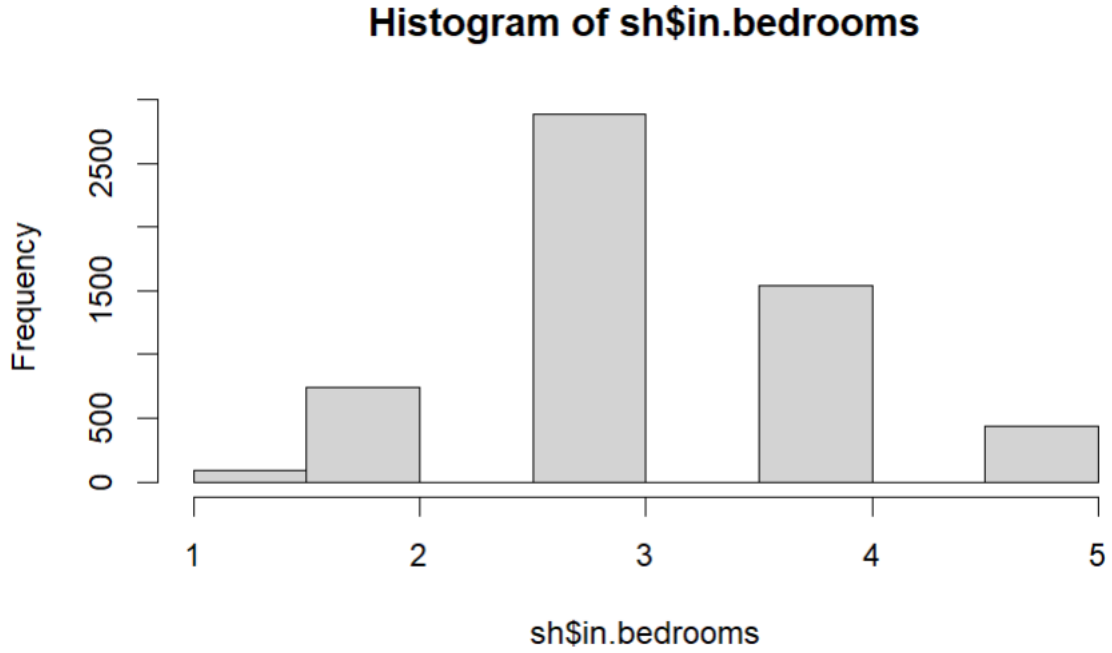
Some of the visualizations that we did are



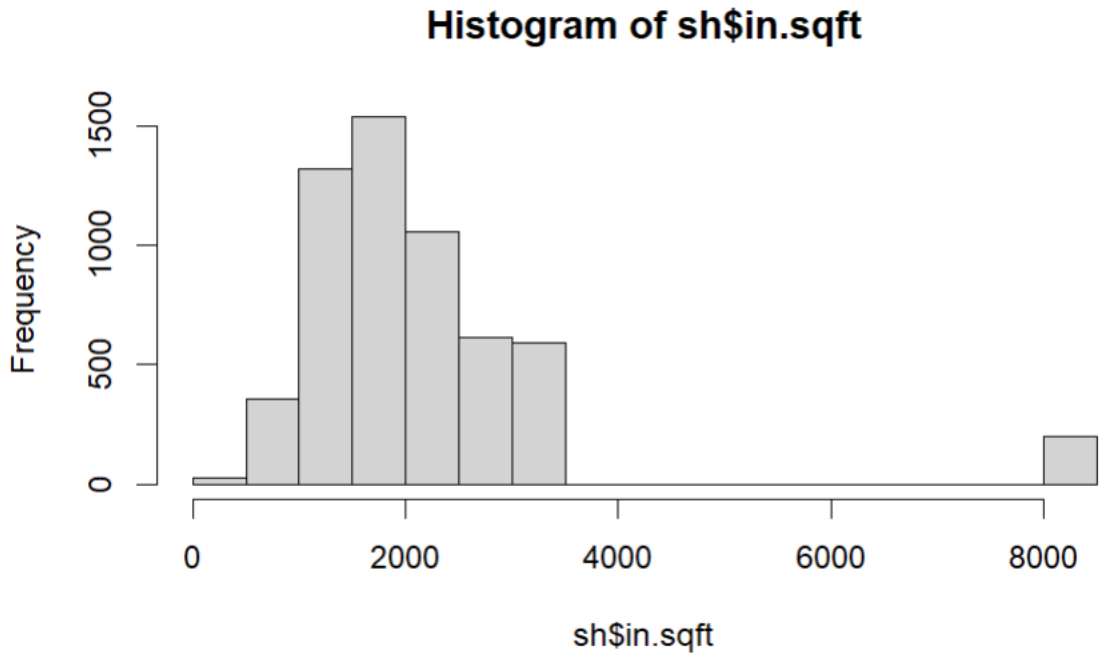


This shows the county's that we are doing analysis on, on the map. We mapped all the counties in both the South and North Carolina map.

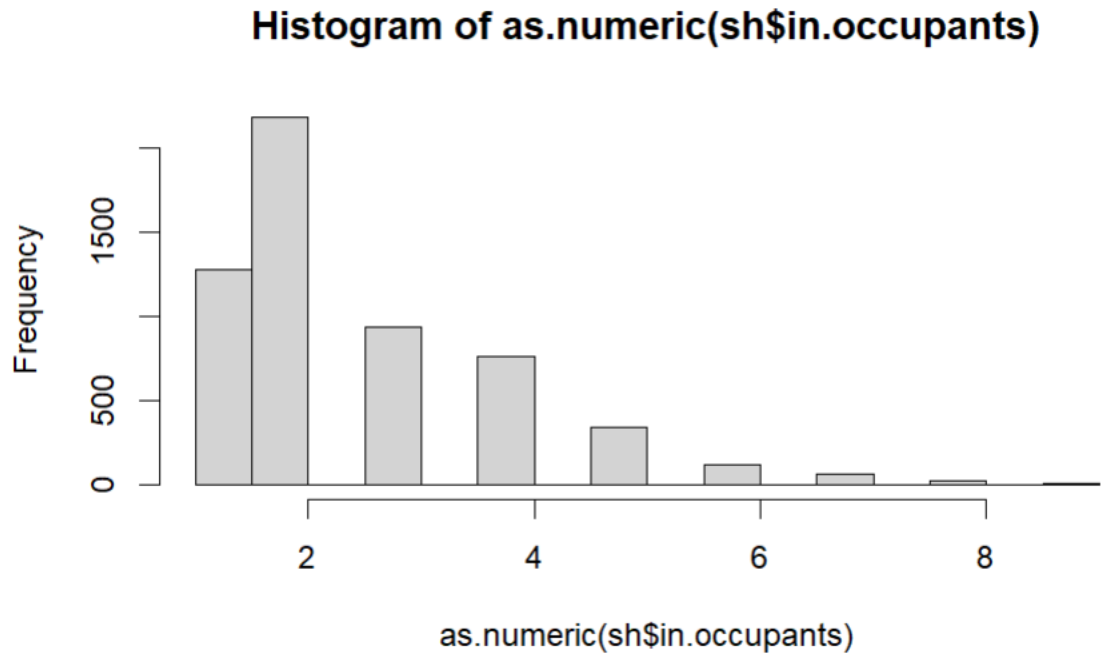
The energy consumption per house depends on the bedrooms present. So we saw a histogram to see how many bedrooms does each house have. We can see how most houses have 3 bedrooms, very few having 1,2 and 5.

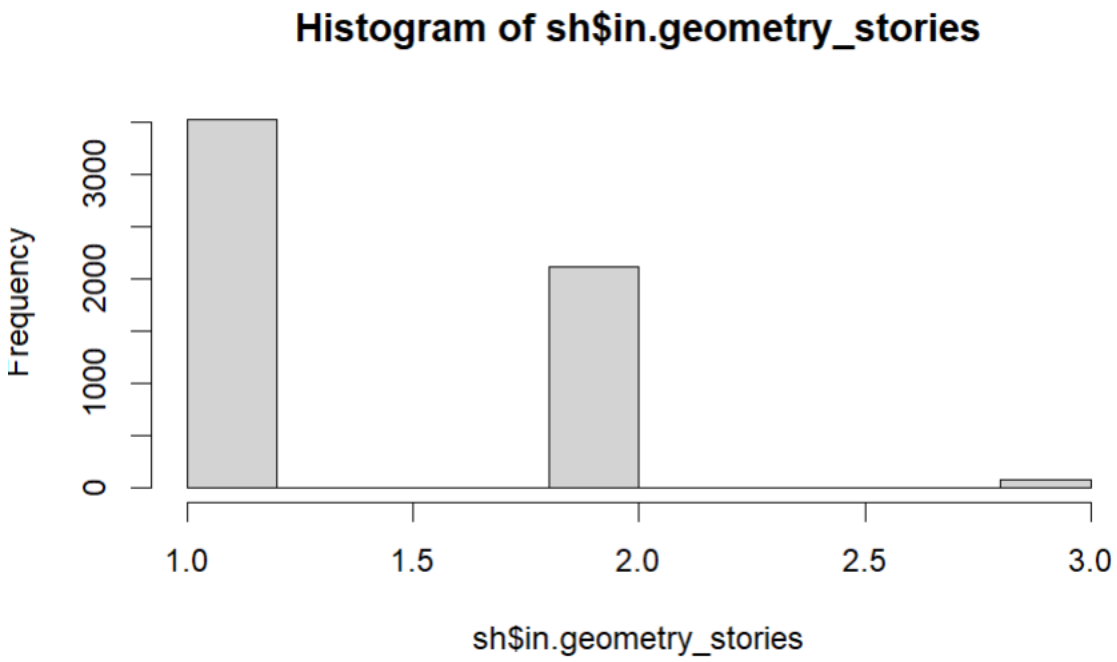


We then realized how regardless of the number of bedrooms the larger the square feet of the house the higher is the energy consumption so:

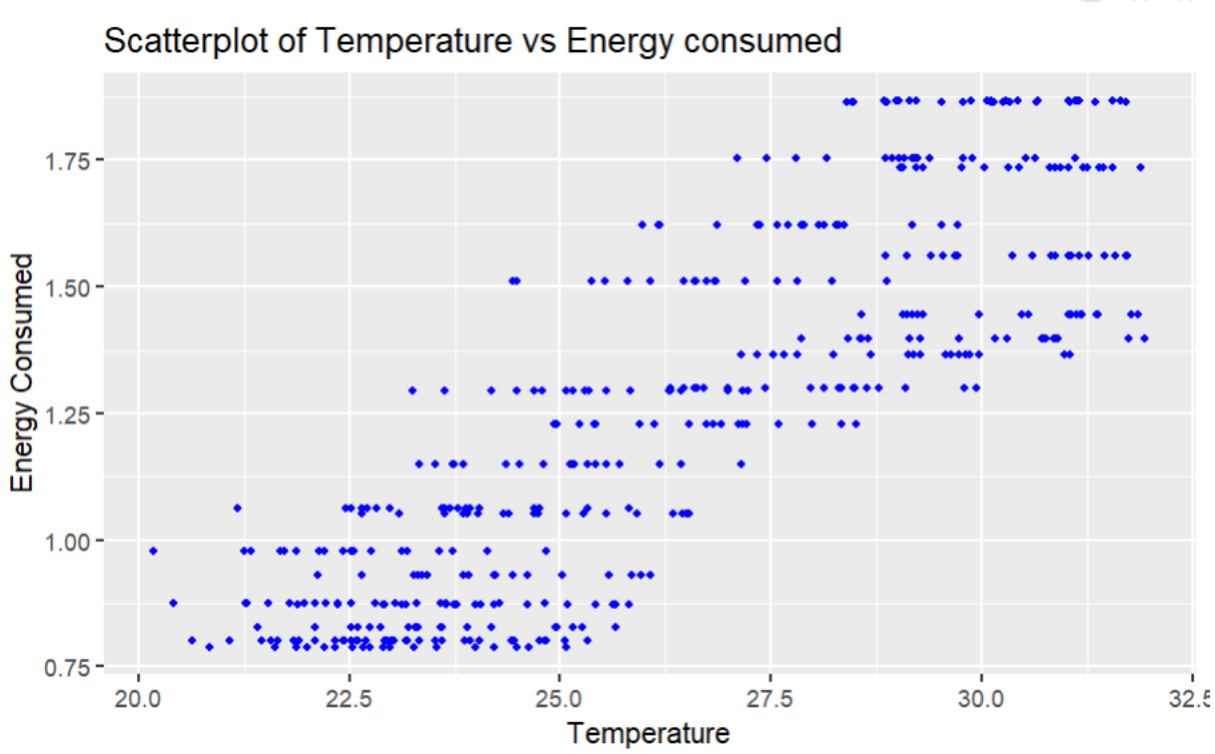


With the same assumption we check the amount of occupants in each house and the house stories

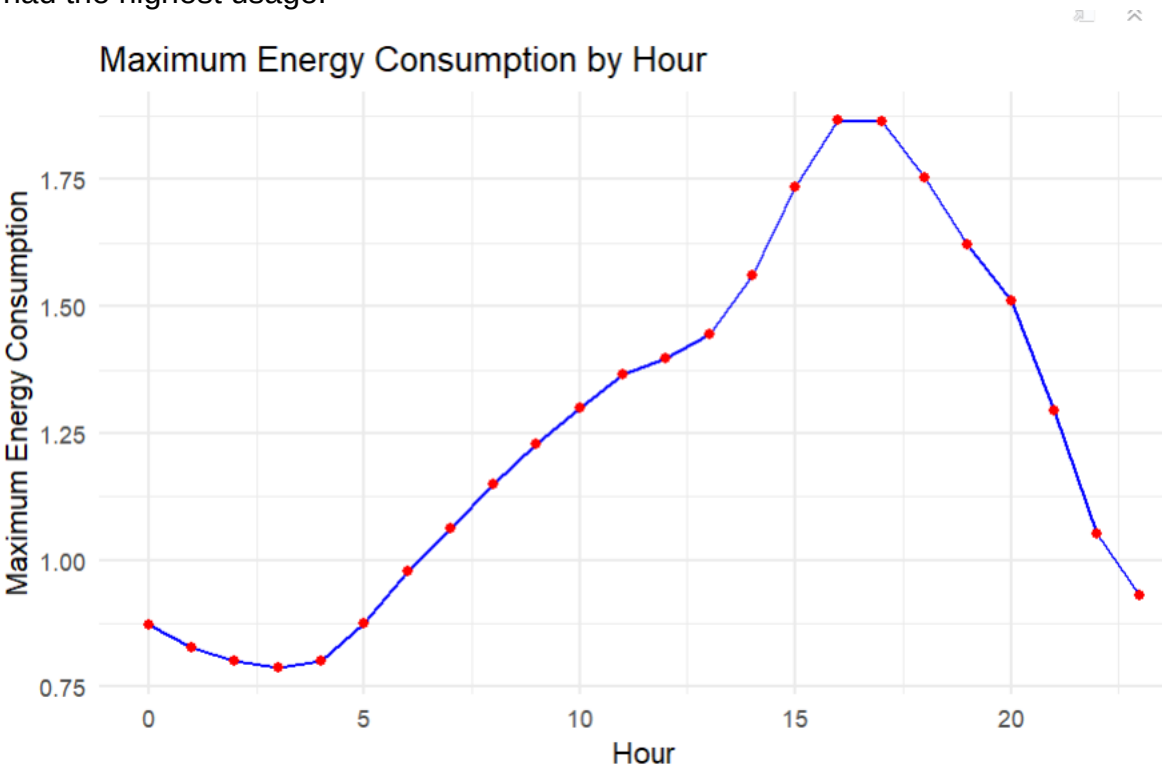




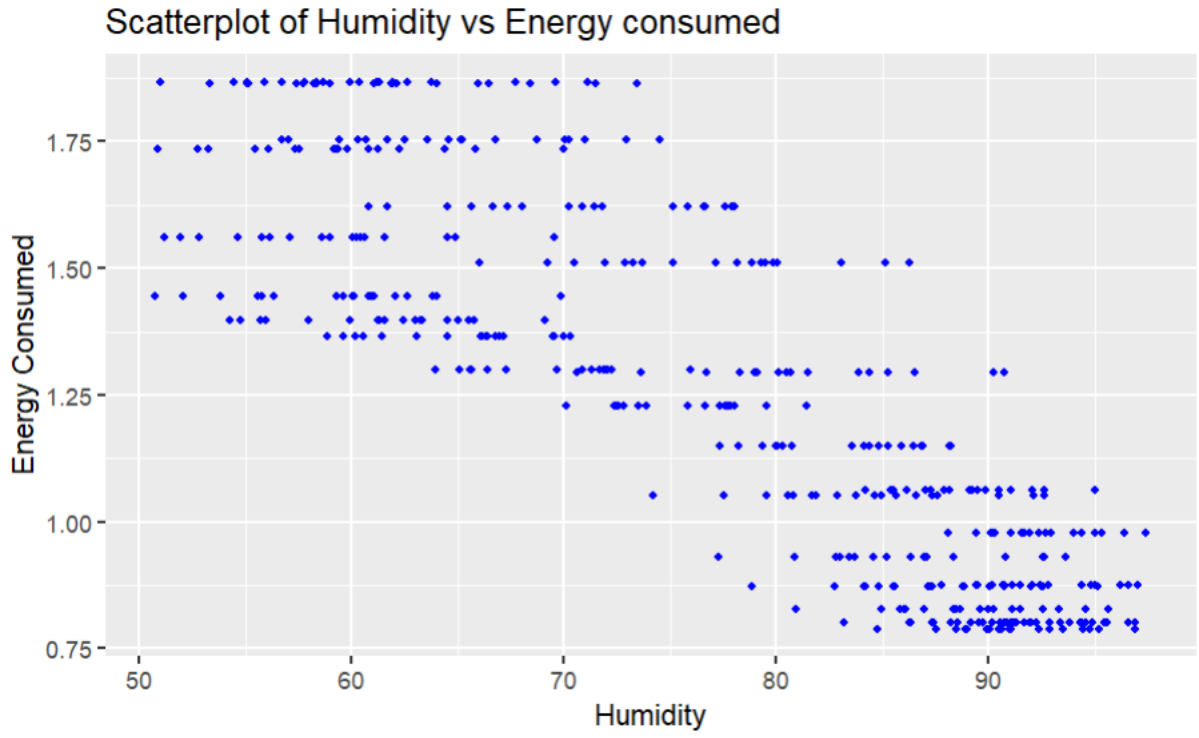
We also analyzed which were several factors affecting energy consumption with respect to weather and found out that as the temperature increases the energy consumption increases as the cooling units are being utilized more and they are being set to lower temperature points.



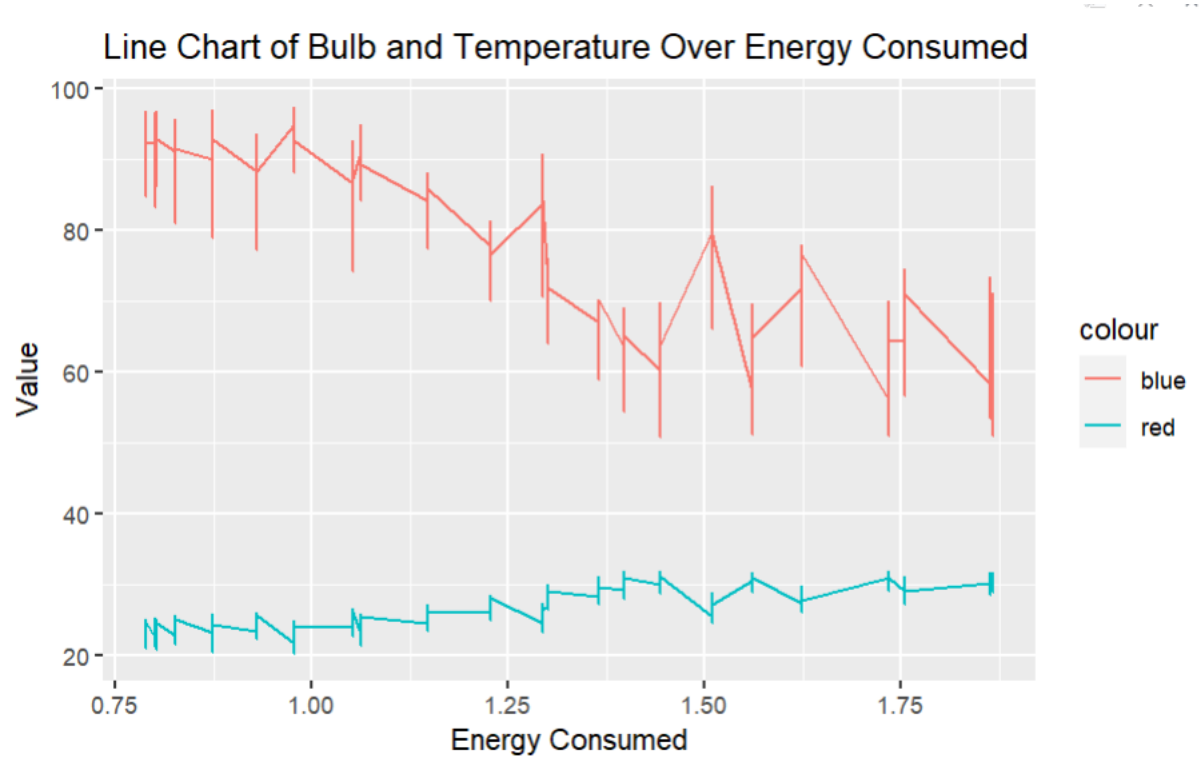
We created a line chart to visualize this energy consumption per hour and saw how 4pm had the highest usage:

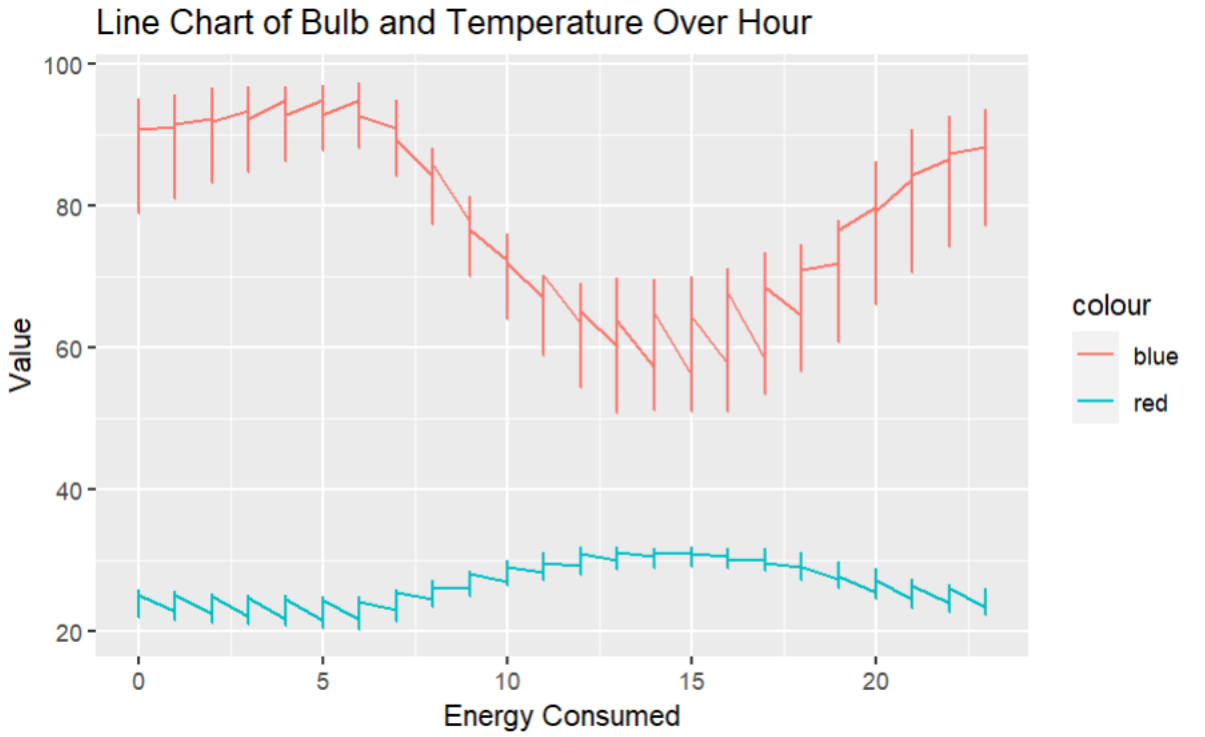


We also observed that there was an inverse relation between the energy consumption and humidity levels in the air.



The scatterplot helped us to learn their interaction and correlation and helped us realize why linear modelling was the way to go. We also used line chart for visualizing and comparison





Modelling and Code Explanation

For our modelling we tried three different models: Linear, Decision Tree and Random Forest

Linear Model:

A linear model is a type of mathematical model that represents a relationship between a dependent variable and one or more independent variables. The basic idea behind a linear model is to assume that this relationship is linear, meaning it can be expressed as a linear combination of the input variables. The general form of a linear model for a single independent variable is: y = mx + b

For our linear model we first started with data cleaning and merging. After that we created a catch loop to take all the relevant and helpful columns for linear modelling.

```r
#Changing column names to make them more accessible
colnames(complete_data) <- make.names(colnames(complete_data))

#Creating a list to store columns that didn't have more than one factors
rm_col_list <- c()

#Removing the outcome column name
columns <- colnames(complete_data)
columns <- columns[!columns %in% 'mean_total_energy_consumed']

for (i in columns){
    tryCatch(
        summary(lm(mean_total_energy_consumed ~ .,data =
complete_data[,c('mean_total_energy_consumed',i)])),
    error = function(e){
        rm_col_list <<- c(rm_col_list, i)
    })
}

#Creating a backup of complete data for future use
complete_data_bkp <- complete_data
```

After that we transformed and cleaned the data more to remove the nulls using tidyverse functions

We individually checked p value of all columns and directly removed the columns if they were not significant.

```
#Now checking the p-value for all remaining individual columns and rejecting
them if they don't satisfy alpha level = 0.05

reject_list = c()

#Removing the outcome column name
columns <- colnames(temp_complete_data)
columns <- columns[!columns %in% 'mean_total_energy_consumed']

for(i in columns){
    a = lm(mean_total_energy_consumed ~ .,data =
temp_complete_data[,c('mean_total_energy_consumed',i)] )
    pVal <- anova(a)$'Pr(>F)'[1]
    if(pVal>0.05 | is.na(pVal)==TRUE){
        reject_list <- c(reject_list,i)
    }
}

#Removing columns which has only one value or one value and null values
delete_list = c()
col <- colnames(temp_complete_data)
col <- col[!col=='mean_total_energy_consumed']
for(i in col){
    if (nrow(unique(temp_complete_data[,i]))<=3 ){
        delete_list = c(delete_list, i)
    }
    tryCatch(
        if(any(unique(temp_complete_data[,i]) == 'None'))
            delete_list = c(delete_list, i)
```

After that we deleted all the models and ran a linear model. We got an r-squared value of 87% which was low so we created a loop to add values that make the adjusted r squared better but still not too high to avoid overfitting.

```r
#Creating a train and test data sets
set.seed(1)
row.number <- sample(1:nrow(transformed_data), 0.95*nrow(transformed_data))
summary(transformed_data)
train = transformed_data[row.number,]
test = transformed_data[-row.number,]

str(transformed_data)
#Checking out there dimensions
dim(train)
dim(test)

one_dim_test <- head(test,1)

one_dim_test_increased_temp <- one_dim_test
one_dim_test_increased_temp$Dry.Bulb.Temperature...C. <-
one_dim_test_increased_temp$Dry.Bulb.Temperature...C. + 5
#Creating a model using linear regression
model.lm <- lm(mean_total_energy_consumed ~ .,data = train )

#Checking out the details of the model
summary(model.lm)

#Checking out the accuracy of the model by predicting values for test data
set
predicted_output <- predict(model.lm, newdata = one_dim_test)
predicted_output_f <- predict(model.lm, newdata =
one_dim_test_increased_temp)
predicted_output
```

After we got the linear model we took the prediction and saw the accuracy to be 91% and got good prediction. From the above cod we also added 5 degrees in temperature to calculate p value and we got both future and current energy prediction

Random Forest:

We used our transformed data to calculate accuracy of random forest model.

A Random Forest is an ensemble learning technique that is used for both classification and regression tasks. It belongs to the class of ensemble methods, which combine the predictions of multiple models to improve overall performance and generalization. Random Forests are particularly powerful and versatile, known for their robustness and ability to handle complex datasets.

We used randomForest() library for that and created a model. We got its accuracy to be 99% which means the model is overfitted so we ended up skipping that

```{r}
# install.packages("randomForest")
library(randomForest)
library(caret)

# Split the data into training and testing sets
sample_index <- sample(1:nrow(transformed_data), 0.7 *
nrow(transformed_data))
train_data <- transformed_data[sample_index, ]
test_data <- transformed_data[-sample_index, ]

# Build the random forest model
rf_model <- randomForest(mean_total_energy_consumed ~ ., data = train_data)

# Make predictions on the test set
predictions <- predict(rf_model, test_data)
# Calculate R-squared using the caret package
rsquared <- R2(predictions, test_data$mean_total_energy_consumed)
cat("R-squared:", rsquared, "\n")
#we skip this because of overfitting
```

```
R-squared: 0.9999829
```

Decision Tree Model:

The third model that we tried was a decision tree model.

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each node represents a decision or a test on a feature, each branch represents an outcome of the test, and each leaf node represents the predicted outcome or class label. The topmost node in a decision tree is known as the root node

For decision tree model we downloaded the rpart library and made the model and checked its accuracy. It had an accuracy of 95% and was again over fitted so we ended up skipping it.

Trying Decision Ter

```{r}
# install.packages("rpart")
library(rpart)
# Train a decision tree model
tree_model <- rpart(mean_total_energy_consumed ~ ., data = transformed_data)
# Make predictions
predictions <- predict(tree_model)

rsquared <- cor(predictions, transformed_data$mean_total_energy_consumed)^2
rsquared
#the model is overfitted
```

```
[1] 0.9517543
```

Shiny Application

Overview

Our Shiny Application serves as an interactive platform for exploring energy consumption patterns and predicting future energy demand. Users, including the CEO of the power company, can leverage this tool to gain insights into the factors influencing energy consumption and make informed decisions to manage peak demand effectively.

Features

1. Linear Regression Model Summary:
- Users can select a specific column for analysis.
- The app provides a summary of the linear regression model, offering insights into the relationships between selected variables and energy consumption.

2. Histogram Visualization:
- Users can visualize the distribution of the selected variable with respect to hours through an interactive histogram.

3. Energy Consumption Prediction:
- The app allows users to input specific conditions, such as temperature, humidity, and wind speed, to predict energy consumption.
- Future energy consumption predictions are provided, considering potential changes in environmental factors.

How to Use

1. Select Column for histogram analysis:
- Choose a column from the dataset to analyze.

2. Generate Model:
- Click on the "Generate Model" button to create a linear regression model based on the selected column.

4. Future and Current Energy Prediction:
- Input environmental conditions for a specific hour and predict both current and future energy consumption.

Usage Scenario

Our Shiny App provides a user-friendly interface for the CEO to interact with the data, understand energy consumption patterns, and make data-driven decisions to optimize resources and address potential challenges associated with increased demand.

Energy Consumption Prediction App

Select Column:

mean_total_energy_consumed

Linear Regression Model Generation:

Generate Model

Select an hour:

0 16 23

0 3 6 9 12 15 18 21 23

Enter Dry Bulb Temperature(C) Value:

35

Enter Relative Humidity Value:

4

Enter Wind Speed (m.s,) Value:

3

Enter Wind.Direction (Deg.) Value:

2

Enter Global.Horizontal.Radiation (W.m2.) Value:

4

Direct Normal Radiation (W.m2.) Value:

6

Diffuse Horizontal Radiation (W.m2.) Value:

5

Select an in.puma:

G45000103

In Reeds Balancing Area Value:

6

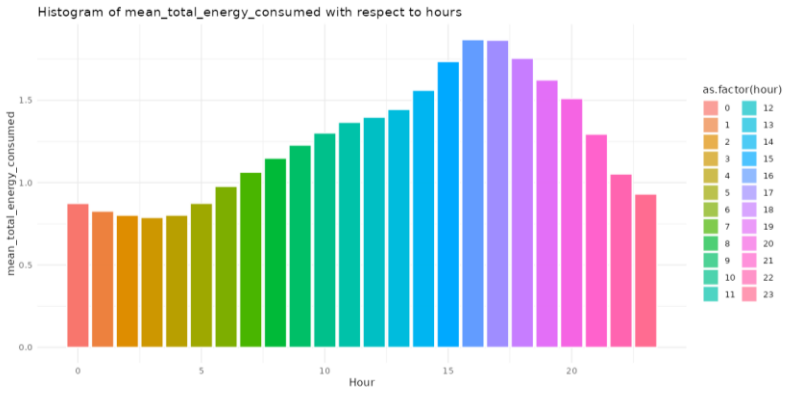
In Weather File Longitude Value:

5

Future Energy Prediction

Increase in Temperature (C):

Histogram:

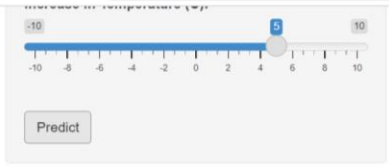


Linear Regression Model Summary:

```
Call:
lm(formula = mean_total_energy_consumed ~ ., data = train_Data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32094 -0.07542 -0.00543  0.06635  0.41817

Coefficients:
                                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)                        3.867e-01  3.399e-01    1.138    0.255
hour                               7.386e-03  6.168e-05  119.741  < 2e-16 ***
Dry.Bulb.Temperature...C.          1.177e-02  9.282e-04   12.680  < 2e-16 ***
Relative.Humidity....             -1.887e-02  2.160e-04  -87.338  < 2e-16 ***
Wind.Speed..m.s.                   1.626e-01  8.378e-04  194.057  < 2e-16 ***
Wind.Direction..Deg.               3.510e-06  1.374e-05    0.255    0.798
Global.Horizontal.Radiation..W.m2. -1.126e-03  9.775e-06 -115.224  < 2e-16 ***
Direct.Normal.Radiation..W.m2.     7.963e-04  4.463e-06  178.420  < 2e-16 ***
Diffuse.Horizontal.Radiation..W.m2. 4.086e-04  2.264e-05   18.047  < 2e-16 ***
in.pumaG45000102                   2.837e-02  2.801e-03   10.128  < 2e-16 ***
in.pumaG45000103                   2.840e-02  2.587e-03   10.979  < 2e-16 ***
in.pumaG45000104                   2.863e-02  2.879e-03    9.944  < 2e-16 ***
in.pumaG45000105                   8.363e-02  2.577e-03   32.455  < 2e-16 ***
in.pumaG45000200                  -1.180e-01  2.106e-03  -56.018  < 2e-16 ***
in.pumaG45000301                   2.829e-02  2.761e-03   10.247  < 2e-16 ***
in.pumaG45000302                   2.749e-02  2.506e-03   10.970  < 2e-16 ***
in.pumaG45000400                   1.428e-01  3.319e-03   43.034  < 2e-16 ***
in.pumaG45000501                   2.050e-01  4.571e-03   44.849  < 2e-16 ***
in.pumaG45000502                   2.059e-01  4.508e-03   45.668  < 2e-16 ***
in.pumaG45000601                   4.915e-02  3.906e-03   12.584  < 2e-16 ***
in.pumaG45000602                   2.535e-02  4.113e-03    6.163 7.17e-10 ***
```




```
in.pumaG45000603                  1.529e-01  4.694e-03   32.580  < 2e-16 ***
in.pumaG45000604                  1.534e-01  4.755e-03   32.259  < 2e-16 ***
in.pumaG45000605                  1.535e-01  5.177e-03   29.659  < 2e-16 ***
in.pumaG45000700                  9.172e-02  6.016e-03   15.247  < 2e-16 ***
in.pumaG45000800                  1.534e-01  5.555e-03   27.618  < 2e-16 ***
in.pumaG45000900                  1.897e-01  6.537e-03   29.028  < 2e-16 ***
in.pumaG45001000                  1.825e-01  7.631e-03   23.920  < 2e-16 ***
in.pumaG45001101                  1.754e-01  7.950e-03   22.056  < 2e-16 ***
in.pumaG45001102                  1.734e-01  7.909e-03   21.926  < 2e-16 ***
in.pumaG45001201                  2.065e-01  5.733e-03   36.021  < 2e-16 ***
in.pumaG45001202                  2.064e-01  5.640e-03   36.594  < 2e-16 ***
in.pumaG45001203                  2.062e-01  5.704e-03   36.155  < 2e-16 ***
in.pumaG45001204                  2.062e-01  5.682e-03   36.293  < 2e-16 ***
in.pumaG45001300                  2.063e-01  4.512e-03   45.716  < 2e-16 ***
in.pumaG45001400                  2.317e-01  4.608e-03   50.293  < 2e-16 ***
in.pumaG45001500                  1.670e-01  3.257e-03   51.271  < 2e-16 ***
in.pumaG45001600                  1.676e-01  2.783e-03   60.221  < 2e-16 ***
in.reeds_balancing_area          -7.657e-02  2.420e-03  -31.643  < 2e-16 ***
in.weather_file_longitude        -1.089e-01  2.218e-03  -49.087  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.107 on 130148 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.9074
F-statistic: 3.27e+04 on 39 and 130148 DF,  p-value: < 2.2e-16
```

Predicted Outcome:

Predicted Energy Consumption: 0.36 kWh

Future Predicted Energy Consumption: 0.41 kWh

URL: https://2fkohm-advait-narvekar.shinyapps.io/Project/

Peak Energy Demand Management

To implement peak energy demand management, eSC (the energy company) can take a multi-faceted approach using insights gained from the analysis and modeling performed. Here are potential strategies:

1. Time-of-Use Pricing:

Implement time-of-use pricing to encourage consumers to shift energy usage to non-peak hours. This involves charging higher rates during peak periods (like at 4pm like we observed) and lower rates during off-peak hours. The Shiny app could be used to communicate these pricing changes and educate consumers about potential cost savings.

2. Demand Response Programs:

Implement demand response programs that incentivize consumers to reduce energy usage during critical peak periods. Notify consumers in advance through the Shiny app and offer rewards or discounts for participating in energy-saving initiatives.

3. Energy Efficiency Campaigns:

Leverage insights from the EDA to identify factors contributing to high energy usage. Launch targeted energy efficiency campaigns through the Shiny app, providing personalized recommendations to consumers on how to reduce energy consumption.

4. Smart Home Devices:

Encourage the adoption of smart home devices that allow users to monitor and control their energy usage. The Shiny app can integrate with these devices, providing real-time data and enabling users to make informed decisions about their energy consumption.

5. Weather-Based Predictions:

Utilize the weather data to predict upcoming temperature increases. Proactively communicate with consumers through the Shiny app, advising them to take energy-saving measures on days with anticipated high temperatures.

6. Public Awareness:

Launch public awareness campaigns through the Shiny app and other channels to educate consumers about the environmental and cost benefits of reducing energy usage during peak periods. Highlight success stories and showcase the positive impact of collective efforts.

9. Continuous Monitoring and Feedback:

Implement continuous monitoring of energy usage patterns and user feedback through the Shiny application. Use this information to refine strategies, improve communication, and adapt to changing consumer behaviors.

Impact

The impact of the project on the energy company (eSC) and the broader community can be significant. Here are potential impacts across various aspects:

1. Operational Efficiency:

The implementation of data-driven models and predictive analytics allows eSC to optimize energy distribution, reducing the risk of blackouts during peak demand periods. This improves the overall operational efficiency of the company.

2. Cost Savings:

Time-of-use pricing and demand response programs can lead to a more balanced distribution of energy demand throughout the day, potentially reducing the need for costly infrastructure upgrades. This results in cost savings for the energy company.

3. Environmental Sustainability:

By promoting energy efficiency, encouraging the use of renewable energy, and reducing overall energy consumption during peak hours, the project contributes to environmental sustainability. This aligns with global efforts to mitigate climate change.

4. Customer Engagement and Satisfaction:

The Shiny app provides a user-friendly interface for customers to monitor and control their energy usage. Educational campaigns and incentives foster a sense of engagement and satisfaction among customers, leading to a positive relationship with the energy company.

5.Adaptability and Resilience:

The use of predictive models and continuous monitoring allows eSC to adapt to changing energy consumption patterns and environmental conditions. This enhances the resilience of the energy infrastructure.

6. Innovation and Technology Adoption:

The project showcases the adoption of innovative technologies, such as smart meters, predictive modeling, and the Shiny app. This positions eSC as an industry leader in embracing technological advancements for the benefit of both the company and its customers.

7. Economic Stimulus:

The cost savings achieved through energy optimization may lead to economic benefits for both the energy company and its customers. Additionally, investments in renewable energy and energy-efficient technologies can contribute to economic growth.

8. Data-Driven Decision Making:

The emphasis on data-driven decision-making ensures that strategies and initiatives are grounded in evidence. This leads to more effective and targeted interventions for energy optimization.

In summary, the impact of this project is multifaceted, encompassing operational efficiency, cost savings, environmental sustainability, customer satisfaction, community empowerment, and more. The positive outcomes contribute to the long-term success and resilience of the energy company while aligning with broader societal goals for sustainability and responsible resource management.

Conclusion

Overall, this project was an excellent exercise which allowed us to practice some real-world data science and allowed us to dive into the nitty gritty of a challenge that one might encounter in the workplace.

Not only did we learn how to make accurate predictive models, which can predict future energy demand based on certain parameters, but also how to clean the data and perform exploratory data analysis. We also learned how to work with different type of data sources like parquet and csv. Another great learning was learning how to display our findings using a shiny web app which we thought was also quite enlightening.

One thing that proved to be a challenge was the sheer number of columns that were present in our data and being able to find the right columns for doing predictive modeling.

Group Members and Tasks

Name	Tasks
Advait Narvekar	-Merging -Modelling
Amol Borkar	-Cleaning -EDA
Harshil	-PPT -EDA
Jill Karia	-Modelling -Shiny Application
Rishikesh Thakker	-Shiny Application -EDA