IST 687: Introduction to Data Science

Prof. Akit Kumar

PROJECT REPORT

ESCs ENERGY CONSUMPTION ANALYSIS

GROUP 09

BAROT PRATIK

MODI KUSHAL

SRIVASTAVA DIVYANSHU

VARTAK MAYURA

CONTENTS

INTRODUCTION

ESC Electrical is a local company that specializes in electrical services for both homes and businesses. Established in 2011, they offer a wide range of services, including quality installations and customer satisfaction. The company provides electricity (power) to residential properties in South Carolina (and a small part North Carolina.

The problem statement assumes us to be in the role of consultant for eSC.

Global warming worries eSC, especially as it relates to how it may affect the demand for their electricity. To put it briefly, they are concerned that the demand placed on their electrical infrastructure, that is their capacity to provide clients with electricity when they need it would be excessive in the upcoming summertime which will result in too much demand on their electrical grid. Blackouts will result from this, which is something eSC wants to prevent.

Instead of expanding their capacity to supply their clients with additional energy meaning constructing a new power plant, they would like to learn more about the main factors influencing energy consumption and how they can best motivate their consumers to reduce their energy use.

To put it briefly, they want to minimize energy use if summer of 2019 is "extra hot" in order to meet demand instead of constructing a new energy production facility. The ecosystem would benefit from this strategy as well!

eSC is focusing on energy usage in July, typically the peak month for energy demand. The research question centers on using weather data to reduce energy consumption in homes, which could lower energy costs and lessen environmental harm. Many elements influence energy consumption efforts, with a primary driver being environmental sustainability. Through better energy management, companies can decrease their ecological impact and support broader efforts to address climate change.

DESIGN OF RESEARCH QUESTION AND METHODOLOGY

This research question explores how weather data can enhance building energy efficiency and cut energy costs. It recognizes that energy requirements vary across buildings and that geographic location significantly affects weather patterns. Occupant behavior also plays a crucial role in energy use, making it essential to account for these variables when devising weather-based optimization strategies. This study aims to uncover how weather data correlates with energy consumption and identify opportunities to optimize building operations, such as adjusting heating and cooling systems, maximizing natural ventilation, and employing smart energy management systems.

The primary objective is to cultivate a thorough comprehension of how weather data can be effectively utilized to reduce energy consumption and environmental effects in buildings through various data modeling techniques, such as Linear Regression, Support Vector Machine, and Associative Rule Mining. Utilizing this data helps in optimizing energy usage and decreasing overall energy costs. By analyzing the interplay between dependent and independent variables and building energy demand, strategies can be developed to minimize energy use during peak demand periods. Furthermore, weather forecasts can be employed to anticipate future energy requirements and adjust building operations accordingly. This approach can result in more efficient energy resource usage and lower greenhouse gas emissions.

Research question:

How can weather data be leveraged to predict energy consumption?

Methodology:

As suggested by the professor we have followed a streamlined methodology to conduct this research where in the data was first analyzed to formulate an appropriate research question, after this the data was again carefully analyzed to have a clean data and then the three datasets were merged to form a data frame on which analysis can be conducted.

After finalization of the data frame, different machine learning models were used to make predictions and were analyzed to understand which was appropriate.

Lastly, visualizations were created for better understanding to draw conclusions and recommendations.

DATA ANALYSIS AND CLEANING

The following were the datasets that we considered to merge our required data frame:

DATA SETS:

1. House data
2. Energy usage data
3. Weather data
4. MetaData (Data Dictionary)

PREPARING THE DATASETS:

As our research question primarily focuses on the usage of weather data for forecasting the energy utilization in a geographic region (in our case SC, Greenville), we accordingly selected columns.

Housing Data:

The focus being on weather conditions, we select limited attributes from the housing data which are either essential for energy consumption calculation (e.g. sqft of the house, the insulation type, etc.) or change significantly with changing weather conditions (e.g. heating and cooling parameters like heating type and cooling type). Additionally, we have kept the identifier bldg_id for the merging of the datasets.

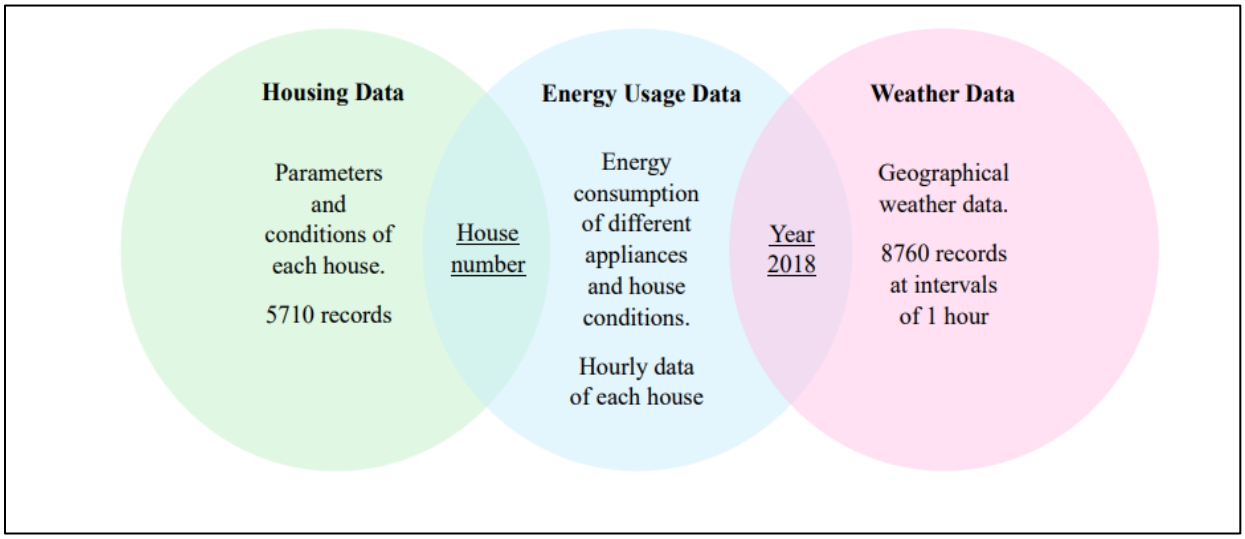
Energy Data:

For each house, we have consolidated the energy output for various types and summed it into 1 value each for electricity, fuel, and gas on an hourly basis. We have selected these 3 columns which are the focus on our energy consumption side in addition to the timestamp (variable time) for merging the data with the weather dataset and we have introduced bldg_id variable to merge the data with the housing dataset.

Weather Data:

We have considered all the 8 columns from weather data so far (subject to change in future iterations as per requirements) as our primary focus is to use various weather parameters to predict the energy consumption. It is joined with the energy consumption of each house based on the datetime and adds the weather parameters to the merged dataset for the county to which the house belongs.

DATA MERGING



As mentioned in the data preparation stage, the merged dataset includes the columns relevant to the research question.

The datasets have been merged based on bldg_id (identifier for housing data and energy data) and timestamp (identifier for weather data and the merged energy and housing dataset). The resultant merged dataset consists of 604233 observations of 22 variables.

This dataset represents the weather conditions, energy consumption totals (for electricity, fuel and gas) and the characteristics of the houses present in the housing dataset which are in the city Greenville. The merged dataset is saved as a CSV file to be used in the future in order to create a checkpoint and help eliminate redundancies and repetitive workflow and optimize memory utilization.

PREDICTIONS USING DIFFERENT MODELS

We will employ predictive models to forecast the primary factors influencing energy consumption for the upcoming year, using input data. In the upcoming sections, we will explore various categorization models, evaluating their accuracy and other characteristics. The models employed will be supervised learning models with discrete, class-based outcomes. Our objective is to predict variations in the total energy consumption, determined by the summation of individual parameters in the dataset. We will train our model using independent parameters extracted from merged datasets, with the remaining attributes serving as predictors for the model.

1. **Support Vector Machines (SVM**):

   Support Vector Machines (SVM) are like smart tools in machine learning that help us figure out how to classify things or make predictions based on data. Imagine each piece of information as a point in space, and SVMs draw a line or plane that best separates different groups or classes. Below is the output from our SVM model.

```
ENERGY
mse 12.47885
rmse 3.532541
sd 15.53745
range 0.941 98.256
summary 0.941 8.2735 17.46316 20.7716 29.0515 98.256

ELECTRIC
mse 2.252647
rmse 1.500882
sd 8.160958
range 0.08 50.13
summary 0.08 3.5135 7.2105 9.77885 13.7475 50.13

NON ELECTRIC
mse NaN
rmse NaN
sd NA
range Nasummary 0 NULL NULL
```

2. XGBOOST Model:

XGBoost builds a predictive model by combining the predictions of multiple individual models, often decision trees, in an iterative manner. The algorithm works by sequentially adding weak learners to the ensemble, with each new learner focusing on correcting the errors made by the existing ones.

```
# Compare predictions to true values
library(Metrics)
rmse <- rmse(test$out_total_energy, preds)
print(paste("RMSE on test set: ", rmse))
```

```
Warning: package 'Metrics' was built under R version 4.3.3[1] "RMSE on test set:  15.0414812283659"
```

```
```{r}
summary(model)
```
```

```
                Length Class              Mode
handle               1 xgb.Booster.handle externalptr
raw             288415 -none-             raw
niter                1 -none-             numeric
evaluation_log       3 data.table         list
call                 6 -none-             call
params               8 -none-             list
callbacks            2 -none-             list
feature_names        1 -none-             character
nfeatures            1 -none-             numeric
```

3. Linear Regression Model:

The merged dataset has been partitioned into training and testing datasets following a random distribution where 80% is used for training and the remaining 20% for testing. We then develop a linear regression model with the aid of the standard linear modeling function. This model aims to predict the total energy consumption using a range of independent variables that encompass building features, weather conditions, and temporal factors. We gather a comprehensive summary from this model that includes the regression coefficients along with other statistical details. Using the testing data, we proceed to predict outcomes and calculate the Mean Squared Error (MSE) to assess how well our model performs. It's critical to analyze the model's coefficients to understand the influence each predictor has on our target variable. Finally, we use our model to foresee the effects on energy consumption when the input data is altered to reflect a scenario where the dry bulb

temperature is raised by 5 degrees Celsius and other weather variables changed proportionally.

```
Call:
lm(formula = out_total_energy ~ time_quarter + in.occupants +
    in.vintage + in.usage_level + in.hvac_heating_type + in.hvac_cooling_type +
    out_hvac_electricity + out_hvac_non_electric + `Dry Bulb Temperature [°C]` +
    `Relative Humidity [%]` + `Wind Speed [m/s]` + `Global Horizontal Radiation [W/m2]` +
    `Direct Normal Radiation [W/m2]` + `Diffuse Horizontal Radiation [W/m2]`,
    data = train1)

Residuals:
    Min      1Q  Median      3Q     Max
-34.933  -3.758  -0.869   2.333  99.736

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -2.317e+02  9.946e+00 -23.291  < 2e-16 ***
time_quarterMorning               -1.598e+00  4.007e-01  -3.989 6.67e-05 ***
time_quarterNight                 -5.516e+00  1.748e-01 -31.555  < 2e-16 ***
time_quarterNoon                  -1.852e+00  3.275e-01  -5.655 1.59e-08 ***
in.occupants                      -1.172e+00  4.914e-02 -23.847  < 2e-16 ***
in.vintage                         1.267e-01  5.064e-03  25.017  < 2e-16 ***
in.usage_levelLow                 -6.226e+00  1.695e-01 -36.727  < 2e-16 ***
in.usage_levelMedium              -7.505e+00  1.642e-01 -45.702  < 2e-16 ***
in.hvac_heating_typeDucted Heating 3.961e+00  1.798e-01  22.034  < 2e-16 ***
in.hvac_heating_typeNon-Ducted Heating 2.011e+00 3.039e-01  6.620 3.72e-11 ***
in.hvac_cooling_typeHeat Pump             NA         NA      NA       NA
in.hvac_cooling_typeNone          -3.699e-01  3.883e-01  -0.953   0.3408
in.hvac_cooling_typeRoom AC        1.721e+01  3.376e-01  50.988  < 2e-16 ***
out_hvac_electricity               1.320e+00  6.298e-03 209.622  < 2e-16 ***
out_hvac_non_electric              1.918e+00  6.204e-02  30.924  < 2e-16 ***
`Dry Bulb Temperature [°C]`       -1.332e-02  8.880e-03  -1.500   0.1338
`Relative Humidity [%]`           -3.542e-02  5.150e-03  -6.878 6.32e-12 ***
`Wind Speed [m/s]`                 8.963e-02  3.782e-02   2.370   0.0178 *
`Global Horizontal Radiation [W/m2]` 1.475e-02 1.493e-03   9.876  < 2e-16 ***
`Direct Normal Radiation [W/m2]`  -1.343e-02  1.100e-03 -12.218  < 2e-16 ***
`Diffuse Horizontal Radiation [W/m2]` -2.095e-02 2.483e-03  -8.437  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.019 on 15058 degrees of freedom
Multiple R-squared:  0.8262,    Adjusted R-squared:  0.826
F-statistic:  3767 on 19 and 15058 DF,  p-value: < 2.2e-16
```

```{r}
pred1 <- predict(lm_model_Energy,test1)
df1 <- data.frame(Predicted = pred1, Actual=test1$out_total_energy)
df1$PercentageError <- (df1$Actual-df1$Predicted)/df1$Actual*100
rmse <- sqrt(mean((df1$Actual-df1$Predicted)^2))
# rmse
# mean(df1$Actual)
# mean(df1$Predicted)
data.frame(RMSE = rmse, Actual = mean(df1$Actual), Predicted = mean(df1$Predicted))
```

Description: df [1 × 3]

RMSE <dbl>	Actual <dbl>	Predicted <dbl>
6.909554	20.27626	20.37472

The model appears to have a good fit to the data, as evidenced by the relatively high R^2.The low RSE suggests that the residuals are small, indicating that the model's predictions are close to the actual values.

RESULTS AND MODEL EVALUATION

After evaluating multiple models for temperature forecasting, it has been determined that linear regression provides the best fit for the given dataset. The decision is based on comprehensive testing and comparison of various regression models, including Support Vector Machines (SVM), XGBoost, and others. The linear regression model consistently demonstrated superior performance in terms of accuracy and predictive power, making it the preferred choice for temperature forecasting in this context.

Linear Regression (LR) achieved the highest accuracy (82.62%) among other models, indicating a strong linear relationship between predictors and the target variable. LR's assumption of linearity aligns well with the underlying relationships present in the dataset. With a Root Mean Squared Error (RMSE) of 7.15601, the model appears to be reasonably accurate in predicting average values that are observed. In contrast, SVM, known for its effectiveness in capturing non-linear patterns, yielded an (RMSE) of 3.53. This discrepancy may be attributed to challenges in kernel selection or an overarching linearity in the dataset. The XGBoost Model, an ensemble method, demonstrated an (RMSE) of 15.04, suggesting that its ensemble nature and sensitivity to hyperparameters may not have provided a substantial advantage in this scenario.

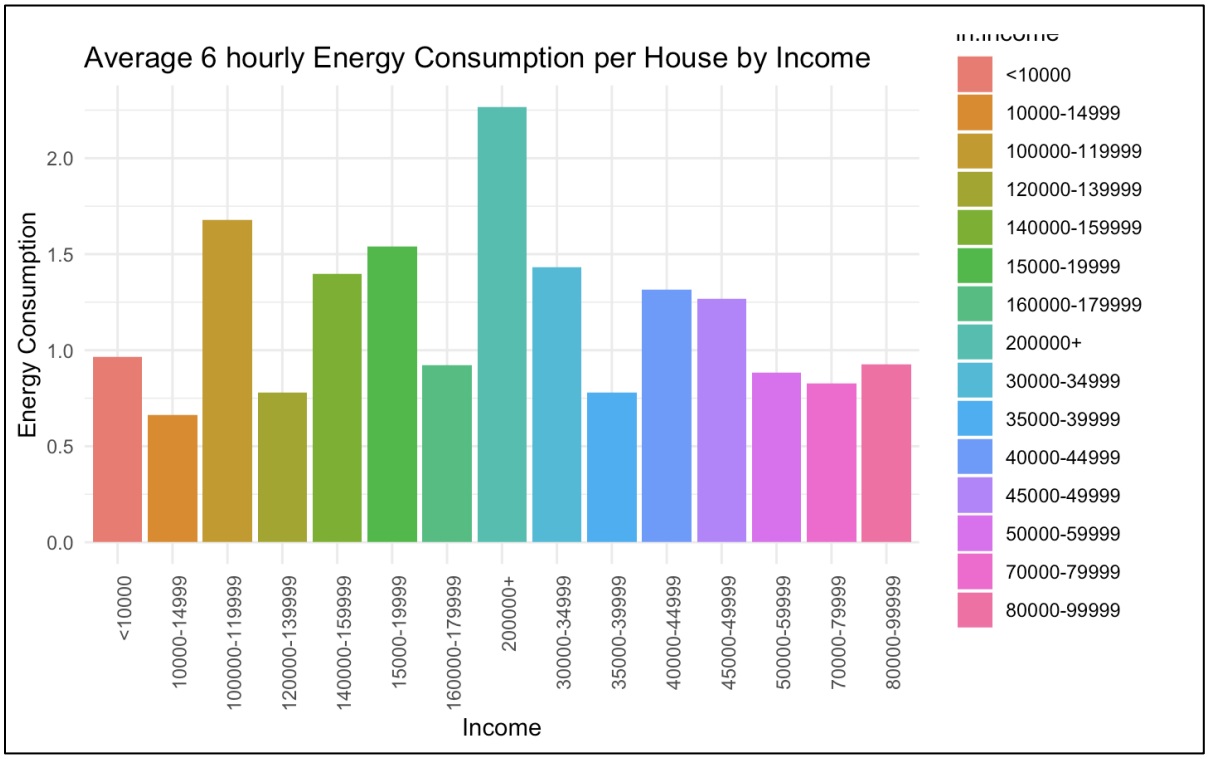
The decreased accuracy of the XGBoost and Support Vector Machines (SVM) models may indicate that the features of the dataset are not ideal for these non-linear models. Although they could perform better with changes to XGBoost parameters or SVM kernel types, Linear Regression is a good option because the data is linear. Think about doing additional research and fine-tuning considering the particular needs and features of your dataset. Because linear regression (LR) presupposes a straight-line relationship between the components under consideration and the anticipated outcome, which is consistent with the behavior of our data, LR works well for our dataset. The LR model showed good performance. This excellent accuracy suggests that the innate linear trends in our data were well-captured by LR.

CONCLUSION

1. ***Energy Consumption VS Income:***

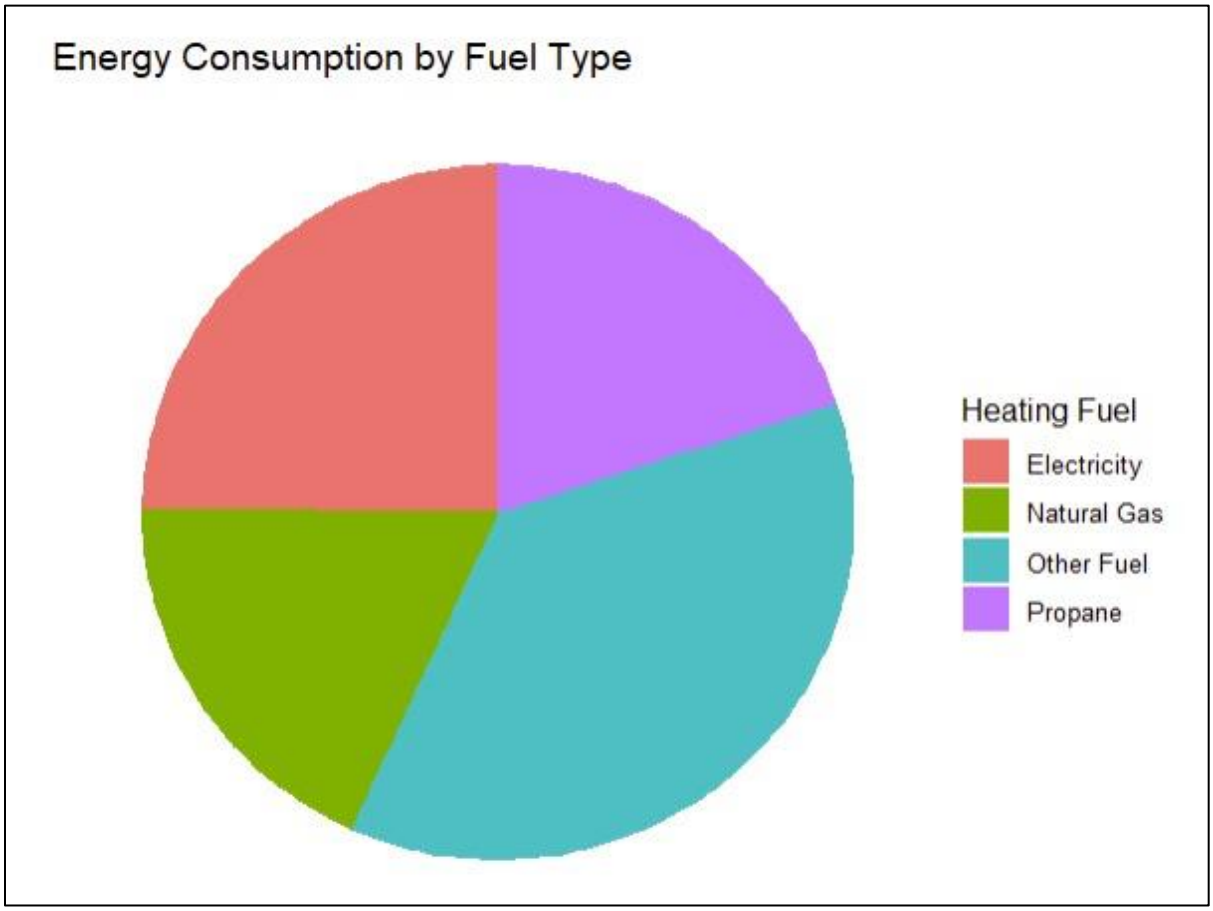
It can be vividly seen that the families having income <10000 are financially weaker and hence their consumption of energy is weaker whereas, the families having income>50000 are the wealthier families and hence they might have installed more sufficient appliances and hence again their energy consumption at a moderate level.

But the families having their income between 15000-34999 can be considered as moderate-income families wherein their electric energy demand is high and even lack the knowledge of more sustainable energy consumption appliances and hence eventually their energy consumption is at the peak.



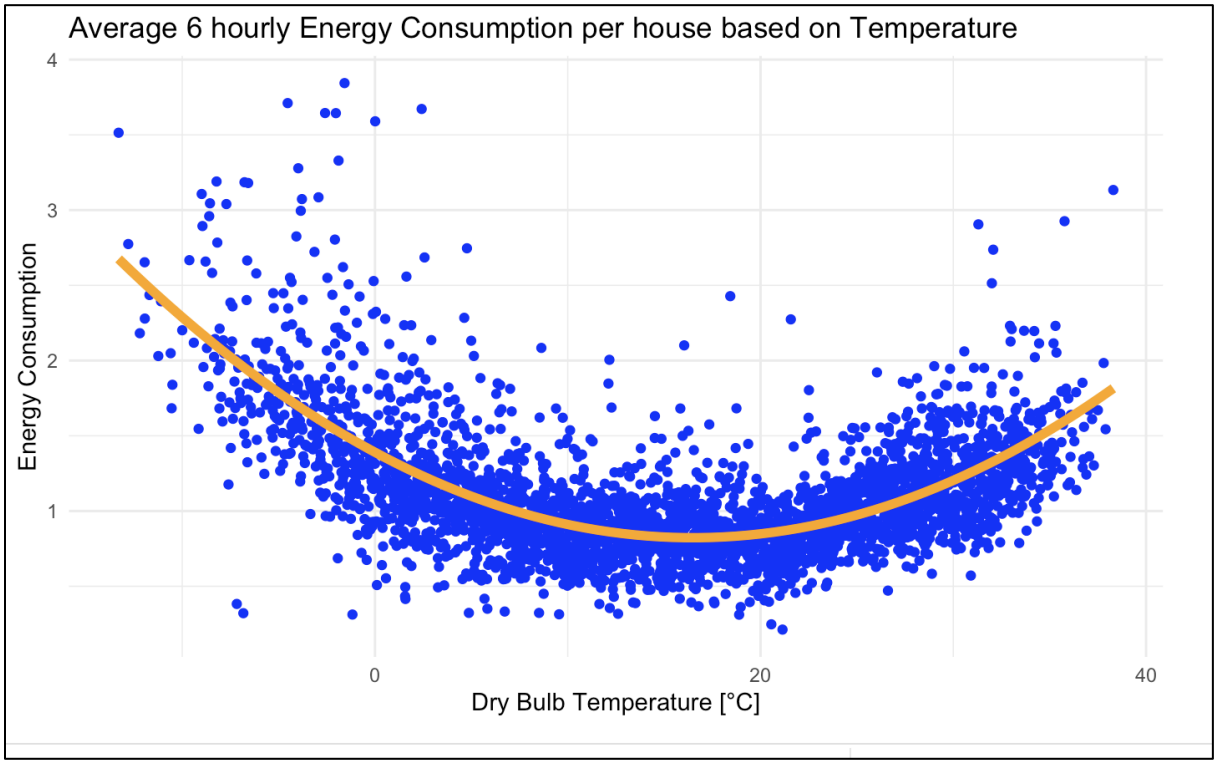
2. ***Energy* consumption vs Fuel Type:**

We have observed that energy consumption is vastly attributed to Other Fuel type which is not environmentally sustainable.



3. ***Energy consumption vs Time:***

It is evident from the image that the consumption of energy increases to an extreme demand when the temperature of air is less than -10 degree celsius. (This be stated since during the peak winter when it starts freeing, heaters are turned on and thereby increasing energy consumption). During moderate temperatures, the energy consumption seems to be moderate. Later moving further, since the temperature starts increasing thereby making days more and more hot the energy consumption starts shooting simultaneously.



SHINY APPLICATION

An interactive web application framework for R called Shiny app enables users to create interactive web apps right out of R. Shiny apps are widely used in businesses that demand dynamic and interactive reporting on datasets, and they are especially useful for data analysis and visualization. dashboard includes multiple tabs, indicating various types of visualizations or analyses available. These might include "Temperature vs. Energy Output," "Electricity Usage vs. Wind Speed," "Energy Output by Cooling Type," and "Energy Output by Time of Day," allowing users to explore different aspects of the HVAC data. The main functionalities of the app include the dynamic presentation of a Data Table showcasing weather information specific to the chosen, a visually informative plot illustrating the correlation between different variables. In essence, this Shiny app provides a sophisticated yet accessible tool for users to delve into and interpret weather data, serving as a valuable resource for meteorological analysis.

We have published our final project ShinyApp and it can be found on the following link: https://prbarot.shinyapps.io/eSC_finProj_Grp9/

R CODE:

```
library(shiny)
library(tidyverse)
library(readr)
library(ggplot2)

# Load the data (make sure to adjust the path to where your file is located)
summarizedData <- read_csv("/Users/kushal/Documents/IST
687/project/Final_App/summarizedData.csv")

# Define the UI
ui <- fluidPage(
  titlePanel("HVAC Data Visualizations"),
  tabsetPanel(
    tabPanel("Temperature vs. Energy Output",
             sidebarLayout(
               sidebarPanel(
                 sliderInput("tempRange", "Temperature Range:",
                             min = min(summarizedData$`Dry Bulb Temperature [¬∞C]`, na.rm = TRUE),
                             max = max(summarizedData$`Dry Bulb Temperature [¬∞C]`, na.rm = TRUE),
                             value = c(min(summarizedData$`Dry Bulb Temperature [¬∞C]`, na.rm =
TRUE),
                                       max(summarizedData$`Dry Bulb Temperature [¬∞C]`, na.rm = TRUE))
                 )
```

```
            ),
            mainPanel(
              plotOutput("tempPlot")
            ))
    ),
    tabPanel("Electricity Usage vs. Wind Speed",
             sidebarLayout(
              sidebarPanel(
                sliderInput("windRange", "Wind Speed Range:",
                            min = min(summarizedData$`Wind Speed [m/s]`, na.rm = TRUE),
                            max = max(summarizedData$`Wind Speed [m/s]`, na.rm = TRUE),
                            value = c(min(summarizedData$`Wind Speed [m/s]`, na.rm = TRUE),
                                      max(summarizedData$`Wind Speed [m/s]`, na.rm = TRUE))
                )
              ),
              mainPanel(
                plotOutput("windPlot")
              )
             )
    ),
    tabPanel("Energy Output by Cooling Type",
             sidebarLayout(
              sidebarPanel(),
              mainPanel(
                plotOutput("coolingTypePlot")
              )
             )
    ),
    tabPanel("Energy Output by Time of Day",
             sidebarLayout(
              sidebarPanel(),
              mainPanel(
                plotOutput("timeDayPlot")
              ))
    ))
)
# Define server logic
server <- function(input, output) {
  output$tempPlot <- renderPlot({
    summarizedData %>%
      filter(`Dry Bulb Temperature [¬∞C]` >= input$tempRange[1], `Dry Bulb Temperature
[¬∞C]` <= input$tempRange[2]) %>%
      ggplot(aes(x = `Dry Bulb Temperature [¬∞C]`, y = out_total_energy)) +
      geom_point(alpha = 0.5) +
      geom_smooth(method = "nls", formula = y ~ a * x^2 + b * x + c, method.args = list(start =
list(a = 1, b = 1, c = 1)), se = FALSE, size = 2, color = "orange") +
```

```
    labs(title = "Relationship between Temperature and Total Energy Output", x = "Temperature
(¬∞C)", y = "Energy Output (kWh)") +
    theme_minimal()
  })

  output$windPlot <- renderPlot({
   summarizedData %>%
    filter(`Wind Speed [m/s]` >= input$windRange[1], `Wind Speed [m/s]` <=
input$windRange[2]) %>%
    ggplot(aes(x = `Wind Speed [m/s]`, y = out_hvac_electricity)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "nls", formula = y ~ a * x^2 + b * x + c, method.args = list(start =
list(a = 1, b = 1, c = 1)), se = FALSE, size = 2, color = "red") +
    labs(title = "Relationship between Wind Speed and HVAC Electricity Usage", x = "Wind
Speed (m/s)", y = "HVAC Electricity Usage (kWh)") +
    theme_minimal()
  })

  output$coolingTypePlot <- renderPlot({
   summarizedData %>%
    group_by(in.hvac_cooling_type) %>%
    summarise(total_energy = sum(out_total_energy, na.rm = TRUE)) %>%
    ggplot(aes(x = in.hvac_cooling_type, y = total_energy, fill = in.hvac_cooling_type)) +
    geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set3") +
    labs(title = "Total Energy Output by HVAC Cooling Type", x = "Cooling Type", y = "Total
Energy Output (kWh)") +
    theme_minimal()
  })

  output$timeDayPlot <- renderPlot({
   summarizedData %>%
    group_by(time_quarter) %>%
    summarise(total_energy = sum(out_total_energy, na.rm = TRUE)) %>%
    ggplot(aes(x = time_quarter, y = total_energy, fill = time_quarter)) +
    geom_bar(stat = "identity") +
    scale_fill_manual(values = c("Morning" = "#FFD700", "Afternoon" = "#FF8C00",
"Evening" = "#1E90FF", "Night" = "#483D8B")) +
    labs(title = "Total Energy Output by Time of Day", x = "Time of Day", y = "Total Energy
Output (kWh)") +
    theme_minimal()
  })
}

# Run the application
shinyApp(ui = ui, server = server)
```

RECOMMENDATIONS:

1. **Encourage the use of solar energy solutions:** Encouraging the use of renewable energy sources is the goal of solar panel promotion for owners of both residential and commercial properties. This can lower carbon footprints, lessen dependency on non-renewable energy sources, etc.

2. **Educate people on the advantages of utilizing energy-efficient equipment.** Energy-efficient appliances use less electricity than inefficient appliances to do the same activities, which lowers energy costs and lessens the impact on the environment.

3. **Implement a tiered pricing structure that increases rates above the median energy consumption threshold for middle-income households:** This recommendation is about using a progressive pricing model to encourage conservation of energy. By increasing the cost of electricity as consumption increases, particularly beyond a set threshold, households are incentivized to reduce their usage.

4. **Even Odd Rule**: Introducing an Even Odd Rule for energy consumption patterns ensures a balanced distribution, preventing excessive demand during peak periods.

5. **Targeted Approach**: A targeted strategy to reduce energy consumption and mitigate the risk of blackouts can be achieved by focusing on the highest energy consumption month in July.

6. **Environmental Sustainability**: The eSC is aligned with environmental sustainability objectives by proactively managing energy demand, thereby making it a responsible provider of electricity.

7. **Holistic Approach**: Together, the proposed actions will form a comprehensive approach that addresses urgent issues and contributes to a more secure and Eco sensitive future.

ACKNOWLEDGEMENT

We are extremely grateful to our professors Prof. Akit Kumar and Prof. Christopher Dunham for the support and guidance throughout the project and this semester and all the teammates who have contributed to this project. Thanks to Syracuse University for the tools and facilities provided to complete this project.

We would like to extend our gratitude to all the external resources we used during this project.