---
title: "Final-Project"
author: ""
date: "2023-11-28"
output:
  word_document: default
  html_document: default
  pdf_document: default

Research Questions

Does type of insulation is homes have a major influence on the amount of energy being used?
Does type of AC cooling have an effect on the energy being used?
Does income have an effect on the energy being used?

Data Preparation Phase

```{r}
library(tidyverse)
library(arrow)
library(readr)
library(arrow)
library(dplyr)

# Read the static house information
static_info <- arrow::read_parquet("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet")

# Extract unique building IDs from the static information
building_ids <- unique(static_info$bldg_id)

# Initialize an empty list to store filtered energy data for each building ID
filtered_energy_data_list <- list()

# Filter specific datasets within energy_data (adjust conditions as needed)
selected_variables <- c("out.electricity.hot_water.energy_consumption",
"out.electricity.lighting_exterior.energy_consumption",
"out.electricity.plug_loads.energy_consumption",
"out.electricity.refrigerator.energy_consumption",
"out.fuel_pil.hot_water.energy_consumption",
"out.electricity.ceiling_fan.energy_consumption",
"out.electricity.clothes_dryer.energy_consumption",
"out.electricity.clothes_washer.energy_consumption",
"out.electricity.colling_fans_pumps.energy_consumption",
```

```
"out.electricity.freezer.energy_consumption",
"out.electricity.heating_fans_pumps.energy_consumption",
"out.electricity.heating.energy_coonsumption", "time")

# Fetch and filter specific energy data for each building ID
for (id in building_ids) {
  energy_url <- paste0("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-
houseData/", id ,".parquet")

  # Read energy data for the current building ID
  energy_data <- arrow::read_parquet(energy_url)

  selected_energy_data <- energy_data %>%
    select(starts_with(selected_variables))

  # Store filtered data in the list
  filtered_energy_data_list[[id]] <- selected_energy_data

  # Process or store the filtered data as needed
  print(paste("Processed data for Building ID:", id))
}

energy <- do.call(rbind, filtered_energy_data_list)

```

```{r}
# This is bringing weather based on counties.
counties <- unique(static_info$in.county)

# Fetch weather data for each county
combined_weather_data <- data.frame()  # Initialize an empty dataframe

for (county in counties) {
  weather_url <- paste0("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-
data/weather/2023-weather-data/", county, ".csv")
  weather_data <- readr::read_csv(weather_url)
  weather_data$county <- county
  # Store weather data for each county
  combined_weather_data <- bind_rows(combined_weather_data, weather_data)

  # Process or store the weather data as needed
  print(paste("Processed weather data for", county))
}
```

```

```{r}
# This is adding a filter to just focus on the month of july in both data sets before merging the two.
library(dplyr)
july_weather_data <- combined_weather_data %>%
  filter(format(date_time, "%m") == "07")

# For energy data filtering
july_energy_data <- energy %>%
  filter(format(time, "%m") == "07")
```

```{r}
# This is renaming the date_time for the weather data so that is able to merge better
# with the energy data.
july_weather_data <- july_weather_data %>% rename(time = date_time)
```

```{r}
# This is merging the energy dataset with the weather dataset.
merged_data <- merge(july_weather_data, july_energy_data, by = "time", all = TRUE)
```

```{r}
# This is going through and cleaning the data by getting rid of missing values and
# duplicates. It also got rid of variables that only present data that was zero.
cleaned_data <- na.omit(merged_data)
static_info <- na.omit(static_info)
cleaned_data <- unique(merged_data)
static_info<- unique(static_info)
cleaned_data <- na.omit(cleaned_data)
```

Exploratory Analysis Phase

```{r}
summary(cleaned_data)
```

```{r}
library(corrplot)
# This is focused on Cooling Systems
```
```

```
selected_columns <- c("Dry Bulb Temperature [°C]",
"out.electricity.cooling.energy_consumption",
"out.electricity.cooling_fans_pumps.energy_consumption",
"out.electricity.ceiling_fan.energy_consumption", "out.electricity.pv.energy_consumption")
selected_data <- cleaned_data[, selected_columns]
correlation_matrix <- cor(selected_data)
print(correlation_matrix)

# This is focused on Washer and Dryers.
selected_columns2 <- c("Dry Bulb Temperature [°C]",
"out.electricity.clothes_dryer.energy_consumption",
"out.electricity.clothes_washer.energy_consumption")
selected_data2 <- cleaned_data[, selected_columns2]
correlation_matrix2 <- cor(selected_data2)
print(correlation_matrix2)

selected_columns3 <- c("Dry Bulb Temperature [°C]",
"out.electricity.heating_fans_pumps.energy_consumption",
"out.electricity.heating_hp_bkup.energy_consumption","out.electricity.heating.energy_consu
mption", "out.electricity.hot_water.energy_consumption",
"out.fuel_oil.heating_hp_bkup.energy_consumption",
"out.fuel_oil.heating.energy_consumption", "out.fuel_oil.hot_water.energy_consumption" )
selected_data3 <- cleaned_data[, selected_columns3]
correlation_matrix3 <- cor(selected_data3)
print(correlation_matrix3)

selected_columns4 <- c("Dry Bulb Temperature [°C]",
"out.electricity.hot_tub_heater.energy_consumption",
"out.electricity.hot_tub_pump.energy_consumption",
"out.electricity.pool_heater.energy_consumption",
"out.electricity.pool_pump.energy_consumption","out.electricity.well_pump.energy_consumpt
ion")
selected_data4 <- cleaned_data[, selected_columns4]
correlation_matrix4 <- cor(selected_data4)
print(correlation_matrix4)

```

```{r}
library(ggplot2)
ggplot(cleaned_data, aes(x=`Dry Bulb Temperature [°C]` ,
y=out.electricity.cooling.energy_consumption)) + geom_point() + geom_smooth(method =
"lm", se = FALSE, color="blue")+ labs (x="Insulation in Ceiling", y="Cooling Energy
Consumption", title = "Cooling Energy Consumption by Insulation in Ceiling")
```

```
```