# George Mason University

# CS 504

# Spring 2021

**Predicting Power Utilization Based on Historical Smart Meter Energy Consumption Data**

**Draft 3 Energy Project Report**

**Team Carbon**

**Team Members:**

**Jhony Islam**

**Spencer Marlen-Starr**

**Radha Kanuri**

**Cassidy Laskodi**

**Stavros Kalamatianos**

Table of Contents

[Abstract 3](#_Toc69651697)

[1 Introduction 4](#_Toc69651698)

[1.1 Background and Rationale 4](#_Toc69651699)

[1.2 Research 8](#_Toc69651700)

[1.3 Project Objectives 9](#_Toc69651701)

[1.4 Problem Space 9](#_Toc69651702)

[1.5 Primary User Story (-ies): 10](#_Toc69651703)

[1.6 Solution Space 10](#_Toc69651704)

[1.7 Product Vision - Sample scenarios (why would someone want to use this) 11](#_Toc69651705)

[Scenario #1 11](#_Toc69651706)

[Scenario #2 11](#_Toc69651708)

[1.8 Definition of Terms: 11](#_Toc69651709)

[2 Data Acquisition 13](#_Toc69651710)

[2.1 Overview: 13](#_Toc69651711)

[2.2 Field Descriptions: 13](#_Toc69651712)

[2.3 Data Context: 18](#_Toc69651713)

[2.4 Data Conditioning 18](#_Toc69651714)

[2.5 Data Quality Assessment: 20](#_Toc69651715)

[2.6 Other Data Sources 21](#_Toc69651716)

[3 Analytics and Algorithms 21](#_Toc69651717)

[3.1 Time Series Analysis 21](#_Toc69651718)

[3.2 Multiple Linear Regression and Pearson’s Correlation Coefficient Test 23](#_Toc69651719)

[3.3 Analysis of the Weather Time Series and Correlation of Daily Weather Statistics to the Daily Average Energy Consumption 25](#_Toc69651720)

[3.4 Analyzing the Treatment Effects from the Experiment of Randomly Introducing Dynamic Time of Use Energy Pricing 26](#_Toc69651721)

[4 Visualization 27](#_Toc69651722)

[5 Findings 27](#_Toc69651723)

[6 Summary 27](#_Toc69651724)

[7 Future Work 27](#_Toc69651725)

# Abstract

[1-2 paragraph that summarizes your project: a) statement of the problem; b) research methods uses; c) results and findings; and d) conclusions and recommendations.]

# Introduction

## Background and Rationale

Energy consumption worldwide has been increasing at neck breaking speeds during the last 60 years. The overwhelming majority of the fuel chosen by heavily industrialized nations to produce electricity, in the geographical areas of North America, Europe and Japan has been relatively cost effective petrochemicals, the extraction technology of which humanity has mastered over decades of exploration.

The painstaking study of distillation of crude oil to its finest components facilitated the production of fuels specialized to power all types and sizes of motor/air transportation machinery, from vespas to unfathomably massive military aircrafts, such as the Ukrainian manufactured Antonov An-225 Mriya.

The aggressive entrance of a third of the world’s population, summed up in China and India, to the race on energy consumption, challenged the energy resource abundance utopia the world has been living on, and with the expected future depletion of crude oil resources in the near future, poses an immediate threat to our daily lives, adds unexpected items at the top of Governmets’ priority agendas, and sends alarming requests to the scientific community for new types of energy and supporting technology.

Unfortunately, the energy problem has been accompanied by the environmentally-insensitive behavior of a number of constituencies worldwide who via their unsubstantiated and aloof beliefs exacerbated a number of environmental problems. Their multi-year irresponsible activities have started manifesting in the form of global climate change, especially during the last couple of decades.

However, what has also been changing during the last 20 years is the outcome of sluggish, yet constant, research on renewable energy technologies, such as wind, biofuels thermal based on solar, and photovoltaics that finally show sufficient maturity to serve a long term energy supply without losing their cost competitive edge.

The result of progress on renewable technologies and their noticeable adaption shows in the following chart from Wikipedia (World Energy Consumption, 2021) as a positive annual increase of the comparative height among major energy sources. Despite Coal, Oil and Natural Gas dominating the chart [Fig. 1], renewable technologies have started contributing enough to be visible on the chart with an upward trend.

Figure 1: Global energy consumption comparison among major energy sources.

Today, more than 40% of the world’s energy consumption is in the form of electricity, and it is expected to grow to 60% by 2040. Figure 2 shows how according to the projections of the International Energy Agency [IEA] (Bizon, 2017). IEA projects that there will be a significant injection of renewables in the next 20 years to the point that they will account for about 44% of the world’s energy consumption.

Figure 2: Projections of the world’s energy consumption by the International Energy Agency.

As the process of renewable innovations gets more support by Governments, it will help the industry get out of the slow progress from the last 10 to 15 years. Also, private capital will gain more faith to invest in them, their prices will decrease, and economies of scale would progressively emerge.

The presence of power electronics based energy conversion systems is of paramount importance to enable renewable electricity to be pumped into the existing electricity distribution grid.

The three major electricity consuming sectors that require those types of electronics are industry, transportation and residential.

Transportation has made great progress electrifying motor vehicles, speed trains, smaller boats and parts of bigger vessels. That’s possible with the use of power electronics.

Transportation and Industry consumers exist as more compacted business entities, where environmentally favoritistic policies can be implemented quicker through C-level governance as opposed to the last electricity consuming sector, residential.

The residential sector consists of millions of consumers that can be governed through legislature that can be steered towards certain constituencies interests taking advantage of the public's political beliefs. Depending upon that, it may be more difficult, or not, to get the public’s attention and assistance to the greater cause.

That cause is adapting responsible energy consumption habits focusing on energy reduction without jeopardizing quality of life, while promoting the use of relatively inexpensive energy efficient products and services. The overall goal is to decrease greenhouse gases and decelerate, stop or reverse climate change.

The United Kingdom [UK], as a member of the industrialized nations and the G8 is no different. Its investment on energy is dominated by the part devoted to electricity. As Figure 3 (Department for Business, Energy and Industrial Strategy, 2018) shows, it was 55% of the total in 2019.

Figure 3: UK’s Energy Industry Investments By Energy Resource

The UK is also in favor of renewables.Figure 4 displays the UK's primary fuel production constant increase in biofuels during the last 20 years.

Figure 4: UK’s Fuel Production by energy source in tons of oil energy equivalents for Selected Years between 1990 and 2019

The impetus for UK’s elevated environmentally friendliness is the entire European Union commitment to it.

According to the European Environmental Agency, the region has been experiencing impacts of climate change such as rising sea levels and extreme weather. This is caused by greenhouse gases (such as carbon dioxide, methane and nitrous oxide) being emitted into the atmosphere. In Europe, energy consumption is responsible for about 78% of the greenhouse gases emitted in the European Economic Area (EEA) (Energy and Climate Change, 2017).

The Paris Agreement of 2015 is a legally binding international treaty through the United Nations (UN) which commits the signatories to reducing their greenhouse gas emissions, according to the United Nations Framework Convention on Climate Change (UNFCCC) (The Paris Agreement, 2015). Nearly 200 countries signed it in 2015, and since then there has been a growing desire for low-carbon solutions in the energy sectors of the signatories and some (New Zealand for example) have even adopted carbon-neutrality as a medium term national goal.

In the UK, residential energy consumption is responsible for 15% of the country's greenhouse gas emissions (ECIU).This amounts to the average UK household consuming 3,731 KiloWatt per hour (KWh) per year, and that number does not include heating (ovoenergy) (Topping, 2021).

The UK's commitment to energy savings with further focus on decreasing emissions that contributes greatly to a global campaign to stop and/or reverse climate change is shown in Table 1. This data is from a 2020 UK government sponsored publication titled “UK Energy in Brief” (UK Energy In Brief 2020, 2014). Table 1 also displays how UK’s bioenergy topped the list of energy increase by source type, after tripling during the last 10 years even (barely) surpassing the increase on nuclear energy use.

UK’s energy consumption changes by energy supply type

Table 1: Percentage increase in UK’s energy consumption by type between 2010 and 2019.

UK’s progress adapting cleaner energy is more evident from another statistic showing that energy from renewables, such as biofuels, solar and wind were also increased, leaving nuclear energy supply behind in the list of energy resources.

Figure 5 stresses the magnitude of the change and shows the energy and the carbon ratio falling dramatically during the last 20 years. Both ratios are calculated per unit of Gross Domestic Product (GDP). Despite the increase of GDP, both energy consumption and carbon emissions decreased making the ratio drop more dramatically as it can be quickly realized from Table 2.

Figure 5: Table 2 - Energy and Carbon ratios during the period of 1990-2019 in graphic display and explicit numerical format.

One more evidence of the nation’s incredible job is Table 3 in Figure 6 where energy production generated gas emissions have also dropped in hundreds of millions of tons since 1990.

Fig. 6 - Table 3: Greenhouse gas emissions in carbon monoxide mass equivalents expressed in millions of tons.

But what drives that uninterrupted pace of progress?

Simple, UK parliamentary democracy committed to reducing greenhouse gasses expedited their course of action in June 2019 when they passed legislation to become a 0% emitter relative to 1990 levels by 2050. That means the UK will be able to remove from the low atmosphere as many tons of carbon emissions as it produces for energy consumption and other industrial activity, to become a net zero mean emitter (UK net zero target, 2020).

The mandates of the UK government became bullet points in the agenda of priorities of the country’s energy suppliers who strive to meet the government’s expectations with programs like the smart meter installation. Such programs will guide their efforts to collect consumer data. This analysis will help develop ways to educate the public on how to save energy and the emergence of products and services to serve that cause. Smart meters, unlike traditional analogue meters, are more interactive and multifunctional. They are easier for customers to read and share information over wireless frequency networks.

London, as the UK's capital and Europe’s widest, has been in a “greener is cleaner” state of mind for a while.

The LowCarbon London page of UK Power Networks (Innovation at UK Power Networks) advertises a few dozen innovation projects developed to decrease carbon emissions. Some of those that stuck out are:

* “Above and Beyond” uses drones for visual power line and asset inspection, instead of helicopters.
* “Arc-Aid” features quick-response sensors for overhead power line fault detection to minimize repair crews driving around trying to locate the electrical network culprit.
* “Edge FCLi'' collaborative project uses machine learning technology, called “PowerFactory” to ensure the substation [electricity distribution station] equipment prone to fault at, or after, 95% of their rating limits, do not reach that level of usage. Preventing that from happening minimizes the fault current coming from distributed generators connected to the greater network, thus supporting the UK's Net Zero carbon emissions target.

In this study, a representative sample of a few thousands of households spread out in the urban London area is split into two different groups.

1. The first control group was subject to: “dynamic time of use” pricing with the prior knowledge of the tiered pricing structure per KWh as total consumption increased during the day.
2. The second control group was subject to receiving price signals when pricing was getting adjusted during the day, so they could regulate their house’s energy use to save money and contribute their small quota to a lower emissions London.

The experimental group of “non-time of use” customers remained on a flat price rate per KWh.

The goal of the study was dual.

1. The first goal was to test the types of signals the supplier could send to the public to control and manage high renewable sources supply times.
2. The second goal was to adjust prices purposefully to throttle down consumption at certain parts of the grid under stress during high utilization times.

The data sets involved carry useful information about the participants' way of living, income and spending habits indexed in an index called ACORN. It shows whether the level of the social stratum of the people who belong in the ACORN group lies above or below the national scale in the UK.

The accompanying weather data completes the set and provides two major variables that influence energy consumption in the residential sector of the UK, where more than half of the UK’s energy investment went to.

The Kaggle data set “Smart Meters in London” will be the analysis resource we’ll use for the rest of this study.

## Research

The problem we look to address is whether there are effective methods to cut down energy consumption for individuals. A main point of focus is utilizing smart meters within the capital of the UK, London. There have been additional studies revolving around the tracking data through smart meters with one to state in particular done by four authors with electrical engineering backgrounds.

In the study, “Probabilistic Peak Load Estimation in Smart Cities Using Smart Meter Data” performed by Mingyang Sun, Yi Wang, Goran Strbac, and Chongqing Kang, the authors seeked to estimate peak demand for customers within London using smart meters similarly to what our team is looking to accomplish. They identify two major challenges surrounding the project being, “Different types of properties with various future customers exhibit different consumption behaviors” and “The demand diversity among customer loads significantly increases the difficulties in estimating the group peak demand” when they are defining their problem (Sun, Wang, Strbac, Kang, 1609).

It is helpful to see the challenges they identified in their problem statement as we may face similar challenges as well. The study proposes a solution where they utilize historical smart meter data in order to predict potential consumption behavior for future customers (Sun, Wang, Strbac, Kang, 1611). They go in depth on the model they chose to best represent and predict the data by identifying the factors used, identifying the model to predict data, and describing the dataset as a whole. This study chose to focus on using probability to determine peak demand for future customers which will aid in our study as we can see the factors identified in consumption as well as the data description from a previous study. By giving us a starting point in determining which factors are most effective in understanding the data, we can take a more effective approach to our problem. This study helps set up an excellent starting point for beginning our study.

## Project Objectives

This project aims to understand important aspects of energy consumption and more specifically how stakeholders are affected due to weather phenomena and other factors. A few questions we look to answer are:

* How are individuals at the household level affected by weather phenomena?
* Are there policies in place pertaining to weather phenomena and energy consumption?
* What role do price and cost play in energy output by energy suppliers?
* Does climate change play a significant role in energy consumption patterns?

These questions identify different stakeholders and their stances on energy consumption, while also focusing on energy consumption as a whole. Through our analysis of the UK dataset, we hope to get a general understanding of how energy consumption works worldwide.

## Problem Space

In this study, we look at homes in London, England and attempt to find out how people consume energy in their homes. The data is a sample of 5,567 homes in the London area from November 2011 to February 2014. This data comes from “smart meters'' placed in people’s homes by the British Government at the European Union’s recommendation. The weather data is from a darksky Application Programming Interface (API).

The energy consumption habits of the monitored houses is correlated to a number of factors of residents' daily lives, such as using more heating energy at night during the winter months when temperatures are lower, or the reverse in the summer days when temperatures soar upwards. Both scenarios vary by the house size, its location and its surroundings (surrounding trees will block a morning sun that would warm the house naturally, in lieu of heating), they vary by whether they plan outings during holidays or not, by how many appliances they run in their house and at what time of the day, their spending habits and plainly speaking whether they want to or even care to be more vigilant when it comes to energy savings.

Moreover, a greater problem we will try to address in this study is how to exploit the multi-variable demographic dependency of energy utilization and local weather data, to analyze patterns of energy consumption at the household level and make recommendations for how to potentially expand it to energy forecasting at the level of a greater metropolitan area that tends to be the energy consumption hog of a geographical region.

All the above wonders will stimulate our minds while examining the data more closely in the next few weeks. Exploratory analysis and examination of data relationships will help us extract meaningful information from the data to answer the above questions.

## Primary User Story (-ies):

* As a head of a household, I want to understand how my energy usage rate works, so that I can be more cost effective.
  + Understand what usage rates are accessible to households
  + Determine what households can change to affect usage rates
  + Determine if changes can make households more cost effective
  + Identifying if economic status plays a significant role and if so its effect
* As an energy supplier, I want to determine which factors impact energy usage, so that I can improve my revenue.
  + Which factors can energy suppliers manipulate (e.g. household income, long term investments, family size, occupations and others that collectively define pretty accurately the social strata of the city of London where the energy consumption data was collected from.)
  + How do I manipulate these factors to improve revenue?
  + How do I balance supply and demand to determine pricing?
  + What services are offered to promote reduction in costs, i.e. savings for energy usage?
* As a government official, I want to identify the different factors that affect energy usage, so that I can create relevant policies which will reduce energy usage.
  + Discover the factors that are impacted through government policies
  + Will government officials be more aggressive or passive with their policies?
  + Understand what drives the consumption of energy and it will perhaps stimulate ideas for possible changes in how to decrease useless energy consumption.

## Solution Space

Considering the intense pressure of time to deliver a workable system in less than 6 weeks we will claim that our system, at least from its design, will provide warning scenarios where certain events or combinations of independent variables will trigger a phase of higher energy consumption. The basis of such a situation will be examined with predictive and possibly prescriptive statistics to the point the quality of the data supports it. The reader then will be left with a few takeaway points about what can be done to suppress energy consumption, smooth the distribution of energy consumption throughout the day (reduce the sharpness of the spike in daily energy consumption which probably occurs around 5 - 7 pm) or the reasons why it is not possible to do either.

## Product Vision - Sample scenarios (why would someone want to use this)

### Scenario #1

### One interested party would be a startup with innovative products in the works. Such a business entity would like to know what’s their market segment and the weather conditions in their geography to adjust energy saving product development by wallet capacity and scope, i.e., for a residence or for an individual. Lifestyle information of potential customers will also help marketing the product properly. Some obstacles the startup might face could be availability of resources to produce their goods as well as availability of data to market to their customer base.

### Scenario #2

Policy makers would be interested in this energy data analysis. Being informed about the current usage of energy according to demographic will help them establish rules about energy consumption. They can form a team who closely monitor different demographics such as age, gender, annual income, etc. For example, if a certain demographic area is consuming more energy than needed, policy makers can enforce restrictions to control the area. A major obstacle policy makers may face is gaining traction and support from individuals for their particular policies. Without support, it would be difficult to enact any policy created by officials.

## Definition of Terms:

**ACORN**- Acorn is a powerful consumer classification that segments the population into 62 different types, providing a detailed understanding of the consumer characteristics of people and places across the UK.

**Apparent Temperature** - The perceived temperature in degrees Fahrenheit derived from either a combination of temperature and wind (Wind Chill) or temperature and humidity (Heat Index) for the indicated hour.

**Bank Holidays** - A bank holiday is a business day during which financial institutions are closed.

**Cloud Cover** - forecasts total cloud cover. The colors are picked from what color the sky is likely to be, with Dark blue being clear. Lighter shades of blue are increasing cloudiness and white is overcast. This forecast may miss low cloud and afternoon thunderstorms. When the forecast is clear, the sky may still be hazy, if the transparency forecast is poor.

**Dew Point** -  the temperature the air needs to be cooled to (at constant pressure) in order to achieve a relative humidity (RH) of 100%.

**Energy** - Energy, in physics, the capacity for doing work. It may exist in potential, kinetic, thermal, electrical, chemical, nuclear, or other various forms.

**Icon** - This is the icon that shows up on the weather app to show the user what the weather would be on that particular day or hour

**LCLid** - LCLids are the ID numbers of Power Meters.

**Pressure** - Pressure is the ratio of force to the surface area over which it is exerted.

**Energy Tariff** - An energy tariff is the pricing scheme an energy provider uses to charge a customer for gas and electricity consumption. A fixed rate tariff defines the cost of a unit of energy for a certain amount of time, typically per hour, while prices of a variable tariff fluctuate according to the market.

**Temperature Min & Max** - Extreme observed temperatures during the last 24 hours between 7pm CST - 8amCST per USA’s National Weather Service [NWS] definition. NWS’s defintion may very well differ from UK’s Weather Service, however, this quick description of the variable provides an idea of how it’s defined.

**Temperature Min & Max Time** - denotes the time a temperature minimum or maximum gets recorded in contract to the expected time. The weather phenomenon finds its explanation in the abrupt movements of air masses that influence the weather dynamics in the region of measurement, henceforth, the  skewness in the time of manifestation.

**Wind Bearing** - the generally accepted definition of the term dictates that bearing is associated with the direction of movement of an object within the zone by an active wind pressure differential. For example, a southern wind on a lake will favor a northbound sail the most.

**Wind Speed** - The speed of air at a fixed set of coordinates, measured either as the integral of speed variation during a prespecified period of time or as an instantaneous speed measurement, usually in cases the wind speed fluctuates insignificantly for all intents and purposes of the measurement.

# Data Acquisition

## Overview:

Our data set consists of 6 main files and 112 half hour block files we finally consolidated into a single set. All raw data was provided in xls or csv format and they span from November 2011 through February 2014. The files of the set consists of the  following list of items:

Households information, such as a sample residence Smart Meter ID, the residents ACORN group, whether they belong to the control or the experimental group of participants and some energy descriptive statistics.

Daily half-hour energy consumption measurement set per household’s Smart Meter through the period of the study

An index of the ACORN lifestyle defined hierarchy that shows the number of people in the ACORN group comparing it with the national scale baselined at 100

Detailed weather hourly/daily pressure, temperature and water/air content/velocity/direction meteorological variables

## Field Descriptions:

The files in Kaggle have been provided in a way that allowed us to create the relations we needed by consolidating or transforming the raw data.

Our database contains a total of seven relations. In this section we’ll describe the attributes of those relations created for our analysis from the greater Kaggle data set.

**daily\_dataset:**

LCLid (Type: string) – Identification Number of Smart Meter installed in a household selected to participate in the study

day (Type: date) – Date a measurement was taken

energy\_count (Type: int)  – is the number of measurements collected every day. In 98.8% of the days 48 measurements were collected daily, once per half an hour.

energy\_median, energy\_mean, energy\_min, energy\_max, energy\_std, energy\_sum (Type: float)  – are the median, mean, min, max, standard deviation and sum of the daily measurements.

**information\_households:**

LCLid (Type: string) – Identification Number of Smart Meter installed in a household selected  to participate in the study

stdorToU (Type: string) – Standard tariff vs Time of Use pricing, i.e., a fixed price per KWh throughout the hours of the day against a tiered price structure that varies depending the hour of the day

ACORN (Type: string) – Describes consumer’s social best fit by occupation, salary and spending behavior.

ACORN\_grouped (Type: string) – Single word description of ACORN’s social status

File: it’s a pointer to the file that includes the smart meter readings and energy statistics of the specific smart meter

**hhblock:**

LCLid (Type: string) – Identification Number of Smart Meter installed in a household selected  to participate in the study

day (Type: date) – Date a measurement was taken

hh\_x where x = 0,1,2,...,46,47  (Type: float)  –  measurement at a half hour interval throughout the 24 hrs of the day

**weather\_daily\_darksky:**

[please refer to the Section 1.8 “Definition of Terms” for more details about some scientific attributes listed below]

|  |
| --- |
| temperatureMax (Type: double) – Maximum daily temperature |
| temperatureMaxTime (Type: datetime) – Time of maximum daily temperature |
| windBearing (Type: integer) – Wind direction during the day |
| Icon (Type: string) – The icon used on darksky to describe the day’s general weather |
| dewPoint (Type: float) – The temperature the air needs to be to have 100% humidity |
| temperatureMinTime (Type: datetime) – Time of minimum daily temperature |
| cloudCover (Type: float) – Percentage of the sky that is covered by clouds |
| windSpeed (Type: float) – The speed of the wind in knots (nautical mile per hour = 0.51 m sec-1 = 1.15 mph) |
| pressure (Type: float) – Atmospheric pressure in Millibars |
| apparentTemperatureMinTime (Type: datetime) – The time that the weather will feel the coldest |
| apparentTemperatureHigh (Type: float) – The perceived temperature tantamount to the forecasted high in degrees Celsius |
| precipType (Type: string) – The type of forecasted precipitation |
| visibility (Type: float) – The distance that an object can be clearly observed, measured in miles. |
| humidity (Type: float) – The concentration of water droplets in the air expressed as a percentage |
| apparentTemperatureHighTime (Type: datetime) – The time when the forecasted high temperature was perceived |
| apparentTemperatureLow (Type: float) – The perceived temperature tantamount to the forecasted low in degrees Celsius |
| apparentTemperatureMax (Type: float) – The highest perceived temperature in degrees Celsius |
| uvIndex (Type: float) – The prediction of exposure to the sun’s uv rays |
| time (Type: datetime) – The time this information was recorded |
| sunsetTime (Type: datetime) – The time when the sun will set |
| temperatureLow (Type: float) – The forecast of low temperature for the day in degrees Celsius |
| temperatureMin (Type: float) – The minimum temperature for the day in degrees Celsius |
| temperatureHigh (Type: float) – The forecast of high temperature for the day in degrees Celsius |
| sunriseTime (Type: datetime) – The time when the sun will rise |
| temperatureHighTime (Type: datetime) – The time when the forecasted high temperature will be observed |
| uvIndexTime (Type: datetime) – The time which the UVIndex applies |
| summary (Type: string) – A brief description of the day’s weather |
| temperatureLowTime (Type: datetime) – The time when the forecasted low temperature will be observed |
| apparentTemperatureMin (Type: float) – The lowest perceived temperature in degrees Celsius |
| apparentTemperatureMaxTime (Type: datetime) – The time when the temperature will perceived as the highest |
| apparentTemperatureLowTime (Type: datetime) – The time when the forecasted low temperature was perceived |
| moonPhase (Type: float) – The observed percentage of a full moon |

**weather\_daily\_darksky:**

[please refer to the Section 1.8 “Definition of Terms” for more details about some scientific attributes listed below]

|  |
| --- |
| visibility (Type: float) – The farthest one can identify an object, measured in miles |
| Wind bearing (Type: int) – The direction in degrees in which the wind is blowing |
| temperature (Type: float) – The temperature of the hour |
| time (Type: datetime) – The time which the data was recorded |
| dewPoint (Type: float) – The air temperature, expressed in degrees Celsius, at which the air will be saturated with water, i.e., humidity will be at 100%. |
| pressure (Type: float) – Atmospheric pressure in Millibars |
| apparentTemperature (Type: float) – The perceived temperature in degrees Celsius |
| windSpeed (Type: float) – The speed of the wind measured in knots  (nautical mile per hour = 0.51 m sec-1 = 1.15 mph) |
| precipType (Type: string) – The predicted amount of precipitation in centimeters |
| icon (Type: string) – The icon of the darksky app’n used to report the weather |
| humidity (Type: float) – The air water content expressed as a percentage |
| summary (Type: string) – A brief description of the day’s weather |

**acorn\_details:**

MAIN CATEGORIES (Type: string) – main categorization of ACORN groups characteristics, such as housing, salary, population, etc

CATEGORIES (Type: string) – subgrouping within the main categories, type of housing such as single family, townhouse or age groups within the population category

REFERENCE (Type: string) – ACORN group category details, e.g., house bedrooms or whether the ACORN participant owns [pays off or mortgages] or rents the property

ACORN - x where x = A, B, C, …., P, Q  (Type: float) – ACORN group index numeric value. In essence the index is the percentage of how many people, compared to the national baseline of 100, are in the ACORN group. An index of less than 100 means the ACORN group’s population is less than 100% of the national baseline value for the specific combination of the initial three attributes of this relation.

**acorn\_details\_transp:**

MAIN CATEGORIES (Type: string) – main categorization of ACORN groups characteristics, such as housing, salary, population, etc

CATEGORIES (Type: string) – subgrouping within the main categories, type of housing such as single family, townhouse or age groups within the population category

REFERENCE (Type: string) – ACORN group category details, e.g., house bedrooms or whether the ACORN participant owns [pays off or mortgages] or rents the property

ACORN (Type: float) – ACORN group index numeric value. In essence the index is the percentage of how many people, compared to the national baseline of 100, are in the ACORN group. An index of less than 100 means the ACORN group’s population is less than 100% of the national baseline value for the specific combination of the initial three attributes of this relation.

ACORN - x where x = A, B, C, …., P, Q  (Type: string) – letter designator of an ACORN group

**tariffs:**

TariffDateTime (Type: datetime) – it’s the date and time during the entire period of collection of the data set at half an hour intervals

Tariff (Type: string) – the pricing range, namely, Low, Medium, High for every value of TariffDateTime

**uk\_bank\_holidays:**

Bank holidays (Type: datetime) – the counterpart of the US federal holidays throughout the period of data collection.

Type (Type: string)  – short holiday description

contains the information of the participating households, including the Smart Meter ID installed on premise, the group they fall in, namely, “dynamic time of use” or “non-time of use”,  their ACORN category and its ACORN\_grouped description and a block of energy statistics associated with the Smart Meter installed.

ACORN details: carries the

## Data Context:

As the government in the UK looks to better track energy consumption, they are imploring energy suppliers to install smart meters in all homes within England, Wales, and Scotland. As of September 30th, 2020, there were 22.2 million households and small businesses with smart meters. This statistic demonstrates the UK’s success in following through with their plans of installing smart meters through all homes. Although the coronavirus (COVID-19) halted the installation of these smart meters quite a bit, the UK installed 856,000 smart meters in Q3 of 2020 alone, which was a six-fold increase from Q2 of 2020. The UK prioritizes the work of installing these smart meters even through a global pandemic.

The driving force behind wanting to install these smart meters is actually branched out all the way from the European Union, which wanted nations to find plans to tackle climate change. The smart meter was the British government’s solution to this task and allows for them to track energy consumption in a more accurate and effective manner. Through this dataset that tracked electrical consumption of smart meters from 5,567 London Households, the UK’s smart meter solution is displayed in full effect. This data also looks to analyze data through ACORNs which are geo-demographic segmentations of the UK’s population. They provide analysis on different types of people within the UK by studying social factors and population behavior. By utilizing these ACORNs, the dataset seeks to track the energy consumption of UK households in a more in-depth manner.

## Data Conditioning

Although the term Data Conditioning is based upon data routing optimization among systems and it’s concerned about the security of data in transit and at rest, it is also closely related to data wrangling techniques responsible for bringing the data closer to a more accurate state.

Data conditioning requires solid state components and L2 cache memory devices supported by fast I/O buses within the mobos of multi-way servers, in the realm of storage area networks multi-GBit communication between clusters of servers and storage arrays through multiple zones of fiber channel fabric switches and dark fiber connected WANs in great scale cloud computing networks where data movements need to move at jaw dropping speeds.

Because of the simplicity of our case in the following lines data conditioning will be intertwined with data cleansing addressing the former when the conversation converges more on that topic.

Following is a cradle-to-grave description of our data adventure from its source to the final schema it currently resides in a MySQL installation awaiting for data exploration and analysis.

At first we downloaded the data from the Kaggle site in its raw status, that was either in xls or csv format. Then we wrote a series of load statements in SQL and imported the raw data into tables (relations). The importing process demanded to specify the attribute delimiter and the line separator so we had to match the field’ format. Likewise on encoding that, thankfully, most of the time it was the expected utf-8.

To do the above we used text editors like Notepad ++ with features like all characters and encoding turned on. Then we quickly browsed the source file to ensure we had the right reading parameters specified.

The initial DDL of the hosting tables had all the attributes data typed as text. That allowed us to circumvent the problem of breaking the importing process every time a tuple’s attribute had the wrong content. After the data was successfully loaded on the table we slowly cleaned up every field we identified illegal values that didn’t match the expected data type and then we converted the attribute to it.

That was the process we followed for the entire set of data. For the most part those illegal values were white spaces we identified by searching for attributes of zero length after trimming, bad dates with improperly formatted timestamps,  or numbers with a high number of digits that we cast as data of Numeric data type of certain accuracy.

We also encountered some cases where the data had a combination of the above problems so we had to apply a greater number of scripts to clean up before converting. To facilitate the data conversion we developed a stored procedure packed with the cleaning and conversion statements we called repeatedly for multiple attributes to expedite scripting.

We also had to re-process some loaded data because of their geometry that would not even abide to 1NF.

The acorn detail data was set up as a crosstab with acorn index names as columns. In plain relational logic, that would make a join with other tables that had the same attribute stored in a column very difficult or impossible. To resolve this issue we transposed the original table using a series of unions and created another table with a suffix to distinguish it from the original. In that table we placed every acorn index value and its description in each one column by keeping the rest of the column values constant for every transposed tuple.

The date formatted columns across the schema are not consistent. We didn’t want to lose the timestamp, that most of the time was the differentiator [because we have ideas how to use it in our analysis] thus, the entity relationships on our ERD will be based on the date portion of the date attribute in case it carries a date and time component.

For example, in the weather\_daily\_darksy table the dates are displayed as dates with an hourly time component, whereas in the daily dataset table they show as dates at midnight a join statement will require substring the date component from both tables.

In the case the time element is necessary in the calculation it will still be used for any type of analysis, however, based only upon the data of that table.

To populate the hhblock table we had to read and import 112 xls files [data conditioning] we downloaded on the Windows host on purpose as a data conditioning exercise. Python on Jupyter notebook did the trick. We hid the MySQL credentials into a text file we read with Python, then we opened firewall port 3306 on both ends and established a connection to MySQL schema from the Windows host. Reading the importing the files in hhblock table took only a few mins.

We could not change the directory secure-file-priv [can’t be dynamically changed, only during installation] points to, so we sent all files on the Linux filesystem in var/lib/mysql before we loaded them [adta conditioning] into the tables.

Bidirectional drag and drop between the VBox and the Windows host didn’t work out of the box so we used WinSCP to move the data back and forth between the two, whenever there was  need.

Later on, we realized that logging in the Ubuntu VM was using the one connection VBox allows to a specific user and that was blocking any other RDP attempt, perceived as a second connection by the same user. That was enough of a hint to understand that logging off and trying to only RDP would suffice to get access to Ubuntu.

We logged on the Ubuntu box to use MySQL WorkBench for database access and scripting to pull data for analysis. The lack of a good amount of RAM of the Windows host where the Ubuntu VM was running, created significant communication latency and rendered our development efforts sluggish. To alleviate the problem we installed DBeaver on the Windows host, that made our communication with MySQL much faster and expedited our development work.

## Data Quality Assessment:

As the group reviewed the dataset, we wanted to ensure that quality of the data was up to standards. After careful examination of the attributes of the provided data, and thorough research of its scientific interpretation, we ensured that the data was complete and covered all the aspects required to address the project’s objectives.

We also ensured that there were no Null values within the dataset where there shouldn’t be and we replaced missing attribute values with Nulls. The uniqueness of the data can be demonstrated through the unique identifiers provided for the individual records.

A quick look at the ERD shows that all relations follow 1NF and they have one or more candidate keys, either a single attribute primary key or a composite key. Thus, all relations abide to 2NF, however, there’s at least one table with a transitive functional dependency.

In informations\_households relation changing in the value of non-key attribute acorn\_grouped may cause attribute acorn to change in turn, thus there’s further possible decomposition to attain even higher forms of normalization. In this case we could create a relation that assigns a primary key acorrn\_group\_ID to Acorn-A, Acorn-B, … Acorn-Q groups, and then replace the acorn attribute in informations\_households relation with the acorrn\_group\_ ID of the new relation.

Therefore our ERD remains in 2NF and requires more work before it gets to 3NF, something we won’t venture to do because of time constraints. We remain confident that at 2NF we won’t be hindered from running any SQL scripts we’ll use to feed our analysis and visualizations.

While it is difficult to ensure an absolute accuracy within this dataset, as we are not subject matter experts in this field, the 8.2 usability rating score of the data acquired from Kaggle justifies its acceptable quality.

The 2NF relations of our database do not allow referential integrity. Regardless, atomicity is a principle that must be followed to ensure the integrity of the database so we commit our statement upon successful completion of a running statement.

While the data had some minor hiccups such as using one specified date format, it still conforms to the correct practices in gathering and describing the topic of analysis. Overall, our team has identified the dataset to hold up to good standards to be utilized for the purpose of this project.

## Other Data Sources

For this project, we will not be using another data source. We have come to the conclusion that this database has a sufficient amount of data for us to work with, and we have formed our project around it. To find another dataset that would provide us with the same range of data would be difficult and time consuming, and most importantly, there is no guarantee that we would be able to find another valid data set which is commensurate with ours at all.

# Analytics and Algorithms

It is quite obvious from browsing our ERD that the most prominent feature of the data set under study is the element of time. Our data contains several time fields that are deeply integrated with the data, temperatures specifically. The daily darksky data set has recorded a lot of variations of temperatures with special meaning in meteorology accompanied by their date/time of observation.

Our analysis uses many tools to help us study the data and draw meaningful answers to stated objectives.

We used a correlation matrix to identify and evaluate the statistical significance of the relationship between temperature and energy consumption.

We ran a Shapiro-Wilk test to detect whether daily temperature half hour data readings are normally distributed, throughout the time span of the data set and produced visualizations in the form of heat maps to examine how much variables like temperature and humidity drive energy consumption.

## Time Series Analysis

Due to the time driven nature of the data, we picked two main algorithms for energy forecasting from the pool of time series analysis. Before providing an outline of their inner workings let’s quickly review the four main attributes of a time series.

**Time Series main features**

There are four characteristics in a time series: Level, Trend, Seasonality or Cyclicality and Noise.

1. **Level** is a frequently observed value in the spiky history of a time series. Typically, there are many of these values along the time series, hence, the Level will fluctuate erratically unless the Trend characteristic is present in the Time Series, as the next section describes.

2. When the Level of a time series changes in a certain direction (upwards or downwards) we say that the time series exhibits a **Trend**. However, if the Level fluctuates around a mean then the Trend is random and that renders it useless for further analysis.

When the Trend is not random it can typically be described by a mathematical function such as linear, exponential, inverse, and others.  In that case, trend is perceived as a vector with a magnitude of change and a direction.

3. **Seasonality** is easily described and understood as periodic peaks and valleys in the time series. An example is the familiar retail cycle that maxes out around the holiday season because of the general public’s shopping sprees and quiets down for a few weeks right after New Year’s.

4. **Noise,** is the meaningless or, inexplicable by logic, or science, part of the time series.

These four properties  are considered to interact in an additive or multiplicative manner to produce the final value of the time series we usually observe.

**Holt-Winter Time series technique**

The **Holt-Winters Exponential Smoothing (HWES)**, also called the Triple Exponential Smoothing method, is suitable for studying univariate time series with trend and/or seasonal components, like the energy-temperature or energy-humidity pairs, and it models the next time step as an exponentially weighted linear function of observations at prior time steps, taking trends and seasonality into account.

Holt-Winter is a layered methodology of forecasting techniques where every layer is theoretically based on the one below and provides a correction to a flaw of the previous step. At its foundation sits the **Weighted Average** where weights are picked from frequently used weighing functions, such  as logarithmic, linear, quadratic etc. to smooth out historical data, or to emphasize on the time series values the scientist determines they should carry more weight in the forecasting.

The next layer up is **Exponential Smoothing** that applies yet another layer of weighted averaging to all past values with the weights declining exponentially from more recent to older values. The latter statement makes exponential smoothing unusable when the time series exhibits a trend and/or seasonality.

The next technique called **Holt Exponential Smoothing** corrects regular exponential smoothing’s flaw on trend, whereas the seasonality shortcoming gets addressed by the **Holt-Winters Exponential Smoothing** [HWES], hence, the term assigned to coin the time series technique.

**Seasonal Autoregressive Integrated Moving Average**

ARIMA [Autoregressive Integrated Moving Average] is an analysis model applied to time series data for exploratory purposes or to predict future values.

Regular ARIMA model’s forecast is based upon a number of lagged historical values [autoregression], the differencing of historical values at a certain lag, and the lagged prediction errors as input variables. The reason the model does not pick the past values themselves is to ensure it maintains more stable predictability.

When it is apparent that seasonality is driving a great deal of the observed variance in the data we need to take it into account using SARIMA, an upgraded edition of ARIMA. This improved method forecasts the next step in the sequence as a linear function of the differenced observations, errors, differenced seasonal observations, and seasonal errors at prior time steps.

In a nutshell, SARIMA combines the ARIMA model with the ability to perform the same autoregression, differencing, and moving average modeling at the seasonal level.

The notation for the model outlines the order for the AR(p), I(d), and MA(q) models as corresponding parameters to an ARIMA function and AR(P), I(D), MA(Q) and m parameters at the seasonal level, e.g., **S**ARIMA(p, d, q)**(P, D, Q)m** where “m” is the number of time steps in the seasonal period. The SARIMA model can be used as an overarching model to derive simpler models, such as AR, MA, ARMA and ARIMA, by simply changing the values of its parameters from 0 to greater than 1, depending upon the characteristics of the time series.

Like Holt-Winters SARIMA is also suitable for univariate time series with trend and/or seasonal components.

As with HWES, SARIMA will be used to forecast energy consumption by changing the seasonal lag, namely the AR and m components, as defined above. The data set’s detailed data down to half hour intervals for a total of more than 880 days provides fertile ground for analysis at different time levels, e.g., months, weeks, days, even hours and half-hour periods.

## Regression and Modeling Daily Consumption

Linear Regression is a statistical method that measures the relationship between an independent variable and a dependent variable. In this study, we’ll be using multiple linear regression, which is a method to predict an outcome based on several independent variables. There are plenty of variables in the historical Smart Meter energy consumption data. Through multiple linear regression, we want to explore how the sum of TemperatureMax and the dewPoint affects the total energy consumption. TemperatureMax and the dewPoint are the explanatory variables that are used to predict the outcome which is the energy\_sum. Upon applying the formula in R: [*lm(formula = energy\_sum ~ temperatureMax + dewPoint, data = wdataset])*, we have retrieved a summary of results. The best way to analyze whether to reject or accept your hypothesis is by checking the p-value. In this scenario, the p-value is 9.342e-08 which means we reject the null hypothesis. The min residual is -14.4895 and the Max is 25.7105. When plotted, the residuals will be random which could be an advantage for further exploration in case there’s a hidden pattern we didn’t consider. The most useful R functions were lm(), summary.lm(), coef(), formula(), residuals(), and plot(). The code wasn’t complex as R provided all the logic necessary to analyze the results.

In addition, we also conducted the correlation test between multiple variables to find the correlation coefficients using the Pearson parametric correlation test. The variables used were energy\_sum, temperatureMax, dewPoint, windSpeed, and pressure from the daily\_dataset and weather\_daily\_darksky data sets. We computed the correlation Matrix using the cor() function and passing the method type “Pearson”. In order to find the correlation p-values, we used the rcorr() function from the Hmisc package to calculate the significance levels for Pearson correlations. The p values were under 0.05, which means that there was a strong relationship between all variables and the Null hypothesis is rejected. The correlation coefficient displays that the temperatureMax and dewpoint has a value of 0.86 which shows the highest relationship among all variables that were examined. A strong positive linear relationship is represented by a value of 1 and a negative linear relationship is represented with a value of -1. The Pearson’s correlation coefficient of 0.86 is a strong positive linear relationship. The Pearson parametric correlation test was conducted in R. The functions used were: cor(), rcorr(), corplot(), chart.Correlation(), and heatmap(). The code in R was made simpler because of the packages and functions provided.

**Modeling Daily Energy Consumption**

An important aspect of the objective we’ll address is whether our data for energy consumption follows a normal distribution and can we model or predict the energy consumption for customers for a given date. To determine whether the data follows a normal distribution, we decided to implement the Jarque-Bera test for normality. It is reliable for large data sets, like ours, and designed to determine if our data follows a normal distribution. The formula for the test-statistic of the Jarque-Bera test is JB =n[(b1)26+(b2-3)224]where n is the sample size, √b1 is the sample skewness coefficient, and b2 is the kurtosis coefficient. With this test statistic we state the Null and alternative hypotheses as such:

Ho: The data is distributed normally, Ha: The data is not distributed normally

The test statistic of total energy consumption across time results in a test statistic of 170992448.01 and a p-value which is virtually zero. As such, we reject the null hypothesis and state that the data does not follow a normal distribution. The test was implemented using Python’s Scipy library that supports in-depth statistical analysis calculations.

Now we know the data does not follow a normal distribution; we should model the energy consumption data with an algorithm that does not require a normal distribution. One such algorithm is the machine learning algorithm of random forest regression. Random forest regression is a supervised learning algorithm that builds multiple, decision trees and combines them together to make more accurate and stable predictions. Utilizing the daily\_dataset and hh\_block tables the random forest algorithm will identify the importance of features supplied. The random forest and querying of data was done through libraries in python such as pandas and sci-kit learn.

The steps taken to implement the random forest algorithm are as follows:

1. Join the daily energy consumption column from the daily\_dataset table to the hh\_block table through a date and unique identifier key.
2. Calculate and initialize attributes for the energy consumption the day before on each record, the midday energy consumption on each record, the average energy consumption on a specific date, and a column for each of the categorical variables on what day of the week the energy consumption was collected on (i.e. the check\_mon column would be zero if it wasn’t a Monday or the number one if it was a Monday).
3. Drop all records with Null and zero values as they cause harm in the algorithm’s functionality
4. Identify the energy consumption for each customer per day as the labels and the other columns as the features (since we have limited computing resources we restricted the features to a few columns).
5. We split the data into training and test partitions so that we can measure the accuracy of our data.
6. Fit the random forest regression with 200 trees and all the features. Proceed to test the fit with our test partition and identified the mean absolute error as well as the accuracy
7. Identified the most important features and then reran the random forest regression with only those most important features. We also reidentified the mean absolute error and the accuracy again
8. Plotted the predicted values vs the actual values

Overall the complexity of the code could be described as a medium as it utilized queries that are not too difficult to implement or understand. It also did not take many lines of code to implement the random forest algorithm as most of the work is done through the library itself. The accuracy of the algorithm is tested utilizing the test partition of the dataset and calculating the mean absolute error.

Through our initial regression we received an accuracy level of 73% and identified the two most important features as the total energy consumption through the first 12 hours of the day and the energy consumption from the previous day. After rerunning the algorithm again with only those two variables we received a 71% accuracy rate, but saved much more in computing resources by only utilizing two features. We noticed that the algorithm was able to predict most of the data of energy consumption, however for higher energy consumption its performance worsened. If we were to rerun the algorithm again we would search for features that could better predict energy consumption outputs at higher levels.

## Analysis of the Weather Time Series and Correlation of Daily Weather Statistics to the Daily Average Energy Consumption

This dataset only spanned over the course of two years, so we do not expect any long term effects of climate change to show up in the data. However, we wanted to see how steady some of the weather statistics were, so we performed a Dickey-Fuller test on temperatureMax and pressure from the weather\_daily\_darksky dataset.

This test checks if there is a unit root present, which would mean that the series is non-stationary. The Null hypothesis states if there is a unit root present then the series is not stationary, and the alternative hypothesis is that the series is stationary. For the temperatureMax test, the Dickey-Fuller number was -2.9886, and the p-value was 0.1598. Since the p-value was greater than alpha, or 0.05, we cannot reject the null hypothesis and, thus, we conclude that temperatureMax is not a stationary time series. However, for pressure we found the Dickey-Fuller number to be -5.8983 and the p-value was 0.01. With the p-value being less than 0.05, we reject the Null hypothesis and assume the pressure is a stationary time series. These conclusions show that in short periods of time, the max daily temperature is not a steady state series.

We also performed a Shapiro-Wilks test on the temperatureMax and pressure columns of the weather\_daily\_darksky dataset to test for normality. This test simply takes a set of data to discern if that population is normally distributed. The Null hypothesis is that the sample set is normally distributed, while the alternative hypothesis is that it is not normally distributed. TemperatureMax gave a W-statistic of 0.98656 and a p-value of 3.112e-07. Since the p-value is significantly lower than the alpha of 0.05, we must reject the null and accept that the temperature is not normally distributed. Similarly, the pressure statistics value equaled 0.99115 and its p-value was 3.822e-05, indicating a not normally distributed population.

The effects of climate change on energy consumption is difficult to analyze with only two years of data, however, we decided to do a correlation analysis to see if any of the weather and energy variables had any hidden meaningful correlations. The correlation analysis showed that energy consumption was not significantly correlated to any of the weather data. The highest correlation from energy\_mean in the daily\_dataset was the temperature variables, temperatureMax with a correlation coefficient of -0.1742. The negative correlation makes sense because as energy usage gets higher, the colder you would expect it to be outside, heating demand would increase when outdoor temperatures get lower. The variables not being significantly correlated show that although the weather plays some role in daily energy usage fluctuations.

## Analyzing the Treatment Effects from the Experiment of Randomly Introducing Dynamic Time of Use Energy Pricing

Treatment Group – houses who were subjected to dynamic time of use energy prices

Mean of the maximum daily energy consumption (over a half hour period): 0.79 kWh

Standard deviation of maximum daily energy consumption: 0.60 kWh

Mean of energy\_sum, i.e. total daily energy consumption: 9.50 kWh

Standard deviation of total daily energy consumption: 8.08 kWh

Treatment group sample size: 1,100 houses, 706,031 half hourly energy readings

Control Group – houses on a flat rate tariff of 14.288 pence/kWh

Mean maximum daily energy consumption: 0.85 kWh

Standard deviation of maximum daily energy consumption: 0.68 kWh

Mean of total daily energy consumption: 10.28 kWh

Standard deviation of total daily energy consumption: 9.37 kWh

Control group sample size: 4,467 houses, 2,804,402 half hourly energy readings

Because the daily readings of energy usage collected by the smart meters has plenty of extremely large outliers in it compared to what you would expect it to have if it followed a Gaussian distribution, I performed a two-sample Wilcoxon test rather than a t-test, the non-parametric equivalent test for a difference in means, on the join of the daily\_dataset and the informations\_households tables. The results of that hypothesis test are shown below:

While this decrease is extremely statistically significant, I think it is arguable that the magnitude of this decrease in average daily energy usage of 7.6% is not as large as environmentally minded policy makers would have hoped. That being said, in some ways, the difference in peak daily energy usage is even more important than the difference in total daily energy usage.

So, I also ran a Wilcoxon rank sum test of two independent samples on the difference in the average values of the peak daily energy usage in the treatment versus the control groups. The test statistic for that test was W = 1.0179e+12 with a corresponding p-value smaller than 2.2e-16 which is basically zero.

However, as before, the practical significance here is more important than the statistical significance and the difference in the average maximum daily energy usage for the treatment and control groups was only 7.1 % which was good but not good enough.

Both of the previous tests were useful, but neither of them controlled for omitted variable bias, so in addition to them and because we have access to panel data aka longitudinal data, I ran both a random effects regression and a fixed effects regression on it to try to quantify the effect of time of use pricing on total daily energy use when controlling for other factors. I then conducted a Hausman test to see whether I could use the results of my fixed effects model or I had to stick with my random effects model, the null hypothesis of the Hausman test is that the random effects is more efficient and thus should be used while the alternative hypothesis is that fixed effects is at least as consistent as random effects, so it should be used instead. The results of the Hausman test on the FE vs RE specifications is shown below:

Because the null hypothesis was rejected, I used the results from my FE model specification. The Beta coefficient on the dynamic time of use dummy variable was -1.067 with a p-value below 2.2e-16 which again is essentially zero. That means that the (partial) effect of implementing time of use energy pricing on a given home in London causes a decrease on average of about 1.07 kWh per day of electricity usage in that house. Annually, that amounts to an average reduction of 390.55 kWh used per home in London. According to Statista, there were around 3.4 million dwelling units in London as of 2019. Therefore, if time of use had been implemented city wide, there could have been a decrease of something on the order of about 1.33 billion kWhs in London annually!

# Visualization

[Describe and show findings and results based on a mix of figures and descriptive text. If you have video, it will be limited to presentation, however, it can also be reference as media file in your Blackboard file exchange.]

# Findings

# Summary

[Summarize your findings and results for the reader. What did you discover, prove, disprove, etc.]

# Future Work

[Critical section! Propose future work or next step(s) for this project.]

# Appendix A

Code references – any code used for the analysis

# Appendix B: Risks

# Risk: The sample data is too small to represent the population.

# Probability: low

# Impact: high

# Mitigations:

# Identify a different dataset that better represents the population.

# Risk: Computing power may not be enough to process and analyze the data.

# Probability: low

# Impact: high

# Mitigations:

# Partitioning the dataset to split up the computing power required.

# Risk: The integrity of the data may not be up to standards.

# Probability: medium

# Impact: high

# Mitigation:

# Scrub the data so that it can be utilized appropriately.

# Appendix C

Agile Development

# Appendix D: References

1. What is acorn? (n.d.). Retrieved March 25, 2021, from https://acorn.caci.co.uk/
2. What is apparent temperature? (n.d.). Retrieved March 25, 2021, from https://meteor.geol.iastate.edu/~ckarsten/bufkit/apparent\_temperature.html
3. Kagan, J. (2021, March 04). Bank holiday. Retrieved March 25, 2021, from https://www.investopedia.com/terms/b/bank-holiday.asp
4. Washington clear sky chart. (n.d.). Retrieved March 25, 2021, from https://www.cleardarksky.com/c/WashingtonDCkey.html
5. US Department of Commerce, N. (2021, January 26). Dew point vs Humidity. Retrieved March 25, 2021, from https://www.weather.gov/arx/why\_dewpoint\_vs\_humidity
6. Energy. (n.d.). Retrieved March 25, 2021, from https://www.britannica.com/science/energy
7. European Environmental Agency. (2020, November 23). Energy and climate change. Retrieved March 27, 2021, from https://www.eea.europa.eu/signals/signals-2017/articles/energy-and-climate-change
8. London\_powermeters. (n.d.). Retrieved March 25, 2021, from https://cagdasyetkin.github.io/London\_powermeters.html
9. The science of everyday things. ENCYCLOPEDIA.COM. 22 MAR. 2021 . (2021, March 25). Retrieved March 25, 2021, from https://www.encyclopedia.com/science-and-technology/physics/physics/pressure
10. Compare the Market. (2020, February 11). Energy tariffs explained. Retrieved March 25, 2021, from https://www.comparethemarket.com/energy/content/energy-tariffs-explained/
11. Local observed maximum and minimum images (2015, March 13). US Department of Commerce. Retrieved March 25, 2021, from https://www.weather.gov/fgf/gfeobsmaximin
12. United Nations Climate Change. “What Is the Paris Agreement?” Accessed March 27, 2021. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.
13. “How Is the UK Tackling Climate Change?” Energy & Climate Intelligence Unit. Accessed March 27, 2021. https://eciu.net/analysis/briefings/uk-energy-policies-and-prices/how-is-the-uk-tackling-climate-change.
14. M. Sun, Y. Wang, G. Strbac and C. Kang, "Probabilistic Peak Load Estimation in Smart Cities Using Smart Meter Data," in *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1608-1618, Feb. 2019, Accessed March 27th 2021. https://ieeexplore.ieee.org/abstract/document/8304659/authors#authors
15. World Energy Consumption. (2021, January 17). In *Wikipedia*. https://en.wikipedia.org/wiki/World\_energy\_consumption#/media/File:Global\_Energy\_Consumption.svg
16. Bizon, N., Mahdavi Tabatabaei, N., Blaabjerg, F., Kurt, E. (Eds.). (2017). Energy Harvesting and Energy Efficiency Technology, Methods, and Applications. Springer.
17. Department for Business, Energy and Industrial Strategy. (2014 June). UK Energy In Brief 2020.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/904503/UK\_Energy\_in\_Brief\_2020.pdf
18. Energy and Climate Change. (2017, August 29). European Environmental Agency.https://www.eea.europa.eu/signals/signals-2017/articles/energy-and-climate-change
19. The Paris Agreement. (2015, December 12). United Nations Climate Change. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
20. Topping, C. (2021, March 18). Average electricity usage in the UK: how many kWh does your home use?. OVOEnergy https://www.ovoenergy.com/guides/energy-guides/how-much-electricity-does-a-home-use.html#:~:text=Just%20because%20an%20average%20UK,using%20large%20amounts%20of%20electricity.
21. UK net zero target. (2020, April 20).  Institute for Government. https://www.instituteforgovernment.org.uk/explainers/net-zero-target#:~:text=In%20June%202019%2C%20parliament%20passed,a%20'net%20zero'%20emitte.
22. Innovation at UK Power Networks. (n.d.). UK Power Networks. https://innovation.ukpowernetworks.co.uk/
23. hplindia. n.d. Difference between Smart and Analogue Meters | HPL India. [online] Available at: https://www.hplindia.com/blog/details.php?class=difference-between-Smart-and-analogue-meters#:~:text=Analogue%20meters%20are%20the%20basic,accurate%20amount%20of%20energy%20consumed.&text=Smart%20meters%2C%20on%20the%20other,multifunctional%20than%20the%20analogue%20meters. Retrieved March 28 2021
24. gov.uk. 2020. *Smart Meter Statistics in Great Britain: Quarterly Report to end September 2020*. [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/937577/Q3\_2020\_Smart\_Meters\_Statistics\_Report\_FINAL.pdf Retrieved March 31 2021.
25. caci.co.uk. n.d. *The Acorn User Guide*. [online] Available at: https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf Retrieved March 31 2021
26. Ciplani, J. (2014, May 14). Dark Sky Review: Dark Sky is a weather app you won't mind paying for, Retrieved April 1, 2021 from https://www.cnet.com/reviews/dark-sky-review/
27. Atomic Data Definition Statement Support. (n.d.). MySQL 8.0 Reference Manual https://dev.mysql.com/doc/refman/8.0/en/atomic-ddl.html
28. Danjou, J. (2019, January 7). Why you should care that your SQL DDL is transactional, Retrieved April 2, 2021 from https://julien.danjou.info/why-you-should-care-that-your-sql-ddl-is-transactional/
29. Bakshi, C., 2020. Random Forest Regression. [online] Medium. Retrieved April 14 2021 from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84