# George Mason University

# CS 504

# Spring 2021

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

**Draft 4**

**Team Carbon**

**Team Members:**

**Jhony Islam**

**Spencer Marlen-Starr**

**Radha Kanuri**

**Cassidy Laskodi**

**Stavros Kalamatianos**

Table of Contents

[Abstract 4](#_Toc70780416)

[1 Introduction 5](#_Toc70780417)

[1.1 Background and Rationale 5](#_Toc70780418)

[1.2 Research 16](#_Toc70780419)

[1.3 Project Objectives 17](#_Toc70780420)

[1.4 Problem Space 17](#_Toc70780421)

[1.5 Primary User Story (-ies): 18](#_Toc70780422)

[1.6 Solution Space 18](#_Toc70780423)

[1.7 Product Vision 19](#_Toc70780424)

[Scenario #1 19](#_Toc70780425)

[Scenario #2 19](#_Toc70780427)

[2 Data Acquisition 19](#_Toc70780428)

[2.1 Overview: 19](#_Toc70780429)

[2.2 Field Descriptions: 19](#_Toc70780430)

[2.2.1  ACORN Details 26](#_Toc70780431)

[2.3 Data Context: 27](#_Toc70780432)

[2.4 Data Conditioning 28](#_Toc70780433)

[2.5 Data Quality Assessment: 29](#_Toc70780434)

[2.6 Other Data Sources 31](#_Toc70780435)

[3 Analytics and Algorithms 31](#_Toc70780436)

[Overview 31](#_Toc70780437)

[3.1 Time Series Analysis 53](#_Toc70780438)

[3.2 Regression and Modeling Daily Consumption 70](#_Toc70780439)

[3.3 Analyzing the Longitudinal Data on Treatment Effects from the Experiment of Randomly Introducing Variable Time of Use Energy Tariffs on London Households in an Attempt to Decrease both Overall Daily Energy Consumption and Peak Daily Energy Consumption 78](#_Toc70780440)

[4 Visualization 87](#_Toc70780441)

[5 Findings 89](#_Toc70780442)

[6 Summary 94](#_Toc70780443)

[7 Future Work 95](#_Toc70780444)

[7.1 Adding More Data 95](#_Toc70780445)

[7.2 Possible Real Time Advice for Households 96](#_Toc70780446)

[7.3 Expansion Beyond Households 96](#_Toc70780447)

[Appendix A: Code References 98](#_Toc70780448)

[Appendix B: Risks 100](#_Toc70780449)

[Appendix C: Definiton of Terms 101](#_Toc70780450)

[Appendix D: References 103](#_Toc70780451)

# Abstract

This research report examines the relationships between multiple aspects of weather and residential energy consumption patterns, the effects price charged per unit of energy consumption has on it, and also the role that socioeconomic status and demographic factors play in average household energy consumption behavior. We investigate these many relationships in an attempt to gain greater insights we can use in the development of effective strategies that can be implemented to mitigate the negative impacts of future climate change on humanity. We begin our analysis with a data visualization heavy exploratory analysis of our source data. From there we proceed with various techniques of time series analysis, including the Holt-Winters Exponential Smoothing (HWES) and Seasonal Autoregressive Integrated Moving Average (SARIMA), and then made forecasts using HWES and SARIMA regarding certain aspects of residential energy usage. Beyond those, we also performed one-sided Mann-Whitney U Tests for two samples, and, we ran complex regressions of three different varieties: multiple linear regression, random forests regression, and both fixed effects and random effects regressions for panel data analysis.

# 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.

Energy consumption worldwide has been increasing at breakneck speeds over 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 the distillation of crude oil to its finest components has facilitated the production of fuels specialized to power all types and sizes of motor and air transportation vehicles, 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.

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 producing technology.

Unfortunately, this energy problem has been accompanied by the environmentally-insensitive behavior of a number of constituencies worldwide who due to their unsubstantiated and self-serving beliefs have failed to change course which has made a number of environmental problems worse right now than they otherwise could have been. Their irresponsible inaction has already begun to result in global climate change, especially during the last few decades.

What has also been changing during the last 20 years are the results of sluggish, yet assiduous research on renewable energy technologies, such as wind, biofuels, geothermal, tidal power, solar thermal (to heat up water for hot water use), concentrated solar thermal power plants and solar photovoltaics; some of which are finally showing sufficient maturity to potentially be able to serve as major components of a new long term energy supply portfolio in developed countries without losing their cost competitiveness.

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 [Figure 1a], renewable energy technologies have started contributing enough to be visible on the chart with an upward trend [Figure 1b].

Chart, histogram

Description automatically generated

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

A picture containing text, electronics, screenshot

Description automatically generated

Figure 1b: Global capacity in renewable energy between 2004-2018. Graph is a snapshot overview of how global capacity for renewable power has evolved since 2004 and it does not include large hydro-electric projects of more than 50MW.

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.

Diagram

Description automatically generated

Figure 2a: 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 will begin to emerge in renewable energy production over time. A promising trend shown in Figure 2b.

Chart, line chart

Description automatically generated

Figure 2b: Capital Invested and Deal Count over Time

Power electronics is the engineering study of converting electrical power from one form to another. Some examples of uses for power electronic systems are DC/DC converters used in mobile devices, such as cell phones and AC/DC converters in computers and televisions. Larger scale power electronics are used to control hundreds of megawatt of power flow across our nation’s power grid.

It is estimated that the power wasted in desktop PCs sold in one year is equivalent to seventeen 500MW power plants! Under the premise how ubiquitous and important power electronics are, it is critical to improve their efficiency and one more reason to enable more renewable electricity pumped in 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 compact business entities, where environmentally friendly policies can be implemented more quickly through C-level governance as opposed to the last electricity consuming sector, residential.

The residential sector consists of millions of consumers who are represented and governed by a legislature which they can steer towards what they see as their informed long term interests taking advantage of the system to supersede the rest of the public's uninformed policy preferences. Depending upon that, it may be more difficult, or not, to get the public’s attention and assistance to the greater cause of anthropogenic global warming mitigation.

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 while also minimizing the costs of the mitigation of greenhouse gas emissions so that everyone is made better off than if that mitigation had been done inefficiently.  
The United Kingdom (UK), which is a member of the industrialized nations and the G8, is typical of those nations in that its investment on energy is dominated by the portion devoted to electricity. As Figure 3 (Department for Business, Energy and Industrial Strategy, 2018) shows, it was 55% of the total in 2019.

Chart, bar chart

Description automatically generated

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.

Chart, bar chart

Description automatically generated

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 the UK’s increased environmentally friendliness is the entire European Union's commitment to that goal. According to the European Environmental Agency, the region has already begun experiencing impacts of climate change recently such as rising sea levels and increases in extreme weather. This is caused by greenhouse gases (such as carbon dioxide, methane and nitrous oxide) being emitted into the atmosphere which increases their relative composition within our atmosphere. Once these greenhouse gases humans emit accumulate in the atmosphere, they trap the heat of UV rays from the sun which hit the earth and bounce off towards space, the GHGs deflect them back towards the Earth’s surface which warms the planet’s surface.

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 KiloWatts per hour (KWh) per year, and that number does not include heating (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 the UK’s bioenergy topped the list of energy increase by source type, after tripling during the last 10 years, even (barely) surpassing the increase in nuclear energy use.

Table

Description automatically generated

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

The UK’s progress in adopting cleaner energy is more evident from another statistic showing that energy from renewables, such as biofuels, solar and wind 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.

Graphical user interface, chart, line chart

Description automatically generated

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.

Chart

Description automatically generated

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, the UK's parliamentary democracy which had 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 has committed itself to the goal of emitting no more tons of GHGs in 2050 than they did in 1990.

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 program. Such programs will allow for the mass collection of energy consumption data from British citizens. Analysis of the data from these smart meters will help enable energy analysts to discover simple strategies for saving energy which they can then educate the public on. Furthermore, their findings could also open up market opportunities for environmentally conscious entrepreneurs to create products and services which can help make saving energy easier for British citizens in all sorts of ways. 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, the UK's capital and biggest metropolitan region as well as Europe’s largest urban area, 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 beyond 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 thousand households spread out in the urban area of London is split into two different groups.

1. The first was the treatment group which was subjected to: “dynamic time of use” pricing  as the treatment (whose effects on energy consumption were the goal of the study) with the prior knowledge of the tiered pricing structure per KWh as total consumption increased during peak use hours each day.
2. The second group was the control group of “non-time of use” customers which remained on a flat rate per KWh (Std) throughout each day.

The goal of the study was dual.

1. The first goal was to test the effectiveness of the various signals the power supplier could send to the public to help it control and manage the intermittency of the electricity generated by its renewable energy sources as they become a larger and larger part of its overall electricity production portfolio over time.
2. The second goal was to adjust prices purposefully to throttle down consumption throughout certain parts of the grid that undergo stress during peak usage hours.

The data sets involved carry useful information about the participants' way of living, income and spending habits organized 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 average 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 (which is where more than half of the UK’s energy investment resources go to).

## Research

The problem we look to address is whether there are effective methods to entice British individuals and households to cut down on their residential energy consumption. A main point of focus is utilizing smart meters within the capital of the United Kingdom, London so as to have good quality data on residential energy usage patterns there to go off of. There have been additional studies which revolve around tracking data from smart meters, with one 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 sought 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 useful for understanding the data, we can take a more effective approach to our problem.

## Project Objectives

This project aims to understand important aspects of residential energy consumption patterns and more specifically how stakeholders are affected by weather phenomena and other factors. This project addresses the following questions:

* How does London’s weather affect residential energy consumption behavior?
* Can electrical utility providers forecast future energy consumption and if they can, how far into the future can they do so?
* Do geo-demographics play a role in London’s energy consumption?
* Do London residents contribute to energy conservation and if they do, how?
* What effects if any did the Variable Time-of-Use energy pricing scheme have on energy consumption in London overall and on the acuteness of peak daily usage specifically?
* How does energy consumption vary from day to day and does it follow a certain statistical distribution?

These questions identify different stakeholders and their stances on energy consumption, while also focusing on energy consumption as a whole. Our analysis is scoped to the London dataset. We hope to get a general understanding of energy consumption in a specific geographical location and use it as a starting point for expanding, extending, and contributing this work for implementations in other metropolitan areas around the world.

## 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 dataset comes from a sample of 5,567 homes in the London area from November 2011 to February 2014. This data comes from “smart meters'' placed on people’s homes by the British Government at the behest of the European Union. 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 and during the winter months when temperatures are lower, or using more air conditioning in the middle of 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 also vary by whether they plan outings during holidays or not, by how many appliances they usually run in their house and during what times of the day, their spending habits and to put it bluntly, whether even care enough about the environment and the climate to try to be more vigilant when it comes to energy savings and efficiency.

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 any large geographical region.

Exploratory analysis and the examination of specific relationships within the data will help us extract meaningful information from our dataset 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.

# Data Acquisition

## Overview:

Our data set consists of 6 main files and 112 half hour block files, which was consolidated into a single dataset. 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 allowed for the creation of the relations by consolidating or transforming the raw data.

This resulted in a database containing a total of eleven relations. In this section we’ll describe the attributes of those relations created for our analysis from the greater Kaggle data set.

**daily\_dataset:**

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| LCLid | string | Identification Number of Smart Meter installed in a household selected to participate in the study |
| day | date | Date a measurement was taken |
| energy\_count | 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 | int | are the median, mean, min, max, standard deviation and sum of the daily measurements. |

**information\_households:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| LCLid | string | Identification Number of Smart Meter installed in a household selected  to participate in the study |
| stdorToU | 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 | string | Describes consumer’s social best fit by occupation, salary and spending behavior. |
| ACORN\_grouped | string | Single word description of ACORN’s social status |
| file | string | it’s a pointer to the file that includes the smart meter readings and energy statistics of the specific smart meter |

**hhblock:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| LCLid | string | Identification Number of Smart Meter installed in a household selected  to participate in the study |
| day | date | Date measurement was taken |
| hh\_n, where n = 0,1,2,...,46,47 | float | measurement at a half hour interval throughout the 24 hrs of the day |

Important limitations regarding the hhblock data table: because the half hourly smart meter readings from the houses in our dataset are included as columns/fields, not rows/records and due to the fact that if we transposed those columns into rows by date for every one of the 5,567 houses in our data set, the number of rows in the hhblock table would balloon from about 3.5 million, which is how big it is now, to well over 200 million which we thought was way too large to work with for this project. As a result, most of our analysis of this dataset does not make use of the data in the hhblock table, except for a few charts. One other issue with the hhblock table is that the half hourly energy use readings from thousands less than 5.5 thousand households are included for the entirety of several months out of the November 2011 to February 2014 time frame of the smart meter data collection included in the study with some days, weeks, and months having readings from only several hundred houses.

**weather\_daily\_darksky:**

[please refer to Appendix C “Definition of Terms” for more details about some scientific attributes listed below]

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

**weather\_hourly\_darksky:**

[please refer to Appendix C “Definition of Terms” for more details about some scientific attributes listed below]

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| visibility | float | The farthest one can identify an object, measured in miles |
| windBearing | int | Wind direction during the day |
| temperature | float | temperature of the hour |
| time | datetime | The time which the data was recorded |
| dewPoint | 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 | float | Atmospheric pressure in Millibars |
| apparentTemperature | float | The perceived temperature in degrees Celsius |
| windSpeed | float | The speed of the wind measured in knots  (nautical mile per hour = 0.51 m sec-1 = 1.15 mph) |
| precipType | string | The predicted amount of precipitation in centimeters |
| icon | string | The icon from the darksky app used to report the weather |
| humidity | float | The air water content expressed as a percentage |
| summary | string | A brief description of the hour’s weather |

**acorn\_details:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| MAIN CATEGORIES | string | main categorization of ACORN groups characteristics, such as housing, salary, population, etc |
| CATEGORIES | string | Subgrouping within the main categories, type of housing such as single family, townhouse or age groups within the population category |
| REFERENCE | 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 | 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:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| MAIN CATEGORIES | string | main categorization of ACORN groups characteristics, such as housing, salary, population, etc |
| CATEGORIES | string | Subgrouping within the main categories, type of housing such as single family, townhouse or age groups within the population category |
| REFERENCE | 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 | 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 | string | letter designator of an ACORN group |

**tariffs:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| TariffDateTime | datetime | it’s the date and time during the entire period of collection of the data set at half an hour intervals |
| Tariff | string | the pricing range, namely, Low, Medium, High for every value of TariffDateTime |

**uk\_bank\_holidays:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| Bank Holidays | datetime | the counterpart of the US federal holidays throughout the period of data collection. |
| Type | string | Short holiday description |

## 2.2.1  ACORN Details

The following list is a concise overview of Acorn's 2013 UK population profile of demographics percentages.

|  |  |  |  |
| --- | --- | --- | --- |
| Acorn Group Description | | Population | % |
| 1 | Affluent Achievers |  |  |
|  | 1.A - Lavish Lifestyles | 820,947 | 1.3 |
|  | 1.B - Executive Wealth | 7,788,972 | 12 |
|  | 1.C - Mature Money | 5,663,939 | 8.8 |
| 2 | Rising Prosperity |  |  |
|  | 2.D - City Sophisticates | 2,024,721 | 3.2 |
|  | 2.E - Career Climbers | 3,579,716 | 5.6 |
| 3 | Comfortable Communities |  |  |
|  | 3.F - Countryside Communities | 4,160,615 | 6.5 |
|  | 3.G - Successful Suburbs | 3,844,002 | 6 |
|  | 3.H - Steady Neighbourhoods | 5,376,958 | 8.4 |
|  | 3.I - Comfortable Seniors | 1,645,668 | 2.6 |
|  | 3.J - Starting Out | 2,569,813 | 4 |
| 4 | Financially Stretched |  |  |
|  | 4.K - Student Life | 1,550,112 | 2.4 |
|  | 4.L - Modest Means | 5,078,729 | 7.9 |
|  | 4.M - Striving Families | 5,564,601 | 8.7 |
|  | 4.N - Poorer Pensioners | 3,128,512 | 4.9 |
| 5 | Urban Adversity |  |  |
|  | 5.O - Young Hardship | 3,222,867 | 5 |
|  | 5.P - Struggling Estates | 4,730,766 | 7.4 |
|  | 5.Q - Difficult Circumstances | 2,962,375 | 4.6 |
| 6 | Not Private Households |  |  |
|  | 6.R - Not Private Households | 550,486 | 0.9 |
|  | Total | 64,263,799 | 100 |

## 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

The source data that was downloaded from the owner’s Kaggle web page was already packaged in xls and csv formats.

During database importing the most frequent problem was values that didn’t match the attribute’s data type, such as blank spaces in a numerical attribute that had to be removed to allow a change to the right data type. We used the Notepad++ text editor to identify these data flaws and we corrected them during the ETL process using MySQL.

In order to render the data usable and ready for querying, we had to reformat a lot of date fields, particularly to strip out unnecessary midnight hour (00:00:00) leftover strings in date attributes. Date fields are ubiquitous in our data set and necessary for joining relations so we had to make sure they’re in the proper condition before analysis commenced.

The half-hour block energy readings came in latitudinally in records per day per smart meter in the raw files, making it problematic to use them with joins with other relation attributes that had dates with half hour times in a format like 03-06-2013 15:30:00 stored in columns.

We partially circumvented that problem by summing all half hour intervals grouped by date for all smart meters. That made the data set shorter, more manageable (880 days x 48 sums, with each half-hour sum produced from more than 5,500 residences, resulted in a relation of a few tens of thousands records) and easier to transpose. Instead, transposing 48 measures of half hour intervals from over 5,500 smart meters for over 880 days would result in a table with hundreds of millions of records.

The above trick didn’t hurt our logic because during the transposition we created date fields to carry the date and time that corresponded to each half hour interval.

During our efforts to bring our dataset closer to second normal form (2NF) we had to break down a couple of many-to-many relationships with the creation of intermediate tables, so we used SQL statements to program the logic that performed this task and double checked the results.

Another problem we had to resolve was the ill-conditioned raw format of the ACORN detail data.

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 the new 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 were not consistent. We didn’t want to lose the timestamp because most of the time, that was the differentiator [due to various 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. That was another quiet agreement we came up to for on-the-fly data conditioning when it comes to date based joining.

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. In this case,  a join statement will require a substring to isolate the date component from both tables.

## Data Quality Assessment:

Below are the summaries of our assessments of the dataset we chose with respect to each of the attributes of data quality:

* Completeness: because there are no missing or null values in our dataset, it has an excellent completeness degree of 100%.
* Uniqueness: there are no duplicate values in our dataset which means that each data record in it is unique.
* Accuracy: while the smart meters installed in the London homes were no doubt of high quality, there do appear to be some significant outliers in our data set which means that there may be either accuracy issues or energy spikes induced by the weather or triggered by an event at the particular property. The presence of outliers has been addressed in the Exploratory Data Analysis section to make sure they don’t pose any significant risks in the validity of statistical analysis.
* Atomicity: By default, MySQL runs in **autocommit** mode. Therefore, as soon as an update gets executed, MySQL will store the update on disk.
* Conformity: our dataset does appear to conform to generally accepted formats.
* Overall Quality: the overall quality of our source data is high because it is all from smart meter readings which generates objective, high quality data as opposed to a self report survey which is an opt in sampling method which generates low quality data.

According to its manufacturer MySQL’s InnoDB [the storage engine we will use on this project] also supports atomic DDL statements but not transactional ones. For example, in the case when more than one DDL statements get executed then each one must run as a separate transaction, otherwise if all get executed as one transaction and a rollback takes place before they all complete, the DDL up to the point of rolling back will persist in the database and the rollback won’t take place in full.

Atomicity is important when dealing with concurrent database modifications. Consistency in ACID depends on atomicity’s success, therefore, to present the data while transactions are in progress the concept of isolation level was introduced. Isolating the transactions to ensure their successful execution in the world of databases where hundreds of them may be taking place simultaneously would introduce a huge performance hit. For that reason, there are four isolation levels MySQL supports like other popular database platforms to present the data at a certain state when a Select statement runs. The default InnoDB level is Repeatable Read.

In MySQL’s documentation words: Consistent Reads within the same transaction read thesnapshot established by the first read. This means that if you issue several plain (non-locking)SELECT statements within the same transaction, theseSELECT statements are consistent also with respect to each other.

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 does hold up to good standards commonly 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

## Overview

It is quite obvious from browsing our ERD [Figure 7] that the most prominent feature of the data set under study is the element of time. Our data contains several date/time fields [showing in red on the ERD] that are deeply integrated with the data and temperatures. The daily darksky data set contains variations of temperatures with special meaning in meteorology accompanied by their date/time of observation.

Diagram

Description automatically generated

Figure 7:  Entity Relationship Diagram

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 find out whether the daily temperature half hour data readings we used are normally distributed throughout the time span covered by our dataset and produced visualizations in the form of heat maps to examine how much variables like temperature and humidity drive energy consumption.

We also quantified and analyzed the treatment effects of the experiment which was performed on London residents by taking a random sample of 5,567 households in the city which was then split up into a treatment group of 1,100 people subject to the implementation of dynamic energy (Time of Use or ToU) pricing during the day, while leaving the remaining roughly 4,500 households on the standard (Std) flat rate tariff of 14.23 pence per kWh, the energy price everyone in the study was on, before the dynamic time of use pricing was introduced to the treatment group. For this purpose, we ran what is known as a Mann-Whitney U Test for the difference in distribution between two independent samples of both total daily energy usage and peak daily energy usage between the treatment and control groups. The Mann-Whitney U test, which is also known at the Wilcoxon Rank Sum Test for two independent samples is the non-parametric equivalent of the two sample t-test for a difference in means, we did run one of those for a difference in means between treatment and control groups for both enery\_sum and energy\_max as well, but we did not include them in this project report because the distribution of both of those energy consumption aggregates in our dataset were non-Gaussian, so the t-test results did not mean anything, that is to say, they were not externally valid or even internally valid for that matter.

After analyzing the results of our two Mann-Whitney U tests, we then ran a fixed effects regression specification and a random effects regression specification on our daily energy usage data because our dataset is a longitudinal aka a panel dataset, followed by a Hausman test to determine which of the previous results was more legitimate to use to make inferences from, and draw conclusions about our data. We ran both an FE and a RE regression with energy\_sum, i.e. total daily energy consumption as the dependent variable and a dummy variable for time of use energy pricing as the key independent variable while controlling for all of the other independent variables which could influence energy usage that we had data on. This was followed by another set of FE and RE regression models but with energy\_max, i.e. peak daily energy consumption as the dependent variable while again controlling for all other factors which could impact energy consumption behavior.

Our analysis ran out of RStudio or a collection of Jupyter notebooks from a connection to a live AWS RDS MySQL instance protected behind a public faced firewall on which we set up the inbound rules of connectivity for every member of the team’s WAN IP address. Then we let the connection continue upstream to the RDS by creating a similar set of inbound rules on the subnet’s security group configuration. We permitted egress traffic to any destination IP address.

Several popular R and *Python* packages have been employed in this effort, such as ggplot2, dplyr, dbConnect, libridate, tidyverse, corrr, psych, plm, plotly, numpy, purrr, lubridate, *pandas*, *statsmodels*, *sqlalchemy, math* to name a few.

**Exploratory Data Analysis**

The first data stress test during Exploratory Data Analysis [EDA] was the exploration of outliers in the daily 48 half-hour intervals of energy consumption readings collected from all the participating smart meters for the entire lifespan of UK Energy’s project that lasted about 2.3 years.

Due to the half-hour interval data set’s longitudinal format, we had to develop an R function that takes a vector of values as an input, calculates their IQR and quantiles, and then the cut-off ranges beyond which all data points are considered outliers. Our R function can also test the vector against the Shapiro Normality Test. Thus, its vector output consists of 5 values: upper and lower outlier cut-off values, the mean of the vector values excluding [or including] the outliers, the Shapiro statistic and its p-value.

Following we provide 4 graphs that show the plots of Shapiro statistic vs its p-value in two different formats. The first 2 graphs show the plots of the statistic vs p-value pairs for the vector of the 48 intervals of every smart meter across all time, whereas, the next two graphs show the plots of the same pair of statistical results for the sum of the energy readings in every day’s half-hour interval across meters.

The half hour intervals start from the first 30 minutes after midnight of a new day’s start to the last 30 minutes of that day right before the midnight of the next day’s start. Therefore, Oct 1st 12:00 AM - Oct 1st 12:30 AM is the 1st half hour interval or half hour interval 0, whereas, the last interval runs between Oct 1st 11:30 PM - Oct 2nd 12:00 AM

The data set we ran the statistical tests on looked like this [Table 4], for the first 2 plots:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Day **1** Date | Smart Meter 1 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **1** Date | Smart Meter 2 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| ... | ... | ... | ... | ... | ... | ... |
| Day **1** Date | Smart Meter n | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **2** Date | Smart Meter 1 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **2** Date | Smart Meter 2 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| ... | ... | ... | ... | ... | ... | ... |
| Day **2** Date | Smart Meter n | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **3** Date | Smart Meter 1 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **3** Date | Smart Meter 2 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| ... | ... | ... | ... | ... | ... | ... |
| Day **3** Date | Smart Meter n | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| ... | ... | ... | ... | ... | ... | ... |
| Day **n** Date | Smart Meter 1 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| Day **n** Date | Smart Meter 2 | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |
| ... | ... | ... | ... | ... | ... | ... |
| Day **n** Date | Smart Meter n | Half-hour interval 0 | Half-hour interval 1 | Half-hour interval 2 | …. | Half-hour interval 47 |

Table 4: Structure of the data set used to check for Shapiro normality of daily half hour intervals **without** aggregation

and for the next 2 plots the dataset looked like Table 5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day **1** Date | Sum(Half-hour interval 0) for all smart meters | Sum(Half-hour interval 1) for all smart meters | Sum(Half-hour interval 2) for all smart meters | …. | Sum(Half-hour interval 47) for all smart meters |
| Day **2** Date | Sum(Half-hour interval 0) for all smart meters | Sum(Half-hour interval 1) for all smart meters | Sum(Half-hour interval 2) for all smart meters | …. | Sum(Half-hour interval 47) for all smart meters |
| ... | ... | ... | ... | ... | ... |
| Day **n** Date | Sum(Half-hour interval 0) for all smart meters | Sum(Half-hour interval 1) for all smart meters | Sum(Half-hour interval 2) for all smart meters | …. | Sum(Half-hour interval 47) for all smart meters |

Table 5: Structure of the data set used to check for Shapiro normality of daily half hour intervals **with** aggregation

Figure 8 shows a plot of Shapiro statistic and the p-values independent of each other as two different variables. We see a great deal of points with large test statistics and a lot of p-values above 0.1 and 0.05. But that doesn’t tell us anything until we plot the two numbers as a data point in a scatterplot to identify how many are statistically significant.

A picture containing chart

Description automatically generated

Figure 8: Shapiro normality statistic and p-value vs. time from daily half hour intervals **without smart meter** aggregation

Figure 9 shows Shapiro’s results pair values on an x-y coordinate system. It’s easy to conclude that only a handful of points are statistically significant above thresholds 0.1 [magenta dotted line] or 0.05 [red dotted line]. Those are located at the upper right tip of the line that carries only a handful among the millions of points [over 5,500 smart meter daily intervals x 2.3 yrs x 365 days/yr] collected.

Note that Figure 9 has its y-axis in logarithmic scale to increase the graph’s readability.

Chart, line chart

Description automatically generated

Figure 9: Shapiro normality statistic vs p-value pairs of daily half hour intervals plotted as x-y data points from data **without** smart meter aggregation

|  |  |  |  |
| --- | --- | --- | --- |
| Above/Below p-value | Numb of points | % of points | p-value |
| *Below* | *3,354,441* | *96.68%* | 0.05 |
| **Above** | **99,656** | **2.87%** | 0.05 |
| NA | 15,255 | 0.45% | 0.05 |
| *Below* | *3,389,328* | *97.69%* | 0.1 |
| **Above** | **64,769** | **1.86%** | 0.1 |
| NA | 15,255 | 0.51% | 0.1 |

Table 6: The number of data points with p-values above and below p-value thresholds 0.05 and 0.1 for smart-meter-non-aggregated data

Continuing our data exploration, as mentioned above, we’ll run the same Shapiro normality exercise, only this time we’ll aggregate our data across smart meters.

Figure 10 shows the results of scatter plotting the Shapiro statistic and the p-values independently vs time. As expected we have a much lower number of points, than in Figure 9, with very few surpassing the p-value threshold of 0.05, although most of the test points show a high statistic value above 0.85.

Chart

Description automatically generated with medium confidence

Figure 10: Shapiro normality statistic and p-value vs. time from daily half hour intervals **with smart meter** aggregation

Again, the above chart doesn’t tell us much about the number of data points that make it through the 95% or 90% confidence interval.

Figure 11 shows the results of Shapiro statistic vs p-value in a scatter plot where we observe a higher number of data points surpassing the 95% confidence level and barely a few more the 90% level, compared with the non-smart-meter-aggregated data in Figure 9 in percentage terms.

Note that Figure 11 has its y-axis in logarithmic scale to increase the graph’s readability.

Chart, line chart

Description automatically generated

Figure 11: Shapiro normality statistic vs p-value pairs of daily half hour intervals plotted as x-y data points from data **with smart meter** aggregation [p-value=0.05 in red and p-value=0.1 in magenta]

Table 7 shows the increased number of data points that made it the 95% and 90% confidence interval of the normality test.

|  |  |  |  |
| --- | --- | --- | --- |
| Above/Below p-value | Numb of points | % of points | p-value |
| *Below* | *768* | *92.87%* | 0.05 |
| **Above** | **59** | **7.13%** | 0.05 |
| *Below* | *806* | *97.47%* | 0.1 |
| **Above** | **21** | **2.53%** | 0.1 |

Table 7: Data Points with p-values over 0.05 and 0.1 thresholds for smart-meter-aggregated data

We also created two charts which enabled us to ascertain if London has any clear peak hours of energy consumption every day on average and if it does, when they are. Both of the following graphs were created in Microsoft Excel, they are included below:

Chart, line chart

Description automatically generated

Figure 12: A chart which plots the mean energy use for all households over each half hour interval as it varies throughout the day

Chart, scatter chart

Description automatically generated

Figure 13: A chart of the patterns of energy use by households when different levels of the variable time of use tariffs apply

It is clear upon inspection of the previous two charts that there is a distinct peak time of day for energy use, but when that window is exactly looks different in the two charts. In the first chart, it appears to be between around 6 and 9 pm, while in the second, it looks like it is between 5 and 8 pm, or maybe 5 and 9 pm. So, it might be safe to say that peak daily energy usage hours in London include the time between 6 and 8 pm, that much is clear.

**Energy Consumption vs Weather Elements**

In this section we explored the correlation of daily energy consumption (in half-hour or hourly intervals) with some of the weather variables available; it shows a strong dependency.

Figure 14 shows the correlation matrix of average energy consumption vs a number of weather variables such as temperature, humidity, visibility, wind bearing, dew point, wind speed and apparent temperature.

The diagonal shows the frequency charts of each variable in the above order with average energy consumption on the upper left and apparent temperature at the bottom right.

The bottom triangular part of the matrix shows a bivariate scatter plot of the diagonal column and diagonal row variables that cross at that cell, with a red fitted line.

For example the green bordered cell in the matrix shows the scatterplot between humidity (diagonal variable from straight up) and wind bearing (diagonal variable from straight to the right of the cell). Following the same convention the red bordered cell shows the scatter plot of temperature vs wind speed.

The upper triangular part of the matrix shows the correlation coefficient of the mirrored cell below the diagonal and its significance level as a number of stars.

Each significance level is associated to a symbol : p-values(0, 0.001, 0.01, 0.05, 0.1, 1) <=> symbols(“\*\*\*”, “\*\*”, “\*”, “.”, " “). The 3 stars indicate that the probability to reject a true Null hypothesis, i.e., the probability to make a Type I error is virtually zero.

Thus, the yellow cell that shows the scatter plot of temperature vs dew point has a correlation coefficient 0.95 at 99.9% significance level, therefore, the Null hypothesis that states there is a strong positive correlation between temperature and dew point cannot be rejected.

Other statistically significant correlations (>=absolute(0.8)) at the 99.9% significance level showing in Figure 14 are a negative correlation between temperature and average energy consumption (-0.84), a negative correlation between apparent temperature and average energy consumption (-0.85), a positive correlation between temperature and apparent temperature (0.98) and a positive relationship between dew point and apparent temperature (0.95).

Chart

Description automatically generated

Figure 14: Correlation matrix of average energy consumption and selected weather variables

A picture containing diagram

Description automatically generated

Figure 15: Same variable associations with Figure 14, only under a different visual format and comes as a visual aid complement.

The source data set for the following two figures was the hhblock relation after cleaning up the outliers on every day’s 48 half-hour interval readings for every smart meter across the data collection time horizon.

In Figure 16 we plotted two variables vs time. These are the daily energy consumption average of the half hour intervals and the average of max and min temperatures on that date.

The line graph shows clearly how average energy consumption and average temperature waves exhibit similar amplitudes but inverted phases, thus when the temperatures are lower than 10 degrees Celsius the average half-hour energy consumption increases at the most by 0.1 KWhs at a peak value of 0.3 KWhs to keep London residences warm.

Therefore, energy consumption is correlated to temperature so if the weather forecasts for lower temperatures an increase of the city’s energy load should be expected.

To study the sensitivity of energy consumption to daily outliers we re-plotted [Figure 17] the same pair of variables with the outliers included, only to discover a non-existent difference in consumption’s behavior against temperature, nor any significant change in the range of average values.

From this point on we’ll be using energy readings without removing the outliers the impact of which gets smoothened out in the average energy calculations.

Chart

Description automatically generated

Figure 16: Daily Half-Hour Avg of Energy Consumption vs Avg Daily Temperature (**outliers excluded**)

Chart

Description automatically generated

Figure 17: Daily Half-Hour Avg of Energy Consumption vs Avg Daily Temperature (**outliers included**)

Despite the fact water weather variables, such as humidity and precipitation don't show strong correlation with average energy consumption, perhaps because London residents don’t rely on it for cooling as much as heating, we decided to run a quick analysis in search of a multidimensional relationship among average energy consumption, temperature, humidity and precipitation.

In an effort to conceptualize how those four variables sway through time we created four heat maps that display the values of all variables color-coded across annually, monthly, weekly and daily time frames compressed into one chart per variable.

In Figure 18 temperature in London shows the typical patterns of North Hemisphere weather with warmer weather picking up in May and expiring in October. We observe very similar temperature profiles, close or above 20 degrees Celsius during the two full years of 2012-2013 provided in the data set, as the color-coded cells show, regardless of which time format the viewer picks to look at them.

Next, in Figure 19 Humidity looks relatively consistent throughout the 2.3 year time frame. It occasionally drops to levels between 40% and 60%, however, in its overwhelming majority it stays close or above 80%.

Figure 20 explains why humidity is so prevalent in London’s weather. According to the data, mostly rain, and a few times snow in the winter months of Dec-Febr, with some surprising outliers in March, or even early April, is an inseparable part of London’s daily weather regardless of season.

Usually, humidity levels are higher in the summer months because warm air holds more moisture. For example air at 68 degrees Fahrenheit can contain 10 times more water than at water’s freezing point, namely 32 degrees Fahrenheit. In London, humidity levels do not change much even in the winter, because of the daily dose of rain and in the summer it gets worse.

In reality the air moisture can not be exactly characterized as rain. London is humid because of its maritime climate, that's basically a temperate oceanic weather pattern. The continuous inbound breeze from the Atlantic ocean acts as a permanent moisture supplier keeping the city’s air water content higher than normal but, it also has a significant cooling effect that diverts London’s weather away from extremes such as the US mid-Atlantic unbearable hot humid summer nights.

That partially explains why in Figure 21 energy consumption is only picking up during the colder times of the year, and not so much in the summer, when London residents enjoy the natural cooling effect described above to keep them within their levels of comfort in conjunction with supplementary low energy cooling fans, that are very common in Europe. Hence, the absence of expensive AC units from most of the London homes that would boost energy consumption during the warmer months.

Chart

Description automatically generated

Figure 18: Temperature Time Series heatmap for all time

Chart, timeline, box and whisker chart

Description automatically generated

Figure 19: Humidity Time Series heatmap for all time

**Chart

Description automatically generated with medium confidence**

Figure 20: Precipitation Time Series heatmap for all time

**Chart

Description automatically generated with medium confidence**

Figure 21: Average Energy Consumption Time Series heatmap for all time

**Social Stratification of Energy Consumption**

Our search for factors correlated with energy consumption continues with Figure 22 which shows how it varies among the three socioeconomic levels of ACORN classifications in the city of London. Listed in descending order of social stratification, these are: the Affluent class, the Comfortable class and the Adversity class.

The chart that displays the average daily energy consumption vs time coincides with the intuitive guess that energy is utilized at rates closely related to the standard of living and fiscal convenience.

Chart, histogram

Description automatically generated

Figure 22: Daily average energy use vs time for the three main social classes in London

What is also really interesting in Figure 22 is that all three classes are completely coordinated in their energy utilization habits. We’ll dare to guess this means they are having very similar habits keeping the lights on and daily time intervals they keep their residences at comfortable temperatures, therefore the difference among the waves of utilization is strictly dependent upon the size of their residences.

For example, affluent households who can afford 2-story townhouses will by default spend more energy than households facing financial adversity who probably live in smaller apartments, even if they follow the same life routine that ensures similar daily utilization rates. We believe that's exactly what we observe in Figure 22.

**Energy Consumption vs Standard or Time of Use Pricing**

Figure 23 shows the variation of daily average across different participant houses divided in two categories. Those who knew about the schedule of pricing tariff (Time of Use or ToU) and those who only knew about the standard pricing (Std for standard pricing) and not the daily variations according to ToU. The price change warning signals were communicated to ToU participants via pop up messages sent by UK servers to their smart meters.

Chart, histogram

Description automatically generated

Figure 23: Daily energy consumption average vs time by ToU and Std pricing (green line of ToU group and red line of Std group)

Although the ToU group’s energy use started frugally it quickly worsened by catching up to the Std pricing group but started to differentiate again around February 2012. A much clearer distinction between the two groups' energy use started in March 2012 and continued throughout the entire 2.3 year period, staying in two occasionally overlapping lines during the second half of the time period.

**Energy Consumption by combination of social stratification and type of pricing [Std vs ToU]**

Figure 24 family of plots combines the hourly energy consumption for every hour of the day, for the three major ACORN groups and their ToU and Std subgroups. A latitudinal comparison across tariff subgroups for all ACORNs reveals Affluents and Comfortable ToU subgroups as the most energy sensitive ones throughout the 24 hours of the day with Affluents maximizing savings between the two. Surprisingly the Adversity ToU’s group energy utilization doesn’t look much different than its std’s counterpart, as if they hadn't been informed.

Chart

Description automatically generated

Figure 24: Average hourly Energy Consumption by Tariff and ACORN

The next two plots are zooming out in time to plot the energy usage range of values for the entire time frame in a box plot format, with box plots for households which remained on the standard flat rate price per kWh for energy use and those that were put on the variable time of use rate tariffs during 2013, and for each of the three major socioeconomic status categories. The boxplots reveal the energy usage habits of the ACORN groups during all the periods of tariff range. Again, the Affluents were the most sensitive/reactive group when they had time use of energy pricing imposed on their homes.

In Figure 25 the top of the Affluents’ boxplot is in close proximity to energy consumption 0.3 KWhs and in Figure 26 it drops by almost 0.04 KWhs, below 0.26 KWhs. The Comfortables boxplots drop by amounts that vary between 0.02 and 0.03 KWhs across the tariff ranges, whereas the Adversity group had the worst performance by barely approaching a drop of 0.015 KWhs during the time of normal tariff pricing.

Chart, box and whisker chart

Description automatically generated

Figure 25: Daily average values annual spread by ACORN on Std Tariff pricing

Chart, box and whisker chart

Description automatically generated

Figure 26: Daily average values annual spread by ACORN on ToU Tariff pricing

A confirmation of the observations made so far comes from the next 2 bar charts. In order to produce Figures 27a and 27b we calculated how much the hourly average energy consumption dropped between ACORN groups on Std and ToU pricing and then we ranked, grouped and counted all energy drops by ACORN group.

For example, Figure 27a shows that the Affluents came first dropping their energy consumption by the biggest amount sixteen times and in second place another eight times, in the total of 24 times we calculated the energy drops between ACORN groups on Std and ToU pricing.

As discussed earlier in the study, the Affluent group has been the most responsive in the presence of tariff information, with the Comfortable second and, counter-intuitively the Adversity one the least responsive of all.

Chart, bar chart

Description automatically generated

Figure 27a: Frequency of first, second or third best drop in hourly energy consumption between ACORN groups on Std and ToU pricing (stacked bar format)

Chart, bar chart

Description automatically generated

Figure 27b: Frequency of first, second or third best drop in hourly energy consumption between ACORN groups on Std and ToU pricing (adjacent bars format)

## 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 giving an outline of their inner workings the next section will provide a review of four main attributes of a time series and the two analytic techniques that were used for the analysis.

**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 continues to fluctuate within the same range of variation around a mean over time then there is no Trend in the data and the phenomenon the data represents is considered stationary.

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.

1. **Seasonality** is described and understood as periodic peaks and valleys in the time series. For example, 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 after New Year’s.
2. **Noise,** is the meaningless or, inexplicable by logic or science, part of the time series.

These four properties may interact in an additive or multiplicative manner to produce the final values and shape of the time series that we 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 weighting functions, such  as logarithmic, linear, quadratic etc. to smooth out historical data, or to emphasize the time series values which the scientist determines should carry more weight in their forecasts.

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 title of this 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.

The standard 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 version of ARIMA. This seasonal effects robust 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 up 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.

**Time Series Analysis with Monthly energy data**

In this study we used both time series analysis algorithms for a single value and time interval forecasts. Our single value experiment was a forecast of the last value in the data set. In particular, after calculating the daily average of energy consumption across smart meters by ACORN group [Affluent, Comfortable, Adversity, ACORN- ACORN-U], Tariff period [Low, Medium, High] and whether a household was on the time of use tariffs or not [std, ToU] we created a subset of the whole data frame, filtering for every combination of the three attributes [5 x 3 x 2] and we applied HWES and SARIMA algorithms to predict what the next day’s average energy consumption would be. Then we compared the forecast of the real last day’s average energy consumption with the predicted one by matching the 3-variable values to the right subset.

The last day in the data with a Tariff value defined in the tariffs table was 12/31/2013. From the tariffs table we observe that the tariff was Normal during the entire day. Thus, we expect to compare the forecasts of the HWES and SARIMA techniques from the subsets with Tariff equal to Normal for both Std and ToU tariff groups [since UK Energy decided not to adjust the tariff during that day, perhaps because it was the last day of 2013, hence a good day for profit] and all ACORN categories.

The following table [Table 8] summarizes our results with brief error percentage calculations for both techniques.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Std/ToU** | **Acorn Group** | **Tariff** | **SARIMA** | **HWES** | **Avg\_Energy\_Consump** | **%SARIMAError** | **%HWES Error** |
| ToU | Affluent | Normal | 0.257 | 0.257 | 0.265 | *-3.02* | *-3.02* |
| ToU | Comfortable | Normal | 0.215 | 0.215 | 0.220 | *-2.27* | *-2.27* |
| ToU | Adversity | Normal | 0.197 | 0.198 | 0.194 | 1.55 | 2.06 |
| ToU | ACORN- | Normal | 0.206 | 0.177 | 0.196 | 5.10 | *-9.69* |
| ToU | ACORN-U | Normal | 0.283 | 0.290 | 0.300 | *-5.67* | *-3.33* |
| Std | Affluent | Normal | 0.286 | 0.286 | 0.286 | 0.00 | 0.00 |
| Std | Comfortable | Normal | 0.256 | 0.256 | 0.256 | 0.00 | 0.00 |
| Std | Adversity | Normal | 0.207 | 0.207 | 0.211 | *-1.90* | *-1.90* |
| Std | ACORN-U | Normal | 0.281 | 0.268 | 0.285 | *-1.40* | *-5.96* |

Table 8: Percentage errors of current data last day’s average energy consumption projection using SARIMA and HWES time series forecasting techniques

Our analysis continued with an attempt to make longer timespan projections employing both time series techniques.

**Holt-Winter’s Forecasting**

After collecting the average monthly hourly values for energy consumption for all months we plotted [Figure 28] them vs time for a visual inspection of the time series.

Chart, line chart

Description automatically generated

Figure 28: Average Monthly Energy Consumption across all smart meters vs time

Next we started working on the multiplicative HWES model by decomposing [Figure 29] the time series to search for Levels, Trends and Seasonality in the data.

Chart

Description automatically generated with medium confidence

Figure 29: Monthly Average of Hourly Energy Use Time Series decomposition

Figure 29 shows 4 charts. Their sequence from top to bottom is the original [Observed] set, the extracted Trend, the Seasonal element and the Residual, noisy or unexplained part of the series.

It is quite clear that the data contains all three features, namely, Levels, a downward Trend and strong Seasonal component.

Then, we fit the data into a model of Holt-Winters triple Exponential Smoothing algorithm implemented by statsmodels. Our plot [Figure 30] contains both Additive and Multiplicative versions of the model.

Chart, line chart

Description automatically generated

Figure 30: Forecasting with Holt-Winter’s Additive and Multiplicative model

The results are quite promising, so we proceeded with forecasting the energy average for the latest 6-month window of our data set to test the model’s accuracy.

To do that we divided the dataset into a 80/20 Train and a Test set and then we ran a model fit.  Figure 31 shows the results. The model has done a decent forecasting job as it shows in the latter segment of the series.

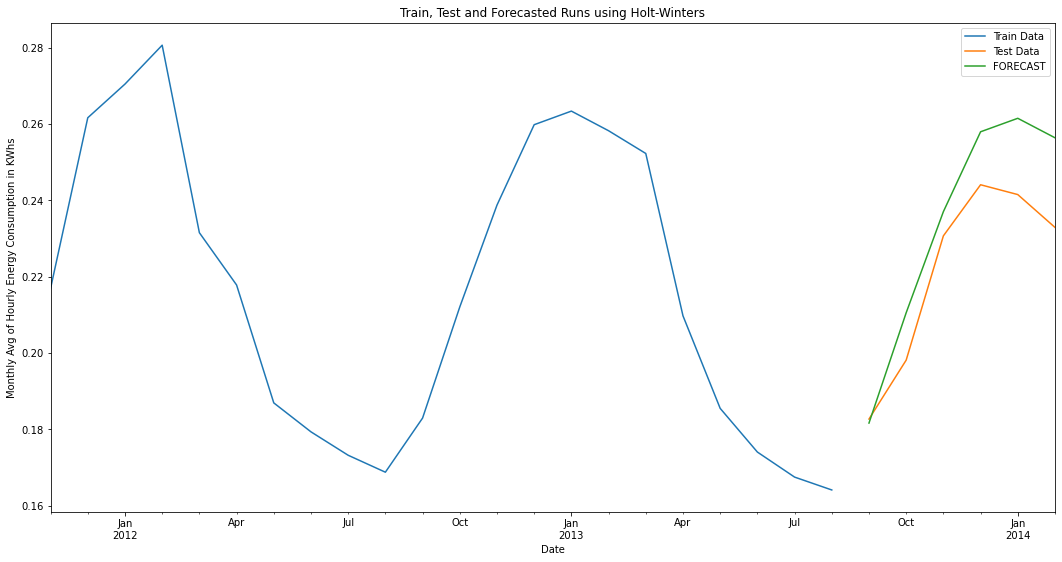


Figure 31: Train, Test and Forecast Runs [most current 6 months] - Holt-Winter’s Model

Figure 32 zooms in the forecasted part of Figure 31 to show a visual of the discrepancy between the Test and Forecast lines. As we move further away from the start of the forecasted period the gap between the Test and Forecasted values widens to reach a maximum of about 0.02 KWhs at the end of the forecasted period, in Febr. 2014.

Chart, line chart

Description automatically generated

Figure 32: Test vs Forecast Runs of Holt-Winter’s Model [last 6 months]

We’ll conclude this forecasting exercise with a couple evaluation metrics that measure the prediction error, providing a more rigorous evaluation of HWES model’s performance.

The two evaluation errors we calculated were MAE [Mean Absolute error] and MSE [Mean Squared error] showing below.

Mean Absolute Error = 0.013

Mean Squared Error = 0.0002

Overall, HWES performed reasonably well, considering we used only 2.3 years worth of data. A data set that spanned a longer time frame is expected to render a more accurate HWES model.

**SARIMA Forecasting**

Just like with HWES, in this section of our forecasting analysis we’ll rely on statsmodel’s implementation of SARIMA. There are 3 major components in the SARIMA function:

1. The data
2. The function order argument that informs about the number of AR terms [*p*], the number of time steps for nonseasonal differencing needed for stationarity [*d*] and the number of MA lagged forecast errors [*q*] to take into account in the regression.
3. The function’s seasonal\_order argument that’s similar to order argument only; it's meant to specify the seasonality component. The first three values of seasonal\_order are the same with order argument’s, except the last value that defines the seasonality period’s length of time. Thus, in the case the repeating pattern is a week, the last argument of seasonal\_order would be equal to 7.

In this analysis we’ll follow a methodology that finds the model’s parameters by grid search.

A grid search, also known as the hyperparameter optimization method, iteratively explores and evaluates different combinations of parameters for fitting an ARIMA model. In each iteration we fit the seasonal ARIMA model with the SARIMAX() function of the statsmodels module, returning an object of the MLEResults class, that, in turn, holds an aic attribute that carries the AIC value. The model with the lowest AIC value gives us the best-fitting model that determines our parameters of *p*, *d*, and *q*.

The Akaike Information Criterion [AIC] is a mathematical method to evaluate how well a model fits the data it was generated from. In statistics, the AIC is used for model comparison purposes and it helps determine which model best fits the data.

With the grid search function defined, we can call our monthly data and print out the model parameters with the lowest AIC value. The AIC results for our data set show below:

*ARIMA(0, 1, 0)x(1, 0, 0, 12)*

*Lowest AIC: -136.559*

The grid search function resulted in an ARIMA(1,0,0,12) seasonal component model with a lowest AIC value equal to -136.559. Next, we used these parameters to fit our SARIMAX model [Table 9].

Table

Description automatically generated

Table 9: SARIMAX model fitting results the Monthly Average of Hourly Energy Consumption time series

In general terms lower values for AIC and BIC means that a model is considered to be more likely to be the True model. HQIC is yet another criterion for model selection. Although often cited, HQIC has seen very little use in practice. Nevertheless, its value is still up to par with the other two measures so there’s nothing unusual here.

The log-likelihood value is pretty low, a good indication of the model’s good fit.

The low p-value of Ljung-Box hypothesis test fails to reject the Null hypothesis that states the time series residuals are not autocorrelated; an expected outcome, at least  based upon visual intuition that dictates the time series model does not exhibit a lack of fit.

The low value of the skewness and kurtosis joint hypothesis Jarque-Bera test indicates that residuals are normally distributed. No wonder since the value of skewness is a perfect 0 and the kurtosis is very close to the standard value of 3 for a univariate normal distribution.

It is important to run model diagnostics to investigate whether the model assumptions have been violated.

The top-right plot in Figure 33 shows the kernel density estimate (KDE) of the standardized residuals, which suggests the errors are normally distributed with a mean approaching zero.

Table 10, right below Figure 33, shows a more detailed statistic of the residuals. From the description of the residuals, the approximate zero mean suggests that the prediction is unbiased.

A picture containing graphical user interface

Description automatically generated

Figure 33: SARIMA Forecasting model diagnostics

--------------------------------------------------------------------------------------

count    28.000000

mean      0.007247

std       0.045472

min      -0.049146

25%      -0.006501

50%      -0.002476

75%       0.010941

max       0.216953

Table 10: SARIMA Forecasting model validity - Residual Statistics

SARIMAX Results object of the statsmodel module comes with a lot of methods that provide great functionality; here we'll use a method to perform in-sample prediction and out-of-sample forecasting and another that returns the lower- and upper-bounded confidence intervals of the predictions of the fitted parameters, at a 95% confidence interval by default.

Here we’ll perform an in-sample prediction of the most recent 6 months and an out-of-sample forecasting of the next 6 months.

The forecasted intervals values are showing in Table 11.

---------------------------------------------------------------------------------------------------------------------

Lower confidence interval of all\_meter\_monthly\_energy\_mean

2013-09-01                             0.144678

2013-10-01                             0.174109

2013-11-01                             0.187792

2013-12-01                             0.216292

2014-01-01                             0.216964

2014-02-01                             0.208066

Upper confidence interval of all\_meter\_monthly\_energy\_mean

2013-09-01                             0.204144

2013-10-01                             0.233575

2013-11-01                             0.247258

2013-12-01                             0.275758

2014-01-01                             0.276429

2014-02-01                             0.267532

Table 11: Lower and Upper confidence intervals of Monthly Average Energy Consumption across all smart meters

The plot of the predicted and forecasted energy values against our original dataset, for the timespan of the data shows below in Figure 34.

The solid line plot shows the observed values, while the dotted lines show the last current six-month rolling forecasts following closely, while staying bounded by the confidence intervals in the pink-shaded area. As the six-month forecast moves beyond the last month of our data set [Febr. 2014] , the confidence interval widens to reflect the uncertainty of the outlook.

Chart, line chart

Description automatically generated

Figure 34: Current and forecasted hourly energy consumption monthly averages with their Upper and Lower confidence intervals

**Time Series Analysis with Daily energy data**

The above analysis was based on monthly basis energy totals. Out of curiosity we decided to continue our forecasting adventure retracing the exact same analysis steps only by using daily averages.

As you’ll find out we never crossed the finish line due to a problem we couldn’t overcome because of reasons beyond our control.

The time series candidate for analysis shows much busier this time. Figure 35 shows the daily average of hourly energy consumption across smart meters vs time and Figure 36 is its decomposition.

Chart, histogram

Description automatically generated

Figure 35: All Smart Meter Daily Average of Hourly Energy Consumption vs time

In Figure 36 we observe a non-existent trend that’s fluctuating sinusoidally, a very clear high frequency seasonality and significant noise. Maybe by moving one order of magnitude lower in the dimension of time we lost the time series components; they were much more well defined in the monthly data set.

Chart, histogram

Description automatically generated

Figure 36: Daily Average of Hourly Energy Consumption Series decomposition

**HWES Forecasting Model**

Despite the discouraging results from decomposition we continued with a modeling fit of a Triple Exponential Smoothing or HWES that resulted in a very promising outcome [Figure 37].

Figure 38 shows how HWES did forecasting the last 6 months of the data set after an 80/20 split of the data for training and testing. The forecast line exaggerated by approximately 0.025 KWhs till Jan. 2014 and then it kept widening for the rest of the predicting horizon, reaching the max overshoot of 0.075 KWhs. Considering that the sinusoidal of the observed curve has an amplitude of about 0.08 the max error on the edge of the outlook, close to Febr. 2014, comes to be as much as one amplitude.

The above back-of-the-envelope calculations describe an extreme discrepancy observed in the predicting region that changes continuously between 0.075 KWhs and lower values throughout. This is strike two that shakes our faith to the model, however, it won’t deter us moving forward in our forecasting modeling exercise.

Chart, histogram

Description automatically generated

Figure 37: HWES model fit on Daily Energy Consumption Averages

Chart

Description automatically generated

Figure 38:  Forecasting of last six-month Daily Energy Consumption Averages using HWES

Zooming in the forecasting segment of Figure 38 produces Figure 39 that clearly shows the model’s over-predictability. Notice that around Febr. 2014 forecasting hits its maximum value of discrepancy at about 0.08 KWhs.

Text

Description automatically generated

Figure 39: Close up of HWES predicting horizon

The evaluation metrics, we also used for the Monthly data, show how the model has fared in terms of errors. These metrics show below:

Mean Absolute Error = 0.04  [that’s half of the observed curve’s approximate amplitude]

Mean Squared Error = 0.002

If we had more time we would try a weekly model or we would experiment with a forecasting horizon between two and four months and observe how the error evaluation metrics change, before making a decision on its final value.

**SARIMA Forecasting**

Our efforts to continue forecasting with SARIMA ended quickly because of a hardware error triggered by the abruptly increased complexity of the calculations required in the initial steps of the methodology.

In Appendix A the reader will find a screenshot of the recorded error that delineates the hardware limitation.

We tried to mitigate this issue by looking for a GPU implementation of SARIMA algorithm that would allow our NVIDIA video card’s GPU to engage and support our CPU [something similar has been done with DBScan clustering as we experienced in another GMU class] but we couldn’t find anything available, so, we ended our forecasting calculations in this section of the study.

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

We also performed a Shapiro-Wilk test on the temperatureMax and pressure columns of the weather\_daily\_darksky dataset to test for normality. This test is used to discern if a variable of interest in a dataset is normally distributed. The null hypothesis is that the data 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 standard threshold 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, again indicating a non-normal distribution of atmospheric pressure.

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 variable that has the highest correlation with energy\_mean in the daily\_dataset was the temperature variable, 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, it is not a large role, at least not over this short time scale.

## Regression and Modeling Daily Consumption

**Linear Regression**

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.

Chart

Description automatically generated

Figure 40: Energy Sum vs. Temperature Max.

Chart, scatter chart

Description automatically generated

Figure 41: Dew Point vs. Temperature Max

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 for most variables were under 0.05, which means that the null hypothesis is rejected. The only exception was the relationship between the temperatureMax and dewpoint which has a p value of 0.86 which shows a strong 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.

A picture containing diagram

Description automatically generated

Figure 42: Scatterplot matrix with energy sum and TemperatureMax, dew Point, wind speed, and pressure.

Chart, bubble chart

Description automatically generated

Figure 43: Correlation Matrix using the Pearson Method with variables including Energy Sum, TemperatureMax, dew Point, wind speed, and pressure.

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

The test statistic of total energy consumption across time results in a test statistic of 170,992,448.01 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.

We can see support for this in Figure 44 which visualizes the distribution of daily energy consumption. It demonstrates a positively skewed distribution which supports our claim that the daily energy consumption is not normally distributed.

Shape, rectangle

Description automatically generated

Figure 44: Distribution of Daily Energy Consumption

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% utilizing 13 features we believed would best predict daily energy consumptions. From there we also 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. This can be seen in Figure 45 which displays all the features all utilized initially and their importance in the algorithm.

Chart, waterfall chart

Description automatically generated

Figure 45: Variable Importance from Random Forest Algorithm

We also notice that the day of the week does not play a significant role in predicting daily energy consumption in our dataset, along with year, month, day, and the average daily consumption for a particular date giving very little importance within the algorithm.

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. This can be seen in Figure 46 which visualizes the actual values for daily energy consumption and the predicted values from the algorithm. We can see the lower values of energy consumption being fit better than the higher outputs. If we were to rerun the algorithm again we would search for features that could better predict energy consumption outputs at higher levels.

Chart, scatter chart

Description automatically generated

Figure 46: Actual vs Predicted Values of Daily Energy Consumption

## Estimating the Average Treatment Effects of Variable Time of Use Energy Prices on Total Daily Consumption and Consumption during Peak Hours using Data from a Longitudinal Field Experiment on a Representative Sample of London Households in 2013

**Key Descriptive Statistics for the Treatment and Control Groups**

Treatment Group – homes subjected to dynamic time of use energy prices in 2013

Median of the max daily energy consumption over a half hour period: 0.67 kWh

Sample mean of the maximum daily energy consumption over a half hour period: 0.79 kWh

Sample standard deviation of maximum daily energy consumption: 0.60 kWh

Kurtosis of energy\_max: 5.03 which is leptokurtic, meaning its distribution has thick tails

Median of energy\_sum: 7.42 kWh

Sample mean, i.e. average of energy\_sum, i.e. total daily energy consumption: 9.44 kWh

Sample standard deviation of energy\_sum: 8.26 kWh

Kurtosis of energy\_sum: 29.05 which is extremely leptokurtic, i.e. it has fat tails

Treatment group sample size: 1,100 houses, 393,612 records of daily energy aggregates

Control Group – homes which remained on a flat energy rate of 14.29 pence/kWh in 2013

Median of the max daily energy consumption over a half hour period: 0.69 kWh

Sample mean of the maximum daily energy consumption over a half hour period: 0.84 kWh

Sample standard deviation of maximum daily energy consumption: 0.68 kWh

Kurtosis of energy\_max: 6.53 which is leptokurtic, meaning its distribution has thick tails

Median of energy\_sum: 7.90 kWh

Sample mean, i.e the average of total daily energy consumption: 10.28 kWh

Sample standard deviation of total daily energy consumption: 9.40 kWh

Kurtosis of energy\_sum: 29.20 which means its distribution is very leptokurtic, i.e. it has fat tails

Control group sample size: 4,467 houses, 1,546,619 records of daily energy aggregates

**Estimating the Average Treatment Effect of Variable Time of Use Prices on Total Daily Energy Consumption**

We will start with several Q-Q plots to see if our data follows a Gaussian distribution. A quantile-quantile plot or Q-Q plot is an exploratory graphical tool used by statisticians, statistical data analysts, and data scientists to check the validity of a distributional assumption for a data set. The idea is to calculate the value one would theoretically expect for each data point in a dataset if they followed a certain probability distribution, which in this situation is the misnamed “Normal” distribution which I prefer to call by its other name, the Gaussian distribution. If our energy\_sum data do in fact follow a Gaussian probability distribution, then the points on the following Q-Q plots will approximately fall on straight diagonal lines.

A picture containing diagram

Description automatically generated

Figure 47: Q-Q Plot of the observed distribution of the entire total daily energy consumption column vs what it would look like if it followed a Gaussian distribution

A picture containing diagram

Description automatically generated

Figure 48: Q-Q Plot of the observed distribution of total daily energy consumption for time of use rate homes vs what one would expect it to be if it followed a Gaussian distribution

Chart

Description automatically generated with medium confidence

Figure 49: Q-Q Plot of the observed distribution of peak daily energy consumption for flat rate homes vs what one would expect look like if it followed a Gaussian distribution

As can be seen in the above Q-Q plots of the total daily energy consumption numbers in our dataset, the readings corresponding to them collected by the smart meters do not follow a Gaussian distribution closely or approximately, in fact, they do not even come close to following the bell curve.

**Mann-Whitney U Test for Decrease in Total Daily Energy Use in the Treatment Group**

Because our data is non-Gaussian and does not follow a student’s t distribution even approximately either, we could not rely on the results of a two independent samples t-test for a difference in means which is the standard parametric statistical test implemented in this situation, so we performed its non-parametric equivalent which is a one-sided Mann-Whitney U Test, also known as a two independent samples Wilcoxon Rank-Sum Test, on the energy\_sum and standard rate or time of use rate fields in a left outer join of the daily\_dataset and the informations\_households tables in RStudio. The null hypothesis for a one-sided Mann-Whitney U Test in this situation is that the two populations have equal distributions and the alternative hypothesis is that the distribution for the ToU households is shifted to the right of the distribution for the standard (std) households. The results of that hypothesis test are included below:

Text

Description automatically generated

Table 12: A snip of the console in RStudio after running the wilcox.test() function

It is important to point out that the test statistic, W, of 3.20e+11, an enormous number (over three hundred brillion), is not an estimate of the difference in medians, that would be patently absurd, rather it is value of the U statistic which is where the “U” in Mann-Whitney U comes from. The U statistic corresponds to a summation of the ranks of all the values in the first sample, which in this situation is the households on the standard flat rates, with the minimum value subtracted (what this statistic is measuring is hard to explain in a comprehensible way, so we’ve put in a second explanation). Put differently, the W, i.e. the U statistic is a count of the number of energy\_sum values for standard rate households that each daily energy\_sum value for households on time of use rate tariffs is less than when it is less than at least one daily energy\_sum value for households on standard flat energy rates.

The corresponding p-value to that test statistic is less than 2.2e-16 which is extremely statistically significant at any common or reasonable threshold because which is basically zero. While this decrease is extremely statistically significant, it is arguable that the practical significance of the raw magnitude of the decrease in the median of total daily energy consumption in the treatment group of around 0.45 kWh (the difference in location of 0.4500106), which in percentage terms is a decrease of roughly 5.7%. This decrease is somewhat underwhelming if the goal is dramatic action to help mitigate future climate change.

**Panel Data Econometrics**

The Mann-Whitney U Test was useful as a starting tool for our analysis of the treatment effects of this experiment, but it does not account for the potential impact of omitted variable bias because it is the non-parametric equivalent of running a single variable regression. So in addition to it and because we have access to longitudinal data (aka panel data as econometricians call it), we ran both a random effects regression and a fixed effects regression with energy\_sum as the dependent variable and our time of use dummy variable as the key regressor it to try to quantify the effect of time of use pricing on total daily energy use when controlling for other all other fixed and random factors that can impact energy usage behavior that we have data on.

We then conducted a Hausman specification test to see whether we should use the results of our fixed effects model or we should stick with our random effects results. The null hypothesis of the Hausman test is that random effects is more consistent than fixed effects, so it should be the preferred model out of the two to implement while the alternative hypothesis is that fixed effects is at least as efficient and consistent as random effects, so it should be preferred. The results of the Hausman Test on the FE vs RE model specifications we ran is included below:

Text

Description automatically generated with medium confidence

Table 13: A snip of the console in RStudio showing the results of the Hausman Test

We found that we did not have sufficient evidence to reject the null hypothesis that our random effects specification is more consistent than our fixed effects specification, so we include only the results from our RE regression; however, to be honest, it would not really make a difference if we used the results of our FE regression we ran anyway because the coefficient on the ToU\_dummy, its standard error, its t-statistic and corresponding p-value in the results of the FE regression were all identical up to two decimal places with the results of the RE regression.

Our overall Oneway (individual) effect random effects model had an Chi-Square statistic of 116,123 with a p-value of less than 2.2e-16 which again is essentially zero and thus highly statistically significant. The coefficient on the time of use dummy variable was -1.13 with a standard error of around 0.02, a single regressor z-value of roughly -71.8 and a p-value below 2.2e-16 which is easily significant at the 0.01 level. That means that what is known in statistical jargon as the partial effect of introducing variable time of use energy pricing on a given household in London (while holding all other factors which impact energy usage behavior constant) is a decrease on average of about 1.13 kWh per day of electricity usage by that household. This is actually larger than the raw difference in means between the mean daily energy use of the treatment and control groups which is only 0.84 kWh and it amounts to roughly an 11% reduction in total daily energy consumption

**Estimating the Treatment Effect of Variable Time of Use Prices on Maximum, i.e. Peak Daily Energy Consumption**

We’ll start this analysis with an assessment of the distribution of the daily maximum energy consumption (within a half hour period) data, specifically, whether or not they can be said to approximately follow a Gaussian bell shaped curve. Instead of clogging up this section with another 3 large Q-Q plots though, we will simply state the results of two informal hypothesis tests we conducted on the energy\_max variable which together are as convincing or more than a Q-Q plot as to the non-Gaussian nature of energy consumption during peak use hours. The first was to run a query in MySQL to see what the maximum value of the energy\_max field was, the result was 10.76 kWh used within a half hour period.

Given that the mean of the energy\_max field in the daily energy use aggregates table is 0.83 and the standard deviation of it is 0.60, that means that a value of 10.76 kWh is 15 sigmas away from the mean which is frankly so ridiculous that it could be taken as a falsification of a Gaussian distributional assumption in and of itself! The second was to make sure that the 7.6 kWh reading wasn’t a fluke by making sure it was not the only record more than 4 or 5 sigmas away from the mean, so we ran another query in MySQLWorkbench to determine how many records in the dataset contain max daily energy readings more than 6 sigmas away from the mean which would require a half hourly smart meter reading of 4.8 kWh or more and found that there were 2,345 such readings! Looking at both of those findings together to paint a fuller picture of the data, it is clear that it cannot be said to approximately follow a Gaussian.

**Mann-Whitney U Test for Decrease in Peak Daily Energy Use in the Treatment Group**

As a result, we once again ran a Wilcoxon Rank Sum Test of two independent samples on the difference in the overall distributions and the median values of the max, i.e. peak daily energy usage in the treatment vs the control groups. And because this is a one-sided version of the Mann-Whitney U Test, the alternative hypothesis is that the distribution of the max daily energy readings for the ToU group is shifted to the right of the distribution for the standard flat rate group. The results of that Wilcoxon Rank Sum Test are:

Text

Description automatically generated

Table 14: A snip of the results of the wilcoxon.test() function ran on the energy\_max and type of energy pricing for the household variables

As you can see, the null hypothesis that the two populations were equal, i.e. there was no difference in the locations of the two distributions or their median peak hours energy usage was rejected unambiguously. Due to the fact that the test statistic U for the Mann-Whitney U Test is so much more complicated than the majority of standard parametric statistical tests which are the ones students in statistics courses are generally taught, I will include another explanation of what the W in the above snip (which is actually the U) means, this time by illustrating the formula used to calculate it. The formula for the test statistic for the Mann Whitney U Test distribution free statistical test is the smaller of U1 and U2, defined below:

where R1 is the sum of the rankings of all the values in the first group, which is households on a standard flat rate, and R2 is the sum of the ranks for the second group, which is households on ToU prices.

However, as before, the practical significance here is more important than the statistical significance and the reduction in the median maximum daily energy usage in the treatment as compared to the control group was only 0.029 kWh. In percentage terms, that comes out to around 4.4% which is in the right direction because it is a decrease, but not a very large one in any practical sense of a reduction in strain on the power grid during peak use hours.

**Panel Data Econometrics**

For the same reasons as highlighted before of controlling for the potential of omitted variables bias, we also ran both fixed effects and random effects regression specifications in RStudio with the mean of peak daily energy usage as the dependent variable and whether the homeowner is being charged a standard flat rate per kWh for their energy consumption or being charged differential time of use rate tariffs where they are charged a low, medium, or high rate per kWh used depending on what time of day it is used by them. And as before, once we had run both the RE and FE regressions, we then used a Hausman test to decide which results to include here in our report and to base our inferences, interpretations, and conclusions off of. The result of that Hausman test is included below:

Text

Description automatically generated

Table 17: A snip of the results of the Hausman Specification Test

We failed to reject the null hypothesis that the preferred model is the random effects specification, thus we report the results of our Oneway (individual) effect random effects regression specification with energy\_max as the dependent variable and the time of use tariff dummy variable as the key independent variable.

The overall model Chi-Square statistic for our Oneway random effects regression is 64,156 with a p-value of less than 2.2e-16. The coefficient on the time of use tariff dummy variable is -0.07, its standard error is 0.001, and its test statistic, i.e. its z-value is -62.9 with a corresponding p-value of less than 2.2e-16. This means that on average and holding other potential causal factors constant, the decrease in the energy consumption per half hour during peak daily usage hours for London homes when they are on time of use rate tariffs is about 0.07 kWh. Given that the mean value of the max daily energy usage over a half hour period in the control group was 0.85 kWh, a decrease of 0.07 is a roughly 8% reduction in energy consumption during peak hours which is a good amount.

# Visualization

Figures utilized to model daily energy consumption with the random forest algorithm were done through the open-source python IDE, Jupyter Notebooks. Within the notebooks we leveraged python libraries such as pandas and matplotlib to create the visualizations shown. The overall complexity to utilize these libraries is not very difficult as there is documentation throughout the internet on its use, so there are many examples available to leverage. The code to create these visualizations were placed in a github repository for version control so that we could expand upon the code utilized when needed.

One risk worth mentioning that we identified when utilizing python libraries was the sheer amount of computing resources required for the code developed to run. As our dataset is fairly large it is more difficult to create visualizations that accurately represent our data as it is. Adding on that, any code we ran had to process millions of records of data, making it difficult to create effective visualizations. In order to mitigate this risk, we focused on only attributes that we felt were relevant in predicting the data through our research. Doing so helped us process our algorithms faster and focus only important features within the algorithm. While the number of records was still large, this helped better visualize the factors in predicting the values for those records.

EDA and Time Series Analysis visualizations depended  heavily on several  libraries, ggplot, qplot, ggplot2, matplotlib and plotly to name a few, we used with Python and R.

Our analysis took place entirely in Jupyter notebooks using R for Exploratory Data Analysis and Python for Time Series modeling and forecasting.

During our analysis work we made sure we leveraged the power of SQL to apply data transformations, such as transpositions, partitioning with sorting, combinations of grouped aggregations and conditional summarizations to deliver the data set just as expected for a visualization that would provide clues to justify our observations or suggest otherwise.

We view the above strategy as risk averse because it passes the hard work to SQL instead of embedding into the visualization itself. Our setup allows us to execute any SQL logic in the RDS database that uses AWS resources leaving the easier task of plotting data points straight out of the SQL delivered resultset, to the locally installed plot library that would rely on scarcer local hardware resources.

The risks we encountered with visualizations per se had to do mostly with the complexity of the graph we wanted to produce and how to manipulate the universe of features of the plot library to render the graph. For example, Figures 25a and 25b are quite complex to calculate. In short, the Figures show how many times the Affluent, Comfortable and Adversity groups ranked first, second or third dropping their energy use, after they got notified of the electrical utility provider’s  daily tariff schedule. The values of energy decrease were calculated for each group on each of the 24 hours of the day across the timespan of the project.

SQL took care of the data heavy lifting but when it came to plotting the results, the risk got elevated when we decided to display it as a pair of a stacked bar and a clustered bar graph, diverting away from our original thought to deliver a table. After careful examination of ggplot’s commands we identified the right ones to meet our expectations, nevertheless, that was a risky moment during the visualization phase.

Throughout our efforts we used an assortment of plots depending upon the layout of the data we wanted to analyze.

With our R calculations we used scatter and line plots to observe how statistical variables change independently or in tandem. Bar and line charts revealed correlations among data variables, heat maps helped us identify weather variable dependencies and family of line and stacked bar charts helped us detect social group habits on energy utilization.

Visualization was particularly helpful in Python based analysis because it made the results of time series forecasting, that usually come in the form of matrices of coefficients and model accuracy statistics, more palatable to human intuition. Via plotting we were able to see how the time series components varied with time, assess the predictability of the final model and watch its recommendations how the future of energy consumption will play out.

# Findings

**Limitations of Our Analysis and Findings**

Due to issues involving the structure certain parts of our source data came in and the magnitude of it, some things that we wanted to investigate were not possible to calculate directly. One example is the average energy use for the 1-3 hour period every day which constitutes the peak daily usage hours for energy in London. As we explained in a paragraph directly below our table outlining the structure and describing the contents of the hhblock table in our dataset in section 2.2 of this report, the half hourly energy readings from each smart meter in our sample are arranged as columns because if they had been included as rows instead, the hhblock table would have been in the hundreds of millions of rows long instead of the roughly 3.5 million rows it came with. A table of this size would clearly not be feasible to work with, so while we were able to determine when peak daily energy use hours in London are from the two data visualizations of average energy usage by time of day in section 4, namely, from about 6 to 8 in the evening (right after most people are getting home from work), we could not run statistical tests on the differences in the distributions of the entire treatment and control groups for either the sum or the mean of those six half hourly readings for every day in calendar year 2013 (still over 12 million rows) because that would still be computationally infeasible or at least highly impractical due to the seven week deadline for this project using the equipment at our disposal.

As a result of our inability to calculate the reduction in energy usage for households in the treatment group during peak usage hours throughout 2013, in other words, the treatment effect, directly, we settled for a second best solution of instead estimating the reduction in the energy use during peak hours by using the decrease in maximum daily energy use values over a half hour period per from the daily\_dataset table. We justify this substitution because in order for the concept of peak daily energy use hours to have any concrete meaning, it must be the case that a majority of the maximum hours and half hours of energy usage in any given day fall within those hours. If they did not, then either those hours would not actually be the peak energy use hours on average or we would see a single hour or half hour period in Figures 48 or 49 when average energy use jumps up above the levels within the peak hours which we do not see.

**Exploratory Data Analysis**

The amount of EDA we managed to run against the data sets was not exhaustive, however, it was sufficient to draw some concrete conclusions about lifestyle practices of London residents when it comes to energy utilization.

At first, running the Shapiro test against every day's aggregates of smart meter readings and every smart meter’s 48 half-hour energy consumption data we concluded that most of the daily series of readings are not normally distributed. Checking again for normality on daily data, after summing the half hour energy consumption from all smart meters showed a slightly greater percentage of days with normally distributed readings at 95% significance level, and a way much higher fraction at 90% significance level.

This observation lends itself as a good indication that the daily energy consumption data is not normally distributed and needs to be re-examined for normality at a higher level of aggregation, perhaps monthly or annually, if we had a lot more years worth of data available. Only then, that is if the statistical test is statistically significant, will it be more accurate to start making conclusions about energy conservation based upon normality.

In the analysis section we plotted energy consumption vs temperature with and without outliers to prove the reversely directed (out of phase) correlation with temperature persists.

Then we explained why energy increases are observed only in the winter and not in the summer although humidity remains at higher levels throughout the year. The driving force of the constant airflow from the Atlantic is the main phenomenon responsible for cooling, that works concurrently with the humidity injection maintaining temperatures in the low 20Cs during the summer, rendering a regular ceiling fan sufficient to keep London residences at comfortable temperatures.

We also examined the energy consumption of the three most populous ACORN groups and we discovered it is layered in terms of magnitude with Affluents spending distinctively more than Comfortables who spend more than those suffering from Adversity, with the pattern of consumption (peaks and troughs) remaining the same across groups.

The same is true if we categorize the participant residences in ToU and Std tariff pricing.

But when we examined the difference in energy consumption between the ToU and Std subgroups within the same ACORN group, we observed that the variable time of use tariffs were adapted to quite well by Affluent residences. It dropped their consumption in a more cost-effective manner than Comfortables, who followed suit, and did much better than the Adversity group that lagged behind significantly, to our surprise.

The same was also discerned when we explored the drops in energy consumption for each ACORN group during the times when the time of use tariffs were in the Low, Normal and High ranges, after each group was informed what the tariff pricing would be during the day.

Affluents dropped their energy consumption about 1.5 times more than the Comfortables in all tariff ranges, followed by a negligible change of the Adversity group participants.The greatest drop for all groups was during the times tariff was in Normal charging rate.

**Time Series Analysis and Forecasting**

After getting a clearer view of the data from EDA we engaged the methods of Holt-Winters (Triple) Exponential Smoothing (HWES) and Seasonal Autoregressive Integrated Moving Average (SARIMA) to forecast future average energy utilization. Both techniques are suitable for univariate data and they add support for trend and seasonality.

We used both methods with monthly and daily data, with success on both data sets only with HWES. SARIMA failed to complete running against the daily data set due to the lack of local hardware resources.

Starting with monthly data, we presented all monthly average values except the most recent 6 months to build-train/test the models. Both methods provided a clear trend-seasonality decomposition of the energy consumption time series with accurate predictions of energy values to the second decimal point during testing.

SARIMA proved to be more rewarding by providing a forecasting band of low and high forecasts that widened as the forecasting horizon continued into the future. SARIMA was also more informative in its results matrix where many statistics were offered to prove the residual errors of the model obeyed the assumption of normality.

Recognizing that the entire data set spanned only 2.3 years, the model outcome was decent however, a data set covering a greater time period is highly recommended for good forecasting results. From the current model’s success we would expect that having access to more years of data to build and train our model with and to forecast a monthly data model would add to accuracy and could add the ability to make longer term forecasts.

With daily data utilized we had more data points to work with, however, considering the lag of seasonality increases to 365 we would still need more years for a more accurate model and trustworthy forecast.

HWES discovered no trend, a lot of seasonality hidden in the data with noticeably large residual errors. Training resulted in a HWES model with a good fit but only for short scoped predictions. Its forecasting daily average energy consumption value was off up to 25% close to the end of the 6-month testing time horizon, thus weakening our trust in the model’s ability to make extended time horizon forecasts.

**Distribution and Short-term Projections of Daily Energy Consumption**

To reinforce a particularly interesting and important finding that was discovered was that the daily energy consumption data did not follow a normal distribution. As seen in our analysis, the daily data consumption for these households were positively skewed as we believe that the higher range of daily consumption was more one-offs rather than their typical daily usage. For example, a household may have kept some of their electronics on longer for a particular day which would lead to a higher energy consumption. These are days we can consider as atypical usage and deviate from the expected energy consumption for households on a daily basis.

These atypical days of consumption are more difficult to predict because they are often due to extraneous factors that occur at random. We can see that in our model for predicting daily energy consumption it performs well at the typical consumption level where most of the data occurs and finds it harder to predict those higher atypical values. Doing more research in understanding what extraneous factors cause these higher energy consumption rates may provide better context into our analysis and understanding if and how we can predict atypical consumption from households.

Another thing to note is the lack of available computing resources limited our ability to utilize more factors to predict daily consumption. As adding any additional column to our algorithm would result in adding a vast amount of records we had to limit the number of columns we believed were relevant to our analysis. With additional computing resources we would be able to handle the volume of data in a more timely manner and be able to focus more time on our analysis rather than the processing of the data.

Overall, we determined that predicting data at lower consumption rates was more accurate than the higher consumption rates as the majority of data falls within the lower ranges. As higher consumption rates are usually caused by idiosyncratic factors, we would look to find a variable that would better signal these higher consumption rates.

**Effects of the Random Introduction of Variable Time of Use Prices per kWh for Residential Energy Consumption in London, Including a Steep Premium During Peak Usage Hours**

In section 3.3, we found that according to a one-tailed **Mann-Whitney U** test for a difference in the distribution between two samples, there was a median of the difference in the total daily energy consumption between a sample from the treatment group and a sample from the control group of about 0.45 kWh, or 5.7%. And because the data on households and their energy consumption patterns we had in our dataset was panel data (aka longitudinal), we also ran a **random effects regression** model on energy\_sum against a key independent variable of ToU\_dummy. The results we got from running that RE regression were a mean decrease in total daily usage of 1.13 kWh, or 11%. That amounts to an average reduction of about 34 kWh in the mean of energy consumed per London household per month and annually, it amounts to an average reduction of around 412.5 kWh per household. According to Statista, there were around 3.4 million dwelling units in London as of 2019. Therefore, if time of use pricing for energy were implemented city wide and our results generalize to the rest of London, there would be a decrease of something on the order of 1.4 billion kWhs in London annually!

Furthermore, that 11% reduction in average daily energy use is only the case when English utility providers change their energy prices from a flat rate of 14.3 p/kWh for residential energy consumption during all hours to 67.2 pence per kWh during peak usage hours, 11.8 pence per kWh during normal usage hours, and 3.99 pence per kWh during low usage hours. The law of demand in economics states that the higher the price charged for a good or service, the lower the quantity demanded will be for that good or service. Thence, it is clear that a variable time of use energy pricing scheme of, for instance, 75 p/kWh during peak usage hours, 20 p/kWh during normal hours, and 7 p/kWh during low usage hours would result in an even greater reduction in residential energy consumption among households than 11% and an even higher array of rates per kWh would result in reductions in total household energy consumption which are larger still.

For peak daily use hours, the results of our second one-tailed **Mann-Whitney U** test indicated a median of the difference in energy use between a sample from the treatment group and a sample from the control group of 0.029 kWh, or around 4.4%. And again, because we had the privilege of access to panel data (data on a fixed set multiple entities over multiple time consecutive periods, so that if the fixed set of entities are a specific group of people, panel data follows them over time) which enables us to use powerful techniques from panel data econometrics, we also ran a **random effects regression** specification with energy\_max as the dependent variable in the model and a time of use dummy variable as the key independent variable whose effects on the dependent variable we were estimating. The results of this second RE regression indicated that the decrease in the mean of energy consumption during half hour periods within peak daily use hours for households on variable time of use pricing (while holding all other factors which impact energy use constant) is 0.07 kWh, or around 8%.

If that rough estimate of an 8% decrease in energy consumption per half hour during peak use hours is an externally valid finding, that is to say, if it will continue to hold if variable time of use tariffs at the rates imposed in this study (a high tariff of 67.20 pence/kWh during peak hours, a normal tariff of 11.76 pence/kWh during most of the day, and a low tariff of 3.99 pence/kWh) were extended to the rest of the households in London, there would be a corresponding 8% decrease in the strain on the city’s power grid during peak use hours. That means that some of the power generators for London which emit lots of greenhouse gases and are currently only kept around as backups which are kept running on idle or are turned off completely until they are needed which only happens occasionally and only during peak daily usage hours, could be decommissioned to prevent the emission of more GHG into the atmosphere.

# Summary

The first most evident conclusion from our analysis was that London residences' energy utilization is coupled with outdoor temperature more than humidity that doesn’t change as much throughout the year. This finding is hardly a discovery and quite common to all of us, however, what makes it important in this study is how well synchronized is London’s electricity demand to the particular physical variable, i.e. temperature, that has the strongest statistically significant impact on energy.

We extended the above realization with the development of short-term energy-temperature forecasting models of acceptable accuracy. Despite a possible requirement to confirm it with data from more years, based on the temperatures profiles of the season and our analysis results city officials and energy providers alike, could still predict more accurately the demand of electrical load and, in turn, plan its efficient distribution across the city of London, balancing residential areas and, let’s say downtown business districts that demand lower energy after hours.

Another analysis finding that complicates the implementation of the above objective is that London residents do not respond with the same eagerness to energy cost-reduction motives even with the benefit of dynamic energy pricing.

The Adversity group demand for energy hardly reacted to any of the tariff ranges, Low, Normal or High. The Adversity group amounts to above 41% of the general population of London according to the listing in section 2.2.1, so, needless to say that laying down a Government plan to manage the group’s energy needs will have a great impact in the city’s overall demand.

For example, roof solutions like thin-high efficiency solar cells that don’t need a bright sun for many hours of the day can still accumulate the necessary amount of energy to provide hot water for part of the day in Adversity areas.

It’s also evident that the Adversity group needs to be more educated before another motivational boost with the proper marketing/tax cutting techniques to start responding to energy price inducements with the same passion as the other two groups.

In spite of the greater response of the Affluent group to electricity cost incentives, their daily average is still the highest among the Affluent, Comfortable and Adversity groups and, because of that they are still good candidates for energy reduction.

Affluents are about one third of London’s consumers [see section 2.2.1] and those who most likely own the bigger residences in London, hence, they are excellent contenders to afford the financial outlay of alternative energy resources, such as solar energy home battery systems, with a short term payback period, or reside in sustainable architecture constructions.

# Future Work

## Adding More Data

The main issue we faced during our analysis was that there was not enough data from enough different places to make our work externally valid, i.e. generalizable. In order to truly see the general effects of weather and time on energy consumption, there would need to be a wider variety of data sources to work with. In order to achieve this, we recommend that the scale of the smart meter project expands to a worldwide platform. We would also like to see a replication in dozens or even hundreds of different cities in developed nations with advanced economies around the world of the experiment we analyzed here of randomly introducing variable time of use energy pricing to some residents, while keeping whatever the current local energy pricing scheme is the same for others, then studying the differences in energy consumption patterns between these “treatment” and “control” groups afterwards. Preferably, some of these follow up experiments of randomly introducing variable time of use energy pricing could try surcharges other than the peak usage hours rate of 67.2 pence per kWh used in London during their peak daily energy usage hours in order to see if imposing higher surcharges could lead to an even greater reduction in residential energy consumption during those times of the day when the power grid is most strained and how much greater.

This would help because it would then cover people with different cultures, living in different climates, earn different average and median incomes, and have different average levels of wealth which could all lead to a wide variety of different patterns of energy consumption which in turn could lead to differential price elasticities of demand for household energy. Informed considerations of factors such as these could cause some places to perhaps require steeper peak hours energy usage premiums than others to cause a large enough decrease in energy use to make a dent in GHG emissions.

Another way that more data could be added is searching for similar datasets which cover much longer stretches of time, such as 5-10 years or even more if they are out there to be found. This could help us do a broader analysis on the weather and with it we would be able to see the shorter-to-medium effects of climate change (on a geological time scale) within the weather data. We would also be able to more clearly see how energy consumption is changing overtime. This data would be valuable in identifying more variables affecting energy consumption so we would be able to give better advice on how to incentivize and enable more energy efficient consumer behavior.

## Possible Real Time Advice for Households

Another possibility for future work could be giving real time advice to households. With constant incoming data from the smart meters, an algorithm could recommend adjustments on how to more efficiently manage their energy use. There is evidence this works for the conservation of energy as well. In 2004, The Hydro One Pilot was a study that tested the effects of real time feedback on residential energy consumption over a 2.5 year period (Hydro One). Over 400 people (including the control group) in Canada participated in this study where they tested the use of Hydro One PowerCost Monitor on energy consumption in these households. This monitor showed the energy consumption of the household and cost in real time (Hydro One). Not only did more efficient energy consumption take place, but that effect was persistent throughout the 2.5 year period (Hydro One) of the study. This shows that giving people access to relevant information about their energy consumption, positive change can take place.

This can even be taken a step farther if we provide them with more specific data and analysis of the breakdown of their energy consumption in terms of where it is coming from within the house (i.e. refrigerator, showers, lights). There are many things people know they could be doing to save energy and money, but if they could see their day to day changes and receive sound advice based on that, then it could serve as a constant reminder to actually implement those energy saving strategies. Even having a more specific energy document than what companies give you (just how much energy a household uses and how much money they owe for that usage), for instance, a customized feedback message that could tell people where and when they are using too much energy or where energy consumption has changed month to month and provide recommendations on how to improve it.

This recommendation system could be a seperate business, or the government’s way to encourage ways to reduce unnecessary energy consumption. There are many alternative appliances that are relatively inexpensive and much more energy efficient, and there could be a way to show how people’s energy usage and as a result, their monthly energy bills would change after replacing some of their old, inefficient appliances. Doing daily analysis in real time could show households how they compare to other houses around them. If it is possible to distribute smart metering devices around the world, then this advice can also be distributed to inform and encourage greener households across the globe.

## Expansion Beyond Households

The city of London could split the cost of upgrading apartment buildings with their apartment owners to increase their environmentally friendly operation, by painting white the building roofs and insulating those that do not face the sun as much during the day. A before and after comparison of energy use by the tenants on a couple of test projects will give the city an idea how more energy efficient it can become by expanding the project to the entire city.

Gardens or thin film solar cells can also be placed around an edifice in areas with greater sunshine throughout the year. The gardens can be either a form of landscaping for energy-efficient houses to help the city decrease its carbon-emissions or a place for tenants to grow natural products for their own consumption. In the case of solar cells the energy collected can be stored in batteries to light common hallways inside the building during the night with motion detector regulated lights to turn on, only if there’s walking in and out building traffic.

Again, the city can collect energy usage information before and after the revamping project and pass along the results to the owners to inform them when any investment amount they made together with the city for building upgrades paid them back in the form of energy savings and how much they’ll be automatically saving from the point on.

# Appendix A: Code References

For more information, code, and supporting data please visit the Github repository linked below.

<https://github.com/LnxRls/CS504_Project>

SARIMA’s implementation against the daily data set sets off the error recorded as shown in the screenshot below. That hardware limitation prevented us from continuing further with forecasting using the SARIMA method.

Text

Description automatically generated

*The following paragraphs were originally included in the sub section of section 3.1 labelled “Analysis of the Weather Time Series Data and the Correlation of Daily Weather Statistics to the Daily Average Energy Consumption” but were moved to the appendix following feedback from the professor as to the inability of a 28 month long period of data collection to be able to say anything about climate change or anthropogenic global warming.*

This dataset only spanned over the course of two years, so we did not necessarily expect any strong 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 in means. The Null hypothesis states if there is a unit root present then the maximum temperature is not a stationary process, 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 a non-stationary phenomenon over this time period of 28 months. For pressure however, we found the Dickey-Fuller test statistic 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, so we can assume that pressure is stationary. These results show that even over relatively short periods of time on a geologic scale, the max daily temperature is not in a steady state. This was not surprising even though we did not necessarily expect to observe the effects of anthropogenic global warming over such a short time period of 28 months. Our Dicky-Fuller test of max temperatures was a simple way of confirming our priors of the reality of human caused or enhanced climate change empirically rather than just assuming are priors are correct and going from there. And because we could not reject the null hypothesis that the max temperatures are non-stationary, we interpret that as meaning that the climate is indeed changing.

*The following estimates and explanations were originally in section 3.3, but were moved to the appendix due to difficulties in their interpretation.*

**Wilcoxon Effect Size estimates for effect of time of use tariffs on energy\_sum**

In order to more formally estimate the magnitude of the effect size for the difference in total daily energy consumption between the treatment group and control group, we calculated the Wilcoxon Effect Size in RStudio as well. The results of this test are included below:



This result, an effect size, r, of 0.035, is generally seen as a small effect in the statistical research literature because it is below 0.3.

**Wilcoxon Effect Size estimates for effect of time of use tariffs on energy\_max**

As we did with the total daily energy consumption, we ran a function in RStudio to calculate the Wilcoxon Effect Size for the maximum daily energy consumption. The results are included below:



An effect size, r, of 0.024 is even smaller than the previous effect size for the difference in total daily energy consumption, thus, this result too is considered “small”.

# 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.

Chart, treemap chart

Description automatically generated

# Appendix C: Definiton 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 | A variable that 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 Smart 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 (Compare the Market, 2020). |
| Temperature Min & Max | Extreme observed temperatures during the last 24 hours between 7pm CST - 8amCST per USA’s National Weather Service [NWS] definition (Local observed maximum and minimum images, 2015). 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 | It 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. |

# 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
30. Hydro One. (2006). “The Impact of Real-Time Feedback on Residential Electricity Consumption : The Hydro One Pilot.” (2006). Retrieved 30 April 2021 from https://www.reducemyenergy.com/PDF/Summary%20Results%20Hydro%20One%20Pilot%20-%20Real-Time%20Feedback.pdf