How to Use Python and Mathematical Modeling to Better Understand the Impact of Electricity Pricing on Consumption

Saba Nejad

November 2, 2023

PyData NYC 2023

Hi! I'm Saba. Thank you for coming to my talk.

My name is Saba. I'm a Data Engineer at Point72. Today, I'll be walking you through a portion of my master's thesis.

Goals for this talk

- The importance of knowing your data before starting your analysis.
  - how it was collected
  - its biases
  - dimensionality, fill-rate, etc.

Goals for this talk

- The importance of knowing your data before starting your analysis.
  - how it was collected
  - its biases
  - dimensionality, fill-rate, etc.
- The importance of starting with a mathematical model or framework, rooted in the real world, before trying to fit a model to your data.

What are we going to talk about?

- a data modeling problem

What are we going to talk about?

- a data modeling problem
- the data from a trial in London that implemented a particular pricing model on a subset of houses to quantify the impact of said pricing model on electricity consumption.

Overview of the Talk

- Introduce the Problem Space; Define Terminology

Overview of the Talk

- Introduce the Problem Space; Define Terminology
- Introduce the Trial and the Dataset

Overview of the Talk

- Introduce the Problem Space; Define Terminology
- Introduce the Trial and the Dataset
- The Mathematical Model

Overview of the Talk

- Introduce the Problem Space; Define Terminology
- Introduce the Trial and the Dataset
- The Mathematical Model
- Data Prep, Cleaning, Processing, and Analysis

Overview of the Talk

- Introduce the Problem Space; Define Terminology
- Introduce the Trial and the Dataset
- The Mathematical Model
- Data Prep, Cleaning, Processing, and Analysis
- Results, Conclusion, Takeaways

Useful Details about the Problem Space

In electicity, there is a supply and demand constaint.

- Electricity is unique in that its **storage is prohibitively costly**.

In electicity, there is a supply and demand constaint.

- Electricity is unique in that its **storage is prohibitively costly**.
- As a result, **supply must meet demand** at all times.

In electicity, there is a supply and demand constaint.

- Electricity is unique in that its **storage is prohibitively costly**.
- As a result, **supply must meet demand** at all times.
- If demand exceeds supply, there will be a power outage.

In electicity, there is a supply and demand constaint.

- Electricity is unique in that its **storage is prohibitively costly**.
- As a result, **supply must meet demand** at all times.
- If demand exceeds supply, there will be a power outage.
- Power outages are extremely costly and we try to prevent them as much as possible.

Demand Response is a tool that helps lower demand.

- This unique feature of electricity, poses a **challenge**.

Demand Response is a tool that helps lower demand.

- This unique feature of electricity, poses a **challenge**.
- Imagine an unpredictably hot day where there is more demand than expected.

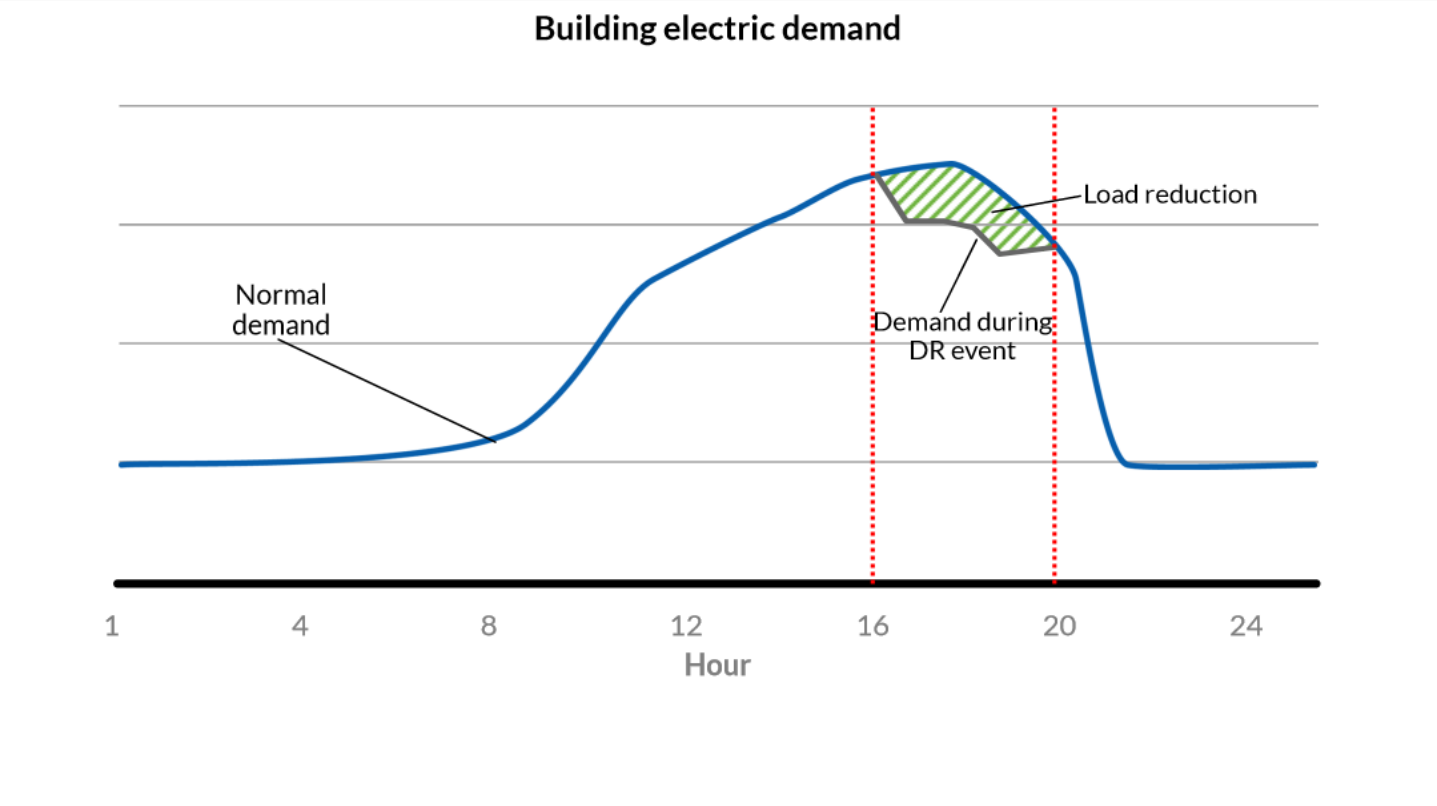
Demand Response is a tool that helps lower demand.

- This unique feature of electricity, poses a **challenge**.
- Imagine an unpredictably hot day where there is more demand than expected.
- To prevent a power outage, we need close to real-time methods to lower demand.

Demand Response is a tool that helps lower demand.

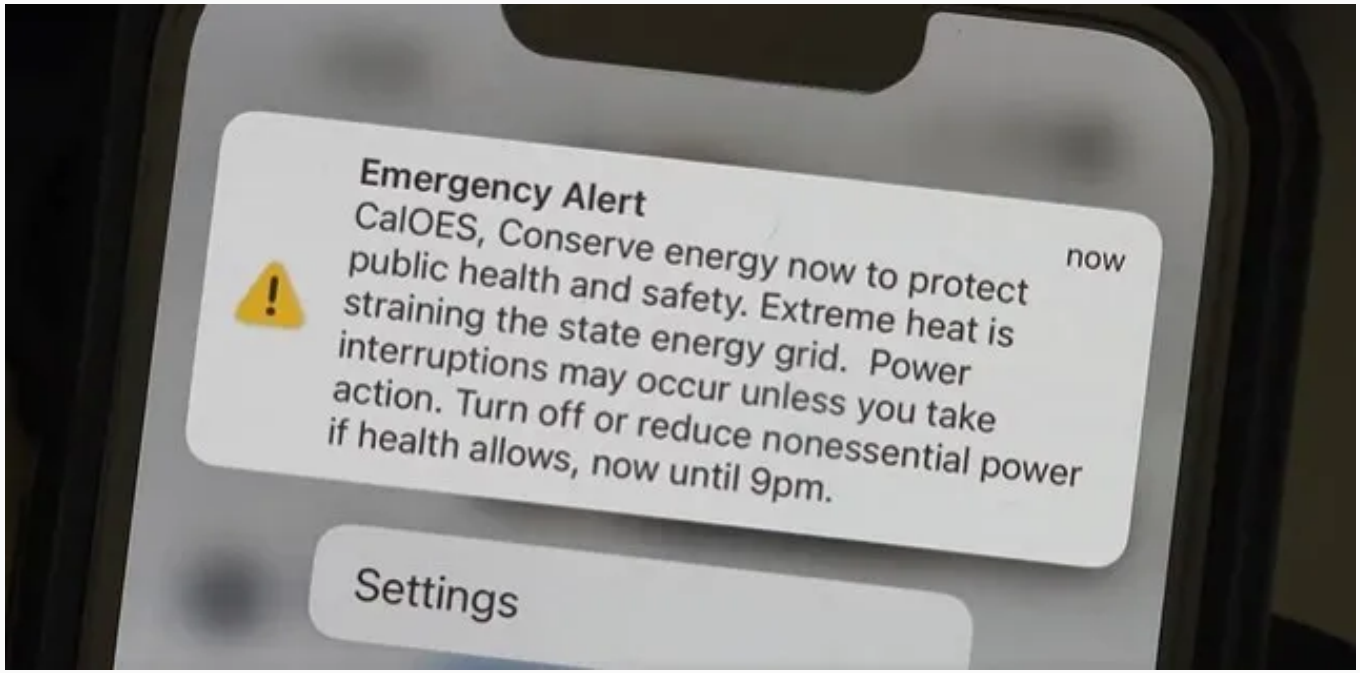
- This unique feature of electricity, poses a **challenge**.
- Imagine an unpredictably hot day where there is more demand than expected.
- To prevent a power outage, we need close to real-time methods to lower demand.
- **Demand Response** is one such method.

Demand Response visual example



One demand response method are blast texts.

- Assumption: Demand can be shifted around.



Another demand response method is using pricing models.

These pricing models **increase** price, to incentivize people to

- Assumption: Consumers are price sensitive i.e. they will respond to price incentives.

Another demand response method is using pricing models.

These pricing models **increase** price, to incentivize people to

- Assumption: Consumers are price sensitive i.e. they will respond to price incentives.
- Two pricing models:

Another demand response method is using pricing models.

These pricing models **increase** price, to incentivize people to

- Assumption: Consumers are price sensitive i.e. they will respond to price incentives.
- Two pricing models:
  - Time of Use Pricing (ToU)

Another demand response method is using pricing models.

These pricing models **increase** price, to incentivize people to

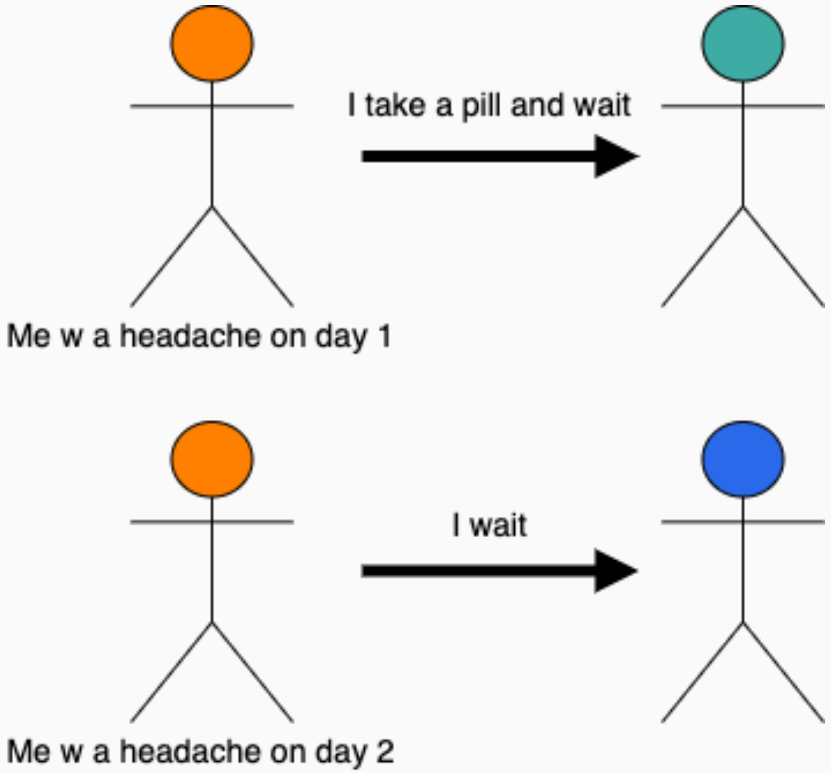
- Assumption: Consumers are price sensitive i.e. they will respond to price incentives.
- Two pricing models:
  - Time of Use Pricing (ToU)
  - Dynamic Time of Use Pricing (dToU)

Problem Statement

Our goal is to understand if **Dynamic Time of Use Pricing** was effective in lowering consumption **in time of high demand**.

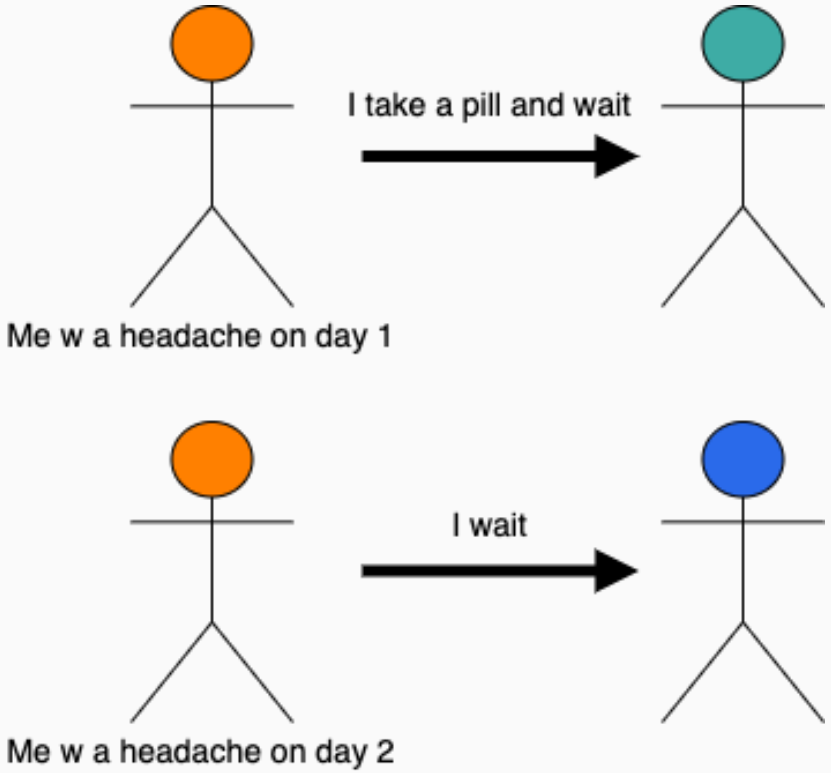
Understanding causality and some useful terminology

- What is a treatment? What is the treatment effect?



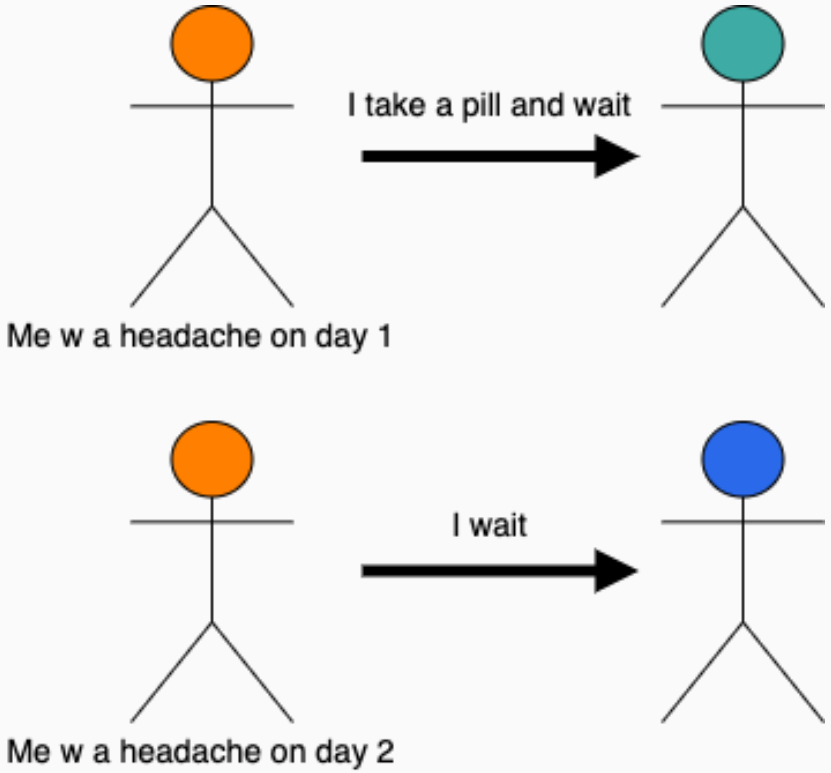
Understanding causality and some useful terminology

- What is a treatment? What is the treatment effect?
- What is a 'unit' undergoing treatment?



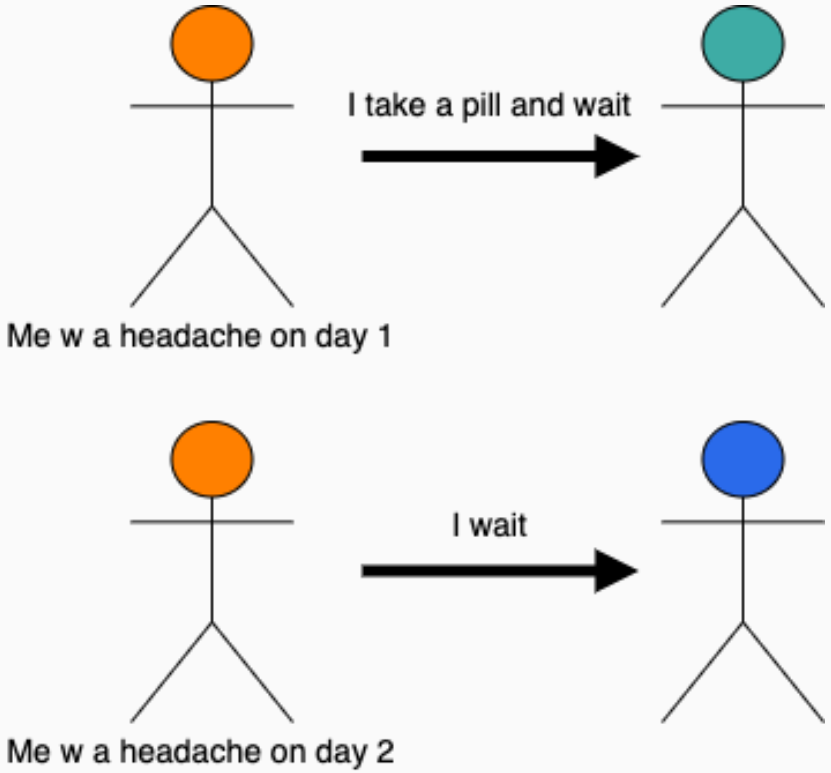
Understanding causality and some useful terminology

- What is a treatment? What is the treatment effect?
- What is a 'unit' undergoing treatment?
- What does it mean for something to have a causal effect on something else?



Understanding causality and some useful terminology

- What is a treatment? What is the treatment effect?
- What is a 'unit' undergoing treatment?
- What does it mean for something to have a causal effect on something else?
- What is the counterfactual?



The Fundamental Problem of Causal Inference (FPCI)

- Assume $X \in \{0, 1\}$ is a binary causal variable and Y is a response variable (which may be continuous).

The Fundamental Problem of Causal Inference (FPCI)

- Assume $X \in \{0, 1\}$ is a binary causal variable and Y is a response variable (which may be continuous).
- Assume X has a causal effect on Y.
  - $Y_0 :=$ the value of Y when $X = 0$ (untreated unit)
  - $Y_1 :=$ the value of Y when $X = 1$ (treated unit)
  - $T = Y_1 - Y_0 :=$ the treatment effect

The Fundamental Problem of Causal Inference (FPCI)

- Assume $X \in \{0, 1\}$ is a binary causal variable and Y is a response variable (which may be continuous).
- Assume X has a causal effect on Y.
  - $Y_0 :=$ the value of Y when $X = 0$ (untreated unit)
  - $Y_1 :=$ the value of Y when $X = 1$ (treated unit)
  - $T = Y_1 - Y_0 :=$ the treatment effect
- Y_0 and Y_1 are **counterfactuals** of one another. We observe **only** Y_1 or Y_0, but not both at the same time.

Solutions Around FPCI

1. Temporal Stability & Causal Transience

Solutions Around FPCI

1. Temporal Stability & Causal Transience
2. Unit Homogeneity

Solutions Around FPCI

1. Temporal Stability & Causal Transience
2. Unit Homogeneity
3. Estimate causal effects for populations rather than units

Solutions Around FPCI for this Trial

1. Temporal Stability & Causal Transience X
2. Unit Homogeneity X
3. Estimate causal effects for populations rather than units ✓

Random samples of populations are studied to estimate treatment effects.

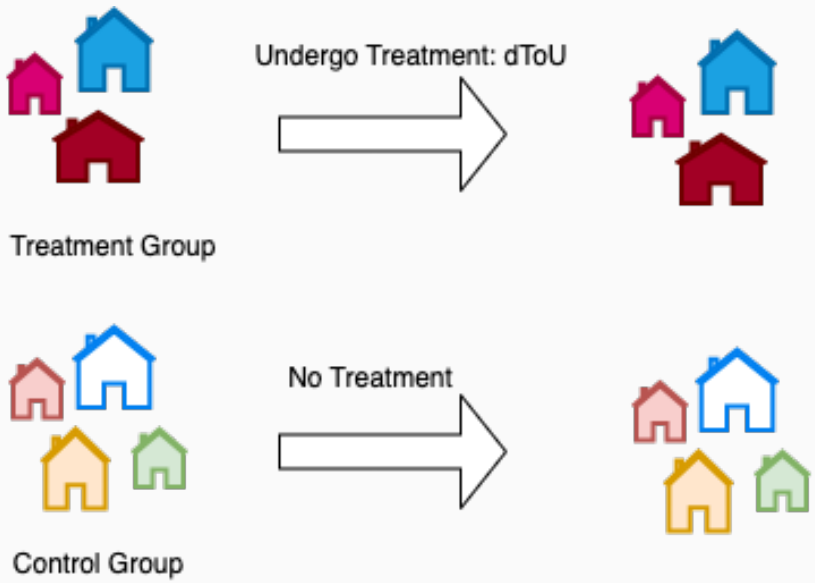
- It's difficult to estimate the treatment effect for a unit.
  $\implies$ FPCI makes it **necessary** to study a **population vs a unit**. The **larger the population**, the easier to robustly estimate treatment effects.

Random samples of populations are studied to estimate treatment effects.

- It's difficult to estimate the treatment effect for a unit.
  $\Longrightarrow$ FPCI makes it **necessary** to study a **population vs a unit**. The **larger the population**, the easier to robustly estimate treatment effects.
- When estimating treatment effect for populations, to remove bias terms, we want groups that behave similarly out of sample.
  $\Longrightarrow$ analyses are done on **random samples** of populations.

The Details of the Trial Studied

High Level Overview of the Trial



Low Carbon London Smart Meter Trial

- Motivation: The Climate Change Act of 2008 sets the target of reducing carbon emissions to 20% of 1990 levels by 2050.

Low Carbon London Smart Meter Trial

- Motivation: The Climate Change Act of 2008 sets the target of reducing carbon emissions to 20% of 1990 levels by 2050.
- Data spans November 2011 and February 2014.

Low Carbon London Smart Meter Trial

- Motivation: The Climate Change Act of 2008 sets the target of reducing carbon emissions to 20% of 1990 levels by 2050.
- Data spans November 2011 and February 2014.
- Treatment took place in the calendar year 2013.

Low Carbon London Smart Meter Trial

- Motivation: The Climate Change Act of 2008 sets the target of reducing carbon emissions to 20% of 1990 levels by 2050.
- Data spans November 2011 and February 2014.
- Treatment took place in the calendar year 2013.
- Readings were taken at half hourly intervals, for the entire duration of the trial, and for 5600 houses in total (around 167M records).

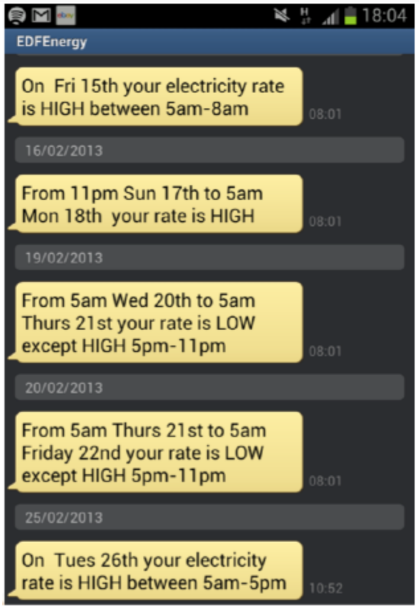
dToU: pre-set prices, implemented at different times of day

- dToU price bands vs static pricing model:

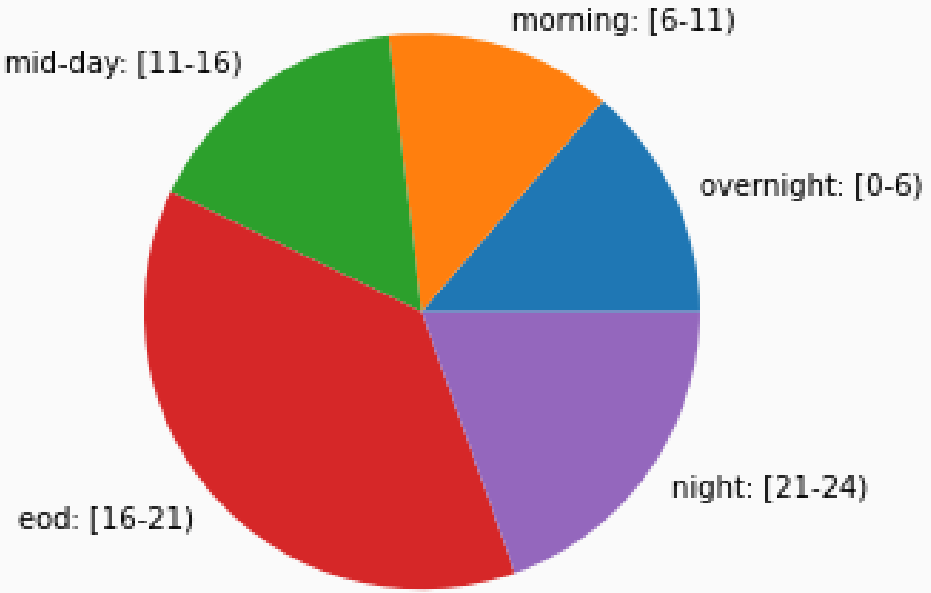
dToU Pricing Model	Static Pricing Model
High (67.20 p/kWh)	14.228 p/kWh
Normal (11.76 p/kWh)	14.228 p/kWh
Low (3.99 p/kWh)	14.228 p/kWh

dToU: pre-set prices, implemented at different times of day

- "Dynamic": Times of day when households would be subject to high price point was communicated a day ahead via the Smart Meter In Home Display or text message.



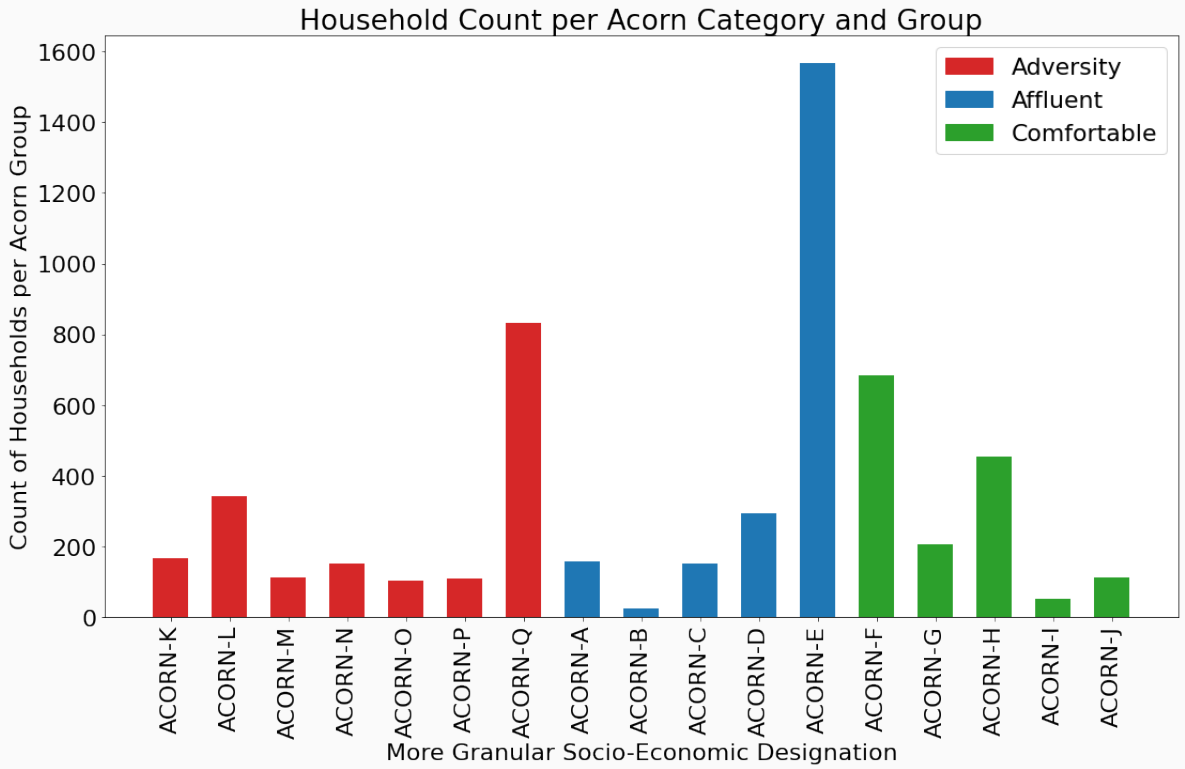
Distribution of High Price Periods



Features: Two Levels of Discrete Socio-Economic Categories

Label	Acorn group	Acorn category
A	Wealthy executives	Wealthy achievers
B	Affluent greys	Wealthy achievers
C	Flourishing families	Wealthy achievers
D	Prosperous professionals	Urban prosperity
E	Educated urbanites	Urban prosperity
F	Aspiring singles	Urban prosperity
G	Starting out	Comfortably off
H	Secure families	Comfortably off
I	Settled suburbia	Comfortably off
J	Prudent pensioners	Comfortably off
K	Asian communities	Moderate means
L	Post industrial families	Moderate means
M	Blue collar roots	Moderate means
N	Struggling families	Hard pressed
O	Burdened singles	Hard pressed
P	High rise hardship	Hard pressed
Q	Inner city adversity	Hard pressed

Features: Two Levels of Discrete Socio-Economic Categories



Biases I: Treatment group opted in.

The trial is double opt-in:

1. Households opt into sharing their data with the trial at all.

Biases I: Treatment group opted in.

The trial is double opt-in:

1. Households opt into sharing their data with the trial at all.
2. Some opted into undergoing dToU pricing in 2013.

Biases II: There were cash incentives for the treatment group.

The treatment group was given some incentives:

- A guarantee that they will be reimbursed at the end of trial if they are worse off on the dToU tariff than they would have been on their previous tariff.
- Assurances regarding how many hours would be charged at the high price band.
- £100 for signing up to the dToU tariff.
- Another £50 for staying on the dToU tariff until the end of trial.
- Entry into a prize draw after completion of the post trial survey.

The two groups were not biased geographically

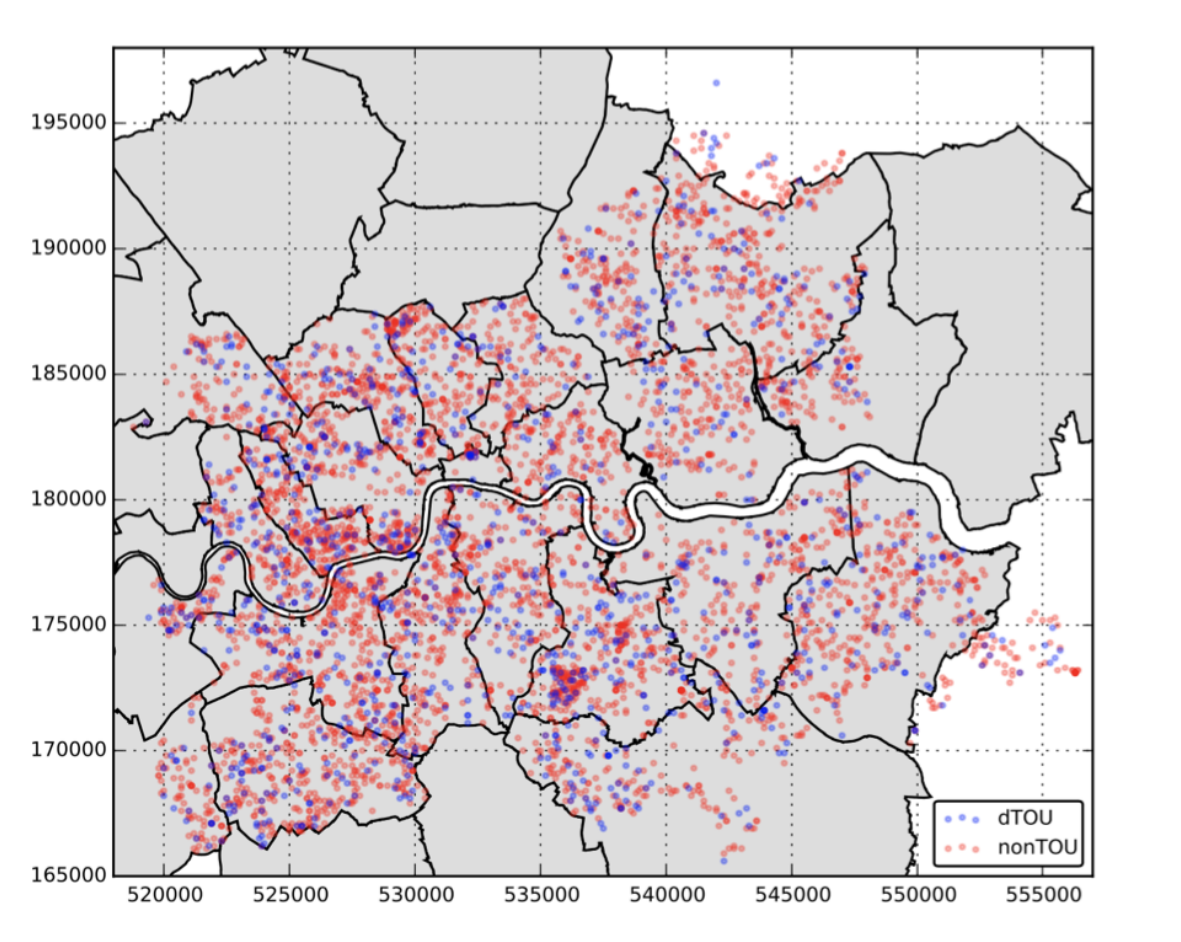


Figure 1: Trial household sample locations overlaid on the borough boundary map of Greater London. This shows that the treatment and control group were representative samples of Greater London.

The two groups were not biased socio-economically

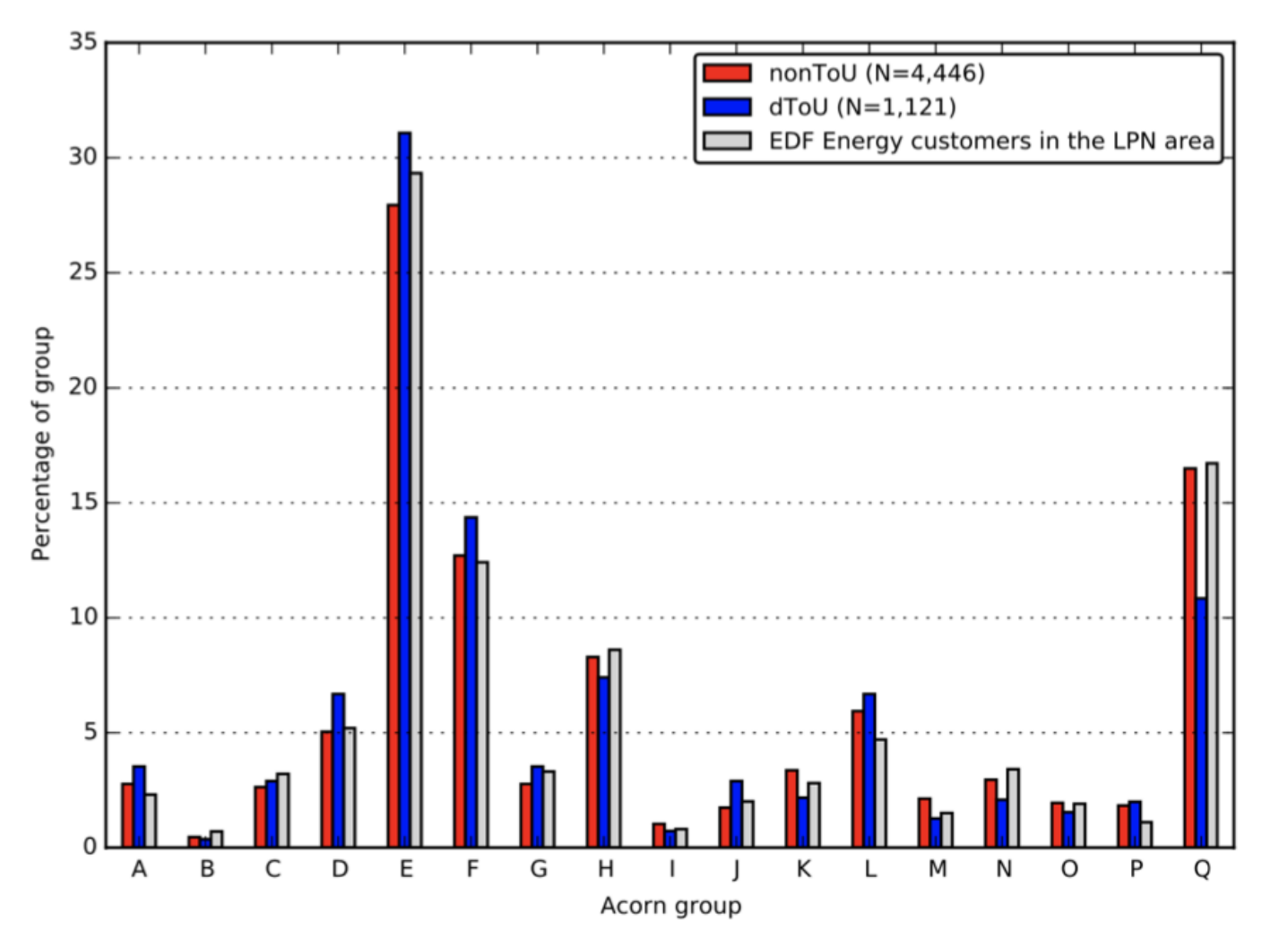
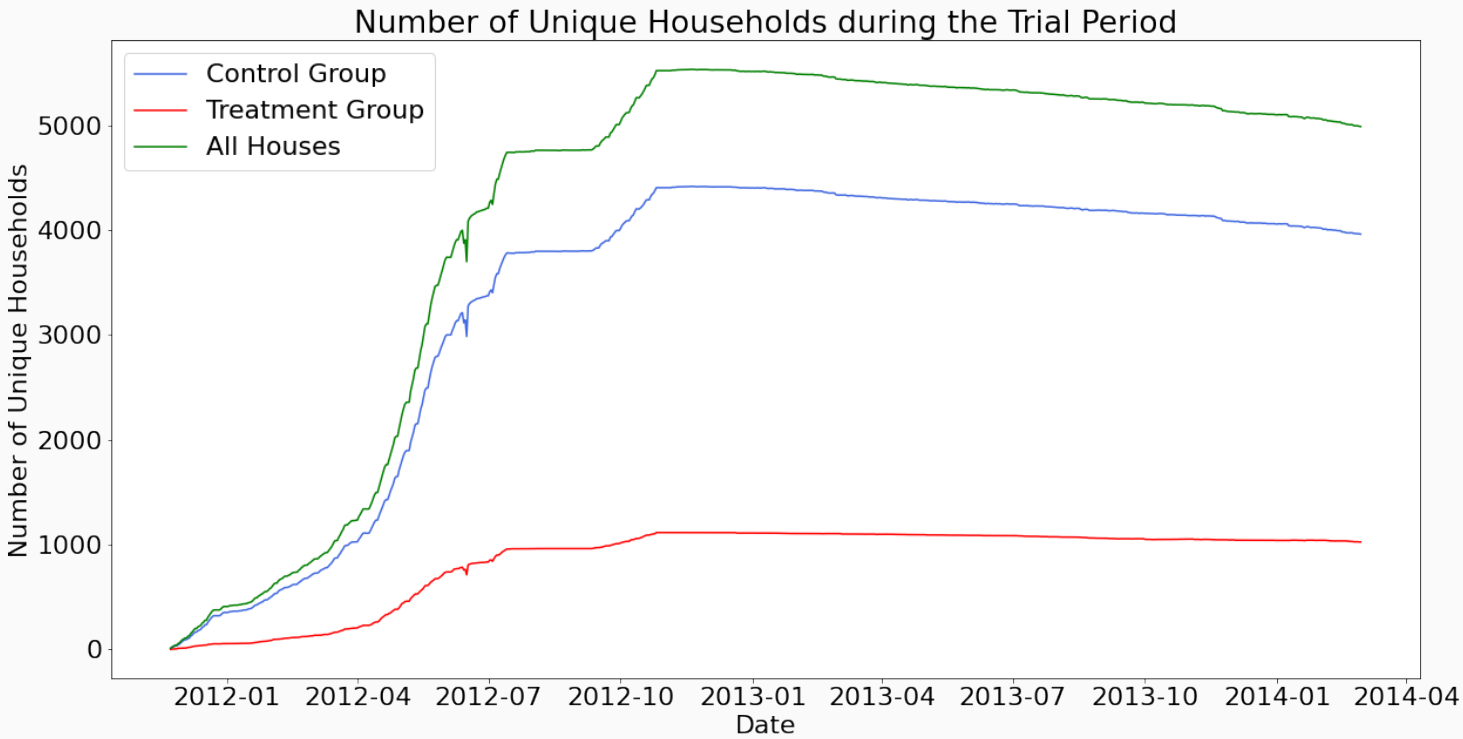


Figure 2: The treatment and control groups were also representative socio-economic samples of Greater London.

Onboarding households creates a missing data issue.

- 2012: Users were still onboarding.
- 2013: Some users dropped out of both the treatment and the control groups
- ~5,600 total households participated: ~4,500 were in the control group, ~1,100 were in the treatment group.



Treatment vs Control Group Baselines

- Though our groups are **geographically and socio-economically balanced**, one group opted into dToU pricing and was offered some incentives.

Treatment vs Control Group Baselines

- Though our groups are **geographically and socio-economically balanced**, one group opted into dToU pricing and was offered some incentives.
- What are some reasons the baselines might differ given what we know?

Treatment vs Control Group Baselines

- Though our groups are **geographically and socio-economically balanced**, one group opted into dToU pricing and was offered some incentives.
- What are some reasons the baselines might differ given what we know?
    - price sensitive group opted in

Treatment vs Control Group Baselines

- Though our groups are **geographically and socio-economically balanced**, one group opted into dToU pricing and was offered some incentives.
- What are some reasons the baselines might differ given what we know?
    - price sensitive group opted in
    - low consuming group opted in

Treatment vs Control Group Baselines

- Though our groups are **geographically and socio-economically balanced**, one group opted into dToU pricing and was offered some incentives.
- What are some reasons the baselines might differ given what we know?
  - price sensitive group opted in
  - low consuming group opted in
  - those households are flexible time-wise: particular jobs, hours, etc.

Let's formulate the mathematical model that will inform our analysis.

Let's establish some symbology standards

Consider the following segmentation of the data.

- $\alpha_y :=$ control group's consumption matrix during year y
- $\beta_y :=$ treatment group's consumption matrix during year y

What are these matrics and what is in them?

$$\alpha_y = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_c} \\ a_{21} & a_{22} & \cdots & a_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tn_c} \end{bmatrix} \quad \beta_y = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n_t} \\ b_{21} & b_{22} & \cdots & b_{2n_t} \\ \vdots & \vdots & \ddots & \vdots \\ b_{t1} & b_{t2} & \cdots & b_{tn_t} \end{bmatrix}$$

α matrics are of size $\alpha_{t \times n_c}$; β matrics are of size $\beta_{t \times n_t}$.

What are these matrics and what is in them?

$$\alpha_y = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_c} \\ a_{21} & a_{22} & \cdots & a_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tn_c} \end{bmatrix} \quad \beta_y = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n_t} \\ b_{21} & b_{22} & \cdots & b_{2n_t} \\ \vdots & \vdots & \ddots & \vdots \\ b_{t1} & b_{t2} & \cdots & b_{tn_t} \end{bmatrix}$$

α matrics are of size $\alpha_{t \times n_c}$; β matrics are of size $\beta_{t \times n_t}$.

t is the number of half-hour measurements in a year ($t = 365 \times 48 = 17,520$ for 2012 and 2013).

What are these matrics and what is in them?

$$\alpha_y = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_c} \\ a_{21} & a_{22} & \cdots & a_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tn_c} \end{bmatrix} \quad \beta_y = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n_t} \\ b_{21} & b_{22} & \cdots & b_{2n_t} \\ \vdots & \vdots & \ddots & \vdots \\ b_{t1} & b_{t2} & \cdots & b_{tn_t} \end{bmatrix}$$

α matrics are of size $\alpha_{t \times n_c}$; β matrics are of size $\beta_{t \times n_t}$.

t is the number of half-hour measurements in a year ($t = 365 \times 48 = 17,520$ for 2012 and 2013).

n_c is the number of households in the control group; n_t is the number of households in the treatment group.

Our goal is to find a good estimate for the counterfactual consumption.

$\hat{\beta}_{2013}$ is the counterfactual consumption for β_{2013}.

Our goal is to find a good estimate for the counterfactual consumption.

$\hat{\beta}_{2013}$ is the counterfactual consumption for β_{2013}.

$T = \hat{\beta}_{2013} - \beta_{2013}$ is the treatment effect.

Our goal is to find a good estimate for the counterfactual consumption.

$\hat{\beta}_{2013}$ is the counterfactual consumption for β_{2013}.

$T = \hat{\beta}_{2013} - \beta_{2013}$ is the treatment effect.

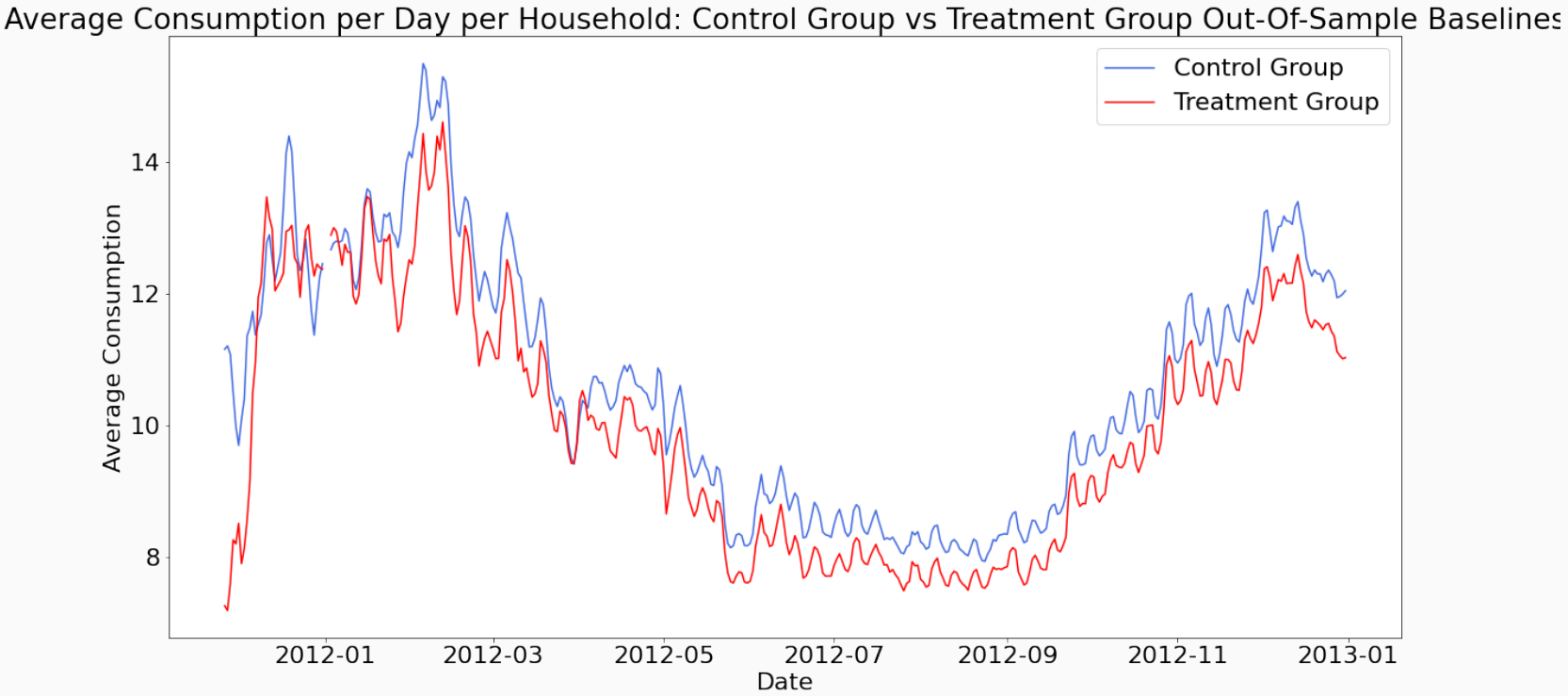
Goal is to **find a good estimate** for $\hat{\beta}_{2013}$.

Naive Model: Use Treatment Group in 2013 as Counterfactual

If we didn't know about all these biases, we *could* use control group in 2013 as the counterfactual estimate for treatment group in 2013.

$$\begin{aligned}\hat{\beta}_{2013} &\approx \alpha_{2013} \\ T &= \beta_{2013} - \hat{\beta}_{2013} \\ &= \beta_{2013} - \alpha_{2013}\end{aligned}$$

Treatment vs Control Group Baselines



$\Longrightarrow$ the two groups don't have the same baseline electricity consumption patterns.

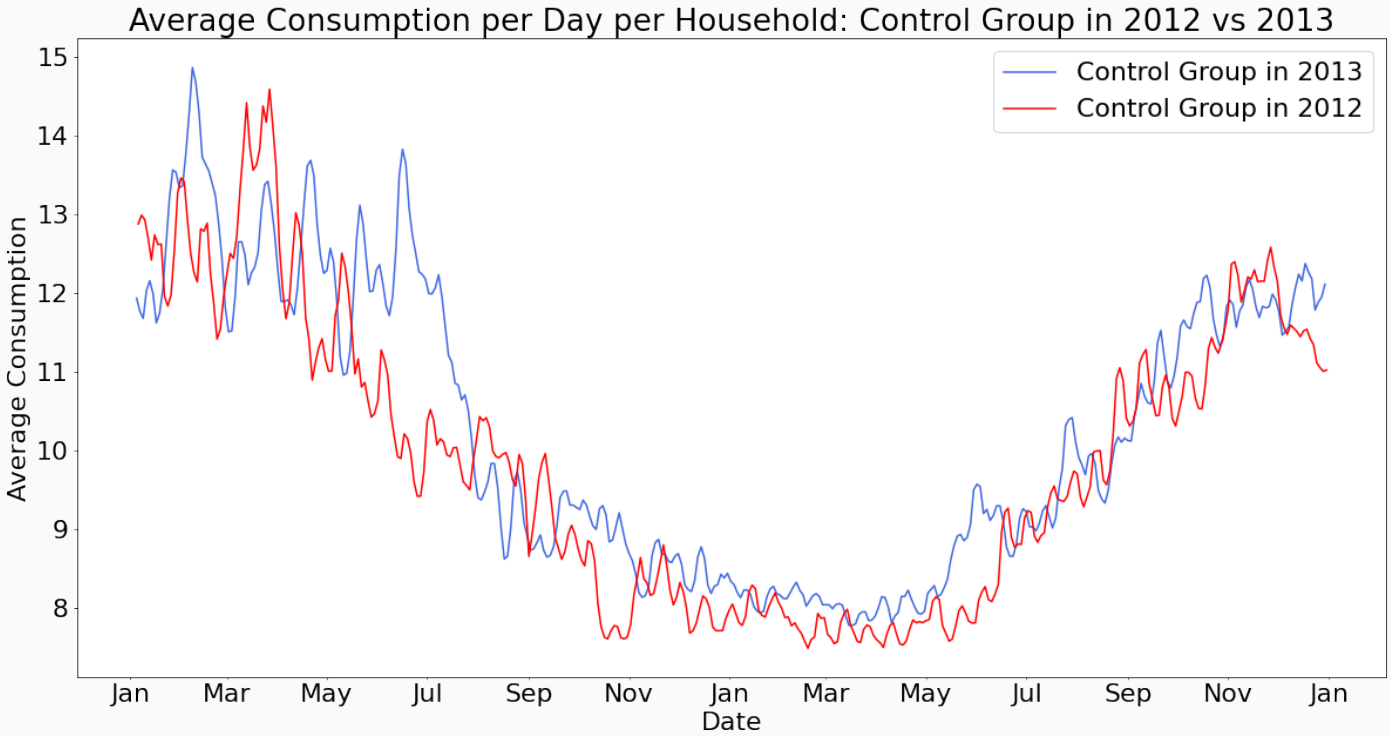
Sophisticated Naive Model

Let's instead use treatment group in 2012 as the counterfactual estimate for treatment group in 2013.

$$\hat{\beta}_{2013} \approx \beta_{2012}$$

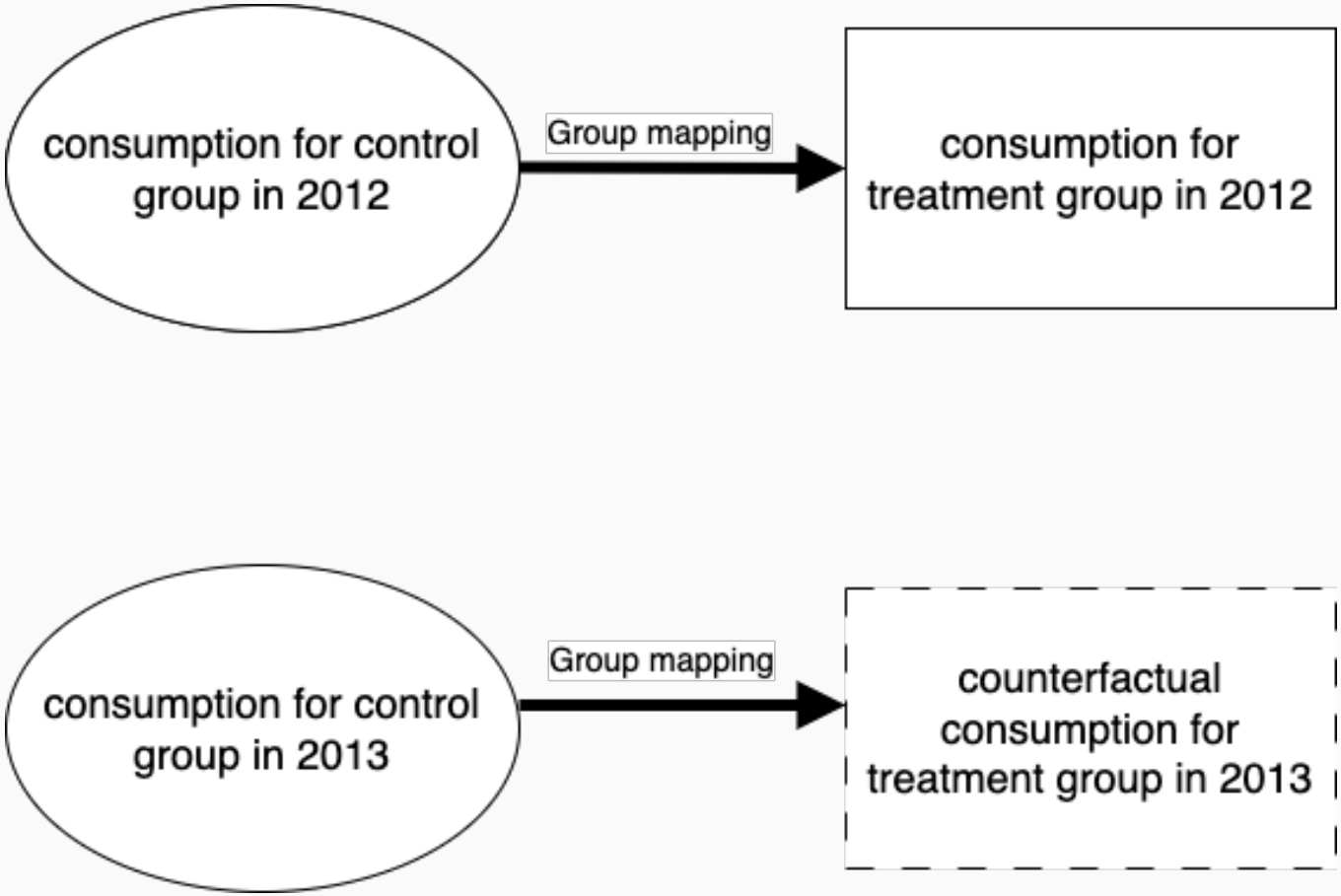
$$\begin{aligned} T &= \beta_{2013} - \hat{\beta}_{2013} \\ &= \beta_{2013} - \beta_{2012} \end{aligned}$$

Changes from Time: 2012 vs 2013

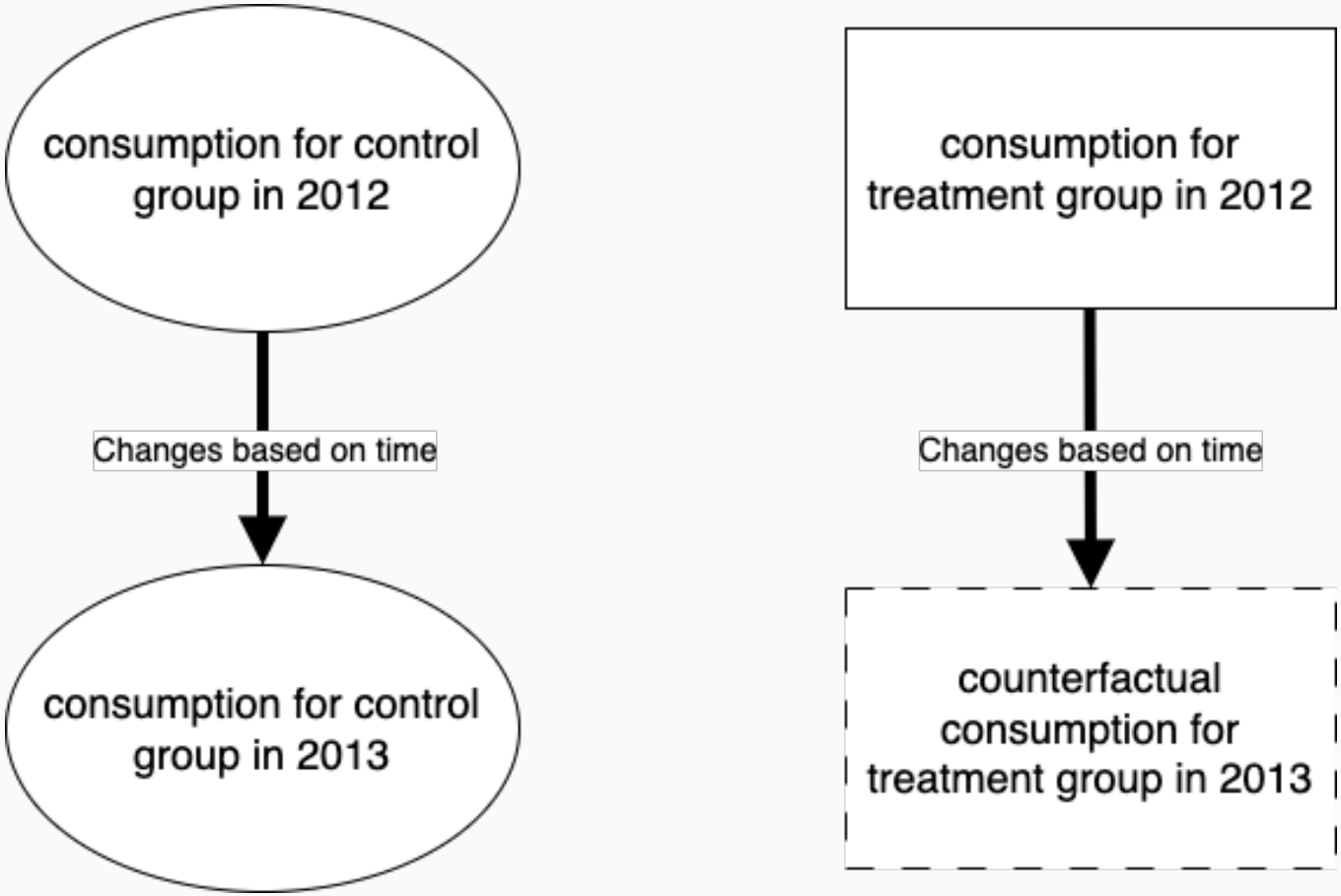


$\Longrightarrow$ Things changed between the two years even for the control group.

Treatment vs Control Group Baselines



Treatment vs Control Group Baselines

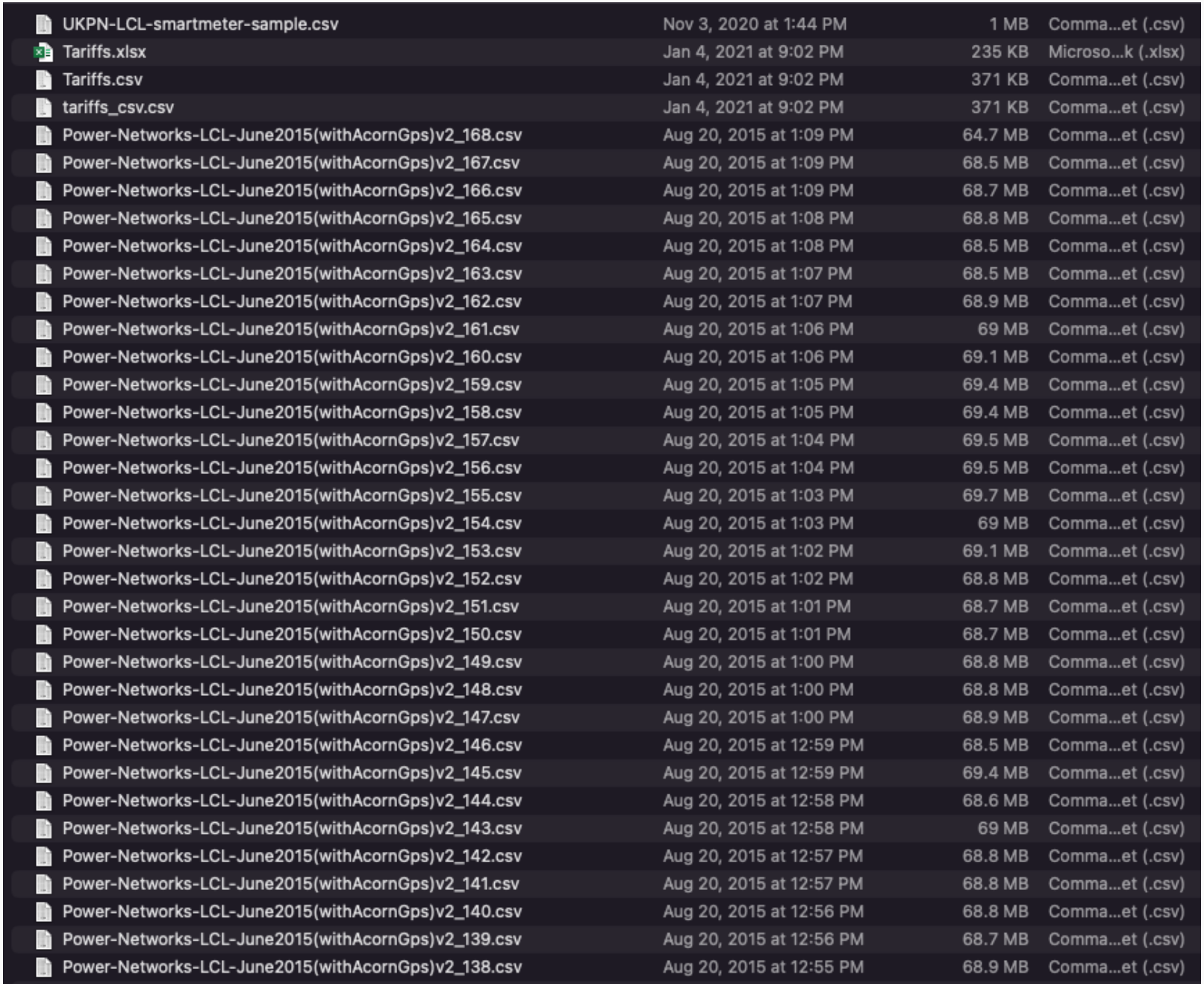


Mathematical Model Using Group Mapping

$$\alpha_{2012} X = \beta_{2012}$$
$$\alpha_{2013} X = \hat{\beta}_{2013}$$
$$\Delta \text{treatment} = \beta_{2013} - \hat{\beta}_{2013}$$

Now we have a model, let's use run the analysis based on the model

Data Prep, Cleaning, Processing



UKPN-LCL-smartmeter-sample.csv	Nov 3, 2020 at 1:44 PM	1 MB	Comma...et (.csv)
Tariffs.xlsx	Jan 4, 2021 at 9:02 PM	235 KB	Microso...k (.xlsx)
Tariffs.csv	Jan 4, 2021 at 9:02 PM	371 KB	Comma...et (.csv)
tariffs_csv.csv	Jan 4, 2021 at 9:02 PM	371 KB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_168.csv	Aug 20, 2015 at 1:09 PM	64.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_167.csv	Aug 20, 2015 at 1:09 PM	68.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_166.csv	Aug 20, 2015 at 1:09 PM	68.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_165.csv	Aug 20, 2015 at 1:08 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_164.csv	Aug 20, 2015 at 1:08 PM	68.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_163.csv	Aug 20, 2015 at 1:07 PM	68.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_162.csv	Aug 20, 2015 at 1:07 PM	68.9 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_161.csv	Aug 20, 2015 at 1:06 PM	69 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_160.csv	Aug 20, 2015 at 1:06 PM	69.1 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_159.csv	Aug 20, 2015 at 1:05 PM	69.4 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_158.csv	Aug 20, 2015 at 1:05 PM	69.4 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_157.csv	Aug 20, 2015 at 1:04 PM	69.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_156.csv	Aug 20, 2015 at 1:04 PM	69.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_155.csv	Aug 20, 2015 at 1:03 PM	69.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_154.csv	Aug 20, 2015 at 1:03 PM	69 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_153.csv	Aug 20, 2015 at 1:02 PM	69.1 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_152.csv	Aug 20, 2015 at 1:02 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_151.csv	Aug 20, 2015 at 1:01 PM	68.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_150.csv	Aug 20, 2015 at 1:01 PM	68.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_149.csv	Aug 20, 2015 at 1:00 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_148.csv	Aug 20, 2015 at 1:00 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_147.csv	Aug 20, 2015 at 1:00 PM	68.9 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_146.csv	Aug 20, 2015 at 12:59 PM	68.5 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_145.csv	Aug 20, 2015 at 12:59 PM	69.4 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_144.csv	Aug 20, 2015 at 12:58 PM	68.6 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_143.csv	Aug 20, 2015 at 12:58 PM	69 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_142.csv	Aug 20, 2015 at 12:57 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_141.csv	Aug 20, 2015 at 12:57 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_140.csv	Aug 20, 2015 at 12:56 PM	68.8 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_139.csv	Aug 20, 2015 at 12:56 PM	68.7 MB	Comma...et (.csv)
Power-Networks-LCL-June2015(withAcornGps)v2_138.csv	Aug 20, 2015 at 12:55 PM	68.9 MB	Comma...et (.csv)

Data Prep, Cleaning, Processing

1. first set of files, segmented by year, contain the `house_id` (str), `treated` (bool), `date_time` (datetime), `KWH/hh` (float).

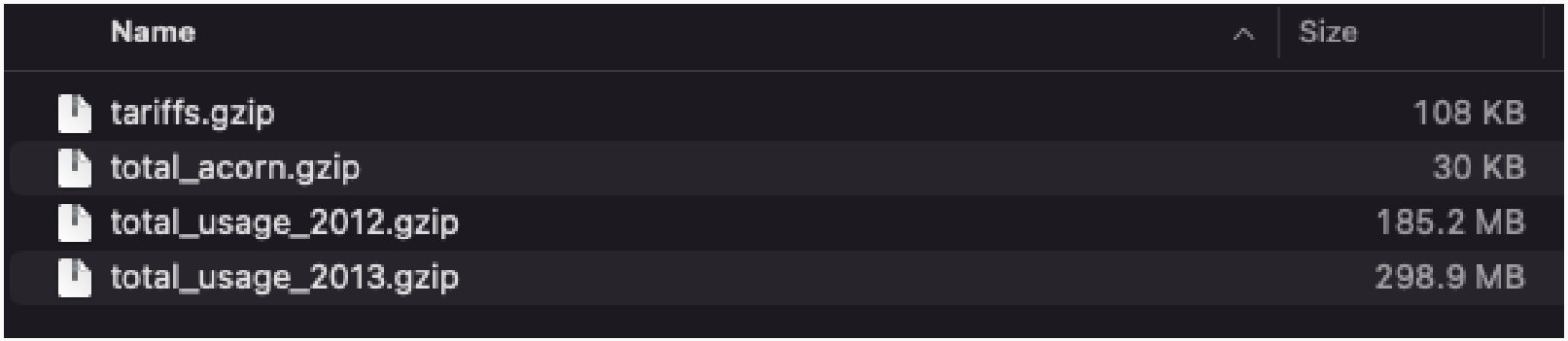
Data Prep, Cleaning, Processing

1. first set of files, segmented by year, contain the `house_id` (str), `treated` (bool), `date_time` (datetime), `KWH/hh` (float).
2. second contains `house_id` (str), `acorn` (str), `acorn_group` (str).

Data Prep, Cleaning, Processing

1. first set of files, segmented by year, contain the `house_id` (str), `treated` (bool), `date_time` (datetime), `KWH/hh` (float).
2. second contains `house_id` (str), `acorn` (str), `acorn_group` (str).
3. third is the tarrif file contains `date_time` (datetime), price per designation for that hh (float).

Data Prep, Cleaning, Processing



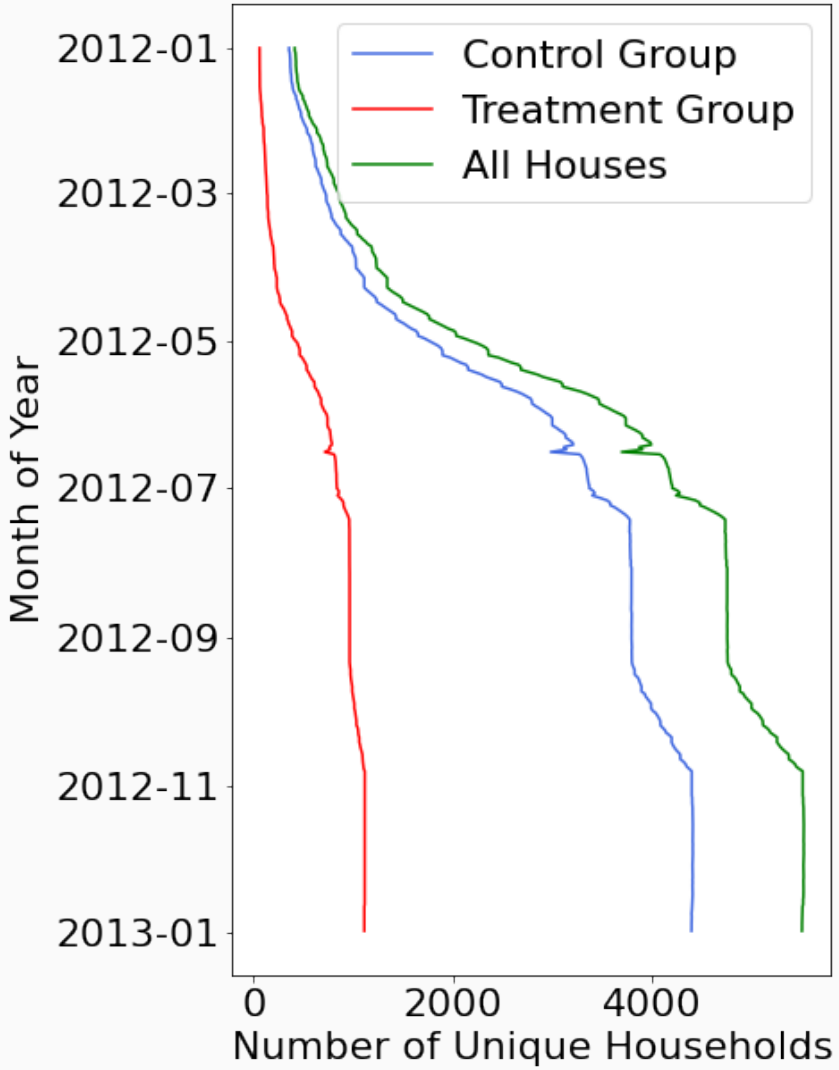
Our 2012 matrics are half full.

$$\alpha_{2012} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_c} \\ a_{21} & a_{22} & \cdots & a_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tn_c} \end{bmatrix}$$

$$\alpha_{2012} X = \beta_{2012}$$

$$\alpha_{2013} X = \hat{\beta}_{2013}$$

$$\Delta \text{treatment} = \beta_{2013} - \hat{\beta}_{2013}$$



We have some paths forward

- Impute values

We have some paths forward

- Impute values
  - Pro: You are keeping all of your data and you can capture the full trend.

We have some paths forward

- Impute values
  - Pro: You are keeping all of your data and you can capture the full trend.
  - How do you impute values?

We have some paths forward

- Impute values
  - Pro: You are keeping all of your data and you can capture the full trend.
  - How do you impute values?
  - Are you introducing error?

We have some paths forward

- Limit the timeframe to, for example, latter half of 2012

We have some paths forward

- Limit the timeframe to, for example, latter half of 2012
    - Con: Can't capture annual seasonality.

We have some paths forward

- Use a fixed panel: only keep the houses that have values for all of 2012.

We have some paths forward

- Use a fixed panel: only keep the houses that have values for all of 2012.
    - Con: You are making your sample size small.

We have some paths forward

- Reduce dimensionality

We have some paths forward

- Reduce dimensionality
  - Go from matrix of $t \times n$ to vector of size t.

We have some paths forward

- Reduce dimensionality
  - Go from matrix of $t \times n$ to vector of size t.
  - Take the mean over all houses that you have at any given t to remove this issue.

We have some paths forward

- Reduce dimensionality
  - Go from matrix of $t \times n$ to vector of size t.
  - Take the mean over all houses that you have at any given t to remove this issue.
  - Con: You are reducing dimensionality and removing valuable datapoints that might teach you something.

The path forward depends on your usecase.

There is no wrong answer here.

The path forward depends on your usecase.

There is no wrong answer here.

The choice of methodology depends on the assumptions you are comfortable making.

The path forward depends on your usecase.

There is no wrong answer here.

The choice of methodology depends on the assumptions you are comfortable making.

It's important to make informed decisions grounded in the type of analysis and with full knowledge of your data and its shortfalls.

Reducing dimensionality: Matrices $\rightarrow$ Vectors

$$\alpha_y = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n_c} \\ a_{21} & a_{22} & \cdots & a_{2n_c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tn_c} \end{bmatrix} \longrightarrow \overline{\alpha_y} = \begin{bmatrix} \overline{a_1} \\ \overline{a_2} \\ \vdots \\ \overline{a_t} \end{bmatrix}$$

where $\overline{a_i} = \frac{\Sigma a_{i,m}}{n}$ where $a_{i,m}$ has a value and n is the total number of elements with a value.

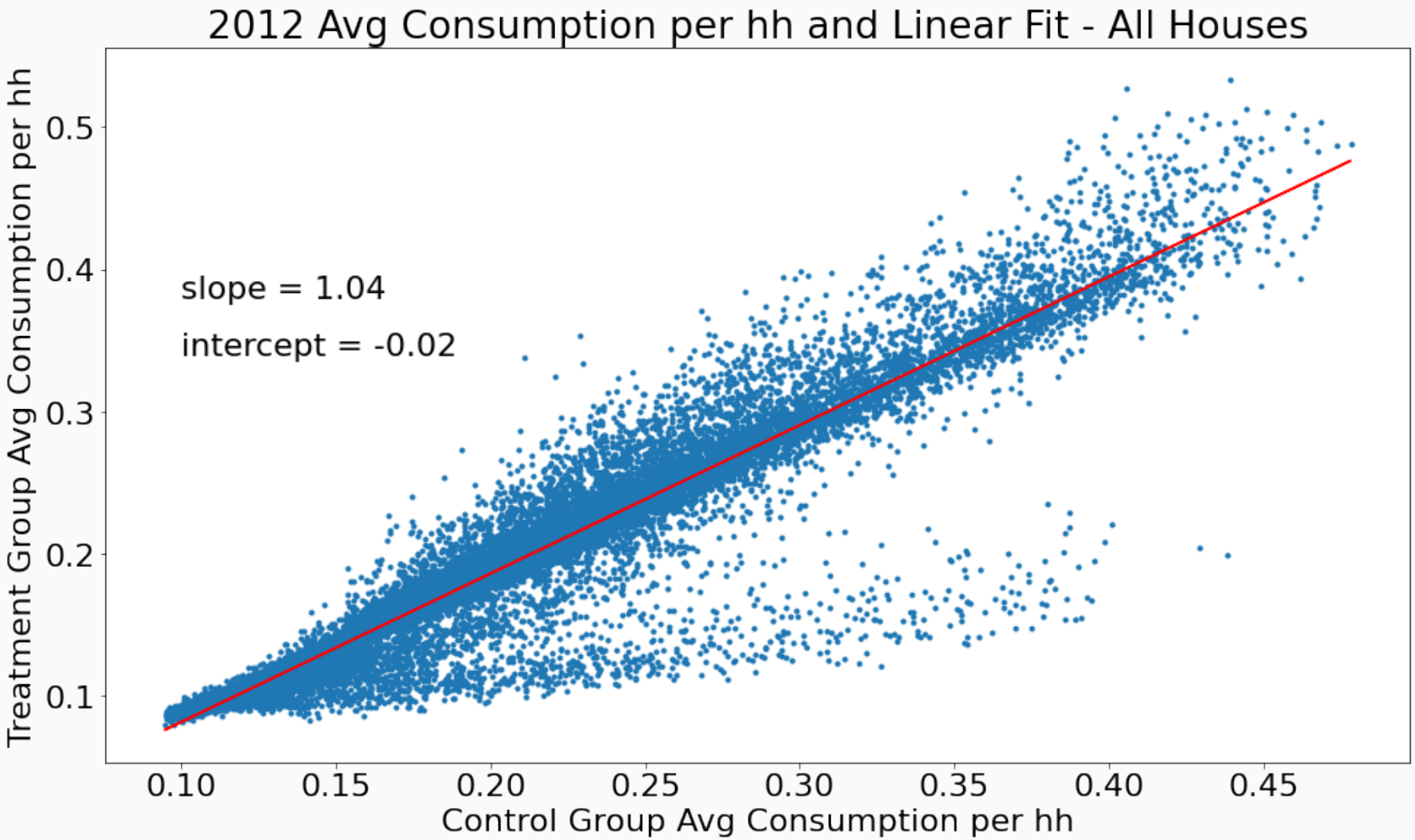
Aggregated Linear Regression Group Mapping Model

$$\overline{\beta_{2012}} = a \times \overline{\alpha_{2012}} + b$$

$$\overline{\hat{\beta}_{2013}} = a \times \overline{\alpha_{2013}} + b$$

$$\overline{\Delta \text{treatment}} = \overline{\beta_{2013}} - \overline{\hat{\beta}_{2013}}$$

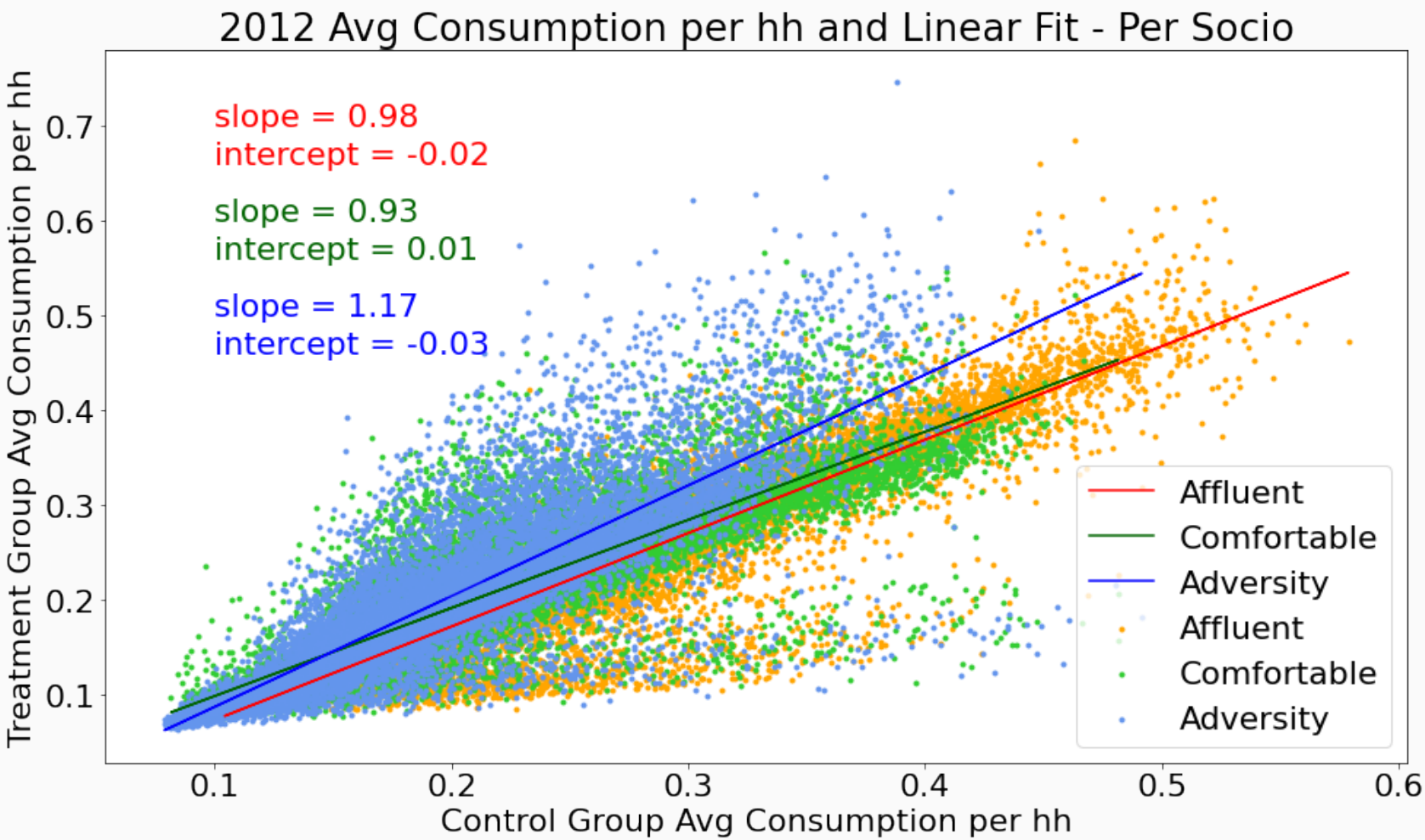
Learned linear relationship from 2012



Let's rerun the linear regression per socio-economic group

$$\alpha_y = \begin{bmatrix} \overbrace{\begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{t1} & a_{t2} \end{matrix}}^{\text{Adversity}} & \overbrace{\begin{matrix} \cdots\cdots \\ \cdots\cdots \\ \ddots \\ \cdots\cdots \end{matrix}}^{\text{Comfortable}} & \overbrace{\begin{matrix} a_{1n_c} \\ a_{2n_c} \\ \vdots \\ a_{tn_c} \end{matrix}}^{\text{Affluent}} \end{bmatrix}$$

Learned linear relationship per socio from 2012



Interpret the estimated treatment effect

$$\overline{T} = \begin{bmatrix} \overline{t_1} \\ \overline{t_2} \\ \overline{t_3} \\ \overline{t_4} \\ \vdots \\ \overline{t_i} \\ \overline{t_{i+1}} \\ \vdots \\ \overline{t_{t-1}} \\ \overline{t_t} \end{bmatrix}$$

Interpret the estimated treatment effect

$$\overline{T} = \begin{bmatrix} \overline{t_1} \\ \overline{t_2} \\ \overline{t_3} \\ \overline{t_4} \\ \vdots \\ \overline{t_i} \\ \overline{t_{i+1}} \\ \vdots \\ \overline{t_{t-1}} \\ \overline{t_t} \end{bmatrix}$$

MEDIUM

HIGH

LOW

HIGH

But where is the python?

```
>>> t_adversity.shape
(17520, 1)
>>> highs[:4]
['2013-01-04T14:00:00', '2013-01-04T14:30:00',
'2013-01-04T15:00:00', '2013-01-04T15:30:00', ...]
```

But where is the python?

```
>>> t_adversity.shape
(17520, 1)
>>> highs[:4]
['2013-01-04T14:00:00', '2013-01-04T14:30:00',
'2013-01-04T15:00:00', '2013-01-04T15:30:00', ...]

>>> t_adver_w_dates = pd.Series(t_adversity.reshape(-1),
  index=usage_vector_2013_adversity.index)
2013-01-01 00:00:00   -0.231607
2013-01-01 00:30:00   -0.212789
2013-01-01 01:00:00   -0.215968
2013-01-01 01:30:00   -0.206459
>>> t_adver_w_dates.loc[highs].mean() * 100
-12.066
```

Results

	Lows	Normals	Highs	Overall
Adversity	-1.11%	-6.41%	-12.07%	-6.17%
Comfortable	-1.11%	-7.22%	-12.44%	-6.82%
Affluent	3.88%	1.26%	-4.51%	1.24%

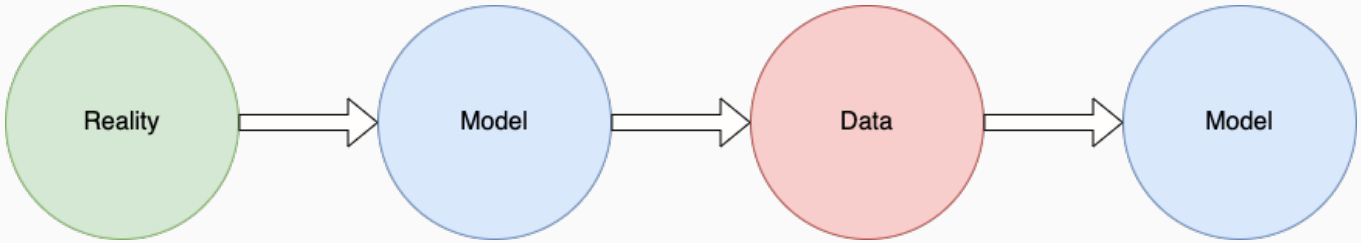
$\Longrightarrow$ The treatment was effective in lowering consumption **in times of high demand**.

Takeaways

- KYD: know your data: learn how it was collected, its biases, shortcomings, feature space, problem space.

Takeaways

- KYD: know your data: learn how it was collected, its biases, shortcomings, feature space, problem space.
- Before diving into fitting a model to your data, it's important to base your analysis on some mathematical model



Thank You! Questions?

- https://github.com/sabanejad
- https://www.linkedin.com/in/sabanejad/
- https://dspace.mit.edu/handle/1721.1/144969