CARLSON SCHOOL
OF MANAGEMENT
UNIVERSITY OF MINNESOTA

MEASURING THE CAUSAL EFFECT OF DATA CAPTURE ON RETAIL SALES

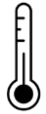
Group 3

Michael Brucek, Kevin Grady, Anthony Meyers, Danny Moncada, Jack Quick

Data Generating Process

- **Sales** data from **45** WalMart stores
- Stores located in **different regions**
- **1,000** days of data (2010/02/05 to 2012/11/01)
- Promotional data captured starting day **651**
- Data also consists of **economic** and **geographic** variables:

$ Sales 

 Temperature

Unemployment 

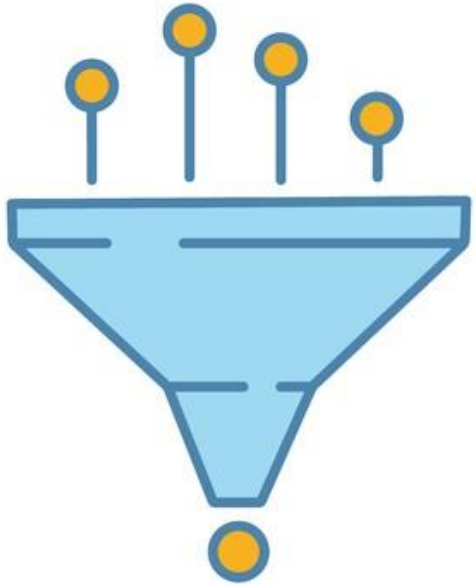
 Gasoline Price

Consumer Price Index 

Problem Statement

Did WalMart effectively leverage data assets to generate an accretive ROI in the post-treatment period (after day 651)?

Data is cheap



Information is expensive

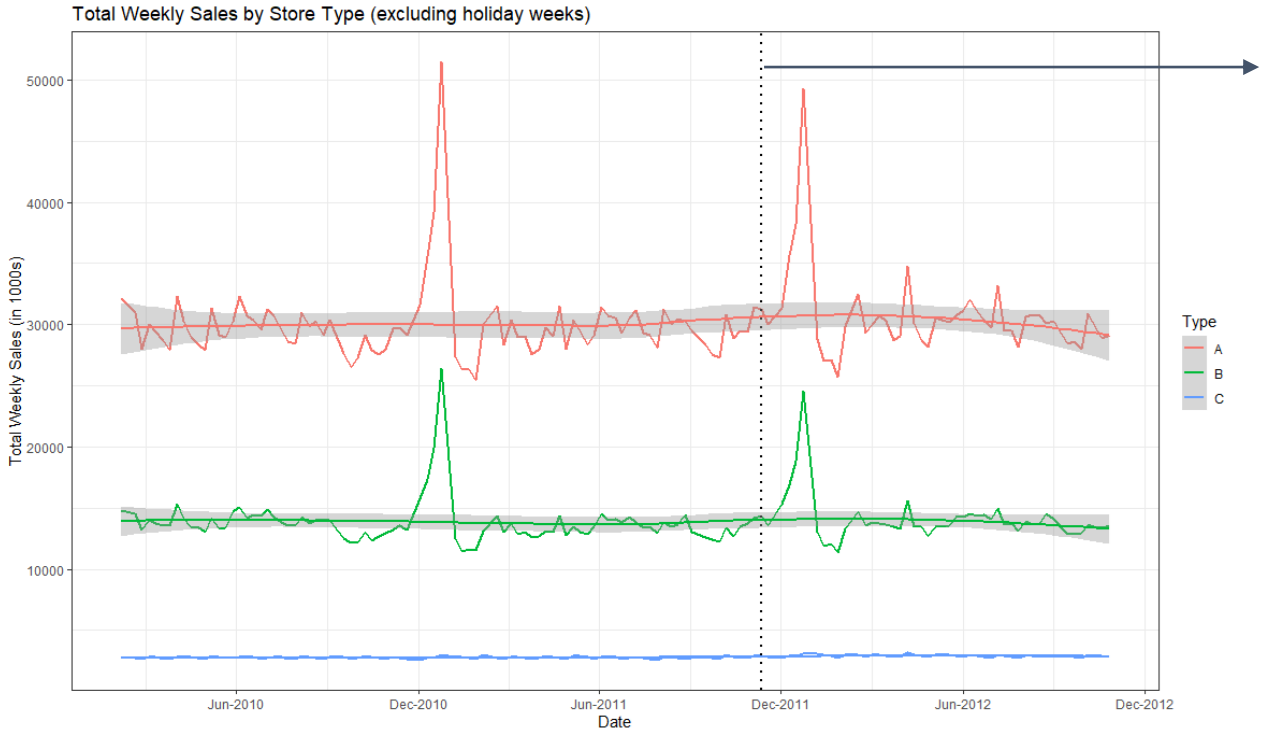
Importance of the Causal Question

- Data analytics platforms are expensive and need to drive ROI
- When inventory level is miscalculated, it's costly to the bottom line
- Customers become conditioned to always expect discounts when they are used too frequently
  - 2017 NRF Study: 1 in 3 shoppers only purchased sale items during holidays
- Effective markdown strategies are hard to master - even small casual gains can help

Threats to Causal Inference

- Data generating process may change due to strategy shifts
- Omitted variable bias and mismeasurement are clearly present, as economic forecasting requires extreme number of features
- Selection bias may be present due to:
  - The data was collected during the financial crisis
  - Only 45 stores were included
- Simultaneity bias also present because high performing stores do not need strategic actions applied to them

Data Summary



- Promotion data capture began in October 2011
- Store type clearly driving differentiated sales.

- Power testing indicates the data is slightly underpowered, only achieving significance at the 61.5% CI.
  - Would need an additional 407 weeks to achieve 85% CI

Fixed Effects Model

- Fixed Effect model on the impact of the post treatment period shows positive increase in sales during the treatment period at a significant level.

Feature	Coefficient
Treatment	$56,749.90
Temperature	-$825.312
Fuel Price	-$28,924.70
CPI	-$3,632.31
Unemployment	-$9,505.96

All significant at the 0.05 level

- However, model displays serial correlation meaning the coefficient results should **not** be used.

Test	p-value
Endogeneity (*Hausman*)	0.9353
Serial Correlation (*Breusch-Godfrey*)	2.2e-16
Heteroskedasticity (*Breusch-Pagan*)	2.2e-16

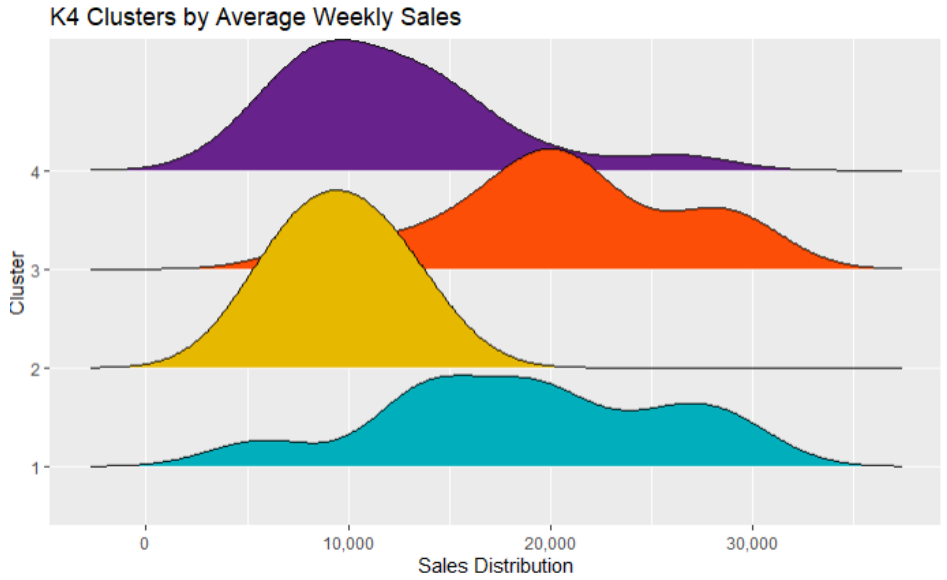
Synthetic Control Method

- **Synthetic Control Method** was utilized as well
- Predictions generated using store level promotional data in addition to economic and regional data
- In order to generate more accurate predictors, stores were clustered using K-Means algorithm to control for unobserved confounders such as:
  - Regional shopping habits
  - Regional cost levels
  - Regional distribution of WalMart stores
  - Microeconomic considerations

Synthetic Control Method: Clustering Methods

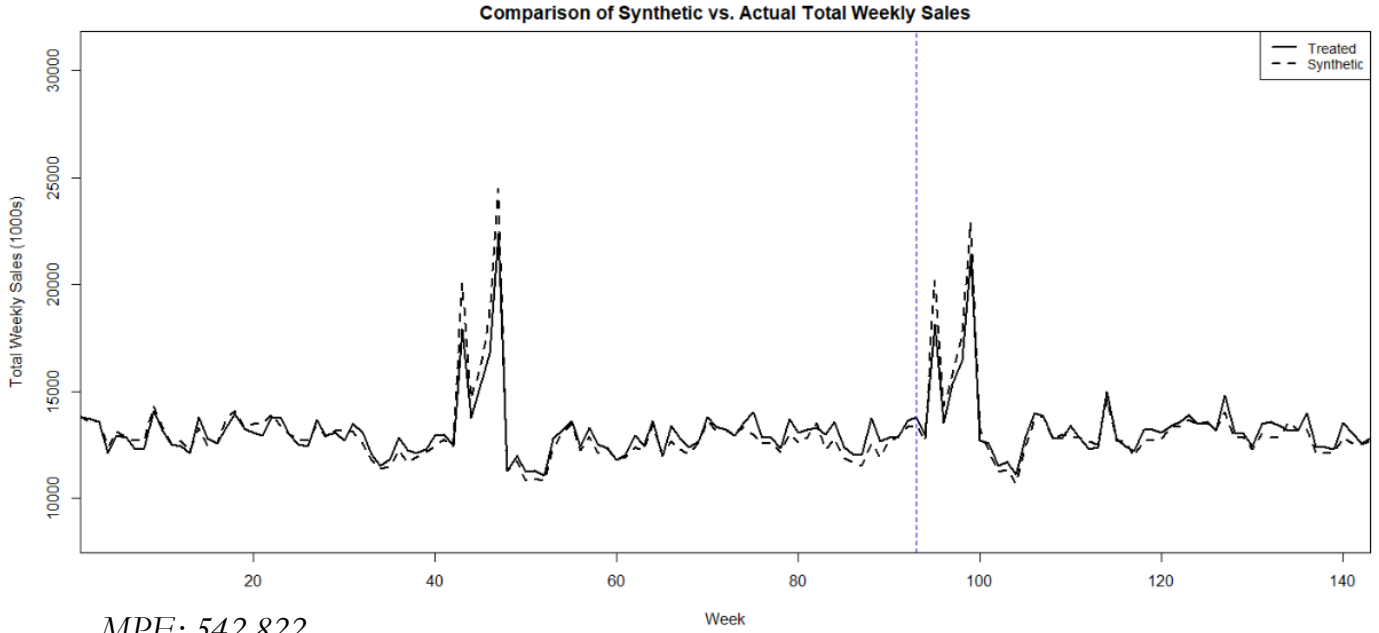
- To control for a number of regional differences, stores were clustered together
- K-Means algorithm clustered stores based on normalization of the average of the feature variables.

- **Four** distinct clusters identified



Synthetic Control Method: Results

- Synthetic Control Method generated **good fit** to actual data (*mean prediction error around 4%*)
- Synthetic model showing **no increase** post-treatment



MPE: 542,822

Conclusions

No causal relationship found between capturing promotional data and increasing sales.

Actions:

Cost of acquiring, retaining, and analyzing store level promotion data should be considered.

Data science teams to investigate further promotional level opportunities.