Candy Bar Ads Campaign Analysis

Background

What is it?

- An online ads campaign to increase the sales of a candy bar

What kind of data are available?

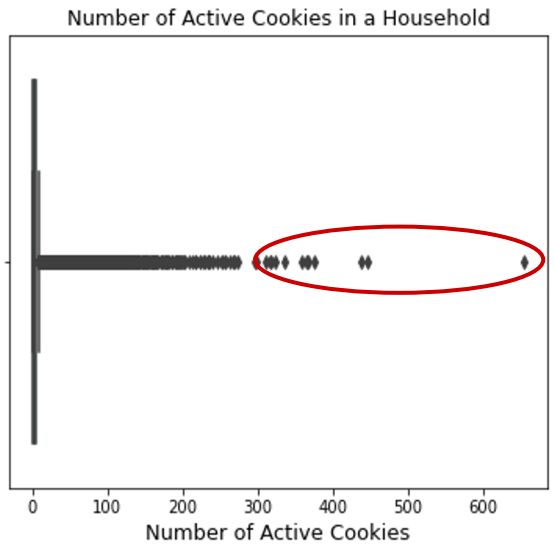
- 1,048,575 observations
- Aggregated to household level
- Demographics (Age, Income, No. of Children, etc.)
- Cookies (online activities)
- Sales (5 **quarters** back in time - 4 **weeks** following the campaign)

What is the question?

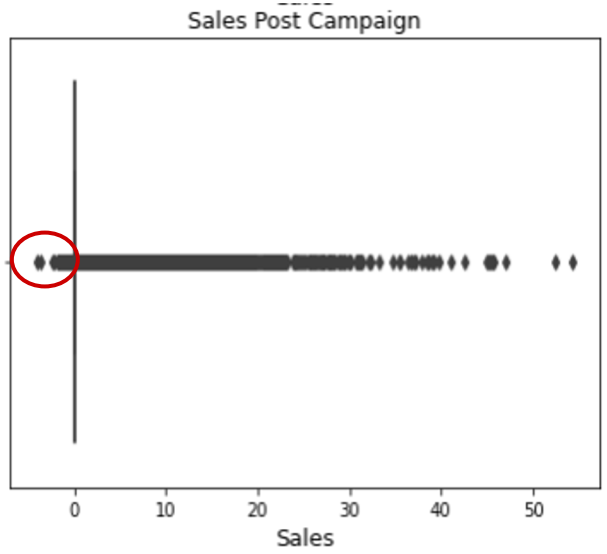
- Did the ads campaign **increase** sales?

Further Data Cleaning is Needed..

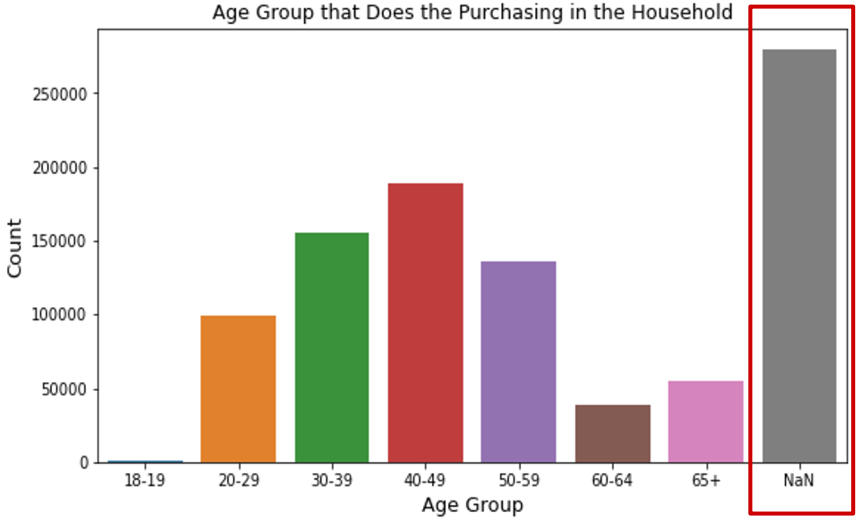
- Extreme values
  - Cookies
  - Sales
- Not Outliers



- Negative sales values
  - Post campaign sales
- Dropped

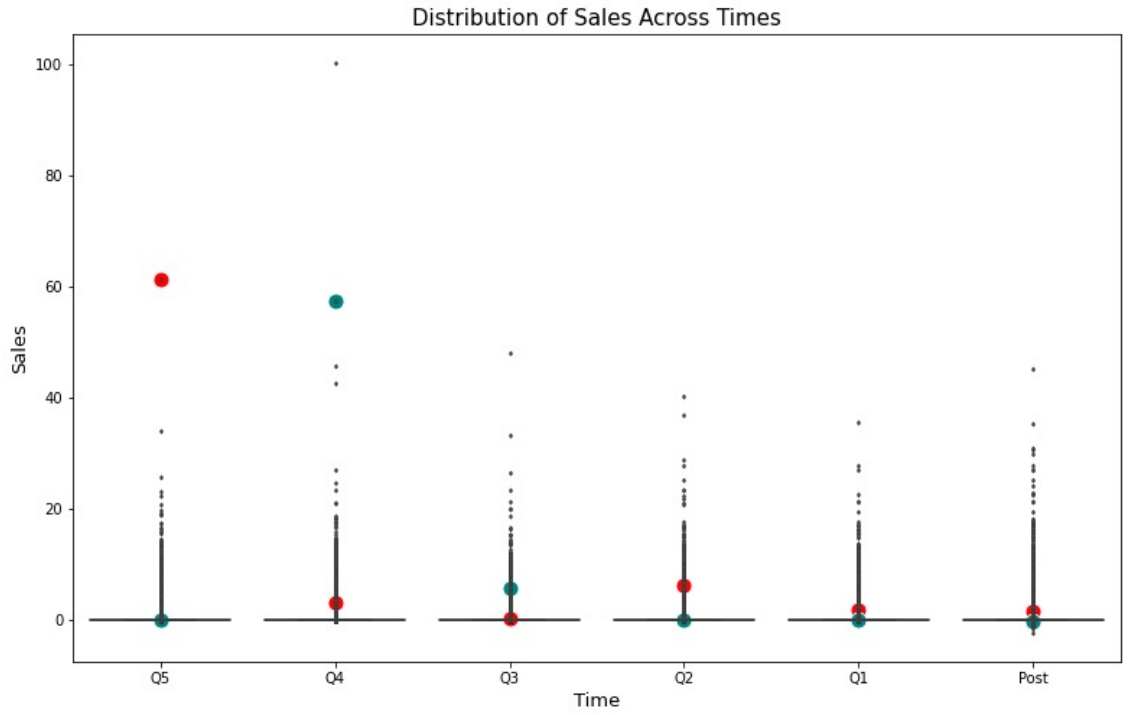


- Missing values
  - Household Income
  - Age that does the purchasing
  - Home value
- Imputed & Dropped



No Household Has Consistently Extreme Sales Values

- Certain households experience sudden spikes in sales for one quarter
- Households with consistent high sales were not found
- Assumptions:
  - Spikes are due to seasonal effect, other campaigns
  - Extreme values are not outliers

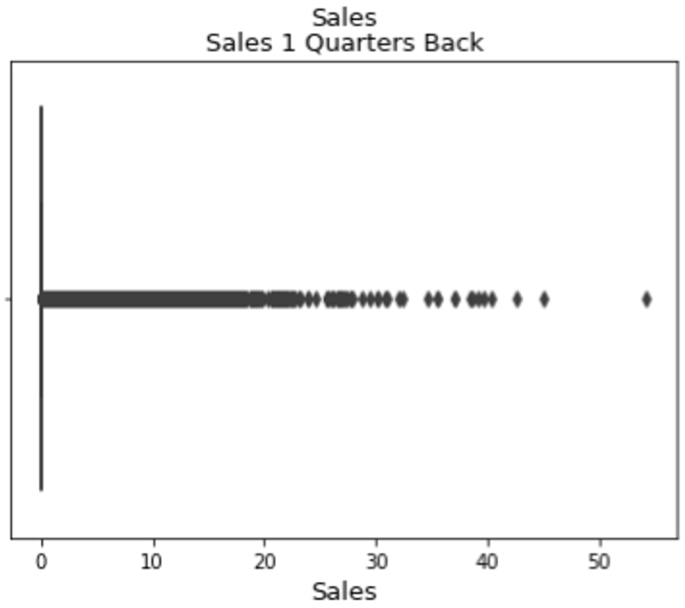


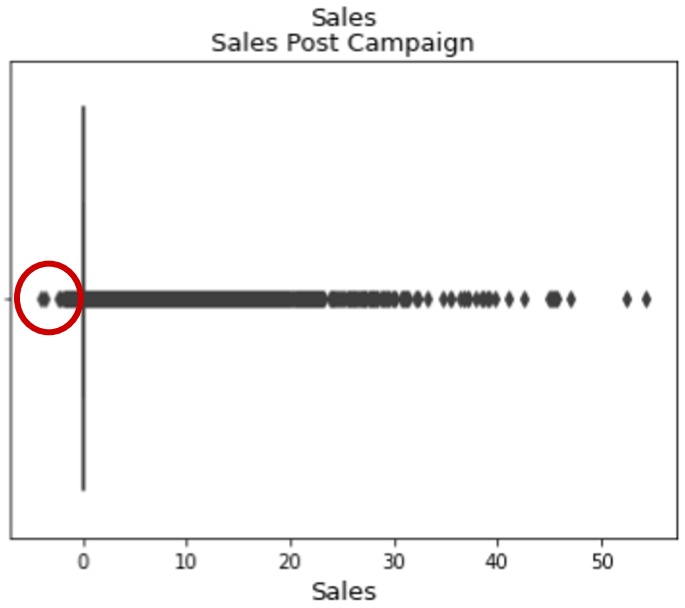
Certain Households Were Extremely Active with Few Cookies

Some households have 1 cookie only, but were more active than the average			
	No. of Cookies	**No. of Days Online**	**No. of Events**
Average	3.54	16.48	102.22
Household ID: 436609	1	6.0	681

- Moderate positive correlation (**0.38**) between No. of Cookies and No. of Events
- Further investigation on events needed
- Assumption:
  - These households are not anomalous

Households with Negative Post-Campaign Sales are Dropped

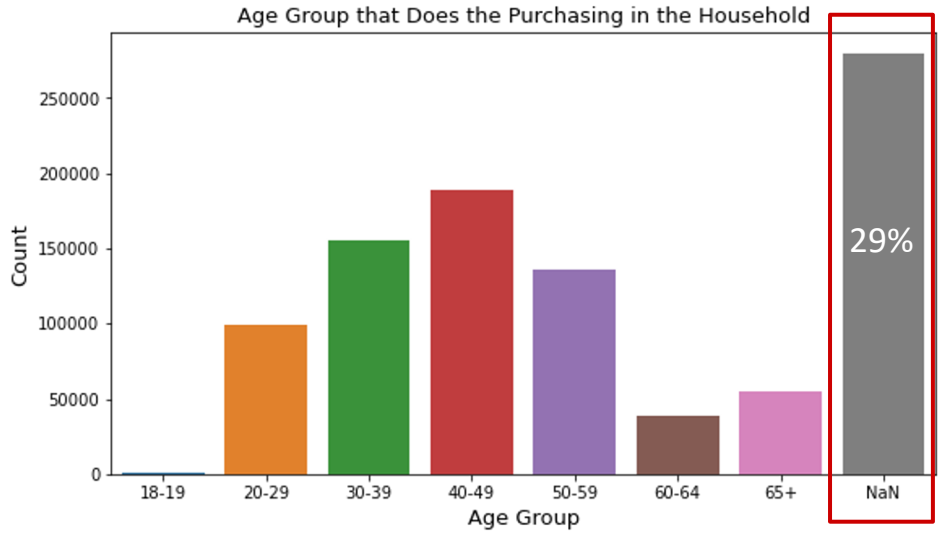


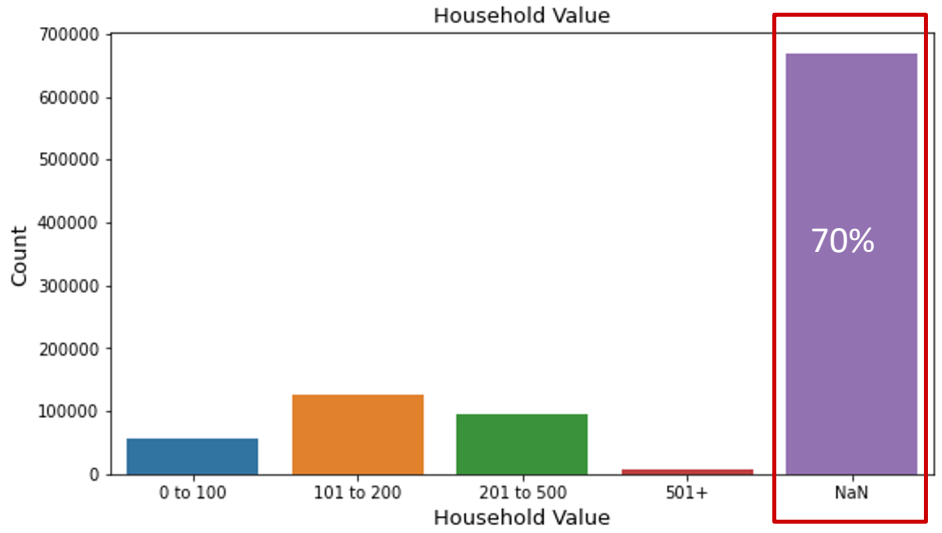


	Count
Original data	1,048,575
Negative Sales	95,122
After Dropping	953,453

- Post campaign sales range is inconsistent with pre campaign sales
- Sales values shouldn't be negative
- Assumption: negative values were made by errors

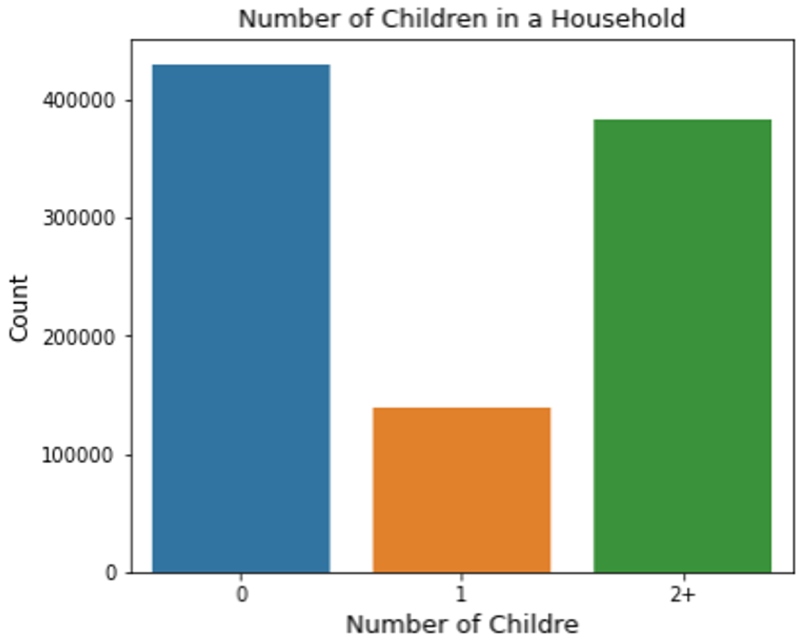
Missing Values are Dropped or Imputed with KNN Imputation

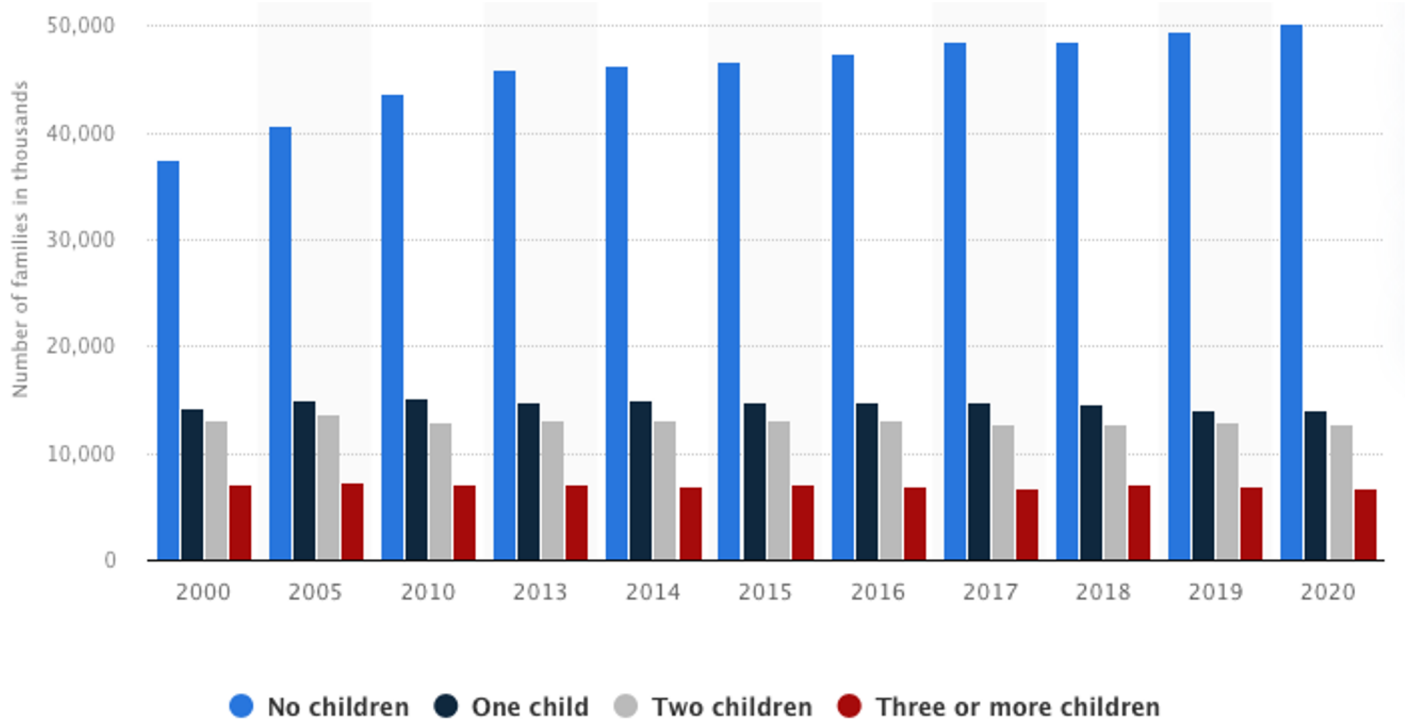




- Large amount of missing data
- Dropping Home Value
- Imputation on Age and Household Income
  - KNN Imputation based on **10%** of original data
  - Faster computation
  - Not doing prediction
- Assumption: data were missing completely at random

The Diversity of Target Audience Represents National Demographic



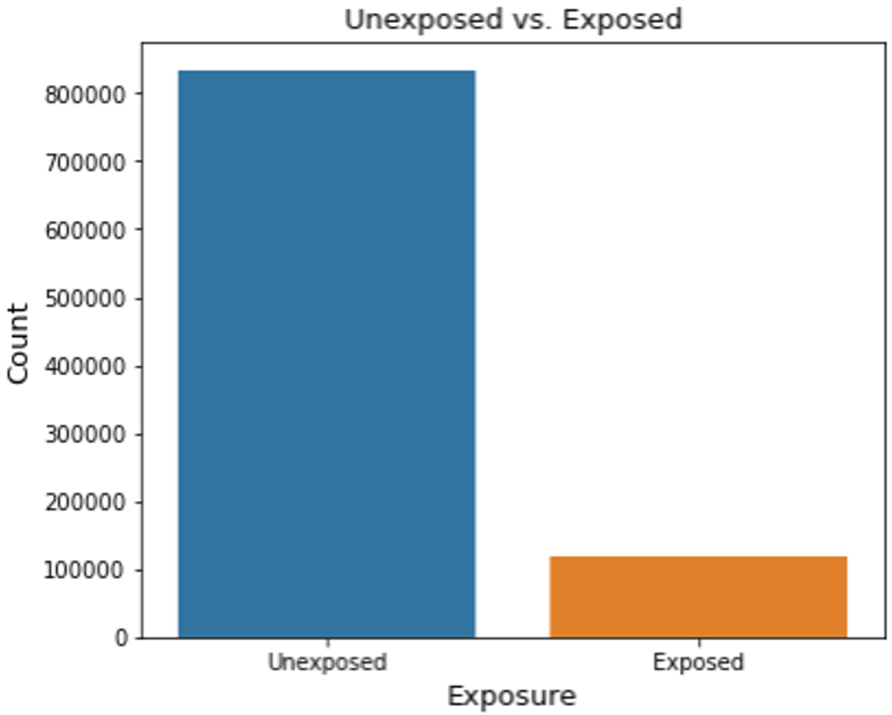


Source: US Census Bureau

- Visual inspection
- statistical tests (e.g. chi-square)
- Assumption: campaign tried to target groups that are representative of national demographic

Assume the Campaign Reached All Intended Audiences

- Exposed to Campaign vs. Actually saw the campaign
- More data needed
- Assumption: households who are exposed will have the treatment effect



Perform Matching To Balance Covariates

- Use continues variables to check for balance
- Perform propensity score-based matching

		No. of Individuals	No. of Cookies	No. of Days Online	No. of Events
Before Matching	Unexposed	10.52	3.60	16.98	108.17
	Exposed	10.45	3.07	13.03	61.49
After Matching	Unexposed	10.28	2.98	**11.84**	**46.21**
	Exposed	10.45	3.07	**13.03**	**61.49**

- Assumption: After Matching, Control and Treatment groups are reasonably balanced in their covariates.

Exposed Households Are NOT Significantly Higher Post-Campaign

Exposed and unexposed households do not show significant difference before the campaign

- Assumption: Sales refers to revenue, NOT number of items sold or people sold to
- Results of t-test

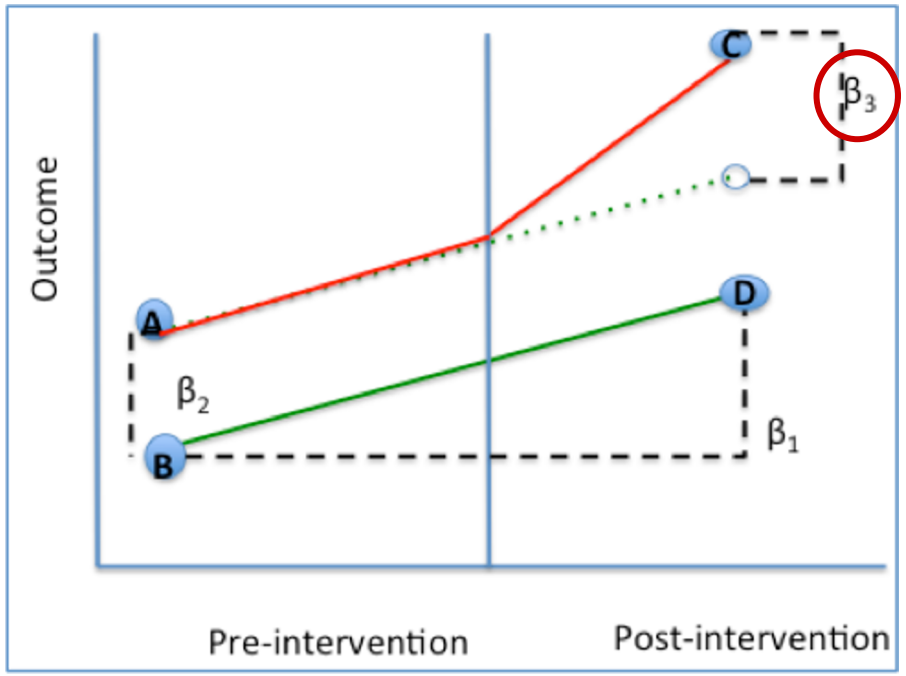
	Q1	Post-campaign
Unexposed	0.106	0.188
Exposed	0.103	0.205
p-value	**0.611**	**0.144**

- Assumption: independence and normal distribution are met

Treatment Effect is Estimated with Difference-in-Difference Regression

$$Y = \beta 0 + \beta 1 * [Pre/Post] + \beta 2 * [Control/Treat] + \beta 3 * [Pre/Post * Control/Treat] + \beta 4 * [Covariates] + \varepsilon$$

Coefficient	Interpretation
β0	Baseline
β1	Effect of time
β2	Effect of Exposure
β3	**Treatment Effect**
β4	Effect of other covariates



Source: Columbia University

- Regress on Q1 and Post Campaign sales data
- Sales from Q5 - Q2 treated as covariates
- Assumption:
  - Regression assumptions are met (normality, linearity, equal variance, independence)
  - D-in-D assumptions are met (parallel trends, positivity, SUTVA)

No Evidence to Support the Campaign Increased Sales

- Treatment effect (β3): 0.0212
  - p-value: **0.122**

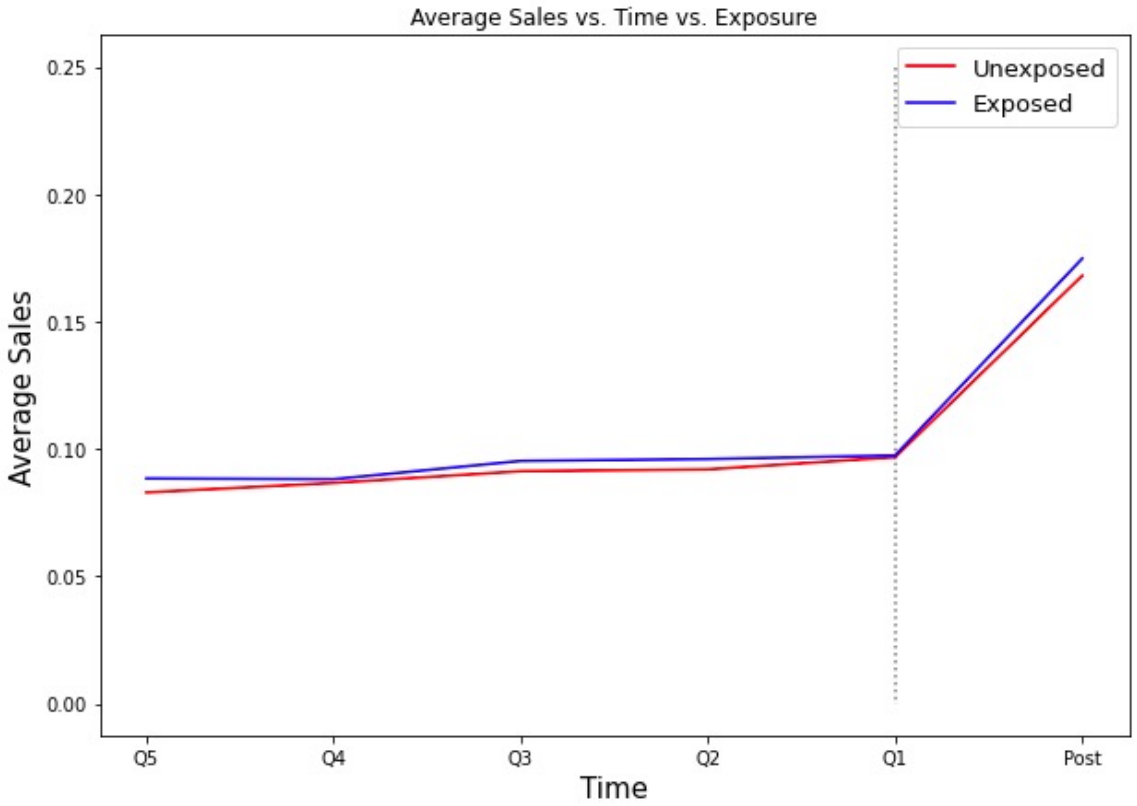
- **↑ sales during and after campaign**
- ↑ sales in Q4 to Q2 ⇒ ↑ Sales
- Number of Children, Income, Age and states are NOT significantly associated with sales

Baseline Sales = -0.0784		
Features	**Coefficient**	**P-value**
Number of Individuals	0.0005	0.486
1 children	0.0110	0.290
2+ children	-0.0043	0.584
Household Income: 51 – 100k	-0.002	0.796
Household Income: 101 – 150k	0.0119	0.325
Household Income: 151k+	-0.0128	0.545
Purchaser Age: 20 - 29	0.0378	0.701
Purchaser Age: 30 - 39	0.0506	0.608
Purchaser Age: 40 - 49	0.0584	0.553
Purchaser Age: 50 - 59	0.0597	0.545
Purchaser Age: 60 - 64	0.0291	0.769
Purchaser Age: 65 +	0.0512	0.606
During and After Campaign (β1)	0.0814	< 0.000
Q4 Sales	0.0652	< 0.000
Q3 Sales	0.2225	< 0.000
Q2 Sales	0.0979	< 0.000
Exposed to Campaign (β2)	-0.0051	0.599

The Effect of the Campaign is Difficult to Capture

- There are other factors that can impact sales
- Sales increased after campaign started, but not due to campaign alone
- Further investigation needed

- β2 minimized through matching



Limitations

Limitation on Knowledge

- Were there other campaigns
- Negative post-campaign sales?
- Detailed break down of cookies (e.g. add-to-cart)

Limitation on Techniques

- Potential multicollinearity issue
- Computational power during imputation
- Imbalanced covariates
- Test assumptions were difficult to validate

Next Steps

- Investigate abnormal and missing values
- Use the whole dataset for analysis
- Collect more features (e.g. breakdown of cookies)
- More feature engineering
- More thorough matching
- Explore historical campaigns

Keep practical significance (time & cost) in mind