

# A Multi-Arm A/B Testing and Causal Evaluation of Marketing E-Mails with Synthetic Validation and Heterogeneous Treatment Effects

Alexander Yu, Cameron Fong, Nathan Gin  
Department of Statistics, University of California, Irvine

December 12, 2025

# Contents

|           |   |           |
|-----------|---|-----------|
| <b>1</b>  | <b>Introduction</b>                             | <b>3</b>  |
| <b>2</b>  | <b>Data Description</b>                         | <b>3</b>  |
| <b>3</b>  | <b>Descriptive Statistics</b>                   | <b>4</b>  |
| <b>4</b>  | <b>Causal Identification</b>                    | <b>5</b>  |
| <b>5</b>  | <b>ATE Results on Email Dataset</b>             | <b>6</b>  |
| 5.1       | Difference-in-Means . . . . .                   | 6         |
| 5.2       | Direct Method Estimates . . . . .               | 7         |
| 5.3       | Double Robust Estimator . . . . .               | 8         |
| <b>6</b>  | <b>CATE and Heterogeneous Treatment Effects</b> | <b>9</b>  |
| 6.1       | CATE by Customer Type . . . . .                 | 9         |
| 6.2       | CATE by History Segment . . . . .               | 10        |
| 6.3       | CATE by Zip Code . . . . .                      | 11        |
| 6.4       | CATE by Channel . . . . .                       | 11        |
| <b>7</b>  | <b>Synthetic Dataset Validation</b>             | <b>12</b> |
| <b>8</b>  | <b>Discussion</b>                               | <b>13</b> |
| <b>9</b>  | <b>Limitations</b>                              | <b>13</b> |
| <b>10</b> | <b>Conclusion</b>                               | <b>13</b> |
| <b>11</b> | <b>References</b>                               | <b>14</b> |

# 1 Introduction

Marketing e-mail campaigns are widely used to drive customer engagement and revenue, but evaluating their effectiveness is challenging because customer behavior is influenced by prior purchasing history, preferences, and exposure to other marketing activities. Randomized A/B testing provides a reliable way to isolate the causal impact of campaign messages by ensuring that differences in outcomes can be attributed to the treatment itself rather than selection bias. In this study, we analyze a large-scale marketing experiment involving 64,000 customers who were randomly assigned to receive a Mens promotional e-mail, a Womens promotional e-mail, or no e-mail, with customer behavior tracked over a two week period.

While A/B tests are often used to identify which campaign performs best on average, average effects alone offer limited guidance for real world marketing decisions. Campaign effectiveness frequently varies across customer segments defined by prior spending behavior, geography, and purchasing channels. To address this, we evaluate both overall campaign performance and heterogeneity in treatment effects across key customer groups. Our goal is to provide insight not only into which e-mail performs best on average, but also into how campaign messaging can be better aligned with different types of customers to improve marketing effectiveness.

## 2 Data Description

The data in this study contains 64,000 customers enrolled in a randomized marketing experiment designed to evaluate the effectiveness of targeted e-mail campaigns on short-term purchasing behavior. Each customer was randomly assigned to one of three intervention groups: Mens E-Mail (promotional content featuring men’s merchandise), Womens E-Mail (promotional content featuring women’s merchandise), or a No E-Mail control condition. Following treatment assignment, customer behavior was tracked for a two week post campaign period, during which three key customer outcome measures were recorded. Below we describe the variables included in the dataset.

### Post-Treatment Outcomes (Measured Over Two Weeks)

- **Visit:** Indicator variable equal to 1 if the customer visited the website at least once, 0 otherwise.
- **Conversion:** Binary indicator equal to 1 if the customer made any purchase during the follow-up period.
- **Spend:** Total dollar amount spent during the two-week window (zero for non-purchasers).

### Baseline Customer Attributes (Collected Prior to Randomization)

- **Recency:** Number of months since the customer’s most recent purchase.
- **History:** Dollar amount spent in the prior twelve months.
- **History\_Segment:** Categorical variable grouping annual spending into predefined bins (e.g., \$0–\$100, \$100–\$200, \$200–\$350, etc.).
- **Mens:** Indicator equal to 1 if the customer purchased men’s merchandise in the past year.
- **Womens:** Indicator equal to 1 if the customer purchased women’s merchandise in the past year.

- **Zip\_Code:** Geographic classification of the customer’s location (Urban, Suburban, or Rural).
- **Newbie:** Binary indicator equal to 1 for customers who joined within the last 12 months.
- **Channel:** Purchasing channel(s) used by the customer in the previous year (e.g., Web, Phone, Multichannel).

## Treatment Assignment Variable

- **Segment:** Randomly assigned intervention condition: Mens E-Mail, Womens E-Mail, or No E-Mail. This variable defines the treatment exposure whose causal effect is evaluated.

The e-mail experiment was implemented using true randomization, with each of the 64,000 customers assigned independently and with equal probability(1/3) to the Mens E-Mail, Womens E-Mail, or No E-Mail group. The randomization was not conditioned on any customer characteristics, so a customer may receive an E-mail completely unrelated such as opposite sex. In addition, the dataset is complete as all customers have fully observed treatment assignments, outcome measures, and baseline covariates. No meaningful missingness was detected in key variables such as Recency, History, product-category indicators, or demographic fields. Since each customer contributes exactly one observation and no post-randomization attrition occurred, the dataset reflects the full randomized sample, ensuring that no selection bias or data loss threatens the validity of the experiment.

## 3 Descriptive Statistics

The Table 1 below summarizes customer characteristics across the three treatment groups. The Control, Mens E-Mail, and Womens E-Mail arms are nearly identical in size, each roughly one-third of the sample which reflects the intended randomization. The continuous covariates show strong balance as well with Recency averages from 5.7–5.8 months and prior-year purchase History display similar means and variability across groups. The categorical characteristics are similarly well balanced as a pproximately 55 percent of customers in each group previously purchased Mens merchandise, and a comparable proportion purchased Womens merchandise. History Segment distributions differ by at most a few tenths of a percentage point across arms, and geographic composition (Urban, Suburban, Rural), Newbie status, and purchasing Channel are nearly identical.

Table 1: Baseline Characteristics by Treatment Group

| Characteristic             | Control       | Mens E-Mail   | Womens E-Mail |
|----------------------------|---------------|---------------|---------------|
| <b>Number of Customers</b> | 21306 (33.3%) | 21307 (33.3%) | 21387 (33.4%) |
| <b>Recency (mean (SD))</b> | 5.7 (3.5)     | 5.8 (3.5)     | 5.8 (3.5)     |
| <b>History (mean (SD))</b> | 240.9 (252.7) | 242.8 (260.4) | 242.5 (255.3) |
| <b>History Segment</b>     |               |               |               |
| 1) \$0 – \$100             | 7612 (35.7%)  | 7724 (36.3%)  | 7634 (35.7%)  |
| 2) \$100 – \$200           | 4836 (22.7%)  | 4691 (22.0%)  | 4727 (22.1%)  |
| 3) \$200 – \$350           | 4044 (19.0%)  | 4090 (19.2%)  | 4155 (19.4%)  |
| 4) \$350 – \$500           | 2124 (10.0%)  | 2097 (9.8%)   | 2188 (10.2%)  |
| 5) \$500 – \$750           | 1652 (7.8%)   | 1597 (7.5%)   | 1662 (7.8%)   |
| 6) \$750 – \$1,000         | 622 (2.9%)    | 644 (3.0%)    | 593 (2.8%)    |
| 7) \$1,000+                | 416 (2.0%)    | 464 (2.2%)    | 428 (2.0%)    |
| <b>Mens (%)</b>            |               |               |               |
| 0                          | 9519 (44.7%)  | 9568 (44.9%)  | 9647 (45.1%)  |
| 1                          | 11787 (55.3%) | 11739 (55.1%) | 11740 (54.9%) |
| <b>Womens (%)</b>          |               |               |               |
| 0                          | 9638 (45.2%)  | 9558 (44.9%)  | 9622 (45.0%)  |
| 1                          | 11668 (54.8%) | 11749 (55.1%) | 11765 (55.0%) |
| <b>Zip Code</b>            |               |               |               |
| Rural                      | 3139 (14.7%)  | 3243 (15.2%)  | 3181 (14.9%)  |
| Suburban                   | 9625 (45.2%)  | 9501 (44.6%)  | 9650 (45.1%)  |
| Urban                      | 8542 (40.1%)  | 8563 (40.2%)  | 8556 (40.0%)  |
| <b>Newbie (%)</b>          |               |               |               |
| 0                          | 10611 (49.8%) | 10621 (49.8%) | 10624 (49.7%) |
| 1                          | 10695 (50.2%) | 10686 (50.2%) | 10763 (50.3%) |
| <b>Channel</b>             |               |               |               |
| Multichannel               | 2606 (12.2%)  | 2577 (12.1%)  | 2579 (12.1%)  |
| Phone                      | 9327 (43.8%)  | 9240 (43.4%)  | 9454 (44.2%)  |
| Web                        | 9373 (44.0%)  | 9490 (44.5%)  | 9354 (43.7%)  |

## 4 Causal Identification

Before estimating treatment effects we need to evaluate whether the required causal identification assumptions are satisfied in our dataset so we can proceed with the requisite statistical approaches. These causal identification in this potential outcomes framework requires three assumptions: SUTVA, ignorability, and positivity.

SUTVA requires that the observed outcome satisfies the following

$$Y_i = \sum_{a \in \mathcal{A}} Y_i^*(a) \mathbb{I}(A_i = a),$$

and that customers do not interfere with one another and each treatment corresponds to a unique, well-defined intervention. In our setting, each customer independently receives and responds to an e-mail, and there is no mechanism by which one customer’s treatment affects another’s behavior. Therefore, SUTVA is highly plausible.

Ignorability requires

$$\{Y_i^*(a) : a \in \mathcal{A}\} \perp\!\!\!\perp A_i \mid X_i,$$

meaning treatment assignment depends only on observed covariates  $X$  and not on unmeasured factors. Table 1 provides strong support for ignorability as across all baseline covariates including recency, purchase history, gender, zip-code category, newbie status, and marketing channel, the three arms show nearly identical means and proportions. The number of customers per treatment is identical (approximately 33.3% each) and continuous covariates are similar. History-segment and gender distributions also overlap almost perfectly. This balance implies that the process generating treatment assignment is independent of baseline covariates and a good chance that ignorability is well-supported.

Positivity requires that every covariate profile has a non-zero probability of receiving each treatment:

$$0 < P(A = a | X) < 1, \quad \forall a \in \mathcal{A},$$

ensuring that no subgroup is structurally excluded from any treatment. Since the distributions of all covariates are nearly identical across treatment groups, there is no evidence of exclusion (e.g., men restricted from Women’s E-Mail). Each category—gender, region, history segment, channel—appears in all arms at similar frequencies so once again this balance across all covariates is good evidence that positivity is met.

Overall, we can kind of see how the empirical covariate balance in Table 1 strongly supports all three identification assumptions from causal inference and clinical trial theory. The e-mail a/b test style experiment behaves similarly as a well-executed multi-arm randomized controlled trial which is the gold standard.

## 5 ATE Results on Email Dataset

### 5.1 Difference-in-Means

Since the e-mail campaign was implemented using true random assignment, the analysis closely mirrors what would be done in a classical randomized controlled trial. In an RCT, the key property is that treatment groups are comparable on average, both in observed characteristics and in unobserved factors that might influence outcomes. This balance means that simple differences in mean outcomes between groups provide unbiased estimates of causal effects without requiring covariate adjustment. In our case we are solving the following:

$$\begin{aligned}\widehat{ATE}_{\text{Mens, Control}}(Y) &= \bar{Y}_{\text{Mens}} - \bar{Y}_{\text{Control}}, \\ \widehat{ATE}_{\text{Womens, Control}}(Y) &= \bar{Y}_{\text{Womens}} - \bar{Y}_{\text{Control}}, \\ \widehat{ATE}_{\text{Mens, Womens}}(Y) &= \bar{Y}_{\text{Mens}} - \bar{Y}_{\text{Womens}}.\end{aligned}$$

Using this framework, we find that the Mens e-mail produces the largest improvements across all three customer-behavior outcomes. In the table below we see that relative to the Control group, customers receiving the Mens e-mail spend 77 cents more on average, and they are also more likely to visit the website and convert, with increases of 7.6 percent in visit probability and 0.68 percent in conversion probability. The Womens e-mail also shows average spending rises by about 42 cents, and both visit and conversion probabilities increase, though less sharply than for the Mens treatment. The direct comparison between the two active treatments customers exposed to the Mens e-mail exhibit higher spending, higher likelihood of visiting, and higher likelihood of converting than those who received the Womens e-mail.

These findings establish great insight and a benchmark against which the precision and robustness of more advanced causal estimators can be evaluated in subsequent analyses.

Table 2: ATE (Difference in Means) Estimates for All Outcomes

| Treatment Comparison | Spend    | Visit    | Conversion |
|----------------------|----------|----------|------------|
| Mens vs Control      | 0.769827 | 0.076590 | 0.006805   |
| Womens vs Control    | 0.424412 | 0.045233 | 0.003111   |
| Mens vs Women        | 0.345415 | 0.031356 | 0.003694   |

## 5.2 Direct Method Estimates

Although randomization guarantees that difference-in-means yields unbiased estimates, the Direct Method offers additional advantages by producing more efficient, covariate-adjusted treatment-effect estimates and enabling individual-level counterfactual predictions. To complement our baseline analysis, we therefore apply the Direct Method (also known as the Outcome Regression approach) to estimate the causal effects of the Mens and Womens e-mail campaigns on spending, visit behavior, and conversion. The main idea here is to fit a predictive model for the outcome as a function of customer characteristics and treatment assignment, use this model to estimate each customer’s potential outcome under all possible treatments, and then average these individual-level counterfactual differences to obtain overall ATE estimates.

$$\widehat{\text{ATE}}_{DM}^{(M-C)} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}(X_i, \text{Mens}) - \hat{\mu}(X_i, \text{Control})),$$

$$\widehat{\text{ATE}}_{DM}^{(W-C)} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}(X_i, \text{Womens}) - \hat{\mu}(X_i, \text{Control})),$$

$$\widehat{\text{ATE}}_{DM}^{(M-W)} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}(X_i, \text{Mens}) - \hat{\mu}(X_i, \text{Womens})).$$

In this case we will estimate ATEs using an S-Learner framework(similar to proccess seen in CausalDM), in which a single predictive model is trained to map customer features and treatment indicators to observed outcomes while using a Gradient Boosting Regressor. This model leverages customer-level covariates such as recency, purchase history, and others to produce stable outcome predictions.

After the model is trained we construct counterfactual predictions for every customer under all three treatment assignments (Control, Mens E-Mail, Womens E-Mail). These counterfactuals represent how each customer would have behaved had they been assigned to any of the other treatment conditions. The Average Treatment Effects are then computed by averaging the predicted differences in outcomes across the entire population which are the following results.

Table 3: Direct Method ATE Estimates for All Outcomes

| Treatment Comparison | Spend  | Visit  | Conversion |
|----------------------|--------|--------|------------|
| Mens vs Control      | 0.6083 | 0.0712 | 0.00575    |
| Women vs Control     | 0.3071 | 0.0411 | 0.00244    |
| Mens vs Women        | 0.3012 | 0.0301 | 0.00331    |

Across all three outcomes we see that the Mens Email campaign delivers the strongest effect. It increases expected spending by approximately 61 cents per customer and leads to noticeably higher visit and conversion rates relative to both the Control group and the Womens campaign. The Womens E-Mail campaign also produces positive effects, though consistently smaller in magnitude where it is roughly half that of the Mens

campaign. The Mens and Womens comparison further confirms that the Mens campaign is more effective on average at driving customer engagement and purchasing behavior.

The Direct Method results follow similar patterns to the simple difference-in-means estimates but provide a more stable and precise measure of the campaign effects.

### 5.3 Double Robust Estimator

In addition to the Direct Method, we estimate treatment effects using the double robust (DR) estimator to see whether it can further improve efficiency and provide a robustness check on our results. Since the e-mail experiment was conducted as a fully randomized trial, unbiased estimation does not require adjustment for treatment assignment. However, the DR approach extends the Direct Method by incorporating information from both the outcome model and the treatment assignment mechanism. In this setting, the DR estimator may reduce variability and assess the stability of our estimates rather than to correct for bias. This motivates the DR estimating equation presented below

$$\begin{aligned}\widehat{\text{ATE}}_{\text{DR}}^{(M-C)} &= \frac{1}{n} \sum_{i=1}^n \left[ \hat{\mu}(X_i, \text{Mens}) - \hat{\mu}(X_i, \text{Control}) \right. \\ &\quad \left. + \frac{\mathbb{I}(A_i = \text{Mens})(Y_i - \hat{\mu}(X_i, \text{Mens}))}{\hat{\pi}_{\text{Mens}}(X_i)} - \frac{\mathbb{I}(A_i = \text{Control})(Y_i - \hat{\mu}(X_i, \text{Control}))}{\hat{\pi}_{\text{Control}}(X_i)} \right], \\ \widehat{\text{ATE}}_{\text{DR}}^{(W-C)} &= \frac{1}{n} \sum_{i=1}^n \left[ \hat{\mu}(X_i, \text{Womens}) - \hat{\mu}(X_i, \text{Control}) \right. \\ &\quad \left. + \frac{\mathbb{I}(A_i = \text{Womens})(Y_i - \hat{\mu}(X_i, \text{Womens}))}{\hat{\pi}_{\text{Womens}}(X_i)} - \frac{\mathbb{I}(A_i = \text{Control})(Y_i - \hat{\mu}(X_i, \text{Control}))}{\hat{\pi}_{\text{Control}}(X_i)} \right], \\ \widehat{\text{ATE}}_{\text{DR}}^{(M-W)} &= \frac{1}{n} \sum_{i=1}^n \left[ \hat{\mu}(X_i, \text{Mens}) - \hat{\mu}(X_i, \text{Womens}) \right. \\ &\quad \left. + \frac{\mathbb{I}(A_i = \text{Mens})(Y_i - \hat{\mu}(X_i, \text{Mens}))}{\hat{\pi}_{\text{Mens}}(X_i)} - \frac{\mathbb{I}(A_i = \text{Womens})(Y_i - \hat{\mu}(X_i, \text{Womens}))}{\hat{\pi}_{\text{Womens}}(X_i)} \right].\end{aligned}$$

Table 4: Double Robust ATE Estimates for All Outcomes

| Treatment Comparison | Spend  | Visit  | Conversion |
|----------------------|--------|--------|------------|
| Mens vs Control      | 0.8148 | 0.0768 | 0.0071     |
| Womens vs Control    | 0.4538 | 0.0443 | 0.0034     |
| Mens vs Womens       | 0.3873 | 0.0323 | 0.0040     |

Using a double robust estimator, we see that both mens and womens email campaigns increase visit rates, conversion, and spending relative to control or no email. Men’s email shows the largest effects increasing visit probability by 7.68 percent, conversion by 0.71 percent, and spending by 0.81 dollars on average. Women’s email also improves all 3 outcomes although to a lesser extent with a 4.43 percent increase in visit probability, 0.34 percent increase in conversion, and 0.45 dollar increase in spending. Direct comparison between the two active treatments confirms that mens email outperforms womens email across all outcomes. The double

robust estimates are consistent with both the difference-in-means and direct method results, reinforcing the conclusion that the mens email is the most effective campaign overall.

## 6 CATE and Heterogeneous Treatment Effects

Treatment effect heterogeneity or HTE refers to the variation in causal effects across different subpopulations defined by baseline characteristics. Understanding HTE is crucial for tailoring marketing strategies to specific customer segments and optimizing resource allocation. In this section, we employ Causal Forests to estimate Conditional Average Treatment Effects (CATEs) for the Mens and Womens email campaigns, allowing us to identify which customer features drive differential responses to the interventions.

Formally:

$$E[Y(1) - Y(0) \mid X] \text{ varies with } X$$

Where  $X$  represents the covariates such as recency, purchase history, etc. If treatment effect changes depending on  $X$ , then there is heterogeneity. The variables of interest chosen to investigate are history segment to see if past purchasing habits affect the response to the email, customer type or whether a person is new to the site in question if those more familiar with the site are more likely to purchase products, zip code if certain areas (notably urban) have a higher chance to both visit and purchase items, and channel whether certain devices make it more convenient to click email ads. The next subsection will go over the CATE by various key variables. This motivates further exploration of treatment effect heterogeneity, which we address in the next section using CATE modeling.

### 6.1 CATE by Customer Type

Table 5: CATE by Newbie Status

| Newbie                   | Visit    | Conversion | Spending  |
|--------------------------|----------|------------|-----------|
| <b>Mens vs Control</b>   |          |            |           |
| No                       | 0.057619 | 0.004945   | 0.518407  |
| Yes                      | 0.053243 | 0.004929   | 0.646432  |
| <b>Womens vs Control</b> |          |            |           |
| No                       | 0.000352 | -0.001585  | -0.176583 |
| Yes                      | 0.007458 | 0.001408   | 0.214128  |
| <b>Mens vs Womens</b>    |          |            |           |
| No                       | 0.034611 | 0.004421   | 0.399963  |
| Yes                      | 0.025931 | 0.002755   | 0.269286  |

Across customer types, the men’s email consistently generates positive effects relative to the control group, with both newbies and non-newbies showing higher visit rates, conversion rates, and especially higher spending as compared to both the control and women’s email. The spending lift is substantial for between men’s email and control with about \$ 0.52 for non-newbies and \$ 0.65 for newbies compared to control, indicating that the Men’s email is effective across new and old customers. In contrast, the Women’s email shows much weaker performance. Among non-newbies, its effect is essentially neutral on visits and slightly negative on spending, whereas among newbies it produces a small positive impact on visits and a modest spending lift

of about 0.21 compared to control but both still less effective compared to mens email. Directly comparing Men’s vs Women’s emails, the Men’s version clearly outperforms the Women’s version for both customer types: Men’s emails drive stronger increases in visit rate, conversion, and spending. Overall, these results indicate that the Men’s email is the more effective treatment across both newbies and non-newbies, while the Women’s email provides meaningful value only for newer customers.

## 6.2 CATE by History Segment

| History Segment         | Visit     | Conversion | Spending  |
|-------------------------|-----------|------------|-----------|
| <b>Men vs Control</b>   |           |            |           |
| \$0–\$100               | 0.045739  | 0.003387   | 0.315019  |
| \$100–\$200             | 0.051290  | 0.002892   | 0.443773  |
| \$200–\$350             | 0.062001  | 0.006962   | 0.425057  |
| \$350–\$500             | 0.074945  | 0.008891   | 1.500787  |
| \$500–\$750             | 0.071382  | 0.005843   | 1.110247  |
| \$750–\$1000            | 0.050762  | 0.013146   | 1.457034  |
| \$1000+                 | 0.072808  | 0.007819   | 0.940359  |
| <b>Women vs Control</b> |           |            |           |
| \$0–\$100               | 0.009992  | 0.000718   | 0.309802  |
| \$100–\$200             | 0.004250  | 0.001624   | 0.171597  |
| \$200–\$350             | -0.007047 | -0.004721  | -0.580140 |
| \$350–\$500             | -0.006826 | -0.003871  | -0.905179 |
| \$500–\$750             | 0.005505  | 0.005003   | 0.357493  |
| \$750–\$1000            | 0.004189  | 0.003930   | 0.102003  |
| \$1000+                 | 0.028177  | 0.005494   | 1.744044  |
| <b>Men vs Women</b>     |           |            |           |
| \$0–\$100               | 0.016437  | 0.000894   | -0.029689 |
| \$100–\$200             | 0.028005  | 0.001710   | 0.335770  |
| \$200–\$350             | 0.044326  | 0.009174   | 0.882761  |
| \$350–\$500             | 0.057440  | 0.006311   | 1.027141  |
| \$500–\$750             | 0.034880  | 0.002711   | 0.384652  |
| \$750–\$1000            | 0.019401  | 0.006856   | 0.566638  |
| \$1000+                 | 0.025488  | -0.000020  | -0.749106 |

Table 6: CATE Estimates by History Segment for Visit, Conversion, and Spending

CATE analysis shows substantial treatment effect heterogeneity across purchase history segments. Men’s email performs strongly across all outcomes, particularly for customers with \$350–\$1000 in prior spending with an average 1.50 per person that received a mens email vs no email or average 1 dollar per person increase that received a mens email vs womens email. Women’s email exhibits mixed effects, with positive impacts for low and very high spenders but negative impacts for the \$200–\$500 segments. Direct comparison (Men vs Women) confirms that Men’s email consistently outperforms Women’s email for visit rate, conversion rate, and spending. These results imply that targeting strategies should prioritize Men’s email for mid- and upper-history customers while avoiding Women’s email in the \$200–\$500 range.

### 6.3 CATE by Zip Code

Table 7: CATE by Zip Code Type

| Zip Code Type            | Visit    | Conversion | Spending  |
|--------------------------|----------|------------|-----------|
| <b>Mens vs Control</b>   |          |            |           |
| Rural                    | 0.055512 | 0.004312   | 0.595440  |
| Suburban                 | 0.054358 | 0.005377   | 0.640372  |
| Urban                    | 0.056579 | 0.004676   | 0.513299  |
| <b>Womens vs Control</b> |          |            |           |
| Rural                    | 0.000521 | 0.000350   | 0.094205  |
| Suburban                 | 0.005481 | -0.000678  | -0.130201 |
| Urban                    | 0.056579 | 0.000427   | 0.159911  |
| <b>Mens vs Womens</b>    |          |            |           |
| Rural                    | 0.030956 | 0.003347   | 0.391729  |
| Suburban                 | 0.029166 | 0.004049   | 0.465542  |
| Urban                    | 0.031199 | 0.003153   | 0.165960  |

Much like the previous variables, the men’s email consistently outperforms the control group across all zip code types (Rural, Suburban, Urban) with positive effects on visit rates, conversion rates, and spending. The spending lift is particularly notable in suburban areas where men’s email drives an average increase of about \$0.64 per person compared to control. In contrast, the women’s email shows mixed results: it has a small positive effect on visits in rural and urban areas but a negative effect in suburban areas, and its impact on spending is generally weak or negative across all zip code types. Direct comparisons between men’s and women’s emails further highlight the improved effect of the men’s version, which consistently yields higher visit rates, conversion rates, and spending increases regardless of geographic location. Overall, these findings suggest that the men’s email is the more effective treatment across all zip code.

### 6.4 CATE by Channel

Table 8: CATE by Channel

| Channel                  | Visit    | Conversion | Spending  |
|--------------------------|----------|------------|-----------|
| <b>Mens vs Control</b>   |          |            |           |
| Multichannel             | 0.066020 | 0.006723   | 0.920752  |
| Phone                    | 0.054383 | 0.004506   | 0.523518  |
| Web                      | 0.053537 | 0.004873   | 0.548496  |
| <b>Womens vs Control</b> |          |            |           |
| Multichannel             | 0.001797 | 0.000523   | -0.056822 |
| Phone                    | 0.004806 | -0.000475  | 0.027197  |
| Web                      | 0.003627 | 0.000142   | 0.033195  |
| <b>Mens vs Womens</b>    |          |            |           |
| Multichannel             | 0.040143 | 0.004830   | 0.504127  |
| Phone                    | 0.030702 | 0.003526   | 0.301339  |
| Web                      | 0.027096 | 0.003301   | 0.320560  |

Finally looking into channels, the men’s email consistently outperforms the control group across all channels (Multichannel, Phone, Web) with positive effects on visit rates, conversion rates, and spending. The spending lift is particularly notable in the Multichannel segment where men’s email drives an average increase of about \$0.92 per person compared to control . In contrast, the women’s email shows weak or negative results across all channels, with minimal impact on visits and spending. Direct comparisons between men’s and women’s emails further highlight the superior effect of the men’s version, which consistently yields higher visit rates, conversion rates, and spending increases regardless of channel. Overall, these findings suggest that the men’s email is the more effective treatment across all purchasing channels with having the largest difference in visit, conversion, and spending compared to the previous 3 variables of zip code, history segment, and customer type (exception being the high tier past spenders).

## 7 Synthetic Dataset Validation

To validate our analysis of the real e-commerce A/B test, we constructed a synthetic dataset. This enables a direct comparison of causal estimators that is not possible with the real data alone. We evaluate all of our ATE estimators, Difference-in-Means (DiM), the Direct Method (DM), and the Double Robust (DR) estimator using Mean Squared Error (MSE), which captures both bias and variance. The table below reports MSE values across all treatment comparisons and outcomes, along with an overall MSE for each method.

| <b>Treatment Comparison</b>      | <b>Spend</b> | <b>Visit</b>  | <b>Conversion</b> |
|----------------------------------|--------------|---------------|-------------------|
| <b>Difference-in-Means (DiM)</b> |              |               |                   |
| Mens vs Control                  | 0.0760       | 0.0787        | 0.00044           |
| Womens vs Control                | 0.0095       | 0.0736        | 0.00031           |
| Mens vs Womens                   | 0.0319       | 0.0001        | 0.00001           |
| <b>Overall MSE</b>               |              | <b>0.0300</b> |                   |
| <b>Direct Method (DM)</b>        |              |               |                   |
| Mens vs Control                  | 0.00126      | 0.06209       | 0.00005           |
| Womens vs Control                | 0.03791      | 0.05642       | 0.00009           |
| Mens vs Womens                   | 0.02534      | 0.00014       | 0.00001           |
| <b>Overall MSE</b>               |              | <b>0.0204</b> |                   |
| <b>Double Robust (DR)</b>        |              |               |                   |
| Mens vs Control                  | 0.03323      | 0.09036       | 0.00043           |
| Womens vs Control                | 0.00017      | 0.08335       | 0.00030           |
| Mens vs Womens                   | 0.01812      | 0.00009       | 0.00001           |
| <b>Overall MSE</b>               |              | <b>0.0251</b> |                   |

Table 9: Mean Squared Error (MSE) of ATE Estimates Across Methods and Outcomes

The results show that Difference-in-Means exhibits the highest overall MSE, showing higher variance, particularly for the spend and visit outcomes. In contrast, the Direct Method achieves the lowest overall MSE, demonstrating clear efficiency gains from covariate adjustment even under randomized treatment assignment. The Double Robust estimator improves upon Difference-in-Means but does not outperform the Direct Method, as its robustness advantages are less relevant in a fully randomized setting. Overall, these results confirm that covariate-adjusted estimators, especially the Direct Method, provide more precise treatment effect estimates in this context, supporting their use in the main analysis.

## 8 Discussion

Across all outcomes, the Mens e-mail produces the largest improvements on average, increasing visits, conversions, and spending relative to both control and the Womens e-mail. The Direct Method and DR estimates follow the same pattern as difference-in-means, suggesting the main conclusions are stable across estimators.

The CATE analysis shows substantial heterogeneity. Mens e-mails perform strongly across most segments, with especially large spending gains among moderate to high prior spenders and multichannel customers, while Womens e-mails are weaker and more variable, including midrange history segments where effects are near zero or negative. These results suggest that although the experiment supports clean causal identification, the most actionable insight is that targeting strategies informed by heterogeneity could outperform uniform deployment.

The synthetic validation further clarifies the statistical properties of our estimators. Overall, the synthetic experiments show that while all estimators recover the true treatment effects reasonably well, covariate adjusted approaches like the Direct Method and the double robust estimator achieve much lower mean squared error than the simple difference in means benchmark. This highlights meaningful efficiency gains even in a randomized experimental setting.

## 9 Limitations

Although the e-mail campaign was implemented as a purely randomized, multi-arm experiment which preserves the validity of a classical RCT and A/B test, the design may be less effective from a practical marketing perspective. The Mens and Womens e-mails were randomly assigned without regard to prior purchasing behavior or customer preferences, meaning some customers received messages weakly aligned with their interests. While this strengthens causal identification, it can weaken observed treatment effects and reduce the contrast between experimental arms.

As statisticians we like the fully randomized design but from a business and decision-making standpoint, the experiment may not be the most optimal targeting strategy. The estimated effects reflect intent-to-treat impacts under random exposure rather than the effectiveness of targeted or personalized campaigns. As a result, while the design supports rigorous causal inference, it is less informative for guiding practical marketing deployment.

Another potential limitation is that even though the dataset includes several key observed covariates such as recency, purchase history, and acquisition channel, unobserved covariates remain a limitation when interpreting differential responses across treatments. Latent factors such as individual preferences, baseline interest in product categories, prior exposure to marketing content, or underlying engagement propensity are not observed but may systematically affect how customers respond to different e-mail messages. While randomization ensures unbiased estimation of effects relative to the control group, it does not eliminate the influence of these unobserved characteristics on treatment effect heterogeneity.

## 10 Conclusion

Using a randomized marketing experiment, we find that the Mens e-mail campaign consistently outperforms the Womens e-mail and the no e-mail control in increasing short-term website visits, conversions, and customer spending. These results indicate that e-mail content matters for customer engagement and that message design can have a measurable impact on purchasing behavior even over a short follow up period.

Beyond average performance, our analysis shows that campaign effects vary meaningfully across customer segments. The strongest gains from the Mens e-mail are concentrated among customers with higher prior spending and multichannel purchasing behavior, while the Womens e-mail exhibits weaker and more variable effects across segments. Taken together, these findings highlight the value of combining randomized experimentation with segment level analysis to inform more targeted and effective marketing strategies, rather than relying solely on average A/B test results.

## 11 References

Hillstrom, K. (2008, March 20). The MineThatData e-mail analytics and data mining challenge [Data set]. MineThatData. <https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and.html>