Statistical Advice on the Effect of Interventions on Beverage Sales

Parham Pishrobat (71097927),
Johnson Chen (85784080),
Sarah Masri (97415681)

March 28, 2024

1. Introduction

Concerns around sugar consumption and its health implications have prompted many interventions to change consumer behaviour cut down on sugary beverages. The current study investigates the effectiveness of two types of intervention strategies to motivate and incentivize consumers to choose zero-calorie beverages over sugary alternatives. In particular, the research question focuses on the impact of two strategies to inform consumers about calorie content through visual presentations: posters highlighting either the calorie content or the physical activity required to burn calories. Furthermore, the effectiveness of price discounts on behaviour is explored, both independently and in conjunction with explanatory messaging. This study also seek to understand if the effectiveness of those strategies varies across different sites.

2. Data Description and Summaries

The study adopts an interrupted time-series multi-site quasi-experimental design to assess the effectiveness of the five interventions on the purchase patterns of bottled sugary and zero-calorie beverages. The data are recorded from cafeterias and convenience shops at three hospital sites, denoted by A, B, and C. Hospital A is urban and has two cafeterias and two convenience shops. Hospital B is also an urban setting but it has only one cafeteria. Hospital C is a suburban setting, having one cafeteria and one convenience shop. Both interventions and data collection were automatic at site A and by trained personnel in sites B and C. Sugary beverages include regular soft drinks and iced teas, sweetened with natural sugars like sucrose and corn syrup, and zero-calorie beverages include diet soft drinks and teas, and water. Other beverages, such as juice and milk, are excluded from the study due to challenges in identifying their sugar contents. Nevertheless, the total sales information is kept in the study to represent the overall patterns of beverage sales irrespective of their sugar content.

The experiment, starting on October 27, 2009 and ending 32 weeks later, measure daily sales of bottled sugary and zero-calorie beverages. The study period included a baseline data collection phase, intervention phases to elicit behaviour change, and washout periods to assess the persistence of intervention effects. The dataset contains 631 sales counts for zero-calorie, sugary, and all drinks, sale location, day of the week, and the type of intervention applied. The recorded variables consist of sales count, day of the week, site location, type of intervention, and beverage category (zero-calorie and sugary options). The price interventions consist of two periods of 10% discount on zero-calorie beverages, with one phase providing additional explanatory messaging about the discount. The calorie messaging interventions provided information on the caloric content of sugary drinks, the physical activity required to burn off these calories, and a combination of both strategies.

The day of week, site and intervention covariates are each considered categorical data types. Other observations are classified under numerical data types as they measure sales counts. Missing data is observed over some control periods of the study.

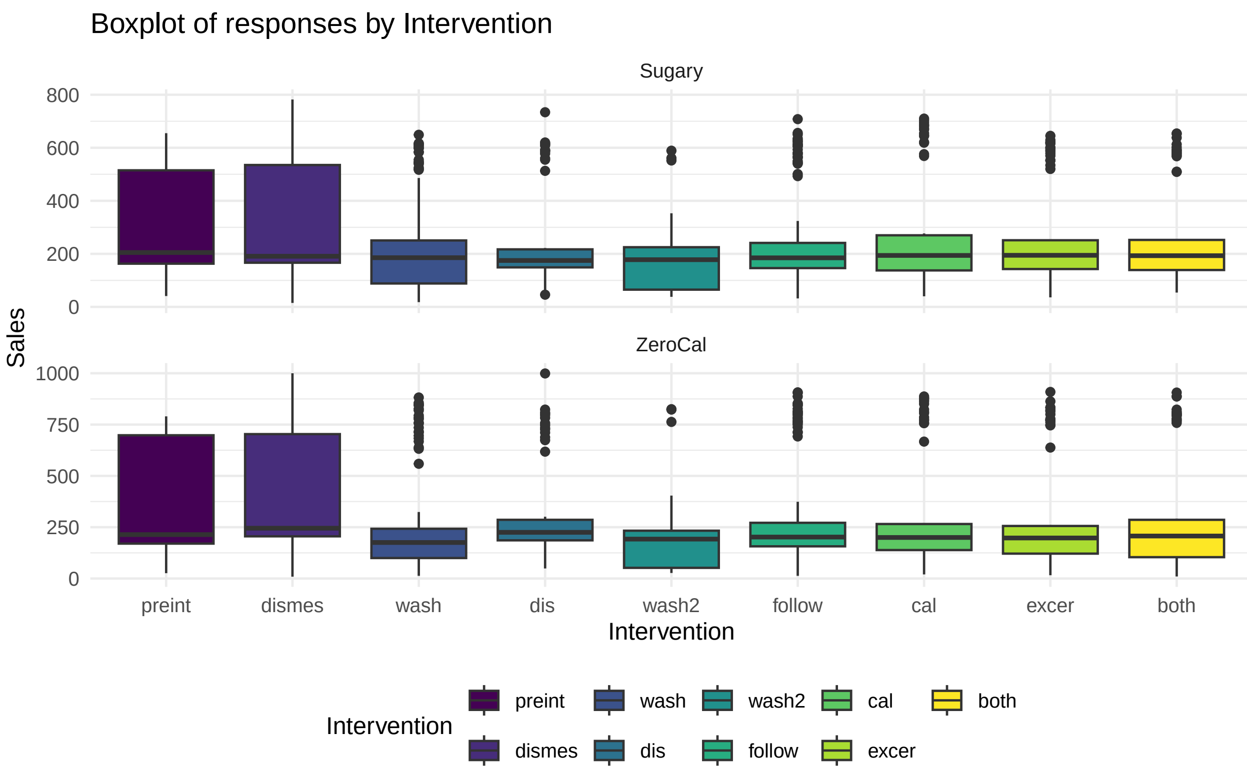
3. Exploratory Data Analysis

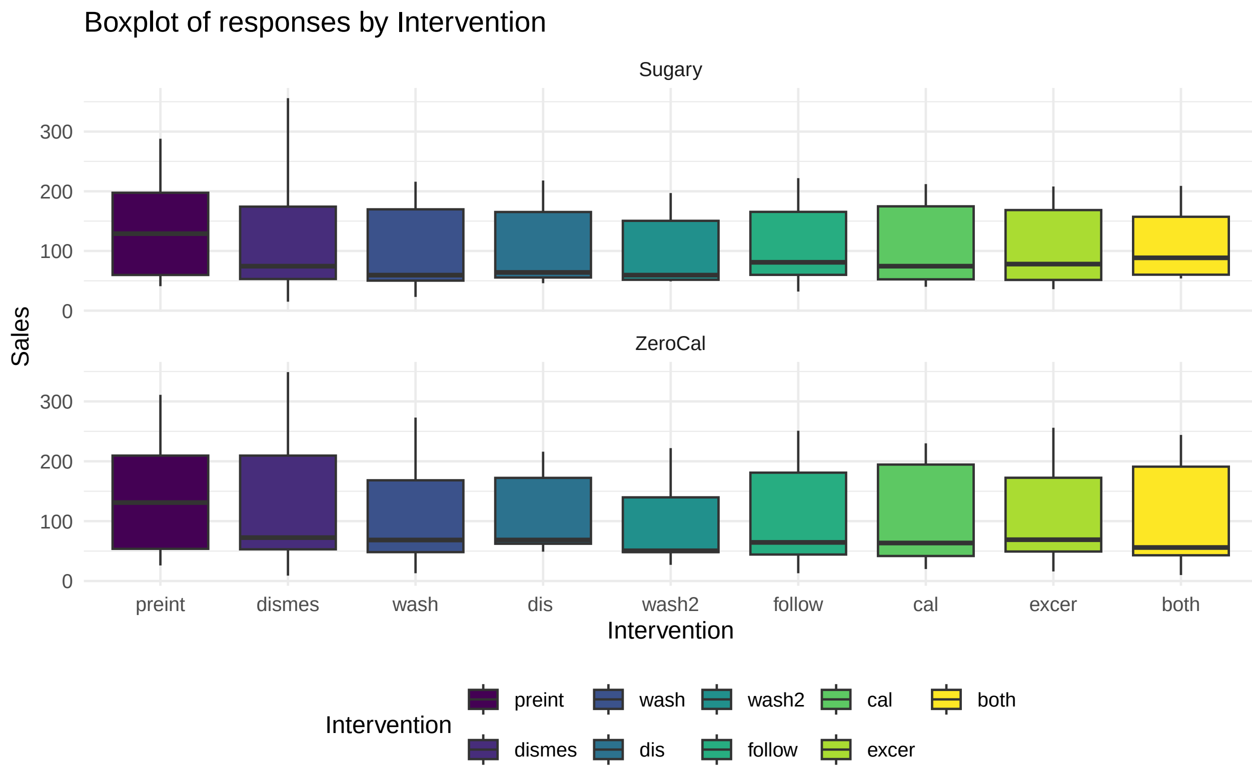
Exploratory Data Analysis (EDA) is an important first step in data analysis in uncovering underlying patterns, relationships, and outliers in the data. It ensures the subsequent analysis is built on a solid understanding of the data, thereby enhancing the reliability of the findings. To explore the data of the study, box plots and time series plots are highly important and included in the main report. The visualization in particular provide insight into the distribution of sugary and zero-calorie beverage sales, as well as their temporal trends. Additionally, Appendix A provides supplementary visualizations encompassing missing values, histograms and a correlation plot. Collectively, these exploratory techniques build a foundation and justifies the development of our formal analysis, aiming to explore the impact product labeling on beverage sales.

3.1 Boxplot

A boxplot is a method for graphically demonstrating the property of statistical distribution of numerical data. This following boxplot shows the distribution of sugary and zero-calorie beverage sales across different intervention strategies. Each boxplot captures the sales data variability with the central line denoting the median, the edges of the box indicating the interquartile range (IQR), and the whiskers extending to the furthest points that are not considered outliers. Outliers are individual points beyond the whiskers.

Note that due to the bimodality of sales, the boxplot incorrectly indicates many outliers. The interventions, labelled on the x-axis, include baseline (preint), discount & messaging (dismes), washout period (wash), the discount only (dis), second washout (wash2), follow up (follow), calorie content poster (cal), the exercise required to burn the calorie content (excer), and combination of the two (both).





3.2 Correlation Plot

The following plot investigates the correlation structure between the day of the week (`DofW`), the number of zero-calorie drinks sold (`ZeroCal`), and the number of sugary drinks sold (`Sugary`). In this plot, the size, color of the circles and number represent the strength of the correlation coefficients between the variables. The `ZeroCal` and `Sugary` variables exhibit a very strong positive correlation with a correlation coefficient of 0.97. This suggests that sales of zero-calorie and sugary drinks are closely related; when sales of one type increase, sales of the other type tend to increase in a similar fashion. Conversely, both `ZeroCal` and `Sugary` drinks show a negative correlation with `DofW`, as indicated by the coefficient of -0.39. This negative correlation suggests a tendency for the sales of both drink types to decrease on certain days of the week.

3.3 Time Series Plot

The following stacked line plot represents the sales time series of zero-calorie and sugary beverages across different sites. Each line represents the sales trajectory of one beverage type—green for zero-calorie and blue for sugary drinks. The x-axis represents time (in days), and the y-axis represents the sales volume. Dashed vertical lines indicate the start of different interventions, labelled as dismes (discount & messaging), dis (only discount), cal (calorie content poster), exer (exercise-based posters), and both. The interventions appear to influence sales, as suggested by changes in the lines' trajectories post-intervention. The plots are faceted by site, allowing for a comparative view of sales patterns across different locations.

3.3 Missing Values and Data Imbalance

The dataset contains some missing values. It is important to identify what kind of missing data exists within a dataset to better understand how to handle missingness during formal analysis. Failing to address missing data may lead to a reduction of statistical power and biased results. It appears that the missing data are missing not at random (MNAR),

Figure 1: *This plot illustrates the daily sales volumes of zero-calorie (in green) and sugary (in blue) beverages across three hospital sites over 30 weeks. The dashed and their corresponding shorthand labels mark the interventions to allow for visual assessment of their impact on beverage sales trends.*

meaning that the probability of any given observation being missing varies for unidentified reasons (e.g., hospital closures, public events). It is also important to note from the observations counts whether or not the data appear to be balanced since imbalanced data can hinder model accuracy. Balance between sites appears to be reasonable, where some imbalance is present between interventions. Namely, the 'follow', 'wash', and 'wash2' levels are imbalanced when compared to the other interventions.

4. Formal Analysis

Informed by the exploratory analysis above, three models are suggested for the formal analysis: interrupted times series, generalized estimating equations, and a linear mixed effects model.

4.1 Interrupted Times Series Analysis

The study's design, an interrupted time series across multiple sites, naturally lends itself to Interrupted Time Series Analysis (ITSA). This method effectively handles the challenges posed by non-randomized designs, isolating intervention impacts from pre-existing trends.

The ITSA method can characterize the immediate and ongoing effects of visual calorie content displays and price discounts on beverage selection. It segments the data across different intervention phases to quantify changes in zero-calorie and sugary beverage sales, offering a detailed view of each strategy's effectiveness over time.

Moreover, ITSA's adaptability enables the analysis of variations in intervention effects across sites, enabling site-specific nuances. This approach reveals whether an intervention's success is uniform or site-dependent by comparing the impact of combined interventions with that of singular strategies. Therefore, employing ITSA can address the study's core questions, demonstrating its suitability for unravelling the effects of multifaceted interventions on consumer behaviour.

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step
```

4.2 Generalized Estimating Equations

An alternative approach might be forming The Generalized Estimating Equations (GEE) model. GEE is a convenient and relatively easy to interpret method to model longitudinal data. GEE is suitable for analyzing the data of from this study since daily sales of bottled sugared beverages and zero-calorie beverages are measured repeatedly over time. GEE can be thought of as an extension of the Generalized Linear Model to longitudinal data (Columbia University). This method is particularly convenient due to its high statistical power, built-in handling of missing at random data, and its ability to account for within-subject correlation in non-normal data.

Since the study aims to investigate the amount of beverages sold at each site, this method assumes an outcome of zero-calorie and sugary beverages sold. Predictors include intervention type, site, day of the week, and total beverage sales. Site, day of the week, and total beverage sales predictors allow the model to adjust for any extraneous effects and possible sale or time trends independent of the studies interventions. Wash periods are excluded from the model and total sales are used as a control instead. Since this method models count data, a log link function is most appropriate, such as the Poisson or Negative Binomial. Models may be fitted over all sites simultaneously, or as one model per site. In the latter case, the sit factor may be excluded from the model. It is appropriate to try both to examine comparable results. Once the models are fitted, the GEE method will return coefficients for every intervention or combination of interventions taken during the study. These can be interpreted to help answer the studies main objectives. Namely, to examine how each intervention affects zero-calorie and sugary beverage sales, how sales differed by site, and comparing the impacts between different interventions on zero-calorie and sugary beverages sales. Hypothesis tests can be performed on each coefficient to test intervention and site effects. A Bonferroni correction is needed to adjust for increased risk of Type I error when making multiple statistical tests.

4.3 Linear Mixed Effects Model

Linear Mixed Effects (LME) models are useful for analyzing data structured in clusters in a longitudinal study. Within an LME model, fixed effects are those that are consistent across all observations, such as the global influence of intervention and the day of week. These effects are assumed to be the baseline of impact across all sites. Random effects, on the other hand, account for differences between sites or temporal fluctuations within a site that are not captured by the fixed effects.

In this dataset, the five intervention methods across sites could be transformed into four indicator variables, representing the presence and absence of Discount, Additional Messaging, Calorie Display, and Exercise Display. The pre-intervension period and follow-up period are treated as baseline reference observations. Variables that would be included in the LME model would be day of week, four intervention indicators and duration into the intervention, clustered by site. The model development process involves selection of fixed effect and random effect parameters, typically guided by statistical tests like the log-likelihood test.

5. Conclusions

Both exploratory and formal analyses are recommended for investigating the effectiveness of intervention strategies to promote zero-calorie beverages over their sugary alternatives involves both exploratory and formal analyses. An exploratory data analysis will help identify underlying patterns in the data, such as correlation and missingness. The formal analysis is recommended to includes three models: interrupted time series analysis, generalized estimating equations, and the linear mixed-effects model. Each of these models is capable of handling time-series and longitudinal data. Results from these models can be tested and compared for security. These analyses will answer if the data may indicate an impact on beverage sales by various labeling and discount strategies.

6. References

Columbia University Mailman School of Public Health. (n.d.). Repeated Measures Analysis. Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/repeated-measures-analysis

UCLA Statistical Consulting Group. (n.d.). Introduction to Linear Mixed Models. Retrieved March 1, 2024, from https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

University of Virginia Library. (n.d.). Getting Started with Generalized Estimating Equations. Retrieved March 1, 2024, from https://library.virginia.edu/data/articles/getting-started-with-generalized-estimating-equations

Other Resources:

- Interrupted Time Series Analysis
- Getting Started with Generalized Estimating Equations
- Linear Mixed Effect Model Tutorials:
  - LMEM tutorial (illinois.edu)
  - Chapter 17: Mixed Effects Modeling (uic.edu)
  - Introduction to Linear Mixed Models (ucla.edu)
  - Mixed Effects Models (bodywinter.com)
  - Mixed Effects, lme4 Tutorial
  - Fitting Linear Mixed-Effects Models Using lme4
- ASDA Teaching Recordings:
  - Exploratory Data Analysis
  - Study Design and Data Collection Essentials
  - Mixed Effects Models
- Multiple Testing Correction Techniques:
  - Multiple Testing Correction Techniques
  - An Article on the Multiple Testing Problem
- Model Assumptions and Interpretations:
  - Linear Mixed Effects Model Assumptions
  - DHARMa: a Residual Diagnostics Tool for LMEM
  - Interactions Explained
  - Understanding Interactions between Categorical and Continuous Variables
  - Interaction Plots
  - Interaction Plots
  - Testing for Interaction Significance
  - Package ‘emmeans‘ for Post-hoc Analysis and Pairwise Comparisons of the Interaction Effect
  - Interaction Analysis in R
  - Comparisons and Contrasts in emmeans

Appendix A: Figures

The following sub section contain some additional information and analysis results.

Histogram Plots

The following histogram plots show the frequency distribution of sales for Sugary (in purple), Zero-Calorie (ZeroCal, in teal), and Total (in yellow) beverages. The x-axis of each histogram represents the sales volume, while the y-axis indicates the count of observations within each sales range. The pattern in all histograms is similar: most sales numbers cluster at the lower end of the scale, suggesting a higher frequency of days with fewer sales; however, the sales histograms exhibit a second weaker mode, indicating two common sales volumes across the observed period.

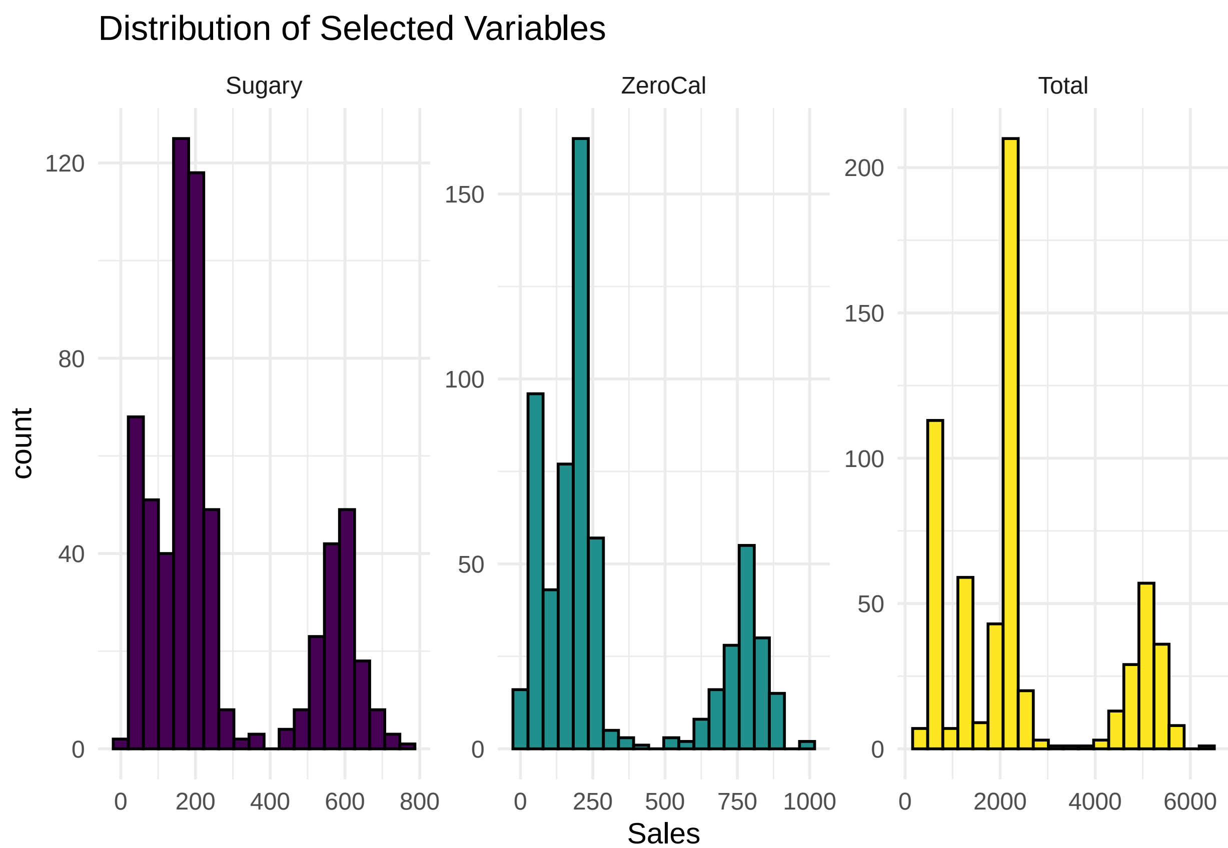


Figure 2: *Sales Distribution Analysis: Histograms displaying the frequency of sales for Sugary (purple), Zero-Calorie (teal), and Total combined (yellow) beverages. Each histogram reveals the distribution pattern of sales volumes, highlighting the bimodal nature of sales across all types.*

Scatter Plot

The following scatter plot depicts the relationship between zero-calorie and sugary beverage sales at three different hospital sites: A or chop (purple), B or HF (blue), and C or NS (yellow). The x-axis represents zero-calorie beverage sales, and the y-axis represents sugary beverage sales. A dashed line, suggesting the line of equality, indicates where the sales for both types would be equal. Points above the line indicate higher sugary beverage sales when compared to zero-calorie ones, and points below the line indicate the opposite. The clustering of points towards the upper right suggests that for higher sales volumes, sugary beverages tend to sell as much as or more than zero-calorie options, particularly in site A (chop). The plot reveals variability in the sales patterns across sites, with the HF site having a more direct correlation between ZeroCal and Sugary sales when compared to other sites.

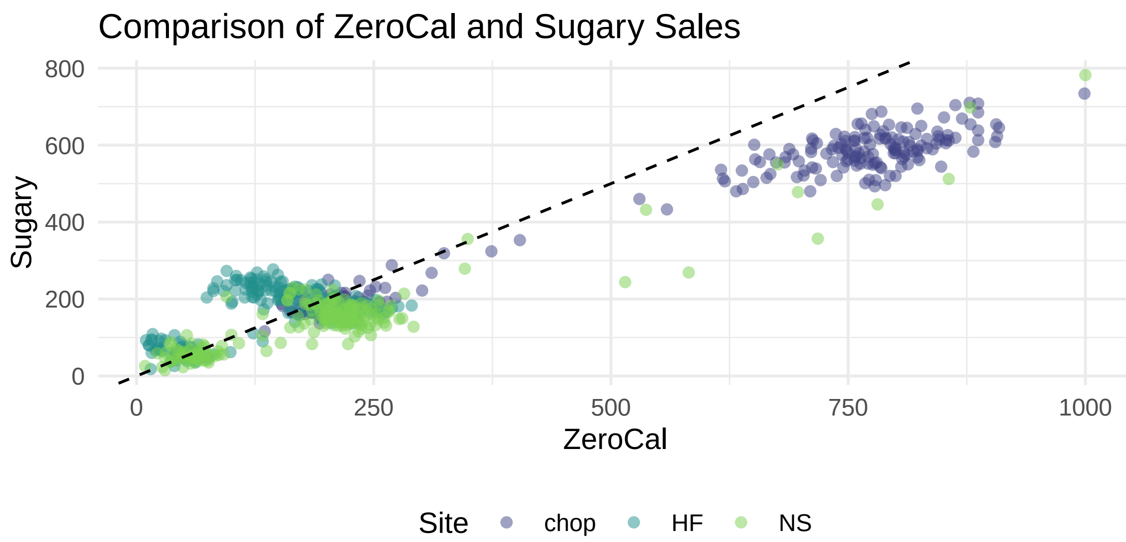
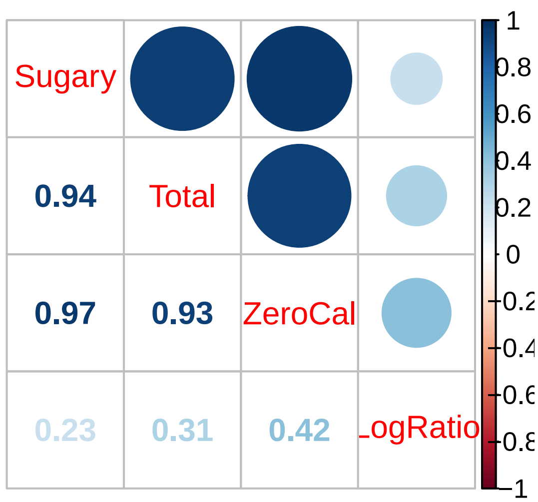


Figure 3: This scatter plot contrasts zero-calorie and sugary beverage sales, colour-coded by the site. Each point represents the paired sales data for a given day, with the site-specific colour coding (chop in purple, HF in blue, NS in yellow) illustrating the sales trend at each location. The dashed diagonal line marks the parity where the sales of both beverage types are equal. Deviations from this line highlight the predominance of one beverage type over the other in daily sales.

Correlation Plot

The following plot visualizes the correlation between numeric variables.



Missing Values

The following plot visualizes the missing values.

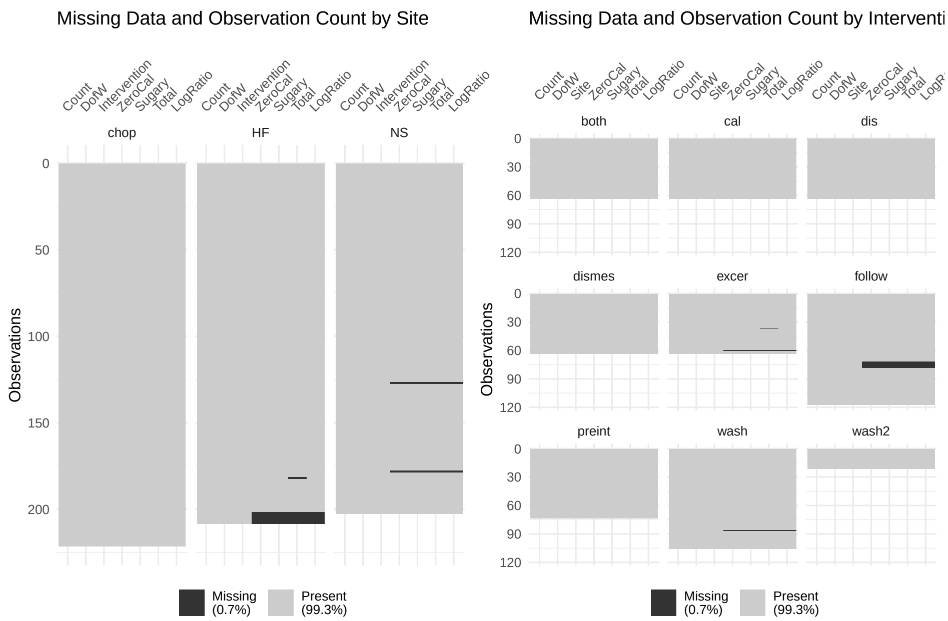


Figure 4: *This plot provides insight into the frequency of missingness within the dataset. Black indicates missing data. Additionally it shows the quanitity of data available by site and by intervention.*

Appendix B: Models

ITSA Model

Interrupted Time Series Analysis (ITSA) with segmented regression is a statistical technique tailored for quasi-experimental designs that involve interventions at known time points. ITSA is particularly suited for this study where interventions are sequentially introduced in a multi-site setting and where the main interest lies in the impact on sales of zero-calorie (ZeroCal) and sugary (Sugary) beverages.

The general form of the segmented regression model for ITSA applied to this context can be expressed as:

$$Y_t = \beta_0 + \beta_1 T_t + \sum_{k=1}^{K} (\beta_{2k} I_{kt} + \beta_{3k} T_{kt} I_{kt}) + \epsilon_t$$

Where: - Y_t is the sales of beverages at time t. - T_t is the time since the start of the study (time trend). - I_{kt} is an indicator for intervention k (0 before intervention k, 1 after intervention k). - T_{kt} is the time since intervention k started, multiplied by the intervention indicator. - β_0 is the intercept, representing the baseline level of sales. - β_1 is the coefficient for the time trend, representing the pre-intervention trend of sales. - β_{2k} is the change in level immediately after intervention k. - β_{3k} is the change in trend after intervention k. - K is the total number of interventions. - ϵ_t is the error term which is assumed to be normally distributed with mean zero and constant variance.

This model can be fitted separately for ZeroCal and Sugary sales to ascertain the unique effects of interventions on each type of beverage. The model can also be expanded to account for auto-correlated errors which are common in time series data, by incorporating an AR(1) process or other suitable autocorrelation structures.

For the investigation of site-specific effects, random effects or fixed effects models can be used. A random effects model would be suitable if we assume that the sites are a random sample from a larger population, with the model taking the form:

$$Y_{it} = \beta_0 + \beta_1 T_t + u_i + \sum_{k=1}^{K} (\beta_{2k} I_{kt} + \beta_{3k} T_{kt} I_{kt}) + \epsilon_{it}$$

Where u_i is the random effect for site i and ϵ_{it} is the within-site error term.

By contrast, a fixed effects model would treat each site as a unique entity and estimate site-specific intercepts:

$$Y_{it} = \beta_{0i} + \beta_1 T_t + \sum_{k=1}^{K} (\beta_{2k} I_{kt} + \beta_{3k} T_{kt} I_{kt}) + \epsilon_{it}$$

With β_{0i} being the intercept for site i, allowing for different baseline sales levels at each site.

The interaction terms $\beta_{3k} T_{kt} I_{kt}$ are critical for evaluating the sustained impact of interventions over time. If these coefficients are significantly different from zero, it suggests that the interventions had an effect beyond an immediate jump or drop in sales, altering the underlying trend of beverage sales.

To evaluate the combined effect of interventions, interaction terms between interventions can be included:

$$Y_{it} = \beta_0 + \beta_1 T_t + u_i + \sum_{k=1}^{K} \beta_{2k} I_{kt} + \sum_{k=1}^{K} \beta_{3k} T_{kt} I_{kt} + \sum_{k<l} \beta_{4kl} I_{kt} I_{lt} + \epsilon_{it}$$

Here, β_{4kl} captures the combined effect of interventions k and l when both are in effect.

Lastly, the model can be augmented with covariates to control for other factors that may influence sales, such as seasonal effects or marketing campaigns. These covariates can be time-varying and should be included in the model if they are thought to confound the relationship between the interventions and sales.

GEE Model

Let $\mathbf{Y_i}$ be the outcome variable for beverage i (zero-calorie or sugared) sales.

Let $g(\cdot)$ be the log link function (Poisson or Negative Binomial). Then, for design matrix **X** including all revenant predictors, the model can be written more explicitly as

$$g(\mathbb{E}[Y_i]) = \beta_{0i} + \beta_{1i}(\text{Discount}) + \beta_{2i}(\text{Discount + Messaging}) + \beta_{3i}(\text{Calorie Messaging})$$

$$+\beta_{4i}(\text{Exercise Messaging}) + \beta_{5i}(\text{Both Calorie Messaging}) + \beta_{6i}(\text{site B}) + \beta_{7i}(\text{site C})$$

$$+\beta_{8i}(\text{Day of Week}) + \beta_{9i}(\text{Total sales})$$

Then β_{0i} is the intercept. (Discount) and (Discount + Messaging) are each dummy variables to represent the discount intervention without messaging, and discount with messaging respectively. (Site B) and (Site C) are also dummy variables to indicate the site, with the baseline being site A.

LME model

Let y_i be the vector of measurements of response for site i, then a Linear Mixed Effect model is written as

$$y_i = X_i\beta + Z_i b_i + e_i$$

where β contains fixed effects, b_i contains random effects parameter for site i. In this case, the expanded general parameters would be

$$y_{ij} = \beta_0 + b_{ij}^0 + (\beta_1 + b_{ij}^1)Treatment + (\beta_2 + b_{ij}^2)DayofWeek + (\beta_3 + b_{ij}^3)Duration + e_{ij}$$

where y_{ij} is the sales of beverage at site i day j, $ □^0 + b_{ij}^0$ is the fixed and random terms for intercept, $(\beta_1 + b_{ij}^1)Treatment$ consists of the fixed and random terms for all four indicator variable for treatment, $(\beta_2 + b_{ij}^2)DayofWeek$ consists of the fixed and random terms for each day of the week, $(\beta_3 + b_{ij}^3)Duration$ represents the fixed and random effect of slope for duration into a specific treatment.

Contributions

Parham Pishrobat (71097927): Introduction, Data, ITSA, Other EDA Plots, Formatting

Johnson Chen (85784080): LMEM, Correlation, Appendix, Formatting

Sarah Masri (97415681): GEE, Missing Values, Conclusion, Formatting