Statistical Advice on the Effect of Interventions on Beverage Sales

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

March 29, 2024

Contents

1 Introduction **2**

2 Data Description and Summaries **2**

3 Exploratory Data Analysis **2**

3.1 Missing Values and Data Imbalance 4

4 Formal Analysis **5**

5 Conclusions **5**

6 References **6**

A Appendix A **7**

A.1 Histogram Plots 7

A.2 Scatter Plot 7

A.3 Boxplots 8

A.4 Correlation Plot 8

A.5 Missing Values 8

B Appendix B **8**

B.1 ITSA Model 8

C Contributions **12**

1 Introduction

Concerns around sugar consumption and its health implications have prompted many interventions to encourage consumers to 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: laebls 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 seeks to understand if the effectiveness of those strategies varies across different hospital 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.

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 variables zero-calorie, sugary, and all drinks, sale location, day of the week, and the type of intervention applied. 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. The total sales information is kept in the study to represent the overall patterns of beverage sales irrespective of their sugar content. 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. To explore the data of the study, box plots and time series plots are highly important and included in the main report. These visualizations 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 justification to the development of the formal analysis, aiming to explore the impact product labeling on beverage sales.

Figure 1 represents the sales time series of zero-calorie and sugary beverages across different sites, colored by the Intervention. Both sugary drink and zero calorie drink sales trends are different across different interventions, indicating that different interventions do have different impacts on sales. The sales trends are different across sites, and the effect of intervention appears to be different across sites as well.

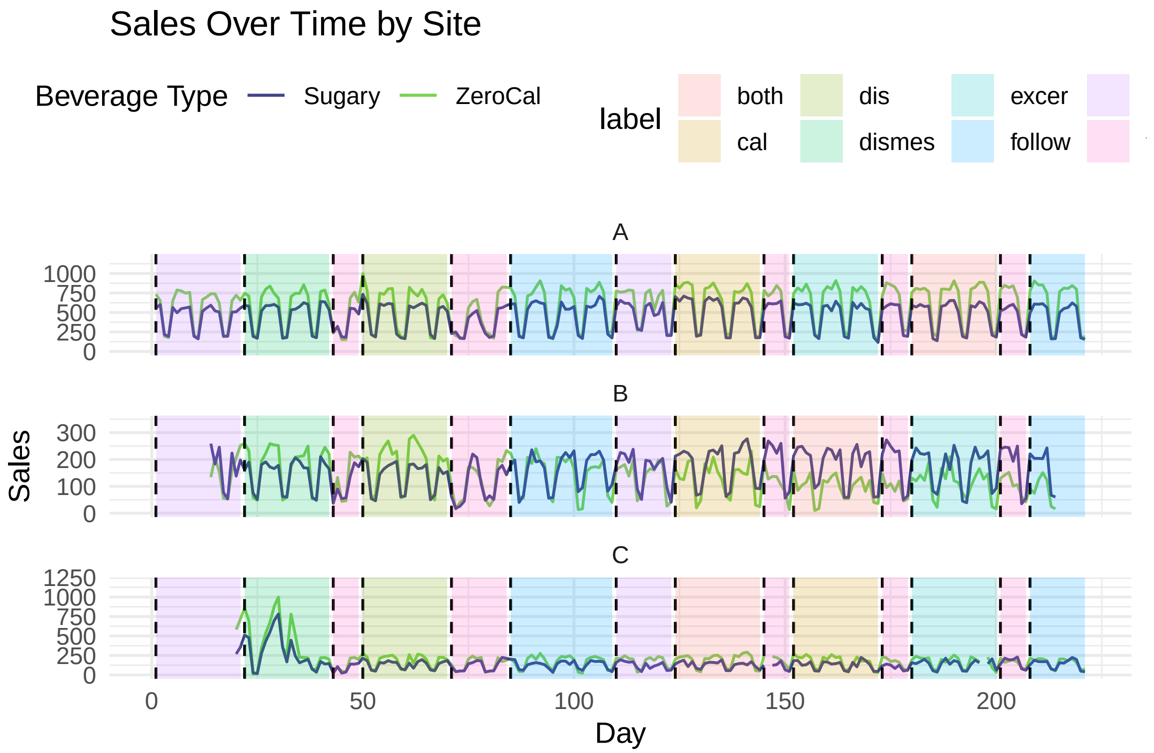

Figure 1: *Time Series plot illustrating the daily sales volumes of zero-calorie (in green) and sugary (in blue) beverages across three hospital sites over 30 weeks. The dashed and their corresponding shorthand labels mark the interventions to allow for visual assessment of their impact on beverage sales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

A boxplot is a method for graphically demonstrating the property of statistical distribution of numerical data. Figure 2 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, and the whiskers extending to the furthest points that are not considered outliers.

Note that bimodality of sales exists in the data. Specifically, there are considerably more beverage sales on weekdays than on the weekend. To better represent the data, the weekend sales are excluded from this plot. Boxplots for all and weekend sales respectively are present in the appendix (Figure 5). The interventions, labelled on the x-axis, include baseline, discount & messaging, washout period, discount only, follow up, calorie content poster, exercise required to burn the calorie content poster, and combination of the two poster types. It is observable that the maximum and median number of sales during the week varies with intervention.

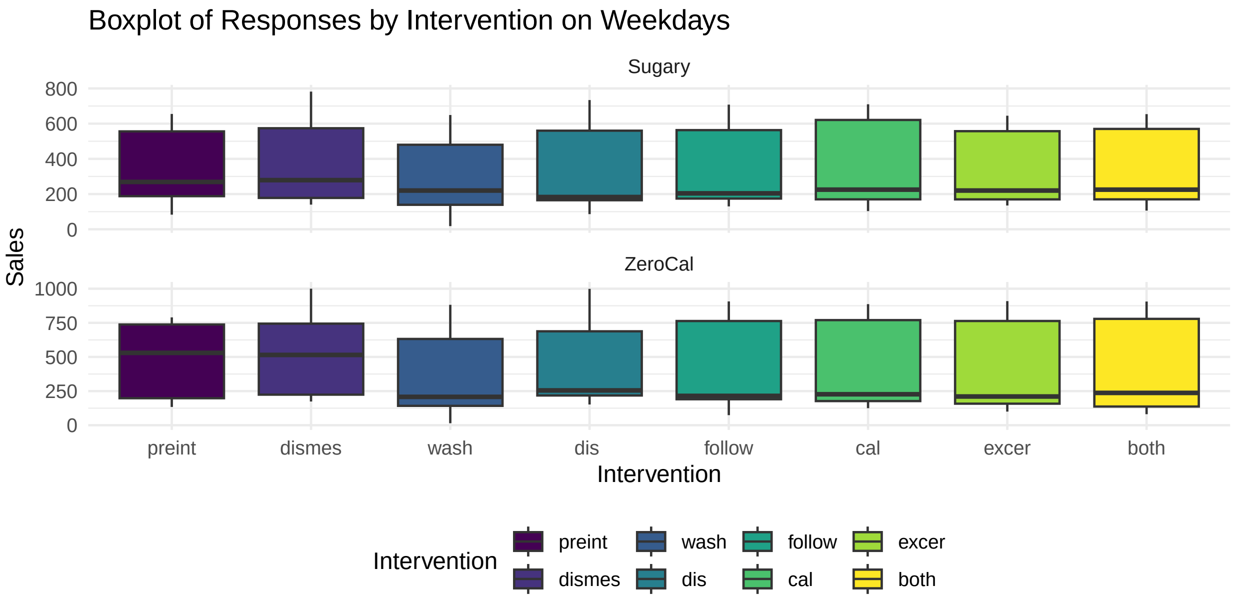

Figure 2: *Boxplots display the sales distribution of Sugary and Zero-Calorie beverages across various interventions. Note that the plot only illustrates weekday sales.*

3.1 Missing Values and Data Imbalance

The dataset contains some missing values, as obseved in Figure 7. It is important to identify what kind of missing data exists within a dataset to better understand how to handle missingness during formal analysis. There are 9 days of unrecorded zero-calorie and sugary beverage sales, 7 of which represent the last week of the study at site B. Failing to address missing data may lead to a reduction of statistical power and biased results. This report removed the last week of study at site B, treating it as not recorded entirely. The rest of missing values are considered Missing At Random (MAR) due to the small amount.

It is also important to note whether of not the dataset reflects a roughly equal number of observations between the sites and interventions respectively. If one class in either the site or interventions are represented disproportionately, then models used may become biased towards the most frequently seen class. Balance between sites appears to be reasonable, where some imbalance is present between interventions. Namely, some of the no-intervention levels are imbalanced when compared to the other interventions, this does not have a significant impact on the analysis of the interventions.

4 Formal Analysis

This study employs Interrupted Time Series Analysis (ITSA) with Linear Mixed Models (LMMs) to assess the impact of labelling interventions on beverage selection. ITSA is a statistical technique to characterize the temporal changes before and after interventions. This approach is especially pertinent to the current analysis as the study employs an interrupted time-series multi-site quasi-experimental design. Although ITSA can be carried out using a variety of linear models, LMMs are preferred to allow for the inclusion a random effect to accommodate for variability across collection sites. LMMs manage complex data, handle missing values, and capture both the fixed effects of interventions and the random effect of variability across sites. The logarithm of the ratio between zero-calorie and sugary beverage sales is considered as outcome of interest because it symmetrically quantifies relative preference shifts, where a higher log ratio indicates a preference for healthier options and vice versa. Correlation structure is defined based on the temporal spacing between observations, reflecting the notion that as time between sales data points increases, their correlation decreases in a manner that may not follow a simple pattern.

Upon initial analysis, which focuses solely on the direct before-and-after impacts of interventions, the study does not identify any statistically significant intervention effects. This outcome highlights the inadequacy of the method in capturing the dynamic and evolving nature of consumer responses over time. Subsequently, the analysis is refined to incorporate interactions between interventions and time, providing a richer understanding of how consumer preferences for beverages evolve in response to interventions. This nuanced approach reveals that exercise-focused messaging, in particular, has a significant and increasingly positive influence on consumer preferences towards healthier beverage options over time. Furthermore, while discount strategies initially appear effective, their impacts do not sustain over time. Moreover, the analysis points to the significance of the day of the week, with consumer behaviour exhibiting notable shifts during weekends, suggesting either variations in beverage preferences between weekdays and weekends or change in the population during weekends.

These insights draw attention to the challenges of influencing dietary choices and the variable success rates of different intervention strategies. Specifically, the effectiveness of visually linking calorie consumption with physical activity emerged as a significant factor in promoting zero-calorie beverage choices. This variation in intervention effectiveness, coupled with the importance of intervention timing demonstrate the strategic value of timing in intervention planning. While the methodology employed provides valuable perspectives on intervention impacts, it operates under certain assumptions, such as linear temporal changes and uniform intervention effects across sites, which may simplify the complex dynamics of actual consumer behaviour. This complexity points to the challenges involved in designing universally effective health promotion interventions, underscoring the need for targeted, contextually aware strategies to encourage healthier consumer choices effectively.

5 Conclusions

In conclusion, our analysis demonstrates the value of integrating ITSA with LMM for evaluating public health interventions in shifting consumer behaviour. While our findings highlight the potential of certain interventions to positively influence consumer behaviour, they also point to the importance of considering temporal patterns in the design and implementation of health promotion strategies. The refined analysis, which incorporates interactions between interventions and time, reveal the progressive influence of exercise-related messaging of the calorie content on shifting consumer preferences towards zero-calorie beverages. Furthermore, the discounting strategies show a significant impact after commencing the intervention but fail to sustain the impact. Variations in preference between weekdays and weekends further highlighted the importance of timing and pointing to different target populations.

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

A Appendix A

The following sub section contains additional figures and their analysis results.

A.1 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.

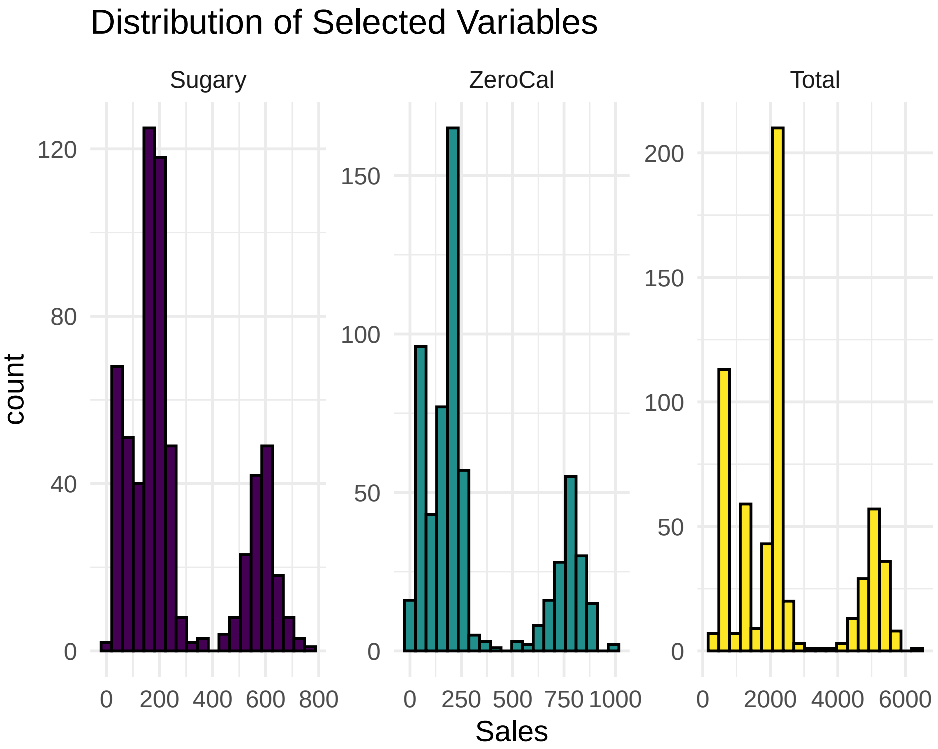

Figure 3: *Sales Distribution Analysis: Histograms displaying the frequency of sales for Sugary (purple), Zero-Calorie (teal), and Total combined (yellow) beverages. Each histogram reveals the distribution pattern of sales volumes, highlighting the bimodal nature of sales across all types.*

A.2 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.

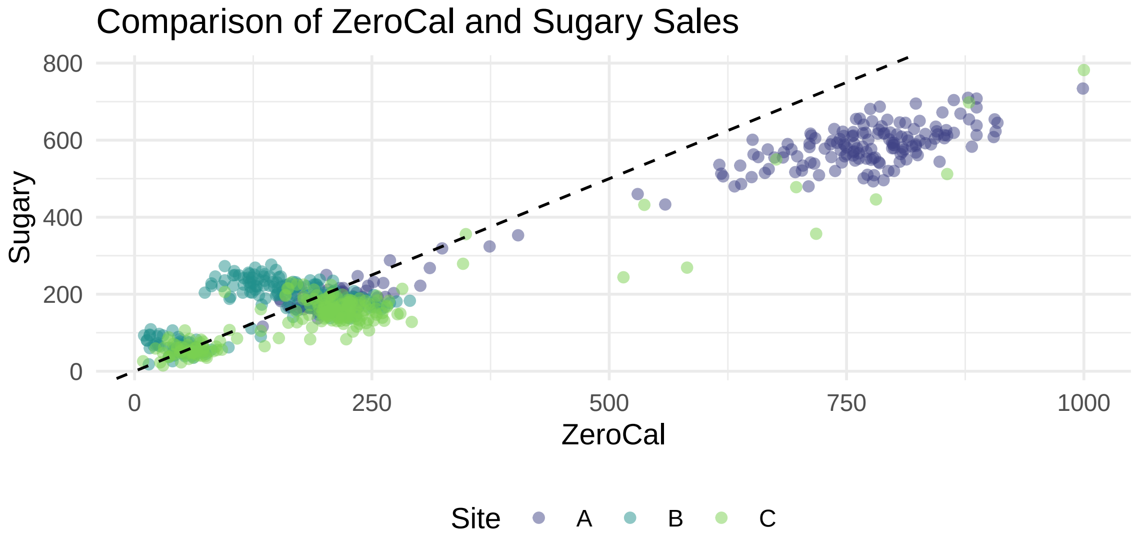

Figure 4: This scatter plot contrasts zero-calorie and sugary beverage sales, colour-coded by the site. Each point represents the paired sales data for a given day, with the site-specific colour coding (chop in purple, HF in blue, NS in yellow) illustrating the sales trend at each location. The dashed 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.

A.3 Boxplots

This boxplot shows the sales distribution data from weekends across sites. (Figure 5).

A.4 Correlation Plot

This correlation plot (Figure **??**) investigates the correlation structure between numeric variables within the dataset. In this plot, the size, color of the circles and number represent the strength of the correlation coefficients between the variables.

A.5 Missing Values

The following plot visualizes the missing values.

B Appendix B

B.1 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

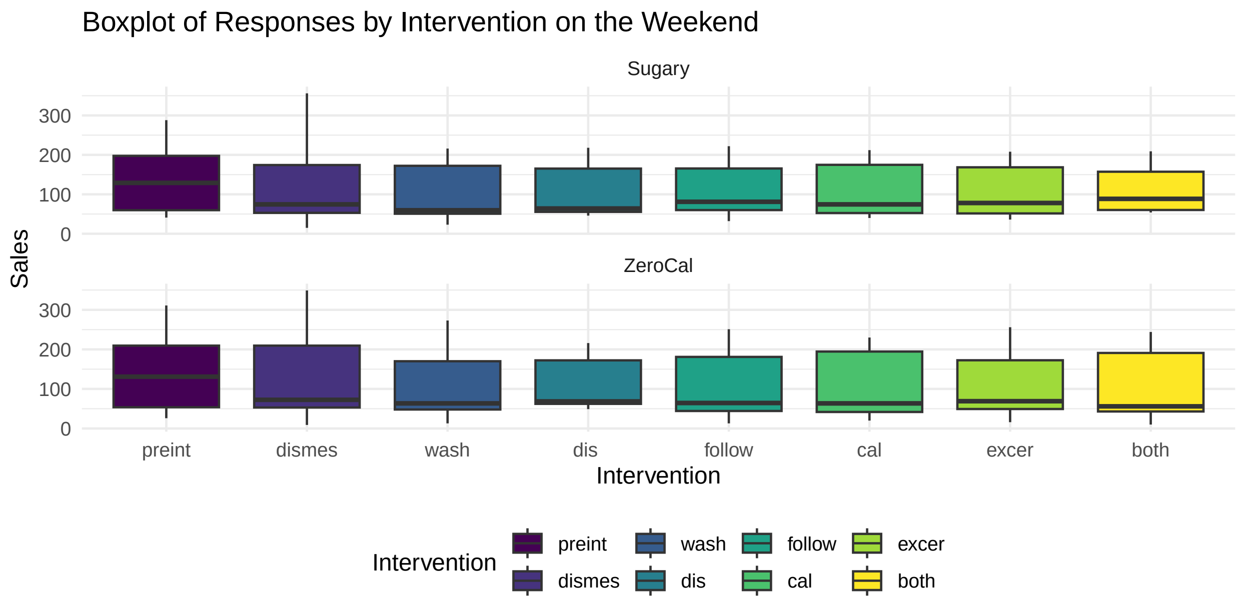

Figure 5: *The boxplots display the sales distribution of Sugary and Zero-Calorie beverages across various interventions. Each intervention is colour-coded and shows the range of sales data with the central line representing the median sales. Note that the plots only illustrates weekday sales.*

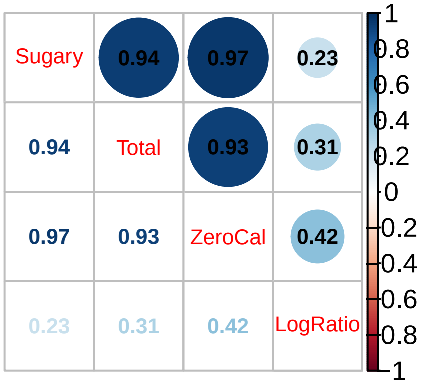

Figure 6: *Correlation plot displaying the correlation between numeric variables*

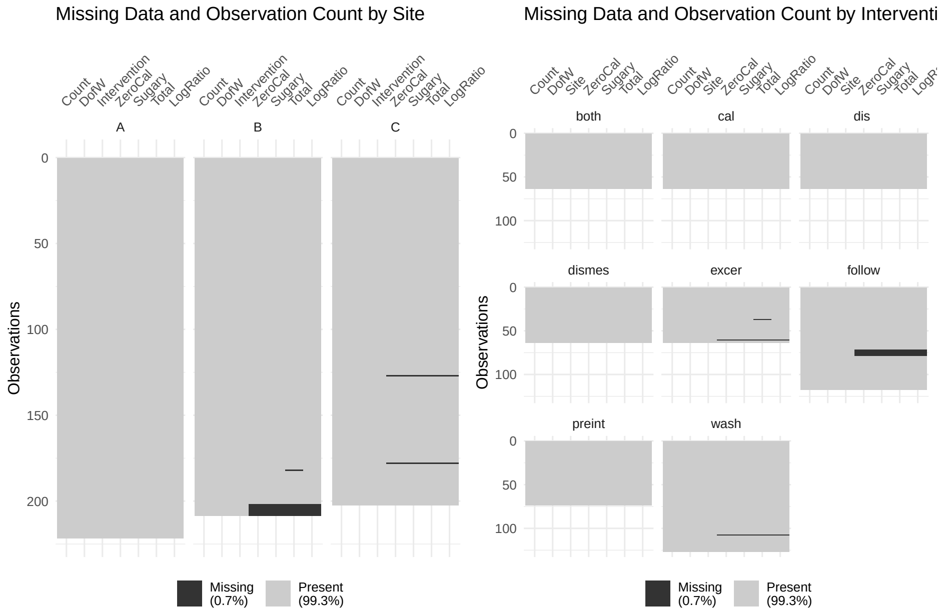

Figure 7: *This plot provides insight into the frequency of missingness within the dataset. Black indicates missing data. Additionally it shows the quanitity of data available by site and by intervention.*

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.

C 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, edits after peer review, proofread.