STAT550 Homework No 2: Advise for Evaluating Interventions on Sugary and Zero-Calorie Beverage Consumption

Son Luu (id), Xihan Qian (54285556) and Javier Mtz.-Rdz. (94785938)

March 1, 2024

1 Introduction

Zero-calorie beverages offer an alternative to sugary drinks that can help to avoid the harmful effects of artificial sweeteners. Therefore, it is important to understand what actions can motivate people to switch to these products. The aim of this study is to assess the impact of messaging and discounts on the buying behaviour of zero-calorie and sugary drinks. In particular, it evaluates different interventions, such as discounts with and without explanation, messaging that displays the caloric content, messaging that shows the equivalent fiscal activity, and a combination of both.

The primary question under consideration is the effect of each intervention on the consumption of sugary and zero-calorie drinks. Specifically, the study will explore the direction, size, comparison, and impact of interventions on each site, as well as how different interventions interact and compare with each other. To answer these questions, this document discusses the characteristics of the data collected, explores its behaviour, and performs a statistical assessment.

2 Data Description and Summaries

To evaluate the effects of interventions, the study gathered data on beverages sold at four cafeterias and three convenience stores across three sites. The dataset records daily sales of these beverages for a period of 221 days, from October 27 to May 23 (#TODO: corroborate dates), and summarizes the data by site. Nevertheless, the observations for each site start on different dates: site A starts on day 1, site B on day 14, and site C on day 20. In total, there are 631 observations in the dataset.

The dataset includes variables related to time, sales, place, and intervention. The time variables are the count of days since the start of the study and the day of the week. The sales variables include zero-calorie, sugary, 100% juice, orange juice, sports, and total beverages sold, but only zero-calorie and sugary beverages are considered for this analysis. As for place and intervention, there is a variable for each one. Table 1 summarizes the variables available in the dataset, their classification, and how they are measured.

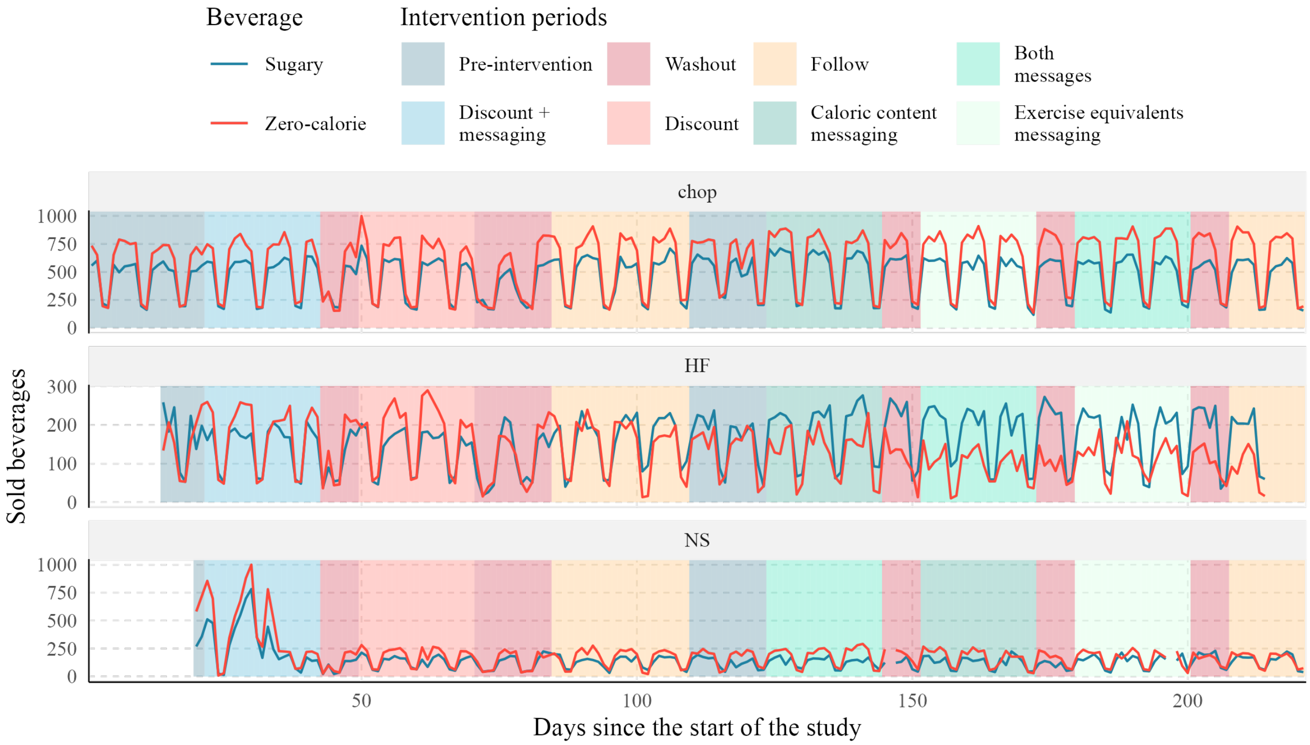
Table 1: **Description of variables**

Variable	Type	Unit
Day of the quasi-experiment	Continuous	-
Day of the week	Continuous	-
Site	Categorical	-
Intervention	Categorical	-
Sugary beverages sold	Continuous	-
Zero-calorie beverages sold	Continuous	-
Other beverages sold	Continuous	-

In addition to the periods that were not recorded at the beginning of the study in sites B and C, there are nine missing values for sales of zero-calorie and sugary beverages. The missing observations correspond to the last week of site B and two days of site C. Aside from the missing information at the beginning and end of the study, the missing values are unrelated to any specific factor. Furthermore, the sales data for other beverages and the total amount have several missing values, but they do not affect this analysis.

3 Exploratory Analysis

Given that the data consists of a time series of sales across three sites, it was necessary to carry out a time-based analysis. In that sense, Figure 1 helps visualize the beverages sold and the shadows behind the lines display the distinct intervention periods. Additionally, Section 7.1 visualization and tests about the relationship among variables, the effect from past observations in the new data points and the decomposition of sugary and zero-calorie sales series in the change by the mean level (trend), the periodicity of the data (seasonal variation) and factors that do not show a pattern (random variation).



Source: client submission.

Figure 1: **Sale of sugary and zero-calorie drinks by intervention**

In particular, Figure 1 shows some important characteristics of the dataset. Firstly, the measurements for each site began at different times. Secondly, the third site experienced a significant increase in sales during most of the first intervention, but afterwards, sales remained at a lower and more stable level. Thirdly, the order of the three calorie messaging interventions was different for each site. Lastly, it is evident from the data that there is a weekly seasonal effect.

4 Formal analysis

4.1 Pre-analysis

Before conducting the statistical analysis, it is important to address some elements that were identified during the Exploratory Analysis that could potentially impact our analysis. These include the weekly effect, different lengths of the pre-intervention observations, and unnecessary information. Firstly, we aim to capture the intervention effects and not the number of day's effects. Therefore, we can remove the weekly impact by either using a percentage of sales or decomposing the time series, as identified in the Exploratory Analysis, and removing the seasonal variation as indicated by Chatfield and Xing (2019). While the percentage approach may be easier, it is recommended to use the decomposition method as the variable containing the total items sold may have inconsistencies that suggest unreliable information. For example, day 199 in the NS site has more zero-calorie and sugary beverages sold than total items, and days 50-53, 107, 184, and 196 in the NS site have decimals.

Secondly, as sites two and three have a short pre-intervention period, it is necessary to compensate for the missing information. This can be done by combining the follow-up and pre-intervention categories into a new category that will serve as the baseline for the study. Lastly, we are not interested in the washout periods that are used to reduce the effects of previous interventions. Hence, this data can be ignored in the study.

4.2 Difference-in-Difference Estimation

DID, or Difference-in-Differences, is a quasi-experimental method that utilizes longitudinal data from intervention and control groups (Angrist and Pischke 2008). This approach is particularly effective in this design, as it aligns with the quasi-experimental nature of the study. This method aims to create a suitable counterfactual for estimating causal effects by comparing changes over time between the groups. The key assumption behind DID is that in the absence of the intervention, the average change in the outcome for the intervention group would have been the same as the average change for the control group.

Before applying the model, there are a few assumptions to be checked. Most importantly, there need to be parallel trends between the intervention and the control group. Moreover, at the baseline, the allocation of the intervention should not depend on the outcome. The model itself would be formulated as: $Y = \beta 0 + \beta 1 * [\text{ Time }] + \beta 2^{*}[\text{ Intervention }] + \beta 3^{*}[\text{ Time * Intervention }] + \varepsilon$. As there are different forms of interventions in the study, including discounts, calorie messaging, and combinations of the two, a linear model can be run first to determine which interaction terms (Time * Intervention) should be included.

In the proposed model, the coefficients of the interaction terms are crucial for assessing the efficacy of interventions. The analysis will concentrate on these coefficients, which vary depending on the study's goals. For instance, to determine if interventions result in higher sales of zero-calorie beverages, all interventions can be collectively analyzed. A positive and significant coefficient for this interaction term would signify a successful increase in sales attributed to the intervention. Similarly, to explore whether combined interventions have a more substantial effect, the focus would shift to the coefficient of the interaction term representing the combined interventions.

5 Conclusion

6 References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Chatfield, Christopher, and Haipeng Xing. 2019. *The Analysis of Time Series: An Introduction with R.* Seventh edition. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton: CRC Press, Taylor & Francis Group.

7 Appendices

7.1 Detailed Explorarory Analysis

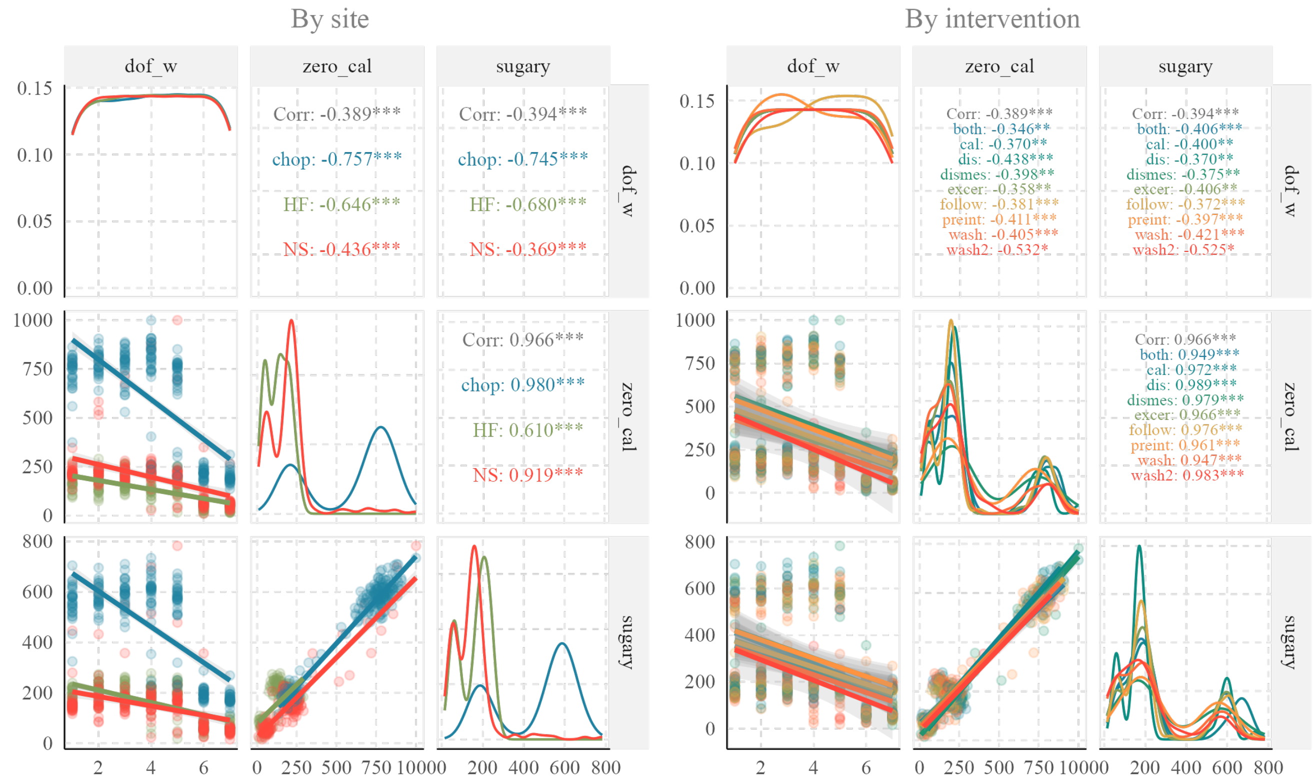
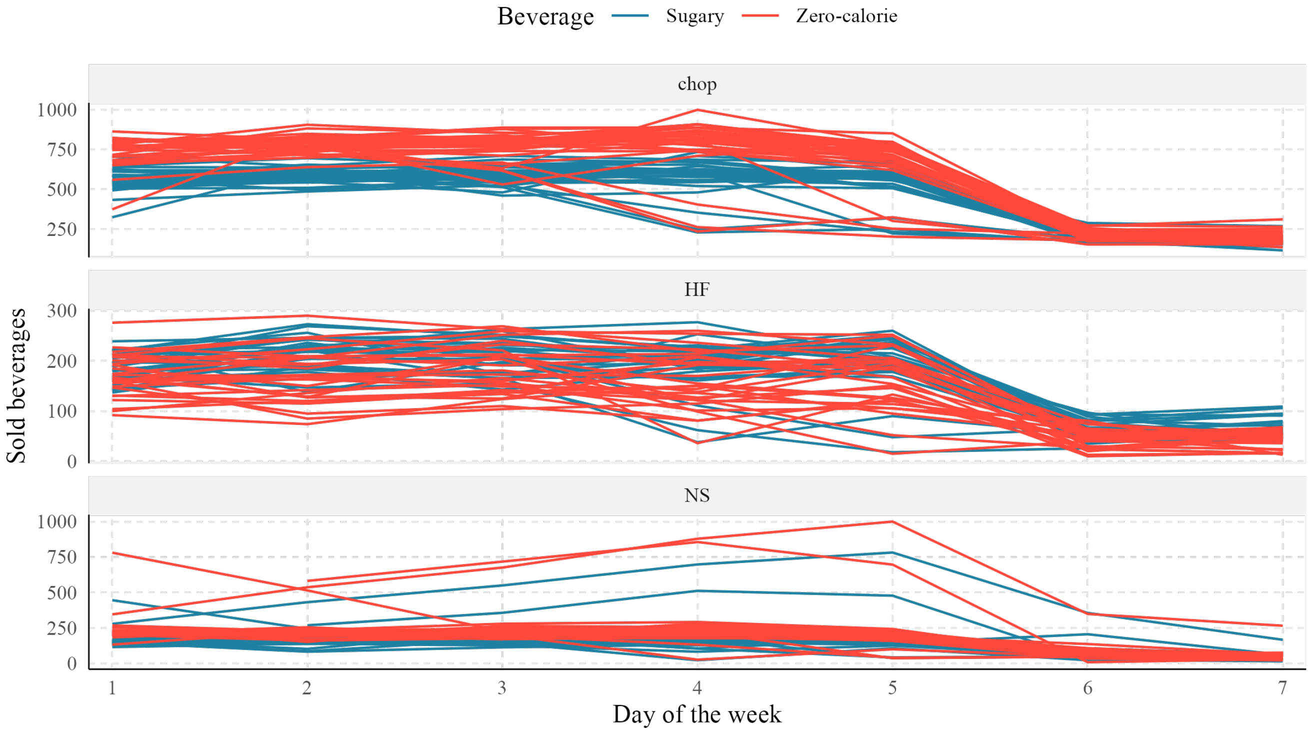


Figure A.1: **Correlation matrix of variables**



Source: client submission.

Figure A.2: **Sale of sugary and zero-calorie drinks by week and site**

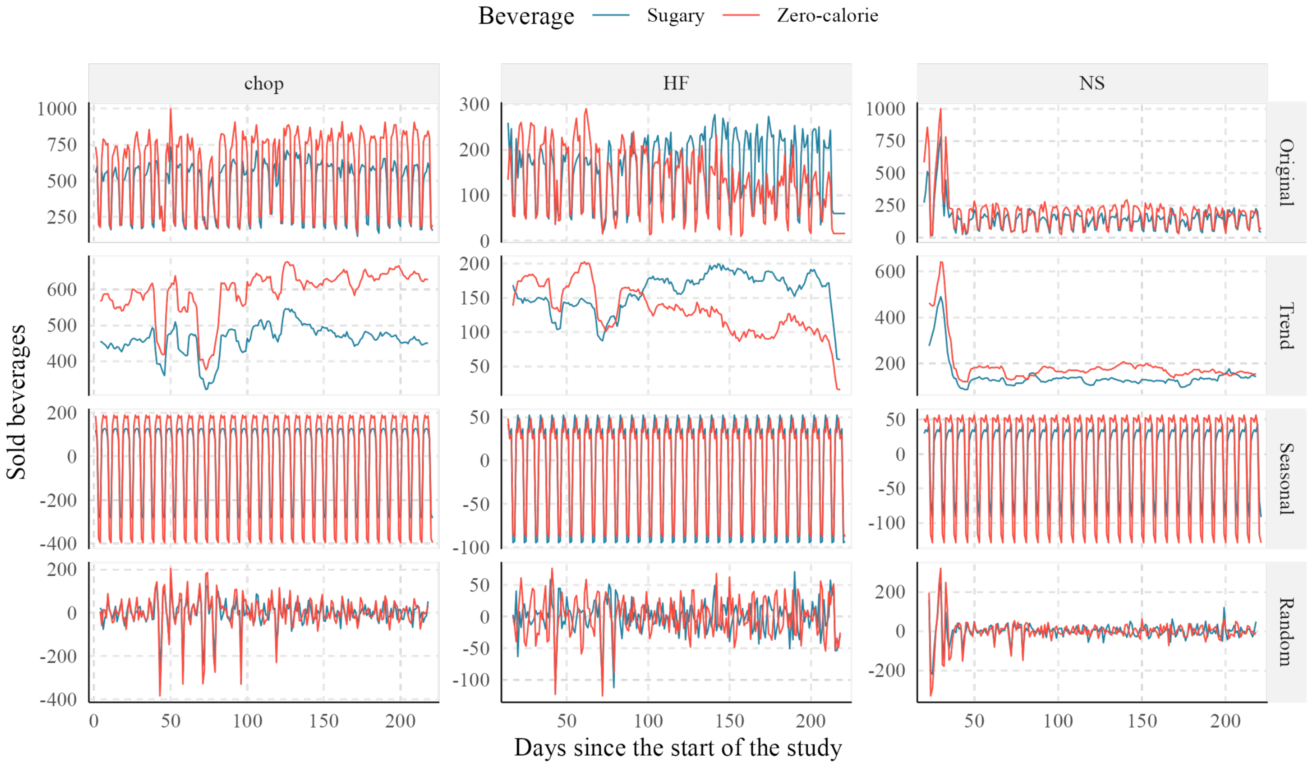


Figure A.3: **Decomposition Analysis of Sales for Sugary and Zero-Calorie Beverages**

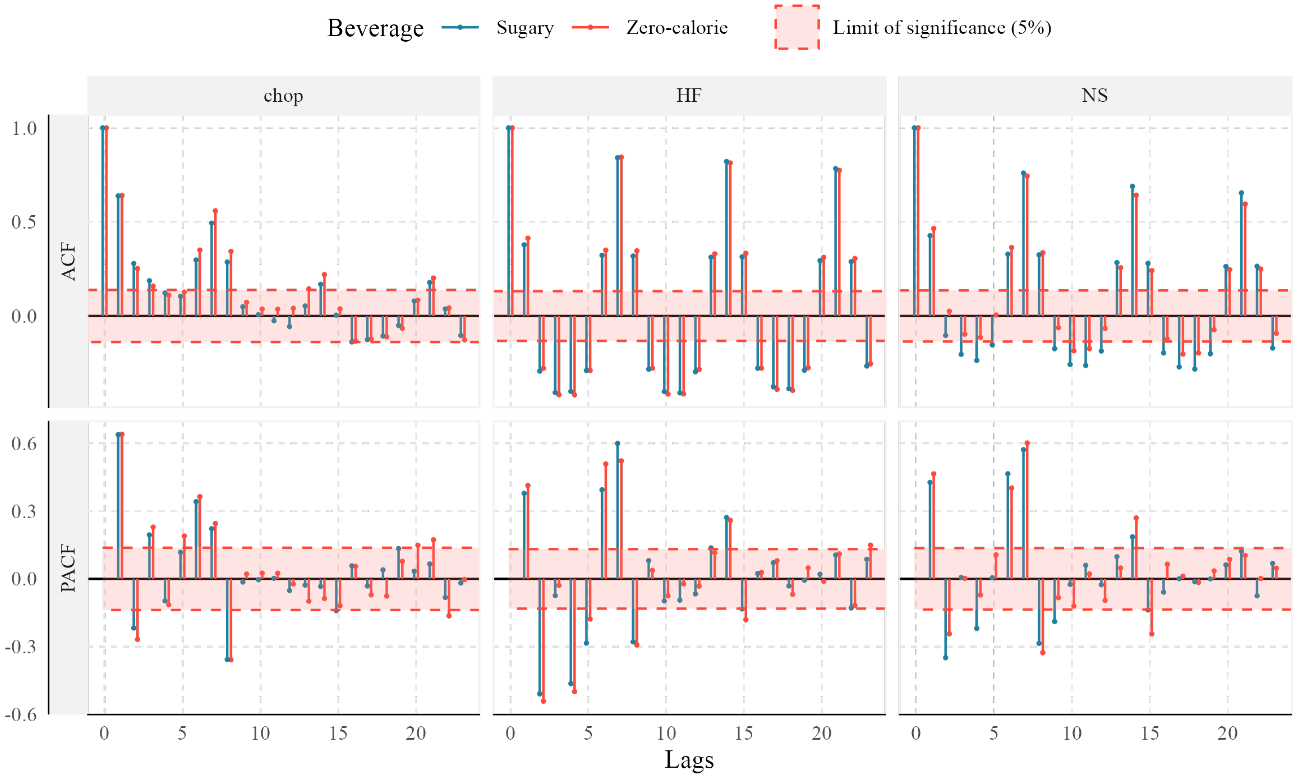


Figure A.4: **ACF and PACF by Beverage and Site**