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

Son Luu (71843379), 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**

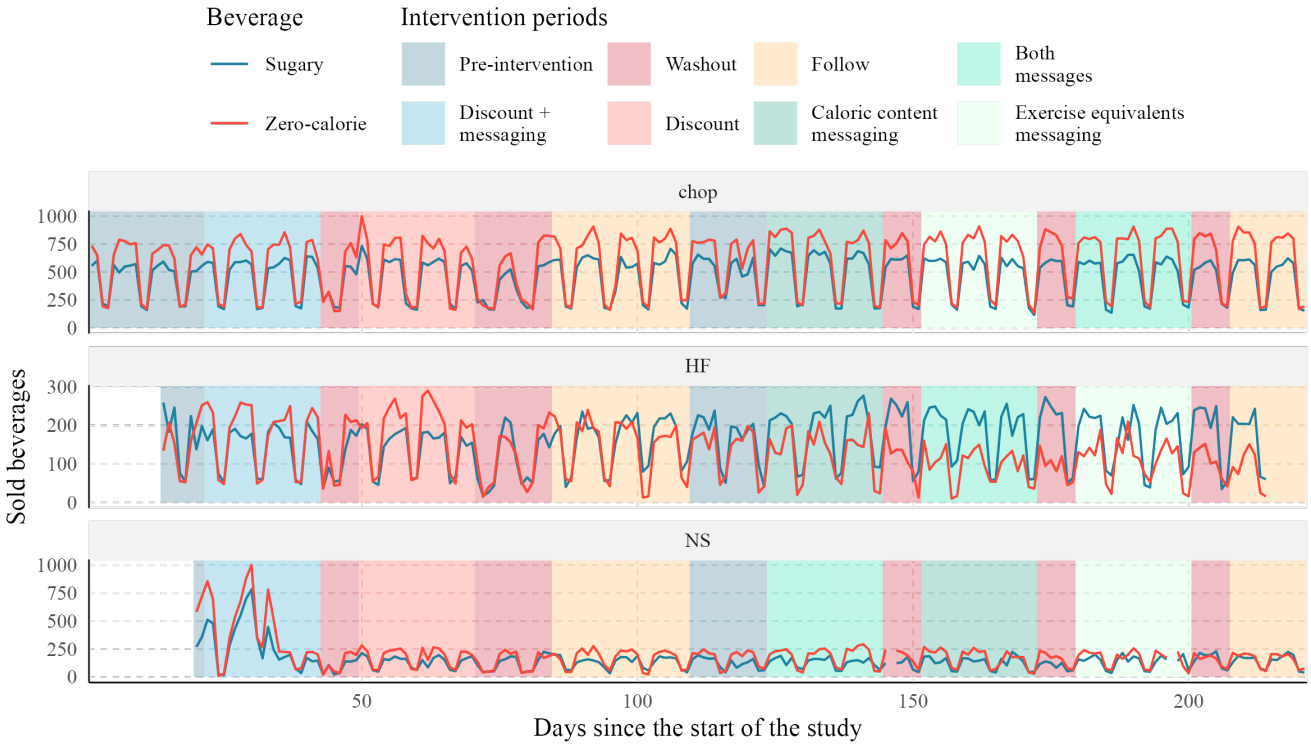
Variable	Type	Unit
Day of the quasi-experiment	Continuous	-
Day of the week	Continuous	-
Site	Categorical	-
Intervention	Categorical	-
Sugary beverages sold	Continuous	-
Zero-calorie beverages sold	Continuous	-
Other beverages sold	Continuous	-

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 ANOVA/ANCOVA

ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more samples, determining if at least one sample mean significantly differs from the others. It's instrumental in experiments aimed at evaluating the effectiveness of different interventions, such as those designed to encourage the selection of zero-calorie beverages over sugary options in hospital settings. By examining the variance within and between intervention groups, ANOVA facilitates understanding whether observed differences in beverage consumption are attributable to the interventions or occur by chance.

This approach offers the advantage of simplicity and clear interpretation, allowing researchers to assess the relative effectiveness of each intervention. However, ANOVA's limitation lies in its inability to account for external variables that could influence outcomes, such as baseline consumption patterns or hospital site characteristics. Unlike ANCOVA, which adjusts for these covariates, ANOVA assumes all group differences result from the interventions themselves, which could lead to oversimplified conclusions. Despite this, ANOVA remains valuable for initial analysis, with its limitations highlighting the need for a comprehensive approach, possibly incorporating ANCOVA, to fully understand the interventions' impacts on beverage selection.

4.3 Linear mixed effect model

Linear Mixed Effects (LME) model is a statistical model that accounts for fixed effects, common trends that are present at all levels of the data, as well as random effects, correlation within groups and variation between groups. The inclusion of random effects makes LME especially adept at handling hierarchical and longitudinal data, where observations are grouped into different levels.

The error terms in LME models are assumed to be independent identically distributed Normal variables with mean zero.

For the data at hand, it is recommended to run a linear regression for each site - intervention combination and compare the estimated parameters. If there are significant differences in estimation for a factor, then a random effect specific to that factor needs to be included in the final LME model.

4.4 Recommendation

For the first question, we recommend using the following ANOVA model

$$\text{Sales} = \beta_0 + \beta_1 \times \text{Intervention}$$

for each of the sites with the pre-intervention period being the reference. The estimated coefficient for each intervention method will then be compared across sites. If there are any significant differences between these estimations, a LME model with site dependent random effects is recommended. For the second question, we recommend using the following ANOVA model

$$\text{Sales} = \beta_0 + \beta_1 \times \text{Site}$$

for each of the intervention. The estimated coefficient for each site will then be compared across interventions. If there are any significant differences between these estimations, a LME model with intervention dependent random effects is recommended. For the third question, we recommend using the following ANOVA model

$$\text{Sales} = \beta_0 + \beta_1 \times \text{Calorie} \ + \beta_2 \times \text{Excercise}$$

for each of the sites with the combination period being the reference. Using the same method as the first question, if there are any significant differences between the estimated coefficients, a LME model with site dependent random effects is recommended. For the final question, we recommend using the following ANOVA model

$$\text{Sales} = \beta_0 + \beta_1 \times \text{Calorie}$$

for each of the sites with the calorie admonishment period being the reference. Again, if there are any significant differences between the estimated coefficients, a LME model with site dependent random effects is recommended. Note that the sales in these model refer to the trend extracted in the EDA step.

If the sales trends are found to be linear for each intervention period, the time coviariate can be added to the models above, turning them into ANCOVA models. In this case, only the pre-intervention period before the third intervention should be kept for the analysis as the time covariate needs to be reset for each intervention period as well as the pre-intervention period.

5 Conclusion

The study employs ANOVA/ANCOVA and LME models to analyze the effectiveness of interventions on beverage sales across hospitals, addressing four research questions. ANOVA models initially

identify differences in intervention effects, while LME models are recommended for deeper analysis when significant variations are observed. This approach ensures a thorough investigation into how each intervention influences beverage choices, the impact of site characteristics, and the effectiveness of calorie information and exercise prompts. The study's methodology provides a clear path to understanding which interventions work best, promoting healthier beverage consumptions.

6 References

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

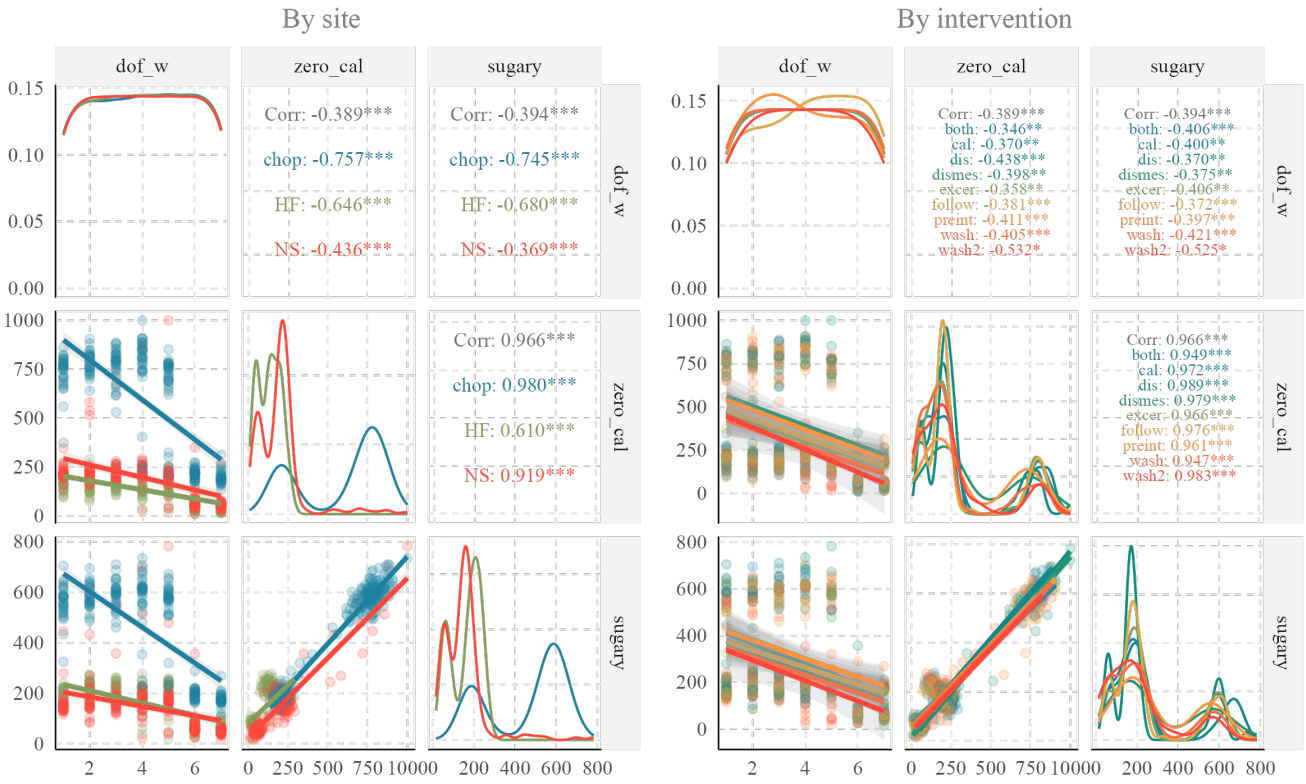
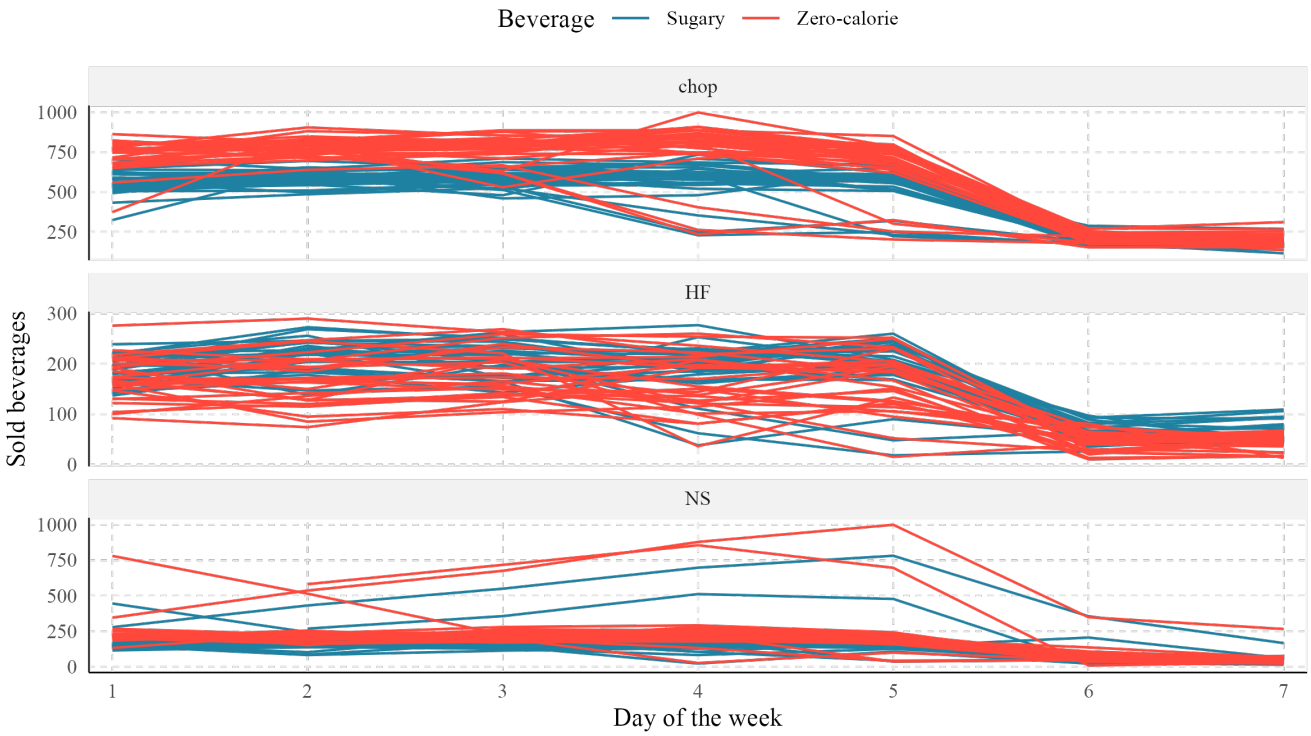


Figure A.1: **Correlation matrix of variables**



Source: client submission.

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

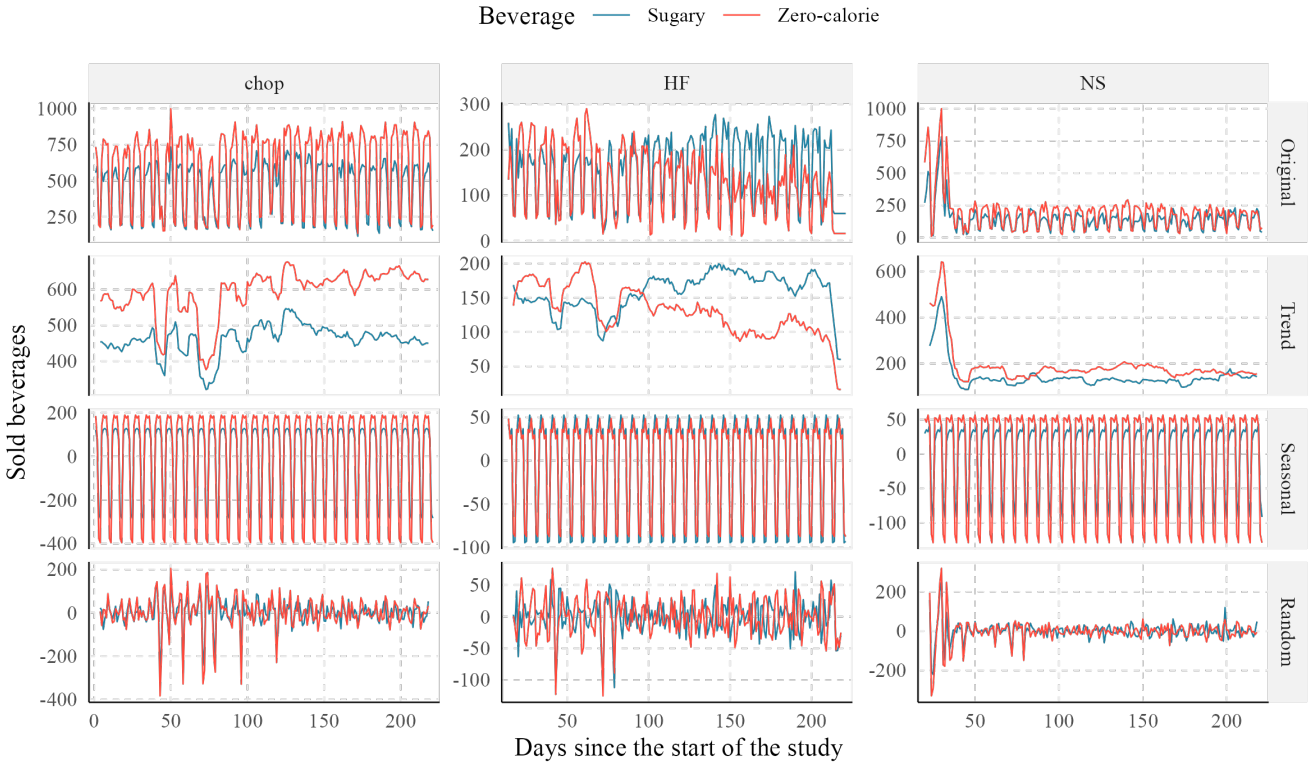


Figure A.3: **Decomposition Analysis of Sales for Sugary and Zero-Calorie Beverages**

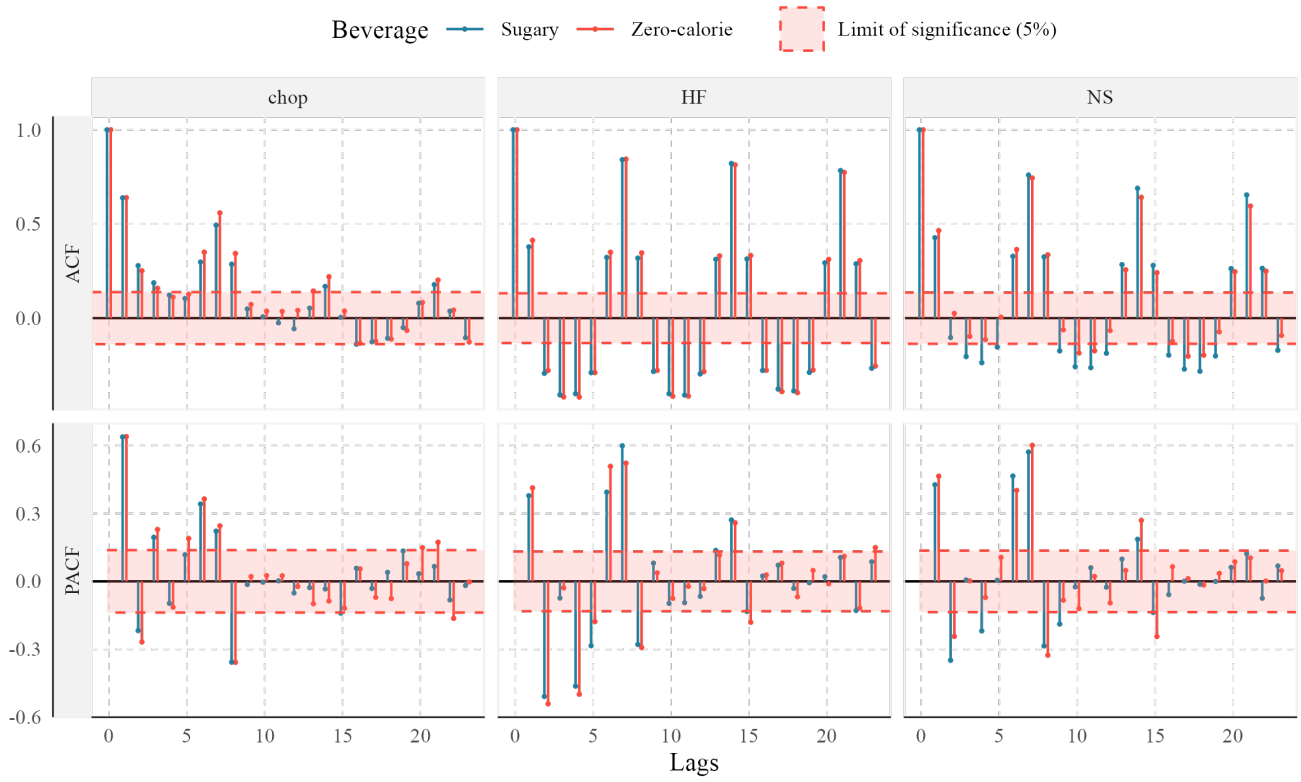


Figure A.4: **ACF and PACF by Beverage and Site**