Client Report (Placeholder)

Parham Pishrobat, Sarah Masri, Johnson Chen

2024-02-29

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(rlang)
```

```
##
## Attaching package: 'rlang'
##
## The following objects are masked from 'package:purrr':
##
##     %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
##     flatten_raw, invoke, splice
```

```
library(visdat)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
source("../code/EDA.R")
source("../code/ITSA.R")
source("../code/GEE.R")
source("../code/LMEM.R")
```

Introduction

- Background of the study.
- Objective(s) of the study.
- Statistical questions to answer.

Concerns around sugar consumption and its health implications have prompted an array of interventions aimed at modifying consumer behaviours in relation to sugary beverages. The current study investigates the effectiveness of various strategies to encourage consumers to choose zero-calorie beverages over sugary alternatives. In particular, the research question focuses on the impact of two types of visual presentations of calorie content through posters that highlight either the calorie content or the physical activity required to burn these calories. Furthermore, the effectiveness of price discounts, both independently and in conjunction with explanatory messaging, as a means to influence consumer choices is explored.

Data Description and Summaries

- Data collection method.
- Study design.
- Sample size.
- Variables measured.
- Missing data.

The data has been collected data from cafeterias and convenience shops within three hospital sites, denoted by A, B, C, after conducting the interventions. Hospitals A is urban and has two cafeterias and two convenience shops. Hospital B is also urban setting and but has only one cafeteria. And finally hospital C is suburban setting, and has one cafeteria and one convenience shop. Both interventions (delivery of discount or the messaging type) and data collection (recording the sales) was automatic at site A and by trained personnel in sites B and C. In the context of this study, sugary beverages include regular soft drinks and iced teas that are sweetened with natural sugars like sucrose and corn syrup, and zero-calorie beverages include diet soft drinks and teas, and water. Note that other beverage types like juices, milk, coffee, and fountain-dispensed drinks are excluded due to categorization challenges.

The study adopts an interrupted time-series multi-site quasi-experimental design to assess the outcomes of five distinct interventions on the purchase patterns of bottled sugary and zero-calorie beverages. The interventions consist of two price discounts and three calorie messaging strategies, each designed to influence consumer purchasing behaviour towards healthier beverage options. The price interventions involved a 10% discount on zero-calorie beverages, with one intervention additionally providing explanatory messaging about the discount. The calorie messaging interventions varied in their approach, providing information on the caloric content of sugary beverages, the physical activity required to burn off these calories, and a combination of both strategies.

The primary outcome of interest is daily sales of bottled sugary and zero-calorie beverages over a span of 30 weeks, from October 27 to May 23. The study periods include baseline data collection phases, intervention phases for both price discounts and calorie messaging, and washout periods to assess the persistence of intervention effects.

(INCOMPLETE)

```
# summarize_data(beverage_sales)$numerical
# unique(beverage_sales$Site)
# unique(beverage_sales$Intervention)
```

3. Exploratory Analysis

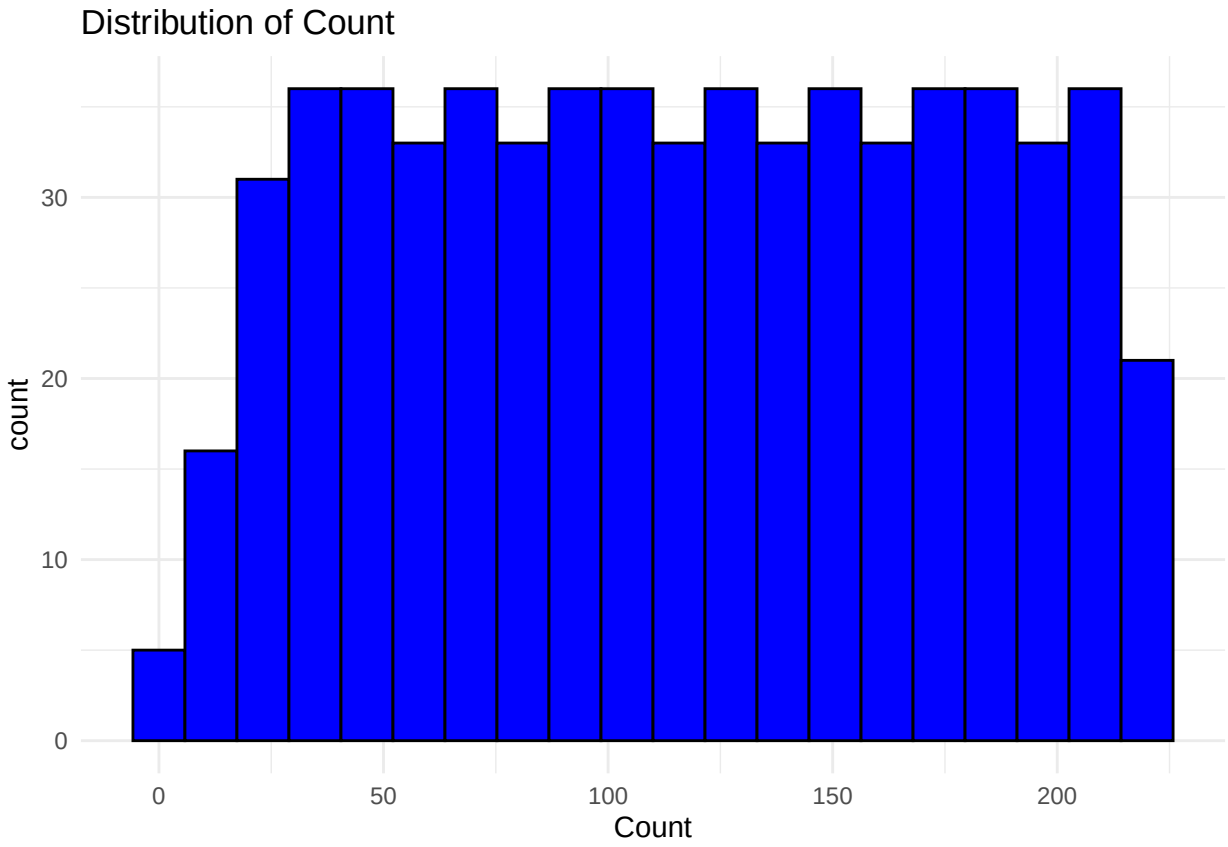
NOTE: we will figure out the order and title of these sub sections at the end

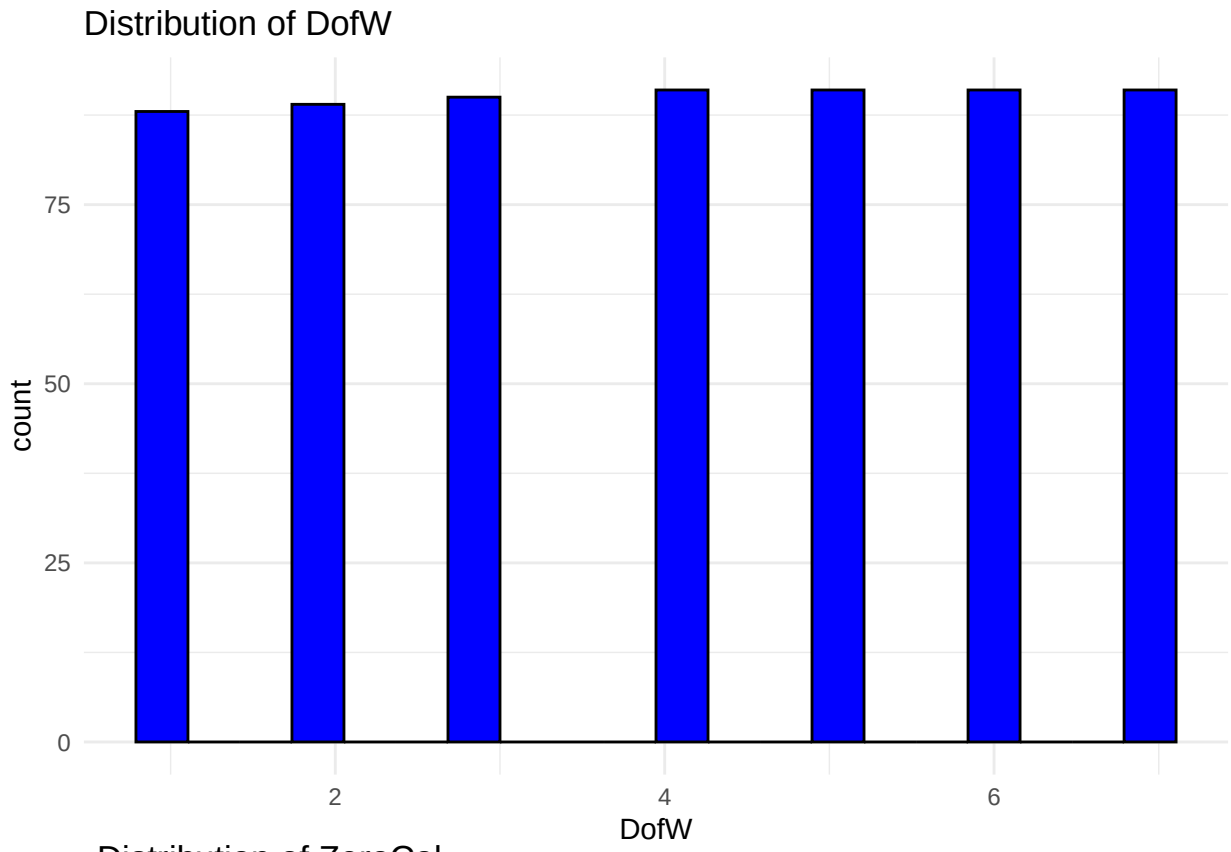
histogram and scatter plots, boxplot (Par; almost done)

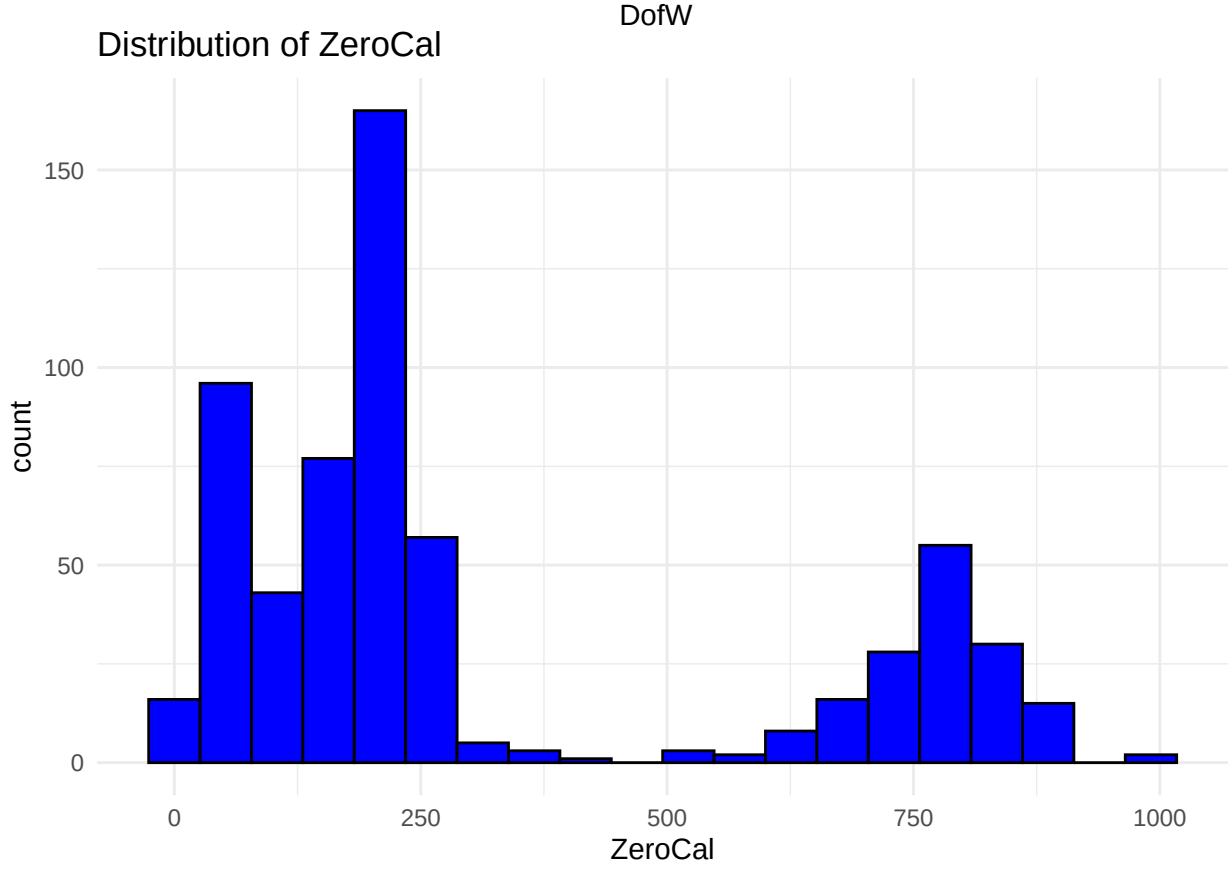
```
plot_histograms(beverage_sales)
```

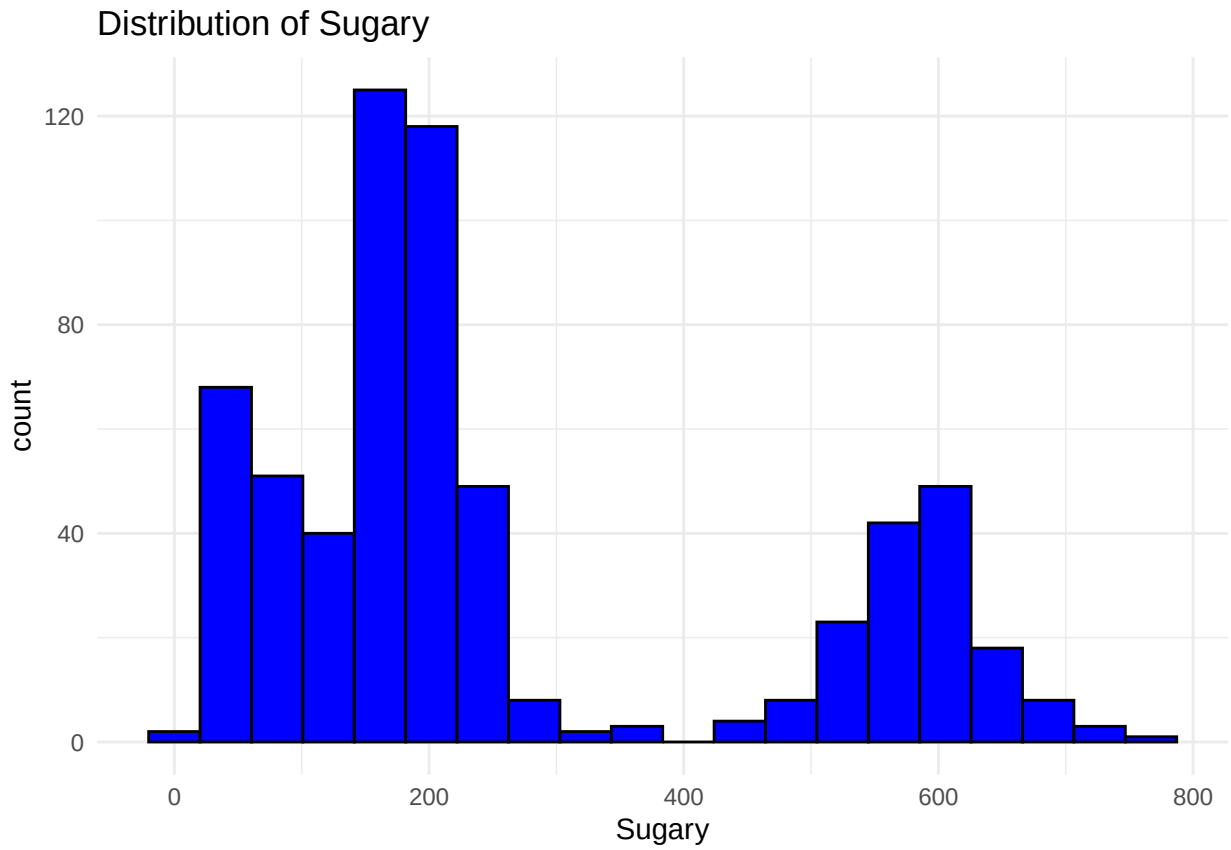
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
```

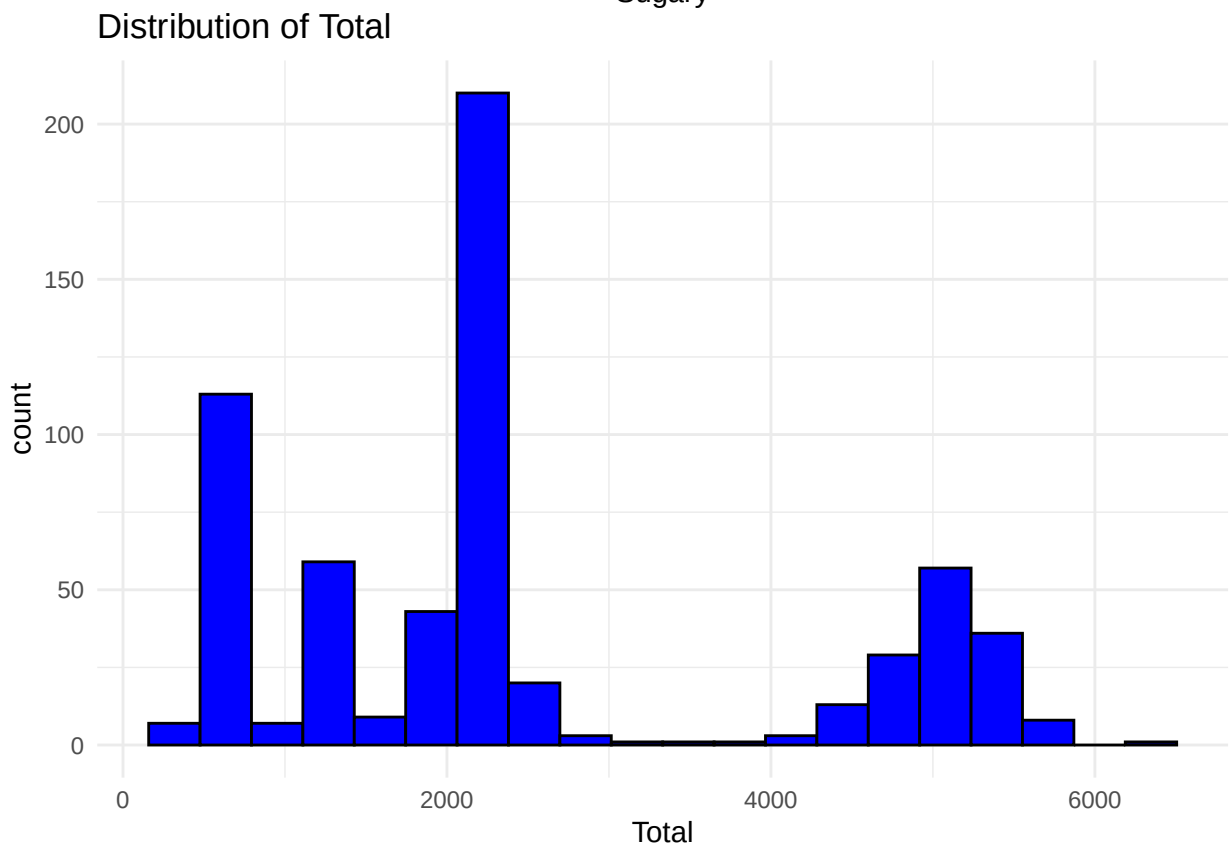
```
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



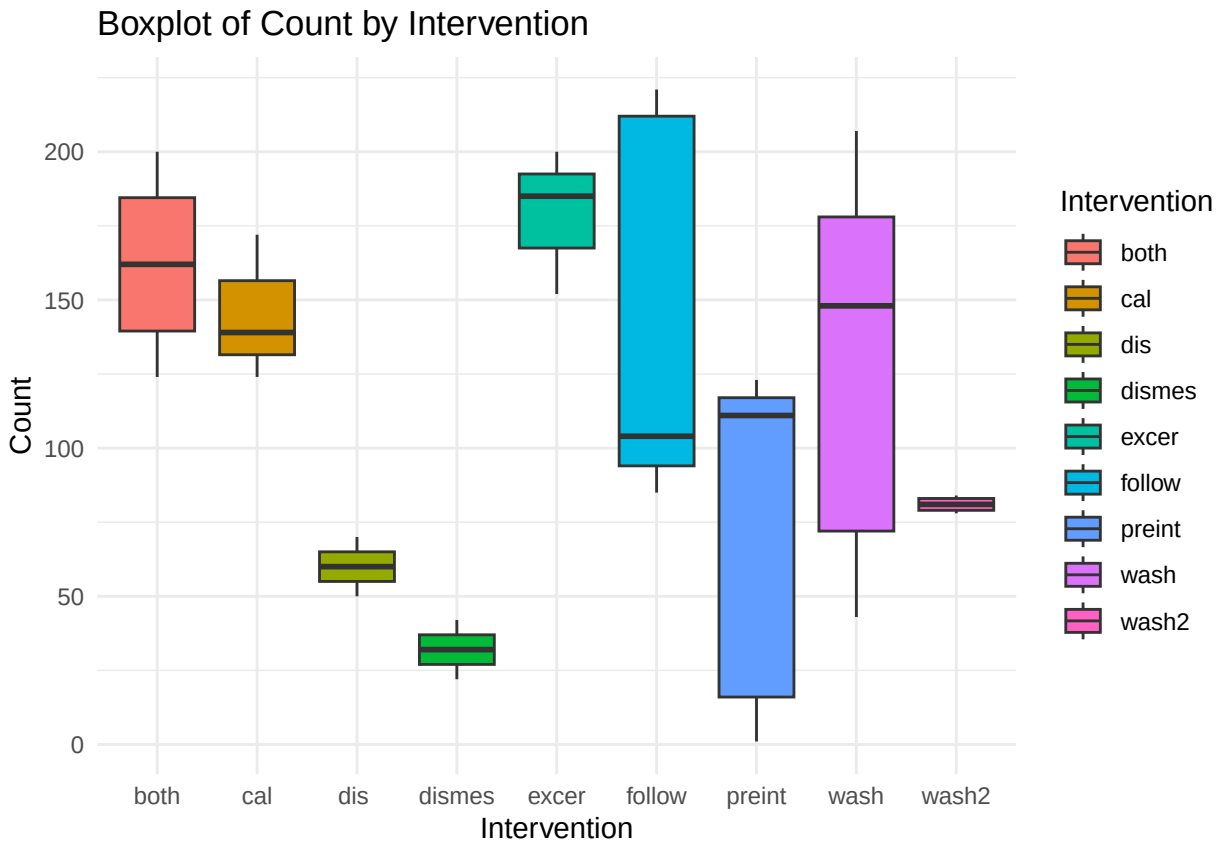

Distribution of DofW
count
75
50
25
0
2
4
6
DofW

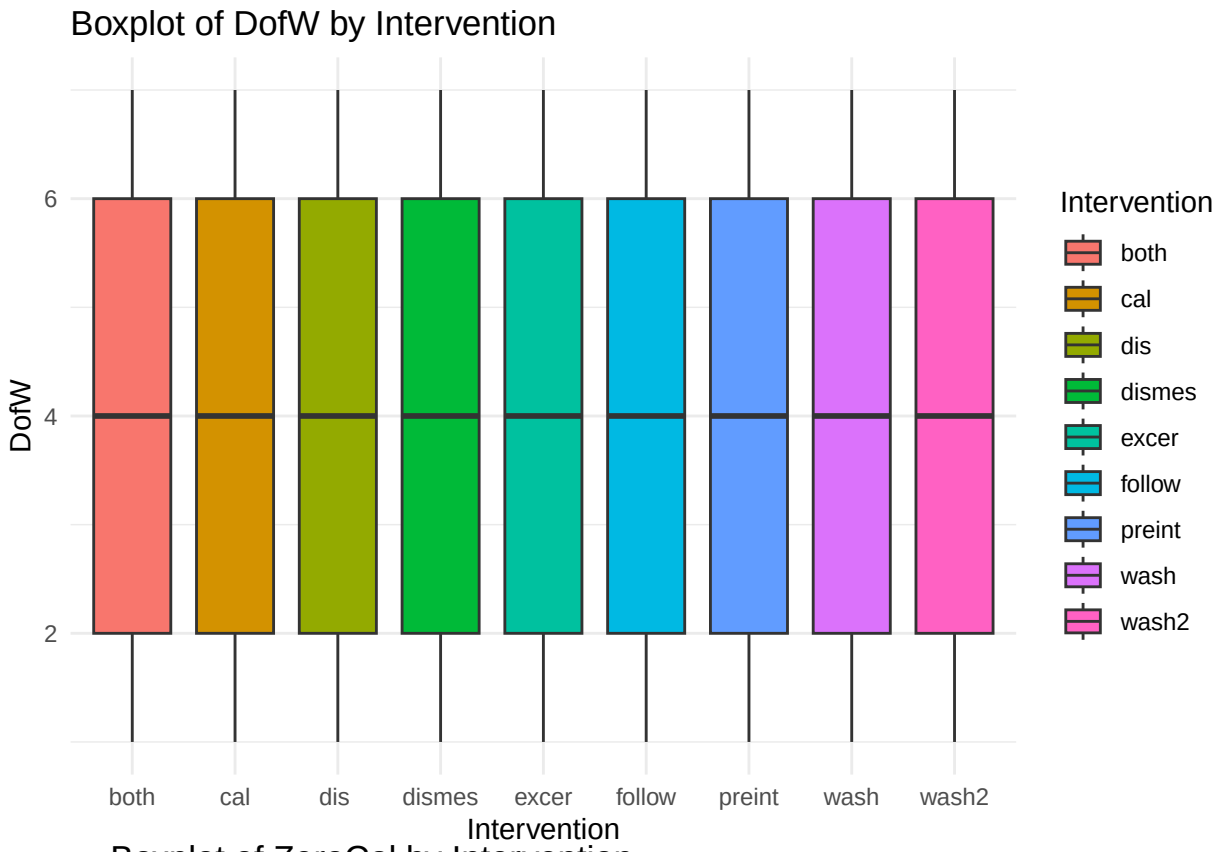

Distribution of ZeroCal
count
150
100
50
0
0
250
500
750
1000
ZeroCal

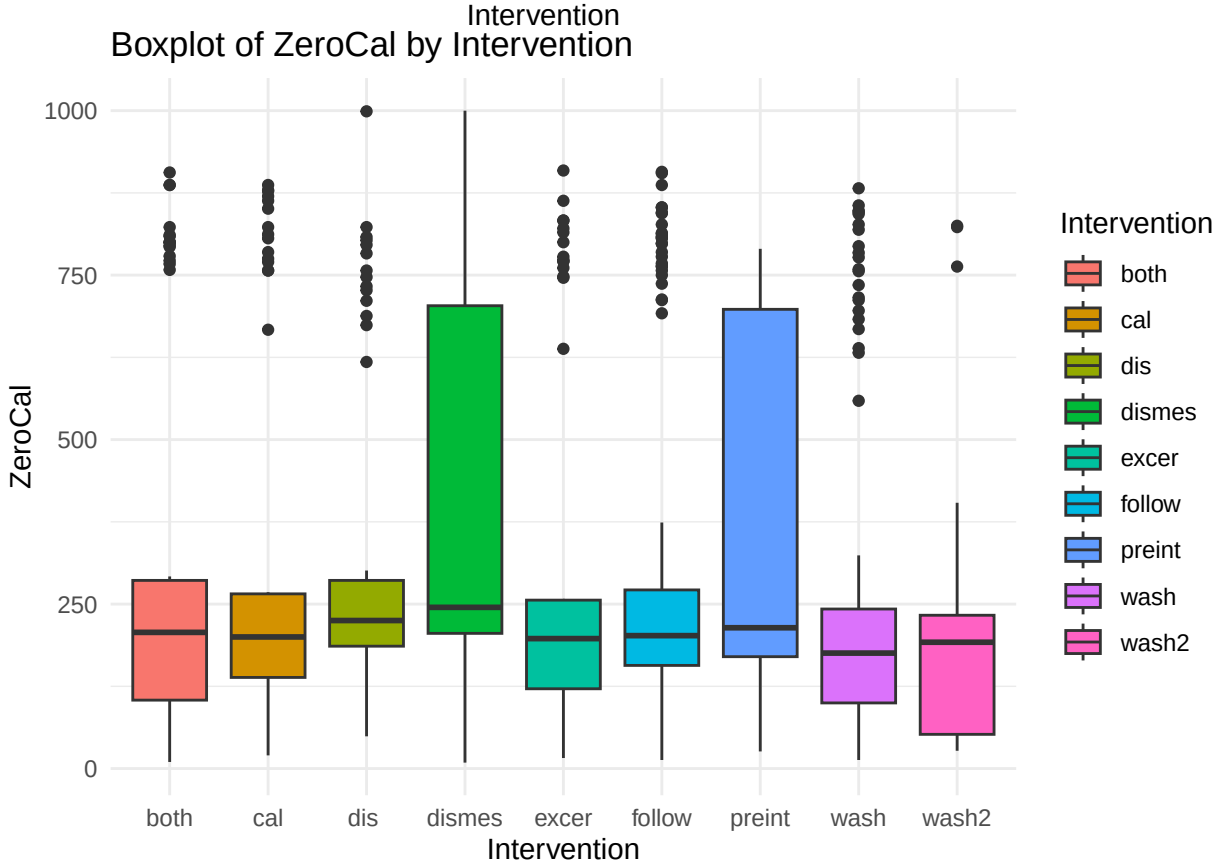

Distribution of Sugary
count
120
80
40
0
0
200
400
600
800
Sugary

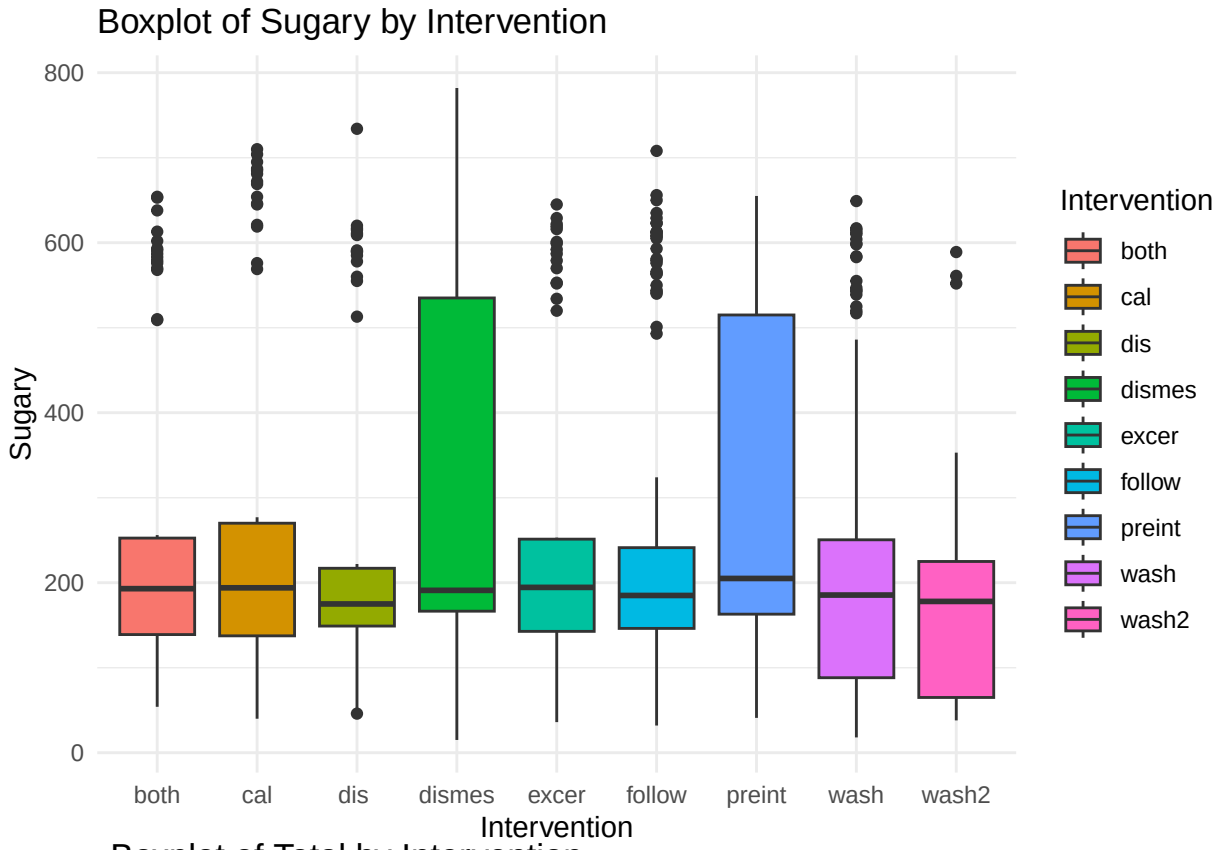

Distribution of Total
count
200
150
100
50
0
0
2000
4000
6000
Total

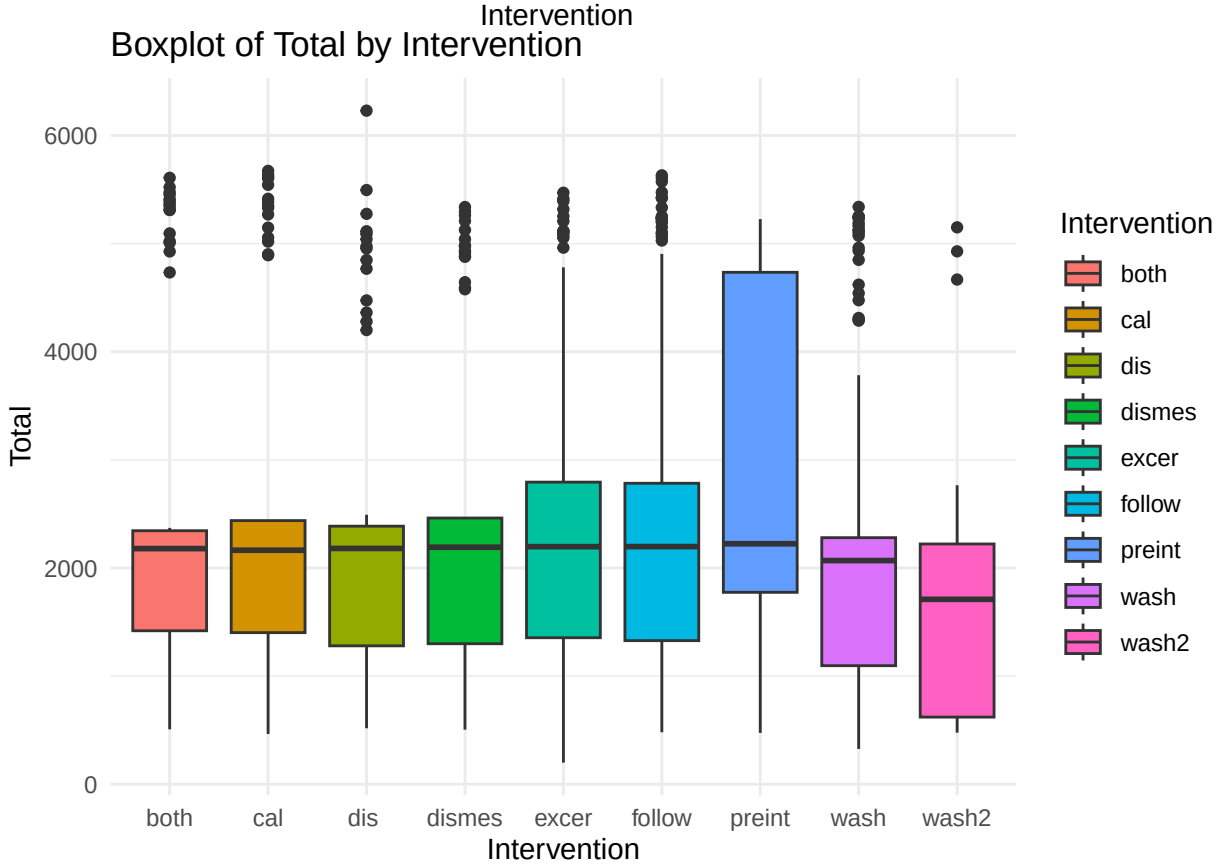
```
boxplots_by_category(beverage_sales, "Intervention")
```



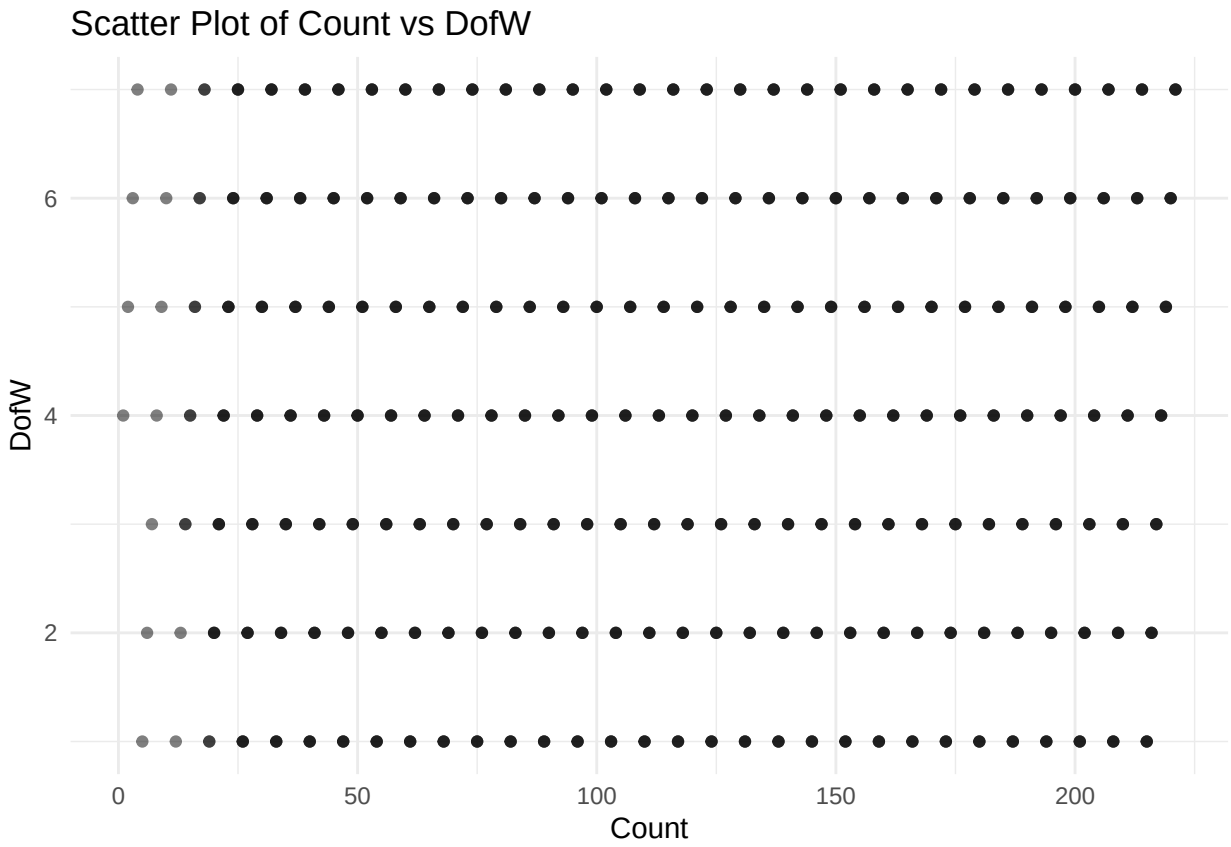

Boxplot of DofW by Intervention
DofW
6
4
2
both
cal
dis
dismes
excer
follow
preint
wash
wash2
Intervention
Intervention
both
cal
dis
dismes
excer
follow
preint
wash
wash2

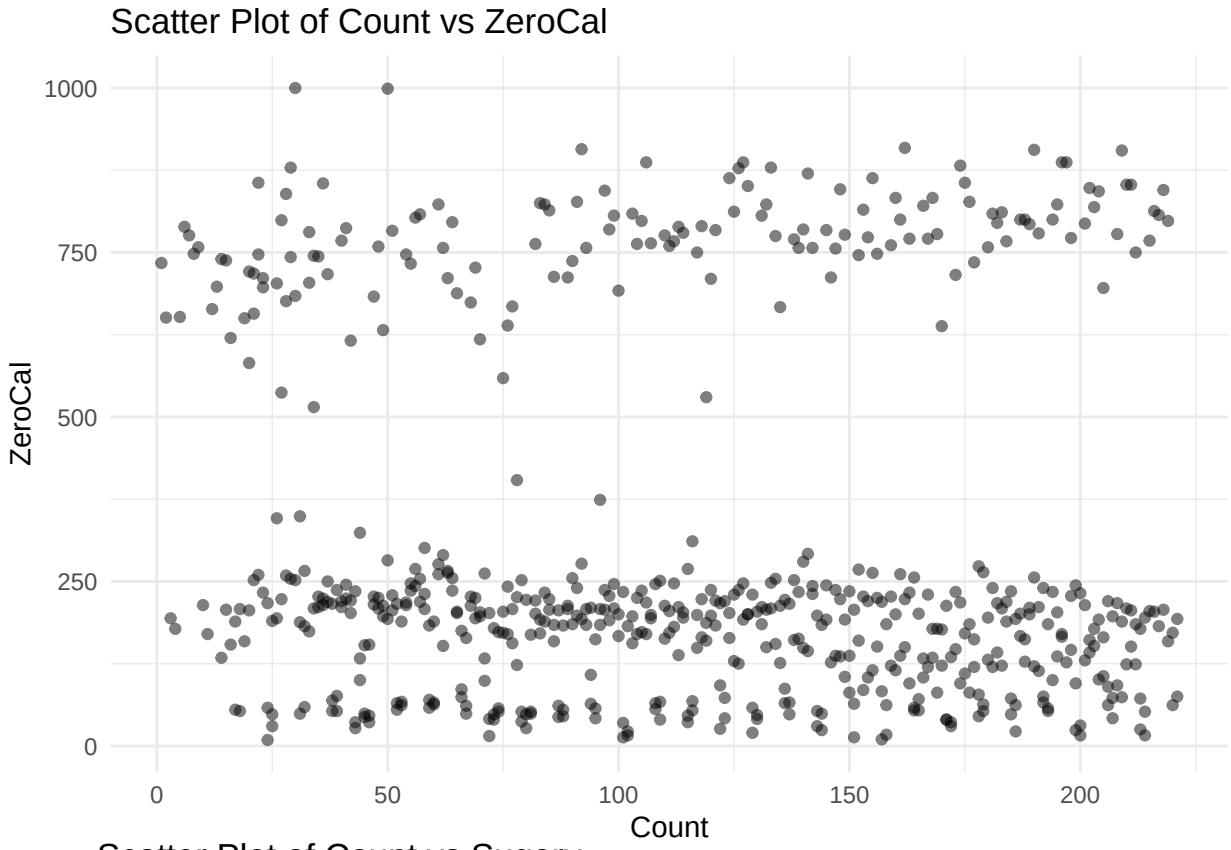

Boxplot of ZeroCal by Intervention
ZeroCal
1000
750
500
250
0
both
cal
dis
dismes
excer
follow
preint
wash
wash2
Intervention
Intervention
both
cal
dis
dismes
excer
follow
preint
wash
wash2

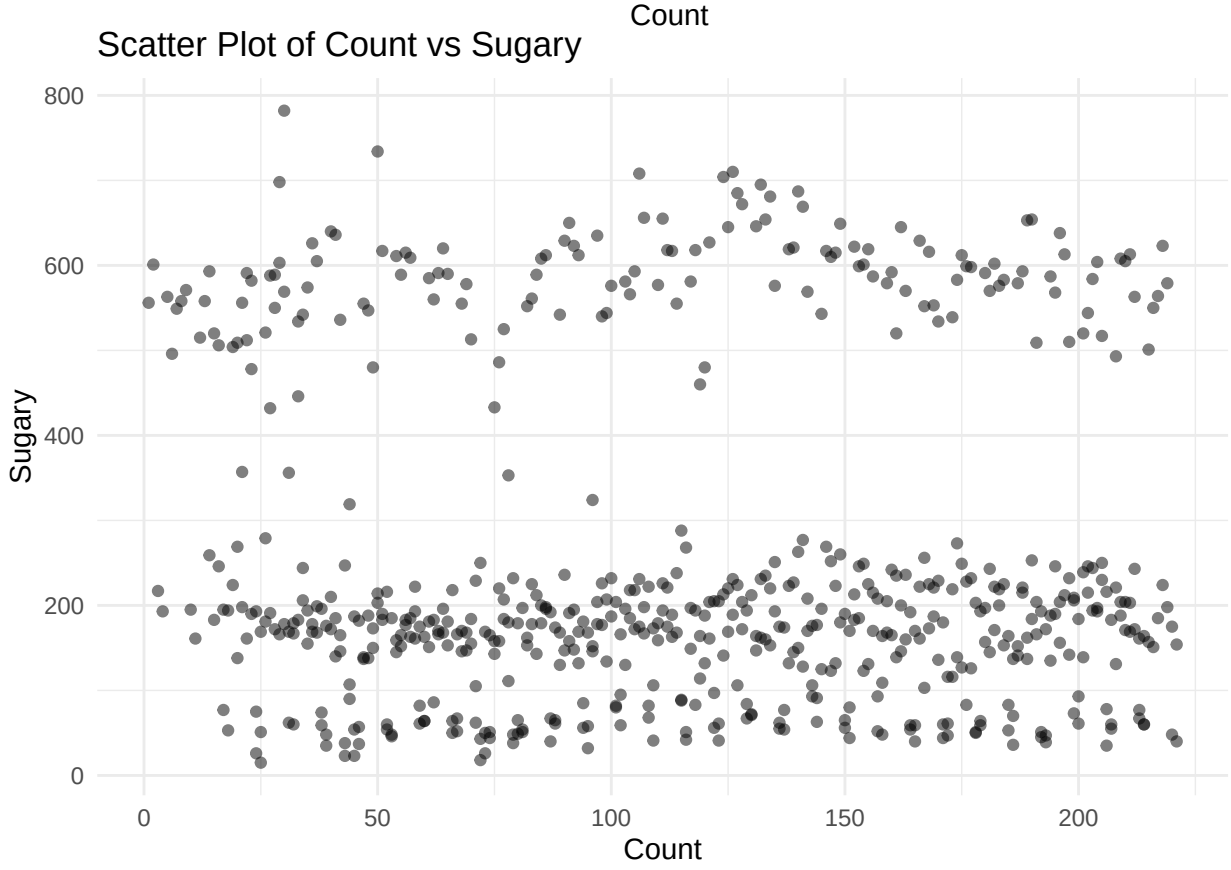

Boxplot of Sugary by Intervention
Sugary
800
600
400
200
0
both
cal
dis
dismes
excer
follow
preint
wash
wash2
Intervention
Intervention
both
cal
dis
dismes
excer
follow
preint
wash
wash2

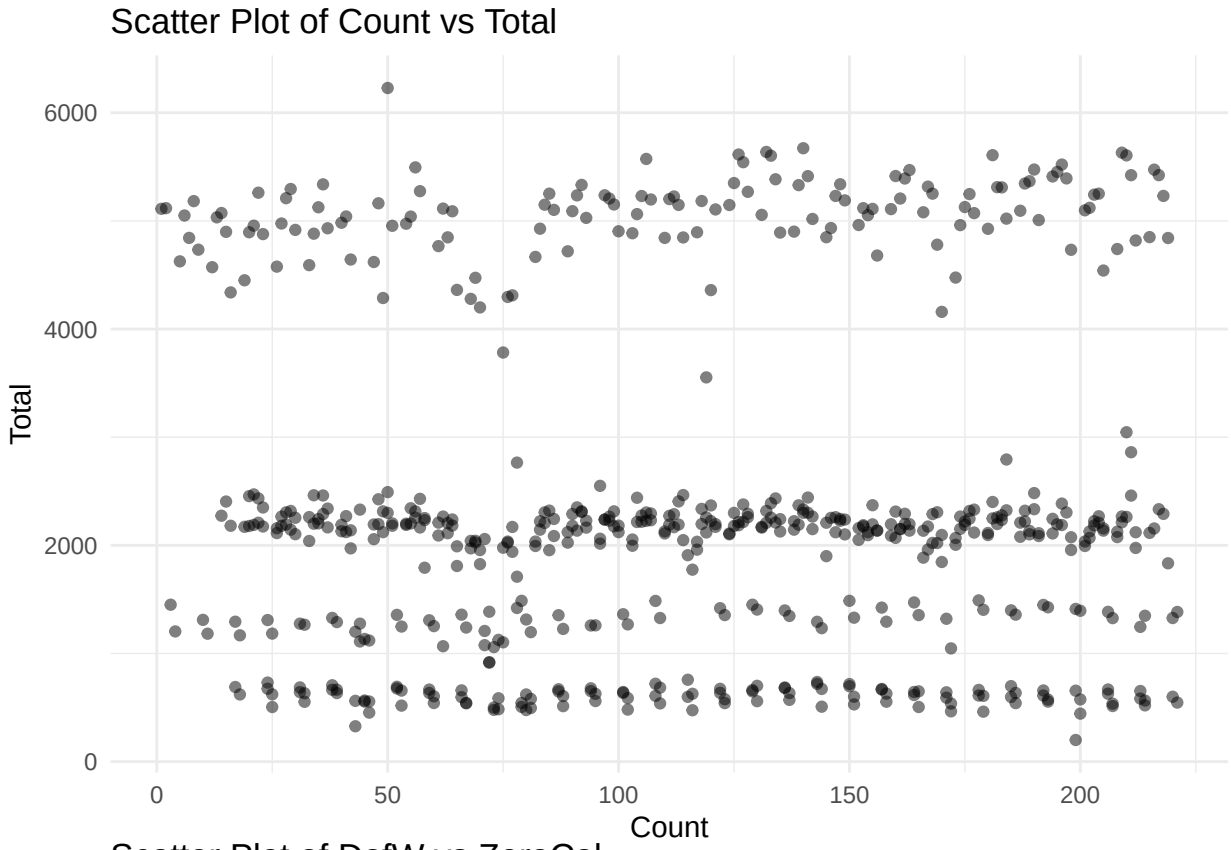

Boxplot of Total by Intervention
Total
6000
4000
2000
0
both
cal
dis
dismes
excer
follow
preint
wash
wash2
Intervention
Intervention
both
cal
dis
dismes
excer
follow
preint
wash
wash2

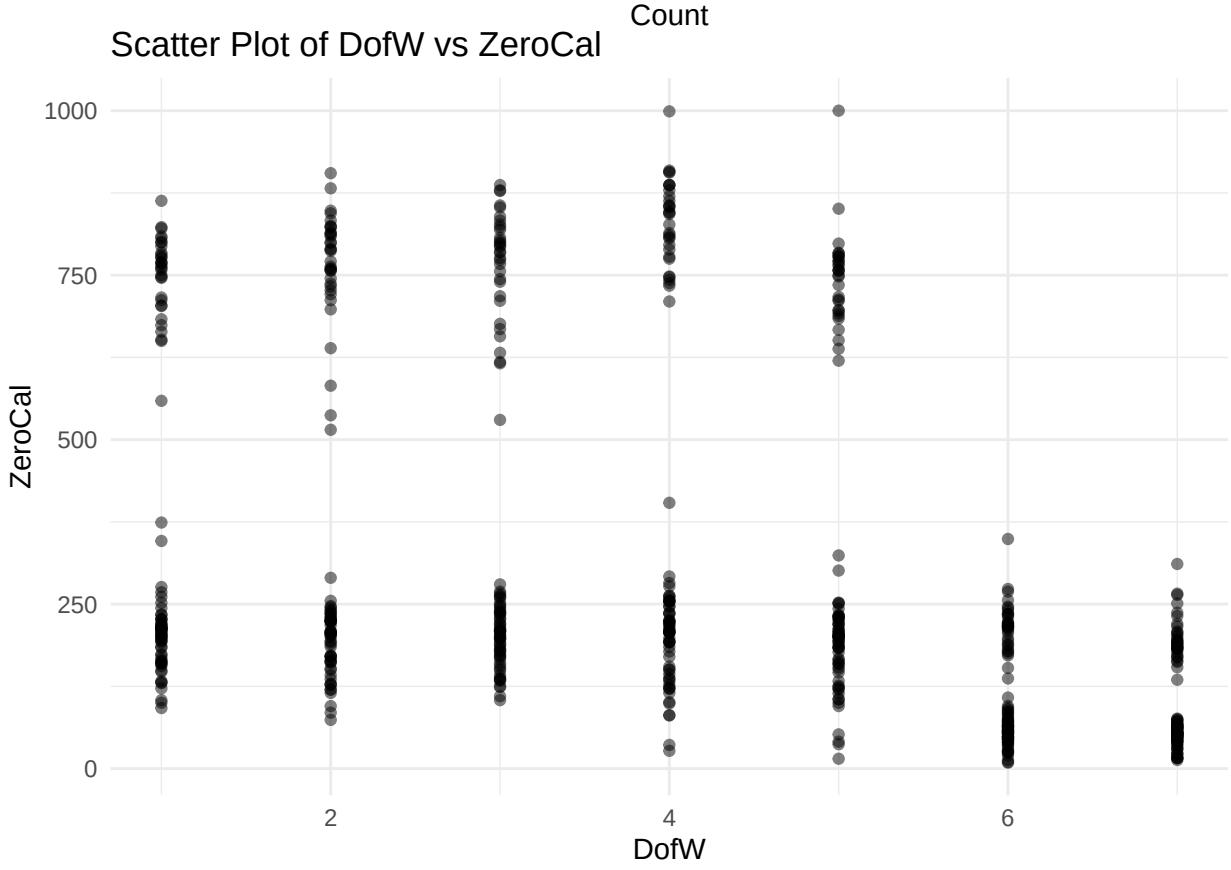
```
scatter_plots(beverage_sales)
```

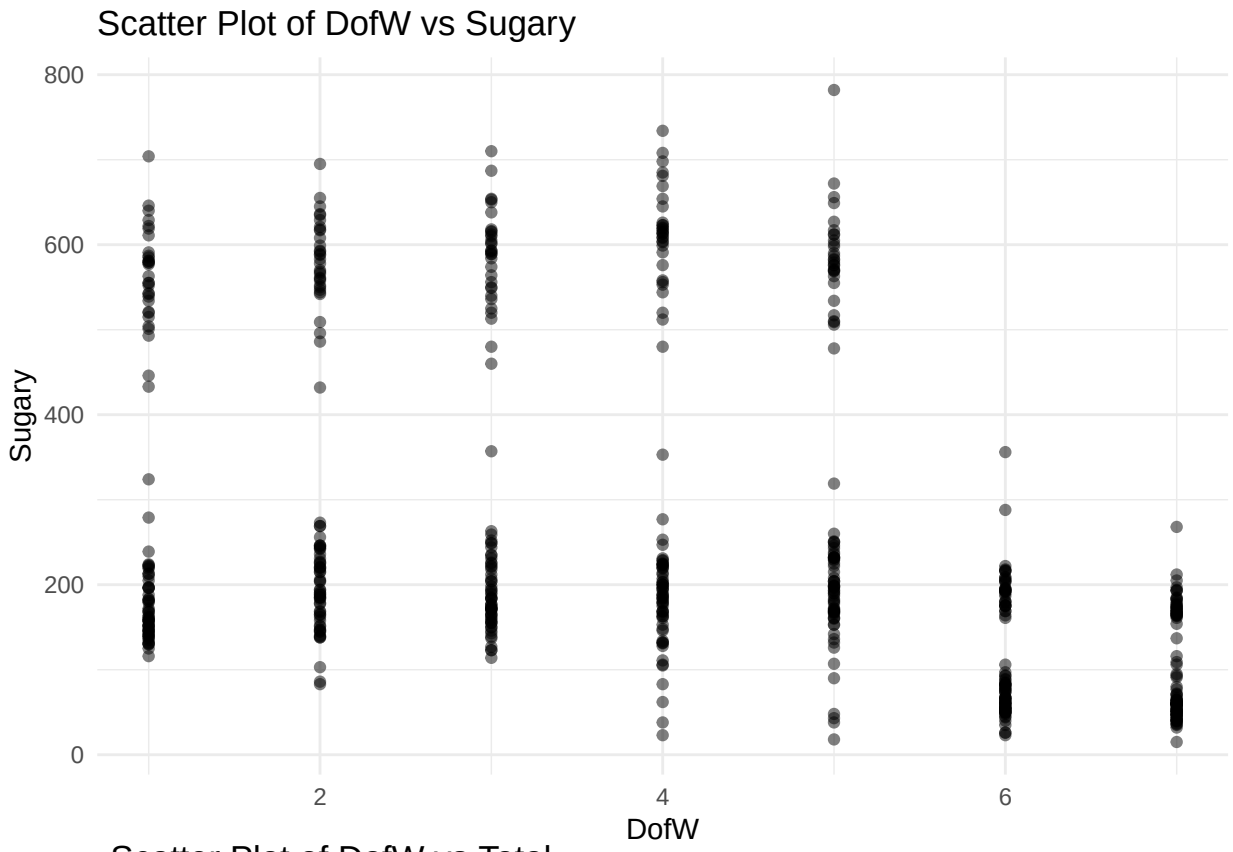


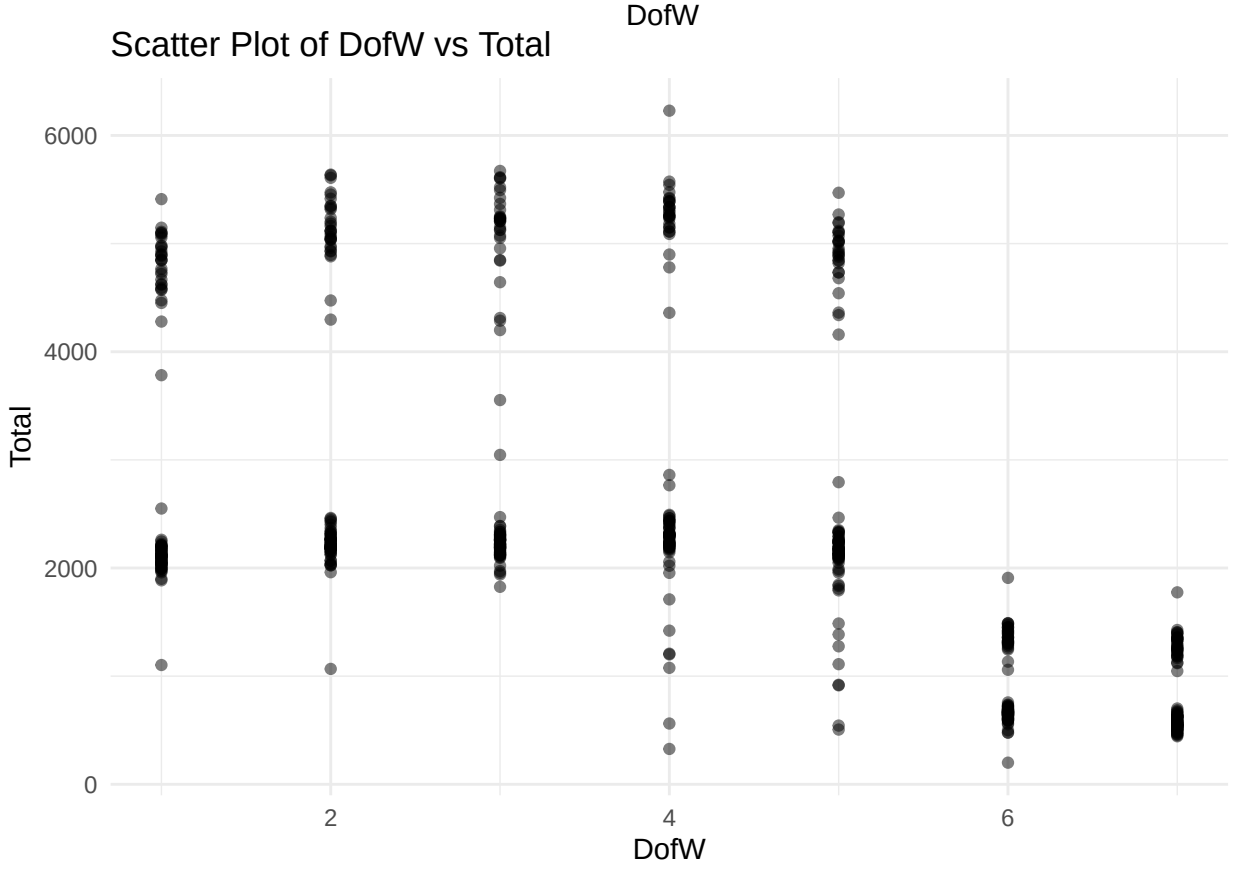

Scatter Plot of Count vs ZeroCal
ZeroCal
1000
750
500
250
0
0
50
100
150
200
Count

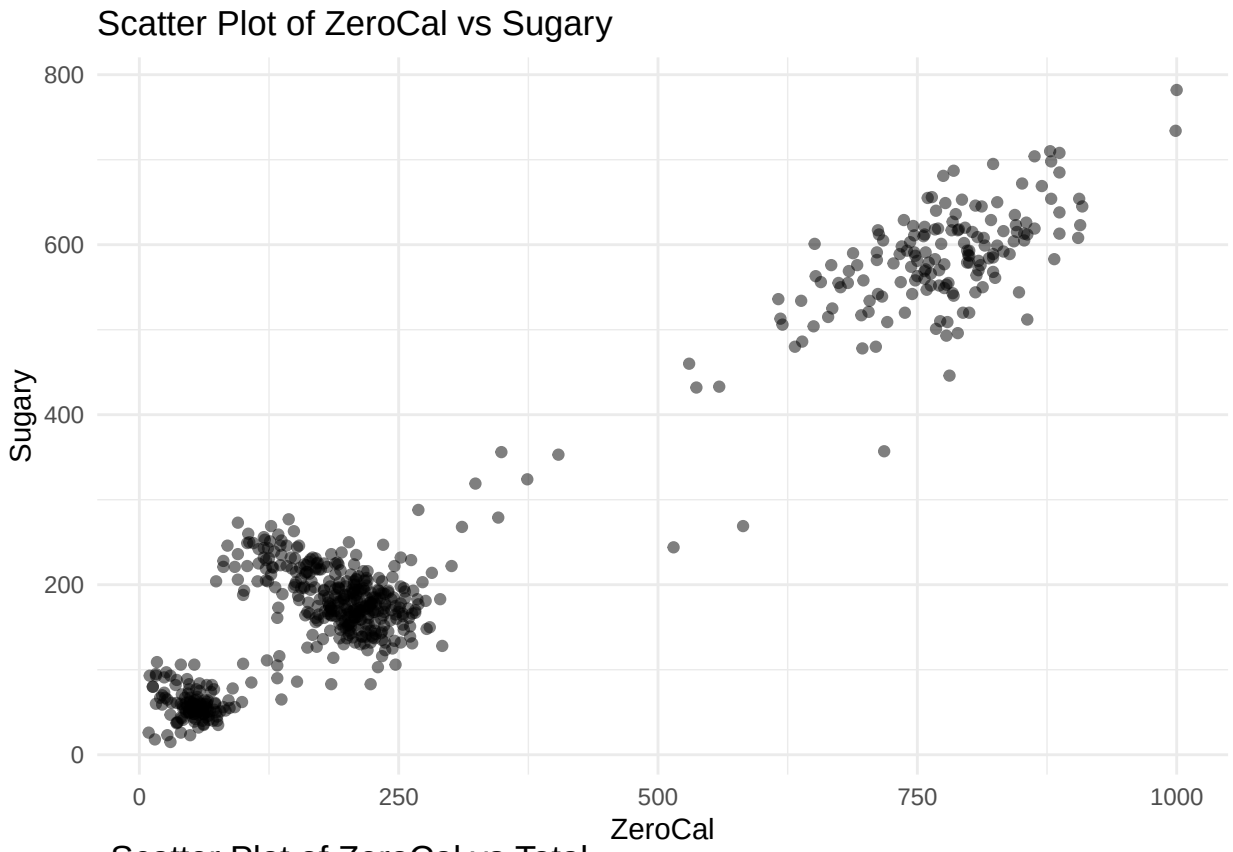

Scatter Plot of Count vs Sugary
Sugary
800
600
400
200
0
0
50
100
150
200
Count

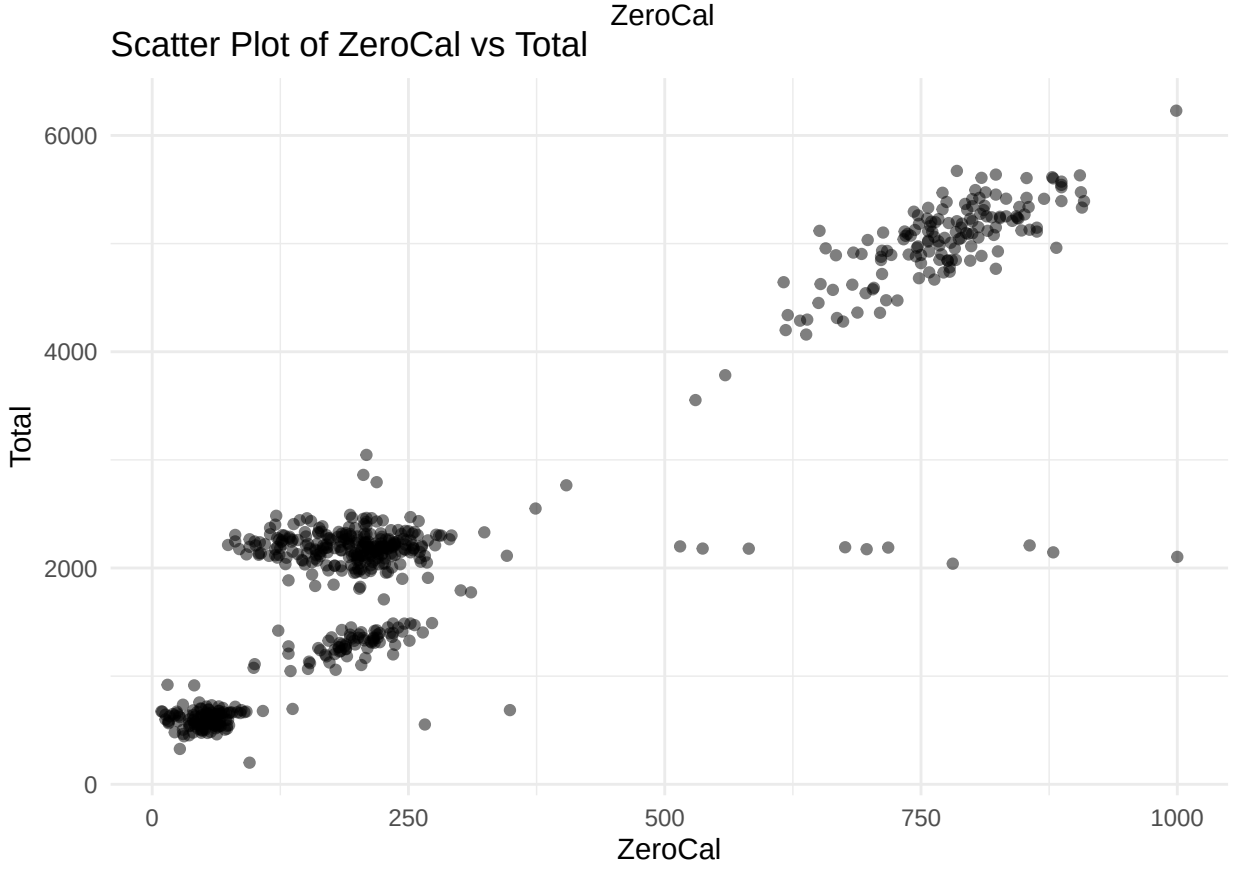

Scatter Plot of Count vs Total
Total
Count
0
2000
4000
6000
0
50
100
150
200

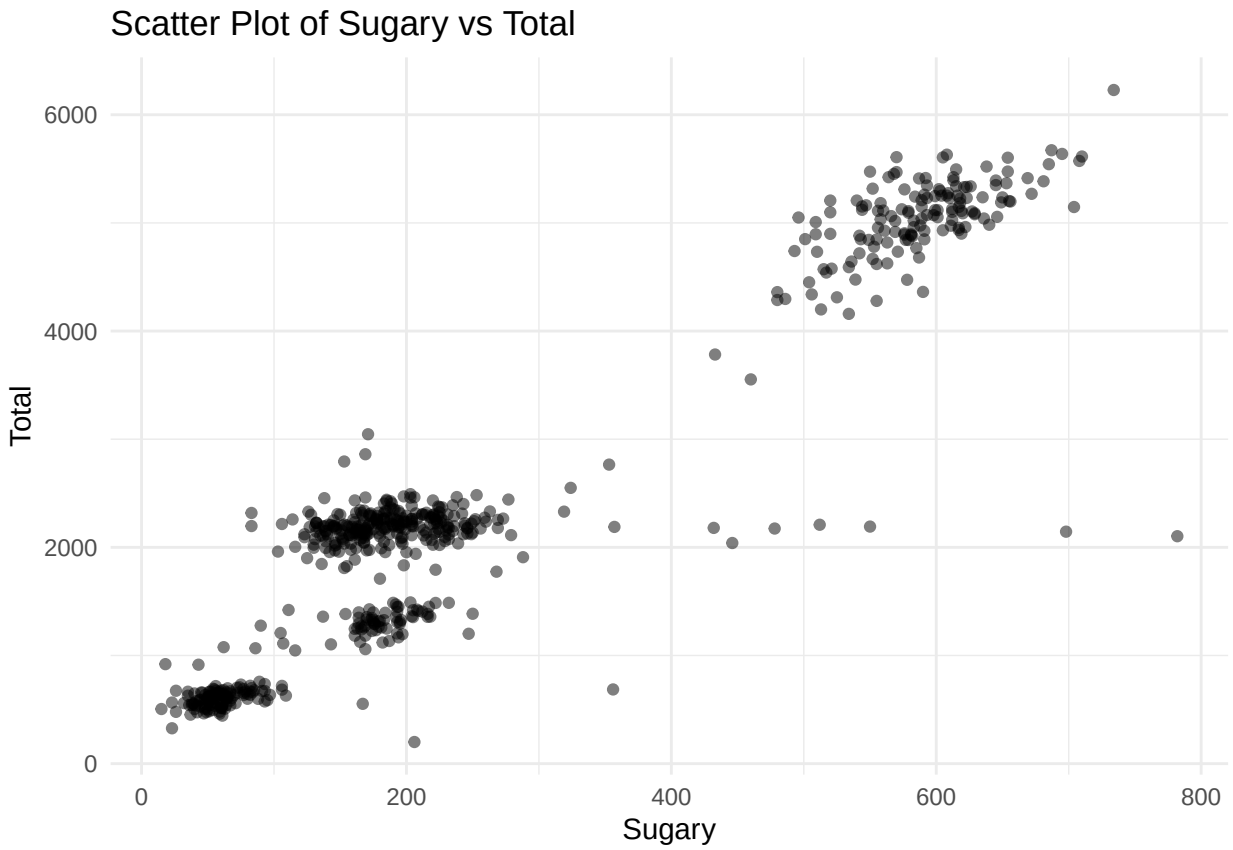

Scatter Plot of DofW vs ZeroCal
ZeroCal
DofW
0
250
500
750
1000
2
4
6


Scatter Plot of DofW vs Sugary
800
600
400
200
0
Sugary
2
4
6
DofW


Scatter Plot of DofW vs Total
6000
4000
2000
0
Total
2
4
6
DofW

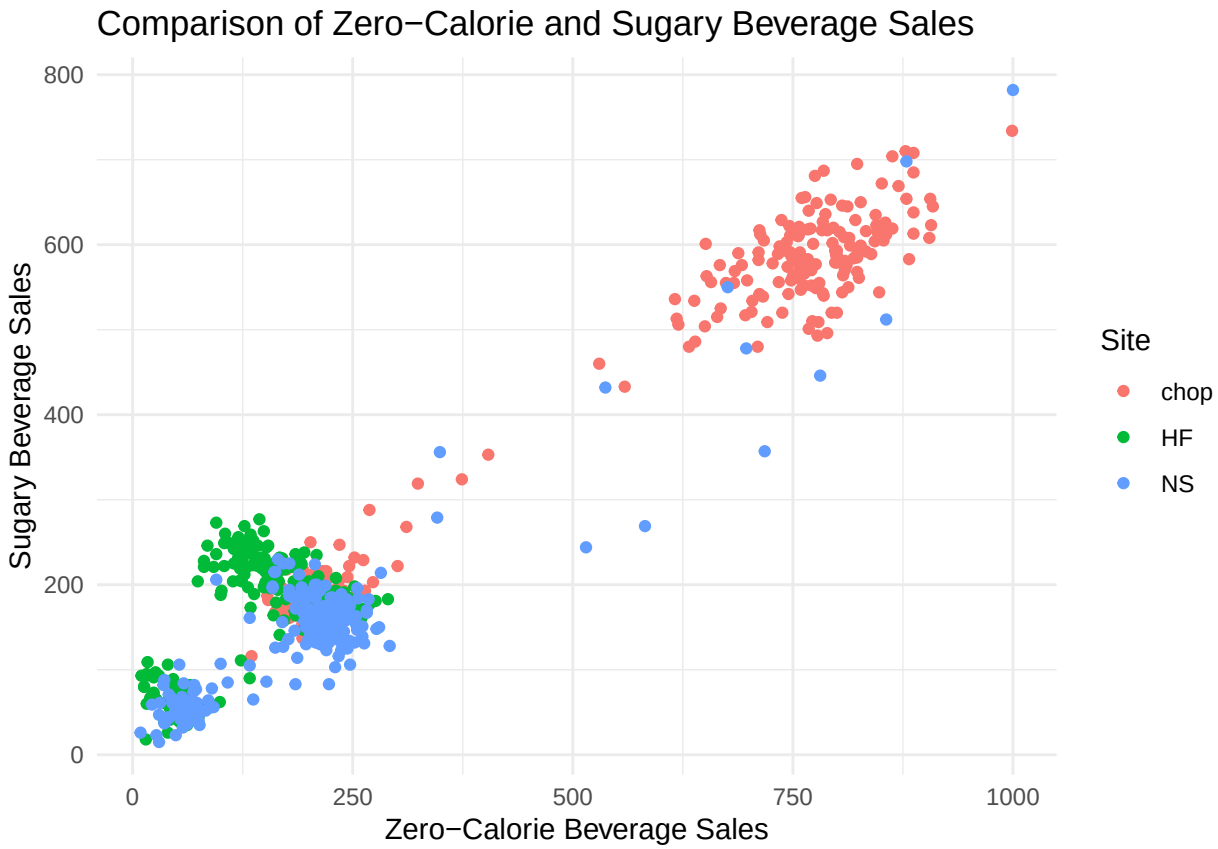

Scatter Plot of ZeroCal vs Sugary
Sugary
ZeroCal
0
200
400
600
800
0
250
500
750
1000


Scatter Plot of ZeroCal vs Total
Total
ZeroCal
0
2000
4000
6000
0
250
500
750
1000




```
plot_beverage_sales_comparison(beverage_sales, "ZeroCal", "Sugary", "Site")
```

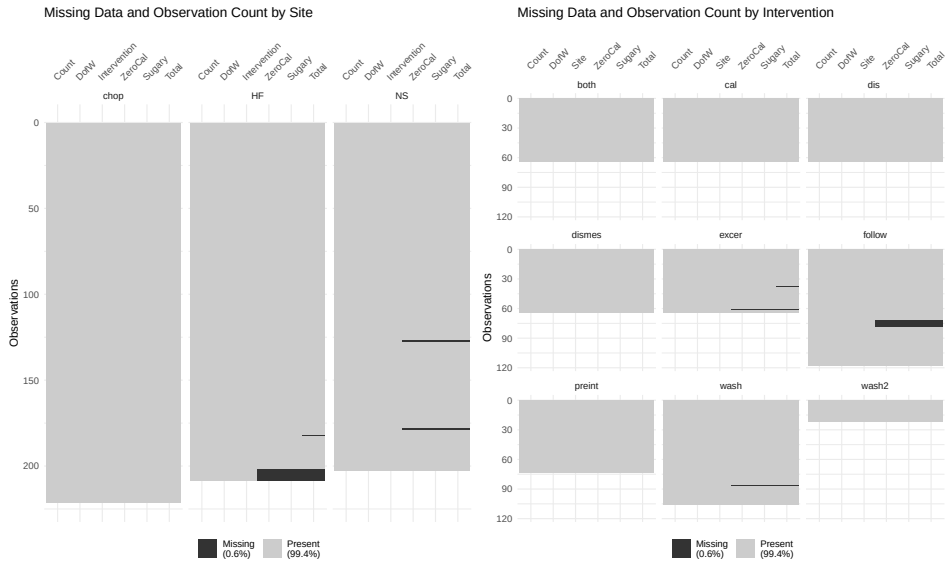
```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



correlation plot (Johnson)

Missing Values and Data Imbalance

The data has some missing data. It is important to identify what kind of missing data exists within a dataset to better understand how to handle missingness during formal analysis. It appears that the missing data qualifies as missing not at random (MNAR), meaning that the probability of any given observation being missing varies for unidentified reasons. It is also important to note from the observations counts whether or not the data appears 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.



line plot for trajectory (Par)

4. Formal Analysis

```
# handle_missing_data(beverage_sales)$MissingOverview
```

ITSA (Par)

GEE (Sarah)

The Generalized Estimating Equations (GEE) approach is a convenient and relatively easy to interpret method to model longitudinal data. GEE is suitable for analyzing the data from this study since daily sales of bottled sugared beverages and zero-calorie beverages were 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 number 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 affected 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.

LMEM (Johnson)

5. Conclusions

- Recommendations to the clients.

6. References

- Properly formatted citations.

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

7. Statistical Appendix

- Mathematical formulas.
- Additional tables/figures.

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 $\mathbf{X}$ including all revenant predictors, the model can be written more explicitly as

$$\begin{aligned} g(\mathbb{E}[Y_i]) = \beta_{0i} &+ \beta_{1i}(\text{Discount}) + \beta_{2i}(\text{Discount + Messaging}) + \beta_{3i}(\text{Calorie Messaging 1}) \\ &+ \beta_{4i}(\text{Calorie Messaging 2}) + \beta_{5i}(\text{Calorie Messaging 3}) + \beta_{6i}(\text{site B}) + \beta_{7i}(\text{site C}) \\ &+ \beta_{8i}(\text{Day of Week}) + \beta_{9i}(\text{Total sales}) \end{aligned}$$

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.