ASSIGNMENT 1

Exploring lemonade data and model specification

Course:

Market Models

(EBM077A05.2022-2023.1)

○ Redacted for privacy reasons

Introduction

This report consists of two main parts. In the first part, we provide information about the market for lemonade and we discuss some important findings about key variables that affect the market outcomes. In the second part, we will introduce a predictive model for the lemonade brand Raak, including the most relevant variables that affect Sales.

The market for lemonade

Our dataset contains data about 7 different lemonade brands (Karvan Cevitam, Raak, Slimpie, Teisseire, EuroShopper and PrivateLabel) and 6 different supermarket chains (Albert Heijn, Plus, Jumbo, Coop, Deen and Hoogvliet). Our original dataset consists of the following variables:

- ***UnitSales:*** The total weekly sales per chain, per brand expressed in Units. The maximum number of Sales was 741177 pieces in one week, while the mean was 23783 and the median was 10394. The maximum is really high as the 3rd quartile is at 24789 pieces. The top 30 highest levels of units sold were of the brand Karvan Cevitam.
- ***PricePU:*** Actual price of 1 unit expressed in euros. Since the 1st and 3rd quartile are at €1.42 and €2.59 respectively, it can be concluded that the price ranges in the market for lemonade are really small.
- ***PricePL:*** Actual price of 1 litre of lemonade in euros. Although quartile ranges are in the same order of magnitude, the maximum price per litre of €9.36 is 3 times as high as the maximum of the PricePU. This means that the lemonade size of the bottles varies a lot. Furthermore, for both the PricePU and PricePL, the median and mean are close to one another indicating a mostly symmetrical distribution in the price data.
- ***BasePricePU:*** Base price of 1 unit expressed in euros. Although the 1st quartile is almost equal to the actual PricePU (€1.43), the 3rd quartile level is €3.18, more than 20% higher than that of the actual PricePU.
- ***BasePricePL:*** Base price of 1 litre of lemonade in euros. A peculiar finding about this variable is that the maximum of €4.94 is almost half of the actual PricePL. This would mean that products have become more expensive during promotional events.
- ***FeatOnly:*** Weighted distribution figure at the chain level for a 'feature-only promotion' in a certain week. A 'feature-only' means outside-store attention for the brand, either in the store flier, an ad in a newspaper or a magazine. With a 3rd quartile level of 0,00, it

becomes clear that Feature Only promotions do not happen very regularly. The mean is 6.43.

- ***DispOnly:*** Weighted distribution figure at the chain level for a "Display-only promotion' in a certain week. A "display-only" means special in-store attention for a brand: a temporary shelf in one of the aisles or a change in the brand's regular shelf. Again, the 3rd quartile level is 0,00 and the mean is 0.98.
- ***FeatDisp:*** Weighted distribution figure at the chain level for when both a Feature and a Display promotion occurs. Again a 3rd quartile level of 0,00 indicates this does not happen regularly. With a mean of 3.14, it becomes clear that a Feature Only promotion occurs either most often or occurs in a larger proportion of a chain when they occur. The Display Only promotion happens either the least or the depth of the promotion is the lowest.

Irregularities and outliers in the data

While working with the dataset we identified some irregularities and outliers. Inconsistencies of this nature can have big effects on the results of the analysis. In the part below we present the issues we have found and how we addressed them.

- There were 128 NAs in the dataset, with missing values for UnitSales, PricePU, PricePL, BasePricePU and BasePricePL. We found these 128 observations belong to the supermarket formula, Hoogvliet, and brand, Teisseire in the time period between 2017 and 2020. This occurred because Hoogvliet didn't offer the Teisseire brand in this period. We decided to not deal with these NAs as they do not affect our brand of interest, Raak.
- The sum of *FeatOnly*, *DispOnly* and *FeatDisp* exceeds 100 for 401 observations, which is theoretically not possible since the numbers represent percentages. This is a significant problem because it will affect our results on the promotional variables. We decided to account for these irregularities by calculating the weighted distribution for the observations where the sum exceeded 100. For example, when *FeatOnly* and *FeatDisp* both have a value of 80, the weighted distribution would be 50/50 after our transformation.
- We also found that for some observations the PricePU exceeded the BasePricePU. This is irregular since we expect PricePU to always be lower than the BasePrice because

promotions generally lower a product's price. As a result, we found negative values for the discount percentages. To account for this, we adjusted the observations for which the discount percentage was negative. We replaced the negative value with a zero, therefore keeping the original intact but making sure there were no negative discounts.

- In the UnitSales data, we identified some outliers per brand. The KarvanCevitam brand has a lot of positive outliers. However, since we expect these peaks to be a logical consequence of promotions and other effects of independent variables we did not replace these outliers to prevent losing potentially important variation in our data.

Extending our dataset

Based on our initial understanding of the dataset and our foresight for further analysis, we have created a number of new variables. In the following section, we will break down the meaning and method of creating these variables.

- ***Date*****:** we used sub(), providing the RegEx pattern to read *Week* and substitute its value, then used ISOweek2date() to finalize the transformation and create the *Date* variable, which is a date object.
- ***Quarter*****:** using lubridate's quarter() on our newly created *Date* variable, we created the *Quarter* variable, which indicates in which quarter an observation takes place, with values ranging from 1 to 4.
- ***Sales*****:** We added another variable named *Sales* which is the result of multiplying *UnitSales* by *PricePU*.
- ***Discount*****:** We created two discount variables: (1) the raw discount (*discount_raw*), calculated as *BasePricePU - PricePU* and (2) the discount percentage (*discount_perc*), calculated as (BasePricePU - PricePU) / BasePricePU * 100.
- ***Adjusted Promotion Variables:*** to address the irregularity in the original promotion variables, we created 3 new adjusted variables indicating the ratio of the promotion features' distribution among the 3 options. To achieve this we employed ifelse() statements which check if the sum of the three original promotional variables is more than 100 per observation. If that is the case then we calculate the adjusted promotional variable, characterized by the following general formula: (Promotional variable / Sum of all 3 Promotional variables) * 100. If the condition was not present then we kept the

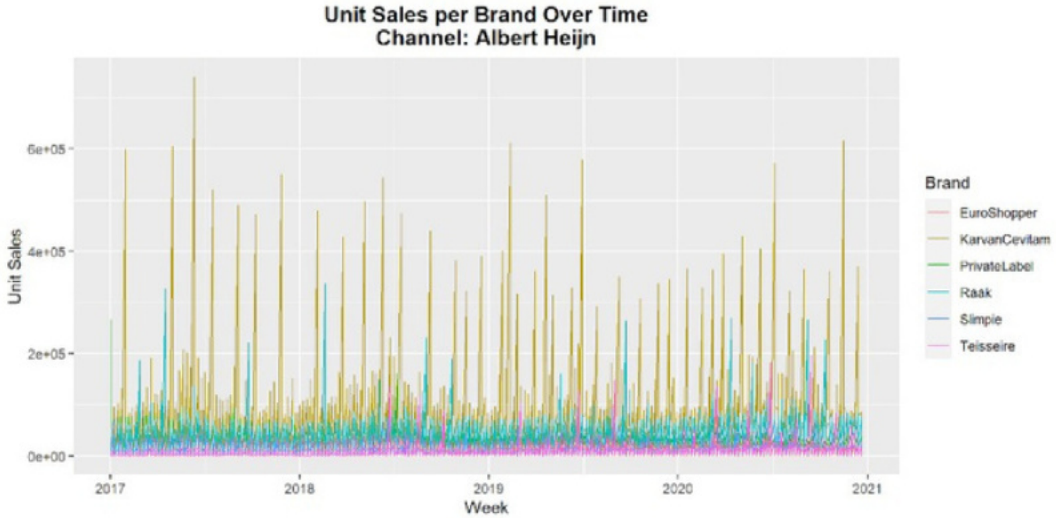
original values. This resulted in the three new variables *FeatOnly_Adjusted, DispOnly_Adjusted* and *FeatDisp_Adjusted.*

- ***Dummy Promotion Variables:*** we created three dummy variables from the promotion variables with 1 indicating if there was a promotion and 0 when there is not.
- ***Promotion Dominance:*** we created a categorical variable that indicates which of the three promotional modes is dominant in that given observation. To achieve this we employed if statements assigning '1' to observations where feature promotion was higher than display and higher than feature+display promotions; '2' to observations where display promotion was higher than feature and higher than feature+display promotions; '3' to observations where feature+display promotion was higher than feature and higher than display promotions; finally if there is no dominant promotion, we assign an NA so that the observation is not used in the analysis.

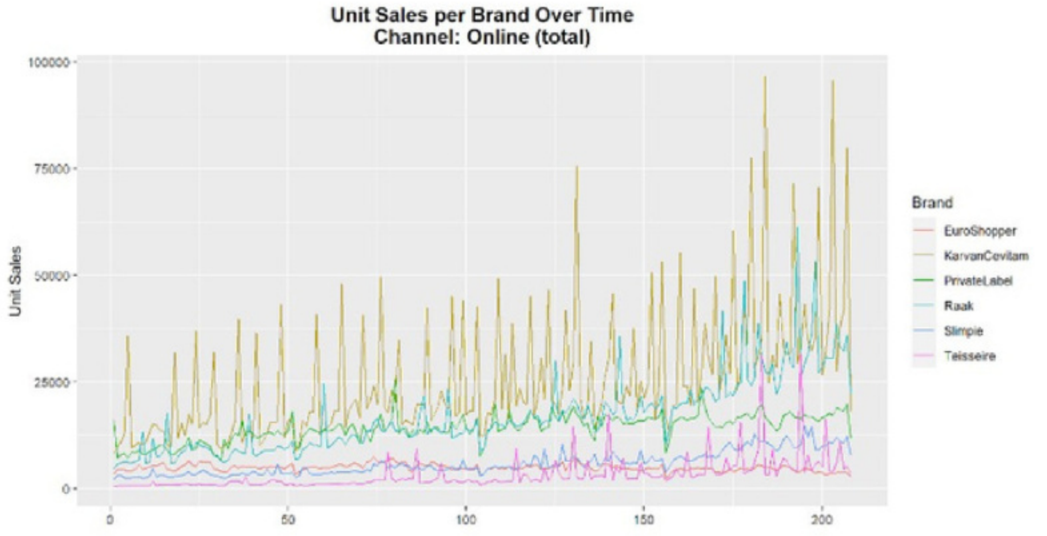
Relevant insights into the variables

1. UnitSales

The figure below shows the Unit Sales per Brand over time. It instantly becomes clear there is fluctuation in sales for most of the brands. However, for Karvan Cevitam the peaks are the largest. The graph also shows that Karvan Cevitam is the most popular brand over time and that Raak and PrivateLabel are also quite popular.



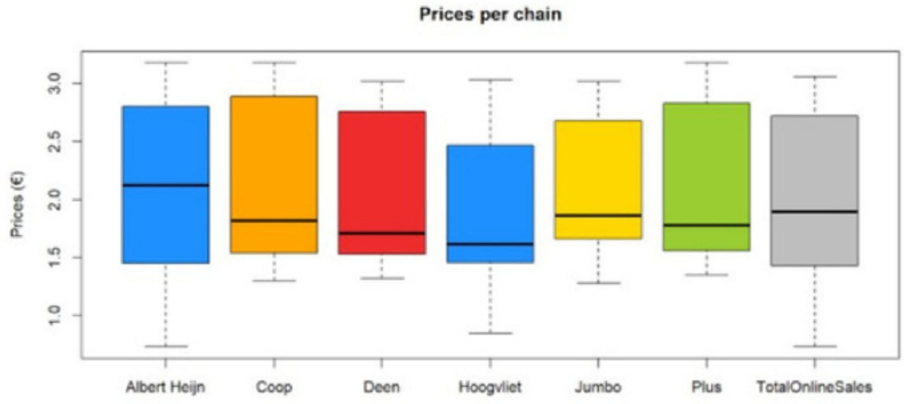
Another interesting insight we found was about online Sales. Total Online Sales show an increasing trend over time for all brands, as shown in the figure below.


Unit Sales per Brand Over Time
Channel: Online (total)

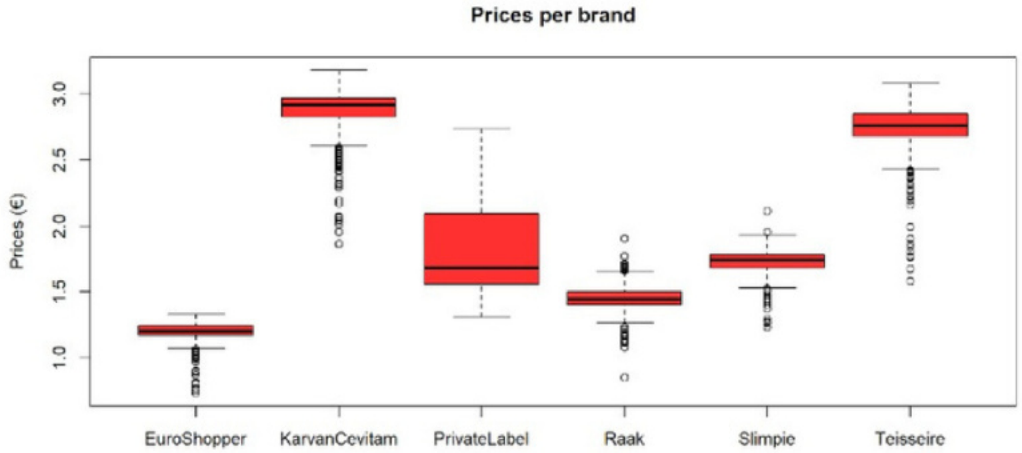
Additionally, we found that the brand EuroShopper is only sold by Albert Heijn, while the other brands are sold in all the chains.

2. Price

As the figure below shows, there are some clear differences between the prices per chain. The Albert Heijn chain is clearly a more premium chain since the average price is much higher (around €2.15), while Hoogvliet has the lowest price on average. The other chains are more equal in terms of prices.


Prices per chain

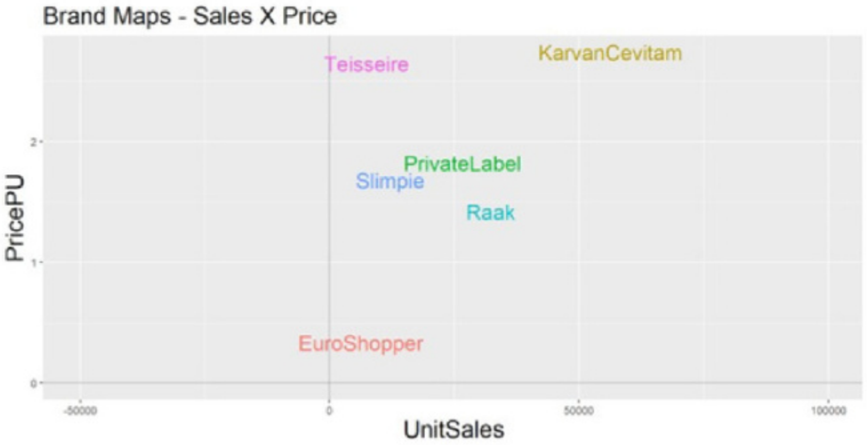
Regarding the prices per brand, we identified clear differences between brands. Both KarvanCevitam and Teisseire are premium brands with average prices between €2.50 and €3.00. EuroShopper is clearly the cheapest brand with an average price of around €1.20. The other 3 brands are relatively equal in terms of price. As we did not have any data on the sizes of the unit, we decided to use per-unit prices as our main price index instead of per liter.



3. Brand positioning

Regarding the positioning of brands, we found multiple insights based on the figure below.

- KarvanCevitam and Teisseire are the "premium" brands with the highest PricePU. However, Karvan Cevitam is much more successful in terms of UnitSales.
- EuroShopper is the real price fighter with clearly the lowest price.
- The other 3 brands are direct competitors in terms of quantities and PricePU.

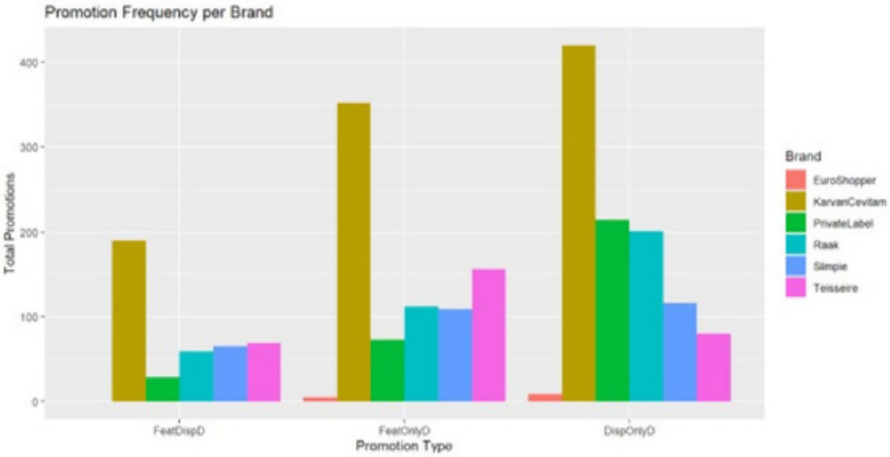


4. Promotional frequency and discount depth per brand

Regarding the promotional frequency and discount size, we found large differences between brands and chains. *Insights frequency per brand:*

- The Karvan Cevitam brand has promotions much more frequently than the other brands.
- The EuroShopper brand almost never offers a promotion.

- The Display Only promotions occur more frequently in general compared to the Feature Only and Feature & Display promotions. The combined promotions occur the least frequently.



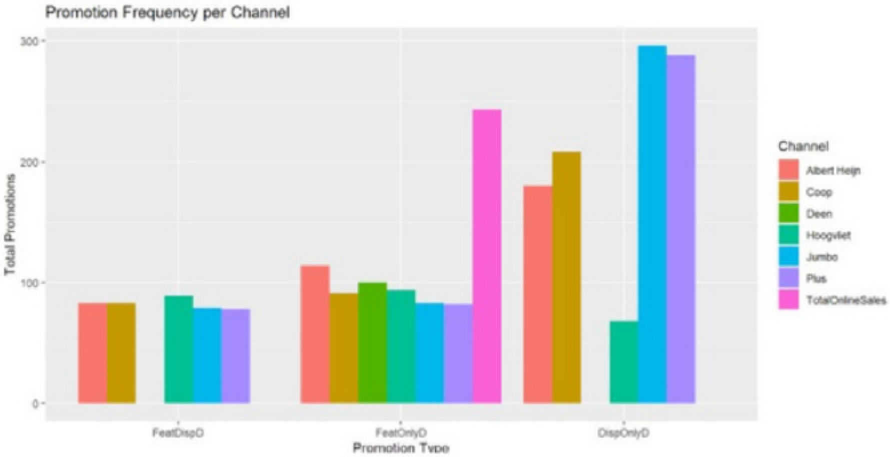
Insights discount size per brand:

- The mean discounts vary between 0% and 5% per unit. Karvan Cevitam (4,9%) and Teisseirre (3,5%) offer the largest discounts, while EuroShopper (0,2%) and PrivateLabel (0,7%) offer the smallest discounts. Raak and Slimpie are in the middle with 2,0% and 2,5% respectively.

5. Promotional frequency and discount size per chain

Insights frequency per chain:

- The number of Feature Only promotions is quite equal for all chains. Only for Online Sales the Feature Only promotions occur more frequently.
- The number of Display Only promotions occurs most frequently, which is generally caused by Jumbo and Plus. Albert Heijn and Coop also mostly do Display Only promotions, while Deen and Hoogvliet almost never do Display Only promotions.
- The Feature and Display promotions occur quite equally for all chains, except for the Deen chain, which does not do any Display promotions.



Insights discount size per chain

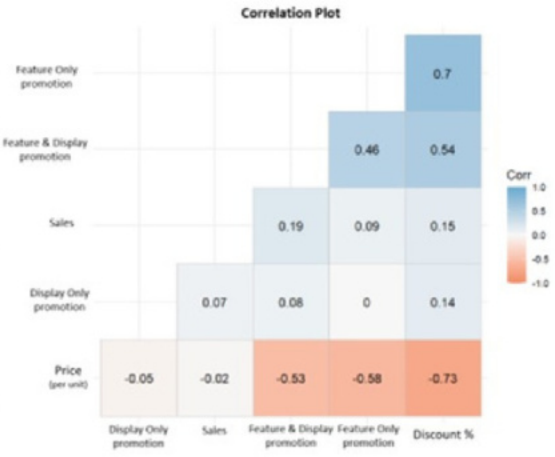
- The differences between the chains are much smaller than the differences between brands regarding the discount size.
- Deen (3,9%) and Plus (3,4%) are the chains with the largest discounts on average. Albert Heijn and Online Sales are equal around 3%, while Coop has a mean of 2%. Hoogvliet (1,2%) and Jumbo (1,3%) are the chains with the smallest discounts.

6. Seasonality

We also found significant differences between the 4 seasons. For all the different chains, Unit Sales were much higher during the second and third quarter compared to the first and fourth quarter. It seems like there is a strong relationship between seasonality and lemonade consumption, which has also been seen in long-standing research by Hoos (1956).

Relationships among the variables

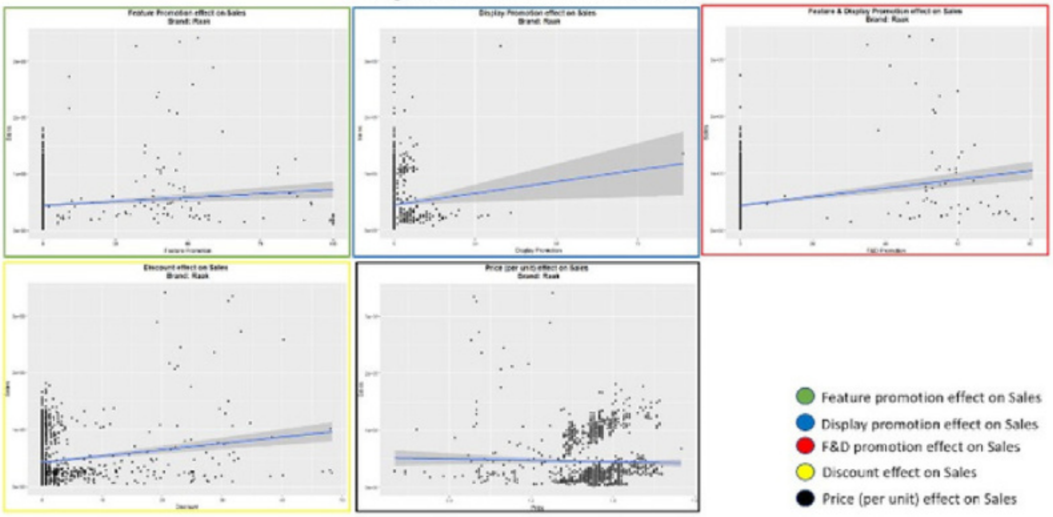
1. Relational plots



To get an initial idea of possible relationships among the variables for our model, we first looked at their correlations, as illustrated in the figure to the right.

Firstly, we examine our KPI variable, sales, and find that all promotional types, as well as discounts, are positively correlated with sales. To confirm this, we consulted with past research and found that both price discounts and in-store promotions have been well-documented to have a positive effect on sales (Blattberg et al., 1995; Blattberg & Neslin, 1993; Martínez-Ruiz et al., 2006; Nijs et al., 2001). Next, we notice that discounts and the three promotional modes are also positively correlated, this could be because retailers promote the products they have discounted, therefore when there are promotions there are also more discounts. The effectiveness of this is demonstrated when shelf promotion is combined with discounts (Garrido-Morgado et al., 2021). Finally, we note the negative correlation between price and feature-only, feature+display promotions, as well as discounts. This is consistent with the previously-described positive correlations.

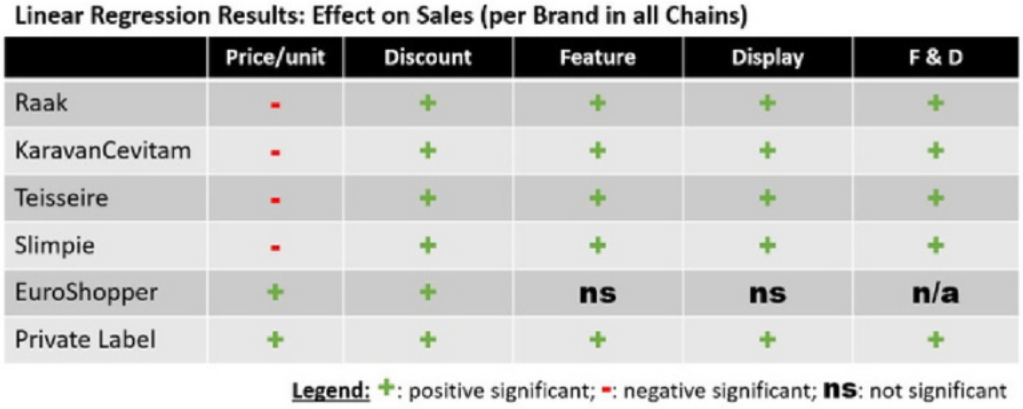
Regression Line Plots: Effects on Sales



Next, we investigated how Sales are affected specifically for our brand, Raak. To do this we fit linear regression lines for each of the three promotion variables, discount and price versus sales, as shown in the figure above. We found that feature-only promotions seem ineffective, as there are high sales both when promotions are present as well as when they are not. Additionally, the regression line is nearly flat. Next, for display-only, we can see that Raak does not do this type of promotion very much, as there are few observations in the plot. Next, the effect on sales, when there is both display and feature promotion, is more visibly positive. Both the observations and the regression line showcase that sales are higher when there is more promotion. However, we also notice that towards the maximal levels of promotion the sales are lower. This can be explained in multiple ways, but one reason is that there is an optimal level of promotion, which once surpassed does not contribute to sales any further. Finally, for both price and discount, we see that sales are not dramatically affected. This can be explained by the fact that Raak is a brand which maintains good-value prices and has stable demand at their chosen price levels.

2. Numerical Analysis & Relationships

To extend our previous findings from the linear regression line plots and literature, we performed several one-to-one regressions to see if the relationships between the variables are significant and in what direction. We did this per brand in all chains, illustrated in the table below:



Linear Regression Results: Effect on Sales (per Brand in all Chains)

	Price/unit	Discount	Feature	Display	F & D
Raak	-	+	+	+	+
KaravanCevitam	-	+	+	+	+
Teisseire	-	+	+	+	+
Slimpie	-	+	+	+	+
EuroShopper	+	+	**ns**	**ns**	**n/a**
Private Label	+	+	+	+	+

Legend: +: positive significant; -: negative significant; **ns**: not significant

Across all brands, besides EuroShopper & Private labels, we see identical effects between sales and the other variables. According to the results, the price has a significant negative effect on sales, while discount and promotion types have significant positive effects. It is important to note EuroShopper and Private labels have constant value prices, hence the positive effects of price on sales. Furthermore, the promotional variables are insignificant for EuroShopper, this is because the brand is online-only and these three promotional variables are not applicable to it.

Next, we performed a one-way ANOVA to test whether observations with one dominant promotion type have variance in their sales. We found the results to be significant, indicating that the means between groups are different. Finally, to understand which of the groups has a different mean we performed a TukeyHSD test. Here we found the difference in the mean sales to be statistically significant only between the feature promotion dominant group and the display promotion dominant group.

Finally, we performed a one-way ANOVA to test whether there is a seasonality effect. We found the results to be significant, indicating that the mean sales between quarters are different. Finally, to understand which of the quarters had a different mean we performed a TukeyHSD test. Here we found the difference in the mean sales to be statistically significant between the first and second quarters, the third and fourth, as well as the second and fourth.

Based on all of these findings, we preliminarily conclude to use price/unit, discount, the three promotion variables and the quarter variables in our model.

Model specification

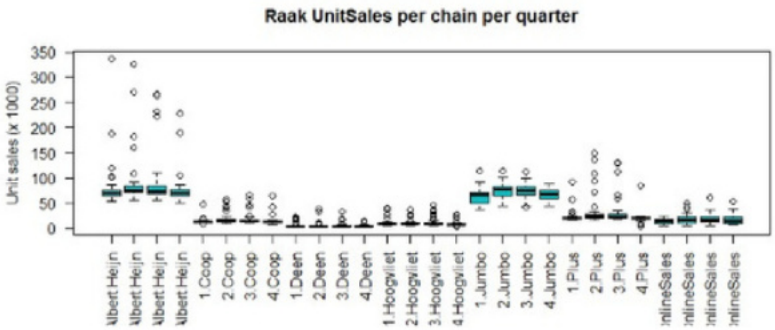
1. Model type

We are specifying a predictive model that forecasts Unit Sales of the brand Raak. This means that our model aims to understand what will happen in the future. We are analyzing the Brand Sales level of demand since we are aiming to predict the future sales of one specific Brand.

2. Linear or Multiplicative

Despite the log transformation being needed before estimating with the multiplicative model, the research of Dai et al. (2022) has demonstrated that interaction effects exist between display-related promotions and price discounts when modelling sales, suggesting that a multiplicative model may be more appropriate if we want to measure interaction effects and elasticity. Besides, from our correlation analysis among independent variables, they seem to have interesting interactions across the years for Raak data. So we decided to build a simple multiplicative model (non-linear in parameters but linearized) in the following sections.

3. Model elements



By performing one-to-one linear regressions of our dependent variables and unit sales, we found that log(*Price), adjusted display, adjusted feature,* and *adjusted display+feature* have a significant effect on *Unit Sales* for Raak data, also seen in the correlation plot below and earlier. Furthermore, unit sales vary across the four quarters according to our ANOVA test. However, the unit sales do not significantly vary across quarters for Raak data at a specific chain. Additionally, we find that quarter 2 and quarter 3 contribute to the most unit sales across the year on Raak data of each Chain.

Hence, we decide to choose 4 variables as the independent variables, excluding *Quarter* due to the ANOVA test results above. Furthermore, we want specific estimates from Raak at each chain. The independent variables are *PricePU, adjusted Feature, adjusted Display and adjusted Feature +Display*, and the dependent variable is *Unit Sales.*

4. Multiplicative Model

In this section, we conduct both not-pooled and pooled methods for the multiplicative modelling.

Step 1: First, by applying the unit-to-unit (not-pooled) method, we model unit sales of Raak as a function of own price, own adjusted display and own adjusted feature and adjusted display+feature, on the dataset of Raak in each chain. This results in *7* models in total, with independent variables that have only positive values being the base and those with 0 or negative values being the exponent**:**

$$S_{it} = \alpha_i \, {P_{it}}^{\beta 1i} \, \beta_{2i}^{F_{it}} \beta_{3i}^{D_{it}} \beta_{4i}^{FD_{it}} \varepsilon_{it} \quad (*)$$

- *with i = 1, …, 6 brands, t = 1, …, 208 weeks*
- *where*

S_{it} = sales of Raak at Chain i, in week t

P_{it} = price of Raak at Chain i, in week t

F_{it} = adjusted use of feature for Raak at Chain i, in week t

D_{it} = adjusted use of display for Raak at Chain i, in week t

FD_{it} = adjusted use of F+D for Raak at Chain i, in week t

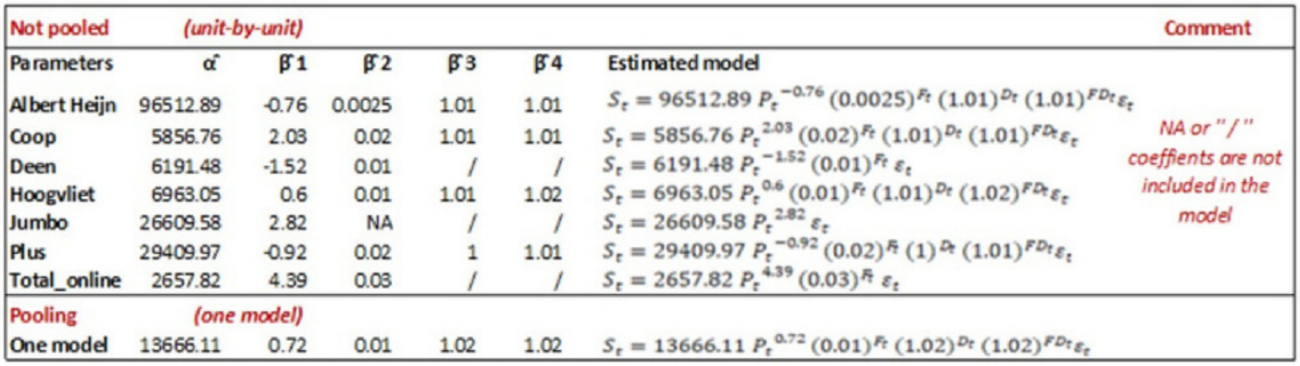
Step 2: For a certain chain, after natural log transformation on the first equation (*), we are able to run a linear regression in the following equation (^):

$\ln S_{it} = \ln(\alpha_i) + \beta_{1i} \ln P_{it} + F_{it} \ln \beta_{2i} + D_{it} \ln \beta_{3i} + FD_{it} \ln \beta_{4i} + \ln \varepsilon_{it}$, *which is also written as*
$S_{it}^{*} = \alpha_i^{*} + \beta_{1i} P_{it}^{*} + \beta_{2i}^{*} F_{it} + \beta_{3i}^{*} D_{it} + \beta_{4i}^{*} FD_{it} + \varepsilon_{it}^{*}$ (^)

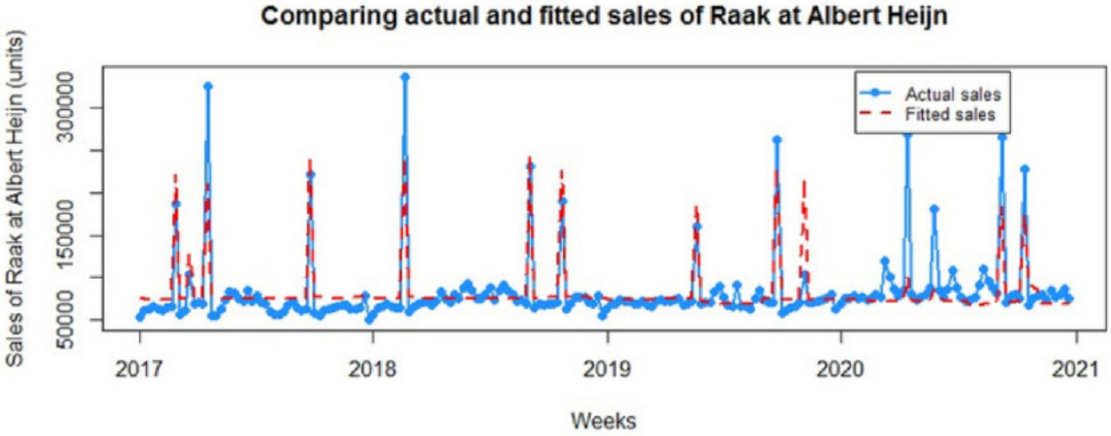
Step 3: With antilog transformation, we can get the original parameters in equation *. For example, we can get α^ by applying $\hat{\alpha} = \exp(\ln(\hat{\alpha}))\ \exp(-\frac{1}{2} var(\ln(\hat{\alpha})))$

with ln (α ̂) = α ̂ ̂*. In a similar fashion, we can get β_2, β_3, β_4. Specifically, β_1 is an elasticity estimate that indicated the percentage change in sales according to the percentage change in price.

Finally, we arrive at a table of estimates and plots of actual vs. fitted sales in the estimated model of Raak at each chain:



Not pooled	(unit-by-unit)						Comment
Parameters	$\hat{\alpha}$	$\hat{\beta}$1	$\hat{\beta}$2	$\hat{\beta}$3	$\hat{\beta}$4	Estimated model	
Albert Heijn	96512.89	-0.76	0.0025	1.01	1.01	$S_t = 96512.89\ P_t^{-0.76}\,(0.0025)^{F_t}\,(1.01)^{D_t}\,(1.01)^{FD_t}\varepsilon_t$	
Coop	5856.76	2.03	0.02	1.01	1.01	$S_t = 5856.76\ P_t^{2.03}\,(0.02)^{F_t}\,(1.01)^{D_t}\,(1.01)^{FD_t}\varepsilon_t$	NA or "/" coeffients are not included in the model
Deen	6191.48	-1.52	0.01	/	/	$S_t = 6191.48\ P_t^{-1.52}\,(0.01)^{F_t}\,\varepsilon_t$	
Hoogvliet	6963.05	0.6	0.01	1.01	1.02	$S_t = 6963.05\ P_t^{0.6}\,(0.01)^{F_t}\,(1.01)^{D_t}\,(1.02)^{FD_t}\varepsilon_t$	
Jumbo	26609.58	2.82	NA	/	/	$S_t = 26609.58\ P_t^{2.82}\,\varepsilon_t$	
Plus	29409.97	-0.92	0.02	1	1.01	$S_t = 29409.97\ P_t^{-0.92}\,(0.02)^{F_t}\,(1)^{D_t}\,(1.01)^{FD_t}\varepsilon_t$	
Total_online	2657.82	4.39	0.03	/	/	$S_t = 2657.82\ P_t^{4.39}\,(0.03)^{F_t}\,\varepsilon_t$	
Pooling	(one model)						
One model	13666.11	0.72	0.01	1.02	1.02	$S_t = 13666.11\ P_t^{0.72}\,(0.01)^{F_t}\,(1.02)^{D_t}\,(1.02)^{FD_t}\varepsilon_t$	



5. Pooled or Not pooled

In addition to the unit-by-unit method (not-pooled), one general model (*pool) is estimated for all chains on Raak sales by applying the pooling method, which requires regression on only one single dataset containing all independent variables and dependent variables' information across 7 chains from a top to bottom manner. And here is the functional form of the pooled model without chain dimension i**:**

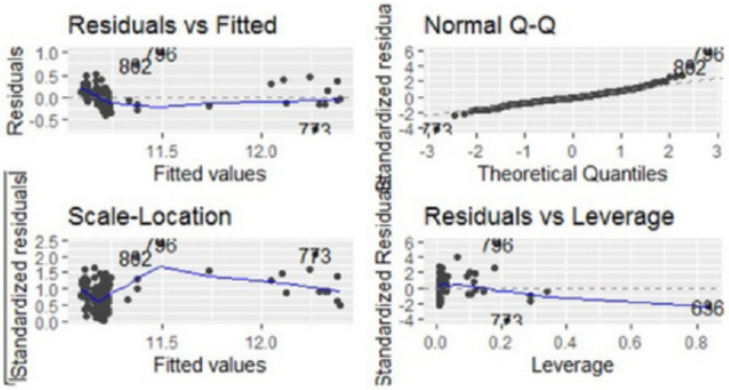
$$S_t = 13666.11\ P_t^{0.72}\,(0.01)^{F_t}\,(1.02)^{D_t}\,(1.02)^{FD_t}\varepsilon_t \qquad \text{(*pool)}$$

- where

  S_t = unit sales of Raak at all chains, in week t
  P_t = price of Raak at all chains, in week t
  F_t = adjusted use of feature for Raak at all chains, in week t
  D_t = adjusted use of display for Raak at all chains, in week t
  FD_t = adjusted use of F+D for Raak at all chains, in week t

However, we decide to choose the unit-by-unit method for the following reasons. Although we only have a few observations (208 at each chain) from the not-pooled way, we can get chain-specific estimates based on Raak data per chain. That is, detailed insights of Raak from estimates can be seen for different chains that vary a lot in terms of chain size, characteristics and targets. Besides, the Raak brand also fluctuates differently per chain in terms of Sales.

So, we believe it is generally a simple and basic model for initial analysis from the plots and statistics, while further improvement and robustness need to be done in the future regarding the NAs in the coefficients, positive estimates of *Price* and model assumption in some chains, as seen in the figure below.



6. Dynamic effect

The current effects model would not be the most ideal one as promotional activities might have a pre- and post-sales effect. Furthermore, in-store promotional activities are associated with creating physical marketing elements and setting them up. This may create delayed-response effects, such as execution and noting delays (Leeflang et al., 2016). Additionally, as lemonade is a specialty product with seasonal effects (Hoos, 1956), we expect there to be customer-holdover effects. More specifically an increased purchase holdover effect, where the marketing stimulus increases the average quantity purchased per period for some time. We believe the Partial Adjustment Model is the most appropriate one since we don't have to include an extra variable, while in the Direct Lag Model we would have to make adjustments to each variable (adding lags). Finally, in the research of Paul et al. (2009), a Partial Adjustment Model was used to measure short- and long-run price elasticity per customer group, region and season. Therefore it may be appropriate in our case in measuring short- and long-run price elasticity for each chain.

7. Error term

As our model indicates, we included an error term for estimating the linear model. To make sure our model is valid, the following assumptions need to be true. In our estimation we need to check for:

a. Mean zero. The mean of the error term needs to be zero.
b. Normality. The error should be normally distributed.
c. Independence. The covariance of the error term should be zero.
d. Homoscedasticity. The variance should not increase or decrease when the dependent variable increases or decreases.

Since we do not have data or detailed information on how the data is gathered, we can not exclude the option that the dataset contains no errors in measurement such as sampling error or poor measurements instrument. The variation in the error term could also be caused by missing values. However, since we do not have missing values in our dataset for model specification the variation of our error term would not be affected by this. For the Deen chain, for example, we do not have a coefficient for the Display Only and Feature + Display variable but this is because these values are 0 and not missing. Next to these assumptions, we will also need to make sure the covariance of the independent variables is zero and the parameters should be constant over time. All of these assumptions will be checked in a later phase.

REFERENCES

Blattberg, R. C., Briesch, R., & Fox, E. J. (1995). How promotions work. Marketing science, 14(3), G122-G132.

Blattberg, R. C., & Neslin, S. A. (1993). Sales promotion models. Handbooks in operations research and management science, 5, 553-609.

Dai, H., Ge, L., Li, C., & Wen, Y. (2022). The interaction of discount promotion and display-related promotion on on-demand platforms. Information Systems and e-Business Management, 20(2), 285-302.

Garrido-Morgado, Á., González-Benito, Ó., Martos-Partal, M., & Campo, K. (2021). Which products are more responsive to in-store displays: Utilitarian or hedonic?. Journal of Retailing, 97(3), 477-491.

Hoos, S. (1956). Lemon industry in California: Long-term projection of market potential for lemon juice products based on variable determinants of summer demand. California Agriculture, 10(10), 2-15.

Leeflang, P. S. H., Wieringa, J. E., Bijmolt, T. H. A., & Pauwels, K. H. (2016). Modeling Markets. Springer-Verlag New York.

Martínez-Ruiz, M. P., Mollá-Descals, A., Gómez-Borja, M. A., & Rojo-Álvarez, J. L. (2006). Using daily store-level data to understand price promotion effects in a semiparametric regression model. Journal of Retailing and Consumer Services, 13(3), 193-204.

Nijs, V. R., Dekimpe, M. G., Steenkamps, J. B. E., & Hanssens, D. M. (2001). The category-demand effects of price promotions. Marketing science, 20(1), 1-22.

Paul, A. C., Myers, E. C., & Palmer, K. L. (2009). A partial adjustment model of US electricity demand by region, season, and sector.