Report

SECTION I

We are interested in understanding how variables influence the monthly sales of cold cereal for a supermarket chain. Understanding how factors are associated with sales will allow the chain to efficiently allocate its scarce resources in order to maximise profits. In the EDA, we will identify crucial variables to build a model that can accurately predict the number of cereal sales. This should allow stores to minimise costs or missed profits in an industry characterised by narrow profit margins.

The dataset, collected over three years, contains 24 variables and gives results for 7500 values of units sold monthly. There is no missing data.

The outcome variable is the units of monthly sales of cold cereal across stores in the US. The number of units range from 1 to 2260, with mean 145.54 and standard deviation 133.56.

We identified critical factors on the outset from knowledge gained from the literature that may guide our EDA:

Promotional methods directly impact sales [1]. Therefore, understanding the influence of various promotion styles is crucial for predicting sales. The effectiveness of price-promotions is heavily influenced by socio-economic factors, where more price-sensitive shoppers are likely to respond more to promotions [3]. Thus, the accuracy of our model may benefit from us identifying a way to cluster consumers into groups that reflect this.

There are a multitude of factors influencing the characteristics of observed sales data and underlying demand. Before starting the EDA, we categorised the potential predictors as one of two kinds;

Endogenous Factors (within store control): UPC, MANUFACTURER, SUBCATEGORY, BASE PRICE, PRICE, promotion methods (TPR, FEATURE, DISPLAY), and availability of the product (NWEEK).

Exogenous Factors (not within stores' control): STORE_NUM, CITY, STATE, AREA_CODE, STORE_TYPE, AVG_WEEKLY_BASKETS, time (MONTH, YEAR).

Our EDA considers covariates in each category sequentially. Categorical covariate analysis uses boxplots to describe the variability of monthly sales for each factor of the covariate in question.

From our analysis of the endogenous covariates, we conclude that CATEGORY is not a meaningful predictor as all observations belong to one category. Figures 7 and 8 of PRICE and BASE PRICE have similar shapes. These two covariates alone do not tell us much about the consumers' perception of the product, so it may be helpful to introduce additional covariates incorporating both these covariates.

Figure 3 describes how the availability of a product limits sales, wherein observations valued at less than 4 reflect a lack of availability. This highlights an important issue as it may hinder our investigation by not accurately reflecting consumer demand and hence the true impact of certain covariates on the sales.

Conclusions drawn from covariates with uneven sample groups should be taken cautiously, and Figures 1-10 suggest that all covariates influencing monthly sales should be considered as potential predictors in our model.

From the exogenous factors analysis, we identified three covariates to cluster. Although STORE_TYPE in Figure 12 loosely groups consumers by shopper profiles, a consumer's geographical location reflects their economic status more explicitly. Given that the CITY covariate has the most factors, it is the optimal covariate to regroup. Likewise, STORE_NUM also has a large number of factors and may benefit from regrouping.

Figures 14-16 were plotted to identify seasonal trends that are consistent across years. Unexpected drivers of sales such as abnormal events manifest as random disturbances to the time series. This may explain the isolated period with a large sales volatility in February 2009. Without access to weather conditions, we can regroup MONTH as SEASONS to trivially reflect the impact of weather on sales.

Figure 13, implies a positive relationship between the size of the store and the number of units sold. This is intuitive as stores that accommodate larger amounts of shoppers are more likely to sell more units of a product. Another issue is highlighted as units will be proportionally related to the size of the store.

The plots for ID and CATEGORY are omitted as they do not offer substantial information. All other covariates influence sales and are potentially viable predictors to consider for our model.

SECTION II

Before building the model, we first construct new covariates from those identified in the EDA.

While analysing endogenous factors, we found that introducing a covariate that measures the percentage discrepancy in price may be meaningful for our investigation. This is the DISCOUNT covariate, computed as the percentage decrease of PRICE from BASE_PRICE. The values are rounded to 3 decimal points. We theorised that the percentage reduction in price directly influences a shopper's decision to make a purchase. This is corroborated by Figure 9. (Note that there are some anomalous values with negative discounts, likely due to human error in the price labelling.)

The EDA suggests grouping MONTH into SEASON, but this was found to be unhelpful.

CITY was identified as a categorical covariate that would benefit from regrouping. The notion that the geographical location of a shopper may attest to their socioeconomic status, and hence we use STORE_TYPE and AVG_WEEKLY_BASKETS covariates as numerical covariates to cluster CITY. Although STORE_TYPE is categorical, we remedied this by introducing new covariates of counts of each STORE_TYPE in each CITY. Cities with a similar number of stores and a similar distribution of store types were grouped. The AVG_WEEKLY_BASKETS of all the stores in the same city were accumulated, to reflect the aggregate transactions of a city. This clustering provides us with relatively even samples while also reducing the number of groups from 51 to 4, alleviating the impracticality of covariates with huge numbers of groups. In Table 1, we can see the results of clustering with the mean proportion of each STORE_TYPE and AVG_WEEKLY_BASKETS.

Different promotional methods have different modalities, and thus the interplay of strategies may lead to differing promotional efficacies. To introduce this into our model without using a three-way interaction term, we decided to cluster STORE_NUM by their predominant promotional strategies, i.e. either TPR_ONLY, DISPLAY, FEATURE, or even a mix of all three. These covariates were provided as proportions of the months active, so we multiplied them by NWEEKS to adjust them to monthly data. Thus, we use these as the numerical covariates to cluster STORE_NUM. In Table 2, we can see the results of the STORE_NUM clustering with the mean proportion of each promotional style.

Hence, we omitted STORE_NUM, MONTH, and CITY from our set of covariates for model building as they were clustered in a manner that allowed for more purposeful interpretation. UPC was omitted as the information it provided was better encapsulated by the other covariates such as MANUFACTURER and SUB_CATEGORY. Additionally, as already mentioned in the EDA, ID and CATEGORY are omitted.

Thus we proceed to build the model with our new set of covariates. The first step to building a model is to determine the type of model we will use.

We first considered the linear model. The assumption of normality, whereby residuals are normally distributed (a prerequisite of linear models), was found to not be satisfied by the cereal data. To show this, a linear model consisting of our chosen set of covariates was built and a diagnostic plot was drawn from it. From the Normal Q-Q plot in Figure 17, it is evident that the data is positively skewed as the tail is distinctively departed from the 45-degree line. This is further corroborated by the plot of Residuals vs. Fitted where the variance grows with the magnitude of the fitted value. These points led us to conclude that there was insufficient evidence to use a linear model.

This left us with two equally viable contenders, the generalised linear model (GLM) and the generalised additive model (GAM). The purpose of the report is to explain the reasons and magnitude by which each covariate influences the monthly sales. Although a GAM is likely to produce a better fit than a GLM, we decided it would not be in the interest of the report to compromise interpretability for potential improvement of accuracy. Thus, we proceed with the GLM, as it strikes a good balance between the prediction accuracy and interpretability.

An intuitive GLM family to use would be Poisson, since the response covariate UNITS is in counts and the Poisson distribution, like the data we have, is positively skewed.

The summary of the Poisson GLM revealed that all covariates had very small p-values, which usually is an indication of their relevancy and evidence for their retention. However, this does not seem to be the case here. We checked the deviance of the model, which turned out to be around 33, and is far larger than 1 - the theoretical deviance of Poisson GLM. This indicates that the model is overdispersed. [2]

Overdispersion can be accounted for by changing the family from Poisson to QuasiPoisson, by including an additional deviance parameter in the model. However, QuasiPoisson does not have an AIC, and predicting and comparing them are complicated, thus not that desirable.

Thus, we considered the Negative Binomial family, which is somewhat an upgrade of Poisson GLM [2]. After looking at the summary of the built Negative Binomial model, the deviance becomes close to one. The Q-Q plots of the Poisson and Negative Binomial (Figures 18 and 19) illustrate the improvement graphically. It is also easier to predict and compare Negative Binomial GLMs, which is good.

With the finalised model, we began to selectively remove covariates. The algorithm of covariates removal follows:

(1) find a covariate in summary with large p-value
(2) build a model without that covariate
(3) run ANOVA with the nested models
(4) remove the covariate if the p-value of ANOVA is large and not remove otherwise.

Out of the covariates with large p-values (`YEAR`, `Grp_STORE_NUM1`, `STORE_TYPE`, `STATE`), our process only removed `YEAR`.

When considering the implementation of interactions, we have made our judgments based on the principle of parsimony, by trying to balance the weight of an additional interaction; and the additional complexity, in terms of the number of factors the categorical covariate adds, creates to interpret the model overall. One interaction we tried was between `Grp_CITY` and `discount` to investigate whether our socio-economic groups have an impact on the effectiveness of a promotion, but we found its p-value to be insignificant. The interaction that proved to be significant was the interaction between `MANUFACTURER` and `discount`.

Figure 20 shows that our assumptions are corroborated. The Q-Q plot shows that the points are mostly on the diagonal. The residual plot shows that approximately 95% of our standardised residuals lie between the values -2 and 2, and the Cooks plot reveals only few leverage points. After looking into those specific cases, there were not any clear reasons for removing the points from the training data, despite the marginal improvement in AIC (-42) in the resulting model.

SECTION III

Our final model is given by:

```
glm.nb(formula = UNITS ~ factor(NWEEKS) + factor(TPRmonth) + factor(DISPLAYmonth) +
factor(FEATUREmonth) + PRICE + BASE_PRICE + MANUFACTURER + SUB_CATEGORY + STORE_TYPE +
AVG_WEEKLY_BASKETS + discount + factor(Grp_CITY) + factor(Grp_STORE_NUM1) + factor(MONTH)
+ discount * MANUFACTURER, data = train1, init.theta = 3.973729105, link = log)
```

Overall, the p-values we obtained are more than reasonable. Over 54 different covariates, only 6 have a p-value superior to 0.05, and the standard errors of our estimates are all at most of order 10^-1, which is reasonable as our estimates are generally of the same order.

We will interpret the coefficients with respect to either endogenous or exogenous categories as defined in the EDA.

Factors within the control of retailers (endogenous) are integral to understand in order to implement effectively. Most of the endogenous factors have a p-value inferior to 10^-3, suggesting the selected covariates are all highly correlated to sales. Two of the endogenous covariates deserve special mention: `NWEEKS` and `discount`.

`NWEEKS` has the largest estimated difference between factors compared to other categorical covariates. For instance, the expected difference in the number of units sold between four-week display and one-week display is approximately 23 units (exp(3.145)). Retailers should thus be conscious of exhausting promotional methods during periods of understocking, as they will not yield comparatively fruitful results. The unavailability of a product intuitively inhibits sales, however, we must note that observations reflecting a product being out of stock do not accurately reflect consumer response to other factors, they only reflect the number of products available. This poses the first limitation wherein analysis of the effectiveness of other factors is hindered by understocking.

`discount` has an estimated coefficient of 2.006, the highest positive estimated coefficient, and also the highest magnitude of the numerical data. Its interpretation details that a 1% increase in discount accounts for an ~ 2% (exp(002006)) increase in the number of units sold, yet this does not indicate one should increase discount endlessly, as that would reduce overall profits. Thus, stores should utilise this finding on a case-by-case basis relating to exogenous factors to determine the discount to apply.

Note that the summary reveals that a higher base price relates to lower sales whereas a higher selling price relates to more sales. Taking into account how `discount` has the largest estimated impact on sales, we can infer that consumers value `discount` higher than `PRICE`. Meaning consumers would rather purchase a product with a higher selling price over one that is cheaper if the perceived "money saved" i.e discount is substantial.

The coefficient estimation tells us that although General MI is the most popular brand, discounts applied to private labels are more effective. This tells us stores should aim to stock more General MI cereal but will receive the best value for money in advertising Private Label cereals.

For exogenous factors, cereal purchases are more frequent in certain periods i.e. January, February, and August. However, our model suggested the removal of `YEAR`, which may be a consequence of not accounting for time series.

The clustering of `CITY` reveals that the aggregate transactions of a city take precedence over the proportion of store types, where Group 4 has the largest average impact on sales, followed by Group 2 which has the highest proportion of upscale stores. Our attempt to segment consumers into socio-economic classes based on the types of stores in their location proved unsuccessful. Access to information on median income per city might have supplemented this issue and revealed the relationship between price sensitivity and promotions more explicitly.

In conclusion, the model adequately presents the main factors affecting the number of units. The most notable issue is the unaccounted time series and the failure to accurately segment consumers into groups reflecting socio-economic status.

Table 1: Group Summary of CITY

Table 2: Group Summary of STORE_NUM

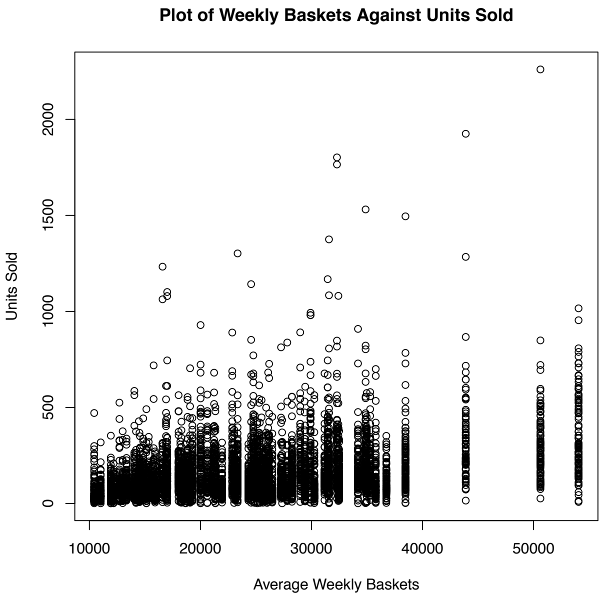

Figure 13: Plot showing relationship between # of units sold and avg. weekly baskets.

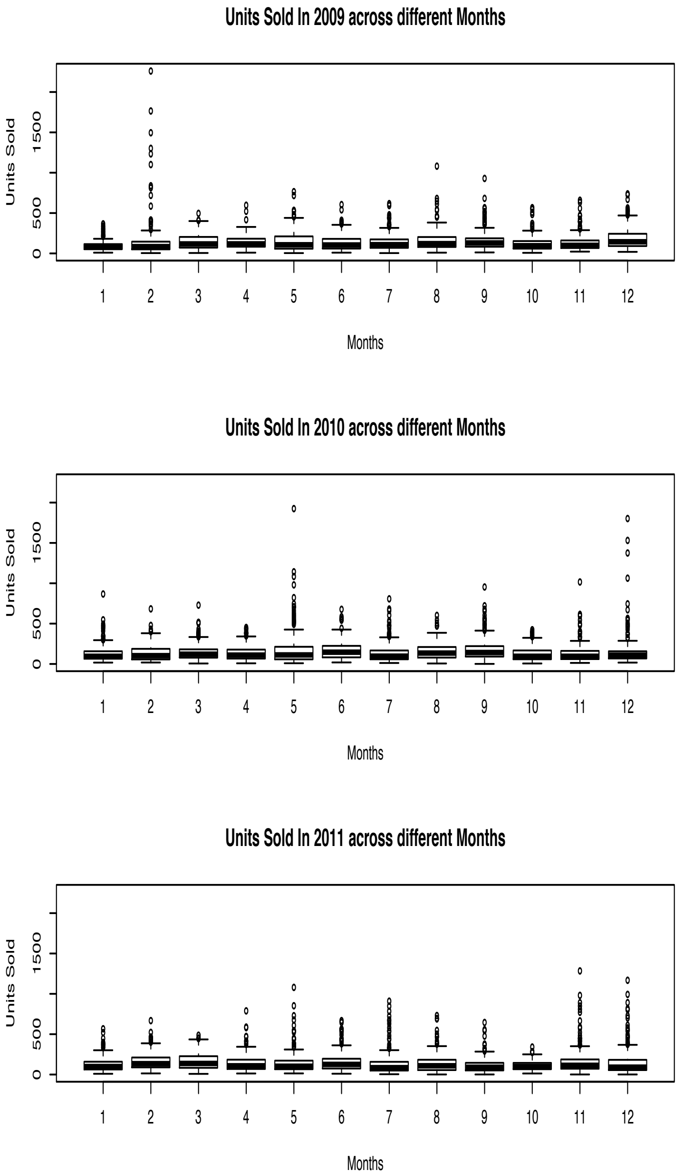

Figure 14-16: Boxplots showing the number of units of cereal sold in each year across the months.

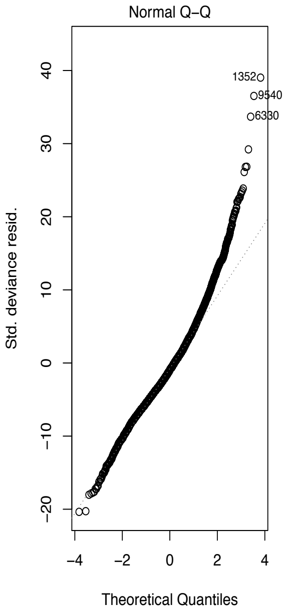

Figure 18: QQ Plot of Poisson GLM

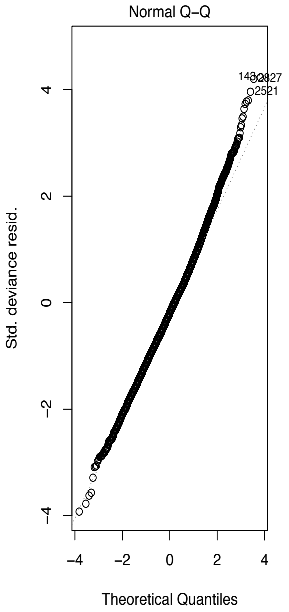

Figure 19: QQ Plot of Negative Binomial GLM

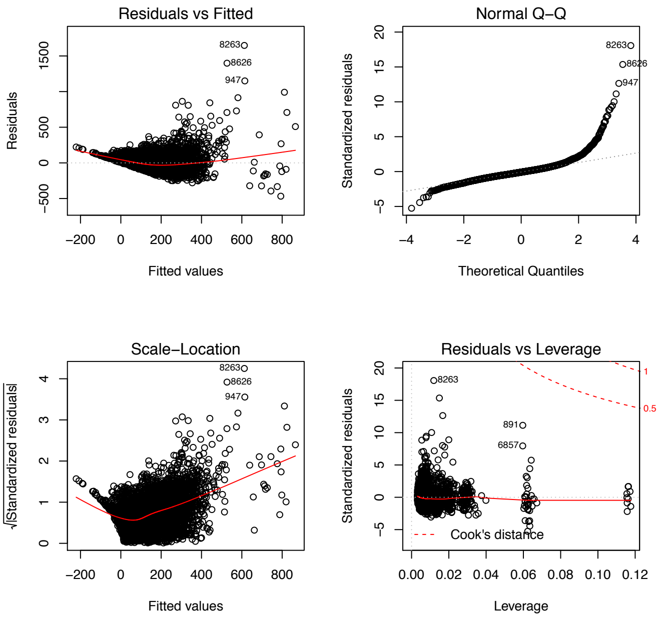

Figure 17: Diagnostic plots of Linear Model

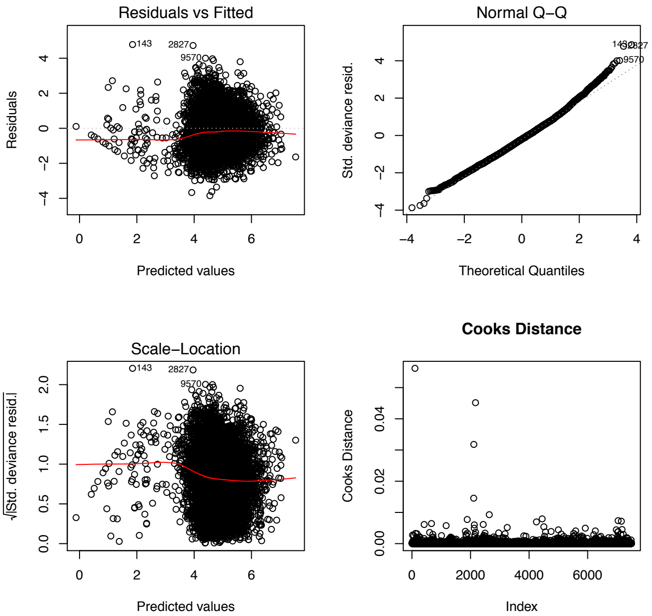

Figure 20: Diagnostic plots of final model

REFERENCE

[1] Ailawadi, K. L., Harlam, B. A., César, J., Trounce, D. (2006). Promotion Profitability for a Retailer. Journal of Marketing Research, 43(4), 518–535.

[2] Introduction: What is overdispersion? http://biometry.github.io/APES//LectureNotes/2016-JAGS/Overdispersion/OverdispersionJAGS.html.

[3] Malik, S. A., Fearne, A., & O Hanley, J. (2019). The use of disaggregated demand information to improve forecasts and stock allocation during sales promotions. International Journal of Value Chain Management, 10(4), 339-357.

The work is equally contributed by all memebers of the group.