Report

abc

Purpose: -Understand the role of price, promotion, brand, product and store characteristics and time of the year in product sales; -Estimate the sales of each of the 2,500 records where you don't have this information.

Section I: Describe briefly what aspects of the problem context you considered at the outset, how you used these to start your exploratory analysis, and what were the important points to emerge from this exploratory analysis.

Carry out an exploratory analysis that will help you to start building a sensible statistical model to understand what drives the numbers of sales for each product, in each store, at each time point . Identify an appropriate set of candidate variables to take into the subsequent modelling exercise, as well as to identify any important features of the data that may have some implications for the modelling. You will need to consider the context of the problem to guide your choice of exploratory analysis.

Section II: Describe briefly (without too many technical details) what models you considered in step (3) above, and why you chose the model that you did.

Using EDA as a starting point, develop a statistical model that enables you to predict the number of sales on (a subset of) the other variables in the dataset, and also to understand the variation in sales. Consider a range of models and use an appropriate suite of diagnostics to assess them. Ultimately however, you are required to recommend a single model that is suitable for interpretation, and to justify your recommendation. Your chosen model should be either a linear model, a generalized linear model or a generalized additive model.

Section III: State your final model clearly, summarise what your model tells you about the factors associated with product sales, and discuss any potential limitations of the model.

Use your chosen model to predict the number of sales of each product at each store and time point where this information is missing, and also to estimate the standard deviation of your prediction errors.

Process: Cluster variables — to identify naturally occurring groups within a data set, prior to building predictive models/model, and to hence improve accuracy of models/model Choose which model by reference to diagnostic plots Finalise chosen model Decide which covariates to remove Include interaction terms Any plots to use to show goodness of model ? Reference context throughout a model with a smaller dispersion parameter, it means that you're capturing more of the variability through your chosen covariates.

Other: Leverage points (?) only 3 - keep or leave and why? Mention about how some sold for more than less - anomalies More reason to give advantage and disadvantage Conclusion more for each ones Ensure code is consistent Ensure each graph saves Larger stores with more sales? Incorporate or

To Do: Consider transformation of units/SWB*NWK Use literature to pick which covariates Plots of clustering Discuss collinearity and serial correlation in each model Check cooks distance and examine the observations with high value Explain possible interactions we consider Interpret the coefficients in glm

Variable	Description
UPC	Unique Product Code of the Cereal
Manufacturer	Brand Of the Cereal (Kellog, Post Foods, Quaker, General Mi, Private Label)
Category	Product category

Variable	Description
Sub_Category	Subcategory of product (All family cereal, Kids cereal, Adult Cereal)
Store_Num	Unique Store number
City	Location of the Store
State	Location of the City
Area_Code	Region of the store
Store_type	Type of the store (VALUE, UPSCALE or MAINSTREAM).
Avg_Weekly_Baskets	average number of weekly baskets sold in the store.
ID	the record ID, from 1 to 10,000.
Month	Month product was sold
Year	Year product was sold
NWeeks	Number of weeks in the data whose first day falls in MONTH (maximum 5) Some entries have NWEEKS equal to a smaller number (eg 1 or 2) because some

products may have not been available at that store throughout the month. | Feature | Proportion of weeks in MONTH that product was promoted through marketing circular | Display | Proportion of weeks in MONTH that the product was on a special display in store. | TPR_Only | Proportion of weeks in MONTH that product was only on temporary price reduction | Price | the average price in that month. | Base_Price | the average "regular price" (i.e., without any promotions) in that month. | Units | The number of items of each product sold in any week whose first day of MONTH

Important features of data to note in eda : disproportionate sample sizes, some products sell for more than base price, large abnormally in date feb2009, units are directly proportional to store size, nweeks denote number of weeks product is available ,

START: We are interested in understanding how a variety of variables influence the monthly sales of cold cereal for a supermarket chain. A clear understanding of the factors associated with higher sales will allow the chain to efficiently allocate their constrained resources to the factors that maximise profits. We will hence identify the crucial variables in this EDA, so as to build a model that can accurately predict the number of cereal sales, to help prevent the overstocking or understocking of inventory. Helping to minimise costs or missed profits in an industry characterised by narrow profit margins.

The dataset, collated over a three year span, contains 20 variables and gives results for 7500 values of units sold monthly. There is no missing data in the dataset.

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 145.5401333 and standard deviation 133.5550642.

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

Firstly, sales can increase several fold in the presence of displays or other promotional methods (Ailawadi, Harlam, César, and Trounce, 2006). Thus understanding the influence of various methods of promotion styles is crucial for understanding sales.

It follows that the effectiveness of price-promotions is heavily influenced by socio-economic factors, where shoppers who are more price sensitive are more likely to respond to promotions than those who are not. Thus, the accuracy of our model may benefit from us identifying a way to segment consumers into homogenous groups that reflect this.

Additionally, geographical location impacts sales by way of either weather or local competitor retailers. As the data provided to us is a subset of the whole data, the information of competitors is not available to use and so this point is left out.

Although cereal does not appear to be a seasonal product, weather can affect the frequency by which consumers visit retail stores, thus the months may benefit from clustering. Additionally, unexpected drivers of retail sales, such as abnormal events like a financial crisis or natural disasters, manifest themselves as random disturbances to the time series data and are correlated across categories and stores that share a sensitivity to those events.

Before starting the EDA, we categorised the potential predictors as one of two kinds; Endogenous factors, i.e within stores control: UPC, manufacturer, sub category, base price, price, promotion methods (tpr, feature, display), nweeks (availability of the product). Or exogenous factors, not within stores' control: store_num, city, state, area code, store type, avg weekly baskets, time (month, year). This will guide our EDA by allowing us to consider the variables in each category sequentially.

From our analysis of the endogenous variables, we conclude that category is not a meaningful predictor as all our observations belong to one category - cold cereal. Additionally, a new potential predictor was introduced , discount, which is the percentage discrepancy between the base price and selling price. Figure 10 of the discount reveals a more linear relationship with units than Figure 8 and 9 of price and base price respectively.

However we must note that there are potential anomalies revealed by Figure 10, where some products were sold for higher than the base price, relating to a negative discount which is not practically plausible.

Also any conclusions drawn from variables with uneven sample groups should be taken cautiously, such as in the comparison of promotional methods where no products were featured for 5 weeks whereas some were displayed or on temporary price reduction for 5 weeks. Also, Figure 4 describes how the availability of a product limits sales, wherein any week valued less than 4 reflects a lack of availability. This highlights an important issue as it hinders our investigation by not accurately reflecting consumer demand and hence the true impact of certain variables on the sales of cereal.

Thus, Figures 1-10 suggest that all the variables influence monthly sales and should be considered as potential predictors in our model.

The following categorical covariate analysis makes use of boxplots to describe the variability of monthly sales for each factor of the variable in question. Any further reference to "variability" indicates a reference to the spread of units of monthly sales of the cold cereal for the covariate in question.

Figure 1 describes the variability for each unique product. Different shoppers may have favoured products that they are likely to repurchase repeatedly. Particularly, one of the products has a significantly larger spread of sales when compared to the others.

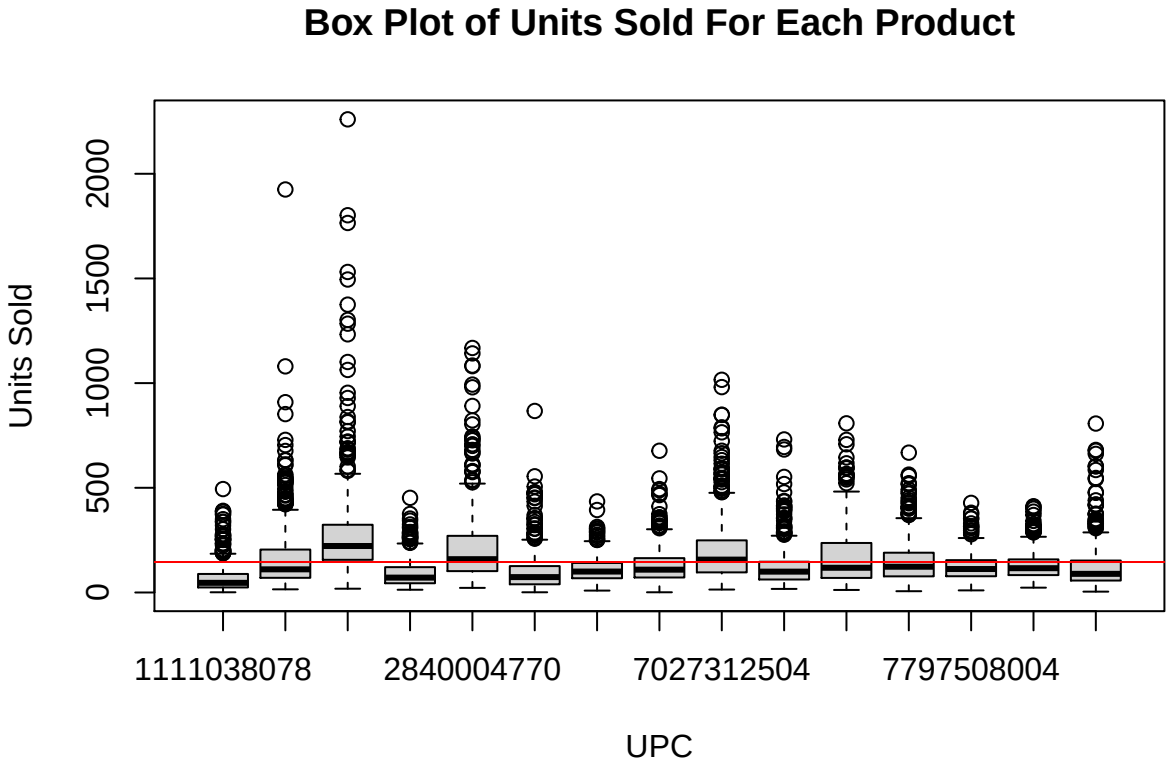


Figure 1: Boxplots showing the number of units of cereal sold for each unique product. The red line indicates the mean units sold across all products.

Figure 2 describes variability for each type of cereal. Note that the sample sizes are uneven, however the figure indicates that we can weakly assume family cereal is more popular as compared to other cereal types.

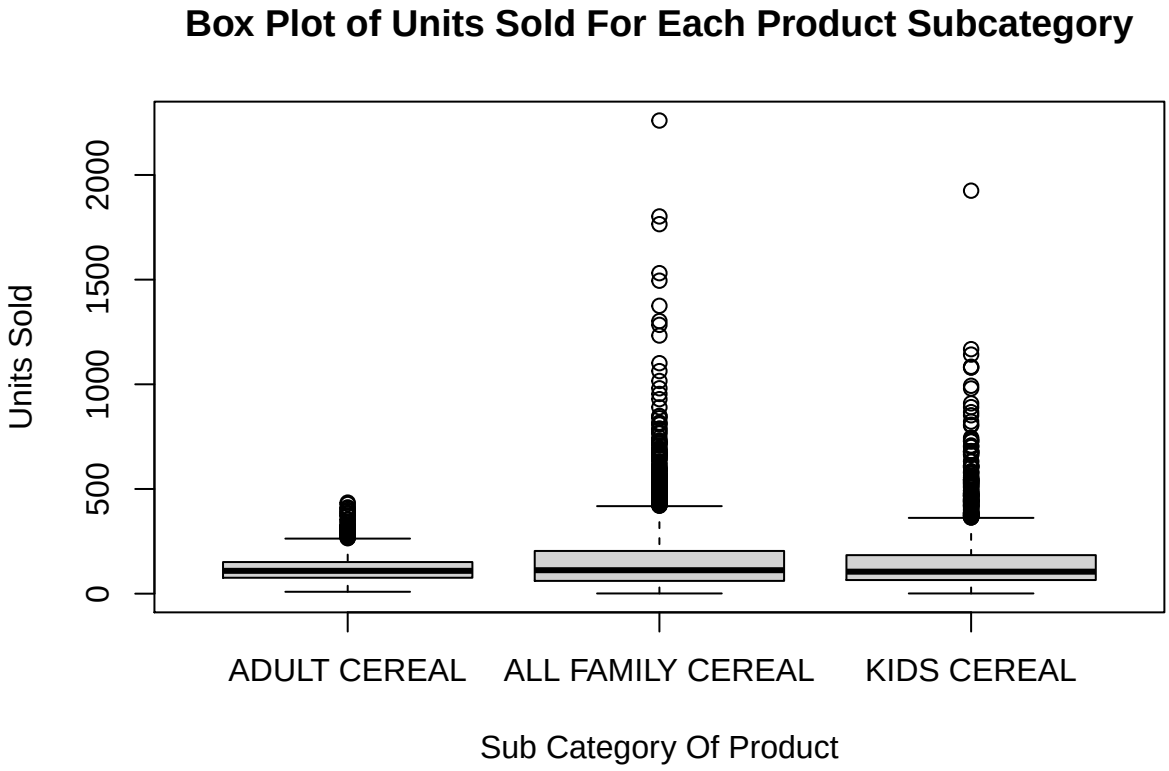


Figure 2: Boxplots showing the number of units of cereal sold for each sub category of the product.

Figure 3 describes variability for manufacturers. Different manufacturers may have differing brand images, whereby more established or "popular" brands may entice a larger consumer base to purchase their products.

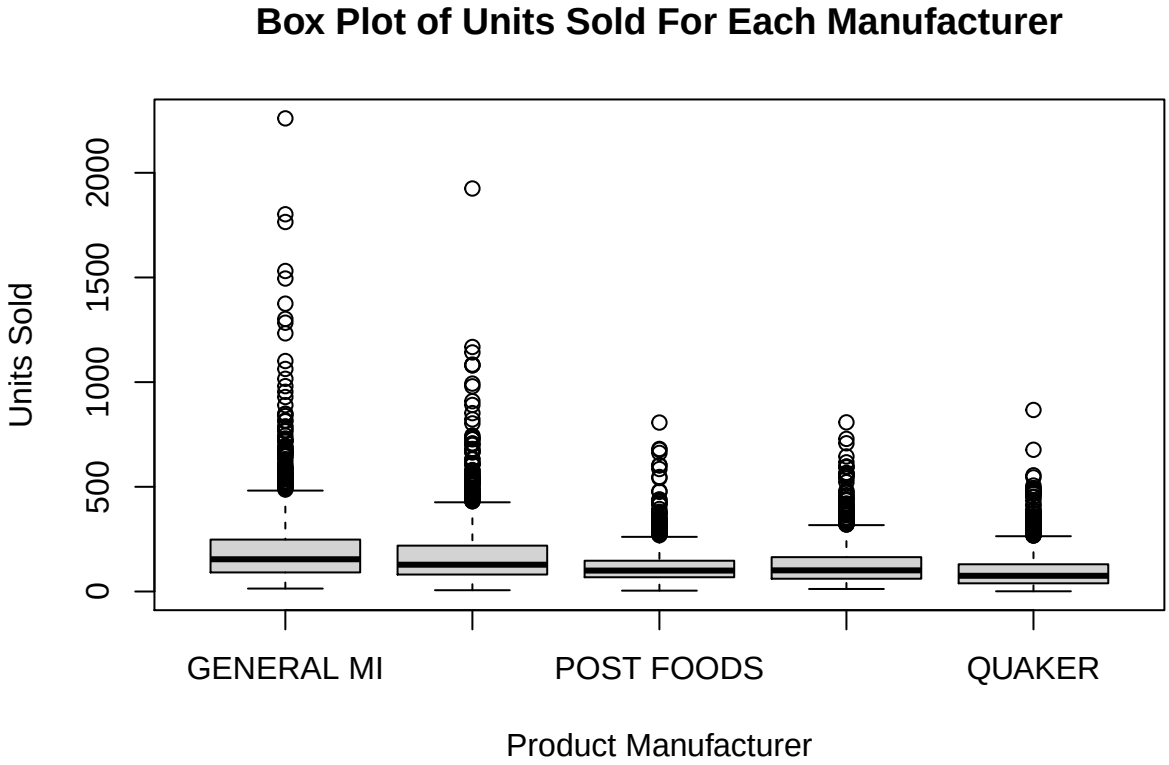


Figure 3: Boxplots showing the number of units of cereal sold for each manufacturer.

Figure 4 describes variability between each store. Different stores are selling various amounts of cereal. There may be confounding variables that help to explain why. The boxplot suggests we should consider clustering the groups in a more meaningful way.

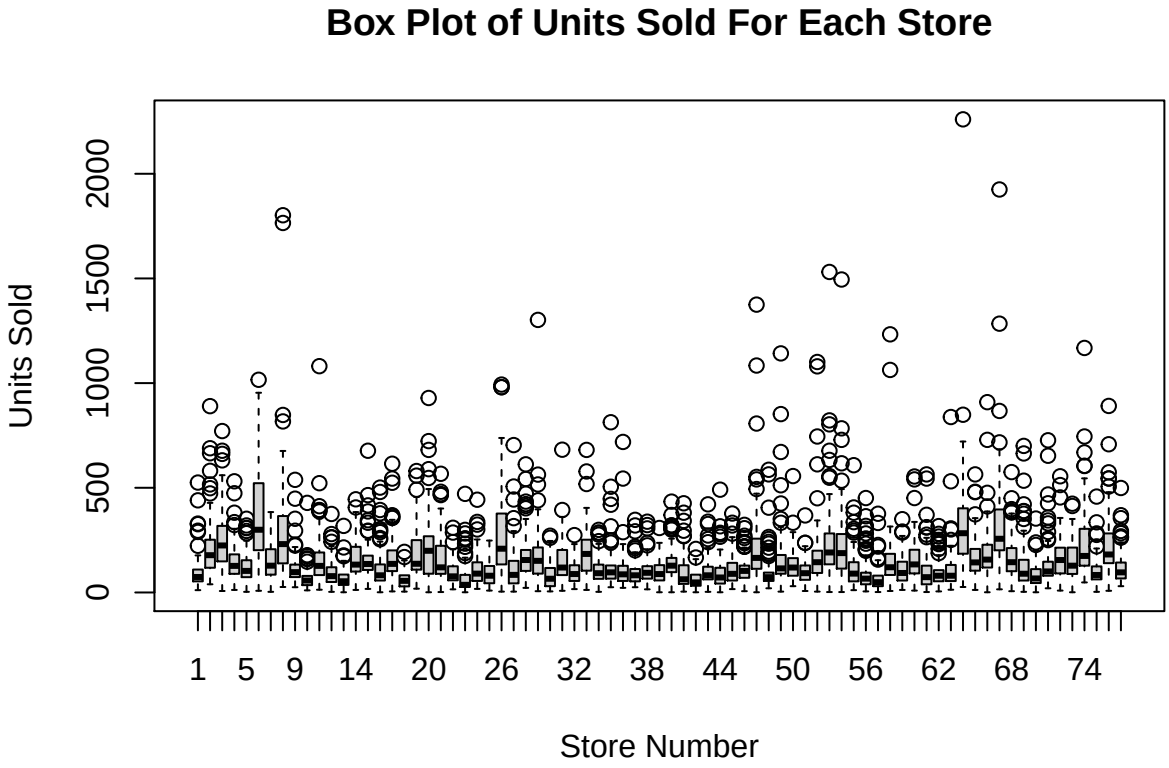


Figure 4: Boxplots showing the number of units of cereal sold for each unique store.

The type of store can be interpreted as a categorisation of the shopper, loosely grouping different shopper profiles to store types allow us to investigate which factors motivate certain shoppers to purchase cereal. Note that there is an imbalance in the data set available, however we can somewhat observe from Figure 5 that mainstream stores sell more cereal on average.

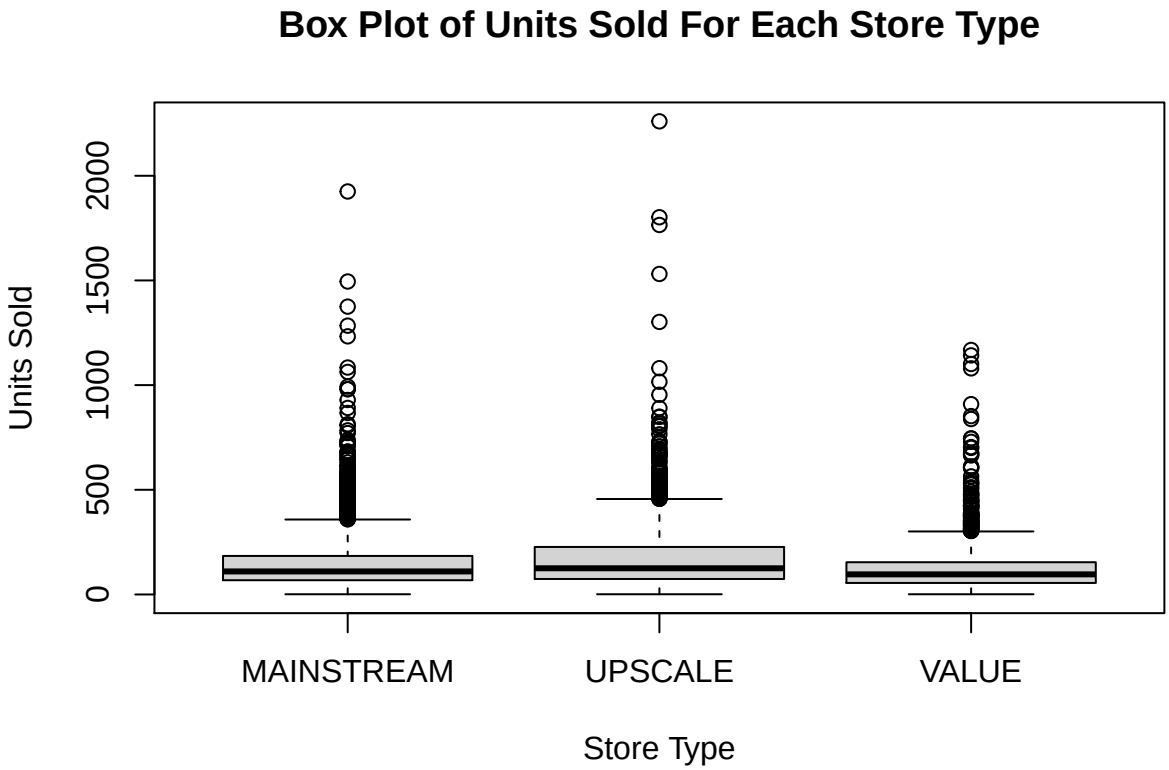


Figure 5: Boxplots showing the number of units of cereal sold for each type of store.

By grouping the stores according to their store type and investigating the spread of monthly sales for each store, Figure 6 shows us firstly that there is an imbalance in the number of stores in each category and secondly that within each store type, there are varying degrees of spread observed in each store. This is indicative of other confounding variables influencing monthly sales more prominently.

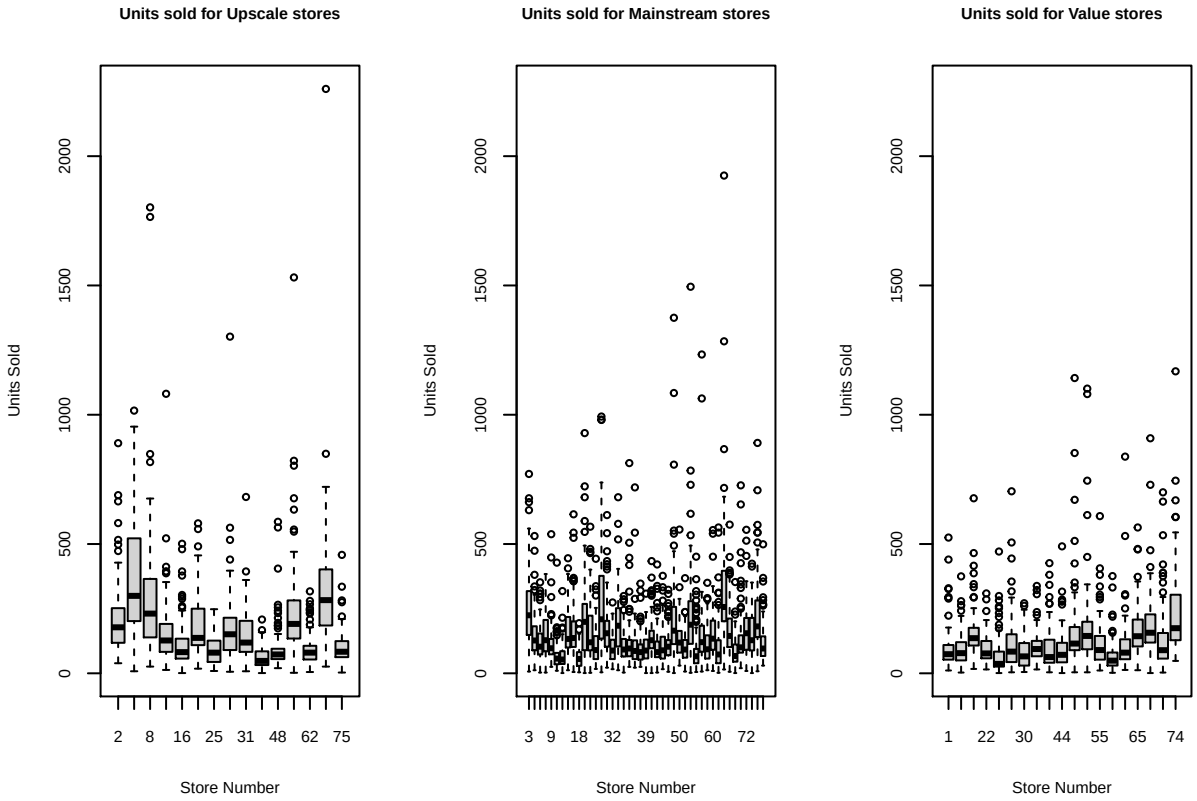
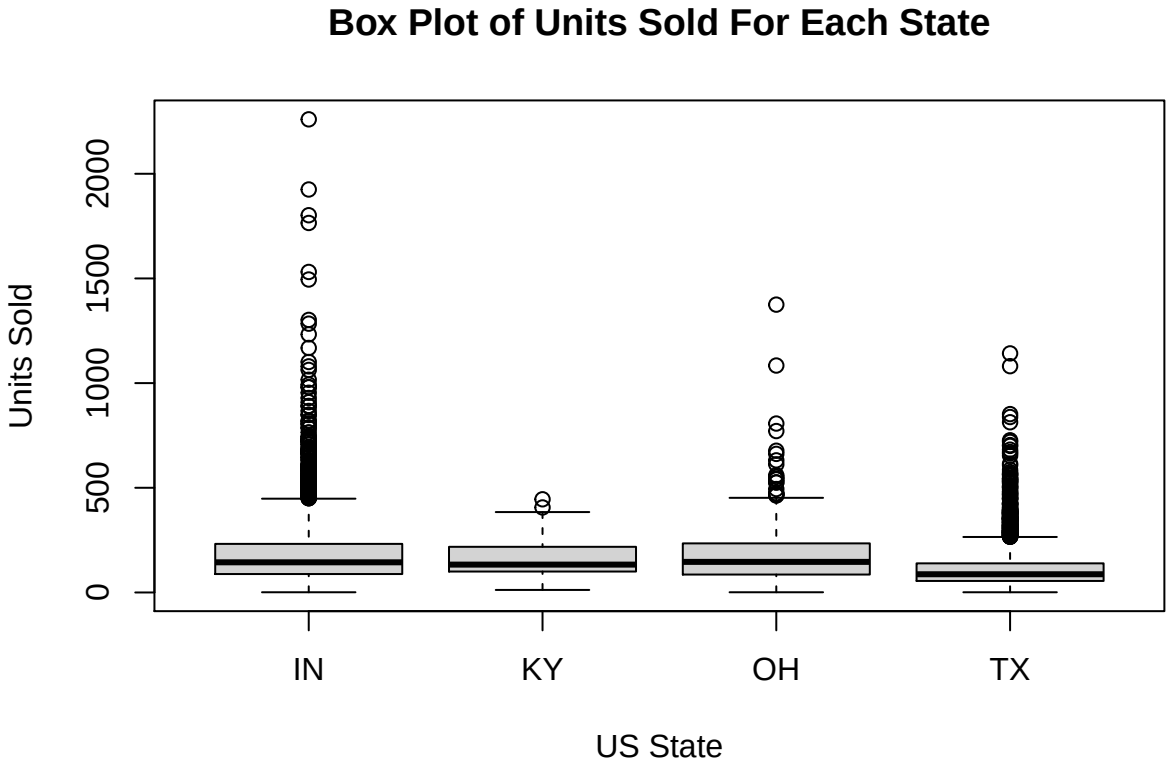
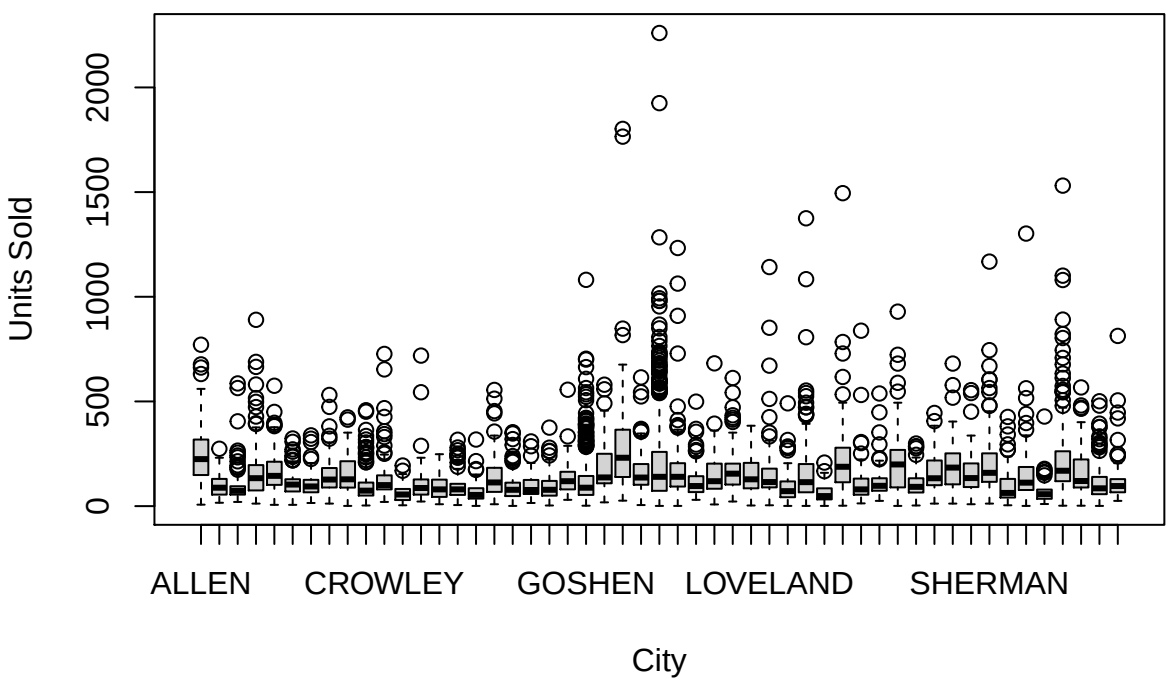


Figure 6: Boxplots showing the number of units of cereal sold for each store in each store type.

Another key factor appears to be the geography, since the state, city and area code a consumer lives in may relate to their economic status. Economic status is an important variable as it directly relates to a shopper's disposable income, and in turn their sensitivity to promotional methods. Note that there is a deficit in the available data from stores in Kentucky, however Figure 7 suggests Indiana has the largest variability in sales. The city box plot suggests that this area of investigation may benefit from a different grouping method. Although the area code box plot suggests that the area code may impact the monthly units, it is intuitive that area code does not reveal any pertinent information regarding our investigation.


Box Plot of Units Sold For Each State
Units Sold
0
500
1000
1500
2000
IN
KY
OH
TX
US State


Box Plot of Units Sold For Each City
Units Sold
0
500
1000
1500
2000
ALLEN
CROWLEY
GOSHEN
LOVELAND
SHERMAN
City

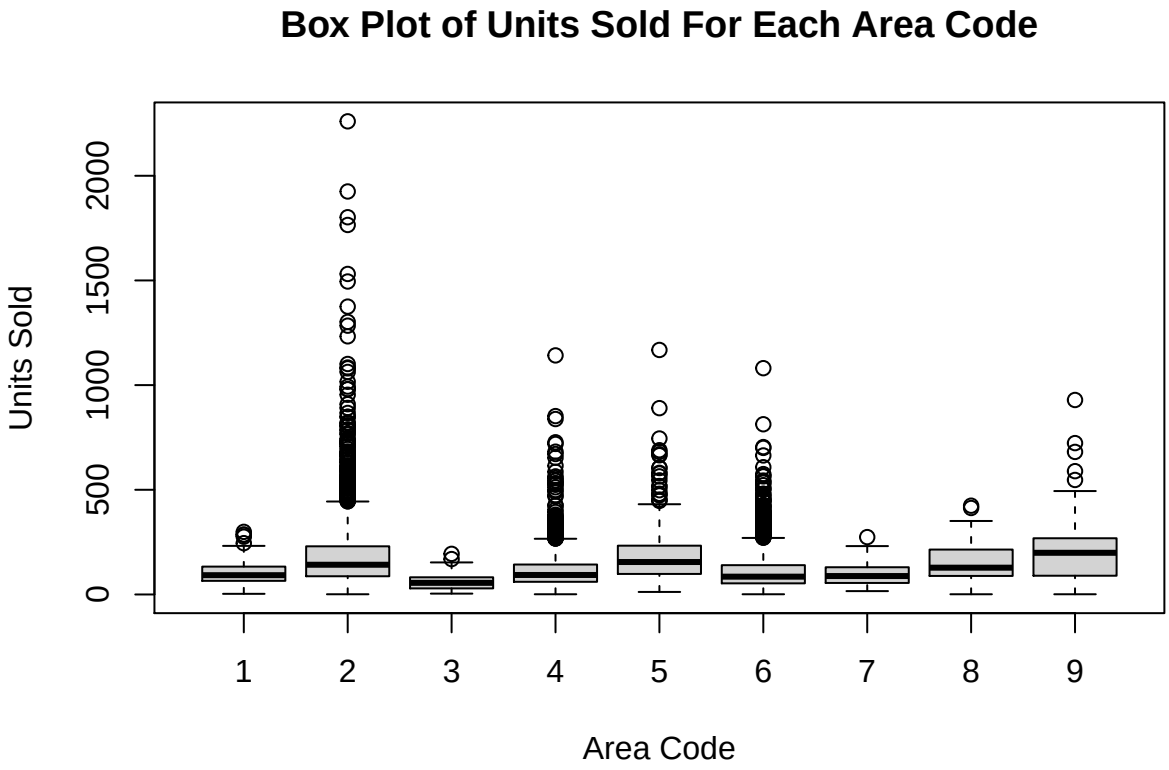


Figure 7: Boxplots showing the number of units of cereal sold for each US state, each city and each area code respectively.

*example of abnormal events We plotted the sales for each month, across the years of data we have, to help identify any trends which are consistent across those years. This would indicate the presence of seasonal factors such as weather impacting consumer demand rather than popularity simply increasing over the years. Although cereal does not appear to be a seasonal product, weather can affect the frequency by which consumers visit retail stores, thus the months may benefit from clustering. Additionally, unexpected drivers of retail sales, such as abnormal events like a financial crisis or natural disasters, manifest themselves as random disturbances to the time series data and are correlated across categories and stores that share a sensitivity to those events. This may explain the isolated period with a large volatility in sales in February 2009 shown in Figure 8.

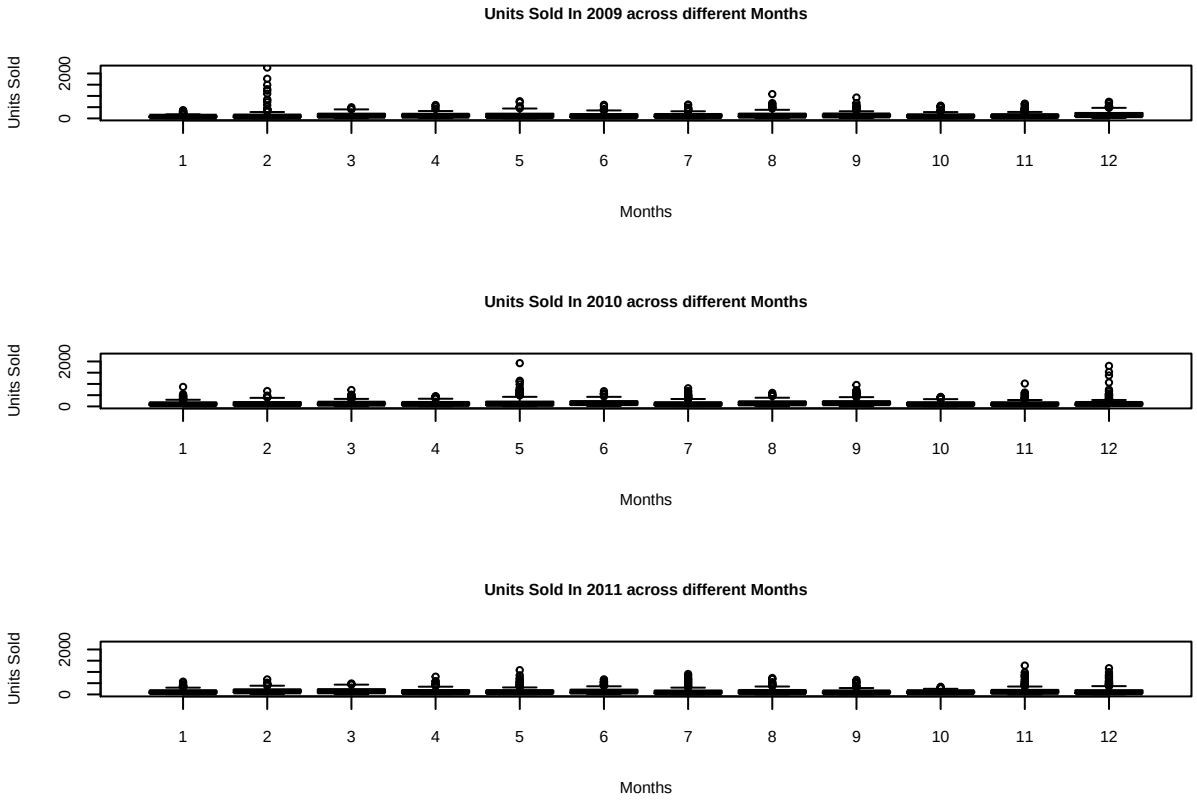


Figure 8: Boxplots showing the number of units of cereal sold in each year across the months.

The number of weeks indicate the availability of the products in question, meaning a lack of availability, i.e any week valued less than 4, would limit the potential sales of the product. This poses an issue as it hinders our investigation by not accurately reflecting consumer demand and hence the true impact of certain variables on the sales of cereal.

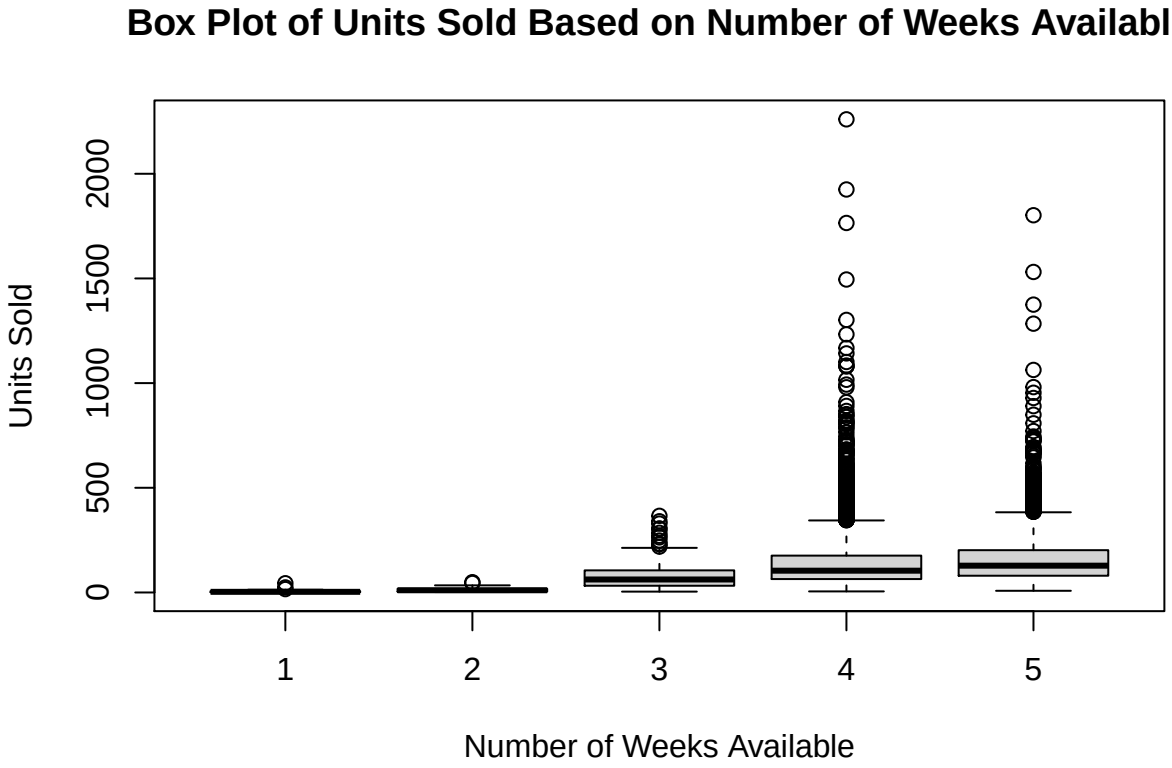


Figure 9: Boxplots showing the number of units of cereal sold for the number of weeks of the month the cereal was available.

A general conclusion is that sales can increase several fold in the presence of displays or other promotional methods (Ailawadi, Harlam, César, and Trounce, 2006). Understanding the influence of various methods of promotion styles only serve to better inform us on its relationship with sales. Any conclusions drawn should consider that no products were featured for 5 weeks whereas some were displayed or on temporary price reduction for 5 weeks. It seems figures 10 reveal a positive relationship between monthly sales and weeks using each promotional method.

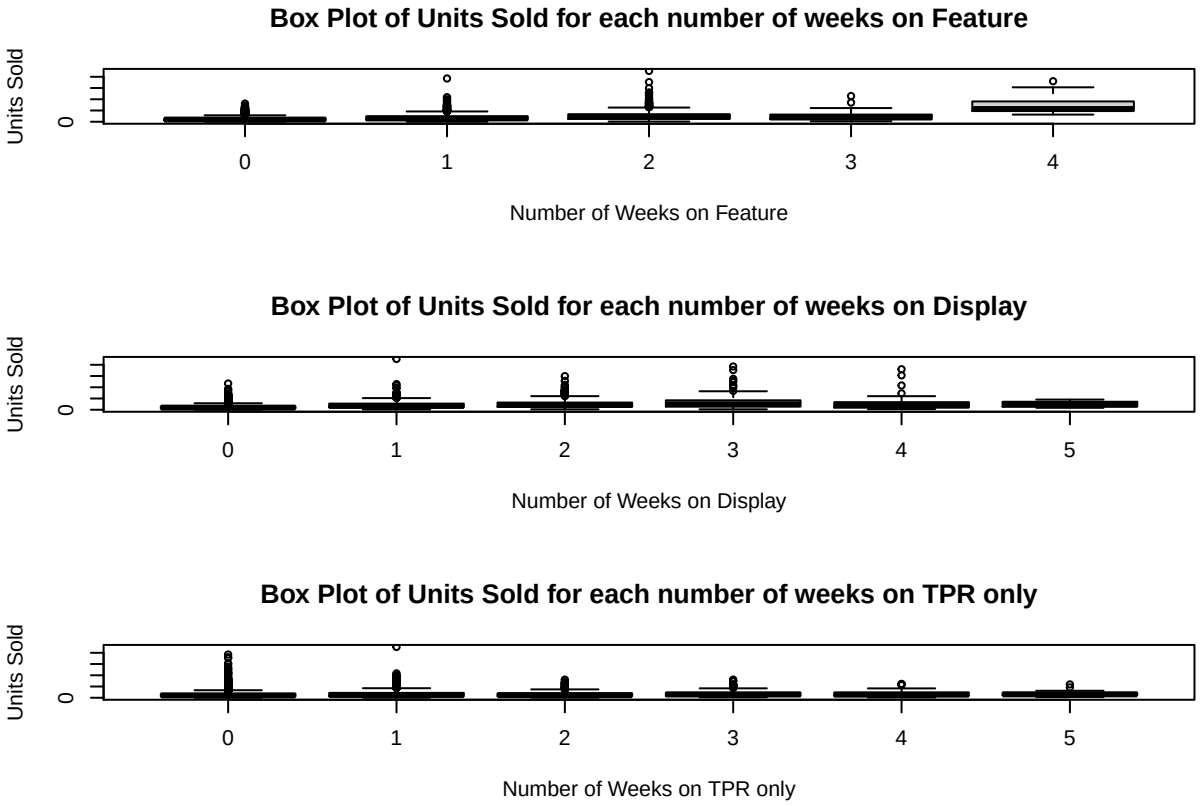


Figure 10: Boxplots showing the number of units of cereal sold based on the number of weeks using each promotional method.)

Idk if remove lol maybe ya #### Feature week distribution for diff types of store. differing consumer groups # may respond differently to promotion methods. the graph suggests further investigation par(mfrow=c(1,3))

boxplot(training$UNITS[training$STORE_TYPE ==“UPSCALE”] ~ featureweek[training$STORE_TYPE ==“UPSCALE”] , main="“, xlab =”“, ylab="", ylim=range(training$UNITS), cex.lab=0.8, cex.axis=0.8, cex.main=0.8)

boxplot(training$UNITS[training$STORE_TYPE ==“MAINSTREAM”] ~ featureweek[training$STORE_TYPE ==“MAINSTREAM”] , main="“, xlab =”“, ylab="", ylim=range(training$UNITS), cex.lab=0.8, cex.axis=0.8, cex.main=0.8)

boxplot(training$UNITS[training$STORE_TYPE ==“VALUE”] ~ featureweek[training$STORE_TYPE ==“VALUE”] , main="“, xlab =”“, ylab="", ylim=range(training$UNITS), cex.lab=0.8, cex.axis=0.8, cex.main=0.8)

Moving on to the numerical analysis, the first two plots describe the relationship between the price and base-price respectively with the monthly sales of cereal. These two plots do not reveal much. However, as seen in the third plot, plotting the percentage reduction reveals a potential relationship between the two.

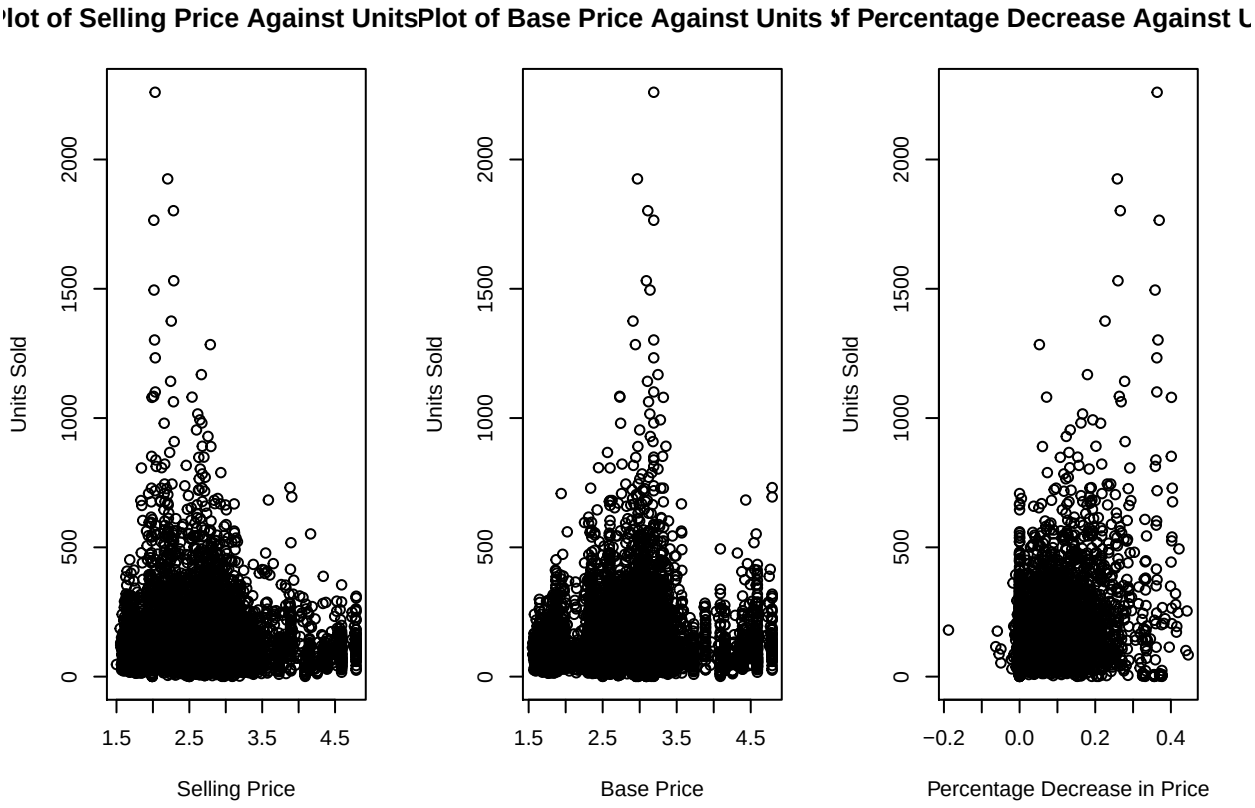


Figure 11: Plots showing the relationship between the number of units of cereal sold and the selling price, original price and discount respectively.

The plot for average weekly baskets, Figure 12, reveals a positive relationship between the size of the store and the number of units sold. This is intuitive as stores who accommodate larger amounts of shoppers are more likely to sell more units of a product.

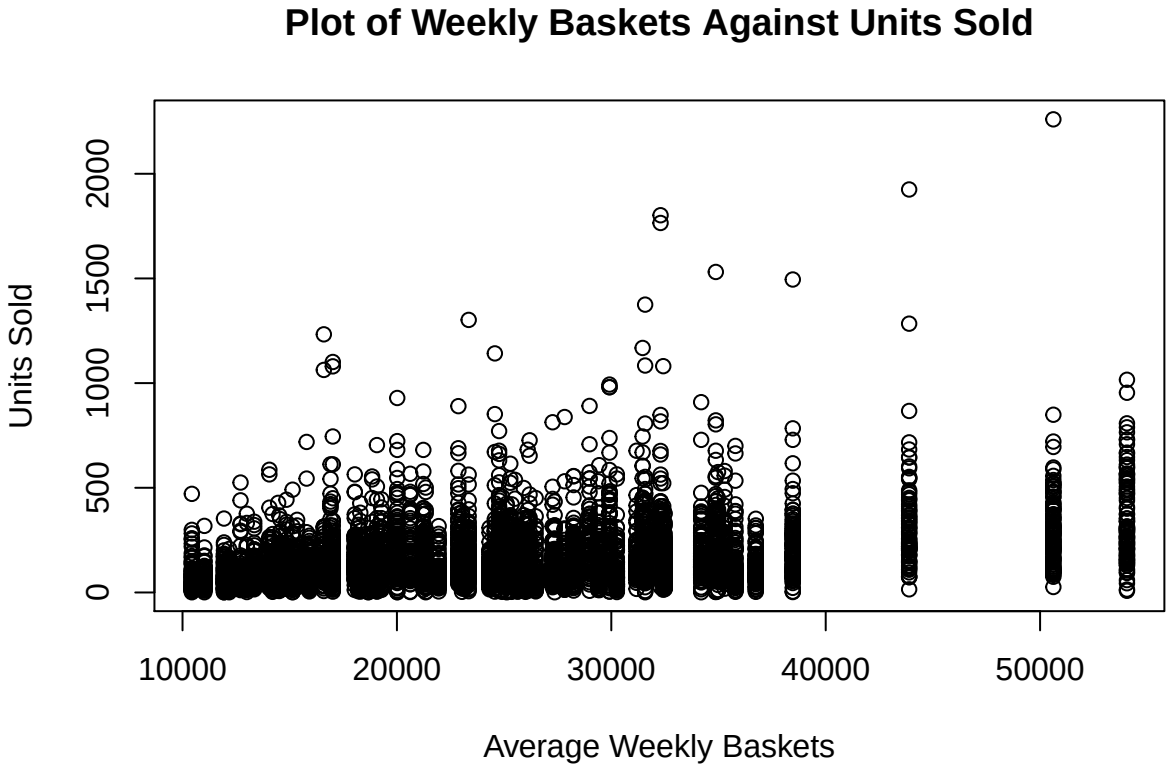


Figure 12: Plots showing the relationship between the number of units of cereal sold and the average weekly baskets.

The graphs for ID and CATEGORY were omitted as they do not offer any substantial information to the investigation. All other variables discussed are potentially viable predictors for our model and should be considered.

MODEL BUILDING

Process: Cluster variables — to identify naturally occurring groups within a data set, prior to building predictive models/model, and to hence improve accuracy of models/model Choose which model by reference to diagnostic plots Finalise chosen model Decide which covariates to remove Include interaction terms Any plots to use to show goodness of model ? Reference context throughout

Other: Leverage points (?) only 3 - keep or leave and why? Mention about how some sold for more than less - anomalies More reason to give advantage and disadvantage Conclusion more for each ones Ensure code is consistent Ensure each graph saves Larger stores with more sales? Incorporate or

If we dont cluster we cant even make the model as it is too many

Before building the model, we will start by constructing new covariates from the given covariates to provide us with more information and maximise predicting power.

There are a multitude of factors influencing characteristics of observed sales data and underlying demand. Factors within the control of retailers (such as promotional pricing) are integral to understand in order to implement effectively. One such factor would be the discount covariate, which is computed as the percentage decrease of PRICE from BASE_PRICE. The values are rounded to 3 decimal points. We believe that the percentage reduction in price directly influences a shoppers decision to make a purchase. This is corroborated by a plot between discount and monthly sales. (Note that there are some anomalous values with negative discounts, likely due to human error in the price labelling.)

Weather was identified to be a universally critical exogenous factor directly impacting consumer demand (ref). In winter people are logically less likely to be inclined to go out than during the summer, so the amount

bought will likely change. We believed the best way to incorporate weather into our model was by grouping the MONTH covariate into SEASON; whereby for instance March, April, and May will be SPRING. Another benefit from this clustering comes from reducing the number of dummy variables from 12 factors to 4.

Likewise, another categorical variable which would benefit from regrouping is CITY. In order to reduce the number of variables and group them in a purposeful way, we implemented hierarchical clustering to CITY. Shoppers can be broadly categorised as price sensitive or up-market, with price sensitive shoppers being more likely to respond to promotions than up-market shoppers. Thus,to reflect this notion, we suggest that the geographical location of a shopper may attest to their socioeconomic status, and hence their sensitivity to price. This led to the utilisation of STORE_TYPE and AVG_WEEKLY_BASKETS covariates as numerical covariates in the clustering of CITY. Although STORE_TYPE is categorical, we remedy this by introducing new variables of counts of each store type in each city. Where two cities with similar number of stores and a similar distribution of store types are grouped together, aligning with our intuition. The AVG_WEEKLY_BASKETS of all the stores in the same city were accumulated, to reflect the aggregate transactions of a city. A portion of the adjusted data set for clustering CITY is shown below. After clustering, CITY was reduced from 51 groups into 4 groups. It is also worth noting that this clustering provides us with relatively even sample groups, while also alleviating the impracticality of a variable with a large number of variables.

To have a more direct idea of each of the four clustered groups of `CITY`, we present below an averaged value of `MAINSTREAM` count, `UPSCALE` count, `VALUE` count, and `AVG_WEEKLY_BASKETS` for each of the groups. Though not immediately drastic, one can still see the groups are varied in compared values.

```
##           CITY MAINSTREAM UPSCALE VALUE   BASKETS
## 1   MAINEVILLE          1       0     1  44284.67
## 2     BEAUMONT          1       1     0  37423.26
## 3        ALLEN          1       0     0  24766.81
## 4        CLUTE          1       0     0  29386.42
## 5       GOSHEN          3       1     4 224979.79
## 6 INDEPENDENCE          4       2     3 274961.46
```

```
##         MAINSTREAM UPSCALE VALUE  BASKETS
## Group 1        1.1     0.0   0.0  24284.1
## Group 2        0.4     1.0   0.1  35521.1
## Group 3        0.2     0.0   1.1  28224.6
## Group 4        3.5     1.5   3.5 249970.6
```

Different promotional methods have different modalities and it is therefore natural to assume that the interplay of the strategies may lead to differing promotional efficacies. Thus, in order to introduce this into our model without the use of a three way interaction term, we decided to cluster the stores by their predominant promotional strategies, i.e either `TPR_ONLY`, `DISPLAY FEATURE` or an even mix of all three. These three covariates were provided as proportions of the months they were active, thus we multiplied them by `NWEEKS` to adjust them back to monthly data. Thus, we incorporated these new variables into our data set to use as the numerical variables to cluster `STORE_NUM` with respect to.

A portion of the adjusted data set for clustering `STORE_NUM` is shown below.

```
##   STORE_NUM TPRmonth DISPLAYmonth FEATUREmonth
## 1         1        0            0           1
## 2         1        1            0           1
## 3         1        0            0           0
## 4         1        3            0           0
## 5         1        0            0           0
## 6         1        2            0           0
```

To have a more direct idea of each of the four clustered groups of `STORE_NUM`, we present below an averaged value of `TPRmonth`, `DISPLAYmonth`, and `FEATUREmonth` for each of the groups. Though not immediately drastic, one can still see the groups are varied in compared values.

```
##          TPRmonth DISPLAYmonth FEATUREmonth
## Group 1    0.661        0.353        0.391
## Group 2    0.403        0.411        0.458
## Group 3    0.380        0.337        0.366
## Group 4    0.408        0.531        0.479
```

We omitted variables such as `STORE_NUM`, `MONTH` & `CITY` as they had been clustered in a manner that allowed for more purposeful interpretation.

The reason for omitting variables such as UPC is because the information they provide is better encapsulated by the other covariates in the model such as manufacturer and sub category. We also decided to omit variables that did not seem to be causally linked to the units sold such as the ID.

After adjusting our set of covariates, we will now proceed to build the model. The first step to building a model is to decide the type of model we will build.

The most standard model is the linear model, we will not be using this. 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, 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 accuracy. Thus, as a GLM provides a good balance between the quality of the model and its interpretability, we opted to proceed with the GLM.

An intuitive starting family to try would be that of poisson since the response variable UNITS is in counts and the poisson distribution, like the data we have, is positively skewed, thus allowing for a more accurate modelling of residuals. This is demonstrated in the following plot.

Thus, we built a GLM with Poisson using again the predetermined set of covariates. The summary is included below.

All the covariates of the model have 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. These extremely small p-values may serve as a warning against the use of the Poisson family. The residual deviance is 231935 with degree of freedom 7447. Since the residual deviance of 231935 is significantly larger than the degrees of freedom of 7447, there is indication of overdispersion in the model. We computed the deviance of the model, shown below, and it is notably far from one.

(REFERENCE: http://biometry.github.io/APES//LectureNotes/2016-JAGS/Overdispersion/OverdispersionJAGS. html) Overdispersion can be alleviated by changing the family from Poisson to QuasiPoisson, which is done below.

Incorporates the deviance into the model - how does this improve overdispersion and how can we show this through the diagnostic.

The QuasiPoisson model incorporates the deviance into the model in order to improve the overdispersion issue, however, it also makes the model harder to use for prediction and comparisons (with other models). It doesn't have aic , which is a problem for us because ....

Thus we consider an alternative way to deal with overdispersion, which is to use a Negative Binomial Family. The summary is again shown below. Negative Binomial GLM can be viewed as a somewhat upgrade of Poisson GLM. (REFERENCE: http://biometry.github.io/APES/ /LectureNotes/2016-

JAGS/Overdispersion/OverdispersionJAGS.html) but we will use some method to discern the superior of the two within the context of our investigation.

Refer to summary and compute it from the deviance and dof The ratio of residual deviance and degrees of freedom, as shown in the summary above, is close to 1 which is consequent as it improves the quality of the Q-Q plot too. The plot on the left below is of Poisson GLM, while the one on the right is of Negative Binomial GLM. It is rather obvious to see that the one on the right is closer to the ideal diagonal line.

Now that we have our model, we will start to remove some covariates that we think are less helpful. The algorithm of covariates removal goes like this: (1) we look at the summary of the model and pick out one covariate with big p-value (2) we build a model without it (3) run an ANOVA with the nested models, and we remove the covariate if the p-value of the ANOVA is big and not remove otherwise. First, we try to remove YEAR

Explain why each model improvement we choose is also good in relation to the context of retail. Remove covariates now with model we have chosen

How did we choose interaction , tried a bunch and see which one was significant without too many factors .- principle of parsimony

Once we had our final model, we calculated the cooks distance for each data point to search for peculiar cases, we could see from the plotting the values that a few points were exceptionally impactful on the model, but after looking into those specific cases there weren't any clear reasons for removing the points from the training data, despite the marginal improvement in AIC (-42) in the resulting model, it did not seem sufficient to warrant removal.

When looking for potential issues of multicollinearity in the model, we analysed the correlation matrix of the numeric covariates and found that the highest correlations were found amongst the different promotional styles. Since they are each sufficiently different from each other, we would not be justified in removing them on grounds of collinearity.

Exploratory analysis: 10 marks. These marks are for (a) tackling the problem in a sensible way that is justified by the context (b) carrying out analyses that are designed to inform the subsequent modelling. Awareness of context: 5 marks.

Model-building: 10 marks. The marks are for (a) starting in a sensible place that is justified from the exploratory analysis (b) appropriate use of model output and diagnostics to identify potential areas for improvement (c) awareness of different modelling options and their advantages and disadvantages (d) consideration of the retail and marketing context during the model-building process.

Quality of argument: 5 marks. The marks are for assembling a coherent 'narrative,' for example by drawing together the results of the exploratory analysis so as to provide a clear starting point for model development, presenting the model-building exercise in a structured and systematic way and, at each stage, linking the development to what has gone before.

What we conclude from our model

Overall, the p-values we obtain for our models are more than reasonable. Over 54 different covariates, only 6 of them 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 logical as our estimates are generally of the same order.

Regarding the interpretation, we could separate our covariates in 2 different categories : endogenous and exogenous.

· Endogenous factors, which are those relating directly to the product, such as marketing process with covariates 'display' and 'tpr', the discount applied, the manufacturer, or the number of weeks on display, at the exception of some of them, have a p-value inferior to 10^-3. Aongst those endogeneous factor, wa also to point out that there are two covariates of which deserve special mention : the discount and the number of weeks. The first, which has a coefficient of 1.977 is the one with highest positive estimation and also the one with the highest magnitude. Interpreting numerically, this means that a 1% increase in discount accounts for

a ~ 2% (exp(0.0197)) increase in the number of units sold. With respect to the second, which differs as it is categorical, it consists of four factors with magnitude of estimate at least 1 (in absolute values). There are in total only seven such covariates in our model, and four of them are accounted for by NWEEK. So, the longer a product is available, the more it will be bought by consumers. As for instance the expected difference in the number of units sold between a display of 1 week and a display of 4 weeks is approximately 23 (exp(3.145)).

· Now, transitioning to the exogenous factors, such as the type of store, the month of the year, or the city, we may conclude that the time influences the consumer habits, especially during January, February, and August, where people appear more inclined to buy products than during the other months.

""" Also, the socio-economic status of the consumer seems to reflect on the model. The households with lower income seem to consume more, and surely be more price-sensitive than the others, as the cities where we expect a bigger concentration of mainstream stores have a stronger influence on the units sold."""""" NOO

.

As a consequence, before all the company may first think about the availability of the product during the month, keeping in mind that a discount, no matter the product, is more effective than any processes applied. By evidence, we can't exclude the impacts of the other covariates as a discount will have a greater effect on a specific product if it is displayed. To increase the efficiency of those campaigns, we could also focus on the fact there are periods where it is more appropriate to conduct them such as during the winter or at the beginning of the school year when mechanically the sales increase.

Being the objective, a focus could be put on lower-income households in order to increase the sales, as they have a bigger propension to consume. """"And so larger marketing campaigns could be conducted in cities where the average income per household is lower."""""