Data Analytics I Group Project

Word count: 1,987

Eugene Ong

Anna Schyberg

Gaspard Rosset

Edoardo Lorenzetti

Zeynep Andsoy

Zheng Luo

Table of contents

Introduction

This report will analyze part of a grocery retailer's sales data in order to provide insight about their performance, as well as how the type of promotion affects the sale of products.

The data, gathered from a sample of stores over 156 weeks, contains sales and promotion information on the top five products from each of the top three brands within four selected categories (mouthwash, pretzels, frozen pizza, and boxed cereal). (CITE)

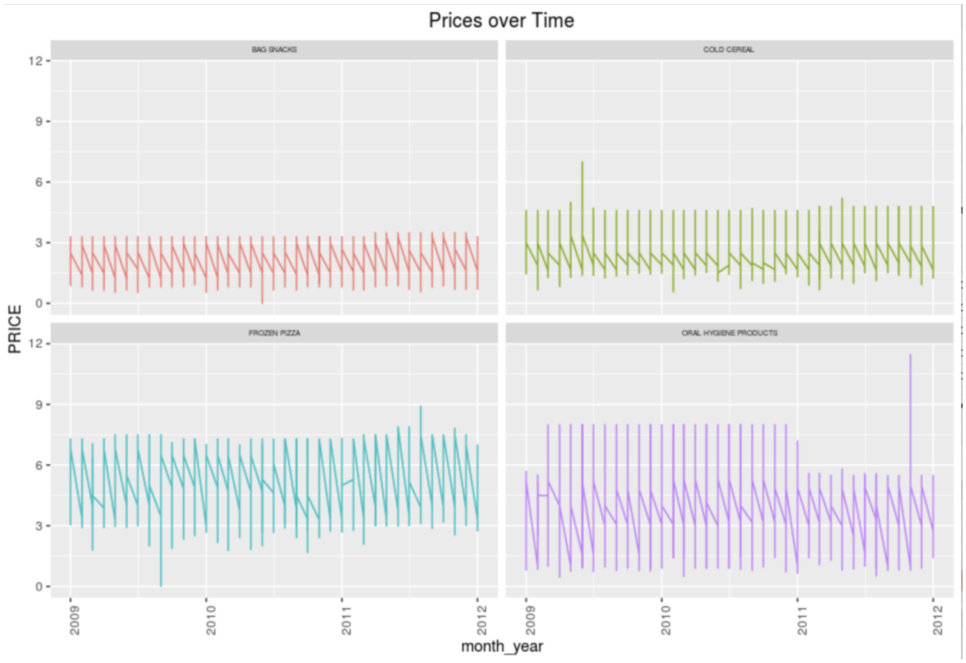
We will start by visualizing the data in order to make it more understandable and accessible; we then make an analysis that touches on price elasticity, seasonal variation, as well as the impact of advertising. As the aim of this project is also to improve future sales performance, we will test three different prediction models in order to estimate how future sales might look like.

Throughout the report, we will see that analyzing data is very useful when taking decisions. By using data collected from previous stores (sales and promotion information), we can understand the impact of different strategies (advertising methods, price changes) on different product categories, and how to use them accordingly in order to boost future sales.

1. Data Visualization

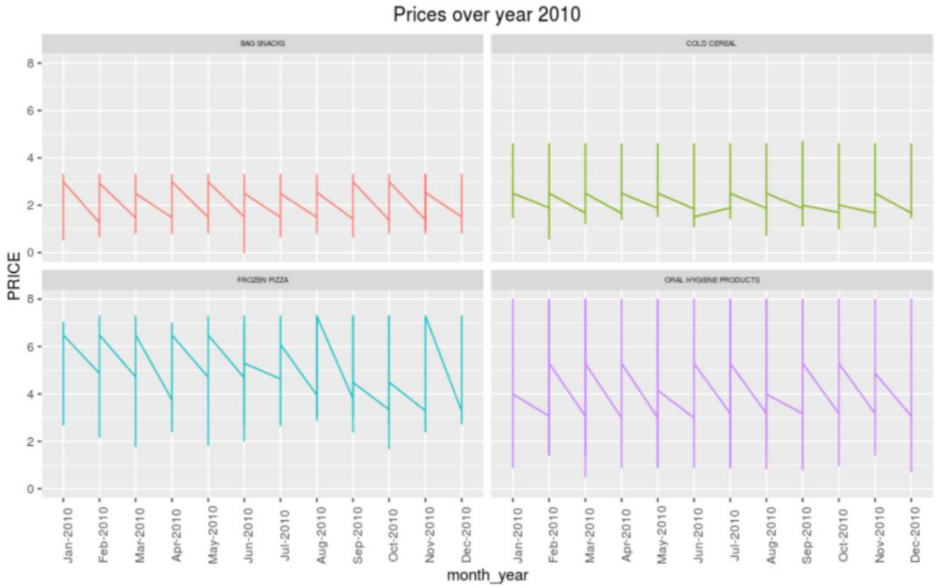
Seasonal changes

We can make use of time-series plots to observe price changes over-time and clearly communicate insight about why certain categories exhibit more seasonal changes in prices - a great approach to visualize the evolution of variables through time.



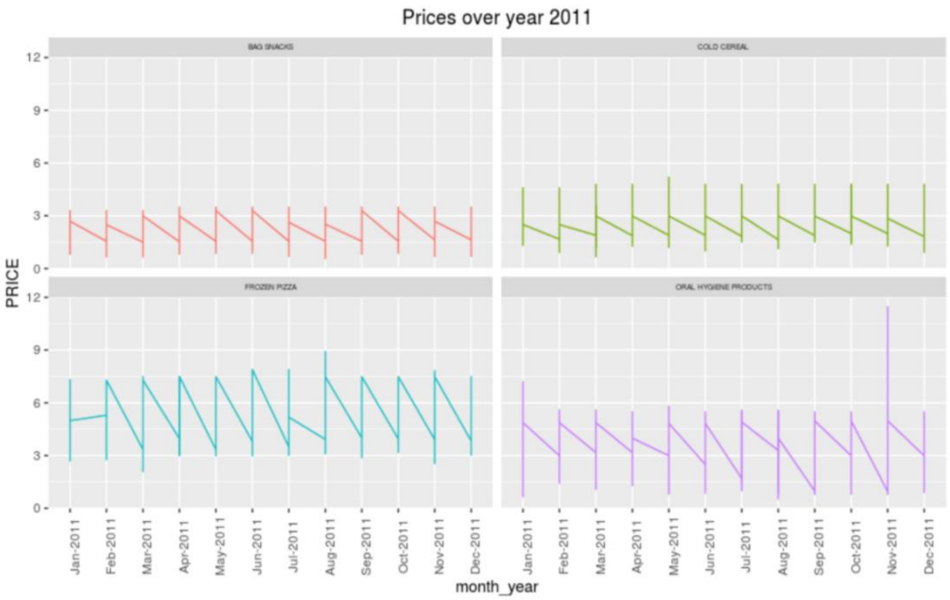
1. The prices of all products over the entire time frame are displayed in subplots for each category

The prices of all products sub-plotted into categories over the timeframe of 156 weeks is shown. Looking at this, it’s possible to make generalizations of price ranges of each category and thus their overall distributions of price over the timeframe. On average, bag snacks have a steady price range of around 1-3$ with relatively weak oscillations, while the price of oral-hygiene products fluctuates more with a higher standard deviation between 1-11$. To be able to estimate on seasonal changes, we focus on individual years as well.



2. The prices of all products over year 2010 are displayed in subplots for each category

We subset the dataset of 2010 to observe changes with respect to months to observe how price varies seasonally throughout the year. Although oral-hygiene products span a broad price range between 1-8$, we see a similar average price between 3-4$ for January, May, and August and an average of 3-5$ for every other month. This clearly indicates that the demand for these products saw a surge in those months. Similarly, the price of frozen pizza fluctuates steeply in August and November but mostly oscillates between roughly 2-7$, and we observe that in warmer months the price is relatively higher than colder months. Bag snacks' and cold cereals' price is relatively more stable with almost no outliers.



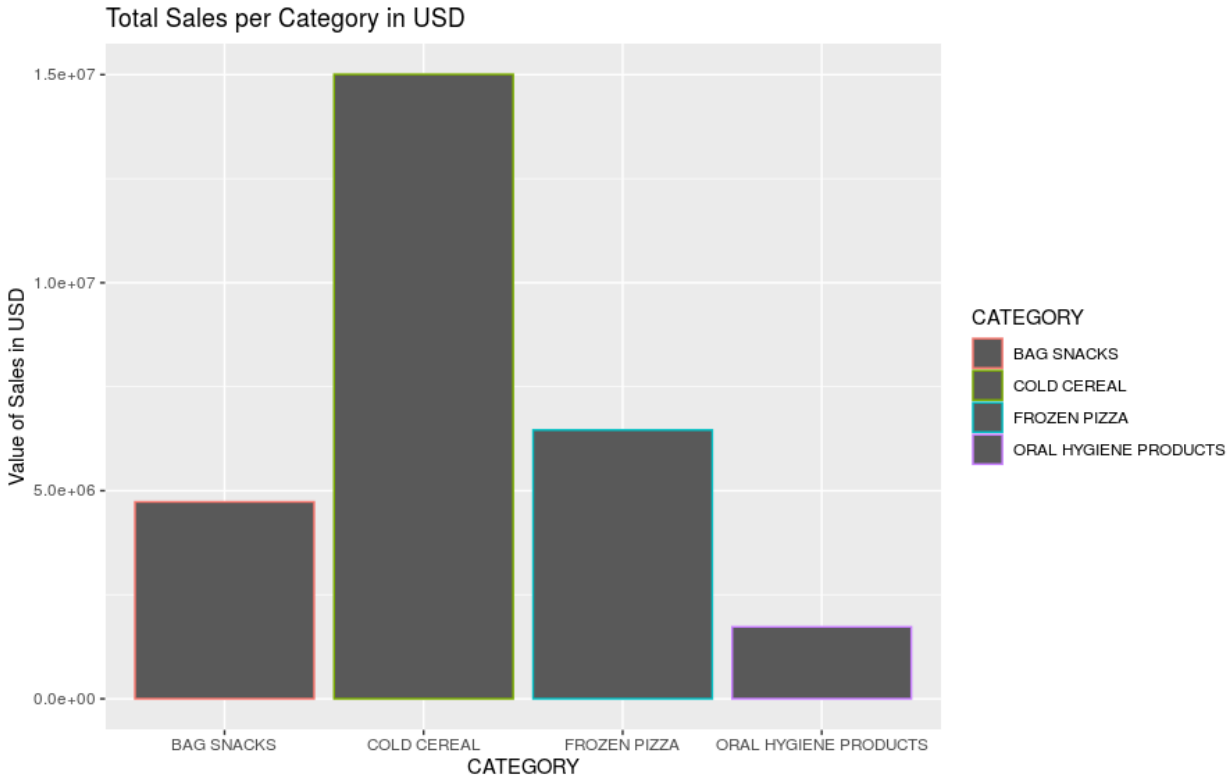
3. The prices of all products over year 2011 are displayed in subplots for each category

To generalize these insights for the entire period, we also examine year 2011 to determine whether similar conditional apply. We see that bag snacks and cold cereal have almost the same price patterns and range as before. For oral-hygiene products we see a decrease in price, thus average price of products drops below 6, suggesting a yearly decrease between year 2010-2011.

Therefore, we conclude that prices of bag snacks and cold cereals are relatively more stable, whereas oral-hygiene products and frozen pizzas price fluctuate more frequently within a year and thus show a rather more seasonal trend.

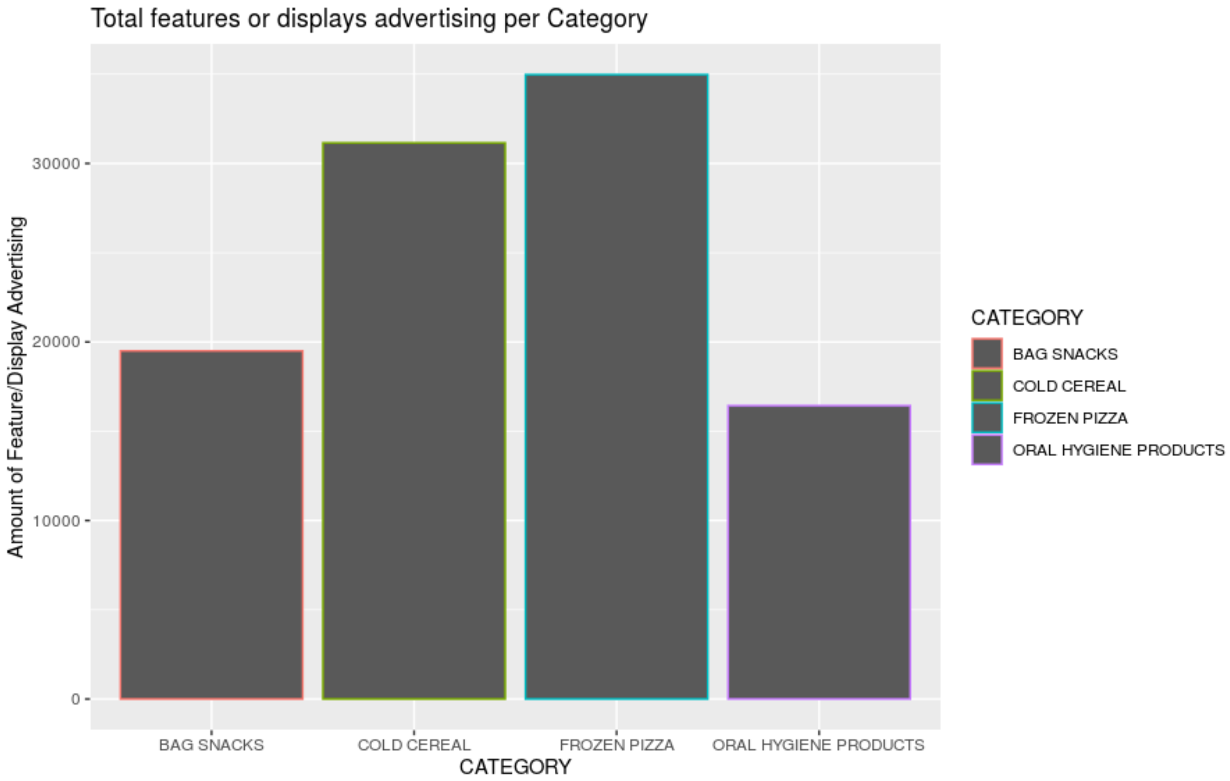
Categorizing Sales

We will now analyze the sales per category and compare them to the number of sales and the features to understand the relativity of sales and provide better insight.



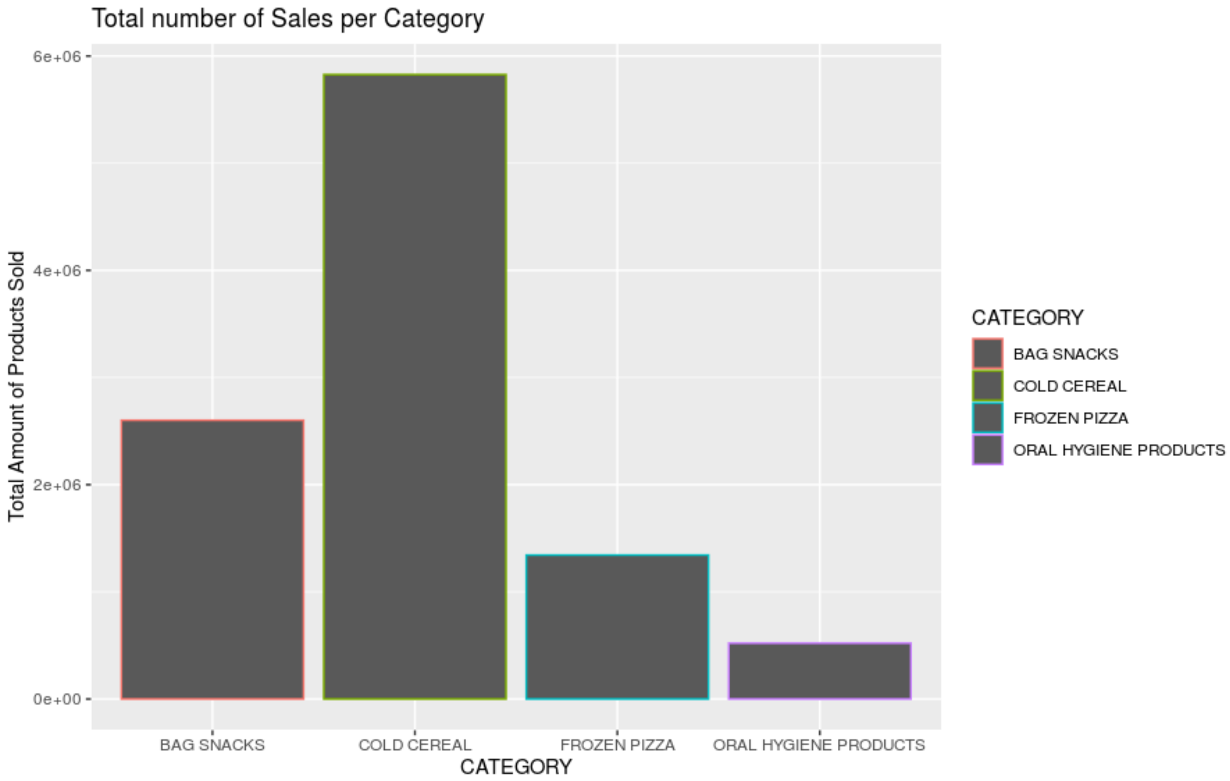
4. This graph presents the total sales of the company over the whole period

In this graph, we see that the sales realized on cold cereals are larger than the rest with a total of over $15M on the period. The other categories' sales are much smaller. Yet, we must look at other factors to understand the reality.



5. This graph presents the total number of sales of the company over the whole period.

With graph two, we see that the numbers generated in sales can't be our only factor to evaluate the company. Indeed, Cold cereals produce more money from sales, however, it also takes much more sales to do so. Despite producing more money than Bag snacks, Frozen pizzas made a smaller number of sales. The graph also comforts us in the idea that Oral hygiene products are less popular.



6. This graph presents the total features or displays advertising of the company over the whole period.

As said earlier, cereals produce more money and sell more products. This seems logical as it has one of the highest number of products displayed. However, for frozen pizza, we are in right to ask ourselves questions. While being the second category with the most sales and the third with the largest number of sales, it is the one with most products displayed.

In conclusion, we suggest the shops to ensure the products displayed are well thought. It seems that less pizzas should be advertised, so can be said about hygiene products. Oppositely, it seems that the number of sales of bag snacks is linked to the number of products advertised and should therefore be increased.

2. Price Elasticity Estimates

We want to estimate price elasticities using linear regression. We are looking to fit the following regression model:

$$\log(UNITS) = \alpha + \beta \log(PRICE) + \gamma FEATURE + \delta DISPLAY + \varepsilon$$

The Linear Regression model is a statistical model that analyses the relationship between the response variable, the units in log, and the explanatory variables, the price in log, the feature and the display. The model assumes that there exists a linear relationship between the variables.

We filtered the data by category and applied the linear model to each category. We obtained the coefficient for each category for the price, feature and display.

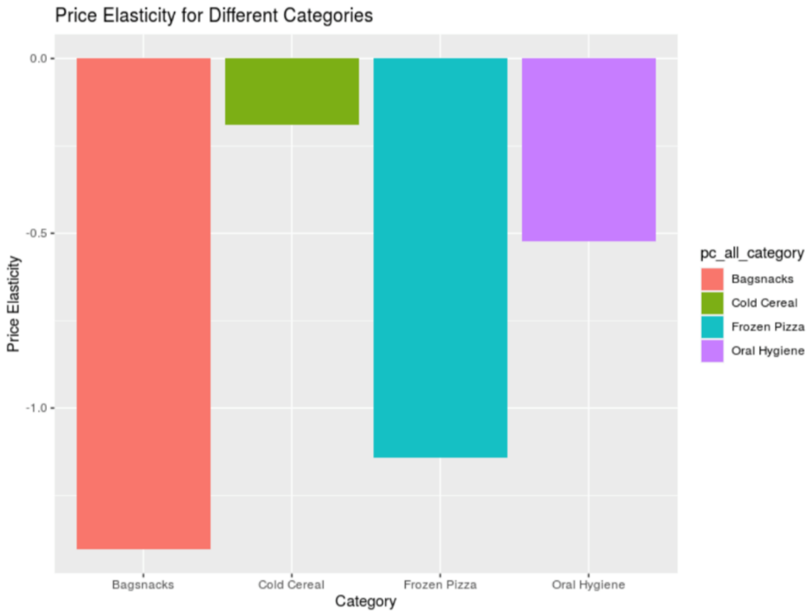
Price elasticity

Price elasticity measures the % change in demand given a 1% increase in price. By the law of demand, we expect elasticities to be negative.

The larger the number in magnitude, the more elastic demand is for that product. If elasticity is between –1 and 0, demand for that product is inelastic. On the other hand, if elasticity is greater in magnitude than –1, demand is elastic.

We notice that elasticity for both bag snacks (-1.4032) and frozen pizza (-1.1430) are greater in magnitude than –1: demand for these products is elastic.

On the other hand, cold cereals (-0.1884) and oral hygiene (-0.5224) products both have a coefficient between –1 and 0: demand is inelastic.



7. The graph shows the difference in price elasticity among different categories

Feature advertising

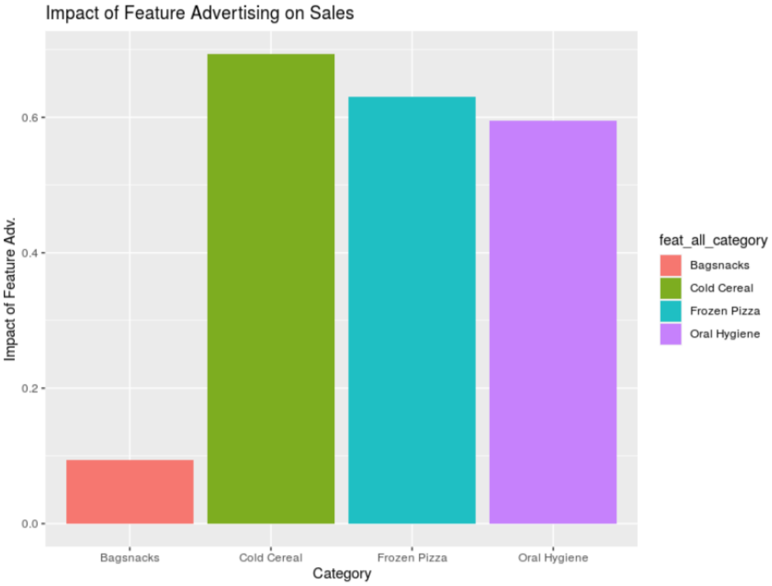
Feature advertising means the product was in in-store-circular.

Here, the greater the coefficient, the more featuring a product gives a boost to demand.

We notice that the category cold cereal has the greatest coefficient (0.6932), meaning it is most affected by feature advertising.

Both Frozen Pizza (0.631) and Oral Hygiene (0.5948) categories are affected by feature advertising as well, but less than Cold Cereal.

Finally, Bag Snacks (0.094) are not really affected by feature advertising.



8. The graph shows the impact of Feature Advertising on sales of different categories

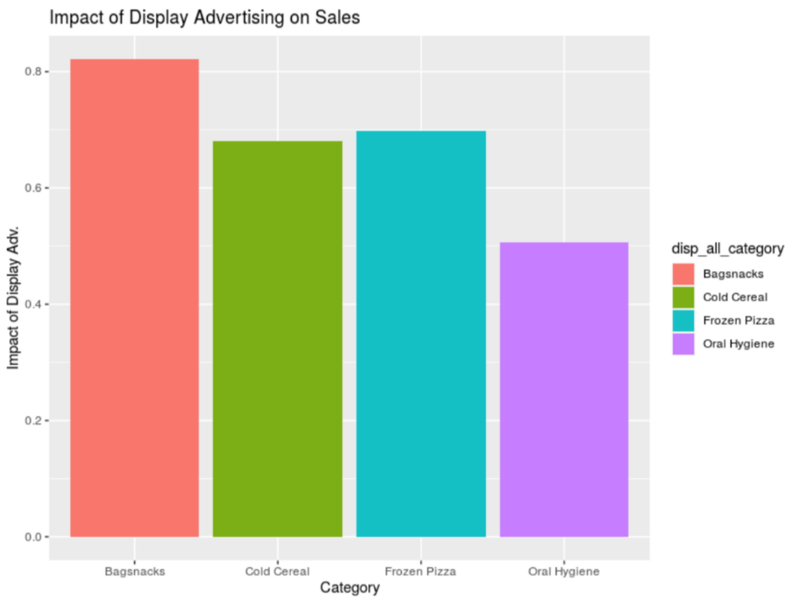
Display advertising

Display advertising means the product was part of an in-store promotional display.

Here, the greater the coefficient, the more display advertising increases demand.

We notice that the bag snacks category (0.8213) is the most affected by feature advertising.

Oral hygiene products (0.507), Cold Cereals (0.68) and Frozen Pizzas (0.699) are also affected, but less.



9. The graph shows the impact of Display Advertising on sales of different categories

The p-value is a good indicator of the quality of our model. The smaller the p-value, the less likely it is we will observe a relationship between the predictor and response variables due to chance. We notice that in every model, the p-value is under 0.05 (the highest p-value being 9.03e^-8). This further supports our model and the accuracy of the results found.

To conclude, according to our models,

- the bag snacks product category is the most affected by an increase in price. It also sees the biggest increase in demand when in display advertising. This makes sense as people will often pick these snacks because they are affordable and on display, for example near the cashier. Feature advertising has a limited impact on the sales of bag snacks.

  => Keeping the prices low and focusing on display advertising is a good strategy for the bag snacks category.

- the cold cereals product category sees the biggest increase with feature advertising. It is also, on a lower scale, affected by display advertising. An increase in price doesn't affect the sales of cold cereals as much as other products.

  => Focusing on feature advertising is a good strategy for the cold cereals category.

- the frozen pizza product category was greatly affected by an increase in price. This make sense as frozen pizzas are cheaper alternative and are often chosen solely due to their affordable price. The products also see an increase both in feature and display advertising.

  => Keeping prices low, as well as moderately featuring and displaying is a good strategy for the frozen pizza category.

- the oral hygiene product category sees an increase both by feature and display advertising. It is less affected by an increase in price than other products.

  => Moderately featuring and displaying is a good strategy for the oral hygiene category.

3. Demand and Forecasting

In order to find the model that best predicts future unit sales, we tested 3 regression models, namely multiple linear regression, regression tree and random forest. With unit sales as the dependent variable, we have to determine the independent variables used in the regression models. There are 12 independent variables, but we have selected 5 variables (PRICE, BASE_PRICE, FEATURE, DISPLAY and TPR_ONLY) that we believe are most appropriate for predicting unit sales.

First, we have to find the best set of variables for each of the regression models. Cross-validation was used on 26 potential sets with the 3 regression models to test their ability to predict on unseen data. Through calculating the in-sample root mean squared errors (RMSE) and other selection criteria, the best predictive set is selected for each of the three models. The best predictive set for the multiple linear regression model is the 21st set while the 11th set is the best for both the regression tree and random forest models (see Appendix, Table 1)

Next, using the selected sets of predictor variables, we apply the 3 models to the test data and calculate the out-sample RMSEs. As seen in the table below, among the three models, the Random Forest model has the lowest out-sample RMSE. Having the lowest out-sample RMSE indicates that the Random Forest model has the best fit and has the most accurate predictive value.

	In-Sample RMSE	Out-Sample RMSE
Linear Regression (21st set)	26.53652	18.56039
Regression Tree (11th set)	25.48388	19.05943
Random Forest (11th set)	31.00492	17.30354

The Random Forest model uses the 11th set of predictor variables, containing PRICE, BASE_PRICE and FEATURE. This reveals that the most important variables for predicting future unit sales are PRICE (the current price at which the unit is sold at), BASE_PRICE (the original price at which the unit is sold at), and FEATURE (whether the unit was in the in-store circular).

Using the Random Forest model allows the grocery retailer to predict how potential changes in PRICE, BASE_PRICE and FEATURE variables of a product would affect its sales. This would enable the grocery retailer to simulate different combinations to achieve the highest unit sales.

Reflection

This project was a way to apply what we learned in class: by practicing the use of R on a specific problem while working together as a group, we got a glimpse into what working as a data analyst for a real company looks like.

The greatest challenge has been coordinating with each other: we had to divide the work, each dedicating to their own part, but then needed a lot of coordination and clarity in order to put everything back together in a seamless way.

The most fun part was finally getting a line of code right after endless attempts to make it work. It was particularly rewarding to finally succeed in something where you put in so much effort.

Overall, we worked very well together. We were organized, dividing our team into smaller ones of two in order to work productively at the same time. We managed to schedule regular meetings, where we brainstormed ideas, discussed the progress we had made, coordinated our findings, encouraged and helped each other out whenever a problem occurred. Everyone did their part with equal enthusiasm and hard work.

Appendix

Table 1

		Linear Regression			Regression Tree		Random Forest	
Set	Variables	Training (In)	Training (Out)	Test	Training	Test	Training	Test
1	PRICE BASE_PRICE	28.40582	28.1292		27.50147		33.44367	
2	PRICE FEATURE	27.40295	27.12568		26.18365		32.0209	
3	PRICE DISPLAY	27.5154	27.19033		26.64742		31.32826	
4	PRICE TPR_ONLY	28.74015	28.4388		28.4498		31.43681	
5	BASE_PRICE FEATURE	27.48928	27.20847		26.58209		30.99726	
6	BASE_PRICE DISPLAY	27.7408	27.40834		26.89731		32.62011	
7	BASE_PRICE TPR_ONLY	29.29714	28.99377		28.9734		31.41356	
8	FEATURE DISPLAY	28.00667	27.71947		27.92321		31.64544	
9	FEATURE TPR_ONLY	28.64343	28.37326		28.59416		31.26696	
10	DISPLAY TPR_ONLY	28.60188	28.29296		28.54613		31.75243	
11	PRICE BASE_PRICE FEATURE	27.40305	27.12577		25.48388	19.05943	31.00492	17.1592
12	PRICE BASE_PRICE DISPLAY	27.45056	27.13653		26.21778		32.35864	
13	PRICE BASE_PRICE TPR_ONLY	27.99796	27.72041		26.27377		31.06308	
14	PRICE FEATURE DISPLAY	26.85653	26.55911		25.92209		31.4687	

15	PRICE FEATURE TPR_ONLY	27.37638	27.09757		26.18365		31.55979	
16	PRICE DISPLAY TPR_ONLY	27.49779	27.17155		26.64742		31.06109	
17	BASE_PRICE FEATURE DISPLAY	26.85557	26.55308		26.29215		31.86553	
18	BASE_PRICE FEATURE TPR_ONLY	27.48916	27.20837		26.58209		31.18031	
19	BASE_PRICE DISPLAY TPR_ONLY	27.73978	27.40742		26.89731		31.03395	
20	FEATURE DISPLAY TPR_ONLY	27.98975	27.70464		27.92321		31.25994	
21	PRICE BASE_PRICE FEATURE DISPLAY	26.83718	26.53652	18.56039	25.59435		31.53957	
22	PRICE BASE_PRICE FEATURE TPR_ONLY	27.36095	27.08372		25.48388		31.55581	
23	PRICE BASE_PRICE DISPLAY TPR_ONLY	27.35528	27.04492		26.21778		31.30416	
24	PRICE FEATURE DISPLAY TPR_ONLY	26.85023	26.55195		25.92209		31.08075	
25	BASE_PRICE FEATURE DISPLAY TPR_ONLY	26.84686	26.54474		26.29215		31.4064	
26	PRICE BASE_PRICE FEATURE	26.83726	26.5364		25.59435		31.52942	

	DISPLAY TPR_ONLY							