Data Analysis for Grocery Sales

Jing Chu jingchu@umich.edu
Junyi Li junyili@umich.edu
Zhen Qin qinzhen@umich.edu
Zhichao Yang yangzhic@umich.edu

Abstract

This report aims to analyze the sales of Corporación Favorita, a large Ecuadorian-based grocery retailer. The dataset contains 80 days' sales data of 1012 items across 54 stores. Data exploration and visualization is performed where we found noteworthy patterns of sales. To verify these patterns and also to testify the taxonomy of items the retailer uses from the perspective of unit sales, we fit a clustering model, where we use Dirichlet Process Mixture Model (DPMM). The clustering results are coherent with what we find in the data exploration process, moreover, the sales of items show some new patterns. Then, to predict unit sales, two regression methods are proposed. We first fit a mixed-effects linear regression model which treats stores and item families as random effects. Furthermore, Bayesian Additive Regression Tree is implemented which is more robust to outliers. Both regression methods generate satisfactory prediction results.

1 Introduction

Corporación Favorita[1] is a large Ecuadorian-based grocery retailer and they operate hundreds of supermarkets with over 200,000 different products. Currently, Corporación Favorita only uses subjective prediction to forecast sales, which are not accurate enough. Mistaken prediction may lead to serious consequences. If the future sales are overestimated, the goods will be overstocked and the perishable goods may go bad. While, when the future sales are underestimated, some popular grocers may be sold out and lead customers to purchase in their competitors' stores. The problem of prediction may be even more complicated if new stores are added, more seasonal goods are transported among various regions and so on. Therefore, Corporación Favorita is interested in applying machine learning to accurately forecast sales. Besides sales forecasting, we are also interested in doing clustering to dates and item classes based on unit sales, these clustering results are very useful to testify our intuition about the effect that each variant has on unit sales, for example, we will guess the unit sales will be higher over the weekend compared to the workdays, based on this assumption, we will want to just include two categories of dates instead of seven in our prediction model. To know if this assumption is appropriate, one of the easiest approach is to cluster unit sales data of each day and see if the sales data over weekends forms one cluster and the sales data of weekdays form another cluster. Furthermore, we can also get new insight to the data if we use a cluster model that can learn the number of clusters automatically, for example, the items are classified into 148 classes without clear definition for each class, we can do clustering to these classes to see if they have some connection to unit sales.

Reminder of the paper is organized as follows. Section 2 performs data description and visualization. Section 3 introduces the clustering model, i.e. DPMM and the insight we got from the dataset with DPMM. Section 4 builds regression models (Mixed Effect Linear Regression Model and Bayesian Additive Regression Tree Model). Conclusion and discussion are provided in Section 5.

2 Data

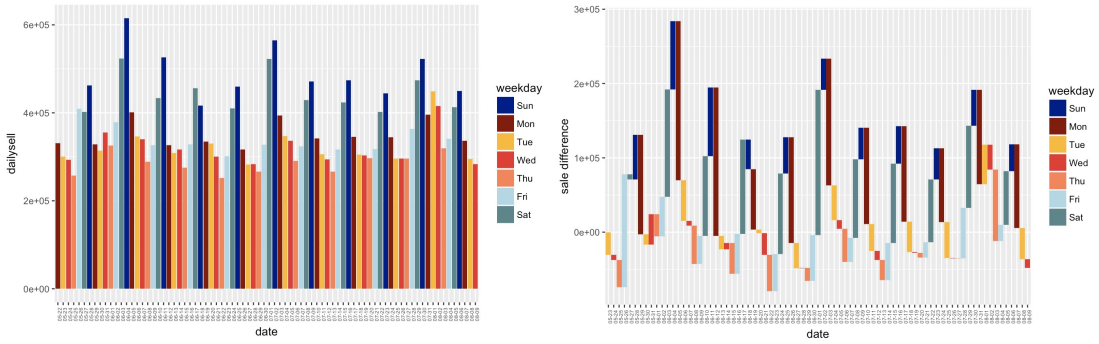
2.1 Data description

The original data is extremely large, which have more than 120 million observations and have too many variables. We resort data by only considering the data from 2017/05/22 to 2017/08/09 (80 days) and select some variables which are *item number* (1012 items in total), *store number* (54 stores in total and each store contains the 1012 items), *family* (23 types of grocers), *city* (22 in total), *perishable* (0 for imperishable, 1 for perishable), whether *on promotion* and *daily sales*. We search additional information such as population[2] and geographic coordinates of cities (2010)[3] to learn

[1] Kaggle: https://www.kaggle.com/c/favorita-grocery-sales-forecasting
[2] Census data: https://kaggle2.blob.core.windows.net/forum-message-attachments/234737/7733/Censo%20Poblacion%202010.xls
[3] Geometrical data: http://www.tageo.com/index-e-ec-cities-EC-lg-ch.html

more about the effect of population density and location to sales.

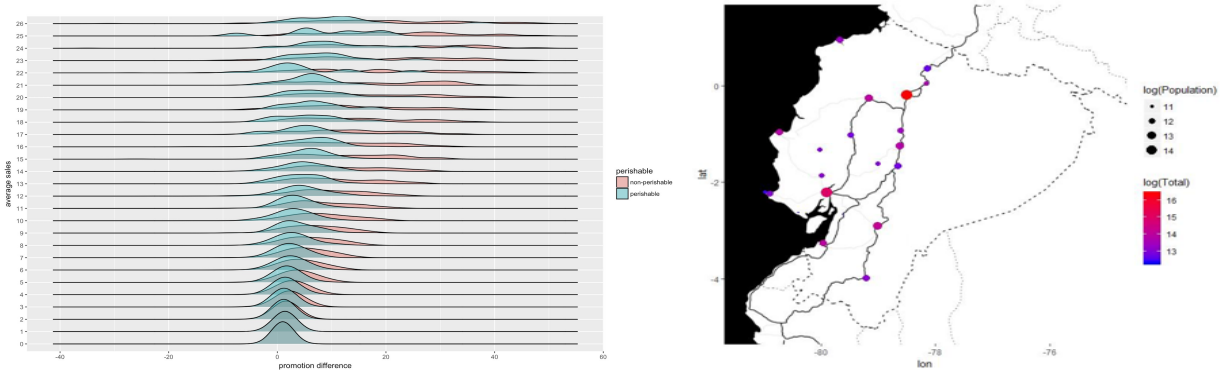
2.2 Data exploration



(a) *(b)*

Figure 1: (a) Total sales per day. (b) Daily sales difference compared with sales in previous day

Both figures are colored according to the days of week. **Fig 1(a)** is the histogram and **Fig 1(b)** is a waterfall plot. Each bar in **Fig 1(b)** starts from either the top or the bottom of previous bar, which indicates the increase and the decrease of sales of previous day respectively. **Fig 1(a)** indicates that the sales on Saturday and Sunday are much higher than the rest days. What's more, there are significant increases of sales from Friday to Sunday in **Fig 1(b)**. This may be caused by the weekend activities and the storage for the following weekdays. After weekends, the sales will drop sharply on Monday and fluctuate slightly until Friday. This trend shows that there is noticeable difference between weekdays and weekend purchase patterns. However, the sales on the date of payoffs may not follow the regular tendency. We do further research and find that, in the region of Corporación Favorita's businesses, people are paid biweekly in the beginning and the middle of each month. On the 1st of June (Thursday), **Fig 1(a)** and **Fig 1(b)** show that the sales compared with that in the previous day decrease but in a much slower way. Despite sales on every Saturday boost, sales on the 1st of July (Saturday) increase much more than usual. The sales increase on the 1st of August (Wednesday), which is contrast to sales on the rest Wednesdays. Different from the pattern at beginning of each month, sales at middle of each month will not display special pattern. Therefore, it seems that the wages paid at the beginning of each month will stimulate people to purchase, while, the wages paid at the middle of each month does not affect the sales.



(a) *(b)*

Figure 2: (a) Effect of promotion on sales. (b) Map labeled by population and total sales

Let S_p and S_u represent the median sales of items when on promotion and median sales of items when not on promotion respectively. Promotion difference is the median sales difference $S_p - S_u$. Average sales are derived by rounding $\frac{S_p+S_u}{2}$. The curves in **Fig 2(a)** show the distribution of promotion difference grouped by perishable/non-perishable items. As the sales become larger, the promotion differences for both perishable and non-perishable items grow steadily since

the center of curves gradually move to the right-hand side. In other words, promotion has a positive effect on sales. Furthermore, the gap between promotion difference of perishable and non-perishable items are larger. It seems that non-perishable items will be affected more by whether there is a promotion.

Fig 2(b) is based on the additional information, i.e. population and geographic coordinates. The black area is sea and the solid lines are the traffic routes. It seems that supermarkets are preferred to be constructed in coastal or traffic convenient areas. The reason to use log transformation of the data is that transformation will lead data to be more easily visually inspected. The larger the dots, the larger the log of population, and the redder the dots are, the larger the number of sales are. Clearly, areas with large population have more sales.

3 Clustering

3.1 Dirichlet Process Mixture Model

3.1.1 Dirichlet Process

The Dirichlet process [2] is an infinite-dimensional generalization of the Dirichlet distribution, and it can be used to set a prior on unknown distributions. More formally, Dirichlet process is defined as follows, suppose P is a random probability measure, S is a measurable set, P_0 is a distribution and α is a real number, then we say:

$$P \sim DP(P_0, \alpha) \tag{1}$$

when for any finite partition $S_1, S_2, ..., S_k$ of S:

$$P(S_1, S_2, ..., S_k) \sim Dirichlet(\alpha P_0(S_1), \alpha P_0(S_2), ..., \alpha P_0(S_k)) \tag{2}$$

We call P_0 the base measure and α the concentration parameter of Dirichlet process.

3.1.2 DPMM

Dirichlet Process Mixture Model is a generalization of Finite mixture model, it can learn the number of clusters automatically, which is a good property we desire in many scenarios, however, this kind of flexibility means we have to sacrifice some kind of efficiency from computational perspective. The formal definition of a generalized mixture model or DPMM is:

$$f(y|P) = \int \kappa(y|\theta) dP(\theta) \tag{3}$$

where $\kappa(\cdot|\theta)$ is a kernel and θ includes location and possibly scale information, P is a mixing probability measure, we set the prior of P as a Dirichlet process in DPMM. More specifically, we use a Gaussian distribution as the kernel distribution in our data analysis. What's more, the base measure of Dirichlet process is set as:

$$P_0(\theta|\gamma) = \text{Normal}(\mu|\mu_0, \frac{\sigma^2}{k_0})\text{Inv} - \text{Gamma}(\sigma^2|\alpha_0, \beta_0) \tag{4}$$

Which is a Normal Inverse Gamma distribution that is conjugate to the posterior distribution and $\gamma = (\mu_0, k_0, \alpha_0, \beta_0)$ is a representation of its parameters. Then the concentration parameter α is set as Gamma distribution:

$$\alpha \sim Gamma(\alpha_\alpha, \beta_\alpha) \tag{5}$$

3.2 Clustering Results

In order to obtain more insight to data, clustering are performed on different variables ahead of other tasks. More specifically, we divide **Date** and **Class** into several groups where instances have significantly different sales. In addition, we explore the relationship between cluster labels and the overall characteristics of sales. For purpose of stability, all numerical variables (sales here) are centered and scaled to zero mean and unit variance as the document suggested. To visualize the clusters of high dimension data in a 2 dimensional plot, we perform Multidimensional Scaling (MDS) to the data set. The process of MDS is as follows, firstly scale the data and compute the distance matrix for each observations by Euclidean distance, then we generate 2-dimension representation of each observation. Finally, we use R package **dirichletprocess** [5] to perform DPMM.

3.2.1 Date

When performing clustering on dates, parameters in **Equation (4)** are set to be default values of the package, but to make DPMM converge to small number of clusters, we set prior of α as $Gamma(4, 50)$, with this prior, we can make α get big values with high probability, therefore, DPMM can converge to small number of clusters.

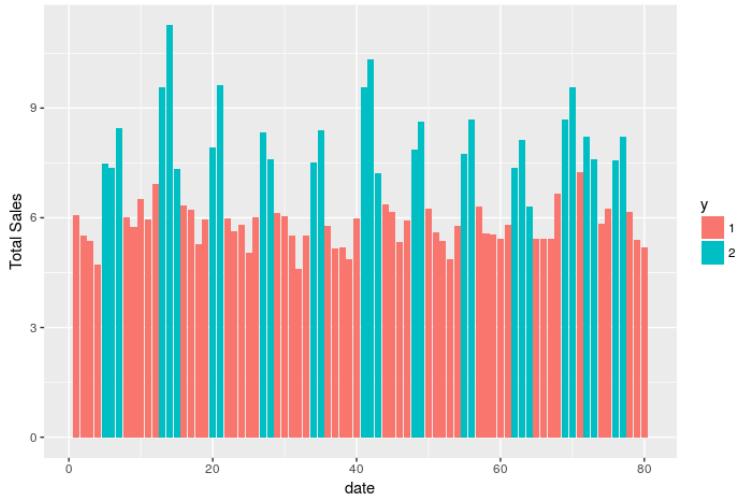


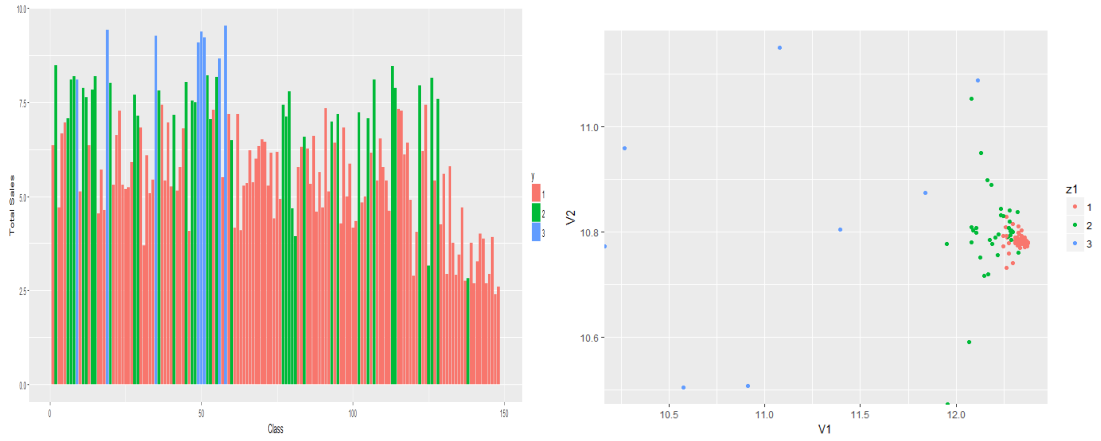
Figure 3: This figure shows clustering results of dates, different colors represent different clusters, th result is based on each day's total sale, the horizon axis stands for dates and the vertical axis stands for total sales.

Fig 3 shows that the 80 days are divided into two groups. It is clear that the light blue group(higher sales) mainly include weekends, sometimes the group also contains Friday or Monday. In contrast, the red group(lower sales) mainly embodies weekdays. It is a very intuitive and reasonable result because many people usually buy groceries on weekends for every week, so we verify our intuition got from the data exploration part.

3.2.2 Class

After clustering on dates, we also do a clustering analysis on **Class**. **Class** is a feature of items, there are 148 classes in total and each class belongs to one family. In order to include all scenarios, we compute the sum of sales per day per **Class**, in other words, if we denote s_{ij} as total sales of class i on date j, then we use $S_i = [s_{i1}, s_{i2}, ..., s_{i80}]$ to represent class i. The prior parameters are all kept the same.

Fig 4(a) shows that item classes are divided into three groups. Generally speaking, the group colored blue has the highest sales, the group colored green has relatively high sales and the group colored red has low sales. In **Fig 4(b)**, points with clustering labels are plot by MDS coordinates. Note that the total sales for different classes differ a lot so we use the log form of the MDS coordinates. There are many blue points which have very high sales. Green points and red points are closer to each other, but there is a clear boundary between them. What's more, we want to interpret the meaning of each cluster, but **Class** is encoded as numbers, so we don't know the exact meaning of each class. We instead compare the relationship between a class's family and its cluster, the result shows there are not strong relation between these two taxonomies, which is very intuitive. **Family** characterizes application of an item class, which differs from its popularity, i.e. sales.



(a) *(b)*

Figure 4: (a) and (b) show clustering results of class,different colors represent different clusters, (a) is based on each class's total sale and (b) is MDS coordinates.

4 Regression

4.1 Data preprocessing

Some items are randomly chosen to show the distribution of unit sales. There are many outliers in **Fig 5**, which are located at around 1500. In fact, several outliers are even at about 8000. It seems that all the outliers are on promotion. By looking for the type of items and whether they are perishable for the outliers, we find that most outliers are non-perishable such as beverage. Outliers will affect the linear assumption; hence logarithm transformation of sales have been taken to reduce the effect of outliers.

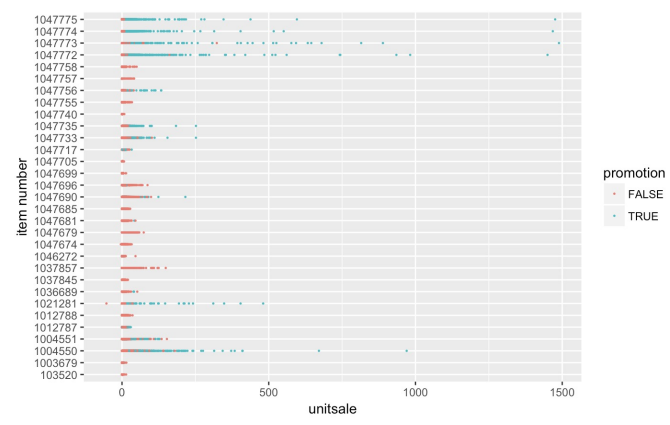


Figure 5: Unit sales of items with or without promotion

What's more, we have already concluded that the larger the population in a city, the larger the unit sales in this city. As *city* is categorical variable, it cannot give us too much information. Then, numerical variable *population* will replace *city*. Population is the additional data which is generated in 2010 and is the latest data we have found online. We also take logarithm of population as the predictor. The influence of population on regression will be further explored.

When building the regression model, variables $\log(unitsales+1)$, *family*, *store_nbr*, *onpromotion*, and $\log(population)$ are selected as predictors. In addition, the interaction terms $\log(unitsales+1) \times onpromotion$ are added in the regression model as promotion affects daily sales significantly.

Data is divided into train and test dataset. Considering the weekly and monthly patterns of daily sales, all the information from June 03^{rd} to June 17^{th} are set to be the variables for the train dataset and the response is the unit sales on June 18^{th}. Similarly, sales from July 1^{st} to July 15^{th} are considered as variables for test dataset and the

response is the sales on July 16^{th}. Both June 03^{rd} and July 1^{st} are the first Sundays of the corresponding month. Since train and test dataset are chosen at comparable period, the bias of prediction can be reduced to the minimum.

4.2 Bayesian Hierarchical Regression Model

Given the hierarchical structure of stores and items, it is natural to consider a hierarchical mixed-effects regression model with the random effects of *family* and *store_nbr*. The model is in the form of $Y = \beta X + uZ + \epsilon$, where β are the fixed effects and u are the random effects. X are $\log(unitsales + 1)$, *onpromotion*, $\log(population)$ and $\log(unitsales + 1) \times onpromotion$; Z are *family* and *store_nbr*; and ϵ is the error term. The model specifies that ϵ i.i.d. follows $N(0, \sigma^2)$, the random effects u follows specified normal distribution as indicated in **Table 1** and $\hat{Y}|u \sim N(\beta X + uZ, \sigma^2)$.

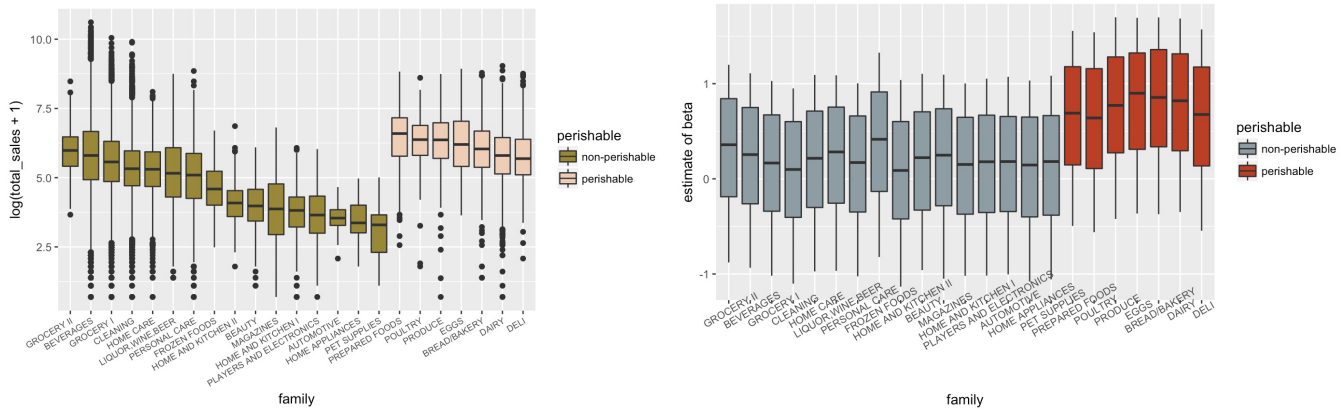
Random effects	$u^{non-perishable}$	$u^{perishable}$	u^{store}
number of coefficients	16	7	54
models	$N(\mu_1, \sigma_1^2)$	$N(\mu_2, \sigma_2^2)$	$N(\mu_3, \sigma_3^2)$

priors for μ_i follows degenerate uniform distribution; priors for $\sigma_i \propto \frac{1}{\sigma_i^2}$

Table 1: Models of random effects

Based on **Fig 6(a)**, we can find that the perishable items have larger average unit sales and the variations of sales within perishable group are smaller compared with that of non-perishable items. Therefore, the constraint $\mu_1 < \mu_2$ is made. The model is fitted using **stan** [4] where Hamiltonian Monte Carlo algorithm is employed.

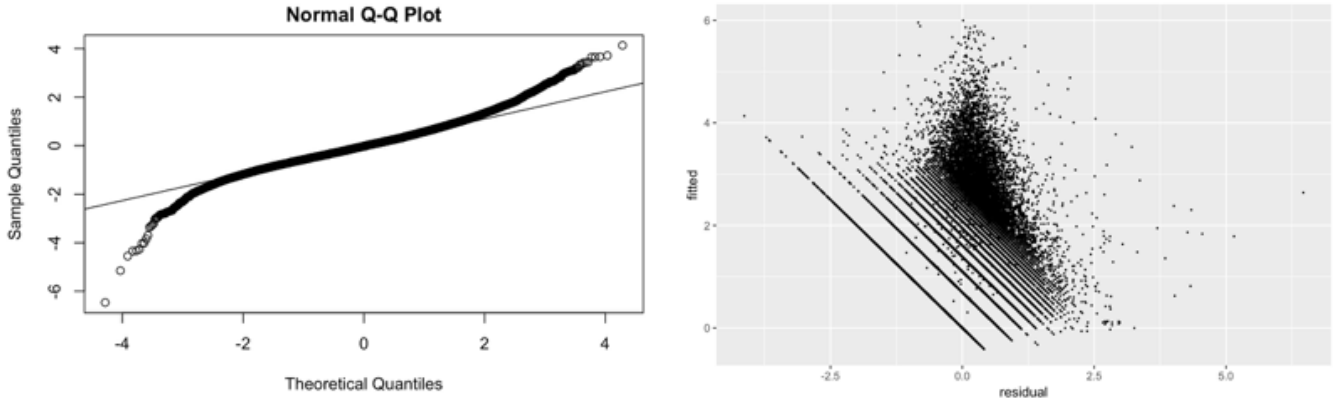
From **Fig 6(b)**, we can find that posterior mean of $u^{non-perishable}$ is about 0.15 and posterior mean of $u^{perishable}$ is about 0.71. There is an obvious difference between posterior means, which confirms that perishable and non-perishable items should have different normal distributions. It agrees with our model assumptions on hyper-parameters above.



(a) *(b)*

Figure 6: (a) Item family total sales. (b) Posterior random effects for family.

Fig 7(a) shows that the residuals have heavier tail compared with normal distribution. In the upper right corner, some points are above the quantile line. The reason is that non-perishable items are affected greatly by promotion. Points in the bottom left corner are below the quantile line, which is caused by large amount of 0 or 1 unit sales. **Fig 7(b)** shows similar results. When sales are 0, the points even form straight lines, due to large amount of 0 sales and possible fitted values. Since log transformation has been taken, all the lines are moved to the left. From **Fig 7(a)** and **Fig 7(b)**, it seems that the mixed effects model does not perform well when unit sales are small. However, there are some evidence to show that this model is still good.



(a) *(b)*

Figure 7: (a) QQ plot of residuals. (b) Fitted values VS residuals.

Min.	1^{st} Qu.	Median	Mean	3^{rd} Qu.	Max.
0.0001	0.6037	1.5510	4.3480	3.8720	2462

Error is in terms of unit sales instead of logarithm of unit sales

Table 2: Summary of root square error (absolute error)

Table 2 shows there is an extremely large outlier 2462. It affects the value of mean error. Hence, median is considered instead. The median is 1.55 which indicates that residual is very small and the median prediction difference between fitted values and actual result is only about 1.55 units. Therefore, mixed effect model can effectively help us predict the sales.

4.3 Bayesian Additive Regression Tree (BART)

Since prediction result of the mixed effect model has some extremely large residuals, we want to try some more complicated model like Bayesian Additive Regression Tree [3]. The advantage is that it is an ensemble tree method and regression tree can properly handle the situation where both categorical variables and continuous variables exist. In the following section, model built by BART will be explored. We consider how perishable and non-perishable items will affect the prediction result separately. We introduce the theorem of BART first, then give details of the result derived from BART model.

4.3.1 Overview of BART

The Bayesian additive regression tree model [3], [1] can be expressed as $Y = f(X) + \epsilon = \sum_{i=1}^{K} T_i^M(X) + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$, where Y is a $n \times 1$ vector of responses, X is a $n \times p$ design matrix and ϵ is the $n \times 1$ vector of noise. T denotes the tree structure and M denotes the parameters at the terminal nodes, so T^M represents an entire tree with both its structures and set of leaf parameters. Besides, the set of the tree's leaf parameters are denoted as $M_i = \{\mu_{i_1}, \mu_{i_2}, ..., \mu_{i_{b_i}}\}$ where b_i is the number of terminal nodes for a given tree.

The prior can be expressed as:

$$\begin{aligned} P(T_1^M, ..., T_k^M, \sigma^2) &= \left[\prod_{i=1}^{k} P(T_i^M)\right] p(\sigma^2) \\ &= \left[\prod_{i=1}^{k} P(M_i|T_i)P(T_i)\right] p(\sigma^2) \\ &= \left[\prod_{i=1}^{k} \prod_{j} P(\mu_{ij}|T_i)P(T_i)\right] p(\sigma^2) \end{aligned} \tag{6}$$

where Tree structure T, Leaf parameters given the tree structure $\mu|T$ and σ^2 are independent with each other.

The first component of the prior,$P(T_i)$ affects the locations of nodes within the tree which contains three components:

(i) The probability that a node at depth d is nonterminal followed by $\alpha(1+d)^{-\beta}$, where $\alpha \in (0,1)$ and $\beta \in [0,\infty]$

(ii) At each interior node, the distribution of the splitting variable assignments

(iii) At each interior node, conditional on the splitting variable, the distribution of the splitting rule assignment.

The default priors for (ii) and (iii) are both uniform. The second component of the prior is expressed as $p(\mu_{ij}|T_i) \sim N(\mu_\mu, \sigma_\mu^2)$ and the third prior is chosen to be $\sigma^2 \sim InvGamma(\nu/2, \nu\lambda/2)$. Additionally, the number of trees also must be chosen by cross-validation.

The last part is for sampling. A Metropolis-within-Gibbs sampler is used to draw samples from posterior distribution: and the sampler's key feature is to employ Bayesian back-fitting, where the j^{th} tree fits iteratively as follows:

$$\begin{aligned} &1: T_1|R_{-1}, \sigma^2, M_1|T_1, R_{-1}, \sigma^2 \\ &2: T_2|R_{-2}, \sigma^2, M_2|T_2, R_{-2}, \sigma^2 \\ &... \\ &k: T_k|R_{-k}, \sigma^2, M_k|T_k, R_{-k}, \sigma^2 \\ &k+1: \sigma^2|T_1, M_1, ..., T_k, M_k, \epsilon \end{aligned} \tag{7}$$

where $R_{-j} := y - \sum_{i \neq j} T_i^M(X), j = 1, ..., k$. The chain is initialized with k simple single node trees and then iterations are repeated until convergence.

We implement this regression tree method by using the package **bartMachine**[3] whose algorithm is substantially faster than package **BayesTree** [1] in **R** since it is fully parallelized at the MCMC iteration level during prediction.

4.3.2 Model Exploration

Models based on perishable and non-perishable items will be performed separately.

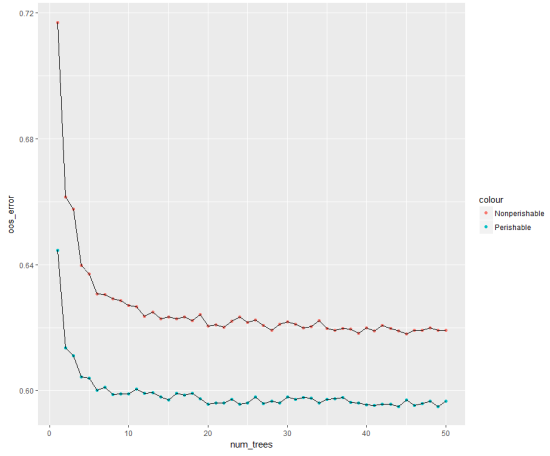


Figure 8: cross-validation of number of trees

To optimize the model result, cross-validation is used to determine the number of trees. The *oos_error* in **Fig 8** stands for out of sample root mean square error (RMSE). From the plot, we can see that RMSEs of non-perishable items are always larger than that of perishable items. The reason for this result may be caused by the great effect of promotion on non-perishable items. In addition, **Fig 8** suggests that we should set 44 trees with perishable items and 45 trees with non-perishable items in our following models.

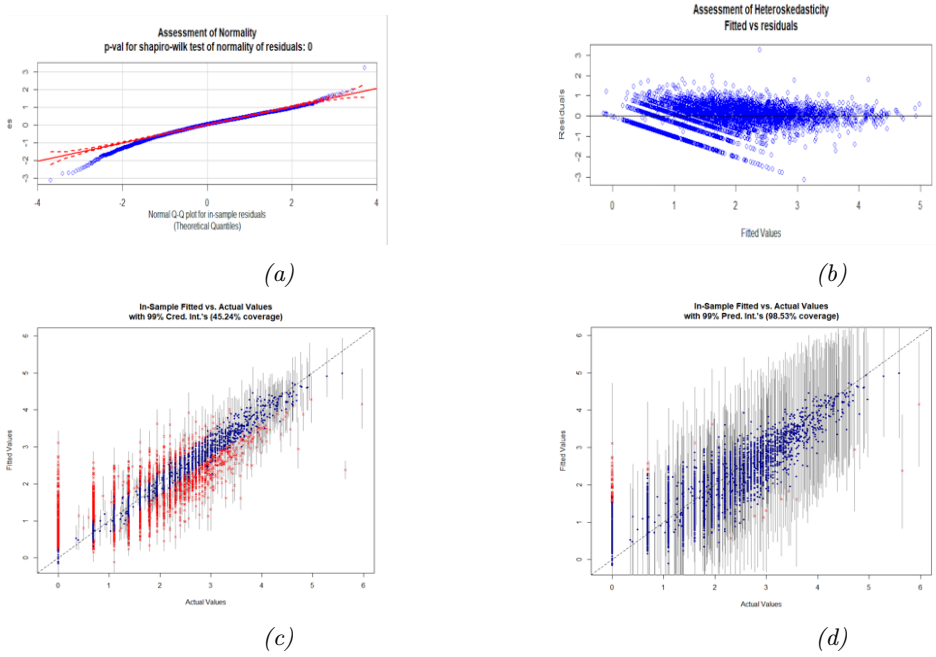


Figure 9: BART results for perishable items: (a) QQ plot of residuals with perishable items. (b) fitted values vs. residuals with perishable items. (c)confidence interval with perishable item. (d)prediction interval with perishable item.

Fig 9 shows the results of BART model for perishable items. The bottom left corner of **Fig 9(a)** shows some points are under the quantile line, which is due to the 0 or 1 unit sales points. It can also be seen from a line at the bottom left corner of **Fig 9(b)** that small unit sales in actual sales may lead to inaccuracy of prediction. The red points in **Fig 9(c)** and **Fig 9(d)** mean that the points are out of the range of confidence interval or prediction interval. While, blue points are within the interval. Since the range of confidence interval is smaller than that of prediction interval, there are more red points in **Fig 9(c)** compared with red points in **Fig 9(d)**. Furthermore, most red points appear at small unit sales, which indicates the same problem mentioned above that there are too many small sales points that affect the normal distribution.

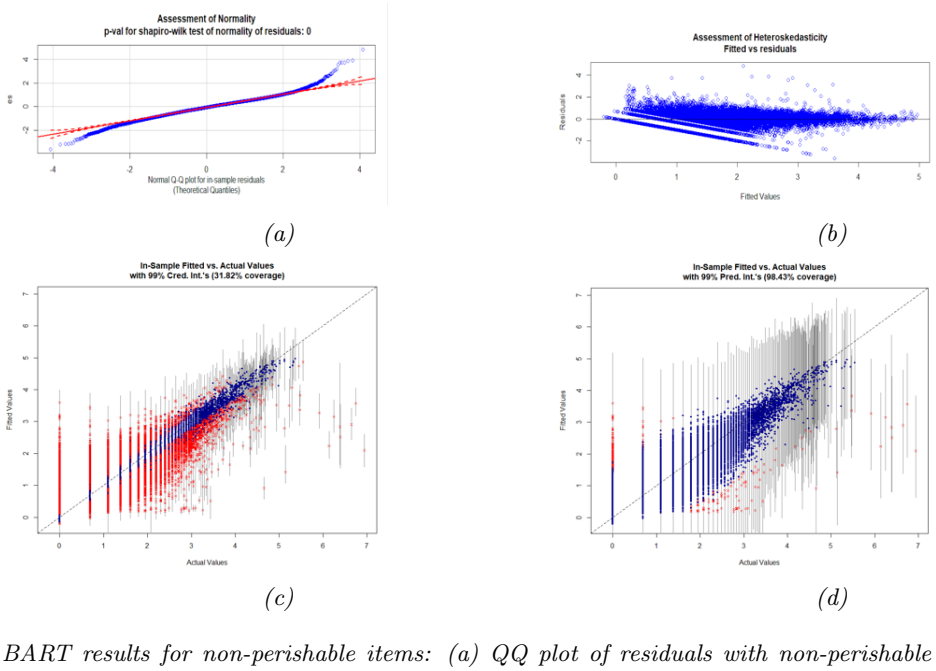


Figure 10: BART results for non-perishable items: (a) QQ plot of residuals with non-perishable items. (b) fitted values vs. residuals with non-perishable items. (c) confidence interval with non-perishable item. (d) prediction interval with non-perishable item

Fig 10 shows the results for non-perishable items. Different to the model with perishable items, the points in upper

right corner in **Fig 10(a)** are above the quantile line. This is due to the significant effect of promotion on non-perishable items. Hence, it is a heavy tail normal distribution. The line in **Fig 10(b)** shows that there are too many small unit sales in the data as well. Similarly, the range of prediction interval is wider than that of confidence interval, then the number of red points in **Fig 10(d)** is smaller. What's more, there are more red points occur at the small unit sales, which indicates the same problem mentioned above.

Overall, both models do not meet the assumption of residuals in normal distributions, which are caused by the large amount of small unit sales and whether there is a promotion. It seems that model with non-perishable items have more accurate prediction results than model with perishable items. Despite normal distributions assumption are not achieved, the models still estimate sales accurately.

	Min.	1^{st} Qu.	Median	Mean	3^{rd} Qu.	Max.
perishable	0.0000	0.8263	2.0870	3.6441	4.8008	177.5592
non-perishable	0.0001	0.6749	1.5837	2.9609	3.6000	273.9518

Table 3: Summary of root square error (absolute error)

Table 3 shows that there are outliers for both perishable and non-perishable items, which are 177.56 and 273.95 respectively. Hence instead of mean, we will explore median. The BART model according to the perishable items have about 2 units of prediction difference comparing to the actual unit sales. What's more, the difference is only about 1.58 units for non-perishable items. It shows that Bart models can predict unit sales precisely.

5 Conclusion

From simple data exploration and visualization, we find that sales show weekly and monthly patterns. At weekends people usually consume more groceries, however, this tendency is violated when the wages are paid in the beginning of each month. The sales are boosted in the beginning of each month no matter what day of the week it is. Items also have great effect on sales. Clearly, promotion stimulates sales significantly, especially for non-perishable items. What's more, population has positive relation with sales, and most supermarkets are built in coastal or traffic convenient areas. Clustering results testify the daily sales pattern we find before, and also show that the sales of item classes can be divided into three groups. Mixed Effects Regression Model and BART model are proposed to predict sales. Although residuals from training model do not meet the normal assumptions, the prediction results are accurate since the root median square error is small. In addition, BART models have much smaller maximum absolute errors than that of Mixed Effects Linear Regression Model in the prediction results.

References

[1] H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[3] A. Kapelner and J. Bleich. bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*, 2013.

[4] S. D. Team et al. Rstan: the r interface to stan. *R package version 2.14. 1*, 2016.

[5] Y. W. Teh. Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer, 2011.