Predicting Walmart Sales: EDA Report

Matthew Wear, Diego Unzueta Ruedas, Gerasimos Marios Vrazalis, Panpan Wu

1 Introduction

Walmart (US) is the biggest retailer in the world by turnover, with a global revenue of $524 billion. [1,2]. With an organisation of this scale, even the smallest percent increase in sales could have immense consequences. Therefore, the understanding of their products and sales is critical. Savings in this field are directly in line with Walmart's goal to "save people money so they can live better". However, with over 12,000 stores and 200,000 products in supercenters, it is difficult to estimate demand trends on individual products [3,4].

The overall goal of this project is to generate predictive models capable of predicting future sales of Walmart items. This report will focus on the data analysis and preprocessing prior to modeling stages.

We utilise the M5 forecasting dataset, a publicly available dataset consisting of sales data of over 30,000 Walmart products [5]. In this report, first the raw data structure is analysed. Then the methodology used in data cleaning and feature extraction is discussed. Finally, some data analysis of the data is presented, analysing the features extracted as well as other elements of the data. If successful, the data preprocessing techniques discussed in this report could be utilised to generate inputs for models estimating demand in real stores.

2 Data Structure

The Kaggle M5 forecasting dataset consists of over 30,000 products sold in Walmart stores across the US, and their sales data over six years. The dataset consists of three primary documents. These are the calendar, the sales data, and the sell price data.

2.1 Calendar

The calendar data consists of information about the days the sales data were recorded. The information about the specific day the sale happened could affect the amount of sales of all products sold that day. The data of each day contains binary indicators on whether events were held on that day, such as religious events, the SuperBowl, Father's day, as well as individual offers or reductions in price for that specific item that day (labeled as SNAP).

2.2 Sales Data

The sales data contains information about the item sold (shown as a product ID, rather than a name), as well as the number of sales of that product everyday across a timespan of over six years. In this dataset, each row indicates a product, and each column contains information about the product. This information involves the department ID that the product belonged to, the store ID it was sold at, the state it was sold at, the number of sales of that product each day and more.

2.3 Sell Price Data

The sale price data contains the price at which every product was sold at throughout the entire time the number of sales data was recorded. This dataset can be joined with the sales data to then perform data analysis on the effect of price on number of sales.

Overall, this dataset is very complete and contains a massive amount of information about the products and their sales over a very large timespan.

3 Data Cleaning

The raw sales data as given shows every day a product was sold for as a feature (column) and every item as an observation (row). While this is a memory efficient way of storing the data, the data must be converted into a format that machine learning algorithms can work with. Specifically, we need to have the number of sales as a single column (target variable) and every observation corresponding to a single item on a single day. Restructuring data in this tabular format enables this time series data to become a supervised learning problem.

Sales data for which there is no corresponding price for that product on the given day was removed since these observations correspond to products that were not on sale at all. (ADD DIMENSIONALITY)

4 Analysis

The objective of this section is to analyse the way each feature affects the number of sales. In particular, the number of sales by item, store and category are examined. In addition, the effect of special events on the number of daily sales is analysed as well as the behavior of sales regarding the sale price of the products.

4.1 The Distribution of Sales

Understanding the overall distribution of sales data is important prior to analysing the effect of other features on total sales.

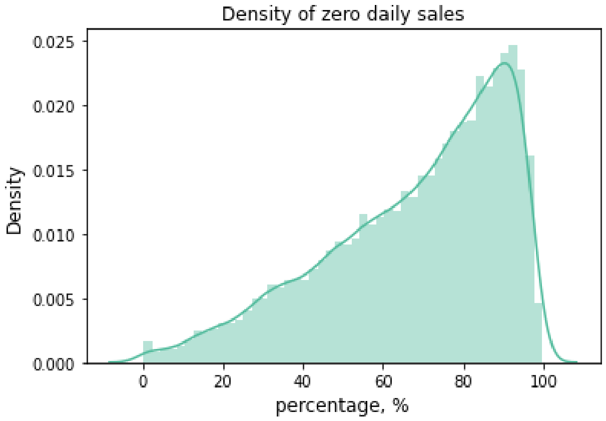


Figure 1. Density of zero sale days as a percentage of total sale days for all products in dataset. The plot is skewed towards higher percentages of zero sale days.

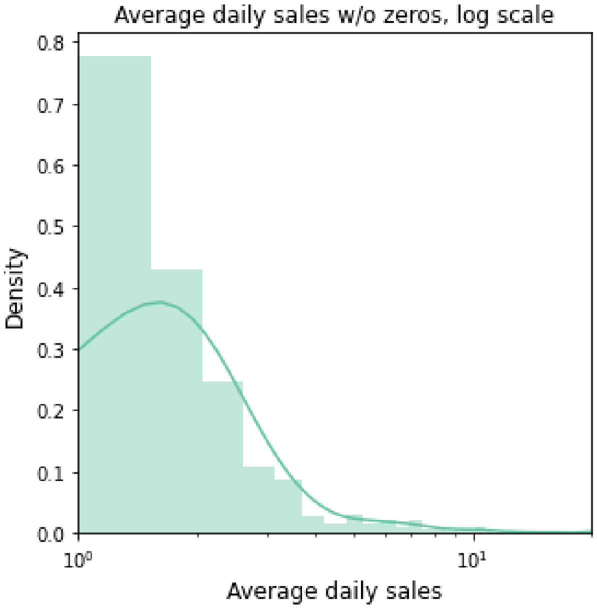


Figure 2. Distribution of product average daily sales. The figure is skewed towards low values of average daily sales.

Figure 1 shows the distribution of the fraction of days that have zero sales and Figure 2 represents the distribution of the average daily sales excluding zero sales for each item. From these figures, it can be noted that most of the items have zero daily sales and the average daily sales mostly lies between 1 and 2. This implies that most days, most items had no sales. When modeling, it will be important to ensure models are not always predicting zero sales. From figure 3, periodic spikes in the sales data are noticeable. These all happen on christmas eve, where all stores close and therefore the total sales of that day each year is zero.

When looking at individual product average sales, it can be seen that different items have different sale averages, this could be due to their specific product characteristics. Moreover, it can be noted that most products have averages between one and two, whilst there are some products with much higher averages.

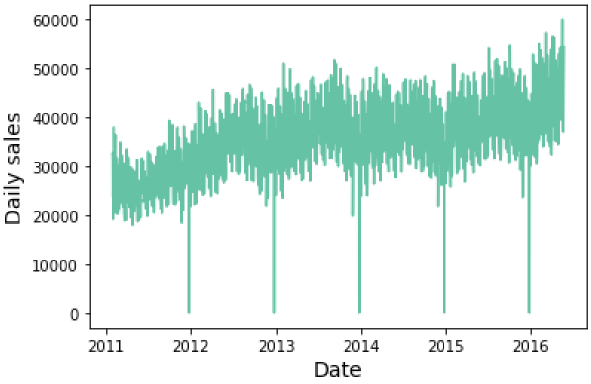


Figure 3. Total number of sales plotted against time. Significant drops in sales can be seen periodically. The overall trend is an increasing number of sales.

4.2 Total Sales by Store/Department

It can be seen from Figures 4 and 5 that the sales vary greatly in both stores and departments. Specifically, the second store in Wisconsin and the third food department have the highest sales. Food is necessary and must be purchased frequently, this could be why the average number of food products are higher than those categorised as household products.

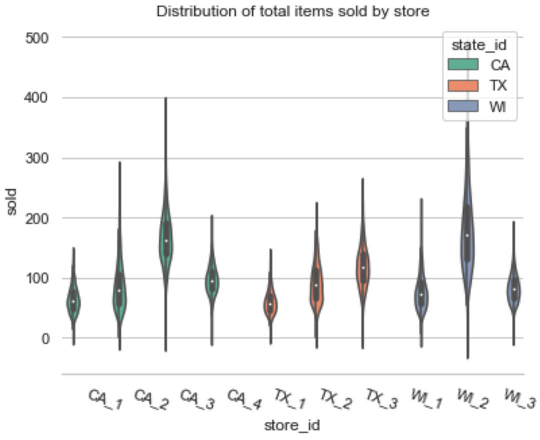


Figure 4. Violin plots of sales data by store for all products. Each store can be seen to have different distributions.

4.3 Average Daily Sales by Category

Figure 6 describes the trend in the average daily sales between the three product departments in the dataset, between 2011 and 2014. The food department has consistently achieved the highest daily sales while the daily sales of the hobbies department remains constant over the years. We can see that there is a natural fluctuation in sales arising from weekend sales being

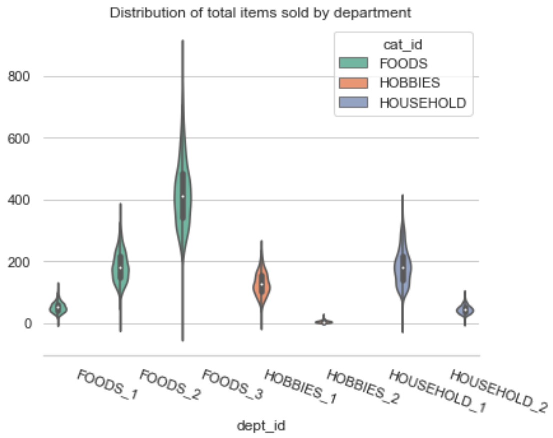


Figure 5. Violin plots of sales data by department. Each department has different distributions of sales data.

higher on average than on a weekday. It's also noticeable that there is a significant decrease in the sales on Christmas day. In fact, no sales are recorded on any Christmas day. In Section 4.5 we analyse whether there are other events that affect the number of daily sales.

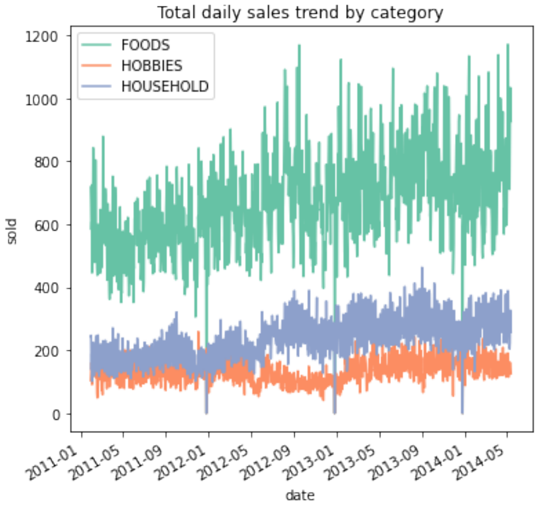


Figure 6. Number of sales made as a function of time. Here, products are grouped by their department name.

Figure 7 analyses the average number of sales by weekday per department. We can see a clear weekly pattern between all the three departments. This figure shows that there are higher average sales on weekends than weekdays and this is what gives rise to sales fluctuations shown in Figure 6.

Figure 8 shows the effect of SNAP events on total average sales. It appears that on days with SNAP, average sales increase. We can also see that there is some evidence to show that SNAP days influence sales more in Wisconsin (WI) than in Texas (TX) or California (CA).

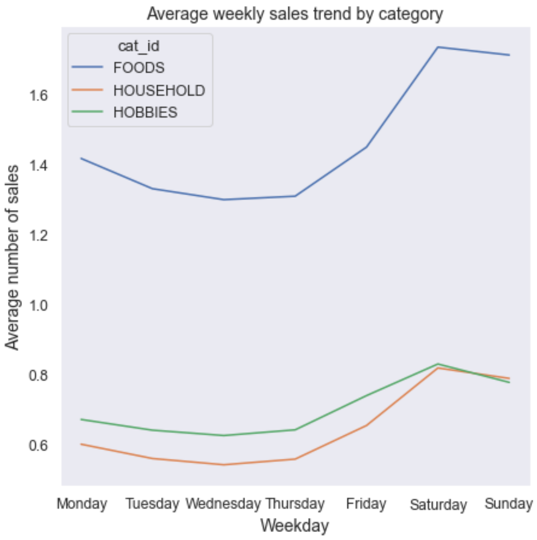


Figure 7. Average number of sales by weekday. Items are grouped by their department name.

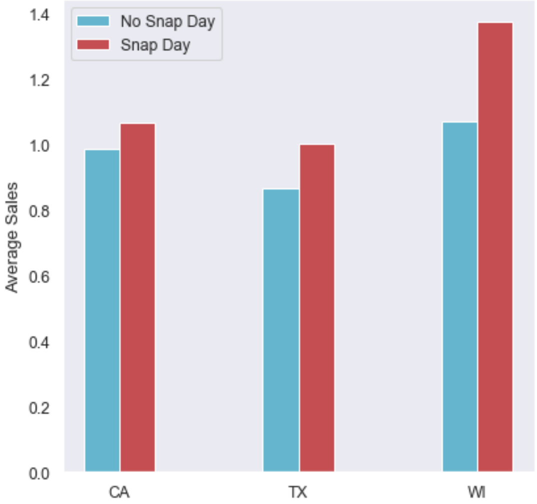


Figure 8. The average sales between normal days and days that are allowed to use SNAP benefits (price discounts).

4.4 Event Analysis

The calendar data provide information regarding the occurrence of any cultural, national, religious or sporting event on a certain day. It is interesting to observe to what extent the presence of such an event affects the number of the daily sales. The most noticeable event is Christmas where there are no sales since it is the only day in the calendar year that Walmart is closed. An important decrease in the daily sales can also be observed on Thanksgiving and New Year's Day due to the limited working hours on these days, figure 10 (https://www.storebusinesshours.com/walmart-hours/). On the contrary, every year from 2011 to 2016, on Superbowl and Labor's day, an increase in sales above the average is recorded with Walmart reaching on these days almost 40 thousand daily sales. The eve of each event was also added in the calendar dataset. By comparing figure 10 and figure 9, it can be observed that the daily sales on event's eve are broadly higher than the ones on the day of the event itself. Especially on the previous day of events such as Superbowl, President's Day, Valentine's Day, Easter, and Mother's Day, every year the daily sales surpass the average number by almost 5 thousand sales.

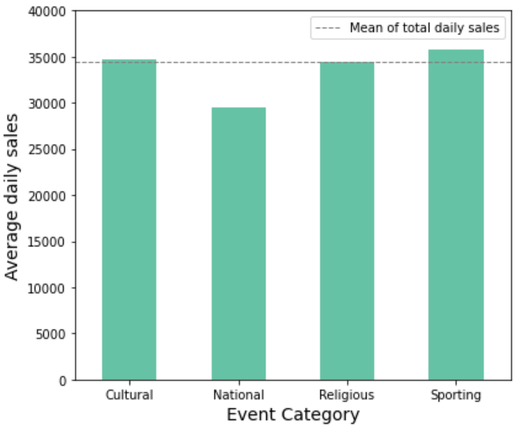
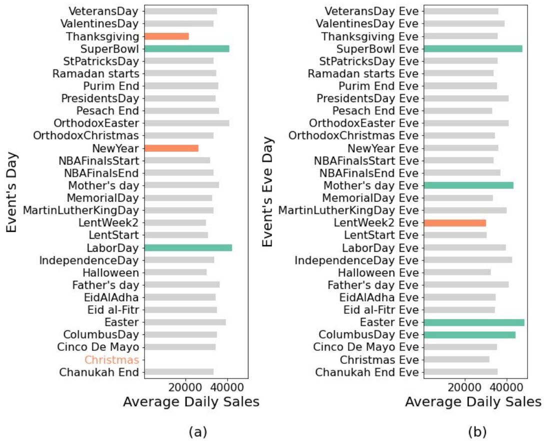


Figure 9. Average daily sales for each one of the four event categories. Cultural, religious and sporting events have almost the same number of daily sales on average with a number close to the total average. National events appear to have the lowest record of daily sales with almost 30 thousand on average.



(a)

(b)

Figure 10. The two bar charts (a) and (b) represent the average daily sales recorded over the 6-years period on days with events and on events' eves respectively. In many cases the average sales on these days differ from the total average daily sales (almost 35 thousand). However, only the days in green and orange shades appeared to have a steady behavior of either increased or decreased daily sales respectively over the years and could possibly be more determinant in predicting future sales on these days.

4.5 Sale Price Analysis

The histogram in Figure 11 shows that the vast majority of items sell for less than $10 and that the distribution of sell prices are skewed towards zero. The log sell price may be a more appropriate model feature because its distribution better fits the Normal distribution.

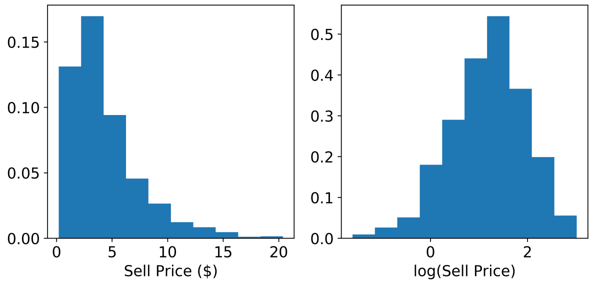


Figure 11. Distribution of sell price and the log sell price.

4.6 Feature Importance

The analysis section discussed the effect of all the data features on average total sales. In order to get an idea of what are the most predictive features already given in the data set, decision trees and linear regression were used to rank features. The top 10 features ranked by importance from both algorithms are given in Figure 12.

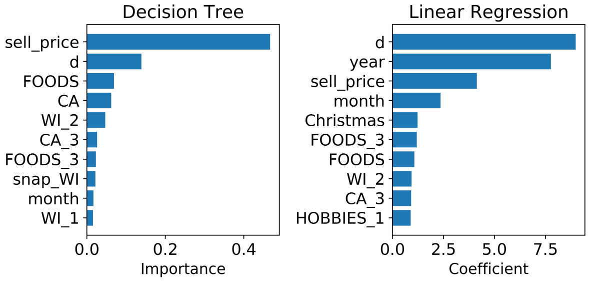


Figure 12. Top 10 most important features as ranked by the decision tree and linear regression.

The implementation of decision trees allows us to rank each feature by their importance. This algorithm shows that sell price, day of the week and FOODS are the most important features. These features also appear in the top 10 most important features as ranked by linear regression.

For linear regression, the absolute value of the coefficient of each feature is used as a ranking. The linear regression algorithm seems to give more weight to time series features such as d, year and month. It also recognises Christmas as a significant feature, since on Christmas day no sales are recorded.

These results are coherent with our findings in the analysis, as these features were found to have a substantial impact on the number of sales.

5 Conclusion

This report proposes a data analysis methodology on the M5 dataset that could later be used prior to the prediction of future sales of products in Walmart. Throughout the analysis, methods were discussed on appropriate ways of cleaning the data, as well as the statistical significance of some of the data provided. Some of the findings involved the importance of product categories, and variations in the distribution of sales in different weekdays. Moreover, feature extraction was discussed, enabling the extraction of even more relevant information from the given data. Some of the proposed features involved the use of event eve data, instead of the event day. The importance of features was extracted using decision trees and linear regression. The results from these are coherent with the data analysis, finding that these selected features can have a substantial effect on sale numbers.