M5 Walmart: Price Prediction Analysis

FA 2023 ECON 491 Final Report

Nimish Mathur
nimishm2@illinois.edu

Randall De Vera
rdevera2@illinois.edu

Faayez Akhtar
faayeza2@illinois.edu

Alexander Asis
aasisi2@illinois.edu

1 INTRODUCTION

Problem. With the current growth of big data in today's modern society, the necessity for efficient data processing is a cardinal requisite for deriving insightful analyses. Retail companies like Walmart are at the forefront of big data analytics, seeking to decipher and better identify retail sale trends to keep the lowest competitive prices in the market. As part, the capacity to predict future pricing is a tedious task based on a thorough examination of multiple contributing elements in addition to historical trends. Understanding the subtleties of how holidays, regional differences between states, and the variety of retail item categories can affect sales and, in turn, forecast future prices, is crucial. Walmart can better meet customer demand and gain a competitive edge in the highly competitive retail industry by combining knowledge from these dimensions to optimize pricing, supply chain logistics, and inventory management.

Background. To better leverage Walmart sales data to understand future price prediction, the Kaggle M5 - Forecasting dataset was used. As part of a previous competition, this dataset provides raw Walmart's sales data over the period of 5 years with the challenge of forecasting 28 days of future sales in different locations. This paper attempts to use a similar framework using an ARIMA regression model to forecast future sales. To do this, let's first analyze the dataset:

(1) **calendar.csv** - Provides insight into the 5 year period in which sales data is provided from April 2011 to April 2016. This also includes descriptions of specific holidays like Easter, Christmas, etc. that may have an impact on sales data. Calendar.csv also contains binary information about if that states Walmart's allow SNAP purchases. This government funded program could contribute to larger unit sales on certain days.
(2) **sales_train_validation.csv** - Contains daily sales data from d_1 to d_1913. This includes state information and product types (hobbies, food, and household). For simplicity of this model, the state information provided is limited to Texas, California, and Wisconsin.
(3) **sell_prices.csv** - Sell prices is a unique data set that contains information about the price point of each sold unit based on the calendar date and State ID.
(4) **weights_validation.csv** - Provides insight into the differences in pricing weights based on the state location and unit type.
(5) **sales_test_evaluation.csv** - This final data set contains the price predictions for the final 28 days. This will be used to test the accuracy of the ARIMA model forecasted results.

In a process to clean and understand the data, the above data sets were combined using a series of query joins on common fields to create a master data set. With this, alignment of calendar dates to sales data is possible.

Proposed Approach (Summary). Understanding the data in its entirety from a first glance can be very tedious. As part, this experimentation used a revised exploratory data analysis approach to identify initial trends and potential errors in the code. The plan below articulates the exact approach used to create our final model:

(1) Retrieved data from the M5 Forecasting Kaggle data set.
(2) Conducted an exploratory data analysis (EDA) to identify initial sales trends related to day of week, major holidays, and differences in unit sales between different product categories.
(3) In addition to the EDA, an analysis was also conducted to confirm and better synthesize the distributions found.
(4) Once a solid understanding of the data and its initial trends had been established, general research was conducted to understand what regression analysis models would best work to predict future sales. As part, research into different models including linear regressions, Random Forest, ARIMA, and Holt-Winters were conducted.
(5) From this analysis, different simple regression models were tested including linear and ARIMA in an attempt to identify which model would best summarize the data. This was plotted against the sales_test_evaluation.csv data which included the actual forecasted data.
(6) Built a conclusive summary of the simple and final regression models, the included dummy variables, and a comprehensive analysis of ARIMA's forecast in comparison to the actual data.

Packages Included. During the course of experimentation, 3 key R-programming packages were used:

(1) **Tidyverse** - Package used to optimize data science related analysis including reading and processing data.
(2) **gridExtra** - Provides a better command and console to plot graphs in a clean and concise manner.
(3) **Forecast** - Used to check the accuracy of residuals in the forecast.

Combined, these packages allowed for key, streamlined data analysis important for forecasting the final regressions.

2 DATA GRAPHS

Conducting an exploratory data analysis provided key insight into initial trends within the sales data. As aforementioned, there are

3 categories of products that Walmart breaks their sales data into. These categories include hobbies, food, and household items. With this knowledge in mind, the first analysis made revolved around understanding how much of the total sales come from each category:

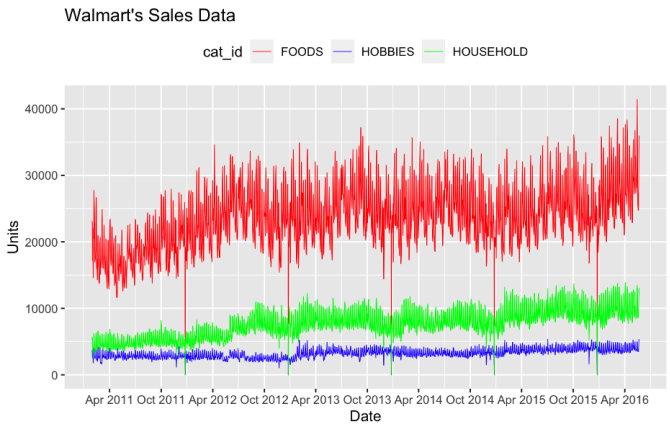

Figure 1: Unit sales by category

From Figure 1, it can easily be identified that food has the greatest number of sales with household and hobbies falling far behind. This concept from a high level makes sense since food is a necessary purchase done on a regular basis. From Figure 1, it can also be seen that all 3 categories follow a trend of rising unit sales from Monday-Saturday and dropping off to its lowest on Sunday. Lastly, it can also be seen that the hobby category has a relatively consistent sales year round.

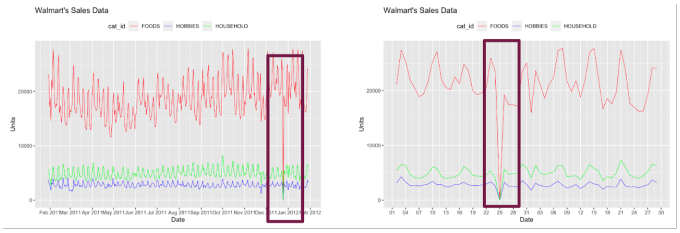

Figure 2: Yearly holiday outlier

Figure 2 focuses on identifying the drastic decrease in sales each year during the December timeframe. After careful analysis, it was concluded that the drop occurs annually on December 25th and can be explained since Walmart is closed nationwide in honor of Christmas. This is the only holiday where Walmart is closed.

Very similar to the last plot, Figure 3 just groups unit sales for each category by week. This makes it a little easier to show that Food and Household sales have grown over the period of 2011 to 2016. Hobbies have remained rather constant.

Analyzing the foods category further, the boxplot for Foods has a mean selling of units around 22,500 units. The lower side of units sold is 0 which makes sense for days like Christmas when the store is closed. The higher side of units sold is 40,000+ which can be

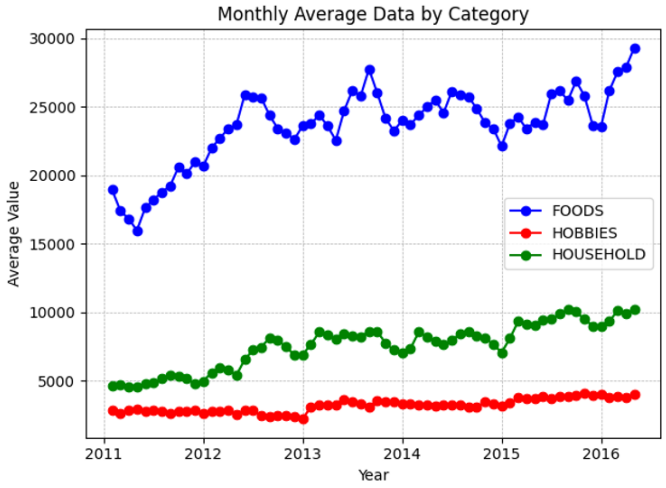

Figure 3: Average Unit Sales by Category

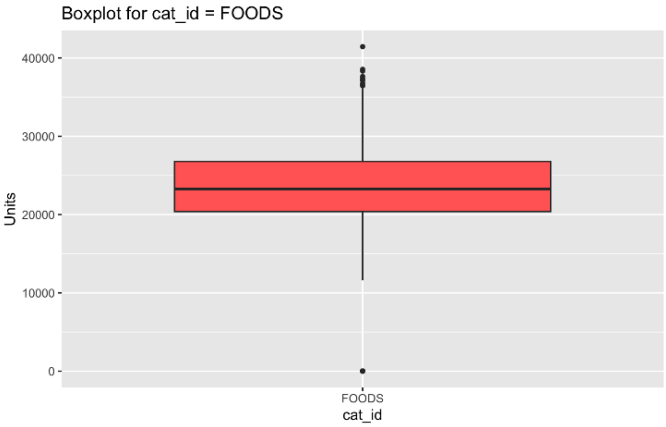

Figure 4: Food Sales Boxplot

attributed by bigger sales days. Maybe further analysis is necessary into big sales, natural disasters, etc. Similar box plots were created for both Hobbies and Household and the same trends exist.

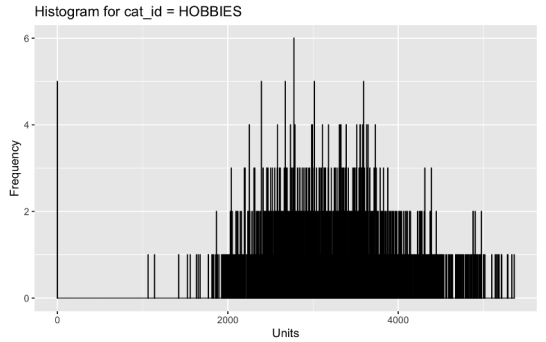

Figure 5: Hobbies Histogram

Figure 5 shows the frequency in which the number of units sold on any given day were the same. Frequencies of 0 units sold can be shown accurate by holidays when Walmart is closed (Christmas). Other days with high frequency sales too can show the peaks and lower sides of units sold with peaks in the middle near the mean. All 3 categories of items sold at Walmart follow a bell-curve, unimodal plot.

From this exploratory data analysis, a better understanding of the data can be established. In creating these plots, proper merging of the different data sets were conducted to create an all-inclusive master table.

3 DIFFERENCES FROM EXISTING WORK

Understanding the Walmart dataset in further detail was important for identifying initial trends. In addition to the analysis conducted in this experiment, it is also important to understand works done by other people in the same topic. In an attempt to do so, let's highlight four papers that assisted our understanding:

Before developing the simple regression model or any other model, our group had to explore the data and its context in greater detail. Makridakis and Spiliotis quantified the objective of M5 competition was to produce the most accurate point forecast for 42,840 time series that represent the hierarchical unit sales of the largest retail company in the world by revenue, Walmart. The data set involves the unit sales of 3049 products, classified into three product categories (Hobbies, Foods, and Households), and seven product departments. This helped better our understanding when attempting to identify the groupings of data for our final forecast. Seeing the data was split up by state and category helped us to narrow down the resulting accuracy points. From this paper, we were also able to further confirm that the products were sold across 10 stores located in three states: California, Texas, and Wisconsin. The data was collected daily and covers the period from 2011-01-29 to 2016-06-19 which is divided into training set, validation set, and test set. The data set also includes exogenous variables e.g. special days, selling price, and SNAP activities (promotions activities). We were able to easily use this information to create dummy variables for our regression.

In this project, our aim is to forecast Walmart's sales using the dataset and structure provided by the M5 Competition. The M5 "accuracy" competition delineates various aggregate levels for grouping sales data. Our team has opted for a level six aggregation, focusing on "total sales of a retail category, a channel, or the whole industry in a country or region" (Fildes, Ma, & Kolassa, 2022). In our forecast context, this entails aggregating and predicting data at a state and category level.

Moreover, we've concluded that working with a **level 6 aggregate** would be the most suitable for our objectives in terms of simplicity compared to the intricacies involved in product-level sales analysis. Our decision aligns with the recommendations made by the authors of "Retail Forecasting: Research and Practice" in the International Journal of Forecasting. They advocated for employing simple exponential smoothing and ARIMA models to effectively capture "strong trends, seasonal variations, serial correlation, and regime shifts" within the data (Fildes, Ma, & Kolassa, 2022).

Additionally, the paper raised doubts about the efficacy of more complex models like ARIMA. Thus, our initial approach will involve using a simple linear regression model, aligning with the authors' suggestions while acknowledging the need to assess the viability of more sophisticated methodologies.

After determining aggregation level, the paper aided greatly in evaluating the accuracy of our model. It also played a role in the decision of creating multiple simple models as opposed to just one that measured the impact of average price on the sales of Walmart's goods. With respect to the former, our team used RMSE as a metric for evaluating the accuracy of our model. This is substantially advocated in the paper "M5 accuracy competition, results, findings, and conclusions" as the best method for accuracy assessment.

With respect to the latter aspect of the decision of creating multiple models; the paper heavily emphasized the importance of creating multiple models due to the possibility of cross-learning across models and obtaining the best predictions. Hence, our team decided to use data from different independent variables to create numerous models and observed the impact on Walmart's product sales from each. This allowed us to understand which variables had the greatest impact, and consequently tweak our predictions to ensure accuracy.

The main findings of the paper indicate three new findings with respect to forecasting. First, the paper argues that LightGBM method, a decision tree machine learning algorithm, is an accurate way to make predictions. Second, the authors contend that making external adjustments to some of the methods is important to increase the model's accuracy. Lastly, the paper found that the explanatory variables are important in time series and its accurate forecasting. In addition to the three new findings, the paper reiterated the importance of combining, "cross-learning", and cross-validation in improving the accuracy of the forecasting model.

For the other model, we consider utilizing time series analysis as a tool to predict the sales unit. The Walmart datasets show the pattern of time series datasets, so we believe the use of time series analysis method would be appropriate. Vyas and AS (2022) compared two approaches in their study which are regression model and time-series model. Based on the result, the study showed the time-series model performed better than the regression model. To be more precise, in the time-series model, ARIMA method gave a lower RMSE value than Holt-Winters method. So, we decided to pick the ARIMA method as our other model.

However, our result showed we should incorporate the seasonality effect in the ARIMA method, so it became the Seasonal ARIMA method. We had tried ARIMA, but after we accommodated the seasonality, the SARIMA model showed much better prediction. Furthermore, our result shows that the SARIMA method was not always the better method compared to the regression method. We have nine models based on the location of the store, and for two models (Food in Texas and Food in Wisconsin), the regression model showed lower RMSE.

4 DATA PREPARATION

The econometric model we chose for our simple model was a linear regression. We created multiple simple linear regression models with different sets of independent variables. The variables we used

were: price, weekdays, holidays/special events, and the SNAP program. This section of the report will contain information on every linear model we constructed and the changes made to the dataset prior model building.

Before we began to write the code to construct our model, we had to prepare the data given to us to obtain the dataset that we would use to train our various models. This dataset had to include information on dependent variables (sales) and all the independent variables we planned to use for our various models. Since all of the independent variables we planned on using were distributed across the datasets provided to us, we had to combine the datasets to obtain our final training data. The datasets provided were named: s**ell_prices.csv, calendar.csv, sales_train.csv, sales_test.csv, weights_validation.csv.** The following steps were used to prepare the data:

(1) Edit **calendar.csv:** We filtered out any empty values in **calendar.csv** (named **calendar** in the code) and narrowed the dataset down to columns relevant to our analysis.
Before moving on to the next step, it is important to understand a prediction decision we made prior to analysis - We decided that observing trends for every single product may hinder the accuracy of our report since there are numerous variables that could affect the demand for a single product. For example, shortage of supply/ingredients and price fluctuations of complements and substitute goods. Accounting for all of these factors for every single good would be an impossible task and make the models unnecessarily complicated. Hence, we decided to build our models from the data set aggregated to the state and category levels (level 6 aggregation). This would be done via the groupby() function.
(2) Merging **sales_train (sales1)** and **sales_test (sales2)**: We used three of the sales datasets provided to us for predictions. The third dataset had 28 values and we decided to use it as our test dataset. **sales1** and **sales2** refer to the first two datasets. We combined these 2 datasets to create the first part of our training dataset: **sales**. After creating **sales**, we realized an issue in the way the data was formatted. The days were columns and the sales for every individual item were organized as rows. This would make it extremely hard to aggregate the data by day and reach the level of aggregation we needed. To remedy this issue, we pivoted the dataset so that the columns became the rows and the rows became the columns. Doing so allowed us to group the data by day and aggregate it to our desired level.
(3) Merge edited **calendar** and **sales**: Our training dataset so far (sales) did not have information on sales by day yet. Since one of our models was going to use the days of the week to predict sales, we had to merge **calendar** and **sales**. The only way to merge the two was through the common 'date' column that they had. However, the format of the date column in calendar was different from the one in **sales**. Therefore, we first changed the format of the date column in calendar and then merged the two on that column to create a new dataset called cal/_sales.
(4) Merge **prices (sell_prices.csv)** and **cal_sales**: At this point we had almost everything we needed in our training dataset to carry out the regression. Since one of the models we were going to run was a linear regression model based on price, the last step was to add the price columns from prices to cal_sales to finally complete our training data set with all the necessary variables for all the models we planned on creating. However, we came across a hurdle at this final step. Prices and cal_sales did not have any common columns to merge on. To overcome this issue, we split the values in the column "item_id" in prices to create three separate columns, one of which was identical to the "cat_id" in cal_sales. Subsequently, we renamed the split column that was identical to "cat_id" in cal_sales to have the same name and merged the two data sets on that column.

5 SIMPLE MODEL

Introduction. Let's introduce Ordinary Least Squares (OLS). OLS is a fundamental statistical method used for estimating the parameters in a linear regression model. The OLS simple regression model focuses on the relationship between a single independent variable and a dependent variable by fitting a linear equation to observed data.

Using a simple line of best fit model, our goal was to forecast the future sales split by category and state.

Econometric Model. The econometric model we used for the "Simple Model" was the OLS regression model as shown below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + U$$

Where:

- Y = dependent/outcome variable
- $\beta_1, \beta_2, \ldots, \beta_K$ = coefficients of respective independent variables $X_1, X_2, \ldots, X_K$
- β_0 = intercept
- U = unobserved random variable

Analysis of Independent and Dependent Variables. We constructed multiple linear models with varying independent variables to investigate which variables have the greatest impact on sales units (outcome variable), and to elucidate if any of the variables had a greater or lesser impact than previously perceived. For all our models, the dependent/outcome variable was sales units for the various goods offered by Walmart. The independent variables and their respective equations were as follows:

- Price
$$Y = \beta_0 + \beta_1 X_1 + U$$
Where: $X_1 = w.avg\_price$
- Event type (National, Religious, Cultural, Sporting)
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$
Where $X_1 = is.National, X_2 = is.Religious$, and $X_3 = is.Cultural$
- All days of the week

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + U$$

Where $X_1 = is.Saturday, X_2 = is.Sunday, X_3 = is.Tuesday, X_4 = is.Wednesday, X_5 = is.Thursday$, and $X_6 = is.Friday$

- Snap programs

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

Where $X_1 = snap\_cr$, $X_2 = snap\_trx$, and $X_3 = snap\_w$

After creating the four models above, we created one final model that used all of the above variables in one model. This is shown below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + U \quad (1)$$

Where X_1 is.isSaturday, X_2 is.isSunday, X_3 is.isTuesday, X_4 is.isWednesday, X_5 is.isThursday, X_6 is.isFriday, X_7 is w.avg_price, X_8 is $snap$ cap, X_9 is $snap$ w, X_{10} is.isNational, X_{11} is.isReligious, and X_{12} is.isCultural.

After creating the final econometric model, we aggregated the data to their respective category and state (nine combinations) and altered the final econometric models to reflect that. The main difference between the three models below is the SNAP variable. We have decided to only include the respective SNAP variable for that specific state.

For Foods and California, Hobbies and California, Household and California:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + U \quad (2)$$

(And similar for the others, just changing the SNAP variable accordingly)

IV and DV Notes. In the independent variables: is_Monday and is_Sporting variables was our attempt to mitigate multi-collinearity which inflates the standard errors of the coefficients; thus, making the coefficients less precise and potentially leading to incorrect results. As a result, the effects of these two variables are implicitly considered within the intercept term.

Explanation of OLS Method. The Ordinary Least Squares (OLS) method is a statistical technique used for estimating values for the outcome variable through a linear regression model. The goal of OLS is to find the best-fitting straight line through a set of points in a manner that minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the linear model.

Here's a step-by-step breakdown of how OLS works:

- **Function:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + U$$

- **Theory:** The OLS method aims to minimize the sum of the squares of the residuals. The residual for each observation is the difference between the observed value of the dependent variable and the value predicted by the model.
- **Minimization:** OLS minimizes the residuals by taking partial derivatives of the function (shown above) with respect to each coefficient and setting them to zero. Then, these equations are solved for the values of the coefficients that minimize the function.
- **Parameter Estimation:** After partially derivative and equating to zero, the resulting equations are a set of normal equations that can be solved to provide the estimates of the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$.
- **Model Fitting:** Once the parameters are estimated, the fitted model can be used to predict the dependent variable Y for given values of independent variables $X_1, X_2, \ldots, X_K$.
- **Assumptions:** OLS relies on several key assumptions, such as linearity, independence, homoscedasticity (constant variance of errors), and normality of error terms. Violation of these assumptions can lead to biased or inefficient estimates.

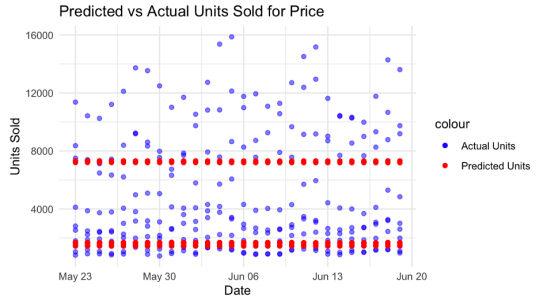

Figure 6: Predicted vs Actual Units Sold for Price

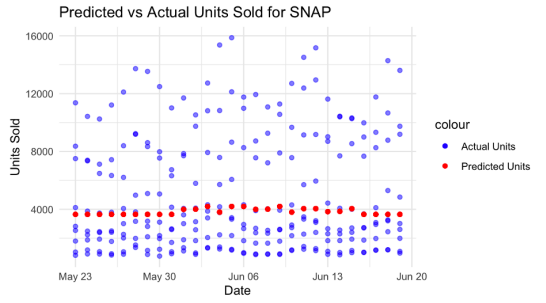

Figure 7: Predicted vs Actual Units Sold for SNAP

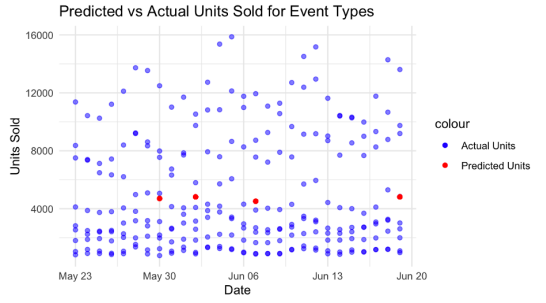

Figure 8: Predicted vs Actual Units Sold for Event Types

Graphs of Forecast vs Actual Data.

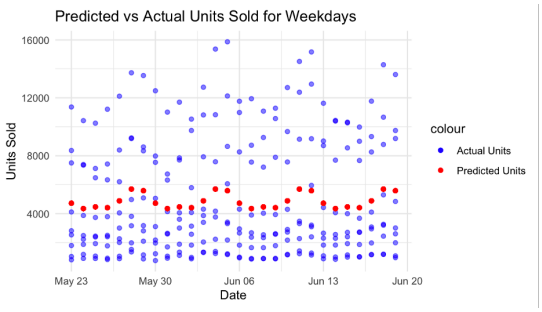

Figure 9: Predicted vs Actual Units Sold for Weekdays

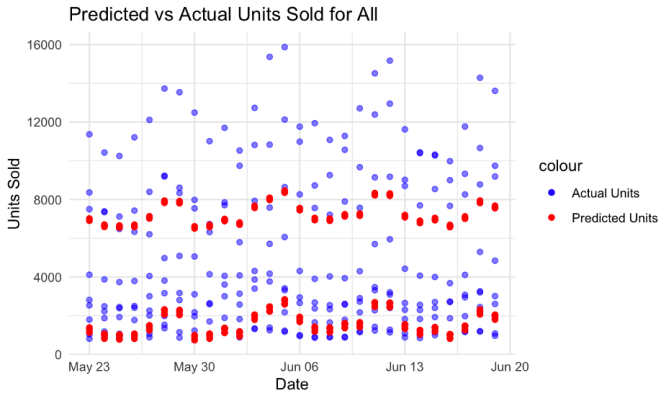

Figure 10: Predicted vs Actual Units Sold for Price, SNAP, Event Type, and Weekdays

Table 1: Your table caption.

Figure	Variable	RMSE
6	Price	1889.451
7	SNAP	3294.912
8	Event Type	3797.791
9	Weekdays	3863.28
10	Price, SNAP, Event Type, Weekdays	1802.947

Root Mean Square Error (RMSE).

Description of Results. We based the result on the Root Mean Square Error and the graphs. To begin, we evaluated each independent variable in the econometric model. Each explanatory variable indicates high RMSE values which means the model fails to accurately predict the outcome (Table 1). This is further proven by Figure 6 - 8 which fails to show variability in the predicted values. Only the explanatory variable, Weekdays, shows some variability or trend which Figure 9 illustrates. We hypothesized that this was due to the fact that more people buy on certain days particularly the weekends.

The independent variable, Event Type, has the highest RMSE value (3797.791). We believe that this may be caused by a limited amount of data or occurrences within the Event Type category,

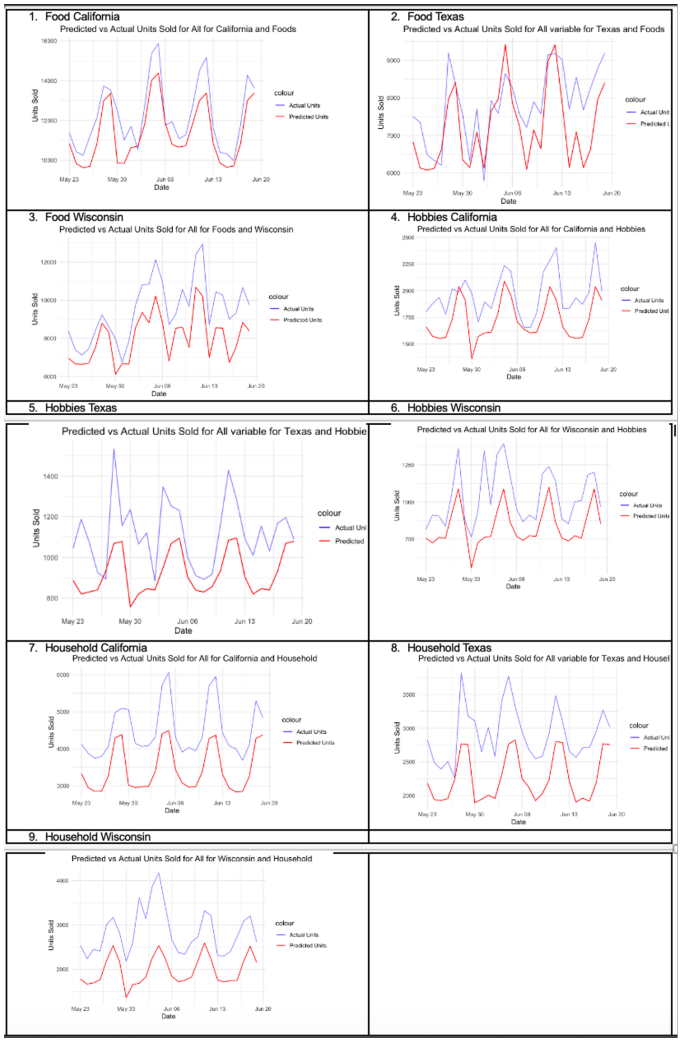

Figure 11: Level 6: Actual vs. Forecasted Sales for Simple Model

Category	State	RMSE
Food	California	1313.375
Household	California	636.8922
Hobbies	California	269.2346
Food	Wisconsin	982.0176
Household	Wisconsin	472.6586
Hobbies	Wisconsin	141.7716
Food	Texas	609.3604
Household	Texas	487.1552
Hobbies	Texas	144.1923

Table 2: RMSE by Category and State

leading to a higher RMSE. In contrast, the explanatory variable, Price, has the lowest RMSE value of 1889.451. We believe that the

abundance of data or variance within the Price variable likely contributes to its lower RMSE value, indicating that it's more closely aligned with the predicted outcome compared to the other variables in the model. When all variables are evaluated, the RMSE value is 1802.947 according to our model which is an improvement compared to RMSE of each independent variable.

Because we aggregated our data at the state and category level, we evaluated the nine possible combinations of the data. As Table 2 illustrates, when the data was aggregated to their respective state and category levels, the simple model improved tremendously. The graphs portrays this since the forecasted value shows quite good fluctuations that closely mirror the trends of actual values. However, we can see some graphs are not good at prediction since they seem to underestimate the predicted value and their respective trends. This is evident in Figure 14 where the predicted values for Wisconsin and Household closely mirror the trend of the actual values but the trend line is below the actual value's trend line.

Regardless of state, the category, Hobbies, seems to perform the best when compared to the other categories. We assume that the Hobbies category consistently demonstrates more predictable patterns or a stronger correlation with the outcome variable across different states, resulting in more reliable predictions compared to other categories within the dataset. Additionally, there are less items within this category which would potentially lead to clearer and more distinct trends or relationships between the items within the Hobbies category and the outcome variable. This smaller set of items might contribute to a more focused and less complex analysis, allowing for a more accurate understanding of how these specific items within Hobbies influence the predicted outcomes. In contrast, the Food category, regardless of state, seems to have the highest RMSE. We assume that the higher number of items in this category contributes to the higher RMSEs since a larger number of items in the Food category could lead to increased variability and complexity in predicting outcomes. The diverse range of food items might introduce more nuanced and diverse factors influencing the outcome variable, making it harder to accurately model and predict. This broader scope within the Food category might result in a wider array of influences, thereby increasing the margin of error in predictions and leading to higher RMSE values.

6 OTHER MODEL

Introduction. After we see the result of the simple regression model, we consider examining another method. Because of the type of the data is time series, we decide to apply a time series method to treat the data. The method is Seasonal Autoregressive Integrated Moving Average or SARIMA.

Econometric Model. This is SARIMA$(1,1,1)(1,1,1)_s$

$$(1-\phi_1 B)(1-\Phi_1 B^s)(1-B)(1-B^s)Y_t = (1+\theta_1 B)(1-\Theta_1 B^s) \quad (3)$$

where:

- Y_t = the time series variable
- B = back notation whose power show the lagged variable of Y_t
- ϕ = the nonseasonal autoregressive coefficient
- Φ = the seasonal autoregressive coefficient
- θ = the nonseasonal moving average coefficient
- Θ = the seasonal moving average coefficient
- s = the seasonal period

Because we consider to take level 6 of aggregate of Walmart data, which is category-state level, we came up with 9 different models. Overall, the parameters is SARIMA(5,0,1)(3,3,3)7, but some model has little different parameters. The model is described in Table 1.

Independent and Dependent Variables. In SARIMA method, there are no independent and dependent variables as in a regression method. However, we can identify some aspects in SARIMA model and consider those as independent and dependent variables.

The dependent variable in SARIMA model above is the unit sales in each day. The independent variables are the lagged unit sales of the data itself, the lagged forecast errors, and the differenced unit sales values. The logic behind this is that the future values of unit sales are presumed to be influenced by its past values and the past errors of prediction.

Explanation of ARIMA Method. ARIMA basically is consisted of three parts, which are:

(1) **Seasonal (S):** Seasonal here means this is an extension of the ARIMA method. This extension has purpose to address specifically seasonal variations in time series data. Seasonality refers to regular, predictable patterns that repeat over a calendar year or any fixed period.
(2) **Autoregressive (AR):** Autoregressive part means there is a relationship between an observation (unit sales in a certain day in Walmart case) and a certain number of lagged observation (unit sales in day before the certain day in Walmart case. The idea is the current values are influenced by the past values.
(3) **Integrated (I):** Integrated refers to the differencing action of raw observation (all unit sales) to make the time series data stationary. Stationery is the condition where the statistical properties of the data, which are mean, variance, auto-correlation, etc., are constant over time. This differencing reduces or eliminates trends and seasonality in the data.
(4) **Moving Average (MA):** This part captures the relationship between an observation (unit sales) and a residual error from a moving average model applied to lagged observation. The purpose of this part is to smooth out fluctuations in the data.

These three parts control the model structure through 7 parameters as follows:

(1) p (AR part): show the number of lag observation in the model.
(2) d (I part): show the how many times the raw observation are differenced.
(3) q (MA part): show the size of moving average window.
(4) P : seasonal lag observation.
(5) D: seasonal differencing.
(6) Q: seasonal size of moving average window.
(7) s: show the length of the seasonal cycle.

Root Mean Square Error (RMSE). The following table holds the RMSE for Level 6 data, split by state and category.

Figure 12: Level 6: Model + Standard Error

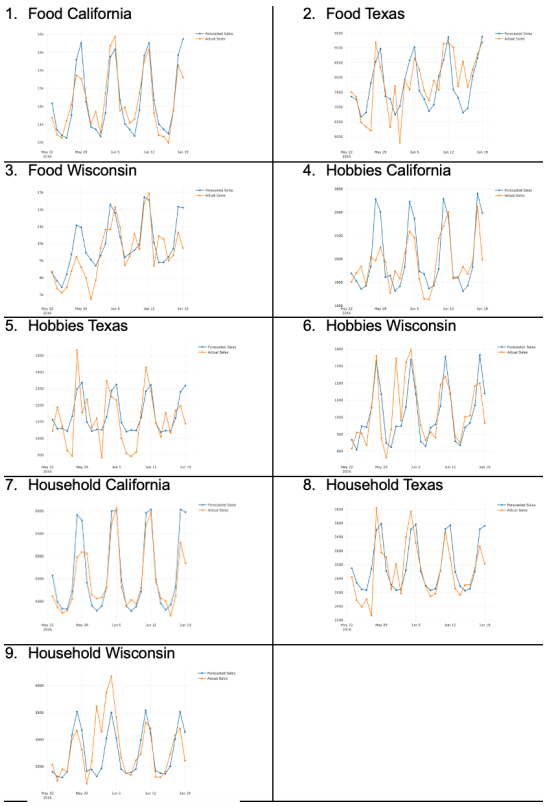

Figure 13: Level 6: Actual vs. Forecasted Sales

Description of The Result. We based the result on the standard error, square of sigma, the graph, and the value of RMSE.

Overall, the standard error show little proportion compared to its coefficient in all models. This show that the coefficient has high level of precision. The problem is there are several NaN (Not available Number) which means the standard error cannot estimated for its specific coefficient. That means there are issues with the estimation process due to convergence problems or issues with the data. We believe this may be caused by overcomplex model which is shown by the high value of parameter of AR and Seasonal part.

The values of square of sigma in all models show quite high variation. At glance, for each category, the values are biggest for store in California and the second is the store in Wisconsin. By category, it is clear that the category Hobbies has the smallest value and category Food comes as the biggest. The value of square of sigma practically tells us how much the unexplained fluctuations that the SARIMA model has not captured.

No.	Category	State	RMSE
1	Food	California	831.0758
2	Food	Texas	754.9589
3	Food	Wisconsin	1167.668
4	Hobbies	California	190.0999
5	Hobbies	Texas	127.5876
6	Hobbies	Wisconsin	129.2514
7	Household	California	397.8572
8	Household	Texas	75.4213
9	Household	Wisconsin	401.1769

Table 3: RMSE values by Category and State

The values log likelihood for all models show quite the same value. It means that overall, all models have similar quality to explain the observed data.

From the graph, the forecasted value show quite good fluctuations that overall rhymes with the actual value. However, we can see some graphs are not good at predicting the actual value's small fluctuations. This is obvious when we look at the graph of Food Texas, Food Wisconsin, and Hobbies Texas. This may be the sign that another variable influenced the sales outside what SARIMA model can predict.

Based on Root Mean Square Error, it can be said that overall value show that the model is better on forecasting compared to the regression model, in line with the shape of the graph. Furthermore, it can be said that store in Texas shows the least RMSE in each category. Store in Wisconsin has the biggest RMSE for category Food and Household, while for category Hobbies, store in California is the biggest.

7 FUTURE DIRECTIONS

One additional step we could take to solve this problem is to try a different aggregation level. Ideally, we would like to increase granularity by aggregating at a lower level, especially at an item ID level. This can reveal finer trends or variations within the data. For example, based on our investigation of the sales data, the Foods category has hundreds more items compared to the other categories which have more pricing variation. Additionally, based on our research and common knowledge, people have different decision making for different products. By increasing our aggregation level, we hope to better capture consumer's diverse economic decision making in terms of buying products.

Another additional step would be to forecast future sales given a specific promotion. For example, we would like to investigate whether "buy-one-get-one" affects future sales. We believe that in order to perform this inquiry, we would need aggregate at an item ID level. In the hypothetical dataset we need, it would contain price elasticities of each item, date, store ID, and item ID which are all to perform the necessary analysis.

In addition to forecasting future sales given a promotion, we can improve the accuracy of the SARIMA model by finding the right parameters. We need to find the best value for p, d, q, P, D, Q, and s.

We might try to conduct a trial and error method to look for the best value. The issue we have in our model is perhaps the number of p. We feel the value is too high so the model is overfitting. So, we believe it is good to lower the value and see what is the result.

8 REFERENCES

(1) Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, The M5 Competition: Background, organization, and implementation, International Journal of Forecasting 38 (2022) 1325-1336 https://doi.org/10.1016/j.ijforecast.2021.07.007 https://www.sciencedirect.com/science/article/pii/S0169207021001187

(2) Robert Fildes, Shaohui Ma, Stephan Kolassa, Retail forecasting: Research and practice, International Journal of Forecasting, Volume 38, Issue 4, 2022, Pages 1283-1318, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast.2019.06.004. (https://www.sciencedirect.com/science/article/pii/S016920701930192X)

(3) Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos, M5 accuracy competition: Results, findings, and conclusions, International Journal of Forecasting, Volume 38, Issue 4, 2022, Pages 1346-1364, ISSN 0169-2070, https://doi.org/10.1016/j.ijforecast (https://www.sciencedirect.com/science/article/pii/S0169207021001874)

(4) Rut Vyas, Revathi AS, Seasonal Sales Prediction and Visualization for Walmart Retail Chain Using Time Series and Regression Analysis: A Comparative Study, 2022 International Conference on Smart Technology and System for Next Generation Computing, https://doi.org/10.1109/ICSTSN53084.2022.9761294 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9761294

(5) Predicting Walmart Sales, Exploratorydataanalysis,and ... - Rit, www.rit.edu/ischoolprojects/sites/rit.edu.ischoolprojects/files/docume _Capstone.pdf. Accessed 3 Oct. 2023.