Superstore Sales

BSTAT-5325-001: Advanced Methods for Analytics

University of Texas at Arlington

Team 08

Fall 2021

Afzal Imaan
Christian Kamarga
Santosh Kusuma
Surendra Mantena
Cameron Royes
Ruchi Shukla

TABLE OF CONTENTS

INTRODUCTION

The industry of retail is a large and competitive space for many companies looking to gain an edge over one another and acquire additional market share, with the goal of optimizing sales and profits. In order to facilitate the growth of a company, businesses collect internal and external data. They use this data to understand their position in the market, improve their company, and use projections to make short and long term decisions. In our study, we made the goal to predict sales for the coming year and more specifically how products will perform based on previous years' data. These measurements and models can be useful for stores in making management operational decisions, customer insight, and inventory related decisions for the business' prospective growth. Both short and long-term decisions can be made from the data that we collected and outcomes projected. The data could facilitate corporate decisions that would shift the focus of the entire company or decisions as small as daily promotions based on store performance that would be better suited for store managers working at the store to have a considerable weight in the decision making for those instances. Also, inventory becomes a large component of efficiency. Product turnover, time products sit on shelves, and other time-based measurements that could be cut down with optimal data provided and easy interpretation of the data. Through these factors and implications, the projections and calculations of the data are critical components to the success of a company.

DATA COLLECTION

This project's data was collected from data.world. The dataset consists of superstore's sales from 2018 through 2021. We collected data from overall sales as well as sales from individual products. These products were categorized into three categories of office supplies, technology, and furniture. In order to provide greater context to the overall sales, the collection of sales based on

products offers greater insight into how the company can improve and become more efficient in strategies across many company aspects. The overall product sales were calculated quarterly as opposed to annual or monthly because the time frame was optimal in displaying easily interpretable graphs, sales calculations, and consistent predictions based on previous years' data. Once the dataset showed the optimal format for a regression analysis, the first determination was the accuracy of the correlation. A high R-Squared measurement was wanted in order to give validity to our data and projections.

Before conducting calculations and projections, questions were compiled to give internal and external company usage. "Can the data be used to predict sales of certain products?" In order for the business to do its best competing with its competitors, knowing the sales of certain products is essentially important, and can yield a better understanding of the superstore's customer demand. "What categories and subcategories will pose the greatest improvement and trend continuation?" Previous years' data was taken into consideration and used to project the upcoming year. This would allow for the company to adjust promotional strategies and inventory to anticipate periods of lower sales. "Does location and shipping have any direct impact on business sales?" Our dataset includes locations from across the country as well as various modes of shipping. These costs can be significant in times when shortages of components that facilitate shipping or the inputs of products themselves. In addition, this factor may serve as a driver of profit margin decreases. "How would this dataset benefit those within and beyond the superstore operations?" It affects various people such as the business' executives, store managers, customers, competitors, and others looking to enter the same market. "According to our sales prediction, how can we determine which state earns greater profit?" Through the analysis with product sales being linked to certain locations, it can be determined where the most sales are coming from. Tracking the geographical

location of sales can help the company determine where they may need to improve. Some of the products may possess better usage in areas where stores are located near more population dense cities and states. The managers can adjust inventory levels of particular areas based on demand and move inventory towards areas with higher value.

The dataset chosen would best fit a linear regression analysis in which the dependent variable is Sales. The independent variables were the Categories/Subcategories of products (Furniture, Technology, and Office Supplies), location variables (City, State, Postal, Region), shipping mode (Standard, Second, First, and Same Day), Segment, Profit, and Quantity.

The dataset contains both quantitative and qualitative variables. The description of the variables are given below:

Segment - Describes whether the customer was an individual or corporate.

City - The City of where the customer resides.

State - The State of where the customer resides.

Postal_code - The Postal/zip code of the customers.

Category - Lists out the types of department the store contains.

Subcategory - Lists out the items present under the Categories.

Sales - The Selling price of the product.

Quantity - The number of items bought.

Discount - The amount of discount given to the particular product.

Profit - The profit that the store has made (in dollars).

METHODOLOGY

For data set exploration, we used various data modeling techniques such as regression, decision trees, and times series forecasting. To answer our business questions, we chose linear time series forecasting to get future predictions and linear regression to calculate the trend component. As a result, we will explain time series methodology and linear regression results obtained in detail throughout this module.

The term "time series" refers to a series of data points that occur in succession over time. One can determine how certain variables are influenced from period to period using a time series. Using historical values and patterns to predict future activity is a common application of time series forecasting.

The time series has multiple components (refer to Appendix: Figure 1). Each component expresses a particular aspect of its movement:

1. Trend - It refers to the upward or downward movement over time characterized by a time series. For instance, long-run trends in the sales of a particular industry may be affected by changes in consumer tastes, growth in the overall population.
2. Seasonal Variation - reflects changes during a particular season. e.g Coca Cola uses historical data from past years to forecast its future sales, the sale of umbrellas and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.
3. Cyclical Fluctuations - Recurring up-and-down movement of trend levels that can last for up to 10 years, measured from peak to peak or trough to trough. Business cycles are one of the most common cyclical fluctuations found in time series data, which are caused by repeated periods of prosperity and recession.

4. Irregular variations - a series may also occur because of irregularities. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, floods, famines, and any other disasters.

Modeling Trend Components

A time series of sales observations that has an essentially straight-line or linear trend is plotted.

"No Trend" Regression Model - When there is no trend, the least squares point estimate b0 of b0 is just the average y value, that is, we have a horizontal line that crosses the y axis at its average value.

$$y_t = \beta_0 + \varepsilon_t$$

"Linear Trend" Regression Model - When sales increase (or decrease) over time, we have a trend. Oftentimes, that trend is linear in nature. Linear trend is modeled using regression, where sales is the dependent variable, time is the independent variable (Weeks, months, quarters, years).

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Modeling Seasonal Components - Within regression, seasonality can be modeled using dummy variables. Consider the equation below:

$$y_t = \beta_0 + \beta_1 t + \beta_{Q2} Q_2 + \beta_{Q3} Q_3 + \beta_{Q4} Q_4 + \varepsilon_t$$

The coefficient will then give us the seasonal impact of that quarter relative to Quarter 1.

Quarter 1, Q2 = 0, Q3 = 0 and Q4 = 0, Quarter 2, Q2 = 1, Q3 = 0 and Q4 = 0

Quarter 3, Q2 = 0, Q3 = 1 and Q4 = 0, Quarter 4, Q2 = 0, Q3 = 0 and Q4 = 1

Multiplicative Decomposition

When a time series exhibits increasing (or decreasing) seasonal variation, we can use the multiplicative decomposition method to decompose the time series into its trend, seasonal, cyclical, and irregular components.

$$y_t = TR_t \times SN_t \times CL_t \times IR_t$$

Here TR_t, SN_t, CL_t, and IR_t represent the trend, seasonal, cyclical, and irregular components of the time series in time t.

Smoothing Techniques

Moving Average methods: It is calculated by adding up all the data points during a specific period and dividing the sum by the number of time periods.

$$\text{Moving Average} = \frac{\text{Sum of the } m \text{ most recent observations}}{m}$$

We will now talk about how to deal with time series data and so let's directly jump into our superstore dataset which is aggregated data broken down into four quarters. Year quarter and sales over here consist of a historic time series data and what we are measuring here specifically is sales and the goal of all is that we want to understand how this sales moves through time and we should be able to extrapolate the sales of perhaps year 2022.

STEP 1)

We plot a graph of Sales vs Time (refer to Appendix: Figure 3). We clearly see that there is a pattern that kind of repeats itself every year. One component that we can visualize is seasonality because this cycle occurs within one year. Second thing here you can see is that the direction of this plot is somewhat increasing and that's what we call a trend component. Final component in

the model that we are using is going to be the irregular component that's always present in the data no matter whether it is a time series or regression.

STEP 2)

We create a timecode column 't' which we will use very often in this analysis. Next, create a column as 'moving average' where we perform smoothing of data or in simple words where we will take care of the irregularities in our data. **As our data is aggregated quarterly, we will take a moving average of 4 periods.** Since we are taking an average of even numbers of periods, we also calculate Centered Moving Averages (CMA). Now plot this CMA, we get the orange line of CMA right on top of the original time series data (refer to Appendix: Figure 3).

STEP 3)

Next step is to divide sales/CMA to see the seasonality factor and irregularity component followed by taking out irregularity and just extract the seasonal (St) component by taking the average of each year Q1 'St' and 'It', then years Q2 St and It and so on. Now deseasonalize the data (Yt, sales/St, seasonal component).

According to classical multiplicative model says that sales (Yt) equal to St* It*Tt. To obtain a trend, we run a simple linear regression using deseasonalized data as response variable (Y) and t as (X) independent variable. To calculate trend components, we need intercept and slope (coefficients) (refer to Appendix: Figure 2).

Finally, we perform the prediction/forecast by combining the components which we removed earlier i.e., **St(seasonal)*Tt(trend).** This way we can go to the year 2022 and perform the forecast. Plot this forecast data and this is the result we were looking for. Error can be calculated by subtracting actual – forecast sales.

Mean absolute deviation (MAD)

$$MAD = \frac{\sum_{t=1}^{n} |e_t|}{n} = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

Mean squared deviation (MSD)

$$MSD = \frac{\sum_{t=1}^{n} e_t^2}{n} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}$$

RESULTS

Through our time series forecast, we were able to see that Q4 posed the highest sales every year and seems to be similar with our predictions in 2022.

Our predictions for 2022 Q4 overall sales increase 12% from 2021 Q4 sales and plan to reach its highest quarterly sales (refer to Appendix: Figure 3).

The Q4 2022 projections suggest that there will be an increase of 7% sales from Q4 2021 actual sales. Furniture poses a decrease of 4.2 % in sales from actual 2021 Q4 sales. While Office supplies show 18% increase in sales from previous year (refer to Appendix: Figure 4a, 4b, 4c).

CONCLUSION

We can conclude that by analyzing the results and data visualization graphs from the superstore data set, we were able to forecast product sales in 2022. Furthermore, in 2022, which product categories and subcategories are most likely to improve and fall. According to our data, there were more sales from office supplies and technology than furniture.

Using data visualization techniques, we were able to determine the number of orders sold and profits made in each state, as well as the overall most successful products and products most likely to go out of stock. Also, explained that the shipping method has no direct impact on business sales and profits, and whether or not to offer discounts on category products.

We anticipate that these findings and recommendations will assist the superstore owner's plan for 2022 more effectively and increase the store's annual sales revenue.

CITATION

Brown, Lorna. “Superstore 2021.” *data.world*, 20 Apr. 2021, 3:49 AM, https://data.world/missdataviz/superstore-2021.

Fernando, Jason. “Moving Average (MA).” *Investopedia*, 05 Oct. 2021, https://www.investopedia.com/terms/m/movingaverage.asp

APPENDIX

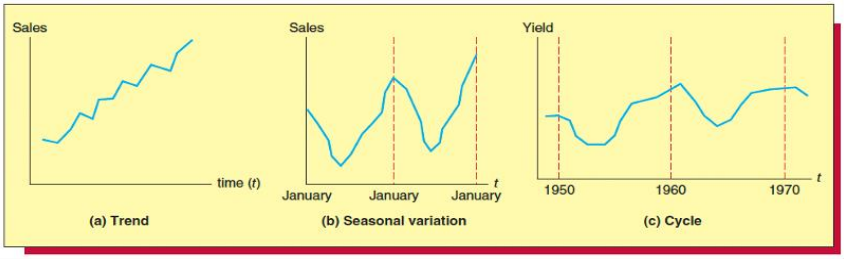

Figure 1: Time Series Components - Trend, Seasonal, and Cyclical

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.882333987
R Square	0.778513264
Adjusted R Sq	0.762692783
Standard Erro	14608.06057
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10501019175	10501019175	49.209203	6.10218E-06
Residual	14	2987536070	213395433.6		
Total	15	13488555245			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	94161.47982	7660.531589	12.29176836	6.867E-09	77731.27364	110591.686	77731.27	110591.686
t	5557.458998	792.2333167	7.014927144	6.102E-06	3858.287527	7256.63047	3858.288	7256.63047

Figure 2: Linear regression between deseasonalize data vs t

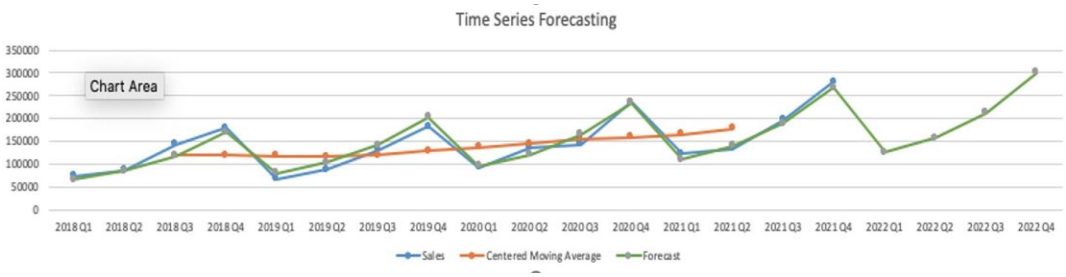

Figure 3: Total sales Forecast

Figure 4a: Office supplies sales forecast

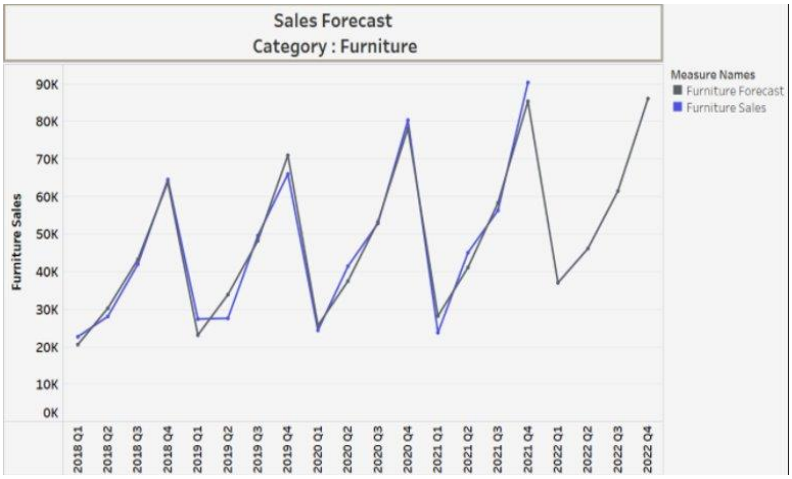

Figure 4b: Furniture sales forecast

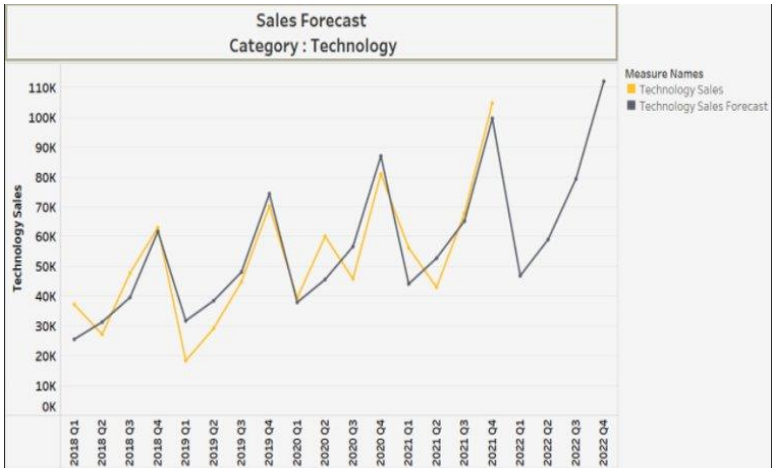

Figure 4c: Technology sales forecast