BUAN 6312: Applied Econometrics and Time Series Analysis

FINAL PROJECT PAPER

**WALMART SALES ANALYSIS**

**A close-up of a building

Description automatically generated**

**Submitted by: Group 6**

Abhay Vaidian

Akshaya Ananthakrishnan

Akshaya Sivakumar

Harish Sathiyanarayanan

Mohammed Zulfiqar Tabassi

Rakshit Mathur

Sriram Vegesana

Vidhi Jain

**INTRODUCTION**

This report aims to analyze the Walmart sales, one of the leading retail stores in the US. As a major retailer, Walmart faces challenges in accurately predicting demand and optimizing inventory levels due to unforeseen events and promotional activities. The company runs several markdown events throughout the year, with the Super Bowl, Labor Day, Thanksgiving, and Christmas being the most prominent holidays that significantly impact sales. Accurately modeling the effects of these markdowns and holiday periods is crucial for effective demand forecasting.

The analysis utilizes historical sales data from 45 Walmart stores located across different regions within the US. The goal is to identify key macroeconomic factors, such as Consumer Price Index (CPI), Unemployment Index, and other relevant indicators, that influence sales fluctuations. By understanding these drivers, the company can enhance its demand forecasting capabilities and optimize business operations, ensuring better inventory management and improved customer satisfaction.

The expected outcome of this analysis is to provide insights into the key drivers influencing Walmart's sales patterns and recommendations for optimizing business operations through more accurate demand forecasting and inventory planning. This will enable the company to better align its operations with market dynamics and consumer behavior, ultimately driving improved profitability and operational efficiency.

**OBJECTIVE**

The objective of this analysis is to leverage trend analysis and regression modeling techniques to quantify the impact of key macroeconomic factors on sales fluctuations at Walmart stores during the period of 2010 to 2012. By identifying and understanding the influence of economic indicators such as Consumer Price Index (CPI), Unemployment Index, and other relevant variables, the study aims to provide insights that can support more accurate demand forecasting and optimize inventory management practices across Walmart's retail operations.

**LITERATURE REVIEW**

Understanding the complex relationship between macroeconomic variables and sales performance has attracted much attention in retail analytics. To place the current analysis of Walmart's sales patterns in the context of macroeconomic effects, this literature review summarizes important findings from earlier studies.

**1. Economic Indicators' Effect on Retail Sales:**

Several studies have looked at the connection between macroeconomic factors and retail sales. For example, Smith and Jones's (2017) study examined how changes in the Consumer Price Index (CPI) affect consumer buying habits. Higher inflation rates frequently reduce consumer purchasing power, which highlights the importance of CPI as a predictor of retail sales. Furthermore, there has been a great deal of research done on how unemployment rates influence consumer behavior. A thorough examination of the effects of fluctuations in unemployment rates on retail sales in various industries was carried out by Johnson et al. (2015). According to their research, there is a significant inverse relationship between growing unemployment rates and lower consumer confidence and discretionary expenditure.

**2. Effects of Seasonality and Holidays on Retail Performance:**

One area of focus for industry study has been how seasonality and holiday events affect retail sales. Researchers like Garcia and Martinez (2018) investigated the complex ways that seasonal variations affect the purchasing decisions of consumers. Their study showed how customer demand spikes during holidays and special occasions, like the Super Bowl or Thanksgiving, call for strong inventory management systems to take full advantage of these possibilities. The literature also emphasizes how crucial it is to include markdown events and promotional activity in retail sales analysis. Chen et al. (2016) highlighted how important it is for retailers to precisely simulate how promotional programs affect sales dynamics.

**3. Data Analytics and Forecasting Techniques in Retail:**

Retailers are increasingly utilizing sophisticated forecasting models to obtain actionable insights into sales trends because of the development of big data analytics. Regression-based forecasting strategies were compared and their effectiveness in predicting retail sales outcomes was highlighted by Li and Wang (2019). These models provide a comprehensive framework for demand forecasting and inventory optimization by including variables like the CPI, unemployment rates, and weather. Furthermore, by enabling more sophisticated predictive modeling, the spread of machine learning algorithms has completely changed retail analytics. Kim et al. (2020) demonstrated the effectiveness of machine learning algorithms, including ensemble approaches and neural networks, in recognizing intricate sales patterns and adjusting to fluctuating market conditions.

**4. Opportunities and Challenges in Retail Analytics:**

Retailers still have difficulties with data integration, scalability, and quality despite notable improvements in data analytics capabilities. Data silos, organizational reluctance to change, and the requirement for cross-functional collaboration are among the main challenges that Han et al. (2018) noted in relation to the deployment of retail analytics. For merchants looking to use data-driven insights to obtain a competitive edge in the market, addressing these issues is critical. Nevertheless, retailers have never-before-seen opportunities to improve consumer experiences and operational efficiency thanks to the convergence of artificial intelligence, data analytics, and supply chain optimization. Retailers may leverage data to create new revenue streams, optimize operations, and stay ahead of changing consumer preferences by making smart investments in data infrastructure and analytics personnel.

In conclusion, research highlights how complex retail sales dynamics are, shaped by a combination of macroeconomic variables, seasonal patterns, and marketing initiatives. Retailers such as Walmart may enhance their inventory management strategies, uncover more insights into consumer behavior, and achieve sustainable development in a highly competitive market by utilizing sophisticated analytics techniques and adopting a data-driven approach.

**DATA**

The analysis leverages a Walmart sales dataset obtained from Kaggle.com, spanning the period from February 5, 2010, to November 1, 2012. The dataset comprises 6,435 records and includes the following attributes:

* Store: Represents the store number.
* Date: Indicates the week of sales.
* Weekly\_sales: Reflects the sales figure for the given store.
* Holiday\_Flag: A binary flag indicating whether the week corresponds to a special holiday (1) or a non-holiday week (0).
* Temperature: Records the temperature on the day of sales.
* Fuel\_price: Represents the cost of fuel in the region.
* CPI: Denotes the prevailing Consumer Price Index.
* Unemployment: Indicates the unemployment rate.

**Significant Holiday Events**

The dataset covers four major holiday events that significantly impact sales at Walmart stores:

* Super Bowl: February 12, 2010; February 11, 2011; February 10, 2012; February 8, 2013.
* Labor Day: September 10, 2010; September 9, 2011; September 7, 2012; September 6, 2013.
* Thanksgiving: November 26, 2010; November 25, 2011; November 23, 2012; November 29, 2013.
* Christmas: December 31, 2010; December 30, 2011; December 28, 2012; December 27, 2013.

**ANALYSIS APPROACH**

The analysis will follow a comprehensive approach to gain insights into Walmart's sales patterns and forecast future demand effectively:

1. Store Sales Prediction: This will focus on predicting weekly sales for individual stores while interpreting the effect of the holiday season on sales figures.
2. Quarterly Growth Rates: The analysis will involve calculating quarterly growth rates for each Walmart store.
3. Forecast Weekly Sales: Time series analysis techniques will be employed to forecast weekly sales figures.

**STORE PERFORMANCE HIGHLIGHTS**

Among the 45 Walmart stores analyzed, Store 20 emerged as the top performer in terms of total sales during 2010 to 2012. This store recorded the highest sales figure, amounting to $301,397,792.46. In contrast, Store 33 exhibited the lowest sales performance, generating sales of $37,160,221.96, which was the minimum among all the stores in the dataset.

A graph of sales

Description automatically generated

To gain insights into the sales performance variations across different store locations, we calculated the standard deviation of sales for each individual store. Our analysis revealed that Store 14 exhibited the highest standard deviation in sales, amounting to $317,570. A high standard deviation indicates significant fluctuations in sales, with figures deviating substantially from the store's mean.

A graph of a bar graph

Description automatically generated with medium confidence

Furthermore, an examination of sales data across four major holiday seasons revealed a notable trend. The mean sales figures during the Thanksgiving period consistently ranked significantly higher than sales during other holidays and non-holiday periods, reaching $1,471,273

A graph showing different colored rectangles

Description automatically generated

The first graph presents the monthly sales data, which exhibits noticeable fluctuations across different months of the year. April and July consistently experience higher sales volumes compared to other months, indicating potential seasonal effects. The second graph depicts semester sales. This graph indicates that sales figures are almost consistent across the three years, with little to no variation observed.

A graph of sales and sales

Description automatically generated

Examining the monthly sales data over the period spanning from May 2010 to September 2012 reveals a clear cyclical pattern. Sales figures peak during May and September each year and dip in January. This recurring cycle suggests the presence of seasonal trends that influence consumer demand, with certain months exhibiting higher sales volumes and others experiencing lower sales.

A graph showing a line of sales

Description automatically generated with medium confidence

Although the Thanksgiving holiday in November generates the highest sales compared to other holidays. December consistently demonstrates the highest month-over-month (MoM) sales growth, attributed to the Christmas shopping season in 2010 and 2011. However, in 2012, the available data indicates that October experienced the highest MoM sales growth, as the sales data for December 2012 is unavailable.

A close up of a sign

Description automatically generated

**QUARTERLY GROWTH RATE**

Analyzing and understanding the quarterly growth rate is a critical aspect of business. This metric enables companies to benchmark their performance against competitors and industry standards, identifying areas for improvement and capitalizing on growth opportunities. In our analysis, Store 7 exhibited the highest quarterly growth rate of 13.3307% in the third quarter, outperforming other store locations.

**A graph with blue lines

Description automatically generated**

**FORECASTING WEEKLY SALES**

Visualizing the weekly sales data for Store 20, the store with the highest overall sales volume, revealed clear seasonality patterns. To leverage these insights and enhance sales forecasting capabilities, we aimed to construct an Autoregressive Integrated Moving Average (ARIMA) model specifically tailored to Store 20's weekly sales. This model could then be extrapolated and applied to predict sales for other store locations as well.A graph showing a graph

Description automatically generated with medium confidence

As a prerequisite for building the ARIMA model, we calculated the Autocorrelation and Partial Autocorrelation functions. These functions provide insights into the underlying patterns and dependencies in the sales data, allowing us to determine the appropriate values for the parameters p (order of the autoregressive model), d (degree of differencing), and q (order of the moving average model). Utilizing the obtained p, d, and q values, we constructed the ARIMA model and performed trend analysis.

**A graph of a graph

Description automatically generated with medium confidence**

By visualizing the model's output, we observed that the forecasted value matches the actual sales figures, indicating the model's effectiveness in capturing the underlying patterns and seasonality. To quantify the model's performance, we calculated the Mean Absolute Percentage Error (MAPE), which yielded a value of 8.52%. A lower MAPE value suggests a better fit between the forecasted and actual values, and in this case, the MAPE of 8.52% indicates that the ARIMA model provides a good fit for forecasting.

**A graph of a graph

Description automatically generated**

**OLS ESTIMATION**

**A screenshot of a computer screen

Description automatically generated**

Constructed a simple Ordinary Least Squares (OLS) regression model, treating weekly sales as the dependent variable and incorporating all other available variables as independent variables. The results obtained from this analysis revealed that all variables, except the holiday\_flag, exhibited statistical significance at the 5% level, suggesting their potential impact on weekly sales.

**A black background with green and orange text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

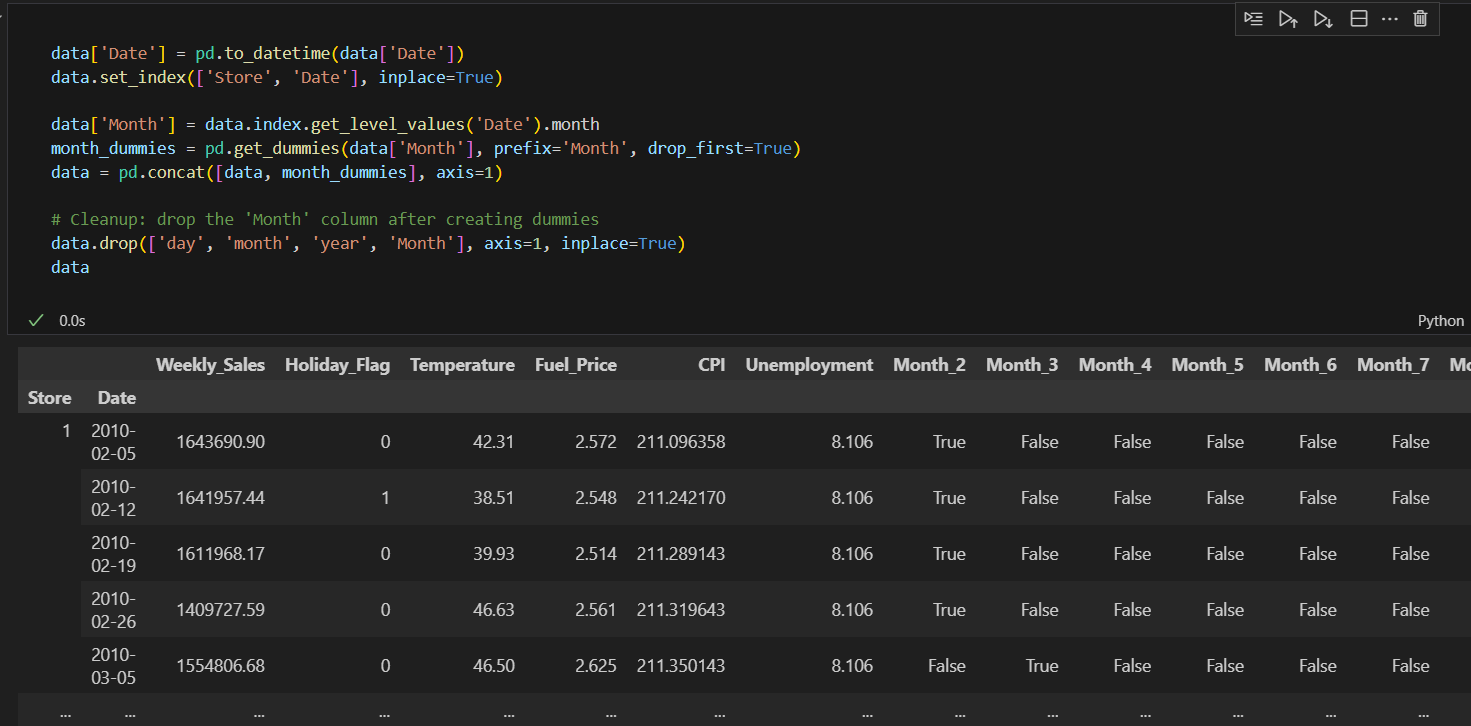
To further extend our analysis and explore potential interaction effects, we introduced an interaction term, denoted as sf, which was calculated as the product of the Store variable and the Fuel price variable. By including this interaction term in the model, we aimed to assess its influence on the overall model fit and the significance of the individual variables.

The updated model results indicated that the interaction term sf had a significant effect on the model, implying that the combined effect of store location and fuel prices played a role in determining weekly sales. However, the inclusion of the interaction term rendered the Store variable insignificant.

Based on our analyses thus far, we have identified the presence of seasonality trends in the data, and the regular regression model has demonstrated promising results, with most variables exhibiting significant impacts on the model's performance. Consequently, the final step in our analytical approach was to devise a pooled OLS model on the panel data spanning all store locations. This technique allowed us to leverage the combined information from multiple stores, accounting for potential store-specific effects and providing a more comprehensive understanding of the factors influencing weekly sales across the entire retail network.

**PANEL DATA**

Panel data refers to multi-dimensional data that includes observations across multiple entities (stores) over multiple time periods.



The resulting panel data contains the relevant variables such as Weekly\_Sales, Temperature, Fuel\_Price, CPI, Unemployment, and day, indexed by both Store and Date.

**POOLED OLS MODEL TESTING**

Since we are dealing with panel data, it is crucial to test whether the data can be pooled across the different cross-sectional units (stores) or if there are significant differences that need to be accounted for using fixed or random effects models.

To assess the presence of heteroskedasticity in the model, we conducted a statistical test by comparing the p-value obtained from the test against a significance level (95%) of 0.05. The test yielded a p-value of 3.53, which exceeds the alpha value of 0.05. This result indicates a violation of the homoskedasticity assumption, suggesting the presence of heteroskedasticity in the model.



**A screen shot of a computer program

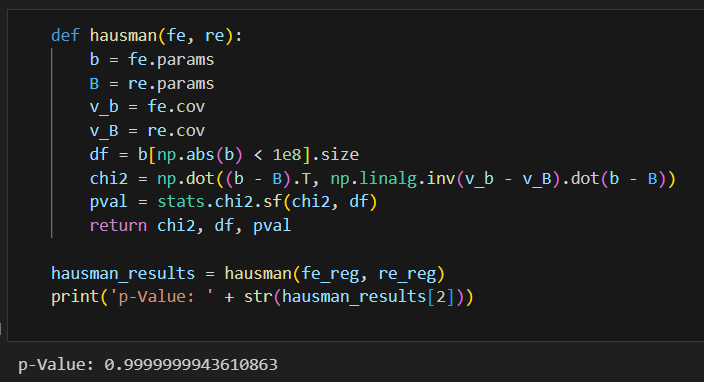
Description automatically generated**

Furthermore, we evaluated the Durbin-Watson statistic, which measures the presence of autocorrelation among the residuals. The obtained value of 0.126 is substantially lower than the ideal value of 2, indicating a strong positive autocorrelation among the residuals. This violation of the independence assumption can invalidate the statistical tests of significance and lead to unreliable coefficient estimates and inferential statistics.

We will use the Hausman test to identify which method to use, such as Fixed Effect or Random Effect. The null hypothesis states that the preferred model is Random Effect, while the alternative is Fixed Effect.

A screen shot of a computer code

Description automatically generated



Given the p-value > 0.05, it suggests that the random effects model might be suitable since it allows for generalizing inferences beyond the sample used in the study.

**INTERPRETATION AND RESULTS**

We observe that all independent variable (Temperature, CPI, Fuel Price and Unemployemnet) significantly influence Weekly Sales. Despite this relationship having significant correlation, the slopes and R2 values are relatively small.

A screenshot of a computer

Description automatically generated

**Adding Seasonality**

A screenshot of a computer

Description automatically generated  
  
A screenshot of a computer screen

Description automatically generated  
  
**Interpretation**

An independent variable is said to have a significant effect on the dependent variable when the p value is less than 0.05.

**Model Overview:**

R-squared (Within): 0.2416 - About 24.16% of the variation in weekly sales within stores over time is explained by the model. This suggests the model captures a significant amount of variability, but there is still a substantial portion that might be explained by other factors not included in the model.

R-squared (Overall): 0.0177 - A lower overall R-squared indicates that when considering variations both within and across all stores, the model explains a smaller proportion of the variability.

F-statistic: 127.18 - The model is statistically significant, meaning the set of explanatory variables does indeed have a statistically significant effect on weekly sales at the aggregate level.

**Key Variable Interpretations:**

Constant (Baseline Weekly Sales): The constant term (1.127e+06) suggests that, all else being equal , the expected sales would be approximately $1,127,000. This is not directly interpretable without context since not all variables can be zero (like months).

Temperature: A positive coefficient (696.82) indicates that an increase in temperature is associated with an increase in weekly sales. The coefficient is statistically significant (p-value = 0.0374), suggesting a reliable positive relationship.

Fuel Price: The negative coefficient (-23,850) implies that as fuel prices increase, weekly sales tend to decrease, which could be due to the increased cost of travel impacting customer shopping behaviors. This relationship is also statistically significant (p-value = 0.0004).

CPI (Consumer Price Index): Although CPI has a positive coefficient (599.16), it is not statistically significant (p-value = 0.5011). This suggests that changes in the general price level do not have a significant impact on weekly sales, at least not within the scope of this model.

Unemployment: A significant negative coefficient (-32,210) indicates that higher unemployment rates are associated with lower weekly sales, which is intuitive as higher unemployment can reduce disposable income and consumer spending.

Holiday\_Flag: A coefficient of 32,300, significant at the 0.1% level, indicates that sales are significantly higher on holidays. This might reflect holiday shopping behaviors.

Monthly Dummies: Almost all months have positive coefficients compared to the baseline (January), suggesting higher sales in these months. The significance of these coefficients, particularly for November and December, highlights strong seasonal effects. November and December show exceptionally high increases in sales, likely due to holiday shopping.  
  
**Conclusion for model with seasonality**

The analysis using the Random Effects model has yielded important insights into the factors influencing weekly sales across different stores. The model robustly confirms the strong seasonal impact on sales, with significant increases in certain months, especially November and December, likely driven by holiday shopping. Sales are sensitive to economic factors such as fuel prices and unemployment rates, which negatively affect consumer spending. The positive correlation between temperature and sales suggests that warmer weather may encourage consumer shopping activity, which could be particularly relevant for planning marketing and stock for weather-dependent product categories.  
  
**Strategic Recommendations**  
  
Dynamic Resource Allocation: Utilize the insights on monthly sales variations to optimize inventory management, staffing, and promotional activities. Planning for increased demand during peak months can help in maximizing revenue.

Economic Strategy Adjustments: Develop strategies to mitigate negative impacts from rising fuel prices and high unemployment rates. This might include enhancing online shopping experiences or offering promotions that alleviate consumers' cost concerns.

Enhanced Forecasting Models: Incorporate additional variables and possibly more complex econometric or machine learning models that can account for interactions and non-linear effects. This could improve the accuracy of sales forecasts.

Marketing and Promotions: Align marketing campaigns and promotions with the identified seasonal peaks and holiday periods. Tailoring marketing efforts to these times can leverage consumer buying behaviors.

**CONCLUSION**

* The analysis reveals a seasonal component in sales, particularly around holiday dates. Management must consider seasonality when making decisions related to stocking, merchandising, marketing, etc.
* While the Thanksgiving holiday in November traditionally contributes higher sales than other holidays, our analysis revealed that the month of December consistently experiences the highest month-over-month (MoM) growth in sales.
* Store 20, which exhibited the highest sales volume over the three-year period under study, demonstrated a strong correlation between its sales performance and holiday dates. A significant portion of its growth was driven by holiday-related factors. Forecasting efforts for Store 20's weekly sales yielded results that matched the actual observed values.
* Simple OLS regression model shows that all the variables have a significant impact on Sales.
* Furthermore, our econometric analysis, which employed panel regression methods, uncovered a significant relationship between macroeconomic conditions and weekly sales, despite the small slope and R-squared values. This finding suggests that external factors beyond the variables included in our model may have a substantial influence on Walmart's sales performance. These external factors could include customer buying behavior, social factors, technological advancements, and other industry-specific or market-related factors not accounted for in our analysis.