ORIE 3120 Final Project

Walmart Retail Sales

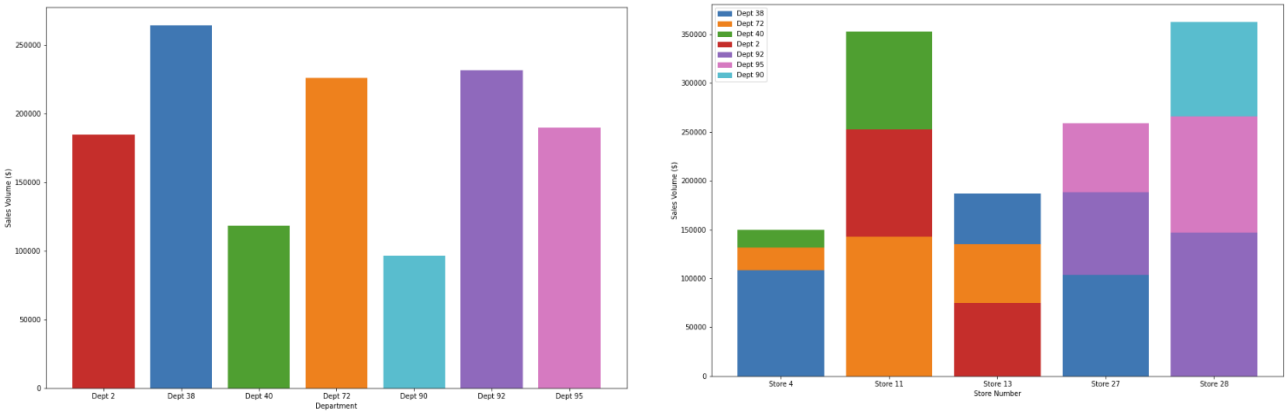
Introduction

Sales continue to post challenges for both economic modeling and for forecasting; as a result, we chose a Walmart historical sales dataset that includes data from 2/5/2010 - 11/1/2012 from stores located in different regions. Due to the anonymized nature of the data, stores and departments are represented only by numbers. The dataset has 3 CSV files: stores, features, and sales. The stores file includes anonymized information about 45 stores (numbered 1-45) and the respective type (A, B, or C) and size of the store. The features file includes the date in week, average temperature, fuel price, five anonymized markdowns (numbered 1-5), Consumer Price Index (CPI), unemployment rate, and whether it is a holiday for a particular date and store. The sales file describes weekly sales separated by each store and department, as well as whether it is a holiday that week.

This project aims to aid executives working in corporate for Walmart. Understanding and forecasting future trends is valuable insight for them as they are the main decision-makers in Walmart sales. In analyzing this dataset, we picked 3 main questions that work together to aid our target audience in 1) providing context and a general overview of their current sales 2) giving executives potential data for room for growth and 3) tying their position into the overall economy as a whole:

Question 1: What is a general breakdown of Walmart sales and how do sales trend over time?

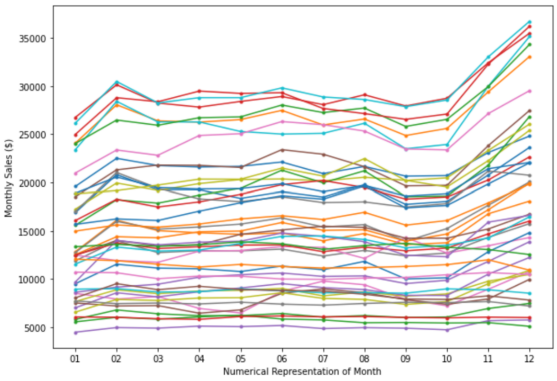
We'd like to utilize forecasting and time-series analysis to understand more about how the sales data provided by Walmart can be predicted and understood by corporate; this could be on both an overall basis as well as per-store or per-department.



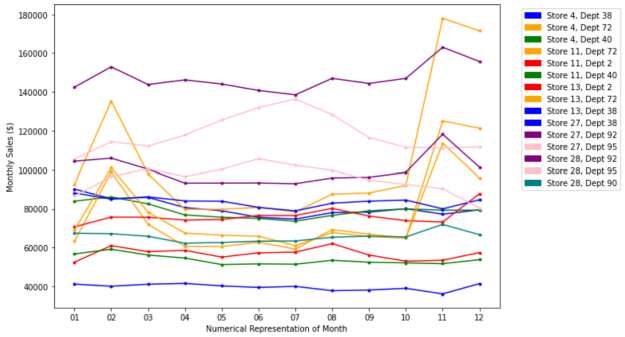
Sales breakdown of the five largest Walmart stores by department contribution (Figure 1.1, 1.2)

The above bar graph (Figure 1.2) informs us which departments contribute the most to the sales of each of the five Walmart stores largest in size. Based on Figure 1.1, it is evident Department 38 contributes the most to the success of these stores both in terms of sales volume and presence. Department 72 and Department 92 also contribute substantially to the sales volume of these large stores. It would be interesting if departments were not anonymous to see what they sell to see which products are most

popular for Walmart sales. Despite this limitation of anonymity of our data that limits our analysis and conclusion, we can still learn about these departments and their trends and impacts, such as how the department sales vary over the course of 1 year.



Monthly sales for each of the 99 departments (Figure 1.3)



Monthly sales for the top 3 departments for the top five largest stores, colored by department (Figure 1.4)

As seen in Figure 1.3, the dataset includes many departments, all with different sales trends. However, it is clear that despite varying trends during the year, they still follow a similar pattern of peaks in February and November-December. To look at the unique patterns of specific departments, Figure 1.4 only plots the 15 store-department combinations that we chose from Figure 1.2. From this, we see similar seasonality per department, with Department 72 peaking in November and February, while Department 95 appears to peak in July. By creating a visual that plots sales by department, we can see that the seasonality highly depends on the department, and presumably the types of products for sale in each department. Although the data is anonymized, we could make inferences about what each department sells based on its seasonality. For example, one might infer that Department 72 (yellow) could sell gift wrap, chocolates, or electronics, while Department 95 (pink) could be beachwear or sporting equipment. Regardless of what they may represent, the distinct seasonality is an important trend to see within this data. We can delve into these trends such as seasonal changes with further analysis.

We can see that, in general, the vast majority of store departments follow a seasonal pattern with regards to sales data. Specifically, the graph above shows that typically, the holiday season, which includes Black Friday in November and Christmas in December, has higher average weekly sales than other periods. A linear regression model does not make sense for predicting and forecasting sales for this data, since it does not take into account seasonality (average weekly sales value depends heavily on month). As such, we chose to build a **Holt-Winters (triple exponential smoothing) model** with an additive trend and 12 seasonal periods, representing each month of the year.

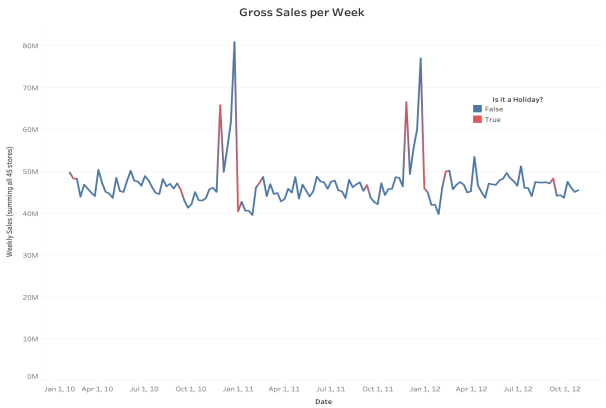
Fitted and forecasted weekly sales from the seasonal Holt-Winters model plotted against the original weekly sales in dataset (Figure 1.5)

The above graph illustrates that the Holt-Winters model generally follows the trends outlined in the department monthly sales line charts (Figure 1.3). This model can aid Walmart in making decisions that increase their average weekly sales over the course of a year. For example, sales decline nearly 28% from December to January on average, according to the model; therefore, Walmart may want to focus on campaigns (such as discounts and promotions) to prevent such substantial drops in sales after holiday periods. Additionally, further consumer investigation (potentially through surveys) would be beneficial to determine why demand tends to drop in September and October, considering the relative consistency in sales throughout the prior several periods (March-August). Finally, understanding seasonal sales trends, which are indicative of demand, can help relevant departments manage and optimize inventory to minimize costs. Thus, this seasonal sales trend analysis can inform marketing and research operations to target certain periods of the year for improvement.

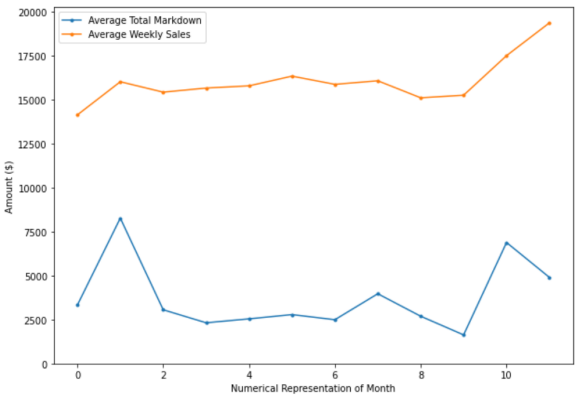
Question 2: How do markdowns and holidays affect sales?

While we've investigated sales and their forecasting as a whole in Question 1, there are some special cases that may also affect sales, which we'd like to investigate as well. Markdowns, for example, are typically used to increase sales by stimulating demand for certain products or during holiday promotions; as such, we wanted to investigate the relationship between markdowns, holidays and weekly sales, and whether (or when) these tools are effective. This could be of great use for companies like Walmart, since it could be used to learn more about their markdown strategy and to propose changes to it; we will dive into the different store types and departments within Walmart, and understand how uneven markdowns affect different stores.

To start, consider the two figures on the top of the next page, Figure 2.1 and Figure 2.2. The former shows combined sales per week over the time period in which data was collected, with Walmart's specified "holiday periods" (the largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas) denoted in red, compared to regular sales weeks which are noted in blue. Looking at it, there appears to be no clear



Sales per Week, Showing Holidays (Figure 2.1)



Difference between Weekly Sales and Avg. Total Markdown over time (Figure 2.2)

trend for most dates: not all holidays tend to drive increases in sales (especially in the middle of the year), and the only holidays that seem truly important to driving sales are Christmas and Thanksgiving.

On the right, in Figure 2.2, is another visualization (over the limited timespan in which any markdowns are available, mostly in 2012) that depicts average weekly sales per month compared to average total markdown in that month. From this, it seems to indicate that there's a mixed association between markdowns and sales: February 2012, for example, had incredibly high markdowns, but little difference in average sales. By contrast, while November markdowns appear to increase sales, it's unclear if this trend is due to markdowns explicitly or not, since in December, the difference in sales is very high but there are no high markdowns occurring. As a result, it seems more important to consider other factors that might influence markdowns: perhaps it's more specific to something like store type, rather than time?

With both of these pictures in mind, and given that time-series forecasting was already completed in Question 1, the decision to make two sets of **predictive models** was chosen: given that being a holiday week is a binary variable, and so are the indicator variables for being each type of store (A, B, or C), we have chosen to model with **logistic regression**, with one set of parameters for each subquestion.

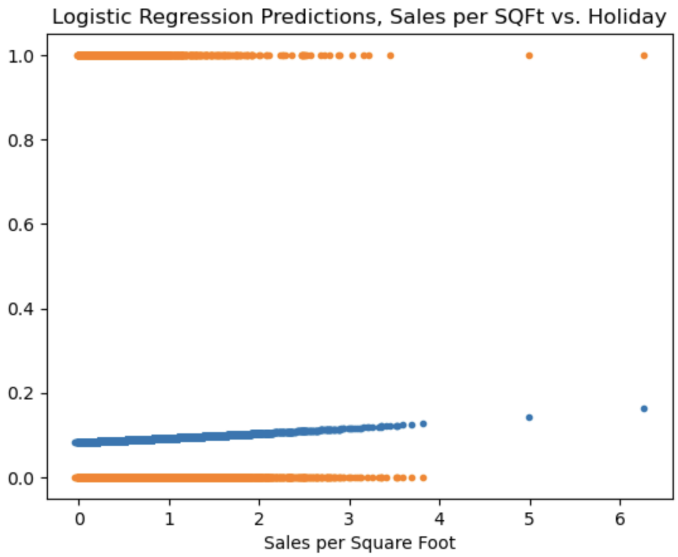
To consider the first model (predicting holiday times), we first consulted a pairplot of each of the variables available to compare them with Holiday times; in this case, it appears that most variables are uncorrelated with it being a holiday, except for *weekly sales* per department and store. Given that there might be hidden confounders in the regression (since there's a lot of spread in sales / department or based on store size), a measure of *sales per square footage of each store* was used; as the only relevant predictor, the model summary and plot are listed on the top of the next page in Figures 2.3 and 2.4.

Of particular interest to this model is that, although the p-value for the *salesperfoot* coefficient is very low (approx. 0), the actual predictions (which are on data for 2011 and 2012; the model was trained on 2010 data for an approximated test-train split) appear to show that the model actually has identified *zero* holidays, indicating that none of the sales values (not even the two far-right points) can be placed to the right of a suitable inflection

```
Optimization terminated successfully.
         Current function value: 0.287672
         Iterations 6
                           Logit Regression Results
```

Dep. Variable:	IsHolidayright	**No. Observations:**	140679
Model:	Logit	**Df Residuals:**	140677
Method:	MLE	**Df Model:**	1
Date:	Mon, 16 May 2022	**Pseudo R-squ.:**	9.741e-05
Time:	15:14:34	**Log-Likelihood:**	-40469.
converged:	True	**LL-Null:**	-40473.
Covariance Type:	nonrobust	**LLR p-value:**	0.004985

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4100	0.011	-212.720	0.000	-2.432	-2.388
salesperfoot	0.1252	0.044	2.860	0.004	0.039	0.211

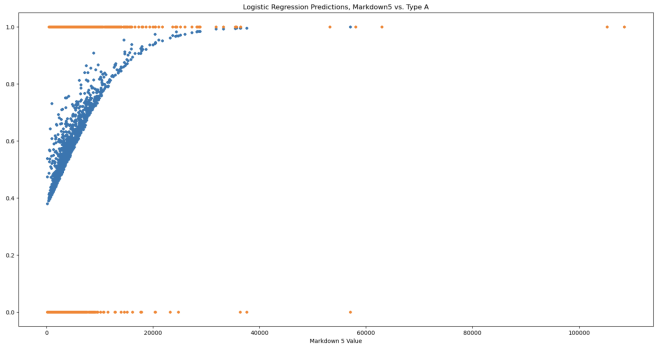


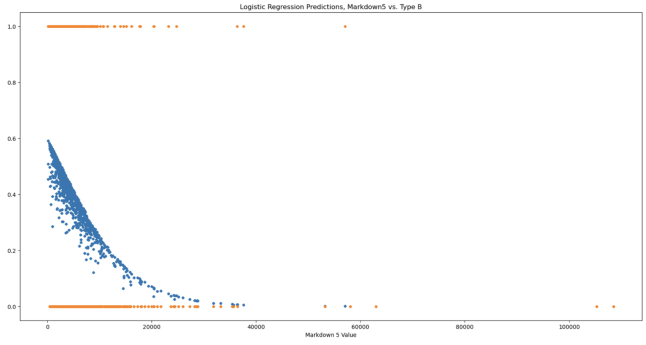
point or cutoff value. Given the wealth of "holiday" data points with low sales per square foot, Walmart might be keen to re-evaluate their definitions of

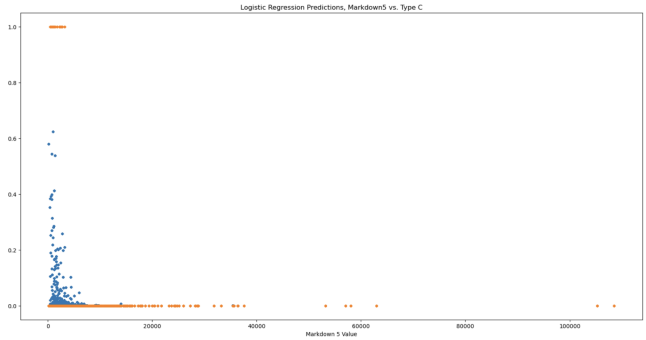
Model Summary (Figure 2.3) and Model Predictions versus Obser. Values of Sales / Ft² vs. Holiday (Figure 2.4)

holidays, or which times of year they should focus on (is Labor Day actually driving or relevant to sales?). Thinking about this model's applicability, it seems that it defied our initial "common-sense" assumption that holidays would be strongly associated with (and thus predicted by) high sales volumes; this is a novel find that should be considered further.

Considering the second set of models, however, we should also consider the next analytical subquestion — whether the available markdown values (separated into MarkDown1 - MarkDown 5) have any sort of value for identifying store or department type. Given that there are 99 listed departments (that are anonymized, so we can't tell which ones each refer to), we considered that an indicator variable encoding for each of the 3 store







Logistic Regression Results, Predicting Whether a Store is Type A (Figure 2.5), Type B (Figure 2.6), or Type C (Figure 2.7) vs. MarkDown 5 Values

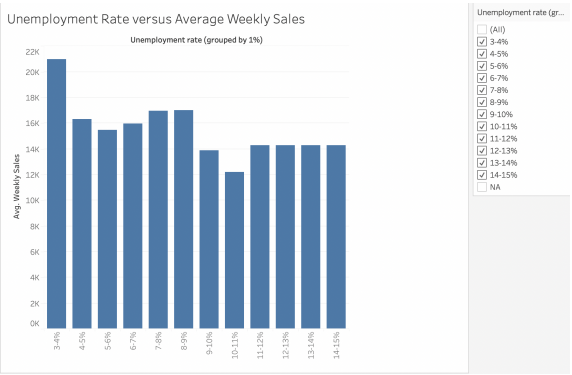
types might be easier to model; graphs for how a logistic function on the MarkDowns could predict or otherwise

identify each type of store (A, B, or C) are above.

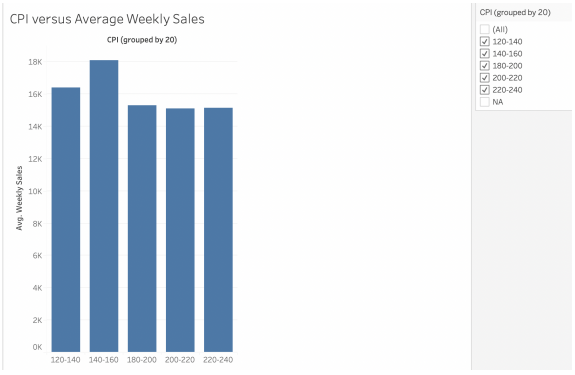
In this case, it seems that the logistic regression models on store type as a function of markdowns (MarkDowns 1, 2, 3, and 5, since MarkDowns 1 and 4 are multicollinear) appear to have some form of descriptive power on identifying store type, not just for MarkDown 5 specifically; this could be relevant for Walmart in identifying where specific markdowns are most used (in this example, it seems higher MarkDowns of type 5 occur most in Type A stores, and rarely in Type C stores), and thus allow Walmart to adapt their markdown strategies accordingly. Perhaps equally importantly, as consumers in turbulent economic conditions, a de-anonymized version of this data might allow us to know where markdowns are strongest, maximize our savings, and make the most of our buying power in otherwise trying economic times. Concluding this subset of the analysis, it seems that it might be possible to understand more about Walmart's markdown strategy and placement using this logistic regression; the relationship between subsets of stores (or even departments) as it relates to markdowns is an area from which we might benefit greatly from even more data than that Walmart provided.

Question 3: How does economic context affect consumer spending at Walmart?

Economic conditions often affect consumer spending, and this may have unique applications for Walmart sales. We want to explore how economic conditions such as CPI, unemployment rate, and fuel price impact sales. Although we have the limitation of older data from 2010-2012, it is interesting to explore sales during different economic periods and Walmart's position as a major department/grocery store. We hope to tie these prediction models into the present day because of their relevance to the current economic situation with consumer prices rising to 8.5% in March of 2022, which is the highest since 1981. Furthermore, we can make an interesting connection as this historical data followed shortly after the 2009 Great Recession, while, in the present day, we are recovering from the COVID-19 pandemic and the effects of the Russia-Ukraine War. Therefore, it is extremely valuable for corporate Walmart decision-makers to understand how economic volatility affects consumer spending, and if they need to adjust their strategy as a result to receive higher sales.

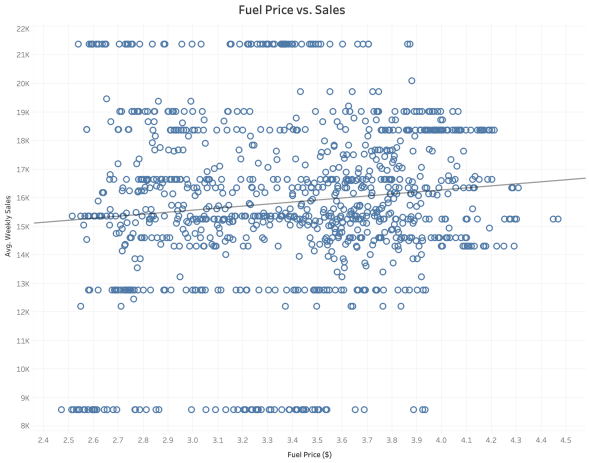


Average Weekly Sales for each percent Unemployment Rate (Figure 3.1)



Average Weekly Sales for each bin of 20 CPI (Figure 3.2)

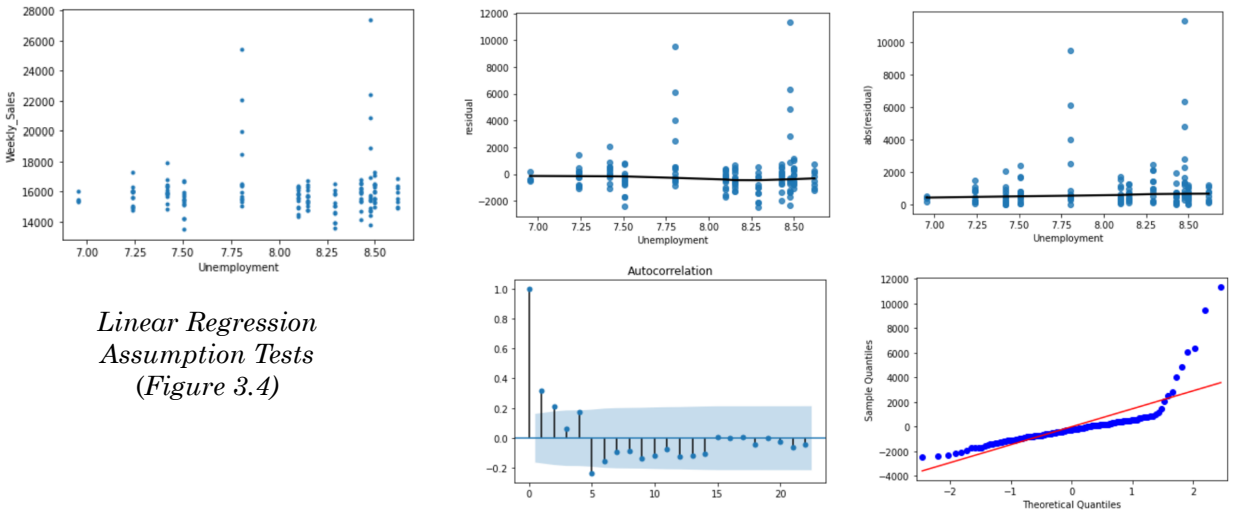
In Figure 3.1, the interesting changes present are the drop between 3-4% and 4-5% as well as the decrease through 8-11%. Although there is not a clear negative correlation, these major jumps and periods of decreasing sales are something we would like to investigate further. In Figure 3.2, we can see that when CPI is between 140-160, there are higher average weekly sales. These are minor fluctuations that we can investigate to see if they are actually correlated.



Considering how fuel price might change sales, we chose to make a scatterplot of fuel price versus average sales/store; this seems to display a mixed (but slightly positive) correlation between the two, with lots of interesting values that require more analysis (like multiple weeks in which average weekly sales are $21,360 / store, regardless of fuel price). This might mean exploring other covariates in the data and relevant trends to see how interconnected the economic context factors we chose are, and how they might affect sales data.

Fuel Price vs Avg. Weekly Sales (Figure 3.3)

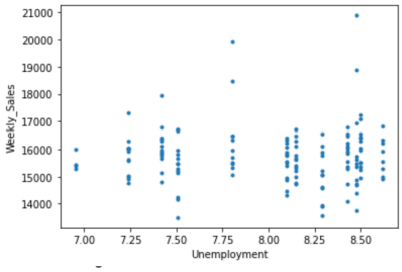
In order to build a model of how economic factors affect Weekly Sales using linear regression, our data must first pass the **four assumptions of linear regression** - a linear relationship, independence, homoscedasticity, and normality. Unfortunately, the economic factors we wished to include such as CPI, unemployment rate, and fuel price all did not pass these four assumptions. For example, Figure 3.4 shows that Unemployment does not meet this criteria, particularly not illustrating a linear relationship or independence.

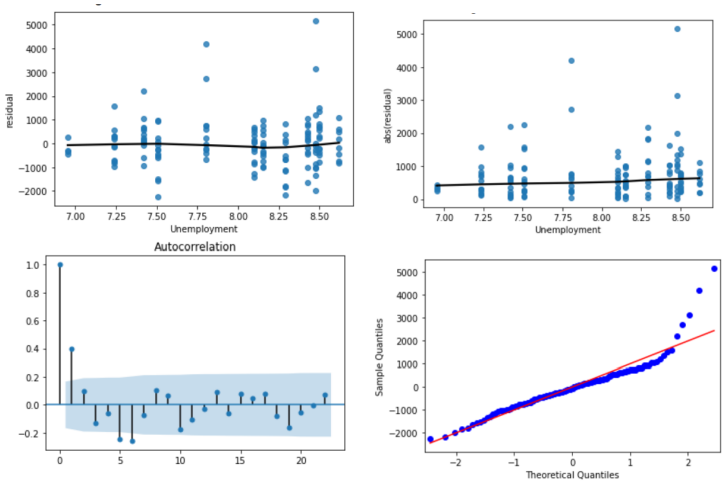


Linear Regression Assumption Tests (Figure 3.4)

Nonlinear transformations of Unemployment such as UnemploymentSquared, UnemploymentSquareRoot, and UnemploymentLog2 did not help fulfill these four

assumptions either. Similarly, removing outliers (Figure 3.5) with high weekly sales (>21000) did not resolve these violations.



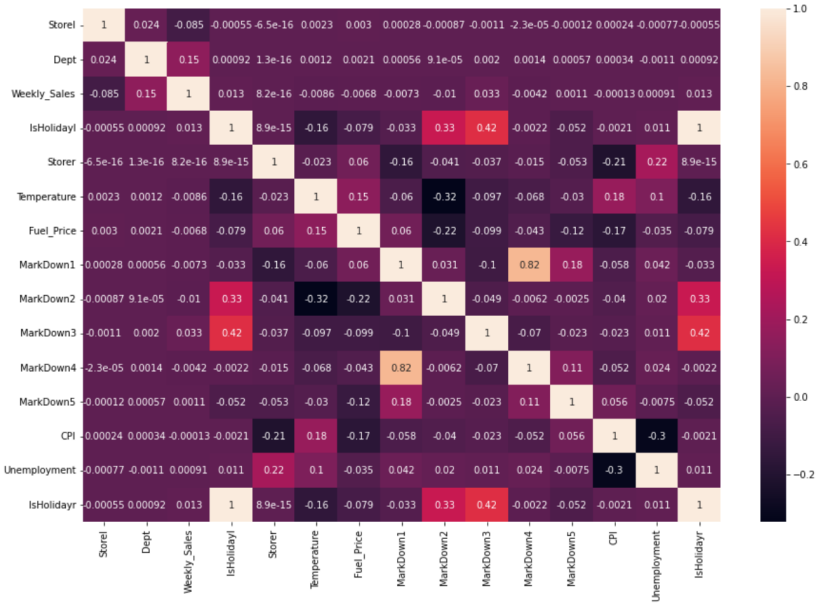


Linear Regression Assumption Tests for Unemployment Rate without outliers (Figure 3.4)

Similar tests on CPI did not pass the required assumptions necessary for linear regression. We decided to create a heat map of the correlations between different covariates (Figure 3.6) to investigate the correlations of possible covariates to include in the linear regression model if we could. This revealed shockingly that there is very little correlation between weekly sales and any of the data provided in the dataset. Walmart's sales not wavering due to differing economic conditions may be due to Walmart's products or place as an affordable, large department and grocery store many rely on. This explains why several of the linear regression assumption tests failed and appeared uncorrelated. It is also important to note that this heatmap was created using all of the data from all of the stores and departments, which could make the data cluttered and difficult to find correlations.

The heatmap for only Department 38 or only Department 95 data is nearly identical to Figture 3.6 (despite different seasonal trends), while the Department 72 data (Nov/Dec peak, see Figure 1.4) has a stronger correlation with Markdown 3. This exploration of covariates illustrates the difficulty of modeling Weekly Sales for all departments given an extremely large and diverse dataset which is an important limitation in our analysis.

Heatmap visualization of correlations between variables (Figure 3.6)



Bibliography

Bachman, Daniel. "Covid-Driven Recession Impact on Retail Industry." *Deloitte United States*, 8 June 2020, https://www2.deloitte.com/us/en/pages/consumer-business/articles/retail-recession.html.

Cox, Jeff. "Consumer Prices Rose 8.5% in March, Slightly Hotter than Expected and the Highest since 1981." *CNBC*, CNBC, 12 Apr. 2022, https://www.cnbc.com/2022/04/12/consumer-prices-rose-8point5percent-in-march-slightly-hotter-than-expected.html.

"The Original Dataset, Containing the 3 CSV Files of Sales, CPI, Markdown History, Department, and Stores." https://www.kaggle.com/Datasets/Manjeetsingh/Retaildataset?Select=Features+Data+Set.csv .

The Investopedia. "The Great Recession Definition." *Investopedia*, Investopedia, 8 Feb. 2022, https://www.investopedia.com/terms/g/great-recession.asp.

Thomas, Lauren. "Retailers Start to Warn of Business Impact from Russia's Invasion of Ukraine." *CNBC*, CNBC, 7 Mar. 2022, https://www.cnbc.com/2022/03/03/ukraine-news-retailers-start-warn-of-business-impact-from-russian-invasion.html.

Wiseman, Paul, and The Associated Press. "U.S. Inflation Surges to Its Highest One-Year Price Hike in over 40 Years." *Fortune*, Fortune, 12 Apr. 2022, https://fortune.com/2022/04/12/us-inflation-march-2022-consumer-price-index-biggest-jump-since-1981/.