Dave Pekerow
Retail Store Sales Project

Task: Clean a dataset containing sales data for a big box retail chain and analyze sales, unemployment rate, Consumer Price Index (CPI), fuel price, and other fields.

Data download: https://www.kaggle.com/datasets/mikhail1681/walmart-sales

Data Cleaning Practice:

Using Python, R, Excel, or another data cleaning tool of your choice, clean the data to meet the following criteria:

Data is sorted first by store number (ascending) and second by date (ascending)

Date is in the format MM-DD-YYYY

Weekly Sales is rounded to the nearest 2 decimal places

Temperature is rounded to the nearest whole number

Fuel Price is rounded to the nearest 2 decimal places

CPI is rounded to the nearest 3 decimal places

Unemployment is rounded to the nearest 3 decimal places

Ensure that there is no missing data

Business Questions:

1. Which holidays affect weekly sales the most?

2. Which stores in the dataset have the lowest and highest unemployment rate? What factors do you think are impacting the unemployment rate?

3. Is there any correlation between CPI and Weekly Sales? How does the correlation differ when the Holiday Flag is 0 versus when the Holiday Flag is 1?

4. Why do you think Fuel Price is included in this dataset? What conclusions can be made about Fuel Price compared to any of the other fields?

Author's Note: For this analysis, I will disregard the data originator's labeling of this dataset as "Wal-Mart" sales. Instead, I will approach it as though it originated from an unnamed big-box retail chain. The methodology will be detailed after addressing the business questions.

ANALYSIS

Business Questions

1. *"Which holidays affect weekly sales the most?"*

- The data confirms that **Black Friday** and the **week leading up to Christmas** are the two most significant drivers of retail sales. Sales during these periods spike significantly, far surpassing those of other holidays or non-holiday weeks.
- Contrary to expectations, **other holidays**—such as **Mother's Day**, **Easter**, and **Presidents' Day**—show no substantial increase in sales, performing similarly to non-holiday periods. While these holidays might be important in other sectors, they do not seem to generate higher sales for this chain, suggesting that consumers are less likely to engage in major shopping sprees around these dates.
- Interestingly, **Memorial Day** and **Labor Day**, both traditionally viewed as strong retail periods for promotions, perform in line with non-holiday weeks. This indicates that either promotional strategies for these holidays are not effective or that customer behavior during these periods does not prioritize high-ticket retail purchases.

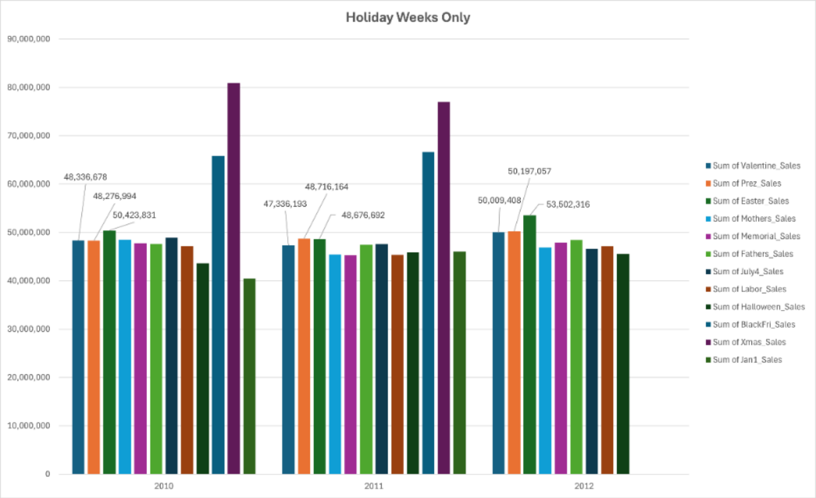


Figure 1: Comparative Holiday Sales Performance – Companywide Retail Store Sales

2. *Which stores in the dataset have the lowest and highest unemployment rate? What factors do you think are impacting the unemployment rate?*

- There is no discernible pattern connecting **regional unemployment rates** to store sales. For instance, stores located in areas with similar unemployment rates show widely varying sales figures, which suggests that unemployment is not a primary driver of retail performance for this company.
- Even when stores are grouped by similar unemployment rates, the differences in sales figures persist, with some stores in high-unemployment regions outperforming those in areas with lower unemployment.
- This lack of correlation implies that the **local economic conditions** represented by unemployment data do not heavily influence consumer behavior in this context. Factors like store location, customer base, or product availability may have more weight in determining sales outcomes than the broader employment landscape.
- The second part of the question, concerning factors affecting the general unemployment rate, falls outside the scope of this analysis. Answering such macroeconomic questions would require more diverse data from multiple sources. Retail sales data from a single

chain, regardless of size, does not reflect broader economic trends. Additionally, the question seems to imply that understanding unemployment is crucial to interpreting retail sales, but for this dataset, it is not relevant.

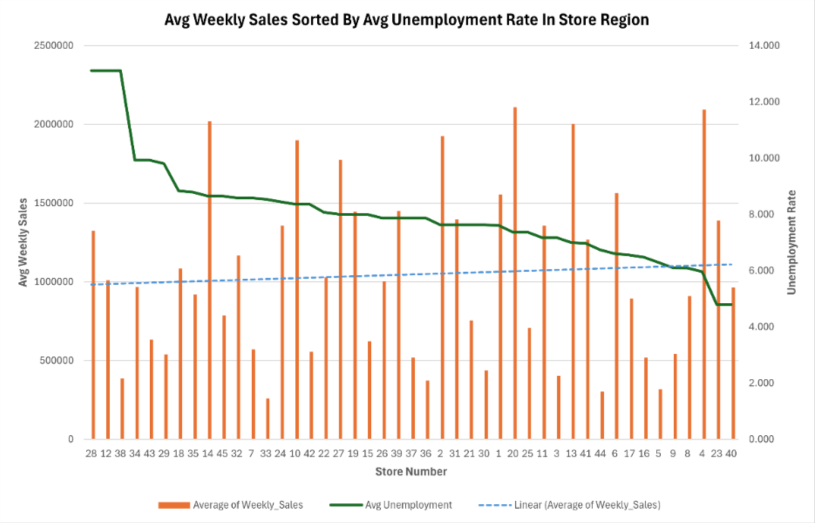


Figure 2: Store Sales Ordered by Regional Unemployment Rate, Highest to Lowest

3. *"Is there any correlation between CPI and Weekly Sales? How does the correlation differ when the Holiday Flag is 0 versus when the Holiday Flag is 1?"*

- The analysis of **Consumer Price Index (CPI)** data similarly shows no meaningful relationship with **weekly sales**. While CPI gradually rises over time, sales remain relatively stable, with the most notable fluctuations occurring around major holidays like **Black Friday** and **Christmas**, rather than being driven by inflationary pressures.
- This suggests that customers' spending habits are less influenced by overall inflation and more by seasonal factors or specific promotions tied to holidays.
- For instance, even when CPI increases, consumers seem willing to spend, especially around the holiday season, underscoring the importance of targeted sales events over the broader economic environment.

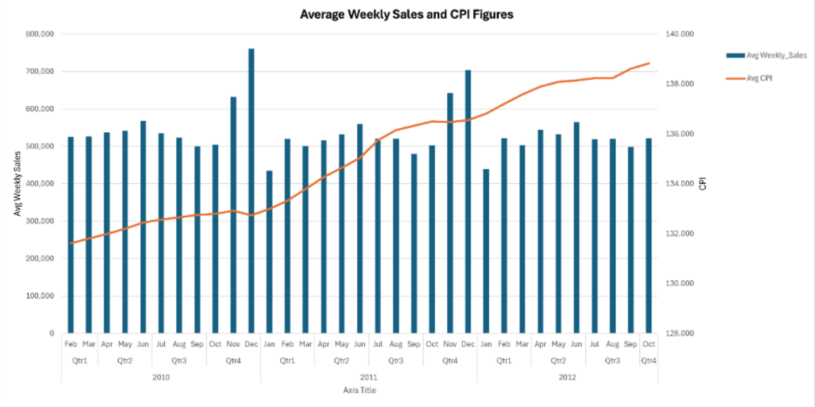


Figure 3: CPI vs. Weekly Sales Over Time

4. *"Why do you think Fuel Price is included in this dataset? What conclusions can be made about Fuel Price compared to any of the other fields?"*

- The inclusion of **fuel price data** in the dataset might suggest an assumption that fuel costs affect consumers' ability to shop, particularly in regions where long commutes are necessary. However, the analysis shows that **fuel price has little to no impact on store performance**.
- Regression analysis returned an **R^2 value close to zero**, confirming the absence of a relationship between fuel price and sales. Even when stores were grouped by **fuel price** and **temperature**, no clear pattern emerged to suggest that fuel price was a significant factor.

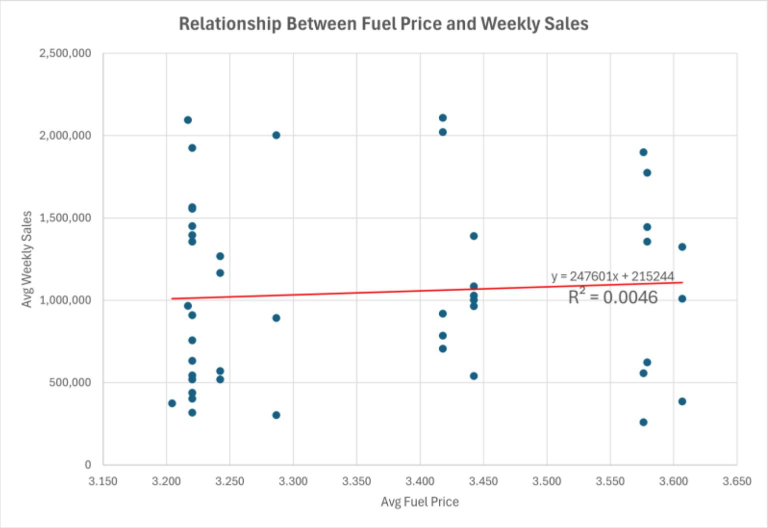
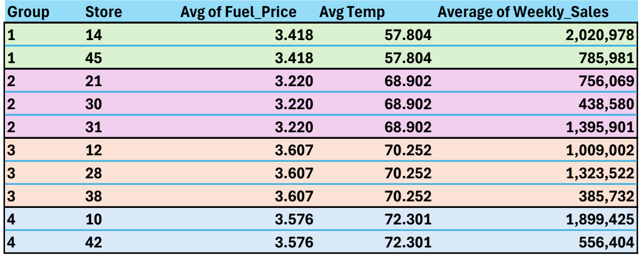


Figure 4: Regression Analysis of Fuel Price Correlation with Weekly Sales

- For instance, stores in regions with similar fuel prices had vastly different sales figures, demonstrating that local fuel costs are not a key driver of retail performance for this chain.



Group	Store	Avg of Fuel_Price	Avg Temp	Average of Weekly_Sales
1	14	3.418	57.804	2,020,978
1	45	3.418	57.804	785,981
2	21	3.220	68.902	756,069
2	30	3.220	68.902	438,580
2	31	3.220	68.902	1,395,901
3	12	3.607	70.252	1,009,002
3	28	3.607	70.252	1,323,522
3	38	3.607	70.252	385,732
4	10	3.576	72.301	1,899,425
4	42	3.576	72.301	556,404

Figure 5: Monthly Store Sales by Fuel Price Grouping

CONCLUSIONS

- The dataset provides valuable insights into **holiday sales trends**, with **Black Friday** and **Christmas week** standing out as the primary revenue drivers. These holidays significantly boost sales, while other holidays—such as **Mother's Day**, **Easter**, **Labor Day**, and **Memorial Day—**produce only marginal increases or no noticeable impact on sales at all. This suggests that the store's product offerings and customer base may not align with holiday-specific shopping trends outside of the major end-of-year events.
- Attempts to correlate **store performance** with **macroeconomic factors—**such as **unemployment rates**, **fuel prices**, and **CPI—**were largely unsuccessful. The analysis showed no strong connection between these variables and retail sales, suggesting that external economic conditions do not play a significant role in driving store-level performance.
    - For instance, despite significant fluctuations in regional unemployment, store sales remained largely unaffected, and the same held true for variations in fuel price and inflation.
    - While external factors might influence consumer spending in other industries or types of businesses, they appear to be less relevant in the retail sector reflected in this dataset.
- **Key takeaway**: To understand the **discrepancies in store performance**, executives and decision-makers should focus on **internal factors** such as **inventory management**, **customer demographics**, **marketing strategies**, and **staffing levels**. Data related to supply chain efficiency, customer preferences, and local competition would likely provide more actionable insights than macroeconomic indicators, which have shown little explanatory power in this context.
- The analysis concludes that **macroeconomic explanations** for why certain stores outperform others are not well-supported by the data. Instead, the focus should shift toward operational and strategic factors that can be directly controlled and optimized within the company.

METHODOLOGY

Data Cleaning

For this analysis, I chose to use **Excel** as the primary tool. While more powerful tools like **SQL** or **Python** could be used, Excel is sufficient here. SQL is best for data extraction once the dataset is clean, and using Python would be overkill given the scope of this project. Excel strikes a balance by allowing for complex conditional formulas, quick data visualization, and ease of use with pivot tables and charts.

The dataset was relatively clean, with no missing fields. However, the **date formatting** was inconsistent, with some rows using the European **DD-MM-YYYY** style and others using **DD/MM/YYYY**. I converted all dates to the **US format (MM-DD-YYYY)** and removed formulas used during cleaning for clarity.

The key issue was the **"Holiday_Flag"** column, which lacked a clear definition of what constituted a holiday and contained inaccuracies. After thorough review, I identified the following holidays as relevant to the analysis:

- Valentine’s Day
- Presidents’ Day
- Easter
- Mother’s Day
- Memorial Day
- Father’s Day
- Independence Day
- Labor Day
- Halloween
- Black Friday
- Christmas (week leading up)
- New Year’s Eve/Day

These holidays were chosen based on their potential impact on retail sales. I created additional columns in the workbook to flag each holiday, marking occurrences with a "1" and all other weeks with a "0." The **"Holiday_Name"** column simplifies filtering, while a new **"Holiday_Flag"** column replaces the original, accurately indicating holiday weeks.

Analysis

Pivot tables were constructed to address each of the core business questions, allowing for quick filtering and visualization of the data. While the tables themselves provide valuable insights, visualizing the data through charts makes patterns and conclusions more apparent and actionable.

The slide deck with visualizations can be viewed here: https://www.canva.com/design/DAGQWefRyb0/--6YXH912mTGPjMm1a-4tQ/edit?utm_content=DAGQWefRyb0&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton.

Data Cleaning and Preparation

Dates:

To standardize the date format, I created a new column with a custom format of **"mm-dd-yyyy"**. Using the formula =B2, I copied the date from column B into the new column and applied it throughout the dataset. Once the conversion was complete, I copied the entire workbook by value to a fresh Excel file and removed the original date column (B).

Holiday Identification:

Holiday identification was managed through a series of complex Excel formulas designed to check if a holiday occurred within the 7-day period starting from the date in column B. Except for Easter, variable-date holidays were identified by checking if the date aligned with the corresponding weekday of the observed holiday.

Valentine's Day

=IF(AND(MONTH(B2)=2,DAY(B2)<=14,DAY(B2)+6>=14),1,0)

Presidents' Day

=IF(AND(MONTH(B2)=2,DAY(B2)+6>=15+7-WEEKDAY(DATE(YEAR(B2),2,15)),DAY(B2)<=21-WEEKDAY(DATE(YEAR(B2),2,15))),1,0)

Easter

Easter posed a challenge due to its lunar calendar dependency. Since we're only dealing with three occurrences in the dataset, I opted to manually screen for the actual dates instead of relying on algorithmic methods like the **Gaussian algorithm**, which did not yield consistent results in this case.

=IF(OR(AND(DATE(2010,4,4)>=B2, DATE(2010,4,4)<=B2+6),AND(DATE(2011,4,24)>=B2, DATE(2011,4,24)<=B2+6),AND(DATE(2012,4,8)>=B2, DATE(2012,4,8)<=B2+6)), 1, 0)

Mother's Day

=IF(AND(MONTH(B2)=5,OR(WEEKDAY(DATE(YEAR(B2),5,1))=1,WEEKDAY(DATE(YEAR(B2),5,1))>2),DAY(B2)<=14-WEEKDAY(DATE(YEAR(B2),5,1)),DAY(B2)+6>=14-WEEKDAY(DATE(YEAR(B2),5,1))),1,0)

Memorial Day

=IF(AND(MONTH(B2)=5,OR(WEEKDAY(B2)=2,WEEKDAY(B2)=3,WEEKDAY(B2)=4,WEEKDAY(B2)=5,WEEKDAY(B2)=6,WEEKDAY(B2)=7,WEEKDAY(B2)=1), DAY(B2)+6>=31-WEEKDAY(DATE(YEAR(B2),5,31),2)),1,0)

Father's Day

=IF(AND(MONTH(B2)=6,DAY(B2)+6>=15+7-WEEKDAY(DATE(YEAR(B2),6,15)), DAY(B2)<=21-WEEKDAY(DATE(YEAR(B2),6,15))),1, 0)

Independence Day

=IF(AND(MONTH(B2)=7,DAY(B2)<=4,DAY(B2)+6>=4),1,0)

Labor Day

=IF(OR(AND(DATE(YEAR(B2),9,1)+MOD(7-WEEKDAY(DATE(YEAR(B2),9,1))+1,7)>=B2,DATE(YEAR(B2),9,1)+MOD(7-WEEKDAY(DATE(YEAR(B2),9,1))+1,7)<=B2+6),AND(DATE(YEAR(B2)+1,9,1)+MOD(7-WEEKDAY(DATE(YEAR(B2)+1,9,1))+1,7)>=B2,DATE(YEAR(B2)+1,9,1)+MOD(7-WEEKDAY(DATE(YEAR(B2)+1,9,1))+1,7)<=B2+6)),1,0)

Halloween

=IF(AND(MONTH(B2)=10,DAY(B2)<=31,DAY(B2)+6>=31),1,0)

Black Friday

=IF(AND(MONTH(B2)=11,DAY(B2)+6>=22-WEEKDAY(DATE(YEAR(B2),11,1))+6,DAY(B2)<=22-WEEKDAY(DATE(YEAR(B2),11,1))+6),1,0)

Christmas Eve

=IF(AND(MONTH(B2)=12,DAY(B2)<=24,DAY(B2)+6>=24),1,0)

New Year’s Day

=IF(OR(AND(MONTH(B2)=12,DAY(B2)+6>31),AND(MONTH(B2)=1,DAY(B2)<=1)),1,0)