Data Exploration and Econometric Analysis of Walmart Sales

AUTHOR
Vaishnavi Thorat

Introduction

This project analyzes Walmart weekly sales data to understand how temperature affects store-level sales. The primary research question is:

Do higher temperatures causally affect weekly sales?

Because temperature is determined by weather conditions, it is plausibly outside the control of individual stores. In the regression analysis, I include store and date fixed effects to control for permanent store differences and common time shocks. Therefore, the estimated effect is identified from within-store changes in temperature over time, holding constant permanent store characteristics and common time shocks.

Data Preparation

To begin the analysis, I load the necessary R packages for data cleaning (tidyverse), date handling (lubridate), and fixed effects regression estimation (fixest). I then load the Walmart sales dataset using a relative file path so that the analysis can be reproduced on another machine. The Date variable is converted into a proper date format to allow time-based analysis, and Store is treated as a categorical variable because it represents distinct store identifiers. Holiday_Flag is a binary indicator (0/1) capturing holiday weeks, which directly affect sales. ## Load data (relative path)

```
library(tidyverse)
library(lubridate)
library(fixest)


df <- read_csv("data/Walmart_Sales.csv", show_col_types = FALSE)

# Convert Date (day-month-year)
df <- df %>%
  mutate(
    Date = dmy(Date),
    Store = factor(Store),
    Holiday_Flag = factor(Holiday_Flag)
  )

# Quick checks
dim(df)
```

```
[1] 6435    8
```

```
names(df)
```

```
[1] "Store"        "Date"         "Weekly_Sales" "Holiday_Flag" "Temperature"
[6] "Fuel_Price"   "CPI"          "Unemployment"
```

```
glimpse(df)
```

```
Rows: 6,435
Columns: 8
$ Store        <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ Date         <date> 2010-02-05, 2010-02-12, 2010-02-19, 2010-02-26, 2010-03-…
$ Weekly_Sales <dbl> 1643691, 1641957, 1611968, 1409728, 1554807, 1439542, 147…
$ Holiday_Flag <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Temperature  <dbl> 42.31, 38.51, 39.93, 46.63, 46.50, 57.79, 54.58, 51.45, 6…
$ Fuel_Price   <dbl> 2.572, 2.548, 2.514, 2.561, 2.625, 2.667, 2.720, 2.732, 2…
$ CPI          <dbl> 211.0964, 211.2422, 211.2891, 211.3196, 211.3501, 211.380…
$ Unemployment <dbl> 8.106, 8.106, 8.106, 8.106, 8.106, 8.106, 8.106, 8.106, 7…
```

```
summary(df$Date)
```

```
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
"2010-02-05" "2010-10-08" "2011-06-17" "2011-06-17" "2012-02-24" "2012-10-26"
```

```
sum(is.na(df$Date))
```

```
[1] 0
```

The dataset contains 6,435 weekly observations and 8 variables. The date conversion was successful, and no missing values were introduced during the transformation. The data is now correctly structured and ready for exploratory analysis and regression modeling.

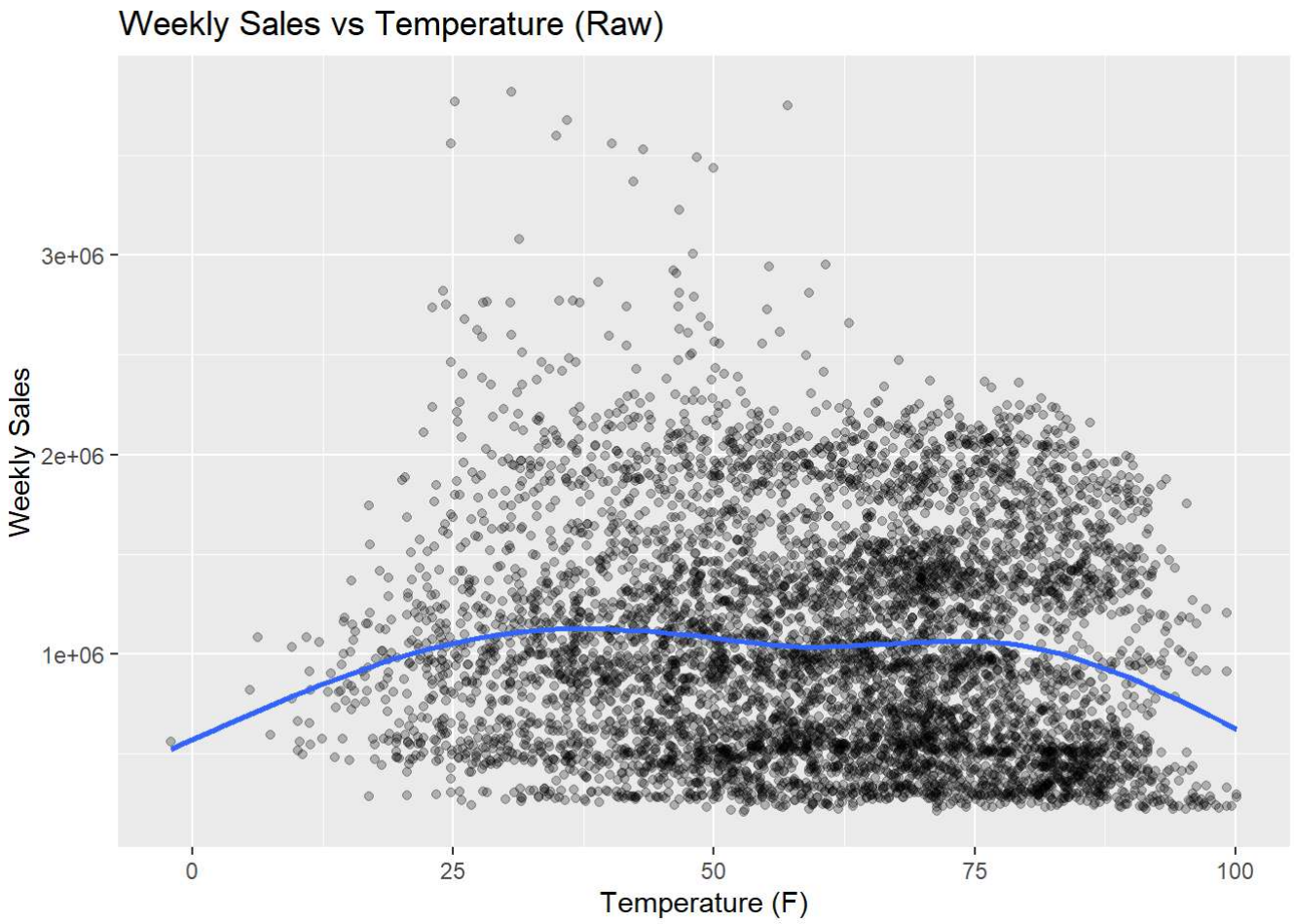
Exploratory Analysis: Temperature and Sales

Before estimating regression models, I first visualize the relationship between temperature and weekly sales. Graphical analysis helps determine whether the relationship appears linear or nonlinear, which is important for choosing the correct functional form in the regression model.

Graph 1: raw relationship (helps decide linear vs nonlinear)

```
ggplot(df, aes(x = Temperature, y = Weekly_Sales)) +
  geom_point(alpha = 0.25) +
  geom_smooth(se = FALSE) +
  labs(
    title = "Weekly Sales vs Temperature (Raw)",
```

```
    x = "Temperature (F)",
    y = "Weekly Sales"
  )
```



The plot suggests a nonlinear (inverted U-shaped) relationship between temperature and sales. Sales appear to increase as temperatures rise from low levels, but decline when temperatures become very high. However, this relationship is descriptive and may reflect other factors such as seasonality or differences across stores. The visual evidence suggests that a quadratic temperature term may be appropriate in the regression model.

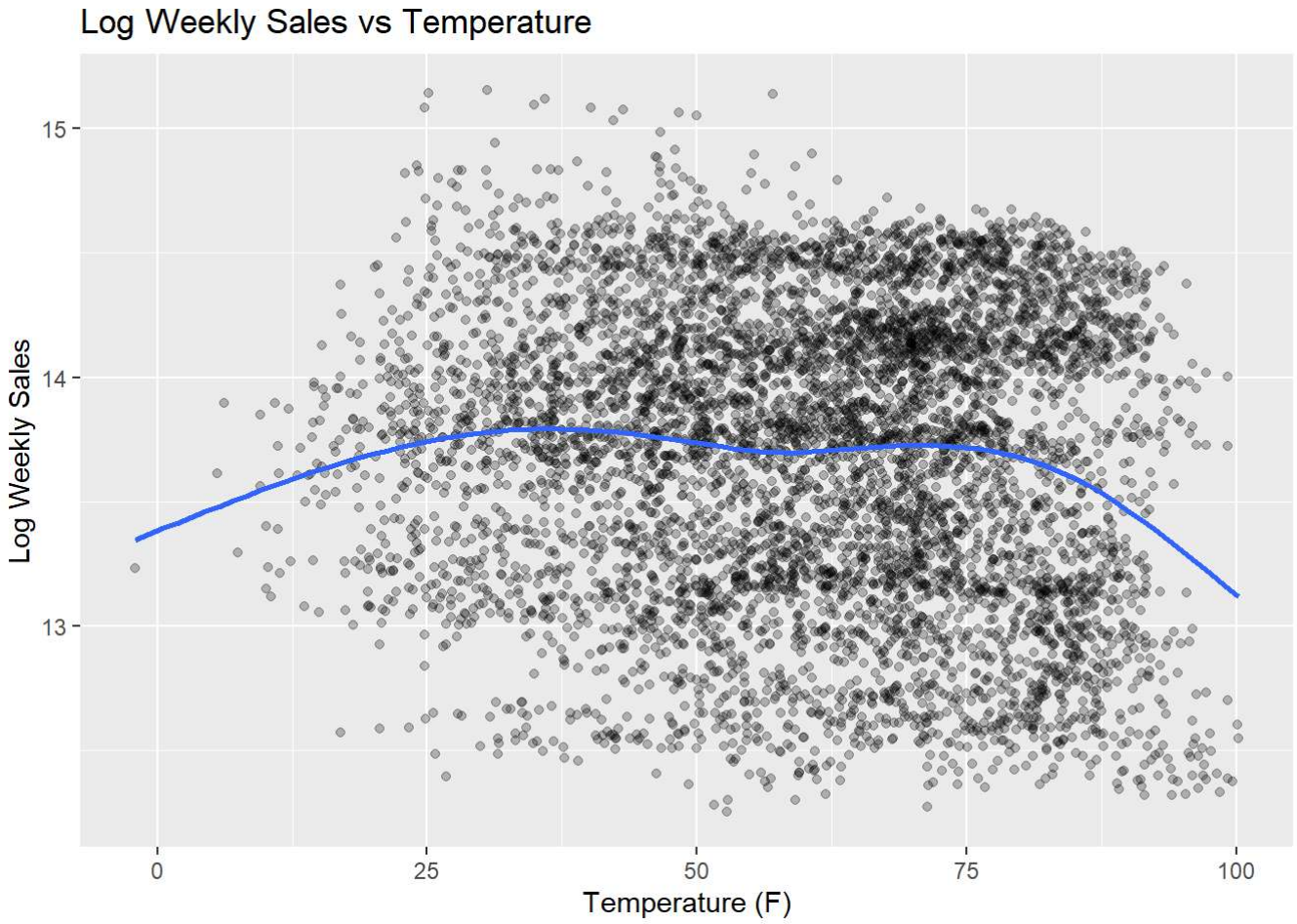
Log Transformation and Functional Form

Because weekly sales vary substantially across stores and contain large numeric values, I transform sales using the natural logarithm. Logging the dependent variable allows coefficients to be interpreted in approximate percentage terms and often stabilizes variance. I then re-examine the relationship between temperature and log sales to determine whether the nonlinear pattern persists.

```
df <- df %>%
  mutate(log_sales = log(Weekly_Sales))

# Plot Temperature vs log(Sales)
ggplot(df, aes(x = Temperature, y = log_sales)) +
```

```
  geom_point(alpha = 0.25) +
  geom_smooth(se = FALSE) +
  labs(
    title = "Log Weekly Sales vs Temperature",
    x = "Temperature (F)",
    y = "Log Weekly Sales"
  )
```



The log-scale plot continues to show a nonlinear relationship between temperature and sales. The inverted U-shaped pattern remains visible, indicating that a quadratic temperature specification is appropriate even after transformation. Therefore, the regression model will include both Temperature and Temperature^2.

Constructing the Quadratic Term

To capture the nonlinear relationship observed in the plots, I construct a squared temperature variable. Including both Temperature and Temperature^2 in the regression allows the marginal effect of temperature to vary across different temperature levels and formally test for an inverted U-shaped relationship.

```
df <- df %>%
  mutate(temp_sq = Temperature^2)
```

Baseline Linear Specification

I begin with a simple linear regression of log weekly sales on temperature. This baseline model provides an initial estimate of the relationship, but it does not yet control for store-level differences or seasonality. Therefore, it should be interpreted cautiously.

```
m1 <- feols(log_sales ~ Temperature, data = df)

summary(m1)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Standard-errors: IID
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 13.880921   0.025106 552.89392  < 2.2e-16 ***
Temperature -0.002951   0.000396  -7.45254 1.0359e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.585733   Adj. R2: 0.008406
```

The linear model estimates a statistically significant negative relationship between temperature and log sales. This implies that, on average, higher temperatures are associated with lower sales in this simple specification. However, this model assumes a constant linear effect, which contradicts the nonlinear pattern observed in the earlier graphs. Therefore, I next estimate a quadratic specification that better matches the visual evidence.

Quadratic Specification

Based on the graphs, the relationship between temperature and sales appears nonlinear. I therefore estimate a quadratic model including both Temperature and Temperature^2 to capture a possible inverted U-shaped pattern.

```
m2 <- feols(log_sales ~ Temperature + temp_sq, data = df)

summary(m2)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Standard-errors: IID
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 13.515732   0.061516 219.70944  < 2.2e-16 ***
Temperature  0.011236   0.002219   5.06472 4.2039e-07 ***
temp_sq     -0.000123   0.000019  -6.49850 8.7203e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.583819   Adj. R2: 0.014721
```

Temperature is positive and significant, while Temperature^2 is negative and significant. This suggests sales increase with temperature at lower level, reach a peak, and then decline at higher temoeratures.

Store Fixed Effects Model

Next, I include Store fixed effects to control for time-invariant differences across stores, such as location, size, or customer base. This helps isolate the effect of temperature within the same store over time.

```
m3 <- feols(log_sales ~ Temperature + temp_sq | Store, data = df)

summary(m3)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Fixed-effects: Store: 45
Standard-errors: IID
             Estimate Std. Error  t value   Pr(>|t|)
Temperature  0.001689 0.00047459  3.55790 3.7654e-04 ***
temp_sq     -0.000019 0.00000408 -4.61248 4.0565e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.121406     Adj. R2: 0.957099
                 Within R2: 0.007119
```

After controlling for store fixed effects, Temperature remains positive and significant, and Temperature^2 remains negative and significant. This suggests the nonlinear relationship is not driven by differences across stores but holds within stores over time.

Two-Way Fixed Effects Model

Next, I include both Store and Date fixed effects. Store fixed effects control for permanent differences across stores, while Date fixed effects control for common shocks affecting all stores in a given week (such as seasonality or macroeconomic conditions). This specification strengthens the causal interpretation.

```
m4 <- feols(log_sales ~ Temperature + temp_sq | Store + Date, data = df)

summary(m4)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Fixed-effects: Store: 45,  Date: 143
Standard-errors: IID
             Estimate Std. Error  t value  Pr(>|t|)
Temperature  0.004821 0.00040739 11.83272 < 2.2e-16 ***
temp_sq     -0.000030 0.00000332 -8.92857 < 2.2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.082353     Adj. R2: 0.979812
                 Within R2: 0.022011
```

Even after controlling for store and time fixed effects, Temperature remains positive and significant, and Temperature^2 remains negative and significant. This suggests a robust nonlinear relationship between temperature and sales that holds within stores over time.

Clustered Standard Errors

To account for potential correlation of errors within stores over time, I cluster standard errors at the store level. This provides more reliable inference when observations within the same store may not be independent.

```
m5 <- feols(
  log_sales ~ Temperature + temp_sq | Store + Date,
  data = df,
  cluster = ~Store
)

summary(m5)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Fixed-effects: Store: 45,  Date: 143
Standard-errors: Clustered (Store)
             Estimate Std. Error  t value   Pr(>|t|)
Temperature  0.004821 0.00089482  5.38717 2.6596e-06 ***
temp_sq     -0.000030 0.00000804 -3.69281 6.0921e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.082353     Adj. R2: 0.979812
                 Within R2: 0.022011
```

After clustering at the store level, Temperature and Temperature^2 remain statistically significant. This indicates that the nonlinear temperature effect is robust to within-store error correlation.

Turning Point Calculation

To interpret the quadratic model, I calculate the turning point of the temperature-sales relationship using the formula $-\beta_1 / (2\beta_2)$. This identifies the temperature level at which sales reach their maximum.

```
beta1 <- coef(m5)["Temperature"]
beta2 <- coef(m5)["temp_sq"]

turning_point <- -beta1 / (2 * beta2)
```

```
turning_point
```

```
Temperature
   81.19951
```

The estimated turning point is approximately 81°F. This means weekly sales increase with temperature up to about 81 degrees, after which sales begin to decline. This confirms the inverted U-shaped relationship.

Full Model with Controls

> Finally, I add economic control variables (Fuel Price, CPI, Unemployment, and Holiday Flag) to account for other factors that may influence sales. This helps reduce omitted variable bias and improves the credibility of the estimated temperature effect.

```
m6 <- feols(
  log_sales ~ Temperature + temp_sq + Fuel_Price + CPI + Unemployment + Holiday_Flag | Store + Da
  data = df,
  cluster = ~Store
)

summary(m6)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Fixed-effects: Store: 45,  Date: 143
Standard-errors: Clustered (Store)
              Estimate Std. Error   t value   Pr(>|t|)
Temperature   0.005190 0.00091700  5.660299 1.0628e-06 ***
temp_sq      -0.000033 0.00000786 -4.168070 1.4155e-04 ***
Fuel_Price   -0.004768 0.02116116 -0.225302 8.2279e-01
CPI          -0.000294 0.00551947 -0.053329 9.5771e-01
Unemployment -0.026207 0.01252625 -2.092137 4.2229e-02 *
... 1 variable was removed because of collinearity (Holiday_Flag1)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.081678     Adj. R2: 0.980131
                 Within R2: 0.037967
```

Even after adding control variables, Temperature remains positive and significant, and Temperature^2 remains negative and significant. This suggests the nonlinear temperature effect on sales is robust. Among the controls, Unemployment has a negative and significant effect, while Fuel Price and CPI are not statistically significant.

Two-Way Clustered Robust Model

Finally, I cluster standard errors at both the Store and Date levels. This accounts for correlation of errors within stores over time and common shocks across stores in the same week, providing the most robust inference.

```
m7 <- feols(
  log_sales ~ Temperature + temp_sq + Fuel_Price + CPI + Unemployment + Holiday_Flag | Store + Da
  data = df,
  cluster = ~Store + Date
)

summary(m7)
```

```
OLS estimation, Dep. Var.: log_sales
Observations: 6,435
Fixed-effects: Store: 45,  Date: 143
Standard-errors: Clustered (Store & Date)
              Estimate Std. Error   t value   Pr(>|t|)
Temperature   0.005190   0.001350  3.845574 0.00038396 ***
temp_sq      -0.000033   0.000011 -3.113932 0.00324262 **
Fuel_Price   -0.004768   0.029267 -0.162904 0.87134015
CPI          -0.000294   0.005464 -0.053876 0.95727820
Unemployment -0.026207   0.012569 -2.085024 0.04290189 *
... 1 variable was removed because of collinearity (Holiday_Flag1)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.081678     Adj. R2: 0.980131
                 Within R2: 0.037967
```

> Even with two-way clustering and full controls, Temperature remains positive and statistically significant, while Temperature^2 remains negative and significant. This confirms a robust nonlinear relationship between temperature and weekly sales.

Model Comparison and Robustness Check

To clearly compare specifications, I present a summary table contrasting the two-way fixed effects model (M4) with the full model including controls and two-way clustered standard errors (M7). This allows evaluation of robustness across specifications.

```
if (!require(modelsummary)) install.packages("modelsummary")
library(modelsummary)

modelsummary(
  list(
    "FE Model" = m4,
    "Full Model (2-way cluster)" = m7
```

```
  ),
  statistic = "({std.error})",
  stars = TRUE
)
```

	FE Model	Full Model (2-way cluster)
Temperature	0.005***	0.005***
	(0.000)	(0.001)
temp_sq	-0.000***	-0.000**
	(0.000)	(0.000)
Fuel_Price		-0.005
		(0.029)
CPI		-0.000
		(0.005)
Unemployment		-0.026*
		(0.013)
Num.Obs.	6435	6435
R2	0.980	0.981
R2 Adj.	0.980	0.980
R2 Within	0.022	0.038
R2 Within Adj.	0.022	0.037
AIC	-13493.3	-13593.2
BIC	-12213.9	-12293.5
RMSE	0.08	0.08
Std.Errors	IID	by: Store & Date
FE: Store	X	X
FE: Date	X	X

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Across both specifications, Temperature remains positive and significant, while Temperature^2 remains negative and significant. The results are stable even after adding controls and two-way clustering, indicating

a robust nonlinear relationship between temperature and weekly sales.

Conclusion

This analysis examined whether temperature affects weekly sales using Walmart store-level panel data. Across multiple specifications including quadratic models, store and date fixed effects, control variables, and clustered standard errors the results consistently show a nonlinear relationship between temperature and sales. Sales increase as temperatures rise up to approximately 81°F, after which sales begin to decline. The stability of the coefficients across models suggests that this nonlinear pattern is robust to different specifications and inference adjustments. Overall, the findings indicate a strong and consistent inverted U-shaped relationship between temperature and weekly sales.