project1\_02\_Analysis

Imaan Ali

2025-08-20

### 1.Loading the data:

The dataset was loaded from an RDS file using readRDS(). The date column was converted to Date format to enable creating time-based features like weekdays, weekends, or seasonal indicators.

#Load dataset from RDS  
df <- readRDS("sales\_top3.rds")  
  
#Inspect the structure of the dataset  
str(df)

## 'data.frame': 124800 obs. of 10 variables:  
## $ date : chr "2012-01-01" "2012-01-01" "2012-01-01" "2012-01-01" ...  
## $ store\_nbr : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ item\_nbr : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ units : int 0 191 214 0 0 0 137 38 131 98 ...  
## $ station\_nbr: int 1 14 7 9 12 14 6 4 17 12 ...  
## $ tavg : int 42 42 42 27 60 42 49 55 47 60 ...  
## $ snowfall : num NA 0 0 NA 0 0 0 0 0 0 ...  
## $ preciptotal: num 0.05 0 0 0.05 0 0 0 0 0 0 ...  
## $ heat : int 23 23 23 38 5 23 16 10 18 5 ...  
## $ cool : int 0 0 0 0 0 0 0 0 0 0 ...

head(df)

## date store\_nbr item\_nbr units station\_nbr tavg snowfall preciptotal  
## 1 2012-01-01 1 5 0 1 42 NA 0.05  
## 2 2012-01-01 2 5 191 14 42 0 0.00  
## 3 2012-01-01 3 5 214 7 42 0 0.00  
## 4 2012-01-01 4 5 0 9 27 NA 0.05  
## 5 2012-01-01 5 5 0 12 60 0 0.00  
## 6 2012-01-01 6 5 0 14 42 0 0.00  
## heat cool  
## 1 23 0  
## 2 23 0  
## 3 23 0  
## 4 38 0  
## 5 5 0  
## 6 23 0

summary(df)

## date store\_nbr item\_nbr units   
## Length:124800 Min. : 1.00 Min. : 5.00 Min. : 0.00   
## Class :character 1st Qu.:12.00 1st Qu.: 5.00 1st Qu.: 0.00   
## Mode :character Median :23.00 Median : 9.00 Median : 0.00   
## Mean :23.09 Mean :19.67 Mean : 22.18   
## 3rd Qu.:34.00 3rd Qu.:45.00 3rd Qu.: 38.00   
## Max. :45.00 Max. :45.00 Max. :5568.00   
##   
## station\_nbr tavg snowfall preciptotal   
## Min. : 1.00 Min. :-16.00 Min. : 0.00 Min. :0.0000   
## 1st Qu.: 7.00 1st Qu.: 48.00 1st Qu.: 0.00 1st Qu.:0.0000   
## Median :12.00 Median : 64.00 Median : 0.00 Median :0.0000   
## Mean :11.32 Mean : 61.18 Mean : 0.02 Mean :0.0793   
## 3rd Qu.:15.00 3rd Qu.: 77.00 3rd Qu.: 0.00 3rd Qu.:0.0500   
## Max. :20.00 Max. :100.00 Max. :16.20 Max. :7.3600   
## NA's :5073 NA's :44778 NA's :2673   
## heat cool   
## Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 1.00 Median : 0.000   
## Mean : 9.63 Mean : 5.813   
## 3rd Qu.:17.00 3rd Qu.:12.000   
## Max. :81.00 Max. :35.000   
## NA's :5073 NA's :5073

#convert date column to Date format  
df$date <- as.Date(df$date, format = "%Y-%m-%d")  
  
#Quick check  
head(df$date)

## [1] "2012-01-01" "2012-01-01" "2012-01-01" "2012-01-01" "2012-01-01"  
## [6] "2012-01-01"

### 2.Data Cleanse

Missing values were addressed thoughtfully rather than dropping rows: tavg was imputed with its mean to preserve the overall temperature trend, while snowfall and preciptotal were set to 0 because missing values likely indicate no precipitation.

Extreme units values were capped at the 99th percentile to reduce the influence of rare outliers without removing valid high-sales events. A log transformation of units was added to stabilize variance and mitigate skewness.

New features were created to improve model performance: is\_weekend captures higher weekend demand patterns, is\_rainy\_day and is\_snowy\_day account for weather effects on sales, and month captures seasonal trends. Each measure helps the model better understand factors that influence product demand.

# --- Check missing values ---  
colSums(is.na(df))

## date store\_nbr item\_nbr units station\_nbr tavg   
## 0 0 0 0 0 5073   
## snowfall preciptotal heat cool   
## 44778 2673 5073 5073

# Impute missing tavg with mean  
df$tavg[is.na(df$tavg)] <- mean(df$tavg, na.rm = TRUE)  
  
# Impute missing snowfall and preciptotal with 0  
df$snowfall[is.na(df$snowfall)] <- 0  
df$preciptotal[is.na(df$preciptotal)] <- 0  
  
# Cap extreme values at 99th percentile  
upper <- quantile(df$units, 0.99, na.rm = TRUE)  
df$units[df$units > upper] <- upper  
  
# Add log-transformed version of units  
df$log\_units <- log1p(df$units)  
  
# Check results  
summary(df[, c("units", "log\_units", "tavg", "snowfall", "preciptotal")])

## units log\_units tavg snowfall   
## Min. : 0.00 Min. :0.000 Min. :-16.00 Min. : 0.00000   
## 1st Qu.: 0.00 1st Qu.:0.000 1st Qu.: 49.00 1st Qu.: 0.00000   
## Median : 0.00 Median :0.000 Median : 63.00 Median : 0.00000   
## Mean : 21.68 Mean :1.453 Mean : 61.18 Mean : 0.01045   
## 3rd Qu.: 38.00 3rd Qu.:3.664 3rd Qu.: 76.00 3rd Qu.: 0.00000   
## Max. :158.00 Max. :5.069 Max. :100.00 Max. :16.20000   
## preciptotal   
## Min. :0.00000   
## 1st Qu.:0.00000   
## Median :0.00000   
## Mean :0.07764   
## 3rd Qu.:0.05000   
## Max. :7.36000

# --New feature creation--  
  
# Create weekend indicator  
df$is\_weekend <- ifelse(weekdays(df$date) %in% c("Saturday", "Sunday"), 1, 0)  
  
# Create rainy day indicator  
df$is\_rainy\_day <- ifelse(df$preciptotal > 0, 1, 0)  
  
# Create snowy day indicator  
df$is\_snowy\_day <- ifelse(df$snowfall > 0, 1, 0)  
  
# Extract month from date  
df$month <- as.integer(format(df$date, "%m"))  
  
# Check first few rows  
head(df[, c("date", "is\_weekend", "is\_rainy\_day", "is\_snowy\_day", "month")])

## date is\_weekend is\_rainy\_day is\_snowy\_day month  
## 1 2012-01-01 1 1 0 1  
## 2 2012-01-01 1 0 0 1  
## 3 2012-01-01 1 0 0 1  
## 4 2012-01-01 1 1 0 1  
## 5 2012-01-01 1 0 0 1  
## 6 2012-01-01 1 0 0 1

#These features help the model account for \*\*weekend effects, weather conditions, and seasonality\*\*, all of which are known to influence product sales.

### 3. Model Building:

Separate linear regression models were fitted for each product by subsetting the dataset and regressing units sold on weather and temporal features.

===== Linear Regression Summary for Product: 5 ===== For the first product, the linear regression shows that sales increase by about 0.07 units for every 1°F rise in temperature (tavg). On weekends, sales rise by about 6 units, while rainy days reduce sales by around 2 units. Each unit of snowfall lowers sales by about 2 units, and sales also decline slightly with increasing month values (≈0.3 units fewer per month). Other variables like total precipitation and the snowy day indicator were not significant. Overall, the model highlights that temperature, snowfall, rainfall, weekends, and seasonality all influence sales, though the explained variance is modest.

===== Linear Regression Summary for Product: 9 ===== For the second product, the regression shows that sales decrease by about 0.12 units for every 1°F rise in temperature (tavg), indicating demand falls slightly in warmer weather. Weekends boost sales by about 5.6 units, while snowy days reduce sales sharply by about 9 units. Each unit of snowfall also lowers sales by ≈1.6 units, though this effect is only marginally significant. Other factors such as total precipitation, rainy days, and month were not statistically significant. The model explains only a small share of variation in sales (Adjusted R² ≈ 0.008), but highlights clear negative effects of snow and positive weekend demand.

===== Linear Regression Summary for Product: 45 =====

For the third product, sales decrease by about 0.22 units for every 1°F rise in temperature (tavg), showing a stronger negative relationship with warmer weather compared to the second product. Weekends are associated with a notable increase of about 7 units in sales. Rainy days reduce sales by about 3.3 units, while snowy days reduce them by nearly 8 units. Additional snowfall also lowers sales by about 2 units per unit increase. Total precipitation has a negative effect of about 2.7 fewer units, reinforcing the weather sensitivity of this product. Month shows a small, marginally significant positive effect. Although the overall fit remains modest (Adjusted R² ≈ 0.023), the model highlights that temperature, precipitation, and weekend effects play a clear role in shaping sales.

# Get unique products  
products <- unique(df$item\_nbr)  
  
# Create an empty list to store models  
models <- list()  
  
# Loop through each product and fit a linear regression  
for (p in products) {  
 # Filter dataset for the product  
 df\_product <- df[df$item\_nbr == p, ]  
   
 # Fit linear regression  
 model <- lm(units ~ tavg + preciptotal + snowfall +   
 is\_weekend + is\_rainy\_day + is\_snowy\_day + month,  
 data = df\_product)  
   
 # Store the model in the list  
 models[[p]] <- model  
   
 # Print product name and model summary  
 cat("\n===== Linear Regression Summary for Product:", p, "=====\n")  
 print(summary(model))  
}

##   
## ===== Linear Regression Summary for Product: 5 =====  
##   
## Call:  
## lm(formula = units ~ tavg + preciptotal + snowfall + is\_weekend +   
## is\_rainy\_day + is\_snowy\_day + month, data = df\_product)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -27.15 -19.85 -16.99 14.23 145.18   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.774313 0.600661 27.926 < 2e-16 \*\*\*  
## tavg 0.065508 0.008985 7.291 3.13e-13 \*\*\*  
## preciptotal -0.457056 0.619453 -0.738 0.4606   
## snowfall -2.176662 0.716998 -3.036 0.0024 \*\*   
## is\_weekend 6.044832 0.344349 17.554 < 2e-16 \*\*\*  
## is\_rainy\_day -2.024297 0.354336 -5.713 1.12e-08 \*\*\*  
## is\_snowy\_day 0.951558 1.340476 0.710 0.4778   
## month -0.288914 0.047603 -6.069 1.30e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 31.72 on 41592 degrees of freedom  
## Multiple R-squared: 0.01028, Adjusted R-squared: 0.01011   
## F-statistic: 61.72 on 7 and 41592 DF, p-value: < 2.2e-16  
##   
##   
## ===== Linear Regression Summary for Product: 9 =====  
##   
## Call:  
## lm(formula = units ~ tavg + preciptotal + snowfall + is\_weekend +   
## is\_rainy\_day + is\_snowy\_day + month, data = df\_product)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33.620 -21.559 -18.041 8.617 146.038   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 26.11817 0.71898 36.327 < 2e-16 \*\*\*  
## tavg -0.11939 0.01075 -11.101 < 2e-16 \*\*\*  
## preciptotal -1.06931 0.74148 -1.442 0.1493   
## snowfall -1.61598 0.85824 -1.883 0.0597 .   
## is\_weekend 5.58512 0.41218 13.550 < 2e-16 \*\*\*  
## is\_rainy\_day 0.68271 0.42413 1.610 0.1075   
## is\_snowy\_day -8.91988 1.60453 -5.559 2.73e-08 \*\*\*  
## month 0.09326 0.05698 1.637 0.1017   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.97 on 41592 degrees of freedom  
## Multiple R-squared: 0.008024, Adjusted R-squared: 0.007857   
## F-statistic: 48.06 on 7 and 41592 DF, p-value: < 2.2e-16  
##   
##   
## ===== Linear Regression Summary for Product: 45 =====  
##   
## Call:  
## lm(formula = units ~ tavg + preciptotal + snowfall + is\_weekend +   
## is\_rainy\_day + is\_snowy\_day + month, data = df\_product)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.60 -24.73 -19.06 25.67 141.80   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.48274 0.67143 54.336 < 2e-16 \*\*\*  
## tavg -0.21971 0.01004 -21.876 < 2e-16 \*\*\*  
## preciptotal -2.66687 0.69243 -3.851 0.000118 \*\*\*  
## snowfall -2.02057 0.80147 -2.521 0.011703 \*   
## is\_weekend 7.01946 0.38492 18.236 < 2e-16 \*\*\*  
## is\_rainy\_day -3.32543 0.39608 -8.396 < 2e-16 \*\*\*  
## is\_snowy\_day -7.86500 1.49840 -5.249 1.54e-07 \*\*\*  
## month 0.10130 0.05321 1.904 0.056956 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35.46 on 41592 degrees of freedom  
## Multiple R-squared: 0.02315, Adjusted R-squared: 0.02299   
## F-statistic: 140.8 on 7 and 41592 DF, p-value: < 2.2e-16

saveRDS(df, "sales\_top3\_cleaned.rds")