# Post Graduate Program - Data Science

## In Partnership With Purdue University

### Course Project - Data Science with R

### Retail Analysis with Walmart Data

Submitted by:

Lavkush Singh

Submitted to:

Purdue University – Simplilearn

# Agenda

- Introduction
- Dataset Summary
- Exploratory Data Analysis – Variables
- Variable's Distribution Summary
- Exploratory Data Analysis – Time Series Graphs of Sales
- Correlation Matrix – Symbolic for better interpretation
- Linear Regression Model Summary
- Appendix

# Introduction

- Walmart - leading retail stores in the US
- Historical data provided of sales from 2010-02-05 to 2012-11-01, with various other features
- Dataset Column names - Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment
- Task is to perform basis statistical analysis and building a predictive model using linear Regression

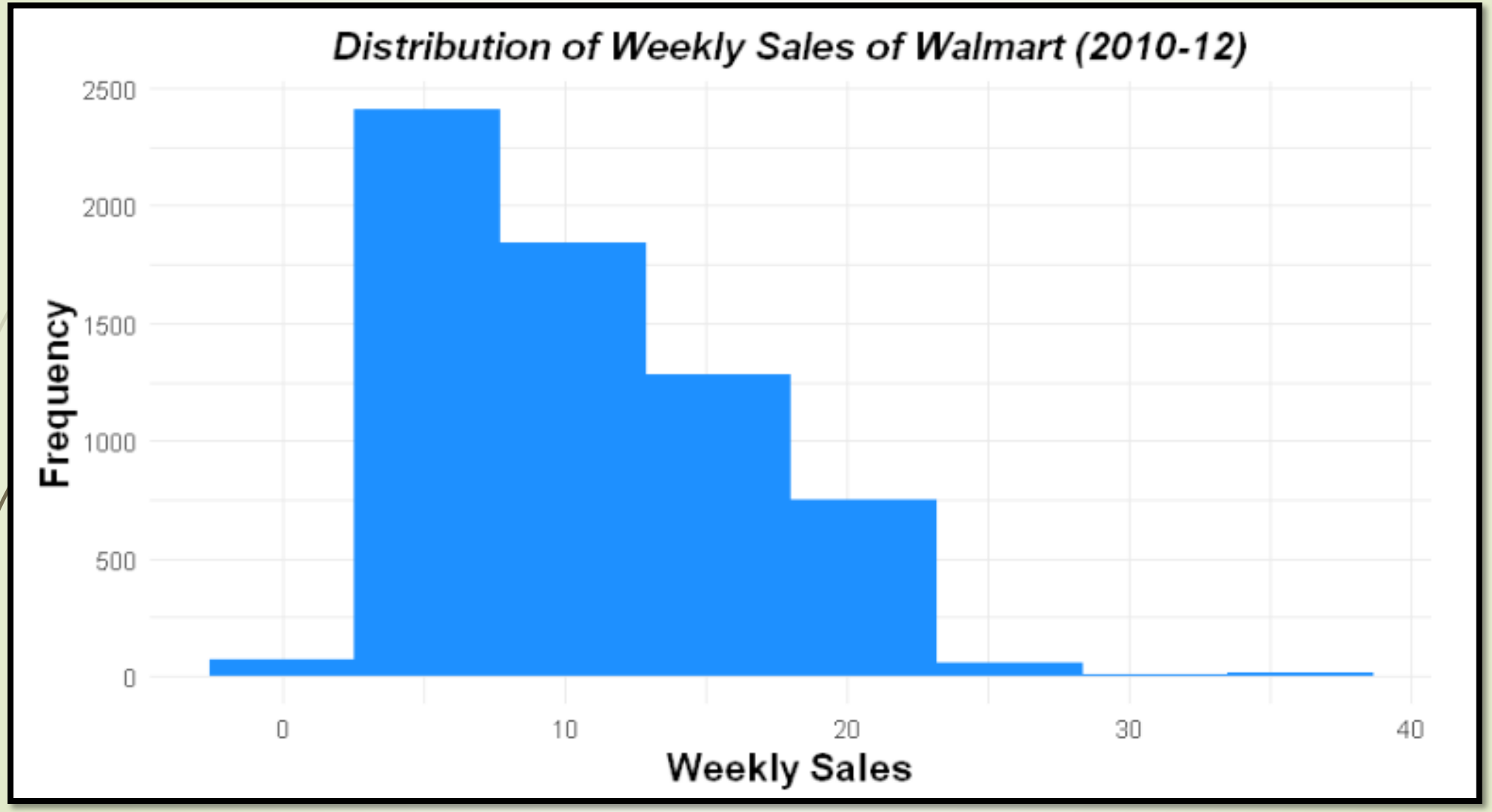
# Dataset Summary

- 6435 observations (rows) of 8 variables (columns)
- No Missing Values
- Data is for 45 stores
- Data is provided for the date range from 2010-02-05 to 2012-11-01
- Holiday flag indicates if the week is a special holiday week, with 1 being holiday week and 0 implies non holiday week
- Most of the variables (columns) are numerical
- Numerical column values varies in scale and range

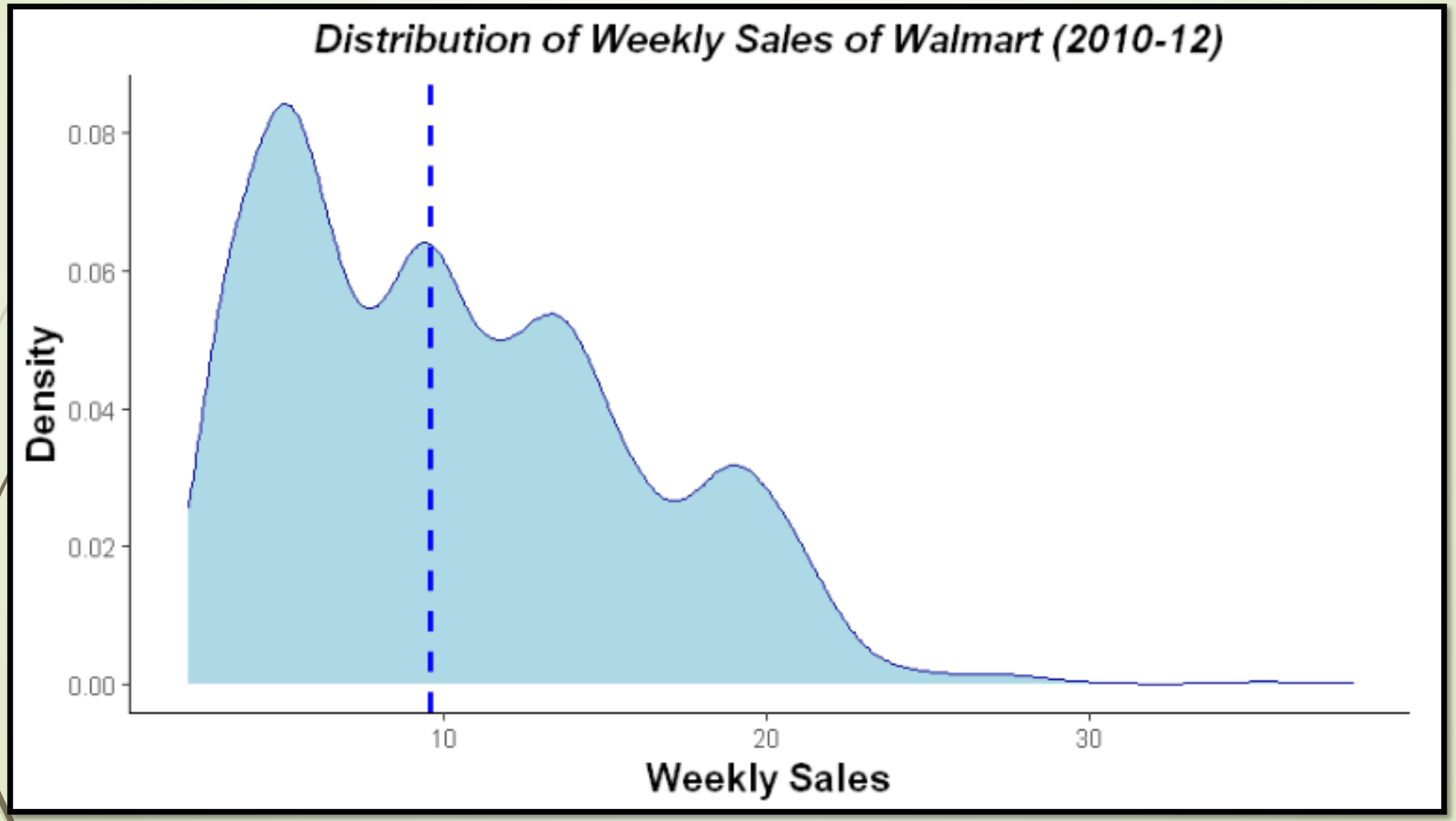
# Exploratory Data Analysis

## Column Variables

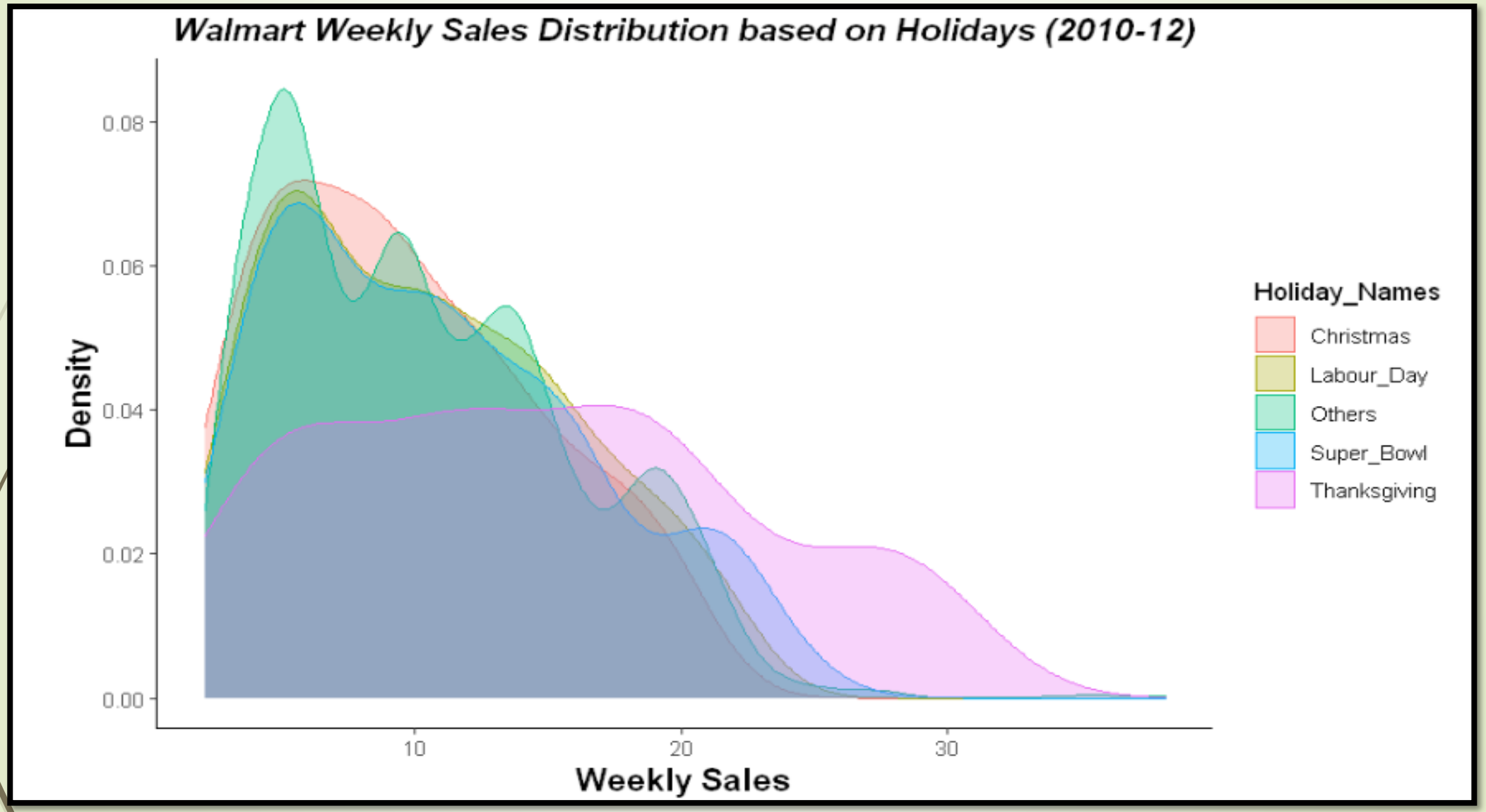
# Exploratory Data Analysis



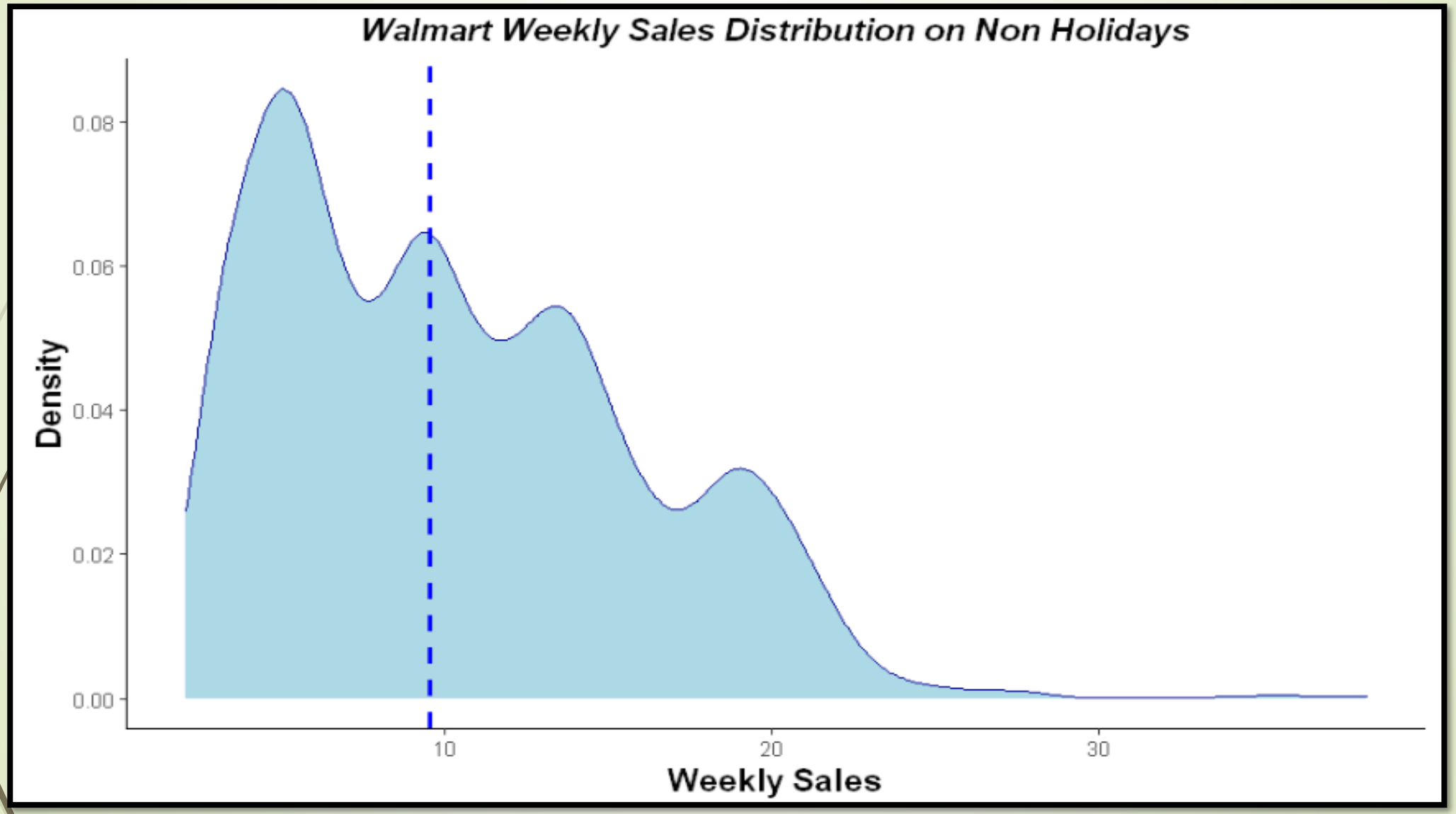
# Exploratory Data Analysis



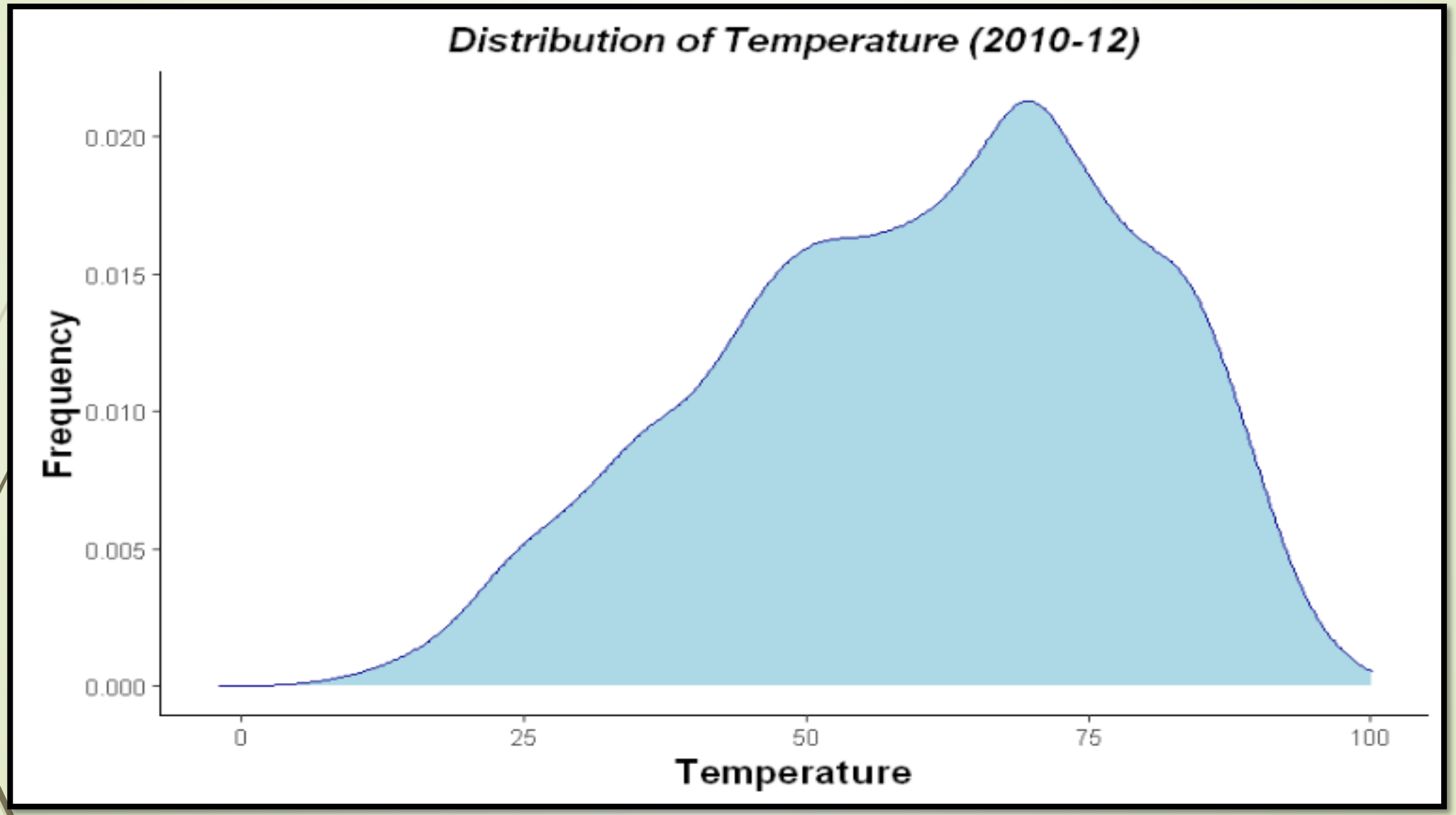
# Exploratory Data Analysis



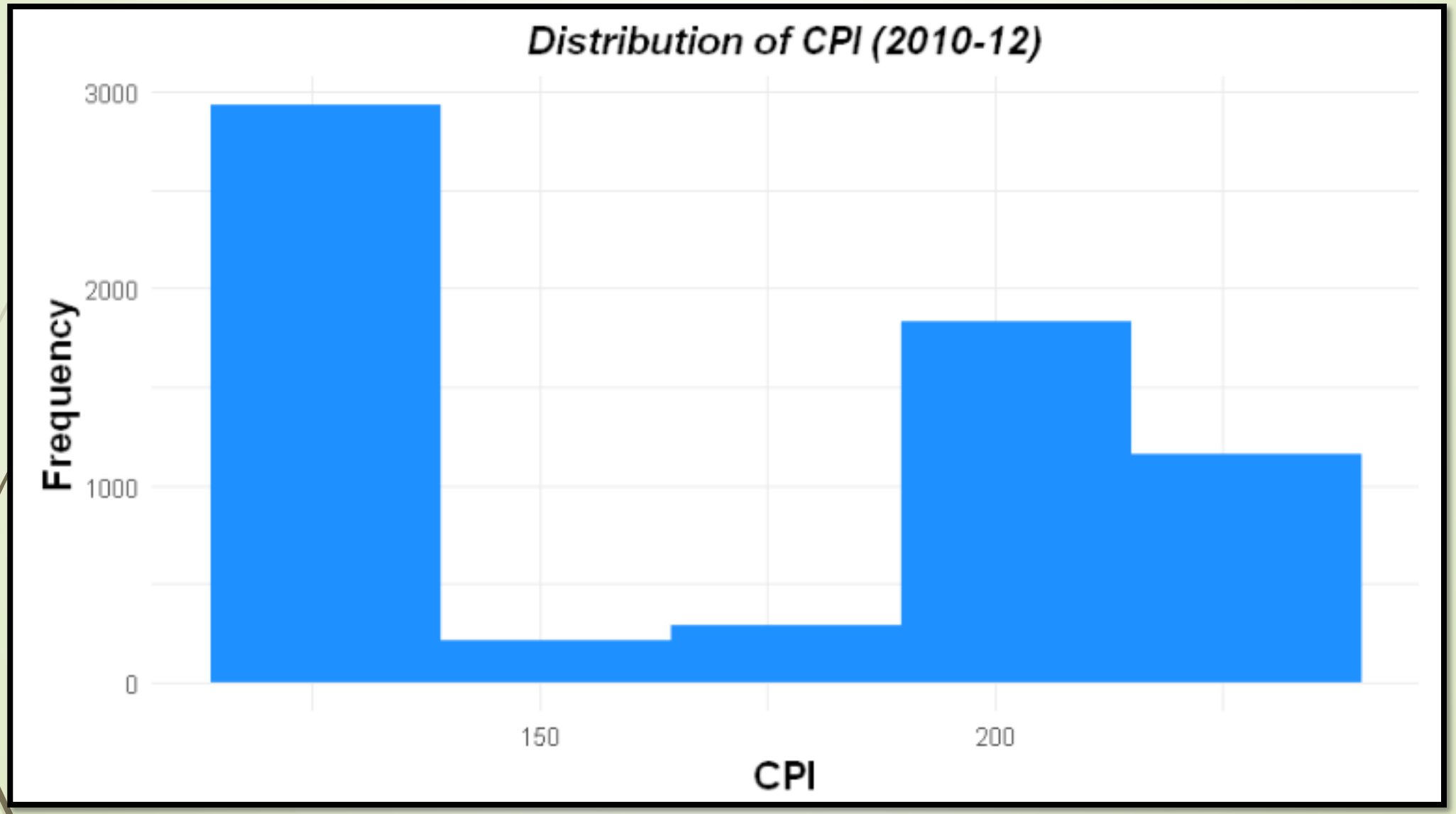
# Exploratory Data Analysis



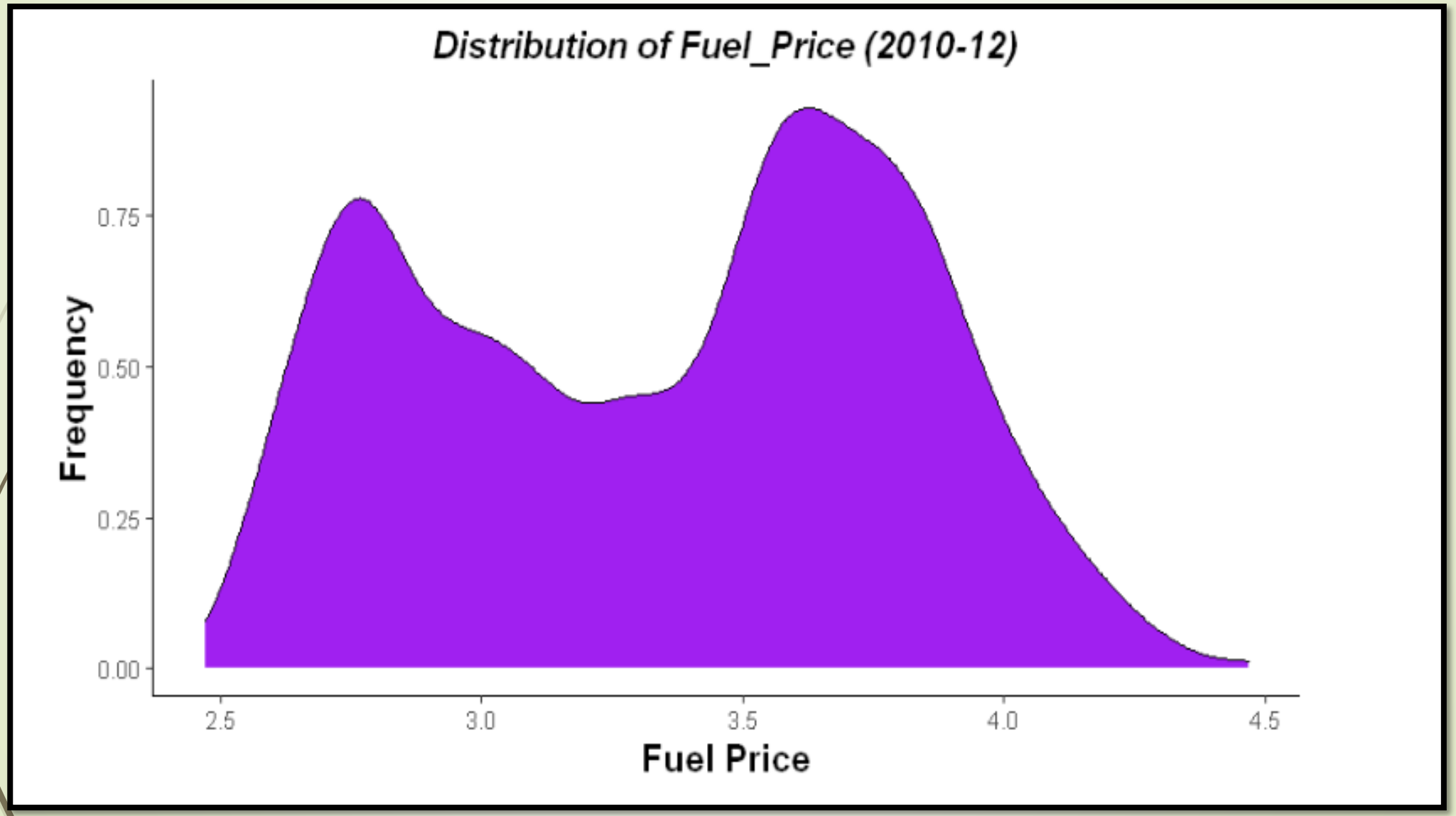
# Exploratory Data Analysis



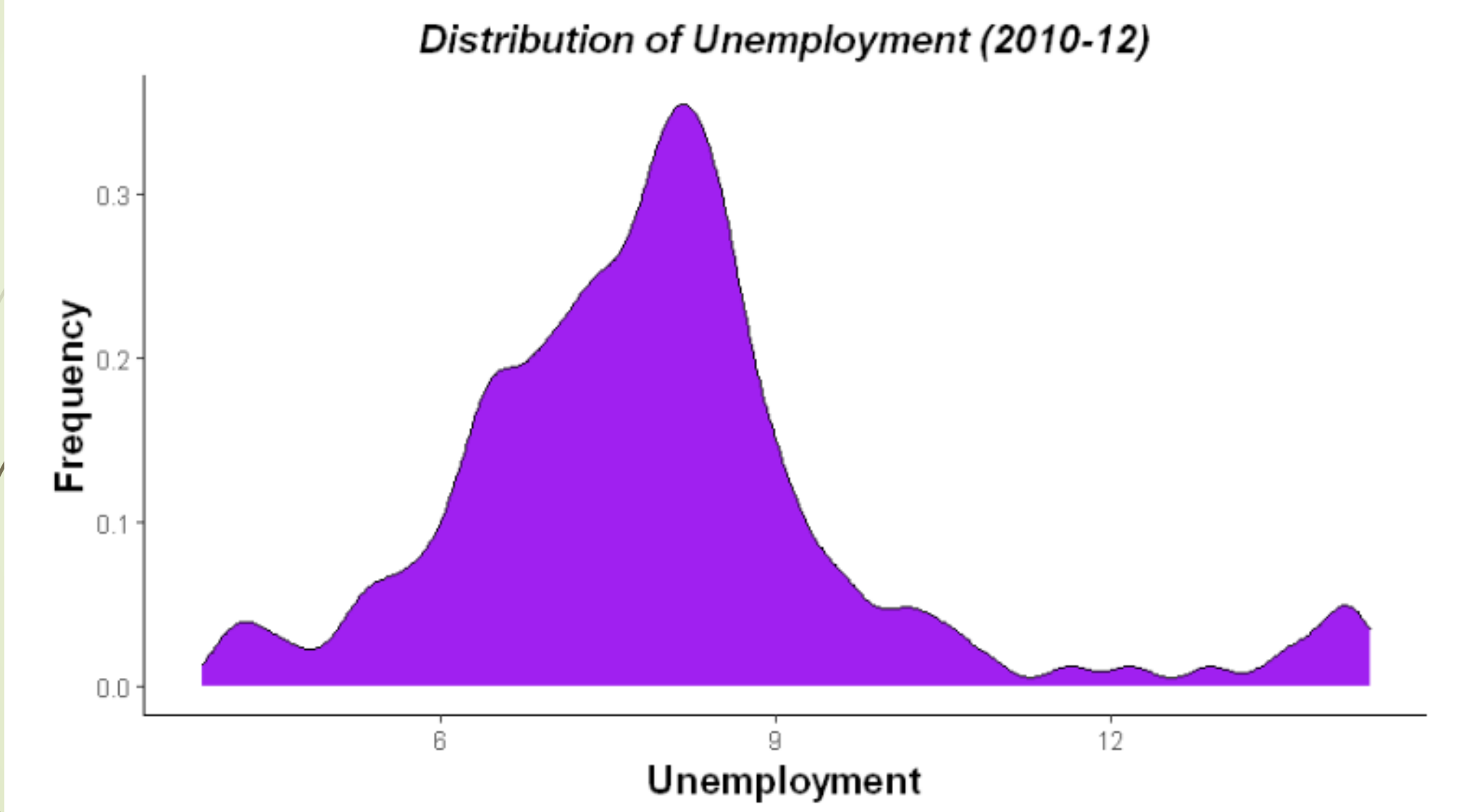
# Exploratory Data Analysis



# Exploratory Data Analysis



# Exploratory Data Analysis



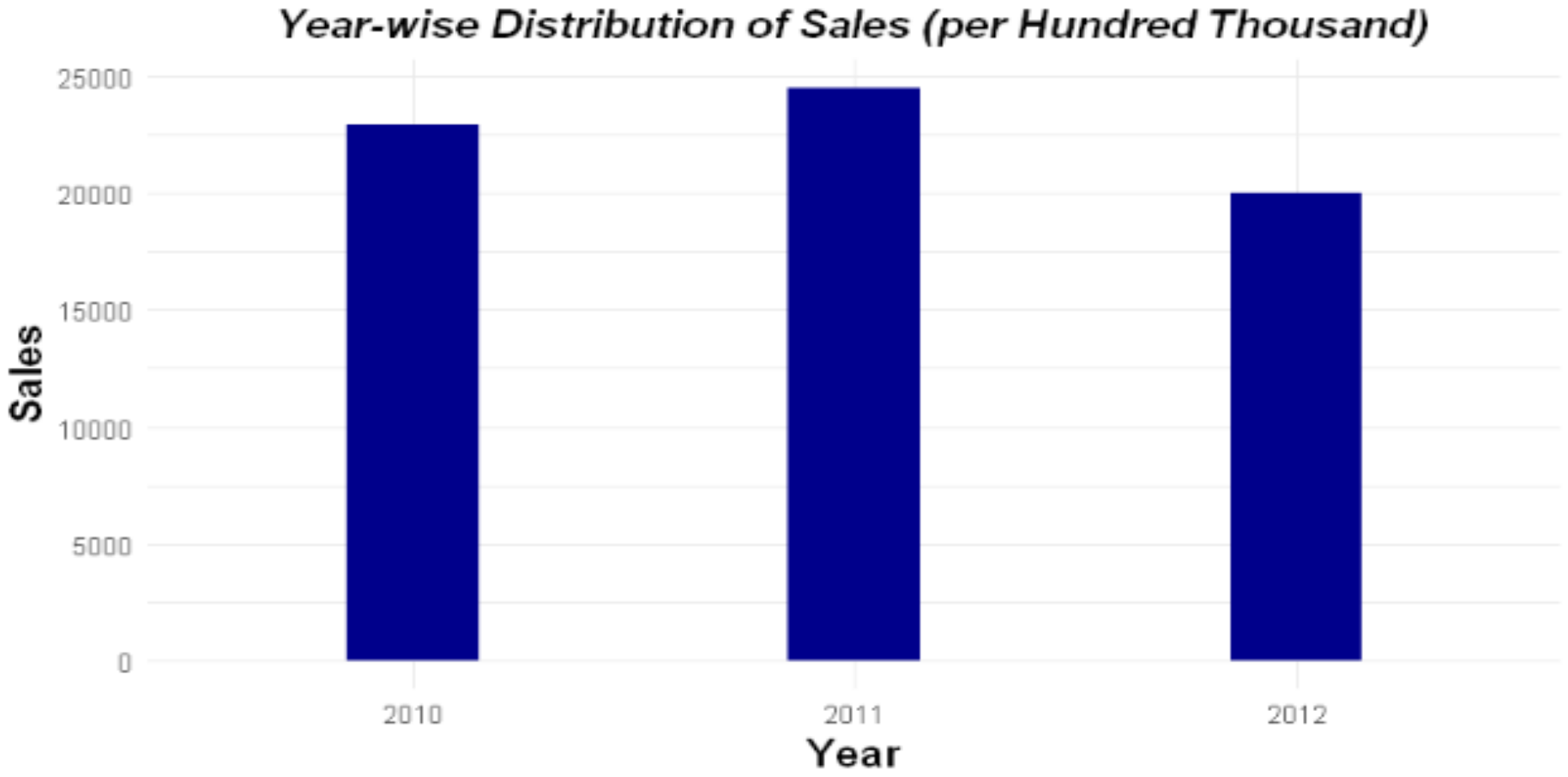
# Variable's Distribution Summary

- Weekly sales of Walmart for Year 2010-12 is not evenly distributed, data is right skewed.
- In all the holiday cases, the Sales distribution is approximately right skewed
- Weekly Sales of Walmart on Non holidays is more or less similar to holiday sales
- It looks like there are few outliers in Temperature column
- Fuel Price is bimodal distribution and Unemployment is approximately normally distributed

# Exploratory Data Analysis

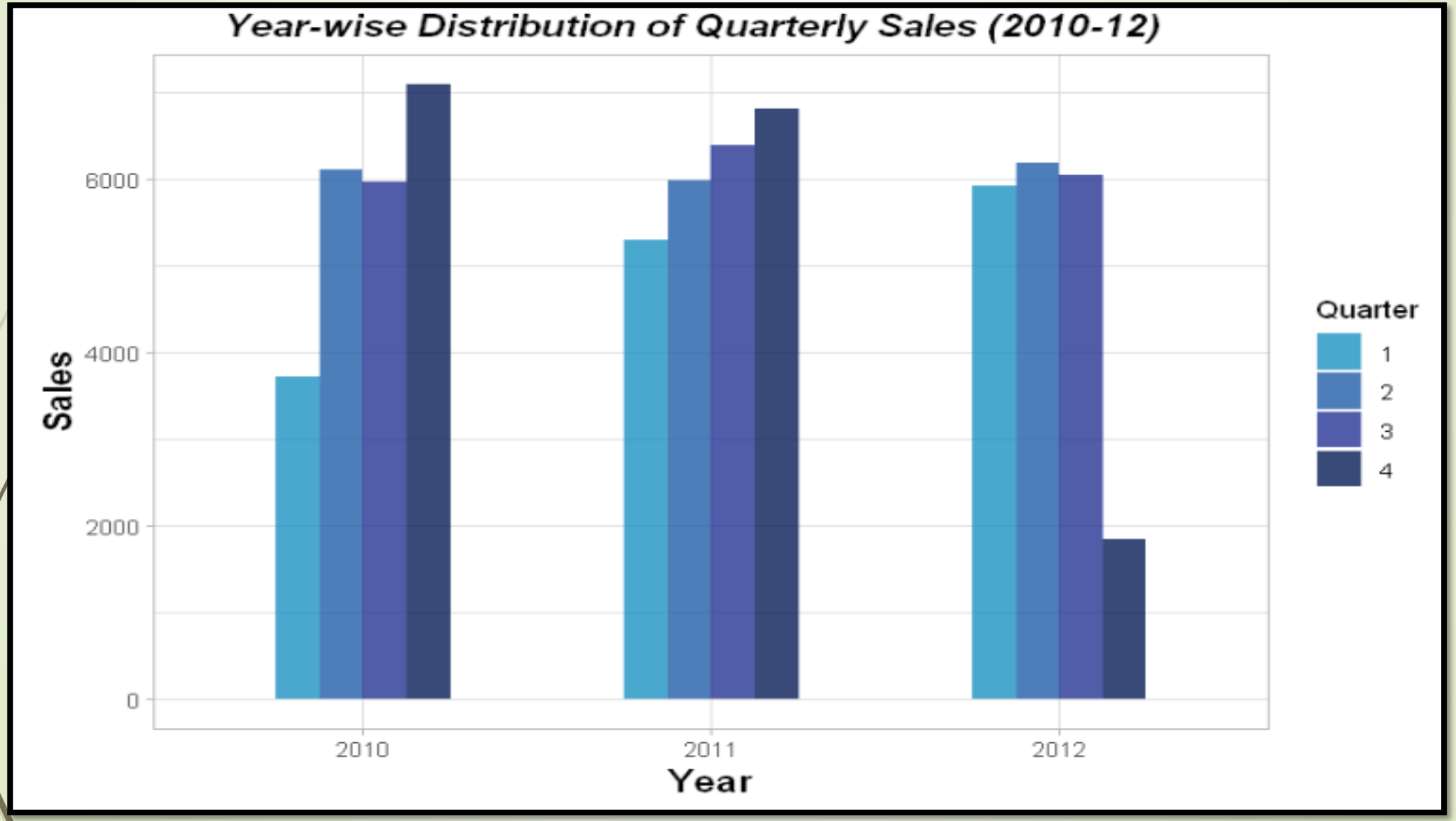
## Time Series of Sales

# Exploratory Data Analysis




- 2011, has recorded the highest sales, higher than the previous year 2010.
- There is a dip in sales in the following year 2012
- However, we are not sure if the sales were actually lower, since we dont have data of Nov & Dec, where sales might have peaked

# Exploratory Data Analysis



# Exploratory Data Analysis

## Year wise Analysis

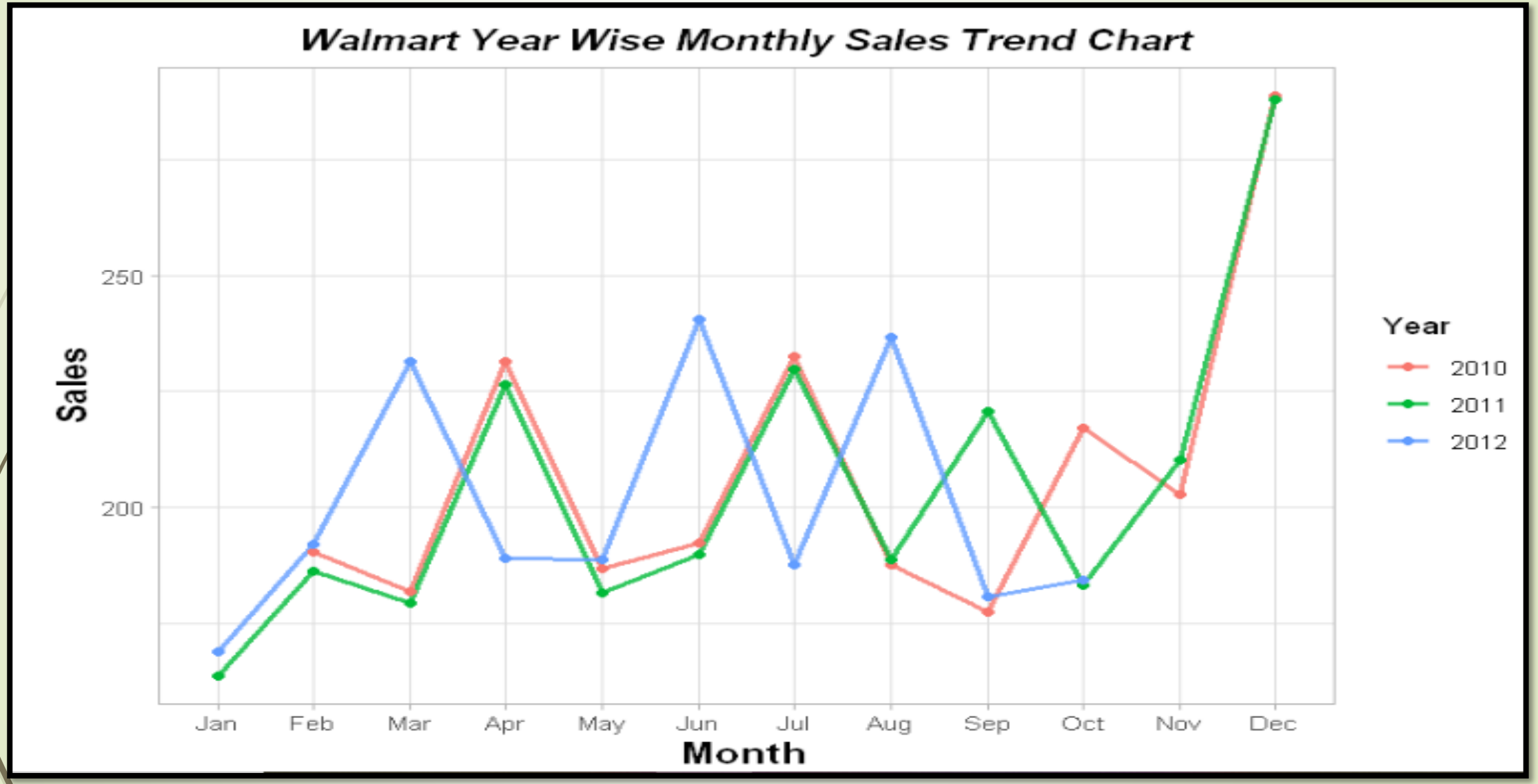
- Year 2010
  - Has growing trend of sales, quarter wise
  - Highest quarter growth was recorded in Q4, highest among all the years
  - Lowest sales was recorded in Q1
- Year 2011
  - Has growing trend of sales, quarter wise
  - Q4 recorded the highest Sales,
  - Lowest sales were recorded Q1
- Year 2012
  - Q2 recorded highest sales, Q1 and Q3 relatively lower
  - Q4 has lowest sales, still we cannot comment on Q4 sales as we don't have data for Nov & Dec month of this Year

## Quarter wise Analysis:

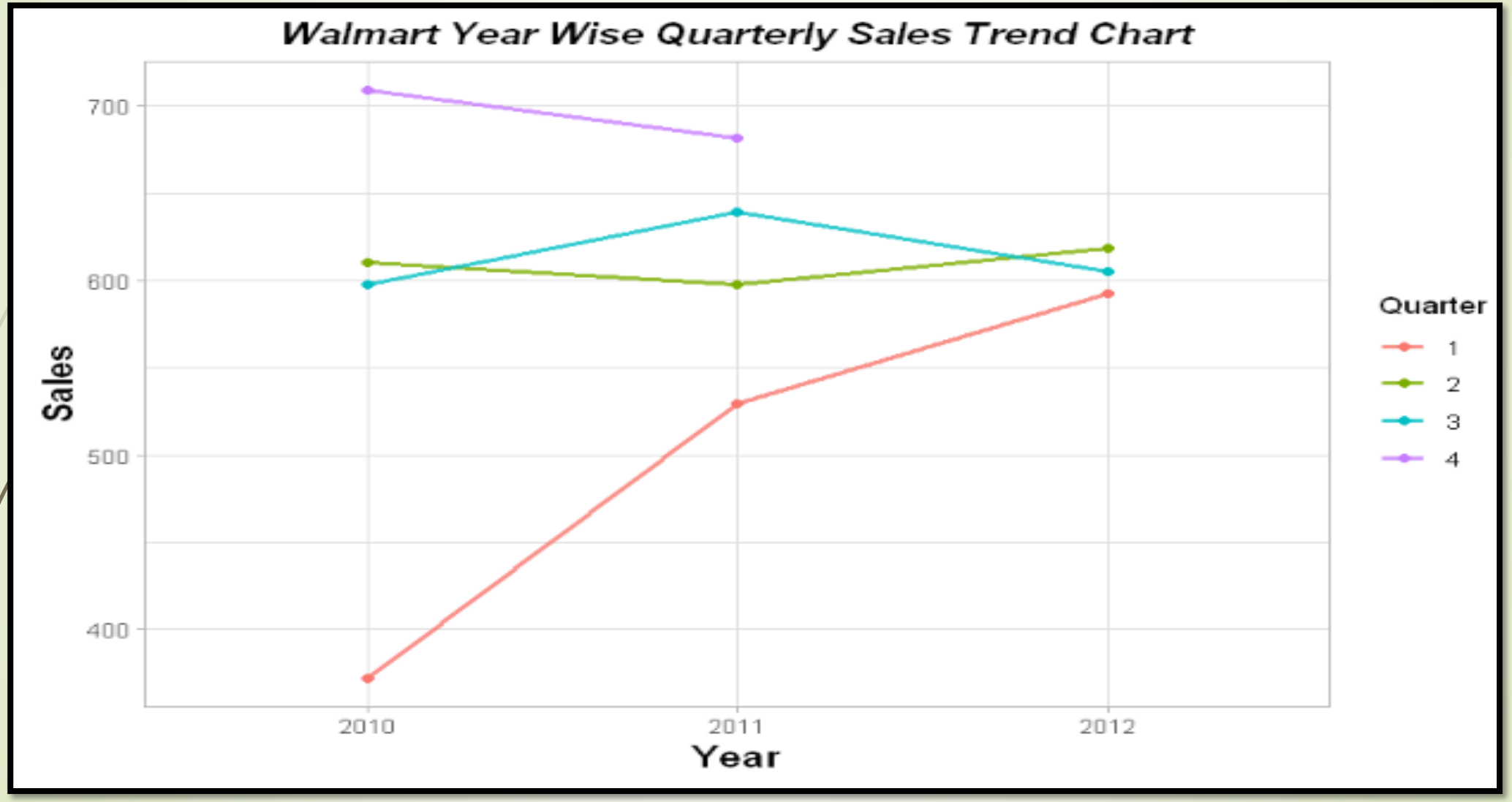
- Quarter 1
  - There has been growing trend of Q1 sales across the years
  - Each year the sales of Q1 was higher than the previous year's Q1 sales
- Quarter 2
  - Q2 sales was almost constant for all the years
  - There was a dip in sales, though not significant, observed in Q2 of year 2011
- Quarter 3
  - Q3 sales was almost constant for year 2010 and 2012
  - The sales rose in 2011 to significantly higher value
- Quarter 4
  - Highest Q4 sales was recorded in year 2010
  - Observed visible drop in sales in the next following year 2011
  - The lowest sales of Q4 was recorded in 2012, based on the available data. We do not have data for Nov and Dec month of this year

# Exploratory Data Analysis

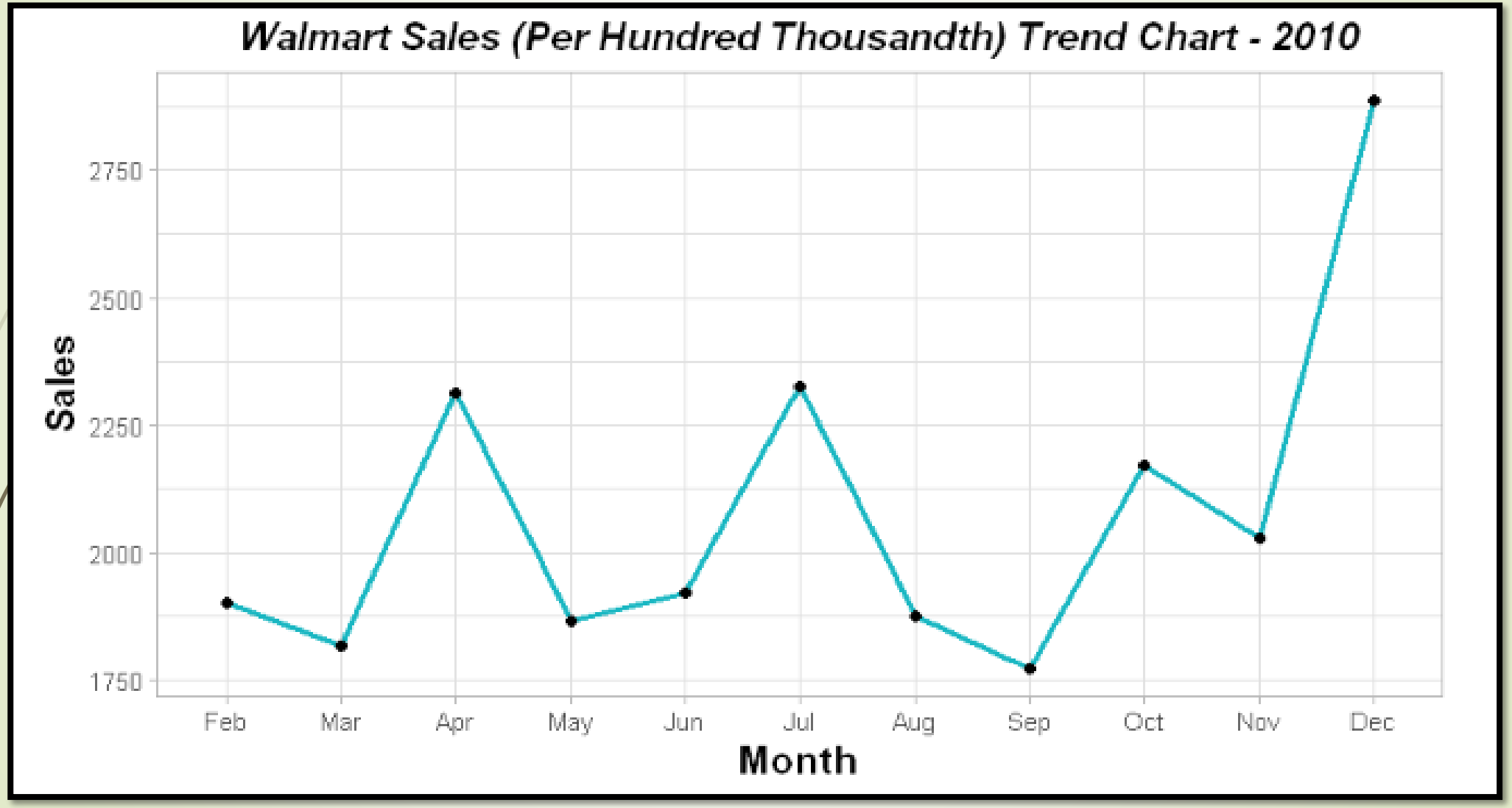
# Exploratory Data Analysis



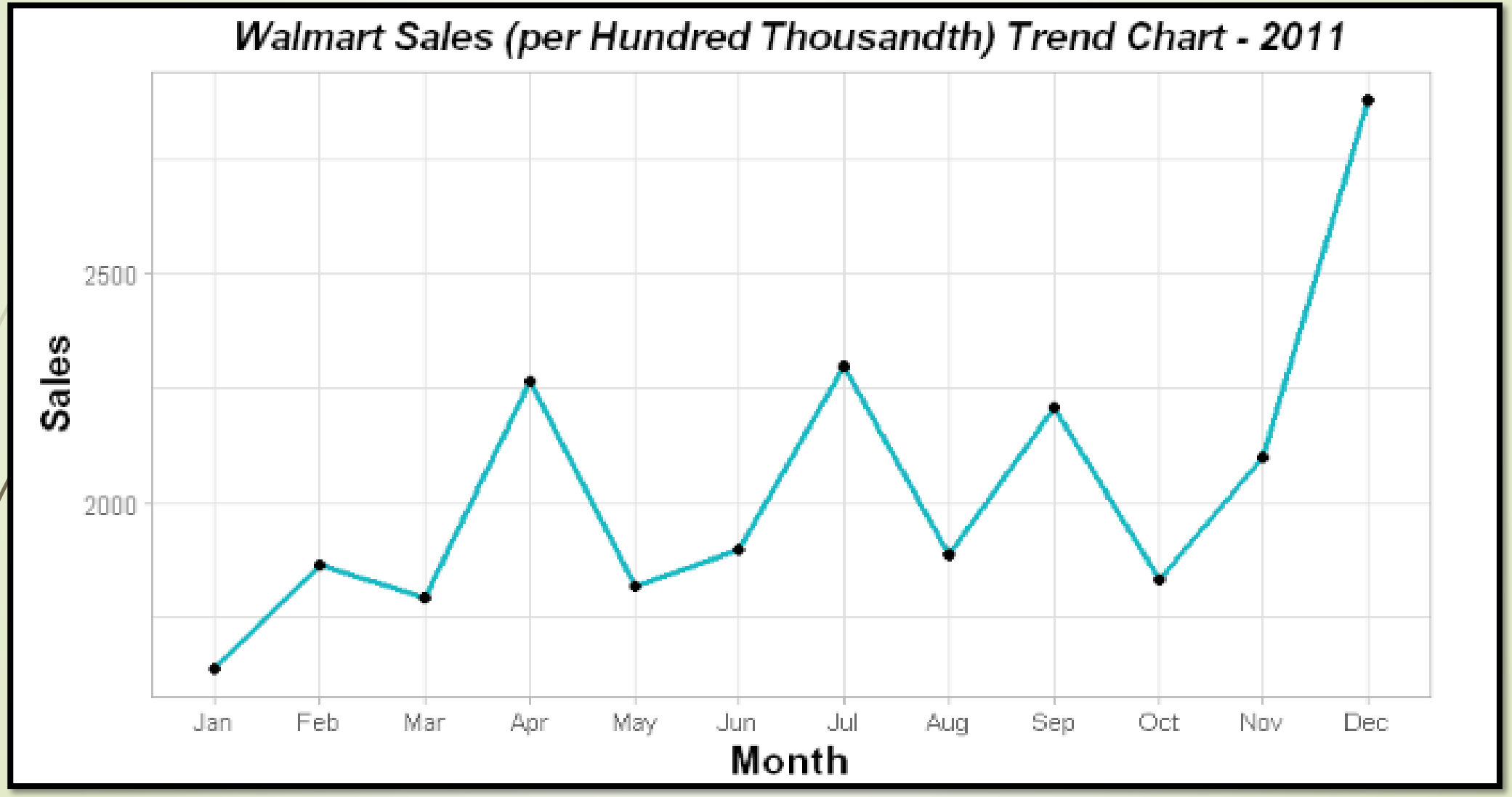
# Exploratory Data Analysis



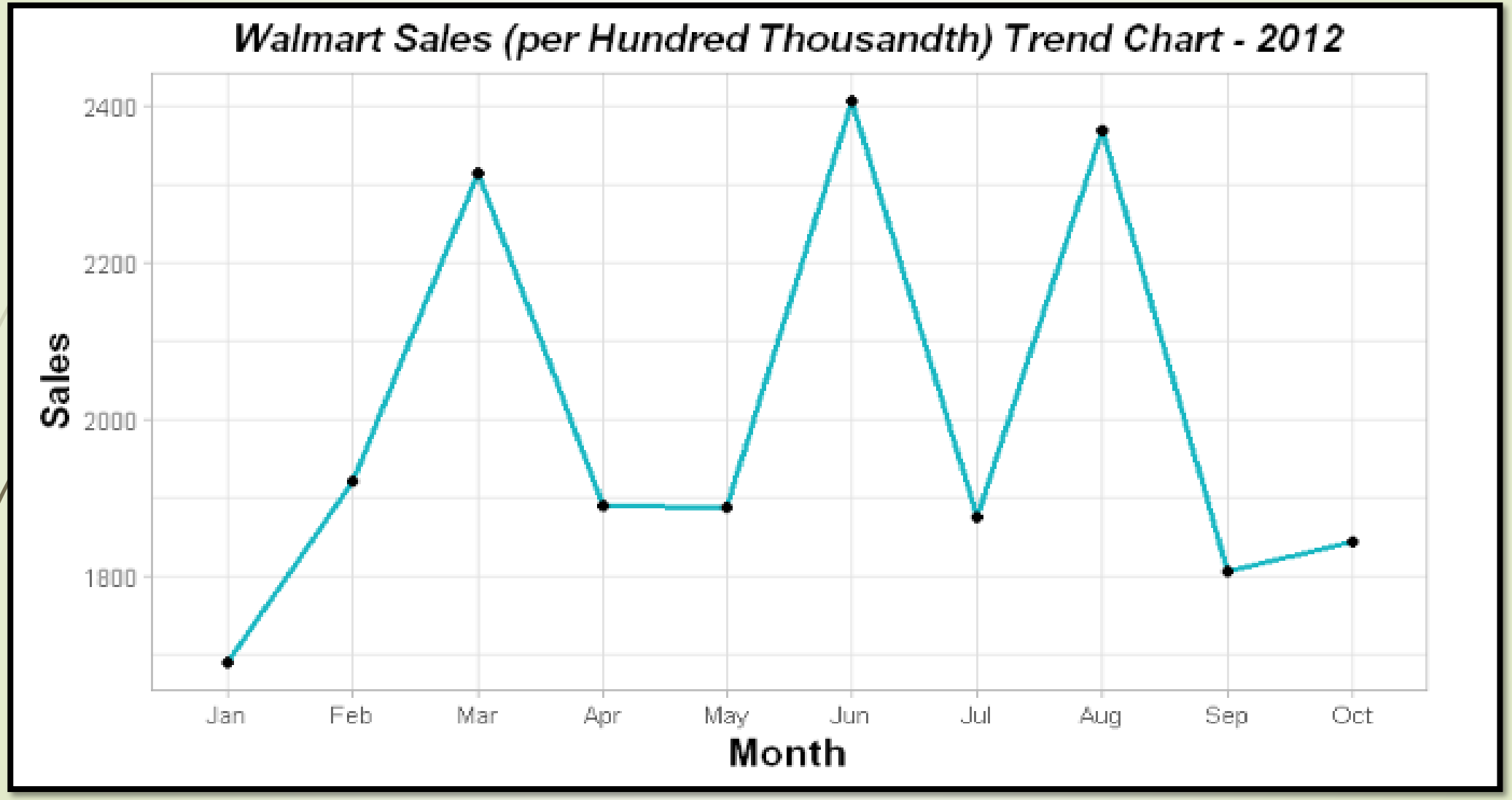
# Exploratory Data Analysis



# Exploratory Data Analysis



# Exploratory Data Analysis



# Correlation Matrix

```
               S  W  H  T  F  C  U  D  M  Y  Q
Store         ++ .  .  .     .     .  .  .  .
Weekly_Sales  .  ++    .     .  .  .     .
Holiday_Flag  .     ++ .  .  .           .
Temperature   .  .  .  ++
Fuel_Price        .     ++ .  .     .  ++ .
CPI           .  .  .     .  ++ .  .  .     .
Unemployment     .        .  .  ++ .  .  .  .
Day           .  .           .  .  ++
Month         .           .  .  .     ++ .  ++
Year          .  .  .     ++    .     .  ++ .
Quarter       .           .  .  .     ++ .  ++
attr(,"legend")
[1] -0.99 '--' -0.6 '.' 0 ' ' 0.6 '++' 1
```

# Linear Regression Model Analysis

```
Call:
lm(formula = Weekly_Sales ~ ., data = trainingSet)

Residuals:
             Min               1Q           Median               3Q
-1.8593459563999 -0.7033787213362 -0.0611230995104  0.6824981528095
             Max
 3.1496180870781

Coefficients:
                      Estimate        Std. Error   t value   Pr(>|t|)
(Intercept)  -0.00832912792241  0.01529840013251  -0.54444  0.5861688
Store        -0.35658882915177  0.01591226538877 -22.40968 < 2.22e-16 ***
Temperature  -0.03966348985775  0.01625051458981  -2.44075  0.0147037 *
CPI          -0.15570365329744  0.01644373448852  -9.46887 < 2.22e-16 ***
Unemployment -0.05080700542226  0.01648367815926  -3.08226  0.0020696 **
Thanksgiving  0.26677343030365  0.13027473797393   2.04778  0.0406530 *
Month         0.06814929556980  0.01597085952582   4.26710 2.0303e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.920626496362 on 3670 degrees of freedom
Multiple R-squared:  0.145539383844,    Adjusted R-squared:  0.144142445507
F-statistic: 104.184544146 on 6 and 3670 DF,  p-value: < 2.220446049e-16
```

- **The model created has all the variables p-value lesser than 0.05, which means that independent variables are significant to the model**
- **Multiple R squared and Adjusted are square are similar, confirming the no irrelavant variables are present in the model**
- **However, the magnitude of R squared is very poor, only 14.5% of variance is explained by the independent variable of dependent variable**
- **This model is not a good fit for this particular data.**

# Thank you!

# Appendix

- Please refer ‘R Project Report - Retail Analysis with Walmart Data’ file, submitted along with this PPT
- Because the code was developed in jupyter notebook, it has source code along with the detailed analysis and report
- All the graphs included in this presentation can also be found in that project report
- This PPT is just a glimpse of the analysis done, for non-tech audience. Detailed work is present in the project report – jupyter notebook file.