Walmart Inc.

Annual Report

Karnik Aswani

Data Science Graduate Student

Table of Contents Page

Summary……………………………………………………………...1

Introduction …………………………………………………………..1

Analysis……………………………………………………………..1-9

Conclusions…………………………………………………………..10

Recommendations……………………………………………………11

References …………………………………………………………...12

Appendices ……………………………………………………...13 - 28

Summary

Walmart is one of the most famous retailers in the world where thousands of customers buy Walmart products that generate the most revenue. More than 10,000 Walmart stores are distributed in 24 countries with different names and more than 2.2 million employees work for Walmart. With this information, Walmart relies on retail data to maximize its profits. This report will explain the different types of analysis that can be used to analyze the Walmart retail dataset.

Introduction

In this report, I will be performing several analytical techniques to answer Walmart’s business questions. The most important business question would be the performance of sales in the Walmart industry. The performance of sales is crucial because it shows if businesses are heading in the right direction or if there need to be some improvements in the business model. The more specific questions will be answered as the analytical techniques are being performed including hypothesis testing.

Analysis

Descriptive Statistical Analysis

The purpose of descriptive statistical analysis is to provide a general statistic of the dataset. Before conducting the analysis, gathering basic statistics should help our audience understand the dataset. Basic descriptive statistics comprise two parts: the measure of central tendency and the measure of dispersion. The measure of central tendency is a summary of the dataset including the mean, mode, median, and outliers. The measure of dispersion is the distribution of the dataset including the standard deviation, mean absolute deviation, variance, range, percentiles, quartiles, skewness, and kurtosis.

The measure of central tendency is the first step of descriptive statistical analysis. The mean is the average of all data values in each variable which is important to determine the distribution of the data. The median is important to determine the distribution of the data by comparing it to the mean. The mode is not too important since the dataset is composed of numeric variables instead of categorical variables. Outliers are values that are far away from normal values. In this dataset, there are no outliers in each of the variables.

The measure of dispersion is the second step of descriptive statistical analysis. The coefficient variation is the most important indicator of the data distribution. Skewness is the measurement of the symmetry of the distribution while kurtosis is the measurement of the tail of a distribution. Below is the table displaying the factors describing the dispersion.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Standard Deviation | Coefficient variation | Conclusion |
| Weekly Sales | 22504.05 | 1.417737 | Data is more spread out |
| Temperature | 18.49718 | 0.3066285 | Data is clustered |
| Fuel Price | 0.4564408 | 0.1358512 | Data is clustered |
| CPI | 39.1675 | 0.2287447 | Data is clustered |
| Unemployment | 1.862796 | 0.2341133 | Data is clustered |
| Volume | 6010904 | 0.5060976 | Data is clustered |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Skewness | Kurtosis | Outliers (Y or N) | Conclusion |
| Weekly Sales | Right side of the distribution | leptokurtic | No | Large clustering on the left side of the distribution |
| Temperature | Left side of the distribution | playkurtic | No | Large clustering on the right side of the distribution |
| Fuel Price | Left side of the distribution | playkurtic | No | Large clustering on the right side of the distribution |
| CPI | Right side of the distribution | playkurtic | No | Large clustering on the left side of the distribution |
| Unemployment | Right side of the distribution | playkurtic | No | Large clustering on the left side of the distribution |
| Volume | Right side of the distribution | leptokurtic | No | Large clustering on the left side of the distribution |

Associational Statistical Analysis

The purpose of associational statistical analysis is to find the relationships of the dataset. There are four types of relationships which are strong positive, weak positive, strong negative, or weak negative. The strongest is strong positive and the weakest is weak negative. Five regression models have been created with each variable.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | R- squared | Coefficient | Strong positive, weak positive, strong negative, or weak negative? |
| Temperature | 5.484858e-06 | 2.849298 | Weak Positive |
| Fuel Price | 2.27136e-06 | 74.3052 | Weak Positive |
| CPI | 0.0004213651 | -11.79408 | Weak Negative |
| Unemployment | 0.0007038275 | -320.5003 | Weak Negative |
| Volume | 4.979265e-08 | -0.0000008354172 | Weak Negative |

Exploratory Data Analysis

The purpose of exploratory data analysis is to find patterns and trends in the dataset. Correlation is an important part of exploratory data analysis since it tells us if there is a positive or negative relationship between the variables including the strength. Below is a chart of the correlation between the weekly sales and the other five variables.

|  |  |  |
| --- | --- | --- |
| Variable | Correlation | Rank |
| Temperature | 0.002341977 | 1 |
| Fuel Price | 0.001507103 | 2 |
| CPI | -0.02052718 | 3 |
| Unemployment | -0.02652975 | 4 |
| Volume | -0.0002231427 | 5 |

A correlation matrix has been conducted in this analysis; however, it is not very useful for determining the correlation among the variables since it is not very precise compared to the correlation tests.

Statistical tests are the second part of the exploratory data analysis because these tests can make a business-informed decision by accepting or rejecting the null hypothesis related to the variables. Below is a table of the different types of tests that were conducted in this analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Statistical Tests | Description | Results | Conclusion |
| Kolmogorov- Smirnov | Determine if two distributions differ | All variables of this test are less than 0.05. | All the variables have different distributions from each other. |
| Cramer-Von Mises Test | Determine if a sample is from a normal distribution or not | All variables of this test are less than 0.05. | All the variables are non-normally distributed. |
| Kruskal-Wallis Test | Determine if there are statistically significant differences between the medians of the groups | All variables of this test are less than 0.05. | All the variables are significantly different across different groups in the data. |
| One Sample T-test | Determine if the means between several groups are equal | All variables of this test are less than 0.05. | All the variables have different means |
| Mann-Whitney U tests | Compare two independent groups of data | All variables of this test are less than 0.05. | All the variables have different distributions among the weekly sales variable. |
| KPSS Test | Determine if the time series in our predictive analysis will be stationary or not | All variables of this test are less than 0.05. | Assume all the variables are not time series stationary |
| Granger- Causality Test | Determine if a time series in each variable with the weekly sales is helpful or not | Temperature and volume variables are greater than 0.05. The rest of the variables are less than 0.05. | Temperature and volume are not useful predictors. The rest of the variables are helpful. |
| Breusch-Godfrey Test | Determine if there is an autocorrelation in each variable | All variables of this test are less than 0.05. | Assume all the variable autocorrelations exist among the residuals. |
| Breusch – Pagan Test | Determine if there is an unequal scatter of residuals. | Temperature, unemployment, and volume are less than 0.05. Fuel price and CPI are higher than 0.05. | Heteroscedasticity is present in temperature, unemployment, and volume. Heteroscedasticity is not present in fuel prices and CPI. |
| Anova Test | To compare the means of three or more groups | IsHoliday, CPI, and Unemployment are less than 0.05. Temperature, Fuel Price, and Volume are greater than 0.05. | < 0.05 = significant differences in the means of these variables.  > 0.05 = no significant differences in the means of these variables. |

Multicollinearity is a concept where many independent variables are correlated in the regression model. The variance inflation factor is a measurement of the multicollinearity in the regression model. AVIF of 1 means there is no correlation among the variables. A VIF between 1 and 5 means there is a correlation among the variables depending on the measurement. A VIF greater than 5 means the variables are highly correlated among the variables. Based on the results of the analysis, all the variables are slightly correlated with each other since all the VIFs are less than 1.25.

Confidence intervals are a range of values that contain the true values with a level of confidence.

The confidence intervals are based on a sample of data, and the confidence level determines if the true values lie within that range. The default value for confidence intervals is 0.95. Based on the plots, temperature and fuel prices have a positive relationship with weekly sales. Unemployment and CPI have a negative relationship with weekly sales. Volume has a neutral relationship with weekly sales.

Predictive Analysis

The purpose of predictive analysis is to predict future events based on historical data. In this section, 7 machine-learning techniques will be conducted to predict the outcome of our data. All six variables conducted earlier will be used in most of the machine learning techniques to show which variable is the best predictor of the performance of sales in Walmart.

The first part of predictive analysis is time series analysis including the Arima model, and Holtz- Winters models. The ARIMA model is a model that measures events that happen over time while Holt-winters is a model that predicts the behavior of a time series. These models are the most important since these models will predict the sales of Walmart. However, the forecasting part of these models will not be accurate, but it will give the audience an idea of which direction sales are headed. Below is a summary table of the time series analysis methods.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Store | ARIMA Model  (Inc, Dec, Constant) | Holt – Winter  (Inc, Dec, Constant) | Box – Test (ARIMA, Holtz - Winter | Conclusion |
| Walmart 1 | Decrease | Increase | (0.6349, 0.2351) | Both models are not a great fit. |
| Walmart 2 | Constant | Decrease | (0.2351, 0.8235) | Both models are not a great fit. |
| Walmart 3 | Constant | Decrease | (0.8019, 0.004019) | Holtz-winter is a great fit |
| Walmart 4 | Constant | Decrease | (0.657, 0.0003157) | Holtz-winter is a great fit |
| Walmart 5 | Decrease | Increase | (0.8193, 0.7648) | Both models are not a great fit. |

The second part of predictive analysis is regression models including linear, multilinear, logistic, XGboost regressor, random forest regressor, and decision tree regressor. These models will determine the accuracy score of the models including R2 , MSE, AIC, and BIC. The goal of this part is to determine which models have the highest accuracy score with fewer errors. Below is a summary table of the regression models.

Linear regression is a regression technique that determines the relationship of a correlation between two variables. In this regression model, one predictor and one response variable are conducted. Below is a summary table of the linear regression model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | R2 | MSE | AIC | BIC | Conclusion |
| Temperature | 0.1417347 | 12625282 | 793.319 | 798.530 | Weak fit |
| Fuel Price | 0.01924174 | 15594188 | 799.207 | 804.420 | Weak fit |
| CPI | 0.03035825 | 700079898 | 798.728 | 803.940 | Weak fit |
| Unemployment | 0.05839966 | 15313188 | 797.418 | 802.631 | Weak fit |
| Volume | 0.0411412 | 15112195 | 798.253 | 803.466 | Weak fit |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Walmart Variables | R2 | MSE | AIC | BIC | Conclusion |
| Temperature | 5.484858e-06 | 506428138 | 9374875 | 9374908 | Weak fit |
| Fuel Price | 2.27136e-06 | 506429765 | 9374877 | 9374909 | Weak fit |
| CPI | 0.0004213651 | 506217523 | 9374705 | 9374738 | Weak fit |
| Unemployment | 0.0007038275 | 506074475 | 9374589 | 9374622 | Moderate fit |
| Volume | 4.979265e-08 | 506430890 | 9374878 | 9374910 | Weak fit |

Multilinear regression is a regression technique that determines the response variable with multiple predictor variables. In this regression model, all the predictor variables are conducted against the response variable. Below is a summary table of the multilinear regression model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Store | R2 | MSE | AIC | BIC | Conclusion |
| Walmart 1 | 0.3598916 | 31665494 | 858.562 | 870.726 | Weak fit |
| Walmart 2 | 0.1432943 | 9614810 | 808.501 | 820.665 | Weak fit |
| Walmart 3 | 0.5942047 | 4538748 | 776.974 | 789.137 | Strong fit |
| Walmart 4 | 0.1490355 | 5186814 | 782.579 | 794.743 | Weak fit |
| Walmart 5 | 0.3100741 | 1415824 | 728.046 | 740.210 | Moderate fit |
| Walmart | 0.001790294 | 505524255 | 9374151 | 9374228 | Poor fit |

Logistic regression is a regression technique that determines the relationship of a correlation if the response variable is binary. In this regression model, the IsHoliday variable is being conducted against all the variables mentioned earlier. Below is a summary table of the logistic regression model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Store | R2 | MSE | AIC | BIC | Conclusion |
| Walmart 1 | 0.4126526 | 0.03333333 | 22.631 | 32.440 | Weak fit |
| Walmart 2 | 0.4068886 | 0.06666667 | 22.716 | 32.525 | Weak fit |
| Walmart 3 | 0.2634638 | 0.07070707 | 54.941 | 73.107 | Weak fit |
| Walmart 4 | 0.2579634 | 0.08080808 | 55.247 | 73.413 | Weak fit |
| Walmart 5 | 0.2877254 | 0.05494505 | 49.155 | 66.731 | Moderate fit |
| Walmart | 0.2063897 | 0.07005684 | 118570 | 118644 | Weak fit |

XGBoost regressor is a machine-learning model that can improve the accuracy of a prediction. This regression model creates training and test datasets to fit into the XGboost model to determine if our data is underfitted or overfitted. Below is a summary of the XGboost regressor model.

|  |  |  |  |
| --- | --- | --- | --- |
| Store | R2 | MSE | Conclusion |
| Walmart 1 | NA | 3.552714e-15 | Failed to fit |
| Walmart 2 | NA | 1.421085e-14 | Failed to fit |
| Walmart 3 | NA | 2505.689 | Failed to fit |
| Walmart 4 | NA | 5.684342e-14 | Failed to fit |
| Walmart 5 | NA | 2.273737e-13 | Failed to fit |
| Walmart | 0.1783555 | 764.762 | Weak fit |

Random forest regressor is a machine-learning model to predict future costs by reducing the variance in the datasets. In this regression model, a randomforest model is adjusted by the tuneRF function to predict accurate sales. Below is a summary table of the random forest regressor model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Store | R2 | MSE | Prediction | Conclusion |
| Walmart 1 | -1.042984 | 3 | 24414.890 | Weak fit |
| Walmart 2 | -0.1009608 | 375 | 66776.480 | Weak fit |
| Walmart 3 | 0.3668836 | 1 | 6070.819 | Strong fit |
| Walmart 4 | -0.4007784 | 21 | 59055.350 | Moderate fit |
| Walmart 5 | -0.3309827 | 178 | 6600.544 | Weak fit |

Decision tree regressor is a machine-learning model that builds decision trees to predict the response variable. There are two types of decision trees which are regression and classification trees. In this regression model, a tree model is pruned to find the lowest test error in the output. Below is a summary table of the decision tree regressor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Store | R2 | MSE | Prediction | Conclusion |
| Walmart 1 | 0.2891249 | 54547183 | 18315.88 | Weak fit |
| Walmart 2 | 0.210525 | 11223002 | 65987.980 | Weak fit |
| Walmart 3 | 0.4176869 | 13990765 | 4920.558 | Strong fit |
| Walmart 4 | 0.2877782 | 6095219 | 58934.810 | Moderate fit |
| Walmart 5 | 0.286847 | 2052139 | 5765.849 | Moderate fit |

Conclusion

Summary of the analyses that were conducted in this report and what were the results. Descriptive analysis was used to understand the overall Walmart dataset by using two techniques. The statistical methods are the measure of central tendency and the measure of dispersion to describe the basic statistics of the dataset that were mentioned in the analysis. It was found that all of the variables had non-normal distributions and no outliers were detected. Also, all of the variables are statistically different across the different groups in the data.

Associational statistical analysis was used to determine the relationships between the variables in the dataset. Basically, a linear regression model was conducted to find the r-squared values, coefficients, and the type of relationships. Furthermore, it was concluded that temperature and fuel price have positive relationships while CPI, unemployment, and volume have weak negative relationships. Therefore, temperature and fuel prices are great indicators of weekly sales whereas CPI, unemployment, and volume are poor indicators of weekly sales.

Exploratory data analysis was used to find any patterns and trends in the dataset. A correlation matrix was conducted in this analysis; however, it did not provide any insights into the datasets. All the statistical tests proved that all the variables are non-normally distributed and significantly different across different groups in the data except for the ANOVA test. The variation inflation factor concluded that all the variables are slightly correlated with each other. Confidence intervals concluded that temperature and fuel price have positive relationships that affect the weekly sales while unemployment and CPI have a negative relationship that affects the weekly sales.

Predictive analysis is the most important analysis for predicting future sales by creating six regression models and two time series models. Multilinear regression and decision tree regressor are the best models while linear regression and random forest regressor are the worst models. Even with parameter tuning, the XGBoost failed to fit the model since all the R2 are NaNs. Although the exact reason is not confirmed, there is a possibility that the data might lack important features that would prevent the failure of the model. There are other reasons why the XGBoost model failed, however, those reasons were concluded in earlier analyses.

Recommendations

There are several recommendations that can improve the sales of Walmart. The business strategies would include improving the in-store experience, increasing product diversity, implementing price optimization, enhancing online presence, optimizing supply chain management, personalized marketing, and enhancing customer service. All of these business strategies are great and each one will make more profits for the retail company. Price optimization will attract more customers since customers love discounts and promotions to save money. Also, customer service would attract more customers by implementing self-service and responding quickly to the customers. Improving the in-store experience can help customers find their products easily, which can reduce frustration among customers. Increasing product diversity will attract customers, however, it would depend on the product and the price. Enhancing our online presence will make the shopping experience with the customers great. Optimizing supply chain management can improve sales by implementing better inventory practices, optimizing transportation and logistics, and improving relationships with suppliers to reduce costs and improve efficiency. Finally, personalized marketing will increase sales as well since Walmart will target and advertise its products to specific groups.

References

“10 Best Retail Merchandising Strategies for Retail Success - Chaindrive.” *ChainDrive.com*, 22 Feb. 2023, https://chaindrive.com/retail-merchandising-strategies/.

robk@statmethods.net, Robert Kabacoff -. “Correlations.” Quick-R: Correlations, https://www.statmethods.net/stats/correlations.html.

robk@statmethods.net, Robert Kabacoff -. “Histograms and Density Plots.” Quick-R: Density Plots, <https://www.statmethods.net/graphs/density.html>.

Auguie, Baptiste. *Arranging Multiple Grobs on a Page*, 9 Sept. 2017, https://cran.r-project.org/web/packages/gridExtra/vignettes/arrangeGrob.html.

“Background Color in GGPLOT2.” R CHARTS | A Collection of Charts and Graphs Made with the R Programming Language, 1 Nov. 2021, https://r-charts.com/ggplot2/background-color/.

Prabhakaran , Selva. Statistical Tests, http://r-statistics.co/Statistical-Tests-in-R.html.

Zach. “How to Perform a Two Sample t-Test in R.” Statology, 3 Aug. 2022, https://www.statology.org/two-sample-t-test-in-r/.

Zach. “How to Perform a Breusch-Pagan Test in R.” Statology, 31 Dec. 2020, https://www.statology.org/breusch-pagan-test-r/.

Zach. “Kolmogorov-Smirnov Test in R (with Examples).” Statology, 10 Dec. 2021, https://www.statology.org/kolmogorov-smirnov-test-r/.

Zach. “How to Conduct an Anderson-Darling Test in R.” Statology, 21 July 2020, https://www.statology.org/anderson-darling-test-r/.

Zach. “How to Conduct a Jarque-Bera Test in R.” Statology, 14 Apr. 2020, <https://www.statology.org/how-to-conduct-a-jarque-bera-test-in-r/>.

Zach. “How to Perform a Cramer-Von Mises Test in R (with Examples).” Statology, 21 Nov. 2022, https://www.statology.org/cramer-von-mises-test-r/.

Zach. “How to Perform a Breusch-Godfrey Test in R.” Statology, 16 Apr. 2021, https://www.statology.org/breusch-godfrey-test-in-r/.

Zach. “How to Perform a KPSS Test in R (Including Example).” Statology, 19 Jan. 2022, <https://www.statology.org/kpss-test-in-r/>.

DataTechNotes. “Regression Example with XGBoost in R.” Regression Example with XGBoost in R, 4 Aug. 2020, https://www.datatechnotes.com/2020/08/regression-example-with-xgboost-in-r.html.

Zach. “How to Perform Logistic Regression in R (Step-by-Step).” Statology, 29 Sept. 2021, https://www.statology.org/logistic-regression-in-r/.

Zach. “How to Calculate AUC (Area under Curve) in R.” Statology, 9 Sept. 2021, <https://www.statology.org/auc-in-r/>.

“Sign In.” RPubs, https://rpubs.com/johansetiawan38/time\_series\_1.

Appendices

Source Code

# Load libraries

library(readr) # reading csv files

Load the datasets

walmart <- read\_csv("walmart\_cleaned.csv")

wmt <- read\_csv("WMT.csv")

# Drop unnecessary columns

walmart <- walmart[c(-1,-9,-10,-11,-12,-13,-16,-17)]

wmt <- wmt[c(-2,-3,-4,-5,-6)]

# Join the two datasets together

walmart <- merge(walmart,wmt, by="Date")

# View the dataset again

View(walmart)

# Sort the date

walmart <- walmart %>%

arrange(mdy(walmart$Date))

View(walmart)

# Check for null values

walmart %>%

is.na()%>%

colSums()

# Descriptive statistics of the dataset

describe(walmart$Weekly\_Sales)

# Calculate distribution of each variable in the dataset

quantile(walmart$Weekly\_Sales)

# Calculate variance of each variable in the dataset

var(walmart$Weekly\_Sales)

# Calculate standard deviation of each variable in the dataset

sd(walmart$Weekly\_Sales)

# Coefficient of Variation

sd(walmart$Weekly\_Sales)/mean(walmart$Weekly\_Sales)

# Calculate the measure of symmetry in each variable in the dataset

skewness(walmart$Weekly\_Sales)

# Calculate the measure of the tails in each variable in the dataset

kurtosis(walmart$Weekly\_Sales)

# Create regression models

regmodel <- lm(Weekly\_Sales ~ Temperature, data = walmart)

summary(regmodel)$r.squared

summary(regmodel)$coefficients

# Plot the Weekly\_Sales variable against the other independent variable

wg <- ggplot(walmart) + aes(x = as.numeric(Weekly\_Sales)) +

geom\_histogram(aes(y = stat(density)), fill = "coral") +

geom\_density(col = "brown4") +

theme(panel.background = element\_rect(fill = "darkslategray1"))

# Determine any outliers in each variable column

out <- boxplot.stats(walmart$Weekly\_Sales)

out\_ind <- which(walmart$Weekly\_Sales %in% c(out))

out\_ind

# Determine outliers by creating boxplots

par(mfrow = c(2, 3))

b <- boxplot(walmart$Weekly\_Sales, ylab = "Weekly Sales", col = "green", main = "Boxplot of Weekly Sales")

# correlation Matrix

walmartc <- walmart[, c(5,6,7,8,9,10,11,12,13,14,15)]

ggcorr(walmartc, palette = "Blues", label = TRUE)

# Correlation Test

cor.test(walmart$Weekly\_Sales, walmart$Temperature)

# Two Sample T Test

t.test(walmart$Weekly\_Sales, walmart$Temperature, var.equal = TRUE)

# Welch Sample T-test

t.test(walmart$Weekly\_Sales ~ walmart$IsHoliday)

# One- Sample Test

t.test(walmart$Weekly\_Sales, mu = 50000)

# Kolmogorov-Smirnov Test

ks.test(walmart$Weekly\_Sales, "pnorm")

# Mann Whitney U Test

wilcox.test(walmart$Weekly\_Sales ~ walmart$IsHoliday, distribution = "exact")

# Cramer-Von Mises Test

cvm.test(walmart$Weekly\_Sales)

# Kruskal-Wallis Test

kruskal.test(walmart$Weekly\_Sales ~ walmart$Temperature)

# KPSS Test

kpss.test(walmart$Weekly\_Sales, null = 'Trend')

# Granger-Causality Test

grangertest(Weekly\_Sales~ Temperature, order = 1, data = walmart)

# Breusch-Godfrey Test

bgtest(Weekly\_Sales ~ Temperature,data = walmart)

# Breusch-Pagan Test

bptest(regmodel)

# Anova test

wfit1 <- aov(walmart$Weekly\_Sales ~ walmart$Temperature + walmart$Fuel\_Price + walmart$CPI

+ walmart$Unemployment + walmart$Volume, data = walmart)

summary(wfit1)

# Anova individual test

waov1 <- aov(walmart$Weekly\_Sales ~ walmart$IsHoliday, data = walmart)

# Summary of the individual anova tests

summary(waov1)

# Variation inflation factor

vif(wfit1)

# Confidence intervals

confint(wfit1, level = 0.95)

# Plot the confidence Intervals

wc <- ggplot(walmart, aes(x = Temperature, y = Weekly\_Sales)) + ggtitle("Weekly Sales vs Temperature")

wc + geom\_smooth(method = "lm", se = TRUE, level = 0.95) +

theme(plot.title = element\_text(hjust = 0.5))

# Filter out store 1- 5 with the number of the Department

walmart1 <- walmart%>%

filter(Store == 1 & Dept == 1) %>%

mutate(Date = as.Date(Date, format = "%m/%d/%Y")) %>%

filter(Date >= "2012-01-01", Date <= "2012-12-31")

# Graph all 5 stores with each variable

# Weekly Sales

g <-ggplot(data = walmart1, mapping = aes(x = Date, y = Weekly\_Sales)) + ggtitle("2012 of Walmart 1 Weekly Sales") +

theme(plot.title = element\_text(hjust = 0.5)) +

geom\_line() +

geom\_point() +

theme(plot.background = element\_rect(fill = "lightpink"))

################## ARIMA Model

## Walmart 1

# TS plot

wts <- ts(walmart1$Weekly\_Sales, start = c(2010,2,5), frequency = 12)

autoplot(wts)

# Decompose the data

wts\_dec <- decompose(wts, type = "additive")

autoplot(wts\_dec)

# Split the dataset into a training and test datasets

w\_tesales <- tail(wts, 15)

w\_trsales <- head(wts, 139)

# Determine if the data is stationary or not

w\_trsales %>%

adf.test()

# Reduce the p-value less than 0.05

w\_trsales %>%

diff()%>%

adf.test()

# Plot the data

w\_trsales %>%

diff()%>%

tsdisplay()

# Plot the Arima model

w\_arima <- auto.arima(y = w\_trsales, seasonal = T)

autoplot(w\_arima$fitted, series = "Sarima") + autolayer(w\_trsales, series = 'Actual')

w\_arima$aic

w\_arima$bic

# plot the Holtz-Winters model

w\_holtz <- HoltWinters(x = w\_trsales, alpha = 0.5, beta = 0.1, gamma = F)

autoplot(w\_holtz$fitted[,1], series = "hw") +

autolayer(w\_trsales, series = "Actual")

# Forecast both models with a 24 month period

w\_forecast <- forecast(w\_arima, h = 24)

w\_forecast1 <- forecast(w\_holtz, h = 24)

# Plot the forecast models

w\_trsales%>%

autoplot() + autolayer(w\_forecast, series = "ARIMA Forecast") +

autolayer(w\_tesales, series = "test data")

w\_trsales%>%

autoplot() + autolayer(w\_forecast1, series = "Holtz Forecast") +

autolayer(w\_tesales, series = "test data")

# Ljung-Box test

Box.test(w\_forecast$residuals, type = "Ljung-Box")

Box.test(w\_forecast1$residuals, type = "Ljung-Box")

plot(w\_forecast$residuals, col = "red")

plot(w\_forecast1$residuals, col = "blue")

################## Linear Regression

# linear regression model of each variable

# Walmart Temperature

regmodel <- lm(Weekly\_Sales ~ Temperature, data = walmart)

# Calculate AIC, BIC, and R2

AIC(regmodel)

BIC(regmodel)

summary(regmodel)$r.squared

# Calculate MSE

walmart$predicted\_sales <- predict(regmodel, newdata = walmart)

wlp <- mean((walmart$Weekly\_Sales - walmart$predicted\_sales)^2)

wlp

# Calculate the average of AIC, BIC, R2, and MSE

sum(summary(regmodelt)$r.squared,

summary(regmodelt2)$r.squared,

summary(regmodelt3)$r.squared,

summary(regmodelt4)$r.squared,

summary(regmodelt5)$r.squared)/5

sum(AIC(regmodelt),

AIC(regmodelt2),

AIC(regmodelt3),

AIC(regmodelt4),

AIC(regmodelt5))/5

sum(BIC(regmodelt),

BIC(regmodelt2),

BIC(regmodelt3),

BIC(regmodelt4),

BIC(regmodelt5))/5

################## Multilinear Regression

wfit <- lm(walmart$Weekly\_Sales ~ walmart$Temperature + walmart$Fuel\_Price + walmart$CPI

+ walmart$Unemployment + walmart$Volume, data = walmart)

# Calculate R2

summary(wfit)$r.squared

# Calculate MSE

mean(wfit$residuals^2)

# Calculate AIC

AIC(wfit)

# Calculate BIC

BIC(wfit)

################## Logistic Regression

# reproduce the dataset

set.seed(100)

# split the dataset into a training and test dataset

wsit <- sample(c(TRUE,FALSE), size = nrow(walmart), replace = TRUE, prob = c(0.7,0.3))

wtrains <- walmart[wsit,]

wtests <- walmart[!wsit,]

# create the logistic model

wlmodel <- glm(IsHoliday ~ Weekly\_Sales + Temperature + Fuel\_Price + CPI + Unemployment

+ Volume, family = "binomial", data = wtrains)

# reduction of scientifc notations

options(scipen = 999)

# summary of the model

summary(wlmodel)

# Calculate R2

pscl::pR2(wlmodel)["McFadden"]

# convert response to 0 and 1

threshold = 0.6

wpredict <- ifelse(predict(wlmodel, type = "response") > threshold,1,0)

wactual <- wlmodel$y

# Create confusion matrix

wconf <- table(wpredict, wactual)

wconf

# Create predict values

wpredicted <- predict(wlmodel, wtests, type = "response")

# Calculate MSE

mse(wactual,wpredict)

# Calculate AIC

AIC(wlmodel)

# Calculate BIC

BIC(wlmodel)

# Calculate area under curve

auc(wtests$IsHoliday,wpredicted)

# plot ROC curve

plot.roc(wpredict,wactual)

################## XGBoost

## Walmart

# reproduce the dataset

set.seed(100)

# split the dataset into a training and test dataset

wd <- createDataPartition(walmart$Weekly\_Sales, p= .75, list = F)

wtrain <- walmart[wd,]

wtest <- walmart[-wd,]

# predictor and response variables in training set

w\_xtrain = data.matrix(wtrain[,-4])

w\_ytrain = (wtrain[,4])

# predictor and response variables in test set

w\_xtest = data.matrix(wtest[,-4])

w\_ytest = (wtest[,4])

# final training and testing sets

w\_xgb = xgb.DMatrix(data = w\_xtrain, label = w\_ytrain)

w\_xgb1 = xgb.DMatrix(data = w\_xtest, label = w\_ytest)

# create the watchlist

wlist = list(wtrain = w\_xgb, wtest = w\_xgb1)

# Fit the XGBoost model and show the training and test datasets of each round

wmodel = xgb.train(data = w\_xgb, max.depth = 7, watchlist = wlist, nrounds = 300)

# create the final model

wfinal = xgboost(data = w\_xgb, max.depth = 7, nrounds = 100, verbose = 0)

# predict the final model from the XGBoost model

w\_ypred <- predict(wfinal, w\_xgb1)

# MSE and R2

mean((w\_ytest - w\_ypred)^2)

caret::R2(w\_ytest,w\_ypred)

# display the model

x = 1:length(w\_ytest)

plot(x, w\_ytest, col = "blue", type = "l")

lines(x, w\_ypred, col = "orange", type = "l")

################ Random Forest

## Walmart

# reproduce the dataset

set.seed(100)

# create the random forest model

wff <- randomForest(formula = Weekly\_Sales ~ Temperature + Fuel\_Price + CPI

+ Unemployment + Volume, data = walmart)

# summary of the model

print(wff)

# trees that have the lowest test MSE

which.min(wff$mse)

# RMSE of the best model

wff$rsq[which.min(wff$rsq)^2]

# plot the model

plot(wff)

# plot the variable important plot

varImpPlot(wff)

# adjust the parameters

wtunef <- tuneRF(x = walmart[,5:15], y = walmart$Weekly\_Sales, ntreeTry = 550,

mtryStart = 5, stepFactor = 0.5, improve = 0.05, trace = FALSE)

# create new data frame

tunef <- data.frame(Temperature = 100, CPI = 220, Fuel\_Price = 5.00,

Unemployment = 11.00, Volume = 160000000)

# predict the sales price

################## Regression Tree

## Walmart 1

# create decision tree model

wrtreec <- rpart(Weekly\_Sales ~ Temperature + CPI + Unemployment + Volume,

data = walmart1, control = rpart.control(cp = 0.00001))

# summary of the complex parameters

printcp(wrtreec)

# find the best cp value

wrbest <- wrtreec$cptable[which.min(wrtreec$cptable[,"xerror"]),"CP"]

# create a prune tree

wptreec1 <- prune(wrtreec , cp = wrbest)

# plot the prune tree

prp(wptreec1, faclen = 1, extra = 1, roundint = F, digits = 3)

# create new dataframe

wframe <- data.frame(Temperature = 74, CPI = 250, Unemployment = 11.50, Volume = 16000000)

# predict the sale price

predict(wptreec1, newdata = wframe)

# calculate the mse

wlp <- predict(wptreec1, newdata = wframe)

wo <- walmart1$Weekly\_Sales

wlpm <- mean((wo - wlp)^2)

wlpm

# calculate the rsquare

wo1 <- walmart1$Weekly\_Sales

wtrpredict <- predict(wptreec1, newdata = walmart1)

wrsse <- sum((wo1 - wtrpredict)^2)

wrsst <- sum((wo1 - mean(wo1))^2)

wrr <- 1 - wrsse/wrsst

wrr

predict(wff, newdata = tunef)