FORECASTING WEEKLY SALES AT WALMART USING REGRESSION ALGORITHM

Neha Cemerla
Sai Durgesh Mukkamala
Sujith Kondreddy
Renuka Meda
Varsha Komalla

INTRODUCTION

Overview of Walmart:

- Prominent U.S. retail chain with 45 stores.
- Faces challenges in sales forecasting due to unexpected demand and stockouts.

Need for Accurate Predictions:

- Lack of foresight and preparedness lead to challenges.
- Pressing need for an effective machine learning algorithm for accurate sales predictions.

Variables Impacting Sales:

- The dataset includes information on events and holidays influencing daily sales.
- Economic conditions such as CPI and Unemployment Index are crucial variables.

DATASET AND CHALLENGES

- Promotional Markdown Events
  - Annual markdown events aligned with key holidays.
  - Sales during holidays carry a fivefold weight.
  - Modeling Markdown Impact
- Key challenge: Modeling markdown impact on holiday weeks
  - Difficulty due to incomplete historical data.
  - Diverse Historical Sales Data
- Dataset overview.
  - Historical sales from 45 Walmart stores in diverse regions
- Primary Goal:
  - Predicting weekly sales in advance
  - Emphasis on regression methods
  - Aiding Walmart's preparation for stock and human resources

OBJECTIVES

- Data Examination and Preprocessing
- Outlier Treatment and Analysis
- Sales Forecasting with Regressors
- KNN Regressor Application
- Model Performance Assessment
- 5-fold Cross-Validation for KNN

1. DATA CLEANING AND EXPLORATION

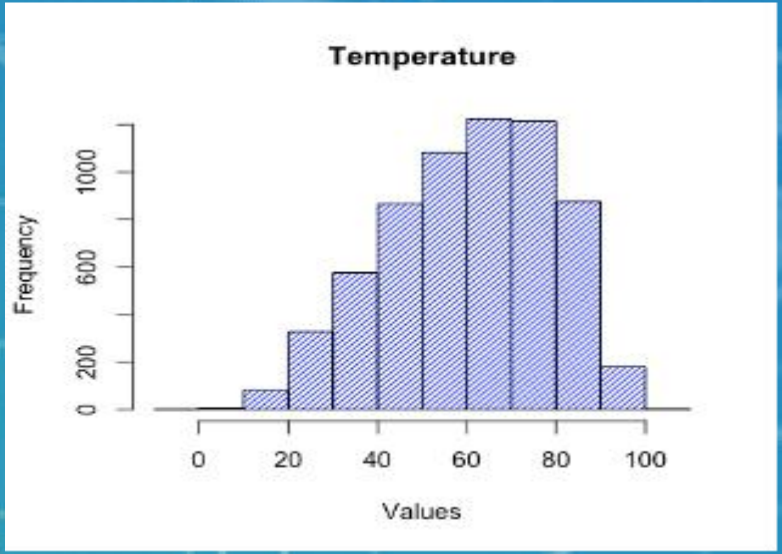
- Identify and eliminate missing values and duplicates
- Create additional features if needed
- Display data distribution of features with graphs

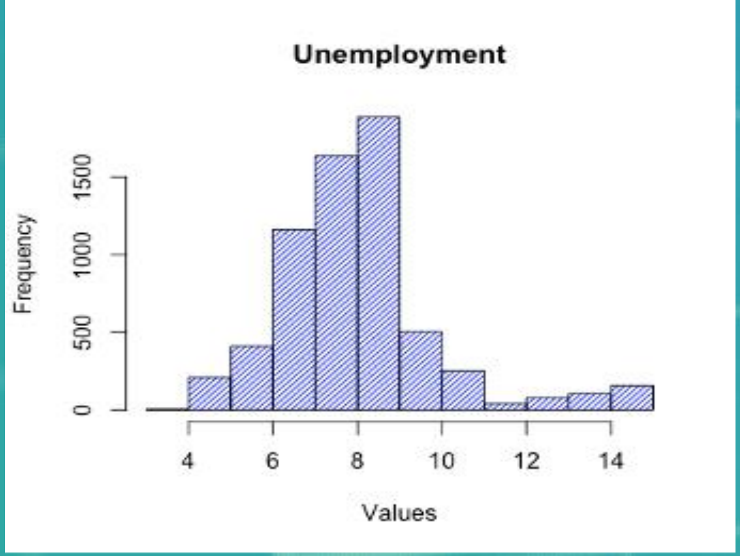
DATA VISUALIZATION ANALYSIS

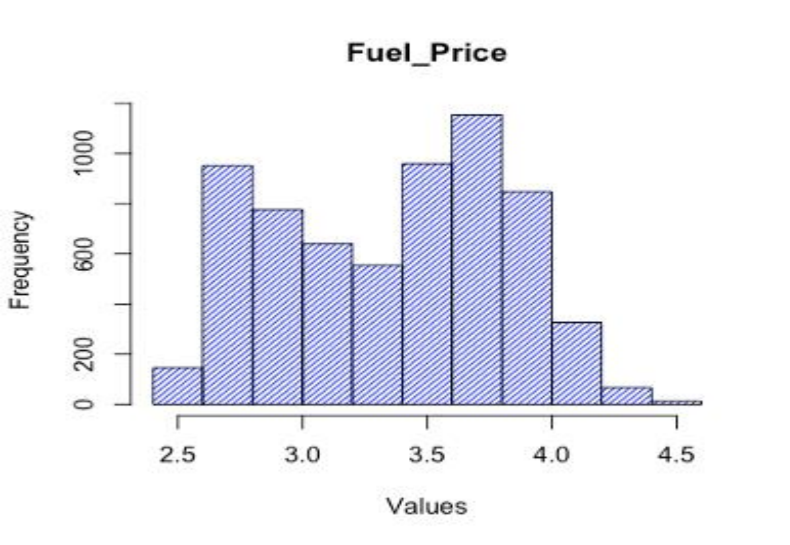
Distribution of Temperature, Unemployment, Fuel Price, CPI, Weekly Sales, Holiday Flags

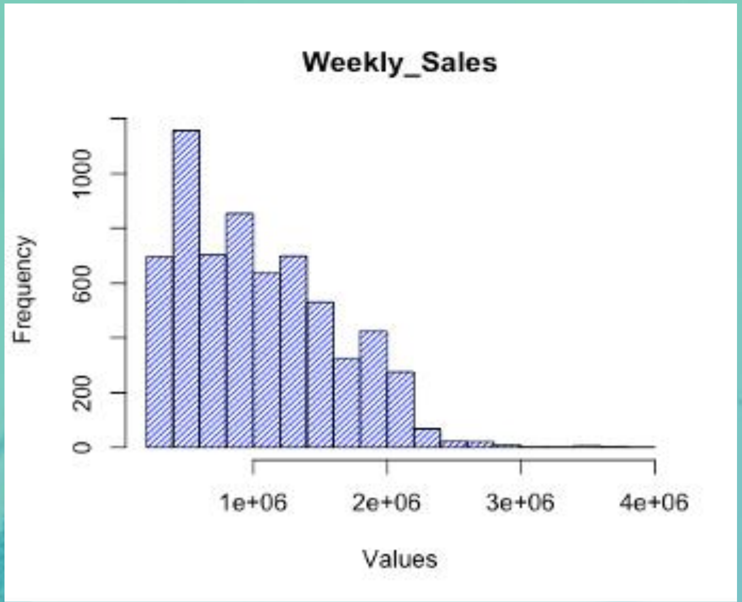
The total no of stores we have analyzed are 45

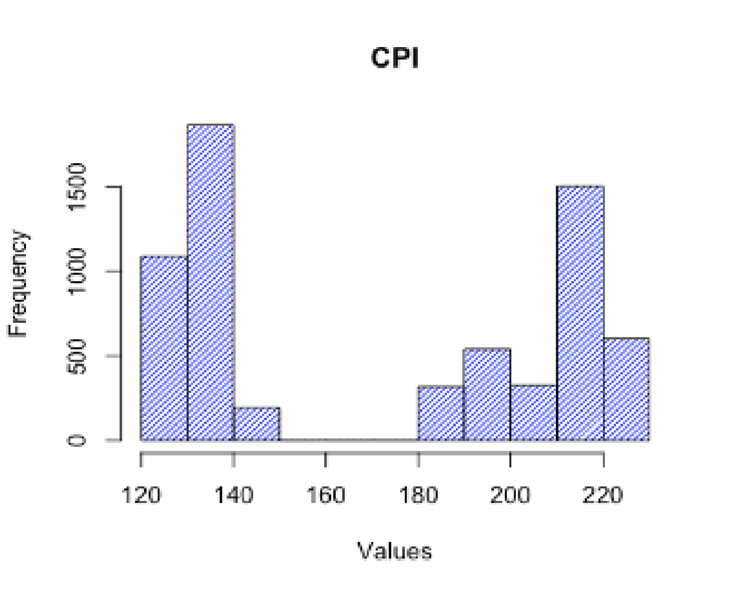
- The Weekly Sales exhibit a right-skewed distribution, which is expected as sales may experience occasional high values.
- Temperature and Unemployment followed a normal distribution
- CPI and Fuel Price display a bimodal distribution.







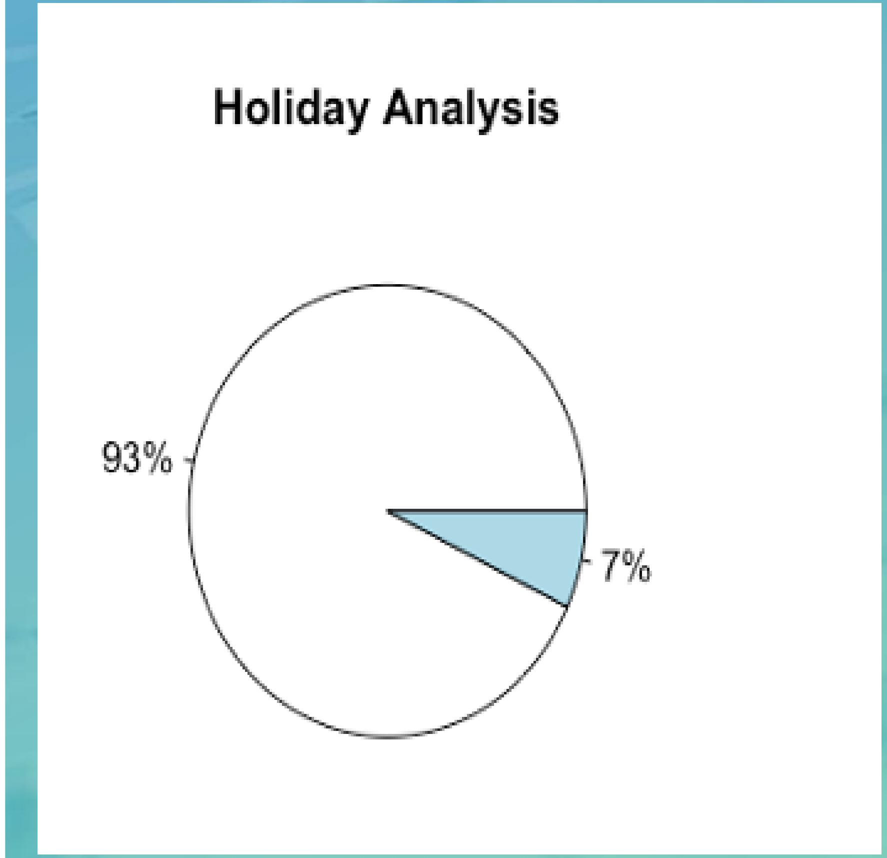




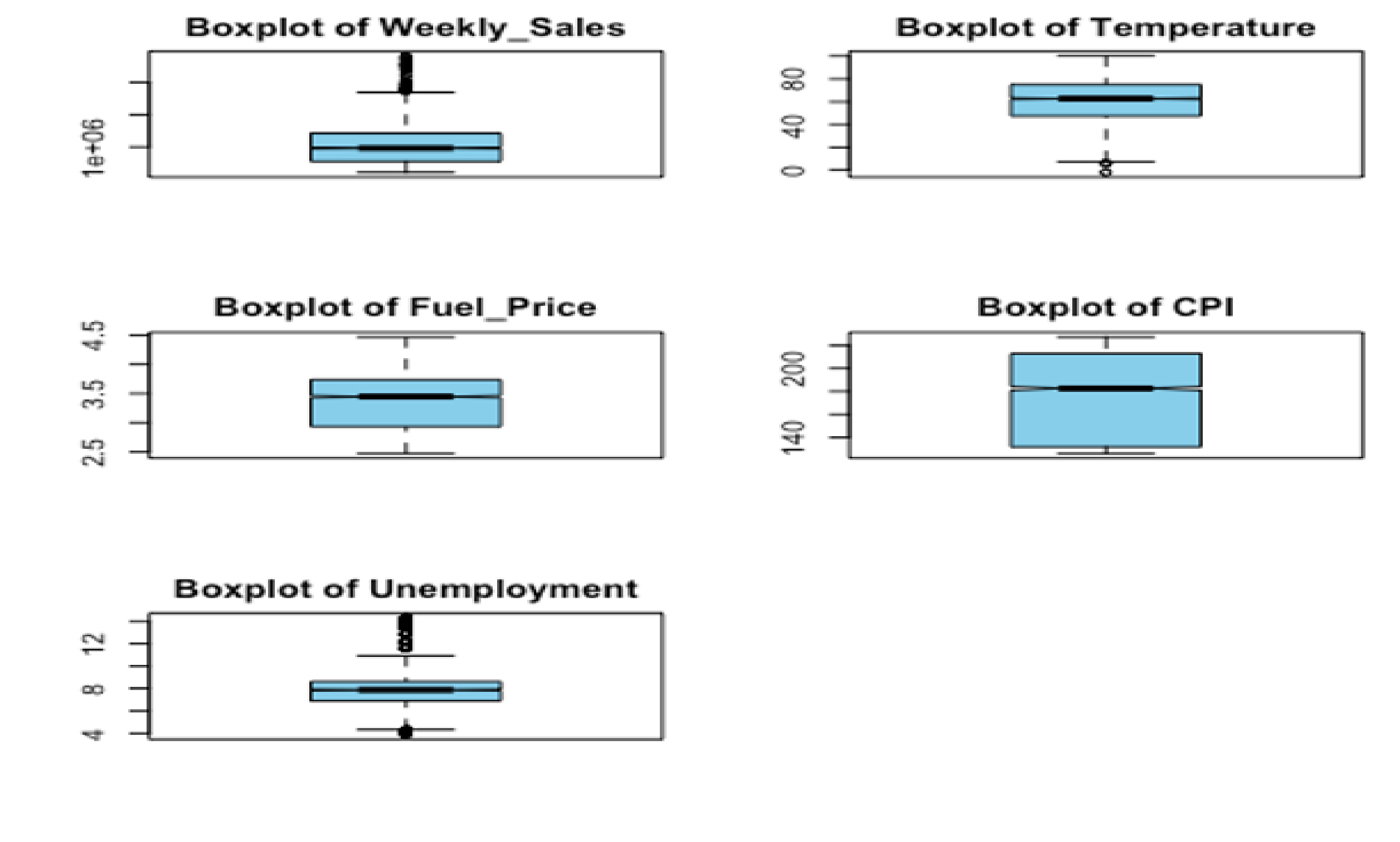
Observations:

- The Weekly Sales exhibit a right-skewed distribution, which is expected as sales may experience occasional high values.
- Temperature and Unemployment followed a normal distribution
- CPI and Fuel Price display a bimodal distribution.

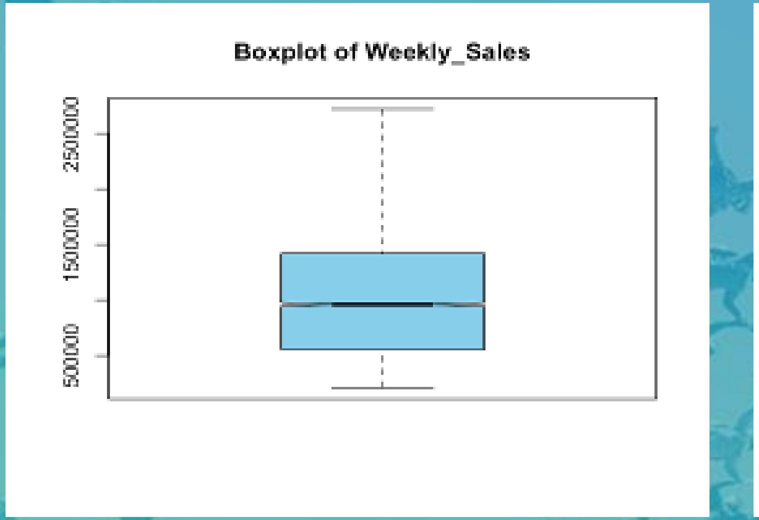
ANALYSIS OF HOLIDAY FLAG

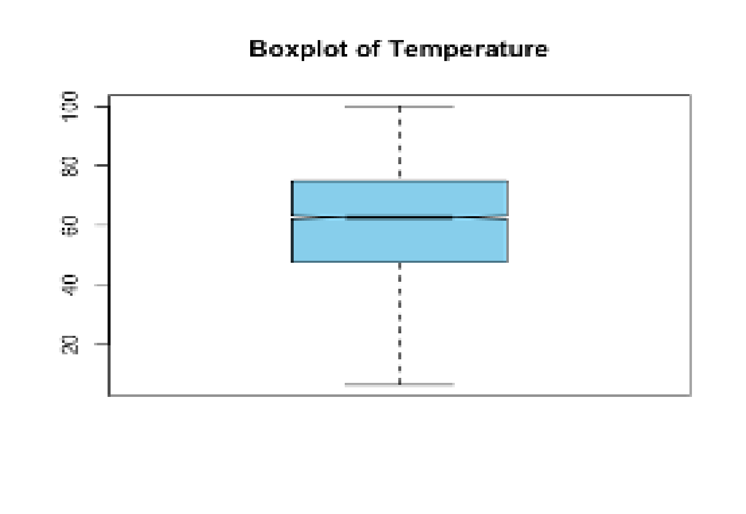


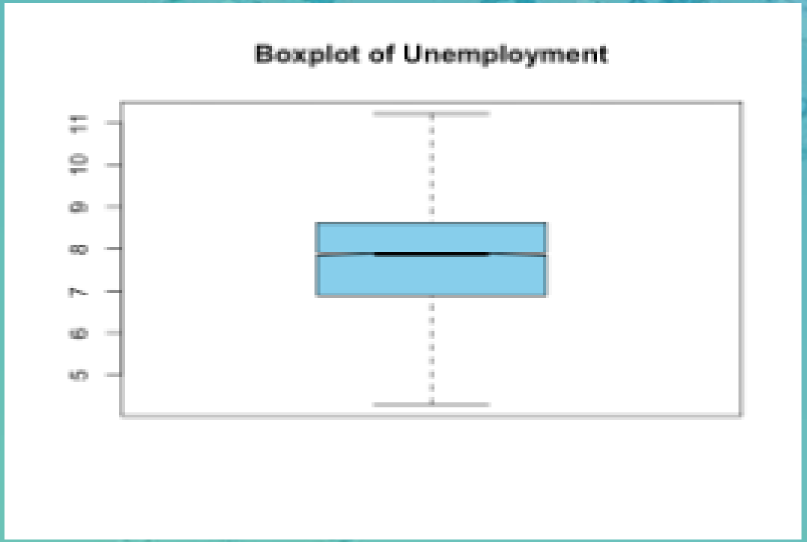
DETECTING OUTLIERS IN THE DATASET



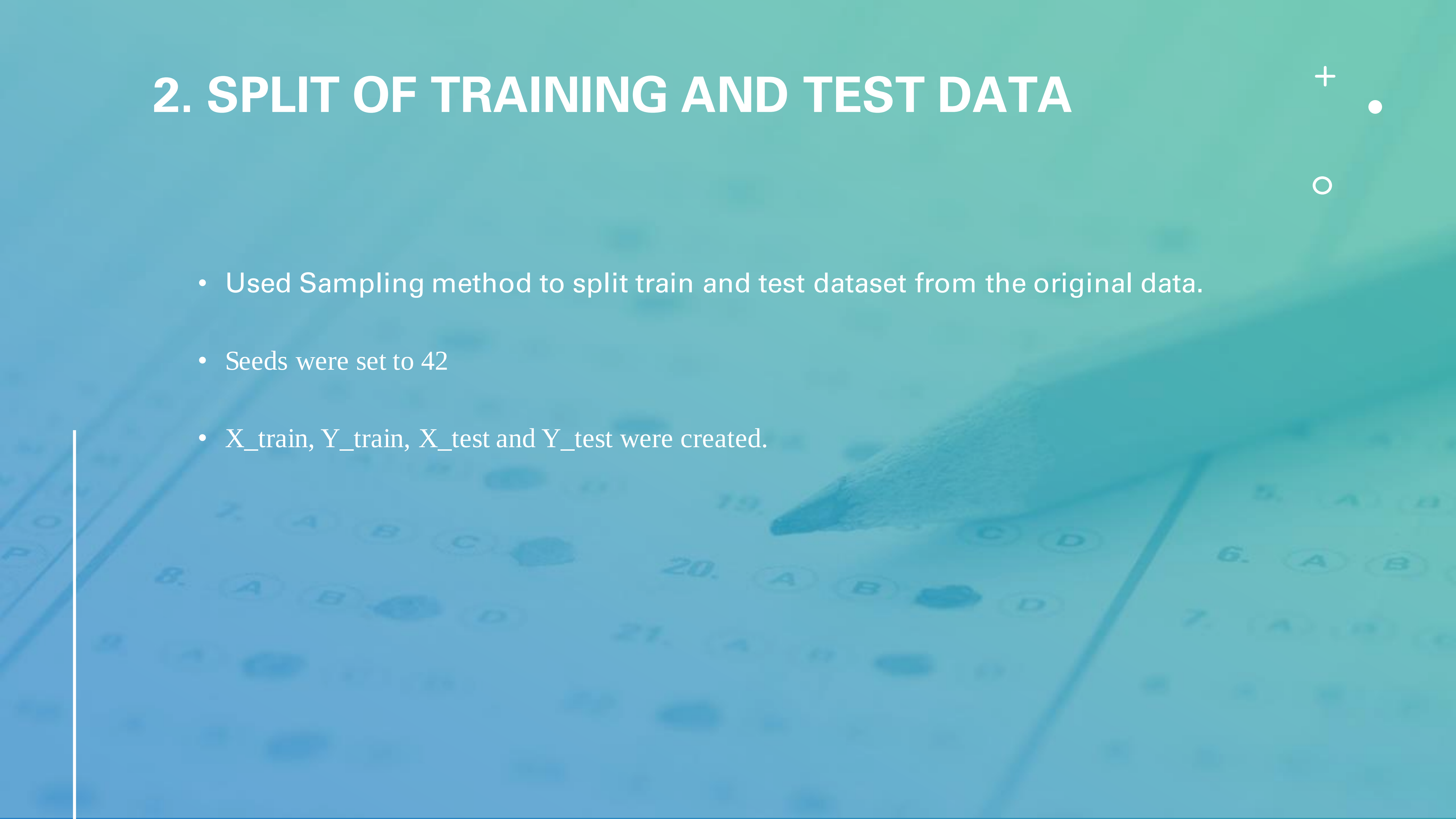
BOXPLOTS AFTER REMOVING OUTLIERS







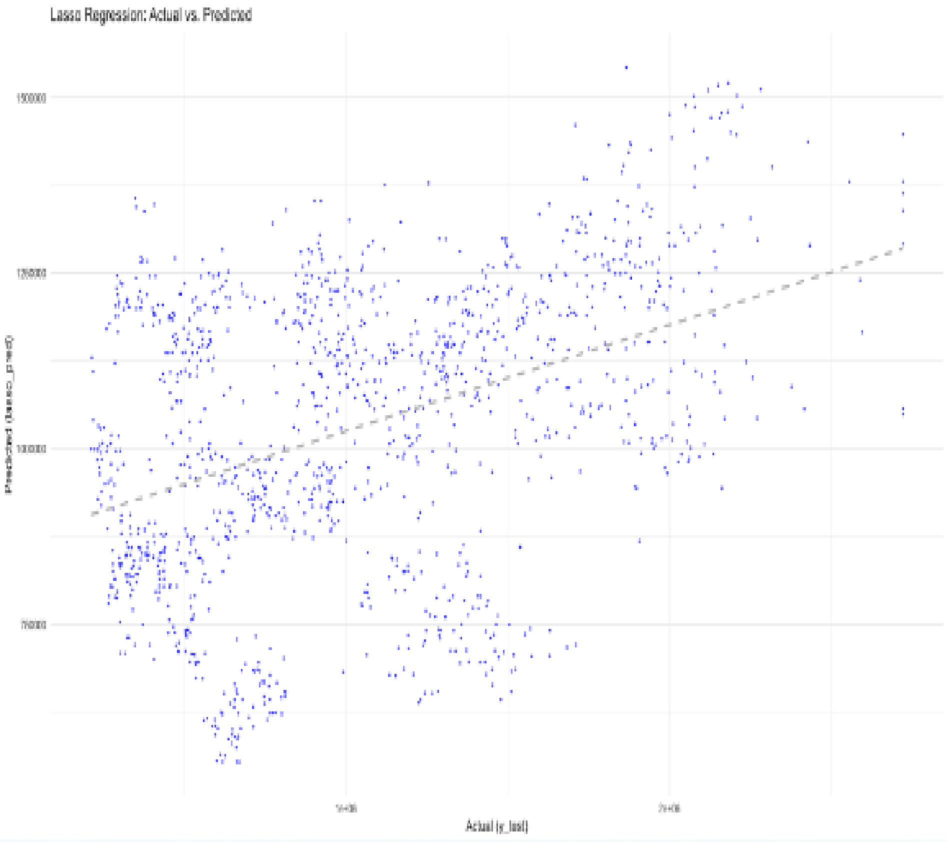
2. SPLIT OF TRAINING AND TEST DATA



- Used Sampling method to split train and test dataset from the original data.
- Seeds were set to 42
- X_train, Y_train, X_test and Y_test were created.

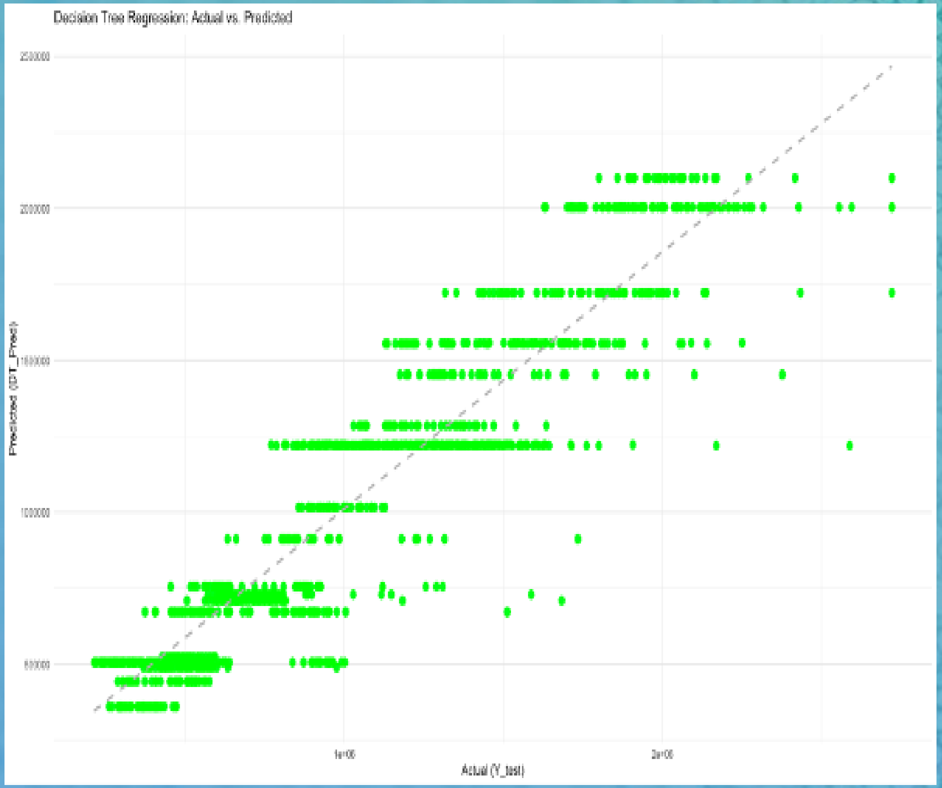
3. LASSO REGRESSION APPLICATION

- Glmnet library was used to implement the lasso regression model.
- X train passed as matrix and alpha selected as 1
- The Root Mean Square for Lasso Regression "511647.015920635"
- R2 score “0.171417631183401”



Actual v/s Predicted for Lasso Regression

4. IMPLEMENTATION OF DECISION TREE



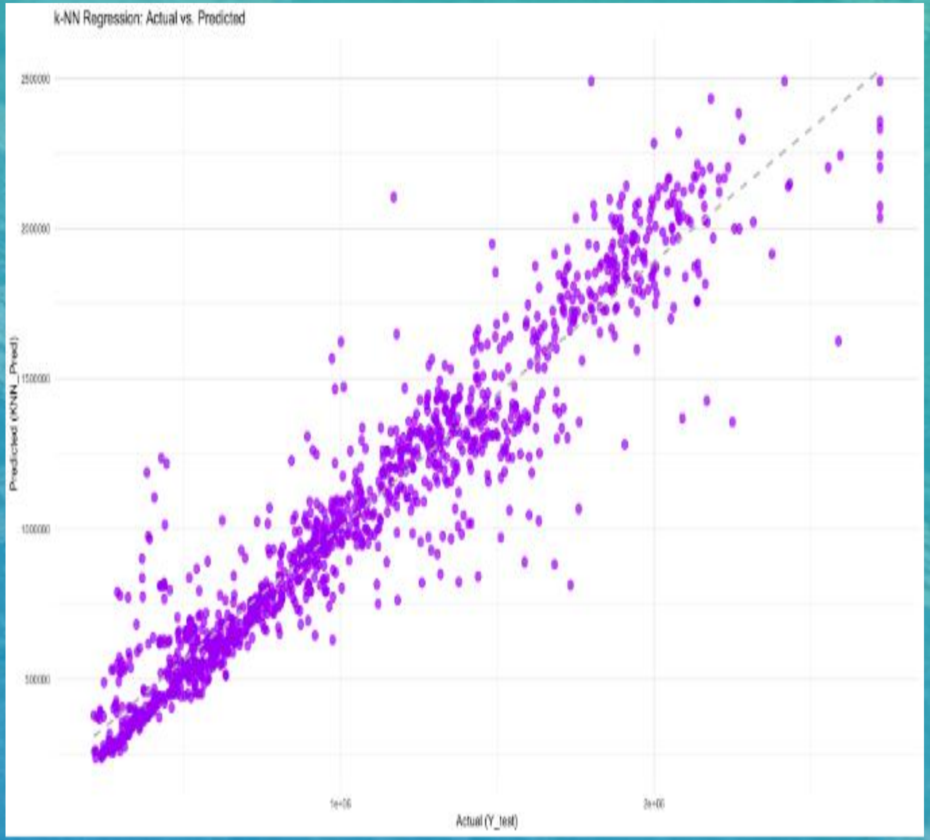
- Rpart library used for Decision tree regression
- Method utilized for implementation “ANOVA”.
- "RMSE for Decision Tree: 211038.267205617" and "R2 score for Decision Tree: 0.858472908977968" respectively.

5. APPLICATION OF KNN REGRESSOR

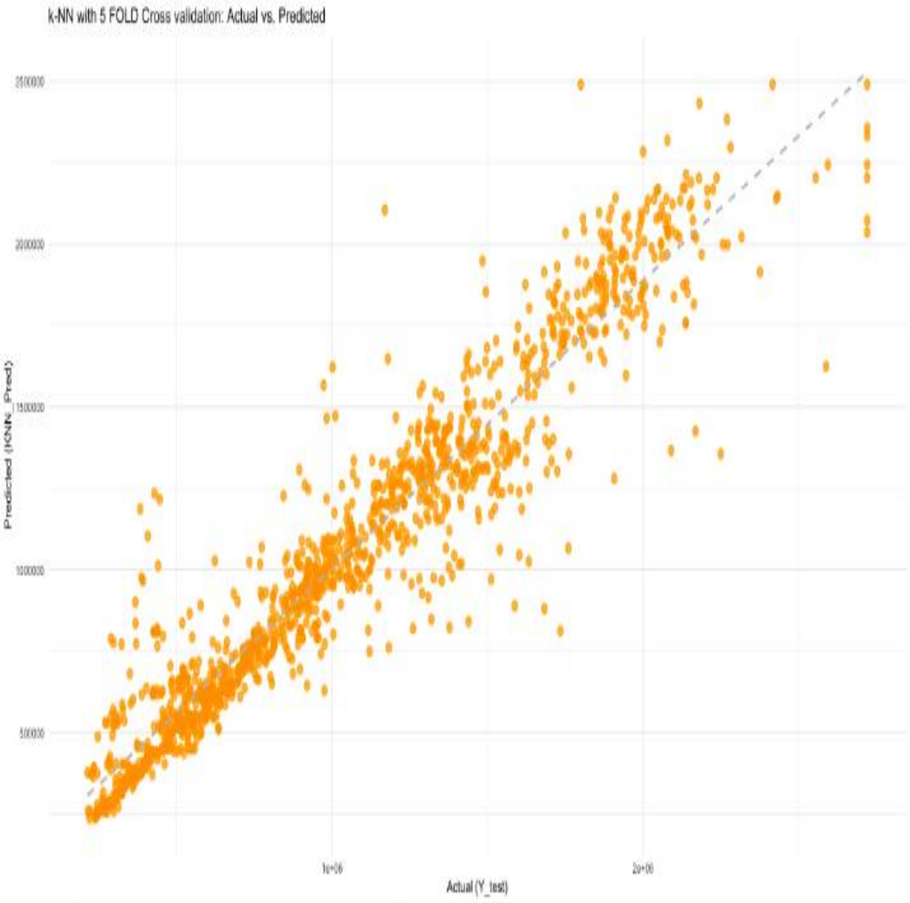
- Library caret was used for training and method used is knn

- Best amongst the three models

- "RMSE for KNN regressor: 170594.876241763" and "R2 score for KNN: 0.907987954336066".


k-NN Regression: Actual vs. Predicted

6. 5-FOLD CROSS VALIDATION APPLIED FOR KNN



- 5 fold cross validation created using cv method:

train_control <- trainControl(method = "cv", number = 5)

- KNN model trained by assigning “Weekly_Sales” as target variable and others as independent variable

model <- train(Weekly_Sales ~., data = data, method = "knn", trControl = train_control)

- Output:

```
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 5149, 5148, 5147, 5148, 5148
Resampling results across tuning parameters:

k  RMSE      Rsquared   MAE
5  176803.9  0.9002398  111261.3
7  179820.7  0.8973014  113559.4
9  186710.1  0.8895853  118799.7
```

PERFOMANCE COMPARSION

This data processing and modelling workflow involves exploring the dataset, applying different regression algorithms, evaluating their performance, and drawing insights from the predictive models to enhance understanding and decision-making in the context of Walmart sales prediction.

```
Resampling: Cross-Validated (5 fold) Summary of sample sizes: 5148, 5147, 5149, 5148, 5148
Resampling results across tuning parameters:

  k  RMSE      Rsquared   MAE
  5  174677.1  0.9027790  110346.3
  7  179134.0  0.8981883  113647.8
  9  186048.7  0.8903392  118087.4
```

OUTCOME:

- The results revealed that KNN exhibited the highest performance with an R-squared score of 0.90, followed by the Decision Tree regressor with a score of 0.85. Conversely, the Lasso regression displayed the least favourable outcome with an R-squared score of only 0.17.

- KNN regressor model demonstrated the most effective predictive capability for Walmart sales.

THANK YOU