# INFLUENTIAL FACTORS AFFECTING SALES AT WALMART AND PREDICTING SALES AT WALMART USING LINEAR REGRESSION.

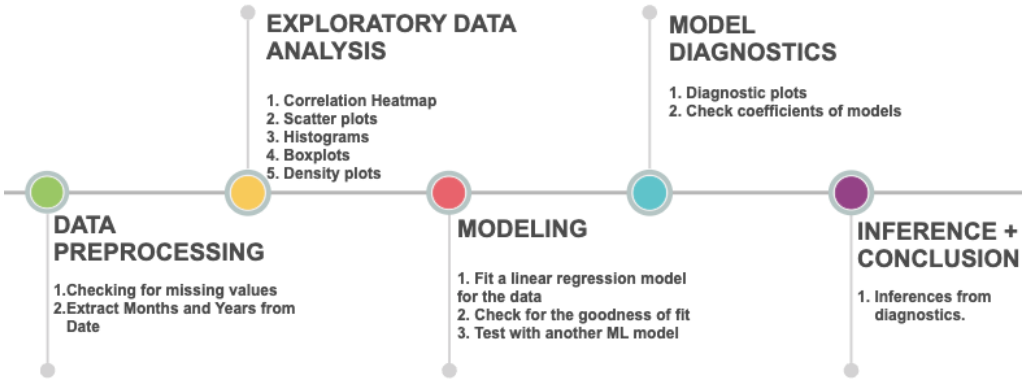
*ANUPAMA KADAMBI* *ASU ID:1222559637*

> **Goal:** To find the most influential factors that contribute to sales at Walmart supercenters.
To fit a linear regression model to the dataset and predict the sales at Walmart.

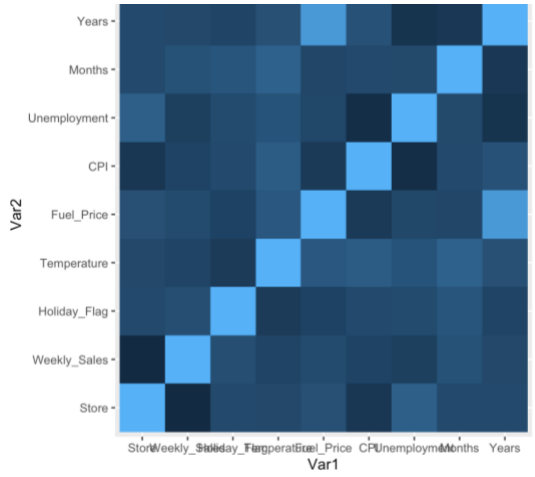
>**Dataset:** Source- Kaggle. Online community for statisticians and data scientists, who publish datasets and are validated by other on the platform.

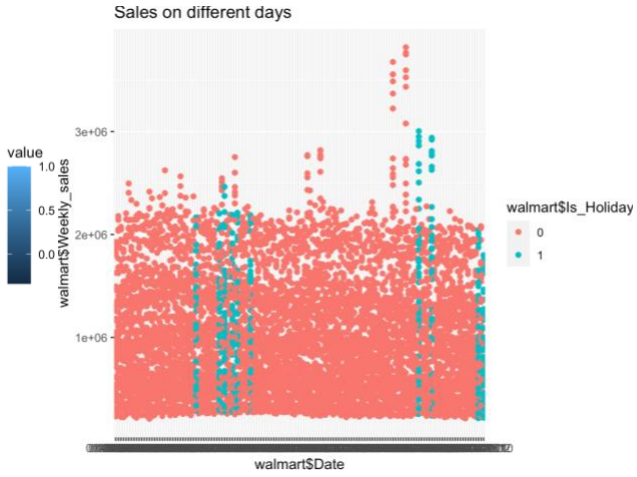
| | | |
|---|---|---|
| Total number of predictors: 8.<br>Response variable- Weekly sales | No. of numerical predictors: 4<br>Temperature, Fuel price, Unemployment rate, CPI. | No. of categorical predictors: 4<br>Store number, Holiday Flag Date (Months, Years) |

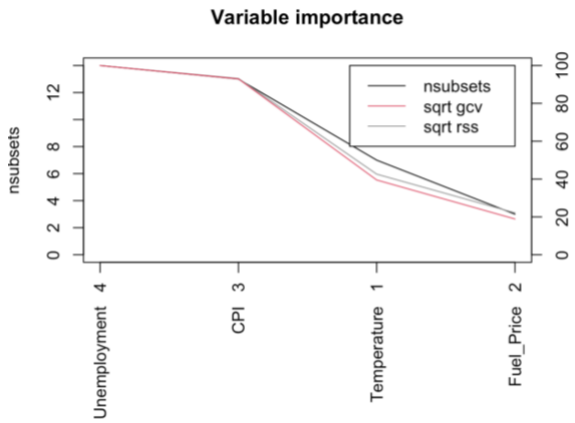
>**Methodology:**




## EXPLORATORY DATA ANALYSIS

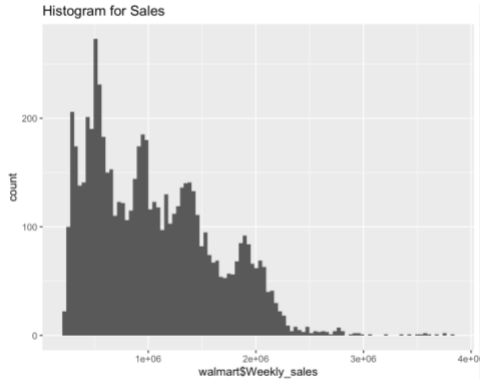


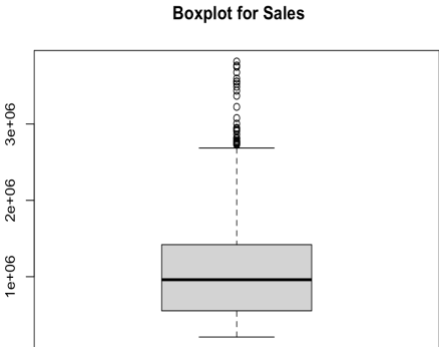



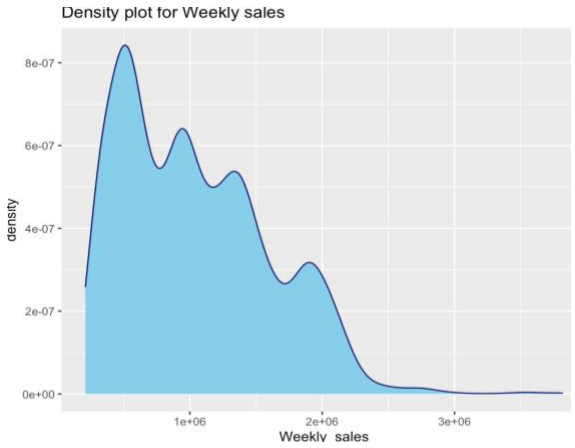



HISTOGRAMS BOXPLOT SALES VS NUM. PREDICTORS.

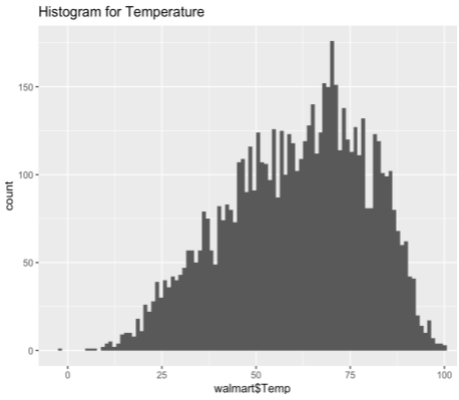
1. Sales

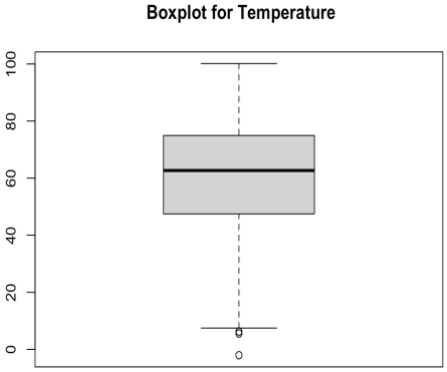


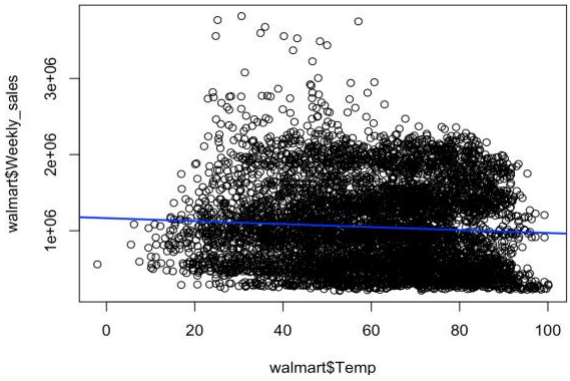



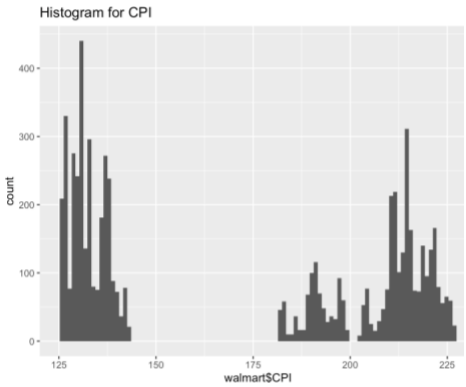

2. Temperature

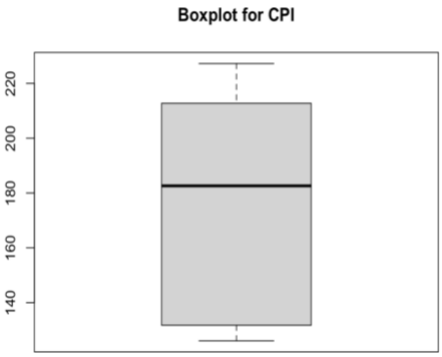


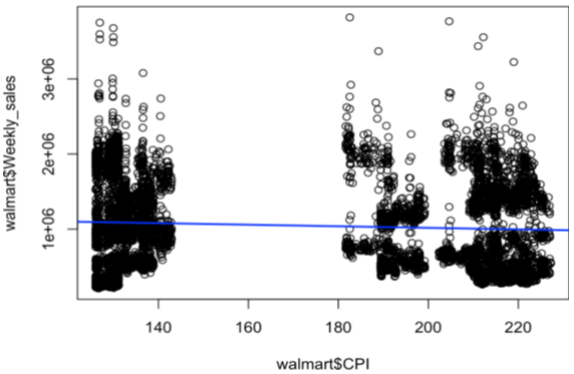



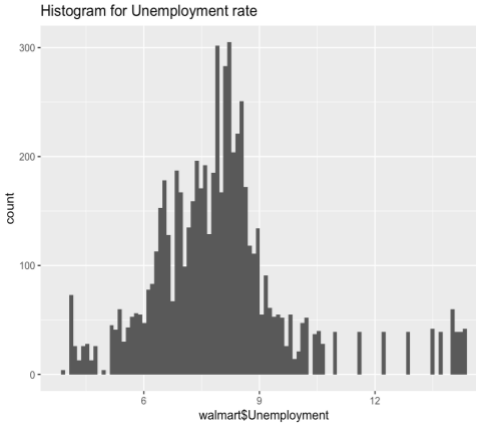

3. CPI

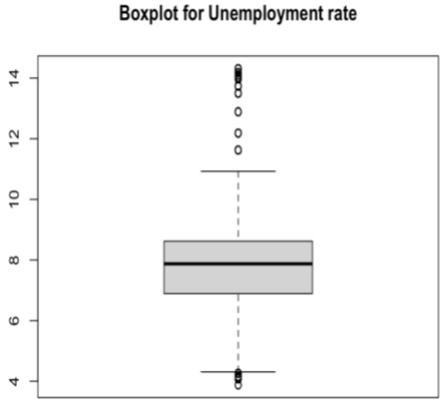


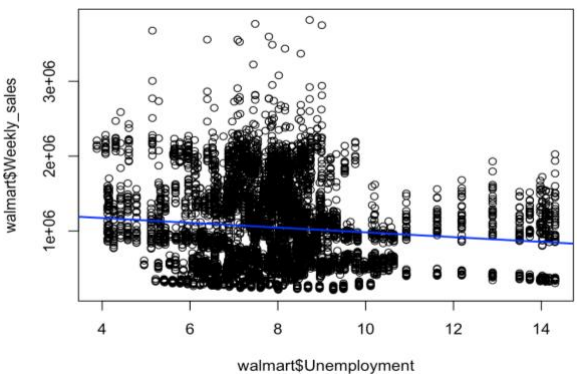



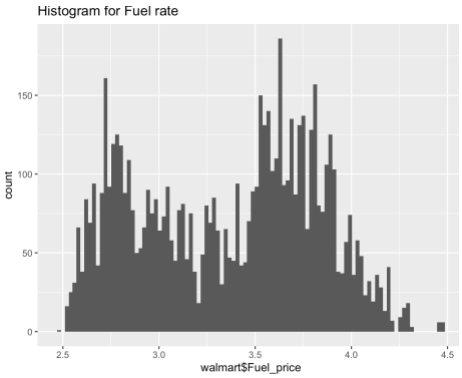

4. Unemployment

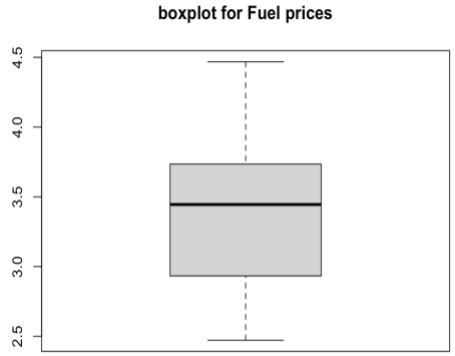


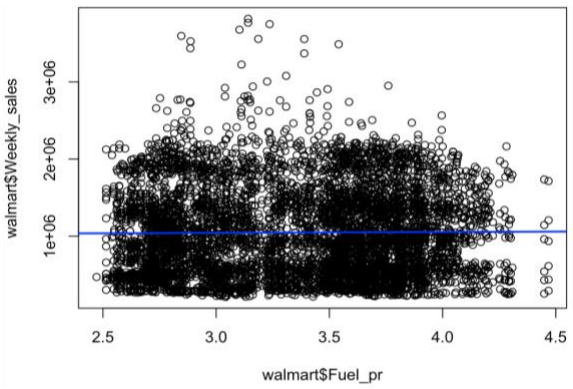



5. Fuel price



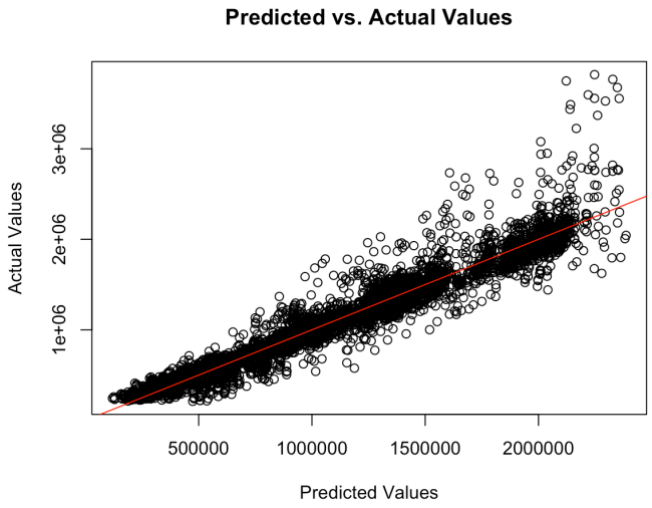




## MODELING USING lm()

model<-lm(Weekly_sales~ Temp+ Fuel_price+CPI+ Unemployment+ Store_number+ Is_Holiday+ Month+ Year, data=walmart)

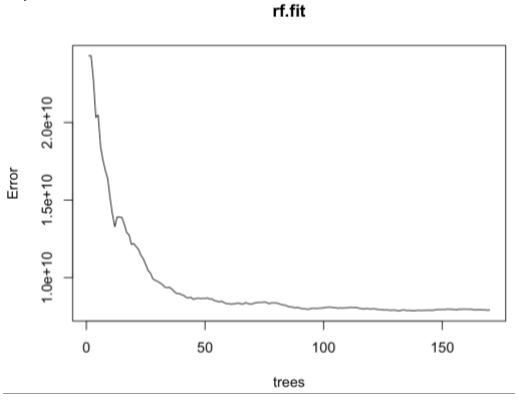
```
Residual standard error: 141900 on 6372 degrees of freedom
Multiple R-squared:  0.9374,    Adjusted R-squared:  0.9368
F-statistic:  1540 on 62 and 6372 DF,  p-value: < 2.2e-16
```

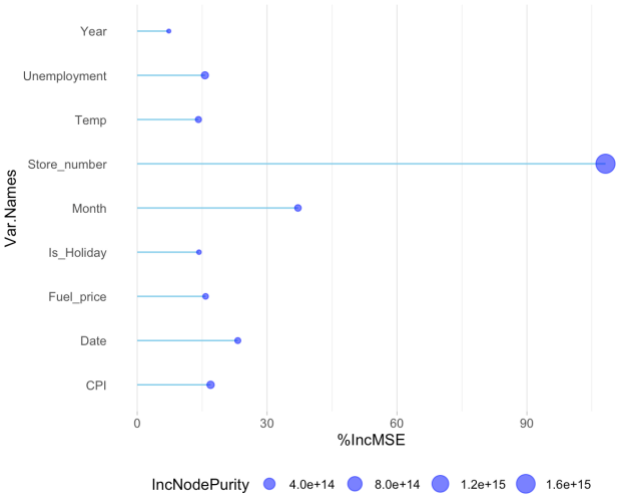


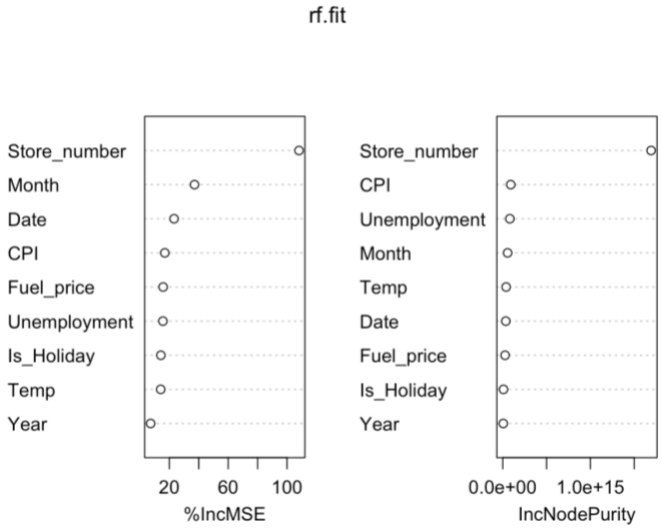

## MODEL USING RANDOM FOREST

rf.fit <- randomForest(Weekly_sales ~ ., data=walmart, ntree=170,keep.forest=FALSE, importance=TRUE)

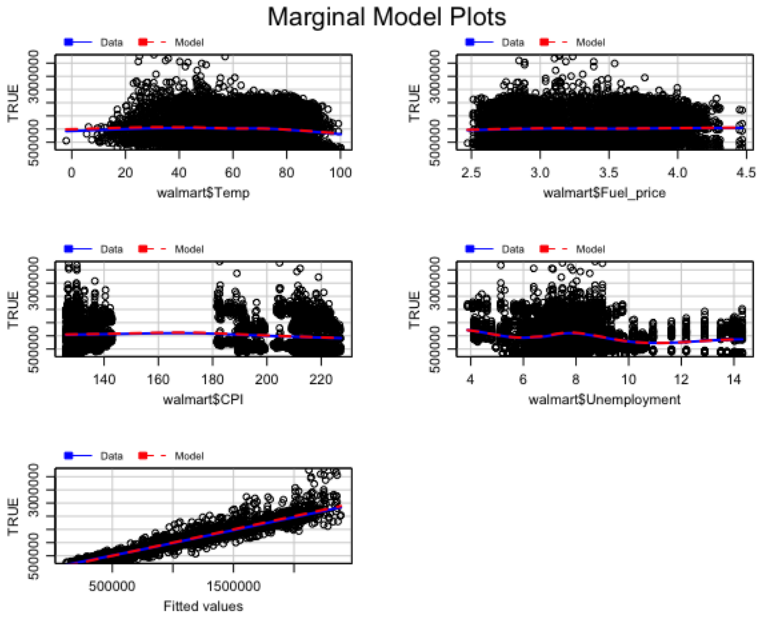
```
Call:
 randomForest(formula = walmart$Weekly_sales ~ ., data = walmart,      ntree = 170, keep.forest = FALSE, impo
rtance = TRUE)
               Type of random forest: regression
                     Number of trees: 170
No. of variables tried at each split: 3

          Mean of squared residuals: 7919188474
                    % Var explained: 97.51
```



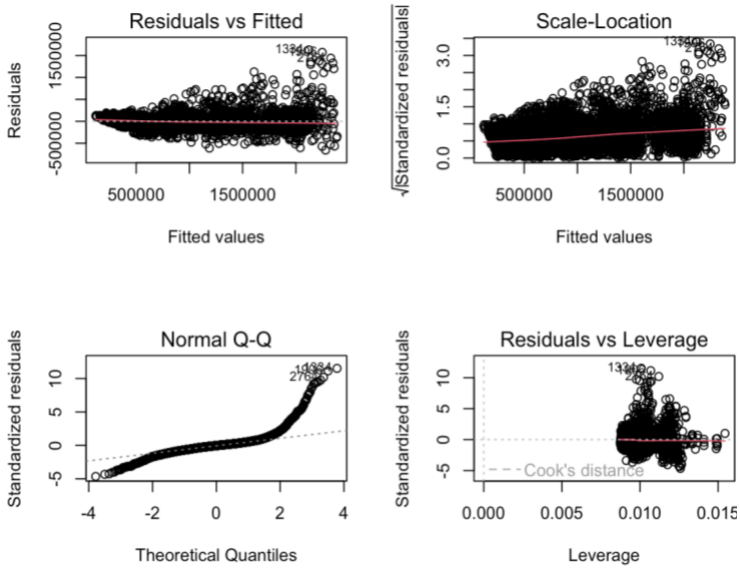


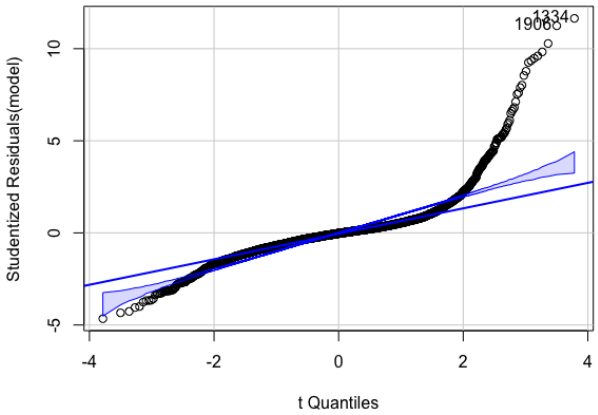

## DIAGNOSTICS




Diagnostic plots




QQ Plot




Residual Plot