Analysis Tasks

Basic Statistics tasks

* Which store has maximum sales

Solution

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated



* Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

Text

Description automatically generated

Chart, histogram

Description automatically generated



Chart, histogram

Description automatically generated



* Which store/s has good quarterly growth rate in Q3’2012

Text

Description automatically generated

Chart, line chart

Description automatically generated

Text

Description automatically generated

* Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together

Text

Description automatically generated

Text

Description automatically generated

* Provide a monthly and semester view of sales in units and give insights

Graphical user interface, text, application

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Statistical Model

For Store 1 – Build  prediction models to forecast demand

* Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.
* Change dates into days by creating new variable.

Graphical user interface, text, application, email

Description automatically generated

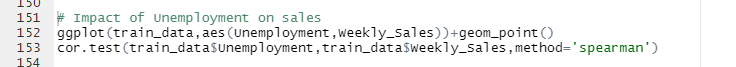
For training data

Text

Description automatically generated

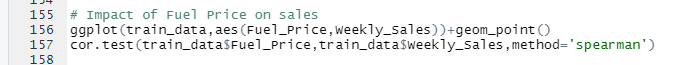
Graphical user interface, text

Description automatically generated



Graphical user interface, text, application

Description automatically generated



Graphical user interface, text, application

Description automatically generated



Text

Description automatically generated

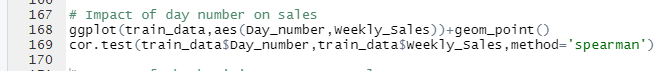


Chart, box and whisker chart

Description automatically generated

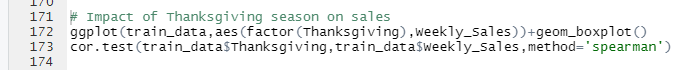
Graphical user interface, text, application

Description automatically generated



Graphical user interface, text, application

Description automatically generated



Chart, box and whisker chart

Description automatically generated

Graphical user interface, text

Description automatically generated



Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, application

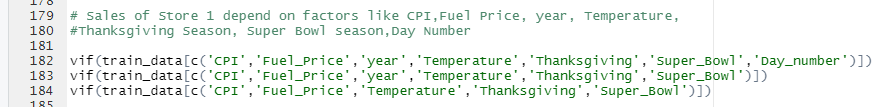
Description automatically generated

Based on training dataset, Weekly Sales of Store 1 has correlation with factors like CPI, Fuel Price, year, Temperature, Thanksgiving Season, Super Bowl Season and Day Number. Since, probability of these features being uncorrelated with Weekly\_Sales<0.05, we can reject the Null Hypothesis which states that they are not correlated with Weekly\_Sales.

However, Unemployment column has p-value >0.05 and therefore we cannot reject Null Hypothesis for Unemployment column for Training Dataset and do not include it in Linear Regaression Model

Now, Variance Inflation Factor (VIF) is calculated to check for Multi-colinearity.

Features with greatest VIF are removed one by one and VIF is recomputed till all remaining features have VIF<10



Graphical user interface, text, application, email

Description automatically generated

After feature reduction using VIF, the remaining Features are CPI, Fuel\_Price, Temperature, Thanksgiving, Super\_Bowl

Now, we perform Linear Regression and remove features with p-val>0.05, till essential features for Linear Regression model having p-value<0.05 are remaining

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Final Linear Regression Model

CPI🡪 Prevailing consumer price index

Temperature🡪Temperature on the day of sale

Thanksgiving🡪 This value is ‘1’ if Thanksgiving holiday lies during the week of sale. Otherwise, it is ‘0’

Predicted Weekly Sales 🡪 Prediction of Weekly Sales

Training Data has R^2 value of 0.248 and Adjusted R^2 value of 0.227

Performance on Test Dataset

Text

Description automatically generated

Text, letter

Description automatically generated

Test Data has R^2 value of 0.359