SALES ANALYSIS OF WALMART DATA

Mayank Gupta, Prerana Ghosh, Deepti Bahel, Anantha Venkata Sai Akhilesh Karumanchi

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

gupta363@purdue.edu, ghoshp@purdue.edu, dbahel@purdue.edu, akaruman@purdue.edu

Abstract

The aim of this project is to enable category managers of Walmart to check the weekly and monthly sales of the departments. Analysis includes the effect of markdowns on the sales and the extent of effect on the sales by fuel prices, temperature, unemployment, CPI etc. has been analyzed using simple and multiple linear regression models. Analytical tools used in the project are Rstudio and Shiny app. The user interface for this project is user friendly as the user has to use the drop down menus and slider inputs to see the variation in graph.

Keywords

weekly sales, monthly sales, temperature, markdowns, store type, store size, department, linear regression

Business Problem

The decision makers of Walmart should be able to analyze the effect of various factors affecting the sales of the products in their 45 stores. The various factors include weather condition i.e., temperature, store size, fuel prices, markdown in prices, unemployment and CPI.

Analytics Problem

In this problem we have analyzed sales across different departments by store type and created weekly and monthly dashboard. We have analyzed the effect of various factors such as temperature, store size, fuel prices, markdown in prices, unemployment and CPI to determine

which factors have a statistical significance in explaining sales in the stores by using simple and multiple linear regression.

Data

The data has been taken from the Kaggle data analytics competition, it contains data of 45 Walmart stores and its various departments. The original data files used for our analysis were **stores.csv**, **train.csv** and **features.csv** which contained the below mentioned fields:

stores.csv: This file contains anonymized information about the 45 stores, indicating the type and size of store.

train.csv: This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week

features.csv: This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.

- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

We merged the data files to our convenience for analysis which have been uploaded in our github page.

Methodology Selection

The methodologies which we used in this project are:

1. Merge different data sets: We had three data files – training, features and stores. We merged all the three data files to see effect of different variables on sale. Since this data was already cleaned, we didn't do any data cleaning work.

2. Study summary descriptive statistics: We have studied how different factors like week, month, store size, temperature effect sales by using ggplot function in R. An interesting observation during descriptive statistics came up where we realized markdowns were also impacting sales. But since markdown data wasn't big enough for linear regression model, we restricted ourselves with just descriptive statistics work.

3. Build linear regression models: We built linear regression in R to predict sales using week of the year, store size and temperature. We have used backward selection model to analyze the effects of various predictors on the sales.

Model Building

We have created separate dashboards to analyze variation of sales for departments with Week, Month, Temperature, Store Size, Markdowns. Next, we tried to get insights by comparing the factors like fuel prices, temperature, unemployment rate, CPI etc., with sales of the stores. We found that there were not much insights by comparing the sales with unemployment, CPI and that

resulted in not much variation in the graphs. Store size, Temperature and Week of the year have shown some interesting findings and it affected the sales by month significantly.

Functionality

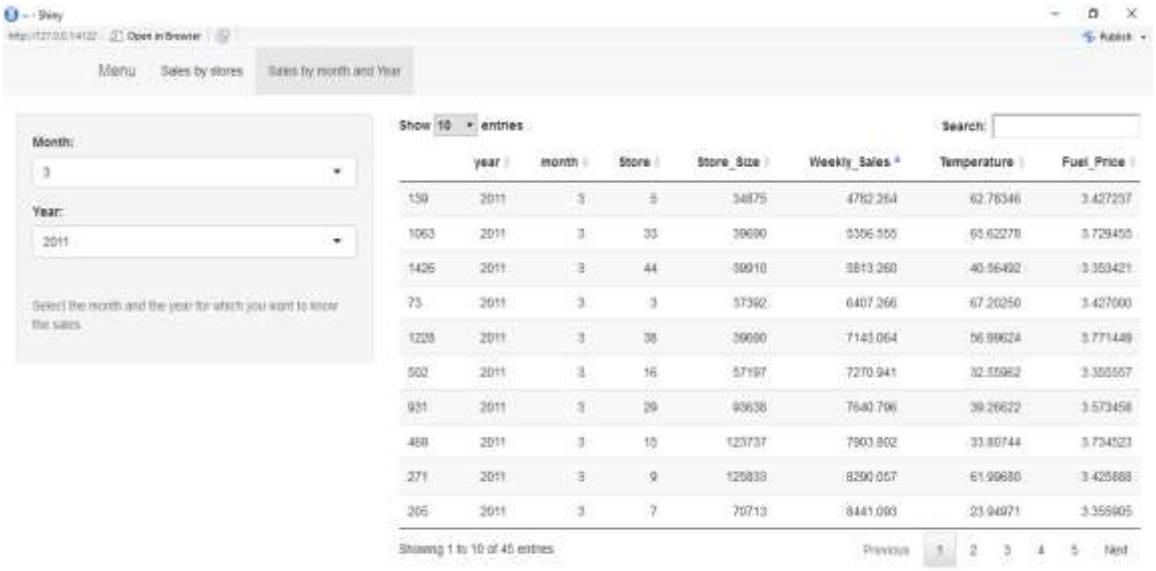
The Shiny apps are built for easy use by a Category Manager to visually analyze the various factors affecting the sales in their stores and to determine which of these factors strongly affect the sales. The first app helps in descriptive analysis. Within the app, the first tab gives the leverage to check the monthly sales by store and its individual departments. Second tab gives the leverage to get to know the performance of the stores and we can also sort the stores by the store size. Next tab will give weekly sales versus the store type for different years. In third tab user can see how sales have been for his department by store and year. In fourth tab, user can see how markdowns impacting sales by store type for different departments. In fifth tab, user can see how sales vary with store sizes.

The second app helps in linear regression analysis. Within this app, the left panel has an option for the user to input the portion of data to be included in training data set; an option to select the regression model to summarize and various other options to change the user preference. The first tab shows the summary of regression model and the statistical metrics of the trained data set such as R^2, adjusted R^2, f-value, t-value, p-value etc. The second tab shows the quality of the prediction of this trained model by showing the R^2 and root mean square error of the training and test data sets. The third and the fourth tabs show the training and the test data sets.

GUI Design and Functionality

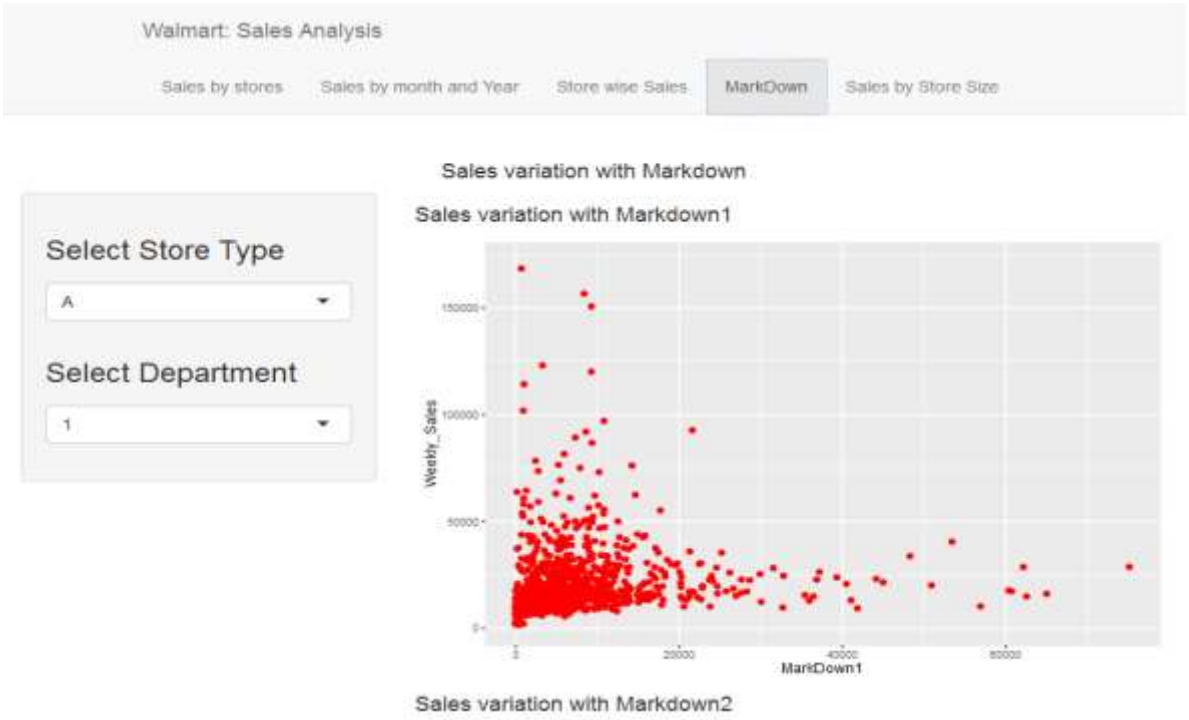
The first app helps in descriptive analysis of the data. The first tab (shown below) depicts the sales by stores. The user inputs the store number from 1 to 45 and the department number from 1 to 99. The corresponding sales vs month of the year, sales vs temperature etc are displayed.

The second tab (below) shows the data with respect to the year and month. It takes the user input of month from 1 to 12 and year from 2010-2012.



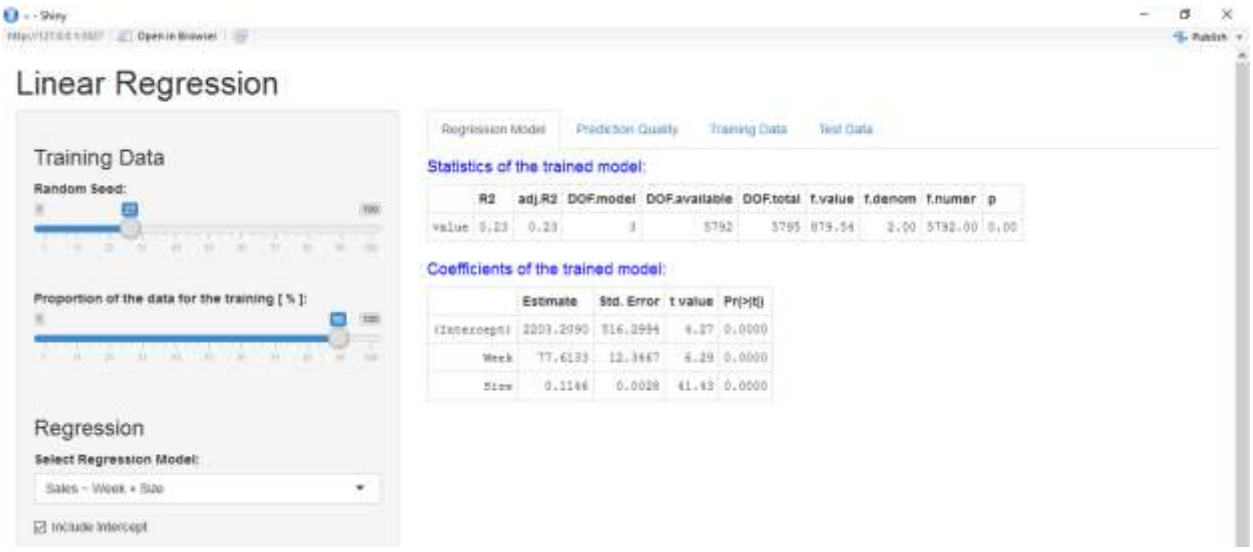
The third tab (below) shows the store-wise sales. It takes input from the user to input the store type of A, B or C; the department number from 1 to 99 and the year from 2010-2012. The corresponding output shows the Weekly Sales categorized by the store numbers.

The foutrth tab (below) shows the effect of markdowns on the sales. The user inputs the store type and the department number to find out how slaes varies with Markdown 1, Markdown 2, Markdown 3, Markdown 4 and markdown 5.

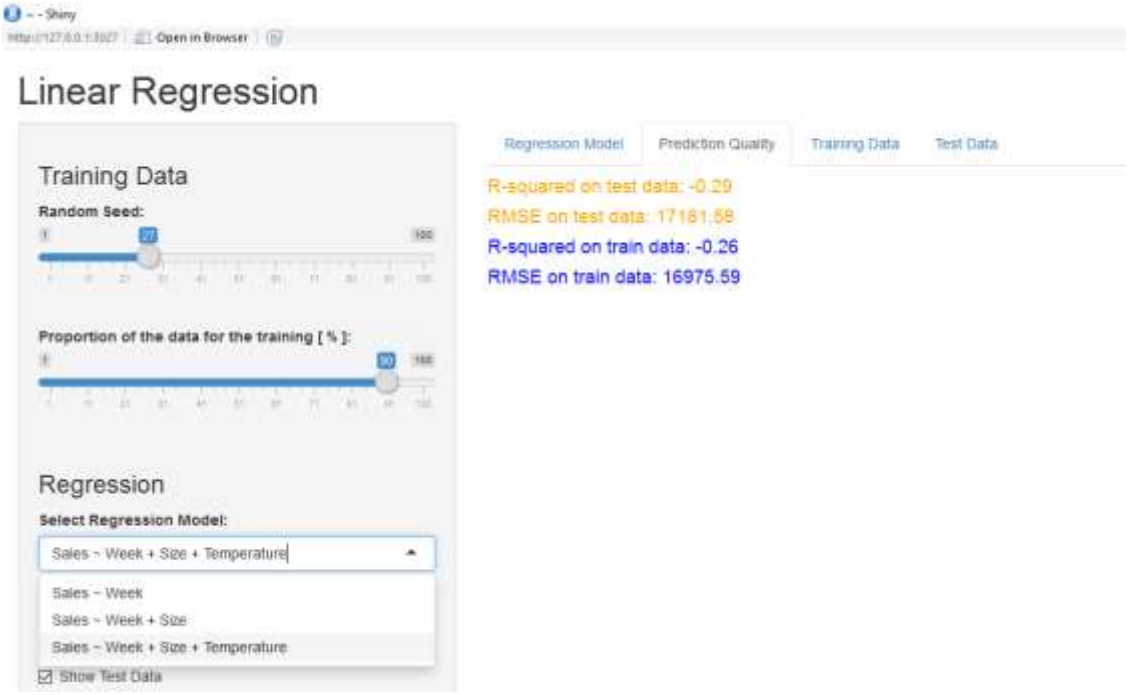


The fifth tab (below) shows how sales is affected by store size. The user inputs the year and the department number and output shows the graph of sales categorized by store type A, B and C using different colors.

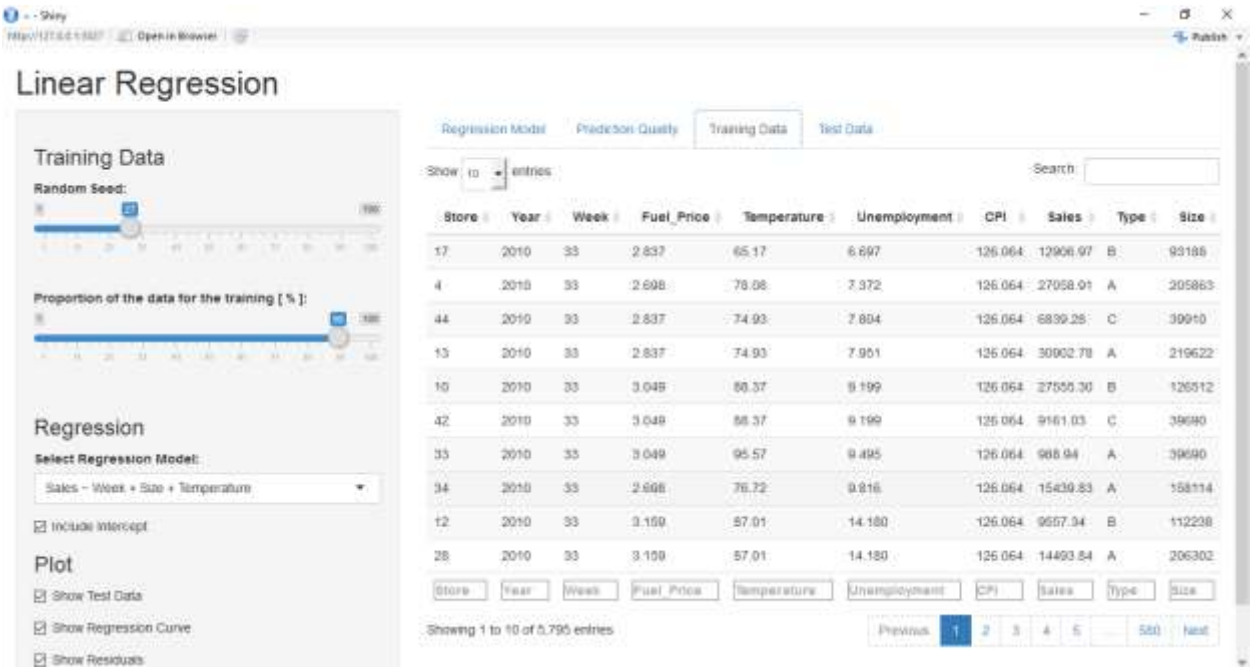
The second app shows the linear regression analysis in effect. With in this app, the left panel has the user input the portion he/she wants to be used for training the data and the user selects the regression model (Sales ~ Week, Sales ~ Week + Size, Sales ~ Week + Size + Temperature) he/she wants to analyse. The first tab (below) shows the summary of the regression analysis and the statistics and coefficients of the trained model.

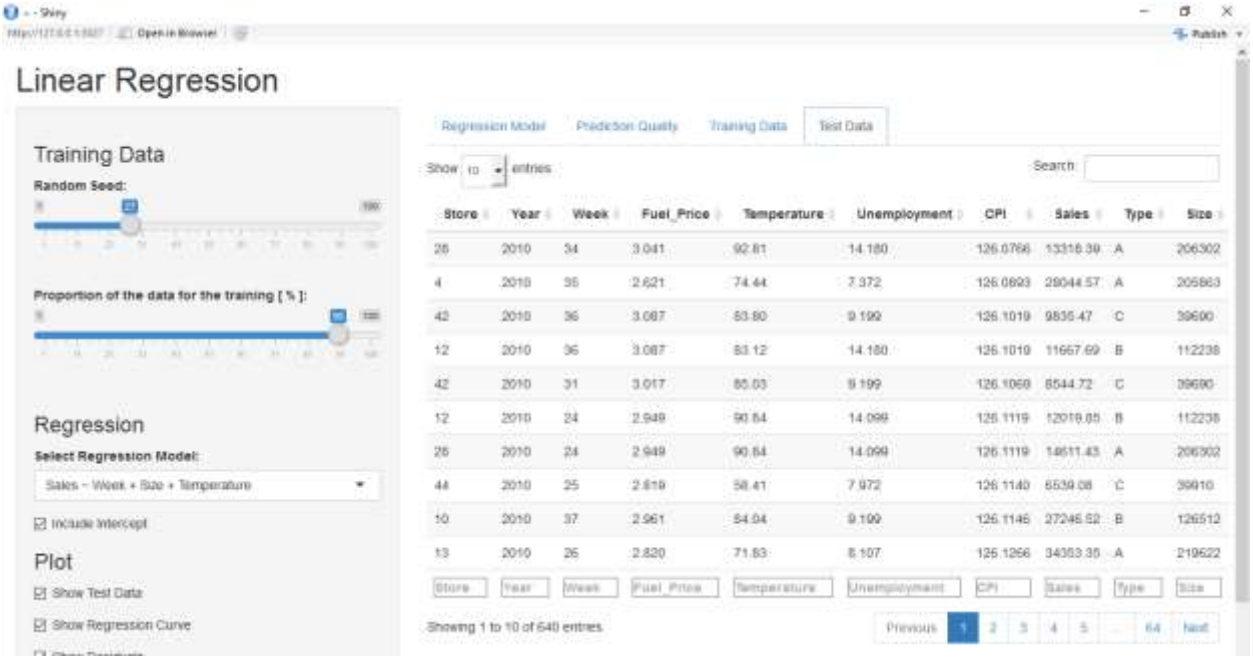


The second tab (below) shows the quality of prediction of the training and test data sets of these three regression models. It shows the R^2 and RMSE values of the test and training data sets.



The third and the fourth tabs (below) show the traing and test data sets based on the portion of data to be included in the training of the data selected by the user.


Linear Regression
Training Data
Random Seed:
Proportion of the data for the training [%]:
Regression
Select Regression Model:
Sales ~ Week + Size + Temperature
Include Intercept
Plot
Show Test Data
Show Regression Curve
Show Residuals
Regression Model
Prediction Quality
Training Data
Test Data
Show 10 entries
Search:
Store Year Week Fuel_Price Temperature Unemployment CPI Sales Type Size
17 2010 33 2.837 65.17 6.697 126.064 12906.97 B 93188
4 2010 33 2.698 78.08 7.372 126.064 27058.01 A 205863
44 2010 33 2.837 74.93 7.804 126.064 6839.26 C 39910
13 2010 33 2.837 74.93 7.951 126.064 30902.78 A 219622
10 2010 33 3.048 68.37 9.199 126.064 27555.30 B 126512
42 2010 33 3.048 68.37 9.199 126.064 9161.03 C 39690
33 2010 33 3.049 95.57 9.495 126.064 988.94 A 39690
34 2010 33 2.698 76.72 9.816 126.064 15439.83 A 158114
12 2010 33 3.159 67.01 14.180 126.064 9557.34 B 112238
28 2010 33 3.159 67.01 14.180 126.064 14493.84 A 206302
Showing 1 to 10 of 5,795 entries
Previous 1 2 3 4 5 580 Next


Linear Regression
Training Data
Random Seed:
Proportion of the data for the training [%]:
Regression
Select Regression Model:
Sales ~ Week + Size + Temperature
Include Intercept
Plot
Show Test Data
Show Regression Curve
Regression Model
Prediction Quality
Training Data
Test Data
Show 10 entries
Search:
Store Year Week Fuel_Price Temperature Unemployment CPI Sales Type Size
28 2010 34 3.041 92.81 14.180 126.0766 13316.39 A 206302
4 2010 35 2.621 74.44 7.372 126.0893 28044.57 A 205863
42 2010 36 3.087 83.80 9.199 126.1019 9836.47 C 39690
12 2010 36 3.087 83.12 14.180 126.1019 11667.69 B 112238
42 2010 31 3.017 85.03 9.199 126.1069 8544.72 C 39690
12 2010 24 2.949 90.64 14.099 126.1119 12019.85 B 112238
28 2010 24 2.949 90.64 14.099 126.1119 14611.43 A 206302
44 2010 25 2.619 58.41 7.972 126.1140 6539.08 C 39910
10 2010 37 2.961 84.04 9.199 126.1146 27246.52 B 126512
13 2010 26 2.820 71.83 8.107 126.1256 34393.35 A 219622
Showing 1 to 10 of 640 entries
Previous 1 2 3 4 5 64 Next

Conclusion

We concluded that Walmart can maintain its store size between 100000 ft. to 150000 ft. to maintain optimum sales and in the month of August as many states in United States offer a tax free on sales, the sales shoot up though there were no markdowns in that month. The analysis shows that in the April and December, store types A & B shows a very high sale because of the special days like Good Friday and Christmas. High markdowns are not helping the sales, the ideal range of markdown 1 is (0 to 18000), markdown 2 is (0 to 10000), markdown 3 is (0 to 500), markdown 4 is (0 to 6000), markdown 5 is (0 to 12500). The relationship between the sales and temperature was varying with each and every department. Hence, the above were the insights drawn by the analysis of the sales data of the Walmart

Next Steps

As a next step, we would like to analyze effect of different markdowns on sales by using advance statistical techniques such as bootstrap method.

References

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

http://shiny.rstudio.com/gallery/basic-datatable.html

http://shiny.rstudio.com/gallery/

https://github.com/FBracun/DevelopDataProducts