****

**PROJECT WRITE-UP – RETAIL ANALYSIS WITH WALMART DATA**

**Submitted on : 12-02-2023**

**Submitted by : Manisankar Baskar**

**PROJECT WRITE-UP**

**INTRODUCTION :**

The project that I chose to work upon is “Retail Analysis with Walmart Data”. I used R Programming executions for this Project to arrive at multiple statistical inferences and to build a Linear model that predicts the Weekly sales of the Walmart Business. The purpose of choosing this topic is that it covers a variety of concepts such as Exploratory Data analysis, Working around dataframes, grouping the data, getting statistical insights, exposure to different approaches to build a model and so on. It also tests my decision making ability to conclude on the best model from business standpoint.

In this write-up, I will elaborate on my approach / Procedural steps that I followed while creating the Prediction model.

**BUSINESS REQUIREMENT:**

The following are the project tasks :

* Analysing the store with maximum sales.
* Analysing the store with maximum Standard Deviation.
* Finding the store with good quarterly growth rate for Q3’2012.
* Arriving at holidays with higher sales than mean sales in non-holiday season for all stores together.
* Getting Monthly and semester sales views from the Dataset.
* Finding weekdays pertaining to Weekly sales across the stores.
* Building Linear models for Store 1, comparing it and decide on the best model for Sales Prediction to Forecast demand.

**OBJECTIVES :**

\*To derive the data and visualize it so as to make data more readable.

\*To get better insights on relationship between the variables and its impact.

\*To transform the data for Linear models and present the data visualizations and sales predictions the most efficient way possible.

\*To analyze the difference between actual and predicted sales for the business so that the business can better plan its inventory accordingly based on demand.

\*To use the model to train similar datasets in future.

\*To help businesses to analyze the best model that gives better Sales predictions

\*To support the business to make important decisions based on predictions.

**MY APPROACH TO BUILD A LINEAR MODEL for STORE 1:**

\*Data extraction for Store 1 from the Walmart Dataset.

\*Check for summary on Store 1 to identify Null values, Missing values and Outliers

\*Perform Data related assumptions

Check for Linear relationship between Dependent and Independent variables by Plotting correlation / graphs that better explains the data.

\*Checking for Multicollinearity between Independent variables using VIF(Variance Inflation Factor) scores and fixing them if it exists.

\*Follow the Model building Approaches :

a.) VIF Approach

b.) Significance Approach

c.) Step Approach

\*Documenting the Statistical outcomes of every model

\*Deciding / Shortlisting the best available model based on Statistical results by model comparisons.

\*Performing Error related assumptions on Shortlisted model:

a.) Homoscedasticity of Error Terms

b.) Normality of Errors

c.) Autocorrelation of Error Terms

\*Finding the Mean Absolute Percentage Error(MAPE)

\*Concluding the best model

\*Sales Predictions to forecast demand based on the best model that we built.

**STATISTICAL INFERENCES** :

For the Dataset that covers sales from 2010-02-05 to 2012-11-01,

1. Store 20 has the Highest overall Sales.
2. Store 14 has the Maximum Standard Deviation for sales
3. Store 7 had the highest quarterly Growth rate for Q3’2012
4. Following are the holiday events with higher sales than mean sales in non Holiday season for all stores together.

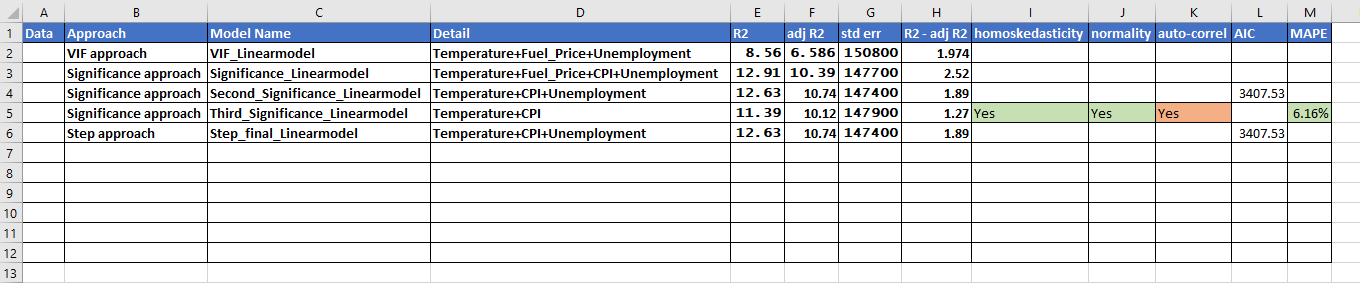
# 2010-02-12, 2011-02-11, 2012-02-10 - Super Bowl

# 2010-11-26, 2011-11-25 - Thanksgiving

# 2012-09-07 - Labour Day

1. There is a steady increase in Sales month on month for Semester 1 and Semester 2 saw a decline in sales month on month.
2. July month recorded the Highest overall sales across the stores.
3. Jan month has the lowest overall sales across the stores compared to other months.
4. Semester 2 has the highest average sales across all the stores. Walmart has to increase the inventory for Semester 2 to meet the consumer demand.
5. There is no linear relationship between Dependent variable “Weekly\_Sales” with other Independent variables such as – Temperature, CPI, Unemployment and Fuel\_Price.
6. There is a Multi-collinearity between Independent variable “CPI” with other Independent variables – Temperature, Unemployment and Fuel\_Price

**MODEL DOCUMENTATION :**



**Model 1 – VIF Approach**

**Model Name – VIF\_Linearmodel :**

\*As you can see, we have an R2 value of 0.0856 which translates to 8.56 percent. R2 is very low than expected.

\*Variable 'Temperature' has a p value less than 0.05 and falls within the range 0 to 0.001.

\*Actual p value for variable 'Temperature' is 0.00156 and is statistically significant.

\*However, p value for variables 'Fuel\_Price' and 'Unemployment' are greater than 0.05 threshold.

\*It means, that the variables 'Fuel\_Price' and 'Unemployment' is not explaining much of dependent variable 'Weekly\_Sales'.

**Model 2 – Significance Approach**

**Model Name – Significance\_Linearmodel :**

\*Variable 'Fuel\_Price' has the highest p value of 0.50696 which is very far from 0.05 threshold.

\*Hence I am dropping the variable 'Fuel\_Price' as it is not explaining much of my dependent variable 'Weekly\_Sales' in this significance based model.

\*Lets build another model by excluding the variable 'Fuel\_Price'.

**Model 3 – Significance Approach**

**Model Name – Second\_Significance\_Linearmodel :**

\*Variable 'Unemployment' has a highest p value of 0.16231 which exceeds 0.05 threshold and are statistically insignificant.

\*Lets drop this variable too and build another model.

**Model 4 – Significance Approach**

**Model Name – Third\_Significance\_Linearmodel :**

\*Now, we only have two variables left in the Third\_Significance\_Linearmodel.

\*In case of Regression :

\*Null Hypothesis(H0) - Independent variable is not influencing dependent variable.(i.e p value>0.05)

\*Alternate Hypothesis(H1) - Independent variable is influencing dependent variable.(i.e p value<0.05)

\*As per this model summary, p value for both the Independent variables 'Temperature' and 'CPI' are well within the p value threshold of 0.05 and are statistically significant.

\*In other words, Independent variables - 'Temperature' and 'CPI' are actually influencing the dependent variable 'Weekly\_Sales'.

\*Since p value is less than 0.05, we can reject the Null Hypothesis and hence Alternate Hypothesis(H1) holds true.

**Model 5 – Step Approach**

**Model Name – Step\_final\_Linearmodel :**

\*The model started iterating and gave me an optimum model (i.e model with a least possible AIC score=3407.53) with variables - Temperature,CPI and Unemployment.

\*Only the following independent variables - Temperature,CPI,Unemployment is contributing to the dependent variable Weekly\_Sales.

**CONCLUDING THE BEST MODEL :**

\*Lets now compare the linear models built using VIF approach,Significance approach and Step Approach.

\*As per the documented results that I infer from the respective model summary,

\* "Third\_Significance\_Linearmodel" is the best model I can suggest to the business.

\***Reason :** Difference between R2 and adjusted R2 is less compared to other Linear models built.

\*What this tells me is that this model has very little insignificance between R2 and adjusted R2 compared to other models.

\*Though this model has a very low R2 of 11.39, based on the available data,this is the best model among the other models that we built.

**ERROR RELATED ASSUMPTIONS ON “Third\_Significance\_Linearmodel”:**

**Assumption 1 : Homoscedasticity and Normal Distribution of Errors :**

\*In the Residuals vs Fitter Values graph, Data is being distributed across the graph and we see a red line in the middle with no trend ( the line is flat).

\*All the Error terms are randomly distributed over the places. What this tells us is that we have homoscedasticity of Error terms.

\*In the QQ plot, I could see the normally distributed data with few data points away from the line, which tells us there are outliers in the data.

\*I am not checking the other 2 plots as we have already confirmed normal distribution and homoscedasticity of error terms based on Residuals vs Fitted graph and QQplot.

**Assumption 2 : Autocorrelation of Error Terms :**

\*Null Hypothesis (H0) is that autocorrelation does not exist.(if p value is >0.05 - H0 is true)

\*Alternate Hypothesis (H1) is that autocorrelation does exist.(if p value is <0.05 then H1 is true)

\*Based on Durbin-Watson test results, we got a p value which is less than 0.05.

\*Hence, Null Hypothesis(H0) gets rejected and alternate Hypothesis(H1) holds true. i.e there is Autocorrelation amongst the Error terms.

**MAPE(Mean Absolute Percentage Error) for “Third\_Significance\_Linearmodel” :**

\*MAPE is an accuracy measure that tells me the relative percentage of errors in my Model.

\*We have MAPE value of almost 6.16%

\*Recommended percentage of MAPE is <=5%. In this case, MAPE value is a little higher than 5%.

**CONCLUSION** :

\*Though this “Third\_Significance\_Linearmodel” is not a very good model, I conclude that this is somewhat a decent model with better MAPE compared to other Linear models we built for the available dataset.