

**Sales Forecasting**

A WHITE PAPER REPORT

*Submitted By*

**PRASHANT MISHRA**

**KAPIL GARG**

**SARVESH BANGAD**

**NATASHA ARORA**





**UNIVERSITY OF CONNECTICUT, SCHOOL OF BUSINESS**

**HARTFORD, CONNECTICUT**

**TABLE OF CONTENTS**

[**INTRODUCTION 3**](#_30j0zll)

**PROBLEM STATEMENT****3**

**BUSINESS CASES****4**

[**DATASET**](#_2et92p0) **6**

**EXPLORATORY DATA ANALYSIS 8**

**[MODEL DEVELOPMENT DETAILS 1](#_tyjcwt)5**

[**BEST MODEL 23**](#_3dy6vkm)

[**RECOMMENDATIONS 2**](#_1t3h5sf)**3**

[**REFERENCES 2**](#_4d34og8)**4**

**INTRODUCTION**

**Walmart Inc.** is an American multinational retail corporation that operates a chain of

hypermarkets, discount department stores, and grocery stores. It is the world’s largest company

by revenue as well as the largest private employer and grocery retailer.

The data which part of our analysis was the Walmart data set obtained from Kaggle. The data contained weekly sales of various departments within different stores over different period of time. Most of the work put into the project evolves around **cleaning and exploring the data** and modeling around different parameters and methodologies.

Using different methodologies, different models were generated and their errors were noted. Time series analysis was done on sales trends and to predict the sales for the end of the year holiday season of 2012, time series forecasting was used.

**PROBLEM STATEMENT**

We are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and we need to project the sales for each department in each store. One of the challenge involves in modeling retail data is the need to make decisions based on limited history. In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge, is modeling the effects of markdowns on these holiday weeks and to predict which departments are affected and the extent of the impact.

**BUSINESS CASES**

In the course of this sales forecasting project, we have focused our strategy based on the below two business cases.

1. Retail companies are currently facing a major challenge of forecasting their sales of products across their stores and departments. This is subjected to numerous factors which include:

* the impact of problem on an organizational level
* the impact of seasonal changes on their sales
* the differences in sales at different stores
* various promotional discounts and offers
* figuring out how and when to introduce new products to be sold

Companies are now-a-days shifting to strategy of forecasting of sales as per the demand keeping in mind the complexity to deal with huge amount of transactional and organizational data. This data is collected from various internal and external sources which paves a way to leverage it to predict future sales along with the buying behavior of different customers. This further results in a better revenue and sales forecast which allows the marketing team to plan their strategy appropriately in an efficient and customer focused manner. We as team took this challenge, aiming to forecast accurate weekly sales for Walmart as its competitive pricing structure and correctly projecting sales is key in its ability to function.

1. Walmart is the leading retail company which runs its business on multiple chains. The implications introduced by holiday season and profit eating discounts is one of the challenge which they are focusing on right now. We aim to address this issue by forecasting weekly sales of 45 stores and their multiple departments across all the states in USA. The impact of markdown prices introduced during a few high revenue weeks of an year is also being considered in our analysis. We will be dealing with this problem utilizing the weekly sales data from 2010 to 2012. Understanding of all the factors which affects the sales at store and department level using the time series forecasting will facilitate the company to plan their hiring process beforehand.

**DATA SET**

The Walmart Store Sales data is published as Walmart recruiting competition on Kaggle. It includes sales data for 45 stores in different regions of United States from 05-02-2010 to 01-11-2012, where each store contains a number of departments.

The different files contained in the data set are –

**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, with the following labels –

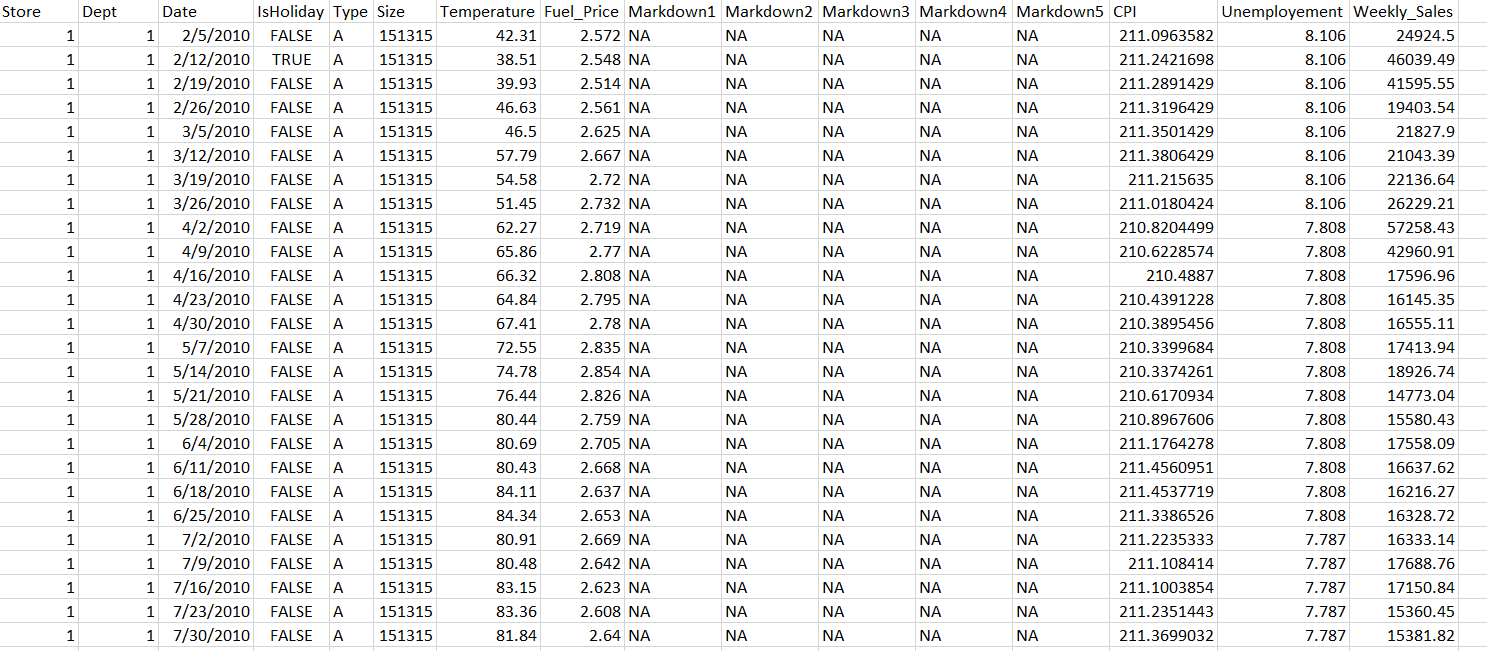
* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales -  sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

**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

**Master dataset-** We merged features, stores, and train datasets to come up with a final master dataset with a total of 16 variables. A glance of master dataset is as shown below.



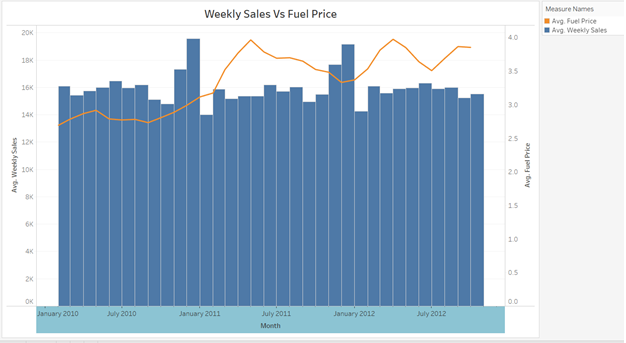
**EXPLORATORY DATA ANALYSIS**

For exploratory data analysis part, we made use of Tableau software. We imported all the datasets and merged them using the join table functionality of Tableau to come up with a master dataset. We tried to figure out the correlation between weekly sales and other factors in the master dataset.

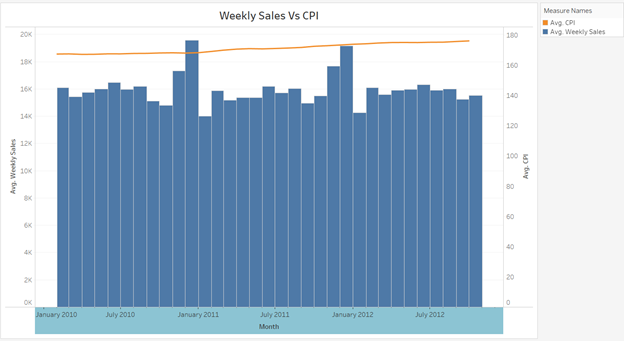
We divided our data analysis part into below three categories:

1. **Finding correlations** between weekly sales and other variables. We implemented two ARIMA models, with and without Fuel Price & CPI. The model with all regressors was better than the other one in terms of RMSE values.

* **Fuel Price vs Weekly Sales**



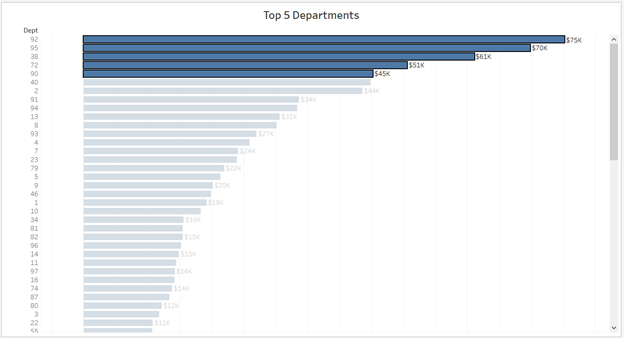
* **CPI vs Weekly Sales**



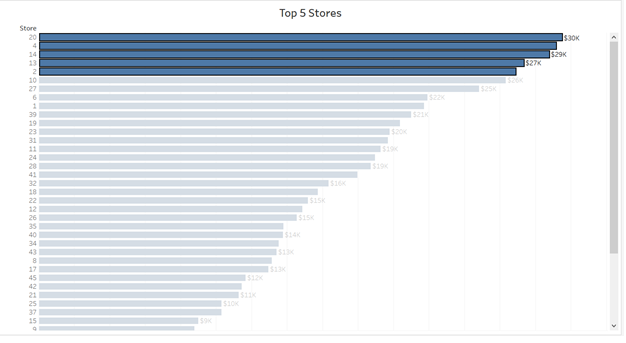
1. **Finding top 5 stores and departments** on the basis of weekly sales

Since our dataset contains a total of 45 stores each having over 90 departments, it was quite difficult to make our models and show their output for each and every combination of store and department. Therefore, we came up with an idea of implementing the forecasting models for presenting weekly sales insights for combinations of top 5 stores and top 5 departments. We used Tableau to determine the top 5 stores and departments as shown below.

* **Top 5 Departments**

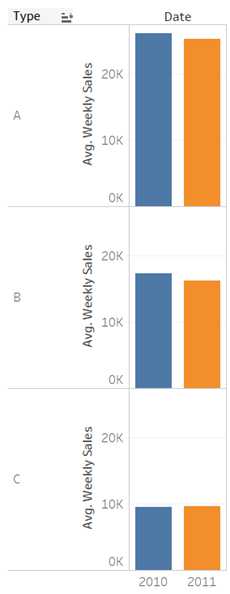
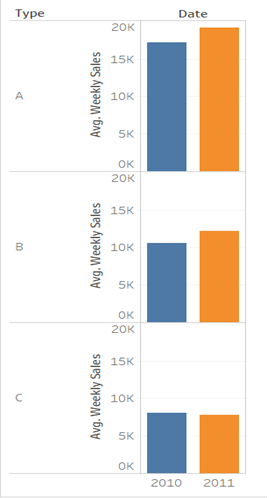


* **Top 5 Stores**



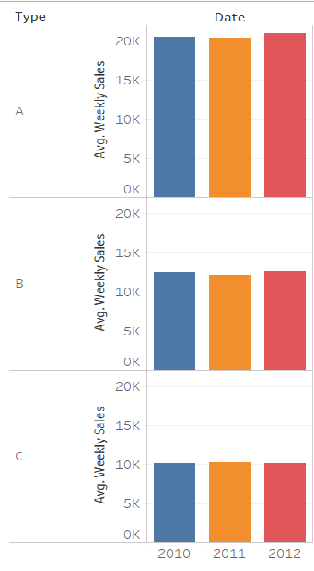
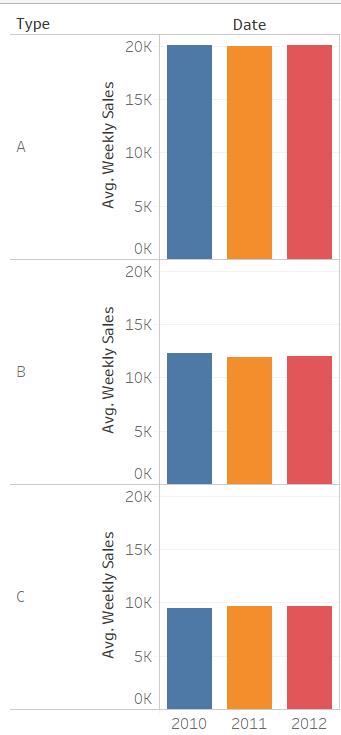
1. **Finding variation in weekly sales for all the years on special occasions** such as Labor day, Christmas, Thanksgiving, and Super Bowl. We observed that for there was a huge difference in weekly sales between that specific week and other three weeks of the same month. This signifies the peak sales during these special weeks throughout an year.

* **Christmas**



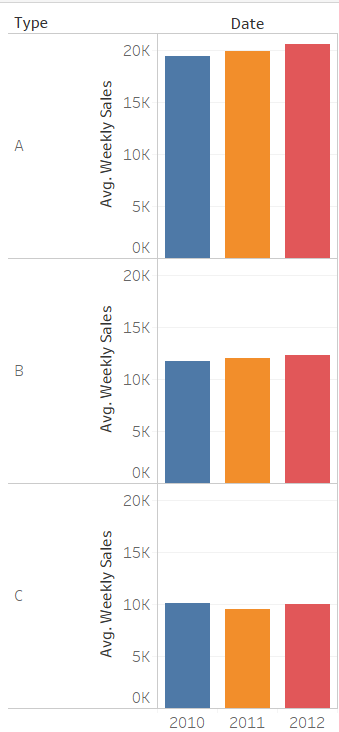
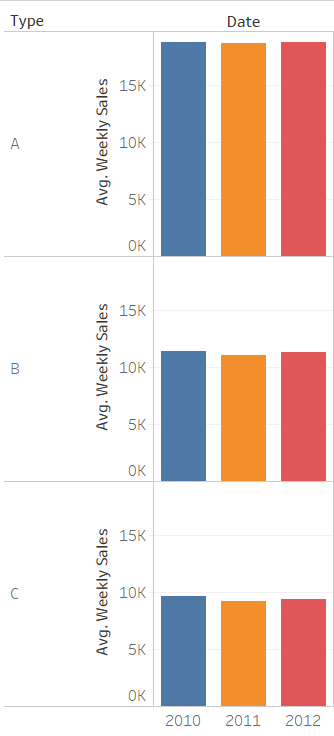
**Christmas Week Weeks other than Christmas**

* **Super Bowl**

** **

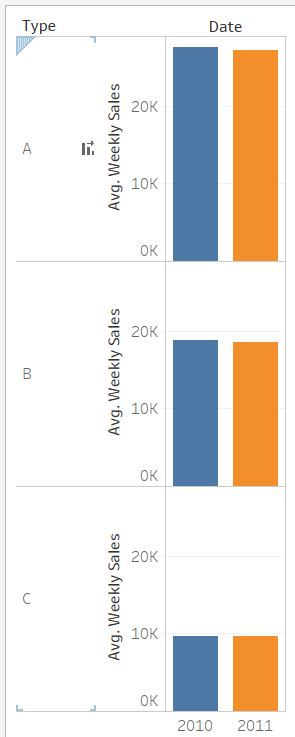
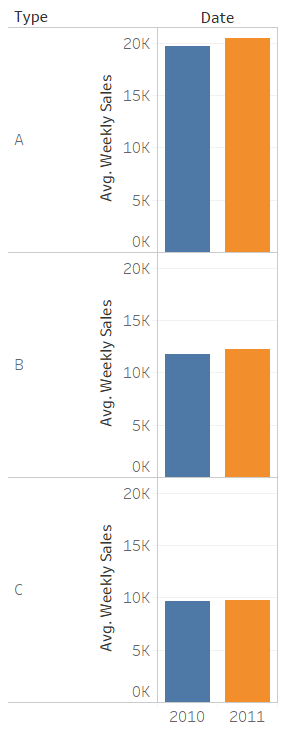
**Super Bowl Week** **Weeks other than Super Bowl**

* **Labor Day**

** **

**Labor Day Week Weeks other than Labor Day**

* **Thanksgiving**

** **

**Thanksgiving Week Weeks other than Thanksgiving**

**[MODEL DEVELOPMENT DETAILS](#_tyjcwt)**

Following models are used in order to predict the weekly sales for each department: -

**Auto Arima**

ARIMA stands for auto-regressive integrated moving average and is specified by these three order parameters: (p, d, q). The process of fitting an ARIMA model is sometimes referred to as the Box-Jenkins method.

The AR component is decided by looking at the lags in PACF and IACF plots and the MA component is decided based on the lags in ACF plots. The d is used to stabilize the timeseries when stationary assumptions are not met.

In our project since we have to predict weekly sales for each department (99) of all the stores (45) there will be 45\*99 = 4455 different time series. It is not practically possible to look at each ACF PACF IACF plot to decide the p, d, and q value for each time series. So, in order to find the p , d, and q automatically for all the departments it is the best way to use the auto arima function.

There are several external regressors like markdowns and CPI, Fuel Price, size and unemployment are used in the time series to predict the weekly sales better. We have done exploratory data analysis to determine what are the significant variable that should be used in the model. We have built time series with different regressors and check the RMSE values to determine which model is performing good.

**ETS (Exponential Smoothing State Space Model)**

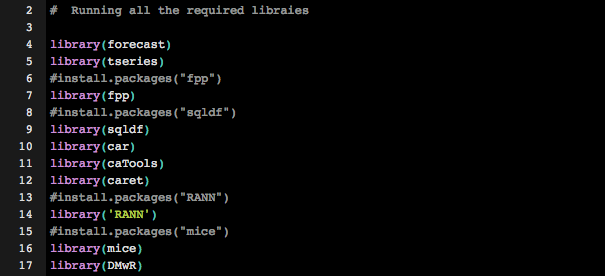
Exponential smoothing model is used to give higher weightage to the recent values.

Smoothed Value(at time t) = Actual value at time t \* alpha + Smoothed Value at the previous time t-1 \* (1 - alpha)

We have built this model and compared its results with the auto arima model. We have used RMSE as the model selection criteria.

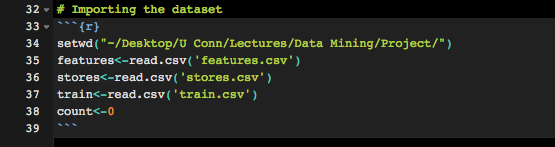
**Explanation of code: -**

**Step 1: -**

****

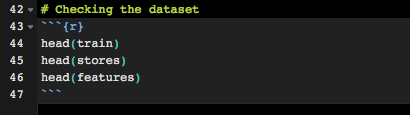
This part of code calls all the required library which are used below.

**Step 2: -**

****

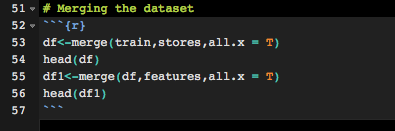
This part of code imports all the dataset from local machine. We have features, stores and train dataset as described in the above section.

**Step 3: -**

****

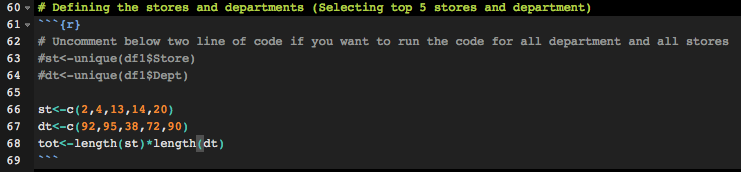
This part of the code checks the data

**Step 4: -**

****

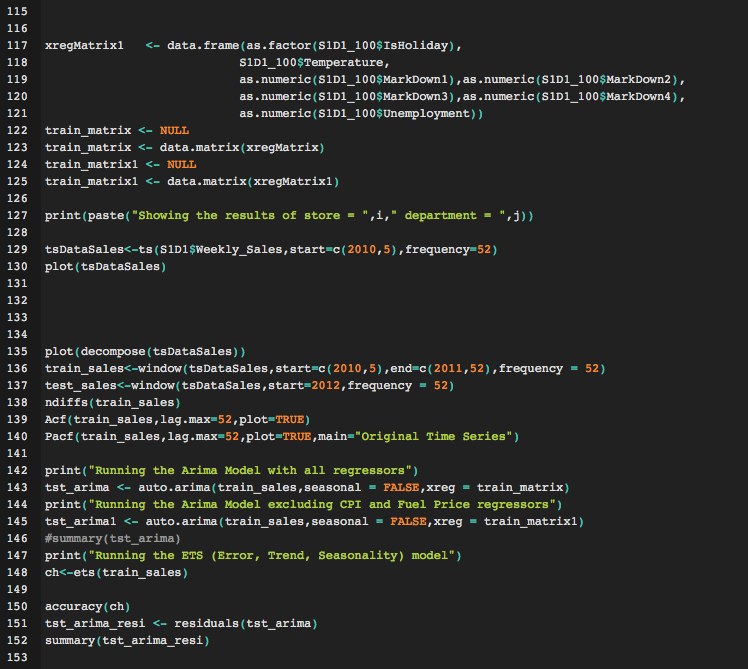
Then we have merged all the three dataset to make one dataset for our modelling purpose. This is the final Analytical Dataset.

**Step 5: -**

****

This part of code selects the store and department for which we want to run the code. Here for the demo purpose we have selected the top 5 stores and top 5 department. Although we can uncomment the code on line 63 and 64 for running it for all the stores and department.

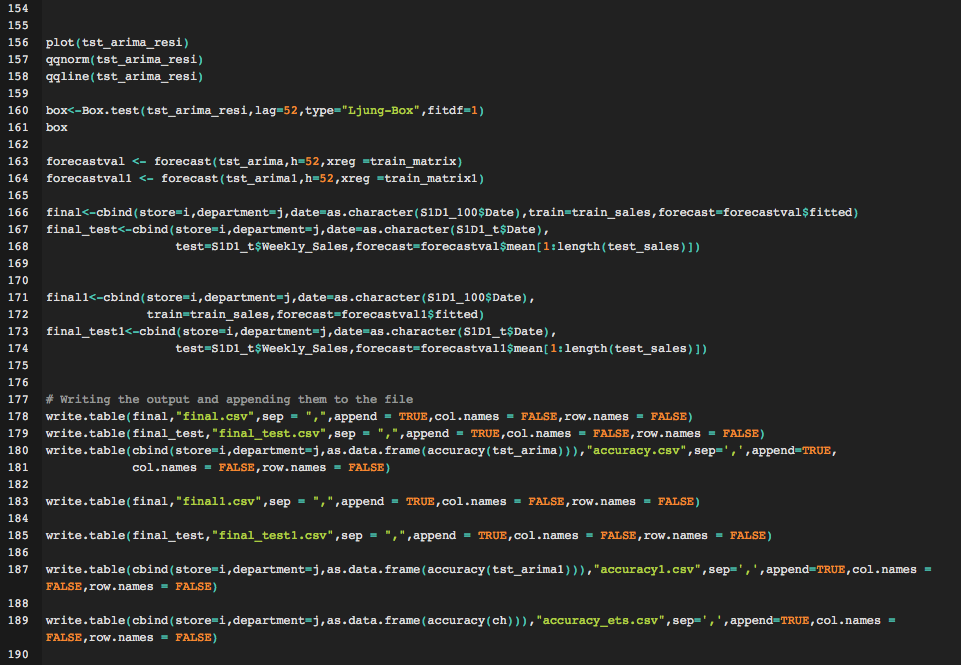
**Step 6: -**

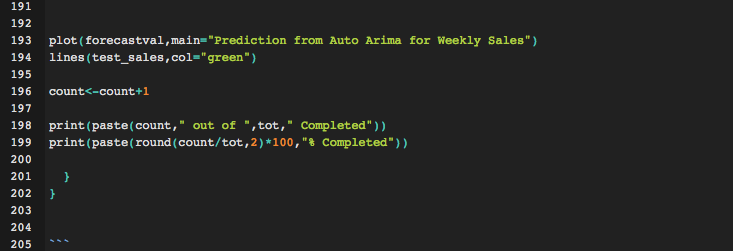
****

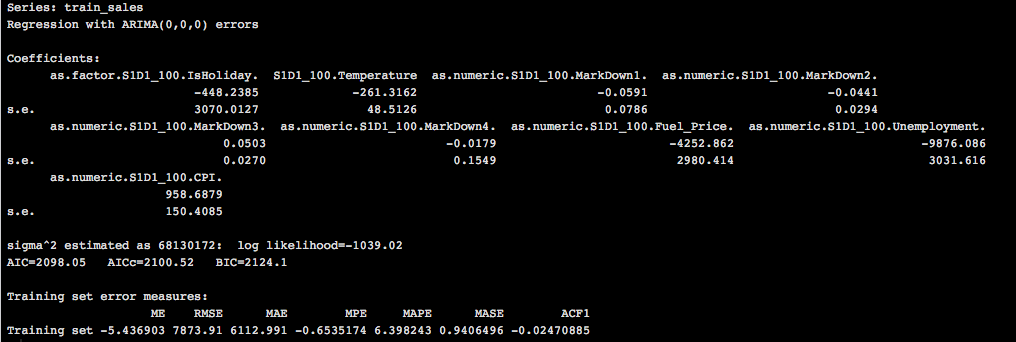
We have created a for loop to run the code for each store and all the departments in the respective store. We have split the training set in to 70:30 to build and out of sample test dataset to verify the model results.

This part of the code changes all the external regressors to the required format. We have use mice library for imputation. This runs random forest (as we have selected random forest as the argument) for the missing values and runs the model to impute the values.

Next, we have built the auto arima and ets model with different combination of regressors.

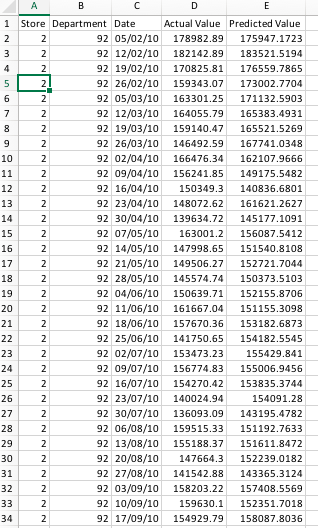
****

****

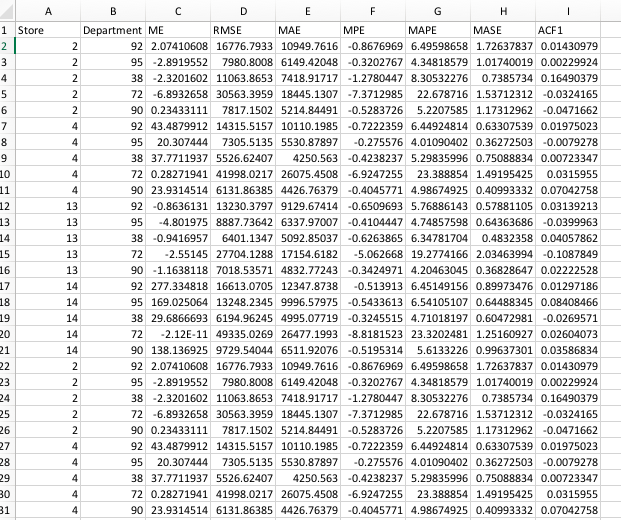
****

This part of the code finally writes all the output and keep on appending them in an excel for all the loops. We have saved the forecasted values and actual values of the test set and the model validation parameters like AIC BIC ME RMSE MAE MPE MAPE MASE values in the csv file.

**Sample Outputs: -**



This is the sample output of the model where we get the actual vs predicted values of the test dataset for all the store and department combinations.



This is another output of the model which saves all the model estimation parameters. This also saves all the model parameters for all the store department combinations.

**BEST MODEL**

We have built two models ARIMA and ETS for each department across each store. For each department, we are comparing RMSE value and selecting the model with lowest RMSE. Using the best model for a particular department, we are predicting weekly sales for that store. For our analysis, we have selected top five departments and top 5 stores. We have built 25 models (5 x 5) of ARIMA and ETS. For most of the cases, ARIMA model is the best model based on RMSE value

**RECOMMENDATIONS**

We did our analysis to come up with the following recommendations

* Hire more people during peak weekly sales
* Pre-plan for inventory management across and within the stores for all departments
* Plan financial budget ahead of the year with the help of forecasted values for the next fiscal year
* Implement the same forecasting models for product level across all departments
* Market basket analysis can be done to promote sales of low selling products by keeping them along with high selling products
* Discounts and promotional offers can be made for promoting weekly sales for departments and stores which have significantly low sales as compared to high value stores.
* Behavior associated with various factors contributing the high sales across the stores and departments can be captured and implemented for the ones with low sales.

**REFERENCES**

1. <https://www.rdocumentation.org/packages/forecast/versions/8.1/topics/auto.arima>

2. <https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials>

3. <https://www.r-bloggers.com/error-trend-seasonality-ets-and-its-forecast-model-friends/>

4. <https://www.rdocumentation.org/packages/forecast/versions/8.1/topics/ets>

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