**Hackathon Report**

**- K.A.N. PRANAVHARSHA**

**1304 Batch – 29**

**Sales Forecasting of Retail Clothing Product Categories**

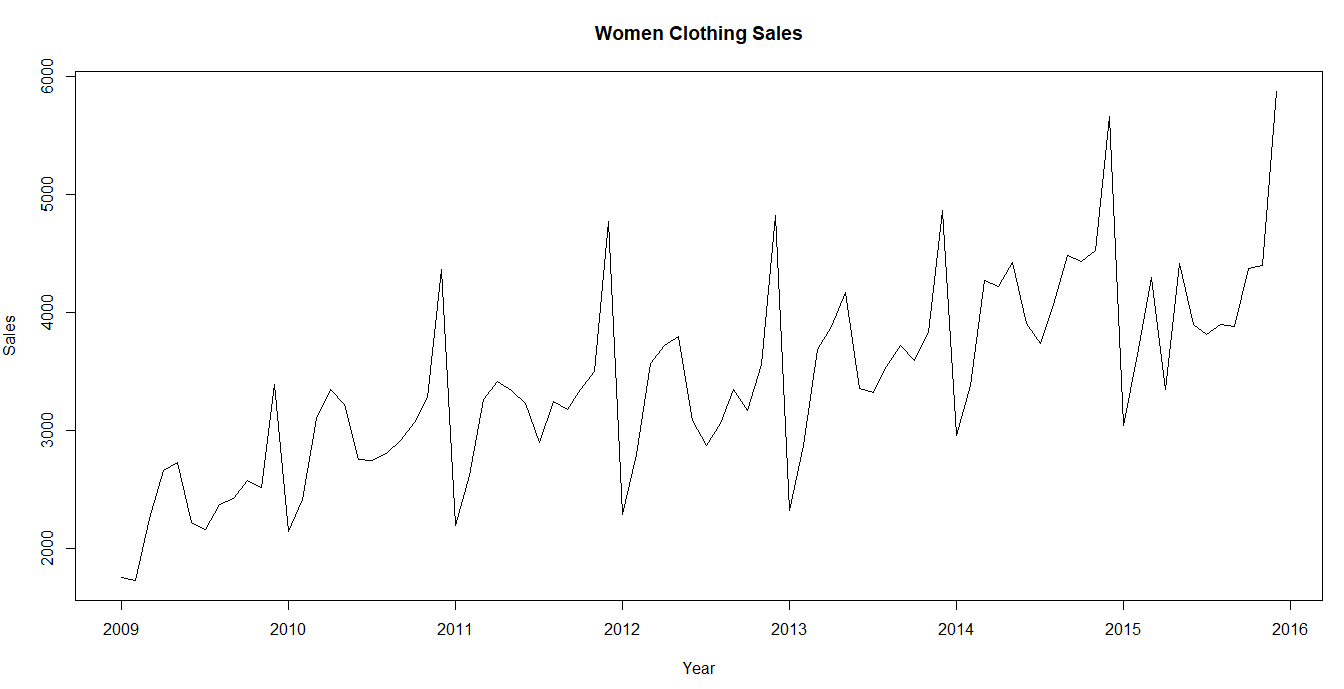
**Purpose:**

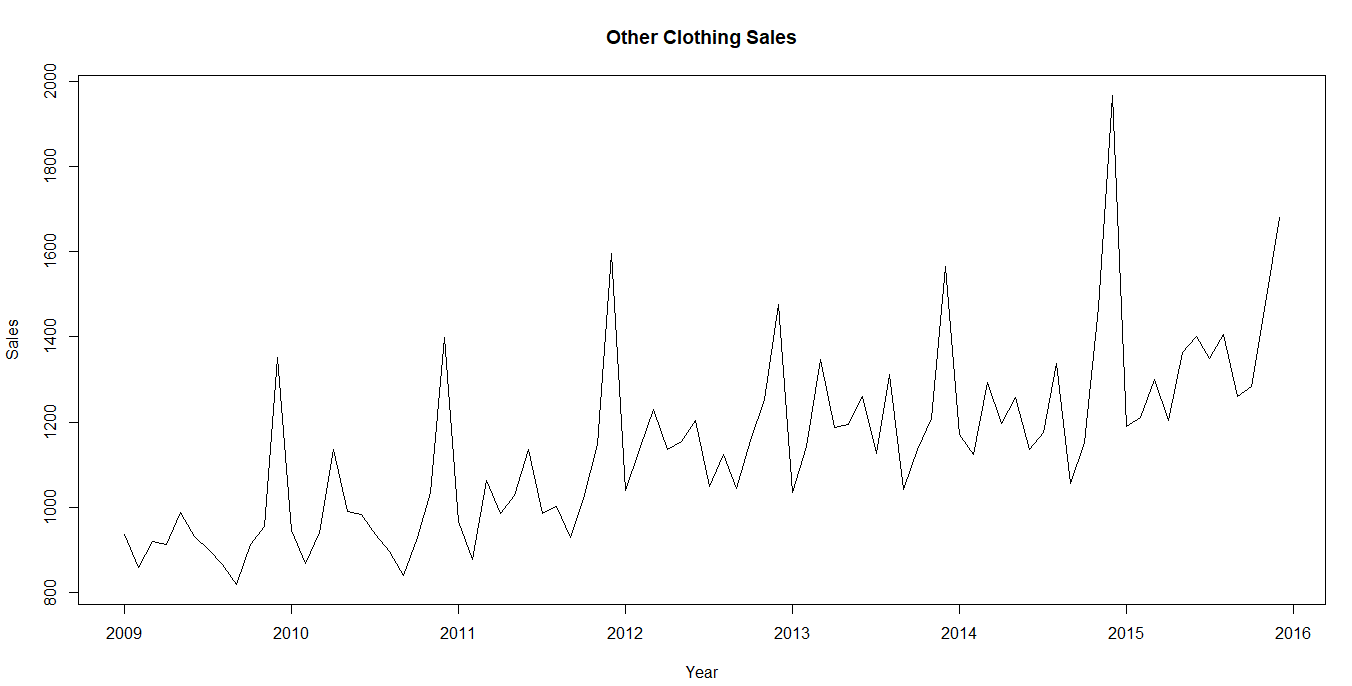
The problem statement is to forecast the sales for the next 12 months on three different clothing categories. The sales forecasting is one of the important thing for the company because they get to know how much amount of raw materials they need to maintain, which category of clothing is making more amount of sales, which category is making less amount of sales and also it gives a chance for the company to improve its performance. It also helps in taking accurate business decisions.

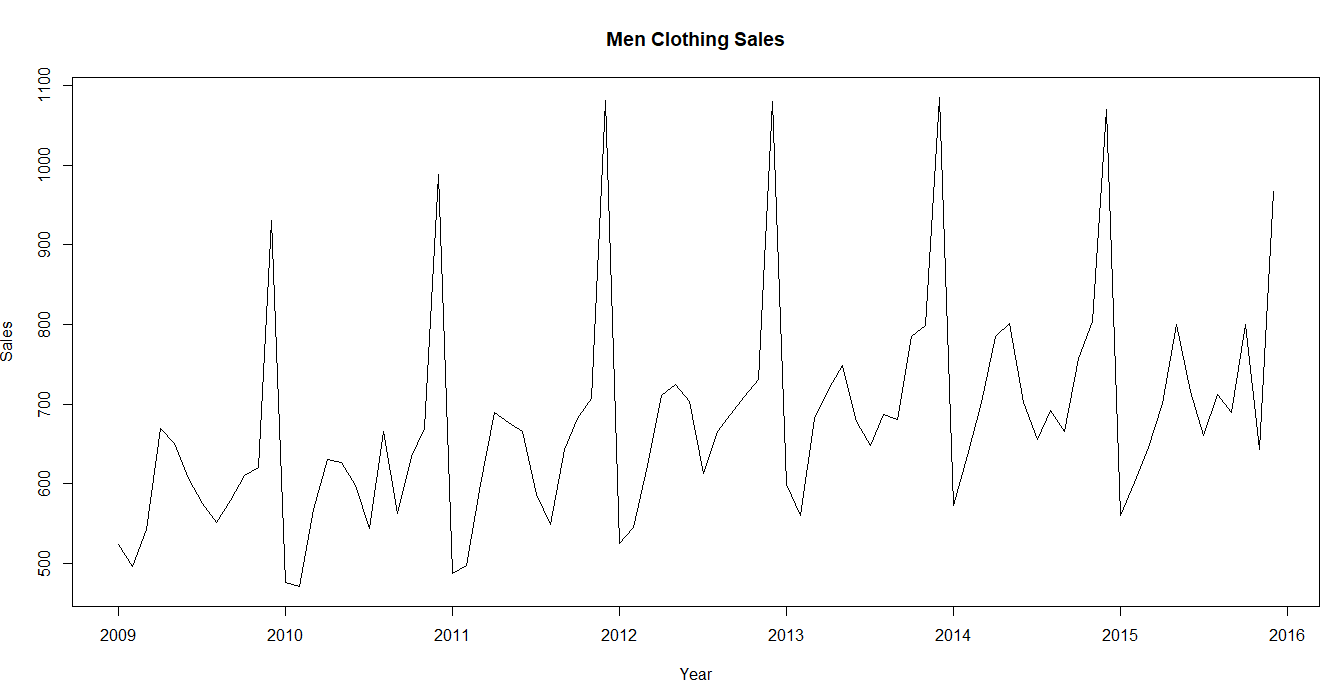
**Data:**

**Sales Data:**

* This contains the details of amount of sales (revenue in thousands of dollars) for every month, all categories (Women Clothing, Men Clothing, Other Clothing) happened from the period 2009 to 2015.
* There were missing values for sales in the data. Imputed them with mean because if we impute them with previous year, same month sales revenue then some the information is missed because we don’t know weather the sales has increased or decreased. So, the safest approach is to impute the missing values with mean.







* In all the three sales series above clearly, we can see that there is trend and seasonality.
* The sales are improving year on year in women clothing. But in Men and Other clothing categories they not improving much overall compared to women clothing.

**Weather Data:**

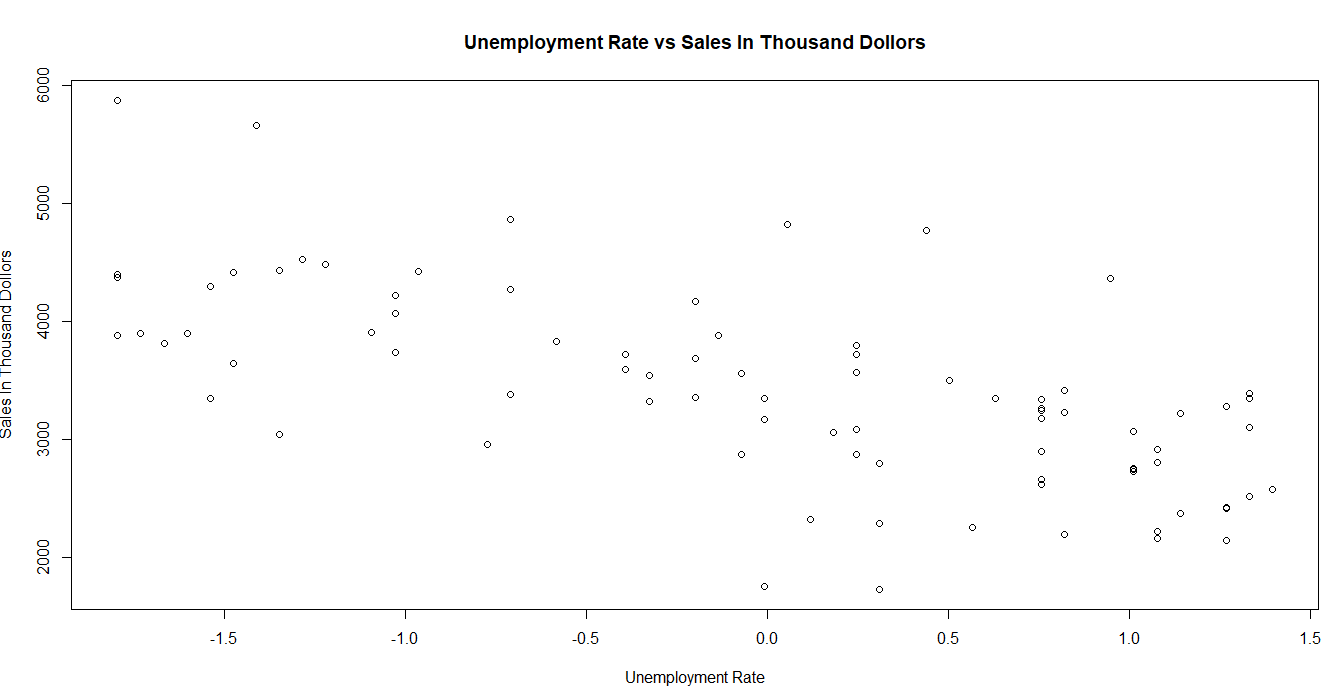
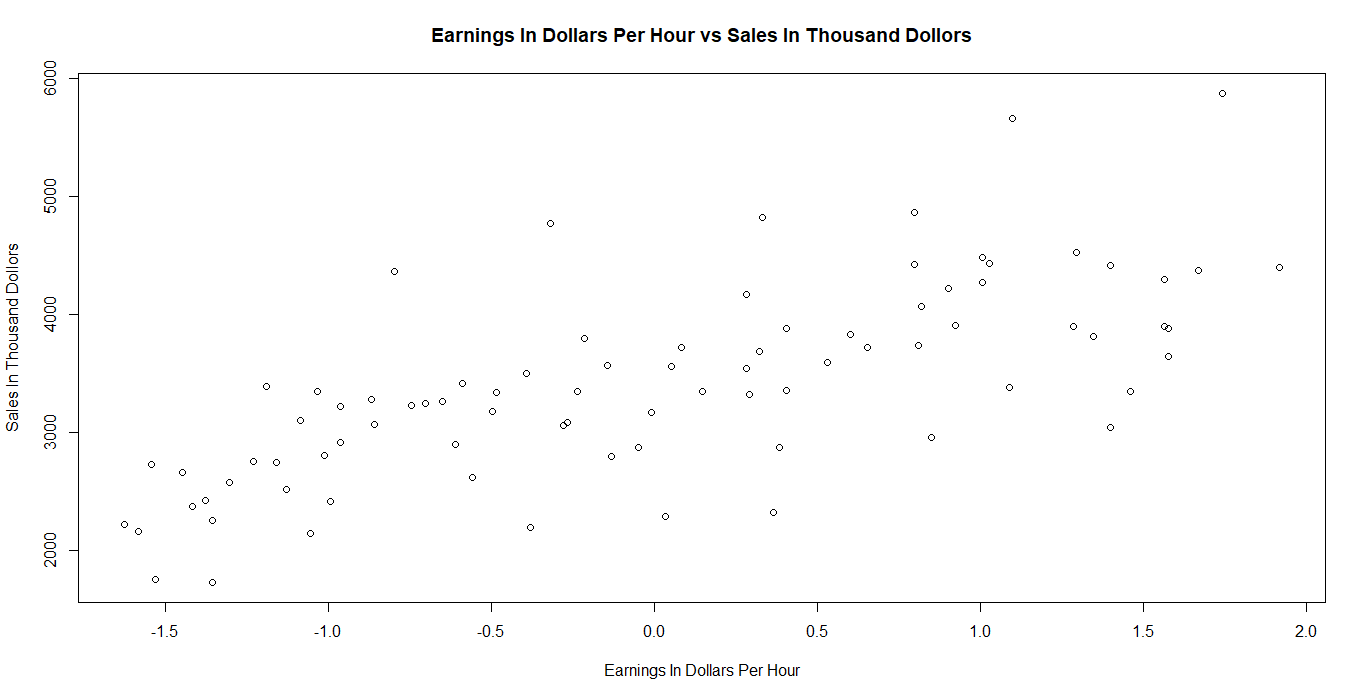
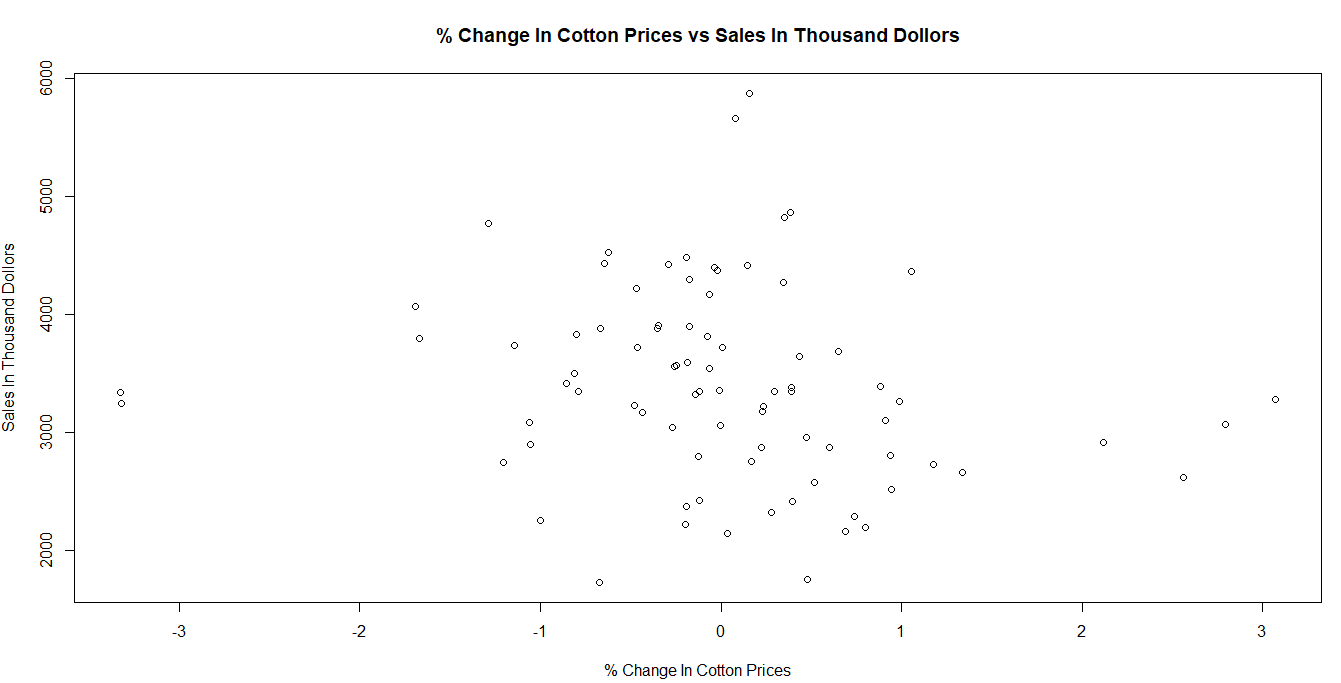
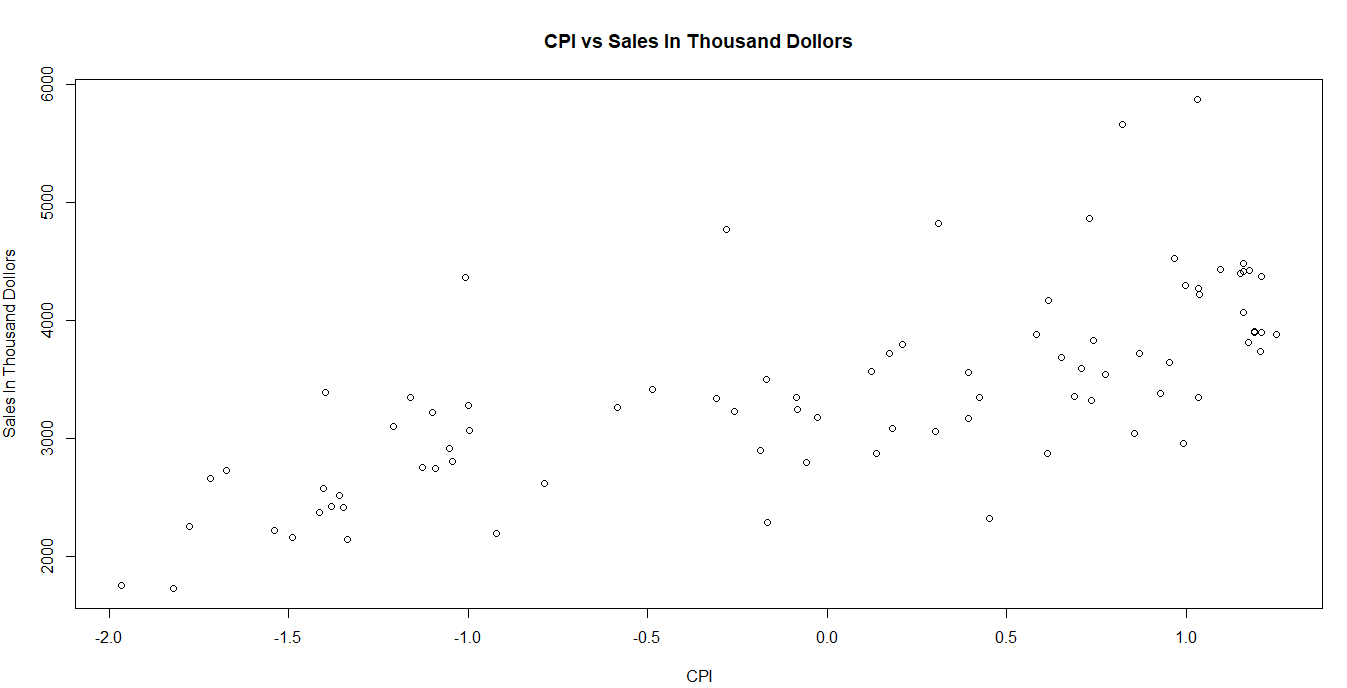
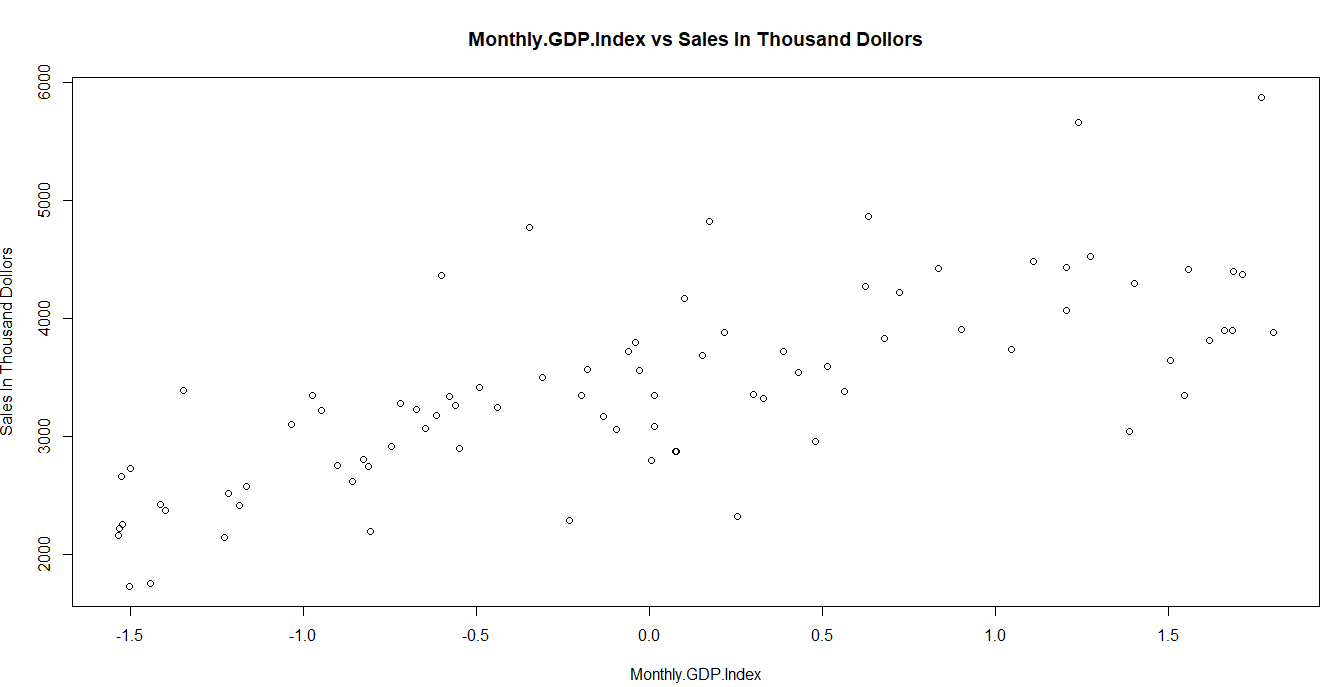
* This contains data about the weather conditions for every day from 2009 to 2016.
* The weather attributes mentioned are humidity, temperature, visibility, sea-level pressure, precipitation, dew point, wind and weather event.
* For many attributes are given in the form of low, high and average for that day. I observed on some days we have high difference between high and low, due to which average is getting impacted. So, I thought not to remove the high and low values for the attributes.
* In each and every sheet the year is mentioned as 2009.
* We have “-” or “?” for some of the values in the data.
* In precipitation “T” is mentioned for some days. Weather event has almost 60% - 70% of null values for every year.

**Events/Holidays Data:**

* This contains data about the holidays in the period of 2009 - 2016.
* The also contains whether the holiday is event holiday or federal holiday and the name of the holiday.

**Macro-Economic Data:**

* This contains data about various Economic influential factors in the period of 2009 - 2016. The data given is for every month.
* The data contains GDP, CPI, Unemployment rate, Interest rates, Party in power, wages per hour, Advertising expenses, Cotton price and details about cotton crop production.
* This economic data has important impact on the sales. I observed linear relationship between the sales(target) and many economic factors.

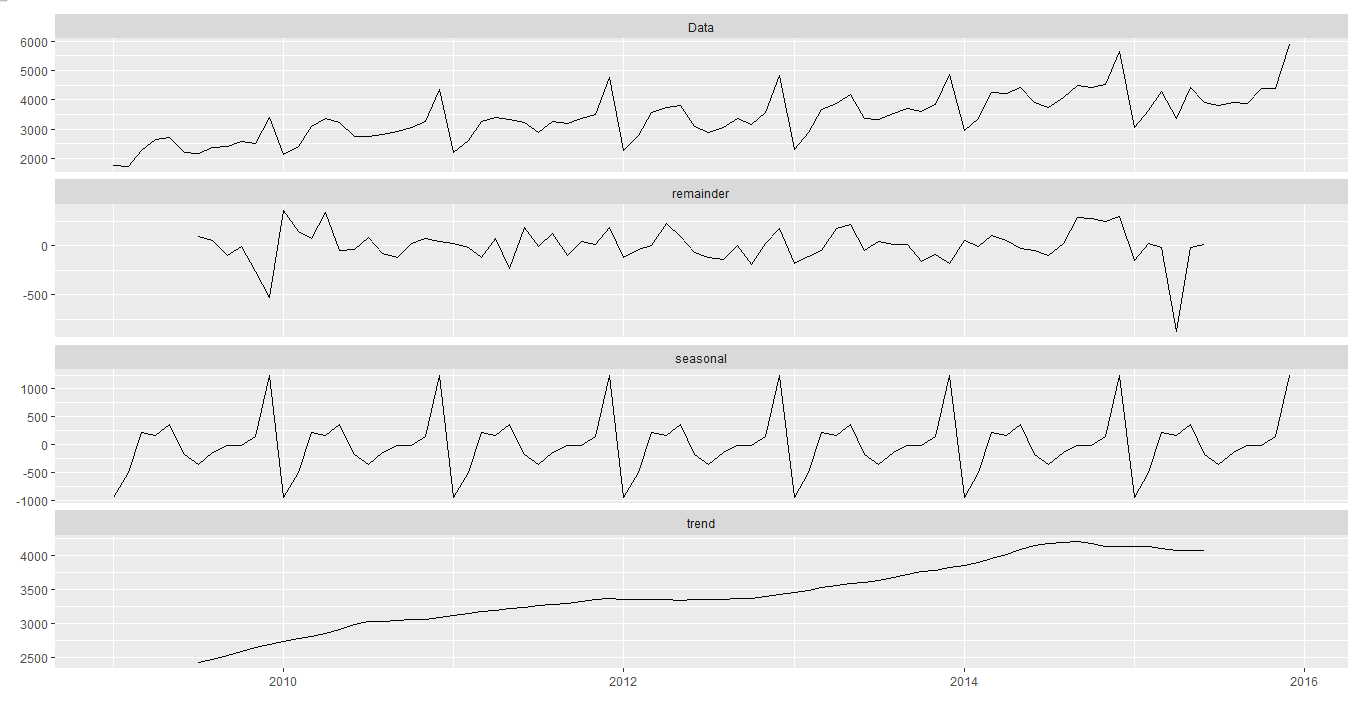
  

**Methodology or Approach:**

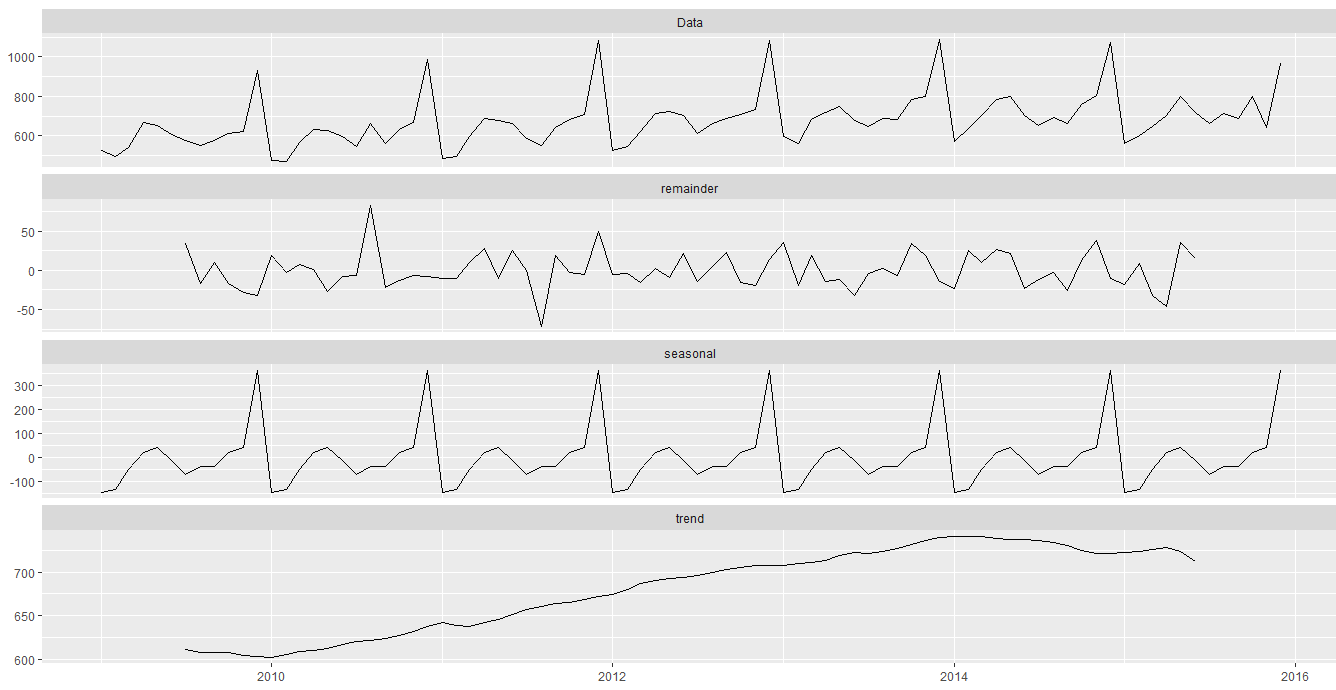
**Time Series:**

* Created Women Clothing, Men Clothing, Other Clothing sets of data frames from the master data based on the category of clothing.
* There were missing values for sales in the data. Imputed them with mean because if we impute them with previous year, same month sales revenue then some the information is missed because we don’t know whether the sales has increased or decreased. So, the safest approach is to impute the missing values with mean.
* Changed the data into time series objects.
* Decomposed the time series to look into Seasonal, Random, trend components.

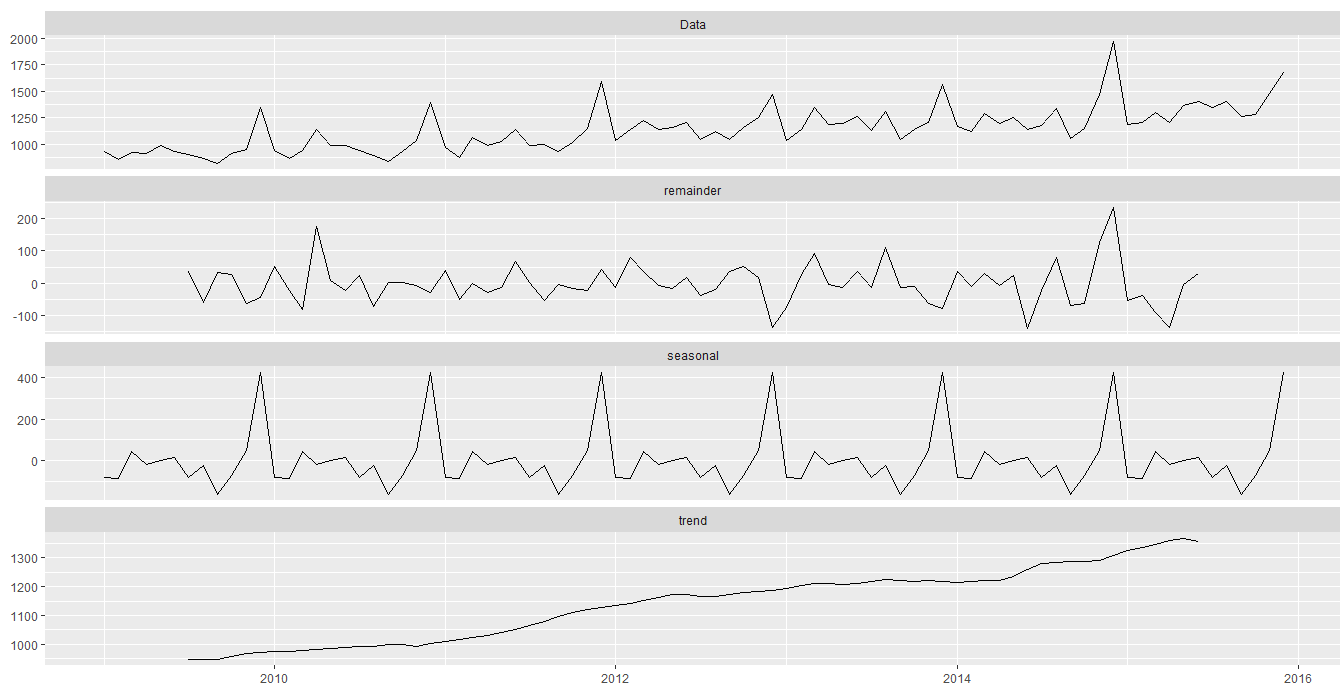
Women Clothing Decomposed Series:



Men Clothing Decomposed Series:



Other Clothing Decomposed Series:

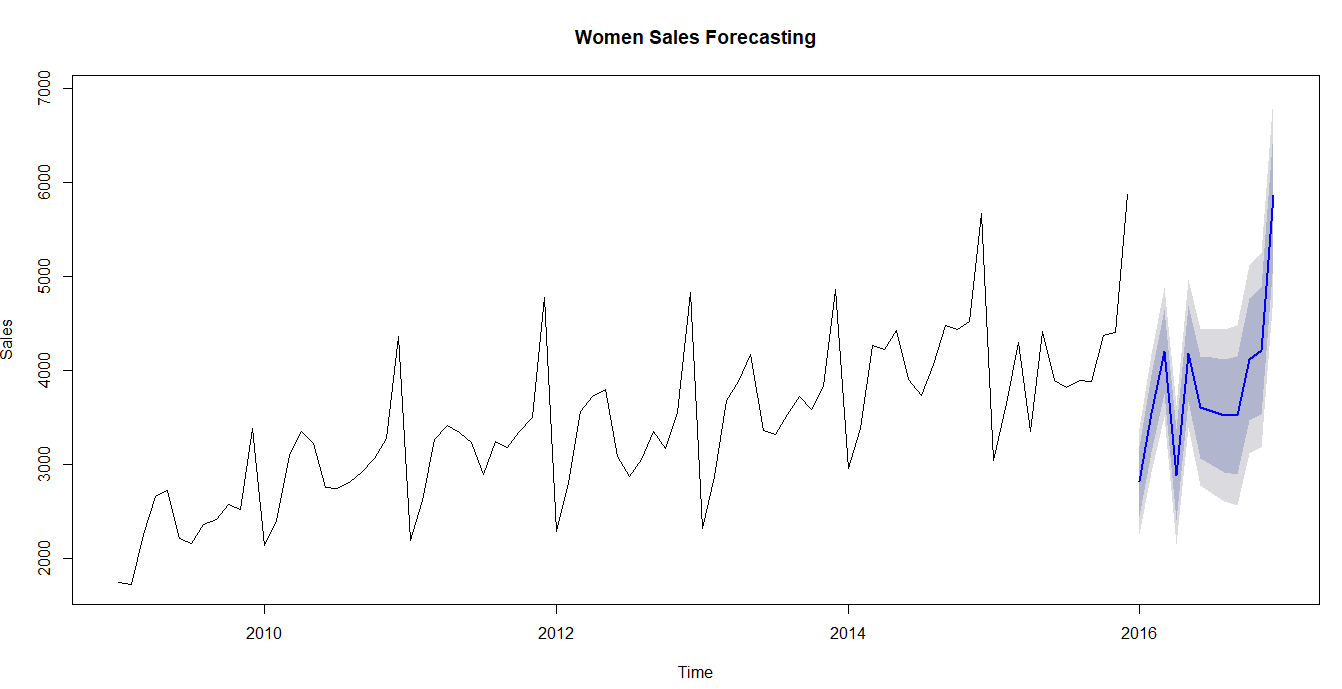


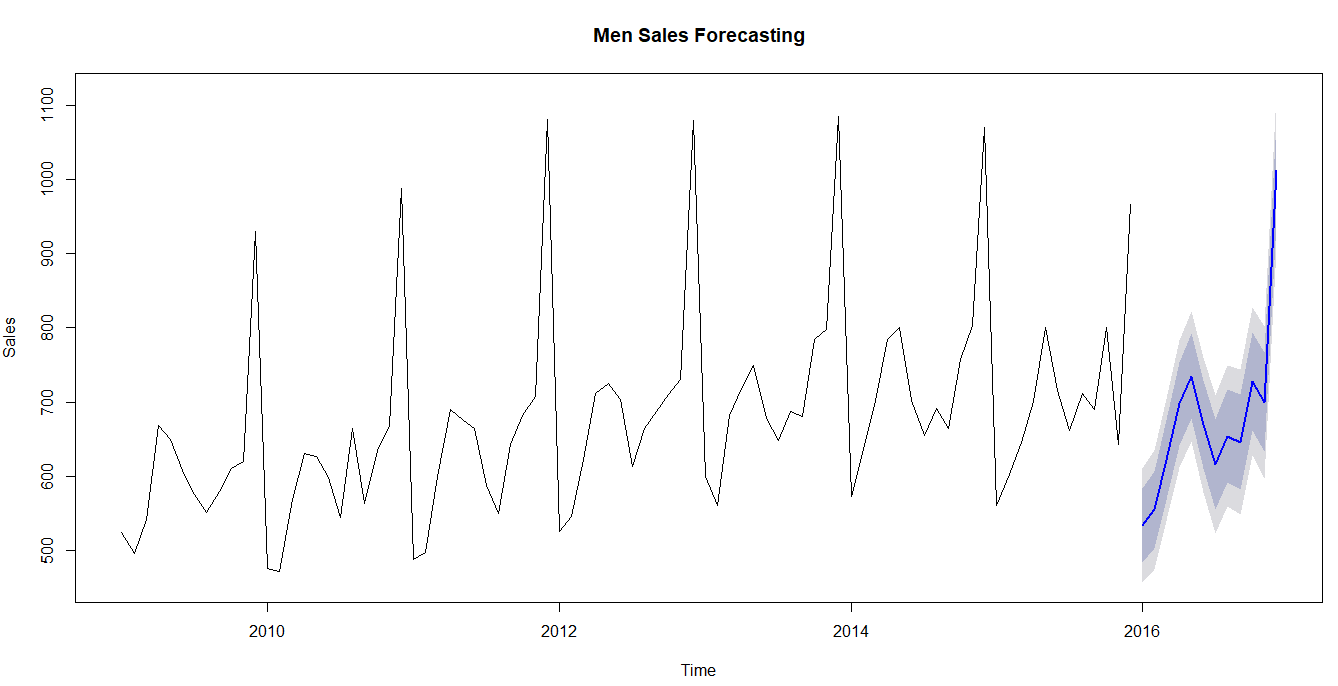
* In all the series clearly, we can see seasonality and trend.

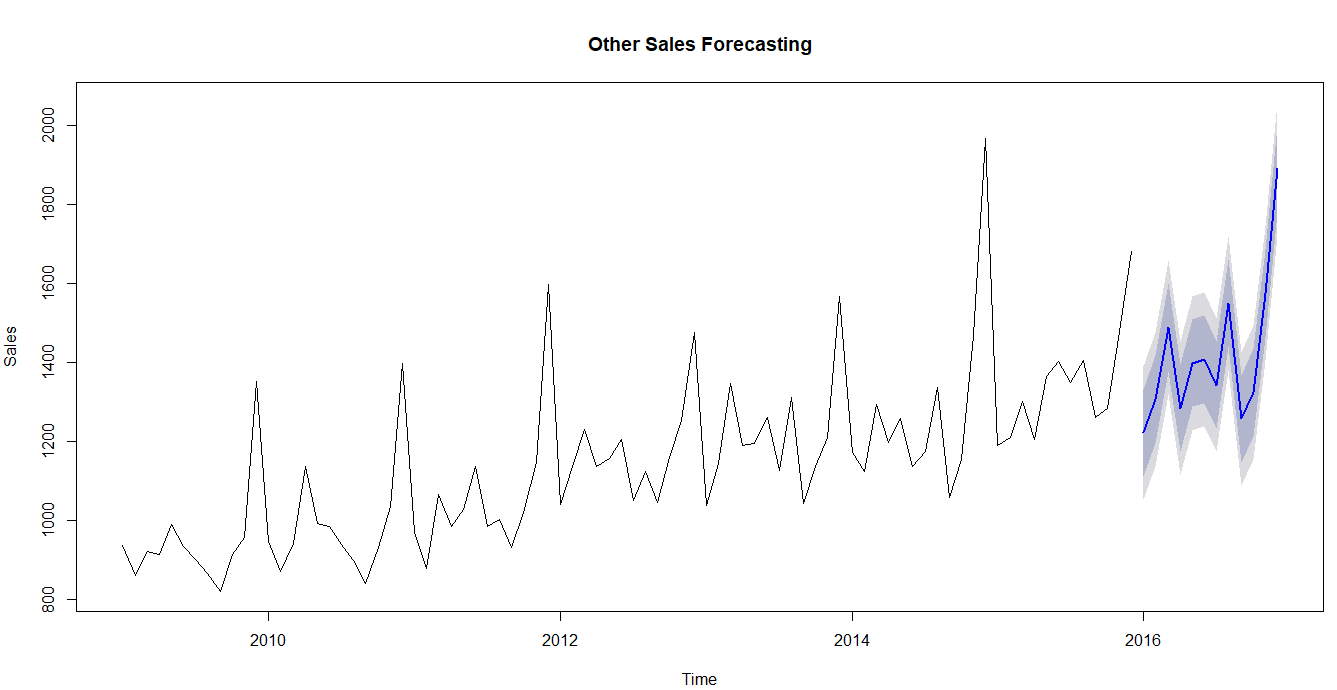
**Models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Women Clothing (Train MAPE)** | **Men Clothing (Train MAPE)** | **Other Clothing (Train MAPE)** | **Overall Test MAPE** |
| **SMA** | 8.06 | 8.08 | 6.27 | 26.06 |
| **Holt Winters** | 15.06 | 16.15 | 14.27 | 18.09 |
| **Auto ARIMA** | 4.80 | 3.53 | 4.24 | 10.329 |
| **Best Tuned ARIMA** | 4.15 | 3.51 | 3.21 | 9.39 |

* The best model is ARIMA with following p, d, q and P, D, Q values
* Women Clothing: (0, 1, 1) (2, 2, 2)
* Men Clothing: (0,1,1) (0,1,1)
* Other Clothing: (0,0,0) (1,2,2)
* Forecasted Series for the best model are below







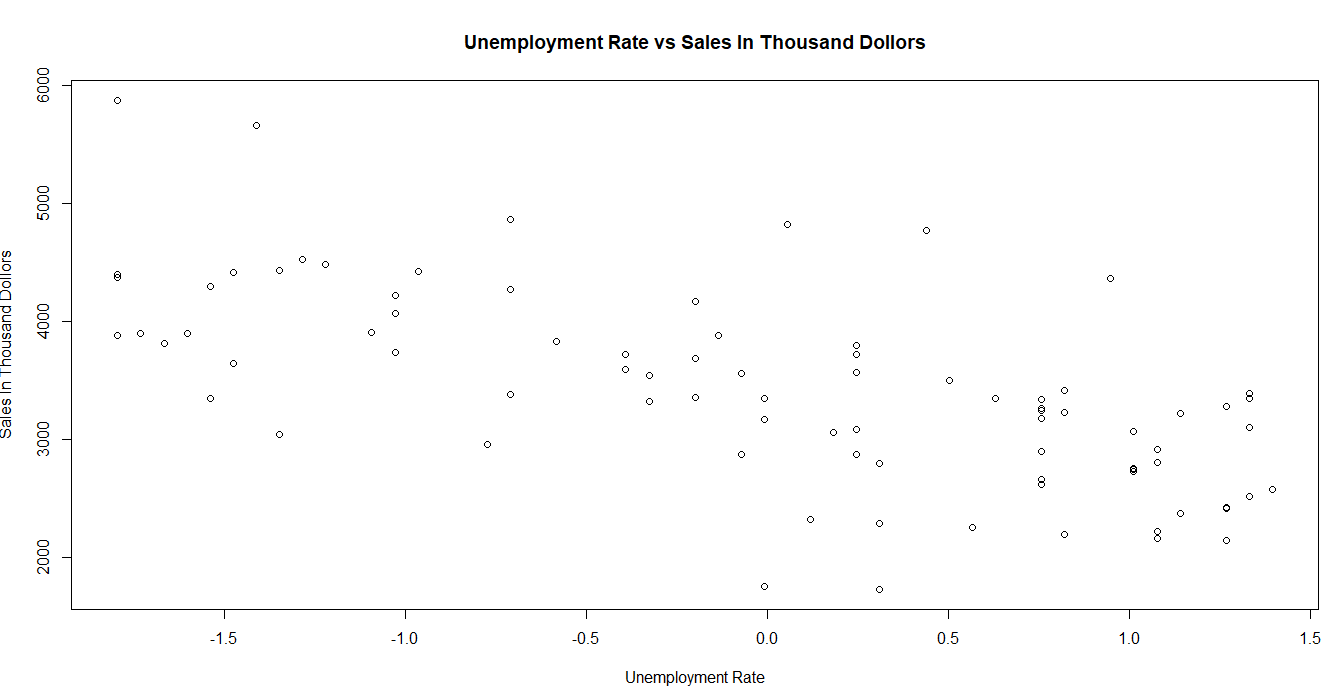
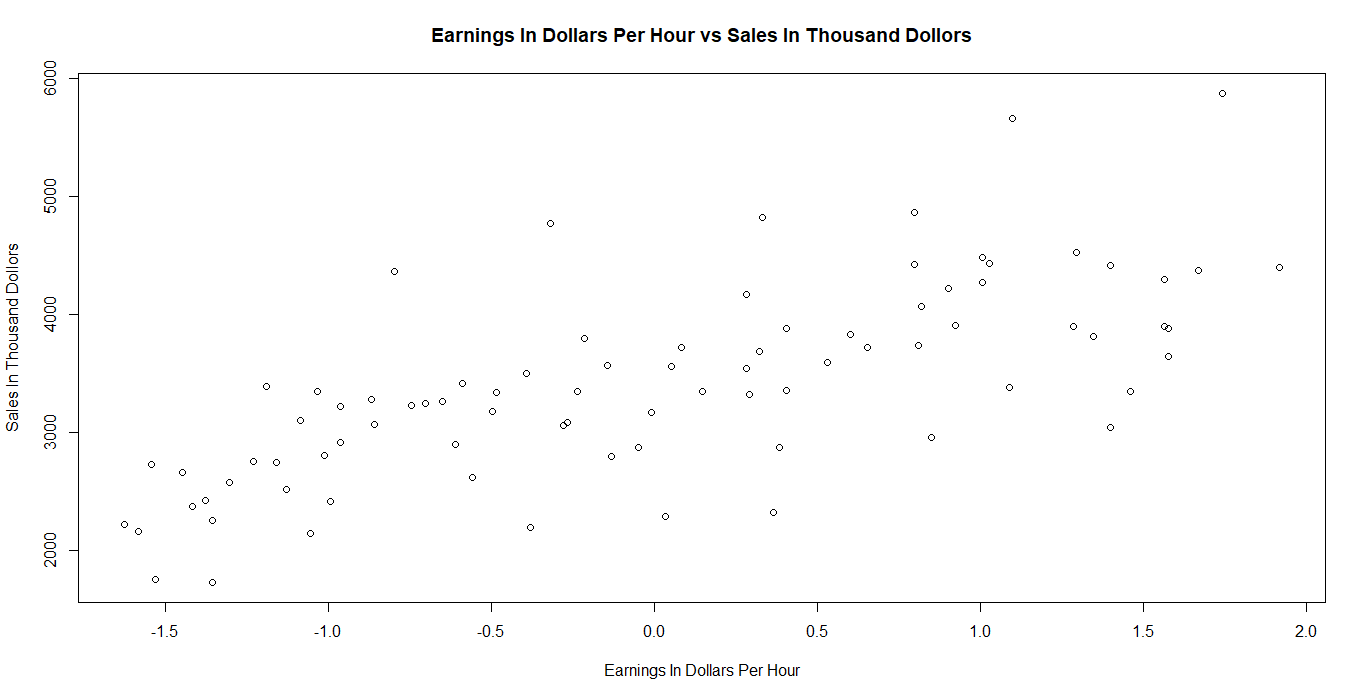
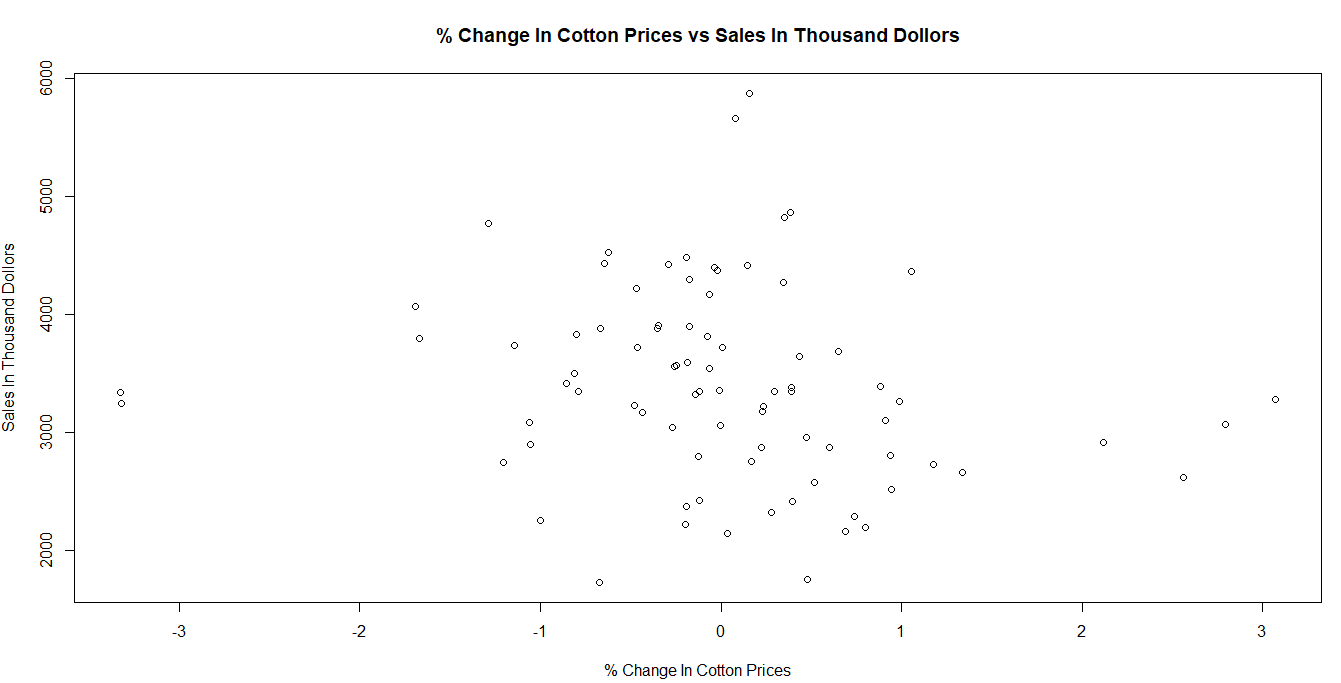
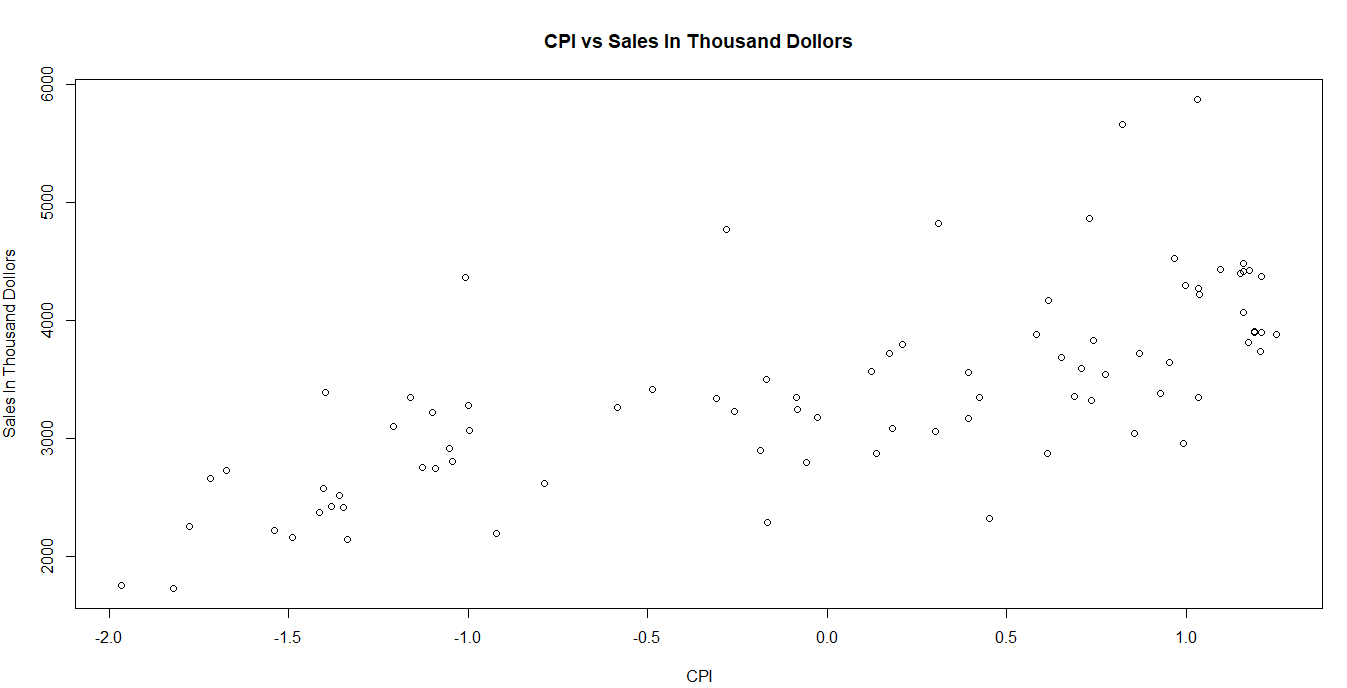
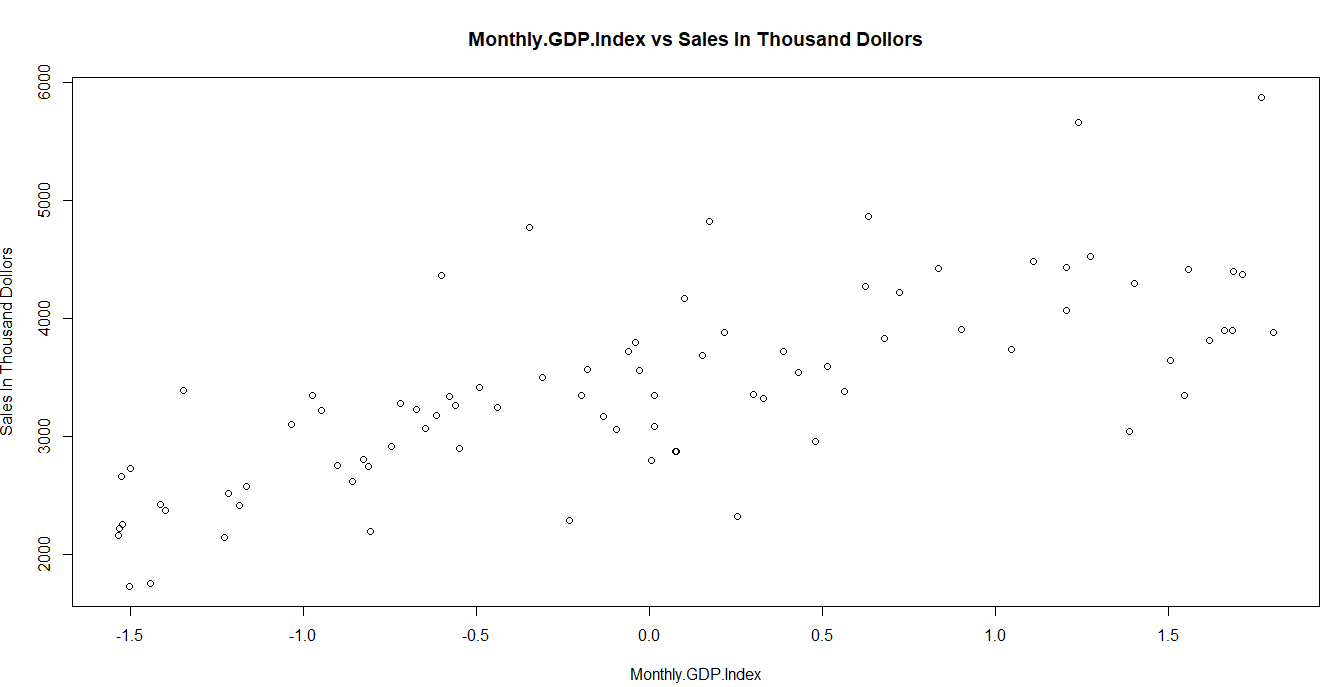
**Regression**:

**Weather Data**:

* Read all the data in excel sheets into separate data-frames because each sheet has its own type of missing values and in all the sheets the year is mentioned as 2009.
* Changed the year from 2009 to respective years.
* Changed “-” and “?” values in some of the attributes to null values.
* Changed the “T” value in Precipitation attribute to zero because T means “Trace” which says that there is very small amount of rain or snow and cannot be measured. So, made the value zero.
* Removed the attribute Weather Event because almost 70% of the values are null and another reason is the same thing is explained by the precipitation. So, removed the attribute.
* Dropped Day from the dataset as it has no significance.
* Aggregated the data by month for each and every year.
* Combined all the year attributes of all years from period 2009-2015 and created train dataset, 2016 with test dataset.

**Economic Data**:

* Separated the Year, Month from the year-month attribute.
* Already the data was aggregated by month.
* Dropped the Party in Power attribute because all the values have same. So, for now it will not have much impact on model.
* Dropped Advertising expenses as 73 values were null.

* Some of the plots between Target and independent variables which are impacting them linearly are mentioned above.

**Holidays Data**:

* Checked for null values. There were no null values.
* Aggerated the data by month and year and created new feature count which holds the number of holidays in that particular year and month irrespective of Holiday category.
* Dropped the attributes Event and Holiday category as it is aggregated to month.
* In some of the months there were no holidays. So, took count as zero.
* Separated the overall data into Train and Test datasets.
* Standardized the data.

**Modeling**:

|  |  |  |
| --- | --- | --- |
| **Model** | **Train MAPE** | **Test MAPE** |
| Linear Regression | 7.97 | 50.87 |
| Linear Regression with Log | 6.56 | 46.40 |
| Linear Regression with Log after Step AIC | 3.53 | 17.93 |
| Lasso Regression | 3.66 | 13.36 |
| Ridge Regression | 6.24 | 16.75 |
| Random Forests | 4.05 | 12.80 |
| GBM | 2.692 | 9.588 |
| Tuned GBM | 1.02 | 7.25 |

**Parameters for the best model**:

GBM - Interaction depth=2, shrinkage=0.01, n-trees=1800

**Results & Analysis:**

* Regression analysis is giving better prediction in sales than Time series predictions because here we are considering other factors from weather, economic factors and holidays data.
* Economic factors are impacting sales than weather and holiday factors.
* Improvements can be done by feature engineering and tuning other models.
* The main lesson learned is how the real data exists and how to pre-process the data and bring it to a position on which we can apply machine learning.
* How to do feature engineering on the data.

**Appendices:**

**Regression:**



**Time series:**

