Name: Varsha Deshpande

I. Analysis Tasks

Question 1: Which store has maximum sales ?

- **R Code** :

```
library("readxl")
work_dir <- "C:/Users/vdvde/Downloads"
setwd(work_dir)

getwd()

walmart_data = read.csv('Walmart_Store_sales.csv')
View(walmart_data)

library('dplyr')
store_sales = aggregate(Weekly_Sales~Store,walmart_data,FUN=sum)
View(store_sales)

numeric_data = c(store_sales$Weekly_Sales)
max_sales = max(numeric_data,na.rm = TRUE)
store_with_max = filter(store_sales,Weekly_Sales == max_sales)
paste("Store with maximum sales is: ",store_with_max$Store ,"with sales of
:",store_with_max$Weekly_Sales)
```

- **Screenshots with output :**

walmart_data :

RegressionAssignment.R | ProjectVarsha.R | walmart_data | store_sales

Filter

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
2	1	12-02-2010	1641957	1	38.51	2.548	211.2422	8.106
3	1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
4	1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
5	1	05-03-2010	1554807	0	46.50	2.625	211.3501	8.106
6	1	12-03-2010	1439542	0	57.79	2.667	211.3806	8.106
7	1	19-03-2010	1472516	0	54.58	2.720	211.2156	8.106
8	1	26-03-2010	1404430	0	51.45	2.732	211.0180	8.106
9	1	02-04-2010	1594968	0	62.27	2.719	210.8204	7.808
10	1	09-04-2010	1545419	0	65.86	2.770	210.6229	7.808

Store_sales (filtered data with store wise total sales):

RegressionAssignment.R | ProjectVarsha.R | walmart_data | store_sales

Filter

	Store	Weekly_Sales
1	1	222402809
2	2	275382441
3	3	57586735
4	4	299543953
5	5	45475689
6	6	223756131
7	7	81598275
8	8	129951181
9	9	77789219
10	10	271617714
11	11	193962787

Maximum store calculation output :

```
> numeric_data = as.numeric(store_sales$Weekly_Sales)
> max_sales = max(numeric_data,na.rm = TRUE)
> store_with_max = filter(store_sales,Weekly_Sales == max_sales)
> paste("Store with maximum sales is: ",store_with_max$Store ,"with sales of :",store_w
ith_max$Weekly_Sales)
[1] "Store with maximum sales is:  20 with sales of : 301397792.46"
```

- **Insights :**

  Method used was to calculate the aggregate of sales over each store and then use max () function to calculate the maximum sales store. Store 20 has the maximum sales in the walmart data set for the given period.

Question 2: Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation ?

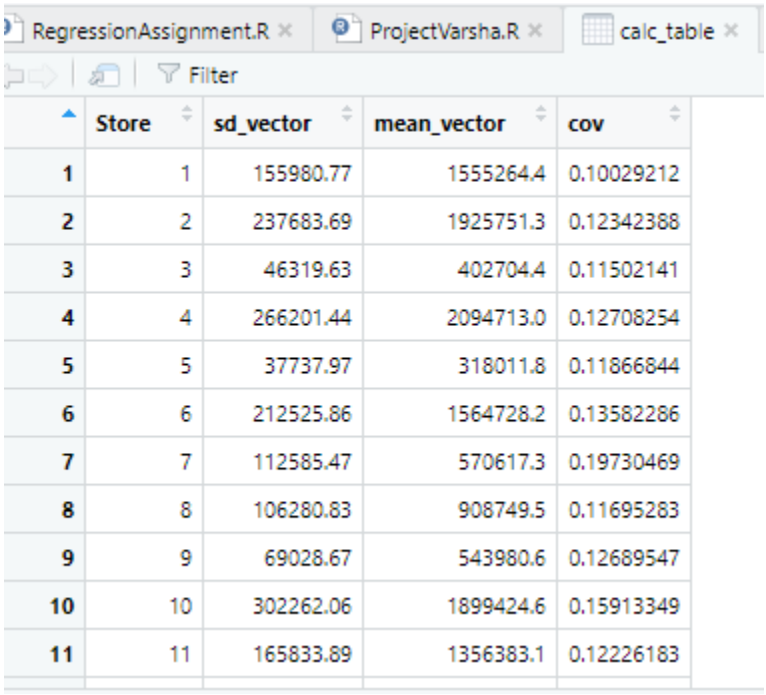
- **R Code:**

```
sd_vector = c()
mean_vector = c()
library('dplyr')
for( i in seq(1,45,by=1)){
  storewise_sales = filter(walmart_data,as.numeric(Store) == i)
  mean_vector <- append(mean_vector, mean(storewise_sales$Weekly_Sales))
  sd_vector <- append(sd_vector,sd(storewise_sales$Weekly_Sales))

}
#Calculate Coff of variances
max_variance = max(sd_vector)
calc_table = data.frame(Store=seq(1:45),sd_vector,mean_vector)
```

```
View(calc_table)

calc_table$Store = unique(walmart_data$Store)
calc_table$cov = calc_table$sd_vector/ calc_table$mean_vector
View(calc_table)
store_with_max_variance = select(filter(calc_table, calc_table$sd_vector ==
max_variance),Store)
paste("Store with maximum Variance is:", store_with_max_variance)
```

- **Screenshots with output :**

Store-wise mean and standard deviation and coefficient of variance vectors:


RegressionAssignment.R | ProjectVarsha.R | calc_table

Filter

	Store	sd_vector	mean_vector	cov
1	1	155980.77	1555264.4	0.10029212
2	2	237683.69	1925751.3	0.12342388
3	3	46319.63	402704.4	0.11502141
4	4	266201.44	2094713.0	0.12708254
5	5	37737.97	318011.8	0.11866844
6	6	212525.86	1564728.2	0.13582286
7	7	112585.47	570617.3	0.19730469
8	8	106280.83	908749.5	0.11695283
9	9	69028.67	543980.6	0.12689547
10	10	302262.06	1899424.6	0.15913349
11	11	165833.89	1356383.1	0.12226183

Filtering the store with maximum variance and displaying:

```
> store_with_max_variance = select(filter(calc_table, calc_table$sd_vector == max_varia
nce),Store)
> paste("Store with maximum Variance is:", store_with_max_variance)
[1] "Store with maximum Variance is: 14"
```

Question 3: Which store has maximum quarterly growth ?

- **R Code:**

```
library(lubridate)
walmart_data$DateOfSales = as.Date(walmart_data$Date,format="%d-%m-%Y")
#Quarter of the Date
walmart_data$Quarter = quarter(walmart_data$DateOfSales)
View(walmart_data)
#Year Of the Date
walmart_data$Year = year(walmart_data$DateOfSales)
```

```
View(walmart_data)
#Year Quarter Column -> eg. 2012-Q2
walmart_data= transform(walmart_data, YearQuarter = paste(Year,"-Q",Quarter))

#Filtering Quarter 3 - 2012 Data
sales_quarterthree = filter(walmart_data, YearQuarter == "2012 -Q 3")
View(sales_quarterthree)

#Filtering Quarter 2 - 2012 Data
sales_quartertwo = filter(walmart_data, YearQuarter == "2012 -Q 2")
View(sales_quartertwo)

#Storewise Quarter 3 - 2012 Data
Storewise_Quarter3Data<-aggregate(Weekly_Sales~Store,sales_quarterthree,FUN=sum)

#Storewise Quarter 2- 2012 Data
Storewise_Quarter2Data<- aggregate(Weekly_Sales~Store,sales_quartertwo,FUN=sum)

#Accumalating Data calculated
accumalated_data <- data.frame(Store = Storewise_Quarter3Data$Store,
          Quarter3Sales=Storewise_Quarter3Data$Weekly_Sales,
           Quarter2Sales = Storewise_Quarter2Data$Weekly_Sales)
View(accumalated_data)

#Calculating growth rate and checking max growth rate store
accumalated_data <- transform(accumalated_data, GrowthRate =
((Quarter3Sales-Quarter2Sales)/Quarter2Sales) *100)
paste("The store which has highest growth rate for Q3/2012 is
Store:",which(accumalated_data$GrowthRate == max(accumalated_data$GrowthRate)))
```

- **Screenshots with output:**

Adding quarter,yearquarter and year column

DateOfSales	Quarter	Year	YearQuarter
2010-02-05	1	2010	2010 -Q 1
2010-02-12	1	2010	2010 -Q 1
2010-02-19	1	2010	2010 -Q 1
2010-02-26	1	2010	2010 -Q 1
2010-03-05	1	2010	2010 -Q 1
2010-03-12	1	2010	2010 -Q 1
2010-03-19	1	2010	2010 -Q 1
2010-03-26	1	2010	2010 -Q 1
2010-04-02	2	2010	2010 -Q 2
2010-04-09	2	2010	2010 -Q 2

Filtering quarter 3 Data:

sales_quarterthree | walmart_data

Fuel_Price	CPI	Unemployment	DateOfSales	Quarter	Year	YearQuarter
3.227	221.8838	6.908	2012-07-06	3	2012	2012 -Q 3
3.256	221.9242	6.908	2012-07-13	3	2012	2012 -Q 3
3.311	221.9327	6.908	2012-07-20	3	2012	2012 -Q 3
3.407	221.9413	6.908	2012-07-27	3	2012	2012 -Q 3
3.417	221.9499	6.908	2012-08-03	3	2012	2012 -Q 3
3.494	221.9584	6.908	2012-08-10	3	2012	2012 -Q 3
3.571	222.0384	6.908	2012-08-17	3	2012	2012 -Q 3
3.620	222.1719	6.908	2012-08-24	3	2012	2012 -Q 3

Filtering quarter 2 data:

sales_quartertwo | sales_quarterthree | walmart_data

Fuel_Price	CPI	Unemployment	DateOfSales	Quarter	Year	YearQuarter
3.891	221.4356	7.143	2012-04-06	2	2012	2012 -Q 2
3.891	221.5102	7.143	2012-04-13	2	2012	2012 -Q 2
3.877	221.5641	7.143	2012-04-20	2	2012	2012 -Q 2
3.814	221.6179	7.143	2012-04-27	2	2012	2012 -Q 2
3.749	221.6718	7.143	2012-05-04	2	2012	2012 -Q 2
3.688	221.7257	7.143	2012-05-11	2	2012	2012 -Q 2
3.630	221.7427	7.143	2012-05-18	2	2012	2012 -Q 2
3.561	221.7449	7.143	2012-05-25	2	2012	2012 -Q 2
3.501	221.7472	7.143	2012-06-01	2	2012	2012 -Q 2

Aggregating storewise and Calculating growth rate and taking max:

```
> Storewise_Quarter3Data<-aggregate(Weekly_Sales~Store,sales_quarterthree,FUN=sum)
>
> Storewise_Quarter2Data<- aggregate(Weekly_Sales~Store,sales_quartertwo,FUN=sum)
>
> accumalated_data <- data.frame(Store = Storewise_Quarter3Data$Store,
+            Quarter3Sales=Storewise_Quarter3Data$Weekly_Sales,
+             Quarter2Sales = Storewise_Quarter2Data$Weekly_Sales)
> View(accumalated_data)
>
> accumalated_data <- transform(accumalated_data, GrowthRate = ((Quarter3Sales-Quarter2
Sales)/Quarter2Sales) *100)
> paste("The store which has highest growth rate for Q3/2012 is Store:",which(accumalat
ed_data$GrowthRate == max(accumalated_data$GrowthRate)))
[1] "The store which has highest growth rate for Q3/2012 is Store: 7"
>
```

	Store	Quarter3Sales	Quarter2Sales	GrowthRate
1	1	20253948	20978760	-3.4549818
2	2	24303355	25083605	-3.1105976
3	3	5298005	5620316	-5.7347486
4	4	27796792	28454364	-2.3109679
5	5	4163791	4466364	-6.7744752
6	6	20167312	20833910	-3.1995803
7	7	8262787	7290859	13.3307760
8	8	11748953	11919631	-1.4319088

- **Insights:**

  Filtering out Quarter Data and Storewise aggregate gives insight that most of the stores have negative growth rate except for the Store number 7. The sales have grown for Store 7 from Q2 to Q3 by 13%.

Question 4: 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

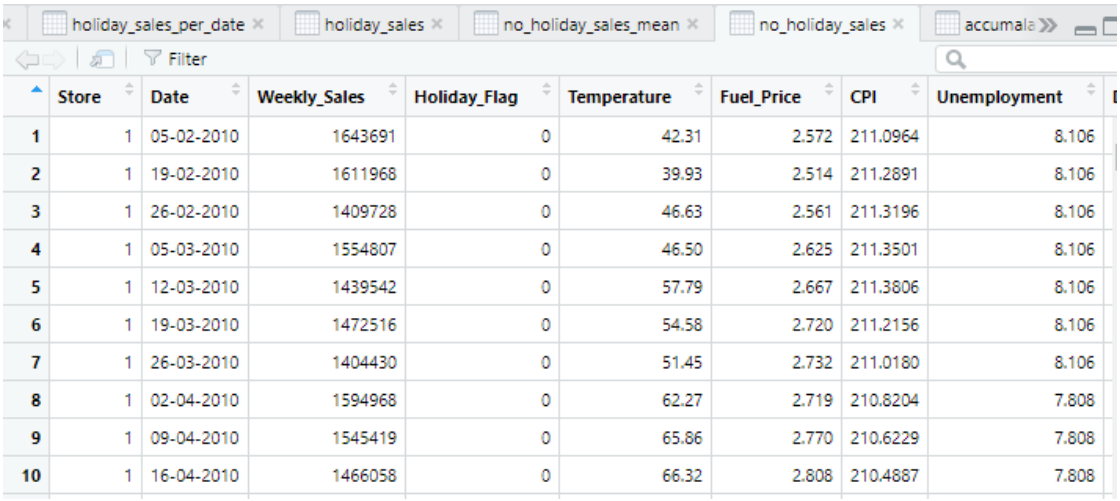
- **R Code:**

```
#Filter non holiday Store sales
no_holiday_sales = filter(walmart_data, as.numeric(Holiday_Flag) == 0)
View(no_holiday_sales)

#Mean per store for non holiday sales
no_holiday_sales_mean = aggregate(Weekly_Sales~Store,no_holiday_sales,
FUN=mean)
View(no_holiday_sales_mean)

#Overall mean for all stores together for non holiday sales
overall_non_holiday_mean =
mean(no_holiday_sales_mean$Weekly_Sales,na.rm=TRUE)
paste("Overall mean for non holiday sales : ",overall_non_holiday_mean)


#Filter holiday Store sales
holiday_sales = filter(walmart_data, as.numeric(Holiday_Flag) == 1)
View(holiday_sales)

#Mean of sales per every date on holidays
holiday_sales_per_date = aggregate(Weekly_Sales~DateOfSales, holiday_sales,
FUN=mean)
```

```
View(holiday_sales_per_date)
#Generated a column for checking if mean sales per holiday is greater or less than mean
#sales calculated for non holiday sales above  (overall_non_holiday_mean)

holiday_sales_per_date = transform(holiday_sales_per_date, ProfitableOrNot =
ifelse(holiday_sales_per_date$Weekly_Sales > overall_non_holiday_mean,"Yes","No"))

#Collect the holidays with more sales than mean non holiday sales
profitable_dates = filter(holiday_sales_per_date,holiday_sales_per_date$ProfitableOrNot
== "Yes")

print("Holidays with Sales more than mean sales on non holidays are: ")
c(profitable_dates$DateOfSales)
```

- **Screenshots with output:**

Filter out non holiday sales

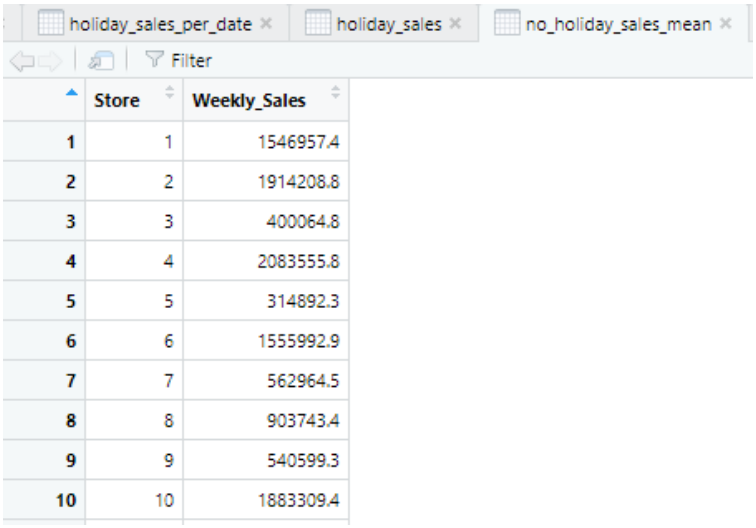


holiday_sales_per_date | holiday_sales | no_holiday_sales_mean | no_holiday_sales | accumula

Filter

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
2	1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
3	1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
4	1	05-03-2010	1554807	0	46.50	2.625	211.3501	8.106
5	1	12-03-2010	1439542	0	57.79	2.667	211.3806	8.106
6	1	19-03-2010	1472516	0	54.58	2.720	211.2156	8.106
7	1	26-03-2010	1404430	0	51.45	2.732	211.0180	8.106
8	1	02-04-2010	1594968	0	62.27	2.719	210.8204	7.808
9	1	09-04-2010	1545419	0	65.86	2.770	210.6229	7.808
10	1	16-04-2010	1466058	0	66.32	2.808	210.4887	7.808

Aggregate non holiday sales storewise



holiday_sales_per_date | holiday_sales | no_holiday_sales_mean

Filter

	Store	Weekly_Sales
1	1	1546957.4
2	2	1914208.8
3	3	400064.8
4	4	2083555.8
5	5	314892.3
6	6	1555992.9
7	7	562964.5
8	8	903743.4
9	9	540599.3
10	10	1883309.4

Overall mean calculation:

```
> overall_non_holiday_mean = mean(no_holiday_sales_mean$Weekly_Sales,na.rm=TRUE
> paste("Overall mean for non holiday sales : ",overall_non_holiday_mean)
[1] "Overall mean for non holiday sales :  1041256.38020886"
```

Filter Holiday Sales:

ProjectVarsha.R* | holiday_sales | no_holiday_sales | accumalated_data | sales_quart

Filter

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI
1	1	12-02-2010	1641957.4	1	38.51	2.548	211.2422
2	1	10-09-2010	1507460.7	1	78.69	2.565	211.4952
3	1	26-11-2010	1955624.1	1	64.52	2.735	211.7484
4	1	31-12-2010	1367320.0	1	48.43	2.943	211.4049
5	1	11-02-2011	1649614.9	1	36.39	3.022	212.9367
6	1	09-09-2011	1540471.2	1	76.00	3.546	215.8611
7	1	25-11-2011	2033320.7	1	60.14	3.236	218.4676
8	1	30-12-2011	1497462.7	1	44.55	3.129	219.5360
9	1	10-02-2012	1802477.4	1	48.02	3.409	220.2652
10	1	07-09-2012	1661767.3	1	83.96	3.730	222.4390

Aggregate holiday sales by date:

ProjectVarsha.R | holiday_sales_per_date

Filter

	DateOfSales	Weekly_Sales
1	2010-02-12	1074148.4
2	2010-09-10	1014097.7
3	2010-11-26	1462689.0
4	2010-12-31	898500.4
5	2011-02-11	1051915.4
6	2011-09-09	1039182.8
7	2011-11-25	1479857.9
8	2011-12-30	1023165.8
9	2012-02-10	1111320.2
10	2012-09-07	1074001.3

Compare with Non holiday sales mean and check if holiday is profitable or not :

ProjectVarsha.R | holiday_sales_per_date | holiday_sal

Filter

	DateOfSales	Weekly_Sales	ProfitableOrNot
1	2010-02-12	1074148.4	Yes
2	2010-09-10	1014097.7	No
3	2010-11-26	1462689.0	Yes
4	2010-12-31	898500.4	No
5	2011-02-11	1051915.4	Yes
6	2011-09-09	1039182.8	No
7	2011-11-25	1479857.9	Yes
8	2011-12-30	1023165.8	No
9	2012-02-10	1111320.2	Yes
10	2012-09-07	1074001.3	Yes

Profitable holidays displayed:

```
[1] "Holidays with Sales more than mean sales on non holidays are: "
> c(profitable_dates$DateOfSales)
[1] "2010-02-12" "2010-11-26" "2011-02-11" "2011-11-25" "2012-02-10" "2012-09-07"
```

- **Insights:**

  The holidays of : SuperBowl, Thanksgiving have more than average non holiday sales for years 2010,2011. However in year 2012, the sales for Labour Day sales were more than average sales of non holiday sales. People visit the store often or more on these holidays.

Question 5: Provide a monthly and semester view of sales in units and give insights

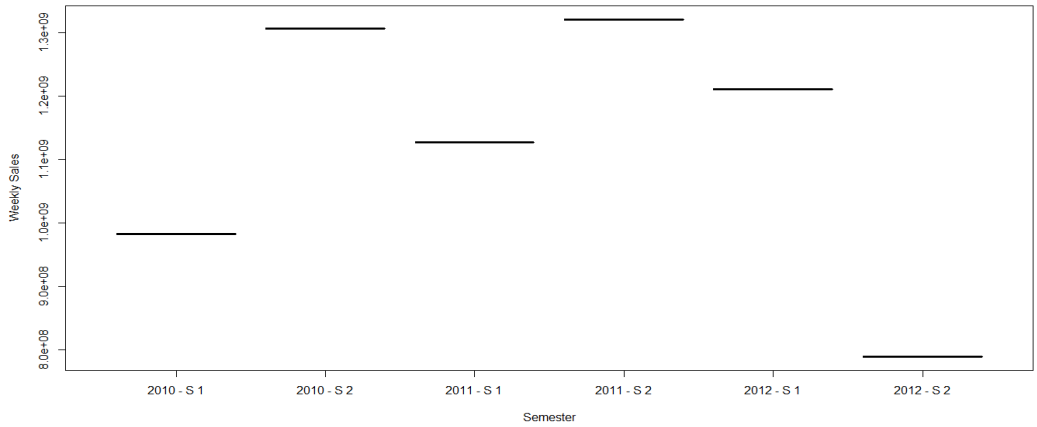
- **R Code:**

```
#Calculating Semester
library("lubridate")
walmart_data$Semester = semester(walmart_data$DateOfSales)
View(walmart_data)

#Year Of the Date
walmart_data$Year = year(walmart_data$DateOfSales)


#Appending calculated Year column with Semester
walmart_data$SemesterYear = paste(walmart_data$Year,"-","S",walmart_data$Semester)
View(walmart_data)

#Aggregating data based on Semester+Year
Semester_wise_group = aggregate(Weekly_Sales~SemesterYear, walmart_data,FUN=sum)
View(Semester_wise_group)

#Plotting Semester+Year Daat
```

```
plot(x = as.factor(Semester_wise_group$SemesterYear),y = Semester_wise_group$Weekly_Sales,
    xlab = "Semester",
    ylab = "Weekly Sales",
    main = ""
)

#Calculating Month of the Sales
walmart_data$Month = month(walmart_data$DateOfSales)
View(walmart_data)

#Aggregating per Month and Year
month_wise_group = aggregate(Weekly_Sales~Month+Year, walmart_data,FUN=sum)
View(month_wise_group)
month_wise_group$MonthYear = paste(month_wise_group$Year,"-","M",month_wise_group$Month)
View(month_wise_group)

#Plotting Data Month Wise per year
#2010 Month Data Plot
month_data_2010 = filter(month_wise_group, month_wise_group$Year == 2010)
plot(x = as.factor(month_data_2010$MonthYear),y = month_data_2010$Weekly_Sales,
    xlab = "Month",
    ylab = "Weekly Sales",
    main = ""
)

#2011 Month Data Plot
month_data_2011 = filter(month_wise_group, month_wise_group$Year == 2011)
plot(x = as.factor(month_data_2011$MonthYear),y = month_data_2011$Weekly_Sales,
    xlab = "Month",
    ylab = "Weekly Sales",
    main = ""
)

#2012 Month Data Plot
month_data_2012 = filter(month_wise_group, month_wise_group$Year == 2012)
plot(x = as.factor(month_data_2012$MonthYear),y = month_data_2012$Weekly_Sales,
    xlab = "Month",
    ylab = "Weekly Sales",
    main = ""
)
```

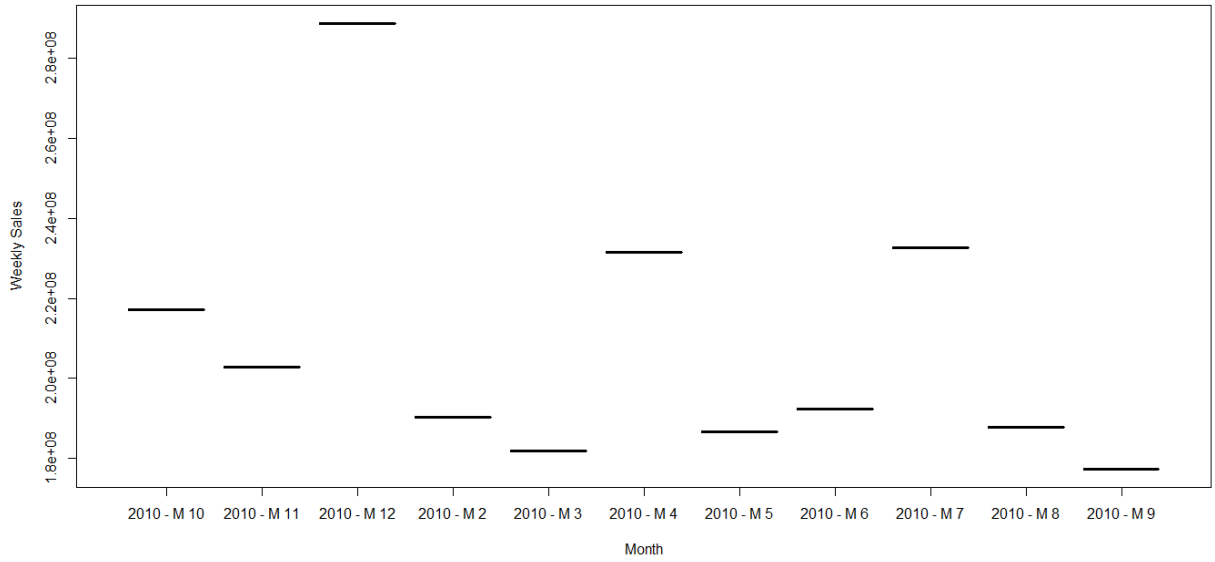
- **Screenshots with output:**

Semester wise sales plot:

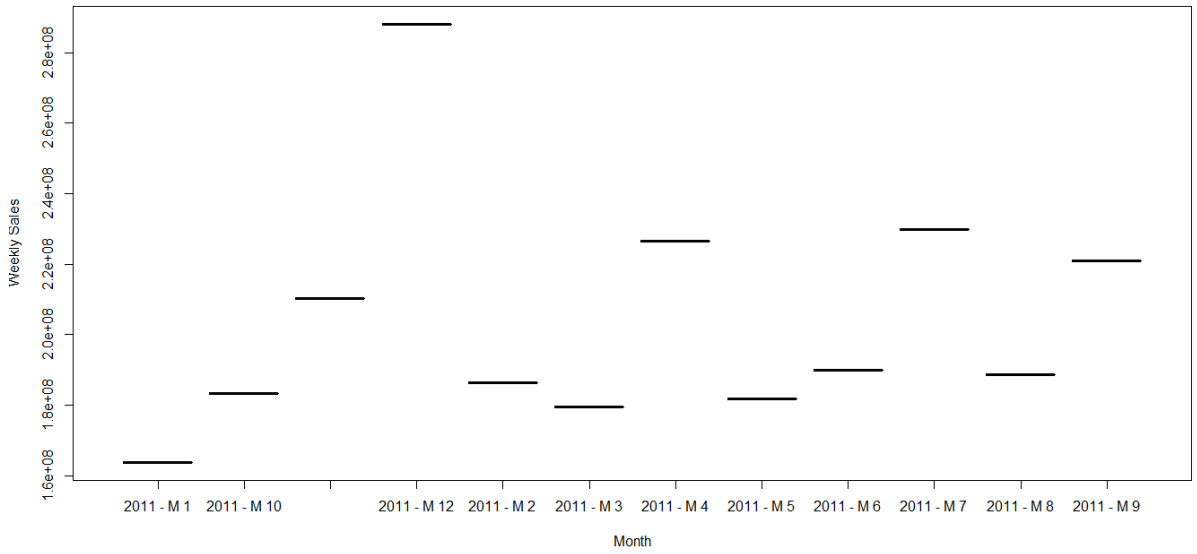


Each Year Monthly plot:

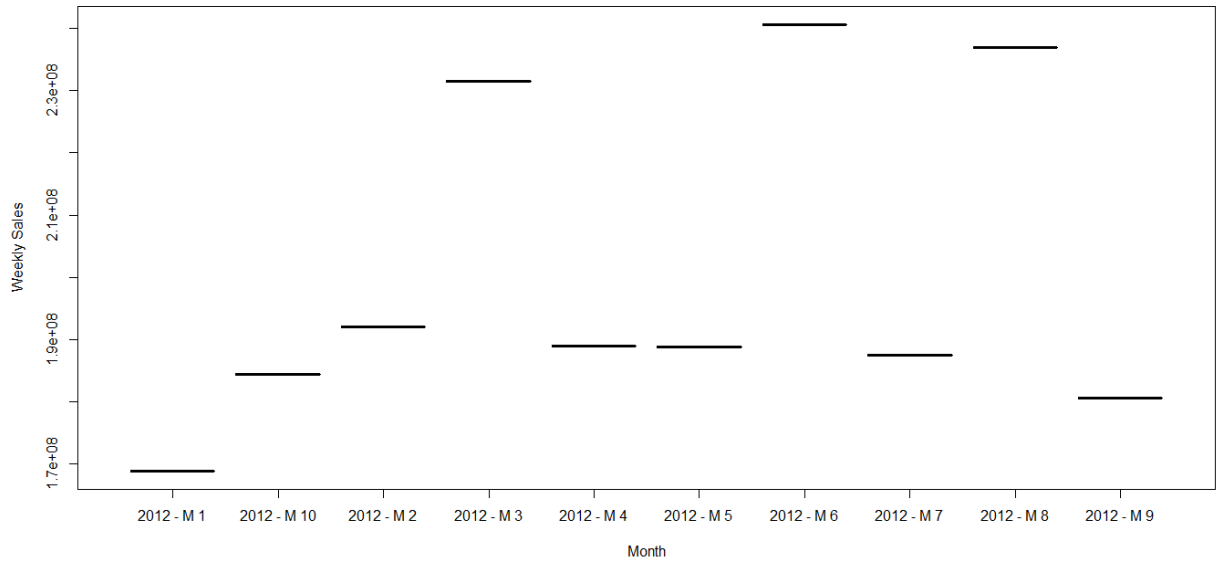
1. 2010



2. 2011



3. 2012



Semester wise aggregate:

	SemesterYear	Weekly_Sales
1	2010 - S 1	982622260
2	2010 - S 2	1306263860
3	2011 - S 1	1127339797
4	2011 - S 2	1320860210
5	2012 - S 1	1210765416
6	2012 - S 2	789367443

Month wise aggregate

Filter

	Month	Year	Weekly_Sales	MonthYear
1	2	2010	190332983	2010 - M 2
2	3	2010	181919803	2010 - M 3
3	4	2010	231412368	2010 - M 4
4	5	2010	186710934	2010 - M 5
5	6	2010	192246172	2010 - M 6
6	7	2010	232580126	2010 - M 7
7	8	2010	187640111	2010 - M 8
8	9	2010	177267896	2010 - M 9
9	10	2010	217161824	2010 - M 10
10	11	2010	202853370	2010 - M 11
11	12	2010	288760533	2010 - M 12

- **Insights:**

Semester-wise:
2011, Semester 2 has the maximum sales among the three years - 2010,2011,2012
The sales have dropped considerably in semester 2 of 2012.
The Semester 1 sales for all three years show an increasing growth.
Semester 2 sales grew for the first two years by a small margin but dropped considerably in 2012.

Monthly for Year 2010:
December month has maximum sales in the Year 2010 as compared to the other months in the year.
The graph does not show any linear relation between months sales.

Monthly for Year 2011:
December month has the highest sales in the Year 2011 as compared to other months.
The graph does not show any linear relation between months sales.

Monthly for Year 2012:
Months March, June and August have the highest or more sales in 2012 compared to other months.

II. Statistics Tasks:

Question 1:

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.**

- **R Code:**

```
library("readxl")
work_dir <- "C:/Users/vdvde/Downloads"
setwd(work_dir)

getwd()

walmart_data_regression = read.csv('Walmart_Store_sales.csv')
View(walmart_data_regression)

library(lubridate)
walmart_data_regression$DateOfSales = as.Date(walmart_data_regression$Date,format="%d-%m-%Y")

#Filter Store 1 data
library("dplyr")
walmart_data_Store1 = filter(walmart_data_regression,as.numeric(Store) == 1)
View(walmart_data_Store1)

#Transforming Dates to ordered numbers
walmart_data_Store1$orderedDates = seq(1:length(unique(walmart_data_Store1$DateOfSales)))

#Capturing the data set relevant for regression by removing insignificant variables - Store,Date,DateOfSales
=> Date is considered as numbered series instead
walmart_data_Store1_with_dates = subset(walmart_data_Store1, select = - c(Store,Date,DateOfSales))

#Hypothesis for CPI
#H0: CPI has no impact on Weekly Sales of Store 1
#Ha: CPI has considerable impact on Weekly Sales of Store 1

#Hypothesis for Fuel Price
#H0: Fuel Price has no impact on Weekly Sales of Store 1
#Ha: Fuel Price has considerable impact on Weekly Sales of Store 1

#Hypothesis for Unemployment
#H0: Unemployment has no impact on Weekly Sales of Store 1
#Ha: Unemployment has considerable impact on Weekly Sales of Store 1


#Performing linear regression on all relevant Data Set
model = lm(Weekly_Sales~.,walmart_data_Store1_with_dates)
summary(model)
```

```
###Conclusion:  Only Temperature is a significant for 0.05 i.e pvalue of only Temperature <0.05
#Hence considering 0.1 cutoff

##Temperature and Holiday Flag are significant for cutoff 0.1

#H0 is true for all other variables, CPI,Unemployment, Fuel Price


#Hypothesis for 0.1 cutoff - temperature
#H0: Temperature has no impact on Weekly Sales of Store 1
#Ha: Temperature has considerable impact on Weekly Sales of Store 1

#Hypothesis for 0.1 cutoff - holiday flag
#H0: Holiday_flag has no impact on Weekly Sales of Store 1
#Ha: Holiday_flag has considerable impact on Weekly Sales of Store 1


model = lm(Weekly_Sales~Holiday_Flag+Temperature,walmart_data_Store1_with_dates)
summary(model)

#Conclusion: As per p-value
#Ha is true for Temperature and for Holiday flag with 0.1 cutoff

#Predicting values with the two significant variables
walmart_data_Store1_with_dates$predicted_val = predict(model,walmart_data_Store1_with_dates)
summary(walmart_data_Store1_with_dates$predicted_val)

walmart_data_Store1_with_dates$difference =
abs((walmart_data_Store1_with_dates$Weekly_Sales-walmart_data_Store1_with_dates$predicted_val)/wal
mart_data_Store1_with_dates$Weekly_Sales)
View(walmart_data_Store1_with_dates)


paste("Error Rate of the model ->", mean(walmart_data_Store1_with_dates$difference) * 100 , "%")
paste("Accuracy Rate of the model ->",(1- mean(walmart_data_Store1_with_dates$difference)) * 100 , "%")
-----------------------------------------------------------------------------------------------------
#Week Days extraction code
walmart_data$Days = weekdays(walmart_data$DateOfSales)
View(walmart_data)
```

- **Screenshots with output:**

Filtered Store 1 Data:

RegressionAssignment.R | walmart_data_Store1 | walmart_data_regression

Filter

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
2	1	12-02-2010	1641957	1	38.51	2.548	211.2422	8.106
3	1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
4	1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
5	1	05-03-2010	1554807	0	46.50	2.625	211.3501	8.106
6	1	12-03-2010	1439542	0	57.79	2.667	211.3806	8.106
7	1	19-03-2010	1472516	0	54.58	2.720	211.2156	8.106
8	1	26-03-2010	1404430	0	51.45	2.732	211.0180	8.106
9	1	02-04-2010	1594968	0	62.27	2.719	210.8204	7.808
10	1	09-04-2010	1545419	0	65.86	2.770	210.6229	7.808

Transformed Date Column in Data:

DateOfSales	orderedDates
2010-02-05	1
2010-02-12	2
2010-02-19	3
2010-02-26	4
2010-03-05	5
2010-03-12	6
2010-03-19	7
2010-03-26	8
2010-04-02	9
2010-04-09	10
2010-04-16	11

Filtered Data relevant for regression model:

Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	orderedDates
1643691	0	42.31	2.572	211.0964	8.106	1
1641957	1	38.51	2.548	211.2422	8.106	2
1611968	0	39.93	2.514	211.2891	8.106	3
1409728	0	46.63	2.561	211.3196	8.106	4
1554807	0	46.50	2.625	211.3501	8.106	5
1439542	0	57.79	2.667	211.3806	8.106	6
1472516	0	54.58	2.720	211.2156	8.106	7
1404430	0	51.45	2.732	211.0180	8.106	8
1594968	0	62.27	2.719	210.8204	7.808	9
1545419	0	65.86	2.770	210.6229	7.808	10
1466058	0	66.32	2.808	210.4887	7.808	11

g 1 to 12 of 143 entries, 7 total columns

Model Summary (Considering all above variables):

```
> model = lm(Weekly_Sales~.,walmart_data_Store1_with_dates)
> summary(model)

Call:
lm(formula = Weekly_Sales ~ ., data = walmart_data_Store1_with_dates)

Residuals:
    Min      1Q  Median      3Q     Max
-304675  -79201  -18223   56433  849204

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1946579.6  2996221.4  -0.650   0.5170
Holiday_Flag    88070.3    49947.1   1.763   0.0801 .
Temperature     -2182.4      931.9  -2.342   0.0206 *
Fuel_Price     -27298.0    49791.0  -0.548   0.5844
CPI             14332.6    13439.1   1.066   0.2881
Unemployment    81043.9    59083.3   1.372   0.1724
orderedDates      278.7     1404.8   0.198   0.8430
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147000 on 136 degrees of freedom
Multiple R-squared:  0.1497,    Adjusted R-squared:  0.1122
F-statistic: 3.991 on 6 and 136 DF,  p-value: 0.001035
```

Only temperature significant for cut off of 0.05 (pvalue of temperature<0.05 -> 0.02)

Hence considering 0.1 cutoff and selecting relevant variables -> Holiday_flag(0.08 < 0.1) and Temperature (0.02 < 0.05)

Model with above filtered variables:

```
> model = lm(Weekly_Sales~Holiday_Flag+Temperature,walmart_data_Store1_with_dates)
> summary(model)

Call:
lm(formula = Weekly_Sales ~ Holiday_Flag + Temperature, data = walmart_data_Store1_with
_dates)

Residuals:
    Min      1Q  Median      3Q     Max
-318302  -87590  -16966   70396  805903

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1691625.9    64161.0  26.365   <2e-16 ***
Holiday_Flag   95407.4    50619.7   1.885   0.0615 .
Temperature    -2094.0      909.1  -2.303   0.0227 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 151200 on 140 degrees of freedom
Multiple R-squared:  0.07311,   Adjusted R-squared:  0.05987
F-statistic: 5.522 on 2 and 140 DF,  p-value: 0.004918
```

Predicted values for the weekly sales with above model:

predicted_val
1603029
1706394
1608013
1593983
1594256
1570615
1577336
1583890
1561233
1553716

Difference between predicted and original values:

predicted_val	difference
1603029	0.0247379094
1706394	0.0392437449
1608013	0.0024535610
1593983	0.1307031369
1594256	0.0253722516
1570615	0.0910518643
1577336	0.0711845931
1583890	0.1277817269
1561233	0.0211507638
1553716	0.0053691258

Error and Accuracy Rate of the model:

```
> paste("Error Rate of the model ->", mean(walmart_data_Store1_with_dates$difference) *
 100 , "%")
[1] "Error Rate of the model -> 6.43344240866695 %"
> paste("Accuracy Rate of the model ->",(1- mean(walmart_data_Store1_with_dates$differe
nce)) * 100 , "%")
[1] "Accuracy Rate of the model -> 93.5665575913331 %"
>
```

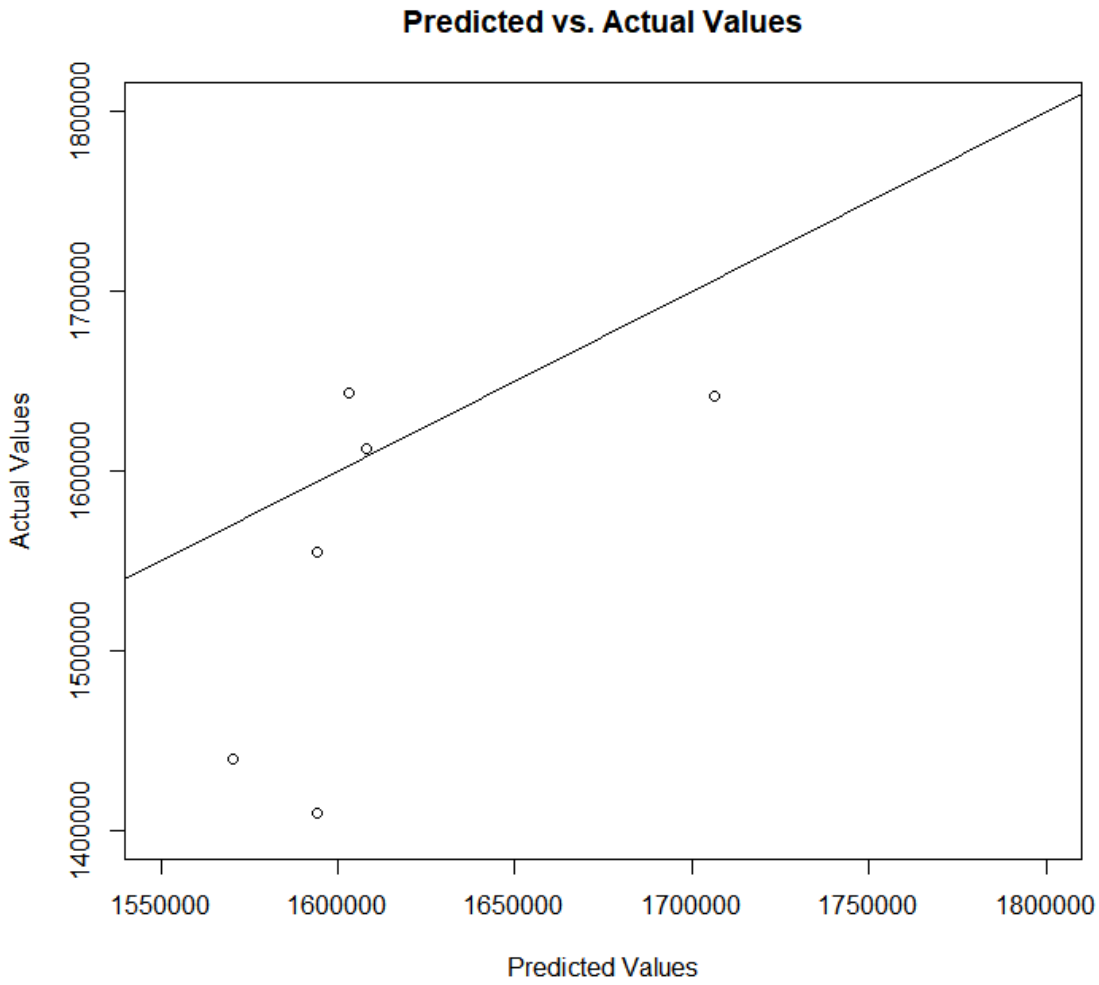
Week Days of Date :

DateOfSales	Days
2010-02-05	Friday
2010-02-12	Friday
2010-02-19	Friday
2010-02-26	Friday
2010-03-05	Friday
2010-03-12	Friday
2010-03-19	Friday
2010-03-26	Friday
2010-04-02	Friday
2010-04-09	Friday

- **Insights:**
  Accuracy rate of the model used is 93%. I.e 93% of the predicted results are similar to the original results.

Accuracy Graph for first 6 rows (for demo):

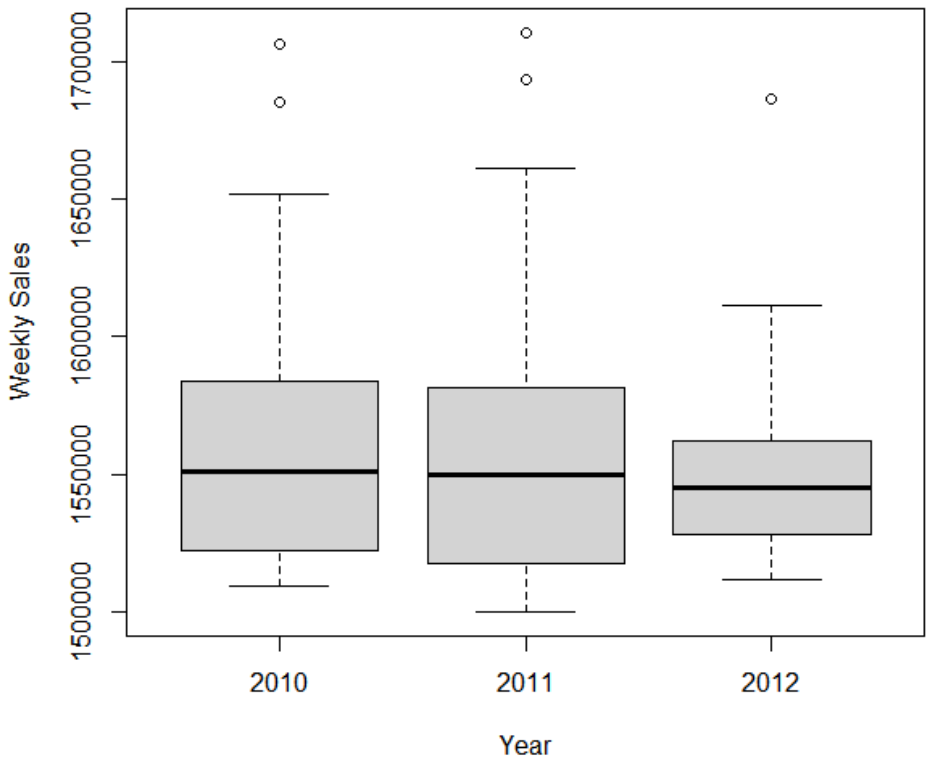
```
plot(x=walmart_data_Store1_subset$predicted_val,
     y=walmart_data_Store1_subset$Weekly_Sales,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs. Actual Values',
     ylim=c(1400000,1800000),
     xlim=c(1550000,1800000)
)
abline(a=0, b=1)
```



Abline to fit a line passing through the data. (Sample intercept)

Plot of predicted values of Weekly Sales for each year.

```
plot(x = as.factor(walmart_data_Store1_with_dates$Year),y =
walmart_data_Store1_with_dates$predicted_val,
    xlab = "Year",
    ylab = "Weekly Sales",
    main = ""
)
```



The overall sales of 2010 and 2011 show similar statistic and graph the mid sales being in range of 1550000.