Lab 3 - Final Project (Walmart)

Chris Bennett, Jackson Lane, Ravneet Ghuman

October 29, 2016

Conducting EDA

```
#Load datasets
setwd("/home/blue/ds/271/assignments/w271_lab3")
stores=read.csv("stores.csv")
features=read.csv("features.csv")
train=read.csv("train.csv")
# review stores dataset
head(stores)
```

```
##   Store Type   Size
## 1     1    A 151315
## 2     2    A 202307
## 3     3    B  37392
## 4     4    A 205863
## 5     5    B  34875
## 6     6    A 202505
```

```
str(stores)
```

```
## 'data.frame':    45 obs. of  3 variables:
##  $ Store: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type : Factor w/ 3 levels "A","B","C": 1 1 2 1 2 1 2 1 2 2 ...
##  $ Size : int  151315 202307 37392 205863 34875 202505 70713 155078 125833 126512 ...
```

```
# stores has an ID for each of the 45 stores along with it's type - A, B or C
# and the size of the store
by(stores, stores$Type, summary)
```

```
## stores$Type: A
##      Store        Type        Size
##  Min.   : 1.00   A:22   Min.   : 39690
##  1st Qu.:11.50   B: 0   1st Qu.:155841
##  Median :25.00   C: 0   Median :202406
##  Mean   :22.23          Mean   :177248
##  3rd Qu.:32.75          3rd Qu.:203819
##  Max.   :41.00          Max.   :219622
## --------------------------------------------------------
## stores$Type: B
##      Store        Type        Size
##  Min.   : 3.00   A: 0   Min.   : 34875
##  1st Qu.:10.00   B:17   1st Qu.: 93188
##  Median :17.00   C: 0   Median :114533
##  Mean   :18.35          Mean   :101191
##   3rd Qu.:23.00          3rd Qu.:123737
##   Max.   :45.00          Max.   :140167
## --------------------------------------------------------
## stores$Type: C
##      Store       Type       Size
##   Min.   :30.00   A:0   Min.   :39690
##   1st Qu.:37.25   B:0   1st Qu.:39745
##   Median :40.00   C:6   Median :39910
##   Mean   :39.00         Mean   :40542
##   3rd Qu.:42.75         3rd Qu.:40774
##   Max.   :44.00         Max.   :42988
```

```
# review shows that type C stores are generally smaller in size compated to the other
# two. Type A and B have a wider range (size wise)

#review features
nrow(features)
```

```
## [1] 8190
```

```
# total number of entries = 8190
# For 45 stores i.e. we have 8190/45=182 entries per store
features$Date=as.Date(features$Date)
head(features)
```

```
##   Store       Date Temperature Fuel_Price MarkDown1 MarkDown2 MarkDown3
## 1     1 2010-02-05       42.31      2.572        NA        NA        NA
## 2     1 2010-02-12       38.51      2.548        NA        NA        NA
## 3     1 2010-02-19       39.93      2.514        NA        NA        NA
## 4     1 2010-02-26       46.63      2.561        NA        NA        NA
## 5     1 2010-03-05       46.50      2.625        NA        NA        NA
## 6     1 2010-03-12       57.79      2.667        NA        NA        NA
##   MarkDown4 MarkDown5      CPI Unemployment IsHoliday
## 1        NA        NA 211.0964        8.106     FALSE
## 2        NA        NA 211.2422        8.106      TRUE
## 3        NA        NA 211.2891        8.106     FALSE
## 4        NA        NA 211.3196        8.106     FALSE
## 5        NA        NA 211.3501        8.106     FALSE
## 6        NA        NA 211.3806        8.106     FALSE
```

```
str(features)
```

```
## 'data.frame':    8190 obs. of  12 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Date, format: "2010-02-05" "2010-02-12" ...
##  $ Temperature : num  42.3 38.5 39.9 46.6 46.5 ...
##  $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
##  $ MarkDown1   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown2   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown3   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown4   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MarkDown5   : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ CPI         : num  211 211 211 211 211 ...
##  $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
##  $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

```
summary(features)
```

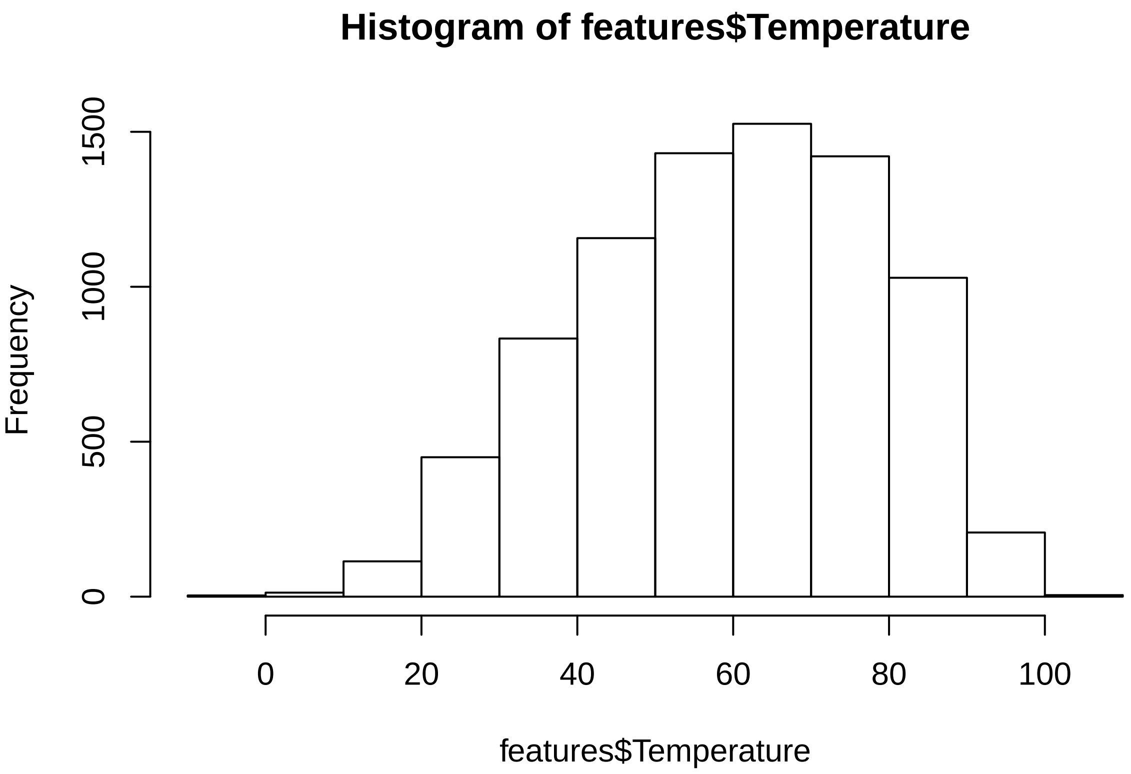
```
##      Store         Date             Temperature       Fuel_Price   
##  Min.   : 1   Min.   :2010-02-05   Min.   : -7.29   Min.   :2.472  
##  1st Qu.:12   1st Qu.:2010-12-17   1st Qu.: 45.90   1st Qu.:3.041  
##  Median :23   Median :2011-10-31   Median : 60.71   Median :3.513  
##  Mean   :23   Mean   :2011-10-31   Mean   : 59.36   Mean   :3.406  
##  3rd Qu.:34   3rd Qu.:2012-09-14   3rd Qu.: 73.88   3rd Qu.:3.743  
##  Max.   :45   Max.   :2013-07-26   Max.   :101.95   Max.   :4.468  
##                                                                    
##    MarkDown1        MarkDown2           MarkDown3        
##  Min.   : -2781   Min.   :  -265.76   Min.   :  -179.26  
##  1st Qu.:  1578   1st Qu.:    68.88   1st Qu.:     6.60  
##  Median :  4744   Median :   364.57   Median :    36.26  
##  Mean   :  7032   Mean   :  3384.18   Mean   :  1760.10  
##  3rd Qu.:  8923   3rd Qu.:  2153.35   3rd Qu.:   163.15  
##  Max.   :103185   Max.   :104519.54   Max.   :149483.31  
##  NA's   :4158     NA's   :5269        NA's   :4577       
##    MarkDown4          MarkDown5             CPI         Unemployment   
##  Min.   :    0.22   Min.   :  -185.2   Min.   :126.1   Min.   : 3.684  
##  1st Qu.:  304.69   1st Qu.:  1440.8   1st Qu.:132.4   1st Qu.: 6.634  
##  Median : 1176.42   Median :  2727.1   Median :182.8   Median : 7.806  
##  Mean   : 3292.94   Mean   :  4132.2   Mean   :172.5   Mean   : 7.827  
##  3rd Qu.: 3310.01   3rd Qu.:  4832.6   3rd Qu.:213.9   3rd Qu.: 8.567  
##  Max.   :67474.85   Max.   :771448.1   Max.   :229.0   Max.   :14.313  
##  NA's   :4726       NA's   :4140       NA's   :585     NA's   :585     
##  IsHoliday      
##  Mode :logical  
##  FALSE:7605     
##  TRUE :585      
##  NA's :0        
##                 
##                 
## 
```

```
get_year_month <- function(d) {
    return(as.integer(format(d, "%m")))
}

#months vector assuming 1st month is Jan.
months <- c("Jan","Feb","Mar",
              "Apr","May","Jun",
              "Jul","Aug","Sep",
              "Oct","Nov","Dec")

#add abbreviated month name
features$monthsText <- months[ get_year_month(features$Date) ]
features$month <- get_year_month(features$Date)
```

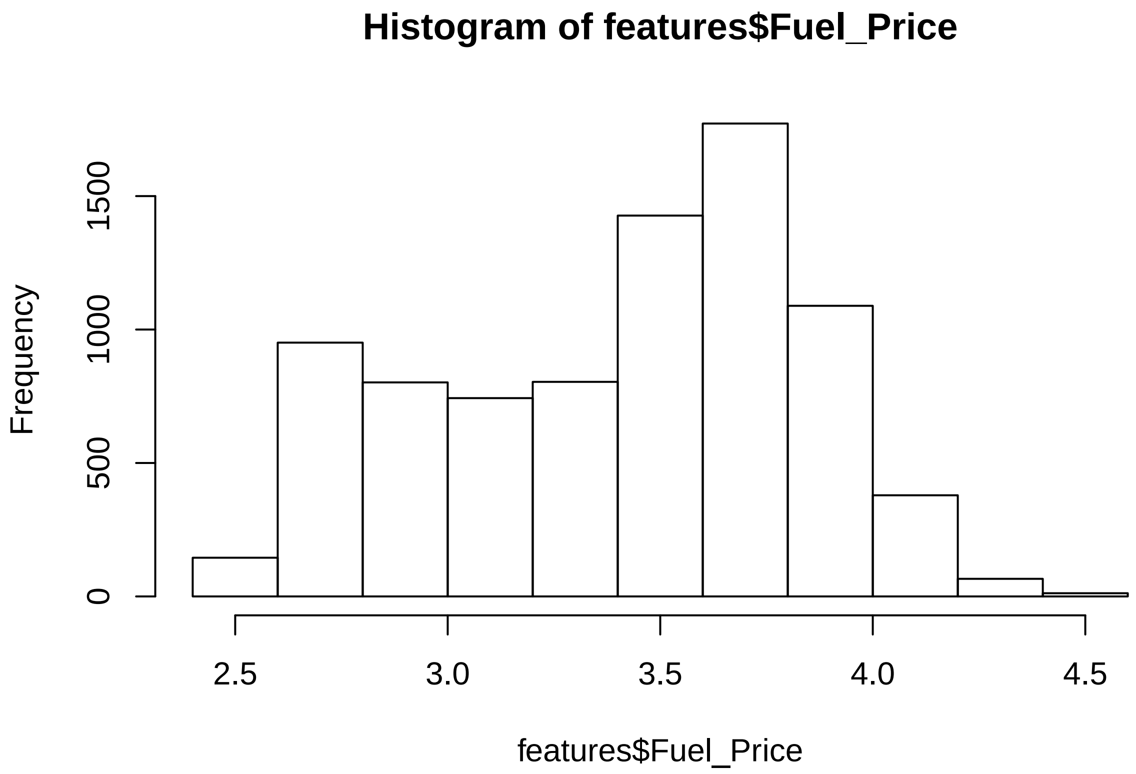
```r
# Review a summary of months vs holidays
table(features$monthsText, features$IsHoliday)
```

```
##
##         FALSE TRUE
##    Apr    810    0
##    Aug    585    0
##    Dec    495  135
##    Feb    540  180
##    Jan    540    0
##    Jul    810    0
##    Jun    765    0
##    Mar    810    0
##    May    765    0
##    Nov    450  135
##    Oct    585    0
##    Sep    450  135
```

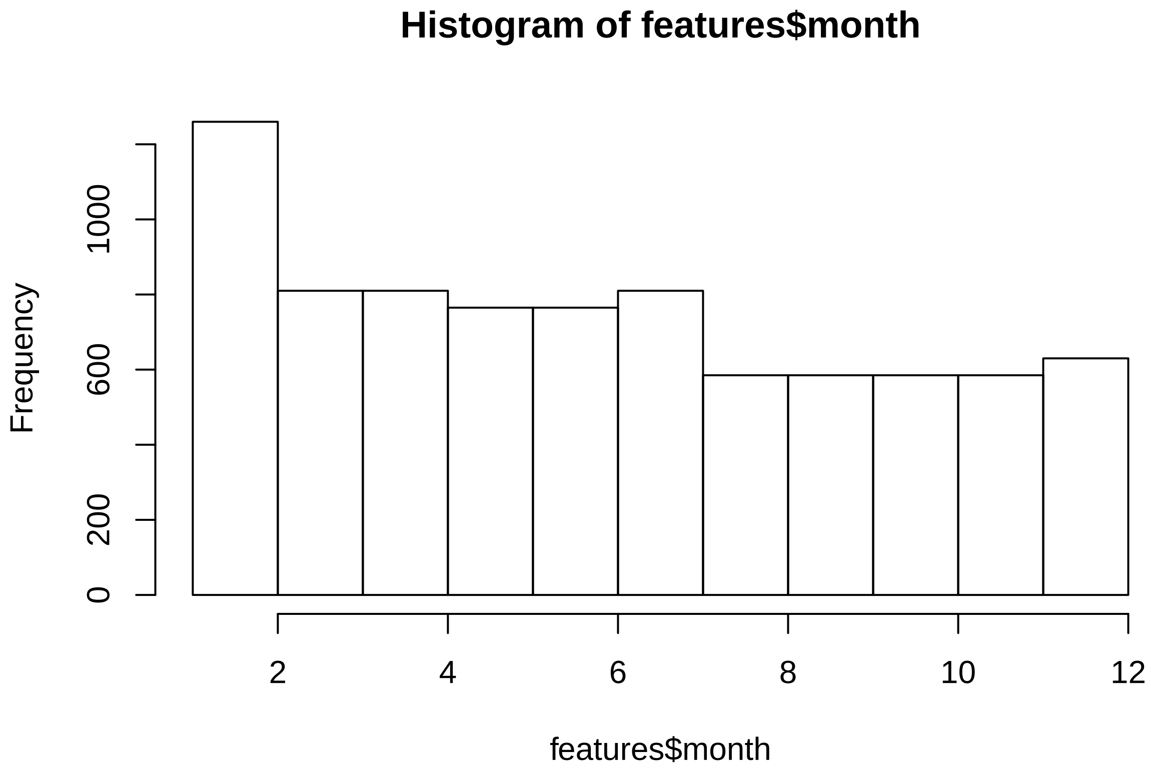
```r
hist(features$Temperature)
```



```r
hist(features$Fuel_Price)
```



```r
hist(features$month)
```



```
#########review train.csv
head(train)
```

```
##   Store Dept       Date Weekly_Sales IsHoliday
## 1     1    1 2010-02-05     24924.50     FALSE
## 2     1    1 2010-02-12     46039.49      TRUE
## 3     1    1 2010-02-19     41595.55     FALSE
## 4     1    1 2010-02-26     19403.54     FALSE
## 5     1    1 2010-03-05     21827.90     FALSE
## 6     1    1 2010-03-12     21043.39     FALSE
```

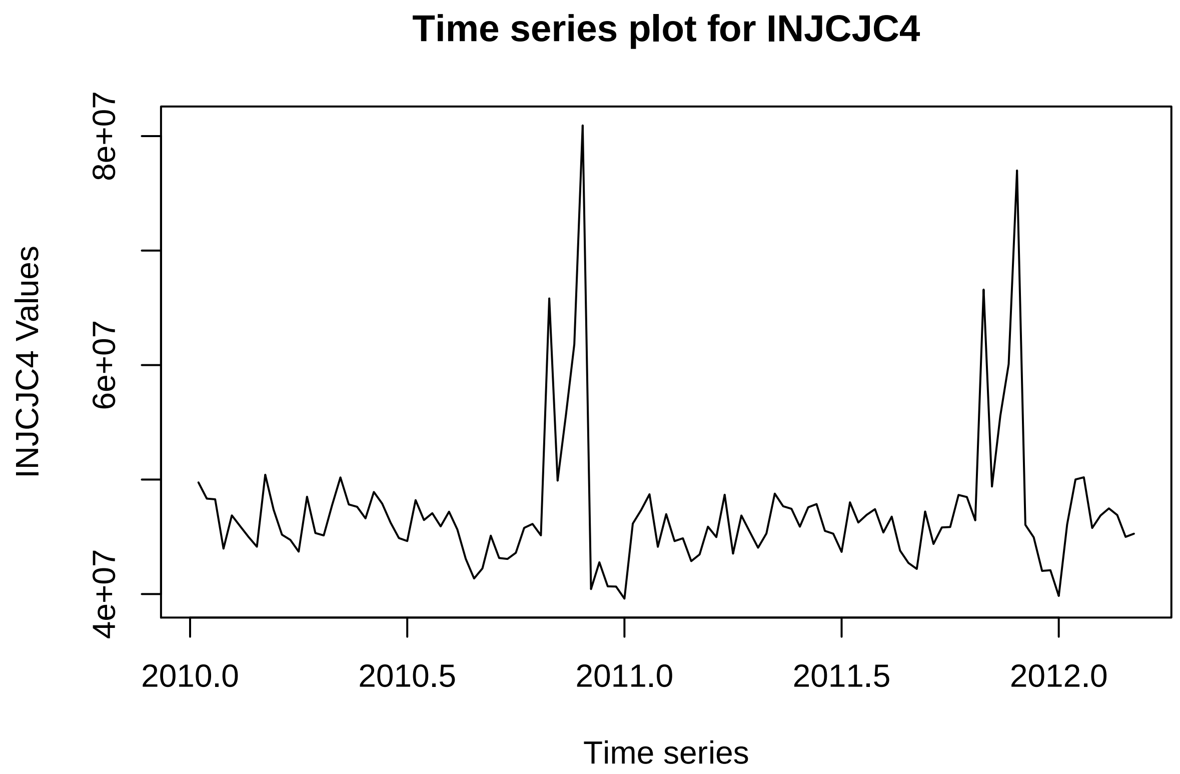
```
str(train)
```

```
## 'data.frame':    421570 obs. of  5 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Dept        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Factor w/ 143 levels "2010-02-05","2010-02-12",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Weekly_Sales: num  24924 46039 41596 19404 21828 ...
##  $ IsHoliday   : logi  FALSE TRUE FALSE FALSE FALSE FALSE ...
```

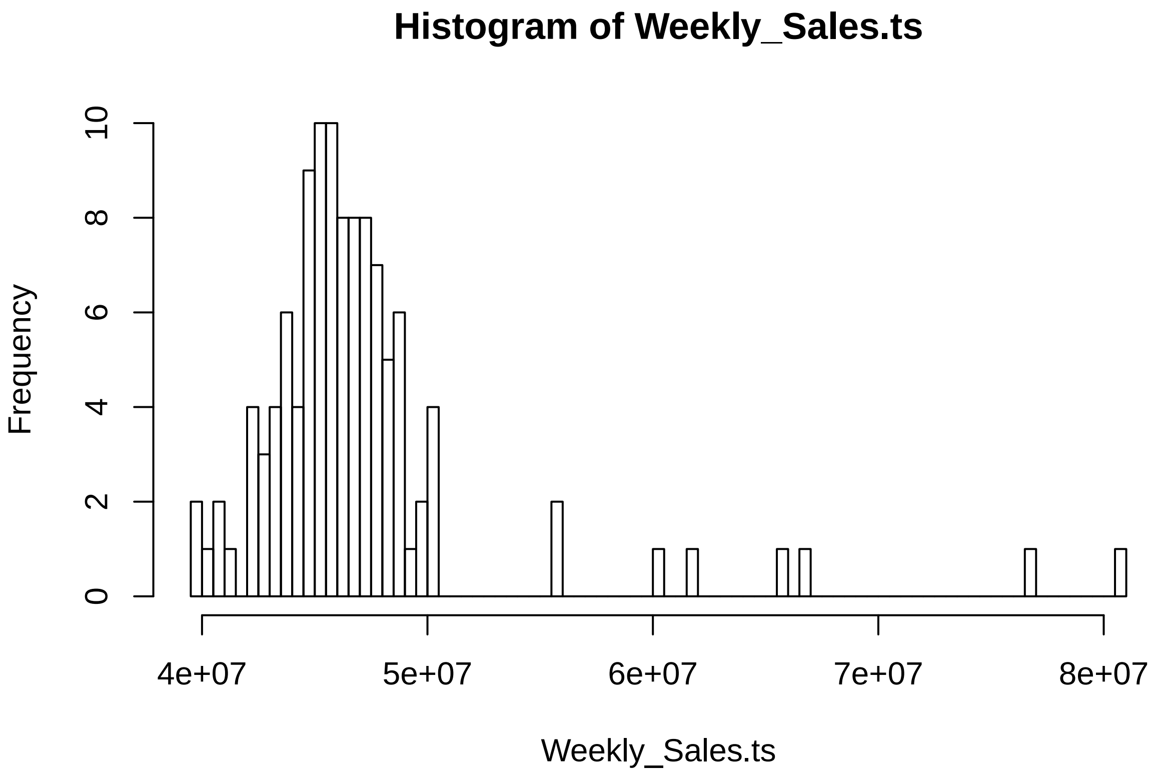
```
train$Date=as.Date(train$Date)
summary(train)
```

```
##      Store           Dept            Date             Weekly_Sales
##  Min.   : 1.0   Min.   : 1.00   Min.   :2010-02-05   Min.   : -4989
##  1st Qu.:11.0   1st Qu.:18.00   1st Qu.:2010-10-08   1st Qu.:  2080
##  Median :22.0   Median :37.00   Median :2011-06-17   Median :  7612
##  Mean   :22.2   Mean   :44.26   Mean   :2011-06-18   Mean   : 15981
##  3rd Qu.:33.0   3rd Qu.:74.00   3rd Qu.:2012-02-24   3rd Qu.: 20206
##  Max.   :45.0   Max.   :99.00   Max.   :2012-10-26   Max.   :693099
##  IsHoliday
##  Mode :logical
##  FALSE:391909
##  TRUE :29661
##  NA's :0
##
##
```

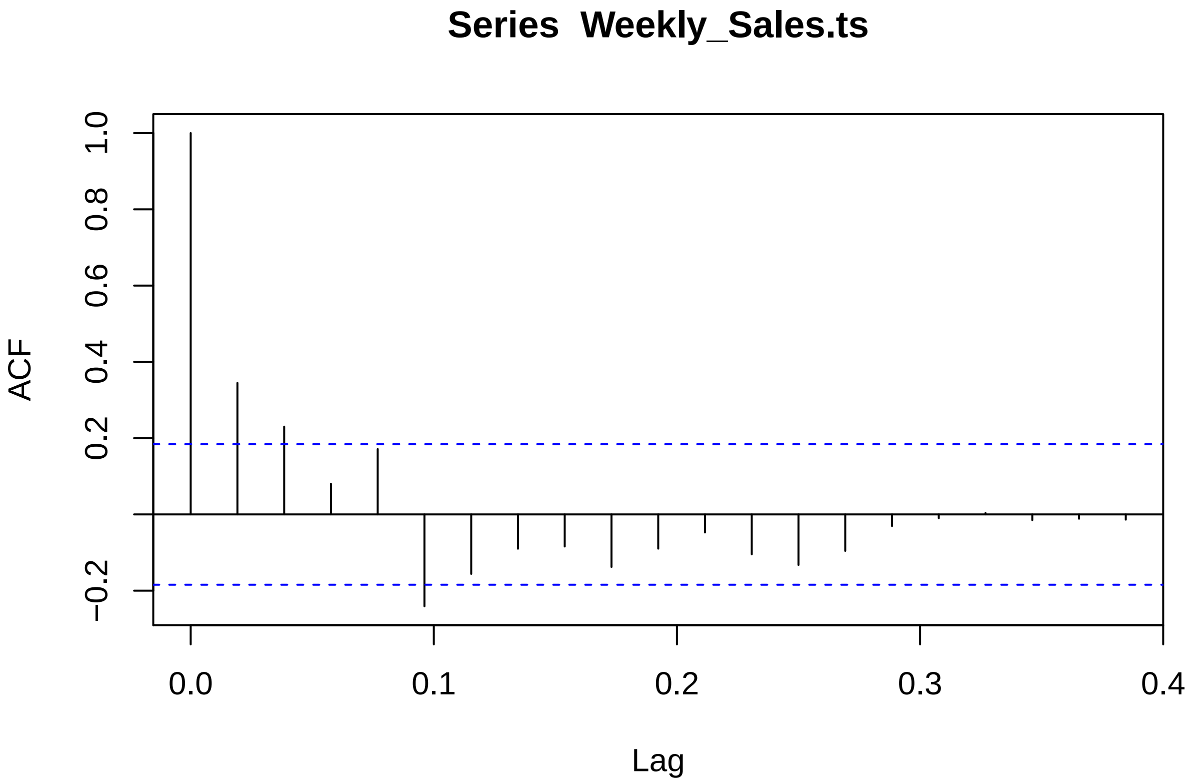
```
# create a new var - total sales by date
salesbydate=aggregate(train$Weekly_Sales,by=list(train$Date), FUN=sum)

Weekly_Sales.ts = ts(salesbydate$x, frequency=52, start=c(2010,2,5), end=c(2012,10,26) )
#plot time series
plot.ts(Weekly_Sales.ts, main="Time series plot for INJCJC4", xlab="Time series", ylab="INJCJC4 Values"
```



```
#plotting histogram for INJCJC4
hist(Weekly_Sales.ts, breaks = 100)
```



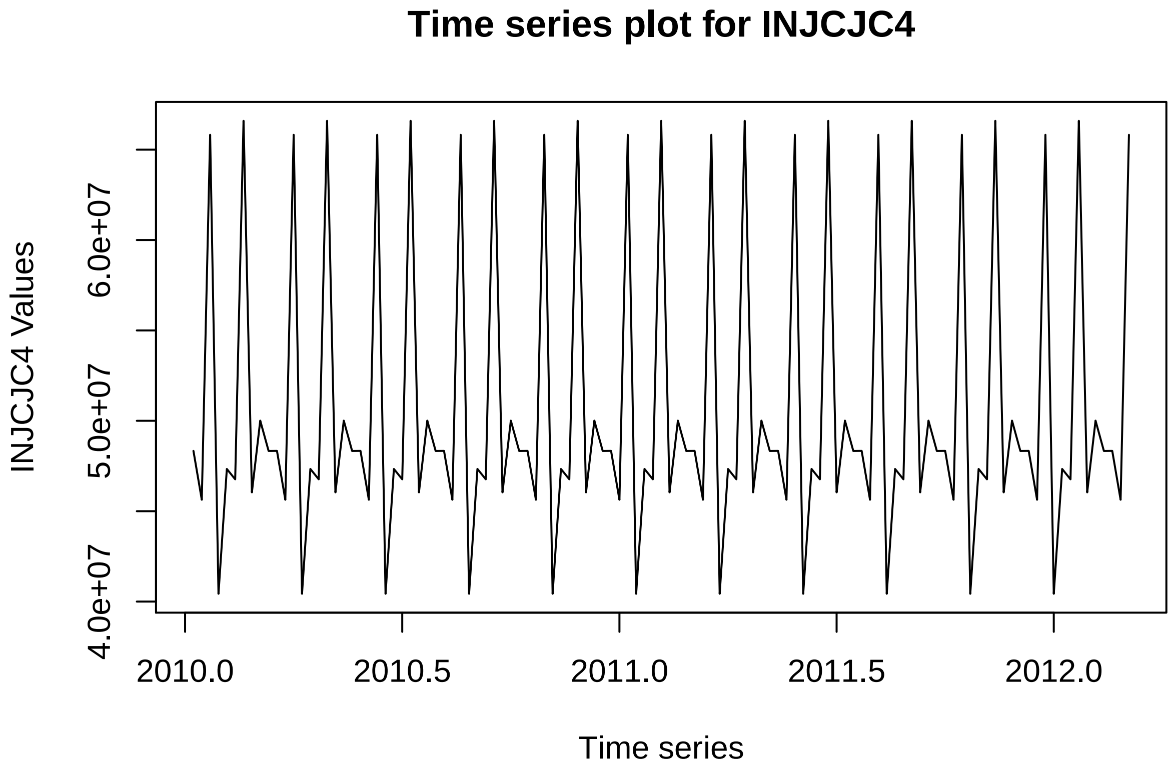
```
acf(Weekly_Sales.ts)
```



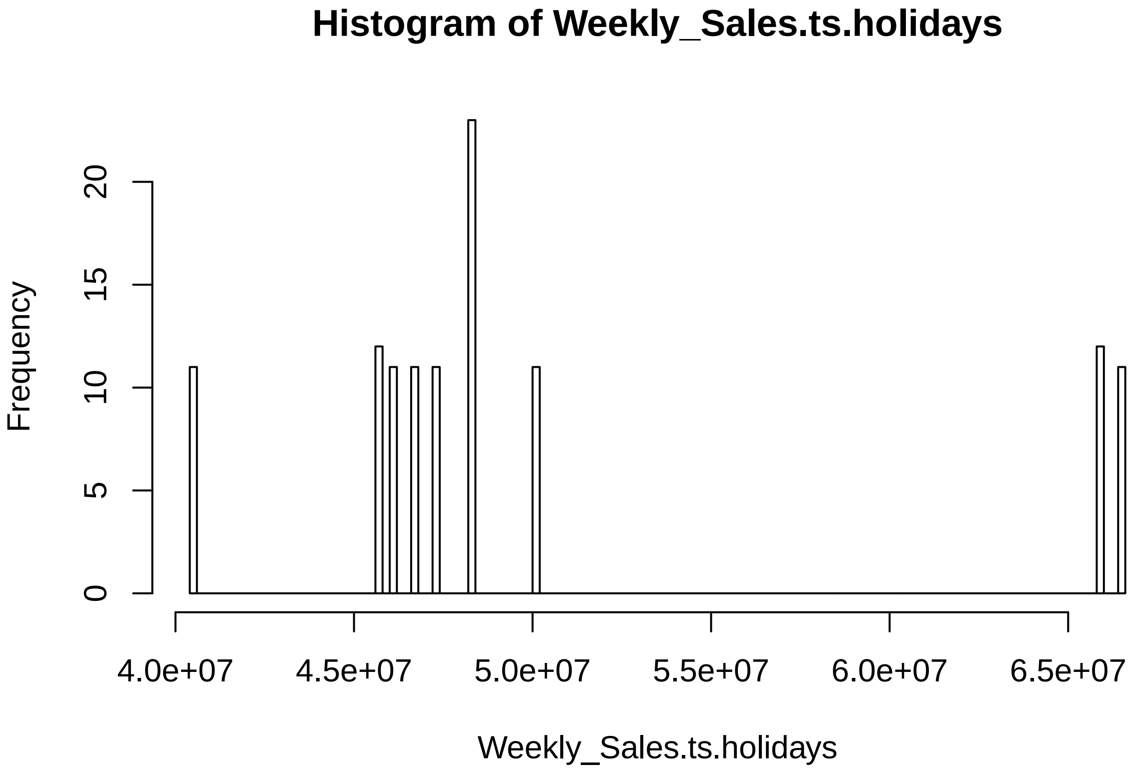
```
############
# create a new var - total sales by date (only  holidays)
salesbydate.holidays=aggregate(train$Weekly_Sales[train$IsHoliday==TRUE],by=list(train$Date[train$IsHol
Weekly_Sales.ts.holidays = ts(salesbydate.holidays$x, frequency=52, start=c(2010,2,5), end=c(2012,10,26
summary(salesbydate.holidays)
```

```
##     Group.1                  x
##  Min.   :2010-02-12   Min.   :40432519
##  1st Qu.:2010-12-04   1st Qu.:46222653
##  Median :2011-05-27   Median :47833126
##  Mean   :2011-06-03   Mean   :50529955
##  3rd Qu.:2011-12-21   3rd Qu.:49591225
##  Max.   :2012-09-07   Max.   :66593605
```

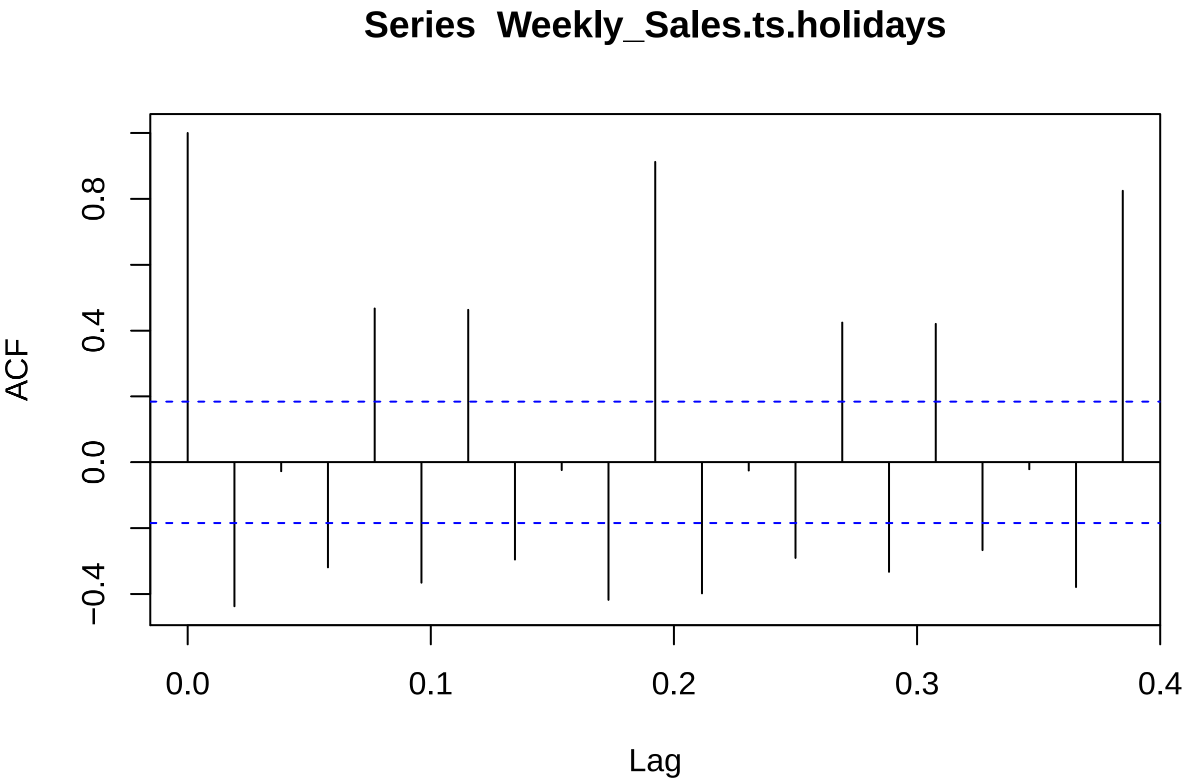
```
#plot time series
plot.ts(Weekly_Sales.ts.holidays, main="Time series plot for INJCJC4", xlab="Time series", ylab="INJCJC
```



```r
#plotting histogram for INJCJC4
hist(Weekly_Sales.ts.holidays, breaks = 100)
```



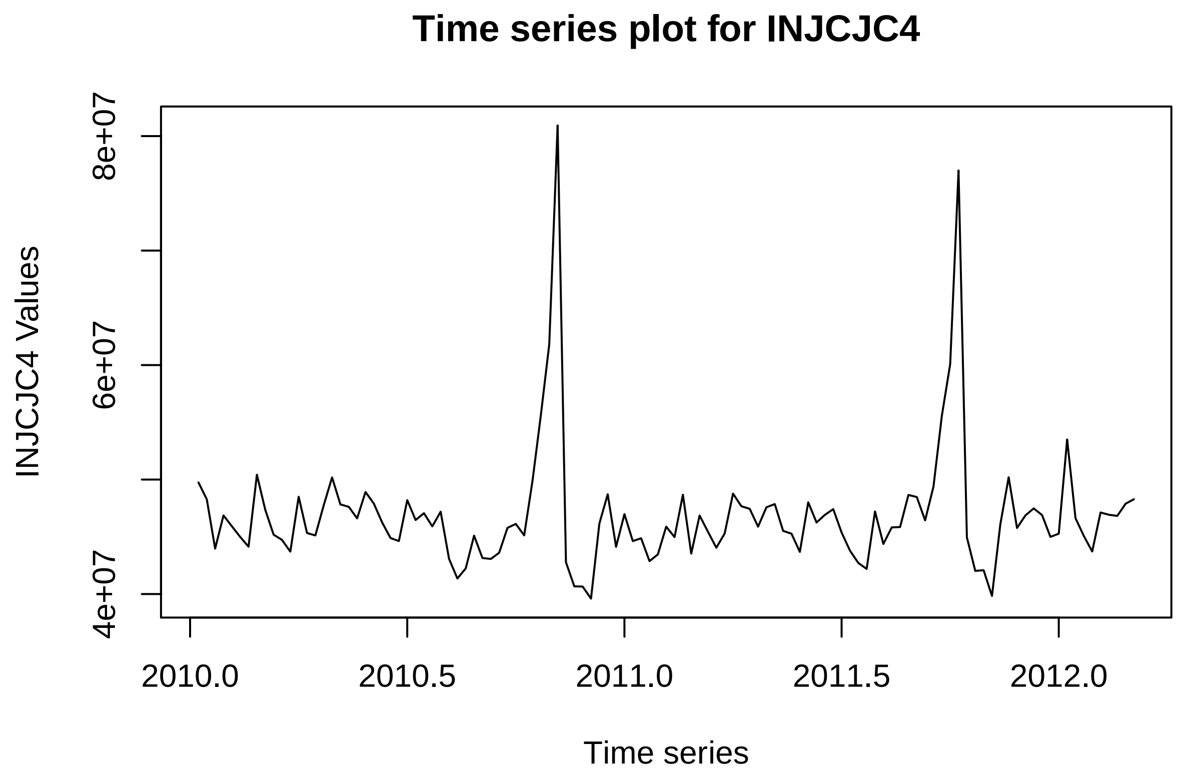
```r
acf(Weekly_Sales.ts.holidays)
```



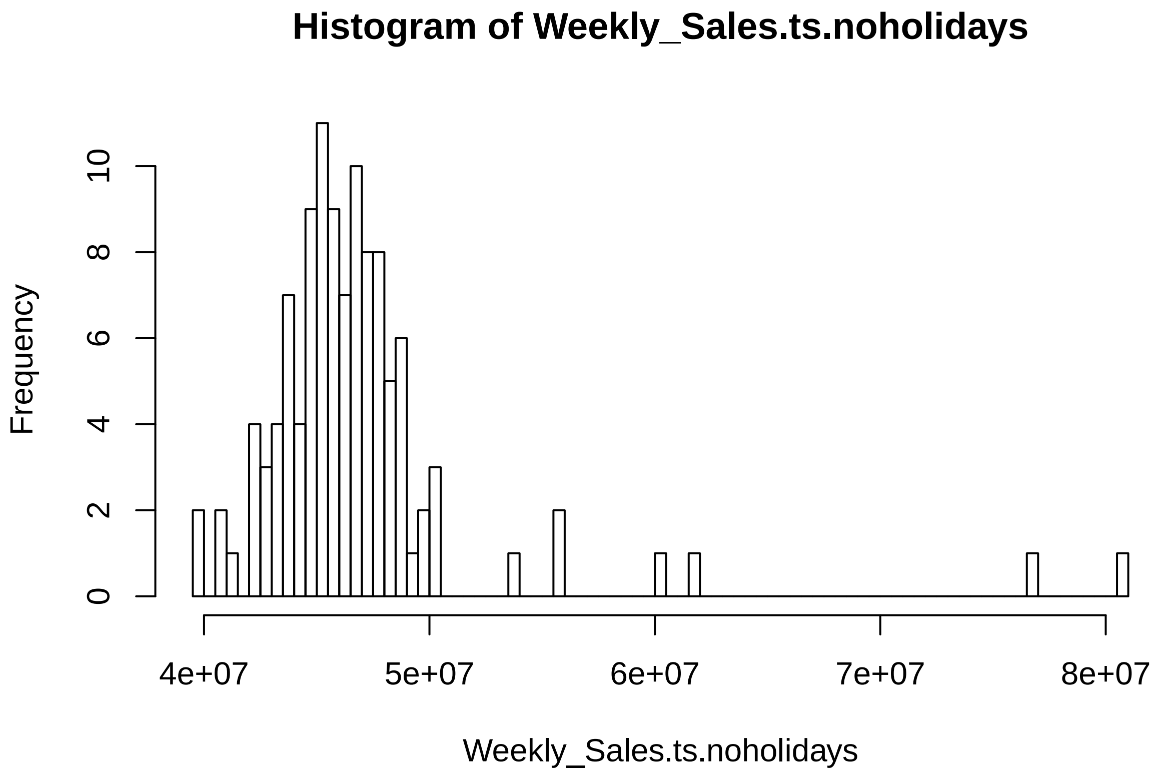
```r
############
# create a new var - total sales by date (without  holidays)
salesbydate.noholidays=aggregate(train$Weekly_Sales[train$IsHoliday==FALSE],by=list(train$Date[train$Is
summary(salesbydate.noholidays)
```

```
##     Group.1                  x
##  Min.   :2010-02-05   Min.   :39599853
##  1st Qu.:2010-10-08   1st Qu.:44734453
##  Median :2011-06-17   Median :46128514
##  Mean   :2011-06-18   Mean   :46856537
##  3rd Qu.:2012-03-02   3rd Qu.:47668285
##  Max.   :2012-10-26   Max.   :80931416
```

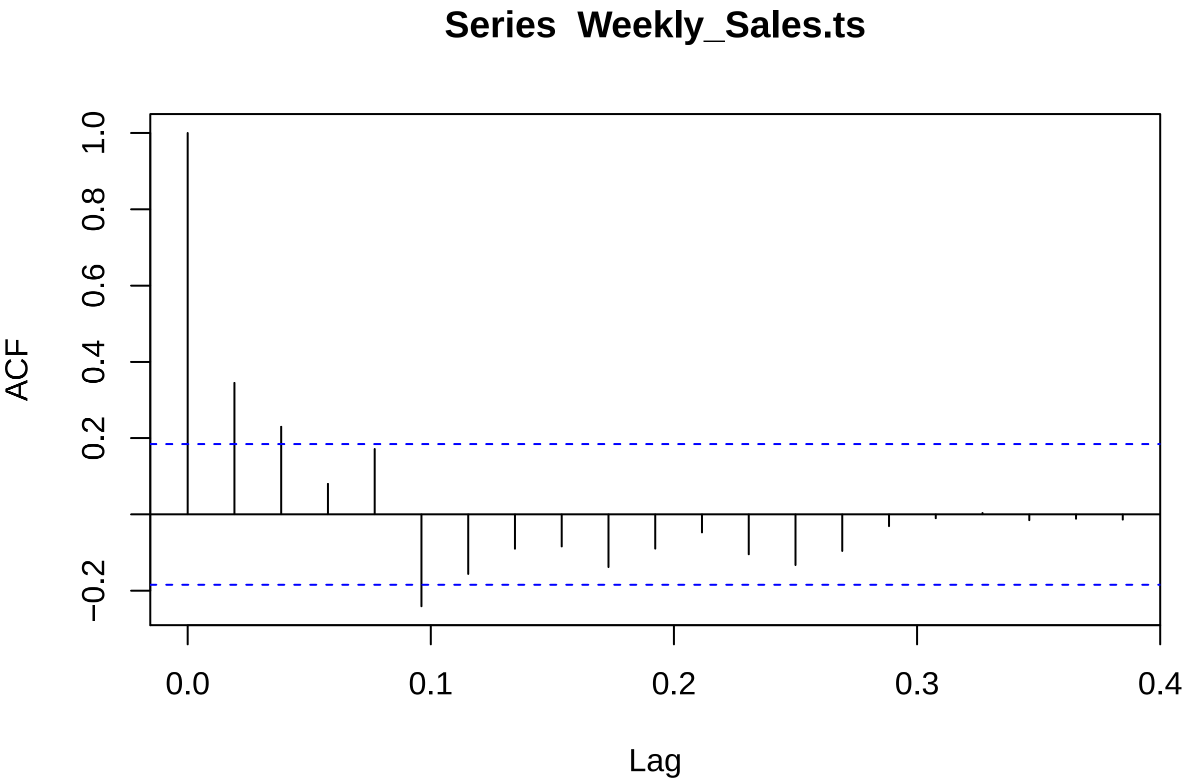
```r
Weekly_Sales.ts.noholidays = ts(salesbydate.noholidays$x, frequency=52, start=c(2010,2,5), end=c(2012,1
#plot time series
plot.ts(Weekly_Sales.ts.noholidays, main="Time series plot for INJCJC4", xlab="Time series", ylab="INJC
```



```
#plotting histogram for INJCJC4
hist(Weekly_Sales.ts.noholidays, breaks = 100)
```



```r
acf(Weekly_Sales.ts)
```


Series Weekly_Sales.ts
ACF
Lag