Forecast Project

BAR Group 4

21/12/2020

```
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Exploratory Data Analysis

```
setwd("C:/Users/amitb/OneDrive/Desktop/placement/Term 5/BAR/project")
train <- fread("train.csv",stringsAsFactors = TRUE)
head(train)
```

```
##    Store DayOfWeek       Date Sales Customers Open Promo StateHoliday
## 1:     1         5 2015-07-31  5263       555    1     1            0
## 2:     2         5 2015-07-31  6064       625    1     1            0
## 3:     3         5 2015-07-31  8314       821    1     1            0
## 4:     4         5 2015-07-31 13995      1498    1     1            0
## 5:     5         5 2015-07-31  4822       559    1     1            0
## 6:     6         5 2015-07-31  5651       589    1     1            0
##    SchoolHoliday
## 1:             1
## 2:             1
## 3:             1
## 4:             1
## 5:             1
## 6:             1
```

```
summary(train)
```

```
##      Store          DayOfWeek            Date             Sales
##  Min.   :   1.0   Min.   :1.000   2013-01-02:   1115   Min.   :    0
##  1st Qu.: 280.0   1st Qu.:2.000   2013-01-03:   1115   1st Qu.: 3727
##  Median : 558.0   Median :4.000   2013-01-04:   1115   Median : 5744
##  Mean   : 558.4   Mean   :3.998   2013-01-05:   1115   Mean   : 5774
##  3rd Qu.: 838.0   3rd Qu.:6.000   2013-01-06:   1115   3rd Qu.: 7856
##  Max.   :1115.0   Max.   :7.000   2013-01-07:   1115   Max.   :41551
##                                   (Other)   :1010519
##    Customers          Open            Promo        StateHoliday
##  Min.   :   0.0   Min.   :0.0000   Min.   :0.0000   0:986159
##  1st Qu.: 405.0   1st Qu.:1.0000   1st Qu.:0.0000   a: 20260
##  Median : 609.0   Median :1.0000   Median :0.0000   b:  6690
##  Mean   : 633.1   Mean   :0.8301   Mean   :0.3815   c:  4100
##  3rd Qu.: 837.0   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :7388.0   Max.   :1.0000   Max.   :1.0000
##
##  SchoolHoliday
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1786
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
str(train)
```

```
## Classes 'data.table' and 'data.frame':   1017209 obs. of  9 variables:
##  $ Store        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DayOfWeek    : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Date         : Factor w/ 942 levels "2013-01-01","2013-01-02",..: 942 942 942 942 942 942 942 942
##  $ Sales        : int  5263 6064 8314 13995 4822 5651 15344 8492 8565 7185 ...
##  $ Customers    : int  555 625 821 1498 559 589 1414 833 687 681 ...
##  $ Open         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Promo        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ StateHoliday : Factor w/ 4 levels "0","a","b","c": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SchoolHoliday: int  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
sum(is.na(train))
```

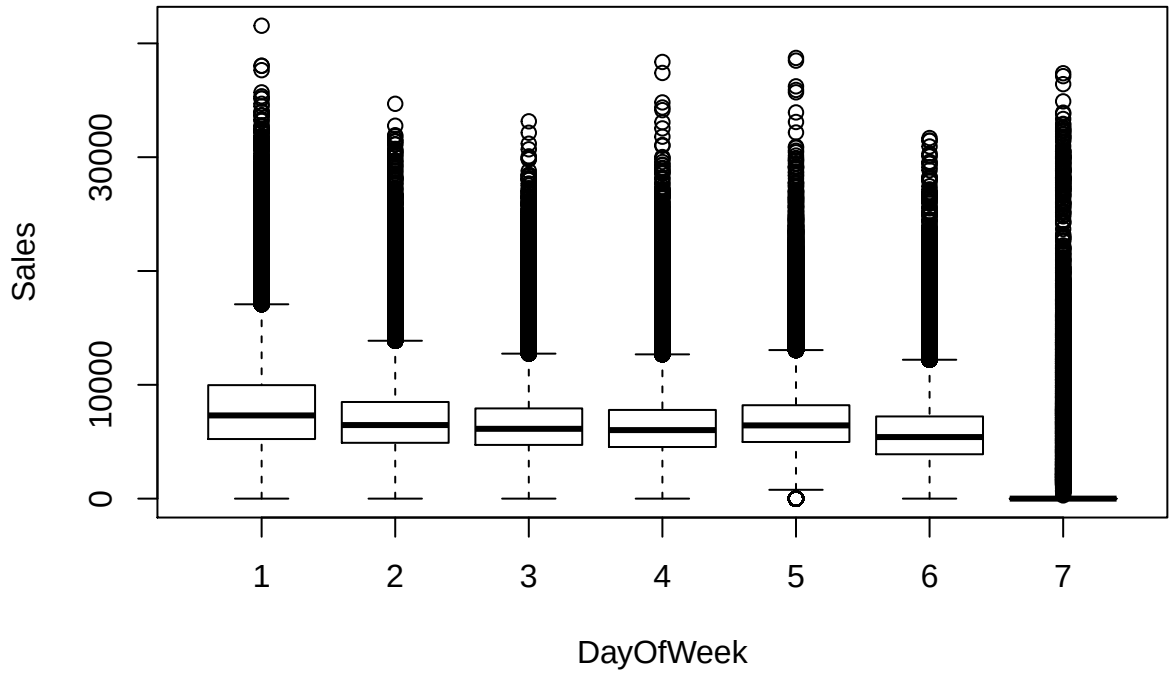
```
## [1] 0
```

```
train$Date <- as.Date(train$Date)
str(train)
```

```
## Classes 'data.table' and 'data.frame':   1017209 obs. of  9 variables:
##  $ Store        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ DayOfWeek    : int  5 5 5 5 5 5 5 5 5 5 ...
```

```
##  $ Date          : Date, format: "2015-07-31" "2015-07-31" ...
##  $ Sales         : int  5263 6064 8314 13995 4822 5651 15344 8492 8565 7185 ...
##  $ Customers     : int  555 625 821 1498 559 589 1414 833 687 681 ...
##  $ Open          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Promo         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ StateHoliday : Factor w/ 4 levels "0","a","b","c": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SchoolHoliday: int  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
boxplot(Sales~DayOfWeek,data = train)
```




```
A<-boxplot(Sales~DayOfWeek,data = train)
```

```
mytable <- A$stats
colnames(mytable)<-A$names
rownames(mytable)<-c('min','lower quartile','median','upper quartile','max')
mytable
```

```
##                   1     2     3     4     5     6 7
## min               0     0     0     0   775     0 0
## lower quartile 5235  4904  4718  4536  4975  3899 0
## median         7310  6463  6133  6020  6434  5410 0
## upper quartile 9972  8491  7926  7792  8206  7220 0
## max           17076 13870 12738 12676 13052 12201 0
```

```
## attr(,"class")
##         1
## "integer"
```

We can observe all the interquartile range for “SUNDAY” contains “0”

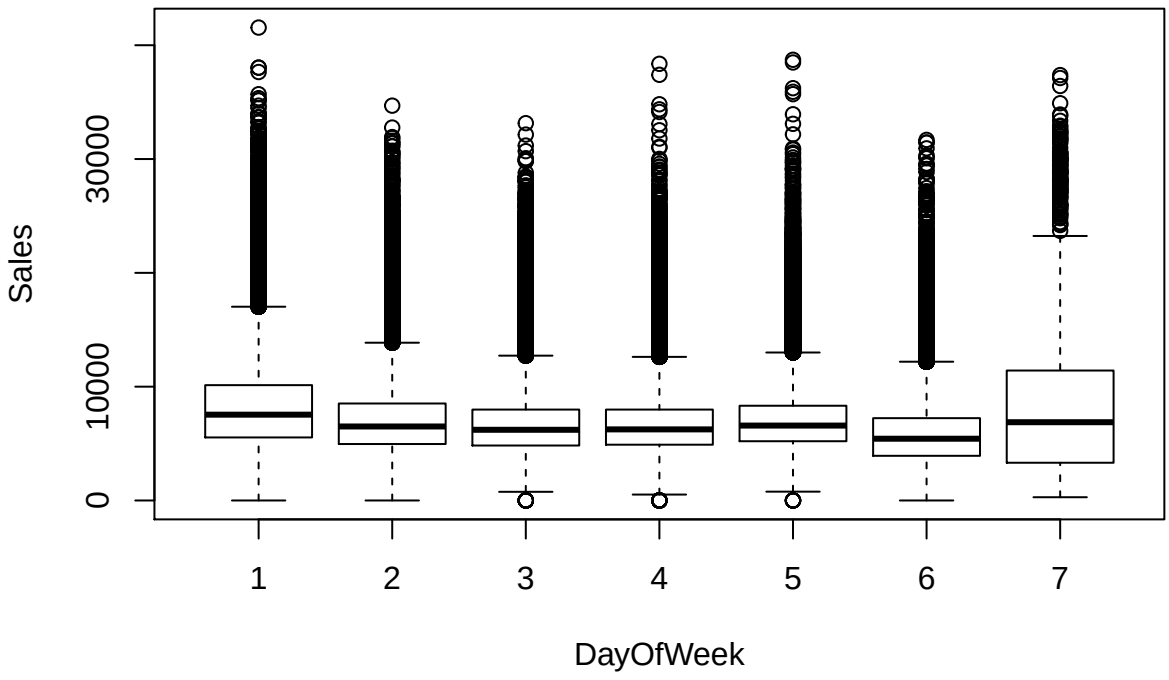
Now we will filter to see data points with Closed Store and 0 Sales.

```
x <- train
y <- filter(x,x$Sales==0)
x <- filter(y,y$Open==0)
str(x)
str(train)
y <- filter(x,x$DayOfWeek==7)
str(y)
```

we can see Major sunday 0 sales are from Store Closures

Next, Analyse Store open data points only

```
x <-train
x <- filter(x,x$Open==1)
boxplot(Sales~DayOfWeek,data = x)
```



```
A <- boxplot(Sales~DayOfWeek,data = x)
```

```
mytable <- A$stats
colnames(mytable)<-A$names
rownames(mytable)<-c('min','lower quartile','median','upper quartile','max')
mytable
```

```
##                    1     2     3     4     5     6     7
## min                0     0   760   520   775     0   286
## lower quartile  5538  4960  4829  4900  5205  3925  3314
## median          7539  6502  6210  6246  6580  5425  6876
## upper quartile 10133  8521  7987  7987  8324  7232 11418
## max            17025 13861 12724 12617 13002 12192 23240
## attr(,"class")
##         1
## "integer"
```

We can Observe the stores opened on sundays have higher sales compared to other days.