Retail Analysis with Walmart Data

2022-04-05

Uploading and Reading the Dataset

```
Wdf = read.csv("~/SimpliLearn Data Analytics/Chapter 4 Data Science with R/Walmart_Store_sales.csv")
```

Data Description

```
View(Wdf)
str(Wdf)
```

```
## 'data.frame':    6435 obs. of  8 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : chr  "05-02-2010" "12-02-2010" "19-02-2010" "26-02-2010"
...
##  $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
##  $ Holiday_Flag: int  0 1 0 0 0 0 0 0 0 0 ...
##  $ 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 ...
##  $ CPI         : num  211 211 211 211 211 ...
##  $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
```

```
head(Wdf)
```

```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price
CPI
## 1     1 05-02-2010      1643691            0       42.31      2.572
211.0964
## 2     1 12-02-2010      1641957            1       38.51      2.548
211.2422
## 3     1 19-02-2010      1611968            0       39.93      2.514
211.2891
## 4     1 26-02-2010      1409728            0       46.63      2.561
211.3196
## 5     1 05-03-2010      1554807            0       46.50      2.625
211.3501
## 6     1 12-03-2010      1439542            0       57.79      2.667
211.3806
##   Unemployment
## 1        8.106
## 2        8.106
## 3        8.106
## 4        8.106
## 5        8.106
## 6        8.106
```

```
class(Wdf)
```

```
## [1] "data.frame"
```

Descriptive Statistics

```
summary(Wdf)
```

```
##      Store        Date            Weekly_Sales      Holiday_Flag     
##  Min.   : 1   Length:6435        Min.   : 209986   Min.   :0.00000  
##  1st Qu.:12   Class :character   1st Qu.: 553350   1st Qu.:0.00000  
##  Median :23   Mode  :character   Median : 960746   Median :0.00000  
##  Mean   :23                      Mean   :1046965   Mean   :0.06993  
##  3rd Qu.:34                      3rd Qu.:1420159   3rd Qu.:0.00000  
##  Max.   :45                      Max.   :3818686   Max.   :1.00000  
##   Temperature       Fuel_Price         CPI         Unemployment    
##  Min.   : -2.06   Min.   :2.472   Min.   :126.1   Min.   : 3.879  
##  1st Qu.: 47.46   1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891  
##  Median : 62.67   Median :3.445   Median :182.6   Median : 7.874  
##  Mean   : 60.66   Mean   :3.359   Mean   :171.6   Mean   : 7.999  
##  3rd Qu.: 74.94   3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622  
##  Max.   :100.14   Max.   :4.468   Max.   :227.2   Max.   :14.313
```

Checking NA values

```
colSums(is.na(Wdf))
```

```
##        Store         Date Weekly_Sales Holiday_Flag  Temperature
Fuel_Price
##            0            0            0            0            0
0
##          CPI Unemployment
##            0            0
```

No null values in the dataset

Loading all the needed libraries

```
library("dplyr")
```

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

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

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

```
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
```

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

library("zoo")

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

#Data Visualization
library("grid")
library("vcd")

## Warning: package 'vcd' was built under R version 4.1.3

library("ggplot2")

## Warning: package 'ggplot2' was built under R version 4.1.3

library("plotly")

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

Converting Date column into Date format also, converting Store and Holiday Flag column into Factor

```
Wdf$Date = as.Date(Wdf$Date,format="%d-%m-%Y")
Wdf$Store = as.factor(Wdf$Store)
Wdf$Holiday_Flag = as.factor(Wdf$Holiday_Flag)
```

Q1- which store has max sales?

```
store_sales = aggregate(Weekly_Sales~Store,data=Wdf, sum)
store_sales

##    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
## 12    12    144287230
## 13    13    286517704
## 14    14    288999911
## 15    15     89133684
## 16    16     74252425
## 17    17    127782139
## 18    18    155114734
## 19    19    206634862
## 20    20    301397792
## 21    21    108117879
## 22    22    147075649
## 23    23    198750618
## 24    24    194016021
## 25    25    101061179
## 26    26    143416394
## 27    27    253855917
## 28    28    189263681
## 29    29     77141554
## 30    30     62716885
## 31    31    199613906
## 32    32    166819246
## 33    33     37160222
## 34    34    138249763
## 35    35    131520672
## 36    36     53412215
## 37    37     74202740
## 38    38     55159626
## 39    39    207445542
## 40    40    137870310
## 41    41    181341935
## 42    42     79565752
## 43    43     90565435
## 44    44     43293088
## 45    45    112395341
```

```
which.max(store_sales$Weekly_Sales)
```

```
## [1] 20
```

```
store_sales[20,]
```

```
##    Store Weekly_Sales
## 20    20    301397792
```

A-Store 20 has highest sale, sale value of 301397792

Q2- Which store has maximum standard deviation i.e., the sales vary a lot?

```
store_sales$sales_mean = aggregate(Weekly_Sales~Store,data=Wdf,
mean)$Weekly_Sales
store_sales$sales_sd = aggregate(Weekly_Sales~Store,data=Wdf,
sd)$Weekly_Sales
str(store_sales)
```

```
## 'data.frame':    45 obs. of  4 variables:
##  $ Store       : Factor w/ 45 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9
10 ...
##  $ Weekly_Sales: num  2.22e+08 2.75e+08 5.76e+07 3.00e+08 4.55e+07 ...
##  $ sales_mean  : num  1555264 1925751 402704 2094713 318012 ...
##  $ sales_sd    : num  155981 237684 46320 266201 37738 ...
```

```
arrange(store_sales, desc(sales_sd))
```

```
##    Store Weekly_Sales sales_mean  sales_sd
## 1     14    288999911  2020978.4 317569.95
## 2     10    271617714  1899424.6 302262.06
## 3     20    301397792  2107676.9 275900.56
## 4      4    299543953  2094713.0 266201.44
## 5     13    286517704  2003620.3 265507.00
## 6     23    198750618  1389864.5 249788.04
## 7     27    253855917  1775216.2 239930.14
## 8      2    275382441  1925751.3 237683.69
## 9     39    207445542  1450668.1 217466.45
## 10     6    223756131  1564728.2 212525.86
## 11    35    131520672   919725.0 211243.46
## 12    19    206634862  1444999.0 191722.64
## 13    41    181341935  1268125.4 187907.16
## 14    28    189263681  1323522.2 181758.97
## 15    18    155114734  1084718.4 176641.51
## 16    24    194016021  1356755.4 167745.68
## 17    11    193962787  1356383.1 165833.89
## 18    22    147075649  1028501.0 161251.35
## 19     1    222402809  1555264.4 155980.77
## 20    12    144287230  1009001.6 139166.87
## 21    32    166819246  1166568.2 138017.25
## 22    45    112395341   785981.4 130168.53
## 23    21    108117879   756069.1 128752.81
## 24    31    199613906  1395901.4 125855.94
## 25    15     89133684   623312.5 120538.65
## 26    40    137870310   964128.0 119002.11
## 27    25    101061179   706721.5 112976.79
## 28     7     81598275   570617.3 112585.47
## 29    17    127782139   893581.4 112162.94
```

```
## 30    26    143416394  1002911.8 110431.29
## 31     8    129951181   908749.5 106280.83
## 32    34    138249763   966781.6 104630.16
## 33    29     77141554   539451.4  99120.14
## 34    16     74252425   519247.7  85769.68
## 35     9     77789219   543980.6  69028.67
## 36    36     53412215   373512.0  60725.17
## 37    42     79565752   556403.9  50262.93
## 38     3     57586735   402704.4  46319.63
## 39    38     55159626   385731.7  42768.17
## 40    43     90565435   633324.7  40598.41
## 41     5     45475689   318011.8  37737.97
## 42    44     43293088   302748.9  24762.83
## 43    33     37160222   259861.7  24132.93
## 44    30     62716885   438579.6  22809.67
## 45    37     74202740   518900.3  21837.46
```

A-Store 14 has highest standard deviation = 317569.95

Q3- Which store/s has good quarterly growth rate in Q3'2012?

```
# creating copy of Wdf
Wdf2 = Wdf

Wdf2$month_Year = substr(Wdf2$Date, 1, 7)
Q3_2012 = filter(Wdf2,month_Year == "2012-07" | month_Year== "2012-08" |
month_Year== "2012-09")
Q2_2012 = filter(Wdf2,month_Year == "2012-04" | month_Year== "2012-05" |
month_Year== "2012-06")

#Aggregating sales by store for Q3-2012
Q3_2012_Sales = summarise(group_by(Q3_2012,Store),sum(Weekly_Sales))

#Aggregating sales by store for Q3-2012
Q2_2012_Sales = summarise(group_by(Q2_2012,Store),sum(Weekly_Sales))

Q3_2012_Growthrate = merge ( Q2_2012_Sales , Q3_2012_Sales , by = 'Store')
Q3_2012_Growthrate = mutate(Q3_2012_Growthrate, Growth_Rate =
((Q3_2012_Sales$`sum(Weekly_Sales)` - Q2_2012_Sales$'sum(Weekly_Sales)')*100)
/ Q2_2012_Sales$'sum(Weekly_Sales)')
gr = arrange(Q3_2012_Growthrate, desc(Growth_Rate))
View(gr)
```

A- Store 15 has highest growth rate in Q3 2012

Q4- 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

```
SuperBowl = as.Date(c("2010-02-12","2011-02-11","2012-02-10","2013-02-08"))
LabourDay = as.Date(c("2010-09-10", "2011-09-09", "2012-09-07", "2013-09-
06"))
Thanksgiving = as.Date(c("2010-11-26", "2011-11-25", "2012-11-23", "2013-11-
```

```
29"))
Christmas = as.Date(c("2010-12-31", "2011-12-30", "2012-12-28", "2013-12-
27"))

Walmart_Holiday = Wdf[1:3]

Walmart_Holiday$hflag = ifelse(Walmart_Holiday$Date %in% SuperBowl, "SB",
ifelse(Walmart_Holiday$Date %in% LabourDay, "LD", ifelse(Walmart_Holiday$Date
%in% Thanksgiving, "TG", ifelse(Walmart_Holiday$Date %in% Christmas,
"CH","None"))))
aggregate(Weekly_Sales~hflag,data=Walmart_Holiday, mean)
```

```
##   hflag Weekly_Sales
## 1    CH     960833.1
## 2    LD    1042427.3
## 3  None    1041256.4
## 4    SB    1079128.0
## 5    TG    1471273.4
```

A- Thanks giving have highest sales than mean. Mean sales in non-holiday season for all stores together is 1041256.4 and except Christmas all holidays have higher sales than average sale in non-holiday sale.

Q5- Provide a monthly and semester view of sales in units and give insights

```
semester_view = Wdf
View(semester_view)
semester_view_month_year = transform(semester_view,Year_Sale =
as.numeric(format(Date,"%Y")),Month_Sale = as.numeric(format(Date,"%m")))
View(semester_view_month_year)

Summarized_View =
aggregate(Weekly_Sales~Month_Sale+Year_Sale,semester_view_month_year,sum)
View(Summarized_View)

Insights = arrange(Summarized_View,desc(Weekly_Sales))
View(Insights)
```

A- The sales are highest in December and Lowest in January and are higher in second semester of every year

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.

```
library(dplyr)
semester_viewtore1 = select(filter(Wdf, Store==1),-1) ## Filtering data for
Store 1 for building linear model
```

```
View(semester_viewtore1)
str(semester_viewtore1)
```

```
## 'data.frame':    143 obs. of  7 variables:
##  $ Date        : Date, format: "2010-02-05" "2010-02-12" ...
##  $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
##  $ Holiday_Flag: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ CPI         : num  211 211 211 211 211 ...
##  $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
```

```
## Linear Model
Wdf_lm = lm(Weekly_Sales ~ Holiday_Flag + Temperature + Fuel_Price+ CPI +
Unemployment , semester_viewtore1)
summary(Wdf_lm)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Holiday_Flag + Temperature + Fuel_Price +
##     CPI + Unemployment, data = semester_viewtore1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -305166  -78247  -18260   53643  854412
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2427856    1752958  -1.385   0.1683
## Holiday_Flag1    89376      49338   1.811   0.0723 .
## Temperature      -2160        922  -2.343   0.0206 *
## Fuel_Price      -24337      47335  -0.514   0.6080
## CPI              16632       6786   2.451   0.0155 *
## Unemployment     80209      58727   1.366   0.1742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146500 on 137 degrees of freedom
## Multiple R-squared:  0.1495, Adjusted R-squared:  0.1184
## F-statistic: 4.815 on 5 and 137 DF,  p-value: 0.0004359
```

```
## Drop most insignificant variables- Unemployment and Fuel Price (p value =
60.80%)
Wdf_lm1 = lm(Weekly_Sales ~ Holiday_Flag + Temperature ++ CPI ,
semester_viewtore1)
summary(Wdf_lm1)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Holiday_Flag + Temperature + +CPI,
##     data = semester_viewtore1)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -300742  -80390  -11862   57057  842876
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -269917.6   610415.6  -0.442  0.65904
## Holiday_Flag1   96246.0    48996.5   1.964  0.05148 .
## Temperature     -2423.3      885.8  -2.736  0.00704 **
## CPI              9185.2     2843.5   3.230  0.00154 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146400 on 139 degrees of freedom
## Multiple R-squared:  0.1378, Adjusted R-squared:  0.1192
## F-statistic: 7.407 on 3 and 139 DF,  p-value: 0.0001222
```

```
## Drop most insignificant variable Holiday_Flag1 (p value = 5.15%)
Wdf_lm3 = lm(Weekly_Sales ~ Temperature + CPI , semester_viewtore1)
summary(Wdf_lm3)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Temperature + CPI, data = semester_viewtore1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -312205  -85704   -9198   57222  830489
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -233190     616327  -0.378  0.70574
## Temperature    -2769        877  -3.157  0.00195 **
## CPI             9156       2872   3.187  0.00177 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147900 on 140 degrees of freedom
## Multiple R-squared:  0.1139, Adjusted R-squared:  0.1012
## F-statistic: 8.998 on 2 and 140 DF,  p-value: 0.0002107
```

We can say only CPI and Temperature are the Variables we can use to build a model as other variables are insignificant

Model can be further improvised by-

1-Considering all Stores data for prediction

2-Using Advanced models like Decision Trees, Random Forest

3-Using K cross validation techniques for Sampling data

Change dates into days by creating new variable

```
Data2 = Wdf
Data2$Weekdays = weekdays(Data2$Date)
View(Data2)
```