Walmart Data Exploration

Terrence Josiah

6/18/2022

1. Reading the CSV file

```
getwd()
```

```
## [1] "C:/Users/terre/Desktop/Personal Projects/Retail analysis with Walmart
sales data/Programming files"
```

```
setwd("C:/Users/terre/OneDrive/Documents")

walmart = read.csv("WALMART_SALES_DATA.csv", header = TRUE)

head(walmart)
```

```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      C
PI
## 1     1 05-02-2010      1643691            0       42.31      2.572 211.09
64
## 2     1 12-02-2010      1641957            1       38.51      2.548 211.24
22
## 3     1 19-02-2010      1611968            0       39.93      2.514 211.28
91
## 4     1 26-02-2010      1409728            0       46.63      2.561 211.31
96
## 5     1 05-03-2010      1554807            0       46.50      2.625 211.35
01
## 6     1 12-03-2010      1439542            0       57.79      2.667 211.38
06
##   Unemployment
## 1        8.106
## 2        8.106
## 3        8.106
## 4        8.106
## 5        8.106
## 6        8.106
```

```
str(walmart)
```

```
## '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 ...
```

2. Check if there's any null value in the data

```
is.null(walmart)

## [1] FALSE
```

FINDINGS: There are no null values in the data.

3. Sum of the Weekly Sales group by the Store, find those with the highest sales

- Formula: aggregate(cbind(x$Frequency$, xMetric2, x$Metric3) ..., by(), FUN = sum)

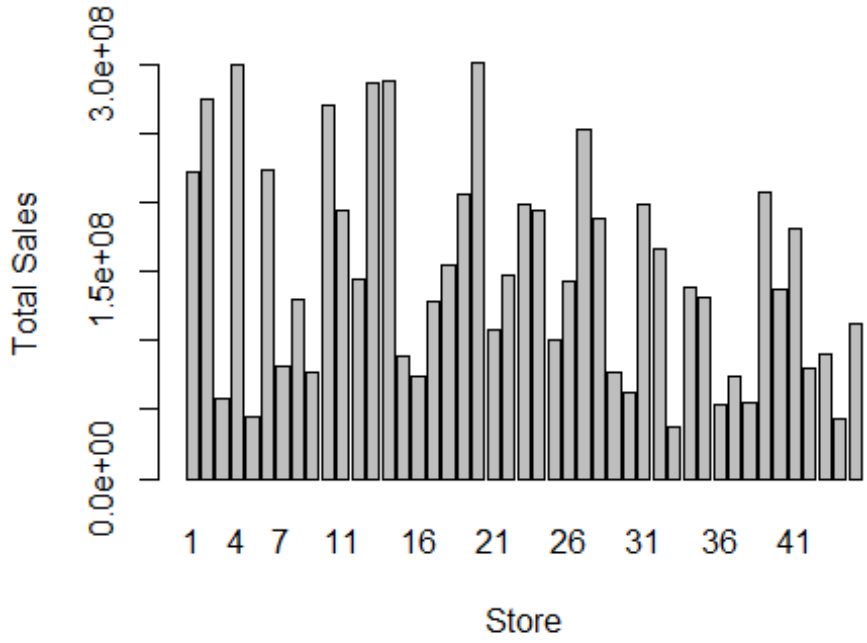
```
sum_WS = aggregate(walmart$Weekly_Sales, by = list(Store = walmart$Store), FU
N = sum)
sum_WS

##    Store         x
## 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
```

- Visualize using barplot

```
barplot(sum_WS$x, names.arg = sum_WS$Store, ylab = "Total Sales", xlab = "Sto
re")
```




```
max(sum_WS$x)

## [1] 301397792
```

- Check which row has the max Weekly_Sales

```
which(sum_WS$x == max(sum_WS$x), arr.ind = TRUE)
```

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

FINDINGS: Store 20 is has the highest total Weekly Sales of 301,397,792 Dollar.

4. Store with the highest Standard Deviation

```
sd_WS = aggregate(walmart$Weekly_Sales, by = list(Store = walmart$Store), FUN
= sd)
sd_WS
```

```
##    Store         x
## 1      1 155980.77
## 2      2 237683.69
## 3      3  46319.63
## 4      4 266201.44
## 5      5  37737.97
## 6      6 212525.86
## 7      7 112585.47
## 8      8 106280.83
## 9      9  69028.67
## 10    10 302262.06
## 11    11 165833.89
## 12    12 139166.87
## 13    13 265507.00
## 14    14 317569.95
## 15    15 120538.65
## 16    16  85769.68
## 17    17 112162.94
## 18    18 176641.51
## 19    19 191722.64
## 20    20 275900.56
## 21    21 128752.81
## 22    22 161251.35
## 23    23 249788.04
## 24    24 167745.68
## 25    25 112976.79
## 26    26 110431.29
## 27    27 239930.14
## 28    28 181758.97
## 29    29  99120.14
## 30    30  22809.67
## 31    31 125855.94
## 32    32 138017.25
## 33    33  24132.93
## 34    34 104630.16
## 35    35 211243.46
## 36    36  60725.17
```

```
## 37    37  21837.46
## 38    38  42768.17
## 39    39 217466.45
## 40    40 119002.11
## 41    41 187907.16
## 42    42  50262.93
## 43    43  40598.41
## 44    44  24762.83
## 45    45 130168.53
```

```
max(sd_WS$x)
```

```
## [1] 317569.9
```

- Check which row has the max SD

```
which(sd_WS$x == max(sd_WS$x), arr.ind = TRUE)
```

```
## [1] 14
```

FINDINGS: Store 14 has the highest Standard Deviation of 317569.9, which indicates that The sales of Store 14 vary a lot.

5. Coefficient of mean to standard deviation (Coefficient of Variation)

```
mean_WS = aggregate(walmart$Weekly_Sales, by = list(Store = walmart$Store), F
UN = mean)
mean_WS
```

```
##    Store         x
## 1      1 1555264.4
## 2      2 1925751.3
## 3      3  402704.4
## 4      4 2094713.0
## 5      5  318011.8
## 6      6 1564728.2
## 7      7  570617.3
## 8      8  908749.5
## 9      9  543980.6
## 10    10 1899424.6
## 11    11 1356383.1
## 12    12 1009001.6
## 13    13 2003620.3
## 14    14 2020978.4
## 15    15  623312.5
## 16    16  519247.7
## 17    17  893581.4
## 18    18 1084718.4
## 19    19 1444999.0
## 20    20 2107676.9
## 21    21  756069.1
```

```
## 22    22 1028501.0
## 23    23 1389864.5
## 24    24 1356755.4
## 25    25  706721.5
## 26    26 1002911.8
## 27    27 1775216.2
## 28    28 1323522.2
## 29    29  539451.4
## 30    30  438579.6
## 31    31 1395901.4
## 32    32 1166568.2
## 33    33  259861.7
## 34    34  966781.6
## 35    35  919725.0
## 36    36  373512.0
## 37    37  518900.3
## 38    38  385731.7
## 39    39 1450668.1
## 40    40  964128.0
## 41    41 1268125.4
## 42    42  556403.9
## 43    43  633324.7
## 44    44  302748.9
## 45    45  785981.4
```

- Coefficient of Variation can be calculated by dividing standard deviation with the mean of the data

```
CoV = sd_WS/mean_WS
CoV
```

```
##    Store          x
## 1      1 0.10029212
## 2      1 0.12342388
## 3      1 0.11502141
## 4      1 0.12708254
## 5      1 0.11866844
## 6      1 0.13582286
## 7      1 0.19730469
## 8      1 0.11695283
## 9      1 0.12689547
## 10     1 0.15913349
## 11     1 0.12226183
## 12     1 0.13792532
## 13     1 0.13251363
## 14     1 0.15713674
## 15     1 0.19338399
## 16     1 0.16518065
## 17     1 0.12552067
## 18     1 0.16284550
## 19     1 0.13268012
```

```
## 20     1 0.13090269
## 21     1 0.17029239
## 22     1 0.15678288
## 23     1 0.17972115
## 24     1 0.12363738
## 25     1 0.15986040
## 26     1 0.11011066
## 27     1 0.13515544
## 28     1 0.13732974
## 29     1 0.18374247
## 30     1 0.05200804
## 31     1 0.09016105
## 32     1 0.11831049
## 33     1 0.09286835
## 34     1 0.10822524
## 35     1 0.22968111
## 36     1 0.16257891
## 37     1 0.04208412
## 38     1 0.11087545
## 39     1 0.14990779
## 40     1 0.12342978
## 41     1 0.14817711
## 42     1 0.09033533
## 43     1 0.06410363
## 44     1 0.08179331
## 45     1 0.16561273
```

```
max(CoV$x)
```

```
## [1] 0.2296811
```

- Check which row has the max Coefficient of Variation

```
which(CoV$x == max(CoV$x), arr.ind = TRUE)
```

```
## [1] 35
```

FINDINGS: Store 35 has the highest Coefficient of Variation, indicating that Store 35 has the highest variability around the mean than other stores.

6. Which store has a good Quarterly growth rate in Q3'2012 (July, August, September)

- As the date column is character, convert it into date
- %d = day
- %m = numeric month
- %b = abbreviated non-numeric month (Aug)

- %B = full non-numeric month (August)
- %y = 2 digit numeric year (08)
- %Y = full numeric year (2008)

```
walmart$Date = as.Date(walmart$Date, format = "%d-%m-%Y")
str(walmart)
```

```
## 'data.frame':    6435 obs. of  8 variables:
##  $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Date        : Date, format: "2010-02-05" "2010-02-12" ...
##  $ 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 ...
```

- Make new month and year columns

```
walmart$Day = format(walmart$Date, "%d")
walmart$Month = format(walmart$Date, "%m")
walmart$Year = format(walmart$Date, "%Y")
head(walmart)
```

```
##   Store       Date Weekly_Sales Holiday_Flag Temperature Fuel_Price      C
PI
## 1     1 2010-02-05      1643691            0       42.31      2.572 211.09
64
## 2     1 2010-02-12      1641957            1       38.51      2.548 211.24
22
## 3     1 2010-02-19      1611968            0       39.93      2.514 211.28
91
## 4     1 2010-02-26      1409728            0       46.63      2.561 211.31
96
## 5     1 2010-03-05      1554807            0       46.50      2.625 211.35
01
## 6     1 2010-03-12      1439542            0       57.79      2.667 211.38
06
##   Unemployment Day Month Year
## 1        8.106  05    02 2010
## 2        8.106  12    02 2010
## 3        8.106  19    02 2010
## 4        8.106  26    02 2010
## 5        8.106  05    03 2010
## 6        8.106  12    03 2010
```

We will move to SQL

- Export the table we want

```
write.table(walmart, file = "WalmartData.csv", row.names = FALSE, sep = ",")
```