Walmart Data Challenge

William Eerdmans

January 31, 2017

Read in the total data and subset the 1MM row data file

```
#read in sales_cust data
sales_cust_tot <- read.csv("sales_cust.csv")
sales_cust_tot <- tbl_df(sales_cust_tot)

#read in store
store_comp <- read.csv("store.csv")
store_comp <- tbl_df(store_comp)
```

Check for missing values, inconsistencies, and column data types

It can be seen in sales_cust that there are 59 NA values in Open and 103 NA values in SchoolHoliday

When looking at Open, NAs are only during 7/5/15 & 7/6/15. However, some of the stores have no sales, whereas some have sales

What can also be seen is that store 384-398 repeat for these NA values

When observing other Day = 7 dates, it was found that they too were all 0. Thus, those NAs can be determined to be Open = 0, whereas, the NAs that actually have sales will be assumed to be Open = 1.

When looking at the School Holiday NA values and observing other values between StateHoliday and SchoolHoliday, I didn't see any repeatable pattern between them, beyond that when there are certain StateHolidays, SchoolHolidays may occur. However, I believe the best course of action would be to delete these rows due to 103 values out of ~1MM is negligible. I created a new dataframe in order to seperate the prior values that were not NA in sc_open_NAs.

```
#Summarize the data for initial look
head(sales_cust_tot)
```

```
## # A tibble: 6 × 9
##   Store DayOfWeek    Date Sales Customers  Open Promo StateHoliday
##   <int>     <int>  <fctr> <int>     <int> <int> <int>       <fctr>
## 1     1         5 7/31/15  5263       555     1     1            0
## 2     2         5 7/31/15  6064       625     1     1            0
## 3     3         5 7/31/15  8314       821     1     1            0
## 4     4         5 7/31/15 13995      1498     1     1            0
## 5     5         5 7/31/15  4822       559     1     1            0
## 6     6         5 7/31/15  5651       589     1     1            0
## # ... with 1 more variables: SchoolHoliday <int>
```

```r
summary(sales_cust_tot)
```

```
##      Store          DayOfWeek          Date             Sales
##  Min.   :   1.0   Min.   :1.000   1/1/14 :   1115   Min.   :    0
##  1st Qu.: 280.0   1st Qu.:2.000   1/1/15 :   1115   1st Qu.: 3727
##  Median : 558.0   Median :4.000   1/10/13:   1115   Median : 5744
##  Mean   : 558.4   Mean   :3.998   1/10/14:   1115   Mean   : 5774
##  3rd Qu.: 838.0   3rd Qu.:6.000   1/10/15:   1115   3rd Qu.: 7856
##  Max.   :1115.0   Max.   :7.000   1/11/13:   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
##                   NA's   :59
##  SchoolHoliday
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1786
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :103
```

```r
#visually observe the data where there are NAs
#Start with Open
sc_open_NAs <- sales_cust_tot %>% filter(is.na(Open))

#Take a look at the values
print(sc_open_NAs, n=59)
```

```
## # A tibble: 59 × 9
##    Store DayOfWeek   Date Sales Customers  Open Promo StateHoliday
##    <int>     <int> <fctr> <int>     <int> <int> <int>       <fctr>
## 1    384         1 7/6/15 10337      1214    NA     0            0
## 2    385         1 7/6/15  6951       620    NA     0            0
## 3    386         1 7/6/15  6479       545    NA     0            0
## 4    387         1 7/6/15  8817      1172    NA     0            0
## 5    388         1 7/6/15  9596      1032    NA     0            0
## 6    389         1 7/6/15 11928      1407    NA     0            0
## 7    390         1 7/6/15 11380      1083    NA     0            0
## 8    391         1 7/6/15  5293       668    NA     0            0
## 9    392         1 7/6/15  7580       721    NA     0            0
## 10   393         1 7/6/15  5780       568    NA     0            0
## 11   394         1 7/6/15  7824       673    NA     0            0
## 12   395         1 7/6/15  4153       505    NA     0            0
## 13   396         1 7/6/15  9726      1038    NA     0            0
## 14   397         1 7/6/15  5947       842    NA     0            0
## 15   398         1 7/6/15  4426       577    NA     0            0
## 16   384         7 7/6/14     0         0    NA     0            0
## 17   386         7 7/6/14     0         0    NA     0            0
```

```
## 18   387         7 7/6/14     0         0    NA     0            0
## 19   388         7 7/6/14     0         0    NA     0            0
## 20   389         7 7/6/14     0         0    NA     0            0
## 21   390         7 7/6/14     0         0    NA     0            0
## 22   391         7 7/6/14     0         0    NA     0            0
## 23   392         7 7/6/14     0         0    NA     0            0
## 24   393         7 7/6/14     0         0    NA     0            0
## 25   394         7 7/6/14     0         0    NA     0            0
## 26   395         7 7/6/14     0         0    NA     0            0
## 27   396         7 7/6/14     0         0    NA     0            0
## 28   397         7 7/6/14     0         0    NA     0            0
## 29   398         7 7/6/14     0         0    NA     0            0
## 30   384         6 7/6/13  4369       495    NA     0            0
## 31   385         6 7/6/13  6634       625    NA     0            0
## 32   386         6 7/6/13  6860       619    NA     0            0
## 33   387         6 7/6/13  5880       938    NA     0            0
## 34   388         6 7/6/13  7409       866    NA     0            0
## 35   389         6 7/6/13  6537       861    NA     0            0
## 36   390         6 7/6/13  9231       952    NA     0            0
## 37   391         6 7/6/13  2677       389    NA     0            0
## 38   392         6 7/6/13  4035       431    NA     0            0
## 39   393         6 7/6/13  5879       533    NA     0            0
## 40   394         6 7/6/13  7993       728    NA     0            0
## 41   395         6 7/6/13  2536       365    NA     0            0
## 42   396         6 7/6/13  4889       554    NA     0            0
## 43   397         6 7/6/13  3993       588    NA     0            0
## 44   398         6 7/6/13  4484       492    NA     0            0
## 45   384         5 7/5/13  7874       886    NA     1            0
## 46   385         5 7/5/13  8277       740    NA     1            0
## 47   386         5 7/5/13  9652       798    NA     1            0
## 48   387         5 7/5/13  9207      1234    NA     1            0
## 49   388         5 7/5/13 11064      1194    NA     1            0
## 50   389         5 7/5/13  9751      1239    NA     1            0
## 51   390         5 7/5/13 11856      1102    NA     1            0
## 52   391         5 7/5/13  6099       831    NA     1            0
## 53   392         5 7/5/13  7515       755    NA     1            0
## 54   393         5 7/5/13  6146       620    NA     1            0
## 55   394         5 7/5/13 10338       786    NA     1            0
## 56   395         5 7/5/13  5274       635    NA     1            0
## 57   396         5 7/5/13 10749      1092    NA     1            0
## 58   397         5 7/5/13  6447       851    NA     1            0
## 59   398         5 7/5/13  6271       668    NA     1            0
## # ... with 1 more variables: SchoolHoliday <int>
```

```r
#For the Open NAs, impute 1's for the Open column and for those where the Day of the week is 7, impute
sales_cust_tot[which(is.na(sales_cust_tot$Open) & sales_cust_tot$DayOfWeek == 7),]$Open <- 0
sales_cust_tot[which(is.na(sales_cust_tot$Open) & sales_cust_tot$Sales > 0),]$Open <- 1

#Now turn attention to the 103 NAs in SchoolHoliday
sc_holiday_NAs <- sales_cust_tot %>% filter(is.na(SchoolHoliday))

#Take a look at the values
print(sc_holiday_NAs, n=59)
```

```
## # A tibble: 103 × 9
```

```
##    Store DayOfWeek     Date Sales Customers  Open Promo StateHoliday
##    <int>     <int>   <fctr> <int>     <int> <dbl> <int>       <fctr>
## 1    398         6  7/25/15  5044       538     1     0            0
## 2    384         5  7/24/15  7459       869     1     0            0
## 3    385         5  7/24/15  5328       560     1     0            0
## 4    386         5  7/24/15  5582       489     1     0            0
## 5    387         5  7/24/15  7513       963     1     0            0
## 6    388         5  7/24/15  7672       913     1     0            0
## 7    389         5  7/24/15  9265      1124     1     0            0
## 8    390         5  7/24/15  9514       899     1     0            0
## 9    391         5  7/24/15  4536       606     1     0            0
## 10   392         5  7/24/15  5718       581     1     0            0
## 11   393         5  7/24/15  5081       515     1     0            0
## 12   394         5  7/24/15  7615       669     1     0            0
## 13   395         5  7/24/15  3177       446     1     0            0
## 14   396         5  7/24/15  7972       862     1     0            0
## 15   397         5  7/24/15  4512       668     1     0            0
## 16   398         5  7/24/15  4427       579     1     0            0
## 17   384         4  7/23/15  8856      1050     1     0            0
## 18   385         4  7/23/15  5549       519     1     0            0
## 19   386         4  7/23/15  4766       450     1     0            0
## 20   387         4  7/23/15  7631       979     1     0            0
## 21   388         4  7/23/15  8329       968     1     0            0
## 22   389         4  7/23/15  9681      1216     1     0            0
## 23   390         4  7/23/15  9138       859     1     0            0
## 24   391         4  7/23/15  4733       590     1     0            0
## 25   392         4  7/23/15  5870       609     1     0            0
## 26   393         4  7/23/15  4877       504     1     0            0
## 27   394         4  7/23/15  7948       648     1     0            0
## 28   395         4  7/23/15  3261       392     1     0            0
## 29   396         4  7/23/15  9983      1081     1     0            0
## 30   397         4  7/23/15  4550       662     1     0            0
## 31   398         4  7/23/15  4585       517     1     0            0
## 32   384         3  7/22/15  6738       804     1     0            0
## 33   385         3  7/22/15  5280       490     1     0            0
## 34   386         3  7/22/15  5305       475     1     0            0
## 35   387         3  7/22/15  7560       997     1     0            0
## 36   388         3  7/22/15  7198       837     1     0            0
## 37   389         3  7/22/15  8989      1118     1     0            0
## 38   390         3  7/22/15  8938       852     1     0            0
## 39   391         3  7/22/15  3766       523     1     0            0
## 40   392         3  7/22/15  5354       576     1     0            0
## 41   393         3  7/22/15  4364       438     1     0            0
## 42   394         3  7/22/15  6649       554     1     0            0
## 43   395         3  7/22/15  2545       391     1     0            0
## 44   396         3  7/22/15  8392       848     1     0            0
## 45   397         3  7/22/15  4649       677     1     0            0
## 46   398         3  7/22/15  3999       456     1     0            0
## 47   384         5 10/10/14  7986       896     1     1            0
## 48   386         5 10/10/14  7254       595     1     1            0
## 49   387         5 10/10/14 14584      1862     1     1            0
## 50   388         5 10/10/14  9744      1101     1     1            0
## 51   389         5 10/10/14 12469      1454     1     1            0
## 52   390         5 10/10/14 11495      1105     1     1            0
```

```
## 53   391         5 10/10/14  6965       865     1     1            0
## 54   392         5 10/10/14  7175       731     1     1            0
## 55   393         5 10/10/14  6197       624     1     1            0
## 56   394         5 10/10/14 10234       792     1     1            0
## 57   395         5 10/10/14  3862       559     1     1            0
## 58   396         5 10/10/14  9502      1029     1     1            0
## 59   397         5 10/10/14  5893       788     1     1            0
## # ... with 44 more rows, and 1 more variables: SchoolHoliday <int>
```

```r
#delete the rows in School Holiday where NA
#make a new dataframe without the rows with NA
sales_cust <- sales_cust_tot[which(!is.na(sales_cust_tot$SchoolHoliday)),]
```