Ticket System Data Analysis

I. Environment Setting

```
# load library
library(ggplot2)
library(data.table)
suppressMessages( library(dplyr) )
suppressMessages( library(lubridate) )
# set time
Sys.setlocale("LC_TIME", "English")
# set working directory
setwd("C:/Users/ASUS/ticket-system/system")
# read in the files
files <- list.files( "data", full.names = TRUE )
data  <- fread( files, stringsAsFactors = FALSE, header = TRUE, sep = ",",
                colClasses = "character" )
```

II. Exploratory Data Analysis

The section exploratory data analysis addresses four main questions.

1. Total Ticket Revenue
2. Mean of SoldPrice by Gender
3. Age Distribution
4. Analyze TicketSiteCode

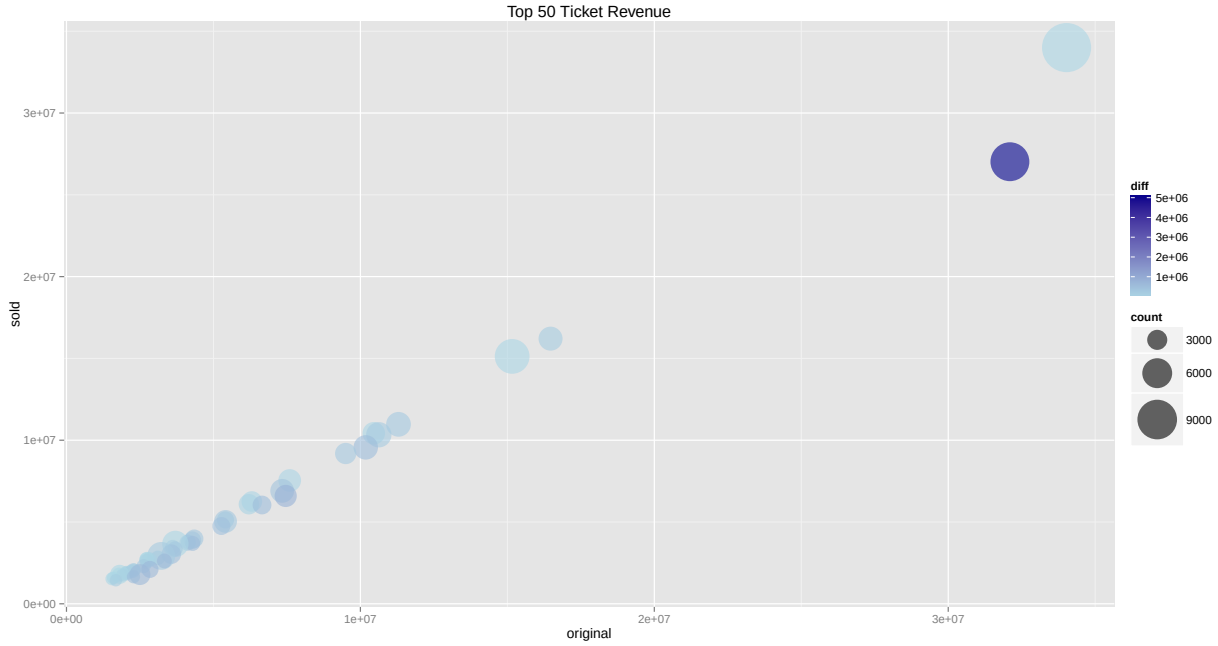
1. Total Ticket Revenue Business always cares about money, do not blaim them that is what business do, they make money. Slight digression there, anyway, let us trace back how much revenue did each TicketCode generated.

- `price` Calculate the total amount of the original ticket price and the price that were sold grouped by each TicketCode. Extract the top 50 ordered by total sold price, also add an additional column that states the difference between the total original and sold price.

```
price <- data[ , .( original = sum( as.numeric(OriginalPrice) ),
                    sold = sum( as.numeric(SoldPrice) ), count = .N ), by = TicketCode ] %>%
         arrange( desc(sold), desc(original), desc(count) ) %>%
         top_n( 50, sold ) %>%
         mutate( diff = original - sold )
head(price)
```

```
##    TicketCode original     sold count    diff
## 1:      10605 34026400 33998920 11921   27480
## 2:      10440 32098590 27022635  8779 5075955
## 3:      10413 16469980 16207980  4190  262000
## 4:      10439 15163400 15122910  7504   40490
## 5:      10430 11296430 10970440  4417  325990
## 6:      10619 10456200 10424210  3747   31990
```

```
# top 50 plot
ggplot( price, aes( original, sold, size = count, color = diff ) ) +
    geom_point( alpha = .6 ) +
    scale_size_continuous( range = c( 5, 20 ) ) +
    scale_color_gradient( low = "lightblue", high = "darkblue" ) +
    ggtitle("Top 50 Ticket Revenue")
```



The analysis of the following section will only be based on part of the dataset for simplicity.

- `highdata` For deeper insight, we will extract the TicketCode in which their total sold revenue are larger than 10^7.

```
# top-saling TicketCode
high <- price$TicketCode[ (price$sold > 10^7) ] ; high
```

```
## [1] "10605" "10440" "10413" "10439" "10430" "10619" "10329"
```

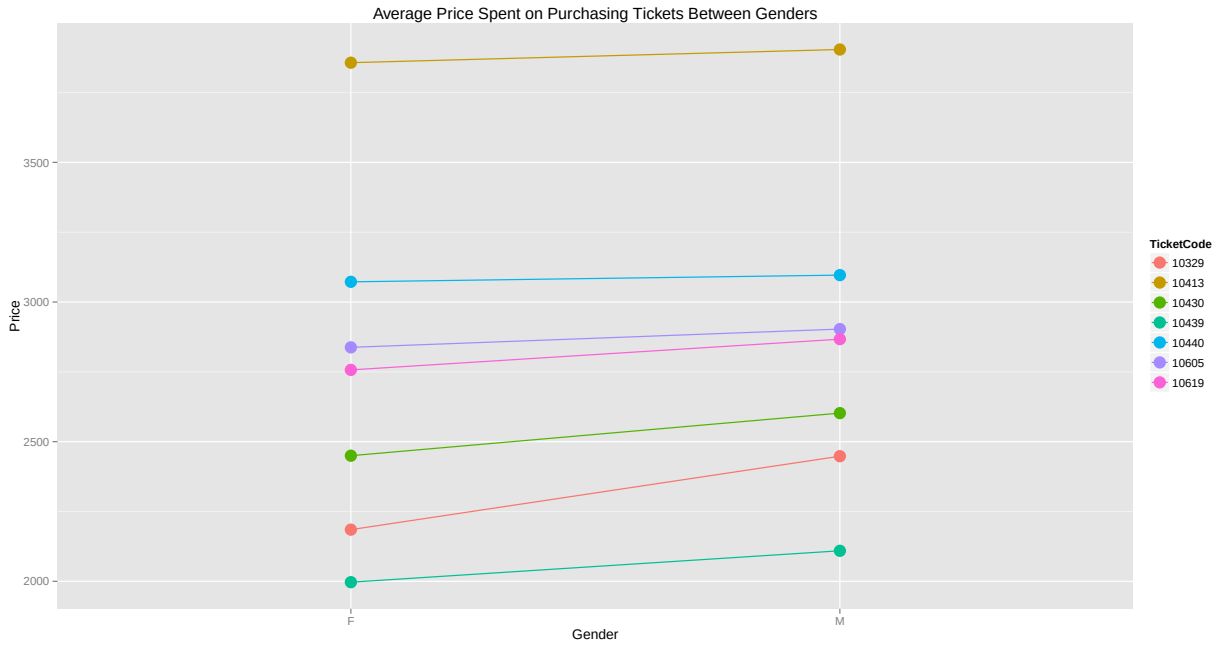
```
highdata <- data[ TicketCode %in% high, ]
```

2. Mean of SoldPrice by Gender The first question we would like to answer for these top-saling tickets is : Do male or female have different behaviorial patterns ?

- `mean1` SoldPrice is the amount of money that the consumer actually spent on purchasing the tickets. Let us start with looking at the average amount of money spent on buying tickets for each top-saling TicketCode and between genders.

```
mean1 <- aggregate( as.numeric(SoldPrice) ~ TicketCode + Gender, data = highdata,
                    FUN = mean ) %>% arrange( TicketCode )
# rename the third column, it was too lengthy
```

```
names(mean1)[3] <- "Price"
# plot of the mean
ggplot( mean1, aes( as.factor(Gender), Price, color = TicketCode, group = TicketCode ) ) +
    geom_point( size = 5 ) + geom_line() + xlab("Gender") +
    ggtitle("Average Price Spent on Purchasing Tickets Between Genders")
```



- **Note:** Based on the plot, it is kind of hard to tell whether there actually is a difference in the mean of the SoldPrice between male and female, therefore we will confirm the notion by conducting a t-test between the two sample for every TicketCode.

```
# rejection level of p-value
alpha <- .05
sapply( high, function(x)
{
    # extract only the needed column from the data
    tmp <- highdata[ TicketCode == x, ] %>% select( SoldPrice, Gender )
    # check the equality of variance for the t-test
    boolean <- var.test( as.numeric(SoldPrice) ~ as.factor(Gender), data = tmp,
              alternative = "two.sided" )$p.value > alpha
    # conduct the t-test, return boolean, true stating that there's a
    # difference between the two gender regarding the mean of amount of sold tickets
    t.test( as.numeric(SoldPrice) ~ as.factor(Gender), data = tmp,
            paired = FALSE, var.equal = boolean )$p.value < alpha
})
```

```
## 10605 10440 10413 10439 10430 10619 10329
##  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
```

- **Note:** Based on the results of the t-test, it seems that for the 7 TicketCode that attributed to more than 10^7 ticket revenues, only two of them are impartial for the average amount of money spent on

purchasing tickets between male and female. If you look back at the plot, it looks like boys tend to spend more on buying. But that was the average per person, what about the total amount ?

```
# gender distribution
table(highdata$Gender)
```

```
##
##     F     M
## 35393  9772
```

```
# total amount of money spent of tickets by gender
aggregate( as.numeric(SoldPrice) ~ TicketCode + Gender, data = highdata, FUN = sum ) %>%
    arrange( TicketCode )
```

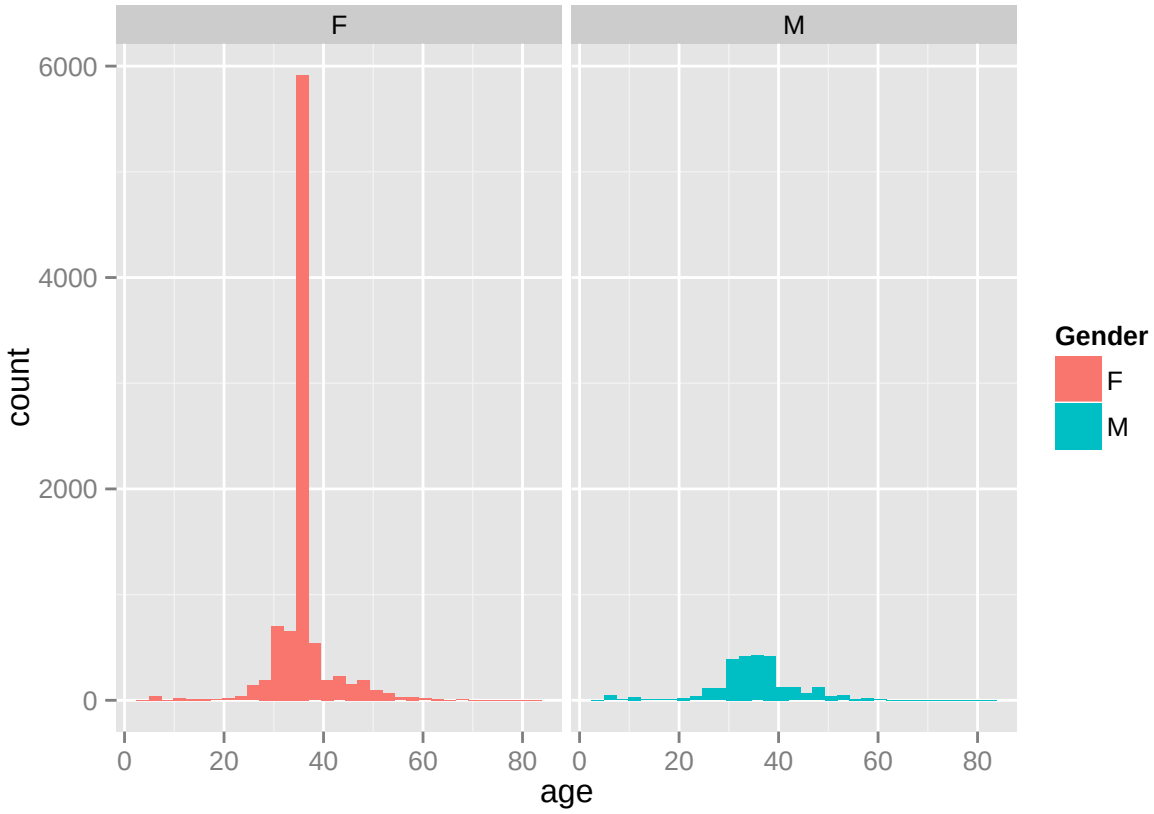
```
##    TicketCode Gender as.numeric(SoldPrice)
## 1       10329      F                7852980
## 2       10329      M                2479800
## 3       10413      F               12311580
## 4       10413      M                3896400
## 5       10430      F                8415040
## 6       10430      M                2555400
## 7       10439      F               12532710
## 8       10439      M                2590200
## 9       10440      F               20532645
## 10      10440      M                6489990
## 11      10605      F               26441920
## 12      10605      M                7557000
## 13      10619      F                7981510
## 14      10619      M                2442700
```

- **Note:** Wow! Despite the previous analysis told us that on average, males spent slightly more than females, the consumers/users for this ticket system are largely dominated by females. That is, tickets that were purchased by female are three times higher than that of male (Inference from the table above) and they account for a large portion of the total ticket revenue (Looking at the aggregated data above, the statement is true for the all 7 top-saling tickets).

3. Age Distribution So that was the discrepancy between the two genders, what about the age ? Which age level is the main target audience for these top-saling tickets. A histogram might be a good place to start.

```
# add the age of the person using the BirthYear column
highdata[ , age := year(today()) - as.numeric(BirthYear) ]
```

```
# extract one of the ticket concert and look at its age distribution
agedata <- highdata[ TicketCode == high[1], ] %>% select( SoldPrice, Gender, age )
# age distribution histogram by gender
ggplot( agedata, aes( age, fill = Gender ) ) + geom_histogram() + facet_grid( ~ Gender )
```



- **Note1:** To avoid making the report too lengthy, the age distribution histogram is depicted for only one of the TicketCode, the first one from the `highdata` to be exact. All other histogram looks quite similar to it, so we will use it to draw some hypotheses.
- **Note2:** From the histogram, we can boldy assume that the major segment of audience for this ticket system is female consumers with the age of $30 \sim 40$. This was the count, so what about the total revenue?

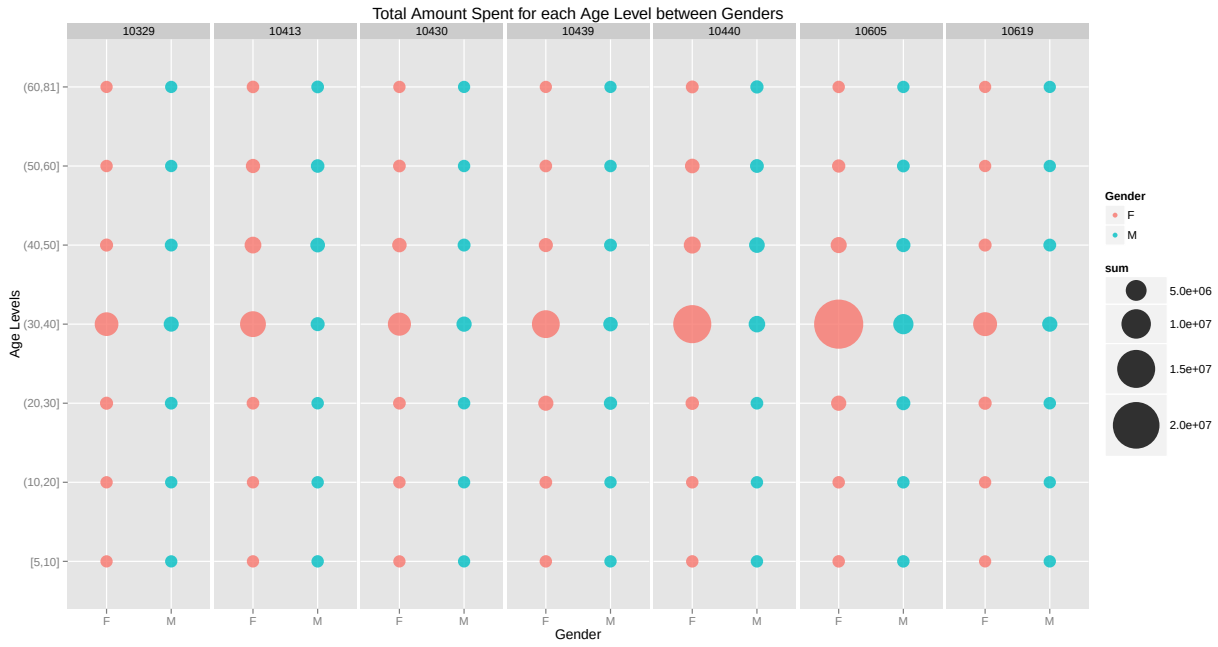
Define 7 age levels for categoring the age column, and add a new column cut to store that level in the `highdata`.

```
# define age levels
breaks <- with( highdata, c( min(age), seq( 10, 60, 10 ) , max(age) ) )
highdata$cut <- cut( highdata$age, breaks = breaks, include.lowest = TRUE )
# age distribution
table(highdata$cut)
```

```
##
##  [5,10] (10,20] (20,30] (30,40] (40,50] (50,60] (60,81]
##     359     543    3178   32978    5647    1936     524
```

```
# the sum of sold price for every ticket, gender and age breaks
sum1 <- highdata[ , .( sum = sum( as.numeric(SoldPrice) ) ),
                by = list( TicketCode, Gender, cut ) ] %>% arrange( TicketCode, cut )
# plot
```

```
ggplot( sum1, aes( Gender, cut, color = Gender, size = sum ) ) +
    geom_point( alpha = .8 ) + facet_grid( ~ TicketCode ) +
    scale_size_continuous( range = c( 5, 20 ) ) +
    labs( y = "Age Levels",
          title = "Total Amount Spent for each Age Level between Genders" )
```



- **Note1:** The table of the cut, and the plot confirms the fact the people of age 30 ~ 40 does in fact buy more tickets and lead to more revenues than other age levels, for clearity, the bigger the point in the plot, the higher the total amount of money was spent for that age level, matching our previous assumption.
- **Note2:** The biggest difference for the amount of money spent between male and female is seen for TicketCode 10605. After looking it up in the TicketName column, it was the concert of Jay Chou. Another minor point, it is quite surprising that there were records of people that falls under the age category 5 ~ 10.

Section Conclusion:

Although we do not know that whether the person that bought the ticket is actually the one that went to the concert, Suggestions can still be made according to these findings. Next time the ticket system is selling tickets that are similar to these top-saling tickets, in other words, concerts held by similar singers or bands, conduct a A/B testing on the website user interface of the ticket system , test that whether changing the user interface to meet the taste of female consumers with the age of 30 ~ 40 will boost its ticket sales. Or the complete opposite strategy is to make it more appealing to males to elevate that bleak sales of theirs. As for which marketing strategies should this ticket system use, it will more likely depend on other market surveys to see the actual reasons and motivations that lead to their different behavorial patterns.

4. Analyze TicketSiteCode In the last part of the analysis, we wish to observe the total revenue generated by each TicketSite ? Also how many TicketSite leads to the majority of the revenue.

- `site` Total revenue generated by each ticketsite.

```
site <- highdata[ , .( sum = sum( as.numeric(SoldPrice) ) ), by = TicketSiteCode ] %>%
            arrange( desc(sum) )
site
```

```
##      TicketSiteCode      sum
##   1:          88888 77718275
##   2:            715  7217800
##   3:            589  2631600
##   4:            248  2077280
##   5:            206  1901935
##  ---
## 350:            654     1800
## 351:            682     1200
## 352:           5405     1200
## 353:            563     1200
## 354:             39      800
```

```
# percentage of the TicketSiteCode that generated 70 and 80% of the total revenue.
sapply( c( .7, .8 ), function(x)
{
    mean( !cumsum( site$sum / sum(site$sum) ) > x ) * 100
})
```

```
## [1] 0.5649718 4.5197740
```

- **Note1:** TicketSiteCode 88888 accounts for 62.64 percent of the ticket system's total revenue. As for the TicketSite that had extremely inferior amount of revenues, maybe it is time to shut them down.
- **Note2:** The long tail theory where a small proportion of the products generates large proportion of the revenues also holds for this ticket system. Where 0.5 percent of the ticketsite contributes to 70 percent of the sales and 4.5 percent contributes to 80.