Homework 1

Ghandilyan Lilit

9/27/2018

Read the dataset movies3.csv into R

```
movies <- read.csv("movies3.csv")
```

The variable gross_adjusted is showing gross box office of the movie adjusted by inflation. The variable budget_adjusted shows the budget of the movie adjusted by inflation

Write a code that will make variables gross_adjusted and budget_adjusted in 1000 USD. Dont create new variable, just overwrite the old ones. (1 point)

```
movies$gross_adjusted <- movies$gross_adjusted/1000
movies$budget_adjusted <- movies$budget_adjusted/1000
```

What is the minimum for box office ? What is the minimum for budget ? (1 point)

```
min(movies$gross_adjusted)
```

```
## [1] 0.973
```

```
min(movies$budget_adjusted)
```

```
## [1] 0.29
```

The variable genre_first is showing which genre was mentioned first on the movies imdb webpage

How many Action movies are there? How many comedies ? (2 point)

```
sum(movies$genre_first=="Action")
```

```
## [1] 721
```

```
sum(movies$genre_first=="Comedy")
```

```
## [1] 844
```

Create a new dataframe with the most popular geners. Take those movies only, whose genre appear in the dataframe more than 100 times.(3 points)

Hint Suppose you want to subset the mtcars dataset in a way that will have only cars who has 6 or 8 cylinder. One way to go with it is the following: df <- mtcars[mtcars$cyl == 6|mtcars$cyl ==8,]

However the most efficient way will be: df <- mtcars[mtcars$cyl %in% c(6,8),]

```
counts <- movies %>% group_by(genre_first) %>% count()
filter<- counts[counts$n > 100, ]
movies <- movies[movies$genre_first %in% filter$genre_first, ]
```

what is the standard deviation of the imdbRating.(2 points) Hint. If you are getting NA after running the function, one reason can be that the variable has NA inside. Look at the help of the function, specifically for the argument na.rm

```
sd(movies$imdbRating, na.rm = TRUE)
```

```
## [1] 1.035619
```

On average, which genre has made the highest box office revenue ? (5 points)

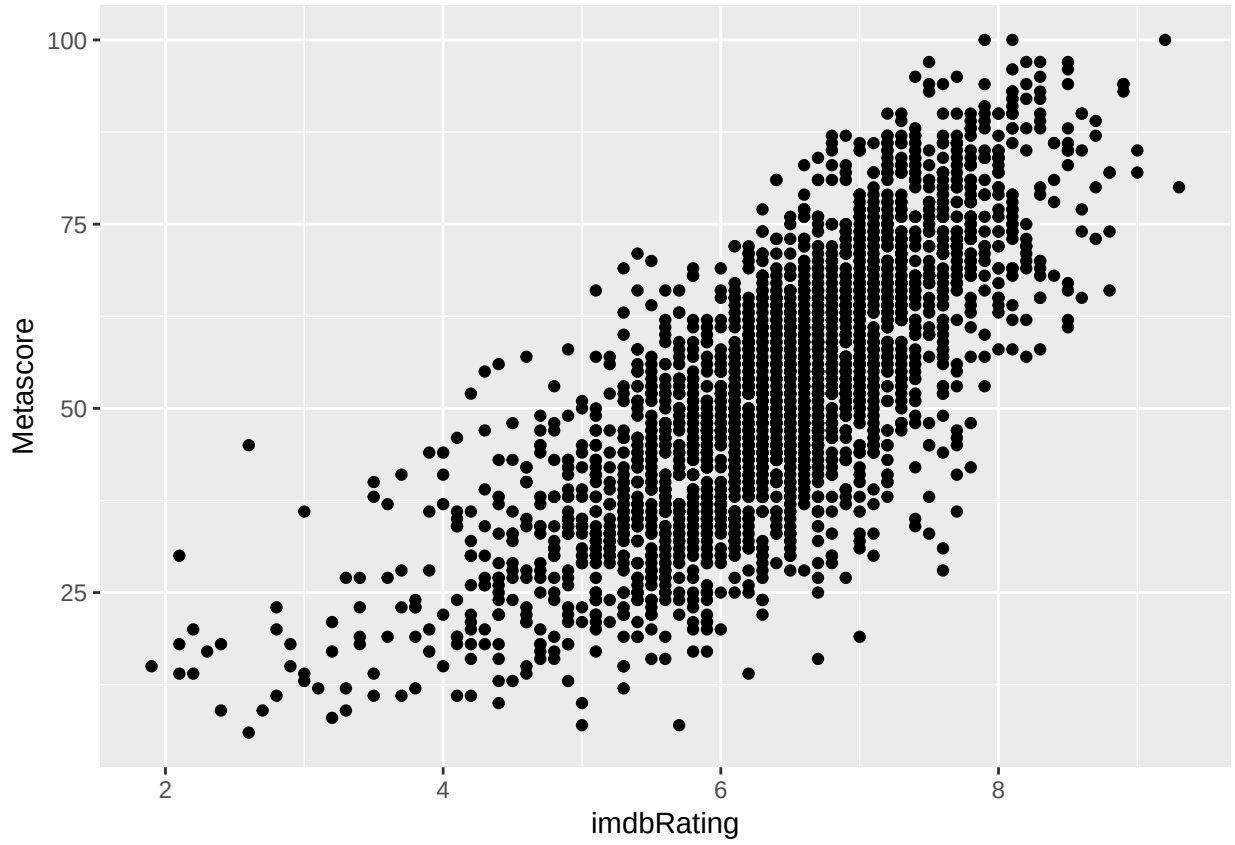
Hint: use function ?aggregate (https://goo.gl/DUyftz) or anything else, but dont give me lenghty code

```
df <- aggregate(x = movies$budget_adjusted, by = list(genre = movies$genre_first), FUN = mean)
df[which.max(df$x), "genre" ]
```

```
## [1] Action
## 17 Levels: Action Adventure Animation Biography Comedy ... Western
```

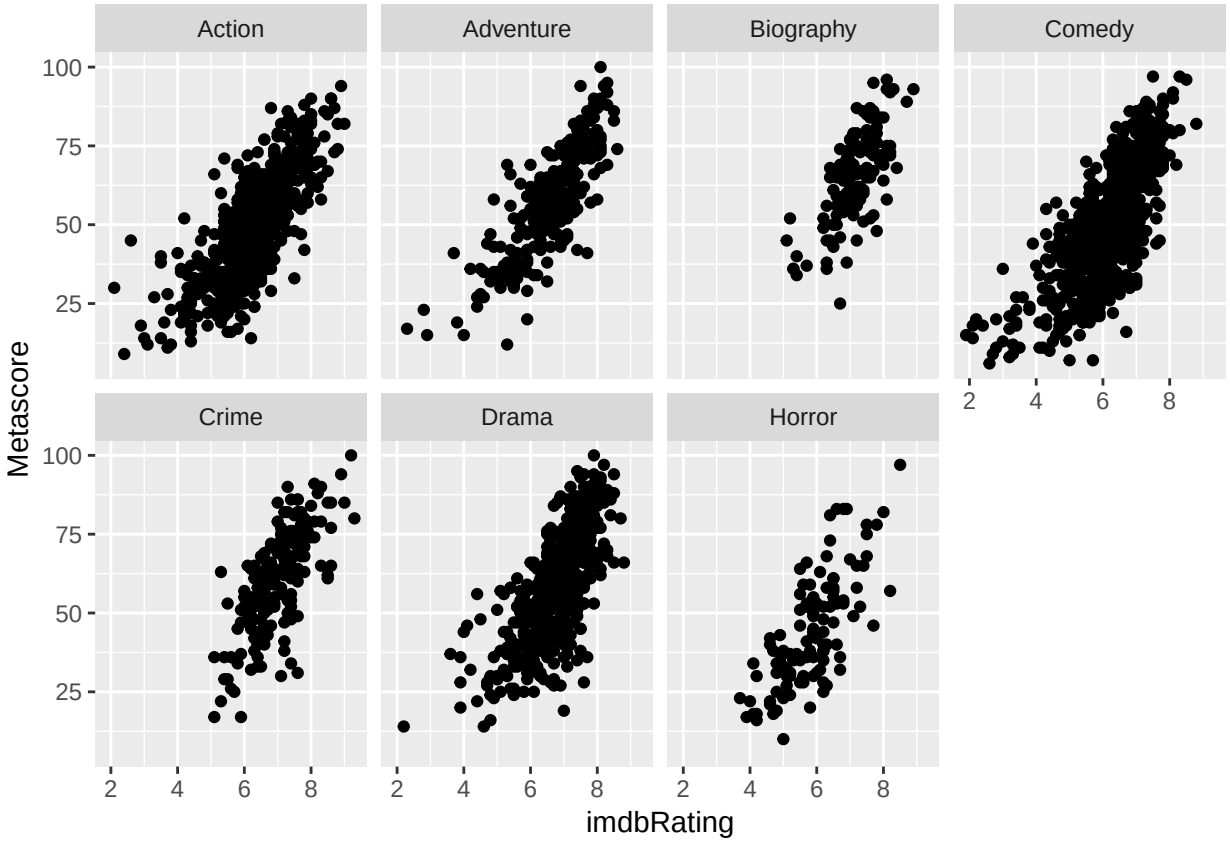
Using ggplot construct scatterplot between imdbRating and Metascore. (3 points)

```
ggplot(data= movies, aes(x=imdbRating, y = Metascore)) + geom_point(na.rm = TRUE)
```



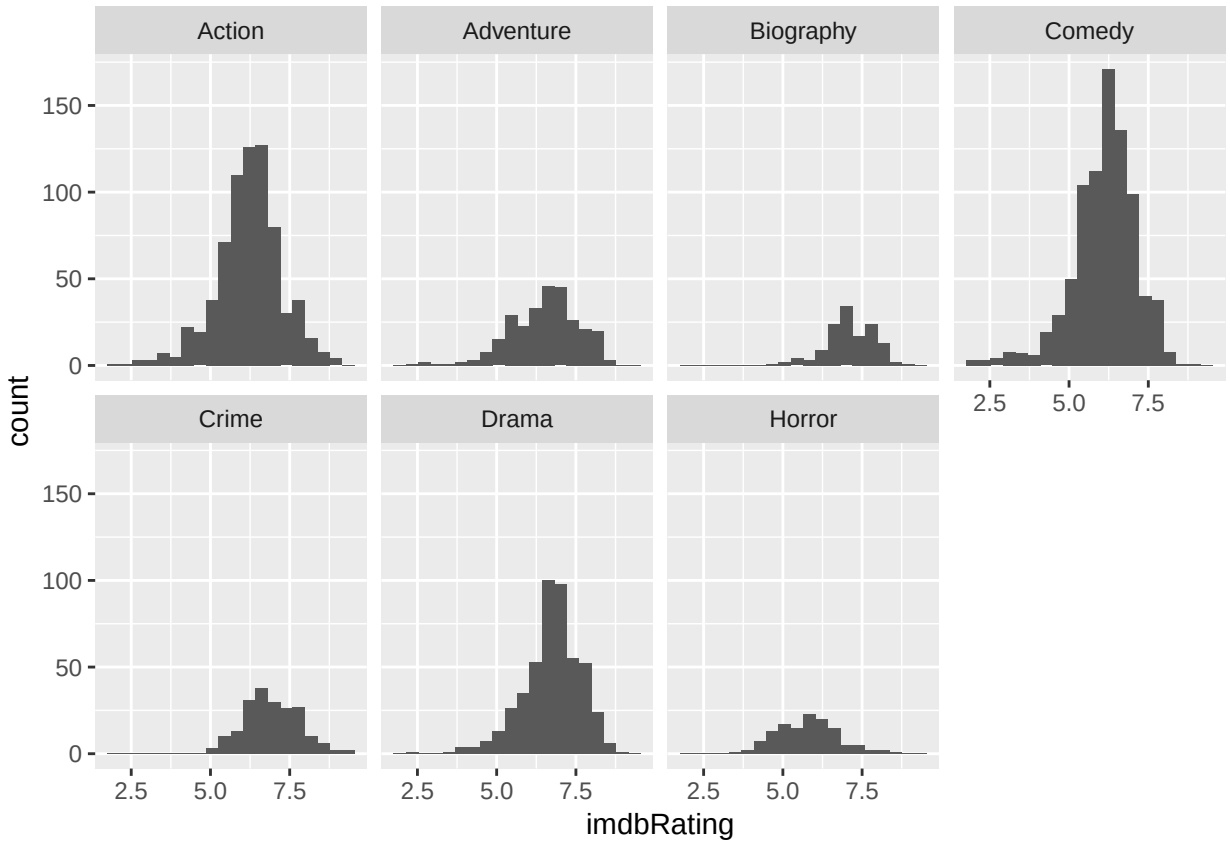
Do the same thing for each genre (using facet_grid() or facet_wrap()) (3 points)

```
plt <- ggplot(data= movies, aes(x=imdbRating, y = Metascore)) + geom_point(na.rm = TRUE)
plt + facet_wrap(~genre_first, ncol = 4)
```



Construct histogram for imdbRating for each genre (3 point)

```
plt <- ggplot(movies, aes(imdbRating)) + geom_histogram(na.rm = TRUE, bins = 20)
plt + facet_wrap(~genre_first, ncol = 4)
```



Calculate mean, median and sdadard deviation of imdbRating and Metascore for each genre.(4 points)

```
df <- movies %>% group_by(genre_first)
df %>% summarise_at(vars(imdbRating, Metascore), funs(mean, median, sd), na.rm = T)
```

```
## # A tibble: 7 x 7
##   genre_first imdbRating_mean Metascore_mean imdbRating_medi~
##   <fct>                 <dbl>          <dbl>            <dbl>
## 1 Action                 6.23           48.8              6.3
## 2 Adventure              6.50           56.8              6.6
## 3 Biography              7.11           65.5              7.2
## 4 Comedy                 6.11           49.7              6.2
## 5 Crime                  6.93           59.8              6.9
## 6 Drama                  6.74           58.0              6.8
## 7 Horror                 5.80           43.9              5.85
## # ... with 3 more variables: Metascore_median <dbl>, imdbRating_sd <dbl>,
## #   Metascore_sd <dbl>
```

Based on the previous 3 questions, describe your findings in one paragraph.(5 poits)

MetaScore and imdbRating have strong positive correlation. In the histograms we can clearly see the differences of the imdbRating medians depending on the genre, but all of them are higher than five. The distrinutions are approximately normal. The third question shows that MetaScores are usually lower than imdbRatings (considering the fact, of course, that imdbRating is on 10 scale, while Metascore on 100). Also there is much more variability in MetaScores than in imdbRatings, as standard deviations are significantly higher. (Movie critics have stronger feelings about movies and average people tend to watch movies which they assume they will like)

Create a binary variable based on the column OscarWon taking the valu 1, if the film got an Oscar and 0 if it does not. You can use ifelse for this.(3 points)

```
movies$OscarBinary <- ifelse(movies$OscarWon == 0, F, T)
```

Write a code to see what genre films have the highest probability of being awarded Oscar(2 points)

```
oscar <- movies %>% group_by(genre_first) %>% summarize(sum(OscarBinary))
#I already have a counts df from above.
total <- merge(counts, oscar, by = "genre_first")
total$probability = total$`sum(OscarBinary)` / total$n *100
total[which.max(total$probability), "genre_first"]
```

```
## [1] Biography
## 17 Levels: Action Adventure Animation Biography Comedy ... Western
```

use visualization to illustrate the difference in means of budgets of films winning Oscar with those not winning one.(3 points)

```
plt <- ggplot(data = movies, aes(factor(movies$OscarBinary), budget_adjusted)) + geom_boxplot()
plt <- plt + labs(title="Budget Differences OscarWinner vs not", x="OscarWon", y = "Budget")
plt + scale_y_continuous(labels = c("0", "100mln $", "200mln $", "300mln $", "400mln $"))
```