# EBC4223 - ASSIGNMENT 1

Quang Phong - Robert Agatić

2/14/2022

## PART 1: INTRODUCTION

The goal of this assignment is to prepare the text data containing ~25 reviews of each of the top 100 movies in 2019 so that we can relate it to movie revenue.

### Import libraries

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 4.1.2
```

```
library(qdap)
```

```
## Warning: package 'qdap' was built under R version 4.1.2
```

```
## Warning: package 'qdapRegex' was built under R version 4.1.2
```

```
## Warning: package 'qdapTools' was built under R version 4.1.2
```

```
library(SnowballC)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

### Import data

```
load("Data movie reviews final.RData")
```

### Have a look of the data for movies

```
df_Movies <- moviedescriptives # rename the data frame
head(df_Movies)
```

```
##   Rank                                       Release     Gross Max Th   Open
## 1    1                             Avengers: Endgame 858373000   4662 Apr 26
## 2    2                                 The Lion King 543638043   4802 Jul 19
## 3    3 Star Wars: Episode IX - The Rise of Skywalker 515202542   4406 Dec 20
## 4    4                                     Frozen II 477373578   4440 Nov 22
## 5    5                                   Toy Story 4 434038008   4575 Jun 21
## 6    6                                Captain Marvel 426829839   4310  Mar 8
##    Close                         Distributor movie_id
## 1 Sep 12 Walt Disney Studios Motion Pictures        1
## 2  Dec 5 Walt Disney Studios Motion Pictures        2
## 3      - Walt Disney Studios Motion Pictures        3
## 4      - Walt Disney Studios Motion Pictures        4
## 5  Dec 5 Walt Disney Studios Motion Pictures        5
## 6  Jul 4 Walt Disney Studios Motion Pictures        6
```

```
str(df_Movies)
```

```
## ’data.frame’:    100 obs. of  8 variables:
##  $ Rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Release    : Factor w/ 200 levels "1917","21 Bridges",..: 22 170 144 65 185 37 142 14 89 91 ...
##  $ Gross      : num  8.58e+08 5.44e+08 5.15e+08 4.77e+08 4.34e+08 ...
##  $ Max Th     : num  4662 4802 4406 4440 4575 ...
##  $ Open       : chr  "Apr 26" "Jul 19" "Dec 20" "Nov 22" ...
##  $ Close      : chr  "Sep 12" "Dec 5" "-" "-" ...
##  $ Distributor: Factor w/ 45 levels "-","101 Studios",..: 42 42 42 42 42 42 35 42 43 35 ...
##  $ movie_id   : int  1 2 3 4 5 6 7 8 9 10 ...
```

This data set is about Rank, Release, Gross, Max Th, Open, Close, Distributor, movie_id of 100 movies.

**Have a look of the data set for reviews**

```
df_Reviews <- reviews # rename the data frame
head(df_Reviews, 1)
```

```
##   movie_id     review_title
## 1        1  Great climax!\n
##
## 1 So here we have it, AVENGERS: ENDGAME, the expansive sequel to not only AVENGERS: INFINITY WAR but
##            Release
## 1 Avengers: Endgame
```

```
summary(df_Reviews)
```

```
##     movie_id      review_title       review_text
##  Min.   :  1.00   Length:2469        Length:2469
##  1st Qu.: 25.00   Class :character   Class :character
##  Median : 50.00   Mode  :character   Mode  :character
##  Mean   : 50.29
##  3rd Qu.: 75.00
##  Max.   :100.00
##
##                                  Release
##  1917                               :  25
##  21 Bridges                         :  25
##  47 Meters Down: Uncaged            :  25
##  A Beautiful Day in the Neighborhood:  25
##  A Dog’s Journey                    :  25
##  A Dog’s Way Home                   :  25
##  (Other)                            :2319
```

```
dim(df_Reviews)
```

```
## [1] 2469    4
```

This data set is about movie_id, review_title, review_text, Release of 2469 reviews.

# PART 2: DATA PREPROCESSING

### Obtain the length of each review and add it to review data set

```
df_Reviews$length <- nchar(df_Reviews$review_text)
head(df_Reviews$length)
```

```
## [1]  942 2133 5605 1336 1616  940
```

### Obtain average review length and add it to movie data set

```
tempdf_ReviewLength <- df_Reviews %>%
  group_by(Release) %>%
  summarize(averageReviewLength = mean(length))
head(tempdf_ReviewLength)
```

```
## # A tibble: 6 x 2
##   Release                             averageReviewLength
##   <fct>                                             <dbl>
## 1 1917                                              1683.
## 2 21 Bridges                                        1187.
## 3 47 Meters Down: Uncaged                           1250.
## 4 A Beautiful Day in the Neighborhood               1278.
## 5 A Dog’s Journey                                   1096.
## 6 A Dog’s Way Home                                   783.
```

```
# Merge the average review length to df_Movies
df_Movies <- df_Movies %>%
  inner_join(tempdf_ReviewLength, by = "Release")

head(df_Movies)
```

```
##   Rank                                       Release     Gross Max Th   Open
## 1    1                             Avengers: Endgame 858373000   4662 Apr 26
## 2    2                                 The Lion King 543638043   4802 Jul 19
## 3    3 Star Wars: Episode IX - The Rise of Skywalker 515202542   4406 Dec 20
## 4    4                                     Frozen II 477373578   4440 Nov 22
## 5    5                                   Toy Story 4 434038008   4575 Jun 21
## 6    6                                Captain Marvel 426829839   4310  Mar 8
##    Close                         Distributor movie_id averageReviewLength
## 1 Sep 12 Walt Disney Studios Motion Pictures        1             2254.16
## 2  Dec 5 Walt Disney Studios Motion Pictures        2             1732.28
## 3      - Walt Disney Studios Motion Pictures        3             2198.16
## 4      - Walt Disney Studios Motion Pictures        4             1809.56
## 5  Dec 5 Walt Disney Studios Motion Pictures        5             1638.60
## 6  Jul 4 Walt Disney Studios Motion Pictures        6             1912.32
```

### Remove punctuations from review data

```
df_Reviews$review_text <- removePunctuation(df_Reviews$review_text)
head(df_Reviews$review_text, 1)
```

```
## [1] "So here we have it AVENGERS ENDGAME the expansive sequel to not only AVENGERS INFINITY WAR but a
```

### Remove numbers from review data

```
df_Reviews$review_text <- removeNumbers(df_Reviews$review_text)
head(df_Reviews$review_text, 1)
```

```
## [1] "So here we have it AVENGERS ENDGAME the expansive sequel to not only AVENGERS INFINITY WAR but a
```

### Lower texts from review data

```
df_Reviews$review_text <- tolower(df_Reviews$review_text)
head(df_Reviews$review_text, 1)
```

```
## [1] "so here we have it avengers endgame the expansive sequel to not only avengers infinity war but a
```

### Remove stop words from review data

```
df_Reviews$review_text <- removeWords(df_Reviews$review_text, stopwords("en"))
head(df_Reviews$review_text, 1)
```

```
## [1] "     avengers endgame  expansive sequel    avengers infinity war  also   whole last  years    ma
```

### Remove excess white space from review data

```
df_Reviews$review_text <- stripWhitespace(df_Reviews$review_text)
head(df_Reviews$review_text, 1)
```

```
## [1] " avengers endgame expansive sequel avengers infinity war also whole last years marvel cinema fil
```

### Remove symbols from review data

```
apply(df_Reviews['review_text'], 1, function(x) gsub("[[:punct:]]", "", x))
```

### Stem review data

```r
df_Reviews$review_text <- stemDocument(df_Reviews$review_text, language = "english")
head(df_Reviews$review_text, 1)
```

```
## [1] "aveng endgam expans sequel aveng infin war also whole last year marvel cinema film big boot fil
```

# PART 3: ANALYSIS

```r
df_Reviews$polarity <- counts(polarity(df_Reviews$review_text))[, "polarity"]
```

### Obtain average review polarity, standard deviation and add them to movie data

```r
tempdf_ReviewPolarity <- df_Reviews %>%
  group_by(Release) %>%
  summarize(reviewPolarity = mean(polarity),
            stdPolarity = sd(polarity))
head(tempdf_ReviewPolarity)
```

```
## # A tibble: 6 x 3
##   Release                             reviewPolarity stdPolarity
##   <fct>                                        <dbl>       <dbl>
## 1 1917                                         0.372       0.555
## 2 21 Bridges                                   0.103       0.533
## 3 47 Meters Down: Uncaged                     -0.375       0.833
## 4 A Beautiful Day in the Neighborhood          0.231       0.339
## 5 A Dog’s Journey                              0.487       0.387
## 6 A Dog’s Way Home                             0.309       0.405
```

```r
# Merge the average polarity to df_Movies
df_Movies <- df_Movies %>%
  inner_join(tempdf_ReviewPolarity, by = "Release")

head(df_Movies)
```

```
##   Rank                                       Release     Gross Max Th   Open
## 1    1                             Avengers: Endgame 858373000   4662 Apr 26
## 2    2                                 The Lion King 543638043   4802 Jul 19
## 3    3 Star Wars: Episode IX - The Rise of Skywalker 515202542   4406 Dec 20
## 4    4                                     Frozen II 477373578   4440 Nov 22
## 5    5                                   Toy Story 4 434038008   4575 Jun 21
## 6    6                                Captain Marvel 426829839   4310  Mar 8
##    Close                          Distributor movie_id averageReviewLength
## 1 Sep 12 Walt Disney Studios Motion Pictures        1             2254.16
## 2  Dec 5 Walt Disney Studios Motion Pictures        2             1732.28
## 3      - Walt Disney Studios Motion Pictures        3             2198.16
## 4      - Walt Disney Studios Motion Pictures        4             1809.56
## 5  Dec 5 Walt Disney Studios Motion Pictures        5             1638.60
## 6   Jul 4 Walt Disney Studios Motion Pictures        6             1912.32
##   reviewPolarity stdPolarity
## 1      0.4683305   0.4943186
## 2      0.2527416   0.5542439
## 3      0.1288735   0.4316331
## 4      0.3518811   0.7113025
## 5      0.4612540   0.4884192
## 6      0.6994693   0.4925952
```

### Run regression models to explain revenue variable

```
model1 <- lm(Gross ~ reviewPolarity, data = df_Movies)
model2  <- lm(Gross ~ reviewPolarity+stdPolarity, data = df_Movies)
model3 <- lm(Gross ~ reviewPolarity+stdPolarity+averageReviewLength, data = df_Movies)
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = Gross ~ reviewPolarity, data = df_Movies)
##
## Residuals:
##        Min          1Q     Median          3Q         Max
## -124703795  -68782150  -40702008    5360191  727011674
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     83273934   17966048   4.635  1.1e-05 ***
## reviewPolarity 102678316   51968183   1.976    0.051 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 134200000 on 98 degrees of freedom
## Multiple R-squared:  0.03831,    Adjusted R-squared:  0.0285
## F-statistic: 3.904 on 1 and 98 DF,  p-value: 0.05099
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = Gross ~ reviewPolarity + stdPolarity, data = df_Movies)
##
## Residuals:
##        Min          1Q     Median          3Q         Max
## -130383034  -72939301  -32382738   -1146458  729371402
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -38361749   86419928  -0.444   0.6581
## reviewPolarity 120158330   53788553   2.234   0.0278 *
## stdPolarity    224732634  154870145   1.451   0.1500
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 134100000 on 96 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.05741,    Adjusted R-squared:  0.03777 
## F-statistic: 2.923 on 2 and 96 DF,  p-value: 0.05855
```

```
summary(model3)
```

```
## 
## Call:
## lm(formula = Gross ~ reviewPolarity + stdPolarity + averageReviewLength, 
##     data = df_Movies)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -218169805  -67646712  -22450201   36130997  617051245 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         -280131930   99577642  -2.813  0.00596 ** 
## reviewPolarity       130785302   49930440   2.619  0.01026 *  
## stdPolarity          284930502  144319952   1.974  0.05125 .  
## averageReviewLength     141674      34656   4.088  9.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 124300000 on 95 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1984, Adjusted R-squared:  0.1731 
## F-statistic: 7.839 on 3 and 95 DF,  p-value: 9.945e-05
```

### Run regression models with ln of revenue variable as outcome variable

```
df_Movies$lnGross <- log(df_Movies$Gross)
```

```
model4 <- lm(lnGross ~ reviewPolarity, data = df_Movies)
model5  <- lm(lnGross ~ reviewPolarity+stdPolarity, data = df_Movies)
model6 <- lm(lnGross ~ reviewPolarity+stdPolarity+averageReviewLength, data = df_Movies)
```

```
summary(model4)
```

```
## 
## Call:
## lm(formula = lnGross ~ reviewPolarity, data = df_Movies)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.3772 -0.6500 -0.1304  0.4990  2.3706 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     17.8567     0.1154 154.679   <2e-16 ***
## reviewPolarity   0.7328     0.3339   2.195   0.0306 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8626 on 98 degrees of freedom
## Multiple R-squared:  0.04684,    Adjusted R-squared:  0.03711
## F-statistic: 4.816 on 1 and 98 DF,  p-value: 0.03056
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = lnGross ~ reviewPolarity + stdPolarity, data = df_Movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3874 -0.7106 -0.1037  0.4390  2.3882
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.9310     0.5515  30.702   <2e-16 ***
## reviewPolarity   0.8576     0.3432   2.499   0.0142 *
## stdPolarity      1.7188     0.9883   1.739   0.0852 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8555 on 96 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.07316,    Adjusted R-squared:  0.05385
## F-statistic: 3.789 on 2 and 96 DF,  p-value: 0.02608
```

```
summary(model6)
```

```
##
## Call:
## lm(formula = lnGross ~ reviewPolarity + stdPolarity + averageReviewLength,
##     data = df_Movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83643 -0.54349 -0.02863  0.52847  1.80933
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.568e+01  6.546e-01  23.962  < 2e-16 ***
## reviewPolarity      9.124e-01  3.282e-01   2.780  0.00656 **
## stdPolarity         2.029e+00  9.487e-01   2.139  0.03501 *
## averageReviewLength 7.302e-04  2.278e-04   3.205  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8169 on 95 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1636, Adjusted R-squared:  0.1372
## F-statistic: 6.194 on 3 and 95 DF,  p-value: 0.0006871
```