Maths & Stats Group Project

Siow Meng Low

26 September 2016

Loading Data & Key Summary

The first step before any analysis is to load the data. The Kaggle movie data is recorded in the form of a CSV file. We shall first load it and display the key metadata of the data set.

```
library(plyr)

movieDF <- read.csv("D:/R/MathsProject/movie_metadata.csv", header = TRUE, stringsAsFactors = FALSE)

names(movieDF)
```

```
##  [1] "color"                     "director_name"
##  [3] "num_critic_for_reviews"    "duration"
##  [5] "director_facebook_likes"   "actor_3_facebook_likes"
##  [7] "actor_2_name"              "actor_1_facebook_likes"
##  [9] "gross"                     "genres"
## [11] "actor_1_name"              "movie_title"
## [13] "num_voted_users"           "cast_total_facebook_likes"
## [15] "actor_3_name"              "facenumber_in_poster"
## [17] "plot_keywords"             "movie_imdb_link"
## [19] "num_user_for_reviews"      "language"
## [21] "country"                   "content_rating"
## [23] "budget"                    "title_year"
## [25] "actor_2_facebook_likes"    "imdb_score"
## [27] "aspect_ratio"              "movie_facebook_likes"
```

```
dim(movieDF)
```

```
## [1] 5043   28
```

```
str(movieDF)
```

```
## 'data.frame':    5043 obs. of  28 variables:
##  $ color                     : chr  "Color" "Color" "Color" "Color" ...
##  $ director_name             : chr  "James Cameron" "Gore Verbinski" "Sam Mendes" "Christopher Nolan"
##  $ num_critic_for_reviews    : int  723 302 602 813 NA 462 392 324 635 375 ...
##  $ duration                  : int  178 169 148 164 NA 132 156 100 141 153 ...
##  $ director_facebook_likes   : int  0 563 0 22000 131 475 0 15 0 282 ...
##  $ actor_3_facebook_likes    : int  855 1000 161 23000 NA 530 4000 284 19000 10000 ...
##  $ actor_2_name              : chr  "Joel David Moore" "Orlando Bloom" "Rory Kinnear" "Christian Bale
##  $ actor_1_facebook_likes    : int  1000 40000 11000 27000 131 640 24000 799 26000 25000 ...
##  $ gross                     : int  760505847 309404152 200074175 448130642 NA 73058679 336530303 200
##  $ genres                    : chr  "Action|Adventure|Fantasy|Sci-Fi" "Action|Adventure|Fantasy" "Acti
##  $ actor_1_name              : chr  "CCH Pounder" "Johnny Depp" "Christoph Waltz" "Tom Hardy" ...
##  $ movie_title               : chr  "AvatarÂ " "Pirates of the Caribbean: At World's EndÂ " "SpectreÂ
##  $ num_voted_users           : int  886204 471220 275868 1144337 8 212204 383056 294810 462669 321795
##  $ cast_total_facebook_likes: int  4834 48350 11700 106759 143 1873 46055 2036 92000 58753 ...
##  $ actor_3_name              : chr  "Wes Studi" "Jack Davenport" "Stephanie Sigman" "Joseph Gordon-Le
##  $ facenumber_in_poster      : int  0 0 1 0 0 1 0 1 4 3 ...
```

```
##  $ plot_keywords            : chr  "avatar|future|marine|native|paraplegic" "goddess|marriage ceremor
##  $ movie_imdb_link          : chr  "http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1" "http://www
##  $ num_user_for_reviews     : int  3054 1238 994 2701 NA 738 1902 387 1117 973 ...
##  $ language                 : chr  "English" "English" "English" "English" ...
##  $ country                  : chr  "USA" "USA" "UK" "USA" ...
##  $ content_rating           : chr  "PG-13" "PG-13" "PG-13" "PG-13" ...
##  $ budget                   : num  2.37e+08 3.00e+08 2.45e+08 2.50e+08 NA ...
##  $ title_year               : int  2009 2007 2015 2012 NA 2012 2007 2010 2015 2009 ...
##  $ actor_2_facebook_likes   : int  936 5000 393 23000 12 632 11000 553 21000 11000 ...
##  $ imdb_score               : num  7.9 7.1 6.8 8.5 7.1 6.6 6.2 7.8 7.5 7.5 ...
##  $ aspect_ratio             : num  1.78 2.35 2.35 2.35 NA 2.35 2.35 1.85 2.35 2.35 ...
##  $ movie_facebook_likes     : int  33000 0 85000 164000 0 24000 0 29000 118000 10000 ...
```

```
summary(movieDF)
```

```
##     color           director_name      num_critic_for_reviews
##  Length:5043        Length:5043        Min.   :  1.0
##  Class :character   Class :character   1st Qu.: 50.0
##  Mode  :character   Mode  :character   Median :110.0
##                                        Mean   :140.2
##                                        3rd Qu.:195.0
##                                        Max.   :813.0
##                                        NA's   :50
##     duration     director_facebook_likes actor_3_facebook_likes
##  Min.   :  7.0   Min.   :    0.0         Min.   :    0.0
##  1st Qu.: 93.0   1st Qu.:    7.0         1st Qu.:  133.0
##  Median :103.0   Median :   49.0         Median :  371.5
##  Mean   :107.2   Mean   :  686.5         Mean   :  645.0
##  3rd Qu.:118.0   3rd Qu.:  194.5         3rd Qu.:  636.0
##  Max.   :511.0   Max.   :23000.0         Max.   :23000.0
##  NA's   :15      NA's   :104             NA's   :23
##  actor_2_name       actor_1_facebook_likes     gross
##  Length:5043        Min.   :     0         Min.   :      162
##  Class :character   1st Qu.:   614         1st Qu.:  5340988
##  Mode  :character   Median :   988         Median : 25517500
##                     Mean   :  6560         Mean   : 48468408
##                     3rd Qu.: 11000         3rd Qu.: 62309438
##                     Max.   :640000         Max.   :760505847
##                     NA's   :7              NA's   :884
##     genres          actor_1_name       movie_title
##  Length:5043        Length:5043        Length:5043
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  num_voted_users   cast_total_facebook_likes actor_3_name
##  Min.   :      5   Min.   :     0            Length:5043
##  1st Qu.:   8594   1st Qu.:  1411            Class :character
##  Median :  34359   Median :  3090            Mode  :character
##  Mean   :  83668   Mean   :  9699
##  3rd Qu.:  96309   3rd Qu.: 13756
##  Max.   :1689764   Max.   :656730
##
```

```
##  facenumber_in_poster plot_keywords      movie_imdb_link
##  Min.   : 0.000       Length:5043        Length:5043
##  1st Qu.: 0.000       Class :character   Class :character
##  Median : 1.000       Mode  :character   Mode  :character
##  Mean   : 1.371
##  3rd Qu.: 2.000
##  Max.   :43.000
##  NA's   :13
##  num_user_for_reviews   language           country
##  Min.   :   1.0       Length:5043        Length:5043
##  1st Qu.:  65.0       Class :character   Class :character
##  Median : 156.0       Mode  :character   Mode  :character
##  Mean   : 272.8
##  3rd Qu.: 326.0
##  Max.   :5060.0
##  NA's   :21
##  content_rating         budget            title_year
##  Length:5043        Min.   :2.180e+02   Min.   :1916
##  Class :character   1st Qu.:6.000e+06   1st Qu.:1999
##  Mode  :character   Median :2.000e+07   Median :2005
##                     Mean   :3.975e+07   Mean   :2002
##                     3rd Qu.:4.500e+07   3rd Qu.:2011
##                     Max.   :1.222e+10   Max.   :2016
##                     NA's   :492         NA's   :108
##  actor_2_facebook_likes   imdb_score     aspect_ratio
##  Min.   :     0         Min.   :1.600   Min.   : 1.18
##  1st Qu.:   281         1st Qu.:5.800   1st Qu.: 1.85
##  Median :   595         Median :6.600   Median : 2.35
##  Mean   :  1652         Mean   :6.442   Mean   : 2.22
##  3rd Qu.:   918         3rd Qu.:7.200   3rd Qu.: 2.35
##  Max.   :137000         Max.   :9.500   Max.   :16.00
##  NA's   :13                             NA's   :329
##  movie_facebook_likes
##  Min.   :     0
##  1st Qu.:     0
##  Median :   166
##  Mean   :  7526
##  3rd Qu.:  3000
##  Max.   :349000
##
```

```
head(movieDF)
```

```
##   color     director_name num_critic_for_reviews duration
## 1 Color     James Cameron                    723      178
## 2 Color    Gore Verbinski                    302      169
## 3 Color        Sam Mendes                    602      148
## 4 Color Christopher Nolan                    813      164
## 5             Doug Walker                     NA       NA
## 6 Color    Andrew Stanton                    462      132
##   director_facebook_likes actor_3_facebook_likes     actor_2_name
## 1                       0                    855 Joel David Moore
## 2                     563                   1000    Orlando Bloom
## 3                       0                    161     Rory Kinnear
## 4                   22000                  23000   Christian Bale
```

```
## 5                    131                      NA       Rob Walker
## 6                    475                     530  Samantha Morton
##   actor_1_facebook_likes     gross                           genres
## 1                   1000 760505847 Action|Adventure|Fantasy|Sci-Fi
## 2                  40000 309404152        Action|Adventure|Fantasy
## 3                  11000 200074175       Action|Adventure|Thriller
## 4                  27000 448130642                 Action|Thriller
## 5                    131        NA                    Documentary
## 6                    640  73058679         Action|Adventure|Sci-Fi
##      actor_1_name                                            movie_title
## 1     CCH Pounder                                               AvatarÂ
## 2     Johnny Depp              Pirates of the Caribbean: At World's EndÂ
## 3 Christoph Waltz                                             SpectreÂ
## 4       Tom Hardy                                  The Dark Knight RisesÂ
## 5     Doug Walker Star Wars: Episode VII - The Force AwakensÂ
## 6    Daryl Sabara                                          John CarterÂ
##   num_voted_users cast_total_facebook_likes         actor_3_name
## 1          886204                     4834            Wes Studi
## 2          471220                    48350       Jack Davenport
## 3          275868                    11700     Stephanie Sigman
## 4         1144337                   106759 Joseph Gordon-Levitt
## 5               8                      143
## 6          212204                     1873         Polly Walker
##   facenumber_in_poster
## 1                    0
## 2                    0
## 3                    1
## 4                    0
## 5                    0
## 6                    1
##                                                 plot_keywords
## 1                        avatar|future|marine|native|paraplegic
## 2    goddess|marriage ceremony|marriage proposal|pirate|singapore
## 3                          bomb|espionage|sequel|spy|terrorist
## 4 deception|imprisonment|lawlessness|police officer|terrorist plot
## 5
## 6              alien|american civil war|male nipple|mars|princess
##                                    movie_imdb_link
## 1 http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1
## 2 http://www.imdb.com/title/tt0449088/?ref_=fn_tt_tt_1
## 3 http://www.imdb.com/title/tt2379713/?ref_=fn_tt_tt_1
## 4 http://www.imdb.com/title/tt1345836/?ref_=fn_tt_tt_1
## 5 http://www.imdb.com/title/tt5289954/?ref_=fn_tt_tt_1
## 6 http://www.imdb.com/title/tt0401729/?ref_=fn_tt_tt_1
##   num_user_for_reviews language country content_rating    budget
## 1                 3054  English     USA          PG-13 237000000
## 2                 1238  English     USA          PG-13 300000000
## 3                  994  English      UK          PG-13 245000000
## 4                 2701  English     USA          PG-13 250000000
## 5                   NA                                        NA
## 6                  738  English     USA          PG-13 263700000
##   title_year actor_2_facebook_likes imdb_score aspect_ratio
## 1       2009                   936        7.9         1.78
## 2       2007                  5000        7.1         2.35
```

```
## 3       2015                  393        6.8         2.35
## 4       2012                23000        8.5         2.35
## 5         NA                   12        7.1           NA
## 6       2012                  632        6.6         2.35
##   movie_facebook_likes
## 1                33000
## 2                    0
## 3                85000
## 4               164000
## 5                    0
## 6                24000
```

```
#movieDF[movieDF$director_name == "James Cameron", ]
#movieDF[movieDF$actor_1_name == "Johnny Depp", ]
```

As usual, the dataset comes with some missing values. The following R code gives an overview on how many missing values are there per column.

```
colNA <- function(dfCol){ sum(is.na(dfCol)) }

apply(movieDF, 2, colNA)
```

```
##                     color             director_name
##                         0                         0
##    num_critic_for_reviews                  duration
##                        50                        15
##   director_facebook_likes    actor_3_facebook_likes
##                       104                        23
##              actor_2_name    actor_1_facebook_likes
##                         0                         7
##                     gross                    genres
##                       884                         0
##              actor_1_name               movie_title
##                         0                         0
##           num_voted_users cast_total_facebook_likes
##                         0                         0
##              actor_3_name      facenumber_in_poster
##                         0                        13
##            plot_keywords           movie_imdb_link
##                         0                         0
##      num_user_for_reviews                  language
##                        21                         0
##                   country            content_rating
##                         0                         0
##                    budget               title_year
##                       492                       108
##    actor_2_facebook_likes                imdb_score
##                        13                         0
##              aspect_ratio      movie_facebook_likes
##                       329                         0
```

Univariate Statistics

The key variable which we are interested in predicting will be the gross revenue. First let's look at the univariate statistics of this data.

```r
mean(movieDF$gross, na.rm = TRUE)
```

```
## [1] 48468408
```

```r
var(movieDF$gross, na.rm = TRUE)
```

```
## [1] 4.685812e+15
```

Bivariate Statistics

In this section, let us inspect how gross revenue correlates with other columns in the dataset. As seen below, most of the variables have very low correlation coefficient with gross revenue.

```r
boolGross <- !is.na(movieDF$gross)
boolBudget <- !is.na(movieDF$budget)

bool <- boolGross & boolBudget

# Covariance and correlation of gross revenue with movie budget
cov(movieDF$gross[bool], movieDF$budget[bool])
```

```
## [1] 1.586166e+15
```

```r
cor(movieDF$gross[bool], movieDF$budget[bool])
```

```
## [1] 0.1021795
```

```r
boolDuration <- !is.na(movieDF$duration)

bool <- boolDuration & boolBudget

# Covariance and correlation of movie duration with movie budget
cov(movieDF$duration[bool], movieDF$budget[bool])
```

```
## [1] 351978600
```

```r
cor(movieDF$duration[bool], movieDF$budget[bool])
```

```
## [1] 0.07427598
```

```r
boolScore <- !is.na(movieDF$imdb_score)

bool <- boolGross & boolScore

# Covariance and correlation of gross revenue with IMDB score
cov(movieDF$gross[bool], movieDF$imdb_score[bool])
```

```
## [1] 14262485
```

```r
cor(movieDF$gross[bool], movieDF$imdb_score[bool])
```

```
## [1] 0.1980212
```

```r
boolDirFB <- !is.na(movieDF$director_facebook_likes) & (movieDF$director_facebook_likes > 0)

bool <- boolGross & boolDirFB
```

```r
# Covariance and correlation of gross revenue with Director facebook likes
cov(movieDF$gross[bool], movieDF$director_facebook_likes[bool])
```

```
## [1] 37458683716
```

```r
cor(movieDF$gross[bool], movieDF$director_facebook_likes[bool])
```

```
## [1] 0.1878235
```

```r
boolCastFB <- !is.na(movieDF$cast_total_facebook_likes) & (movieDF$cast_total_facebook_likes > 0)

bool <- boolGross & boolCastFB

# Covariance and correlation of gross revenue with movie cast facebook likes
cov(movieDF$gross[bool], movieDF$cast_total_facebook_likes[bool])
```

```
## [1] 312890629885
```

```r
cor(movieDF$gross[bool], movieDF$cast_total_facebook_likes[bool])
```

```
## [1] 0.2463148
```

Next, we'll inspect the total number of likes that the movie gathered on facebook. If we use all the data to calculate correlation, the correlation coefficient is pretty low (< 0.4).

Considering the fact that Facebook was only launched to public in early 2004, it might makes more sense if we only consider the movies made after 2004. The below R code calculate the coefficient for movie produced after 2005 and 2010.

We can see that there's a huge improvement in the correlation. Hence, the total number of facebook likes might be a good predictor for gross revenue.

```r
boolMovieFB <- !is.na(movieDF$movie_facebook_likes)

bool <- boolGross & boolMovieFB

cov(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 538018085159
```

```r
cor(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 0.3780823
```

```r
boolYear <- (!is.na(movieDF$title_year) & movieDF$title_year >= 2005)

bool <- bool & boolYear

cov(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 914729444450
```

```r
cor(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 0.4555846
```

```r
boolYear <- (!is.na(movieDF$title_year) & movieDF$title_year >= 2010)

bool <- boolGross & boolMovieFB & boolYear

cov(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 1.490954e+12
```

```
cor(movieDF$gross[bool], movieDF$movie_facebook_likes[bool])
```

```
## [1] 0.5644529
```

Multiple Grouping

This section looks at the average gross revenue for different groups of movies. The following code separates the data into different groups: coloured, production year, and language.

```
ddply(movieDF, ~ color, summarise, mean = mean(gross, na.rm = TRUE), sd = sd(gross, na.rm = TRUE))
```

```
##               color     mean       sd
## 1                   40680673 55626916
## 2  Black and White 32457017 50001772
## 3             Color 49026187 68952259
```

```
ddply(movieDF, ~ title_year, summarise, mean = mean(gross, na.rm = TRUE), sd = sd(gross, na.rm = TRUE))
```

```
##    title_year        mean          sd
## 1        1916         NaN          NA
## 2        1920   3000000.0          NA
## 3        1925         NaN          NA
## 4        1927     26435.0          NA
## 5        1929   1408975.0   1978520.1
## 6        1930         NaN          NA
## 7        1932         NaN          NA
## 8        1933   2300000.0          NA
## 9        1934         NaN          NA
## 10       1935   3000000.0          NA
## 11       1936    163245.0          NA
## 12       1937 184925485.0          NA
## 13       1938         NaN          NA
## 14       1939 110428945.0 124770876.7
## 15       1940  80350000.0   5586143.6
## 16       1941         NaN          NA
## 17       1942 102797150.0          NA
## 18       1943         NaN          NA
## 19       1944         NaN          NA
## 20       1945         NaN          NA
## 21       1946  22025000.0   2298097.0
## 22       1947      7927.0          NA
## 23       1948   2956000.0          NA
## 24       1949         NaN          NA
## 25       1950   8000000.0          NA
## 26       1951         NaN          NA
## 27       1952  36000000.0          NA
## 28       1953  20500000.0  21920310.2
## 29       1954   4934530.5   6597970.2
## 30       1955         NaN          NA
## 31       1956         NaN          NA
## 32       1957  27200000.0          NA
## 33       1958         NaN          NA
## 34       1959  25000000.0          NA
```

```
## 35       1960  32000000.0          NA
## 36       1961  43650000.0          NA
## 37       1962  11033517.5   7118468.7
## 38       1963  42950000.0  16728493.7
## 39       1964  38237907.2  43589425.6
## 40       1965  69310231.8  69300302.1
## 41       1966   6100000.0          NA
## 42       1967  43100000.0          NA
## 43       1968  36757685.5  28224429.5
## 44       1969  41711931.0  53699609.8
## 45       1970  10450000.0   4333205.1
## 46       1971  27247057.8  23659133.9
## 47       1972  67501217.5  95205895.8
## 48       1973 102919529.0  84904377.2
## 49       1974  55052942.9  35836220.8
## 50       1975 124409732.3 129830979.4
## 51       1976  71117623.5  65220168.6
## 52       1977 106290809.3 161470298.8
## 53       1978  71542234.6  64259622.5
## 54       1979  63579571.4  21999554.3
## 55       1980  57697266.7  72549435.3
## 56       1981  41460781.5  60609729.5
## 57       1982  75037552.2 104528364.2
## 58       1983  70192386.4  73489784.6
## 59       1984  62939598.7  58261286.3
## 60       1985  59223134.1  56550863.1
## 61       1986  44436464.0  53814705.1
## 62       1987  40233264.8  47920484.7
## 63       1988  41190351.8  41310713.3
## 64       1989  49678453.2  55312402.1
## 65       1990  78203971.2  70972957.6
## 66       1991  53844501.7  52051112.1
## 67       1992  63665195.1  56836221.4
## 68       1993  45302091.4  67219669.1
## 69       1994  59395666.2  76919182.2
## 70       1995  44909520.0  44557335.4
## 71       1996  42044174.3  51825099.0
## 72       1997  44793772.4  74553913.0
## 73       1998  38377008.0  46224415.6
## 74       1999  38072176.3  58690522.0
## 75       2000  42172627.6  49822040.5
## 76       2001  43255716.9  59509416.9
## 77       2002  43511151.5  62379549.1
## 78       2003  48727746.7  65612965.7
## 79       2004  40726529.1  60213382.0
## 80       2005  41159143.3  58742416.6
## 81       2006  39237856.0  57256445.4
## 82       2007  46267501.0  71147199.5
## 83       2008  44573509.4  65941087.5
## 84       2009  46207440.2  81700033.9
## 85       2010  49908326.0  71138520.5
## 86       2011  45785836.6  57726260.7
## 87       2012  62873527.7  97668976.8
## 88       2013  56158357.8  79283579.4
```

```
## 89       2014  62412136.9  72690931.5
## 90       2015  66530966.5  94728941.1
## 91       2016  76924035.9  93963814.2
## 92         NA    346434.7    174345.3
```

```
ddply(movieDF, ~ language, summarise, mean = mean(gross, na.rm = TRUE), sd = sd(gross, na.rm = TRUE))
```

```
##      language       mean         sd
## 1              1439760.3  1411148.6
## 2  Aboriginal 39340394.5 46916486.1
## 3      Arabic   840915.5   310668.1
## 4     Aramaic   499263.0         NA
## 5     Bosnian   301305.0         NA
## 6   Cantonese  6429425.3 10867327.3
## 7     Chinese    50000.0         NA
## 8       Czech   617228.0         NA
## 9      Danish   801285.7   769201.3
## 10       Dari  8462619.0 10373663.8
## 11      Dutch  1884888.3  2178427.8
## 12   Dzongkha   505295.0         NA
## 13    English 51025518.5 69527710.9
## 14   Filipino 10166502.0         NA
## 15     French  4852976.8 12082983.9
## 16     German  2916575.7  4107392.7
## 17      Greek   110197.0         NA
## 18     Hebrew  1088492.8  1165124.6
## 19      Hindi  2217130.2  3054481.7
## 20  Hungarian   195888.0         NA
## 21  Icelandic    11835.0         NA
## 22 Indonesian  2294672.0  2560364.4
## 23    Italian  4697477.3  5097473.8
## 24   Japanese  4768039.1  5826939.3
## 25    Kannada        NaN         NA
## 26     Kazakh    77231.0         NA
## 27     Korean  1100611.7   919114.6
## 28   Mandarin  9089529.4 29890950.8
## 29       Maya 50859889.0         NA
## 30  Mongolian  5701643.0         NA
## 31       None  2601847.0         NA
## 32  Norwegian   451137.2   510532.7
## 33    Panjabi        NaN         NA
## 34    Persian  2284408.0  3215505.2
## 35     Polish  1573547.0  1951075.6
## 36 Portuguese  2262183.3  3404429.4
## 37   Romanian  1185783.0         NA
## 38    Russian   723720.0   680553.7
## 39  Slovenian        NaN         NA
## 40    Spanish  8577084.2 12768836.4
## 41    Swahili        NaN         NA
## 42    Swedish    99390.0   126543.8
## 43      Tamil        NaN         NA
## 44     Telugu  6498000.0         NA
## 45       Thai  4153943.0  6715371.1
## 46       Urdu        NaN         NA
## 47 Vietnamese   638951.0         NA
```

```
## 48       Zulu  2912363.0         NA
```

```
ddply(movieDF, ~ title_year + language, summarise, mean = mean(gross, na.rm = TRUE), sd = sd(gross, na.
```

```
##     title_year   language        mean           sd
## 1         1916                    NaN           NA
## 2         1920              3000000.0           NA
## 3         1925                    NaN           NA
## 4         1927     German     26435.0           NA
## 5         1929    English   2808000.0           NA
## 6         1929     German      9950.0           NA
## 7         1930    English         NaN           NA
## 8         1932    English         NaN           NA
## 9         1933    English   2300000.0           NA
## 10        1934    English         NaN           NA
## 11        1935    English   3000000.0           NA
## 12        1936    English    163245.0           NA
## 13        1937    English 184925485.0           NA
## 14        1938    English         NaN           NA
## 15        1939    English 110428945.0 1.247709e+08
## 16        1940    English  80350000.0 5.586144e+06
## 17        1941    English         NaN           NA
## 18        1942    English 102797150.0           NA
## 19        1943    English         NaN           NA
## 20        1944    English         NaN           NA
## 21        1945    English         NaN           NA
## 22        1946    English  22025000.0 2.298097e+06
## 23        1947    English      7927.0           NA
## 24        1948    English   2956000.0           NA
## 25        1949    English         NaN           NA
## 26        1950    English   8000000.0           NA
## 27        1951    English         NaN           NA
## 28        1952    English  36000000.0           NA
## 29        1953    English  20500000.0 2.192031e+07
## 30        1954    English   9600000.0           NA
## 31        1954   Japanese    269061.0           NA
## 32        1955     Danish         NaN           NA
## 33        1955    English         NaN           NA
## 34        1956    English         NaN           NA
## 35        1957    English  27200000.0           NA
## 36        1958    English         NaN           NA
## 37        1959    English  25000000.0           NA
## 38        1960    English  32000000.0           NA
## 39        1961    English  43650000.0           NA
## 40        1962    English  11033517.5 7.118469e+06
## 41        1963    English  42950000.0 1.672849e+07
## 42        1964    English  45185488.6 4.486681e+07
## 43        1964     French         NaN           NA
## 44        1964    Italian   3500000.0           NA
## 45        1965    English  69310231.8 6.930030e+07
## 46        1965     French         NaN           NA
## 47        1966    English         NaN           NA
## 48        1966    Italian   6100000.0           NA
## 49        1967    English  43100000.0           NA
## 50        1967     German         NaN           NA
```

```
## 51        1968    English  36757685.5 2.822443e+07
## 52        1969    English  62554450.0 5.622128e+07
## 53        1969     French     26893.0           NA
## 54        1970    English  10450000.0 4.333205e+06
## 55        1970    Italian         NaN           NA
## 56        1971    English  27247057.8 2.365913e+07
## 57        1972    English  67501217.5 9.520590e+07
## 58        1972    Russian         NaN           NA
## 59        1972    Swedish         NaN           NA
## 60        1973    English 102919529.0 8.490438e+07
## 61        1974    English  55052942.9 3.583622e+07
## 62        1975    English 124409732.3 1.298310e+08
## 63        1976                    NaN           NA
## 64        1976    English  71117623.5 6.522017e+07
## 65        1977    English 106290809.3 1.614703e+08
## 66        1978    English  71542234.6 6.425962e+07
## 67        1979    English  63579571.4 2.199955e+07
## 68        1980    English  57697266.7 7.254944e+07
## 69        1981    English  46218251.0 6.314775e+07
## 70        1981     German  11433134.0           NA
## 71        1981    Italian    126387.0           NA
## 72        1981       None         NaN           NA
## 73        1982    English  75037552.2 1.045284e+08
## 74        1983    English  70192386.4 7.348978e+07
## 75        1983 Portuguese         NaN           NA
## 76        1984    English  62939598.7 5.826129e+07
## 77        1985    English  59223134.1 5.655086e+07
## 78        1985     French         NaN           NA
## 79        1986    English  44436464.0 5.381471e+07
## 80        1987    English  40233264.8 4.792048e+07
## 81        1988    English  42595565.3 4.130581e+07
## 82        1988   Japanese    439162.0           NA
## 83        1989    English  49678453.2 5.531240e+07
## 84        1990    English  78203971.2 7.097296e+07
## 85        1991    English  53844501.7 5.205111e+07
## 86        1992    English  65532597.4 5.664835e+07
## 87        1992    Spanish   2040920.0           NA
## 88        1993    English  47344257.9 6.804770e+07
## 89        1993     French    700000.0           NA
## 90        1993   Japanese     48856.0           NA
## 91        1994  Cantonese  11546543.0           NA
## 92        1994    English  60315841.6 7.737451e+07
## 93        1995  Cantonese  32333860.0           NA
## 94        1995    English  45739490.1 4.488169e+07
## 95        1995     French   1877179.0           NA
## 96        1996      Dutch         NaN           NA
## 97        1996    English  42768306.9 5.213510e+07
## 98        1996     French   1652472.0           NA
## 99        1996    Italian  15091542.0           NA
## 100       1997      Dutch    713413.0           NA
## 101       1997    English  47196092.0 7.583941e+07
## 102       1997     French    237941.0 9.226329e+03
## 103       1997   Japanese   1196393.5 1.558177e+06
## 104       1997    Persian    925402.0           NA
```

```
## 105       1998     Danish   1647780.0           NA
## 106       1998    English  40689267.3 4.681324e+07
## 107       1998     French   4809727.0 6.595404e+06
## 108       1998     German   7267324.0           NA
## 109       1998      Hindi    528972.0           NA
## 110       1998 Portuguese   5595428.0           NA
## 111       1998    Spanish   1286860.5 5.663225e+05
## 112       1999    English  38469614.4 5.893553e+07
## 113       1999     German    927107.0           NA
## 114       1999   Japanese  10037390.0           NA
## 115       2000    English  43148793.9 4.982566e+07
## 116       2000     French   3058380.0           NA
## 117       2000     German      5725.0           NA
## 118       2000      Hindi    610991.0           NA
## 119       2000   Japanese         NaN           NA
## 120       2000   Mandarin 128067808.0           NA
## 121       2000    Persian    673780.0           NA
## 122       2000    Spanish   3302547.5 2.943384e+06
## 123       2001  Cantonese    488872.0           NA
## 124       2001    English  45026700.8 6.040629e+07
## 125       2001     French  11163195.3 1.908597e+07
## 126       2001      Hindi  13876974.0           NA
## 127       2001   Japanese  10049886.0           NA
## 128       2001  Norwegian    313436.0           NA
## 129       2001    Spanish  13622333.0           NA
## 130       2001       Thai    454255.0           NA
## 131       2002 Aboriginal   6165429.0           NA
## 132       2002    English  45655622.6 6.327895e+07
## 133       2002     French   2575196.7 1.629945e+06
## 134       2002   Mandarin     84961.0 0.000000e+00
## 135       2002 Portuguese   7563397.0           NA
## 136       2002    Russian    181655.0           NA
## 137       2002    Spanish   2928009.0 3.933786e+06
## 138       2003       Dari   1127331.0           NA
## 139       2003   Dzongkha    505295.0           NA
## 140       2003    English  52179003.3 6.673501e+07
## 141       2003     French   3323112.2 2.472826e+06
## 142       2003     German   4063859.0           NA
## 143       2003    Italian    223878.0           NA
## 144       2003     Korean   2181290.0           NA
## 145       2003    Russian    502028.0           NA
## 146       2004    Aramaic    499263.0           NA
## 147       2004  Cantonese   8683075.0 1.190993e+07
## 148       2004    English  44641508.5 6.200122e+07
## 149       2004     French   2782842.0 2.689080e+06
## 150       2004     German   2798478.0 3.823273e+06
## 151       2004      Hindi   2921738.0           NA
## 152       2004   Japanese   2560421.5 3.040607e+06
## 153       2004     Korean   1110186.0           NA
## 154       2004   Mandarin  11041228.0           NA
## 155       2004    Russian   1487477.0           NA
## 156       2004    Spanish   2969222.3 3.199246e+06
## 157       2004    Swedish      9910.0           NA
## 158       2005 Aboriginal  72515360.0           NA
```

```
## 159       2005  Cantonese     47111.0           NA
## 160       2005    English  43013304.6 6.003906e+07
## 161       2005   Filipino  10166502.0           NA
## 162       2005     French  39231731.0 5.399649e+07
## 163       2005      Hindi   1635928.5 2.244256e+06
## 164       2005  Hungarian    195888.0           NA
## 165       2005     Kazakh     77231.0           NA
## 166       2005     Korean    211667.0           NA
## 167       2005   Mandarin    668171.0           NA
## 168       2005 Portuguese         NaN           NA
## 169       2005    Spanish  45356386.0           NA
## 170       2005       Thai  11905519.0           NA
## 171       2005       Zulu   2912363.0           NA
## 172       2006                252726.0           NA
## 173       2006     Arabic         NaN           NA
## 174       2006  Cantonese     49413.0           NA
## 175       2006      Czech    617228.0           NA
## 176       2006      Dutch   4398392.0           NA
## 177       2006    English  42202993.4 5.898179e+07
## 178       2006     French   2437905.5 3.421648e+06
## 179       2006     German  11284657.0           NA
## 180       2006     Hebrew    155972.0           NA
## 181       2006      Hindi   2736387.0 7.623403e+05
## 182       2006   Japanese  13753931.0           NA
## 183       2006     Korean   2201412.0           NA
## 184       2006   Mandarin   7306592.5 1.048070e+06
## 185       2006       Maya  50859889.0           NA
## 186       2006    Spanish  17404281.3 1.838522e+07
## 187       2006 Vietnamese    638951.0           NA
## 188       2006       Zulu         NaN           NA
## 189       2007                1066555.0           NA
## 190       2007     Arabic   1060591.0           NA
## 191       2007       Dari  15797907.0           NA
## 192       2007    English  50452509.0 7.322978e+07
## 193       2007     French   3689251.7 2.756393e+06
## 194       2007      Hindi    773358.3 4.908801e+05
## 195       2007   Mandarin    128978.0           NA
## 196       2007  Mongolian   5701643.0           NA
## 197       2007 Portuguese      8060.0           NA
## 198       2007   Romanian   1185783.0           NA
## 199       2007    Spanish   6595454.3 6.294459e+06
## 200       2007    Swedish         NaN           NA
## 201       2008  Cantonese    206678.0           NA
## 202       2008     Danish    145109.0           NA
## 203       2008      Dutch    542860.0           NA
## 204       2008    English  47279283.3 6.709260e+07
## 205       2008     French   3766595.0           NA
## 206       2008     German    476270.0           NA
## 207       2008     Hebrew   2283276.0           NA
## 208       2008      Hindi     55202.0           NA
## 209       2008   Japanese  15081783.0           NA
## 210       2008     Korean    128486.0           NA
## 211       2008   Mandarin    626809.0           NA
## 212       2008    Spanish     75727.0           NA
```

```
## 213       2008       Thai    102055.0           NA
## 214       2009     Arabic    621240.0           NA
## 215       2009    English  49890136.2 8.411377e+07
## 216       2009     French   5436584.2 8.183826e+06
## 217       2009     German   1248516.0 1.377629e+06
## 218       2009      Greek    110197.0           NA
## 219       2009      Hindi    199228.0           NA
## 220       2009    Italian   5004648.0           NA
## 221       2009     Korean         NaN           NA
## 222       2009   Mandarin    167084.7 4.084407e+04
## 223       2009  Norwegian     41709.0           NA
## 224       2009    Russian         NaN           NA
## 225       2009    Spanish  10097224.0 1.424141e+07
## 226       2010    English  52946887.2 7.233795e+07
## 227       2010     French   3317157.5 3.363350e+06
## 228       2010     German     59774.0           NA
## 229       2010      Hindi   2591899.5 2.017794e+06
## 230       2010    Italian         NaN           NA
## 231       2010   Japanese         NaN           NA
## 232       2010   Mandarin         NaN           NA
## 233       2010  Norwegian    252652.0           NA
## 234       2010    Russian         NaN           NA
## 235       2010    Spanish   5100937.0           NA
## 236       2010    Swedish    188870.0           NA
## 237       2011    Bosnian    301305.0           NA
## 238       2011    English  48751263.5 5.844605e+07
## 239       2011     French    391514.5 3.114020e+05
## 240       2011      Hindi    563699.0           NA
## 241       2011 Indonesian   4105123.0           NA
## 242       2011   Mandarin     68325.0 8.359699e+04
## 243       2011       None   2601847.0           NA
## 244       2011  Norwegian   1196752.0           NA
## 245       2011    Persian   3769225.0 4.708295e+06
## 246       2011    Spanish   1391770.0           NA
## 247       2012     Danish    610968.0           NA
## 248       2012    English  66054322.1 9.917850e+07
## 249       2012     French    148409.5 1.088485e+05
## 250       2012     Hebrew   2408553.0           NA
## 251       2012      Hindi   1700111.5 1.905550e+06
## 252       2012 Indonesian    484221.0           NA
## 253       2012    Russian         NaN           NA
## 254       2012    Spanish   5782159.5 1.599171e+05
## 255       2013     Arabic         NaN           NA
## 256       2013    Chinese     50000.0           NA
## 257       2013    English  60320093.7 8.096044e+07
## 258       2013     German    100412.0           NA
## 259       2013      Hindi   2718067.3 2.280404e+06
## 260       2013  Icelandic     11835.0           NA
## 261       2013    Italian   2835886.0           NA
## 262       2013   Japanese     22770.0           NA
## 263       2013    Kannada         NaN           NA
## 264       2013   Mandarin   6594136.0           NA
## 265       2013     Polish   3826455.0           NA
## 266       2013    Russian         NaN           NA
```

```
## 267       2013    Spanish  11160249.0 2.219764e+07
## 268       2014                    NaN           NA
## 269       2014  Cantonese    129115.0           NA
## 270       2014    English  64625314.8 7.309849e+07
## 271       2014     French    231186.0           NA
## 272       2014      Hindi         NaN           NA
## 273       2014   Mandarin    377420.0           NA
## 274       2014 Portuguese     15246.0 7.093695e+03
## 275       2014    Russian         NaN           NA
## 276       2014    Spanish  17382982.0           NA
## 277       2014    Swahili         NaN           NA
## 278       2015  Cantonese   2126511.0           NA
## 279       2015    Chinese         NaN           NA
## 280       2015    English  69373779.9 9.578730e+07
## 281       2015     French         NaN           NA
## 282       2015     Hebrew     34151.0           NA
## 283       2015   Mandarin    342984.5 3.826459e+05
## 284       2015    Panjabi         NaN           NA
## 285       2015 Portuguese    375723.0           NA
## 286       2015   Romanian         NaN           NA
## 287       2015    Russian         NaN           NA
## 288       2015  Slovenian         NaN           NA
## 289       2015    Spanish         NaN           NA
## 290       2015      Tamil         NaN           NA
## 291       2015     Telugu   6498000.0           NA
## 292       2015       Urdu         NaN           NA
## 293       2016                    NaN           NA
## 294       2016    English  79042326.6 9.439038e+07
## 295       2016     French         NaN           NA
## 296       2016     Hebrew    560512.0           NA
## 297       2016      Hindi         NaN           NA
## 298       2016   Japanese         NaN           NA
## 299       2016     Korean    770629.0           NA
## 300         NA                    NaN           NA
## 301         NA    English    145118.0           NA
## 302         NA     French         NaN           NA
## 303         NA  Icelandic         NaN           NA
## 304         NA    Italian         NaN           NA
## 305         NA   Japanese         NaN           NA
## 306         NA     Polish    447093.0 0.000000e+00
## 307         NA    Swedish         NaN           NA
```