ECON 424 - Final Project (Winter 2023)

Submitted by Raphael Shawn Gozali (20805288) and Jason Tedjosoesilo (20801292)

Questions

For this project, these are the questions that we will be answering:

1. Without any interactions, which covariate(s) affect revenue the most?
2. Are there any certain genres, actors, or keywords that affect revenue the most?
3. If we do interactions between genres, actors, and keywords, are there any particular combination(s) that affect revenue the most?
4. Do people's preferences on movie genres and actors change over the years/decades?

Data

In this project, we will be using the Movies dataset from Kaggle: https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies?resource=download. We downloaded this dataset on April 2, 2023 on 2pm EST and continue to work with it locally. This dataset consists of around 723 thousand rows of movies, with their descriptions and details separated into columns. The response variable that we are focusing on from this dataset is the "revenue" column, which shows us the revenue of the movie. The covariates that might be useful from this dataset includes: date published, budget, genres, casts, keywords, and runtime.

Here is the first few rows of the data:

```
data <- read.csv("./movies.csv")
head(data)
```

```
##       id                          title                           genres
## 1  76600       Avatar: The Way of Water Science Fiction-Adventure-Action
## 2 631842             Knock at the Cabin          Horror-Mystery-Thriller
## 3 646389                          Plane        Action-Adventure-Thriller
## 4 505642 Black Panther: Wakanda Forever Action-Adventure-Science Fiction
## 5 956101              The Eighth Clause                         Thriller
## 6 603692            John Wick: Chapter 4            Action-Thriller-Crime
##   original_language
## 1                en
## 2                en
## 3                en
## 4                en
## 5                la
## 6                en
##
## 1                                                        Set more than a decade after th
```

```
## 2                                        While vacationing at a remote cabin a young girl and her t
## 3
## 4 Queen Ramonda Shuri M'Baku Okoye and the Dora Milaje fight to protect their nation from intervening
## 5
## 6                                                                  With the price on his h
##   popularity
## 1  10255.685
## 2   3422.537
## 3   2618.646
## 4   2525.408
## 5   2259.303
## 6   2252.114
##                                                                            production_companies
## 1                                                  20th Century Studios-Lightstorm Entertainment
## 2 Blinding Edge Pictures-Universal Pictures-FilmNation Entertainment-Wishmore-Perfect World Pictures
## 3             MadRiver Pictures-Di Bonaventura Pictures-G-BASE-Olive Hill Media-Riverstone Pictures
## 4                                                                                   Marvel Studios
## 5                                                                    SDB Films-El Hombre Orquesta
## 6                      Thunder Road-87Eleven-Lionsgate-Summit Entertainment-El-Torky Art Production
##   release_date  budget    revenue runtime   status
## 1   2022-12-14 4.6e+08 2309660236     192 Released
## 2   2023-02-01 2.0e+07   52000000     100 Released
## 3   2023-01-12 2.5e+07   51000000     107 Released
## 4   2022-11-09 2.5e+08  858535561     162 Released
## 5   2022-04-29 0.0e+00          0       0 Released
## 6   2023-03-22 9.0e+07          0     169 Released
##                                           tagline vote_average vote_count
## 1                              Return to Pandora.        7.739       6227
## 2 Save your family or save humanity. Make the choice.    6.457        888
## 3                    Survive together or die alone.        6.901        785
## 4                                          Forever.        7.338       3922
## 5                                                          4.600         10
## 6                        No way back. One way out.        8.319        202
##
## 1
## 2
## 3
## 4 Letitia Wright-Lupita Nyong'o-Danai Gurira-Winston Duke-Dominique Thorne-Tenoch Huerta Mejía-Angela
## 5
## 6
##
## 1             loss of loved one-dying and death-alien life-form-resurrection-sequel-dysfunctional fam
## 2 based on novel or book-sacrifice-cabin-faith-end of the world-apocalypse-home invasion-lgbt-afterc
## 3
## 4                                 loss of loved one-hero-sequel-superhero-based on comic-mourning-c
## 5
## 6                                                                  new york city-martial arts-hitm
##                      poster_path                    backdrop_path
## 1 /t6HIqrRAclMCA60NsSmeqe9RmNV.jpg /ovM06PdF3M8wvKb06i4sjW3xoww.jpg
## 2 /dm06L9pxDOL9jNSK4Cb6y139rrG.jpg /zWDMQX0sPaW2u0N2pJaYA8bVVaJ.jpg
## 3 /qi9r5xBgcc9KTxlOLjssEbDgO0J.jpg /9Rq14Eyrf7Tu1xk0Pl7VcNbNh1n.jpg
## 4 /sv1xJUazXeYqALzczSZ3O6nkH75.jpg /xDMIl84Qo5Tsu62c9DGWhmPI67A.jpg
## 5 /8tc8eMFAX2SDC1TRu987qFQy8Cl.jpg /kLnqNE9Af5QHyvUxw8cDGhF1ilv.jpg
## 6  /vZloFAK7NmvMGKE7VkF5UHaz0I.jpg /i8dshLvq4LE3s0v8PrkDdUyb1ae.jpg
```

```
##
## 1    183392-111332-702432-505642-1064215-436270-874764-613200-315162-965839-1013870-100287-758009-10
## 2 1058949-646389-772515-505642-143970-667216-1048522-785084-1058617-986054-640146-937278-1001500-717
## 3                                        505642-758769-864692-631842-1058949-925943-758009-315162-6157
## 4               436270-829280-76600-56969-312634-1037858-238-551271-22023-736526-899112-468073-63285
## 5
## 6
```

Before we do some analysis, we will clean the data. First, we will remove unnecessary columns and also data with zero revenues. We are only keeping released movies above 40 minutes as we are not including short movies in the data. This is based on the definition from the Academy of Motion Picture Arts and Sciences, where they define a short film as "an original motion picture that has a running time of 40 minutes or less, including all credits". We also do not include the movies with runtime of value 999. There might also be duplicates in the data, so we need to remove duplicates as well. Since people can add random movies into this database, we try as best as we can to filter those out. To make sure that a movie is legitimate, we filter the movies that do not have any genres, production companies, and credits (all three are empty values).

```r
# Only movies that has revenue, is released, and more than
# 40 minutes long Also remove if runtime = 999 and remove
# the movies with missing genres, production companies, and
# credits
clean_data <- data[data$revenue > 0 & data$status == "Released" &
    data$runtime > 40 & data$runtime != 999 & !(data$genres ==
    "" & data$production_companies == "" & data$credits == ""),
    !names(data) %in% c("overview", "popularity", "status", "tagline",
        "vote_average", "vote_count", "poster_path", "backdrop_path",
        "recommendations")]


# removing empty ID rows
clean_data <- clean_data[!is.na(clean_data$id), ]

# resetting row.names
row.names(clean_data) <- NULL

# remove duplicate data
clean_data <- clean_data[!duplicated(clean_data[, 1]), ]

write.csv(clean_data, file = "./movies_filtered.csv", row.names = FALSE)

n <- length(clean_data[, 1])
```

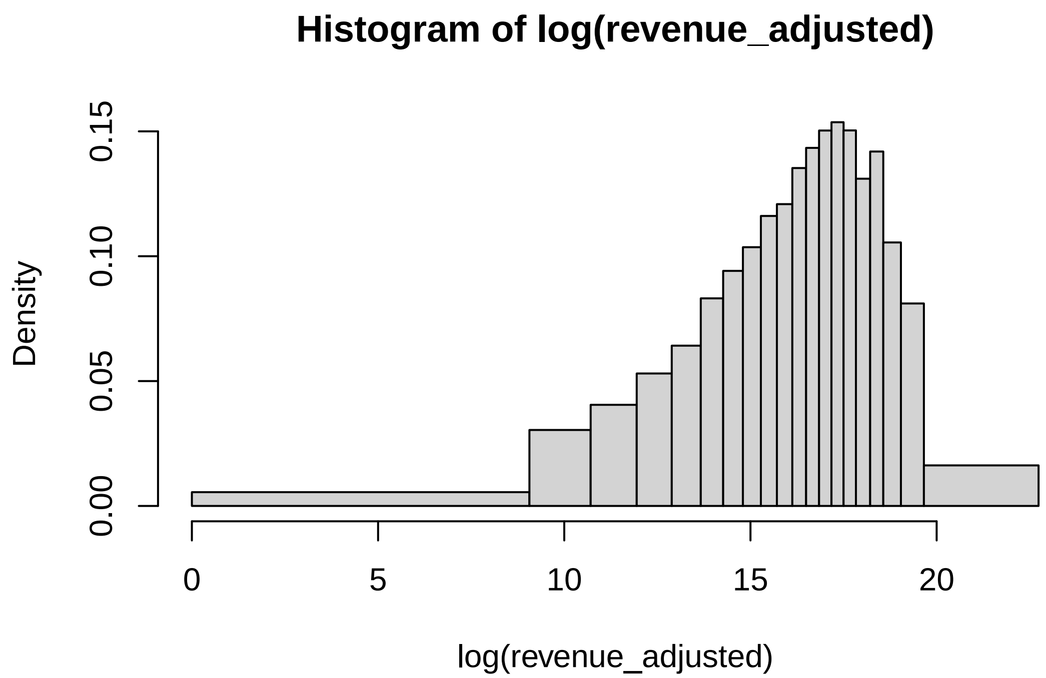
We took this data and use a Python API to calculate the adjusted revenues by adding inflation factors. Here are the Python code below.

```python
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
import requests
import json
from tqdm import tqdm
from pathlib import Path
```

```python
# get movies from movies_filtered.csv
movies = pd.read_csv(r'movies_filtered.csv')
n = len(movies)

# Function to get the inflation rate from the API
# Input: start_time (based on release_date)
# Output: inflation rate (inflated to April 3, 2023)
def get_dollar(start_time):
    api_url = "https://www.statbureau.org/calculate-inflation-price-jsonp?jsoncallback=?"

    headers = {'Content-type': 'application/json'}

    payload = {
        "country": "united-states",
        "start": str(start_time),
        "end": "2023/04/03",
        "amount": "1",
        "format": True
    }
    response = requests.post(api_url,  data=json.dumps(payload), headers=headers)
    my_bytes = response.content

    amount_s = my_bytes.decode('utf8').replace("'", '"')
    amount_s = amount_s[4:-2]
    return float(amount_s)

# getting the inflation rate of each movie
inflation = [0] * n
failed = []
for i in tqdm(range(n)):
    try:
        inflation[i] = get_dollar(movies.iloc[i]['release_date'])
    except:
        failed.append(i)

# data with no release_date will be given inflation = 1 (no inflation)
for fail in failed:
    inflation[fail] = 1

# getting the adjusted revenue for each movie
revenue_adjusted_c = movies['revenue'] * np.array(inflation)
df2 = movies.assign(revenue_adjusted=revenue_adjusted_c)

# exporting the file movies_adjusted.csv
filepath = Path('movies_adjusted.csv')
filepath.parent.mkdir(parents=True, exist_ok=True)
df2.to_csv(filepath)
```

We calculated all the revenues inflated to April 3, 2023 and include them in the data as a new column called “revenue_adjusted”. The data that do not have release_date will have their adjusted revenue be the same as their revenue.

```r
clean_data <- read.csv("./movies_adjusted.csv")
n <- length(clean_data[, 1])

hist(log(clean_data$revenue_adjusted), breaks = quantile(log(clean_data$revenue_adjusted),
    p = seq(0, 1, length.out = 21)), freq = FALSE, xlab = "log(revenue_adjusted)",
    main = "Histogram of log(revenue_adjusted)")
```



From the histogram above, we can see that when we split the dataset into 20 bins of 5% quantiles each, most of them have high values on revenues after being adjusted to inflation.

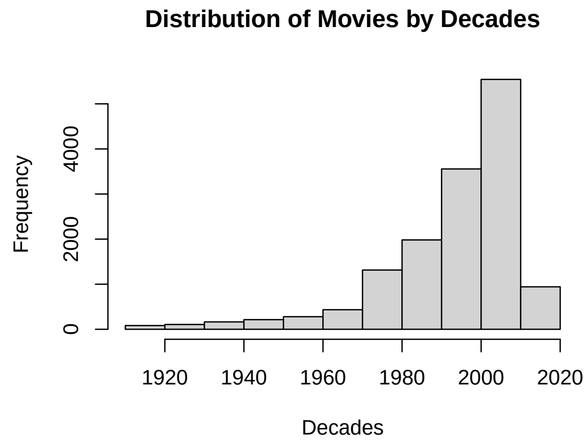
After cleaning the data, we have a total of 14,678 rows of movies.

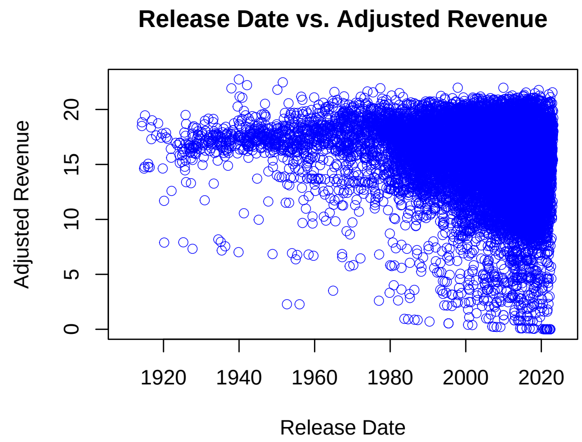
Limitations

Before we start with our methods, we would like to address the limitation of the methods that we are using here.

```r
clean_data$year <- format(as.Date(clean_data$release_date), format = "%Y")
clean_data$decades <- floor(as.numeric(clean_data$year)/10) *
    10
df_list <- split(clean_data, clean_data$decades)
```

```r
par(mfrow = c(1, 2))
hist(clean_data$decades, xlab = "Decades", main = "Distribution of Movies by Decades")
plot(as.Date(clean_data$release_date), log(clean_data$revenue_adjusted),
    xlim = c(as.Date("1910-01-01"), as.Date("2025-01-01")), col = "blue",
    lwd = 0.5, xlab = "Release Date", ylab = "Adjusted Revenue",
    main = "Release Date vs. Adjusted Revenue")
```





Based on the histogram above, we can see that the distribution of the movies are not equal across time. There are more recent movies in the dataset than the older movies. Also, from the scatterplot above, it is visible that the variance of the data changes over time. This shows that our data is non-stationary. In addition, we do not see a lot of data points in the early decades, and not a lot of them have low revenues. Since this is a problem, we do not use time as a variable in our model. We do not change the distribution of the data since it is better to have more data on recent decades. This gives us a notion of weights on the data, with having more movies closer to the present.

Other than that, note that the data cleaning process is not perfect. Since the data source is publicly available to be added, there are some made-up data points in the raw data. We cleaned it as best as we could, but there is no guarantee on whether there are still some made-up data in this dataset.

Methods & Results

To answer the questions above, we will create some models using LASSO method via cv.glmnet function. We use LASSO since we will have a lot of variables coming from the text columns. Thus, we use LASSO to do variable reduction to obtain a balanced model. To build the model, we will have to do text analysis on the texts first. There are five columns which has text variables: production_companies, genres, credits, keywords, and original_language. The terms that we are using from these columns are separated by dashes ("-"). Using this information, we will separate these text columns into a sparse matrix full of terms. For each document, the term will be equal to 1 if it appears on that document, and 0 if it does not appear. After getting each term separated, we added single letter prefixes with "$" on the terms to be able to tell which column the terms come from. Then, we filter the sparse matrix so that only terms that appear in at least 50 movies is included to be in the model.

```
library(pdftools)
```

```
## Using poppler version 22.04.0
```

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(SnowballC)

to_another <- content_transformer(function(x, y, z) gsub(y, z,
```

```r
    x))

add_pre <- function(x, pre) {
    ifelse(!is.na(x) & x != "" & nchar(x) > 0 & x != " ", return(paste(pre,
        x, sep = "")), return(x))
}

add_prefix <- content_transformer(add_pre)
```

For terms coming from production_companies column, we have to change "Metro-Goldwyn-Mayer" to "Metro_Goldwyn_Mayer" due to the name having dashes in it. This is the only company that appears in at least 50 movies that contains "-" in its name. We added "p$" to indicate production_companies terms. From here, we got a result of 71 production companies.

```r
#### PRODUCTION COMPANIES
prod_comp <- clean_data$production_companies
docs <- Corpus(VectorSource(prod_comp))
# Remove '-' from the name since it is the separator of the
# data
docs <- tm_map(docs, to_another, "Metro-Goldwyn-Mayer", "Metro_Goldwyn_Mayer")
docs <- tm_map(docs, to_another, "no_production_companies", "")
docs <- tm_map(docs, to_another, " ", "_")
docs <- tm_map(docs, to_another, "-", " ")
docs <- tm_map(docs, stripWhitespace)
# adding prefix p$ for production_companies
docs <- tm_map(docs, add_prefix, "p$")
docs <- tm_map(docs, to_another, " ", " p$")
dtm <- DocumentTermMatrix(docs)

# get production companies in at least 50 movies
key_pc <- sort(findFreqTerms(dtm, 50))
key_pc
```

```
##  [1] "p$20th_century_fox"              "p$amblin_entertainment"
##  [3] "p$arte_france_cinéma"            "p$atresmedia"
##  [5] "p$bbc_film"                      "p$blumhouse_productions"
##  [7] "p$canal+"                        "p$canal+_españa"
##  [9] "p$cannon_group"                  "p$castle_rock_entertainment"
## [11] "p$ciné+"                         "p$cj_entertainment"
## [13] "p$cnc"                           "p$columbia_pictures"
## [15] "p$constantin_film"               "p$dentsu"
## [17] "p$dimension_films"               "p$dreamworks_pictures"
## [19] "p$dune_entertainment"            "p$europacorp"
## [21] "p$film_i_väst"                   "p$film4_productions"
## [23] "p$filmnation_entertainment"      "p$focus_features"
## [25] "p$fox_2000_pictures"             "p$fox_searchlight_pictures"
## [27] "p$france_2_cinéma"               "p$france_3_cinéma"
## [29] "p$gaumont"                       "p$hollywood_pictures"
## [31] "p$imagine_entertainment"         "p$ingenious_media"
## [33] "p$lakeshore_entertainment"       "p$lionsgate"
## [35] "p$malpaso_productions"           "p$metro_goldwyn_mayer"
## [37] "p$millennium_films"              "p$miramax"
## [39] "p$morgan_creek_productions"      "p$new_line_cinema"
```

```
## [41] "p$new_regency_pictures"          "p$orion_pictures"
## [43] "p$paramount"                     "p$participant"
## [45] "p$pathé"                         "p$polygram_filmed_entertainment"
## [47] "p$regency_enterprises"           "p$relativity_media"
## [49] "p$scott_free_productions"        "p$scott_rudin_productions"
## [51] "p$screen_gems"                   "p$silver_pictures"
## [53] "p$sony_pictures"                 "p$studiocanal"
## [55] "p$summit_entertainment"          "p$téléfilm_canada"
## [57] "p$tf1_films_production"          "p$the_weinstein_company"
## [59] "p$toei_company"                  "p$toho"
## [61] "p$touchstone_pictures"           "p$tristar_pictures"
## [63] "p$tsg_entertainment"             "p$united_artists"
## [65] "p$universal_pictures"            "p$village_roadshow_pictures"
## [67] "p$walt_disney_pictures"          "p$warner_bros._pictures"
## [69] "p$wild_bunch"                    "p$working_title_films"
## [71] "p$zdf"
```

For terms coming from genres column, we added "g$" to indicate genre terms. We decided to use all genres since there are only 19 of them. The only genre with less than 50 movies is "tv_movie", but we decided to include it as a term.

```
#### GENRES
genres <- clean_data$genres
docs2 <- Corpus(VectorSource(genres))
docs2 <- tm_map(docs2, to_another, " ", "_")
docs2 <- tm_map(docs2, to_another, "-", " ")
docs2 <- tm_map(docs2, stripWhitespace)
# adding prefix g$ for genres
docs2 <- tm_map(docs2, add_prefix, "g$")
docs2 <- tm_map(docs2, to_another, " ", " g$")
dtm2 <- DocumentTermMatrix(docs2)

# get all genres
key_gen <- sort(findFreqTerms(dtm2, 0))
key_gen
```

```
##  [1] "g$action"          "g$adventure"       "g$animation"
##  [4] "g$comedy"          "g$crime"           "g$documentary"
##  [7] "g$drama"           "g$family"          "g$fantasy"
## [10] "g$history"         "g$horror"          "g$music"
## [13] "g$mystery"         "g$romance"         "g$science_fiction"
## [16] "g$thriller"        "g$tv_movie"        "g$war"
## [19] "g$western"
```

For terms coming from credits column, there are a lot of names that contains a dash symbol. The names of Korean casts can be dealt with since they follow a regex pattern, as seen in the code. However, this does not cover all names. It is hard to clean this, so we decided to remove one-word names from the results. We added "c$" to indicate cast terms. From this column, we managed to get 67 terms.

```
#### CREDITS
casts <- clean_data$credits
docs3 <- Corpus(VectorSource(casts))
docs3 <- tm_map(docs3, to_another, " ", "_")
```

```r
# deal with Korean names
docs3 <- tm_map(docs3, to_another, "_([[:alpha:]]+)-([[:lower:]]+)$",
    "_\\1_\\2")
docs3 <- tm_map(docs3, to_another, "_([[:alpha:]]+)-([[:lower:]]+)-",
    "_\\1_\\2-")
docs3 <- tm_map(docs3, to_another, "-", " ")
docs3 <- tm_map(docs3, stripWhitespace)
# adding prefix c$ for casts
docs3 <- tm_map(docs3, add_prefix, "c$")
docs3 <- tm_map(docs3, to_another, " ", " c$")
dtm3 <- DocumentTermMatrix(docs3)

# get all casts in at least 50 movies
key_cast <- sort(findFreqTerms(dtm3, 50))

# hard to remove '-' from two-worded names separated by '-'
# so we remove single-word names that comes from them
key_cast <- key_cast[grepl("_", key_cast)]
key_cast
```

```
##  [1] "c$alec_baldwin"         "c$alfred_molina"       "c$anthony_hopkins"
##  [4] "c$antonio_banderas"     "c$ben_kingsley"        "c$ben_stiller"
##  [7] "c$bess_flowers"         "c$bill_murray"         "c$brad_pitt"
## [10] "c$brian_cox"            "c$bruce_willis"        "c$cate_blanchett"
## [13] "c$christopher_plummer" "c$christopher_walken"  "c$clint_eastwood"
## [16] "c$danny_glover"         "c$danny_trejo"         "c$dennis_quaid"
## [19] "c$donald_sutherland"    "c$ethan_hawke"         "c$forest_whitaker"
## [22] "c$frank_welker"         "c$gene_hackman"        "c$harrison_ford"
## [25] "c$harry_dean_stanton"   "c$harvey_keitel"       "c$j.k._simmons"
## [28] "c$james_franco"         "c$joe_chrest"          "c$john_cusack"
## [31] "c$john_goodman"         "c$john_hurt"           "c$john_leguizamo"
## [34] "c$john_turturro"        "c$johnny_depp"         "c$julianne_moore"
## [37] "c$keanu_reeves"         "c$keith_david"         "c$kevin_bacon"
## [40] "c$liam_neeson"          "c$m._emmet_walsh"      "c$matt_damon"
## [43] "c$meryl_streep"         "c$michael_caine"       "c$michael_papajohn"
## [46] "c$morgan_freeman"       "c$nicolas_cage"        "c$nicole_kidman"
## [49] "c$owen_wilson"          "c$paul_giamatti"       "c$richard_jenkins"
## [52] "c$robert_de_niro"       "c$robert_downey_jr."   "c$robert_duvall"
## [55] "c$robin_williams"       "c$samuel_l._jackson"   "c$sigourney_weaver"
## [58] "c$stanley_tucci"        "c$stephen_root"        "c$stephen_tobolowsky"
## [61] "c$steve_buscemi"        "c$susan_sarandon"      "c$sylvester_stallone"
## [64] "c$thomas_rosales_jr."   "c$tom_hanks"           "c$willem_dafoe"
## [67] "c$woody_harrelson"
```

For terms coming from keywords column, we added "k$" to indicate keyword terms. From this column, we managed to get 312 terms.

```r
#### KEYWORDS
keywords <- clean_data$keywords
docs4 <- Corpus(VectorSource(keywords))
docs4 <- tm_map(docs4, to_another, " ", "_")
docs4 <- tm_map(docs4, to_another, "-", " ")
docs4 <- tm_map(docs4, stripWhitespace)
```

```r
# adding prefix k$ for keywords
docs4 <- tm_map(docs4, add_prefix, "k$")
docs4 <- tm_map(docs4, to_another, " ", " k$")
dtm4 <- DocumentTermMatrix(docs4)

# get all keywords in at least 50 movies
key_keys <- sort(findFreqTerms(dtm4, 50))
key_keys
```

```
##   [1] "k$1920s"                          "k$1930s"
##   [3] "k$1940s"                          "k$1950s"
##   [5] "k$1960s"                          "k$1970s"
##   [7] "k$1980s"                          "k$1990s"
##   [9] "k$19th_century"                   "k$action_hero"
##  [11] "k$adultery"                       "k$africa"
##  [13] "k$aftercreditsstinger"            "k$airplane"
##  [15] "k$alcohol"                        "k$alcoholic"
##  [17] "k$alcoholism"                     "k$alien"
##  [19] "k$alien_invasion"                 "k$amnesia"
##  [21] "k$animal"                         "k$anime"
##  [23] "k$anthropomorphism"               "k$anti_hero"
##  [25] "k$apocalyptic_future"             "k$army"
##  [27] "k$artificial_intelligence"        "k$assassin"
##  [29] "k$assassination"                  "k$australia"
##  [31] "k$author"                         "k$baby"
##  [33] "k$bank_robbery"                   "k$baseball"
##  [35] "k$based_on_children's_book"       "k$based_on_comic"
##  [37] "k$based_on_manga"                 "k$based_on_novel_or_book"
##  [39] "k$based_on_play_or_musical"       "k$based_on_short_story"
##  [41] "k$based_on_true_story"            "k$based_on_video_game"
##  [43] "k$based_on_young_adult_novel"     "k$battle"
##  [45] "k$beach"                          "k$best_friend"
##  [47] "k$betrayal"                       "k$biography"
##  [49] "k$black_and_white"                "k$blackmail"
##  [51] "k$bomb"                           "k$brother"
##  [53] "k$brutality"                      "k$buddy_cop"
##  [55] "k$bullying"                       "k$california"
##  [57] "k$cancer"                         "k$car_crash"
##  [59] "k$castle"                         "k$cat"
##  [61] "k$chase"                          "k$chicago_illinois"
##  [63] "k$child_abuse"                    "k$china"
##  [65] "k$christmas"                      "k$church"
##  [67] "k$cia"                            "k$code"
##  [69] "k$college"                        "k$coming_of_age"
##  [71] "k$competition"                    "k$concert"
##  [73] "k$conspiracy"                     "k$cop"
##  [75] "k$corruption"                     "k$creature"
##  [77] "k$criminal"                       "k$cult_film"
##  [79] "k$dance"                          "k$dark_comedy"
##  [81] "k$daughter"                       "k$death"
##  [83] "k$demon"                          "k$depression"
##  [85] "k$desert"                         "k$detective"
##  [87] "k$disaster"                       "k$divorce"
```

```
##  [89] "k$doctor"                        "k$dog"
##  [91] "k$dragon"                        "k$dream"
##  [93] "k$drug_addiction"                "k$drug_dealer"
##  [95] "k$drugs"                         "k$duringcreditsstinger"
##  [97] "k$dying_and_death"               "k$dysfunctional_family"
##  [99] "k$dystopia"                      "k$england"
## [101] "k$epic"                          "k$escape"
## [103] "k$ex"                            "k$experiment"
## [105] "k$explosion"                     "k$extramarital_affair"
## [107] "k$fairy_tale"                    "k$faith"
## [109] "k$falling_in_love"               "k$family"
## [111] "k$family_relationships"          "k$father"
## [113] "k$father_daughter_relationship" "k$father_son_relationship"
## [115] "k$fbi"                           "k$female_friendship"
## [117] "k$female_protagonist"            "k$female_wrestler"
## [119] "k$fight"                         "k$film_noir"
## [121] "k$fire"                          "k$flashback"
## [123] "k$florida"                       "k$forest"
## [125] "k$found_footage"                 "k$france"
## [127] "k$friends"                       "k$friendship"
## [129] "k$funeral"                       "k$future"
## [131] "k$gambling"                      "k$gang"
## [133] "k$gangster"                      "k$gay"
## [135] "k$gay_interest"                  "k$ghost"
## [137] "k$giant_monster"                 "k$good_versus_evil"
## [139] "k$gore"                          "k$grief"
## [141] "k$gun"                           "k$gunfight"
## [143] "k$hallucination"                 "k$haunted_house"
## [145] "k$heist"                         "k$helicopter"
## [147] "k$hero"                          "k$high_school"
## [149] "k$hitman"                        "k$holiday"
## [151] "k$hollywood"                     "k$horror"
## [153] "k$horse"                         "k$hospital"
## [155] "k$hostage"                       "k$hotel"
## [157] "k$husband_wife_relationship"     "k$in"
## [159] "k$infidelity"                    "k$investigation"
## [161] "k$island"                        "k$japan"
## [163] "k$jealousy"                      "k$journalist"
## [165] "k$jungle"                        "k$kidnapping"
## [167] "k$killer"                        "k$kung_fu"
## [169] "k$las_vegas"                     "k$lawyer"
## [171] "k$lgbt"                          "k$lgbt_interest"
## [173] "k$live_action_and_animation"     "k$london_england"
## [175] "k$los_angeles_california"        "k$loss_of_loved_one"
## [177] "k$love"                          "k$love_of_one's_life"
## [179] "k$love_triangle"                 "k$mafia"
## [181] "k$magic"                         "k$male_friendship"
## [183] "k$male_homosexuality"            "k$manhattan_new_york_city"
## [185] "k$marijuana"                     "k$marriage"
## [187] "k$martial_arts"                  "k$mental_illness"
## [189] "k$mexico"                        "k$military"
## [191] "k$money"                         "k$monster"
## [193] "k$mother_daughter_relationship" "k$mother_son_relationship"
## [195] "k$motorcycle"                    "k$movie_business"
```

```
## [197] "k$murder"                      "k$musical"
## [199] "k$musician"                    "k$nazi"
## [201] "k$neighbor"                    "k$neo"
## [203] "k$new_love"                    "k$new_york_city"
## [205] "k$nightclub"                   "k$nightmare"
## [207] "k$noir"                        "k$obsession"
## [209] "k$organized_crime"             "k$orphan"
## [211] "k$paranoia"                    "k$parent_child_relationship"
## [213] "k$paris_france"                "k$parody"
## [215] "k$period_drama"                "k$pets"
## [217] "k$police"                      "k$police_officer"
## [219] "k$politics"                    "k$post"
## [221] "k$pregnancy"                   "k$priest"
## [223] "k$princess"                    "k$prison"
## [225] "k$pro_wrestling"               "k$prostitute"
## [227] "k$psychological_thriller"      "k$psychopath"
## [229] "k$racism"                      "k$rape"
## [231] "k$relationship"                "k$religion"
## [233] "k$remake"                      "k$rescue"
## [235] "k$restaurant"                  "k$revenge"
## [237] "k$rivalry"                     "k$road_trip"
## [239] "k$robbery"                     "k$robot"
## [241] "k$romance"                     "k$romantic_comedy"
## [243] "k$rural_area"                  "k$sadism"
## [245] "k$san_francisco_california"    "k$satire"
## [247] "k$school"                      "k$scientist"
## [249] "k$secret_agent"                "k$secret_identity"
## [251] "k$seduction"                   "k$self"
## [253] "k$sequel"                      "k$serial_killer"
## [255] "k$sheriff"                     "k$ship"
## [257] "k$shootout"                    "k$showdown"
## [259] "k$sibling_relationship"        "k$silent_film"
## [261] "k$singer"                      "k$single_mother"
## [263] "k$slasher"                     "k$small_town"
## [265] "k$snow"                        "k$soldier"
## [267] "k$space"                       "k$space_travel"
## [269] "k$spacecraft"                  "k$spoof"
## [271] "k$sports"                      "k$spy"
## [273] "k$street_gang"                 "k$suicide"
## [275] "k$suicide_attempt"             "k$summer"
## [277] "k$super_power"                 "k$superhero"
## [279] "k$supernatural"                "k$surrealism"
## [281] "k$survival"                    "k$sword_fight"
## [283] "k$teacher"                     "k$teen_movie"
## [285] "k$teenage_girl"                "k$teenager"
## [287] "k$terrorism"                   "k$terrorist"
## [289] "k$texas"                       "k$thief"
## [291] "k$time_travel"                 "k$torture"
## [293] "k$train"                       "k$transformation"
## [295] "k$travel"                      "k$undercover"
## [297] "k$up"                          "k$usa_president"
## [299] "k$vampire"                     "k$vigilante"
## [301] "k$village"                     "k$villain"
## [303] "k$wedding"                     "k$whodunit"
```

```
## [305] "k$widow"                          "k$winter"
## [307] "k$witch"                          "k$woman_director"
## [309] "k$world_war_ii"                   "k$wrestling"
## [311] "k$writer"                         "k$zombie"
```

For terms coming from original_language column, we added "l$" to indicate language terms. From this column, we managed to get 20 terms.

```
#### ORIGINAL LANGUAGE
og_lng <- clean_data$original_language
docs5 <- Corpus(VectorSource(og_lng))
# adding prefix l$ for language
docs5 <- tm_map(docs5, add_prefix, "l$")
docs5 <- tm_map(docs5, to_another, " ", " l$")
dtm5 <- DocumentTermMatrix(docs5)

# get all languages in at least 50 movies
key_lang <- sort(findFreqTerms(dtm5, 50))
key_lang
```

```
##  [1] "l$ar" "l$cn" "l$de" "l$en" "l$es" "l$fa" "l$fr" "l$hi" "l$it" "l$ja"
## [11] "l$ko" "l$ml" "l$pt" "l$ru" "l$sv" "l$ta" "l$te" "l$tr" "l$ur" "l$zh"
```

With all of the terms above combined, we managed to get a total of 489 terms. We used these terms as variables in our model, and include the budget_adjusted as a variable as well. Due to large values of budgets and revenues, we decided to use log transformation on both of these variables and do modelling with them.

```
X <- cbind(dtm[, key_pc], dtm2[, key_gen], dtm3[, key_cast],
    dtm4[, key_keys], dtm5[, key_lang])
y_adj <- log(clean_data$revenue_adjusted)
```

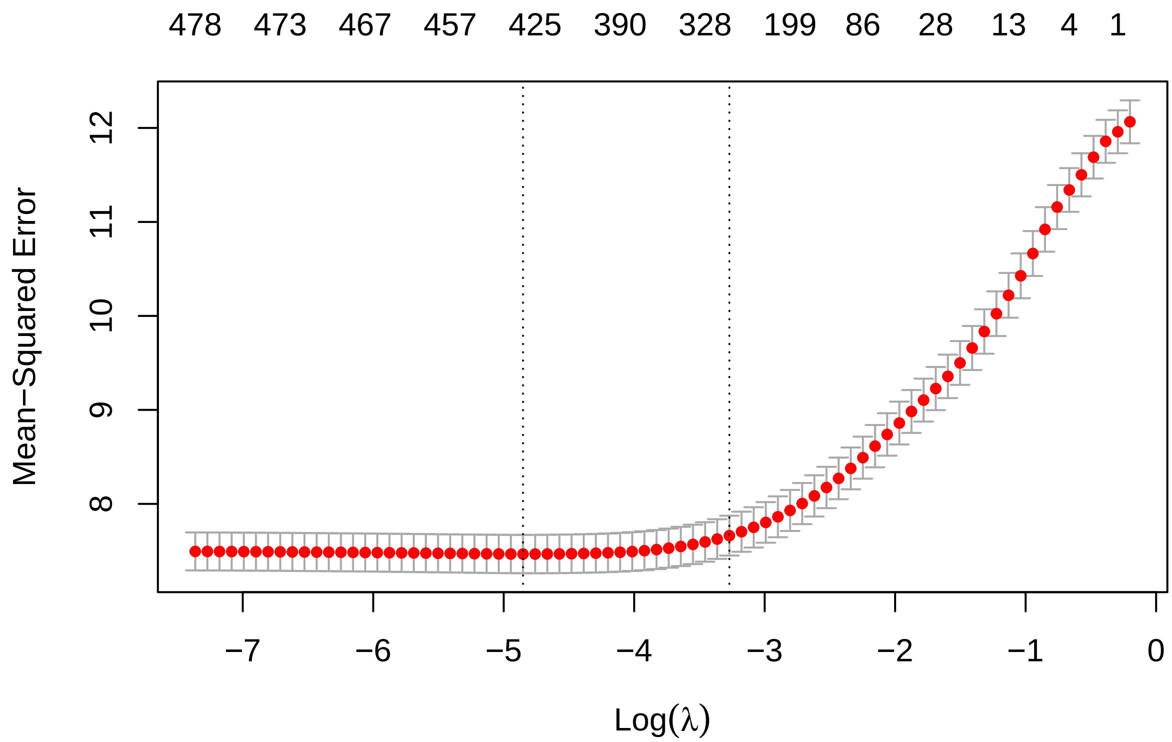
At first, we would like to include the budget column into the model. However, it turns out that 5,178 rows have either zero budget or unspecified amount. So, we decided to not include this column to the model.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-6
```

```
set.seed(424)
model1 <- cv.glmnet(as.matrix(X), y_adj)
plot(model1, xvar = "lambda")
```




```r
coef1 <- coef(model1, s = "lambda.min")
length(coef1[which(coef1 != 0), ][-1])  # -1 to exclude intercept
```

```
## [1] 429
```

```r
sort(coef1[which(coef1 != 0), 1], decreasing = TRUE)
```

```
##                (Intercept)              p$screen_gems
##               12.612435319                2.691182943
##                 p$paramount        p$columbia_pictures
##                2.568606204                2.443533753
##      p$walt_disney_pictures       p$20th_century_fox
##                2.411202846                2.385320720
##       p$touchstone_pictures      p$universal_pictures
##                2.285197787                2.222275215
##           p$new_line_cinema   p$warner_bros._pictures
##                2.183598777                2.139672383
##           p$united_artists       p$tristar_pictures
##                2.124228780                2.053181355
##        p$hollywood_pictures     p$metro_goldwyn_mayer
##                2.044048905                1.900867954
##                       l$ta         p$orion_pictures
##                1.767718471                1.609278386
##         p$fox_2000_pictures                     l$hi
```

```
##                    1.512771358                    1.509731902
##                    p$lionsgate            p$cj_entertainment
##                    1.477667013                    1.467142319
##         p$summit_entertainment                k$haunted_house
##                    1.461621725                    1.423593144
##        p$the_weinstein_company    p$fox_searchlight_pictures
##                    1.414129347                    1.357834505
##               p$focus_features                k$giant_monster
##                    1.339756615                    1.311197780
##                      p$miramax                k$single_mother
##                    1.238019098                    1.227285480
##                      k$slasher        k$duringcreditsstinger
##                    1.222694671                    1.171117731
##                           l$ja          c$sylvester_stallone
##                    1.149753390                    1.148462561
##               c$harrison_ford              p$constantin_film
##                    1.148393558                    1.146232666
##                           l$ko         p$dreamworks_pictures
##                    1.129634686                    1.109828006
##                         k$code                          l$zh
##                    1.105787694                    1.078517569
##                           l$it                  k$spacecraft
##                    1.072918529                    1.071859756
##              k$dying_and_death       p$blumhouse_productions
##                    1.067982777                    1.049773647
##         p$tf1_films_production                     k$florida
##                    1.037376939                    1.033666301
##                  k$silent_film  p$castle_rock_entertainment
##                    1.032217861                    1.029050154
##                 c$bess_flowers    k$based_on_play_or_musical
##                    1.017065439                    0.989306179
##                         k$pets           k$anthropomorphism
##                    0.981832191                    0.979077775
##            k$love_of_one’s_life          c$samuel_l._jackson
##                    0.951116709                    0.941334724
##             p$relativity_media               p$canal+_españa
##                    0.936022730                    0.930793976
##                           l$cn                 p$participant
##                    0.921700596                    0.916153272
##     k$based_on_young_adult_novel                     k$space
##                    0.915551064                    0.914312165
##                         k$epic                     k$villain
##                    0.909096291                    0.891713258
##                 p$cannon_group                   c$tom_hanks
##                    0.890193712                    0.886693133
##                  p$europacorp              p$dimension_films
##                    0.881080099                    0.872209925
##    p$village_roadshow_pictures                   c$brad_pitt
##                    0.864188666                    0.859829314
##                  c$ben_stiller k$father_daughter_relationship
##                    0.857228878                    0.851363777
##                   p$film_i_väst               k$based_on_comic
##                    0.837571688                    0.828496289
##                    k$cult_film                 c$liam_neeson
```

```
##                  0.823800102                 0.823743063
##                     k$sequel     p$new_regency_pictures
##                  0.822446566                 0.814488034
##                   k$creature      k$aftercreditsstinger
##                  0.810802471                 0.805296881
##          p$tsg_entertainment              c$gene_hackman
##                  0.805281425                 0.798126670
##         c$robert_downey_jr.                  k$airplane
##                  0.794558511                 0.759823990
##                 p$atresmedia                 k$buddy_cop
##                  0.752174260                 0.747914568
##                  k$biography                   k$divorce
##                  0.743077567                 0.742053662
##               c$ben_kingsley          k$romantic_comedy
##                  0.738715773                 0.728139844
##                      k$demon            k$usa_president
##                  0.723220144                 0.723048407
##                      k$widow            c$m._emmet_walsh
##                  0.707083437                 0.700591322
##          k$loss_of_loved_one            c$morgan_freeman
##                  0.694099810                 0.690978711
##                    k$kung_fu                    k$remake
##                  0.690430362                 0.689538921
##                c$matt_damon                  k$concert
##                  0.686122540                 0.675649458
##            k$suicide_attempt               k$super_power
##                  0.673597824                 0.672771245
##              c$nicole_kidman               k$california
##                  0.661613779                 0.654948310
##                  k$film_noir                    p$dentsu
##                  0.654199611                 0.649467457
##               c$meryl_streep                    k$sports
##                  0.648053300                 0.640966285
##                     p$pathé     k$psychological_thriller
##                  0.639636399                 0.639279800
##  p$filmnation_entertainment                      k$army
##                  0.639267314                 0.632039765
##                     k$school          c$richard_jenkins
##                  0.626544585                 0.624750611
##               k$secret_agent        c$donald_sutherland
##                  0.616739250                 0.616332900
##             c$robert_de_niro                 g$adventure
##                  0.613826260                 0.613652777
##                     k$orphan                     k$1980s
##                  0.606425428                 0.603262776
##            k$falling_in_love                        l$fr
##                  0.596041718                 0.594918561
##             k$new_york_city               c$joe_chrest
##                  0.589327598                 0.584186598
##        k$extramarital_affair    k$based_on_novel_or_book
##                  0.583446919                 0.580541253
##      p$imagine_entertainment  k$parent_child_relationship
##                  0.580480990                 0.580080372
##             c$clint_eastwood                     k$faith
```

```
##                   0.578775382                     0.577760225
##     k$san_francisco_california             p$france_2_cinéma
##                   0.574673970                     0.572257877
##                         k$cat                           g$war
##                   0.561578398                     0.560689396
##                     k$village                            l$en
##                   0.560103899                     0.558772790
##                     k$musical                  c$alec_baldwin
##                   0.553659059                     0.551340738
##                          l$te           p$dune_entertainment
##                   0.550571694                     0.549662919
##                    k$princess                      k$neighbor
##                   0.545020921                     0.534125917
##                      k$lawyer                 k$lgbt_interest
##                   0.529143632                     0.527023265
##                    k$politics   k$husband_wife_relationship
##                   0.514859553                     0.513289805
##                       k$magic                c$julianne_moore
##                   0.512329539                     0.512214757
##                     g$history                         k$spoof
##                   0.504838574                     0.503307045
##               c$robert_duvall   k$live_action_and_animation
##                   0.502597098                     0.501821393
##                k$supernatural           c$thomas_rosales_jr.
##                   0.500223282                     0.499773389
##              k$london_england                         k$1930s
##                   0.496354031                     0.493747146
##                      k$doctor                 c$stanley_tucci
##                   0.486703058                     0.480722674
##                      g$action                     g$animation
##                   0.480213975                     0.477838597
##   k$artificial_intelligence                      k$disaster
##                   0.476570718                     0.471325215
##            c$anthony_hopkins                        k$castle
##                   0.470647030                     0.469857803
##                    k$whodunit           c$harry_dean_stanton
##                   0.469573728                     0.469380431
##             p$france_3_cinéma              k$male_friendship
##                   0.467894806                     0.465546707
##                k$bank_robbery      p$scott_rudin_productions
##                   0.461693051                     0.461153215
##                      k$travel          p$malpaso_productions
##                   0.459449057                     0.455616245
##                  k$restaurant              c$forest_whitaker
##                   0.454805005                     0.453943923
##                    k$baseball               k$drug_addiction
##                   0.453054961                     0.452543902
##                  k$journalist                        k$author
##                   0.448052346                     0.446966362
##          k$based_on_true_story                c$keanu_reeves
##                   0.442602980                     0.442024564
##                        p$toho                      k$hospital
##                   0.439125826                     0.439081715
##                k$space_travel                        g$comedy
```

```
##                  0.438965684                 0.435660952
##                  k$fairy_tale                k$drug_dealer
##                  0.433408344                 0.431063127
##                       k$self                      k$heist
##                  0.428132959                 0.427673156
##            c$michael_papajohn            c$danny_glover
##                  0.427027751                 0.424232355
##             k$transformation                     k$1960s
##                  0.419793434                 0.419698676
##                    g$romance                   k$brother
##                  0.416884244                 0.415435401
##                    k$england                        k$in
##                  0.415432268                 0.414620787
##                     p$canal+                    k$racism
##                  0.413370048                 0.405786353
##             c$robin_williams              c$johnny_depp
##                  0.404256316                 0.401181785
##                    k$torture                 k$christmas
##                  0.400646837                 0.399926862
##                  k$terrorist              c$frank_welker
##                  0.397029099                 0.389777218
##            k$black_and_white                   k$teacher
##                  0.389279358                 0.386160604
##                        k$fbi                 k$scientist
##                  0.385634453                 0.385535339
##                    k$friends                 k$marijuana
##                  0.382391292                 0.375200131
##                    g$western                  k$teenager
##                  0.372853794                 0.370816486
##             k$movie_business              k$19th_century
##                  0.367527600                 0.367476836
##                   k$marriage   p$morgan_creek_productions
##                  0.367068551                 0.367049911
##                    k$wedding                     k$china
##                  0.366383368                 0.363280079
##                    k$holiday             c$paul_giamatti
##                  0.363253287                 0.362066838
##                     k$forest                   k$monster
##                  0.360105686                 0.355083387
##     p$scott_free_productions                    k$mexico
##                  0.353637278                 0.350716875
##                     k$animal                    k$rescue
##                  0.350016273                 0.349892854
##                     g$family              k$period_drama
##                  0.348288297                 0.347177859
##               k$paris_france                 k$pregnancy
##                  0.345079155                 0.339241781
##                    g$fantasy              k$martial_arts
##                  0.338076474                 0.338001199
##                      k$witch               k$competition
##                  0.337812746                 0.336307965
##                      k$mafia       p$amblin_entertainment
##                  0.331400758                 0.330995848
##       k$sibling_relationship                   k$sheriff
```

```
##                        0.330583750                 0.322438618
##                          g$mystery                    k$horror
##                        0.322238070                 0.321881060
##                           k$summer k$based_on_children's_book
##                        0.320991857                 0.319568145
##                  c$susan_sarandon                    k$family
##                        0.318627014                 0.316478061
##                     p$studiocanal              c$dennis_quaid
##                        0.316100862                 0.315875917
##                      k$time_travel  p$lakeshore_entertainment
##                        0.314597268                 0.313661719
##                          k$rivalry                   k$alcohol
##                        0.313540442                 0.311142331
##                            k$beach              c$bruce_willis
##                        0.309013896                 0.308395077
##          k$los_angeles_california                k$corruption
##                        0.308142247                 0.308134587
##                         k$adultery                   k$vampire
##                        0.307549664                 0.306121941
##                             k$baby                       k$gun
##                        0.305779225                 0.305449797
##                     c$bill_murray                p$wild_bunch
##                        0.301284457                 0.301067487
##               c$sigourney_weaver      k$based_on_video_game
##                        0.298999000                 0.298277983
##                           k$priest                     k$money
##                        0.297611246                 0.296141200
##                       k$psychopath              c$john_goodman
##                        0.294703680                 0.294166812
##                        k$obsession                k$friendship
##                        0.292914200                 0.287960874
##                               l$ar      p$regency_enterprises
##                        0.287300825                 0.284142255
##                   k$love_triangle                  k$showdown
##                        0.280839436                 0.280634714
##              k$female_friendship                k$helicopter
##                        0.279547156                 0.278944485
##                          k$funeral                 k$brutality
##                        0.278816765                 0.278725458
## p$polygram_filmed_entertainment                    k$japan
##                        0.273875483                 0.270839492
##                        k$anti_hero              k$teenage_girl
##                        0.270535178                 0.269096436
##                              k$cop                    k$island
##                        0.266059127                 0.265409138
##                           k$jungle                k$undercover
##                        0.265177066                 0.264990747
##                 k$organized_crime                   k$romance
##                        0.263631073                 0.261991202
##                  k$police_officer                    k$desert
##                        0.261073527                 0.260371604
##                          k$soldier                     k$1920s
##                        0.256549209                 0.250049228
##                  k$alien_invasion               c$willem_dafoe
```

```
##                  0.246961040               0.244126131
##                       k$bomb   p$working_title_films
##                  0.243591007               0.242630257
##                     k$1950s                  k$prison
##                  0.240121370               0.237806528
##                     k$thief           k$investigation
##                  0.235344643               0.233058982
##               k$action_hero                     k$spy
##                  0.231999295               0.229456904
##                 k$flashback                k$survival
##                  0.228602152               0.227799143
##              c$alfred_molina                p$gaumont
##                  0.222651872               0.221014976
##                     k$hotel      c$antonio_banderas
##                  0.217478947               0.216989570
##                 k$road_trip               k$alcoholic
##                  0.211676635               0.209232923
##                      k$ship                   k$anime
##                  0.201443279               0.201190506
##               k$best_friend                g$thriller
##                  0.200489478               0.200393047
##               c$john_cusack                   k$1940s
##                  0.198673300               0.198400518
##                   k$revenge             k$street_gang
##                  0.198009957               0.195663316
##                  k$jealousy               k$seduction
##                  0.194688004               0.194045077
##               k$child_abuse                k$gunfight
##                  0.190649595               0.189773697
##                    k$escape        g$science_fiction
##                  0.189215093               0.188353126
##          c$woody_harrelson                k$daughter
##                  0.187740302               0.186469768
##           c$john_leguizamo                  k$parody
##                  0.183608636               0.182319953
##                  k$dystopia               k$vigilante
##                  0.176868862               0.176747241
##                     k$death               k$car_crash
##                  0.176171208               0.174918239
##            c$michael_caine                  k$dragon
##                  0.173863259               0.173039679
##                    k$satire                    k$nazi
##                  0.171947445               0.168826552
##                        l$de                 k$college
##                  0.167765965               0.158877431
##                   k$amnesia       p$film4_productions
##                  0.155205162               0.152284394
##               k$high_school     k$male_homosexuality
##                  0.152282550               0.151425822
##                     g$crime              c$ethan_hawke
##                  0.150079519               0.146952885
##                        l$ru              k$small_town
##                  0.146874022               0.144087124
##                    k$police                   k$texas
```

```
##                   0.142535546                0.139874712
##               p$sony_pictures             c$stephen_root
##                   0.138766798                0.135392554
##                       k$train                   k$father
##                   0.134358956                0.129813455
##                       k$dream                    k$robot
##                   0.127545804                0.124318723
##                 c$kevin_bacon        k$chicago_illinois
##                   0.119968754                0.119875893
##                    k$musician                   k$sadism
##                   0.119107945                0.117611725
##             p$silver_pictures          c$cate_blanchett
##                   0.115535446                0.114361095
##                    k$gangster            c$j.k._simmons
##                   0.107959088                0.106524815
##                       k$1990s k$mother_son_relationship
##                   0.103342366                0.100960017
##                      k$zombie                    k$fight
##                   0.097795960                0.092483853
##                       g$music                   k$hitman
##                   0.088342979                0.079902514
##                  k$experiment                   k$killer
##                   0.077983741                0.076847967
##                k$world_war_ii                     k$noir
##                   0.076570302                0.073943418
##                    k$shootout k$manhattan_new_york_city
##                   0.063839576                0.055945807
##                        k$love     c$christopher_plummer
##                   0.054833314                0.048787932
##    k$father_son_relationship                k$religion
##                   0.048633572                0.046873197
##                          k$ex                k$gambling
##                   0.046560887                0.040331584
##                          l$sv               k$nightmare
##                   0.035055297                0.034031908
##                    k$new_love                k$assassin
##                   0.031665605                0.030872350
##                  k$kidnapping                p$bbc_film
##                   0.029945822                0.026975868
##                        k$hero               k$explosion
##                   0.024050573                0.018593473
##                   c$brian_cox                   k$1970s
##                   0.018313188                0.016290446
##                  k$conspiracy                  k$church
##                   0.013827079                0.011352618
##       k$family_relationships               c$john_hurt
##                   0.008652759                0.007953874
##                  k$prostitute                      k$up
##                   0.006196392               -0.001942926
##                c$nicolas_cage            k$serial_killer
##                  -0.010852775               -0.023050012
##                         k$gay             k$coming_of_age
##                  -0.036488339               -0.039189843
##                       k$dance           p$ingenious_media
```

```
##                   -0.043309880                  -0.046699749
##                    k$hollywood                       g$drama
##                   -0.069784634                  -0.088935468
##               k$woman_director                  k$surrealism
##                   -0.093459917                  -0.129488168
##                   k$teen_movie                k$gay_interest
##                   -0.131351962                  -0.132369104
##                    k$detective            p$millennium_films
##                   -0.141238641                  -0.142542877
##                 c$james_franco                  c$danny_trejo
##                   -0.143946212                  -0.146672570
##                  c$owen_wilson                     k$robbery
##                   -0.148928681                  -0.159102189
##                   k$rural_area               k$found_footage
##                   -0.189195873                  -0.229293838
##                   k$depression                       p$ciné+
##                   -0.241362479                  -0.241719553
##           c$christopher_walken                        k$gore
##                   -0.244824995                  -0.256591641
##                          p$cnc                      k$writer
##                   -0.263609364                  -0.269504383
##                       g$horror             k$female_wrestler
##                   -0.280295919                  -0.315658525
##                k$assassination             p$téléfilm_canada
##                   -0.355630692                  -0.356301807
##                  k$dark_comedy               k$pro_wrestling
##                   -0.395777237                  -0.524640742
##                           l$ur                       k$grief
##                   -0.582455642                  -0.599084443
##                         k$lgbt                          l$pt
##                   -0.780566881                  -1.028575719
##                           l$fa                 g$documentary
##                   -1.153459063                  -1.552939934
##                    k$wrestling                    g$tv_movie
##                   -2.692157715                  -2.971512671
```

From 489 variables, we have 429 variables in our model. Below are the intercept of the model, the top 10 variables that positively affect revenue, and top 10 variables that negatively affect revenue.

```
ic1 <- coef1[c("(Intercept)"), 1]
paste("The intercept is ", ic1)
```

```
## [1] "The intercept is  12.6124353194507"
```

```
paste("Top 10 variables that positively affect the revenue:")
```

```
## [1] "Top 10 variables that positively affect the revenue:"
```

```
coef1_sort <- sort(coef1[, 1], decreasing = TRUE)[-1]
head(coef1_sort, 10)
```

```
##          p$screen_gems            p$paramount     p$columbia_pictures
```

```
##                2.691183              2.568606               2.443534
##  p$walt_disney_pictures     p$20th_century_fox   p$touchstone_pictures
##                2.411203              2.385321               2.285198
##    p$universal_pictures      p$new_line_cinema p$warner_bros._pictures
##                2.222275              2.183599               2.139672
##        p$united_artists
##                2.124229
```

```r
paste("Top 10 variables that negatively affects the revenue:")
```

```
## [1] "Top 10 variables that negatively affects the revenue:"
```

```r
tail(coef1_sort, 10)
```

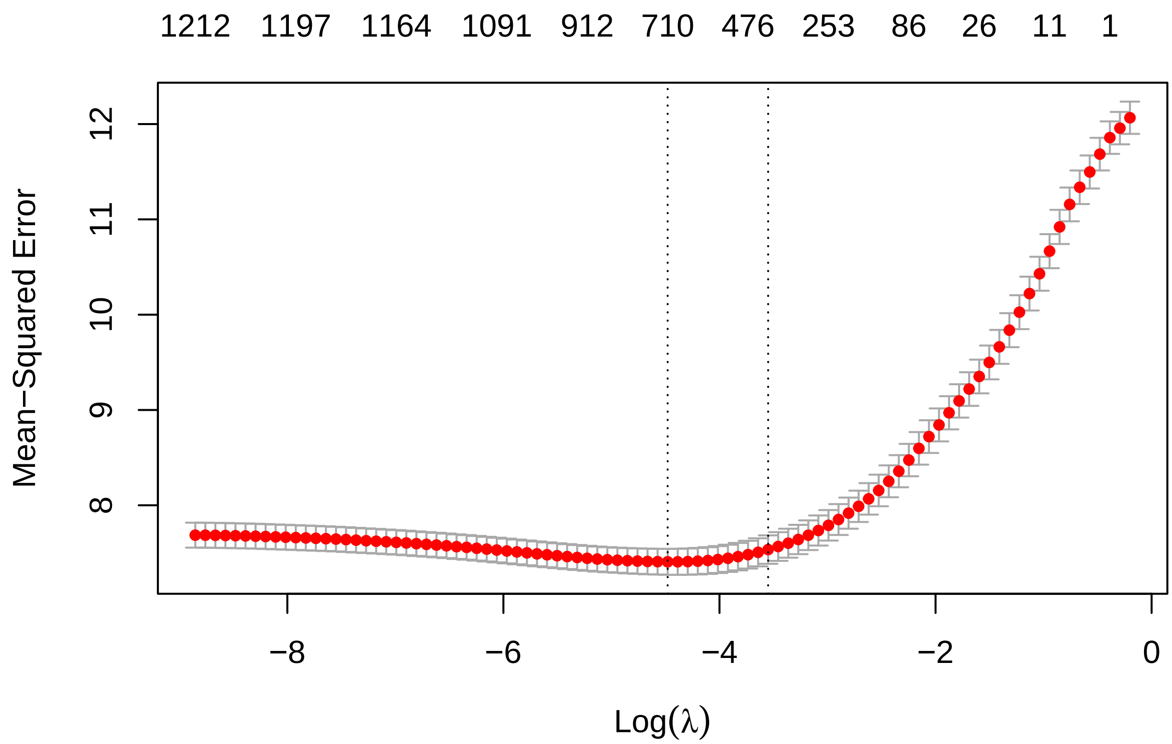
```
##   k$dark_comedy k$pro_wrestling          l$ur        k$grief         k$lgbt
##      -0.3957772      -0.5246407    -0.5824556     -0.5990844     -0.7805669
##            l$pt            l$fa g$documentary     k$wrestling     g$tv_movie
##      -1.0285757      -1.1534591    -1.5529399     -2.6921577     -2.9715127
```

To improve the model more, we want to include interactions between terms in the model. Since it is too much to interact all of the terms that we have, we will get top 10 terms from each column and interact them with each other to make a pair of terms. We do not include the original_language column in the interaction since most of the movies' original language is English. So, we have 40 terms to be paired with each other, giving us 780 interaction variables.

```r
# top 10 of each
key_pc2 <- names(findMostFreqTerms(dtm, 10, INDEX = rep(1, each = n))[[1]])
key_gen2 <- names(findMostFreqTerms(dtm2, 10, INDEX = rep(1,
    each = n))[[1]])
key_cast2 <- names(findMostFreqTerms(dtm3, 14, INDEX = rep(1,
    each = n))[[1]])
key_cast2 <- key_cast2[grepl("_", key_cast2)]  # 4 of them are single names
key_keys2 <- names(findMostFreqTerms(dtm4, 10, INDEX = rep(1,
    each = n))[[1]])
int_vars <- c(key_pc2, key_gen2, key_cast2, key_keys2)
inact <- c()
inact_name <- c()
for (i in 1:(length(int_vars) - 1)) {
    for (j in (i + 1):length(int_vars)) {
        a = as.matrix(X[, int_vars[i]])
        b = as.matrix(X[, int_vars[j]])
        var_name = paste(int_vars[i], ".", int_vars[j])
        v = a * b
        inact <- cbind(inact, v)
        inact_name <- c(inact_name, var_name)
    }
}

df_inact = data.frame(inact)
colnames(df_inact) <- inact_name
```

We add this 780 variables on top of the initial 489 variables in the first model, giving us 1269 variables for the second model.

```r
X2 <- cbind(dtm[, key_pc], dtm2[, key_gen], dtm3[, key_cast],
    dtm4[, key_keys], dtm5[, key_lang], df_inact)
model2 <- cv.glmnet(as.matrix(X2), y_adj)
plot(model2, xvar = "lambda")
```




```r
coef2 <- coef(model2, s = "lambda.min")
length(coef2[which(coef2 != 0), ][-1])  # -1 to exclude intercept
```

```
## [1] 710
```

```r
sort(coef2[which(coef2 != 0), 1], decreasing = TRUE)
```

```
##                                       (Intercept)
##                                      1.249912e+01
##                  c$robert_de_niro . k$woman_director
##                                      3.121845e+00
##                              p$columbia_pictures
##                                      2.643334e+00
##                                     p$screen_gems
##                                      2.615547e+00
##                                       p$paramount
##                                      2.595986e+00
##                            p$touchstone_pictures
##                                      2.561390e+00
```

```
##                                 p$20th_century_fox
##                                       2.514055e+00
##                             p$walt_disney_pictures
##                                       2.420880e+00
##                            p$warner_bros._pictures
##                                       2.343409e+00
##                               p$universal_pictures
##                                       2.313728e+00
##          g$science_fiction . c$robert_de_niro
##                                       2.312593e+00
##                                  p$new_line_cinema
##                                       2.277045e+00
##                                  p$united_artists
##                                       2.156397e+00
##                              p$metro_goldwyn_mayer
##                                       2.145121e+00
##                           c$j.k._simmons . k$love
##                                       2.095606e+00
##                              p$hollywood_pictures
##                                       2.076517e+00
##                                p$tristar_pictures
##                                       2.003320e+00
##           c$samuel_l._jackson . c$liam_neeson
##                                       1.697209e+00
##                  c$nicolas_cage . k$biography
##                                       1.676166e+00
##                                              l$ta
##                                       1.653243e+00
##                                  p$orion_pictures
##                                       1.576818e+00
##                               p$fox_2000_pictures
##                                       1.560967e+00
##                            p$the_weinstein_company
##                                       1.487038e+00
##                             p$summit_entertainment
##                                       1.473376e+00
##                                       p$lionsgate
##                                       1.473322e+00
##                             k$duringcreditsstinger
##                                       1.412087e+00
##                                              l$hi
##                                       1.394998e+00
##                                  p$focus_features
##                                       1.361757e+00
##                      p$canal+ . c$frank_welker
##                                       1.357823e+00
##                        p$fox_searchlight_pictures
##                                       1.343146e+00
##                                p$cj_entertainment
##                                       1.339942e+00
##                                   k$haunted_house
##                                       1.293407e+00
##                      g$romance . c$willem_dafoe
##                                       1.241388e+00
```

```
##                                        p$miramax
##                                     1.212257e+00
##                                  k$single_mother
##                                     1.201212e+00
##            g$science_fiction . c$samuel_l._jackson
##                                     1.159302e+00
##                             p$dreamworks_pictures
##                                     1.143496e+00
##                                  k$giant_monster
##                                     1.060618e+00
##                                             l$ko
##                                     1.055847e+00
##                                      k$buddy_cop
##                                     1.052702e+00
##                                p$constantin_film
##                                     1.052543e+00
##                                      c$brad_pitt
##                                     1.047456e+00
##                              c$sylvester_stallone
##                                     1.045275e+00
##                                k$dying_and_death
##                                     1.036354e+00
##                                    k$silent_film
##                                     1.019610e+00
##                                             l$ja
##                                     1.012502e+00
##                                        k$slasher
##                                     1.008986e+00
##          c$j.k._simmons . k$based_on_novel_or_book
##                                     1.004464e+00
##                                           k$code
##                                     1.002505e+00
##                     g$family . c$steve_buscemi
##                                     9.961378e-01
##                      g$family . c$bruce_willis
##                                     9.874252e-01
##              c$robert_de_niro . c$morgan_freeman
##                                     9.709606e-01
##                                             l$zh
##                                     9.375226e-01
##                            p$tf1_films_production
##                                     9.324411e-01
##                                        k$florida
##                                     9.282380e-01
##                                 c$harrison_ford
##                                     9.281722e-01
##                       k$based_on_play_or_musical
##                                     9.273600e-01
##                                      g$adventure
##                                     9.185655e-01
##                       p$castle_rock_entertainment
##                                     9.172846e-01
##                                           k$pets
##                                     9.150230e-01
```

```
##                       g$science_fiction . k$love
##                                     9.130249e-01
##                                           k$epic
##                                     9.074391e-01
##                                     c$ben_stiller
##                                     9.069395e-01
##                                             l$it
##                                     8.949374e-01
##                                       k$creature
##                                     8.861610e-01
##                                    c$bess_flowers
##                                     8.828913e-01
##                                p$relativity_media
##                                     8.794182e-01
##                  k$father_daughter_relationship
##                                     8.755584e-01
##                               k$anthropomorphism
##                                     8.743774e-01
##          p$columbia_pictures . c$robert_de_niro
##                                     8.739832e-01
##                             k$love_of_one’s_life
##                                     8.715907e-01
##                     g$romance . c$nicolas_cage
##                                     8.709968e-01
##                                  p$canal+_españa
##                                     8.595089e-01
##                     g$horror . c$morgan_freeman
##                                     8.569961e-01
##                     p$canal+ . c$morgan_freeman
##                                     8.523741e-01
##                                    p$participant
##                                     8.504760e-01
##                                             l$cn
##                                     8.448109e-01
##                                         g$comedy
##                                     8.320128e-01
##                                     k$spacecraft
##                                     8.241830e-01
##                          p$blumhouse_productions
##                                     8.192420e-01
##                    p$village_roadshow_pictures
##                                     8.137941e-01
##                                      c$tom_hanks
##                                     8.123906e-01
##                                p$dimension_films
##                                     8.032815e-01
##                                          k$space
##                                     7.992673e-01
##                                        k$villain
##                                     7.865623e-01
##                                   p$cannon_group
##                                     7.853691e-01
##                      k$based_on_young_adult_novel
##                                     7.782826e-01
```

```
##                                      p$film_i_väst
##                                       7.756398e-01
##                                c$samuel_l._jackson
##                                       7.748496e-01
##                           k$based_on_novel_or_book
##                                       7.642640e-01
##                                       c$matt_damon
##                                       7.413751e-01
##                   c$j.k._simmons . k$new_york_city
##                                       7.410187e-01
##                                        k$cult_film
##                                       7.318618e-01
##                                           k$sequel
##                                       7.235152e-01
##                             p$new_regency_pictures
##                                       7.205081e-01
##                                       p$europacorp
##                                       7.193952e-01
##                                    k$usa_president
##                                       7.000752e-01
##                           g$drama . c$frank_welker
##                                       6.974909e-01
##                                      c$liam_neeson
##                                       6.945593e-01
##                                            k$demon
##                                       6.892090e-01
##                                   k$based_on_comic
##                                       6.870760e-01
##                               p$tsg_entertainment
##                                       6.836161e-01
##                                           g$action
##                                       6.793915e-01
##                                          k$concert
##                                       6.792923e-01
##                              k$aftercreditsstinger
##                                       6.791355e-01
##                                         k$airplane
##                                       6.761271e-01
##                                     k$secret_agent
##                                       6.714371e-01
##                                           k$remake
##                                       6.629631e-01
##                                          k$divorce
##                                       6.592426e-01
##                                     c$ben_kingsley
##                                       6.532820e-01
##                c$willem_dafoe . k$woman_director
##                                       6.513428e-01
##                                          k$kung_fu
##                                       6.483162e-01
##                                            k$widow
##                                       6.463279e-01
##                           k$psychological_thriller
##                                       6.411760e-01
```

```
##                       g$horror . c$liam_neeson
##                                   6.405494e-01
##                                  k$super_power
##                                   6.346303e-01
##            p$20th_century_fox . c$bruce_willis
##                                   6.340520e-01
##                               c$m._emmet_walsh
##                                   6.325546e-01
##                                   k$california
##                                   6.322016e-01
##                     p$filmnation_entertainment
##                                   6.225361e-01
##                                       k$sports
##                                   6.211511e-01
##                                       p$dentsu
##                                   6.209699e-01
##                                       k$orphan
##                                   6.198137e-01
##                              k$romantic_comedy
##                                   6.161792e-01
##                                 c$gene_hackman
##                                   6.133935e-01
##                             k$loss_of_loved_one
##                                   6.132478e-01
##                                   c$joe_chrest
##                                   6.122899e-01
##                                       k$school
##                                   6.050066e-01
##                     k$san_francisco_california
##                                   5.982607e-01
##                               c$clint_eastwood
##                                   5.973423e-01
##                   g$family . c$morgan_freeman
##                                   5.939863e-01
##                              k$suicide_attempt
##                                   5.929272e-01
##                                        k$faith
##                                   5.909168e-01
##                                 c$meryl_streep
##                                   5.877919e-01
##                                c$nicole_kidman
##                                   5.857945e-01
##                                        p$pathé
##                                   5.841771e-01
##                                   p$atresmedia
##                                   5.836289e-01
##                      c$liam_neeson . k$murder
##                                   5.803617e-01
##                   k$parent_child_relationship
##                                   5.765675e-01
##                            c$robert_downey_jr.
##                                   5.692035e-01
##                                        k$spoof
##                                   5.689656e-01
```

```
##                        k$based_on_true_story
##                                 5.669564e-01
##                 c$bruce_willis . c$willem_dafoe
##                                 5.618420e-01
##              p$walt_disney_pictures . k$murder
##                                 5.490955e-01
##                                  k$film_noir
##                                 5.451954e-01
##                                        g$war
##                                 5.436421e-01
##                                      k$1980s
##                                 5.415277e-01
##                        k$extramarital_affair
##                                 5.394664e-01
##                     c$nicolas_cage . k$love
##                                 5.385782e-01
##                          c$donald_sutherland
##                                 5.357480e-01
##                    g$horror . c$frank_welker
##                                 5.321484e-01
##                            c$richard_jenkins
##                                 5.320359e-01
##                                    k$musical
##                                 5.306410e-01
##                k$duringcreditsstinger . k$love
##                                 5.299613e-01
##                                        k$cat
##                                 5.264879e-01
##                                    g$history
##                                 5.231505e-01
##                                      k$magic
##                                 5.221871e-01
##                            k$falling_in_love
##                                 5.211684e-01
##                                     k$lawyer
##                                 5.170851e-01
##                             c$morgan_freeman
##                                 5.167982e-01
##                  k$husband_wife_relationship
##                                 5.158077e-01
##                                   k$neighbor
##                                 5.150038e-01
##                     p$imagine_entertainment
##                                 5.048504e-01
##                c$willem_dafoe . c$j.k._simmons
##                                 5.007863e-01
##                                   k$whodunit
##                                 4.973902e-01
##                                     g$family
##                                 4.940985e-01
##                                    k$village
##                                 4.927393e-01
##                                      k$1930s
##                                 4.913076e-01
```

```
##                                        k$new_york_city
##                                           4.897370e-01
##                                     p$france_2_cinéma
##                                           4.876889e-01
##                                      c$julianne_moore
##                                           4.855361e-01
##                                               p$canal+
##                                           4.831596e-01
##                                                   l$fr
##                                           4.807756e-01
##                                 c$thomas_rosales_jr.
##                                           4.783023e-01
##                                             k$princess
##                                           4.743269e-01
##                                                 k$army
##                                           4.692439e-01
##                                              k$vampire
##                                           4.663048e-01
##                                              k$torture
##                                           4.662805e-01
##                                         c$alec_baldwin
##                                           4.629514e-01
##                                            k$biography
##                                           4.628030e-01
##                            g$horror . c$j.k._simmons
##                                           4.609908e-01
##                  k$based_on_novel_or_book . k$love
##                                           4.608640e-01
##                                             k$politics
##                                           4.603056e-01
##                                p$dune_entertainment
##                                           4.593629e-01
##                                      k$london_england
##                                           4.579641e-01
##                                                   l$en
##                                           4.577728e-01
##                                      k$transformation
##                                           4.575241e-01
##                                      k$drug_addiction
##                                           4.567764e-01
##                                                   l$te
##                                           4.539333e-01
##                               g$drama . k$biography
##                                           4.511028e-01
##                           p$scott_rudin_productions
##                                           4.473477e-01
##                           g$action . c$willem_dafoe
##                                           4.470800e-01
##                                        k$bank_robbery
##                                           4.469806e-01
##                                 c$harry_dean_stanton
##                                           4.458753e-01
##                                       k$lgbt_interest
##                                           4.449351e-01
```

```
##                          k$black_and_white
##                               4.439426e-01
##                g$horror . c$willem_dafoe
##                               4.435949e-01
##                                     k$self
##                               4.371899e-01
##                                     p$toho
##                               4.326404e-01
##                           c$paul_giamatti
##                               4.301743e-01
##                                    k$1960s
##                               4.300113e-01
##                                g$animation
##                               4.274920e-01
##                           c$robin_williams
##                               4.263073e-01
##                              k$drug_dealer
##                               4.251101e-01
##                                  g$romance
##                               4.229601e-01
##                                  g$western
##                               4.224298e-01
##                              c$johnny_depp
##                               4.217258e-01
##                                 k$hospital
##                               4.208732e-01
##                             k$supernatural
##                               4.200345e-01
##                          k$male_friendship
##                               4.193118e-01
##                           c$stanley_tucci
##                               4.175308e-01
##                               k$journalist
##                               4.154961e-01
##                k$artificial_intelligence
##                               4.149688e-01
##                                  k$teacher
##                               4.124518e-01
##              k$live_action_and_animation
##                               4.111310e-01
##                             c$keanu_reeves
##                               4.078841e-01
##                                   k$author
##                               4.071627e-01
##                           c$robert_duvall
##                               4.057867e-01
##                          c$forest_whitaker
##                               4.021262e-01
##                          c$anthony_hopkins
##                               3.992732e-01
##                          p$france_3_cinéma
##                               3.974556e-01
##                     k$sibling_relationship
##                               3.954100e-01
```

```
##                                              k$racism
##                                          3.947433e-01
##                                                 k$fbi
##                                          3.926466e-01
##                                          k$restaurant
##                                          3.885058e-01
##                                              k$doctor
##                                          3.869933e-01
##                    p$canal+ . k$based_on_true_story
##                                          3.850616e-01
##                                          k$fairy_tale
##                                          3.838866e-01
##                                              k$travel
##                                          3.829529e-01
##                            g$crime . c$j.k._simmons
##                                          3.813524e-01
##                                        k$martial_arts
##                                          3.808840e-01
##                                        k$space_travel
##                                          3.808256e-01
##                                             g$fantasy
##                                          3.804865e-01
##              p$touchstone_pictures . c$bruce_willis
##                                          3.802644e-01
##                                      k$movie_business
##                                          3.801359e-01
##                            p$lakeshore_entertainment
##                                          3.776525e-01
##                                        c$danny_glover
##                                          3.722250e-01
##                                                  k$in
##                                          3.680238e-01
##                                        c$john_goodman
##                                          3.672527e-01
##                            g$comedy . c$bruce_willis
##                                          3.670100e-01
##                                        k$paris_france
##                                          3.644581e-01
##                                               k$witch
##                                          3.641642e-01
##                                              k$castle
##                                          3.641384e-01
##      p$walt_disney_pictures . c$samuel_l._jackson
##                                          3.619544e-01
##                                            k$disaster
##                                          3.605719e-01
##                                p$malpaso_productions
##                                          3.604743e-01
##                                                 k$gun
##                                          3.541545e-01
##                                             k$brother
##                                          3.540774e-01
##                                               k$china
##                                          3.513248e-01
```

```
##                                  k$adultery
##                                3.498358e-01
##                              k$19th_century
##                                3.463795e-01
##                              k$period_drama
##                                3.443949e-01
##                                   g$mystery
##                                3.443562e-01
##                                  k$baseball
##                                3.428503e-01
##                                 k$marijuana
##                                3.418812e-01
##                    p$morgan_creek_productions
##                                3.415326e-01
##                                p$wild_bunch
##                                3.401405e-01
##                                    k$zombie
##                                3.361284e-01
##                                  k$showdown
##                                3.339270e-01
##                                   k$england
##                                3.338296e-01
##                                  k$marriage
##                                3.315388e-01
##                                   k$monster
##                                3.306447e-01
##                                     k$money
##                                3.298129e-01
##                                    k$rescue
##                                3.268109e-01
##                   c$j.k._simmons . k$revenge
##                                3.267302e-01
##                                 k$anti_hero
##                                3.225618e-01
##                                k$helicopter
##                                3.194652e-01
##                                k$psychopath
##                                3.171130e-01
##                                    k$horror
##                                3.120754e-01
##                                     k$mafia
##                                3.112135e-01
##                   k$based_on_children's_book
##                                3.100598e-01
##                                    k$family
##                                3.095019e-01
##                g$thriller . c$robert_de_niro
##                                3.084982e-01
##                                     k$beach
##                                3.051926e-01
##                                       k$spy
##                                3.041940e-01
##         k$based_on_novel_or_book . k$revenge
##                                3.026899e-01
```

```
##                                     c$bill_murray
##                                      2.994364e-01
##                        g$drama . c$morgan_freeman
##                                      2.993421e-01
##                         g$family . c$willem_dafoe
##                                      2.992950e-01
##                       c$steve_buscemi . k$murder
##                                      2.986310e-01
##                                  c$susan_sarandon
##                                      2.983545e-01
##                               g$science_fiction
##                                      2.977668e-01
##                                c$michael_papajohn
##                                      2.963695e-01
##                          p$scott_free_productions
##                                      2.961675e-01
##                                         k$rivalry
##                                      2.958666e-01
##                                          k$summer
##                                      2.957218e-01
##                                         k$sheriff
##                                      2.935239e-01
##                                       k$scientist
##                                      2.934839e-01
##             g$science_fiction . k$new_york_city
##                                      2.932682e-01
##                                     k$time_travel
##                                      2.925804e-01
##                     p$new_line_cinema . g$horror
##                                      2.879041e-01
##                                         k$funeral
##                                      2.863133e-01
##                                           k$heist
##                                      2.858946e-01
##                            g$thriller . g$horror
##                                      2.848327e-01
##                                       k$christmas
##                                      2.834205e-01
##                           p$amblin_entertainment
##                                      2.806535e-01
##                                    p$studiocanal
##                                      2.787488e-01
##                                       k$pregnancy
##                                      2.750117e-01
##                 p$polygram_filmed_entertainment
##                                      2.732096e-01
##                                      k$corruption
##                                      2.708941e-01
##                                         k$friends
##                                      2.708625e-01
##                                        k$teenager
##                                      2.706081e-01
##                                           g$crime
##                                      2.696396e-01
```

```
##                                    k$japan
##                               2.691719e-01
##                               k$friendship
##                               2.681208e-01
##                             c$dennis_quaid
##                               2.666808e-01
##                                   k$mexico
##                               2.629952e-01
##           p$20th_century_fox . g$drama
##                               2.612212e-01
##                         c$sigourney_weaver
##                               2.598305e-01
##           g$action . g$science_fiction
##                               2.591819e-01
##                     k$based_on_video_game
##                               2.586369e-01
##                               k$terrorist
##                               2.585999e-01
##                                  k$wedding
##                               2.583441e-01
##                             c$stephen_root
##                               2.556472e-01
##                   g$family . k$biography
##                               2.495337e-01
##                          k$police_officer
##                               2.472347e-01
##                                   k$desert
##                               2.453373e-01
##                                   k$parody
##                               2.452577e-01
##                                     k$bomb
##                               2.449883e-01
##                                   k$forest
##                               2.442312e-01
##                              k$competition
##                               2.435827e-01
##                                    k$1950s
##                               2.433793e-01
##                          c$cate_blanchett
##                               2.424496e-01
##                                     k$ship
##                               2.412486e-01
##                    p$working_title_films
##                               2.405223e-01
##                           c$alfred_molina
##                               2.399806e-01
##                                  k$holiday
##                               2.384905e-01
##                                     k$baby
##                               2.381922e-01
##                  k$los_angeles_california
##                               2.374074e-01
##                                   k$prison
##                               2.362325e-01
```

```
##                                 g$thriller . k$sequel
##                                          2.354943e-01
##                               p$regency_enterprises
##                                          2.322330e-01
##              p$touchstone_pictures . g$thriller
##                                          2.318840e-01
##                                           k$romance
##                                          2.292965e-01
##                                         k$road_trip
##                                          2.291921e-01
##                p$touchstone_pictures . g$comedy
##                                          2.287123e-01
##                     g$romance . c$frank_welker
##                                          2.279426e-01
##                p$walt_disney_pictures . g$crime
##                                          2.267159e-01
##                                            k$priest
##                                          2.250638e-01
##                            p$paramount . g$drama
##                                          2.239650e-01
##            p$new_line_cinema . c$steve_buscemi
##                                          2.226642e-01
##                                            k$jungle
##                                          2.196225e-01
##                                       k$action_hero
##                                          2.186515e-01
##                                        k$undercover
##                                          2.178181e-01
##                                     k$investigation
##                                          2.162052e-01
##                                 k$female_friendship
##                                          2.157926e-01
##                               g$comedy . k$revenge
##                                          2.146488e-01
##                                     k$love_triangle
##                                          2.137629e-01
##                     g$science_fiction . k$murder
##                                          2.132116e-01
##                                             k$hotel
##                                          2.108632e-01
##                                             k$anime
##                                          2.108277e-01
##                                      k$teenage_girl
##                                          2.099500e-01
##                                   k$organized_crime
##                                          2.086745e-01
##                                               k$cop
##                                          2.066143e-01
##                                             k$texas
##                                          2.052944e-01
##                                         k$obsession
##                                          2.050783e-01
##                c$bruce_willis . k$woman_director
##                                          1.988474e-01
```

```
##                              g$crime . c$robert_de_niro
##                                            1.958690e-01
##                                              g$thriller
##                                            1.937492e-01
##                                               k$soldier
##                                            1.926180e-01
##                       k$murder . k$based_on_true_story
##                                            1.912208e-01
##                                         c$robert_de_niro
##                                            1.910289e-01
##                                             k$flashback
##                                            1.873762e-01
##                                                k$animal
##                                            1.867742e-01
##                                g$romance . k$sequel
##                                            1.852605e-01
##                              g$comedy . c$frank_welker
##                                            1.837197e-01
##                                                k$father
##                                            1.832356e-01
##                                               k$alcohol
##                                            1.819160e-01
##                                                k$dragon
##                                            1.815044e-01
##                                                 k$thief
##                                            1.809829e-01
##                                              k$survival
##                                            1.807732e-01
##                                           k$street_gang
##                                            1.782187e-01
##                              g$comedy . c$nicolas_cage
##                                            1.772172e-01
##                      g$horror . k$duringcreditsstinger
##                                            1.737471e-01
##                                           k$best_friend
##                                            1.734493e-01
##                            g$thriller . k$new_york_city
##                                            1.722169e-01
##                                     g$horror . k$murder
##                                            1.720703e-01
##                      p$universal_pictures . g$thriller
##                                            1.672134e-01
##                                         p$sony_pictures
##                                            1.671064e-01
##                                                  k$nazi
##                                            1.668462e-01
##                                                k$police
##                                            1.616242e-01
##                                              k$jealousy
##                                            1.607593e-01
##                                                 k$1940s
##                                            1.603090e-01
##                                                 k$robot
##                                            1.598364e-01
```

```
##                                   p$film4_productions
##                                          1.597078e-01
##                                           k$brutality
##                                          1.584958e-01
##                                     p$silver_pictures
##                                          1.564908e-01
##                                  g$comedy . k$sequel
##                                          1.563442e-01
##                                            k$shootout
##                                          1.534936e-01
##                            g$drama . c$robert_de_niro
##                                          1.525313e-01
##                                          k$small_town
##                                          1.512312e-01
##                            g$action . k$new_york_city
##                                          1.505046e-01
##                                   p$canal+ . g$drama
##                                          1.501847e-01
##                          g$romance . k$woman_director
##                                          1.488301e-01
##                                                  l$ar
##                                          1.484047e-01
##                                         c$john_cusack
##                                          1.476543e-01
##                          g$adventure . c$frank_welker
##                                          1.464234e-01
##                                         k$high_school
##                                          1.449714e-01
##                                           k$car_crash
##                                          1.421652e-01
##                                  g$comedy . g$family
##                                          1.402765e-01
##                                    c$antonio_banderas
##                                          1.391933e-01
##                                             p$gaumont
##                                          1.384699e-01
##                                            k$daughter
##                                          1.370441e-01
##                                         c$ethan_hawke
##                                          1.331477e-01
##                                           k$seduction
##                                          1.318896e-01
##                                               k$dream
##                                          1.311723e-01
##                                              k$island
##                                          1.310028e-01
##                          g$adventure . c$nicolas_cage
##                                          1.304604e-01
##                          p$paramount . c$frank_welker
##                                          1.300107e-01
##                                               k$death
##                                          1.263953e-01
##            c$bruce_willis . k$duringcreditsstinger
##                                          1.262757e-01
```

```
##                                        k$child_abuse
##                                         1.250211e-01
##                                           k$gunfight
##                                         1.238922e-01
##                                             k$hitman
##                                         1.235698e-01
##                                               k$noir
##                                         1.205952e-01
##                               g$action . g$thriller
##                                         1.178112e-01
##                                            k$college
##                                         1.161136e-01
##                                              k$1920s
##                                         1.156834e-01
##                           k$mother_son_relationship
##                                         1.152139e-01
##                                 g$drama . g$romance
##                                         1.149837e-01
##                     g$science_fiction . k$biography
##                                         1.135925e-01
##                                 g$thriller . k$love
##                                         1.132899e-01
##                                     c$michael_caine
##                                         1.116359e-01
##                                          k$vigilante
##                                         1.101592e-01
##                                              k$1990s
##                                         1.098731e-01
##                                             k$escape
##                                         1.074891e-01
##               c$samuel_l._jackson . c$steve_buscemi
##                                         1.059085e-01
##                            g$crime . c$bruce_willis
##                                         1.051504e-01
##                           g$family . c$frank_welker
##                                         1.047269e-01
##                                             k$satire
##                                         1.042641e-01
##                                   c$woody_harrelson
##                                         1.020701e-01
##                              p$paramount . g$horror
##                                         1.000907e-01
##                                              g$music
##                                         9.739743e-02
##                                            k$amnesia
##                                         9.325223e-02
##                     p$universal_pictures . g$comedy
##                                         9.317007e-02
##                                          k$dystopia
##                                         9.199477e-02
##                                             k$sadism
##                                         8.973561e-02
##                                           k$gangster
##                                         8.928485e-02
```

```
##                    g$crime . k$new_york_city
##                                 8.912839e-02
##                    g$comedy . c$liam_neeson
##                                 8.811772e-02
##                                    k$revenge
##                                 8.460841e-02
##                                      k$fight
##                                 7.888086e-02
##        c$frank_welker . k$duringcreditsstinger
##                                 7.670475e-02
##             p$warner_bros._pictures . g$horror
##                                 7.660251e-02
##                         k$male_homosexuality
##                                 7.651768e-02
##                 p$touchstone_pictures . k$love
##                                 7.630614e-02
##                             k$alien_invasion
##                                 7.567982e-02
##                    c$willem_dafoe . k$murder
##                                 7.424101e-02
##             p$walt_disney_pictures . g$family
##                                 7.286288e-02
##    g$science_fiction . k$duringcreditsstinger
##                                 7.281725e-02
##                                      k$train
##                                 7.245885e-02
##                  g$thriller . c$steve_buscemi
##                                 7.055254e-02
##                g$drama . k$based_on_true_story
##                                 6.987822e-02
##                           g$drama . k$revenge
##                                 6.745485e-02
##                                     k$killer
##                                 6.739238e-02
##                   g$action . c$morgan_freeman
##                                 6.690972e-02
##                       p$paramount . g$comedy
##                                 6.585300e-02
##           p$warner_bros._pictures . g$thriller
##                                 6.504889e-02
##                              g$drama . k$love
##                                 6.457807e-02
##                                         k$ex
##                                 6.447615e-02
##                                         l$de
##                                 6.433475e-02
##             p$touchstone_pictures . g$romance
##                                 6.067561e-02
##                               k$world_war_ii
##                                 5.654676e-02
##           g$romance . k$based_on_novel_or_book
##                                 5.654305e-02
##           p$columbia_pictures . c$bruce_willis
##                                 5.592149e-02
```

```
##                                 c$john_leguizamo
##                                     5.565507e-02
##                                      k$alcoholic
##                                     5.501019e-02
##              k$woman_director . k$based_on_true_story
##                                     5.027553e-02
##                              g$horror . k$sequel
##                                     4.962329e-02
##                                       k$paranoia
##                                     4.769801e-02
##                       g$romance . c$bruce_willis
##                                     4.614198e-02
##                                            k$cia
##                                     4.504530e-02
##                        k$father_son_relationship
##                                     4.194841e-02
##                           g$adventure . g$family
##                                     4.180987e-02
##                      g$romance . k$new_york_city
##                                     4.111290e-02
##           k$woman_director . k$duringcreditsstinger
##                                     4.105966e-02
##                                     k$experiment
##                                     4.002148e-02
##                                       k$new_love
##                                     3.968769e-02
##                                       k$musician
##                                     3.914901e-02
##                              g$crime . k$revenge
##                                     3.833517e-02
##                           g$thriller . g$romance
##                                     3.594109e-02
##                               k$chicago_illinois
##                                     3.369790e-02
##           c$willem_dafoe . k$based_on_novel_or_book
##                                     3.247067e-02
##                               g$crime . k$sequel
##                                     2.899846e-02
##                                       k$religion
##                                     2.742391e-02
##                                   c$bruce_willis
##                                     2.664769e-02
##                       g$crime . k$woman_director
##                                     2.472756e-02
##                                     k$conspiracy
##                                     1.953909e-02
##                        g$action . c$j.k._simmons
##                                     1.631259e-02
##                p$walt_disney_pictures . g$action
##                                     1.009559e-02
##                                g$action . k$love
##                                     8.846260e-03
##                                           k$rape
##                                     8.429605e-03
```

```
##                        k$manhattan_new_york_city
##                                     7.512869e-03
##                   g$thriller . c$morgan_freeman
##                                     4.900464e-03
##                          k$family_relationships
##                                     4.865821e-03
##                       g$crime . c$steve_buscemi
##                                     4.087352e-03
##                                       k$criminal
##                                     3.423561e-03
##                      g$drama . k$woman_director
##                                     3.283772e-03
##                               g$drama . k$sequel
##                                     7.337859e-04
##              c$robert_de_niro . c$steve_buscemi
##                                    -5.584550e-07
##                                      k$hollywood
##                                    -7.394166e-03
##        p$warner_bros._pictures . c$bruce_willis
##                                    -8.180840e-03
##                                      k$terrorism
##                                    -1.044466e-02
##                      g$crime . c$morgan_freeman
##                                    -1.048818e-02
##      p$warner_bros._pictures . k$woman_director
##                                    -1.052512e-02
##                       g$comedy . c$j.k._simmons
##                                    -1.144162e-02
##                     g$drama . g$science_fiction
##                                    -1.356548e-02
##           p$columbia_pictures . k$new_york_city
##                                    -1.557886e-02
##                                          k$dance
##                                    -1.605197e-02
##                  g$thriller . g$science_fiction
##                                    -1.861654e-02
##                     p$new_line_cinema . g$crime
##                                    -2.047067e-02
##                     p$new_line_cinema . g$drama
##                                    -2.118589e-02
##                                    c$danny_trejo
##                                    -2.144765e-02
##                c$willem_dafoe . c$steve_buscemi
##                                    -2.386744e-02
##                                      k$detective
##                                    -2.713474e-02
##                                             l$tr
##                                    -2.886830e-02
##                p$touchstone_pictures . g$family
##                                    -3.436195e-02
##                                             l$es
##                                    -3.583714e-02
##                    p$paramount . c$j.k._simmons
##                                    -3.659021e-02
```

```
##            p$walt_disney_pictures . c$frank_welker
##                                      -3.700049e-02
##            p$universal_pictures . c$robert_de_niro
##                                      -4.113969e-02
##                                     k$gay_interest
##                                      -4.268127e-02
##                                              k$gay
##                                      -4.488758e-02
##                               k$murder . k$revenge
##                                      -4.510949e-02
##                               c$christopher_walken
##                                      -4.763619e-02
##                           k$new_york_city . k$love
##                                      -5.066212e-02
##                         g$family . c$nicolas_cage
##                                      -5.699956e-02
##                  g$thriller . c$samuel_l._jackson
##                                      -5.771274e-02
##                                       k$teen_movie
##                                      -5.826037e-02
##          p$walt_disney_pictures . k$new_york_city
##                                      -5.952199e-02
##                    g$adventure . c$morgan_freeman
##                                      -5.969365e-02
##                 p$metro_goldwyn_mayer . k$revenge
##                                      -6.027484e-02
##                             g$thriller . k$murder
##                                      -6.029324e-02
##                               p$canal+ . g$horror
##                                      -6.365880e-02
##                    p$20th_century_fox . k$revenge
##                                      -6.452527e-02
##      g$science_fiction . k$based_on_novel_or_book
##                                      -6.603250e-02
##             p$warner_bros._pictures . g$adventure
##                                      -6.734768e-02
##              c$samuel_l._jackson . c$willem_dafoe
##                                      -7.067994e-02
##                           g$romance . k$biography
##                                      -7.191612e-02
##               p$canal+ . k$based_on_novel_or_book
##                                      -7.402640e-02
##       p$columbia_pictures . k$based_on_true_story
##                                      -7.693511e-02
##                     p$columbia_pictures . g$crime
##                                      -7.807017e-02
## p$warner_bros._pictures . k$based_on_novel_or_book
##                                      -7.869858e-02
##          c$nicolas_cage . k$based_on_novel_or_book
##                                      -7.990310e-02
##                                 p$millennium_films
##                                      -8.251056e-02
##                    g$adventure . c$robert_de_niro
##                                      -8.703973e-02
```

```
##                                   g$drama . g$adventure
##                                           -8.832354e-02
##                         p$20th_century_fox . g$adventure
##                                           -9.694978e-02
##                                   p$canal+ . g$thriller
##                                           -9.805932e-02
##                       k$duringcreditsstinger . k$sequel
##                                           -1.016061e-01
##                    g$action . k$based_on_novel_or_book
##                                           -1.016965e-01
##                                     k$murder . k$sequel
##                                           -1.026894e-01
##                                        p$toei_company
##                                           -1.086450e-01
##                                g$action . k$biography
##                                           -1.098864e-01
##            p$warner_bros._pictures . k$new_york_city
##                                           -1.109589e-01
##         p$universal_pictures . k$based_on_true_story
##                                           -1.110706e-01
##                                                  p$cnc
##                                           -1.142615e-01
##           k$based_on_novel_or_book . k$new_york_city
##                                           -1.203583e-01
##                    p$universal_pictures . g$adventure
##                                           -1.216796e-01
##       p$universal_pictures . k$duringcreditsstinger
##                                           -1.261041e-01
##                                  g$comedy . g$romance
##                                           -1.264724e-01
##                       p$touchstone_pictures . k$sequel
##                                           -1.283028e-01
##                                 g$adventure . k$love
##                                           -1.289082e-01
##                                  g$action . g$romance
##                                           -1.319857e-01
##                    p$warner_bros._pictures . k$sequel
##                                           -1.326224e-01
##                             k$murder . k$new_york_city
##                                           -1.359525e-01
##                                           k$rural_area
##                                           -1.452035e-01
##                                                p$ciné+
##                                           -1.473203e-01
##                                              k$robbery
##                                           -1.485616e-01
##                       g$family . k$based_on_true_story
##                                           -1.485994e-01
##                                                 k$gore
##                                           -1.503455e-01
##                             g$romance . c$j.k._simmons
##                                           -1.505677e-01
##                                           k$depression
##                                           -1.543192e-01
```

```
##                                      c$james_franco
##                                       -1.556562e-01
##                p$warner_bros._pictures . p$canal+
##                                       -1.557284e-01
##     p$warner_bros._pictures . c$samuel_l._jackson
##                                       -1.566902e-01
##                              p$canal+ . g$romance
##                                       -1.570256e-01
##                                  g$crime . k$love
##                                       -1.581539e-01
##       k$based_on_novel_or_book . k$woman_director
##                                       -1.712857e-01
##                g$comedy . k$based_on_novel_or_book
##                                       -1.735542e-01
##                      g$thriller . k$woman_director
##                                       -1.759075e-01
##              c$samuel_l._jackson . c$nicolas_cage
##                                       -1.993621e-01
##                          g$adventure . k$biography
##                                       -2.013313e-01
##                  c$frank_welker . k$new_york_city
##                                       -2.147097e-01
##    p$columbia_pictures . k$based_on_novel_or_book
##                                       -2.180465e-01
##         c$steve_buscemi . k$duringcreditsstinger
##                                       -2.201407e-01
##   p$warner_bros._pictures . k$duringcreditsstinger
##                                       -2.267263e-01
##                            g$romance . g$adventure
##                                       -2.275380e-01
##                   p$columbia_pictures . k$sequel
##                                       -2.352130e-01
##                                g$drama . g$comedy
##                                       -2.363450e-01
##                   k$woman_director . k$biography
##                                       -2.376975e-01
##    p$warner_bros._pictures . k$based_on_true_story
##                                       -2.405659e-01
## p$walt_disney_pictures . k$based_on_novel_or_book
##                                       -2.438654e-01
##              g$thriller . k$duringcreditsstinger
##                                       -2.473534e-01
##               p$warner_bros._pictures . k$murder
##                                       -2.566732e-01
##                                g$drama . g$horror
##                                       -2.613771e-01
##                                          k$writer
##                                       -2.642950e-01
##                                   k$found_footage
##                                       -2.644666e-01
##                 p$metro_goldwyn_mayer . g$action
##                                       -2.687279e-01
##                             k$biography . k$love
##                                       -2.779844e-01
```

```
##                                          k$dark_comedy
##                                          -2.785039e-01
##                         g$romance . c$robert_de_niro
##                                          -2.853557e-01
##                               p$paramount . g$family
##                                          -2.921894e-01
##                                   g$drama . g$family
##                                          -2.959880e-01
##                 p$20th_century_fox . c$steve_buscemi
##                                          -2.984458e-01
##                                     k$female_wrestler
##                                          -3.036067e-01
##                                       k$assassination
##                                          -3.153630e-01
##                       p$columbia_pictures . p$canal+
##                                          -3.213863e-01
##                                     p$téléfilm_canada
##                                          -3.323517e-01
##                  p$paramount . p$metro_goldwyn_mayer
##                                          -3.330974e-01
##                      p$universal_pictures . p$canal+
##                                          -3.619353e-01
##                   g$science_fiction . c$willem_dafoe
##                                          -3.753879e-01
##                                               k$grief
##                                          -3.762781e-01
##    p$walt_disney_pictures . k$duringcreditsstinger
##                                          -3.786336e-01
##             p$new_line_cinema . c$samuel_l._jackson
##                                          -3.826168e-01
##                k$duringcreditsstinger . k$biography
##                                          -3.842253e-01
##                                g$crime . k$biography
##                                          -3.937721e-01
##     c$samuel_l._jackson . k$based_on_novel_or_book
##                                          -3.947800e-01
##                      p$touchstone_pictures . g$crime
##                                          -4.110381e-01
##      p$warner_bros._pictures . p$columbia_pictures
##                                          -4.136583e-01
##                          p$new_line_cinema . k$murder
##                                          -4.193294e-01
##                                   g$crime . g$horror
##                                          -4.243871e-01
##                        g$romance . g$science_fiction
##                                          -4.273132e-01
##                               g$comedy . g$adventure
##                                          -4.298255e-01
##                            g$family . c$j.k._simmons
##                                          -4.356424e-01
##            p$paramount . k$based_on_novel_or_book
##                                          -4.677210e-01
##                                   g$crime . g$family
##                                          -4.710831e-01
```

```
##                        g$crime . g$science_fiction
##                                     -4.976461e-01
##                 p$warner_bros._pictures . g$family
##                                     -5.099082e-01
##      p$columbia_pictures . k$duringcreditsstinger
##                                     -5.233493e-01
##                  p$metro_goldwyn_mayer . k$sequel
##                                     -5.408620e-01
##                  g$horror . k$based_on_true_story
##                                     -5.597206e-01
##                         p$paramount . k$biography
##                                     -5.607216e-01
##                               g$comedy . g$action
##                                     -5.617953e-01
##                                   k$pro_wrestling
##                                     -5.648268e-01
##                             g$comedy . g$thriller
##                                     -5.660489e-01
##                             g$adventure . g$crime
##                                     -5.803667e-01
##                    p$metro_goldwyn_mayer . k$love
##                                     -5.866391e-01
##                 p$walt_disney_pictures . k$sequel
##                                     -5.896217e-01
##                g$thriller . k$based_on_true_story
##                                     -5.903012e-01
##                                              l$ur
##                                     -5.913147e-01
##                               g$action . g$family
##                                     -6.114573e-01
##                          c$steve_buscemi . k$love
##                                     -6.127544e-01
##                        g$romance . c$steve_buscemi
##                                     -6.168900e-01
##             p$columbia_pictures . c$j.k._simmons
##                                     -6.215784e-01
##                g$science_fiction . c$bruce_willis
##                                     -6.254616e-01
##             g$science_fiction . k$woman_director
##                                     -6.308524e-01
##                              g$romance . g$horror
##                                     -6.437923e-01
##                               g$action . g$horror
##                                     -6.852733e-01
##                         g$horror . c$steve_buscemi
##                                     -6.860559e-01
##        p$walt_disney_pictures . c$morgan_freeman
##                                     -6.936592e-01
##                               g$comedy . g$horror
##                                     -7.093510e-01
##                    c$robert_de_niro . k$biography
##                                     -7.201388e-01
##                                 g$family . k$love
##                                     -7.450912e-01
```

```
##                                          k$lgbt
##                                   -7.475970e-01
##             p$metro_goldwyn_mayer . p$canal+
##                                   -7.508359e-01
##    p$walt_disney_pictures . p$touchstone_pictures
##                                   -7.528190e-01
##                         g$adventure . g$horror
##                                   -7.532253e-01
##                   g$comedy . g$science_fiction
##                                   -7.561218e-01
##      p$20th_century_fox . k$duringcreditsstinger
##                                   -7.564058e-01
##              p$warner_bros._pictures . k$love
##                                   -7.624349e-01
##    p$20th_century_fox . k$based_on_novel_or_book
##                                   -7.887702e-01
##               c$samuel_l._jackson . k$revenge
##                                   -8.133199e-01
##  p$universal_pictures . k$based_on_novel_or_book
##                                   -8.251867e-01
##                      c$frank_welker . k$sequel
##                                   -8.375187e-01
## p$touchstone_pictures . k$based_on_novel_or_book
##                                   -8.528736e-01
##              p$touchstone_pictures . k$murder
##                                   -8.545799e-01
##            p$columbia_pictures . c$willem_dafoe
##                                   -8.591651e-01
##           p$touchstone_pictures . k$biography
##                                   -9.052055e-01
##     p$metro_goldwyn_mayer . c$samuel_l._jackson
##                                   -9.251117e-01
##   p$touchstone_pictures . k$duringcreditsstinger
##                                   -9.341935e-01
##     p$columbia_pictures . p$metro_goldwyn_mayer
##                                   -9.383670e-01
##    p$universal_pictures . p$touchstone_pictures
##                                   -9.658964e-01
##  p$warner_bros._pictures . p$metro_goldwyn_mayer
##                                   -9.870168e-01
##                                            l$pt
##                                   -9.992032e-01
##                  c$steve_buscemi . k$revenge
##                                   -1.001325e+00
##      p$universal_pictures . p$columbia_pictures
##                                   -1.042375e+00
##                  p$canal+ . k$woman_director
##                                   -1.061219e+00
##                          g$horror . g$family
##                                   -1.068087e+00
##           p$walt_disney_pictures . c$liam_neeson
##                                   -1.076769e+00
##                         p$paramount . p$canal+
##                                   -1.085632e+00
```

```
##       p$warner_bros._pictures . p$new_line_cinema
##                                     -1.143456e+00
##                                              l$fa
##                                     -1.160576e+00
##         p$universal_pictures . p$20th_century_fox
##                                     -1.228540e+00
##         p$warner_bros._pictures . c$steve_buscemi
##                                     -1.235897e+00
##                       g$horror . k$woman_director
##                                     -1.253977e+00
##               p$paramount . p$touchstone_pictures
##                                     -1.338622e+00
##                                     g$documentary
##                                     -1.344474e+00
##      p$warner_bros._pictures . p$20th_century_fox
##                                     -1.356099e+00
##                         p$canal+ . c$willem_dafoe
##                                     -1.401906e+00
##                 p$paramount . p$columbia_pictures
##                                     -1.510158e+00
##                     c$bruce_willis . k$biography
##                                     -1.523262e+00
##                         c$nicolas_cage . k$murder
##                                     -1.654725e+00
##    p$warner_bros._pictures . p$touchstone_pictures
##                                     -1.824671e+00
##             p$warner_bros._pictures . p$paramount
##                                     -1.862702e+00
##               p$columbia_pictures . c$liam_neeson
##                                     -1.921133e+00
##         p$20th_century_fox . p$columbia_pictures
##                                     -2.029013e+00
##                p$universal_pictures . p$paramount
##                                     -2.045971e+00
##          p$universal_pictures . p$new_line_cinema
##                                     -2.049035e+00
##         c$morgan_freeman . k$duringcreditsstinger
##                                     -2.144882e+00
##                 p$canal+ . p$walt_disney_pictures
##                                     -2.162287e+00
##          p$metro_goldwyn_mayer . c$morgan_freeman
##                                     -2.266051e+00
##    p$metro_goldwyn_mayer . k$duringcreditsstinger
##                                     -2.316966e+00
##                     c$steve_buscemi . k$sequel
##                                     -2.371802e+00
##                           g$horror . k$biography
##                                     -2.507689e+00
##                                       k$wrestling
##                                     -2.681321e+00
##                p$paramount . p$20th_century_fox
##                                     -2.687947e+00
##                  c$nicolas_cage . c$willem_dafoe
##                                     -2.692918e+00
```

```
##             c$bruce_willis . k$based_on_true_story
##                                      -2.817947e+00
##                                         g$tv_movie
##                                      -2.837188e+00
##         p$20th_century_fox . p$metro_goldwyn_mayer
##                                      -3.132617e+00
```

From 1269 variables, we have 674 variables in our model with interactions. Below are the intercept of the model, the top 10 variables that positively affect revenue, and top 10 variables that negatively affect revenue.

```r
ic2 <- coef2[c("(Intercept)"), 1]
paste("The intercept is ", ic2)
```

```
## [1] "The intercept is  12.4991157439431"
```

```r
paste("Top 10 variables that positively affect the revenue:")
```

```
## [1] "Top 10 variables that positively affect the revenue:"
```

```r
coef2_sort <- sort(coef2[, 1], decreasing = TRUE)[-1]
head(coef2_sort, 10)
```

```
##  c$robert_de_niro . k$woman_director                  p$columbia_pictures
##                             3.121845                             2.643334
##                        p$screen_gems                          p$paramount
##                             2.615547                             2.595986
##                 p$touchstone_pictures                  p$20th_century_fox
##                             2.561390                             2.514055
##                p$walt_disney_pictures              p$warner_bros._pictures
##                             2.420880                             2.343409
##                 p$universal_pictures g$science_fiction . c$robert_de_niro
##                             2.313728                             2.312593
```

```r
paste("Top 10 variables that negatively affects the revenue:")
```

```
## [1] "Top 10 variables that negatively affects the revenue:"
```

```r
tail(coef2_sort, 10)
```

```
##       p$metro_goldwyn_mayer . c$morgan_freeman
##                                      -2.266051
## p$metro_goldwyn_mayer . k$duringcreditsstinger
##                                      -2.316966
##                     c$steve_buscemi . k$sequel
##                                      -2.371802
##                         g$horror . k$biography
##                                      -2.507689
##                                     k$wrestling
##                                      -2.681321
##                p$paramount . p$20th_century_fox
```

```
##                                      -2.687947
##                 c$nicolas_cage . c$willem_dafoe
##                                      -2.692918
##          c$bruce_willis . k$based_on_true_story
##                                      -2.817947
##                                      g$tv_movie
##                                      -2.837188
##     p$20th_century_fox . p$metro_goldwyn_mayer
##                                      -3.132617
```

Next, we will try to see which terms are included in the model (without interactions) when we split the data set into decades.

```
coefs <- c()

for (df in df_list) {
    #### PRODUCTION COMPANIES
    sub_prod_comp <- df$production_companies
    sub_docs <- Corpus(VectorSource(sub_prod_comp))
    # Remove '-' from the name since it is the splitter
    # symbol of the data
    sub_docs <- tm_map(sub_docs, to_another, "Metro-Goldwyn-Mayer",
        "Metro_Goldwyn_Mayer")
    sub_docs <- tm_map(sub_docs, to_another, "no_production_companies",
        "")
    sub_docs <- tm_map(sub_docs, to_another, " ", "_")
    sub_docs <- tm_map(sub_docs, to_another, "-", " ")
    sub_docs <- tm_map(sub_docs, stripWhitespace)
    # adding prefix p$ for production_companies
    sub_docs <- tm_map(sub_docs, add_prefix, "p$")
    sub_docs <- tm_map(sub_docs, to_another, " ", " p$")
    sub_dtm <- DocumentTermMatrix(sub_docs)

    # get top 10 production companies
    sub_key_pc <- findMostFreqTerms(sub_dtm, 10, INDEX = rep(1,
        each = length(df[, 1])))
    # print(sub_key_pc)

    #### GENRES
    sub_genres <- df$genres
    sub_docs2 <- Corpus(VectorSource(sub_genres))
    sub_docs2 <- tm_map(sub_docs2, to_another, " ", "_")
    sub_docs2 <- tm_map(sub_docs2, to_another, "-", " ")
    sub_docs2 <- tm_map(sub_docs2, stripWhitespace)
    # adding prefix g$ for genres
    sub_docs2 <- tm_map(sub_docs2, add_prefix, "g$")
    sub_docs2 <- tm_map(sub_docs2, to_another, " ", " g$")
    sub_dtm2 <- DocumentTermMatrix(sub_docs2)

    # get top 10 genres
    sub_key_gen <- findMostFreqTerms(sub_dtm2, 10, INDEX = rep(1,
        each = length(df[, 1])))
    # sub_key_gen
```

```r
#### CREDITS
sub_casts <- df$credits
sub_docs3 <- Corpus(VectorSource(sub_casts))
sub_docs3 <- tm_map(sub_docs3, to_another, " ", "_")
# deal with Korean names
sub_docs3 <- tm_map(sub_docs3, to_another, "_([[:alpha:]]+)-([[:lower:]]+)$",
    "_\\1_\\2")
sub_docs3 <- tm_map(sub_docs3, to_another, "_([[:alpha:]]+)-([[:lower:]]+)-",
    "_\\1_\\2-")
sub_docs3 <- tm_map(sub_docs3, to_another, "-", " ")
sub_docs3 <- tm_map(sub_docs3, stripWhitespace)
# adding prefix c$ for casts
sub_docs3 <- tm_map(sub_docs3, add_prefix, "c$")
sub_docs3 <- tm_map(sub_docs3, to_another, " ", " c$")
sub_dtm3 <- DocumentTermMatrix(sub_docs3)

# get top 10 casts
sub_key_cast <- findMostFreqTerms(sub_dtm3, 10, INDEX = rep(1,
    each = length(df[, 1])))

#### KEYWORDS
sub_keywords <- df$keywords
sub_docs4 <- Corpus(VectorSource(sub_keywords))
sub_docs4 <- tm_map(sub_docs4, to_another, " ", "_")
sub_docs4 <- tm_map(sub_docs4, to_another, "-", " ")
sub_docs4 <- tm_map(sub_docs4, stripWhitespace)
# adding prefix k$ for keywords
sub_docs4 <- tm_map(sub_docs4, add_prefix, "k$")
sub_docs4 <- tm_map(sub_docs4, to_another, " ", " k$")
sub_dtm4 <- DocumentTermMatrix(sub_docs4)

# get top 10 keywords
sub_key_keys <- findMostFreqTerms(sub_dtm4, 10, INDEX = rep(1,
    each = length(df[, 1])))
# sub_key_keys

#### ORIGINAL LANGUAGE
sub_og_lng <- df$original_language
sub_docs5 <- Corpus(VectorSource(sub_og_lng))
# adding prefix l$ for language
sub_docs5 <- tm_map(sub_docs5, add_prefix, "l$")
sub_docs5 <- tm_map(sub_docs5, to_another, " ", " l$")
sub_dtm5 <- DocumentTermMatrix(sub_docs5)

# get all languages in at least 50 movies
sub_key_lang <- findMostFreqTerms(sub_dtm5, 10, INDEX = rep(1,
    each = length(df[, 1])))
# sub_key_lang

sub_X <- cbind(sub_dtm[, names(sub_key_pc[[1]])], sub_dtm2[,
    names(sub_key_gen[[1]])], sub_dtm3[, names(sub_key_cast[[1]])],
    sub_dtm4[, names(sub_key_keys[[1]])], sub_dtm5[, names(sub_key_lang[[1]])])
sub_y <- log(df$revenue_adjusted)
```

```r
    sub_model1 <- cv.glmnet(cbind(as.matrix(sub_X), budget = log(df$budget)),
        sub_y)
    sub_coef1 <- coef(sub_model1, s = "lambda.min")
    sub_coef_sort <- sort(sub_coef1[which(sub_coef1 != 0), 1],
        decreasing = TRUE)
    coefs <- c(coefs, sub_coef_sort)

    print(df$decades[1])
    print(sub_coef_sort)

}
```

```
## [1] 1910
##                           (Intercept) p$jesse_l._lasky_feature_play_company
##                             17.619188                             -2.037566
## [1] 1920
##                  (Intercept)                         l$en
##                 12.872678334                  2.515355239
##                      g$music                k$world_war_i
##                  1.572596046                  1.560638871
##      k$based_on_novel_or_book p$charles_chaplin_productions
##                  0.822351799                  0.700643326
##               c$harold_lloyd                 k$silent_film
##                  0.698528915                  0.544411996
##            p$united_artists                     g$romance
##                  0.428811161                  0.391860588
##                       k$code   p$the_vitaphone_corporation
##                  0.347126199                  0.281911891
##                     g$action                c$john_gilbert
##                  0.242734866                  0.219200601
##                  g$adventure                       g$drama
##                  0.212686060                  0.070313929
##              c$noble_johnson       p$warner_bros._pictures
##                  0.057078012                  0.051760059
##           c$nigel_de_brulier                         k$pre
##                  0.047115097                  0.008536016
##                         l$zh
##                 -5.169422687
## [1] 1930
##                 (Intercept)                      l$en              g$adventure
##                 13.35492875                3.82657050               0.39509310
##              c$irving_bacon  k$based_on_novel_or_book                k$musical
##                  0.39501978                0.27075923               0.23081056
##           k$black_and_white               c$clark_gable              c$ward_bond
##                  0.04833436                0.03865860               0.01243344
##     p$warner_bros._pictures                   g$drama      p$rko_radio_pictures
##                 -0.02852272               -0.05415548              -0.66325381
## p$first_national_pictures                      l$zh                     l$sv
##                 -1.80483687               -3.43722103              -4.02346834
## [1] 1940
##                 (Intercept)                      l$en                 g$family
##                  15.5739766                 1.7904567                0.6221444
## p$walt_disney_productions          c$bert_moorhouse        k$black_and_white
```

```
##                0.3285450                 0.3108720                 0.2751412
##                  g$drama                      l$fr
##                0.0116140                -3.1339220
## [1] 1950
##              (Intercept)                      l$en p$walt_disney_productions
##              14.01406765                2.73212189                1.83971429
##                   k$epic             c$bess_flowers                g$romance
##               0.74312169                0.47505151                0.32310133
##               g$thriller                 k$musical  k$based_on_novel_or_book
##               0.23444109                0.22173996                0.19897367
##         c$franklyn_farnum                      l$it                      l$zh
##               0.09930969               -0.30331202               -4.32828693
## [1] 1960
##              (Intercept)                       l$el
##              13.79014960                 2.93332389
##                     l$en                       l$ja
##               2.91435762                 2.12048806
## k$based_on_play_or_musical                     l$it
##               1.36996240                 1.25700882
##             c$frank_baker                    k$epic
##               1.06703617                 0.83283244
##          p$united_artists               g$adventure
##               0.79462128                 0.78341680
##             c$paul_newman    p$warner_bros._pictures
##               0.68694503                 0.66826751
##                 k$musical               c$john_wayne
##               0.63191755                 0.60100324
##           k$new_york_city                   g$action
##               0.49003242                 0.48393764
##         p$20th_century_fox  k$based_on_novel_or_book
##               0.39508430                 0.38180402
##      p$metro_goldwyn_mayer       p$columbia_pictures
##               0.36525624                 0.30889991
##                g$thriller                      g$war
##               0.26071218                 0.24500781
##                  c$al_bain            c$arthur_tovey
##               0.20794112                 0.18202296
##                   g$drama                    g$crime
##               0.14596997                 0.12853139
##            c$bert_stevens                 g$western
##               0.08529632                 0.07598935
##                      l$tr                      l$es
##              -0.06728581                -0.08449426
##                    c$jean                k$cult_film
##              -0.18116967                -0.30000364
##                      l$fa                      l$sv
##              -2.03212288                -2.15777742
## [1] 1970
##              (Intercept)                      l$en                      l$fr
##              13.87015674                3.00268601                2.26708756
##                     l$it                      l$ja                      l$sv
##               2.18202972                1.94060889                1.55480931
##              p$paramount   p$warner_bros._pictures p$walt_disney_productions
##               1.27936799                1.07138399                1.04815586
```

```
##         c$arthur_tovey       p$columbia_pictures        p$20th_century_fox
##             0.92969578                0.87519520                0.75186122
##   p$universal_pictures          p$united_artists           k$new_york_city
##             0.54677529                0.53072478                0.48793989
##              k$musical         g$science_fiction                g$thriller
##             0.43787811                0.39385770                0.38601106
##               g$action                  g$comedy                  k$police
##             0.34800193                0.32897615                0.31015823
##        c$burt_reynolds    p$metro_goldwyn_mayer    p$malpaso_productions
##             0.21870918                0.18933499                0.08099298
##                g$crime           c$robert_duvall               g$adventure
##             0.05604740                0.04842614                0.04752269
##              g$romance              c$ned_beatty                      l$tr
##             0.02824088                0.02440882               -0.21166567
##               k$sports         c$m._emmet_walsh                 g$mystery
##            -0.35207672               -0.44310853               -0.48344011
## [1] 1980
##            (Intercept)                      l$en               p$paramount
##            13.89905010                1.82778249                1.81906516
##   c$sylvester_stallone        p$20th_century_fox                      l$sv
##             1.48692788                1.46321868                1.44100156
##   p$universal_pictures                      l$ja             c$frank_welker
##             1.32452977                1.24234903                1.23232011
##  p$warner_bros._pictures    p$metro_goldwyn_mayer                  k$sequel
##             1.15589328                1.06691026                0.91995750
##     p$tristar_pictures             c$dan_aykroyd      p$columbia_pictures
##             0.85515911                0.81154316                0.77466050
##        k$new_york_city          p$orion_pictures          c$m._emmet_walsh
##             0.71814985                0.71626304                0.70287922
##                   l$it               g$adventure k$based_on_novel_or_book
##             0.61340999                0.58225244                0.52375062
##         k$martial_arts                 k$revenge              c$john_candy
##             0.50697243                0.50348480                0.44735945
##                 c$jean             c$peter_jason k$los_angeles_california
##             0.44558948                0.34885817                0.30536553
##               g$comedy                  k$police           c$robert_loggia
##             0.28300566                0.20666617                0.15690247
##                   l$cn                g$thriller                 g$fantasy
##             0.14726297                0.13083631                0.10518175
##              g$romance         g$science_fiction                   g$crime
##             0.10202353                0.09583170                0.06829004
##         p$cannon_group                      l$es                  g$horror
##            -0.04652501               -0.16066176               -0.21423347
##                g$drama              c$dick_miller          k$woman_director
##            -0.31123069               -0.49666933               -0.53212941
##                   l$tr                      l$zh
##            -0.81733132               -1.18439981
## [1] 1990
##            (Intercept)        p$20th_century_fox               p$paramount
##           13.720757127               2.206850882               2.087479002
##    p$columbia_pictures      p$universal_pictures    p$touchstone_pictures
##            1.886018123               1.881050193               1.738760563
##     p$tristar_pictures  p$warner_bros._pictures                      l$ja
##            1.644264077               1.613253595               1.610805786
```

```
##            c$frank_welker          c$robin_williams     p$hollywood_pictures
##               1.560736571               1.511392863              1.474940512
##          p$new_line_cinema            c$bruce_willis          c$robert_de_niro
##               1.382403003               0.987984202              0.955454456
## k$based_on_novel_or_book            k$martial_arts                     l$en
##               0.889702102               0.836977454              0.828841150
##      c$samuel_l._jackson               g$adventure                p$miramax
##               0.823513621               0.785618621              0.692722691
##               g$thriller           k$new_york_city                     l$it
##               0.685003161               0.621321141              0.591272377
## k$los_angeles_california     c$thomas_rosales_jr.             c$dan_hedaya
##               0.553431907               0.489165535              0.471287020
##                     l$hi                  g$family                 g$comedy
##               0.435066306               0.398755244              0.367024631
##                g$romance                  g$action                  c$jones
##               0.345123882               0.314035654              0.157788389
##                k$revenge                  k$police                 g$horror
##               0.140494122               0.136421170              0.133583726
##                     l$ta         g$science_fiction                   c$jean
##               0.132756870               0.125371487              0.088701763
##                 k$murder                      l$de                  g$drama
##               0.027585913              -0.004287399             -0.057595978
##          k$woman_director                      l$zh
##              -0.279046763              -1.837899989
## [1] 2000
##                    (Intercept)        p$columbia_pictures
##                    12.61244355                 2.74771847
##      p$warner_bros._pictures       p$universal_pictures
##                     2.66962573                 2.60278671
##            p$new_line_cinema     p$walt_disney_pictures
##                     2.57037434                 2.33777122
##        p$dreamworks_pictures          p$20th_century_fox
##                     2.31671042                 2.29660207
##                  p$paramount                       l$ja
##                     2.23948373                 1.99134851
##                         l$hi                       l$ko
##                     1.93195968                 1.92529145
##                         l$de                       l$it
##                     1.82036494                 1.78232946
##                    p$miramax     k$duringcreditsstinger
##                     1.69680067                 1.52386840
##          c$samuel_l._jackson                       l$en
##                     1.38103387                 1.36010170
##                      c$jones             c$bruce_willis
##                     1.34930001                 1.34829750
##                      p$canal+              c$keith_david
##                     1.20485638                 1.15624653
## k$parent_child_relationship         k$loss_of_loved_one
##                     1.02933395                 0.96657352
##                         l$fr                       l$zh
##                     0.95346062                 0.94644874
##                         l$es   k$based_on_novel_or_book
##                     0.92159115                 0.88467866
##                 k$friendship                  g$fantasy
```

```
##                0.86685655                0.77882279
##            c$willem_dafoe                  g$action
##                0.76114084                0.74616265
##           k$new_york_city               g$adventure
##                0.71088120                0.67138589
##          c$morgan_freeman                  g$family
##                0.55706324                0.53629208
##                g$thriller                 g$romance
##                0.52535440                0.49146181
##                      l$ru                  g$comedy
##                0.48953322                0.41782800
##                    c$jean          c$david_koechner
##                0.39149510                0.34113609
##                   c$marie                  k$murder
##                0.33485901                0.33380671
##                  k$sports                   g$crime
##                0.32726149                0.19847428
##                  g$horror          k$woman_director
##                0.19144594                0.05502081
##                   g$drama                 k$revenge
##               -0.09068283               -0.09601916
##             c$justin_long
##               -0.10545322
## [1] 2010
##             (Intercept) p$warner_bros._pictures      p$columbia_pictures
##            12.322133772             3.401782560              3.216685469
##       p$20th_century_fox              p$paramount  p$walt_disney_pictures
##             3.131444226             3.100950661              3.029579880
##     p$universal_pictures              p$lionsgate k$duringcreditsstinger
##             2.881218723             2.275124953              2.134907953
##      c$samuel_l._jackson            c$liam_neeson                 k$sequel
##             1.756017270             1.669748150              1.666026750
##                     l$ja       p$relativity_media                     l$ko
##             1.480024703             1.474050135              1.403209998
## k$based_on_novel_or_book                     l$hi              k$biography
##             1.397301271             1.377900301              1.293368335
##    k$based_on_true_story                     l$zh              g$adventure
##             1.236522516             1.201010232              1.167805572
##             k$friendship             c$joe_chrest                g$fantasy
##             0.932092714             0.924104306              0.876887070
##                  p$canal+            p$studiocanal                 g$comedy
##             0.780160044             0.776022433              0.767181449
##               g$thriller             c$bill_hader                 g$action
##             0.724151839             0.663429052              0.661057856
##                 g$family                g$romance                   k$love
##             0.647299240             0.625278838              0.529420113
##                   c$jean                k$revenge                  c$smith
##             0.519376715             0.492722602              0.487087961
##                  c$jones                     l$en                  c$marie
##             0.398056765             0.384382114              0.337754379
##                     l$fr                  g$crime                     l$ml
##             0.330218603             0.289608133              0.288904415
##          c$j.k._simmons                 k$murder                  g$drama
##             0.139690495             0.129353692              0.113813158
```

```
##                    l$ru                  g$horror          k$woman_director
##             0.008290509              -0.036884321              -0.087694833
##                    l$de             c$james_franco                      l$es
##            -0.139944734              -0.217332030              -0.363613470
## [1] 2020
##             (Intercept)                      l$zh               p$paramount
##            12.087304591               3.921624687               3.180876603
##     p$universal_pictures  p$warner_bros._pictures      p$columbia_pictures
##             2.738808140               2.673959836               2.294189751
##                k$sequel          p$focus_features   k$duringcreditsstinger
##             2.110304894               1.797544324               1.697885299
##                    l$ko                     c$joy           p$bron_studios
##             1.693729477               1.524905682               1.523702809
##         k$based_on_comic       c$anthony_molinari                  k$murder
##             1.422460670               1.363106498               1.306277112
##                g$action               p$lionsgate                      l$ja
##             1.289708784               1.192233090               1.143895503
##                    l$fr               g$adventure     k$based_on_true_story
##             1.068539163               1.047272717               0.991453298
##                 k$anime k$based_on_novel_or_book    k$aftercreditsstinger
##             0.953351926               0.844944935               0.831584008
##                p$canal+                   g$drama                 g$fantasy
##             0.803874310               0.517732182               0.454619018
##                g$comedy               k$superhero                    c$jean
##             0.384640959               0.338885732               0.333984681
##              g$thriller                  g$family           c$michelle_yeoh
##             0.288610630               0.265129662               0.249500713
##  p$blumhouse_productions               g$animation         p$ingenious_media
##             0.221367373               0.198826393               0.118964231
##               g$romance                   c$jones          k$woman_director
##             0.118189999               0.116224343               0.027057797
##                g$horror                      l$de                      l$ru
##            -0.006439627              -0.092382844              -0.193508780
##                    l$es
##            -0.542501300
```

```
# coefs
```