Movie Data

Gözde Nur Özdemir

2023-08-27

- This data set can be download from here

Overview of the Data

1. Index: Index of the each row.
2. Title: Title of the movies.
3. Original Language: Gives us the language of movies.
4. Release Date: When the movie was officially released for public viewing.
5. Popularity: The measure of how well-known or talked-about a particular movie is within a given context.
6. Vote Average: The average rating or score given to the movie by viewers who have voted.
7. Vote Count: The number of votes or ratings that the movie has received from viewers.
8. Overview: Summary or description of the movie plot, themes, and overall content.

Requirements

```
# Please first download required library
library(tidyverse)
library(ggplot2)
library(dplyr)
library(rvest)
library(stringr)
library(corrplot)
```

My Findings

First reorganize the data.

```
dt1$index<-1:10000
which(is.na(dt1))
```

```
## integer(0)
```

```
# There is no any NA values
which(is.null(dt1))
```

```
## integer(0)
```

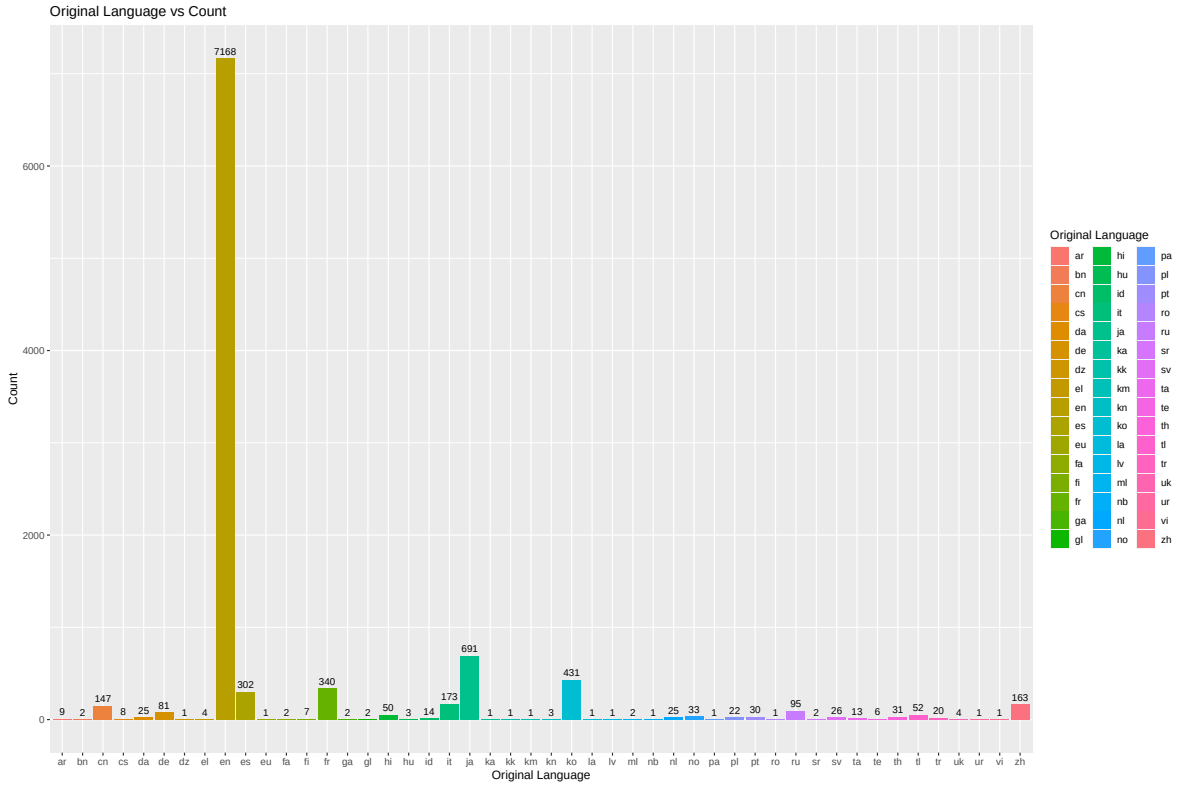
```
# There is no any NULL values
dt1$original_language<-factor(dt1$original_language)
summary(dt1)
```

```
##      index          title           original_language release_date
##  Min.   :    1   Length:10000       en     :7168      Length:10000
##  1st Qu.: 2501   Class :character   ja     : 691      Class :character
##  Median : 5000   Mode  :character   ko     : 431      Mode  :character
##  Mean   : 5000                      fr     : 340
##  3rd Qu.: 7500                      es     : 302
##  Max.   :10000                      it     : 173
##                                     (Other): 895
##    popularity        vote_average     vote_count      overview
##  Min.   :   8.128   Min.   : 0.00   Min.   :    0   Length:10000
##  1st Qu.:  14.931   1st Qu.: 5.90   1st Qu.:  136   Class :character
##  Median :  19.241   Median : 6.50   Median :  510   Mode  :character
##  Mean   :  32.312   Mean   : 6.32   Mean   : 1561
##  3rd Qu.:  28.887   3rd Qu.: 7.10   3rd Qu.: 1602
##  Max.   :3368.627   Max.   :10.00   Max.   :34245
##
```

As we can see there are many English movies were made.

```
plot1 <- ggplot(data = dt1, aes(x = original_language)) +
  geom_bar(aes(fill = original_language),stat = "count") +
  geom_text(stat = "count", aes(label = stat(count)),
            vjust = -0.5, size = 3, color = "black") +
  xlab("Original Language") +
  ylab("Count")+labs(title="Original Language vs Count")+
  guides(fill=guide_legend(title="Original Language"))


plot1
```



Now we can look at the table.

```
popular_film<-ifelse(dt1$vote_average>mean(dt1$vote_average),dt1$index,NA)
popular_film<-na.omit(popular_film)
popular_film_lan<-dt1$original_language[popular_film]
table(popular_film_lan)
```

```
## popular_film_lan
##   ar   bn   cn   cs   da   de   dz   el   en   es   eu   fa   fi   fr   ga   gl
##    7    0   70    5   19   50    1    2 4111  190    1    2    2  204    2    2
##   hi   hu   id   it   ja   ka   kk   km   kn   ko   la   lv   ml   nb   nl   no
##   40    1    9   80  538    1    1    1    3  204    1    0    2    1   11   20
##   pa   pl   pt   ro   ru   sr   sv   ta   te   th   tl   tr   uk   ur   vi   zh
##    0   10   14    1   53    1   15   10    5   21   11   15    4    1    1   96
```

As we can understand from the plot and table popular films are mostly English. This may have 2 reasons first reason is English is worldwide language, but second maybe people speak English are mostly love their

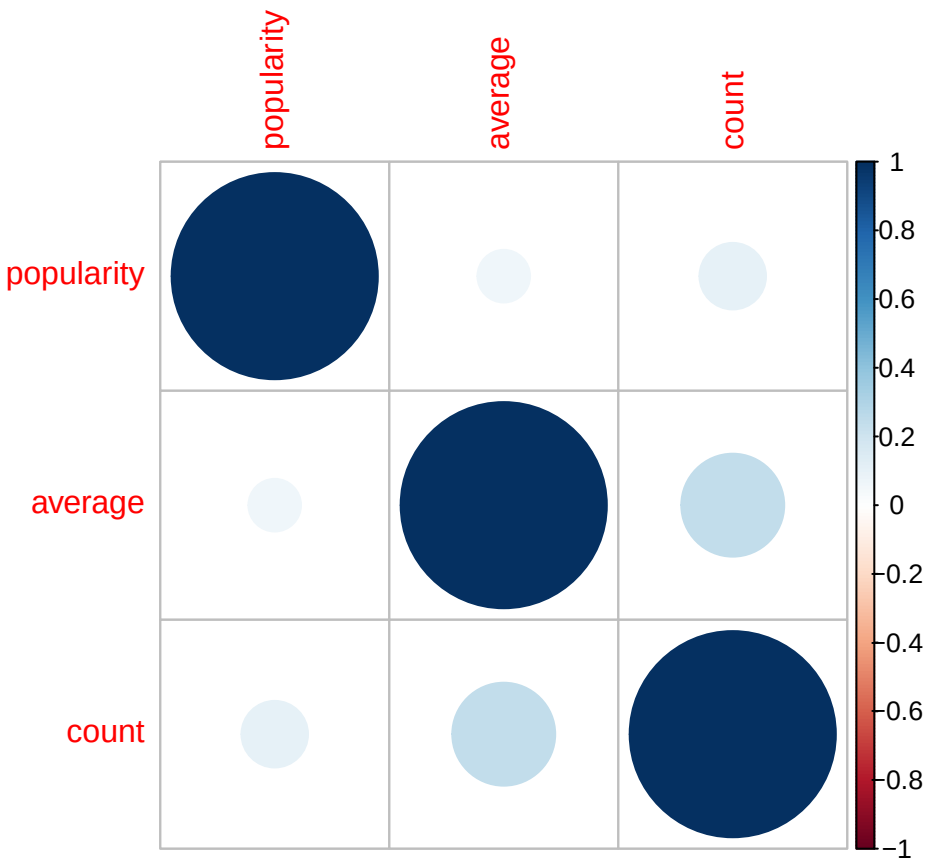
films so they give higher vote. This affect the vote average so to understand which reason is true we also have to look at the vote count to be more robust.

```
popular_film_vote_count<-dt1$vote_count[popular_film]
popular<-data.frame(voteCount=popular_film_vote_count,language=popular_film_lan)
popularVoteCount<-popular[popular$voteCount>mean(popular_film_vote_count),]
rate_of_movie<-na.omit(summary(popularVoteCount$language)/summary(popular$language))
rate_of_movie
```

```
##         ar         cn         cs         da         de         dz         el
## 0.00000000 0.04285714 0.00000000 0.10526316 0.12000000 0.00000000 0.00000000
##         en         es         eu         fa         fi         fr         ga
## 0.35611773 0.05789474 0.00000000 0.00000000 0.00000000 0.09803922 0.00000000
##         gl         hi         hu         id         it         ja         ka
## 0.00000000 0.05000000 0.00000000 0.22222222 0.15000000 0.03717472 0.00000000
##         kk         km         kn         ko         la         ml         nb
## 0.00000000 0.00000000 0.00000000 0.04411765 0.00000000 0.00000000 0.00000000
##         nl         no         pl         pt         ro         ru         sr
## 0.00000000 0.00000000 0.10000000 0.07142857 0.00000000 0.00000000 0.00000000
##         sv         ta         te         th         tl         tr         uk
## 0.20000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000
##         ur         vi         zh
## 0.00000000 0.00000000 0.01041667
## attr(,"na.action")
## bn lv pa
##  2 28 33
## attr(,"class")
## [1] "omit"
```

English has the highest rate outside of other films So English is more popular because of the first reason to be more precise let's look at the statistics As per the research conducted in 2022, 17% of the world's population, i.e., 1.5 billion people, speak the English language worldwide Research conducted by Statista since English is 34% so it is more than the result above.

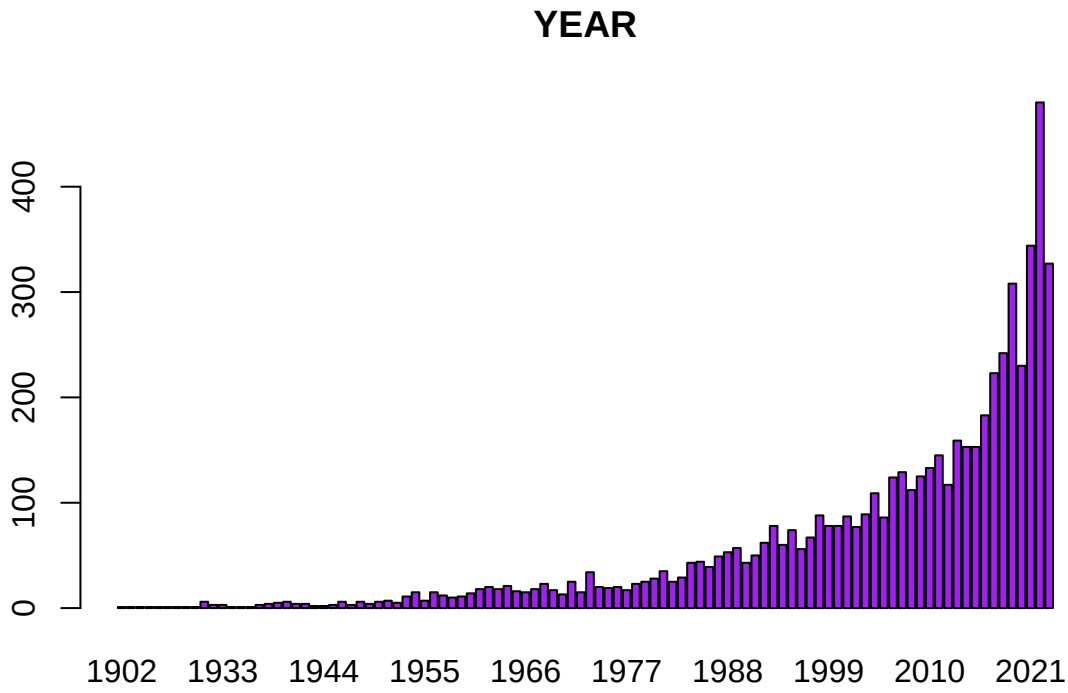
```
cor_matrix<-data.frame(popularity=dt1$popularity,average=dt1$vote_average,count=dt1$vote_count)
cor_matrix<-cor(cor_matrix)
corrplot(cor_matrix)
```



Moreover, this is the correlation matrix.

Which Year The Most Popular Films Were Made?

```
popular_film_date<-dt1$release_date[popular_film]
date<-unlist(strsplit(popular_film_date, split = "-"))
date<-as.integer(date)
year<-ifelse(date>1000,date,NA)
year<-na.omit(year) %>% factor()
plot(year,col="purple",col.main="purple")+title("YEAR")
```




```
## numeric(0)
```

As we can see in 2022, the most films are made. In 2023,2021,2019,2018 also more film made but why in 2022 were made more films? This is mainly about pandemic. After the pandemic film industry came back in 2022. Also blocbuster movies were came back after the covid-19. For more information please check the website

Most Popular Genre in 2022

```
link="https://www.movieinsider.com/movies/2022"
page=read_html(link)
name=page %>% html_nodes('.col-md-5 a , .label-default') %>% html_text()
web_data<-data.frame(title=name[seq(1,length(name)-1,2)],genre=name[seq(2,length(name),2)])
popular_film_title<-dt1$title[popular_film]
popular_data<-data.frame(title=popular_film_title,date=popular_film_date)
web_data$title<-trimws(web_data$title)
popular_data$date <- substr(popular_data$date, start = 7, stop = 10)
popular_data$date<-ifelse(popular_data$date==2022,2022,NA)
popular_dat1a<-na.omit(popular_data)
common_titles <- intersect(popular_data$title, web_data$title)
common_titles_genre <- web_data[web_data$title %in% common_titles, c("title", "genre")]
common_titles_genre$genre<-factor(common_titles_genre$genre)
summary(common_titles_genre)
```

```
##     title                     genre
##  Length:93           Drama      :23
##  Class :character    Thriller   :13
##  Mode  :character    Action     :10
##                      Adaptation : 6
##                      Comedy     : 6
##                      Animation  : 5
##                     (Other)     :30
```

As we can understand in 2022 popular films' genre is drama. And then action,adaption,thriller and so on. But why drama is more popular genre in 2022. Since drama based on real event so world community most likes the drama genre film. So if film makers want to make film, they should choose mainly drama type. For more information about this popular genre type please visit the web site

Recommendation System

First we need to make data frame for the new data.

```
# vote average
v<-vector()
k<-0
m<-0
for(i in common_titles_genre$title){
  m<-m+1
  k<-0
  for(l in dt1$title){
    k<-k+1
    if(i==l){
      v[m]<-dt1$vote_average[k]
    }else{
      next
    }
  }
}
v<-na.omit(v)
common_titles_genre$voteAverage<-v
# language
dt1$original_language<-as.character(dt1$original_language)
v<-vector()
k<-0
m<-0
for(i in common_titles_genre$title){
  m<-m+1
  k<-0
  for(l in dt1$title){
    k<-k+1
    if(i==l){
      v[m]<-dt1$original_language[k]
    }else{
      next
    }
  }
```

```
}
v<-na.omit(v)
common_titles_genre$language<-v
common_titles_genre
```

```
##                                                  title            genre
## 1    Harry Potter 20th Anniversary: Return to Hogwarts      Documentary
## 12                             The Legend of La Llorona         Thriller
## 17                                   Shin Ultraman        Animation
## 18                                           Belle        Animation
## 19                                         Eternals           Sci-Fi
## 20                                          Scream          Mystery
## 27                                           Belle        Animation
## 34                   Hotel Transylvania: Transformania        Animation
## 248                                            Spy  ¡Viva Maestro!
## 566                                           Kids       They/Them
## 570                                           Kids      I Am Groot
## 807                                American Murderer         Thriller
## 808                            Detective Knight: Rogue        Adventure
## 810                                  Wendell & Wild      Stop-Motion
## 819                                The Good Nurse       Adaptation
## 821                                           Till            Drama
## 822                                      Call Jane            Drama
## 823                                Prey for the Devil         Thriller
## 825                                   Armageddon Time            Drama
## 828                                      Holy Spider            Drama
## 831                      All Quiet on the Western Front              War
## 832                                  Wendell & Wild          Comedy
## 837                                      The Wonder         Thriller
## 839                                   Armageddon Time            Drama
## 840                                    Shadow Master         Thriller
## 842                                         Causeway            Drama
## 860                                  Enola Holmes 2           Sequel
## 861                                   My Policeman          Romance
## 864                     Weird: The Al Yankovic Story          Comedy
## 865                              Don't Worry Darling         Thriller
## 866                                    Shadow Master           Action
## 871                                    Lost Bullet 2            Drama
## 874                  Black Panther: Wakanda Forever           Sci-Fi
## 879                                        Spirited          Musical
## 883                                  The Fabelmans            Drama
## 888                  R.I.P.D. 2: Rise of the Damned           Action
## 893                                      The Wonder         Thriller
## 900                                        The Menu           Satire
## 901                                       She Said            Drama
## 902                                  Bones and All        Adventure
## 903                                The Inspection            Drama
## 907                     Scrooge: A Christmas Carol          Musical
## 910                                        Missing         Thriller
## 913                                         Taurus            Drama
## 917                                    Slumberland           Family
## 918                                   Disenchanted           Family
## 921                                        Spirited        Adventure
```

```
## 923                                          Nope        Thriller
## 924                                  Strange World         Family
## 928                                       Devotion         Action
## 929                                  Bones and All      Adaptation
## 930                                  The Fabelmans   Coming-of-Age
## 931                       Lady Chatterley's Lover      Adaptation
## 935                                  The Swimmers       Biography
## 946                                       The Son          Drama
## 950                 Puss in Boots: The Last Wish          Comedy
## 953                          My Name Is Vendetta        Thriller
## 960                                  Violent Night          Crime
## 961                                  Women Talking          Drama
## 965                                   Emancipation        Thriller
## 967                           Top Gun: Maverick          Sequel
## 969                           Savage Salvation          Action
## 973       The Quintessential Quintuplets Movie            Teen
## 975                                          Hunt          Action
## 986                       Lady Chatterley's Lover      Adaptation
## 987                 Scrooge: A Christmas Carol         Holiday
## 990          Diary of a Wimpy Kid: Rodrick Rules       Animation
## 992                                          Bros          Comedy
## 1001                                     Bed Rest        Thriller
## 1003                                    The Whale          Drama
## 1005                                Empire of Light          Drama
## 1013                                     Saint Omer          Drama
## 1018              Guillermo del Toro's Pinocchio      Adaptation
## 1020  Night at the Museum: Kahmunrah Rises Again          Reboot
## 1023                                  Emancipation          Drama
## 1026                          Conan the Barbarian          Action
## 1032                          Conan the Barbarian       Adventure
## 1034                                   The Big 4          Action
## 1035                     Avatar: The Way of Water      Shot-In-3D
## 1037                           As Good as Dead          Action
## 1038                                   Mindcage        Thriller
## 1045                            The Quiet Girl          Drama
## 1053                                Black Adam          Action
## 1057                 Puss in Boots: The Last Wish          Comedy
## 1059                           Top Gun: Maverick          Action
## 1061                                   Babylon          Period
## 1063                                   Corsage          Drama
## 1064                          The Pale Blue Eye      Adaptation
## 1066                                    Living          Drama
## 1069                          After Ever Happy          Drama
## 1070                                    Broker          Drama
## 1074                                    Broker          Drama
## 1075                          A Man Called Otto          Comedy
##      voteAverage language
## 1            7.3       en
## 12           7.8       es
## 17           7.1       ja
## 18           7.2       en
## 19           6.9       en
## 20           7.4       en
## 27           7.2       en
```

```
## 34          7.0       en
## 248         6.8       en
## 566         6.9       en
## 570         6.9       en
## 807         6.6       en
## 808         6.6       en
## 810         6.7       en
## 819         7.0       en
## 821         7.4       en
## 822         6.6       en
## 823         7.1       en
## 825         6.6       en
## 828         7.4       fa
## 831         6.6       en
## 832         6.7       en
## 837         6.7       en
## 839         6.6       en
## 840         6.9       en
## 842         6.7       en
## 860         7.6       en
## 861         7.8       en
## 864         6.8       en
## 865         6.8       en
## 866         6.9       en
## 871         6.6       fr
## 874         7.2       en
## 879         7.0       en
## 883         7.7       en
## 888         6.5       en
## 893         6.7       en
## 900         7.2       en
## 901         7.3       en
## 902         7.2       it
## 903         6.5       en
## 907         6.7       en
## 910         7.6       en
## 913         7.1       en
## 917         7.4       en
## 918         6.9       en
## 921         7.0       en
## 923         6.9       en
## 924         6.4       en
## 928         7.2       en
## 929         7.2       it
## 930         7.7       en
## 931         5.4       en
## 935         7.5       en
## 946         6.6       en
## 950         8.3       en
## 953         6.7       it
## 960         7.5       en
## 961         6.9       en
## 965         7.9       en
## 967         8.3       en
```

```
## 969          6.5       en
## 973          8.4       ja
## 975          6.7       ko
## 986          5.4       en
## 987          6.7       en
## 990          6.6       en
## 992          6.9       en
## 1001         6.6       en
## 1003         8.0       en
## 1005         6.6       en
## 1013         6.4       fr
## 1018         8.2       en
## 1020         6.6       en
## 1023         7.9       en
## 1026         6.8       en
## 1032         6.8       en
## 1034         7.0       id
## 1035         7.7       en
## 1037         6.7       en
## 1038         6.5       en
## 1045         7.5       ga
## 1053         7.0       en
## 1057         8.3       en
## 1059         8.3       en
## 1061         7.5       en
## 1063         6.5       de
## 1064         6.9       en
## 1066         7.0       en
## 1069         6.8       en
## 1070         7.2       ko
## 1074         7.2       ko
## 1075         7.8       en
```

We can find more easily their type and their information.