MOVIE ANALYSIS

YE IN JEON

Problem statement

- What affects the Success of a Movie?

- Current Trend:
  - Expanding Movie Industry
  - Increasing Importance of Movie Revenue

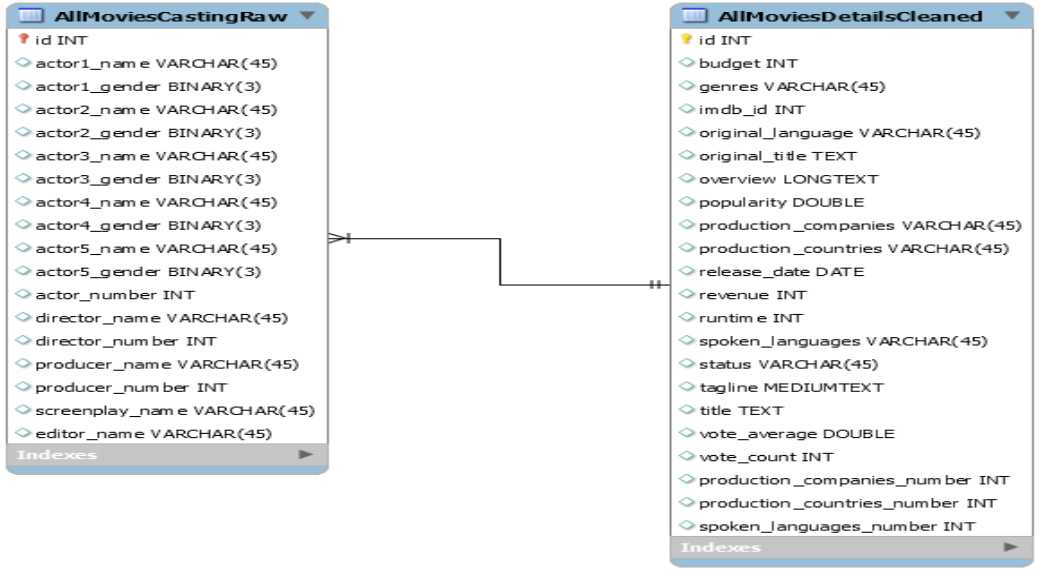
- Definition of success
  - Economic : Revenue
  - Cinematic quality : Vote average

Background of Datasets

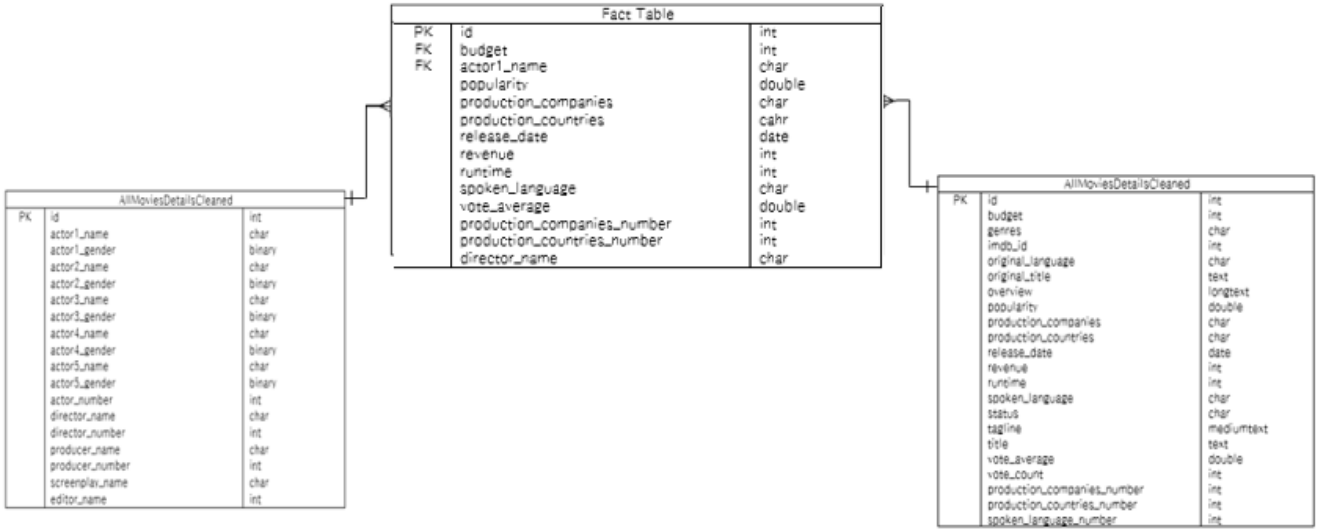
- 2 Datasets from Kaggle, Originally from TMDB(The Movie Database)
- Number of Movies : 350,000+ movies
- Containing Variables
  - AllMoviesDetailsCleaned – Id, budget, genres, imdb_id, original_language, original_title, overview, popularity, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, vote_average, vote_count, production_companies_number, production_countries_number, spoken_languages_number
  - AllMoviesCastingRaw – id, actor1_name, actor1_gender, actor2_name, actor2_gender, actor3_name, actor3_gender, actor4_name, actor4_gender, actor5_name, actor5_gender, actor_number, director_name, director_gender, director_number, producer_name, producer_number, screeplay_name, editor_name

Data Model Diagram (MySQL)

ER Diagram



Data Dimensional Model



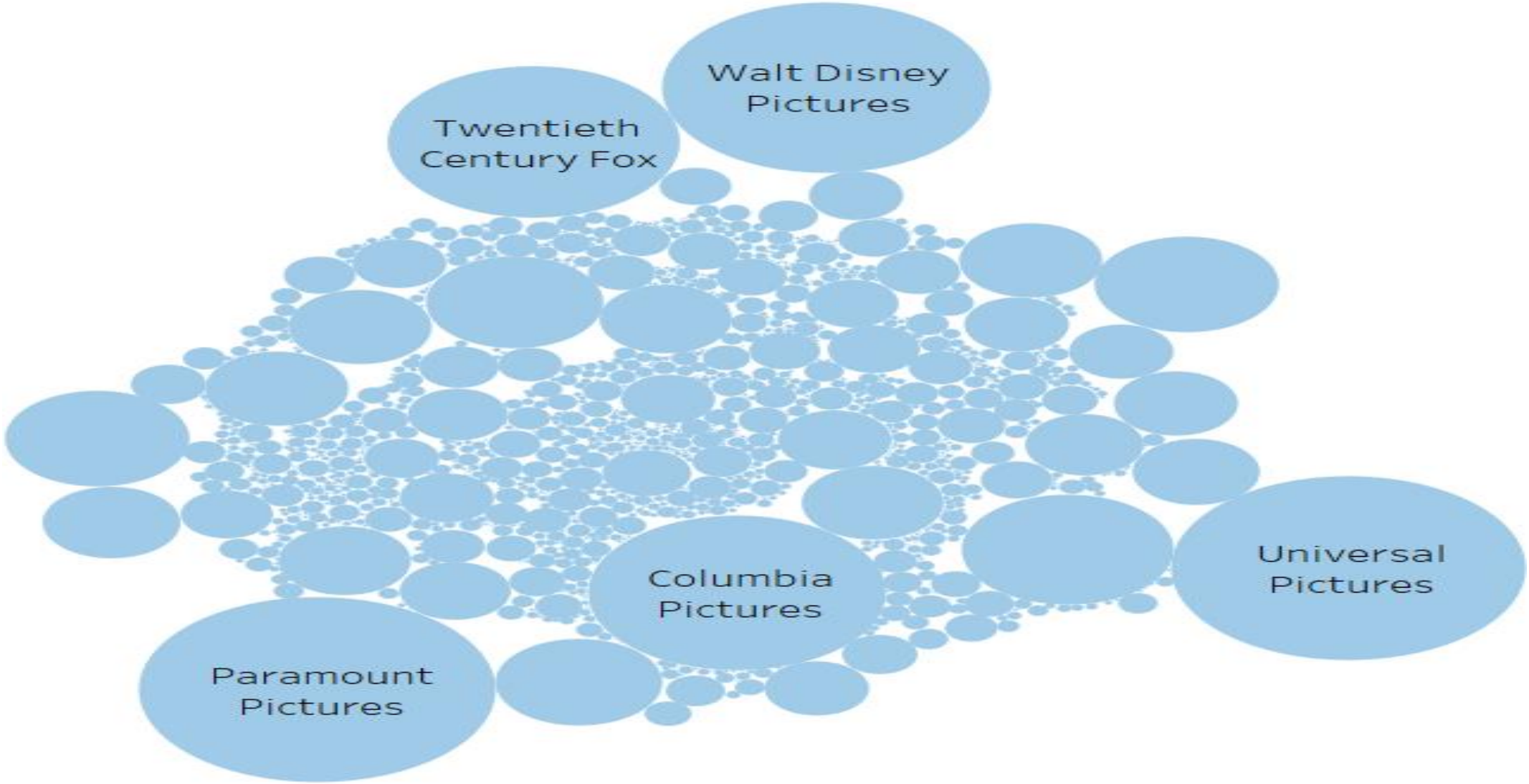
Data Cleaning

- Remove NA variables
- Remove Movies with 0 value of Revenue and Budget
- Drop Unnecessary Columns
  (genres, imdb_id, original_language, original_title, overview, status, tagline, title, vote_count, spoken_languages_number, actor1_gender, actor2_name, actor2_gender, actor3_name, actor3_gender, actor4_name, actor4_gender, actor5_name, actor5_gender, actor_number, director_gender, director_number, producer_name, producer_number, screeplay_name, editor_name)
- Limit period to 2008 to 2017 when analyzing revenue and release month

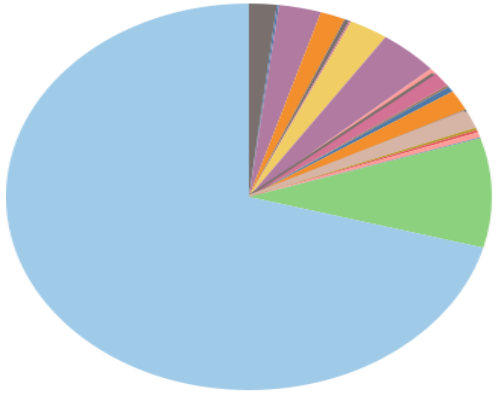
Data Visualization

- With Tableau

- Production companies by Revenue
- Revenue by Production countries & Spoken Language
- Vote Average by Production countries & Spoken Language



Production Companies. Size shows sum of Revenue. The marks are labeled by Production Companies.



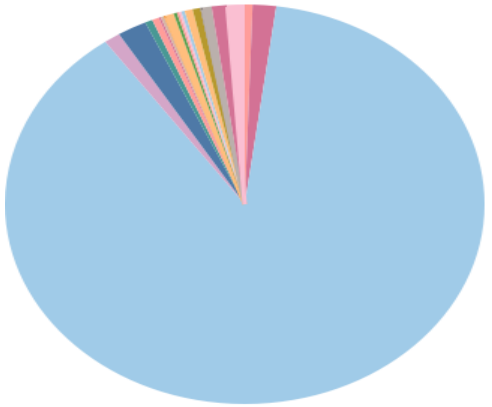
Production Countries (color) and sum of Revenue (size). The view is filtered on Production Countries, which keeps 22 of 235 members.

Revenue

497,393,173,695

Production Countries

- Australia
- Belgium
- Canada
- China
- Czech Republic
- Denmark
- France
- Germany
- Hong Kong
- Hungary
- India
- Ireland
- Italy
- Japan
- Mexico
- New Zealand
- Russia
- South Korea
- Spain
- United Arab Emirates
- United Kingdom
- United States of America



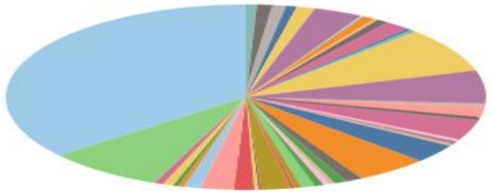
Spoken Languages (color) and sum of Revenue (size). The view is filtered on Spoken Languages, which keeps 21 of 76 members.

Revenue

497,317,448,583

Spoken Languages

- Český
- Deutsch
- English
- Español
- Français
- Italiano
- Latin
- Magyar
- Norsk
- Português
- Русский
- Română
- svenska
- ελληνικά
- עִבְרִית
- العربية
- हिन्दी
- 한국어
- 广州话 / 廣州話
- 日本語
- 普通话



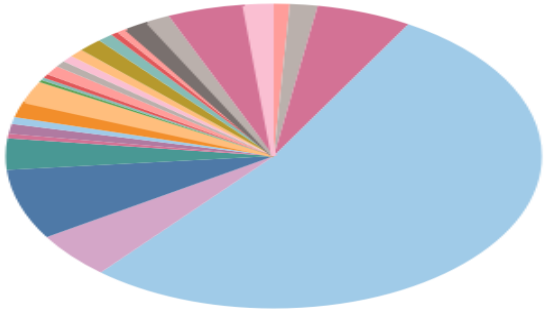
Production Countries (color) and sum of Vote Average (size). The view is filtered on Production Countries, which keeps 45 of 235 members.

Vote Average

539,755

Production Countries

- Argentina
- Australia
- Austria
- Belgium
- Brazil
- Canada
- Chile
- China
- Colombia
- Czech Republic
- Denmark
- Egypt
- Finland
- France
- Germany
- Greece
- Hong Kong
- Hungary
- India
- Indonesia
- Iran
- Ireland
- Israel
- Italy
- Japan
- Mexico
- Netherlands
- New Zealand
- Norway
- Philippines
- Poland
- Portugal
- Romania
- Russia
- Serbia
- South Africa
- South Korea
- Spain
- Sweden
- Switzerland
- Taiwan
- Thailand
- Turkey
- United Kingdom
- United States of America



Spoken Languages (color) and sum of Vote Average (size). The view is filtered on Spoken Languages, which keeps 27 of 76 members.

Vote Average

571,230

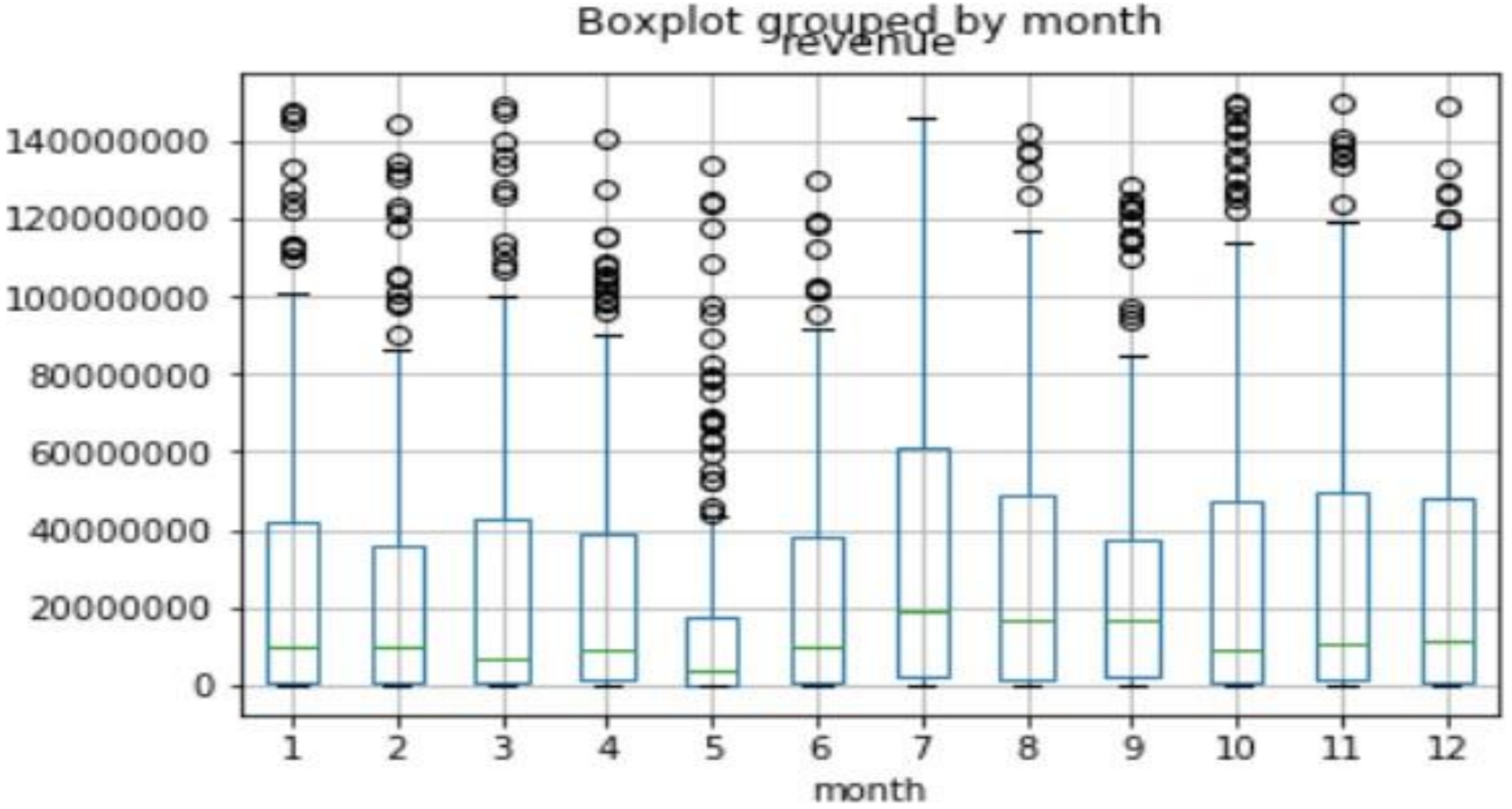
Spoken Languages

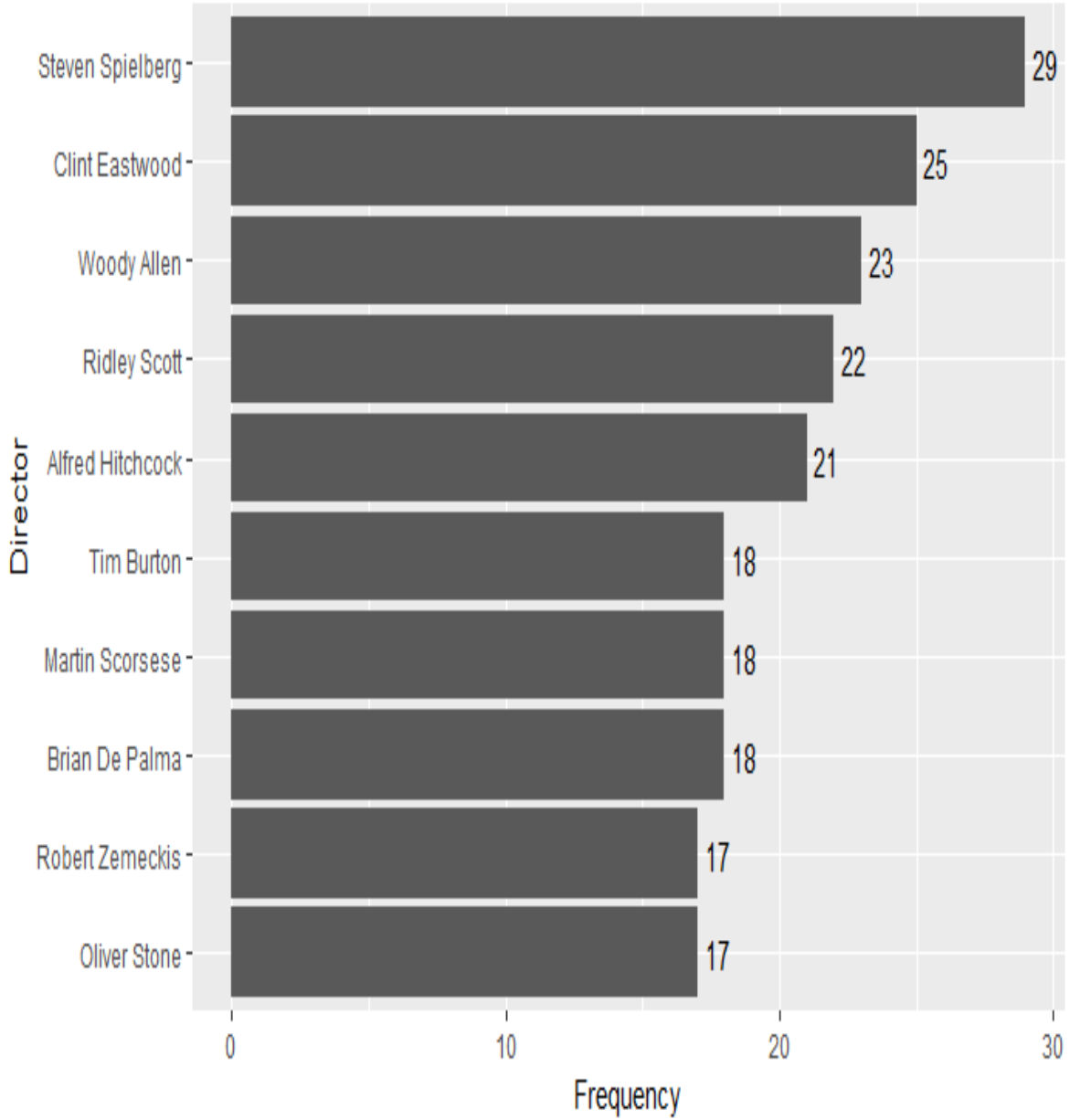
- Český
- Dansk
- Deutsch
- English
- Español
- Français
- Italiano
- Magyar
- Nederlands
- Polski
- Português
- Русский
- Română
- Srpski
- suomi
- svenska
- Türkçe
- ελληνικά
- العربية
- हिन्दी
- தமிழ்
- తెలుగు
- ภาษาไทย
- 한국어
- 广州话 / 廣州話
- 日本語
- 普通话

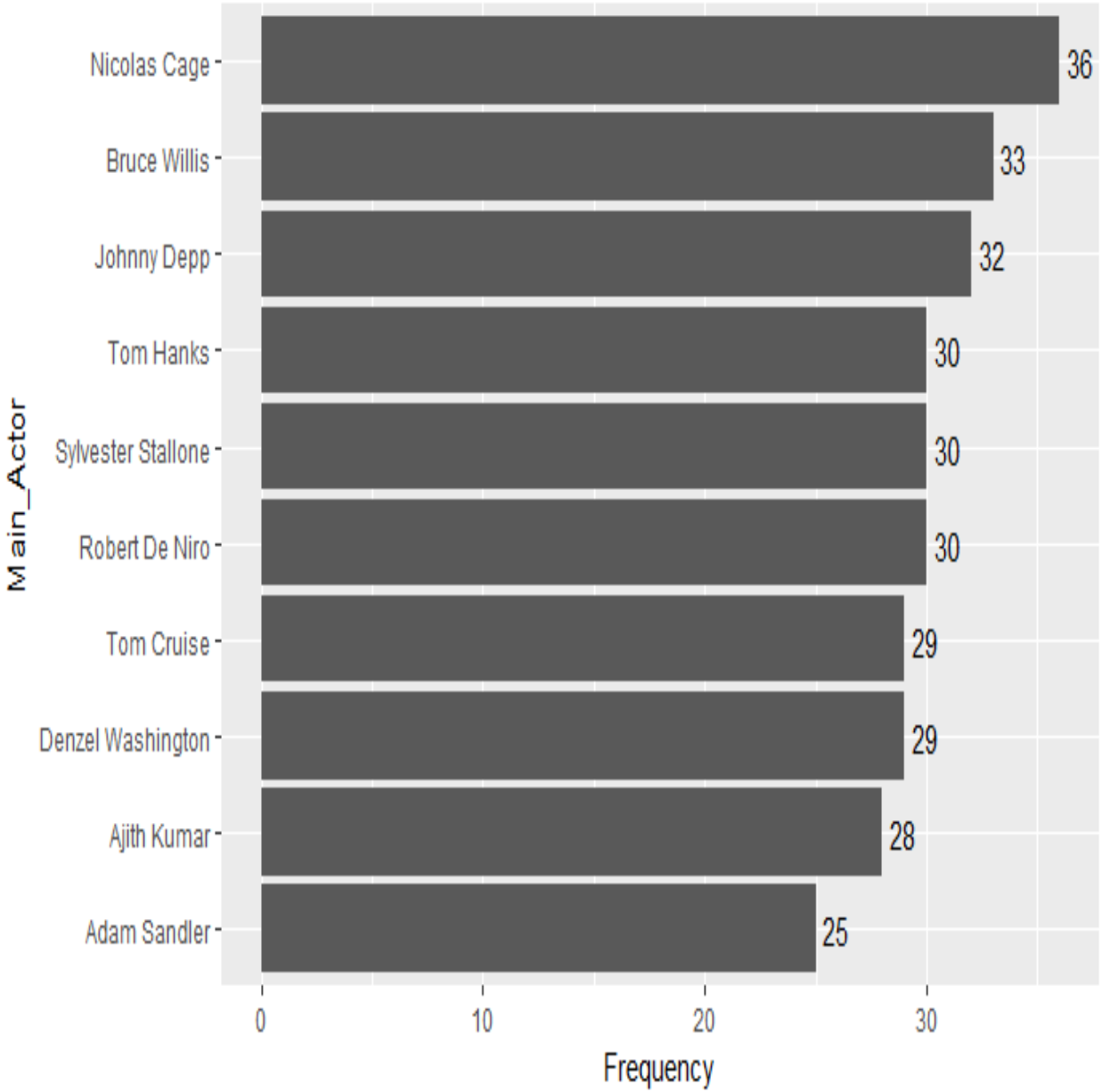
Exploratory Data Analysis

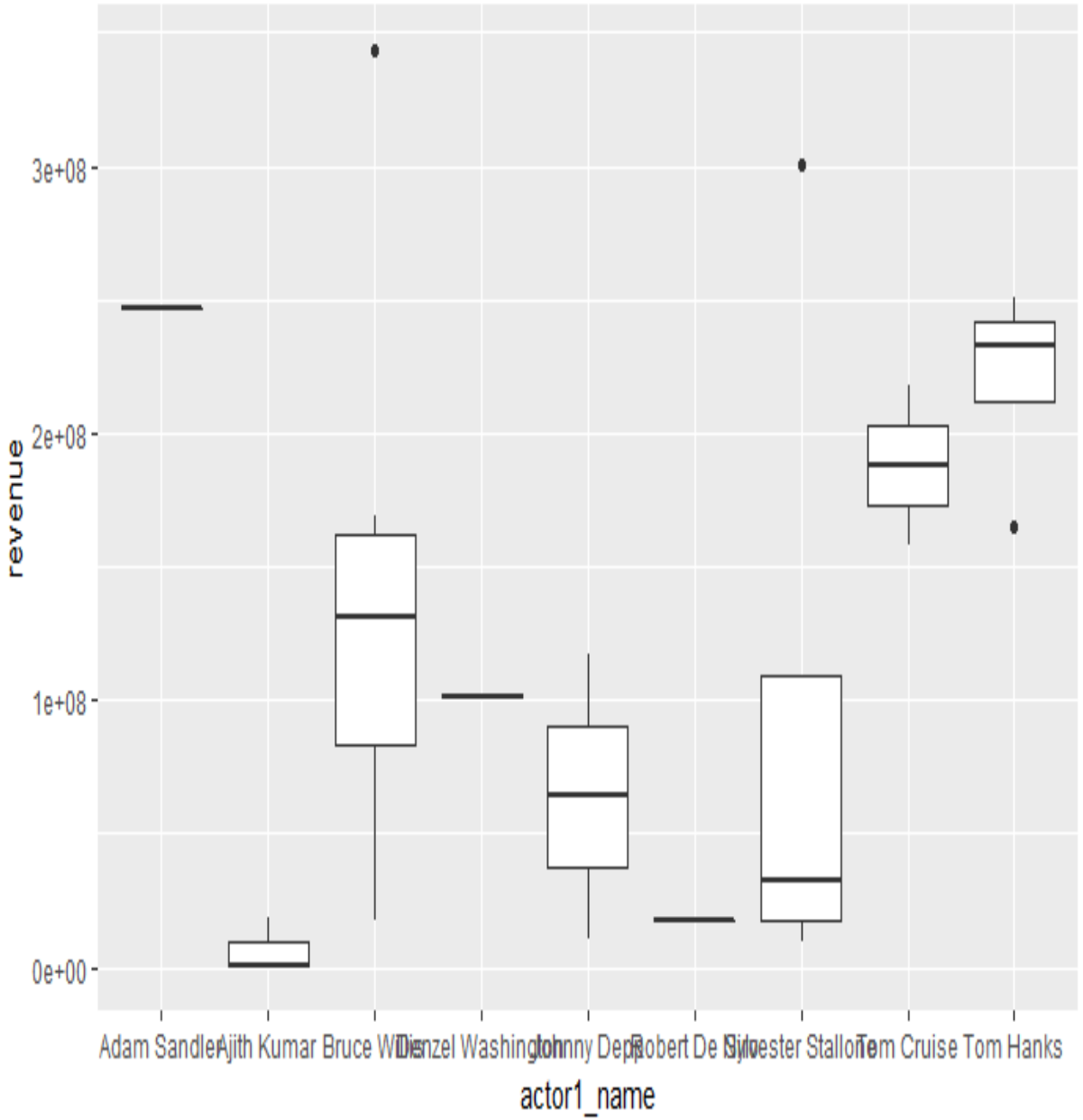
- With Python & R

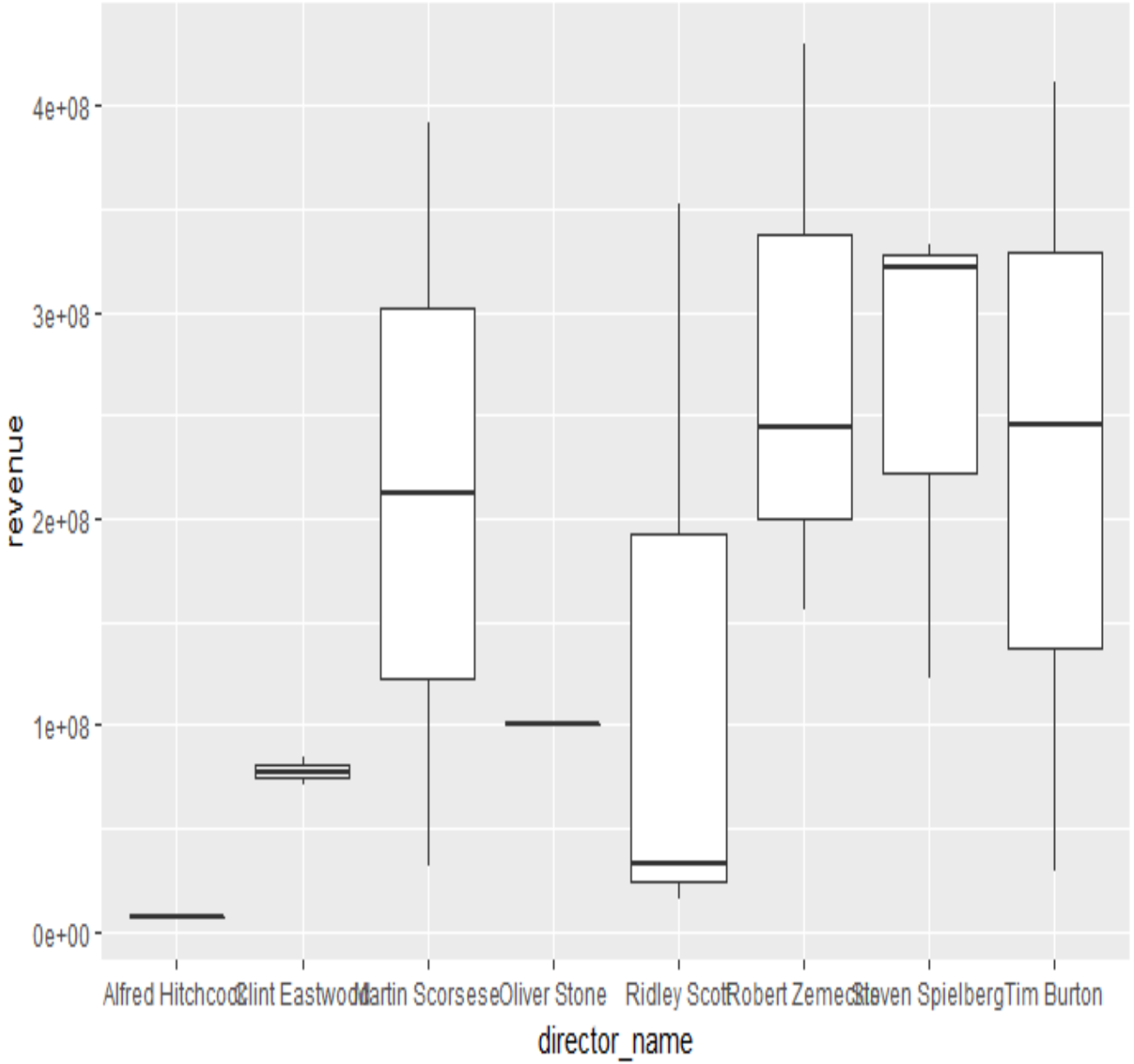
- Boxplot of Revenue by Release Month
- Bar chart of Director & Main Actor Frequency Table
- Boxplot of Revenue & Vote Average by Top Directors
- Boxplot of Revenue & Vote Average by Top Main Actors
- Scatterplot of Revenue by Budget

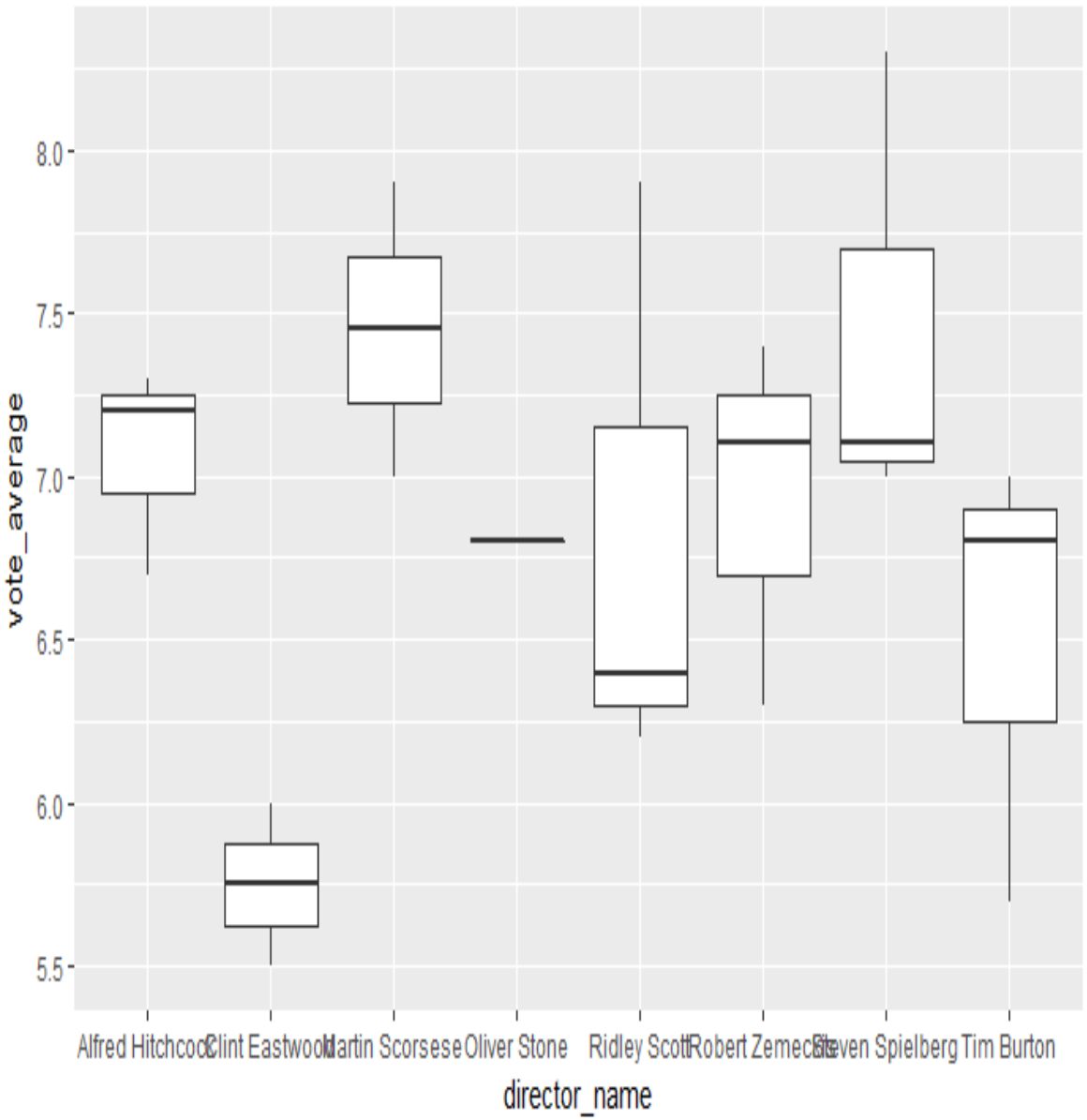

Boxplot grouped by month
revenue
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
1
2
3
4
5
6
7
8
9
10
11
12
month

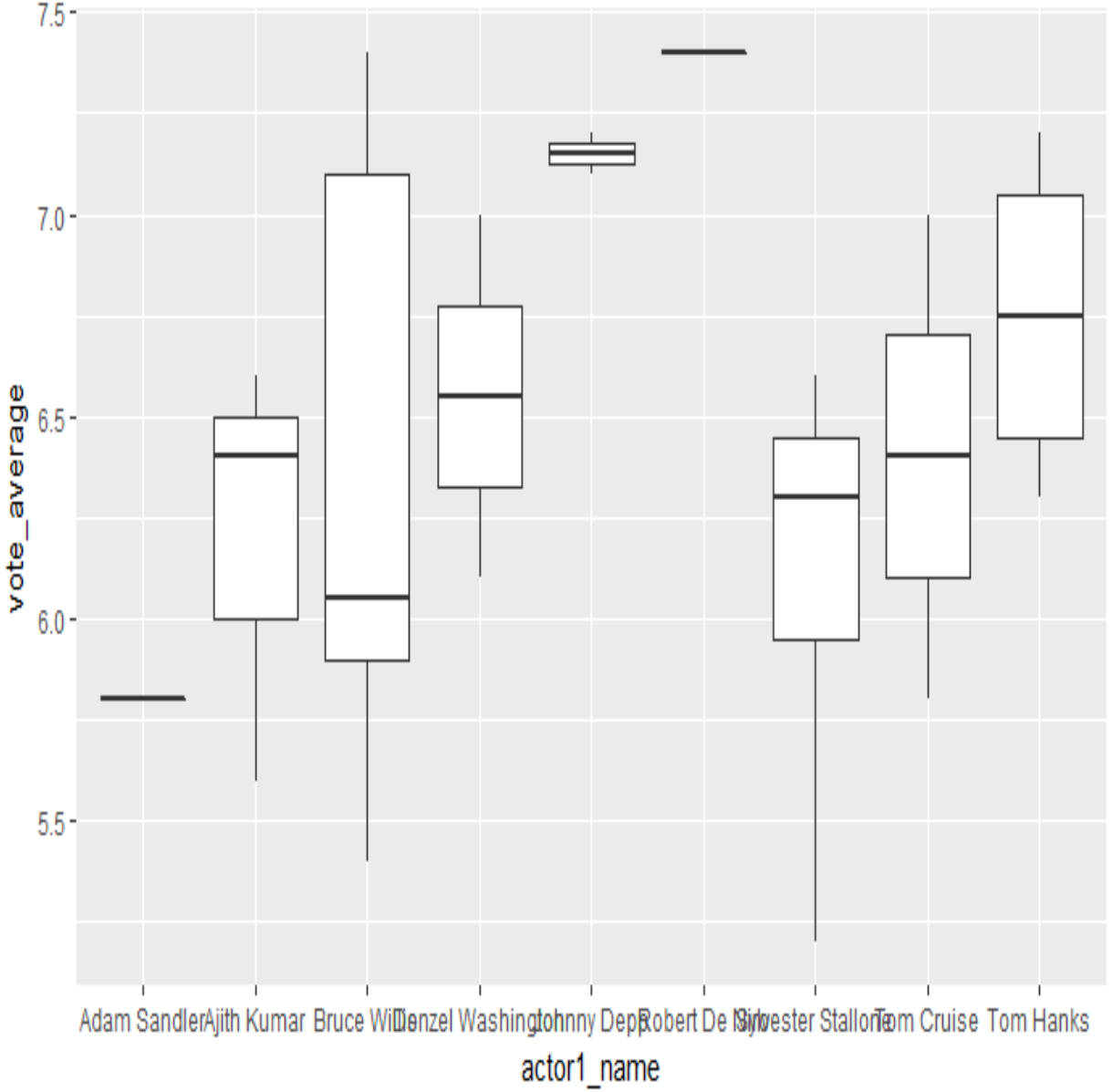

Director
Steven Spielberg
Clint Eastwood
Woody Allen
Ridley Scott
Alfred Hitchcock
Tim Burton
Martin Scorsese
Brian De Palma
Robert Zemeckis
Oliver Stone
29
25
23
22
21
18
18
18
17
17
0
10
20
30
Frequency

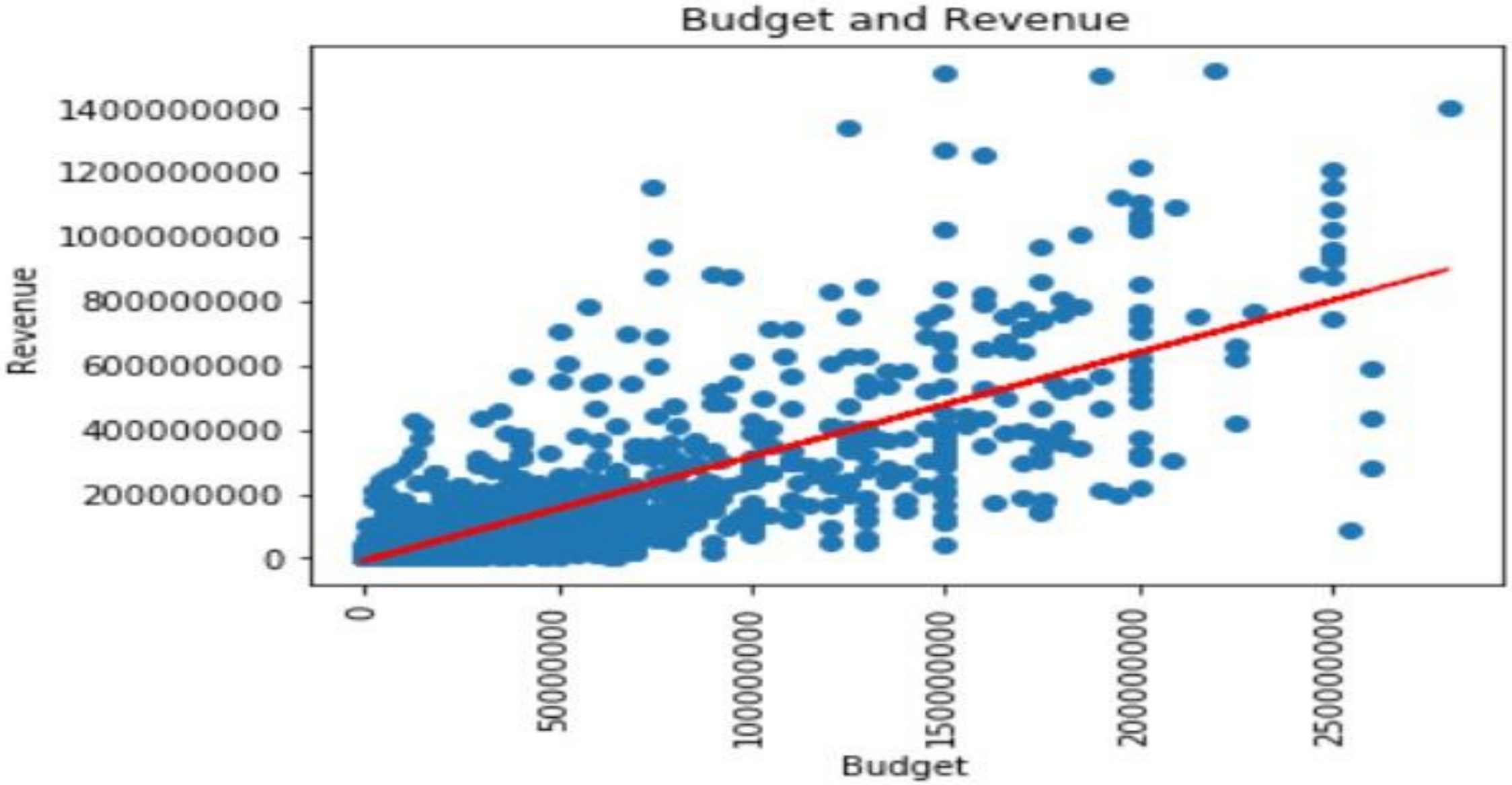

Main_Actor
Nicolas Cage
Bruce Willis
Johnny Depp
Tom Hanks
Sylvester Stallone
Robert De Niro
Tom Cruise
Denzel Washington
Ajith Kumar
Adam Sandler
36
33
32
30
30
30
29
29
28
25
0
10
20
30
Frequency


3e+08
2e+08
1e+08
0e+00
revenue
Adam Sandler
Ajith Kumar
Bruce Willis
Denzel Washington
Johnny Depp
Robert De Niro
Sylvester Stallone
Tom Cruise
Tom Hanks
actor1_name

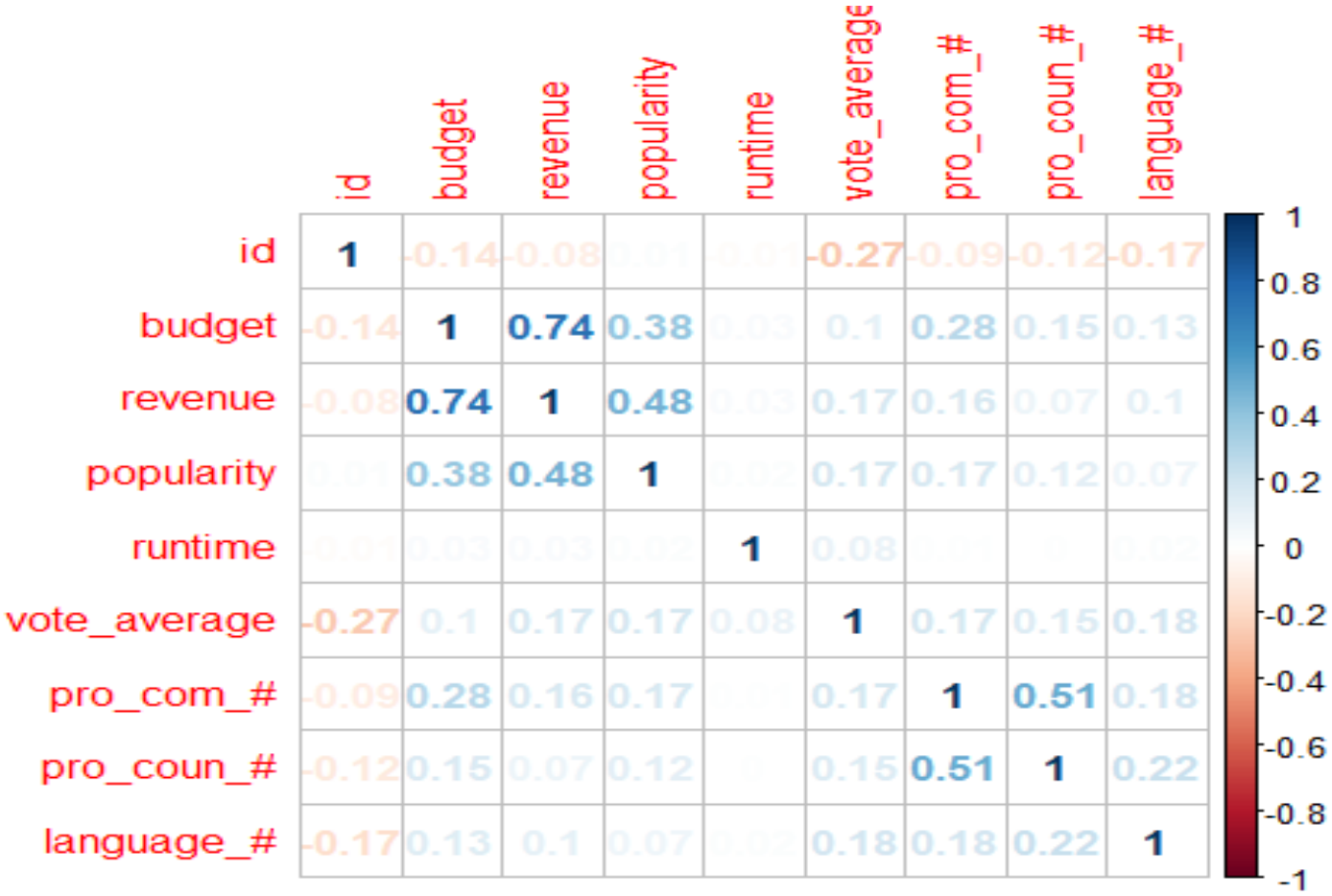

4e+08
3e+08
2e+08
1e+08
0e+00
revenue
Alfred Hitchcock
Clint Eastwood
Martin Scorsese
Oliver Stone
Ridley Scott
Robert Zemeckis
Steven Spielberg
Tim Burton
director_name


vote_average
8.0
7.5
7.0
6.5
6.0
5.5
Alfred Hitchcock
Clint Eastwood
Martin Scorsese
Oliver Stone
Ridley Scott
Robert Zemeckis
Steven Spielberg
Tim Burton
director_name


vote_average
7.5
7.0
6.5
6.0
5.5
Adam Sandler
Ajith Kumar
Bruce Willis
Denzel Washington
Johnny Depp
Robert De Niro
Sylvester Stallone
Tom Cruise
Tom Hanks
actor1_name


Budget and Revenue
Revenue
1400000000
1200000000
1000000000
800000000
600000000
400000000
200000000
0
0
50000000
100000000
150000000
200000000
250000000
Budget

Correlation Analysis (R)



Revenue Prediction Line (R)

```
Call:
lm(formula = revenue ~ ., data = movie_an)

Residuals:
       Min         1Q     Median         3Q        Max
-846368845  -32937301   -6122007   18853286 2043698535

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.922e+07  5.817e+06  -6.743 1.70e-11 ***
id             2.218e+01  1.172e+01   1.892   0.0586 .
budget         2.716e+00  3.733e-02  72.762  < 2e-16 ***
popularity     8.398e+06  3.406e+05  24.655  < 2e-16 ***
runtime       -4.878e+02  8.130e+03  -0.060   0.9522
vote_average   7.372e+06  8.476e+05   8.697  < 2e-16 ***
`pro_com_#`   -3.604e+06  7.131e+05  -5.054 4.45e-07 ***
`pro_coun_#`  -8.919e+06  1.925e+06  -4.633 3.68e-06 ***
`language_#`   1.465e+06  1.472e+06   0.995   0.3197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100200000 on 5955 degrees of freedom
Multiple R-squared:  0.603,     Adjusted R-squared:  0.6024
F-statistic:  1130 on 8 and 5955 DF,  p-value: < 2.2e-16
```

Vote_avg Prediction Line (R)

```
Call:
lm(formula = vote_average ~ ., data = movie_an)

Residuals:
     Min       1Q   Median       3Q      Max
-7.9433 -0.5376   0.1356   0.7855   6.0239

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.544e+00  5.204e-02 106.530  < 2e-16 ***
id            -3.411e-06  1.726e-07 -19.763  < 2e-16 ***
budget        -6.193e-09  7.753e-10  -7.988 1.63e-15 ***
revenue        1.701e-09  1.956e-10   8.697  < 2e-16 ***
popularity     4.820e-02  5.397e-03   8.931  < 2e-16 ***
runtime        7.470e-04  1.231e-04   6.067 1.39e-09 ***
`pro_com_#`    7.722e-02  1.081e-02   7.143 1.02e-12 ***
`pro_coun_#`   7.700e-02  2.928e-02   2.630  0.00857 **
`language_#`   1.890e-01  2.223e-02   8.503  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.522 on 5955 degrees of freedom
Multiple R-squared:  0.1466,     Adjusted R-squared:  0.1455
F-statistic: 127.9 on 8 and 5955 DF,  p-value: < 2.2e-16
```

Conclusion

- Movies tend to have the highest revenue when released on July and the least when released on May
- Revenue, Economic Success, and Budget is highly correlated
- Budget, Popularity, Vote average,
  Number of production country, Number of production company
  affect Revenue which is success(Economic)
- Budget, Popularity, Revenue, Runtime,
  Number of production company, Number of production country,
  Number of language
  affect Vote average, which is success(Cinematic quality)

Further Analysis

- Relationship of Movie Genre and Revenue
- Effect of Budget toward Revenue by Movie Genre