Nama : Demitries Baskhara Rivaldo Tolla
NIM : 123180137
Kelas : B

Latar Belakang :
Web Scraping yang saya buat berdasarkan data yang saya dapatkan melalui web IMDB. Web IMDB adalah sebuah website yang berisi database film yang sudah lama ada. Data yang saya ambil dari website tersebut adalah **Judul , Genre , Pendapatan , Rating , dan Runtime**. Dari data tersebut saya akan melakukan sebuah analisis yang akan menjadi sebuah kesimpulan yang bisa digunakan untuk melakukan perbaikan kedepannya terutama untuk industri perfilman.

Potongan Source Code :

```
library(httr);
library(xml2);
library(rvest);
```

Pada tugas ini saya memakai tiga buah library yaitu httr , xml2 , dan rvest. Httr adalah sebuah tool yang berhubungan dengan URLs and HTTP. Xml2 adalah sebuah tool untuk melakukan pengolahan XML. Rvest adalah tool untuk mempermudah proses Web Scraping.

```
all_data <- data.frame();
```

Pembuatan satu variabel untuk menyimpan semua data yang akan diambil.

```
url <- "https://www.imdb.com/search/title/?groups=top_1000&count=250";
url_web <- read_html(url);
```

Memasukkan url ke variabel kemudian melakukan fungsi read_html untuk membaca seluruh halaman.

```
runtime <- html_nodes(url_web , ".runtime");
data_runtime <- html_text(runtime);
data_runtime <- gsub(" min" , "" , data_runtime);
data_runtime <- as.numeric(data_runtime);
```

Memasukkan data runtime dengan fungsi html_nodes , fungsi ini secara otomatis mencari data yang sesuai dengan input. Selanjutnya dilakukan pembersihan data , data runtime dibersihkan dengan menghilangkan “ min” dan diubah menjadi numeric.

```
judul <- html_nodes(url_web , ".lister-item-header");
data_judul <- html_text(judul);
data_judul <- gsub("\n" , "" , data_judul);
data_judul <- gsub(" " , "" , data_judul);
data_judul <- sub("\\(.*" , "" , data_judul);
data_judul <- sub(".*\\." , "" , data_judul);
```

Memasukkan data judul dengan fungsi html_nodes , fungsi ini secara otomatis mencari data yang sesuai dengan input. Selanjutnya dilakukan pembersihan data , data judul dibersihkan dengan menghilangkan spasi , enter , semua huruf setelah “(“ dan semua huruf sebelum “.”. Sayangnya pada pembersihan ini menimbulkan beberapa kerusakan data , data yang rusak adalah data yang memiliki “.” di antara judul dan mengakibatkan terhapusnya judul.

```
genre <- html_nodes(url_web , ".genre");
data_genre <- html_text(genre);
data_genre <- gsub("\n" , "" , data_genre);
data_genre <- gsub(" " , "" , data_genre);
data_genre <- gsub(",.*" , "" , data_genre);
data_genre <- as.factor(data_genre);
```

Memasukkan data genre dengan fungsi html_nodes , fungsi ini secara otomatis mencari data yang sesuai dengan input. Selanjutnya dilakukan pembersihan data , data genre dibersihkan dengan menghilangkan spasi , enter , semua huruf setelah “,“ dan diubah menjadi factor.

```
rating <- html_nodes(url_web , ".ratings-imdb-rating strong");
data_rating <- html_text(rating);
data_rating <- as.numeric(data_rating);
```

Memasukkan data rating dengan fungsi html_nodes , fungsi ini secara otomatis mencari data yang sesuai dengan input. Selanjutnya dilakukan pembersihan data , data rating dibersihkan dengan diubah menjadi numeric.

```
gross <- html_nodes(url_web , ".sort-num_votes-visible");
data_gross <- html_text(gross);
data_gross <- gsub("\n" , "" , data_gross);
data_gross <- gsub(" " , "" , data_gross);
data_gross <- gsub(".*Gross:" , "" , data_gross);
data_gross <- gsub("Votes:.*" , "" , data_gross);
data_gross <- sub("." , "" , data_gross);
data_gross <- sub("M" , "" , data_gross);
data_gross <- as.numeric(data_gross);
```

Memasukkan data gross dengan fungsi html_nodes , fungsi ini secara otomatis mencari data yang sesuai dengan input. Selanjutnya dilakukan pembersihan data , data gross dibersihkan dengan menghilangkan enter , spasi , semua huruf sebelum “Gross:” , kalimat “Gross:” , semua huruf setelah “Votes:” , kalimat “Votes:” , huruf “.” , huruf “M” , dan diubah menjadi numeric.

```
data <- data.frame("Judul" = data_judul,"Rating" = data_rating,"Runtime" =
data_runtime,"Genre" = data_genre,"Gross" = data_gross);
```

Setelah semua data didapatkan dan sudah dimasukkan ke dalam sebuah variabel , maka selanjutnya adalah memasukkan data data tersebut menjadi satu data frame.

```
all_data <- rbind(all_data , data);
```

Setelah satu url selesai dijalankan maka data dari url tersebut akan dimasukkan kedalam data main. Selanjutnya proses diulangi sampai semua url diambil datanya.

Preview Data :

Data Preview:

Judul (character)	Rating (double)	Runtime (double)	Genre (character)	Gross (double)
HomeAlone	7.6	103	Comedy	285.76
Tenet	7.7	150	Action	53.80
Planes,Trains&Automobiles	7.6	93	Comedy	49.53
TheTrialoftheChicago7	7.9	129	Drama	NA
Spider-Man:IntotheSpider-Verse	8.4	117	Animation	190.24
HarryPotterandtheSorcerer'sStone	7.6	152	Adventure	317.58
KnivesOut	7.9	130	Comedy	165.36
TheGodfather	9.2	175	Crime	134.97
TheGentlemen	7.8	113	Action	NA
1917	8.3	119	Drama	159.23
inHollywood	7.6	161	Comedy	142.50
LoveActually	7.6	135	Comedy	59.70
TheWizardofOz	8.0	102	Adventure	2.08

Link Hasil :
https://drive.google.com/drive/folders/1wNTIRR6mPzRIN-EKNM332EsVnaojzcjd?usp=sharing

Analisis :

	Judul	Rating	Runtime	Genre	Gross
18	TheShawshankRedemption	9.3	142	Drama	28.34
8	TheGodfather	9.2	175	Crime	134.97
30	TheDarkKnight	9.0	152	Action	534.86
53	TheGodfather:PartII	9.0	202	Crime	57.30
151	12AngryMen	9.0	96	Crime	4.36
36	PulpFiction	8.9	154	Crime	107.93
68	TheLordoftheRings:TheReturnoftheKing	8.9	201	Action	377.85
80	Schindler'sList	8.9	195	Biography	96.90
21	SooraraiPottru	8.8	153	Drama	NA
24	TheLordoftheRings:TheFellowshipoftheRing	8.8	178	Action	315.54
35	Inception	8.8	148	Action	292.58
49	FightClub	8.8	139	Drama	37.03
56	ForrestGump	8.8	142	Drama	330.25
112	Ilbuono,ilbrutto,ilcattivo	8.8	161	Western	6.10
25	OneFlewOvertheCuckoo'sNest	8.7	133	Drama	112.00
47	TheMatrix	8.7	136	Action	171.48
52	Goodfellas	8.7	146	Biography	46.84
103	TheLordoftheRings:TheTwoTowers	8.7	179	Action	342.55
171	StarWars:EpisodeV-TheEmpireStrikesBack	8.7	124	Action	290.48
15	Gisaengchung	8.6	132	Comedy	53.37

```
> head(all_data[order(all_data$Rating),],n = 20)
```

	Judul	Rating	Runtime	Genre	Gross
1	HomeAlone	7.6	103	Comedy	285.76
3	Planes,Trains&Automobiles	7.6	93	Comedy	49.53
6	HarryPotterandtheSorcerer'sStone	7.6	152	Adventure	317.58
11	inHollywood	7.6	161	Comedy	142.50
12	LoveActually	7.6	135	Comedy	59.70
31	DazedandConfused	7.6	102	Comedy	7.99
59	ThePeanutButterFalcon	7.6	97	Adventure	13.12
66	TheGodfather:PartIII	7.6	162	Crime	66.67
73	Watchmen	7.6	162	Action	107.51
94	HarryPotterandtheHalf-BloodPrince	7.6	153	Action	301.96
127	MyCousinVinny	7.6	120	Comedy	52.93
130	2	7.6	136	Action	389.81
131	Apollo13	7.6	140	Adventure	173.84
132	Moneyball	7.6	133	Biography	75.61
134	AmericanPsycho	7.6	101	Comedy	15.07
146	Moana	7.6	107	Animation	248.76
148	BabyDriver	7.6	113	Action	107.83
170	AStarIsBorn	7.6	136	Drama	215.29
181	300	7.6	117	Action	210.61
187	Stardust	7.6	127	Adventure	38.63

```
> length(which(all_data$Rating == 7.6))
[1] 129
```

```
> mean(all_data$Rating)
[1] 7.9485
```

```
> length(which(all_data$Rating >= 7.9))
[1] 569
> length(which(all_data$Rating < 7.9))
[1] 431
```

Dari 1000 film yang didapatkan , film “The Shawshank Redemption” menjadi film dengan rating tertinggi yaitu 9.3.Sedangkan untuk rating terendah adalah “Home Alone” dengan rating 7.6. Tapi “Home Alone” bukan satu satunya film yang memiliki rating 7.6 , terdapat 129 film dengan rating 7.6. Rata rata Rating yang didapatkan dari 1000 film adalah 7.9. Film yang berada di atas 7.9 ada 569 film dan yang dibawah ada 431 film.

```
> head(all_data[order(-all_data$Gross),],n = 20)
                                Judul Rating Runtime     Genre  Gross
120   StarWars:EpisodeVII-TheForceAwakens    7.9     138    Action 936.66
16                      Avengers:Endgame    8.4     181    Action 858.37
115                               Avatar    7.8     162    Action 760.51
37                  Avengers:InfinityWar    8.4     149    Action 678.82
28                                Titanic    7.8     194     Drama 659.33
126                           TheAvengers    8.0     143    Action 623.28
294                          Incredibles2    7.6     118 Animation 608.58
30                         TheDarkKnight    9.0     152    Action 534.86
62                               RogueOne    7.8     133    Action 532.18
83                   TheDarkKnightRises    8.4     164    Action 448.14
259                theExtra-Terrestrial    7.8     115    Family 435.11
186                           ToyStory4    7.8     100 Animation 434.04
106                         TheLionKing    8.5      88 Animation 422.78
388                           ToyStory3    8.3     103 Animation 415.00
226               CaptainAmerica:CivilWar    7.8     147    Action 408.08
114                          JurassicPark    8.1     127    Action 402.45
130                                     2    7.6     136    Action 389.81
44  HarryPotterandtheDeathlyHallows:Part2    8.1     130 Adventure 381.01
304                          FindingNemo    8.1     100 Animation 380.84
68   TheLordoftheRings:TheReturnoftheKing    8.9     201    Action 377.85
> |
```

```
> mean(all_data$Gross , na.rm = TRUE)
[1] 68.43785
```

Dari 1000 film yang didapatkan , film “Star Wars : Episode VII - The Force Awakens” menjadi film dengan pendapatan tertinggi yaitu $ 936.66 M. Rata rata pendapatan yang didapatkan dari 1000 film adalah $ 68.43 M.

```
> head(all_data[order(-all_data$Runtime),],n = 20)
                                Judul Rating Runtime     Genre  Gross
577                     GangsofWasseypur    8.2     321    Action     NA
744                              Hamlet    7.7     242     Drama   4.41
287                      GonewiththeWind    8.1     238     Drama 198.68
236                 OnceUponaTimeinAmerica    8.4     229     Crime   5.32
410                      LawrenceofArabia    8.3     228 Adventure  44.82
831           Lagaan:OnceUponaTimeinIndia    8.1     224 Adventure   0.07
527                    TheTenCommandments    7.9     220 Adventure  93.74
511                               Ben-Hur    8.1     212 Adventure  74.70
928                  Swades:We,thePeople    8.2     210     Drama   1.22
42                            TheIrishman    7.9     209 Biography   7.00
347                    Shichininnosamurai    8.6     207    Action   0.27
756                          AndreiRublev    8.1     205 Biography   0.10
974                                Sholay    8.2     204    Action     NA
53                   TheGodfather:PartII    9.0     202     Crime  57.30
597                               MalcolmX    7.7     202 Biography  48.17
68  TheLordoftheRings:TheReturnoftheKing    8.9     201    Action 377.85
693                                Giant    7.6     201     Drama     NA
462                             Spartacus    7.9     197 Adventure  30.00
542                          DoctorZhivago    8.0     197     Drama 111.72
516                              KisUykusu    8.1     196     Drama   0.17
```

```
> head(all_data[order(all_data$Runtime),],n = 20)
                              Judul Rating Runtime     Genre  Gross
968                                     8.2      45    Action   0.98
758                           Freaks    7.9      64     Drama     NA
864                       TheGeneral    8.1      67    Action   1.03
818                          TheKid     8.3      68    Comedy   5.45
571                         DuckSoup    7.8      69    Comedy     NA
719                       Frankenstein    7.8      70     Drama     NA
811                    TheInvisibleMan    7.7      71    Horror     NA
839                   TheSecretofKells    7.6      71 Animation   0.69
954                         TheCircus     8.1      72    Comedy     NA
834                BrideofFrankenstein    7.8      75     Drama   4.36
896                BronenosetsPotemkin    8.0      75     Drama   0.05
107    TheNightmareBeforeChristmas    8.0      76 Animation  75.08
819                           Caligari    8.1      76   Fantasy     NA
934          Batman:MaskofthePhantasm    7.8      76 Animation   5.62
524                    TheJungleBook     7.6      78 Animation 141.84
972                             Zelig    7.7      79    Comedy  11.80
471                     BeforeSunset    8.1      80     Drama   5.82
489                             Rope    8.0      80     Crime     NA
880      InvasionoftheBodySnatchers    7.7      80     Drama     NA
920 Vivresavie:Filmendouzetableaux    8.0      80     Drama     NA
```

```
> mean(all_data$Runtime)
[1] 122.805
```

Dari 1000 film yang didapatkan , film “Gangs of Wasseypur” menjadi film dengan runtime terlama yaitu 321 menit. Rata rata runtime yang didapatkan dari 1000 film adalah 122 menit

```
> length(which(all_data$Judul == ""))
[1] 12
```

Seperti yang dijelaskan pada Potongan Source Code , bahwa terdapat beberapa film yang judulnya hilang setelah proses. Film tersebut berjumlah 12.

```
+ group_by(Genre) %>%
+ summarise(Jumlah = length(Genre))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 14 x 2
   Genre     Jumlah
   <fct>      <int>
 1 Action       174
 2 Adventure     72
 3 Animation     82
 4 Biography     86
 5 Comedy       156
 6 Crime        106
 7 Drama        288
 8 Family         2
 9 Horror        11
10 Mystery       13
11 Western        4
12 Film-Noir      3
13 Fantasy        2
14 Thriller       1
```

Setelah data dibagi berdasarkan genre , didapatkan genre terbanyak adalah Drama dengan total film 288.

```
> head(filter(all_data[order(-all_data$Rating),] , Gross > mean(all_data$Gross , na.rm = TRUE)) , n = 20)
                                Judul Rating Runtime     Genre  Gross
1                         TheGodfather    9.2     175     Crime 134.97
2                        TheDarkKnight    9.0     152    Action 534.86
3                          PulpFiction    8.9     154     Crime 107.93
4     TheLordoftheRings:TheReturnoftheKing    8.9     201    Action 377.85
5                       Schindler'sList    8.9     195 Biography  96.90
6  TheLordoftheRings:TheFellowshipoftheRing    8.8     178    Action 315.54
7                           Inception    8.8     148    Action 292.58
8                         ForrestGump    8.8     142     Drama 330.25
9            OneFlewOvertheCuckoo'sNest    8.7     133     Drama 112.00
10                          TheMatrix    8.7     136    Action 171.48
11        TheLordoftheRings:TheTwoTowers    8.7     179    Action 342.55
12  StarWars:EpisodeV-TheEmpireStrikesBack    8.7     124    Action 290.48
13                       Interstellar    8.6     169 Adventure 188.02
14                           StarWars    8.6     121    Action 322.74
15                              Se7en    8.6     127     Crime 100.13
16                 TheSilenceoftheLambs    8.6     118     Crime 130.74
17                         TheGreenMile    8.6     189     Crime 136.80
18                SavingPrivateRyan    8.6     169     Drama 216.54
19                              Joker    8.5     122     Crime 335.45
20                          Gladiator    8.5     155    Action 187.71
```

Jika data diberikan ketentuan gross diatas rata rata dan rating tertinggi maka “The Godfather” adalah film yang memenuhi kriteria.

```
> head(filter(all_data[order(-all_data$Gross),] , Rating > mean(all_data$Rating)) , n = 20)
                                  Judul Rating Runtime     Genre  Gross
1                          Avengers:Endgame    8.4     181    Action 858.37
2                       Avengers:InfinityWar    8.4     149    Action 678.82
3                                TheAvengers    8.0     143    Action 623.28
4                             TheDarkKnight    9.0     152    Action 534.86
5                         TheDarkKnightRises    8.4     164    Action 448.14
6                                TheLionKing    8.5      88 Animation 422.78
7                                  ToyStory3    8.3     103 Animation 415.00
8                                JurassicPark    8.1     127    Action 402.45
9     HarryPotterandtheDeathlyHallows:Part2    8.1     130 Adventure 381.01
10                               FindingNemo    8.1     100 Animation 380.84
11     TheLordoftheRings:TheReturnoftheKing    8.9     201    Action 377.85
12                                 Deadpool    8.0     108    Action 363.07
13                                InsideOut    8.1      95 Animation 356.46
14          TheLordoftheRings:TheTwoTowers    8.7     179    Action 342.55
15                                 Zootopia    8.0     108 Animation 341.27
16                                    Joker    8.5     122     Crime 335.45
17                    GuardiansoftheGalaxy    8.0     121    Action 333.18
18                              ForrestGump    8.8     142     Drama 330.25
19                                 StarWars    8.6     121    Action 322.74
20 TheLordoftheRings:TheFellowshipoftheRing    8.8     178    Action 315.54
```

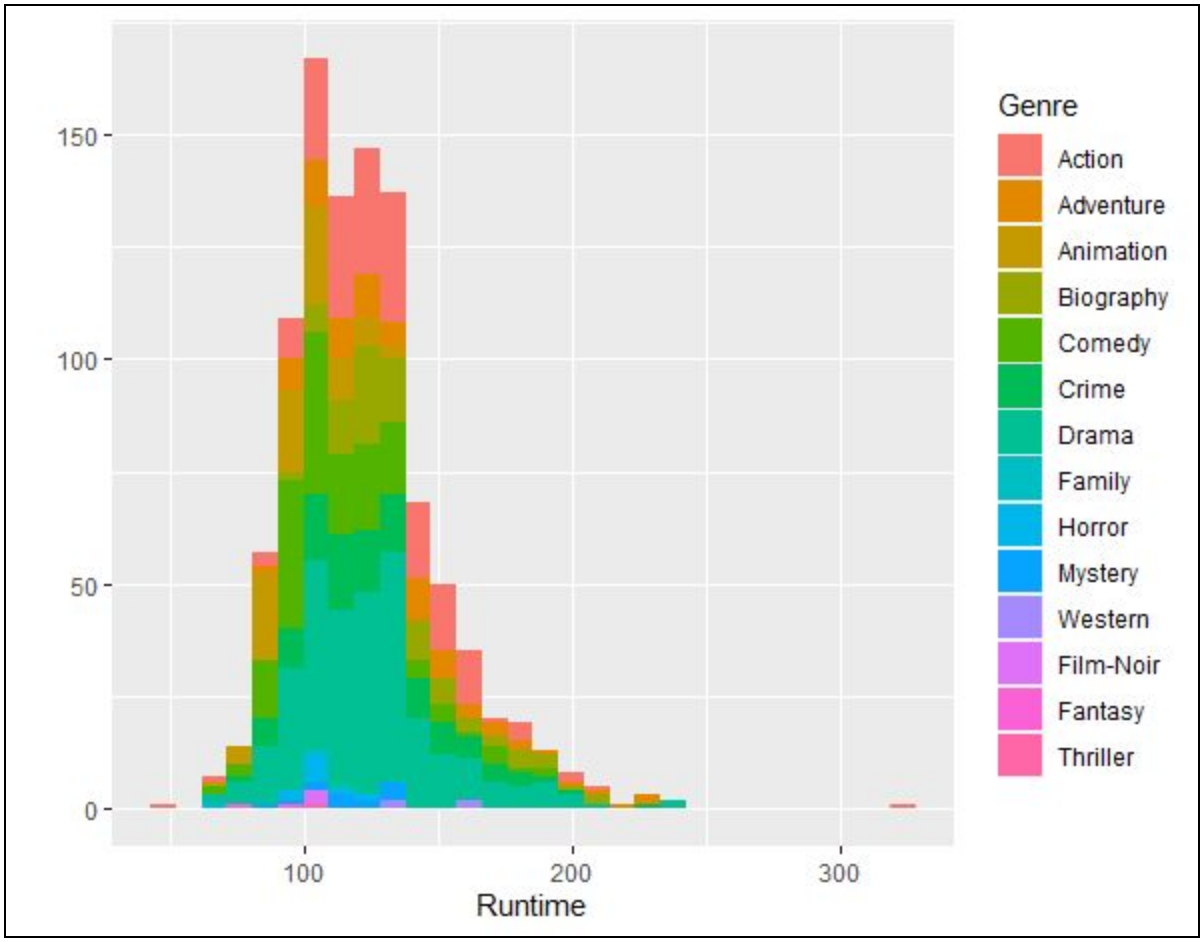
Jika data diberikan ketentuan rating diatas rata rata dan pendapatan tertinggi maka “Avenger : Endgame” adalah film yang memenuhi kriteria.

```
> head(filter(all_data[order(-all_data$Runtime),] , Gross > mean(all_data$Gross , na.rm = TRUE)) , n = 20)
                                  Judul Rating Runtime     Genre  Gross
1                       GonewiththeWind    8.1     238     Drama 198.68
2                     TheTenCommandments    7.9     220 Adventure  93.74
3                                Ben-Hur    8.1     212 Adventure  74.70
4     TheLordoftheRings:TheReturnoftheKing    8.9     201    Action 377.85
5                          DoctorZhivago    8.0     197     Drama 111.72
6                       Schindler'sList    8.9     195 Biography  96.90
7                               Titanic    7.8     194     Drama 659.33
8                           TheGreenMile    8.6     189     Crime 136.80
9                                    JFK    8.0     189     Drama  70.41
10                         Avengers:Endgame    8.4     181    Action 858.37
11                       DanceswithWolves    8.0     181 Adventure 184.21
12                       FiddlerontheRoof    8.0     181     Drama  80.50
13                     TheWolfofWallStreet    8.2     180 Biography 116.90
14          TheLordoftheRings:TheTwoTowers    8.7     179    Action 342.55
15 TheLordoftheRings:TheFellowshipoftheRing    8.8     178    Action 315.54
16                             Braveheart    8.3     178 Biography  75.60
17                          TheGodfather    9.2     175     Crime 134.97
18                       TheSoundofMusic    8.0     172 Biography 163.21
19                           MyFairLady    7.8     170     Drama  72.00
20                          Interstellar    8.6     169 Adventure 188.02
```

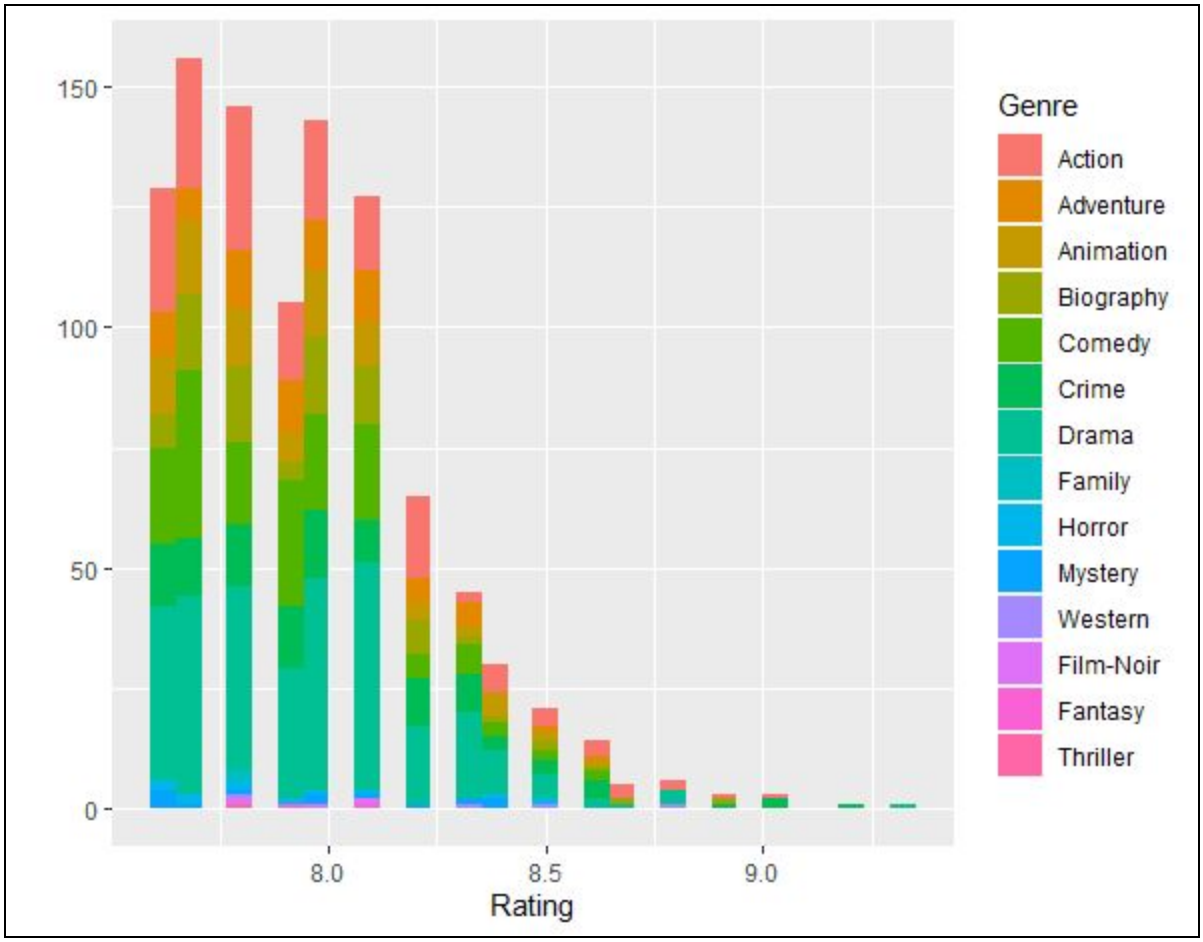
Jika data diberikan ketentuan pendapatan diatas rata rata dan runtime terlama maka “Gone With The Wind” adalah film yang memenuhi kriteria.

```
> head(filter(filter(filter(all_data , Runtime > mean(all_data$Runtime)) , Gross > mean(all_data$Gross , na.
rm = TRUE)) , Rating > mean(all_data$Rating)) , n = 20)
                                  Judul Rating Runtime     Genre  Gross
1                             TheGodfather    9.2     175     Crime 134.97
2                           Avengers:Endgame    8.4     181    Action 858.37
3                         TheWolfofWallStreet    8.2     180 Biography 116.90
4                              Interstellar    8.6     169 Adventure 188.02
5  TheLordoftheRings:TheFellowshipoftheRing    8.8     178    Action 315.54
6                 OneFlewOvertheCuckoo'sNest    8.7     133     Drama 112.00
7                                   DieHard    8.2     132    Action  83.01
8                               TheDarkKnight    9.0     152    Action 534.86
9                         InglouriousBasterds    8.3     153 Adventure 120.54
10                                  GoneGirl    8.1     149     Drama 167.77
11                                Inception    8.8     148    Action 292.58
12                              PulpFiction    8.9     154     Crime 107.93
13                    Avengers:InfinityWar    8.4     149    Action 678.82
14                              FordvFerrari    8.1     152    Action 117.62
15                                Gladiator    8.5     155    Action 187.71
16                            BladeRunner2049    8.0     164    Action  92.05
17    HarryPotterandtheDeathlyHallows:Part2    8.1     130 Adventure 381.01
18                                TheMatrix    8.7     136    Action 171.48
19                              CasinoRoyale    8.0     144    Action 167.45
20                                TheMartian    8.0     144 Adventure 228.43
```

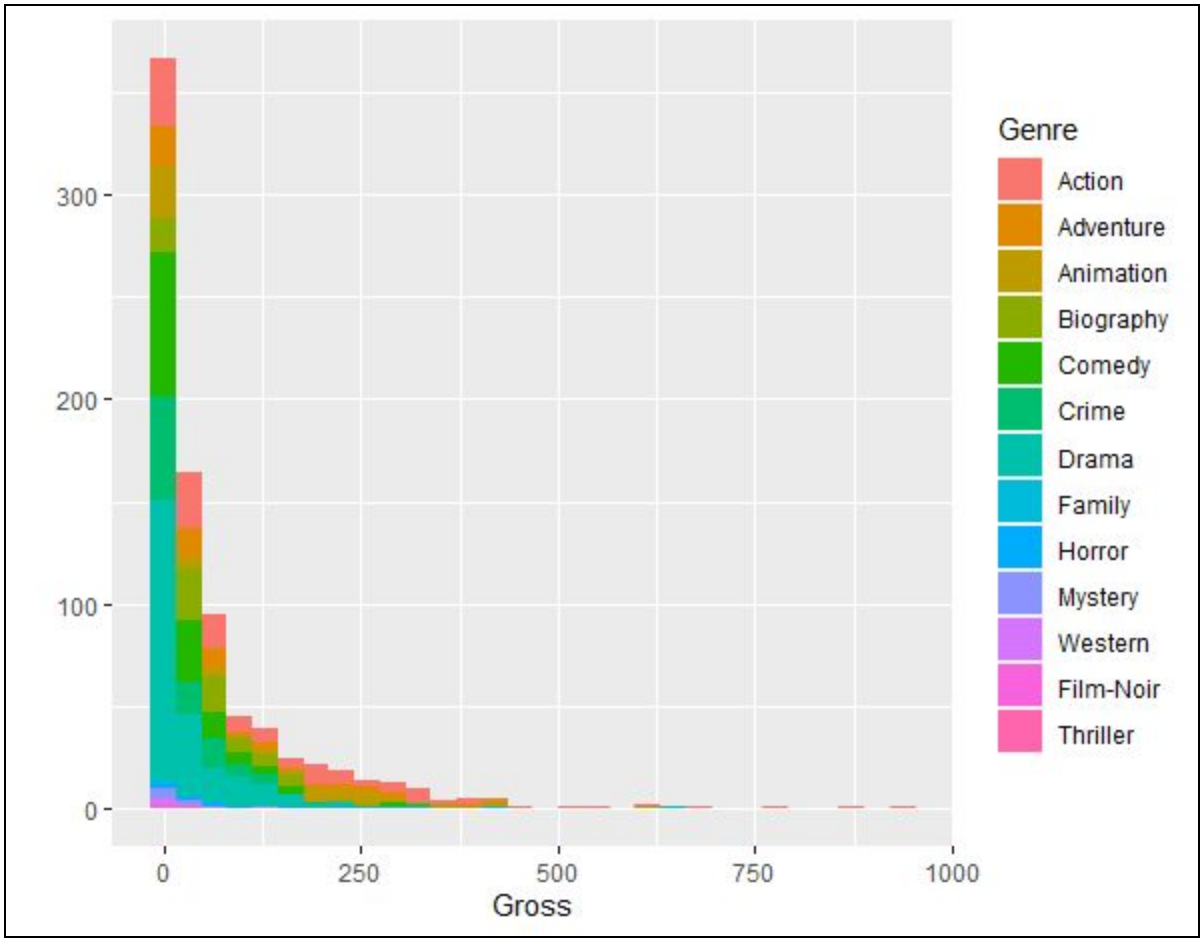
Sedangkan film yang nilai rating , pendapatan dan runtime diatas rata rata adalah “The Godfather”.



Tampilan plot untuk data Runtime.



Tampilan plot untuk data Rating.



Tampilan plot untuk data pendapatan.