ANALIZA PRZYCZYN POPULARNOSCI FILMÓW

Wstęp i pobieranie danych

Chciałam zbadać jakie cechy mają najbardziej popularne filmy. W tym celu zebrałam dane o "top 500" filmów z serwisu Filmweb.

Najpierw pobieram ze strony listę tytułów najpopularniejszych filmów. Określam je jako "najpopularniejsze", gdyż sam ranking opiera się na średniej ocenie (poprawionej o stałą zapewniającą określoną minimalną liczbę wymaganych głosów) nadanej przez użytkowników serwisu.

```
ListaFilmow<-html("http://www.filmweb.pl/rankings/film/world")
ListaFilmowWezly<-html_nodes(ListaFilmow,".place .s-20")
ListaFilmow<-html_text(ListaFilmowWezly)
```

Lista 100 pierwszych Filmów

Miejsce	Tytuł	Miejsce	Tytuł
1	Skazani na Shawshank	51	Gorączka
2	Nietykalni	52	Monty Python i Święty Graal
3	Ojciec chrzestny	53	Mroczny Rycerz
4	Zielona mila	54	Przekręt
5	Forrest Gump	55	Ojciec chrzestny III
6	Lot nad kukułczym gniazdem	56	Siedem dusz
7	Dwunastu gniewnych ludzi	57	Miasto Boga
8	Ojciec chrzestny II	58	Rain Man
9	Pulp Fiction	59	Gwiezdne wojny: Część IV - Nowa nadzieja
10	Władca Pierścieni: Powrót króla	60	Full Metal Jacket
11	Lista Schindlera	61	Adwokat diabła
12	Siedem	62	Psychoza
13	Władca Pierścieni: Dwie wieże	63	Bękarty wojny
14	Życie jest piękne	64	Mój przyjaciel Hachiko
15	Podziemny krąg	65	Przerwana lekcja muzyki
16	Milczenie owiec	66	Wściekłe psy
17	Django	67	Requiem dla snu
18	Chłopcy z ferajny	68	Wróg u bram
19	Piękny umysł	69	Wszystko za życie
20	Incepcja	70	Żywot Briana
21	Pianista	71	Taksówkarz
22	Człowiek z blizną	72	Seksmisja
23	Król Lew	73	Cinema Paradiso
24	Leon zawodowiec	74	Mechaniczna pomarańcza
25	Więzień nienawiści	75	Człowiek w ogniu
26	Gran Torino	76	Katedra
27	Bogowie	77	Uśpieni
28	Chłopiec w pasiastej piżamie	78	Sens życia wg Monty Pythona
29	Braveheart - Waleczne Serce	79	Szósty zmysł
30	Szeregowiec Ryan	80	Lśnienie
31	Wyspa tajemnic	81	Wyścig
32	Gladiator	82	Ściana

Miejsce	Tytuł	Miejsce	Tytuł
33	Prestiż	83	Życie Carlita
34	Gwiezdne wojny: Część V - Imperium kontratakuje	84	Siedmiu samurajów
35	Gwiezdne wojny: Część VI - Powrót Jedi	85	Pan od muzyki
36	Czas Apokalipsy	86	Filadelfia
37	Pamiętnik	87	Jak wytresować smoka
38	W pogoni za szczęściem	88	Blow
39	Władca Pierścieni: Drużyna Pierścienia	89	Przebudzenia
40	Zapach kobiety	90	Spirited Away: W krainie Bogów
41	Dawno temu w Ameryce	91	Żądło
42	Whiplash	92	Interstellar
43	Buntownik z wyboru	93	Ruchomy zamek Hauru
44	Pluton	94	Prawo zemsty
45	Dobry, zły i brzydki	95	Porachunki
46	Służące	96	Iluzjonista
47	Amadeusz	97	Jak rozpętałem drugą wojnę światową
48	Kasyno	98	Gra
49	Infiltracja	99	Święci z Bostonu
50	Łowca jeleni	100	Sami swoi

Chciałam zanalizować jakie cechy danego filmu czynią go popularnym, stąd potrzebowałam podstawowych danych o każdym z filmów.

```
LinkiDoFilmow<-html_attr(ListaFilmowWezly,"href")
n<-length(ListaFilmow)
PodstawoweInformacje<-vector("list",n)

for(i in 1:n){
   Tabele<-readHTMLTable(stri_paste("http://www.filmweb.pl",
      LinkiDoFilmow[i]),stringAsFactors=FALSE)
   PodstInf<-Tabele[[1]]
   PodstawoweInformacje[[i]]<-PodstInf
}
```

Podstawowa tabela z informacjami wygląda mniej więcej tak:

```
##    reżyseria:
## 1 scenariusz:
## 2
## 3    gatunek:
## 4  produkcja:
## 5   premiera:
## 6  boxoffice:
##                                                  Olivier NakacheEric Toledano
## 1                                                Olivier NakacheEric Toledano
## 2                                                              oceń twórców
## 3                                                BiograficznyDramatKomedia
## 4                                                                   Francja
## 5 13 kwietnia 2012  (Polska) 2012-04-13  23 września 2011  (świat) 2011-09-23
## 6                                                    $426 588 510 top #182
```

Uwaga: Reżyserzy wczytali się jako nazwy kolumn stąd ich wydobycie będzie się różniło od wydobycia reszty

Do analizy był jeszcze potrzebny wektor ocen poszczególnych filmów:

```
Oceny<-html("http://www.filmweb.pl/rankings/film/world")
OcenyWezly<-html_nodes(Oceny,".s-16")
Oceny<-html_text(OcenyWezly) #Wczytane w formacie tekstowym - niestety z przecinakmi jako separatory
Oceny<-stri_replace_all_fixed(Oceny[-1],",",".")
Oceny<-as.double(Oceny)
```

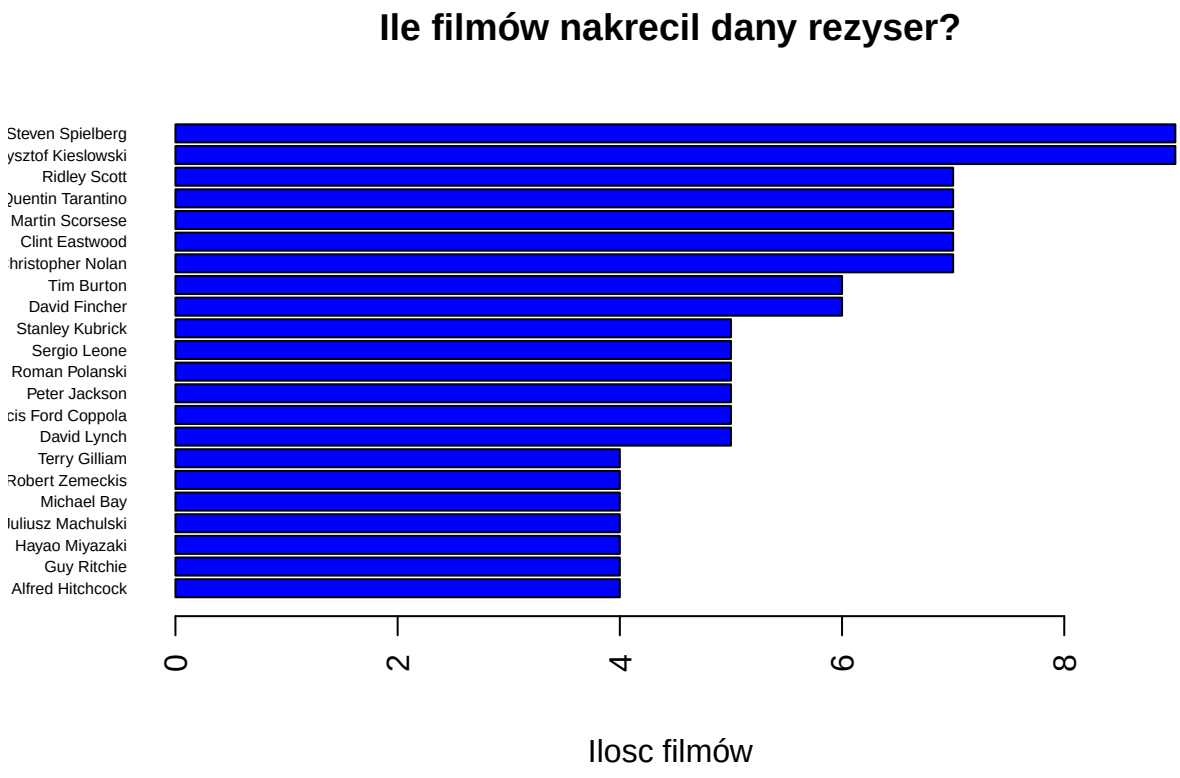
Analiza i wnioski

Najpierw wyciągnęłam odpowiednie dane z tabeli:

```
#Porzadkuje informacje z tabeli
Rezyserzy<-NULL
Scenarzysci<-NULL
Gatunki<-NULL
KrajeProdukcji<-NULL
   for(i in 1:500){
    Rezyser<-names(PodstawoweInformacje[[i]])[2]
    #Moze byc kilku rezyserow sklejonych ze soba na wzor: "Imie NazwiskoImie Nazwisko"
    Rezyser<-unlist(stri_split_regex(Rezyser,"(?<=\\p{Ll})(?=\\p{Lu})"))
    #Dolaczam ich do ogolnej listy
    Rezyserzy<-c(Rezyserzy,Rezyser)

    #Analogicznie kolejne informacje
      #Istnieja tabele,ktore nie maja niektorych wierszy => skomplikowane wyciaganie
    Scenarzysta<-PodstawoweInformacje[[i]][as.character(
       PodstawoweInformacje[[i]][,1])=="scenariusz:",2]
    Scenarzysta<-unlist(stri_split_regex(Scenarzysta,"(?<=\\p{Ll})(?=\\p{Lu})"))
    Scenarzysci<-c(Scenarzysci,Scenarzysta)

    Gatunek<-PodstawoweInformacje[[i]][as.character(
       PodstawoweInformacje[[i]][,1])=="gatunek:",2]
    Gatunek<-unlist(stri_split_regex(Gatunek,"(?<=\\p{Ll})(?=\\p{Lu})"))
    Gatunki<-c(Gatunki,Gatunek)

    KrajProdukcji<-as.character(
       PodstawoweInformacje[[i]][as.character(
          PodstawoweInformacje[[i]][,1])=="produkcja:",2])
    KrajProdukcji<-unlist(stri_split_regex(
       KrajProdukcji,"(?<=(\\p{Ll}|USA))(?=\\p{Lu})"))
    KrajeProdukcji<-c(KrajeProdukcji,KrajProdukcji)
   }
```

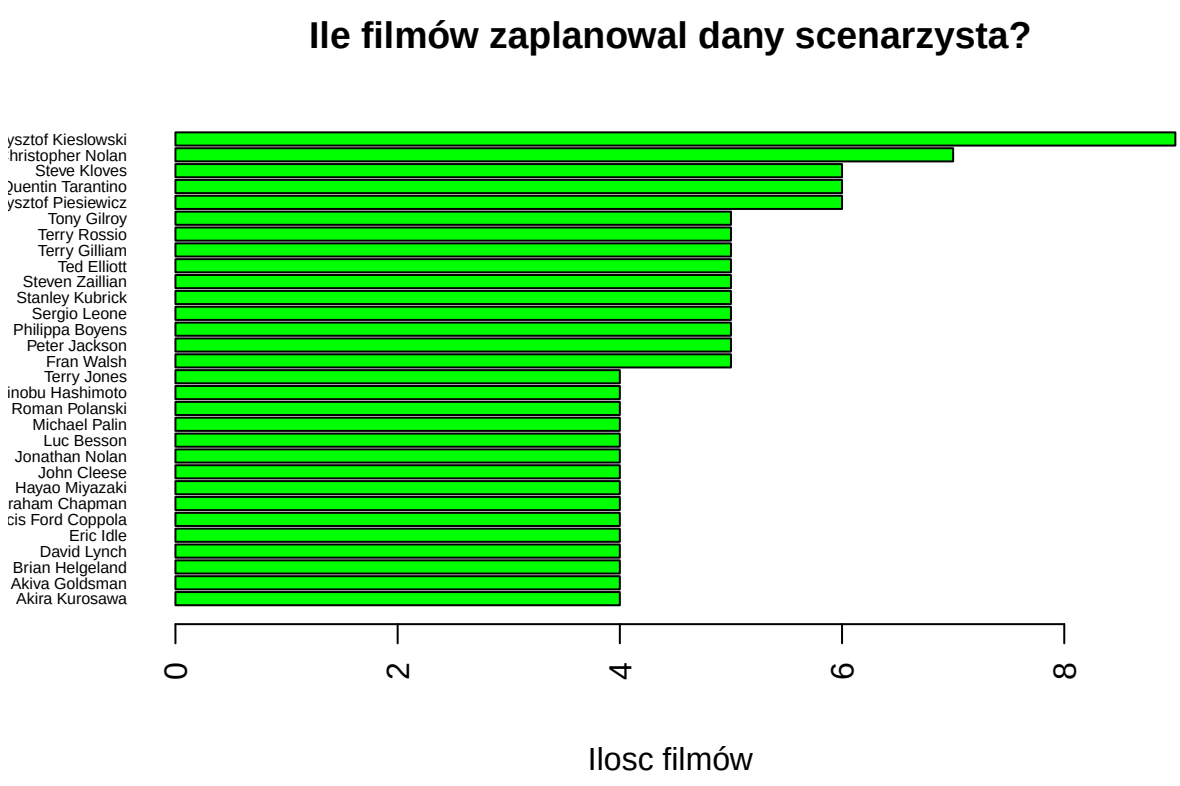
Uzyskane wyniki przedstawiam na wykresach. Dla ich czytelności brałam tylko część najwyżej punktowanych pozycji w danych kategoriach.

```
PodsumowanieRezyserzy<-sort(table(Rezyserzy)[table(Rezyserzy)>3])
barplot(PodsumowanieRezyserzy,
   names.arg=names(PodsumowanieRezyserzy),horiz=TRUE,las=2,col="blue",
   cex.names=0.5,main="Ile filmów nakręcił dany reżyser?",xlab="Ilość filmów")
```



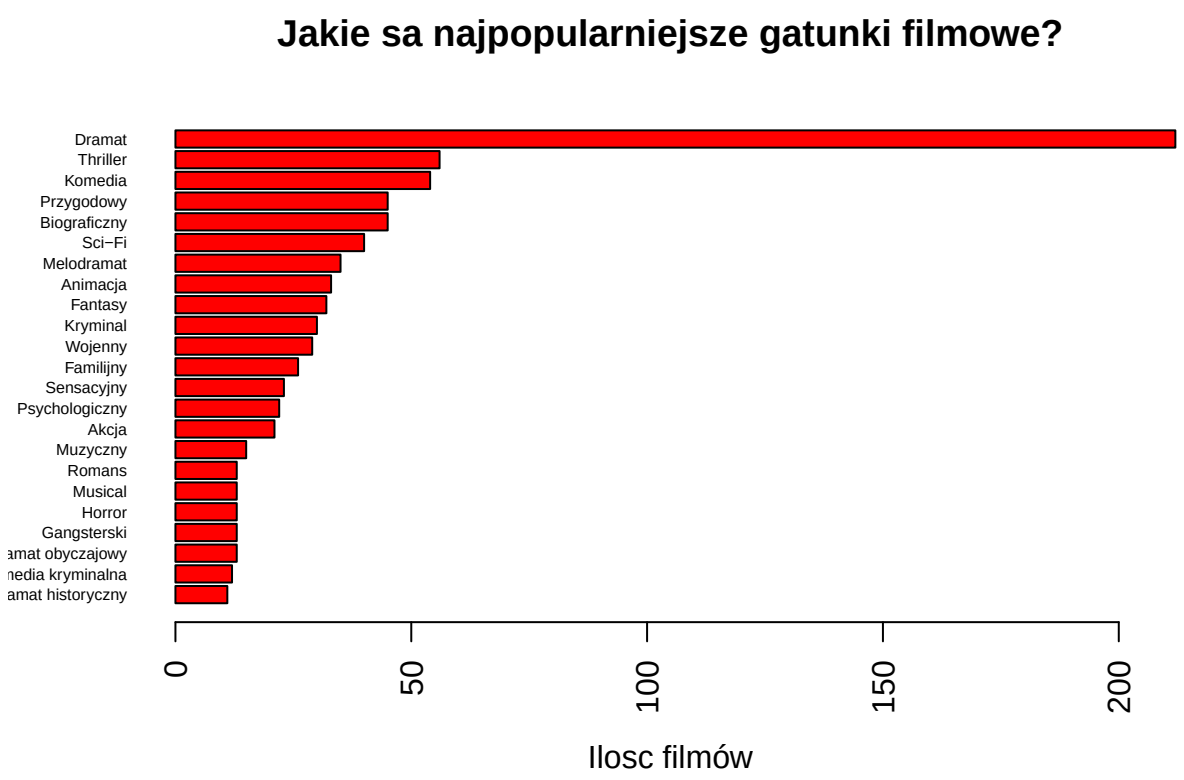
Jak widać wysoko w rankingu plusują się zarówno zagraniczni jak i polscy reżyserzy, co dla mnie było zaskoczeniem, gdyż zwykle mówi się o tym, że Polacy nie doceniają swoich rodaków artystów.

```
PodsumowanieScenarzysci<-sort(table(Scenarzysci)[table(Scenarzysci)>3])
barplot(PodsumowanieScenarzysci,
   names.arg=names(PodsumowanieScenarzysci),horiz=TRUE,las=2,col="green",
   cex.names=0.5,main="Ile filmów zaplanował dany scenarzysta?",xlab="Ilość filmów")
```



Tutaj można zauważyć,że niektórzy reżyserzy są także scenarzystami filmów. Prawdopodobnie dzięki temu, że angażują się także w scenariuszu ich film ma większe szanse na powodzenie.

```
PodsumowanieGatunki<-sort(table(Gatunki)[table(Gatunki)>10])
barplot(PodsumowanieGatunki,
   names.arg=names(PodsumowanieGatunki),horiz=TRUE,las=2,col="red",
   cex.names=0.5,main="Jakie są najpopularniejsze gatunki filmowe?",xlab="Ilość filmów")
```



Kolejne zaskoczenie - prawie połowa filmów z rankingu popularności jest zaliczana do kategorii "Dramat". Komedie plusują się dopiero na trzecim miejscu.

Uwaga! Jeżeli dany film był zaliczany do kilku gatunków lub krajów to został policzony kilkukrotnie.

```
PodsumowanieKraje<-sort(table(KrajeProdukcji)[table(KrajeProdukcji)>7])
barplot(PodsumowanieKraje,
   names.arg=names(PodsumowanieKraje),horiz=TRUE,las=2,col="purple",
   cex.names=0.5,main="W którym kraju wyprodukowano najwięcej filmów?",xlab="Ilość filmów")
```

Powyżej sprawdziły się moje oczekiwania - jesteśmy fanami filmów kręconych w USA :)

Na razie jednak analizowałam wykresy ilościowe. Jak zatem ma się ilość do jakości? Zanalizuję to na ostatnim przykładzie filmów zkategoryzowanych po krajach produkcji.

```
MiejscaProdukcji<-unique(KrajeProdukcji)
OcenaKraju<-vector("numeric",length(MiejscaProdukcji))
names(OcenaKraju)<-MiejscaProdukcji
for(i in 1:500){
     KrajProdukcji<-as.character(
       PodstawoweInformacje[[i]][as.character(
          PodstawoweInformacje[[i]][,1])=="produkcja:",2])
    KrajProdukcji<-unlist(stri_split_regex(
       KrajProdukcji,"(?<=\\p{Ll})(?=\\p{Lu})"))
    OcenaKraju[KrajProdukcji]<-OcenaKraju[KrajProdukcji]+Oceny[i]
}
OcenaKraju<-sort(OcenaKraju/table(KrajeProdukcji)[names(OcenaKraju)],decreasing=TRUE)
```

Tabela pokazująca średnią ocenę filmów wyprodukowanych w danym kraju:

Śrenia ocena

Kraj	Ocena
Nowa Zelandia	7.83125
Irlandia	7.785

Kraj	**Ocena**
Brazylia	7.76
Indie	7.685
Hongkong	7.6825
Ukraina	7.65
Niemcy	7.603962
Grecja	7.6
Czechy	7.595
RPASingapur	7.59
Tajlandia	7.59
Egipt	7.59
Indonezja	7.59
Kenia	7.59
Jordania	7.59
Arabia Saudyjska	7.59
Ghana	7.59
Polska	7.582857
Kanada	7.577143
Francja	7.571373
Hiszpania	7.57
Islandia	7.57
Japonia	7.567647
Bułgaria	7.56
Szwajcaria	7.558333
Australia	7.557778
Dania	7.536667
Szwecja	7.535833
Korea Południowa	7.53
Finlandia	7.5275
Argentyna	7.515
Rosja	7.5075
Holandia	7.5075
Norwegia	7.5
Austria	7.5
Meksyk	7.49
Chiny	7.488
Węgry	7.465
ZSRR	7.43
Słowacja	7.42
Portugalia	7.42
Bahamy	7.38
Estonia	7.37
Jugosławia	7.36
Belgia	7.36
RFN	6.405
USA	6.022332
Włochy	5.432353
Zjednoczone Emiraty Arabskie	3.795
Fed. Rep. Jugosławii	3.78
Wielka Brytania	1.68044
RPAUSA	0
Izrael	0

Porównując z wykresem wyżej widzimy, że raczej mamy zależność odwrotną - im więcej filmów tym są one gorzej oceniane.Gorszej jakości lub może po porstu jest przesyt danym typem filmu? Jedynie Niemcy utrzymali dosyć wysoką pozycję jeśli chodzi o ilość i jakość swoich produkcji.