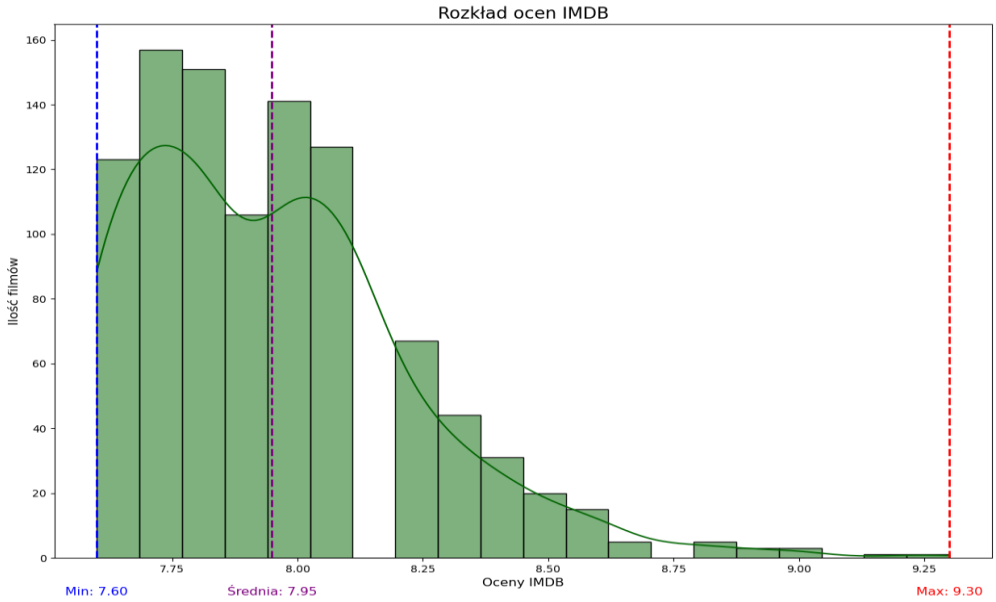
1. Przygotowanie danych

- Dane zacząłem przygotowywać w Excelu (usunąłem min z kolumny runtime, zmieniłem format gross na number, przejrzałem ogólnie wartości w każdej kolumnie po stworzeniu z danych tabelki, wprowadziłem kilka zmian).
- Importowałem dane do Jupyter Notebooka, stworzyłem tam tabelę, sprawdziłem ile jest błędnych wartości w rankingu, usunąłem jedną kolumnę która pojawiła mi się przy importowaniu z excela. Zapisałem i importowałem do PostgreSQL.
- Stworzyłem tabelkę jednak, aby móc załadować cały dataset (program nie opuścił niektórych rzędów z błędną wartością do typu) przypisałem niektórym kolumnom nieodpowiedni typ, który później zmieniłem na poprawny (np. załadowałem gross jako VARCHAR i zmieniłem go na INT).
- Następnie imputowałem dane. Aby uniknąć przerwania ciągłości rankingu filmów, zdecydowałem się na imputację wartości w miejscach, gdzie występują puste wartości null lub blank.
    - Uzupełniłem te braki średnią gross dla całego zbioru danych. Dzięki temu wszystkie filmy zyskały wartość, co zapewnia spójność danych, a jednocześnie nie wpływa na ogólną średnią.
    - Dla brakujących wartości kolumnie meta_score postanowiłem użyć dominanty. Dzięki temu uzupełnione dane lepiej odzwierciedlają typowe oceny filmów.
    - W kolumnie certificate dane uzupełniłem na podstawie wspólnych gatunków filmów (np. dla filmu bez kategorii o gatunku dramat i kryminał wstawiłem najczęściej pojawiającą się kategorie dla tych typów filmów).
- Na koniec sprawdziłem czy występują duplikaty(nie było ich). (Zapytania w SQL w oddzielnym pliku oraz szczegółowy opis operacji w excelu, operacje w JN zapisane na jednym pliku z wykresami i analizą)
- Zaimportowałem dane z SQL do JN i zacząłem tworzyć wykresy, do tworzenia niektórych wykresów musiałem stworzyć nowe tabele, jedna (aktorzy) z nich została stworzona w SQL za pomocą zapytania union, aby mieć spis wszystkich aktorów i ocen filmów w których wystąpili.

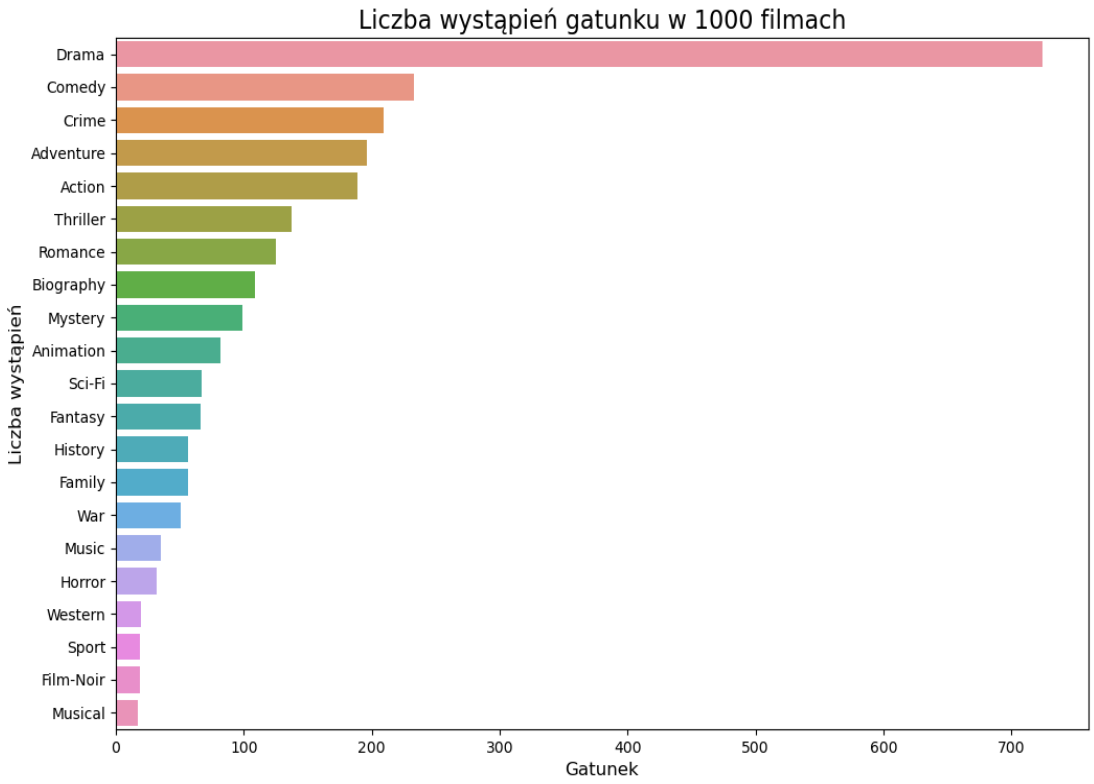
2. Wizualizacja danych.

2.1 Rozkład ocen filmów na IMDb.



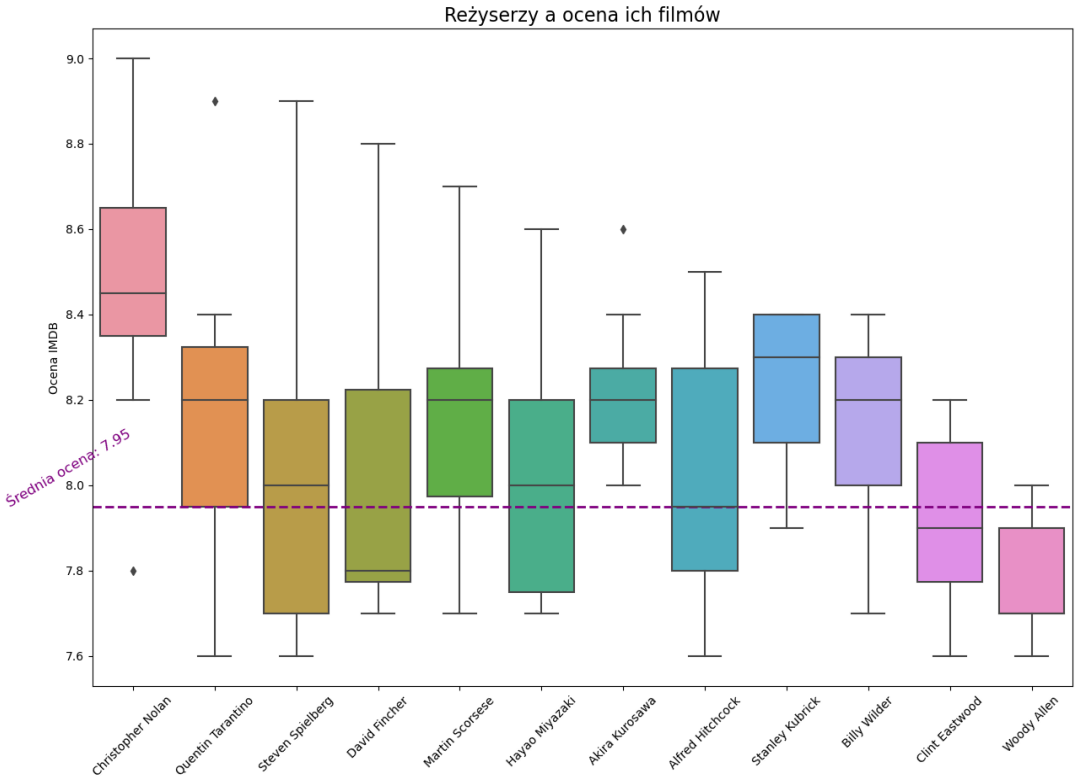
Wykres pokazuje nam ocene w rankingu IMDB, a oś y ilość filmów o takiej samej ocenie, wskazuje liczbę filmów w zależności od ich ocen. Pokazana jest linia, która przedstawia ogólny trend ilości filmów a ocena w rankingu. Wykres ten posiada asymetrię prawostronną co mówi nam o tym, że wieksza część badanych filmów ma ocenę niższą od średniej. Wykres również wskazuje jakią minimalną ocenę musi otrzymać film aby znalazł się w rankingu i ocene maksymalną którą trzeba pobić, aby osiągnąć 1 miejsce w rankingu.

2.2 Liczba wystąpień gatunku w rankingu



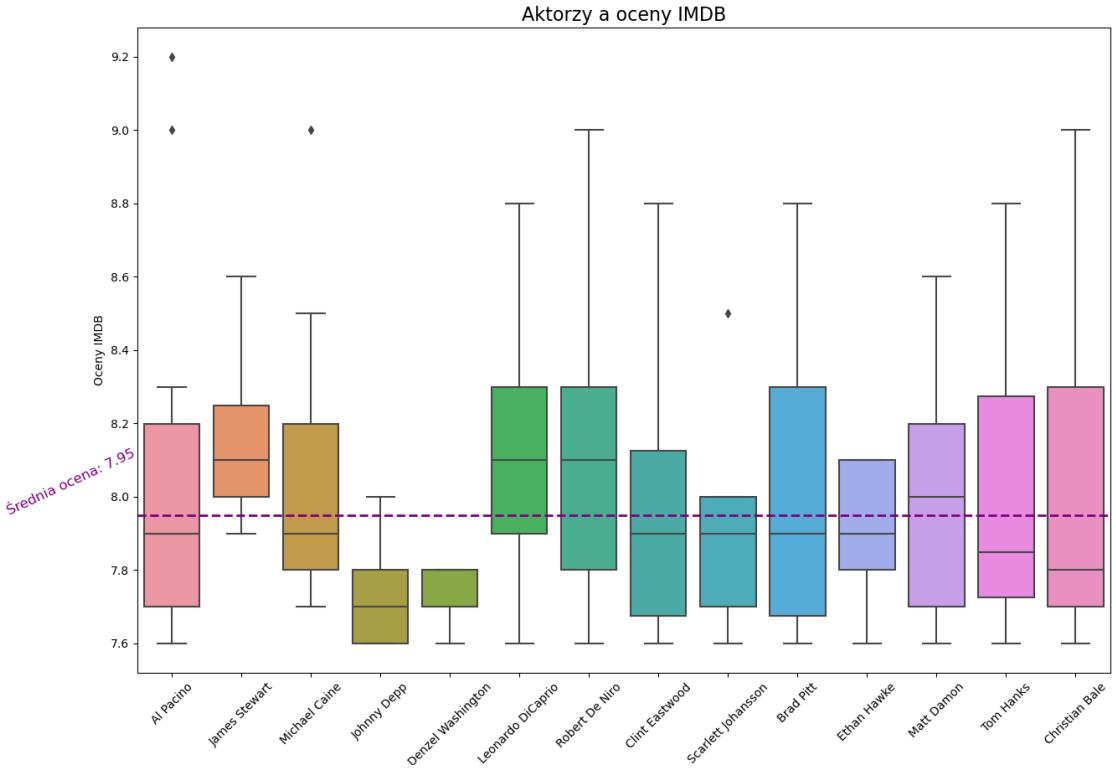
Wykres przedstawia nam jak dużo filmów o danym gatunku lub z jego elementami znajduje się w top 1000 IMDB. Wykres przedstawia, że najczęściej w rankingu występują dramaty lub z jego elementami ponad 700 filmów z całego rankingu, często wysoko oceniane (na miarę pojawienia się w top 1000) są również komedie, kryminały, przygodowe oraz akcji lub posiadające te elementy. Musicale, filmy-noir, westerny oraz sportowe to gatunki, które najrzadziej pojawiają się w w całym rankingu.

2.3 Reżyser a ocena ich filmów.



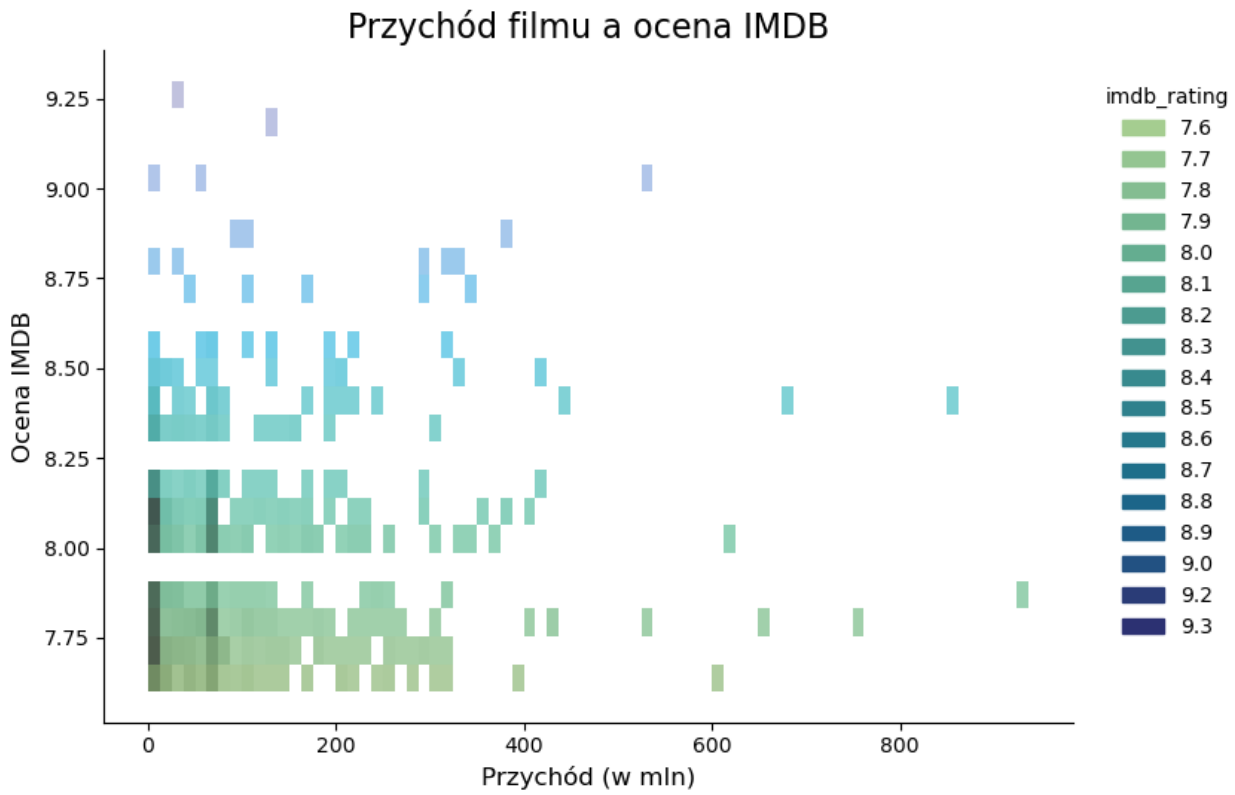
Powyższy wykres przedstawia najczęściej występujących reżyserów w całym rankingu (12 reżyserów, ponieważ od 9 do 12 mieli tyle samo filmów w rankingu). Oceny ich filmów są przedstawione za pomocą wykresu boxplot. Każdy z prostokątów ilustruje rozstęp między pierwszym a trzecim kwartylem, natomiast linia w środku prostokąta oznacza medianę. Z prostokątów wychodzą linie, które wskazują minimalne i maksymalne oceny reżyserów, a wartości te nie zawierają wartości odstających. Kropki nad wykresami reprezentują oceny, które znacząco różnią się od pozostałych. Dzięki temu wykresowi możemy stwierdzić, że reżyser Christopher Nolan ma najwyższą medianę ocen (około 8.4), a zakres ocen jego filmów jest stosunkowo mały, co wskazuje na stabilność wysokich ocen. Akira Kurosawa również należy do grona reżyserów o stabilnych ocenach jako jedyny, którego wszystkie filmy mają oceny wyższe od średniej całego rankingu. Z kolei Steven Spielberg jest reżyserem, którego filmy charakteryzują się większą zmiennością ocen, co świadczy o ich niestabilności.

2.4 Aktor a ocena filmów.



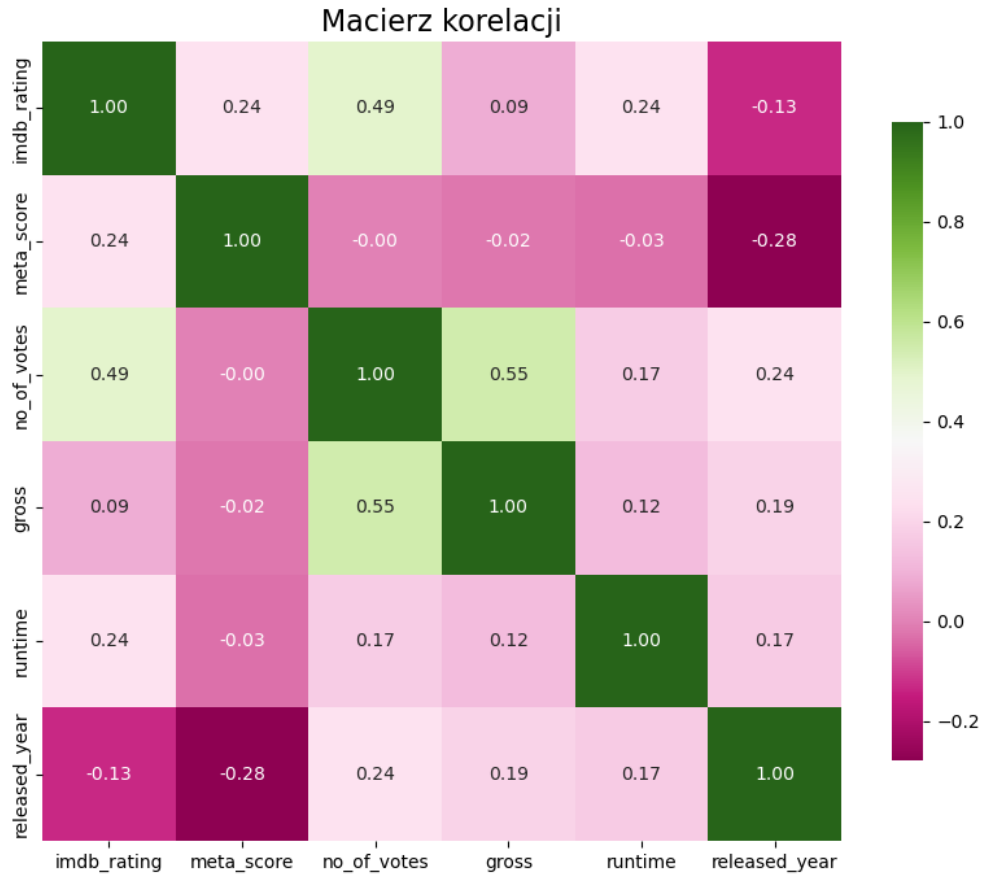
Ten wykres interpretuje się tak samo jak z reżyserami, sposób wyboru aktorów również jest taki sam. Możemy stwierdzić że aktorzy tacy jak Leonardo DiCaprio, Robert De Niro, James Stewart i Matt Damon mają medianę ocen powyżej średniej wartości 7,95, co oznacza, że filmy z ich udziałem są zazwyczaj oceniane wyżej niż przeciętnie. Większość aktorów ma wysokie zróżnicowanie w ocenie ich filmów. Najwyższą oceną dla aktora w tym rankingu należy do Al Pacino – 9.2.

2.5 Przychód filmu a ocena IMDB.



Na przedstawionym wykresie zobrazowano zależność między przychodami filmów a ich ocenami na IMDb, przy czym kolory odpowiadają różnym wartościom ocen. Wynika z tego ,że większość filmów osiąga przychody do około 400 mln USD, nawet te wysoko oceniane. Tylko nieliczne produkcje przekroczyły próg 450 mln USD. Nie widać wyraźnej zależności wskazującej, że najbardziej dochodowe filmy uznawane są za bardzo dobre.

2.6 Macierz korelacji zmiennych



Powyższa macierz korelacji przedstawia zależności między zmiennymi. Im bardziej zielony kolor, tym silniejsza dodatnia korelacja, co oznacza, że wzrost jednej zmiennej wiąże się ze wzrostem drugiej. Siła dodatniej korelacji mieści się w przedziale (0-1], gdzie 0 oznacza brak korelacji, a wartości ujemne do -1 oznaczają korelację ujemną (gdy jedna zmienna rośnie, druga maleje). Ogólnie, zmienne w tym zbiorze danych są słabo skorelowane między sobą. Można jednak zauważyć wyróżniającą się dodatnią korelację między oceną IMDb a liczbą głosów, a także między liczbą głosów a przychodami. Może to sugerować, że im bardziej popularny jest film (czyli otrzymuje więcej opinii/głosów), tym wyższy jest jego przychód. Korelacja ta pokrywa się z wykresem 2.5, mimo podobnych opinii widzów, różne produkcje mogą osiągać zupełnie odmienne wyniki finansowe.

3. Wnioski z analizy danych.

	reżyser character varying (40)	Średnia_ocena_imdb numeric	no_filmów bigint
1	Christopher Nolan	8.46	8
2	Peter Jackson	8.40	5
3	Francis Ford Coppola	8.40	5
4	Charles Chaplin	8.33	6
5	Sergio Leone	8.27	6
6	Stanley Kubrick	8.23	9
7	Akira Kurosawa	8.22	10
8	Quentin Tarantino	8.18	8
9	Martin Scorsese	8.17	10

1. Z analizy tabeli wynika, że reżyserzy mają wpływ na ocen filmów. Skupiłem się na tych twórcach, którzy nie tylko mają najwyższą średnią ocen, ale również wyreżyserowali więcej niż 5 filmów, co pozwala wskazać reżyserów z bogatym dorobkiem filmowym oraz wysokimi ocenami. **Przykłady takie jak Christopher Nolan, Stanley Kubrick czy Akira Kurosawa pokazują, że talent reżyserski odgrywa znaczącą rolę w odbiorze filmów.** Ci reżyserzy mają na swoim koncie wiele produkcji o bardzo wysokich ocenach, co świadczy o ich umiejętnościach.

	gatunek character varying (30)	Średnia_ocena_imdb numeric	no_filmów bigint
1	Action, Adventure, Fantasy	8.20	6
2	Action, Adventure	8.18	5
3	Crime, Drama	8.16	26
4	Action, Adventure, Drama	8.15	14
5	Action, Drama, Mystery	8.08	5
6	Drama, War	8.07	15
7	Comedy, Drama, War	8.06	5
8	Crime, Drama, Film-Noir	8.06	5
9	Drama, Mystery, Sci-Fi	8.04	5

2. Niektóre gatunki wyróżniają się wyższą średnią ocen IMDB, co może wskazywać, że są one odbierane lepiej przez widzów niż pozostałe gatunki. W tej tabeli również zastosowałem selekcje, która wskazuje ocenę tych gatunków które wystąpiły więcej niż pięć razy. Najliczniejszą grupę z bardzo wysoką oceną tworzą filmy z gatunku dramatu kryminalnego, na drugim miejscu dramat u wojennego. **Pokazuje to, że gatunek wpływa na odbiór filmu.**

	reżyser character varying (40)	gatunek character varying (30)	Średnia_ocena_imdb numeric	no_filmów bigint
1	Francis Ford Coppola	Crime, Drama	8.60	3
2	Francis Ford Coppola	Drama, Mystery, Thriller	7.80	1
3	Francis Ford Coppola	Drama, Mystery, War	8.40	1
4	Peter Jackson	Action, Adventure, Drama	8.80	3
5	Peter Jackson	Adventure, Fantasy	7.80	2
6	Sergio Leone	Action, Drama, Western	8.00	1
7	Sergio Leone	Crime, Drama	8.40	1
8	Sergio Leone	Drama, War, Western	7.60	1
9	Sergio Leone	Western	8.53	3

3. **Na podstawie powyższej tabeli możemy zauważyć, że ocena filmów zależy zarówno od reżysera jak i od gatunku w jakim się specjalizują.** Przyglądając się trzem reżyserom z wieloma produkcjami w rankingu, widać, że najwyższe oceny uzyskują filmy należące do gatunków, w których reżyserzy stworzyli najwięcej produkcji. Można zatem wnioskować, że widzowie doceniają filmy, gdy są one tworzone przez reżyserów w ulubionym gatunku reżysera.

	aktor character varying (30)	gatunek character varying (30)	Średnia_ocena_imdb numeric	no_filmów bigint
1	Diane Keaton	Crime, Drama	8.60	3
2	Diane Keaton	Comedy, Romance	8.00	1
3	Diane Keaton	Comedy, Drama, Romance	7.90	1
4	Diane Keaton	Comedy, War	7.70	1
5	Ian McKellen	Action, Adventure, Drama	8.80	3
6	Ian McKellen	Action, Adventure, Sci-Fi	7.90	1
7	Ian McKellen	Adventure, Fantasy	7.80	2
8	Ian McKellen	Adventure, Family, Fantasy	7.60	1
9	Viggo Mortensen	Action, Adventure, Drama	8.80	2
10	Viggo Mortensen	Biography, Comedy, Drama	8.20	1
11	Viggo Mortensen	Comedy, Drama	7.90	1
12	Viggo Mortensen	Action, Crime, Drama	7.60	1

4. **Podobnie jak reżyserzy, aktorzy mają swoje ulubione gatunki filmowe, w których najczęściej występują, a produkcje te zazwyczaj otrzymują najwyższe oceny spośród wszystkich gatunków, w których aktorzy brali udział**. Na przykładzie trzech aktorów: Diane Keaton, Iana McKellena oraz Viggo Mortensena, można zauważyć, że filmy z ich udziałem w gatunkach, w których regularnie się pojawiają, zdobywają wyjątkowo wysokie oceny. To może sugerować, że widzowie lepiej oceniają filmy, w których widzą aktorów z doświadczeniem jak najlepiej zagrać role w danym gatunku filmu.

	aktor character varying (30)	reżyser character varying (40)	Średnia_ocena_imdb numeric	no_filmów bigint
1	Al Pacino	Francis Ford Coppola	8.60	3
2	Al Pacino	Brian De Palma	8.10	2
3	Al Pacino	Michael Mann	8.00	2
4	Al Pacino	Martin Brest	8.00	1
5	Al Pacino	Martin Scorsese	7.90	1
6	Al Pacino	Sidney Lumet	7.85	2
7	Al Pacino	Mike Newell	7.70	1
8	Al Pacino	James Foley	7.70	1
9	Brad Pitt	David Fincher	8.40	3
10	Brad Pitt	Guy Ritchie	8.30	1
11	Brad Pitt	Steve McQueen	8.10	1
12	Brad Pitt	Terry Gilliam	8.00	1
13	Brad Pitt	Quentin Tarantino	7.95	2
14	Brad Pitt	Adam McKay	7.80	1
15	Brad Pitt	Steven Soderbergh	7.70	1
16	Brad Pitt	Bennett Miller	7.60	1
17	Brad Pitt	Barry Levinson	7.60	1
18	Leonardo DiCaprio	Christopher Nolan	8.80	1
19	Leonardo DiCaprio	Martin Scorsese	8.30	3
20	Leonardo DiCaprio	Steven Spielberg	8.10	1
21	Leonardo DiCaprio	Edward Zwick	8.00	1
22	Leonardo DiCaprio	Quentin Tarantino	8.00	2
23	Leonardo DiCaprio	Alejandro G. Iñárritu	8.00	1
24	Leonardo DiCaprio	James Cameron	7.80	1
25	Leonardo DiCaprio	Lasse Hallström	7.80	1

5. **Na ocenę filmów wpływa również zgrana współpraca między reżyserem a aktorami. Z powyższej tabeli wynika, że filmy, w których aktorzy wielokrotnie współpracowali z tym samym reżyserem, otrzymują średnio lepsze oceny niż te, w których grali u danego reżysera po raz pierwszy.** Może to wynikać z lepszego dopasowania aktorów do stylu reżysera, co przekłada się na bardziej spójną współpracę i tworzenie produkcji, które spotykają się z wyższym uznaniem widzów.

Podsumowując:

1. Reżyserzy mają istotny wpływ na ocenę filmów, zwłaszcza ci, którzy stworzyli więcej niż 5 produkcji, jak Christopher Nolan, Stanley Kubrick czy Akira Kurosawa, których filmy regularnie otrzymują wysokie oceny.
2. Niektóre gatunki filmowe, jak dramat kryminalny i wojenny, zdobywają wyższe średnie oceny, co sugeruje, że widzowie bardziej doceniają te kategorie.
3. Zarówno reżyserzy, jak i gatunki, w których się specjalizują, wpływają na oceny filmów – produkcje w ulubionych gatunkach reżyserów otrzymują wyższe noty.
4. Aktorzy, podobnie jak reżyserzy, mają swoje ulubione gatunki, w których ich filmy są lepiej oceniane, co widać na przykładzie Diane Keaton, Iana McKellena i Viggo Mortensena.
5. Zgrana współpraca reżyserów z aktorami również podnosi ocenę filmu – filmy z wielokrotną współpracą reżysera i aktora otrzymują lepsze oceny od tych, w których współpracują po raz pierwszy.