Analiza wybranych danych rzeczywistych z wykorzystaniem metod statystyki opisowej

Statystyka Stosowana

Aleksandra Szczur 276047
Agnieszka Staszkiewicz 268791

6.05.2024

Spis treści

1 Wstęp

1.1 Cel pracy

Celem naszego raportu jest zbadanie zależności ocen IMDb, największej na świecie internetowej bazy danych na temat filmów, a kategorii filmowych. Oceny są wystawiane przez użytkowników platformy w zakresie [1,10]. Dzięki zastosowaniu podstawowych statystyk będziemy w stanie stwierdzić, czy tworzenie filmów z określonej kategorii zwiększa szanse na uzyskanie wysokich ocen od widzów.

1.2 Charakterystyka zbioru danych

Zbiór danych "Movie Industry"[1] obejmuje 7669 rekordów, z których każdy opisuje konkretny film. Filmy wchodzące w skład tego zestawu datowane są kolejno od roku 1986 do 2020, co stanowi średnio około 220 produkcji rocznie. Wyboru filmów dołączonych do zbioru dokonano na podstawie ich popularności w poszczególnych latach według serwisu IMDb.

Każdy rekord zawiera 15 kolumn, reprezentujących różne cechy charakterystyczne filmu:

- budget - budżet przeznaczony na wyprodukowanie filmu
- company - wytwórnia filmowa
- country - kraj produkcji
- director - reżyser
- genre - główny gatunek filmu
- gross - dochód z filmu
- name - nazwa filmu
- rating - oznaczenie wiekowe
- released - data premiery
- runtime - czas trwania filmu
- score - ocena IMDb, jako liczba z zakresu [1-10]
- votes - ilość oddanych głosów
- star - główny aktor
- writer - scenarzysta
- year - rok premiery filmu

[1] `https://www.kaggle.com/datasets/danielgrijalvas/movies/data`

Cechy te można podzielić na dane:

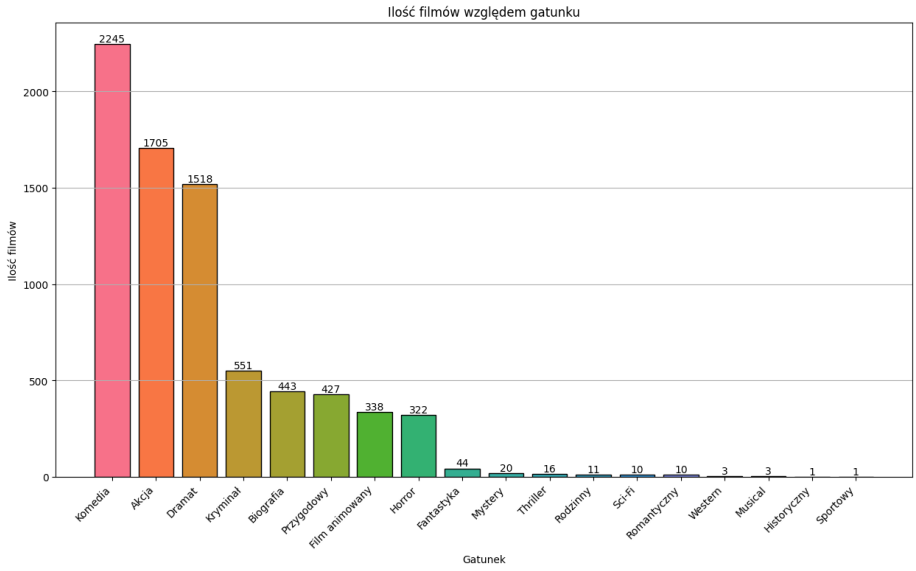
- ilościowe: gross, runtime, score, votes,
- jakościowe: company, country, director, genre, name, rating, released, star, writer, year.

1.3 Źródło

Zestaw danych o nazwie "Movie Industry"został stworzony przez Daniela Grijalve z zamiarem dogłębnego zbadania przemysłu filmowego na przestrzeni ostatnich czterech dekad. Zbiór ten jest dostępny na platformie Kaggle [2], a sam autor gorąco zachęca innych do wspólnej nad nim pracy na platformie GitHub[3], aby stale go rozwijać.

1.4 Przedstawienie danych

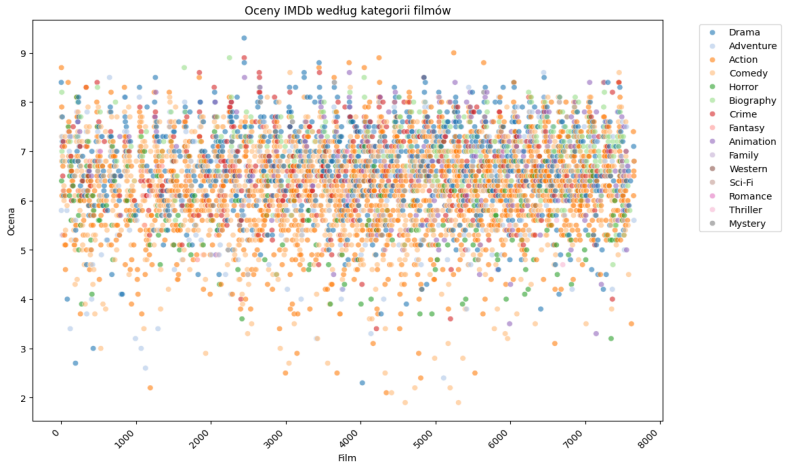
Poniższy wykres przedstawia ilość filmów z danej kategorii. Widzimy, że liczby znacznie różnią się od siebie, przez co wyniki mogą nie być w pełni wiarygodne.



Rysunek 1: Liczba filmów w danej kategorii.

Drugi wykres przedstawia oceny IMDb dla wszystkich filmów z bazy danych. Oś X określa numer wiersza danego filmu, natomiast oś Y to ocena widowni. Dodatkowo dane zostały oznaczone kolorem w zależności od kategorii filmowej.

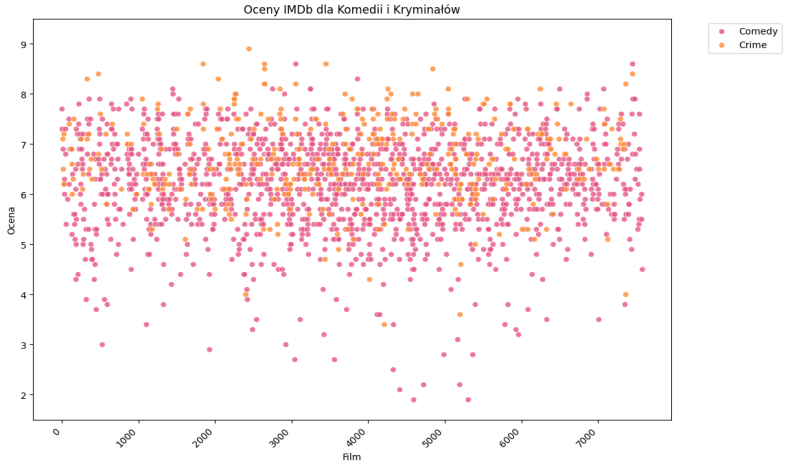
[2] `https://www.kaggle.com/`
[3] `https://github.com/danielgrijalva/movie-stats`



Rysunek 2: Oceny IMDb dla filmów ze zbioru.

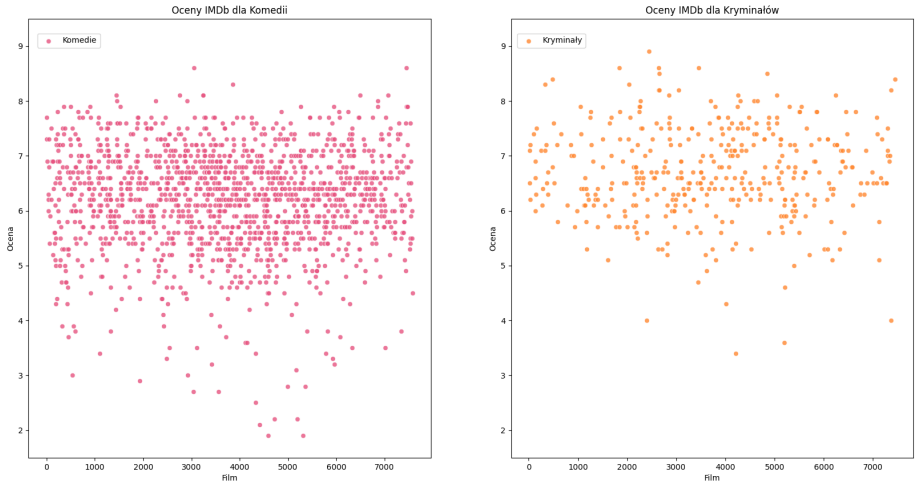
W naszym raporcie skoncentrujemy się na dwóch kategoriach filmowych: komediach i kryminałach. Te dwie kategorie różnią się znacząco od siebie, co pozwoli nam stwierdzić, czy konkretny gatunek filmowy, dzięki swoim charakterystycznym cechom, może być lepiej oceniany przez widzów.

Trzeci wykres ilustruje oceny IMDb dla komedii i kryminałów. Oś X określa numer wiersza danego filmu, natomiast oś Y to ocena widowni. Kolorem różowym oznaczono komedie, natomiast filmy kryminalne są oznaczone kolorem pomarańczowym.



Rysunek 3: Oceny IMDb dla komedii i kryminałów razem.

Aby ułatwić analizę, możemy rozdzielić dane z trzeciego wykresu na dwa osobne wykresy. Poniżej na lewym wykresie przedstawione są oceny filmów z gatunku komedii, natomiast na prawym wykresie znajdują się oceny filmów z gatunku kryminału.



Rysunek 4: Oceny IMDb dla komedii i kryminałów osobno.

2 Podstawowe statystyki

Analiza opisowa jest fundamentalnym zagadnieniem w statystyce, mającym na celu opisanie i zrozumienie danych za pomocą różnorodnych technik i wskaźników statystycznych. Jest to podstawowy proces, który umożliwia poznanie charakterystyki zbioru danych oraz identyfikację podstawowych cech, które są istotne dla dalszej analizy.

2.1 Średnia arytmetyczna

Średnia arytmetyczna $\bar{x}$ jest miarą centralnej tendencji, która oblicza się jako suma wszystkich wartości zmiennej x_i dla badanej zbiorowości, podzielona przez liczbę jednostek w tej zbiorowości (n). Średnią arytmetyczną możemy obliczyć za pomocą wzoru:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

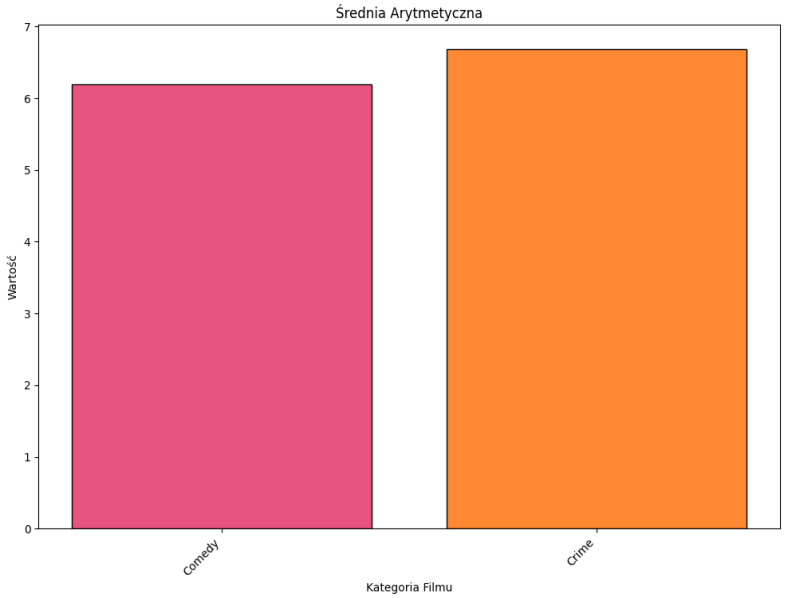
gdzie:

$\bar{x}$ - średnia arytmetyczna,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres średniej arytmetycznej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Średnia arytmetyczna
Komedia	6.19
Kryminał	6.69

Tabela 1: Średnia arytmetyczna ocen IMDb w zależności od kategorii filmowej.



Rysunek 5: Średnia arytmetyczna ocen IMDb w zależności od kategorii filmowej.

Z wykresu i tabeli wynika, że filmy kryminalne otrzymują wyższe oceny od komedii. Możemy więc stwierdzić, że na przestrzeni czterech dekad kryminały podobają się widowni bardziej od filmów komediowych.
Widzimy, że gatunek filmowy ma faktyczny wpływ na to jak zostanie oceniony.

2.2 Średnia harmoniczna

Średnia harmoniczna jest miarą odwrotną do średniej arytmetycznej.Jest zdefiniowana jako iloraz liczby elementów w zbiorze (n) przez sumę odwrotności wartości x_i. Otrzymujemy wzór na średnią harmoniczną:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

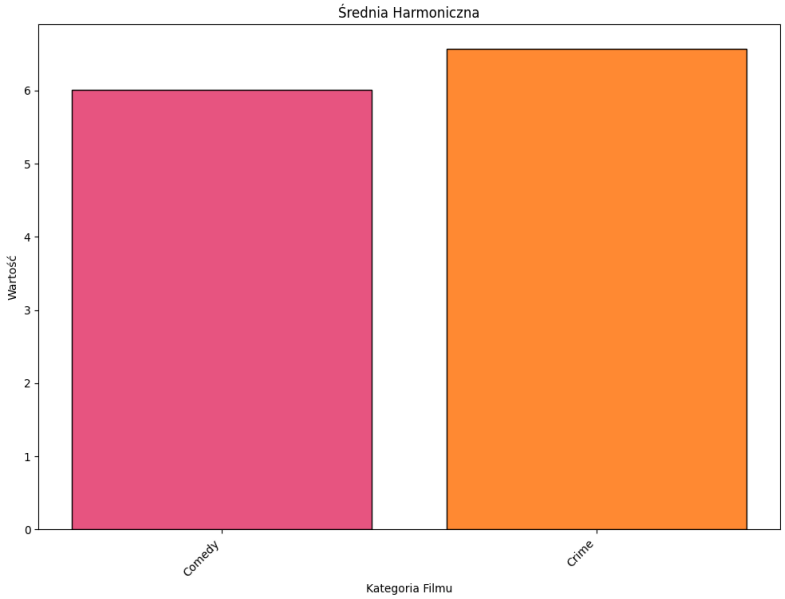
gdzie:

H - średnia harmoniczna,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres średniej harmonicznej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Średnia harmoniczna
Komedia	6.01
Kryminał	6.57

Tabela 2: Średnia harmoniczna ocen IMDb w zależności od kategorii filmowej..



Rysunek 6: Średnia harmoniczna ocen IMDb w zależności od kategorii filmowej.

W tym przypadku, podobnie jak w średniej arytmetycznej, widzimy, że filmy kryminalne zdobywają śrenio lepsze oceny, niż filmy komediowe.

Średnia harmoniczna jest w przypadku ocen IMDb również odpowiednim sposobem na wyznaczanie ich średnich wartości, ponieważ oceny IMDb są wystawiane przez użytkowników w skali [1,10]. Gdyby ocena mogłaby przyjąć wartość 0, wtedy taka średnia byłaby nieokreślona.

2.3 Średnia geometryczna

Średnia geometryczna jest pierwiastkiem n-tego stopnia z iloczynu wszystkich wartości x_i. Jest to sposób mierzenia "średniego tempa wzrostu"w danych, ponieważ bardziej ekstremalne wartości mają większy wpływ na wynik niż w przypadku średniej arytmetycznej. Wzór na średnią geometryczną możemy zapisać jako:

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

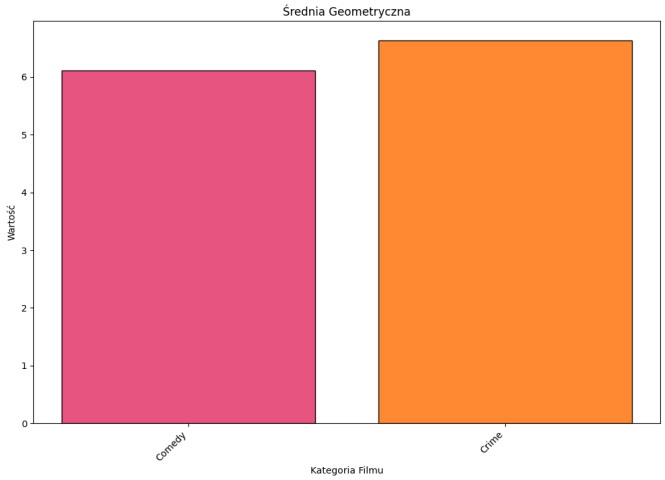
gdzie:

G - średnia geometryczna,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres średniej geometrycznej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Średnia geometryczna
Komedia	6.11
Kryminał	6.63

Tabela 3: Średnia geometryczna ocen IMDb w zależności od kategorii filmowej.



Rysunek 7: Średnia geometryczna ocen IMDb w zależności od kategorii filmowej.

W tym wypadku, ponownie, kryminały otrzymały średnio wyższą ocenę od filmów komediowych.

Znowu dzięki temu, że oceny IMDb są wystawiane w zakresie [1,10] możemy obliczyć ich średnią geometryczną. Gdyby można było wystawić ocenę równą 0, wtedy również średnia geometryczna przyjęłaby wartość 0. W takim wypadku taka statystyka nie byłaby właściwa.

2.4 Średnia winsorowska

Średnia winsorowska jest miarą centralnej tendencji, podobną do średniej arytmetycznej, ale bardziej odporną na wpływ wartości skrajnych. Średnia winsorowska przycina wartości skrajne, zastępując je wartościami największą (dla wartości skrajnych na górze) lub najmniejszą (dla wartości skrajnych na dole), które nie są przycinane. Wzór na średnią winsorowską ma postać:

$$W = \frac{1}{n}\left((k+1)\cdot x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_i + (k+1)\cdot x_{(n-k)}\right)$$

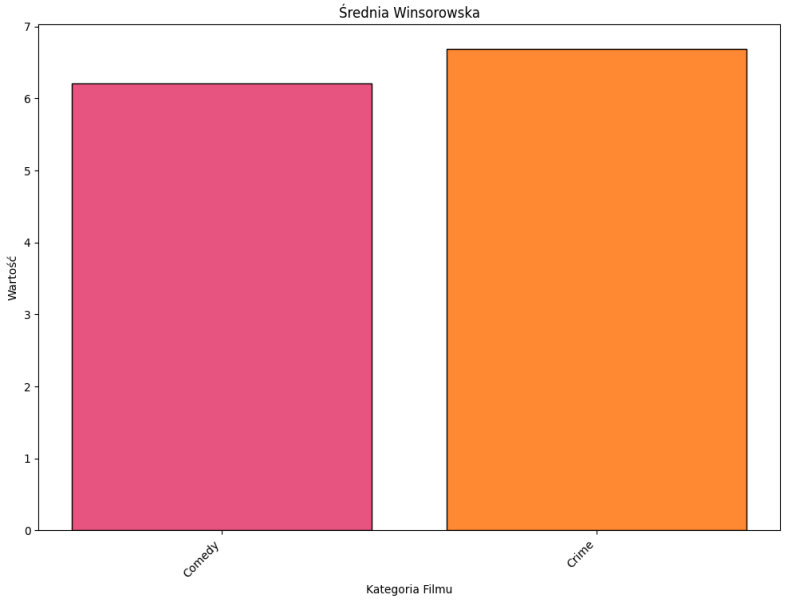
gdzie:

W - średnia winsorowska,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
k - odsetek usuwanych wartości skrajnych,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres średniej winsorowskiej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Średnia winsorowska
Komedia	6.21
Kryminał	6.69

Tabela 4: Średnia winsorowska ocen IMDb w zależności od kategorii filmowej.



Rysunek 8: Średnia winsorowska ocen IMDb w zależności od kategorii filmowej.

W naszym przypadku pomijamy 10% wartości największych i najmniejszych.

Z wykresu i tabeli ponownie zauważamy, że filmy kryminalne cieszą się lepszymi ocenami widzów. Dodatkowo, wartości średniej ucinanej są bardzo podobne do wartości średniej arytmetycznej tych samych kategorii filmowych. Oznacza to, że skrajne wartości z końców przedziałów nie mają znaczącego wpływu na średnią arytmetyczną. Dodatkowo, oceny wystawiane przez użytkowników są zwykle zbliżone do siebie, co sugeruje pewną jednorodność w kwestii ocen.

2.5 Średnia ucinana

Średnia ucinana polega na usunięciu z danych ustalonej liczby najmniejszych i największych wartości, a następnie uśrednieniu pozostałych wartości.

$$\bar{x_t} = \frac{1}{n-2k} \cdot \sum_{i=k+1}^{n-k} x_i$$

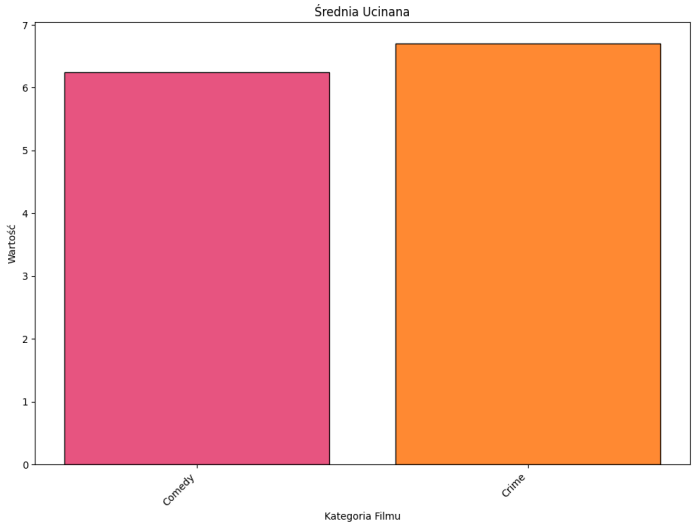
gdzie:

$\bar{x_t}$ - średnia winsorowska,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
k - liczba usuwanych wartości skrajnych z każdej strony,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres średniej ucinanej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Średnia ucinana
Komedia	6.25
Kryminał	6.71

Tabela 5: Średnia ucinana ocen IMDb w zależności od kategorii filmowej.



Rysunek 9: Średnia ucinana ocen IMDb w zależności od kategorii filmowej.

W naszej analizie 'uciętyśmy', tak jak w średniej windsorowskiej, 10% największych i najmniejszych ocen danej kategorii filmowej.

W tym przypadku ponownie widzimy przewagę filmów kryminalnych i wyniki bardzo podobne do wartości średniej arytmetycznej. Możemy znów wyciągnąć wniosek identyczny do tego z punktu poprzedniego, że wartości z końców przedziałów nie mają istotnego wpływu na średnią arytmetyczną, a oceny wystawiane przez użytkowników są do siebie zbliżone.

2.6 Mediana

Mediana służy do określenia wartości centralnej w uporządkowanym zbiorze danych. W przypadku nieparzystej liczby obserwacji, mediana jest środkową wartością w uporządkowanym zbiorze danych. Natomiast dla parzystej liczby obserwacji, mediana jest średnią arytmetyczną dwóch środkowych wartości. W zależnośći od (n) możemy zapisać wzory określające mediane:

$$\text{Mediana} = \begin{cases} \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{dla } n \text{ parzystego} \\ x_{\left(\frac{n+1}{2}\right)} & \text{dla } n \text{ nieparzystego} \end{cases}$$

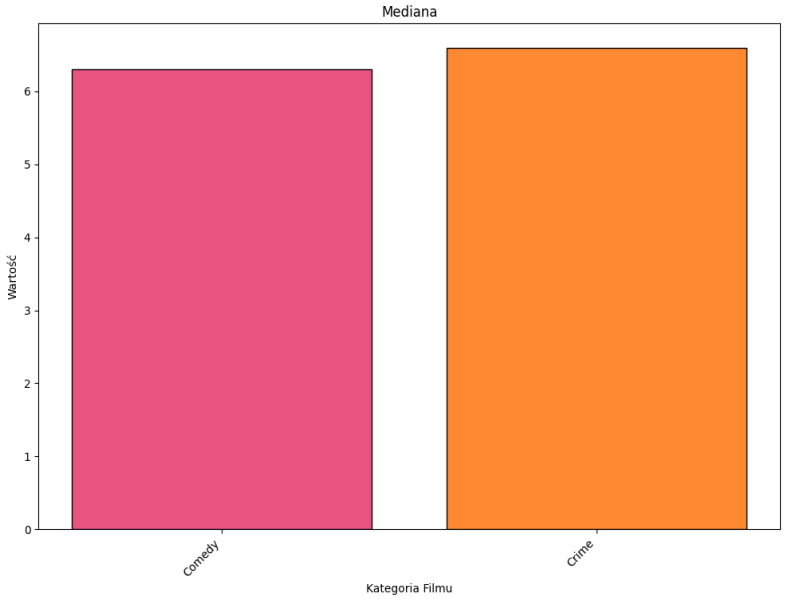
gdzie:

n - liczba jednostek w zbiorowości,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
$x_{\left(\frac{n}{2}\right)}$ - wartość znajdującą się w połowie zbioru danych (dla n parzystego),
$x_{\left(\frac{n}{2}+1\right)}$ - oznacza wartość następującą po $x_{\left(\frac{n}{2}\right)}$ (dla n parzystego),
$x_{\left(\frac{n+1}{2}\right)}$ - oznacza środkową wartość w uporządkowanym zbiorze danych (dla n nieparzystego).

Poniżej przedstawiono tabele i wykres mediany ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Mediana
Komedia	6.3
Kryminał	6.6

Tabela 6: Mediana ocen IMDb w zależności od kategorii filmowej.



Rysunek 10: Mediana ocen IMDb w zależności od kategorii filmowej.

Mediana ocen IMDb są wyższe dla filmów kryminalnych. Ponownie sugeruje nam to, że komedie są mniej lubiane przez widzów. Wartość mediany w przypadku kryminałów jest mniejsza w porównaniu do średniej arytmetycznej, a w przypadku komedii - wyższa. Oznacza to, że w zbiorze ocen filmów komediowych znajdują się kilka bardzo niskich wartości, które wpływają na średnią atrytmetyczną, natomiast w przypadku filmów kryminalnych - wartości wysokich. Ogólnie mówiąc, mediana jest lepszą miarą w przypadku zbiorów z wartościami odstającymi.

2.7 Wariancja

Wariancja jest miarą rozproszenia wartości w zbiorze danych, która określa, jak bardzo te wartości różnią się od średniej arytmetycznej. Wariancja jest obliczana jako średnia arytmetyczna kwadratów odchyleń każdej wartości od średniej arytmetycznej zbioru. Dla populacji, wariancję oznacza się symbolem σ^2, a dla próby symbolem s^2 .Wzór na wariancję ma postać:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

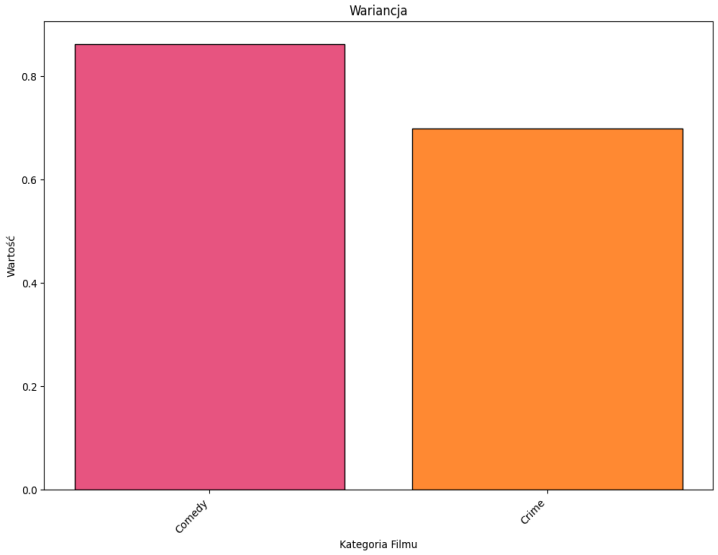
gdzie:

σ^2 - wariancja
n - liczba elementów w próbie (rozmiar próby),
x_i - wartość zmiennej dla i-tej jednostki w próbie,
$\bar{x}$ - średnia arytmetyczna.

Poniżej przedstawiono tabele i wykres wariancji ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Wariancja
Komedia	0.86
Kryminał	0.7

Tabela 7: Wariancja ocen IMDb w zależności od kategorii filmowej.



Rysunek 11: Wariancja ocen IMDb w zależności od kategorii filmowej.

Widzimy, że wartość wariancji ocen dla filmów komediowych jest większa w porównaniu do wariancji ocen filmów kryminalnych. Oznacza to, że oceny filmów komediowych są bardziej zróżnicowane, mniej stabilne w porównaniu z ocenami filmów kryminalnych.

2.8 Odchylenie standardowe z próby

Odchylenie standardowe to miara rozproszenia próbki lub populacji. Jest to pierwiastek kwadratowy z wariancji i wyraża, jak bardzo wartości danych rozpraszają się wokół ich średniej arytmetycznej. Oznaczany jest jako σ. Wzór na odchylenie standardowe to:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

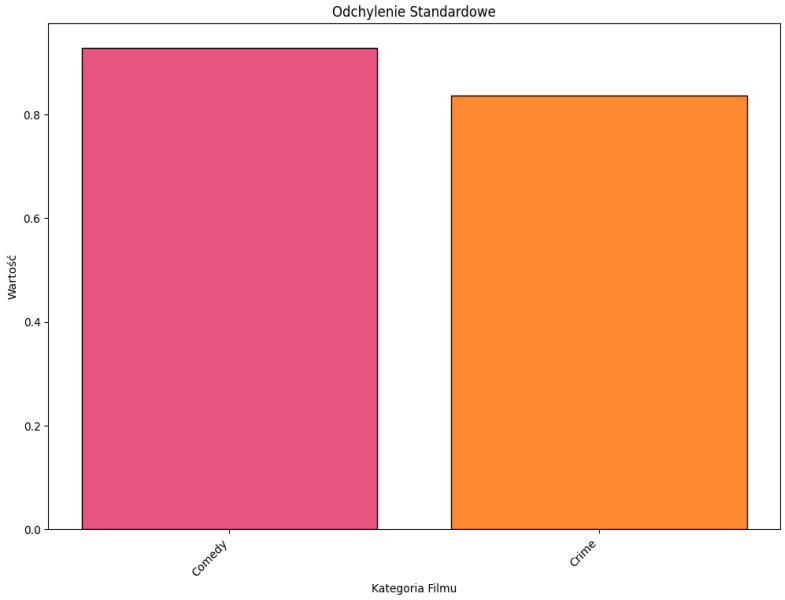
gdzie:

σ - odchylenie standardowe
n - liczba elementów w próbie (rozmiar próby),
x_i - wartość zmiennej dla i-tej jednostki w próbie,
$\bar{x}$ - średnia arytmetyczna.

Poniżej przedstawiono tabele i wykres odchylenia standardowego ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Odchylenie standardowe
Komedia	0.93
Kryminał	0.84

Tabela 8: Odchylenie standardowe ocen IMDb w zależności od kategorii filmowej.



Rysunek 12: Odchylenie standardowe ocen IMDb w zależności od kategorii filmowej.

W tym przypadku, podobnie jak w wariancji, wartość odchylenia standardowego jest większa dla filmów komediowych. Wyciągamy z tego taki sam wniosek, że oceny filmów komediowych wykazują większe zróżnicowanie i są mniej stabilne w porównaniu z ocenami filmów kryminalnych.

2.9 Kwartyle

Kwartyle dzielą wszystkie nasze obserwacje na cztery równe co do ilości obserwacji próby. Określają nam wartości podziału badanej próby w założonych proporcjach. Istnieją trzy kwartyle: pierwszy kwartyl (Q1), drugi kwartyl (Q2), czyli mediana i trzeci kwartyl (Q3).

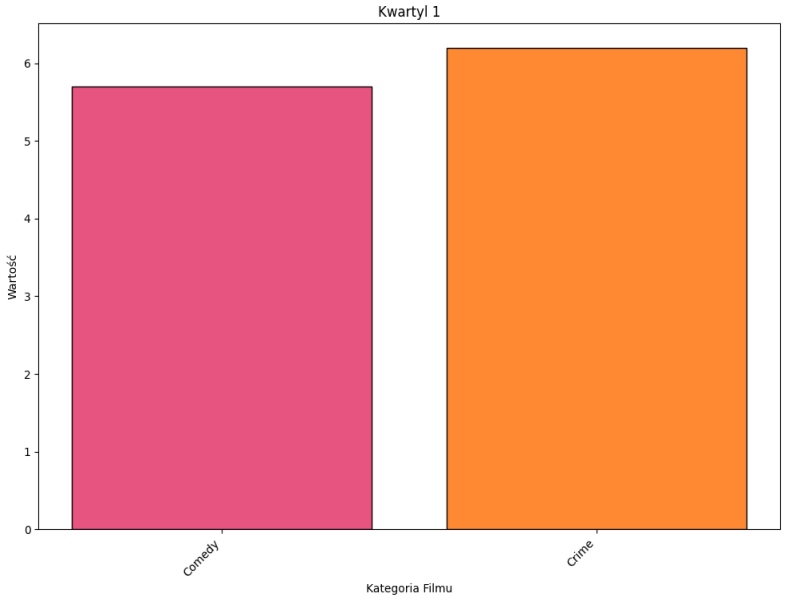
2.9.1 Kwartyl 1

Kwartyl pierwszy (Q1) to punkt, który dzieli najmniejszą część zbioru danych na 25% najniższych wartości. Kwartyl pierwszy jest medianą dolnej połowy próby. Jest to wartość, poniżej której znajduje się 25% obserwacji. Aby obliczyć kwartyl rzędu 1., należy uporządkować dane w porządku rosnącym i wyznaczyć medianę dolnej połowy próby.

Poniżej przedstawiono tabele i wykres kwartylu rzędu 1 ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Kwartyl 1
Komedia	5.7
Kryminał	6.2

Tabela 9: Kwartyl 1 ocen IMDb w zależności od kategorii filmowej.



Rysunek 13: Kwartyl 1 ocen IMDb w zależności od kategorii filmowej.

Widzimy ponownie przewagę filmów kryminalnych. Wartość kwartyla pierwszego filmów komediowych jest niższa w porównaniu z filmami kryminalnymi. Mediana najmniejszej połowy zbioru ocen dla filmów kryminalnych jest większa niż dla filmów komediowych.

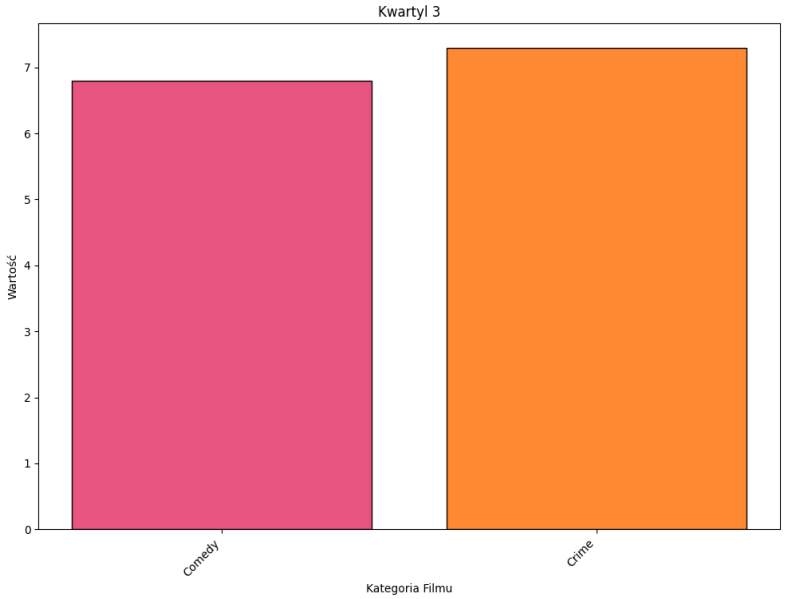
2.9.2 Kwartyl 3

Kwartyl trzeci (Q3) to punkt, który dzieli największą część zbioru danych na 25% najwyższych wartości. Kwartyl trzeci jest medianą górnej połowy próby. Jest to wartość, powyżej której znajduje się 75% obserwacji. Aby obliczyć kwartyl rzędu 3., również należy uporządkować dane w porządku rosnącym i wyznaczyć medianę górnej połowy próby.

Poniżej przedstawiono tabele i wykres kwartylu rzędu 3 ocen IMDb w zależności od kategorii filmowej

Gatunek filmu	Kwartyl 3
Komedia	6.8
Kryminał	7.3

Tabela 10: Kwartyl 3 ocen IMDb w zależności od kategorii filmowej.



Rysunek 14: Kwartyl 3 ocen IMDb w zależności od kategorii filmowej.

Tak jak w przypadku kwartyla pierwszego, wartość kwartyla trzeciego jest większa dla filmów kryminalnych. Mediana największej połowy zbioru ocen dla filmów kryminalnych jest większa niż dla filmów komediowych.

2.10 Rostęp z próby

Rozstęp z próby R to różnica pomiędzy największą a najmniejszą wartością w próbie. Możemy go wyznaczyć za pomocą wzoru:

$$\mathrm{R} = x_{\max} - x_{\min}$$

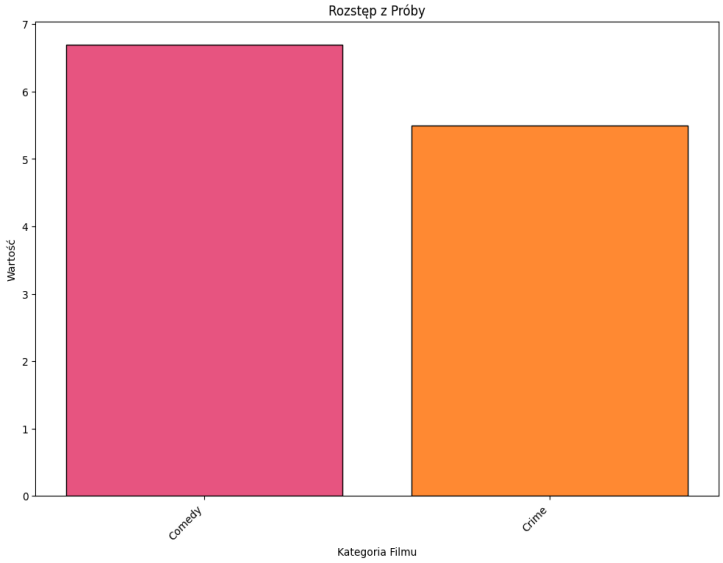
gdzie:

R - rozstęp z próby,
x_{max} - największa wartość w próbie,
x_{min} - najmniejsza wartość w próbie.

Poniżej przedstawiono tabele i wykres rostępu z próby ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Rozstęp z próby
Komedia	6.7
Kryminał	5.5

Tabela 11: Rozstęp z próby ocen IMDb w zależności od kategorii filmowej.



Rysunek 15: Rozstęp z próby ocen IMDb w zależności od kategorii filmowej.

Wartość rozstępu z próby ocen IMDb jest mniejsza dla filmów kryminalnych. Oznacza to, że różnica między najwyższą a najniższą oceną dla filmów komediowych jest niższa niż dla komedii. Czyli, oceny użytkowników są w przypadku kryminałów mniej zróżnicowane i bardziej stabilne.

2.11 Rostęp międzykwartylowy

Rozstęp międzykwartylowy IQR to różnica między trzecim a pierwszym kwartylem w próbie, czyli między wartościami, które oddzielają dolne 25% i górne 25% rozkładu.

$$\text{IQR} = Q_3 - Q_1$$

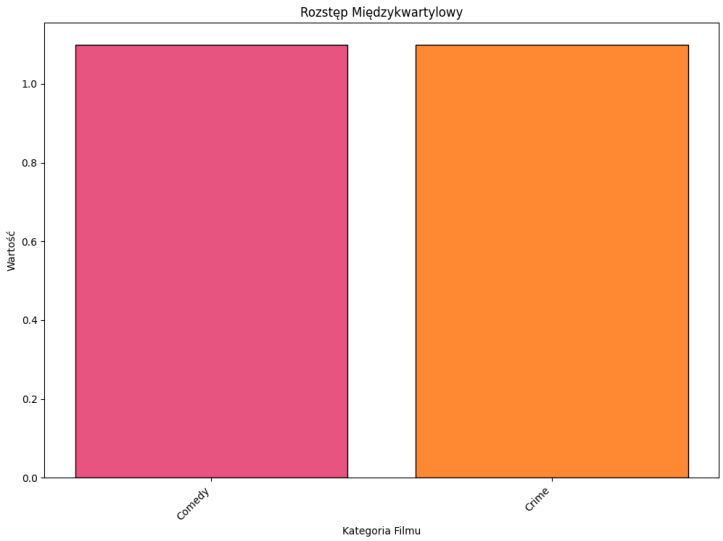
gdzie:

IQR - rozstęp międzykwartylowy,
$Q3$ - kwartyl rzędu 3,
$Q1$ - kwartyl rzędu 1.

Poniżej przedstawiono tabele i wykres rostępu międzykwartylowego ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Rozstęp międzykwartylowy
Komedia	1.1
Kryminał	1.1

Tabela 12: Rozstęp międzykwartylowy ocen IMDb w zależności od kategorii filmowej.



Rysunek 16: Rozstęp międzykwartylowy ocen IMDb w zależności od kategorii filmowej.

Wartość rozstępu międzykwartylowego jest zarówno w przypadku filmów komediowych jak i kryminalnych taka sama. Sugeruje to nam, że rozrzut danych jest w obu przypadkach do siebie podobny, a oceny mają podobną zmienność.

2.12 Odchylenie przeciętne od wartości średniej

Odchylenie przeciętne od wartości średniej jest miarą rozproszenia danych, która informuje o średniej odległości poszczególnych wartości od ich średniej arytmetycznej. Uwzględnia bezwzględne odległości, co oznacza, że uwzględnia zarówno wartości powyżej, jak i poniżej średniej. Wartość odchylenia przeciętnego jest mniejsza od wartości odchylenia standardowego. Wzór na odchylenie przeciętne od wartości średniej to:

$$d_1 = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

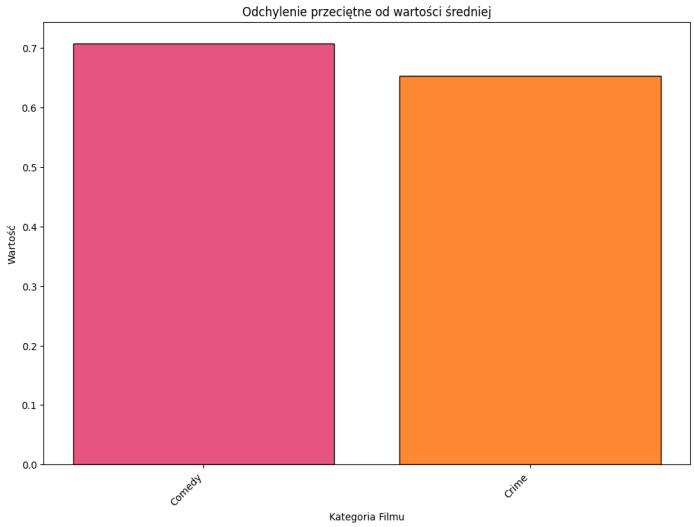
gdzie:

d_1 - odchylenie przeciętne od wartości średniej
n - liczba elementów w próbie (rozmiar próby),
x_i - wartość zmiennej dla i-tej jednostki w próbie,
$\bar{x}$ - średnia arytmetyczna.

Poniżej przedstawiono tabele i wykres odchylenia przeciętnego od wartości średniej ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Odchylenie przeciętne od wartości średniej
Komedia	0.71
Kryminał	0.65

Tabela 13: Odchylenie przeciętne od wartości średniej ocen IMDb w zależności od kategorii filmowej.



Rysunek 17: Odchylenie przeciętne od wartości średniej ocen IMDb w zależności od kategorii filmowej.

Odchylenie przeciętne ocen filmów kryminalnych, tak jak w przypadku odchylenia standardowego, jest mniejsze w porównaniu z filmami komediowymi. Oznacza to, że dla filmów komediowych średnia arytmetyczna odchylania się poszczególnych wyników od wyliczonej średniej arytmetycznej jest większa niż w przypadku filmów kryminalnych.

2.13 Współczynnik zmienności

Współczynnik zmienności jest miarą rozproszenia danych, która pozwala porównywać zmienność między różnymi zbiorami danych o różnych średnich wartościach. Współczynnik określa stosunek odchylenia standardowego do średniej arytmetycznej danych.

Zmienność klasyfikujemy w następujący sposób:
$V \in [0, 20\ (\%)]$ - zmienność mała
$V \in [20, 40\ (\%)]$ - zmienność przeciętna
$V \in [40, 100\ (\%)]$ - zmienność duża
$V \in [100, 150\ (\%)]$ - zmienność bardzo duża
$V \in [150, \infty\ (\%)]$ - zmienność skrajnie duża

Współczynnik zmienności oblicza się wzorem:

$$\mathrm{V} = \frac{s \cdot 100\%}{\bar{x}}$$

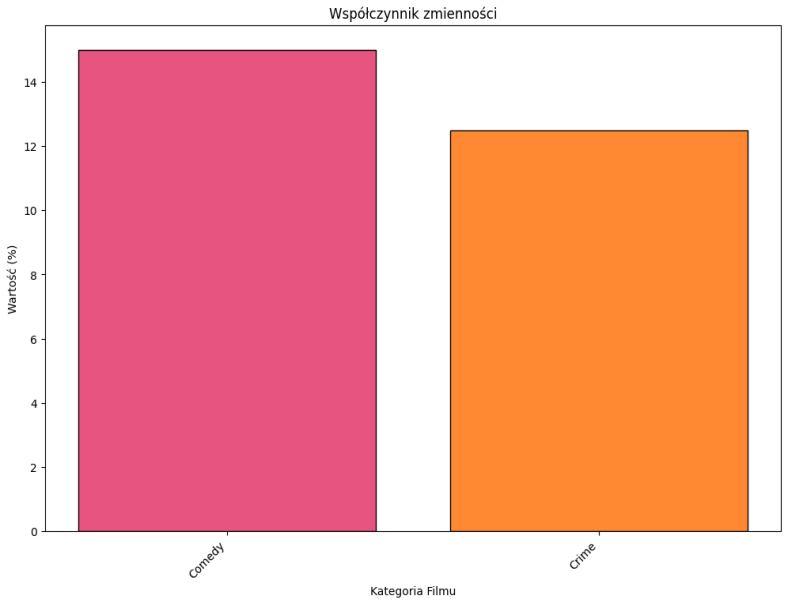
gdzie:

V - odchylenie przeciętne od wartości średniej
s - odchylenie standardowe z próby,
$\bar{x}$ - średnia arytmetyczna.

Poniżej przedstawiono tabele i wykres współczynnika zmienności ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Współczynnik zmienności
Komedia	15%
Kryminał	12.5%

Tabela 14: Współczynnik zmienności ocen IMDb w zależności od kategorii filmowej.



Rysunek 18: Współczynnik zmienności ocen IMDb w zależności od kategorii filmowej.

Z wykresu i tabeli wynika, że zarówno w przypadku filmów komediowych jak i kryminalnych współczynnik zmienności klasyfikuje się do zmienności małej. Oznacza to, że wystawione oceny dla tych dwóch gatunków są mało zróżnicowane.

2.14 Kurtoza

Kurtoza to miara spłaszczenia, która informuje o koncentracji rozkładu.
Gdy K $>$ 3 - rozkład ciężkoogonowy
Gdy K $<$ 3 - rozkład lekkoogonowy
Gdy K $=$ 3 - kurtoza rozkładu normalnego

Kurtozę obliczamy za pomocą wzoru:

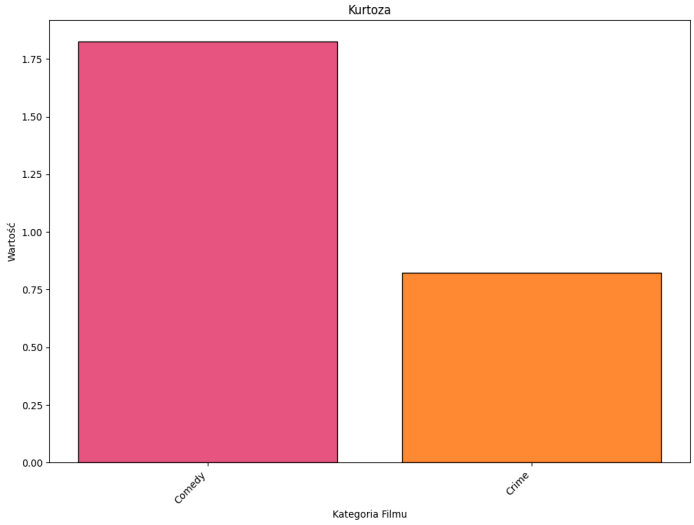
$$\mathrm{K} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

K - kurtoza,
$\bar{x}$ - średnia arytmetyczna,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres kurtozy ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Kurtoza
Komedia	1.83
Kryminał	0.82

Tabela 15: Kurtoza ocen IMDb w zależności od kategorii filmowej.



Rysunek 19: Kurtoza ocen IMDb w zależności od kategorii filmowej.

Widzimy, że zarówno w przypadku komedii jak i kryminałów wartość kurtozy jest mniejsza od 3. Oznacza to, że mamy do czynienia z rozkładami lekkoogonowymi, czyli takimi, w których koncentracja ocen wokół średniej jest niższa niż w przypadku rozkładu normalnego.

2.15 Skośność

Skośność, miara asymetrii obserwacji, informuje jak wyniki dla danej zmiennej kształtują się wokół średniej.

Gdy $\alpha > 0$ - asymetria dodatnia
Gdy $\alpha < 0$ - asymetria ujemna
Gdy $\alpha = 0$ - symetria

Współczynnik skośności obliczamy za pomocą wzoru:

$$\alpha = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

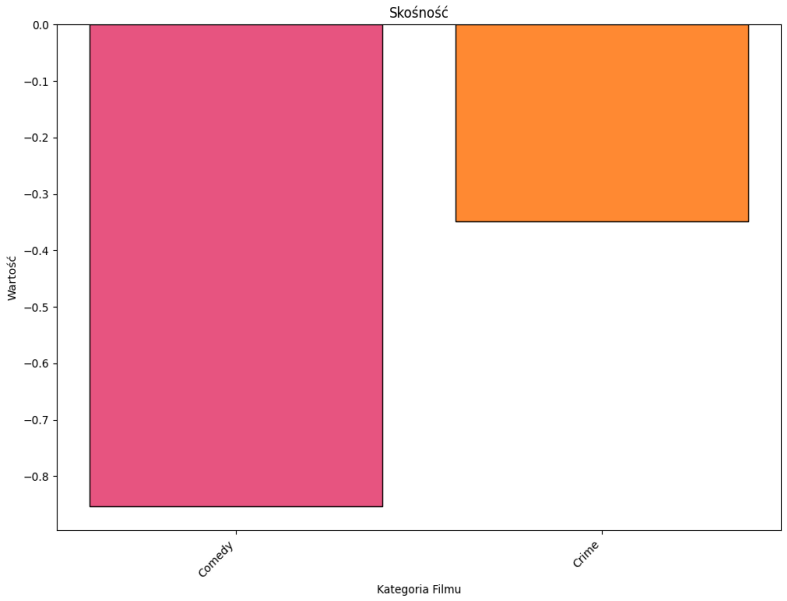
gdzie:

α - współczynnik skośności,
$\bar{x}$ - średnia arytmetyczna,
x_i - wartość zmiennej dla i-tej jednostki w zbiorowości,
n - liczba jednostek w zbiorowości.

Poniżej przedstawiono tabele i wykres skośności ocen IMDb w zależności od kategorii filmowej.

Gatunek filmu	Skośność
Komedia	-0.85
Kryminał	-0.35

Tabela 16: Skośność ocen IMDb w zależności od kategorii filmowej.



Rysunek 20: Skośność ocen IMDb w zależności od kategorii filmowej.

Widzimy, że skośność zarówno w przypadku komedii jak i kryminałów jest wartością ujemną. Oznacza to, że mamy do czynienia z asymetrią ujemną. W takim przypadku badane rozkłady są rozkładami lewostronnie skośnymi.

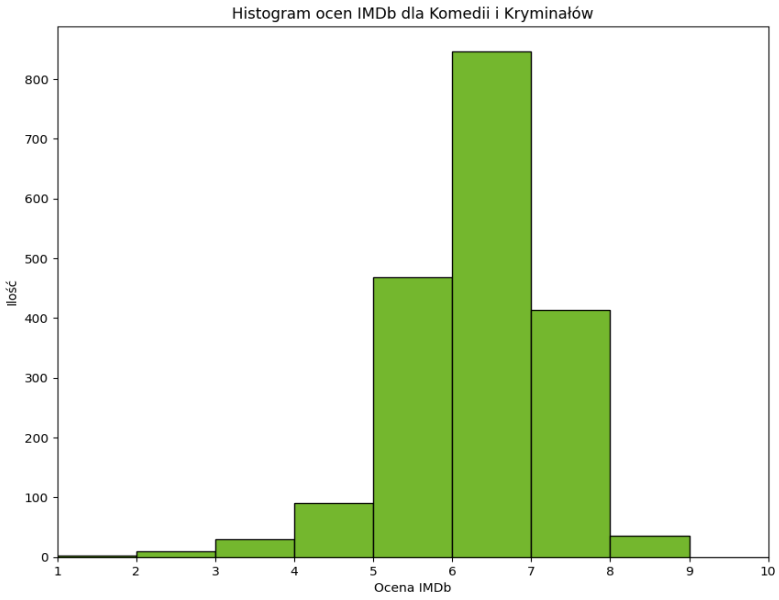
3 Wizualizacja danych

Wizualizacja danych to technika prezentacji informacji i danych za pomocą różnych narzędzi graficznych, takich jak wykresy, diagramy, histogramy czy inne formy graficzne. Ma na celu zrozumienie, analizę i prezentację danych w sposób łatwy do interpretacji i zrozumienia. Wizualizacja danych jest nieodłącznym elementem analizy danych. Dzięki odpowiedniej prezentacji graficznej możemy bardziej efektywnie analizować i interpretować dane, co prowadzi do lepszego zrozumienia problemów i lepszych decyzji opartych na danych.
W tym rozdziale omówione zostaną podstawowe typy wykresów, takie jak: Histogram, Gęstość, Dystrybuanta oraz Wykres pudełkowy.

3.1 Histogram

Histogram to graficzny sposób przedstawienia częstości występowania zmiennej losowej w danym przedziale.

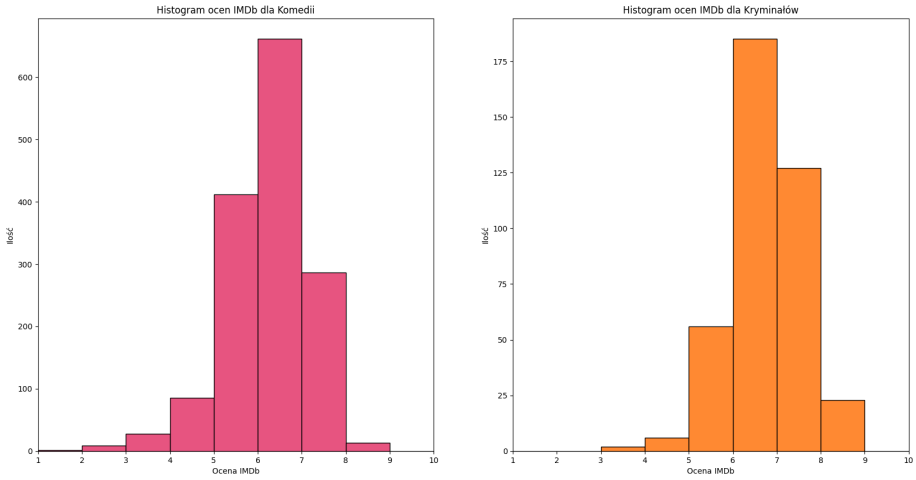
Poniżej przedstawione są histogramy częstości występowania danej oceny IMDb.



Rysunek 21: Histogram ocen IMDb dla komedii i kryminałów razem.

Dla filmów z gatunku komedii i kryminału najczęstsze oceny IMDb to oceny z zakresu [6,7], jest to moda histogramu.

Aby lepiej porównać te dwie kategorie filmowe poniżej przedstawione zostały również osobne histogramy danego gatunku.



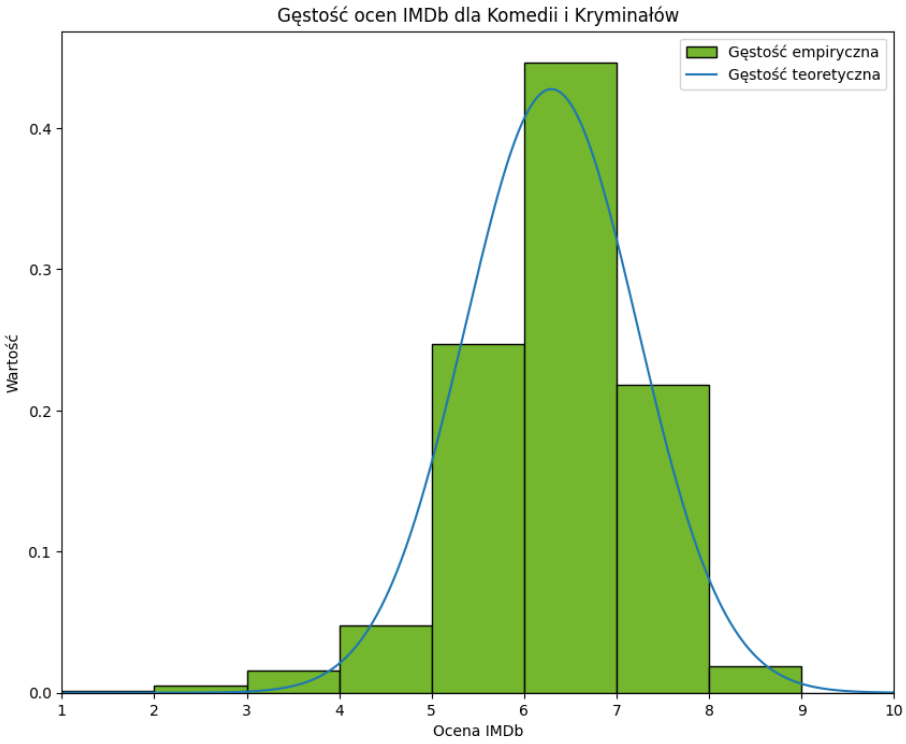
Rysunek 22: Histogram ocen IMDb dla komedii i kryminałów osobno.

Zarówno dla komedii, jak i dla kryminałów najczęstsze oceny IMDb, czyli mody histogramu to oceny z zakresu [6,7].

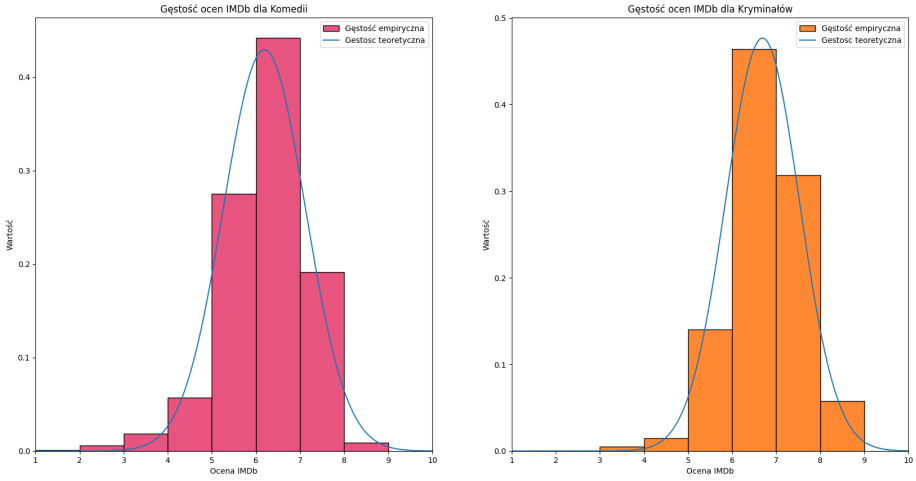
3.2 Gęstość

Gęstość w wizualizacji danych to sposób graficznego przedstawienia, który ilustruje, jak dane są rozmieszczone lub zagęszczone w określonej przestrzeni numerycznej. Jest to narzędzie służące do lepszego zrozumienia rozkładu wartości, wzorców, anomalii oraz relacji między danymi.

Wykres nr 23 przedstawia gęstość ocen dla komedii i kryminałów razem. Natomiast wykres nr 24 prezentuje gęstość ocen osobno dla komedii oraz osobno dla kryminałów.



Rysunek 23: Gęstość ocen IMDb dla komedii i kryminałów razem.



Rysunek 24: Gęstość ocen IMDb dla komedii i kryminałów osobno.

Widzimy, że lewe ogony rozkładów są wydłużone względem prawych. Oznacza to, że rozkłady są lewostronnie skośne, co potwierdza obliczoną wcześniej skośność (2.15).

Wyznaczone gęstośći empiryczne przypominają gęstość rozkładu normalnego, dlatego naniesione zostały na nie obliczone z poniższego wzoru gęstośći teoretyczne:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wartości odchylenia standardowego (σ) i wartości oczekiwanej (μ) zostały obliczone ze wzorów przedstawionych w poprzednich podpunktach i wynoszą:

	Komedie i Kryminały	Komedie	Kryminały
σ	0.93	0.93	0.84
μ	6.3	6.19	6.69

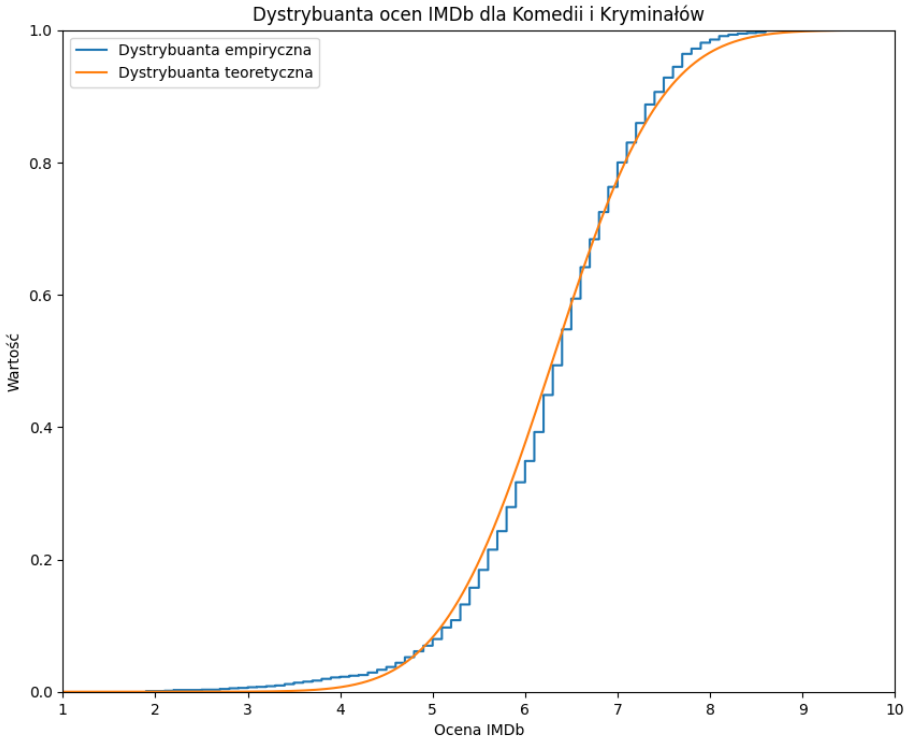
Tabela 17: Wartości odchylenia standardowego i wartości oczekiwanej dla komedii i kryminałów.

Gęstość rozkładu normalnego jest dobrym przybliżeniem gęstości badanego zbioru danych. Można z tego wywnioskować, że więszość ocen wystawionych zarówno dla fimów komediowych, jak i kryminalnych jest zbliżona do średniej arytmetycznej ocen tych filmów. Widzimy również, że o wiele mniej ocen wystawionych jest znacznie mniejszych lub większych od średniej arytmetycznej.

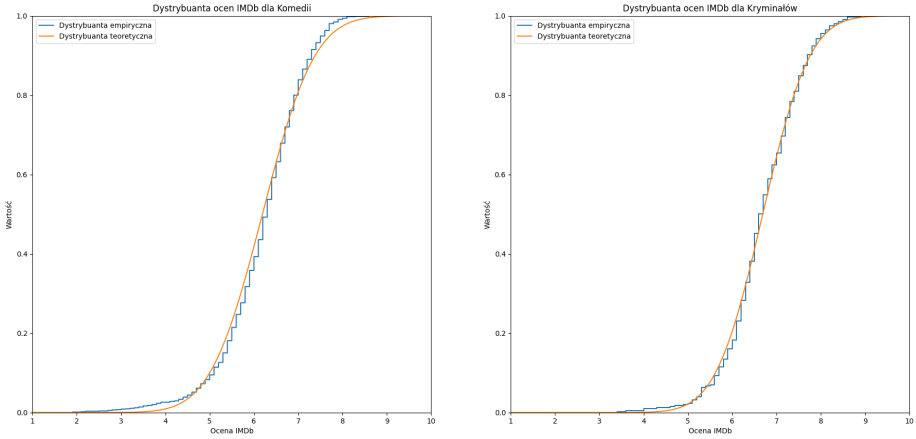
3.3 Dystrybuanta

Dystrybuanta pozwala graficznie ukazać rozkład i rozmieszczenie danych w przestrzeni numerycznej. Poprzez wizualizację dystrybuanty możemy lepiej zrozumieć, jak dane się grupują wokół różnych wartości oraz jak zmienia się ich rozkład.

Wykres nr 25 przedstawia gęstość ocen dla komedii i kryminałów razem. Natomiast wykres nr 26 prezentuje gęstość ocen osobno dla komedii oraz osobno dla kryminałów.



Rysunek 25: Dystrybuanta ocen IMDb dla komedii i kryminałów razem.



Rysunek 26: Dystrybuanta ocen IMDb dla komedii i kryminałów osobno.

Wyznaczone dystrybuanty empiryczne przypominają dytrybuanty rozkładu normalnego, dlatego naniesione na wykresy zostały również teortetyczne dystrybuanty oblizone ze wzoru:

$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

W tym przypadku ponownie wykorzystujemy wartości odchylenia standardowego (σ) i wartości oczekiwanej (μ) z tabeli 17.

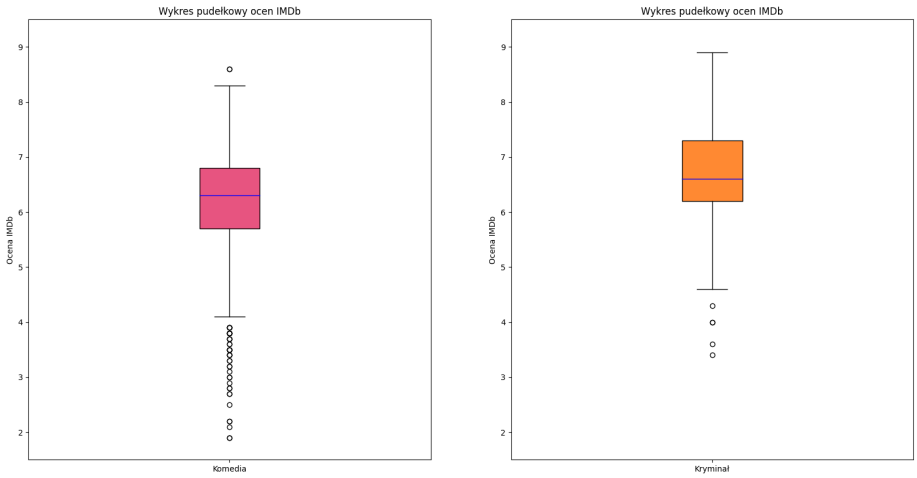
Dystrybuanta rozkładu normalnego jest dobrym przybliżeniem dystrybuanty badanego zbioru danych. Z charakterystycznego kształtu tej dystrybuanty ponownie wnioskujemy, że oceny większości filmów, czy to komediowych, czy kryminalnych, zwykle zbliżają się do średniej arytmetycznej. Większość wyników znajduje się pośrodku przediału, a znaczna mniejszość na jego krańcach.

3.4 Wykres pudełkowy

Wykres pudełkowy pozwala szybko zobaczyć rozkład danych, wykryć wartości odstające i porównać rozkłady między grupami. Składa się z pudełka oraz dwóch tak zwanych "wąsów". Z wykresu pudełkowego możemy odczytać:

- Mediane - linia wewnątrz pudełka
- Kwartyl pierwszy (Q1) - dolna granica pudełka
- Kwartyl trzeci (Q3) - górna granica pudełka
- Rozstęp międzykwartylowy (IQR) - wysokość pudełka (różnica między Q3 a Q1)
- Zakres wartości - wąsy wychodzące z pudełka
- Wartości odstające - punkty nad i pod pudełkiem. Wartości odstające to wartości większe od $Q_3 + 1.5\text{IQR}$ lub mniejsze od $Q_1 - 1.5\text{IQR}$

Poniżej przedstawiono wykresy pudełkowe średnich ocen IMDb dla komedii i kryminału.



Rysunek 27: Wykres pudełkowy ocen IMDb w zależności od kategorii filmowej.

Wykres prawidłowo przedstawia obliczone w poprzednich podpunktach wartości. Ponownie potwierdza zależność ocen wystawionych przez widzów do gatunku filmowego. Kolejny raz widzimy, że filmy kryminalne na przestrzeni czterech dekad były oceniane lepiej od filmów komediowych.

4 Podsumowanie

Z przedstawionych statystyk i wizualizacji wybranego zbioru danych wynika, że filmy kryminalne na przestrzeni lat wypadają lepiej od filmów komediowych. Kryminały zdobrywają średnio lepsze oceny, które dodatkowo są bardziej stabilne niż te wystawione filmom komediowych. Dzięki przybliżeniu rozkładem normalnym widzimy, że więszość ocen wystawionych zarówno dla fimów komediowych, jak i kryminalnych jest zbliżona do średniej arytmetycznej ocen tych filmów. Można stwierdzić, że tworząc film kryminalny mamy większą szansę na sukces niż w przypadku komedii.

Warto dodać, że na wyniki może wpływać znaczna różnica ilości wystawionych opinii, która dla filmów komediowych jest o wiele większa. Dodatkowo, na sukces filmu wpływa cały szereg innych czynników, a nie tylko gatunek. Są to między innymi: główni bohaterowie, reżyserzy, długość filmu czy inne, opisane we wstępie aspekty.

Branża filmowa jest bardzo interesującym polem do wykorzystania statystyki, a nasz raport przybliżył nas do zrozumienia jej dynamiki i trendów. Jej złożoność może nas skłonić do dalszego zgłębiania tego tematu i prowadzenia bardziej zaawansowanych analiz.