*Praktyczne zastowanie testów statystycznych*

Spis treści

[Dane 2](#_Toc156751375)

[Oznaczenia danych w kolumnach 2](#_Toc156751376)

[Wstępne rozpoznanie danych 2](#_Toc156751377)

[Dane numeryczne 3](#_Toc156751378)

[Wykres gęstości punktacji użytkowników 3](#_Toc156751379)

[Wykres gęstości budżetu 4](#_Toc156751380)

[Wykres gęstości przychodu 5](#_Toc156751381)

[Problemy badawcze i ich analiza 5](#_Toc156751382)

[Bibliography 6](#_Toc156751383)

# Dane

## Oznaczenia danych w kolumnach

- names – tytuł filmu

- date\_x – data dnia, w którym odbyła się premiera filmu

- score – ocena użytkowników w skali 0-100

- genre – gatunek lub gatunki, do których zaliczany jest film

- overview – ogólny zarys fabuły filmu

- crew – obsada filmu, aktorzy i aktorki oraz odpowiadające im postać, które grali (aktor/aktorka, postać)

- orig \_title – oryginalny tytuł filmu , w którym film został wypuszczony

- status – status filmu( po premierze, w produkcji, w post produkcji)

- orig\_lang – oryginalny język/języki filmu

- budget\_x – budżet jaki został przeznaczony na produkcje filmu

- revenue – całkowity dochód na całym świecie, który przyniósł film

- country – kraj, w którym został wydany film

## Wstępne rozpoznanie danych

Na początku używamy funkcji „str()” w celu poznania struktury naszych danych. Możemy się dowiedzieć, że mamy do czynienia z 3 kolumnami numerycznymi, a reszta danych to szeroko pojęte napisy.   
Przy pomocy funkcji „head()” oraz „tail()” oglądamy pierwsze i ostatnie 5 rekordów w naszych danych. W ten sposób możemy wysnuć wstępną hipotezę, że potencjalnie można rozpatrzeć innych format danych, niż ten który początkowo wczytał program. Pod dalsze rozpoznanie poddamy kolumny: „status”, „orig\_lang” oraz „country”.  
Aby dowiedzieć się czegoś więcej o w. w. kolumnach oraz co one zawierają, wykorzystamy bilbiotekę „dplyr”. Filtrując dane w tych kolumnach dowiadujemy się, że w kolumnie „status” występują tylko 3 różne wartości. Zmieniamy dane w tej kolumnie na dane faktorowe o 3 stopniach: „Released”, „Post Production” i „In Production”. Pozostałe kolumny niestety miały za dużo unikalnych wartości, żeby ten zabieg powtórzyć.  
Zmieniamy również dane o dacie na odpowiedni format - „Date”  
Za pomocą funkcji „md.pattern()” z pakietu „mice” sprawdzamy potencjalne braki w danych. W naszych danych nie ma braków. Do analizy wybieramy tylko interesujące nas dane, które będziemy wykorzystywać. W takim przypadku odrzucamy „crew”, „overwiev” i „orig\_title”.

## Dane numeryczne

Żeby zbadać dokładniej nasze dane numeryczne posłużymy się funkcją „describe()” (pakiet „psych”) oraz wykresy, do których stworzenia wykorzystamy bibliotekę „ggplot2”. Wyniki w postaci tekstu dają nam pewien pogląd na dane, ale dużo lepiej pokazują to wykresy.

### Wykres gęstości punktacji użytkowników

A graph with a blue line

Description automatically generated

Łącząc wyniki tekstowe z wykresem, możemy powiedzieć, że przeciętna punktacja oscyluje w okolicy 63 i jest ma rozkład ze skośnością lewostronną. Co więcej dzięki wysokiej kurtozie możemy powiedzieć, że dane są ściśle skoncentrowane wokół średniej.

### Wykres gęstości budżetu

A graph showing a green line

Description automatically generated

Wielkość budżetu jest prawostronnie skośny i trochę bardziej skoncentrowany wokół średniej niż rozkład normalny. Średnia jest równa 64882379, a odchylenie standardowe jest równe 57075645, co ze względu na duże wartości jest ciężkie do interpretacji. Po raz kolejny z pomocą przychodzi nam wykres. Można zaobserwować, że dużo filmów ma budżet ok. 45 mln i sukcesywnie spada w dół, ale istnieje jeszcze jedno „wybrzuszenie” w okolicach budżetu równego 270 mln.

### Wykres gęstości przychodu

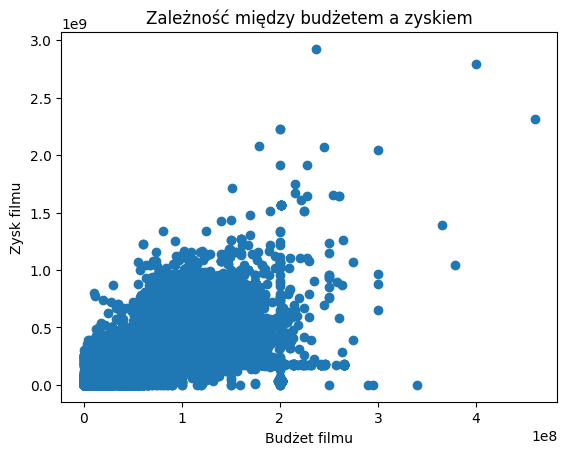
A graph with a purple line

Description automatically generated

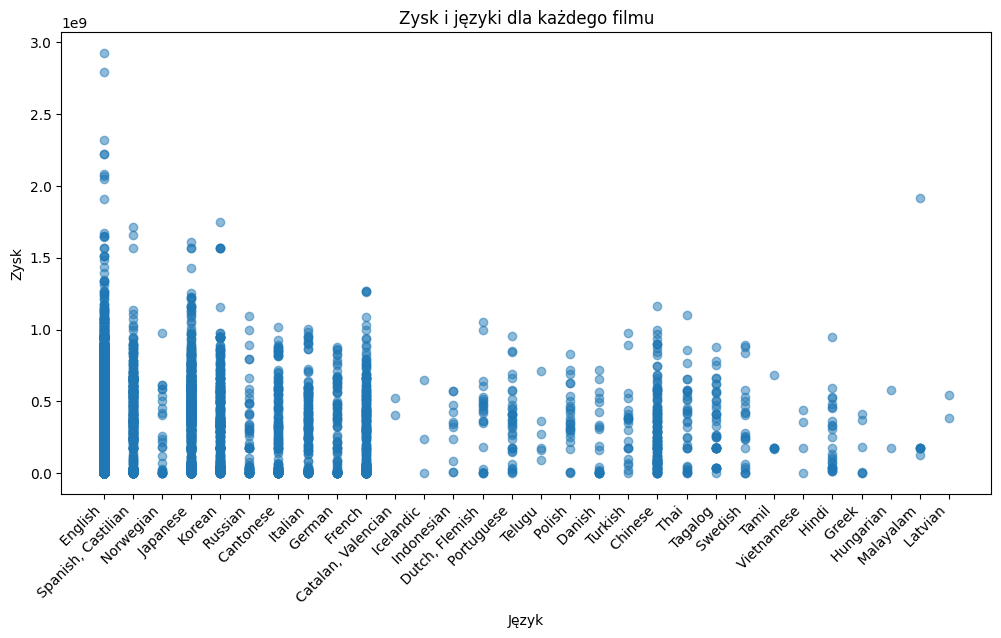
Skośność jest prawostronna, a kurtoza mówi nam, że wyniki są silnie skoncentrowane wokół średniej. Są to wartości z pewnością dużo wyższe niż w przypadku budżetu. Porównując te dwa wykresy możemy dojść do wniosku, że przychód ma mniejsze zróżnicowanie niż budżet. Czy filmy z większym nakładem przynoszą większy zysk? Na to pytanie za pomocą tych dwóch wykresów niestety nie jesteśmy w stanie odpowiedzieć.

# Problemy badawcze i ich analiza

Jednym z pierwszych zagadnień jakie nam się nasuwa jest korelacja pomiędzy budżetem a zyskiem. Czy filmy z większym nakładem przynoszą większe zyski? Korelacja pomiędzy budżetem a zyskiem wynosi 0.67382956917637465 co wyraźnie wskazuje na korelację na granicy silnej i umiarkowanej, co oznacza, że występuje taka zależność. Dodatni znak statystyki intepretujemy tak, że wraz ze wzrostem budżetu wzrasta również zysk.



Kolejnym pytaniem, które sobie postawiliśmy, było czy istnieją jeszcze inne czynniki wpływające na zysk. Postanowiliśmy połączyć to zagadnienie z językami w jakich one oryginalnie powstały.



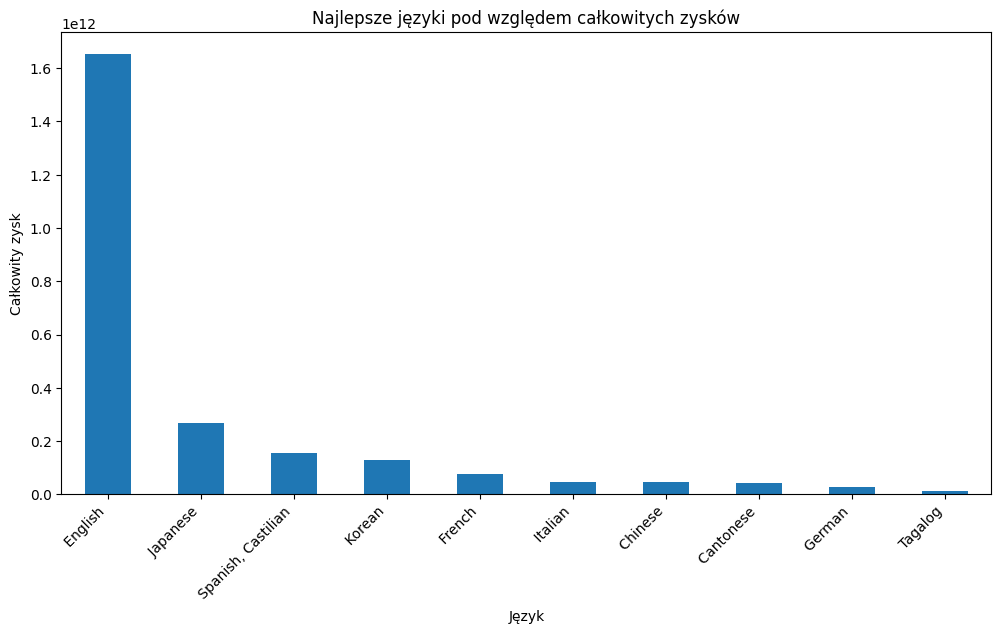
Możemy tutaj wyraźnie dostrzec, że największym zyskiem wyróżniają się filmy w językach angielskim, japońskim, koreańskim, francuskim, chińskim i jego odmianie (cantonese), hinduskim, niemieckim i włoskim. Możemy połączyć to z danymi, które wskazują, że najwięcej filmów powstaje w tych językach.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Patrząc na wykres możemy dostrzec, że filmów, które uzyskały zysk większy niż 1.0 jest niewiele i głównie są to filmy angielskie.

Sprawdzamy również jak wygląda suma zysków dla pojedynczych języków i prezentujemy to na wykresie.



Możemy łatwo zauważyć, że ze względu na swoją liczność filmy w języku angielskim wyraźnie przodują w sumie zysków. Postanawiamy sprawdzić za pomocą testu T-studenta czy istnieje istotna różnica między średnimi zyskiem pierwszych dwóch języków (Angielski i Japoński). Za hipotezę H0 przyjmujemy, że nie ma istotnej różnicy w średnim zysku filmu, jako hipotezę alternatywną, przyjmujemy, że taka istnieje. Za poziom istotności przyjmujemy alfa = 0.05.

Test Statistic: 12.704724686441223

P-value: 7.0358668679091325e-34

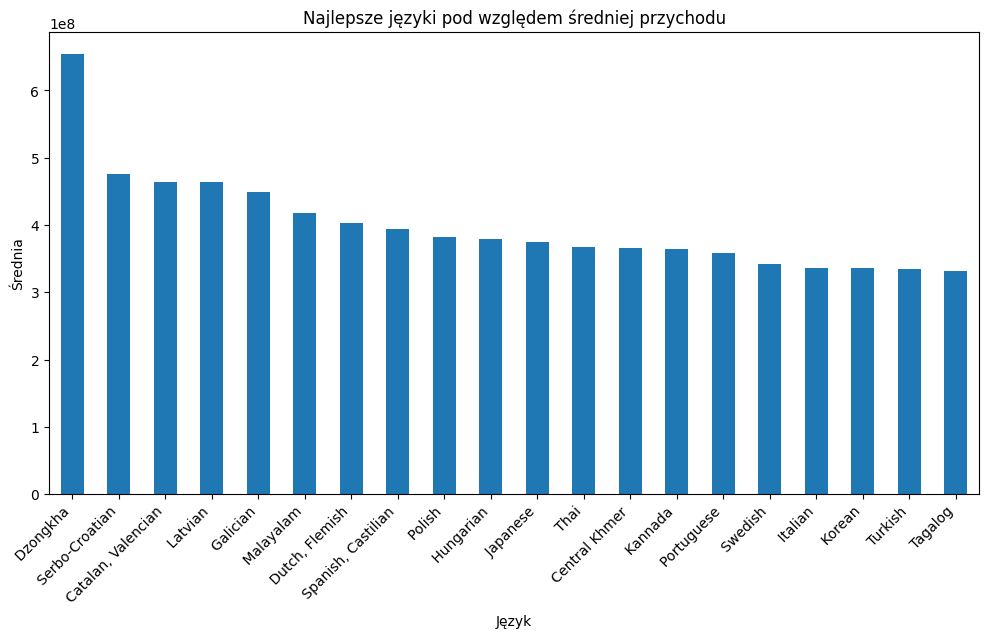
Odrzucamy hipotezę zerową. Istnieje istotna różnica w średnich przychodach między japońskim a angielskim.

Średnia różnica zysków między krajami: 114.146377 mln

Różnica między medianami zysków: 187.574124 mln

Istnieje istotna różnica między tymi językami, jednak co ciekawe, to język japoński osiąga lepsze wyniki. Oznacza to, że filmy produkowane w języku angielskim pomimo posiadania wielu filmów, które osiągają największe zyski posiadają jeszcze więcej filmów, których zyski są bardzo małe i zaniżają średni zysk dla całego języka. Z filmów w języku japońskim, możemy wnioskować, że nie ma aż tak dużej liczby filmów, które przynoszą znikome zyski, przez co średnia jest znacząco większa.

Możemy jeszcze ocenić jak wygląda średni zysk dla pozostałych języków.



Możemy dostrzec, że w czołówce znajdują się nieoczywiste języki, jednak wynika to z niewielkiej liczności filmów w tym języku w naszej bazie – np. Dzonghka 1 film, Catalan 2 filmy itd. Z krajów z większa liczbą filmów można wyróżnić te w języku hiszpańskim, japońskim, portugalskim , koreańskim i włoskim . Możemy wnioskować z tego wykresu, że filmy powstające w tych językach często odnoszą duży sukces i nie posiadają wielu filmów, których wyniki finansowe zaniżają średnią.

Kolejnym pytaniem badawczym na jakie chcielibyśmy odpowiedzieć jest, czy filmy powstałe po 2000 roku dostają coraz lepsze oceny? Stawiamy hipotezę badawczą, że tak – wydaje nam się, że coraz nowsze filmy są coraz lepszej jakości, więc powinny być coraz lepiej ocenianie.

Wykorzystamy do oceniania prawdziwości ponownie test T-studenta. Jako hipotezę H0 przyjmujemy, że średnie ocen nie różnią się między grupami (nowsza i starsza), a jako hipotezę alternatywną przyjmujemy, że są istotnie różne. Za poziom istotności ponownie przyjmujemy alfa = 0.05.

Test Statistic: 8.50399087862706

P-value: 2.089047679871826e-17

Odrzucamy hipotezę zerową. Istnieje istotna różnica w ocenach między starszymi a nowszymi filmami.

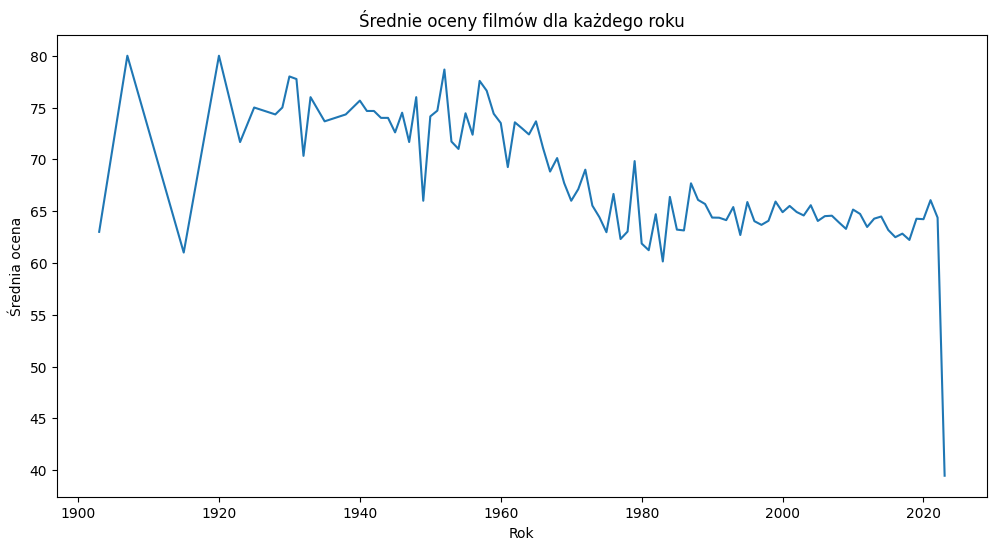
Obliczamy statystki opisowe:

Średnia starych filmów 65.54632721202003

Średnia nowych filmów 62.86610125931637

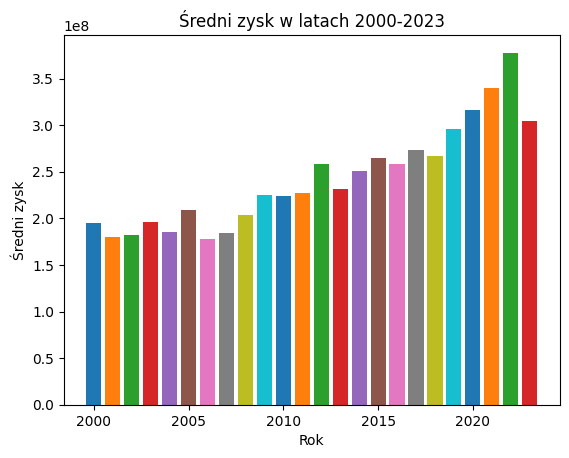
Mediana starych filmów 67.0

Mediana nowych filmów 65.0



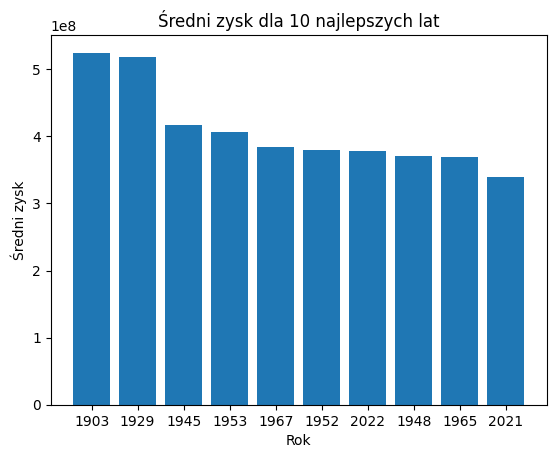
Łatwo zauważyć wyraźny spadek ocen filmów na przestrzeni ostatnich lat. Oceniamy, że może to wynikać z powodu coraz większych wymagań widzów, problemów z tworzeniem oryginalnych filmów i powtarzaniu wzorców, czy chociażby presji finansowej, która zmusza twórców do stawiania zysku na pierwszym miejscu przy tworzeniu materiałów.

Ciekawą sprawą w naszych danych wydaje się drastyczny spadek ocen w 2023. Postanawiamy się temu bliżej przyjrzeć.

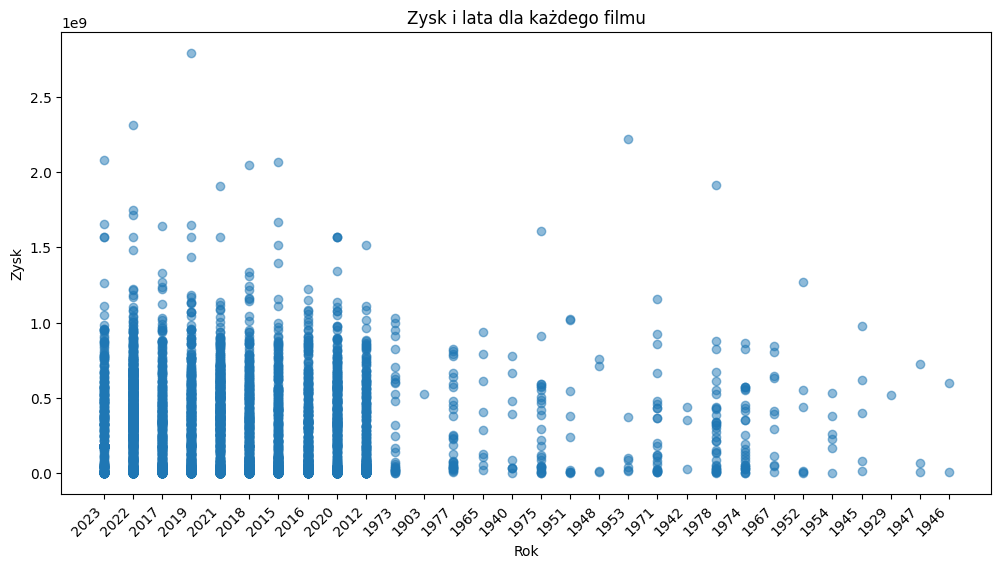


Widzimy, że wraz ze spadkiem ocen znacząco spadł zysk, który do tej pory miał tendencje wzrostową w ostatnich latach, co może oznaczać spadek jakości w tym roku i wzrost niezadowolenia widzów. Zmiana naszym zdaniem jest niewielka w porównaniu do skali spadku ocen, jednak nie jesteśmy w stanie określić jej dokładnego powodu – być może to po prostu specyfika wybranych w tym roku danych, które trafily do naszej bazy.

Postanawiamy jeszcze sprawdzić jakie były największe średnie zyski i w których latach miały miejsce.



Zyski mogą wynikać z ilości filmów jakie zostały wyprodukowane i tak samo jak w przypadku krajów może się okazać , że wynik jest zawyżony przez to, że tylko kilka filmów w danym roku zostały wypuszczonych. Zwizualizujemy dane na szerszym zakresie biorąc 30 najlepszych lat.



Nasza teza się potwierdza, spośród 10 najlepszych lat tylko w 2022 i 2021 roku filmów wyszło dużo i statystyka nie została zawyżona tylko kilkom filmami z największym sukcesem.

Ostatnim pytaniem badawczym jakie postawiliśmy to zbadanie, czy oceny zależą od gatunku filmu. Weźmiemy pod uwagę dwa skrajnie różne gatunki – komedie i horrory. Przewidujemy, że będą istotne różnice między tymi wynikami, ponieważ komedie przez lata były najpopularniejszym gatunkiem filmowym, a horrory stosunkowo niedawno zaczęły zyskiwać na popularności. W związku z tym wydaje nam się, że biorąc pod uwagę spadek wysokich opinii w ostatnich latach, może to jednoznacznie wskazać na różnicę między nimi. Do oceny wykorzystamy test rangowy sum Wilcoxona – gdzie za H0 przyjmujemy, że nie ma różnicy, a za hipotezę alternatywna przyjmujemy, że jest taka różnica. Ustalamy alfa = 0.05. Z racji tego, że dane w kolumnie „genre” są często przedstawiane jako lista gatunków (jeden film może należeć do wielu gatunków) skorzystamy z funkcji „grepl”, które będzie sprawdzać , czy występuje wybrana przez nas fraza.

Wyniki testu:

##

## Wilcoxon rank sum test with continuity correction

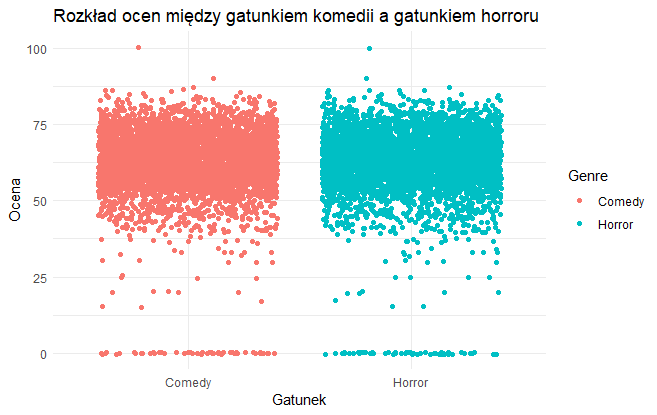
##

## data: comedy\_scores and horror\_scores

## W = 2948371, p-value < 2.2e-16

## alternative hypothesis: true location shift is not equal to 0 w jaki sposob zinterpretowac te wyniki

Niska wartość p-value daje nam podstawy do odrzucenia H0 i stwierdzenia, że mamy podstawy do stwierdzenia, że istnieje istotna różnica między nimi.



Liczba horrorów wynosi 1554, a liczba komedii 2943. Średnia dla komedii jest 63.87632, a mediana 65. Dla horrorów średnia wynosi 59.76126 a mediana 60.

# Bibliography

<https://www.kaggle.com/datasets/ashpalsingh1525/imdb-movies-dataset>