SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSAW SCHOOL OF ECONOMICS

Analiza zbioru odcinków 50 najpopularniejszych seriali na stronie IMDB.com

Projekt zaliczeniowy na zajęcia z przedmiotu:

Prezentacja i wizualizacja danych

Autor

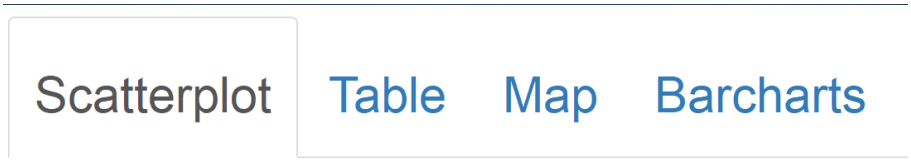
Viet Anh Aleksander Trinh 64364

Cel

Przedmiotem analizy jest zbiór odcinków 50 najczęściej ocenianych seriali na stronie IMDB. Zdecydowałem się dokonać analizy tego zbioru, ponieważ interesuje się tematyką seriali zagranicznych i dość często przeglądam stronę IMDB.com w poszukiwaniu nowego serialu do obejrzenia. Za pomocą danych staram się wywnioskować, jakie czynniki wpływają na to, że dane seriali i ich odcinki są lepiej i częściej oceniane przez użytkowników IMDB. Dane z witryny zostały własnoręcznie przez mnie ściągnięte, przy pomocy scrapera, który napisałem w języku Python. Oprócz tego jedna zmienna została dodana ręcznie.
Zbiór składa się z 12 różnych zmiennych przy czym jedna z nich została rozbita na 3, co miało na celu ułatwić dalsze operacje na zbiorze. Lista zmiennych:
Series- zmienna jakościowa, zawiera nazwy 50 seriali.
Ep.name- zmienna jakościowa, zawiera nazwy każdego z odcinków
Season- zmienna jakościowa zawiera numery sezonów każdego z seriali
Ep.Number- zmienna jakościowa, zawiera numer każdego odcinka serialu
Ep.Rating- zmienna ilościowa, dotyczy oceny poszczególnych odcinków serialu (ocena w skali 1-10)
Length- zmienna ilościowa, w niej zawarte są długości poszczególnych odcinków
Year- zmienna jakościowa, zawiera lata premiery odcinków
Votes- zmienna ilościowa, zawiera liczbę oddanych głosów na każdy z odcinków
Genre.1, Genre2, Genre3- zmienne jakościowe które opisują gatunek serialu, W przypadku Genre2 i Genre3 pojawiają się puste pola, gdyż, niektóre seriale były zaliczane do mniej niż dwóch gatunków
Series.Rating- zmienna ilościowa, zawiera ogólną ocenę serialu na stornie IMDB (ocena w skali 1-10)
TV-Rating- zmienna jakościowa, dotyczy ograniczenia wiekowego dla serialu, możliwe wartości: TV-MA- dla dorosłych, TV-14- dla osób powyżej 14 lat, TV-PG-nieodpowiednie dla małych dzieci, TV-Y7-FV, dla dzieci powyżej 7 roku życia
Location- zmienna jakościowa, zawiera lokalizacje w którym odbywała się akcja serialu
Wszystkie zmienne, oprócz zmiennej Location pochodzą ze strony IMDB.com, natomiast zmienna Location została ręcznie pozyskana ze strony wikipedia.com

Opis aplikacji

Wykresy użyte w tym opracowaniu zostały wykonane przy pomocy aplikacji bazującej na pakiecie Shiny. Aplikacja, którą stworzyłem składa się z 4 głównych komponentów: interaktywnego wykresu punktowego, który umożliwia analizę zależności pomiędzy poszczególnymi zmiennymi. Zakres analizowanych obserwacji można odpowiednio ustalać. Do dyspozycji użytkownika są różne filtry, które pozwalają na odpowiednie dobranie zmiennych i obserwacji do konkretnej analizy. Kolejnym składnikiem aplikacji jest tabela, która pozwoli użytkownikowi na wgląd w dane, posortowanie ich, wyszukanie konkretnych obserwacji. Następnie kolejną częścią aplikacji jest mapa, również w pewnym stopniu interaktywna, natomiast nie jest modyfikowalna jak wykres punktowy z poprzedniej zakładki. Ostatnią częścią aplikacji jest wykres słupkowy, jest on statyczny, natomiast można wybierać zmienne do wizualizacji i filtrować dane.



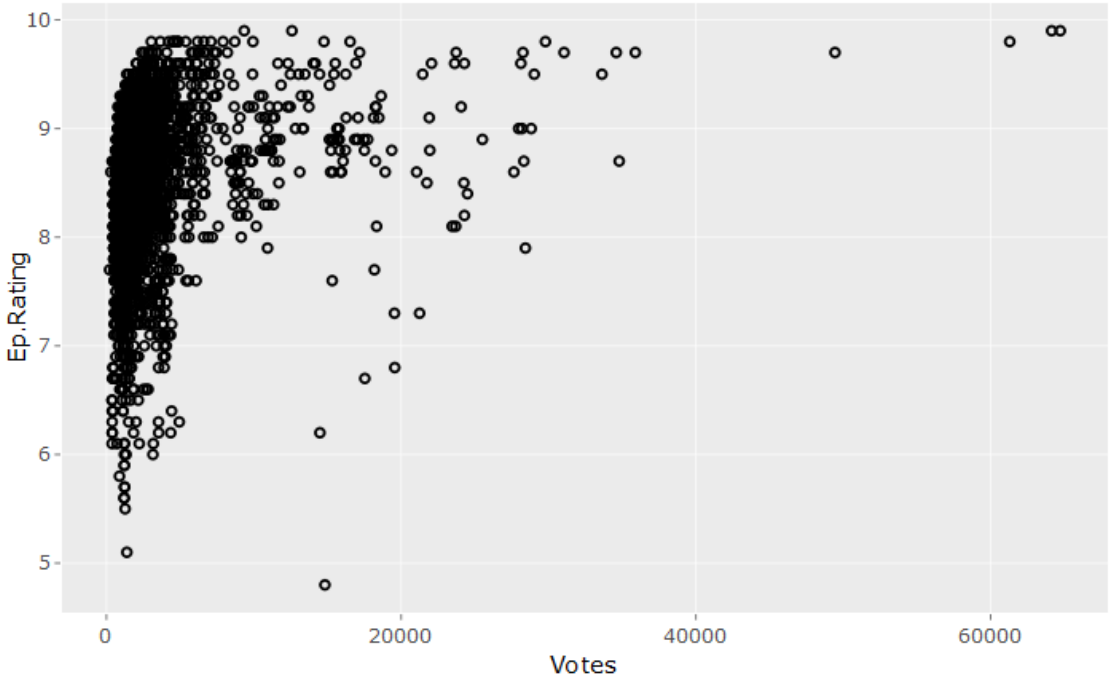
Zrzut ekranu przedstawia pasek zakładek w aplikacji

Analiza zbioru

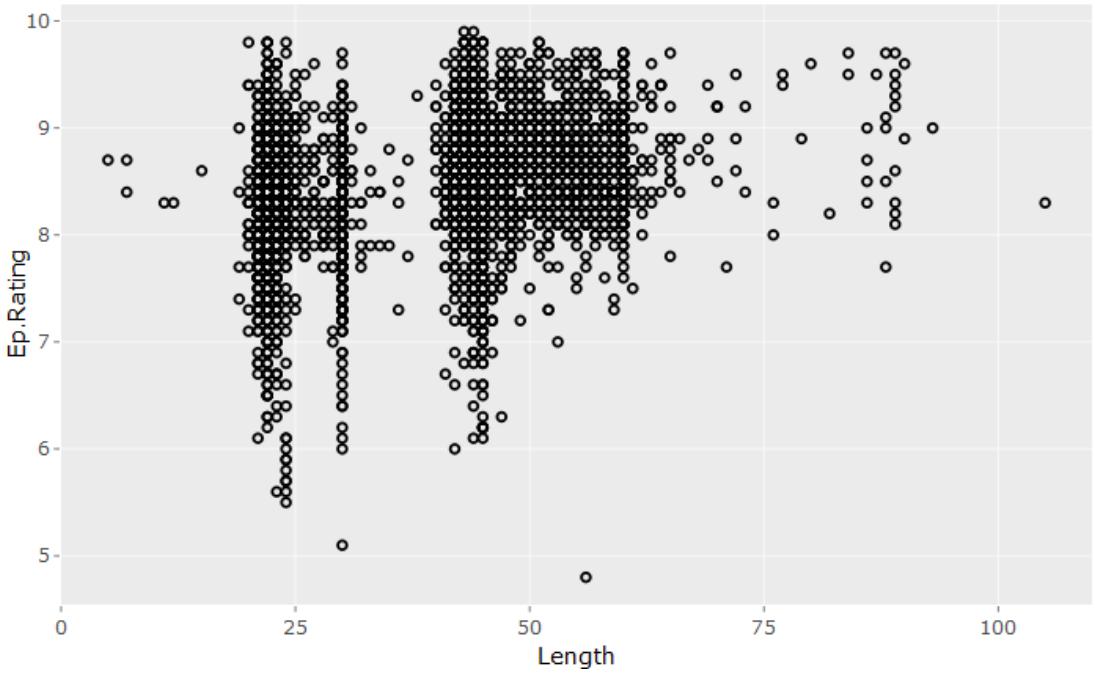
Na bazie stworzonej aplikacji w programie R przy użyciu pakietu Shiny dokonuje następujących analiz zbioru. Na początku przy pomocy filtrów liczby głosów sprawdzam występowanie outlierów.

Widać, że przykładowo odcinek pt. „Battle of the Bastards” serialu Gra o Trony, a także “Ozymandias” z Breaking Bad mają zdecydowanie więcej głosów niż pozostałe odcinki. Redukując w filtrze liczbę głosów to niższej wartości będzie można uzyskać bardziej przejrzysty wykres seriali i zależności pomiędzy liczbą głosów a oceną. Na poniższym

wykresie widać pewną korelacje pomiędzy liczbą głosów, a oceną odcinka. Im więcej ocen tym wyższa zdaje się być ta ocena. Świadczy to najpewniej o tym, że najbardziej popularne odcinki są pozytywnie oceniane przez użytkowników IMDB, ludzie chętniej odwiedzają stronę odcinka, który się im spodobał i wystawiają wysoką ocenę.

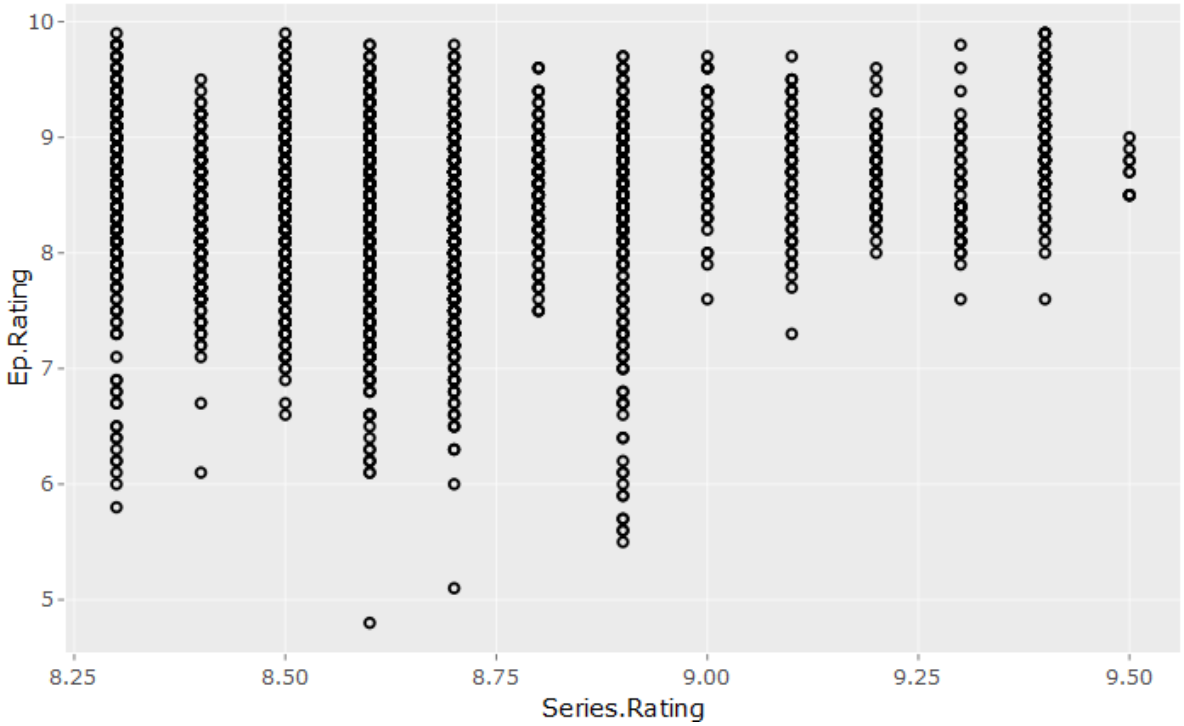


Kolejną zależnością, która chcę zbadać jest zależność czasu trwania odcinka i jego oceny.

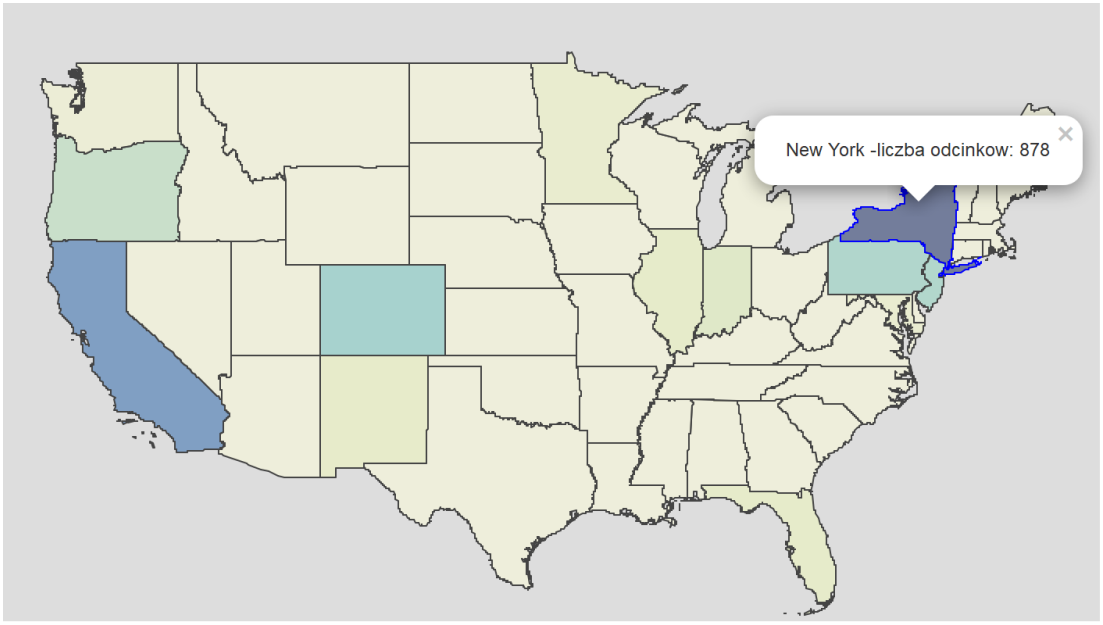


Z powyższego wykresu wynika, że nie występuje zbyt oczywista korelacja pomiędzy tymi zmiennymi. Można zauważyć natomiast, że większość odcinków ma długość pomiędzy 20 a

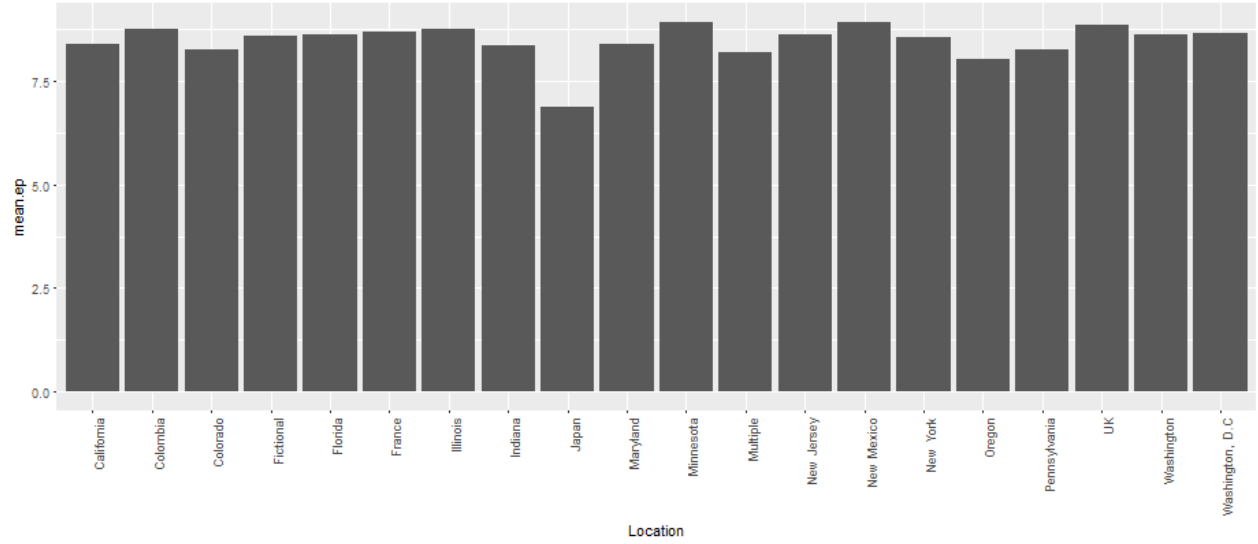
30, lub są dłuższe niż 40 minut, przy czym znowu występuje mała ilość odcinków dłuższy niż godzina. Odcinki dłuższe niż godzina mają nieco wyższe oceny. Najczęściej takie dłuższe odcinki są finałami sezonu/serialu i są one lepszej jakości niż inne odcinki. Krótsze odcinki są najczęściej serialami komediowymi i jak widać na wykresie są one niżej oceniane.
Kolejny wykres służy do przenalizowania zależności oceny wystawionej serialowi ogółem przez użytkownika na stornie IMDB i ocen odcinków. Trzeba zaznaczyć, że ocena ogólna to nie jest średnia ocen z odcinków tylko osobno wystawiana serialowi.



Widać tutaj, że seriale oceniane wyżej czyli ponad 9+, nie miały praktycznie żadnego odcinka z oceną poniżej 7. Widać, że seriale z niższymi ocenami nie mają dość sporo odcinków o niskich ocenach, tzn. prawdopodobnie są to odcinki zapychające, nie wnoszącego niczego ciekawego dla widza. Z drugiej strony, nie ma na dobrą sprawę znaczenia czy dany serial jest wysoko oceniany czy nie i może pojawić się dość wysoko notowany odcinek z oceną ponad 9.5.
Oprócz badania korelacji między danymi zmiennymi chciałbym dowiedzieć się, czy lokacja serialu wpływa jakoś na jego popularność. Aby się tego dowiedzieć stworzę mapę na której zwizualizuje liczbę seriali, których akcja dzieje się w danym miejscu. Jako, lokacja to głownie Stany Zjednoczone lub miejsce fikcyjne posłużę się w tym celu mapą stanów w USA.

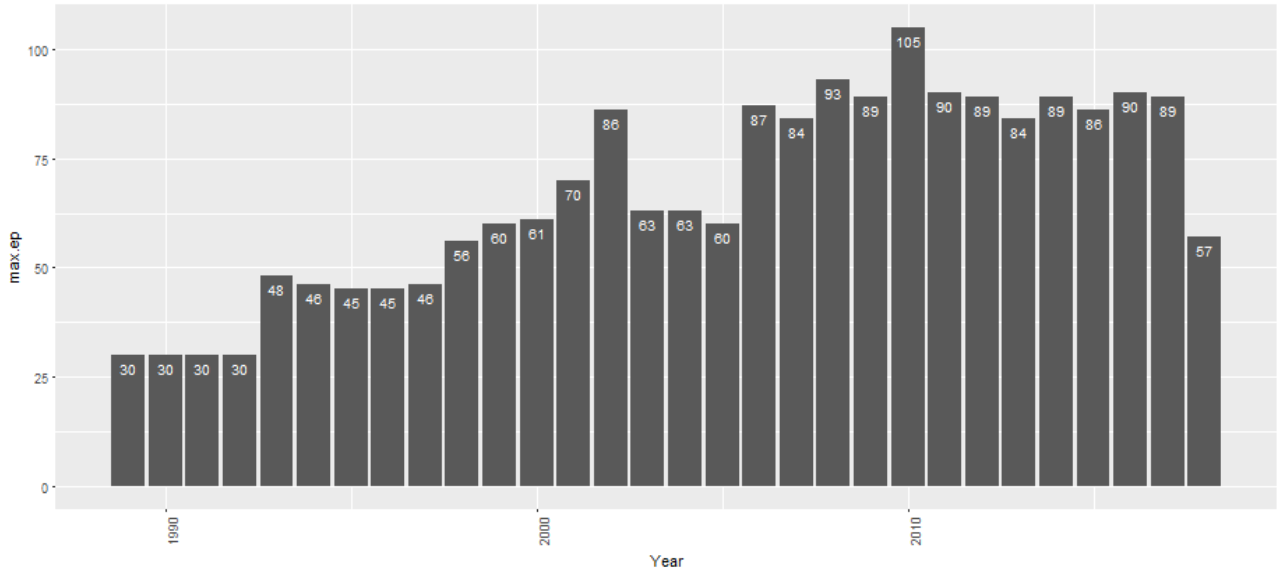


Jak widać, najwięcej jest odcinków w których akcja odgrywała się w Nowym Jorku. Druga pod tym względem jest Kalifornia. Można zatem stwierdzić, że użytkownicy częściej oceniali seriale których fabuła była oparta na wydarzeniach w tych dwóch stanach. W sumie to nic dziwnego gdyż to dwa najbardziej medialne stany w USA. Analizując jeszcze lokalizacje, policzę średnią ocenę odcinków dla każdej z lokalizacji

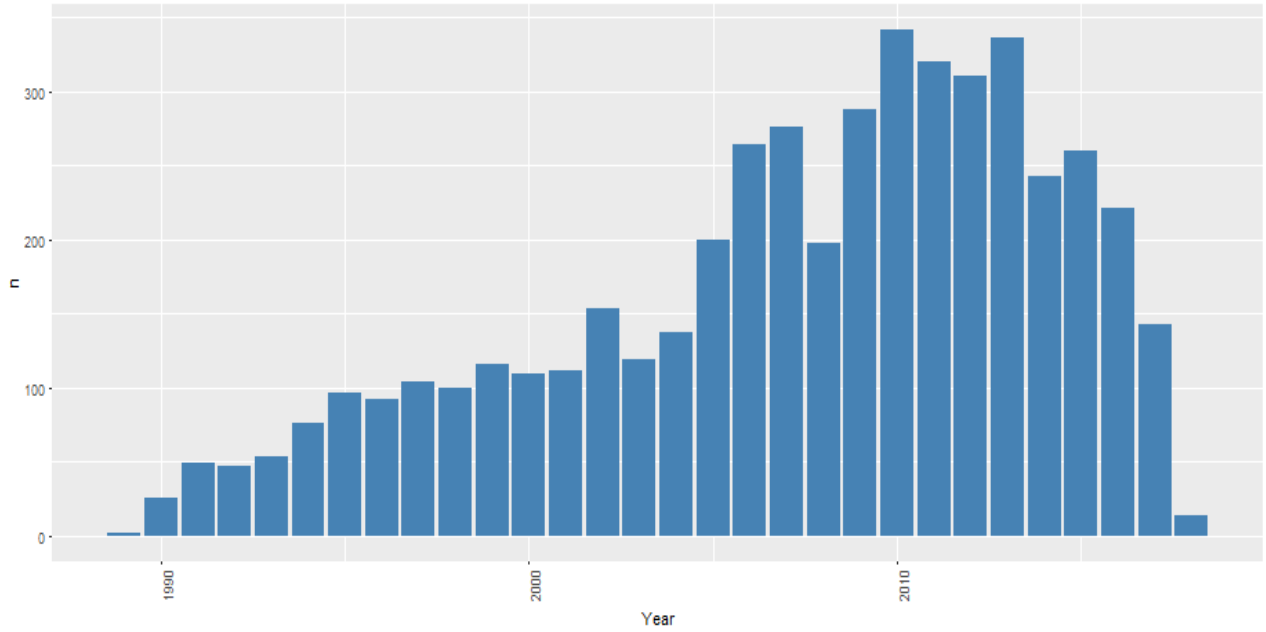


Najlepiej oceniane są seriale, których akcja toczy się w Minnesocie. California również uzyskuje wyższe noty, natomiast Nowy Jork odznacza się niecą niższa średnią oceną, co może świadczyć, ze mimo dużej liczby odcinków w tej lokalizacji, nie wszystkie były one dobrej jakości.

Na koniec chciałbym przeanalizować wpływ czasu na zmianę długości odcinków.



Na wyżej przedstawionym wykresie widać, że wraz z kolejnymi latami odcinki stawały się dłuższe. Być może w końcu producenci mieli odpowiednie środki na produkcje dłuższych epizodów niż dotychczas. Lub też widzowie zaczęli preferować seriale, które nie kończą się szybko, tylko po 30 minutach. Innym wnioskiem, który się nasuwa, jest niska popularność seriali, wyprodukowanych przed latami dziewięćdziesiątymi. Aby sprawdzić, że istnieje zależność czasu od liczby popularnych seriali posłużę się poniższym wykresem.



Jak widać dużo więcej popularnych seriali powstało po 2005 roku. Starsze seriale cieszą się zdecydowanie mniejszym zainteresowaniem ze strony użytkowników IMDB. W liście top 50 seriali nie ma żadnego serialu wyprodukowanego przed rokiem 1989, a odcinki z początku lat 90 to głównie jeden serial (X-Files).

Wnioski

Na podstawie powyższych wykresów można wyciągnąć kilka ciekawych wniosków. Zakładając, że ocena odcinka wyznacza jego poziom, to wysoka liczba ocen oznacza, że dany odcinek był bardzo dobry. Ponadto dłuższe odcinki z reguły dostawały wyższe noty, przy czym sugeruje to dodatkowo, że seriale komediowe, cieszą się mniejszym powodzeniem wśród użytkowników IMDB. Kolejną rzeczą jest to, że wysoko oceniane seriale nie mają pojedynczych odcinków o niskich ocenach. Najlepsze dzieła rzadko zawodzą swoich widzów. Jako, że IMDB to storna amerykańska i głównie korzystają z niej obywatele USA świadczyć to może o preferencji serialów, które opowiadają historie o życiu w USA. Użytkownicy IMDB to są zapewne osoby młodsze, millenialsi, którzy preferują nowsze seriale, a zarazem jakość seriali mogła się poprawić na przestrzeni ostatnich lat.