POLITECHNIKA GDAŃSKA | WYDZIAŁ FIZYKI TECHNICZNEJ I MATEMATYKI STOSOWANEJ

Imię i nazwisko słuchacza studiów podyplomowych: **Artur Karpiński**

Kierunek studiów: **Inżynieria Danych – Data Science**

**PROJEKT** (własny) z Politechniki Gdańskiej

Tytuł pracy w języku polskim: **Analiza danych filmowych na platformie Netflix**

Tytuł pracy w języku angielskim: **Analysis of movie data on the Netflix platform**

## Wstęp

**Projekt dotyczy analizy danych filmowych na platformie Netflix**.

Projekt oparty jest na wybranych danych dotyczących Systemu Rekomendacji Filmów.
- tytuły filmów
- czas wydania, najważniejsze gatunki filmów
- oceny i daty jej wystawienia

Cel projektu nie jest związany z wszystkimi pobranymi danymi. Nie wykorzystuję indywidualnych danych związanych z użytkownikami. Nie chodzi o rekomendowanie, reklamowanie tytułów widzom na podstawie zebranych o nich danych, prefencji.

Najpierw przeprowadziłem prognozowanie wyników w przedziale czasu obejmującym zebrane dane (interpolacja).
Potem prognozowałem przyszłe wyniki (do 2030 r.) na podstawie zebranych danych (ekstrapolacja).
Badania oparłem na regresji wielomianowej. Dla znajdowania wartości pośrednich w obecnych czasach wykorzystałem metodę interpolacji. Prognozowanie wyników w przyszłości oparłem na ekstrapolacji.

Projekt oparty został na otwartych danych z platformy Kaggle.

**System rekomendacji filmów** (poniżej opis zbioru wszystkich danych)
AUTOR zbioru zebranych danych (z Kaggle): Bandi Karthik
**Movie Recommendation System**

# Spis treści

**Wstęp** ..... 1
**Analiza danych** ..... 4
1. Przegląd danych zebranych o filmach platformy ..... 4
2. Popularność gatunków w czasie ..... 5
**Podział kategorii filmów na pojedyncze gatunki** ..... 5
**Rozwój gatunków na przestrzeni czasu** ..... 6
3. Oceny filmów ..... 7
**Badanie ocen filmów** ..... 7
**Najgorsze filmy** ..... 8
**Najlepsze filmy** ..... 9
4. Najbardziej popularne filmy ..... 10
5. Ocena a popularność ..... 10
**Porównanie 2 najpopularniejszych gatunków** ..... 11
**Zmiana ocen filmów w przedziałach czasu** ..... 12
**Zmiana popularności w przedziałach czasu** ..... 12
6. Seria „Star Wars: Episode" ..... 12
7. Zastosowanie regresji wielomianowej (w interpolacji i ekstrapolacji) ..... 14
**„Star Wars: Episode 1" w interpolacji** ..... 14
**Wszystkie dane Netflix w ekstrapolacji** ..... 15
**Porównanie wyników na podstawie regresji wielomianowej** ..... 17
**Zakończenie** ..... 17
**Bibliografia** ..... 17
**Załączniki** ..... 17

## Analiza danych

Ten zbiór danych opisuje działania związane z **oceną pięciogwiazdkową** oraz **tagowaniem dowolnym tekstem** w usłudze **polecania filmów**.
Zawiera 23 mln ocen i 590 tys. tagów w 34 tys. filmów. Te dane zostały utworzone przez 250 tys. użytkowników między 9 stycznia 1995 a 29 stycznia 2016.
Użytkownicy zostali wybrani losowo do włączenia. Wszyscy ocenili co najmniej 1 film.
Dane są zawarte w czterech plikach:
**movies**.csv – filmy (dane o filmach – tytuł, gatunek, rok wydania),
**ratings**.csv – oceny (ocena, data głosu),
**links**.csv – linki (połączenia, relacje danych),
**tags**.csv – tagi (etykiety do danych o użytkownikach).

**Filmy z platformy Netflix.**
Przygotowanie i analiza danych.
Dane z publicznej domeny kaggle.
Dotyczą filmów z platformy Netflix (m.in. oglądalność, oceny).
Wykorzystałem dane dwóch plików:

1. movies – informacje dotyczące filmów (tytuł, gatunek, rok wydania)
2. ratings – informacje dotyczące ocen tych filmów (ocena, data głosu)

### 1. Przegląd danych zebranych o filmach platformy

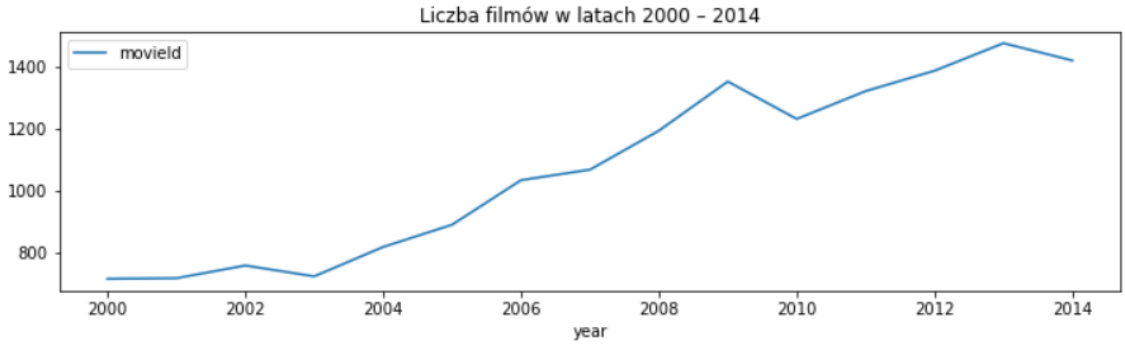
**Przykładowe dane na temat filmów**

| title | genres | year |
|---|---|---|
| Toy Story | [adventure, animation, children, comedy, fantasy] | 1995 |
| Jumanji | [adventure, children, fantasy] | 1995 |
| Grumpier Old Men | [comedy, romance] | 1995 |
| Waiting to Exhale | [comedy, drama, romance] | 1995 |
| Father of the Bride Part II | [comedy] | 1995 |

**Tab. Tytuły, gatunki, lata produkcji przykładowych filmów**

**Liczba filmów w poszczególnych latach**
Wszystkie dane do roku 2014


**Rys. Liczba wszystkich filmów do 2014 roku**




**Rys. Liczba filmów od 2000 do 2014 roku**

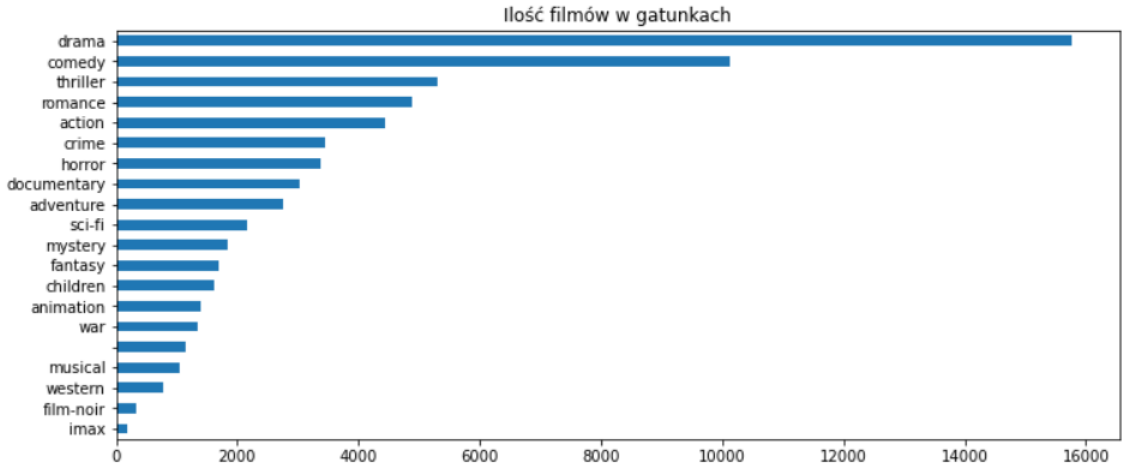
## 2. Popularność gatunków w czasie

### Podział kategorii filmów na pojedyncze gatunki

Ilość filmów w poszczególnych gatunkach (w sposób  rosnący)

| | |
|---|---|
| drama | 15774 |
| comedy | 10124 |
| thriller | 5300 |
| romance | 4875 |
| action | 4445 |
| crime | 3446 |
| horror | 3365 |
| documentary | 3040 |
| adventure | 2763 |
| sci-fi | 2156 |
| mystery | 1837 |
| fantasy | 1692 |
| children | 1609 |
| animation | 1387 |
| war | 1345 |
| | 1145 |
| musical | 1052 |
| western | 779 |
| film-noir | 338 |
| imax | 196 |

**Tab. Ilość filmów w poszczególnych gatunkach**




**Rys. Ilość filmów w poszczególnych gatunkach**

## Często filmy są przydzielone do kilku kategorii

| | |
|---|---|
| 1 | 14372 |
| 2 | 10719 |
| 3 | 6432 |
| 4 | 2023 |
| 5 | 537 |
| 6 | 98 |
| 7 | 21 |
| 8 | 5 |
| 10 | 1 |

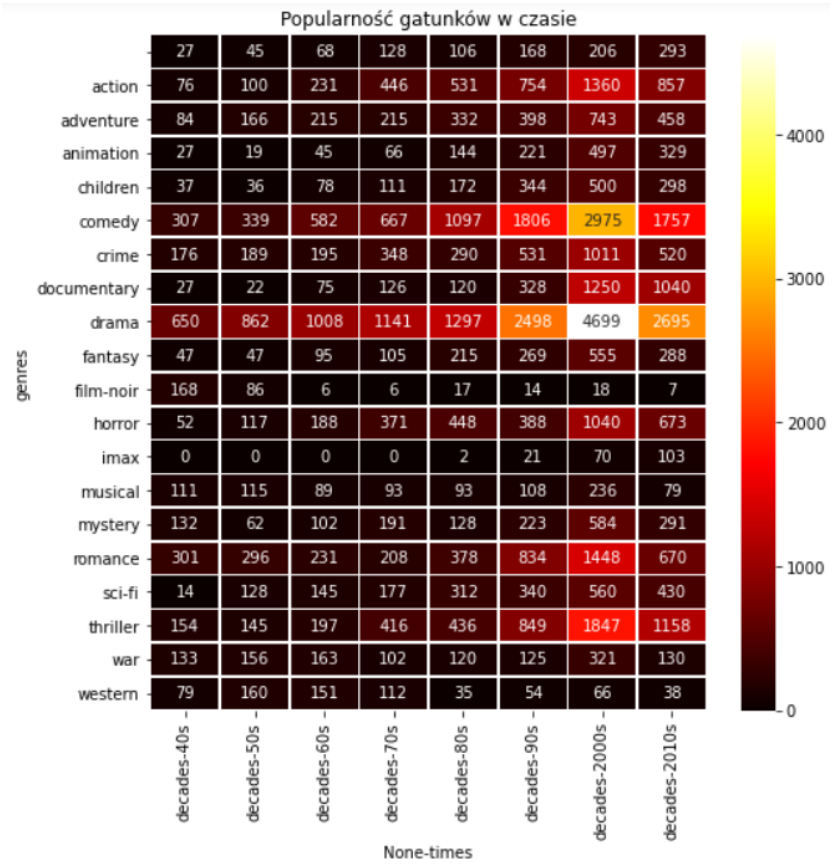
**Tab. Liczba gatunków w wielu kategoriach**


**Rys. Liczba gatunków w wielu kategoriach**

## Rozwój gatunków na przestrzeni czasu

liczba filmów w poszczególnych dekadach

przedział dla lat 1940-2020 co 10 lat




**Rys. Popularność gatunków w czasie, dekadach**

## 3. Oceny filmów

### Badanie ocen filmów

Widzowie oceniali dodając ilość punktów, gwiazdek.

oceny między 0.5 oraz 5.0

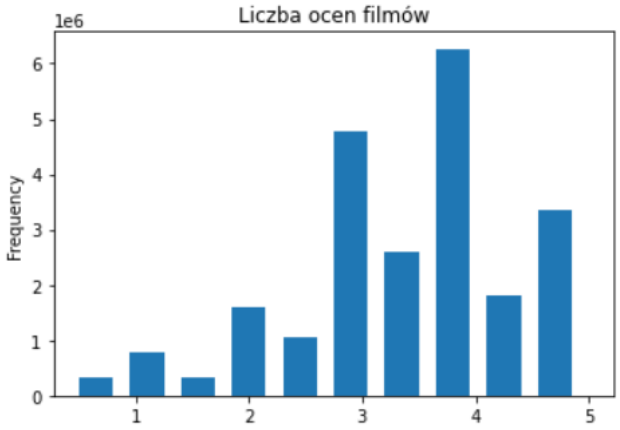
| movieId | rating | timestamp |
|---|---|---|
| 169 | 2.5 | 2008-03-07 22:08:14 |
| 2471 | 3.0 | 2008-03-07 22:03:58 |
| 48516 | 5.0 | 2008-03-07 22:03:55 |
| 2571 | 3.5 | 2015-07-06 06:50:33 |
| 109487 | 4.0 | 2015-07-06 06:51:36 |

**Tab. Przykładowe oceny w czasie**

**Liczba ocen w poszczególnych przedziałach ocen**

| | |
|---|---|
| 4.0 | 6265623 |
| 3.0 | 4783899 |
| 5.0 | 3358218 |
| 3.5 | 2592375 |
| 4.5 | 1813922 |
| 2.0 | 1603254 |
| 2.5 | 1044176 |
| 1.0 | 769654 |
| 1.5 | 337605 |
| 0.5 | 315651 |

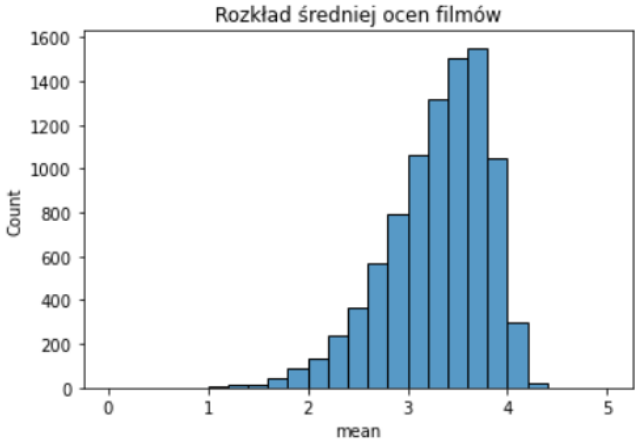
**Tab. Liczba ocen filmów**




**Rys. Liczba ocen filmów**

Rozkład nie jest równomierny, a średnia między 3 i 4.

### Rozkład (przegląd) średnich ocen dla wszystkich filmów




**Rys. Rozkład średniej ocen filmów**

## Najgorsze filmy

Najgorsze filmy z lewej strony wykresu.

Dla nich wartości mniejsze niż w skrajnym kwantylu (od lewej).

Ze związku między mean i std widać, że oceny skrajnie negatywne.

| movieId | mean | std |
|---|---|---|
| 1323 | 1.619639 | 0.999614 |
| 1495 | 1.403794 | 1.082217 |
| 1599 | 1.677817 | 1.071550 |
| 1739 | 1.574730 | 1.087333 |
| 1826 | 1.197917 | 0.930694 |

**Tab. Średnia i odchylenie najgorszych filmów**

### Poniżej lista najgorzej ocenianych filmów

sortowanie względem średniej i w sposób rosnący

| title | genres | year | mean | std |
|---|---|---|---|---|
| SuperBabies: Baby Geniuses 2 | [comedy] | 2004 | 0.851598 | 0.788178 |
| From Justin to Kelly | [musical, romance] | 2003 | 0.985491 | 0.845644 |
| Glitter | [drama, musical, romance] | 2001 | 1.135344 | 0.861708 |
| Gigli | [comedy, crime, romance] | 2003 | 1.170715 | 0.918029 |
| Barney's Great Adventure | [adventure, children] | 1998 | 1.197917 | 0.930694 |

**Tab. Lista najgorszych filmów**

Najgorszym był " SuperBabies: Baby Geniuses 2".

## Najlepsze filmy

Dla najlepszych filmów analogicznie.

Zmiana nierówności i skrajnych kwantyli (od prawej).

Najlepsze filmy były z prawej strony wykresu.

Ze związku między mean i std widać, że oceny skrajnie pozytywne.

| movieId | mean | std |
|---|---|---|
| 50 | 4.318987 | 0.770791 |
| 260 | 4.158052 | 0.958329 |
| 296 | 4.163211 | 0.984206 |
| 318 | 4.441710 | 0.733828 |
| 527 | 4.290952 | 0.849629 |

**Tab. Średnia i odchylenie najgorszych filmów**

### Poniżej lista najlepiej ocenianych filmów

sortowanie względem średniej i w sposób malejący

| title | genres | year | mean | std |
|---|---|---|---|---|
| Shawshank Redemption, The | [crime, drama] | 1994 | 4.441710 | 0.733828 |
| Godfather, The | [crime, drama] | 1972 | 4.353639 | 0.854273 |
| Usual Suspects, The | [crime, mystery, thriller] | 1995 | 4.318987 | 0.770791 |
| Schindler's List | [drama, war] | 1993 | 4.290952 | 0.849629 |
| Godfather: Part II, The | [crime, drama] | 1974 | 4.268878 | 0.875126 |

**Tab. Lista najlepszych filmów**

Najlepszym był "The Shawshank Redemption".

### Średnia ocen

3.2983740005180495

## 4. Najbardziej popularne filmy

**Liczba głosów najpopularniejszych filmów** (w sposób malejący)

| movieId | #votes |
|---|---|
| **356** | 81296 |
| **296** | 79091 |
| **318** | 77887 |
| **593** | 76271 |
| **480** | 69545 |

**Tab. Głosy najpopularniejszych filmów**

**Poniżej lista najpopularniejszych filmów**

sortowanie względem liczby głosów w sposób malejący

| #votes | movieId | title | genres | year |
|---|---|---|---|---|
| 81296 | 356 | Forrest Gump | [comedy, drama, romance, war] | 1994 |
| 79091 | 296 | Pulp Fiction | [comedy, crime, drama, thriller] | 1994 |
| 77887 | 318 | Shawshank Redemption, The | [crime, drama] | 1994 |
| 76271 | 593 | Silence of the Lambs, The | [crime, horror, thriller] | 1991 |
| 69545 | 480 | Jurassic Park | [action, adventure, sci-fi, thriller] | 1993 |

**Tab. Lista najpopularniejszych filmów**

Najpopularniejszym był "Forrest Gump".

## *5.* Ocena a popularność

Nie były to najwyżej oceniane filmy lecz z największą liczbą głosów.

Nie wiemy czy najlepiej oceniane są te najbardziej popularne.

Czyli jednym się podobało, innym nie, ale ogólnie dobre oceny, popularność.

Sortowanie przez odchylenie ' std ' daje pojęcie o spójności ocen.

| #votes | movieId | title | genres | year | mean | std |
|---|---|---|---|---|---|---|
| 230 | 74754 | Room, The | [comedy, drama, romance] | 2003 | 2.400000 | 1.767860 |
| 158 | 1311 | Santa with Muscles | [comedy] | 1996 | 2.598101 | 1.760735 |
| 133 | 59295 | Expelled: No Intelligence Allowed | [documentary] | 2008 | 2.075188 | 1.620765 |
| 141 | 70946 | Troll 2 | [fantasy, horror] | 1990 | 2.092199 | 1.615331 |
| 104 | 50703 | Secret, The | [documentary] | 2006 | 2.673077 | 1.597585 |

**Tab. Filmy popularne dobrze oceniane**

**Średnia ocen**

3.2983740005180495

## Porównanie 2 najpopularniejszych gatunków

sortowanie względem liczby głosów w sposób malejący

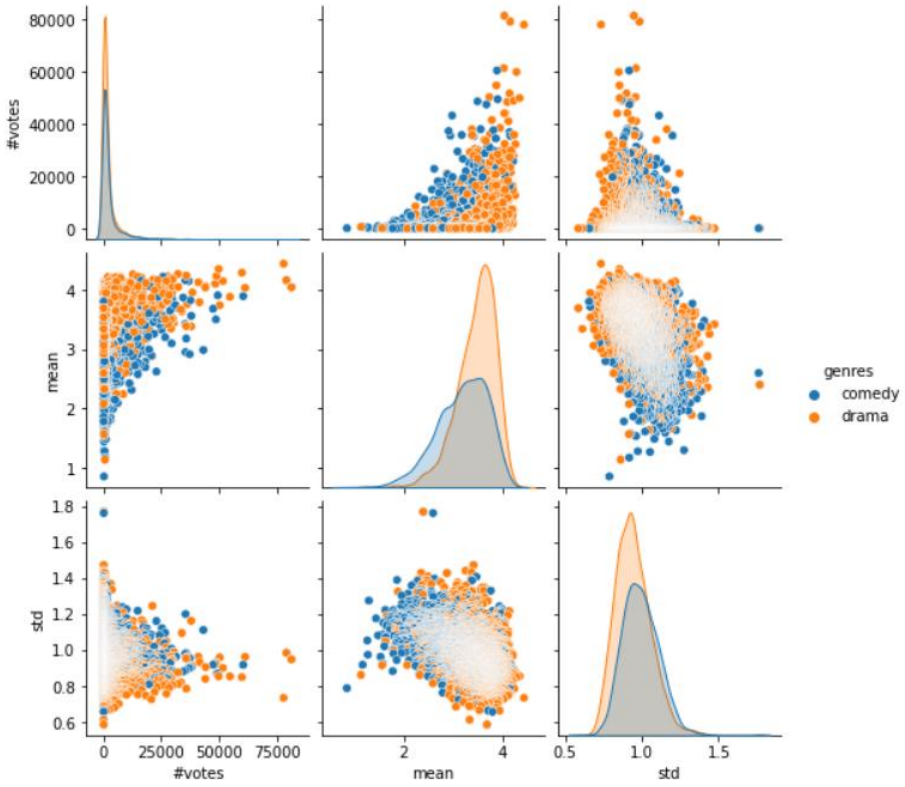
| #votes | mean | std | genres |
|---|---|---|---|
| 230 | 2.400000 | 1.767860 | comedy |
| 230 | 2.400000 | 1.767860 | drama |
| 158 | 2.598101 | 1.760735 | comedy |
| 137 | 3.419708 | 1.471836 | comedy |
| 137 | 3.419708 | 1.471836 | drama |

**Tab. Porównanie komedii i dramatów**

Komedie i dramaty tak samo popularne.

Jednak ocena dramatu statystycznie wyższa niż komedii.

Oceniający też bardziej spójni przy dramatach (mniejsze std) niż komediach.




**Rys. Porównanie komedii i dramatów**

**Wnioski:**

Według widzów lepiej obejrzeć dobry dramat niż średnią komedię.

Czyli niekoniecznie wysoka popularność jest zgodna z oceną.

Widzowie mogą też zwracać uwagę na różne elementy.

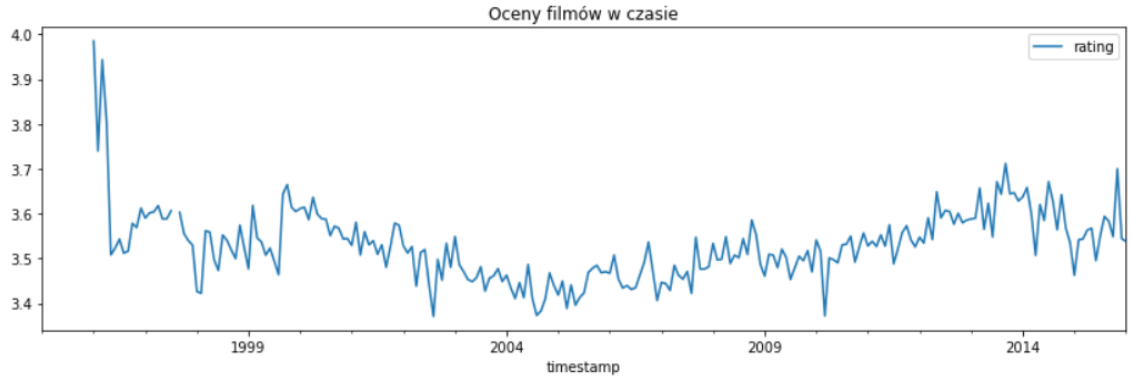
Np. widzowie mają ulubione gatunki filmów, a unikają innych.

### Zmiana ocen filmów w przedziałach czasu

Jak zmieniały się oceny filmów w miesięcznych przedziałach czasu?

rozkład głosów względem czasu za pomocą funkcji ‘mean’.

średnia między 3, a 4



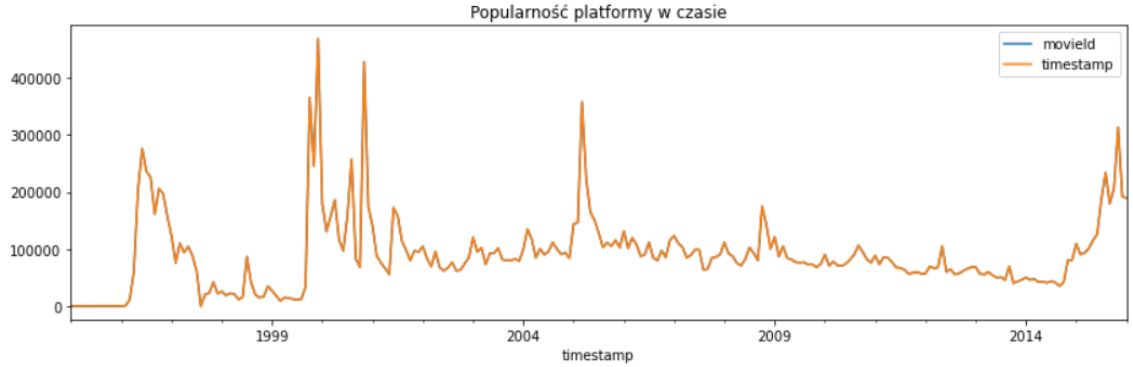

**Rys. Zmiana oceny filmów w czasie**

### Zmiana  popularności w przedziałach czasu

Jak zmieniała się popularność filmów w miesięcznych przedziałach czasu?

rozkład względem liczby głosów w czasie zliczanych za pomocą funkcji ‘count’

mamy film i czas, w którym była ocena




**Rys. Zmiana  popularności platformy w czasie**

## 6. Seria „Star Wars: Episode”

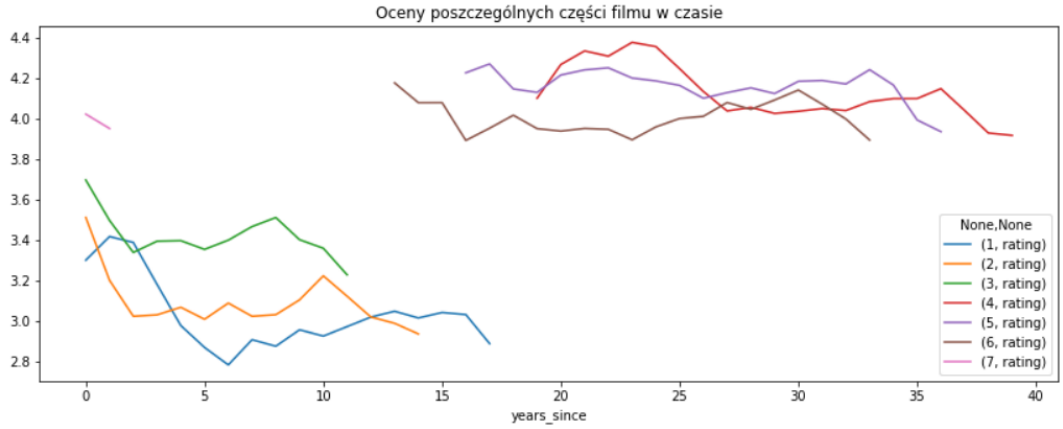
**"Star Wars" – "Gwiezdne Wojny"**

ograniczenie się tylko do serii "Gwiezdne Wojny" (7 części filmów)

| episode | title | year |
|---|---|---|
| 1 | Star Wars: Episode I - The Phantom Menace | 1999 |
| 2 | Star Wars: Episode II - Attack of the Clones | 2002 |
| 3 | Star Wars: Episode III - Revenge of the Sith | 2005 |
| 4 | Star Wars: Episode IV - A New Hope | 1977 |
| 5 | Star Wars: Episode V - The Empire Strikes Back | 1980 |
| 6 | Star Wars: Episode VI - Return of the Jedi | 1983 |
| 7 | Star Wars: Episode VII - The Force Awakens | 2015 |

**Tab. Filmy z serii „Star Wars: Episode”**

Poniżej wykres po dołączeniu i porównaniu ocen z czasem ich wystawienia.




**Rys. Porównanie ocen z czasem ich wystawienia**

Części 4,5,6 powstały najwcześniej (najdawniej).

Części 1, 2, 3, 7 powstały najwcześniej.

| title | year | rating | timestamp |
|---|---|---|---|
| Star Wars: Episode VI - Return of the Jedi | 1983 | 5.0 | 1997-03-26 18:40:32 |
| Star Wars: Episode V - The Empire Strikes Back | 1980 | 5.0 | 2013-05-27 15:10:43 |
| Star Wars: Episode V - The Empire Strikes Back | 1980 | 5.0 | 2008-11-04 21:21:44 |
| Star Wars: Episode V - The Empire Strikes Back | 1980 | 5.0 | 2008-08-02 16:26:28 |
| Star Wars: Episode V - The Empire Strikes Back | 1980 | 5.0 | 2000-01-31 19:36:20 |

**Tab. Oceny filmów z serii „Star Wars: Episode”**

## Wnioski

Dawniejsze były znacznie wyżej oceniane niż te nowsze.

Również dużo bardziej stałe w swojej ocenie.

Nie chodzi w tym porównaniu o ilość lecz wysokość ocen.

## 7. Zastosowanie regresji wielomianowej (w interpolacji i ekstrapolacji)

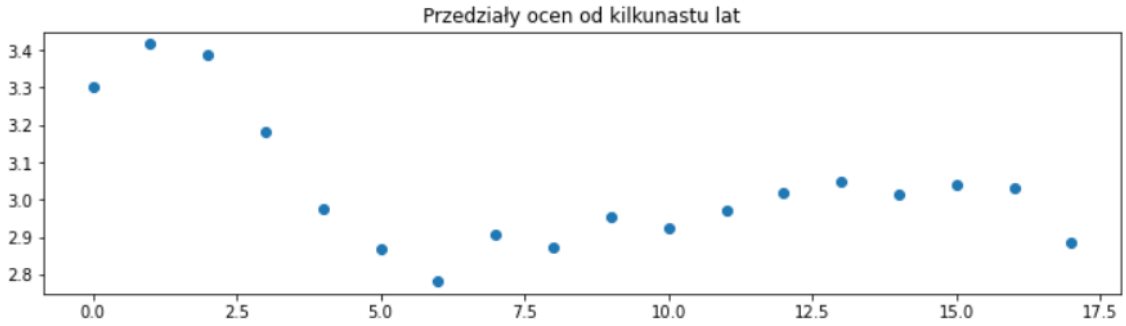
### „Star Wars: Episode 1” w interpolacji

Z **interpolacją** mamy do czynienia, gdy chcemy wyznaczyć wartość y dla konkretnego argumentu x należącego do zakresu istniejących danych X.

Do znajdowania wartości w obecnych czasach, bez prognozowania w przyszłości.

Po dopasowaniu funkcji wyliczamy jej wartości pomiędzy znanymi jej punktami.

Tworzymy wykres oryginalnych danych (kropki), przewidywania czerwoną linią.

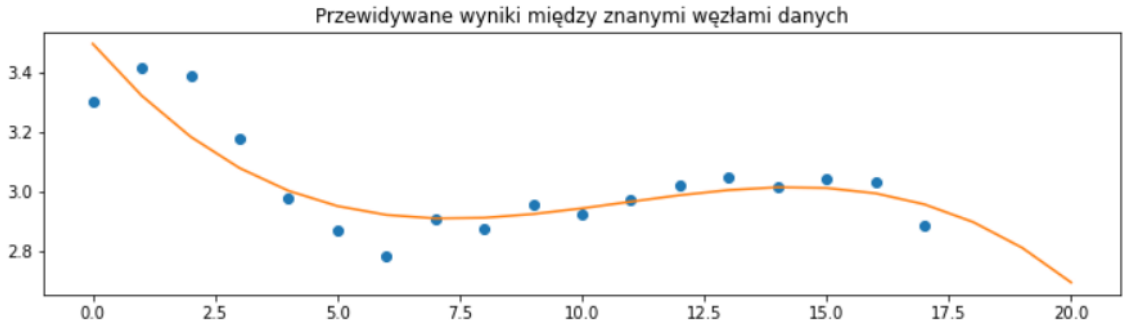



**Rys. Oceny w dotychczasowym przedziale czasu**

Następnie nasze przewidywania na podstawie regresji wielomianowej.

Sprawdzimy optymalny stopień wielomianu (np. 2,3,4).

Na wykresie przewidywane wyniki (linia) między znanymi węzłami danych (kropki).




**Rys. Przewidywane wyniki między znanymi węzłami danych**

Widzimy, że wielomian 3-go stopnia był dobrym dopasowaniem.

Dla 2-go stopnia, to nie mamy dobrego dopasowania (linia zbyt oddalona od punktów).

Dla 4-go stopnia dopasowanie staje się tak dobre, że mało realistyczne.

Nie dałoby się przewidzieć co stałoby się z głosami widzów później.

## Wszystkie dane Netflix w ekstrapolacji

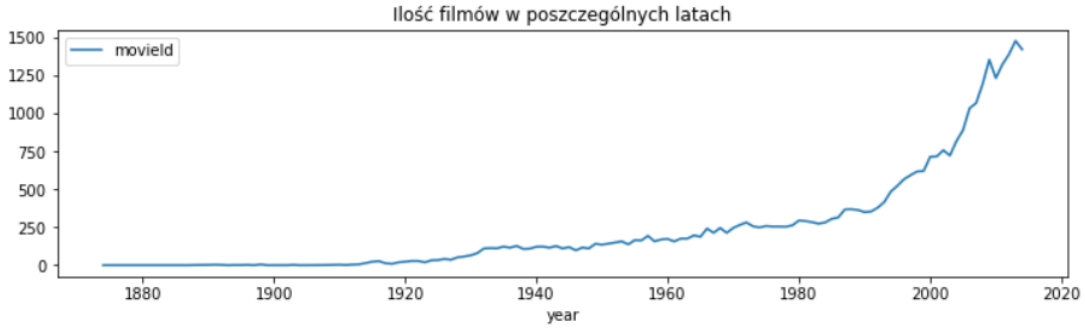
Z **ekstrapolacją** mamy do czynienia gdy chcemy wyznaczyć y dla x nie należącego do zakresu istniejących danych X.

Prognozowanie wartości zmiennej lub funkcji poza zakresem, dla którego mamy dane.

Dopasowanie do istniejących danych pewnej funkcji, następnie wyliczenie jej wartości w szukanym punkcie w przyszłości.

Dobrą praktyką przy ekstrapolacji jest nie podawanie prognoz dla zbyt daleko leżących argumentów.

Poniżej tendencje w formie wykresów.



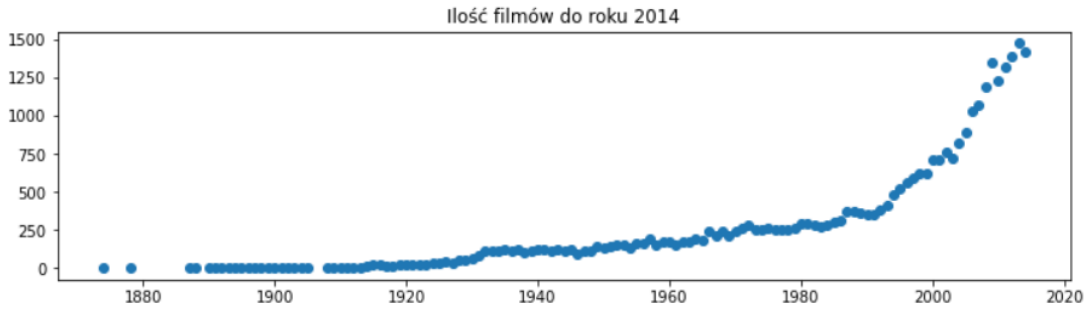

**Rys. Ilość filmów w poszczególnych latach**

Dane ograniczone są do roku 2014 (pełnych), by uniknąć wyników niewiarygodnych.

Można teraz na tych danych dokonać regresji wielomianowej.

Naszymi 'x' będzie rok produkcji, 'y' liczba filmów w konkretnych latach.

Tworzymy wykres oryginalnych danych (kropki).




**Rys. Ilość filmów do roku 2014**

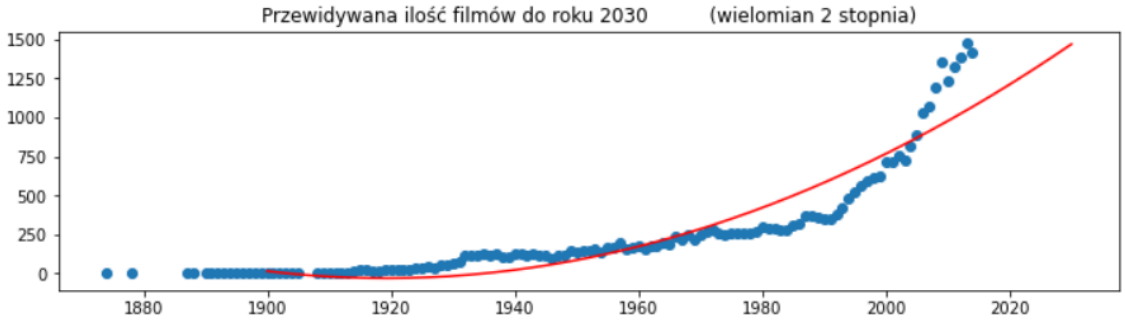
W ekstrapolacji stosujemy algorytm podobny do interpolacji w „Star Wars: Episode 1”.

Zmiana odpowiednich parametrów, wartości zmiennych.

Na wykresach przewidywane wyniki (linia) między znanymi węzłami danych (kropki).

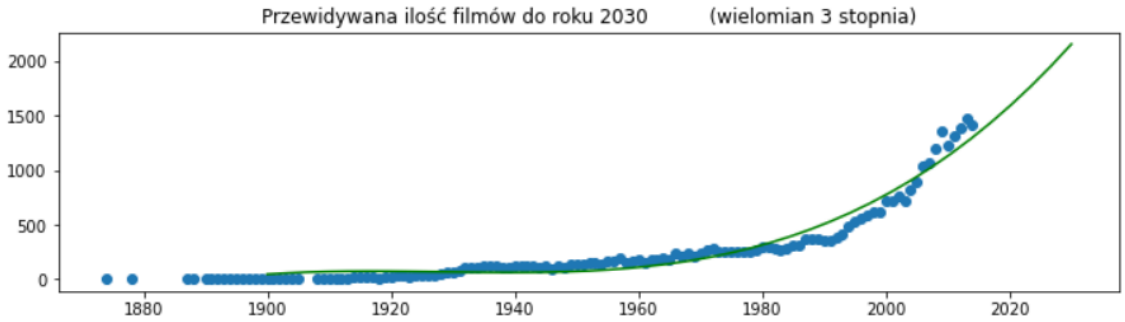
Jak wcześniej sprawdźmy kolejno, czy optymalny wielomian będzie stopnia 2,3,4.

Dla wielomianu 2-go stopnia nie mamy dobrego dopasowania (zbyt odstający).



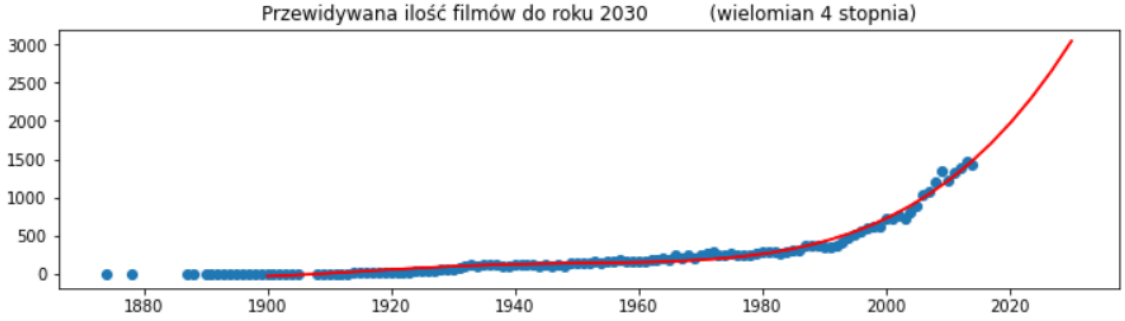

**Rys. Zastosowanie regresji wielomianowej 2 stopnia**

Wielomian 3-go stopnia wydaje się dobry.
Mógłby symulować, w jaki sposób ilość fimów wzrastałaby w przyszłości.
Przewidywania zaznaczone są zieloną linią.




**Rys. Zastosowanie regresji wielomianowej 3 stopnia**

Dla 4-go stopnia dopasowanie staje się tak dobre, że wygląda na mało realistyczne (model nie uczy się).
Przy stopniu wyższym od 3 mniej realne prognozowanie zmian w ilości filmów później.




**Rys. Zastosowanie regresji wielomianowej 4 stopnia**

### Porównanie wyników na podstawie regresji wielomianowej

1. Widzimy, że wielomian 3-go stopnia był dobrym dopasowaniem.
2. Gdy zmienimy na 2-go stopnia, to nie mamy dobrego dopasowania (zbyt odstający).
3. Dla 4-go stopnia (i wyższym) dopasowanie staje się tak dobre, że mało realistyczne.
4. Przy wyższym od 3 trudno przewidzieć co stałoby się z głosami widzów później.

**Mieliśmy 2 przykłady wykorzystania regresji wielomianowej**

Pierwszy - w interpolacji, drugi - w ekstrapolacji.

1. Przy ocenach "Star Wars: Episode 1" oceny można było wykorzystać do znajdowania wartości pośrednich w obecnych czasach (interpolacja).
2. Przy ilości produkowanych filmów, do przewidywania szybkości wzrostu lub spadku w przyszłości (ekstrapolacja).

## Zakończenie

Na podstawie zebranych danych szukałem, analizowałem i pokazywałem zależności między tytułami, gatunkami filmów, ich ocenami i popularnością wśród widzów na przestrzeni czasu.

Kod, wykresy, tabele wykonanywałem w języku Python na notatniku Jupyter Notbook.

## Bibliografia

Dokumentacja biblioteki Numpy: https://numpy.org/doc/stable/

Dokumentacja biblioteki Pandas: https://pandas.pydata.org/docs/

Dokumentacja biblioteki Matplotlib: https://matplotlib.org/stable/index.html

## Załączniki

Załączyłem prezentację, wizualizację projeku z poziomu przeglądarki w postaci notatnika Jupyter Notebook i strony www.