

# **DATAFRAME PROJECT**

**МАТВІЙ ГУРА**

# КОНТЕНТ

- 01** ПРО МЕНЕ ТА МІЙ ДАТАСЕТ
- 02** ДОШКА TRELLO
- 03** ПЕРЕГЛЯД ДАТАСЕТУ
- 04** ОЧИЩЕННЯ ДАТАСЕТУ
- 05** МЕНТАЛЬНА ДОШКА
- 06** ГІПОТЕЗИ ТА ГРАФІКИ
- 07** ЩО НОВОГО Я ДІЗНАВСЯ?
- 08** ЧИМ КОРИСТУВАВСЯ ПРИ СТВОРЕННІ ПРОЕКТУ

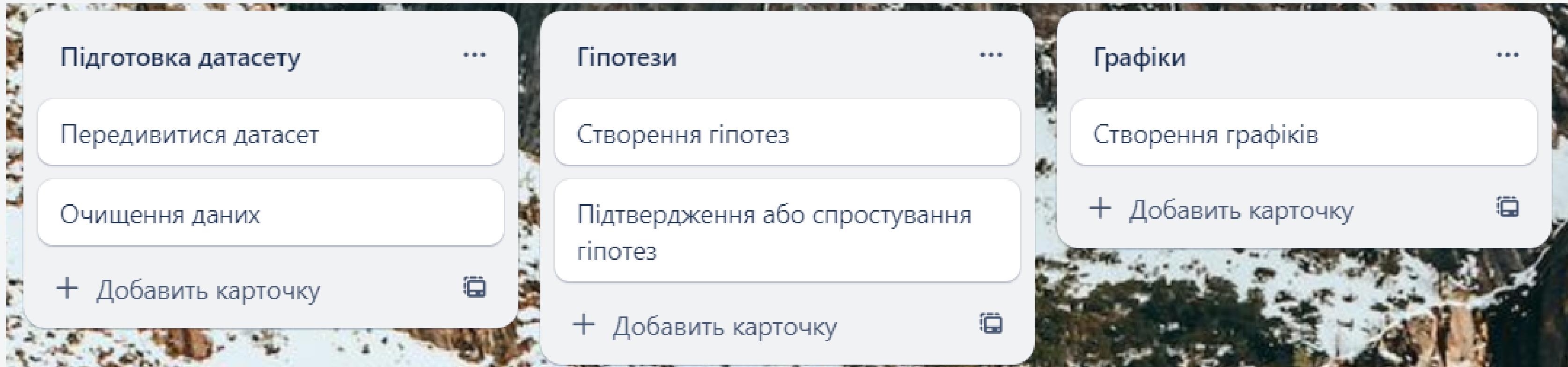
# 1. ПРО МЕНЕ ТА ДАТАСЕТ

Мене звати Матвій, мені 14 років;  
Я цікавлюся ІТ сферою та навчаюсь у школі  
Логіка уже 3-ій рік



Мій датасет – Фільми IMDb – база даних  
фільмів в Інтернеті (Internet Movie Database) –  
спісок художніх фільмів, які отримали  
найвищі середні оцінки користувачів у  
рейтингу сайту

# 2.ДОШКА TRELLO



# 3. ПЕРЕГЛЯД ДАТАСЕТУ

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('IMDB-Movie-Data.csv')
df.info()
print('Від NaN значень неочищенні 2 колонки: 10 і 11')
```

+ Show all

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Rank              1000 non-null    int64  
 1   Title             1000 non-null    object  
 2   Genre             1000 non-null    object  
 3   Description       1000 non-null    object  
 4   Director          1000 non-null    object  
 5   Actors            1000 non-null    object  
 6   Year              1000 non-null    int64  
 7   Runtime (Minutes) 1000 non-null    int64  
 8   Rating            1000 non-null    float64 
 9   Votes              1000 non-null    int64  
 10  Revenue (Millions) 872 non-null    float64 
 11  Metascore         936 non-null    float64 
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
Від NaN значень неочищенні 2 колонки: 10 і 11
```

# 4. ОЧИЩЕННЯ ДАТАСЕТУ

```
print(len(pd.isnull(df['Revenue (Millions)'])))
df['Revenue (Millions)'].fillna(0, inplace = True)
print('Порожні значення заробітку замінено на 0')
```

1000  
Порожні значення заробітку замінено на 0

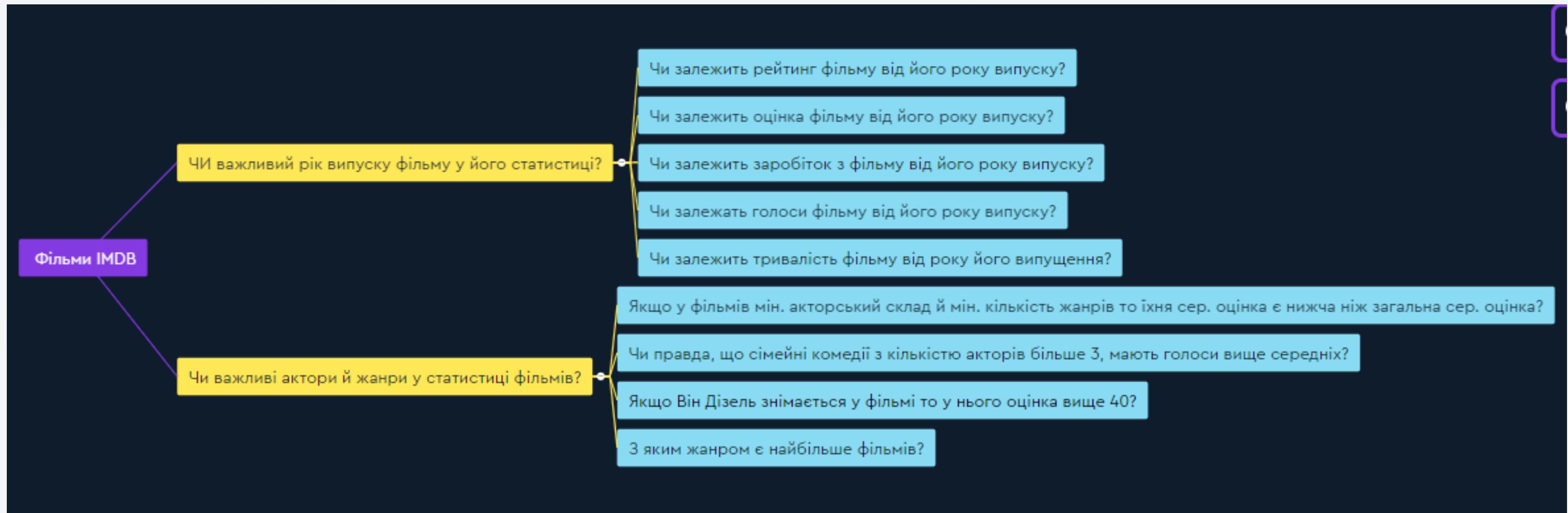
```
print(len(pd.isnull(df['Metascore'])))
df['Metascore'].fillna(df['Metascore'].mean(), inplace=True)
print('Порожні значення оцінки замінено на сер. значення')
```

1000  
Порожні значення оцінки замінено на сер. значення

```
df.info()
print('Як бачимо порожні значення зникли')
```

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   Rank             1000 non-null   int64  
 1   Title            1000 non-null   object  
 2   Genre            1000 non-null   object  
 3   Description      1000 non-null   object  
 4   Director         1000 non-null   object  
 5   Actors           1000 non-null   object  
 6   Year             1000 non-null   int64  
 7   Runtime (Minutes) 1000 non-null   int64  
 8   Rating           1000 non-null   float64 
 9   Votes             1000 non-null   int64  
 10  Revenue (Millions) 1000 non-null   float64 
 11  Metascore        1000 non-null   float64 
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
Як бачимо порожні значення зникли
```

# 5. МЕНТАЛЬНА дошка для створення гіпотез



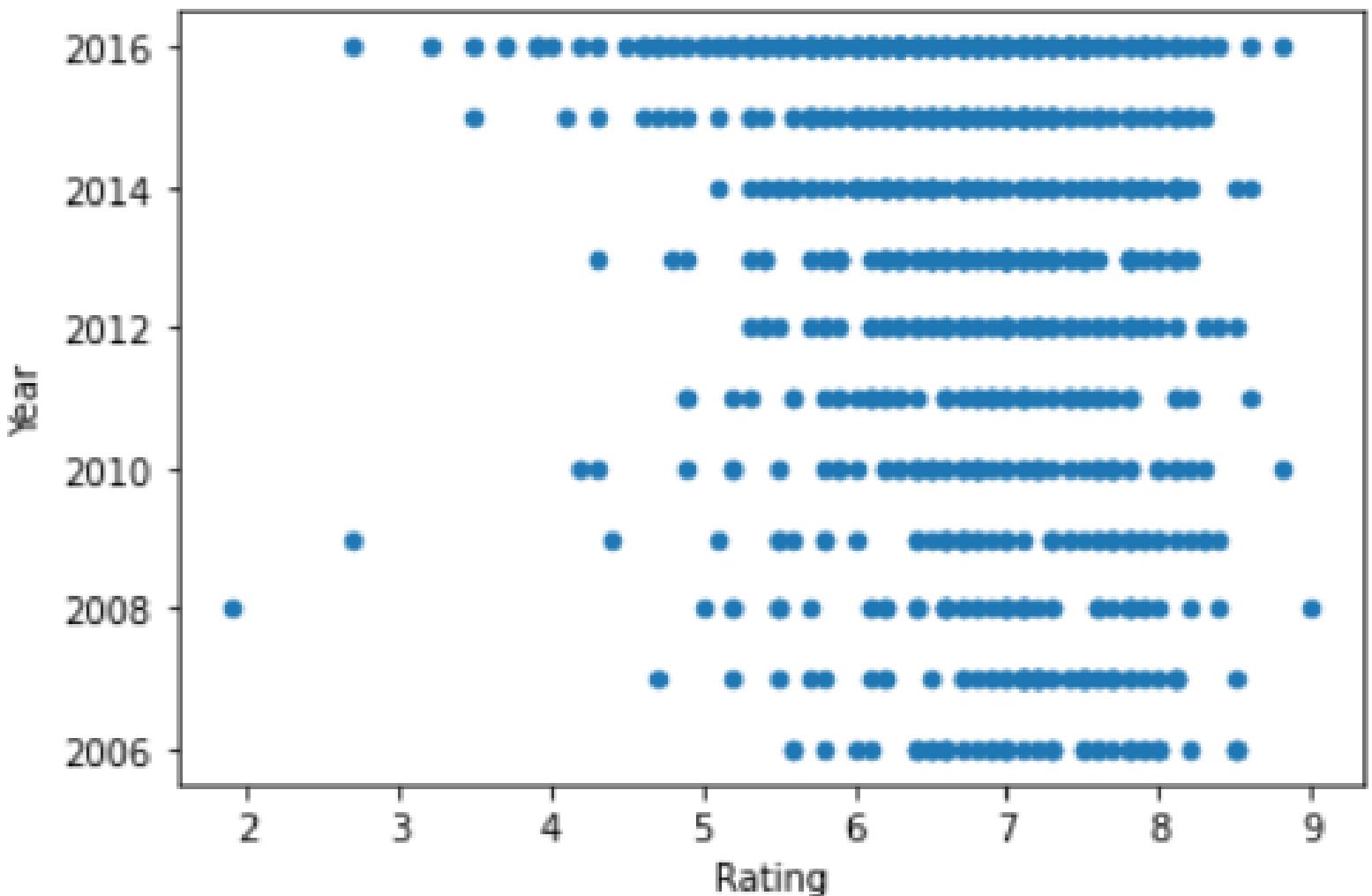
# 6.1 ГІПОТЕЗИ ТА ГРАФІКИ

ЧИ важливий рік випуску фільму у його статистиці?

Чи залежить рейтинг фільму від його року випуску?

```
print("ЧИ ВАЖЛИВИЙ РІК ВИПУСКУ ФІЛЬМУ У ЙОГО СТАТИСТИЦІ")
print("<<Чи залежить рейтингу фільму від року його випущення?>>")
print("Мін. рік фільму:", df["Year"].min())
print("Макс. рік фільму:", df["Year"].max())
print("Сер. рейтинг старих фільмів:", df[df['Year'] <= 2010]['Rating'].mean())
print("Сер. рейтинг нових фільмів:", df[df['Year'] >= 2010]['Rating'].mean())
df.plot(x = 'Rating', y = 'Year', kind = 'scatter')
print("Рейтинг фільмів відносно одинаковий")
```

ЧИ ВАЖЛИВИЙ РІК ВИПУСКУ ФІЛЬМУ У ЙОГО СТАТИСТИЦІ  
<<Чи залежить рейтингу фільму від року його випущення?>>  
Мін. рік фільму: 2006  
Макс. рік фільму: 2016  
Сер. рейтинг старих фільмів: 6.957692307692308  
Сер. рейтинг нових фільмів: 6.654749999999999  
Рейтинг фільмів відносно одинаковий



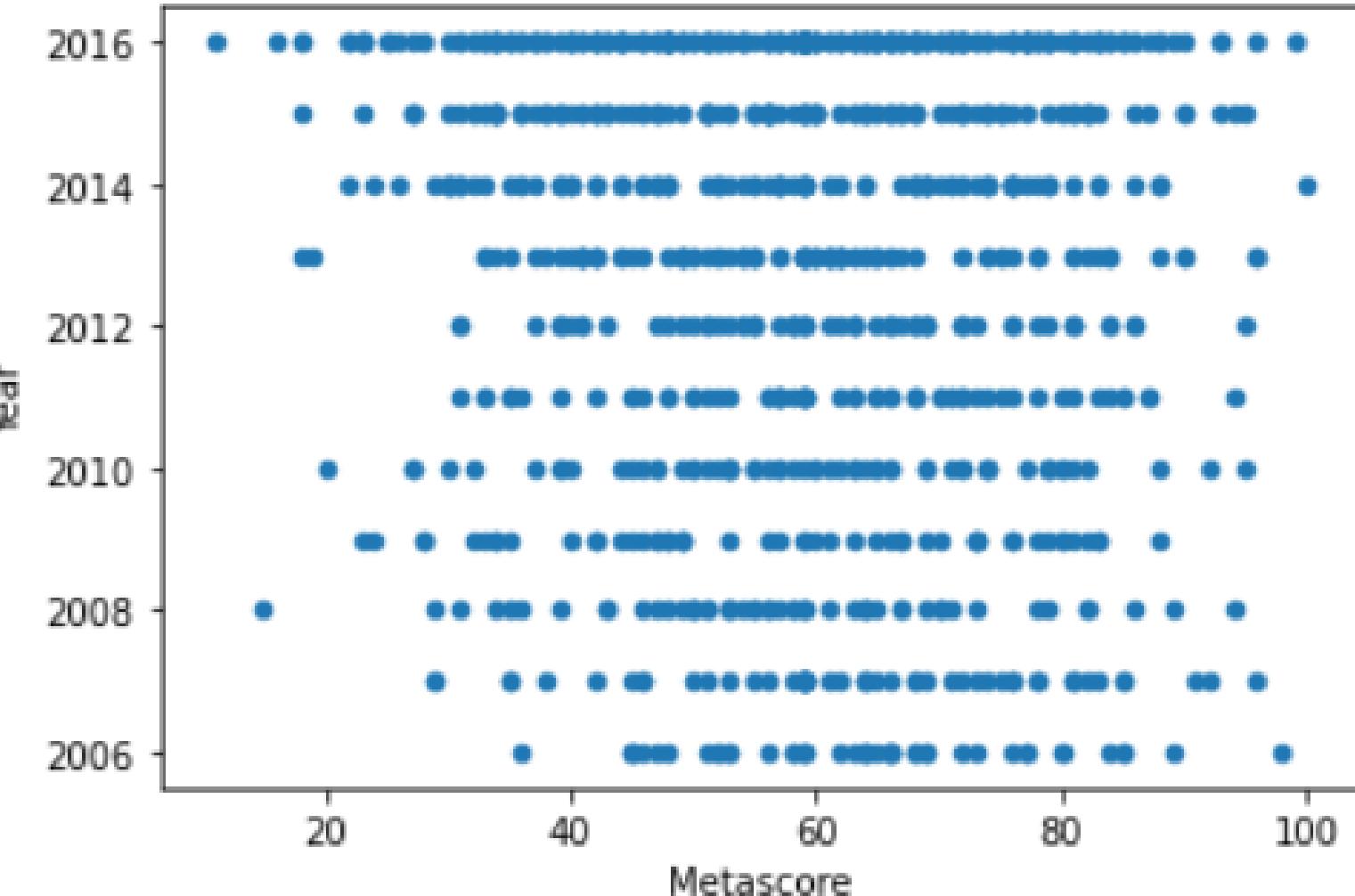
# 6.2 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливий рік випуску фільму у його статистиці?

Чи залежить оцінка фільму від його року випуску?

```
print("<<Чи залежить оцінка фільму від року його випущення?>>")  
print("Сер. оцінка старих фільмів:", df[df['Year'] <= 2010]['Metascore'].mean())  
print("Сер. оцінка нових фільмів:", df[df['Year'] >= 2010]['Metascore'].mean())  
df.plot(x = 'Metascore', y = 'Year', kind = 'scatter')  
print("Оцінка фільмів відносно однакова")
```

```
<<Чи залежить оцінка фільму від року його випущення?>>  
Сер. цінка старих фільмів: 60.26061801446417  
Сер. оцінка нових фільмів: 58.600333867521364  
Оцінка фільмів відносно однакова
```



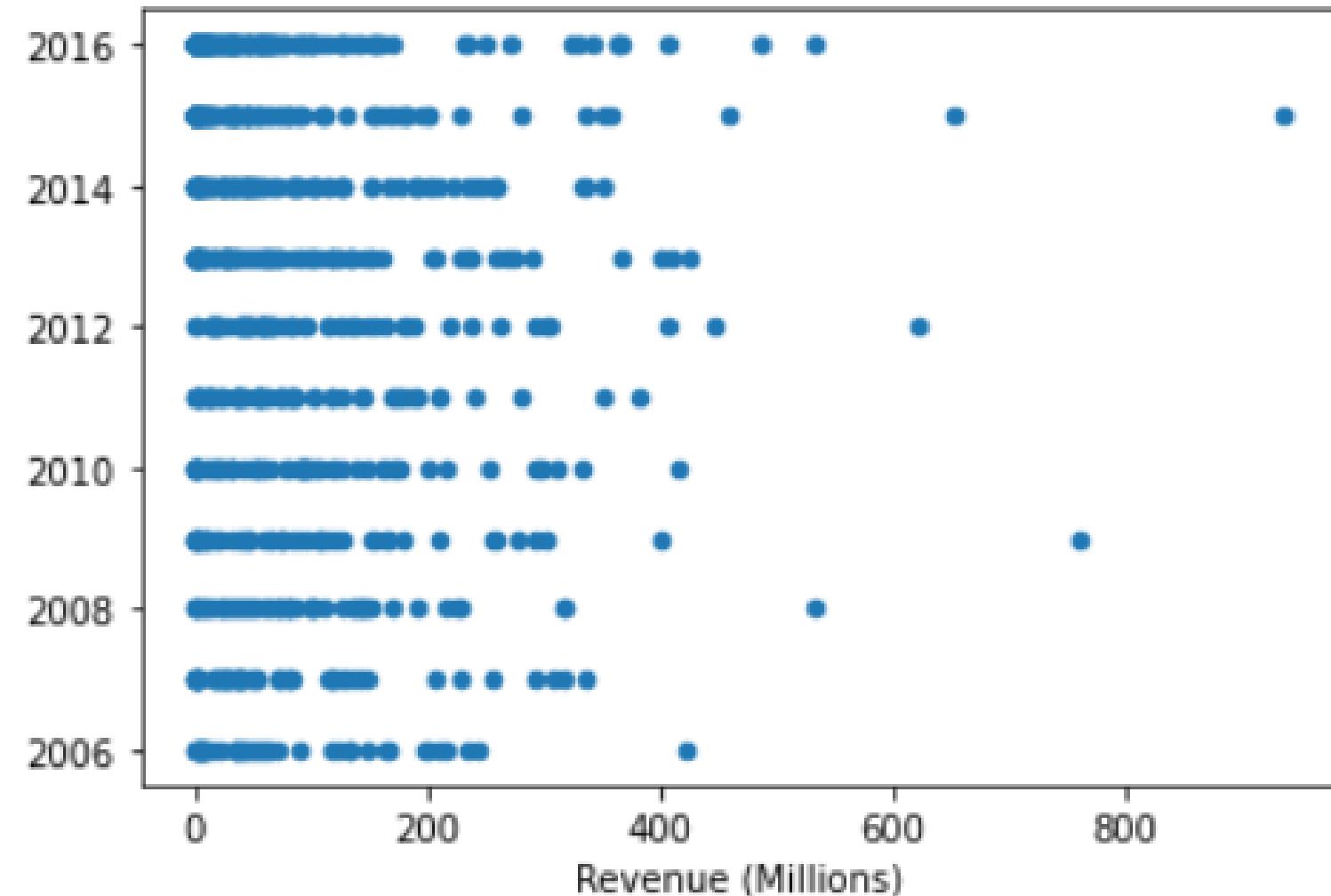
# 6.3 ГІПОТЕЗИ ТА ГРАФІКИ

ЧИ важливий рік випуску фільму у його статистиці?

```
print("<<Чи залежить заробіток з фільму від року його випущення?>>")  
print("Сер. заробіток старих фільмів:", df[df['Year'] <= 2010]['Revenue (Millions)'].mean(), "млн. долларів")  
print("Сер. заробіток нових фільмів:", df[df['Year'] >= 2010]['Revenue (Millions)'].mean(), "млн. долларів")  
df.plot(x = 'Revenue (Millions)', y = 'Year', kind = 'scatter')  
print("Заробіток старих фільмів більший")
```

Чи залежить заробіток з фільму від його року випуску?

```
<<Чи залежить заробіток з фільму від року його випущення?>>  
Сер. заробіток старих фільмів: 93.33007692307692 млн. долларів  
Сер. заробіток нових фільмів: 67.5772375 млн. долларів  
Заробіток старих фільмів більший
```



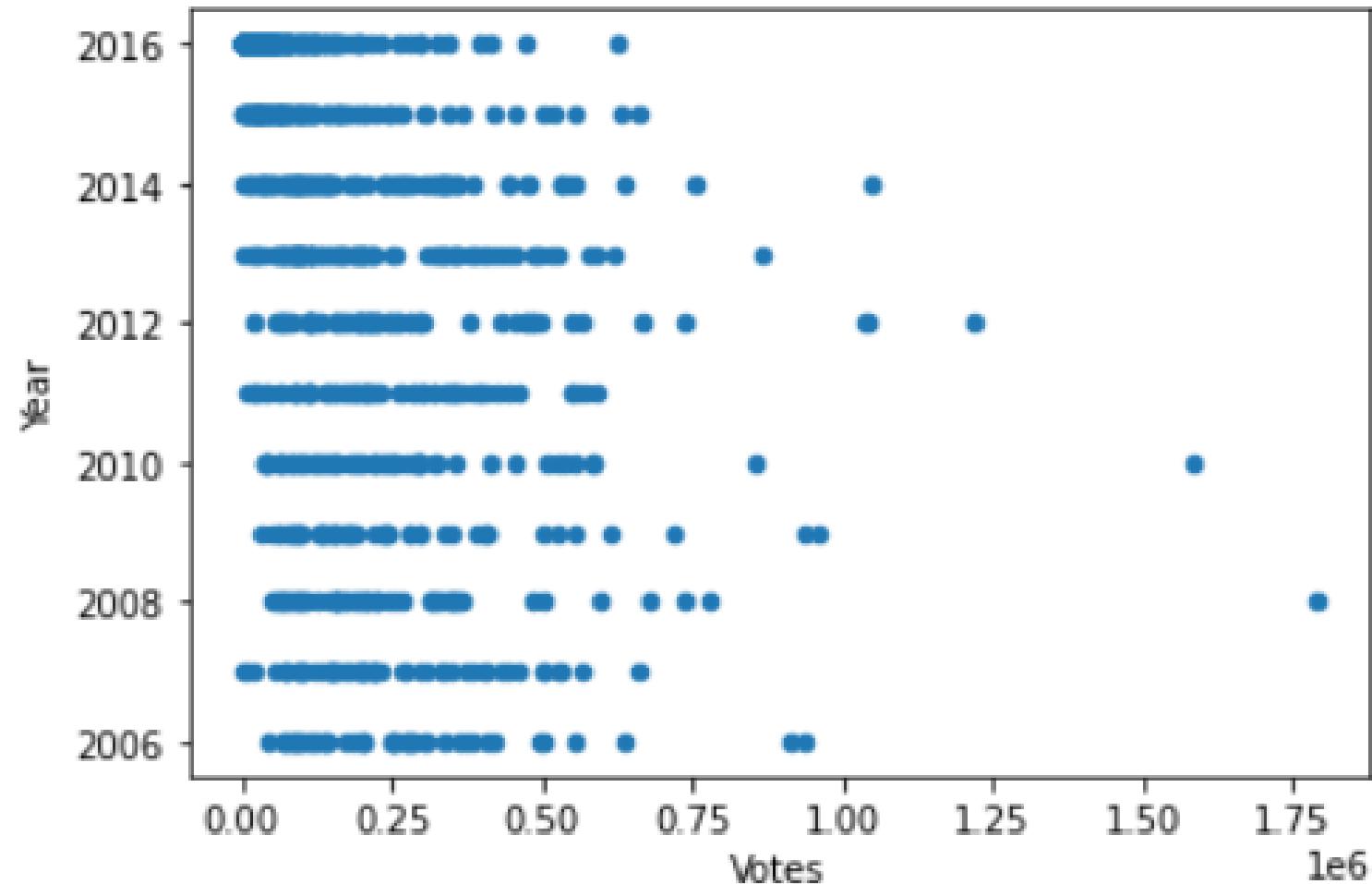
# 6.4 ГІПОТЕЗИ ТА ГРАФІКИ

ЧИ важливий рік випуску фільму у його статистиці?

```
print("<<Чи залежать голоси фільму від року його випущення?>>")  
print("Сер. к-ть голосів старих фільмів:", df[df['Year'] <= 2010]['Votes'].mean())  
print("Сер. к-ть голосів нових фільмів:", df[df['Year'] >= 2010]['Votes'].mean())  
.plot(x = 'Votes', y = 'Year', kind = 'scatter')  
print("К-ть голосів старих фільм перевищують у 2 рази голоси нових фільмів")
```

```
<<Чи залежать голоси фільму від року його випущення?>>  
Сер. к-ть голосів старих фільмів: 258985.90384615384  
Сер. к-ть голосів нових фільмів: 147048.57375  
К-ть голосів старих фільм перевищують у 2 рази голоси нових фільм
```

Чи залежать голоси фільму від його року випуску?



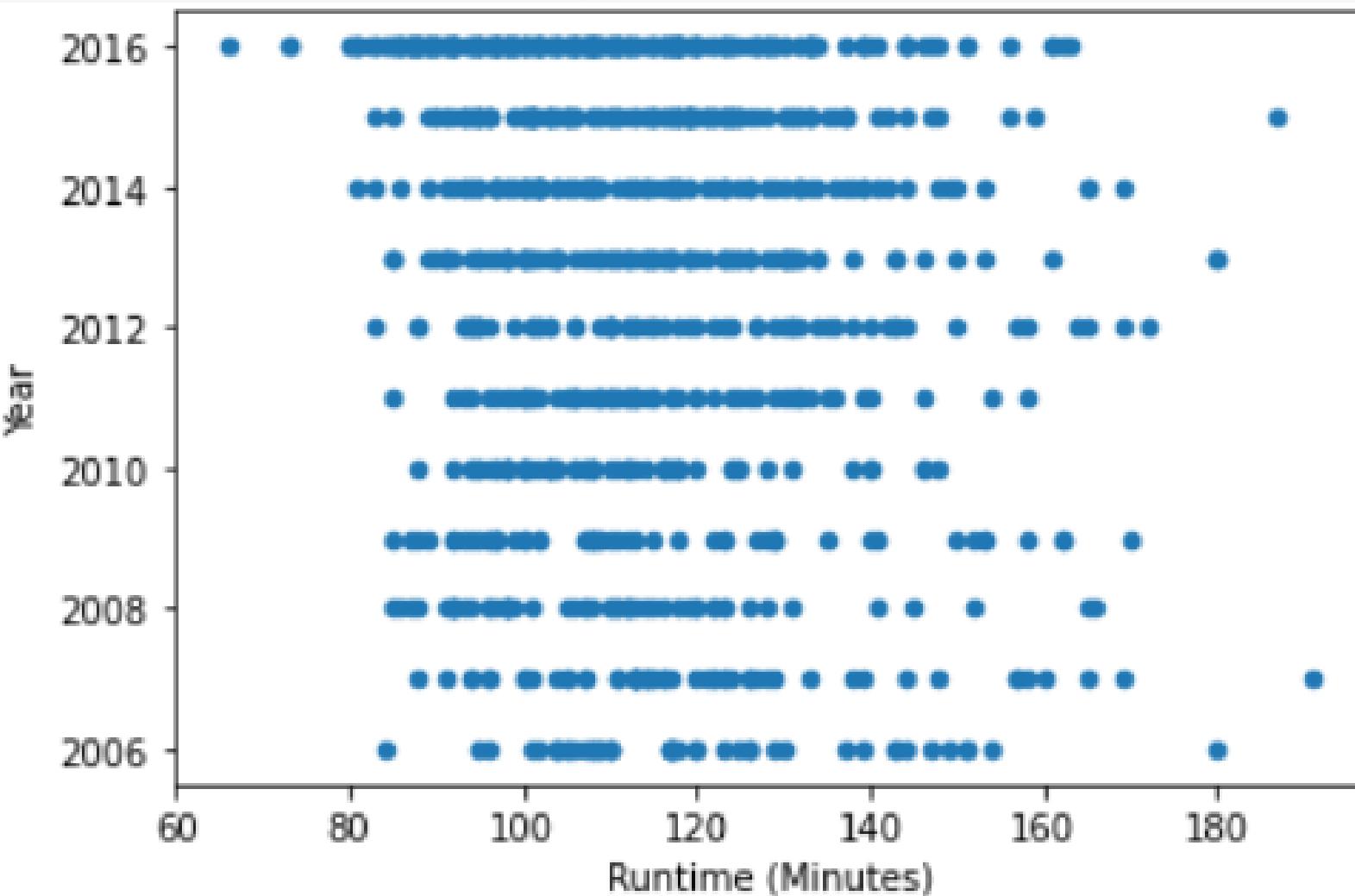
# 6.5 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливий рік випуску фільму у його статистиці?

```
print("<<Чи залежить тривалість фільму від року його випущення?>>")  
print("Сер. тривалість старих фільмів:", df[df['Year'] <= 2010]['Runtime (Minutes)'].mean())  
print("Сер. тривалість нових фільмів:", df[df['Year'] >= 2010]['Runtime (Minutes)'].mean())  
df.plot(x = 'Runtime (Minutes)', y = 'Year', kind = 'scatter')  
print("Тривалість фільмів відносно однакова")
```

<<Чи залежить тривалість фільму від року його випущення?>>  
Сер. тривалість старих фільмів: 115.83076923076923  
Сер. тривалість нових фільмів: 112.155  
Тривалість фільмів відносно однакова

Чи залежить тривалість фільму від року його випущення?



# 6.5 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливий рік випуску фільму у його статистиці?

```
print('Порівняння старих й нових фільмів: \n-Рейтинг одинаковий; \n-Оцінка одна  
print('ВИСНОВОК: По статистиці старі фільми перевищують нові за к-тю голосів, заробітком з ф
```

Порівняння старих й нових фільмів:

- Рейтинг одинаковий;
- Оцінка одна;
- Заробіток старих фільмів дещо більший;
- К-ть голосів старих фільм є два рази більше ніж нових;
- Тривалість фільмів одна

ВИСНОВОК: По статистиці старі фільми перевищують нові за к-тю голосів, заробітком з фільму.

Гіпотеза підтверджена - рік важливий у статистиці фільму.

# 6.6 ГІПОТЕЗИ ТА ШРАФІКИ

Чи важливі актори й жанри у статистиці фільмів?

Якщо у фільмів мін. акторський склад й мін. кількість жанрів то їхня сер. оцінка є нижча ніж загальна сер. оцінка?

```
print('ЧИ ВАЖЛИВІ АКТОРИ Й ЖАНРИ У СТАТИСТИЦІ ФІЛЬМІ?')
print("<<<Якщо у фільмів мін. акторський склад й мін. кількість жанрів то їхня сер. оцінка є нижча ніж загальна сер. оцінка?>>>")
print("Створюю стовпець з кількістю акторів")
def actorsnum2(act):
    act = act.split(',')
    return len(act)
df['Number of actors'] = df['Actors'].apply(actorsnum2)
print("Створюю стовпець з кількістю жанрів")
def genresnum(genre):
    genre = genre.split(',')
    return len(genre)
df["Number of genres"] = df['Genre'].apply(genresnum)
print("Мін. к-ть акторів:", df['Number of actors'].min())
print("Мін. к-ть жанрів:", df['Number of genres'].min())
print("Середня оцінка цих фільмів:", df[(df['Number of genres'] <= 2) & (df['Number of actors'] <= 3)]['Metascore'].mean())
print('Загальна сер. оцінка:', df['Metascore'].mean())
print("Гіпотеза підтверджена. Середня оцінка нижча від загальної середньої оцінки")
```

ЧИ ВАЖЛИВІ АКТОРИ Й ЖАНРИ У СТАТИСТИЦІ ФІЛЬМІ?

<<<Якщо у фільмів мін. акторський склад й мін. кількість жанрів то їхня сер. оцінка є нижча ніж загальна сер. оцінка?>>>

Створюю стовпець з кількістю акторів

Створюю стовпець з кількістю жанрів

Мін. к-ть акторів: 3

Мін. к-ть жанрів: 1

Середня оцінка цих фільмів: 49.0

Загальна сер. оцінка: 58.98504273504273

Гіпотеза підтверджена. Середня оцінка нижча від загальної середньої оцінки

# 6.7 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливі актори й жанри у статистиці фільмів?

Чи правда, що сімейні комедії з кількістю акторів більше 3, мають голоси вище середніх?

```
print("<<<Чи правда, що сімейні комедії з кількістю акторів більше 3, мають голоси вище середніх?>>>")  
print("Середнє значення голосів всіх фільмах", round(df['Votes'].mean(), 0))  
print(round(df[(df['Number of actors'] > 3) & (df['Genre'].str.contains('Comedy')) & (df['Genre'].str.contains('Family'))]['Votes'])  
print('Гіпотеза не підтверджена. Сімейні комедії з кількістю акторів більше 3 мають меншу від середньої кількості голосів.')
```

```
<<<Чи правда, що сімейні комедії з кількістю акторів більше 3, мають голоси вище середніх?>>>
```

```
Середнє значення голосів всіх фільмах 169808.0
```

```
94773.0
```

```
Гіпотеза не підтверджена. Сімейні комедії з кількістю акторів більше 3 мають меншу від середньої кількості голосів.
```

# 6.8 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливі актори й жанри у статистиці фільмів?

Якщо Він Дізель знімається у фільмі то у нього оцінка вище 40?

```
print("<<<Якщо Він Дізель знімається у фільмі то у нього оцінка вище 40?>>>")
print("Оцінка фільму з Він Дізелем:", df[df['Actors'].str.contains('Vin Diesel')]['Metascore'].mean())
print("Гіпотеза підтверджена.")
```

```
<<<Якщо Він Дізель знімається у фільмі то у нього оцінка вище 40?>>>
Оцінка фільму з Він Дізелем: 57.0
Гіпотеза підтверджена.
```

# **6.9 ГІПОТЕЗИ ТА ГРАФІКИ**

## Чи важливі актори й жанри у статистиці фільмів?

# З яким жанром є найбільше фільмів?

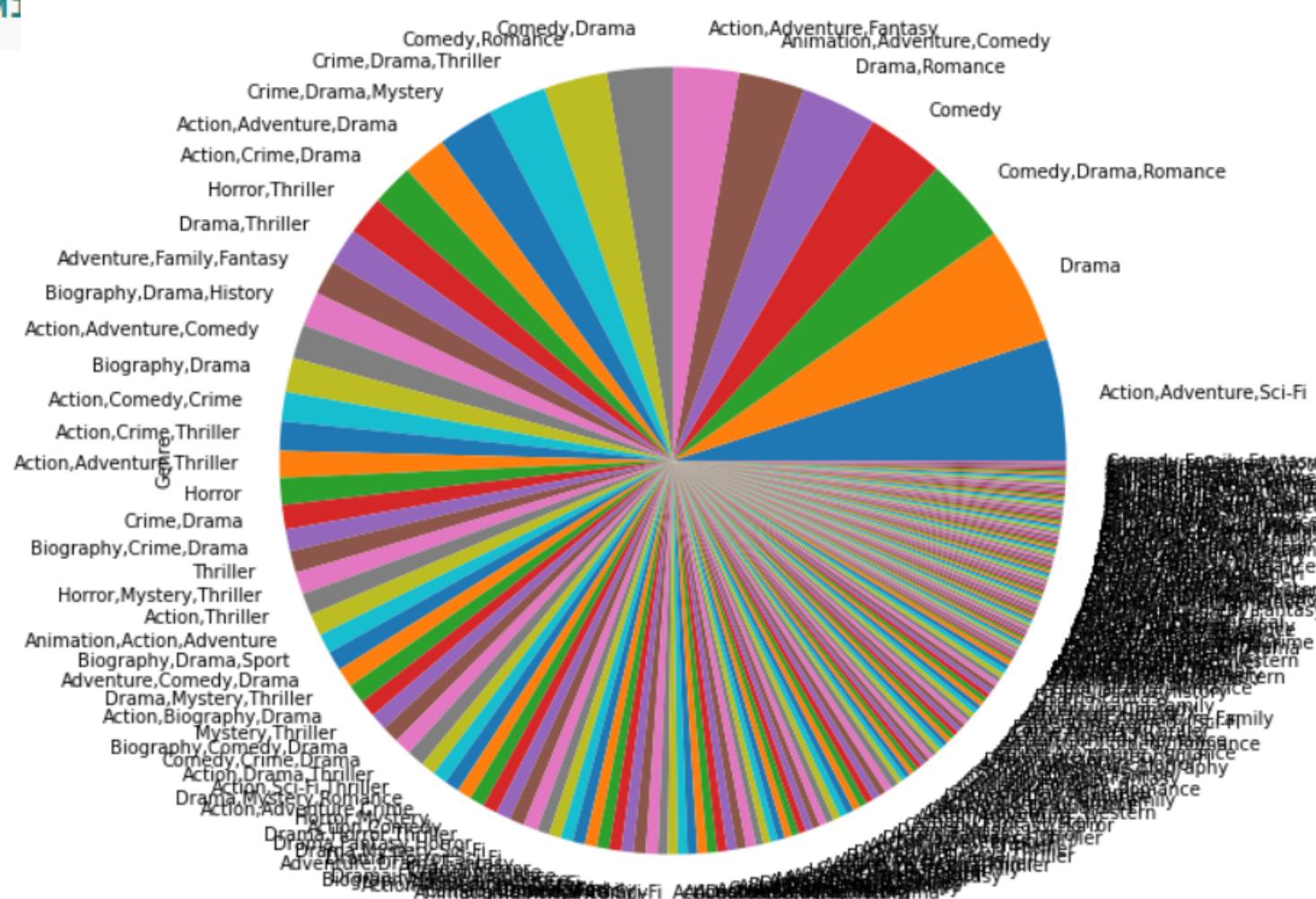
## З ЯКИМ ЖАНРОМ Є НАЙБІЛЬШЕ ФІЛЬМІВ?

## Топ жанрів за кількістю фільмів:

1 місце – Бойовик, Пригоди, Наукова Фантастика;

2 місце – Драма;

3 місце – Комедія, Драма, Романтика.



# 6.9 ГІПОТЕЗИ ТА ГРАФІКИ

Чи важливі актори й жанри у статистиці фільмів?

▶ 0.1s

```
print("ВИСНОВОК: Від жанрів, акторів, певних акторів й жанрів залежить статистика фільмів")
```

ВИСНОВОК: Від жанрів, акторів, певних акторів й жанрів залежить статистика фільмів

# ЩО НОВОГО Я ДІЗНАВСЯ?

За час роботи над датасетом, я багато чого дізнався, а  
саме:

- Старі фільми(2006-2010р.) по статистиці перевищують нові фільми(2010-2016р.);
- Актори й жанри є невід'ємною частиною статистики фільмів;
- Найпопулярніше комбо жанрів за к-тю фільмів - Бойовик, Пригоди, Наукова фантастика.

# Чим користувався при створенні проекту?



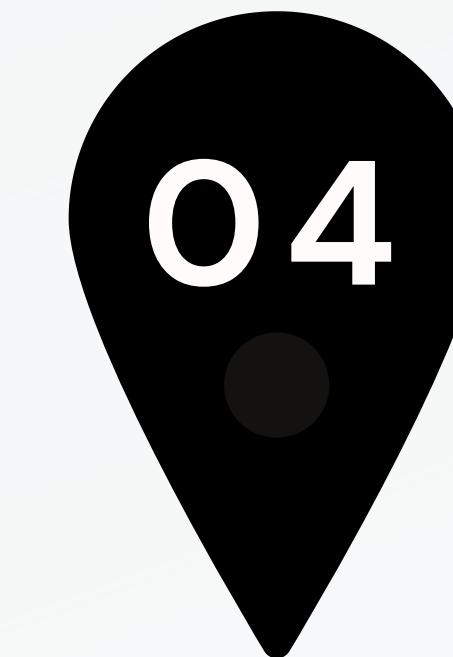
TRELLO



CHATGPT



ПРЕЗЕНТАЦІЇ  
І ЛОГІКИ



DATALORE

# ПОДЯКА

Хотів би подякувати вчителеві, Володимиру  
Ольшанському, за його час вділений мені при  
виконанні проекту!