МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Інститут комп'ютерних наук та інформаційних технологій
Кафедра систем штучного інтелекту

Лабораторна робота №2

З курсу «Візуалізація даних»

Аналіз даних та статистичне виведення

Виконав:
ст. гр. КН-310
Бікєєв Андрій

Викладач:
Бойко Наталія Іванівна

Львів – 2020

1. Хід роботи

1.1. Імпортування бібліотек і даних

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.api as sm

df = pd.read_csv("filmdeathcounts.csv")
```

Ці бібліотеки допоможут в візуалізації даних. та побудові графіків. Дані імпортуються з файлу filmdeathcounts.csv.

1.2. Додавання поля body_per_min, з відношення всіх вбитих у фільмі до довжини фільму у хвилинах

```
data["body_per_min"] = data["Body_Count"] / data["Length_Minutes"]
```

1.3. Побудова гістограми для кількості персонажів, які загинули

```
data['Body_Count'].plot(kind="hist", edgecolor="black", color="cyan", bins=40)
plt.xlabel("Body_Count")
plt.show()
```

На цьому графіку можна побачити 5 розподілів, а також "глобальну"моду усього графіка(а також моду розподілу $0 - 200$), вона дорівнює нулю. Отже, найбільше фільмів у виборці - це таких, в яких ніхто не загинув. Також ми бачимо, що кількість фільмів обернено пропорційна кількості загинувших персонажів, а отже цей графік можливо апроксимувати деякою функцією $y = (a/(bx + d)) + c$, де y - це кількість фільмів, а x - кількість загинувших персонажів, а a, b, c, d - деякі коефіцієнти для апроксимізації.

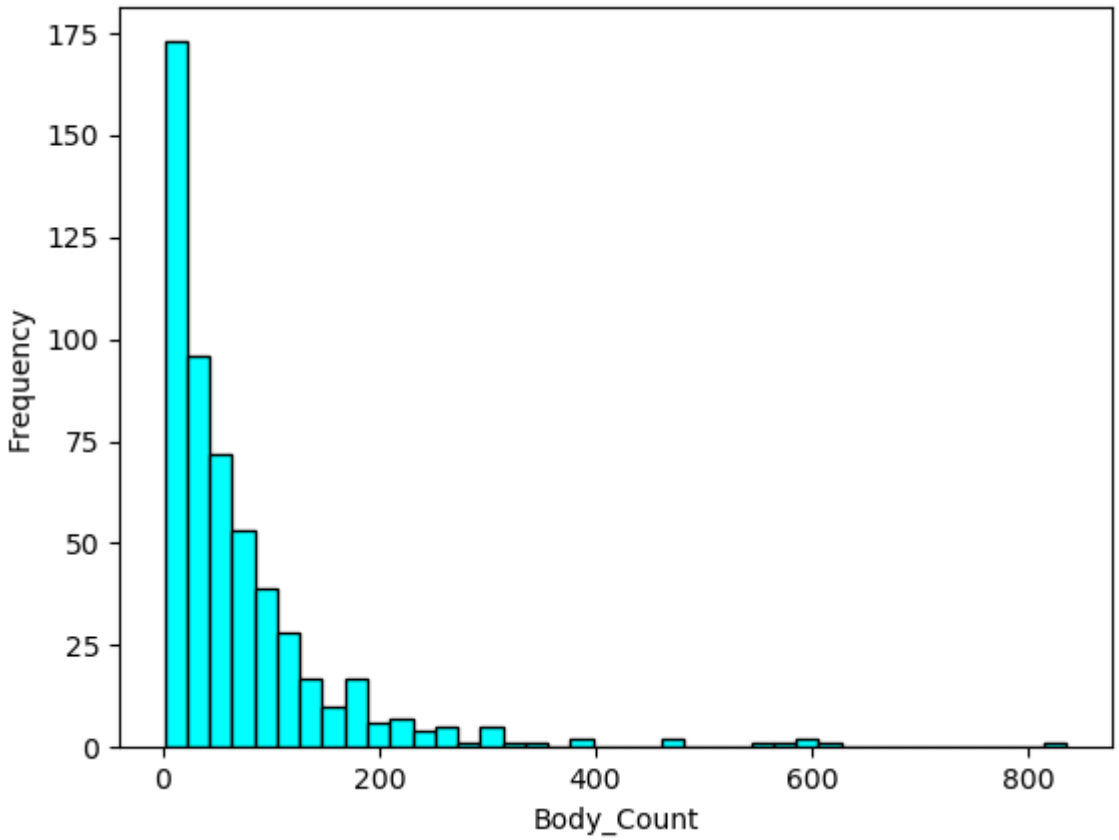

Рис. 1: Гістограма "Загинувші персонажі"

1.4. Топ 10 фільмів в яких загинуло більше всього персонажів

```
print(data.sort_values(by=['Body_Count'], ascending=False).head(10))
print(data.sort_values(by=['body_per_min'], ascending=False).head(10))
```

	Film	Year	Body_Count	MPAA_Rating	\
241	Lord of the Rings: Return of the King	2003	836	PG-13	
217	Kingdom of Heaven	2005	610	R	
4	300	2007	600	R	
388	Tae Guk Gi: The Brotherhood of War	2004	590	R	
509	Troy	2004	572	R	
447	The Last Samurai	2003	558	R	
13	A Fistful of Dynamite	1971	471	PG	
242	Lord of the Rings: Two Towers	2002	468	PG-13	
536	Windtalkers	2002	389	R	
214	King Arthur	2004	378	R	

	Genre	Director	Length_Minutes	\
241	Action|Adventure|Fantasy	Peter Jackson	201	
217	Action|Adventure|Drama|History|War	Ridley Scott	144	
4	Action|Fantasy|History|War	Zack Snyder	117	
388	Action|Drama|War	Je-kyu Kang	140	
509	Adventure|Drama	Wolfgang Petersen	163	
447	Action|Drama|History|War	Edward Zwick	154	
13	Adventure|Western	Sergio Leone	138	
242	Action|Adventure|Fantasy	Peter Jackson	179	
536	Action|Drama|War	John Woo	134	
214	Action|Adventure|Drama	Antoine Fuqua	126	

	IMDB_Rating	body_per_min
241	8.9	4.159204
217	7.1	4.236111
4	7.7	5.128205
388	8.1	4.214286
509	7.1	3.509202
447	7.7	3.623377
13	7.7	3.413043
242	8.7	2.614525
536	5.9	2.902985
214	6.2	3.000000

Рис. 2: Топ 10 за кількістю загинувших персонажів

	Film	Year	Body_Count	MPAA_Rating	Genre	Director	Length_Minutes	IMDB_Rating	body_per_min
4	300	2007	600	R	Action|Fantasy|History|War	Zack Snyder	117	7.7	5.128205
217	Kingdom of Heaven	2005	610	R	Action|Adventure|Drama|History|War	Ridley Scott	144	7.1	4.236111
388	Tae Guk Gi: The Brotherhood of War	2004	590	R	Action|Drama|War	Je-kyu Kang	140	8.1	4.214286
241	Lord of the Rings: Return of the King	2003	836	PG-13	Action|Adventure|Fantasy	Peter Jackson	201	8.9	4.159204
447	The Last Samurai	2003	558	R	Action|Drama|History|War	Edward Zwick	154	7.7	3.623377
509	Troy	2004	572	R	Adventure|Drama	Wolfgang Petersen	163	7.1	3.509202
13	A Fistful of Dynamite	1971	471	PG	Adventure|Western	Sergio Leone	138	7.7	3.413043
214	King Arthur	2004	378	R	Action|Adventure|Drama	Antoine Fuqua	126	6.2	3.000000
405	The Big Red One	1980	338	R	Action|Drama|War	Samuel Fuller	113	7.3	2.991150
536	Windtalkers	2002	389	R	Action|Drama|War	John Woo	134	5.9	2.902985

Рис. 3: Топ 10 за кількістю загинувших персонажів у відношенні до довжини фільму

З цих таблиць, можемо помітити, що за обома критеріями "найжорстокішими"фільмами є фільми жанру Action, Adventure, History, і War. З таблиці два(на малюнку. ??)

1.5. Побудова гістограми для IMDb рейтингу

```
data['IMDB_Rating'].plot(kind="hist", edgecolor="black", color="cyan", bins=20)
plt.xlabel("IMDB_Rating")
plt.show()
```

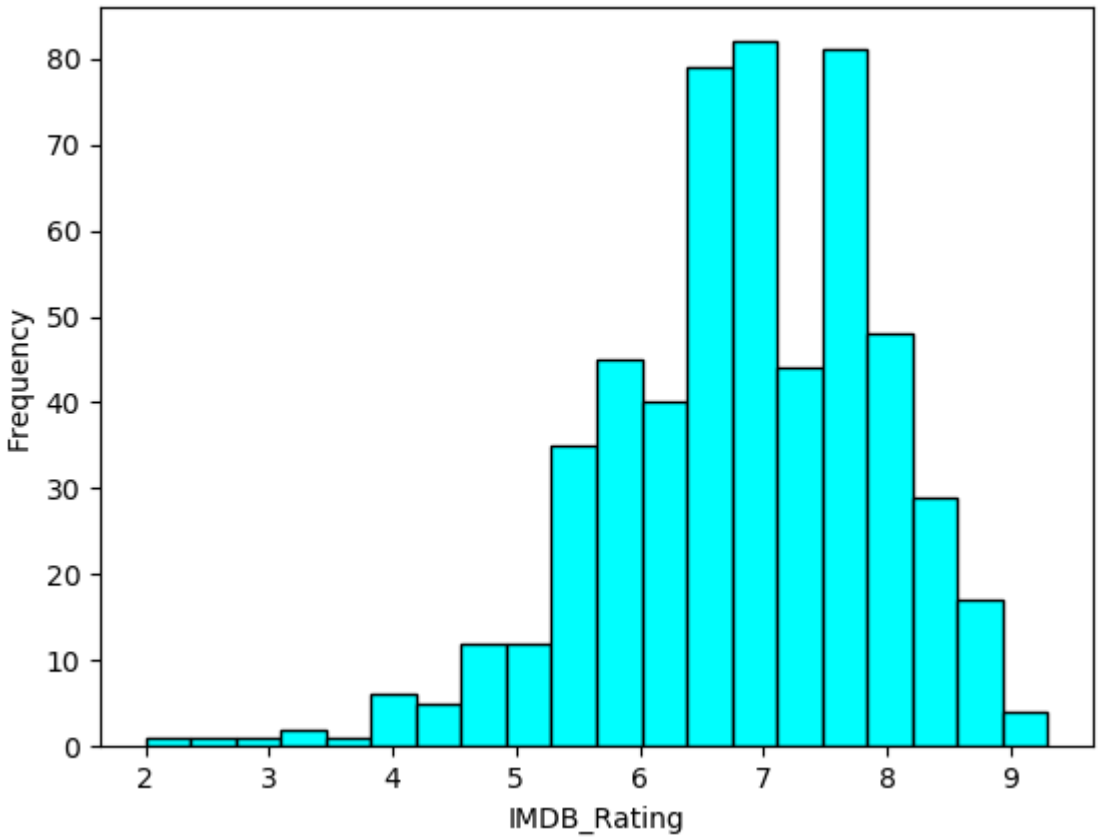

Рис. 4: Гістограма "IMDb рейтинг"

З цієї діаграми можна побачити, що більшість фільмів мають рейтинг більший ніж 5, що є дивним, адже цю оцінку ми вважаємо за "середню". "Глобальною"модою цієї діаграми є 7, у той час, як середнє значення прибзлино 7.5, що означає. що розподіл не є нормальним.

1.6. Симуляція гістограми IMDb рейтингу

Порахуємо середнє значення, та середньоквадратичне відхилення для цієї вибірки, і побудуємо нормальну діаграму з такими значеннями.

```
imdb_mean = data['IMDB_Rating'].mean()
imdb_sd = data['IMDB_Rating'].std()

data['imdb_simulation'] = np.random.normal(imdb_mean, imdb_sd, len(data.index))
data['imdb_simulation'].plot(kind="hist", edgecolor="black", color="cyan", bins=40)
plt.xlabel("imdb_simulation")
plt.show()
```

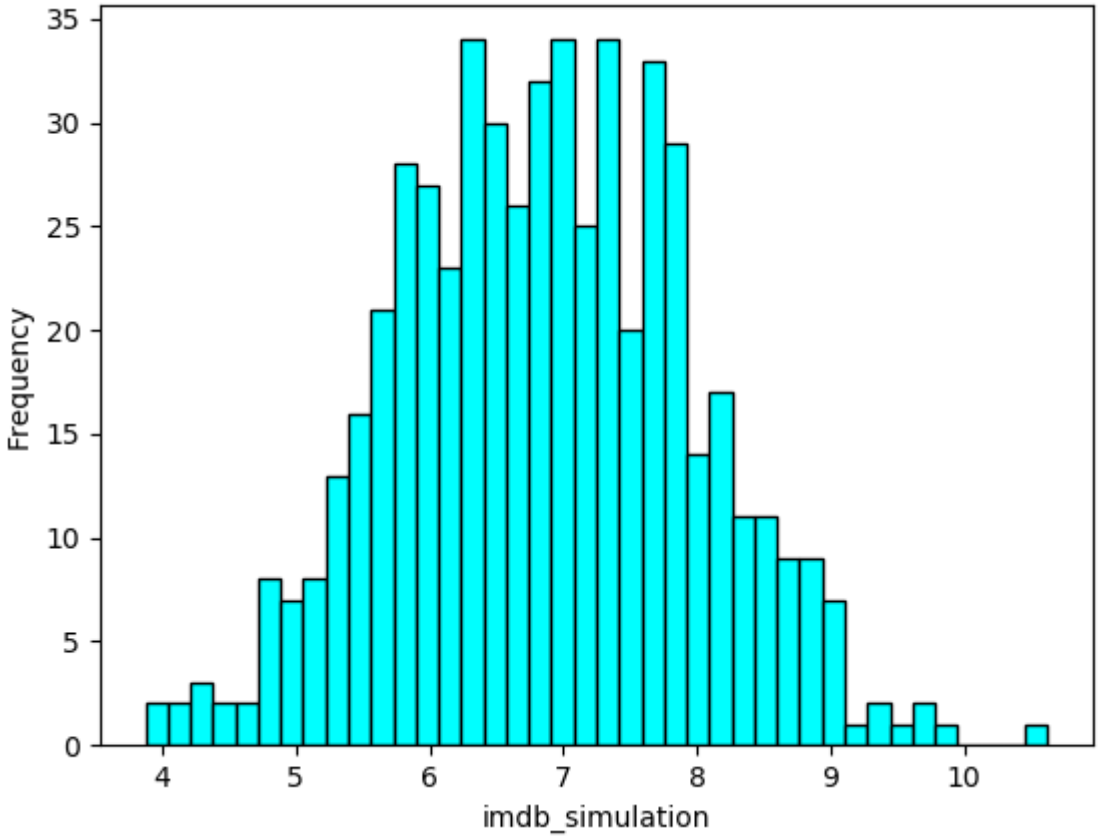

Рис. 5: Гістограма "Симуляція IMDb рейтингу"

1.7. Перевірка нормальності розподілу

Для перевірки нормальності розподілу симуляції, а також даної вибірки побудуємо Q-Q діаграму.

```
sm.qqplot(data['imdb_simulation'])
plt.xlabel("imdb_simulation")
plt.show()
```

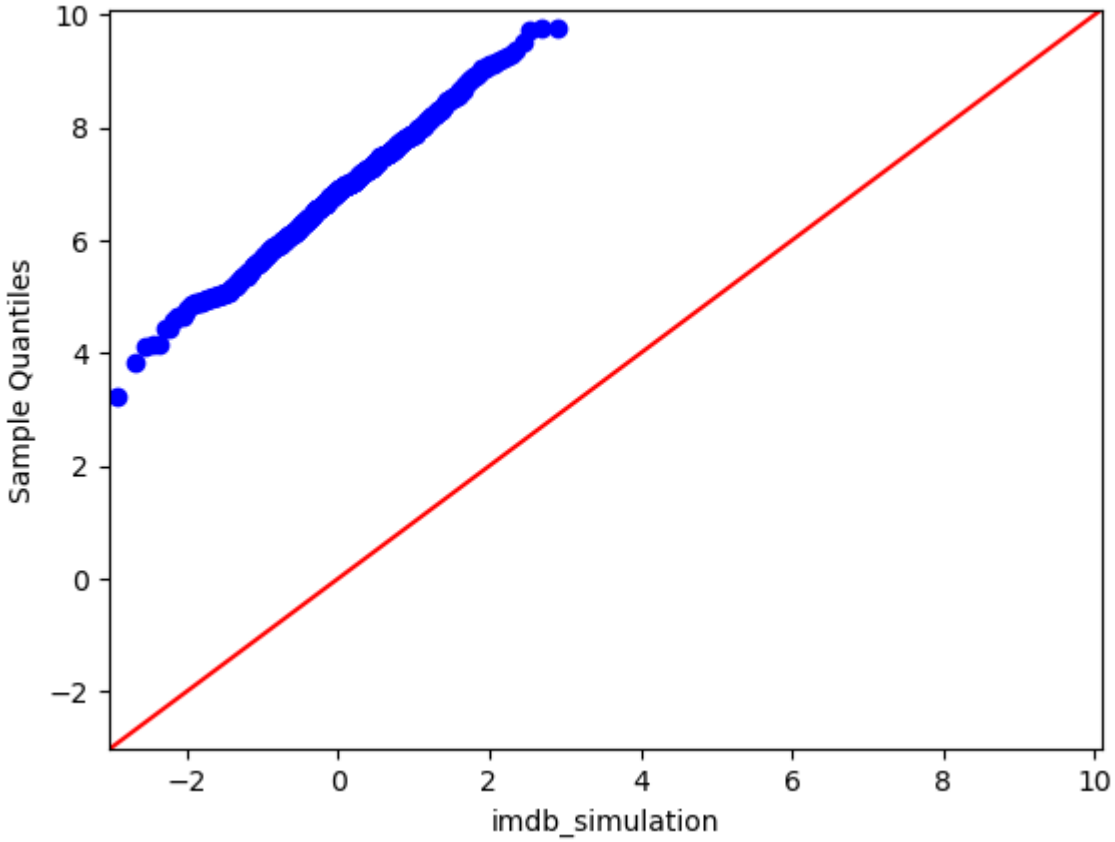

Рис. 6: Q-Q діаграмма для "Симуляції IMDb рейтингу"

Лінійність графіку симуляції показує, що ця вибірка є з нормальни розподілом.

```
sm.qqplot(data['IMDB_Rating'])
plt.xlabel("IMDB_Rating")
plt.show()
```

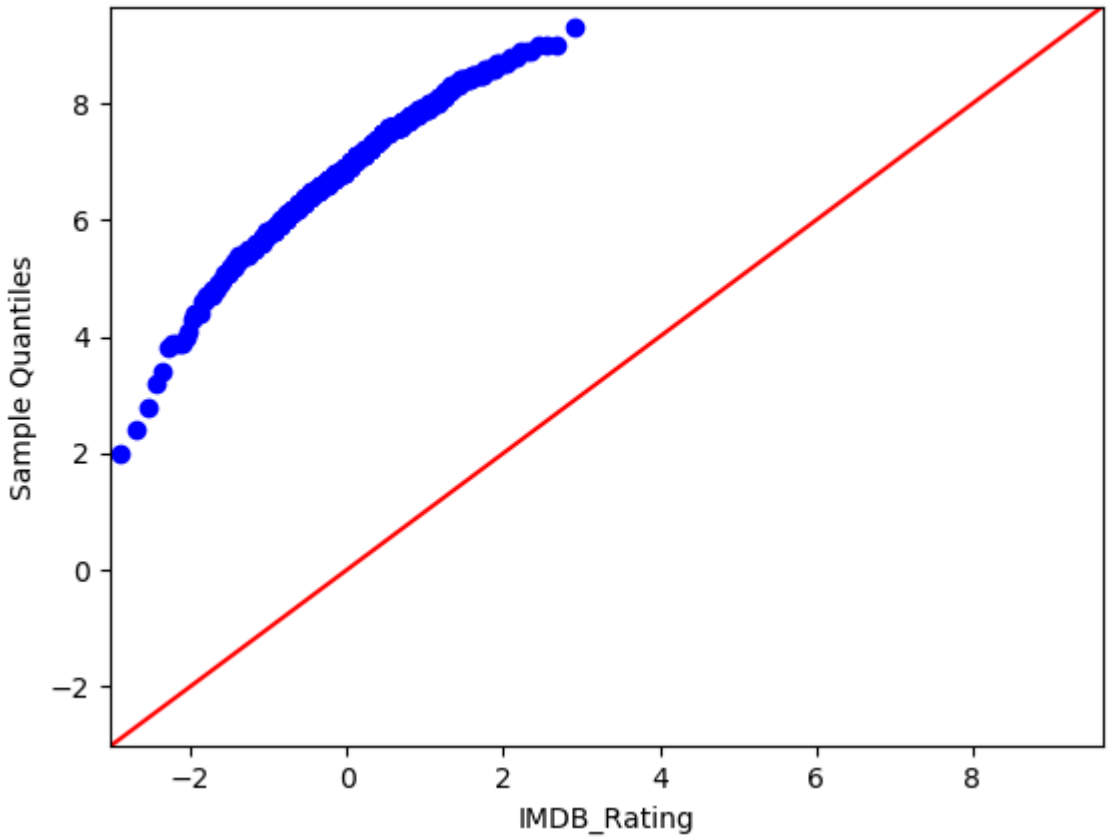

Рис. 7: Q-Q діаграмма для "IMDb рейтингу"

Точки на графіку будують не-лінійний паттерн, отже розподіл цієї вибірки не є нормальним.

2. Висновки

Виконуючи цю лабораторну роботу я навчився зчитувати .csv файли, будувати графіки і діаграми з отриманих даних, а також аналізувати їх.