

The background image shows a wide-angle aerial view of the Chicago city skyline during sunset. The sky is filled with soft, pastel-colored clouds. In the foreground, numerous skyscrapers of various heights are visible, their facades reflecting the warm light of the setting sun. The Chicago River and Lake Michigan are in the background, meeting at the horizon. The city lights are just beginning to turn on, adding small dots of light to the urban landscape.

# Mission: Movie Analytics for Data Scientist

# Instruksi Umum



# Latihan



1. Buka link di <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data>
2. Terdapat dataset *ratings.csv*
3. Buatlah model *machine learning* menggunakan *full data* tersebut.
4. Jika memungkinkan, buatlah kolom baru *genre*, ambil *genre* pertama saja yang muncul apabila terdapat banyak *genre*
5. Buatlah EDA dari *genre* yang didapatkan
6. Buatlah model baru menggunakan kolom *genre* tersebut





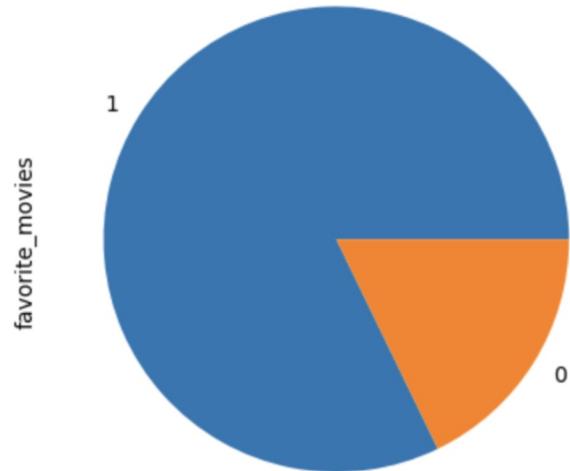
# Data Exploration

# Favorite Movie



```
movies_rating["favorite_movies"] = (movies_rating["median"] >=3).astype("int")
movies_rating["favorite_movies"].value_counts().plot(kind = "pie")
plt.title("Favorite Movies Distribution")
plt.show()
```

Favorite Movies Distribution



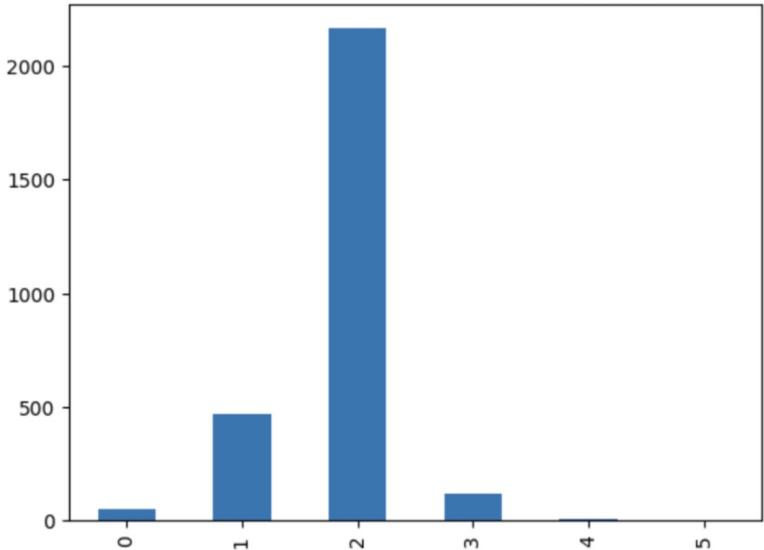
1. Film yang memiliki rating tinggi ( $\geq 3$ ) terdapat sebanyak 80% dari total film



# Movie Duration

```
movies_rating["duration_hours"] = movies_rating["runtime"].div(60).round(0).astype("int")
movies_rating["duration_hours"].value_counts().sort_index().plot(kind = "bar")
plt.title("Movie Duration in Hours")
plt.show()
```

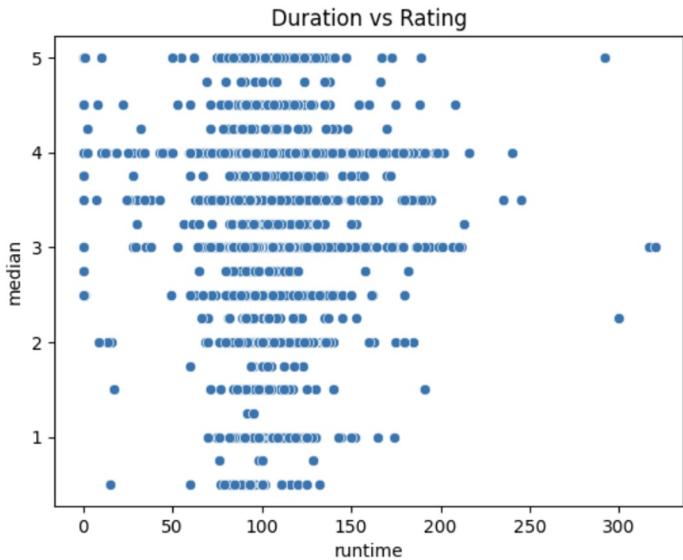
Movie Duration in Hours



1. Durasi film terbanyak terdapat di antara durasi 2-3 jam
2. Terdapat beberapa film panjang yang berdurasi di atas 3 jam

# Movie Duration vs Rating

```
> sns.scatterplot(x = "runtime",
      y = "median",
      data = movies_rating)
plt.title("Duration vs Rating")
plt.show()
```

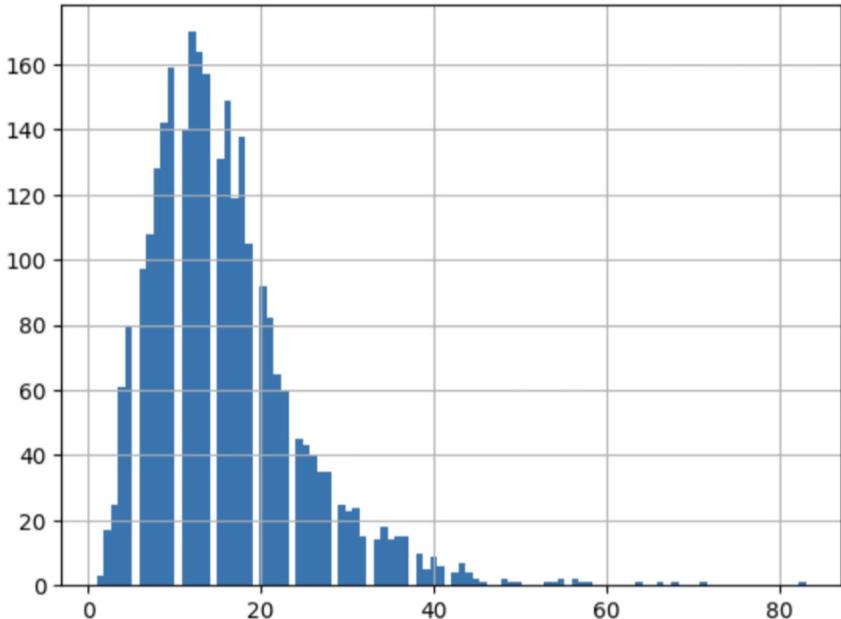


1. Pola berbentuk huruf V. artinya film yang memiliki durasi terlalu pendek atau terlalu panjang justru tidak memiliki rating yang rendah
2. Film rating rendah justru didominasi oleh film yang berdurasi 2-3 jam

# Title Length

```
movies_rating["len_title"] = movies_rating["original_title"].str.len()  
movies_rating["len_title"].hist(bins = 100)
```

<Axes: >

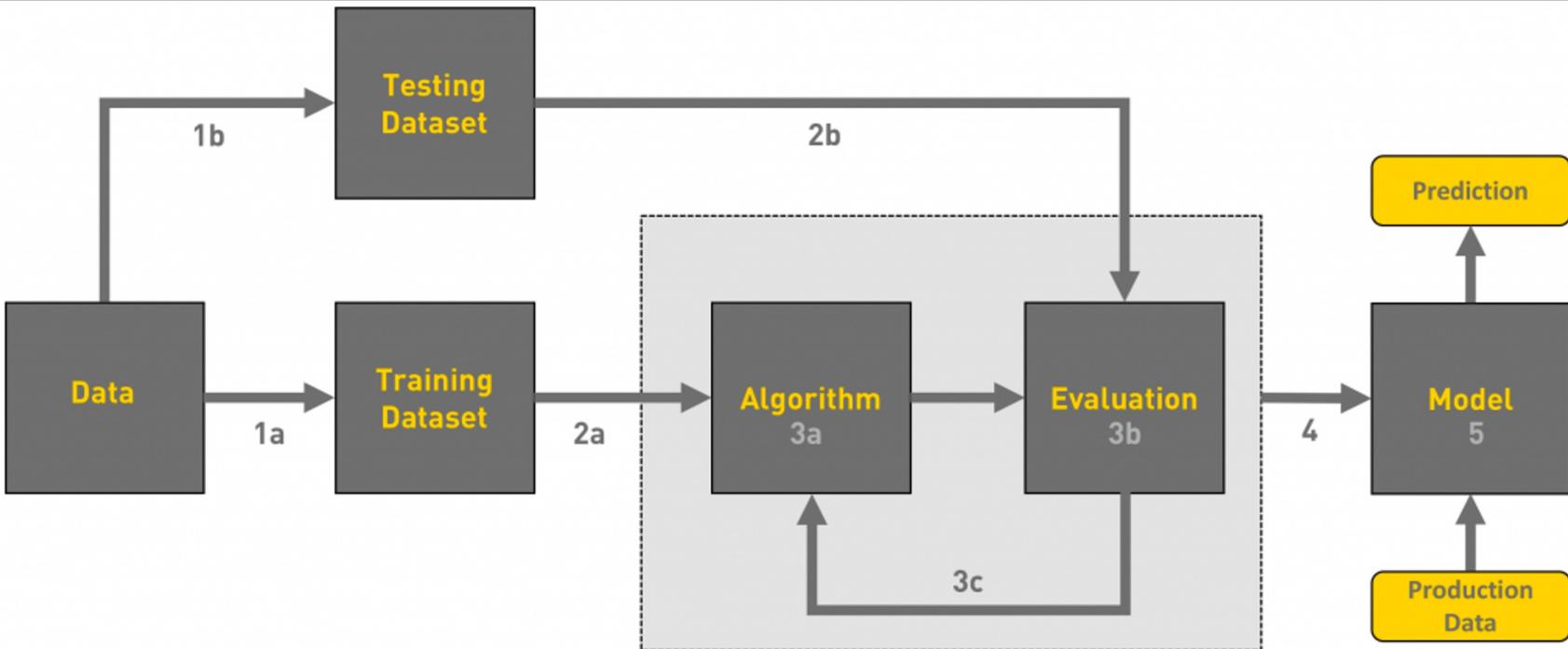


1. Sebagian besar film memiliki panjang judul  $< 30$  huruf
2. Terdapat beberapa film yang memiliki judul yang cukup panjang, yaitu di atas  $>40$  huruf

# Train-Test Data Concept



# Modeling Workflow



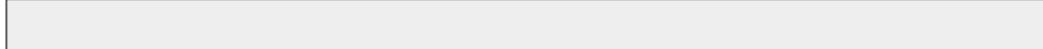
# Holdout Validation



All data



Train



Test



# Cross Validation



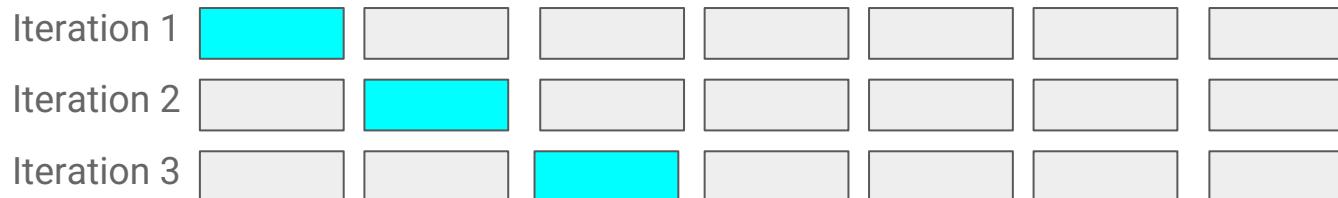
All data



Split into  $k$  parts



Blue one is training dataset



# Machine Learning Workflow



1. Sebagian besar film memiliki panjang judul < 30 huruf
2. Terdapat beberapa film yang memiliki judul yang cukup panjang, yaitu di atas >40 huruf



# Thank You!

