

Exploratory Data Analysis

By : Farry Akbar Pambudi





Introduction

Hello, my name is Farry Akbar Pambudi. I am 16 years old and currently studying in the 11th grade at SMK Telkom Purwokerto, a vocational high school that focuses on technology and telecommunications. I am majoring in Software Engineering, which is known as Rekayasa Perangkat Lunak (RPL) in Indonesian.





About Dataset

TMDB Top Rated Movie Dataset

This dataset contains information about the top-rated movies listed on The Movie Database (TMDB). It includes various features such as movie titles, average ratings, popularity, release dates, genres, and more. This data can be useful for exploratory data analysis, building recommendation systems, and understanding trends in top-rated films.



```
[9] import pandas as pd

df = pd.read_csv('/content/tmdb_top_rated_movies.csv')
```

[10] df

	id	original_language	overview	release_date	title	popularity	vote_average	vote_count
0	278	en	Imprisoned in the 1940s for the double murder ...	1994-09-23	The Shawshank Redemption	35.0440	8.708	28176
1	238	en	Spanning the years 1945 to 1955, a chronicle o...	1972-03-14	The Godfather	36.3040	8.687	21363
2	240	en	In the continuing saga of the Corleone crime f...	1974-12-20	The Godfather Part II	17.2441	8.570	12910
3	424	en	The true story of how businessman Oskar Schind...	1993-12-15	Schindler's List	37.5490	8.564	16380
4	389	en	The defense and the prosecution have rested an...	1957-04-10	12 Angry Men	17.0033	8.548	9091
...
9995	340674	en	Tadek, a Polish detective, becomes suspicious ...	2016-10-12	Dark Crimes	1.7494	4.602	451
9996	16780	en	After a group of Texas teenagers leave prom ni...	1995-09-22	The Return of the Texas Chainsaw Massacre	3.6004	4.595	693
9997	455551	en	When five friends vacation at a remote lake ho...	2017-06-02	The Recall	2.5171	4.585	346
9998	11411	en	With global superpowers engaged in an increasi...	1987-07-24	Superman IV: The Quest for Peace	4.3324	4.583	1281
9999	10285	en	Jason Voorhees is tracked down and blown to bi...	1993-08-13	Jason Goes to Hell: The Final Friday	2.8208	4.600	1027

10000 rows × 8 columns

We start the analysis by reading the dataset into a DataFrame.

I used pandas to read the dataset from a CSV file.

Reading a dataset

Handle Missing Values

```
df.isna().sum()
```

	0
id	0
original_language	0
overview	0
release_date	2
title	0
popularity	0
vote_average	0
vote_count	0

dtype: int64

Checking how many values are missing in each column

Remove Missing Values

```
df=df.dropna()  
df.isna().sum()
```

	0
id	0
original_language	0
overview	0
release_date	0
title	0
popularity	0
vote_average	0
vote_count	0

dtype: int64



Handling Duplicate Data

Check for duplicate rows to ensure the dataset is clean and reliable.

```
check_duplicate = df.duplicated().sum()  
  
print(f"Jumlah data yang duplikat = {check_duplicate}")  
  
Jumlah data yang duplikat = 213
```





Handling Duplicate Data

After identifying duplicates, we proceeded to remove them from the dataset.

```
df = df.drop_duplicates()

handle_duplicate = df.duplicated().sum()

print(f"Jumlah data yang duplikat = {handle_duplicate}")

Jumlah data yang duplikat = 0
```



THANK YOU