**Project Description**

To perform a movie portfolio analysis using Python with libraries like **Pandas**, **NumPy**, and **Matplotlib**, we can follow a few key steps. I'll walk you through the process to analyze movie performance, calculate correlations, and build charts. You’ll need data on movies, such as their budget, revenue, genre, ratings, or box office performance. You can start by gathering a dataset (e.g., from sources like IMDb, Kaggle, or a personal dataset).

Here is a basic step-by-step guide for performing the analysis:

**1. Data Preparation**

* We'll load the data into a Pandas DataFrame.
* Clean the data (handling missing values, incorrect data, etc.).
* Select relevant columns (e.g., budget, revenue, rating, etc.).

**2. Correlation Analysis**

* Calculate correlations between numeric variables (e.g., budget vs. revenue, budget vs. rating).
* Use .corr() in Pandas to find correlations.

**3. Visualization**

* Create charts using **Matplotlib** (scatter plots, bar charts, etc.) to visualize the relationships between variables.
* We'll also look at a heatmap to visualize correlations.

**4. Performing the Analysis in a Jupyter Notebook**

Let me show you the code:

python

CopyEdit

# Import required libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset (example: you might need to load your own CSV file)

# For demonstration, let's use a sample dataset from a CSV file.

# dataset = pd.read\_csv("movies.csv")

# Example dataset for demonstration:

data = {

'Movie': ['Movie A', 'Movie B', 'Movie C', 'Movie D', 'Movie E'],

'Budget': [100, 150, 200, 250, 300], # in millions

'Revenue': [300, 400, 500, 700, 600], # in millions

'Rating': [7.5, 8.0, 6.8, 7.2, 8.5], # out of 10

}

# Create a DataFrame

df = pd.DataFrame(data)

# Show the first few rows of the dataset

print(df)

# Step 1: Clean the data (handling missing values or irrelevant columns)

# In this example, there are no missing values, but in real scenarios, you should use df.isnull().sum() to check for missing data.

# Step 2: Correlation analysis

# Calculate correlations

correlation\_matrix = df[['Budget', 'Revenue', 'Rating']].corr()

# Step 3: Visualize the correlation matrix using a heatmap

plt.figure(figsize=(8, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)

plt.title("Correlation Heatmap")

plt.show()

# Step 4: Visualize budget vs revenue using a scatter plot

plt.figure(figsize=(8, 6))

plt.scatter(df['Budget'], df['Revenue'], color='blue')

plt.title("Budget vs Revenue")

plt.xlabel('Budget (in millions)')

plt.ylabel('Revenue (in millions)')

plt.grid(True)

plt.show()

# Step 5: Visualize Budget vs Rating with a scatter plot

plt.figure(figsize=(8, 6))

plt.scatter(df['Budget'], df['Rating'], color='green')

plt.title("Budget vs Rating")

plt.xlabel('Budget (in millions)')

plt.ylabel('Rating')

plt.grid(True)

plt.show()

# Additional: Bar chart to visualize movie ratings

plt.figure(figsize=(10, 6))

plt.bar(df['Movie'], df['Rating'], color='orange')

plt.title("Movie Ratings")

plt.xlabel('Movie')

plt.ylabel('Rating')

plt.show()

# Step 6: Linear regression (if needed) to predict relationship between Budget and Revenue

from sklearn.linear\_model import LinearRegression

# Prepare data for linear regression (Budget as independent, Revenue as dependent variable)

X = df[['Budget']] # independent variable

y = df['Revenue'] # dependent variable

# Create the regression model

model = LinearRegression()

model.fit(X, y)

# Predict Revenue based on the model

df['Predicted\_Revenue'] = model.predict(X)

# Visualize the linear regression line on Budget vs Revenue

plt.figure(figsize=(8, 6))

plt.scatter(df['Budget'], df['Revenue'], color='blue', label='Actual Revenue')

plt.plot(df['Budget'], df['Predicted\_Revenue'], color='red', label='Predicted Revenue', linewidth=2)

plt.title("Budget vs Revenue with Linear Regression")

plt.xlabel('Budget (in millions)')

plt.ylabel('Revenue (in millions)')

plt.legend()

plt.grid(True)

plt.show()

# You can also perform further analysis such as genre-wise performance, or calculating the ROI (Revenue/Budget)

df['ROI'] = df['Revenue'] / df['Budget']

print(df[['Movie', 'ROI']])

**5. Explanation of Key Steps:**

* **Correlation**: We calculate the correlation between the numerical variables (Budget, Revenue, Rating) to understand how they are related. The .corr() function in Pandas computes pairwise correlations.
* **Heatmap**: The correlation matrix is visualized using a heatmap to give a clearer view of relationships between different features.
* **Scatter Plots**: We plot scatter plots to visually observe the relationship between Budget and Revenue, and Budget and Rating.
* **Linear Regression**: A simple linear regression model is used to predict Revenue from Budget and visualize the line of best fit.

**6. Further Enhancements:**

* You can refine the dataset to include more detailed variables (such as genre, cast, etc.).
* Use more advanced statistical tests or machine learning models (e.g., Random Forest, Gradient Boosting) to predict movie success.
* Visualize more insights like ROI (Revenue/Budget) or genre-wise performance analysis.

**7. Next Steps:**

* You can replace the data dictionary with your real dataset (e.g., load it from a CSV file).
* Experiment with more complex visualizations, such as pair plots or 3D scatter plots if you have multiple dimensions.