```
#numpy -> numerical computation
#pandas ->file handling,reading,manipulation
#matplot ->plot
```

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
#steps know about the file -> pandas https://kaggle.com/datasets/shivamb/netflix-shows
#basic analysis -> numpy,pandas
#visualization ->matplotlib

#load data
df = pd.read_csv('/content/netflix_titles.csv')
# df.info()
# df.head()
# df.tail()
# df.describe()

#analysis
#content type
#count on movies vs tv shows
content_type = df['type'].value_counts()
print(content_type)
#top 5 countries
top_countries = df['country'].value_counts().head(5)
print(top_countries)
#top 5 directors
top_directors = df['director'].value_counts().head(5)
print(top_directors)

#most common genre
top_genre = df['listed_in'].value_counts().head(5)
print("Top Genre : \n",top_genre)

#most common rating
top_rating = df['rating'].value_counts().head(5)
print(top_rating)
#visualization
#pie chart
plt.pie(content_type,labels=content_type.index,autopct='%1.1f%%')
plt.title("Movies vs TV Shows")
plt.show()
#barchart
plt.bar(top_countries.index,top_countries.values)
plt.title("Top 5 Countries")
plt.xlabel("Country")
plt.ylabel("Number of Movies")
plt.show()
#distribution plot
plt.hist(df['release_year'],bins=20)
plt.title("Distribution of Release Year")
plt.xlabel("Release Year")
plt.ylabel("Number of Movies")
plt.show()
```

```
#EDA -exploratory data analysis
#https://www.kaggle.com/datasets/yasserh/titanic-dataset
#steps 1 - data - data cleaning - variation- missing values - visualisation
#import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
#load data
df = pd.read_csv('/content/Titanic-Dataset.csv')
#first look
df.head()
# df.tail()
# df.info()
# df.describe()

#clean data
print(df.isnull().sum())

#fill the missing value
df['Age'].fillna(df['Age'].mean(),inplace=True)
#fill embarked missing values
df['Embarked'].fillna(df['Embarked'].mode()[0],inplace=True)
# #fill with zero
# df['Cabin'].fillna(0,inplace=True)
# #check again
# print(df.isnull().sum())

#drop cabin (too many nulls -687 nulls)
df.drop('Cabin',axis=1,inplace=True)
df.head()

#univariate analysis(whats happening inside one value)  and bivariate analysis(how 2 variables interact)
#univariate age distribution
plt.hist(df['Age'],bins=20)
plt.title("Distribution of Age")
plt.xlabel("Age")
plt.ylabel("Number of Passengers")
plt.show()
#fare distribution
plt.hist(df['Fare'],bins=20,edgecolor='black')
plt.title("Distribution of Fare")
plt.xlabel("Fare")
plt.ylabel("Number of Passengers")
plt.show()
#survival
plt.pie(df['Survived'].value_counts(),labels=df['Survived'].value_counts().index,autopct='%1.1f%%')
#labels
plt.legend(['Not Survived','Survived'])
plt.title("Survival")
plt.show()
#bivariate
#survival by gender bar chart
df.groupby('Sex')['Survived'].value_counts().unstack().plot(kind='bar',stacked=True)
plt.title("Survival by Gender")
plt.xlabel("Gender")
plt.ylabel("Number of Passengers")
plt.show()
#suirvival by class
pivot = pd.pivot_table(df,values='Survived',index='Pclass')
print(pivot)
pivot.plot(kind='bar')
plt.title("Survival by Class")
plt.xlabel("Class")
plt.ylabel("Number of Passengers")
plt.show()
#multivariate
#survival by gender and class with pivot table
pd.pivot_table(df,values='Survived',index='Sex',columns='Pclass').plot(kind='bar',stacked=True)
#correlation matrix
corr =df.corr(numeric_only=True)

#plot heatmap
plt.figure(figsize=(9,7))
im = plt.imshow(corr,cmap="RdYlBu",vmin=-1,vmax=1)

#add colorbar with labels
cbar = plt.colorbar(im)
```