INVESTIGATING MOVIE DATASET

EXPLORATORY DATA ANALYSIS (EDA)

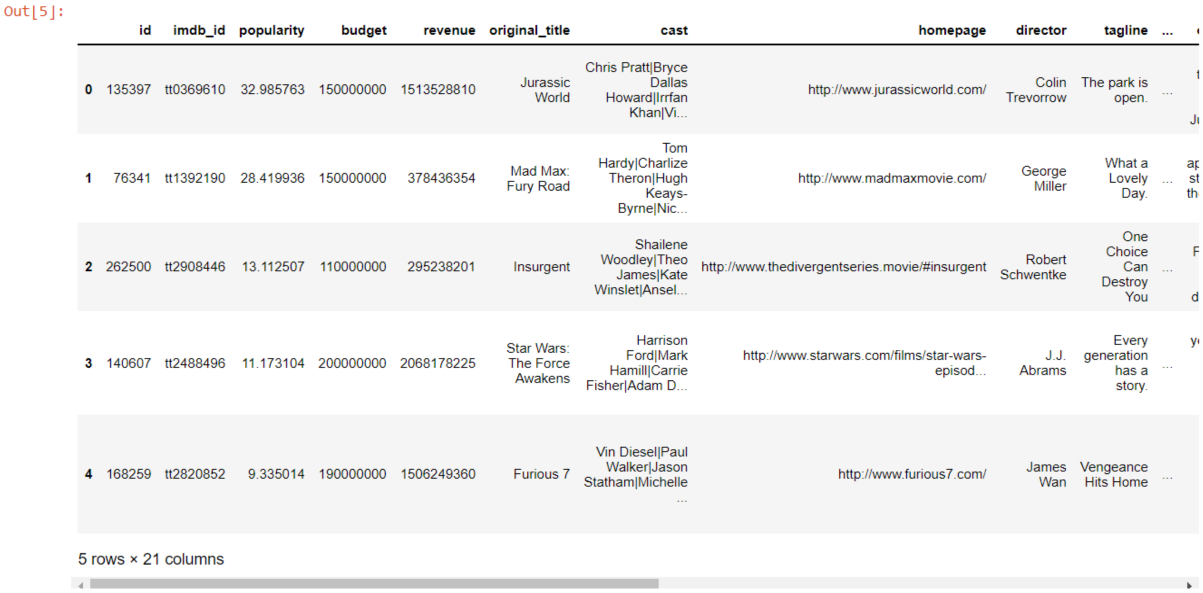
Md. Sayeed Akram	Parth Praveen Shetty	Raghav Chugh
PES2UG20CS201	PES2UG20CS240	PES2UG20CS260

Q. How many rows and attributes?

Code:

```
import csv

import pandas as pd

df=pd.read_csv("tmdb_movies_data.csv")

df.head()

print("Number of rows in dataset: ",len(df.index))

print("Number of columns in dataset: ",len(df.columns))

print('\n\n')

print("Attributes of dataset are", df.columns)
```

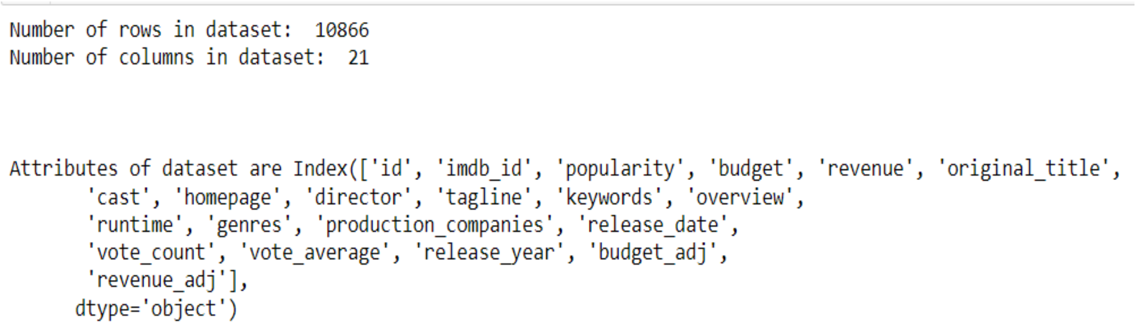
Output:



Out[5]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	...
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	...
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	George Miller	What a Lovely Day.	...
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke	One Choice Can Destroy You	...
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-wars-episod...	J.J. Abrams	Every generation has a story.	...
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	James Wan	Vengeance Hits Home	...

5 rows × 21 columns



```
Number of rows in dataset:  10866
Number of columns in dataset:  21


Attributes of dataset are Index(['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title',
       'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
       'runtime', 'genres', 'production_companies', 'release_date',
       'vote_count', 'vote_average', 'release_year', 'budget_adj',
       'revenue_adj'],
      dtype='object')
```

Conclusion:

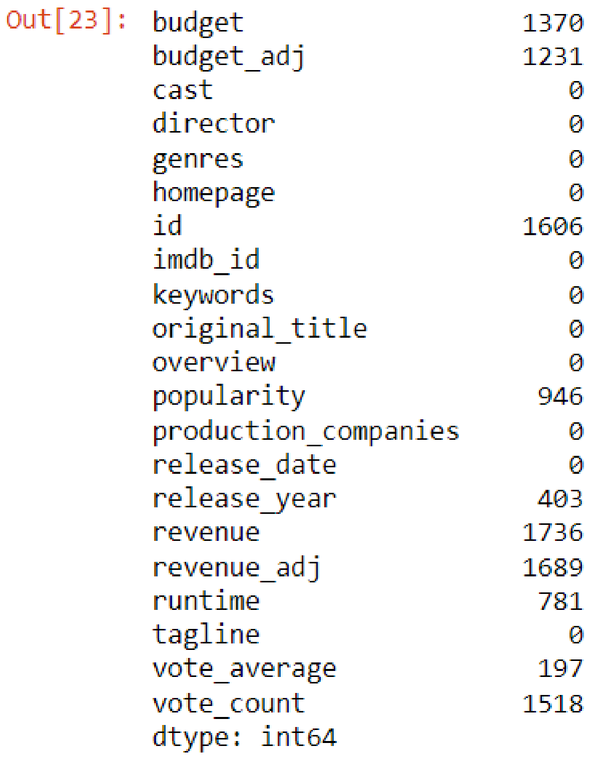
We have chosen Movie Dataset for our data analytics project. This database has 21 attributes and has close to 10,866 rows.

Q2. How many missing data and outliers?

Code for Outliers:

```
Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

print("OUTLIERS")

((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).sum()  #sum of outliers in each attribute
```

Output & Conclusion for Outliers:

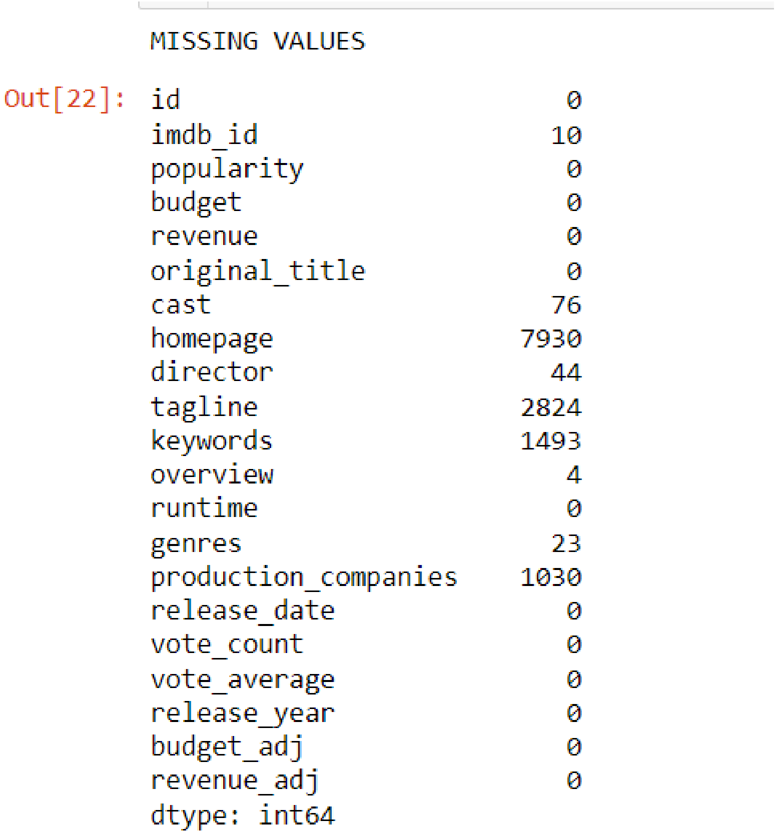


```
Out[23]: budget                  1370
         budget_adj              1231
         cast                       0
         director                   0
         genres                     0
         homepage                   0
         id                      1606
         imdb_id                    0
         keywords                   0
         original_title             0
         overview                   0
         popularity               946
         production_companies       0
         release_date               0
         release_year             403
         revenue                 1736
         revenue_adj             1689
         runtime                  781
         tagline                    0
         vote_average             197
         vote_count              1518
         dtype: int64
```

Code for Missing values:

```
print("MISSING VALUES")

df.isnull().sum()
```

Output & Conclusion for Missing Values:



```
           MISSING VALUES

Out[22]:   id                        0
           imdb_id                  10
           popularity                0
           budget                    0
           revenue                   0
           original_title            0
           cast                     76
           homepage               7930
           director                 44
           tagline                2824
           keywords               1493
           overview                  4
           runtime                   0
           genres                   23
           production_companies   1030
           release_date              0
           vote_count                0
           vote_average              0
           release_year              0
           budget_adj                0
           revenue_adj               0
           dtype: int64
```

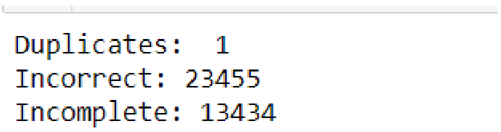
Q. Any inconsistent, incomplete, duplicate or incorrect data?

There a lot many inconsistent, incomplete, duplicate or incorrect data in our data set

```
print("Duplicates: ",df.duplicated().sum())


n=len(df.columns)

sm=0

for i in range (0, n):

   k=df.columns[i]

   sm+=(df[k]==0).sum()


#sm=0

#for i in range (0,n):

#    sm+=(df2[i])

print("Incorrect:",sm)
```

```
df3=df.isnull().sum()

sm=0

for i in range (0,n):

   sm+=(df3[i])

print("Incomplete:",sm)
```

Output



```
Duplicates:  1
Incorrect: 23455
Incomplete: 13434
```

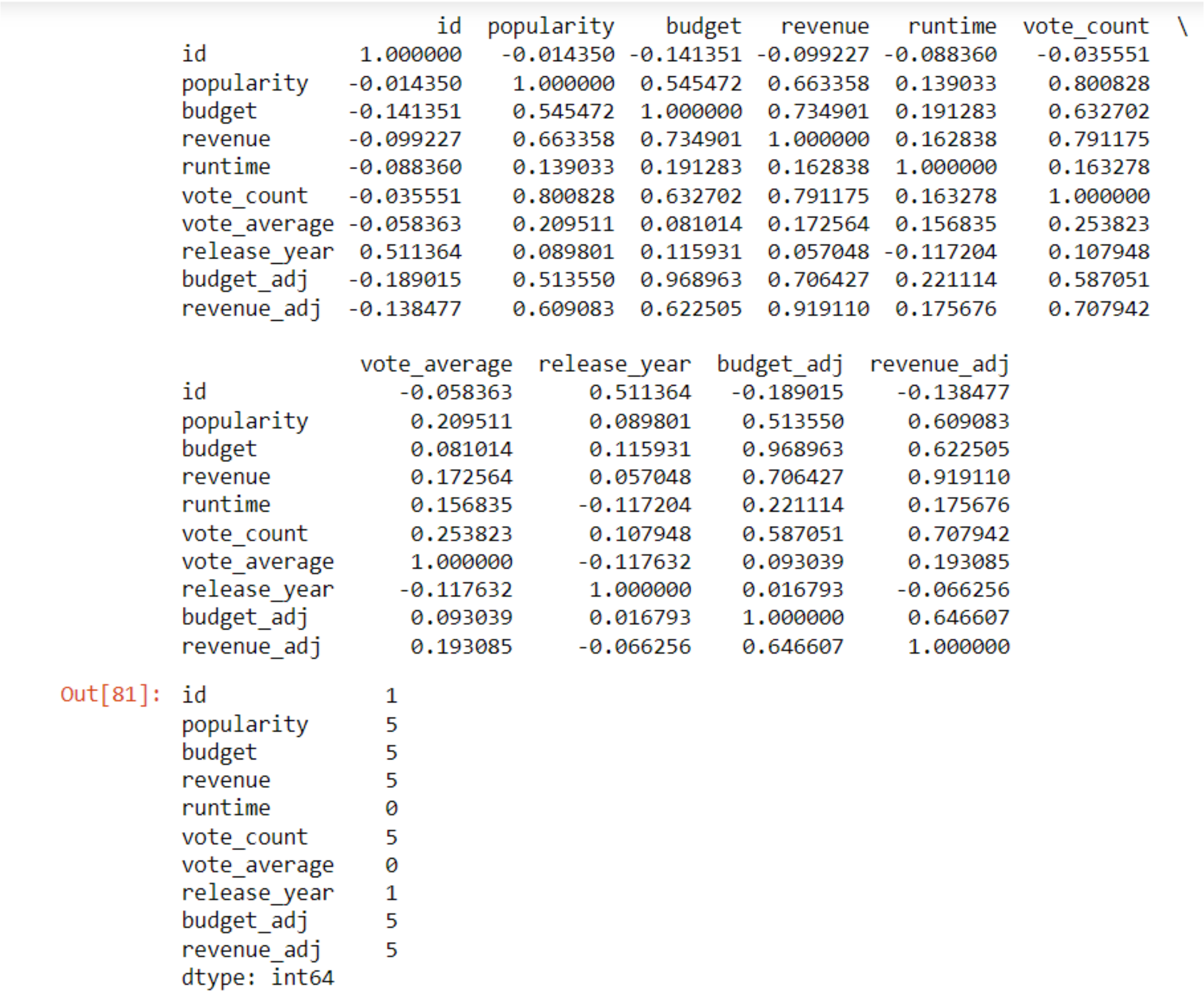
Q. Are the variables correlated to each other?

Yes, a few variables are positively correlated

Code:

```
df7=df.corr()

print(df7)

((df7 > 0.5)).sum()-1
```

Output using numbers:

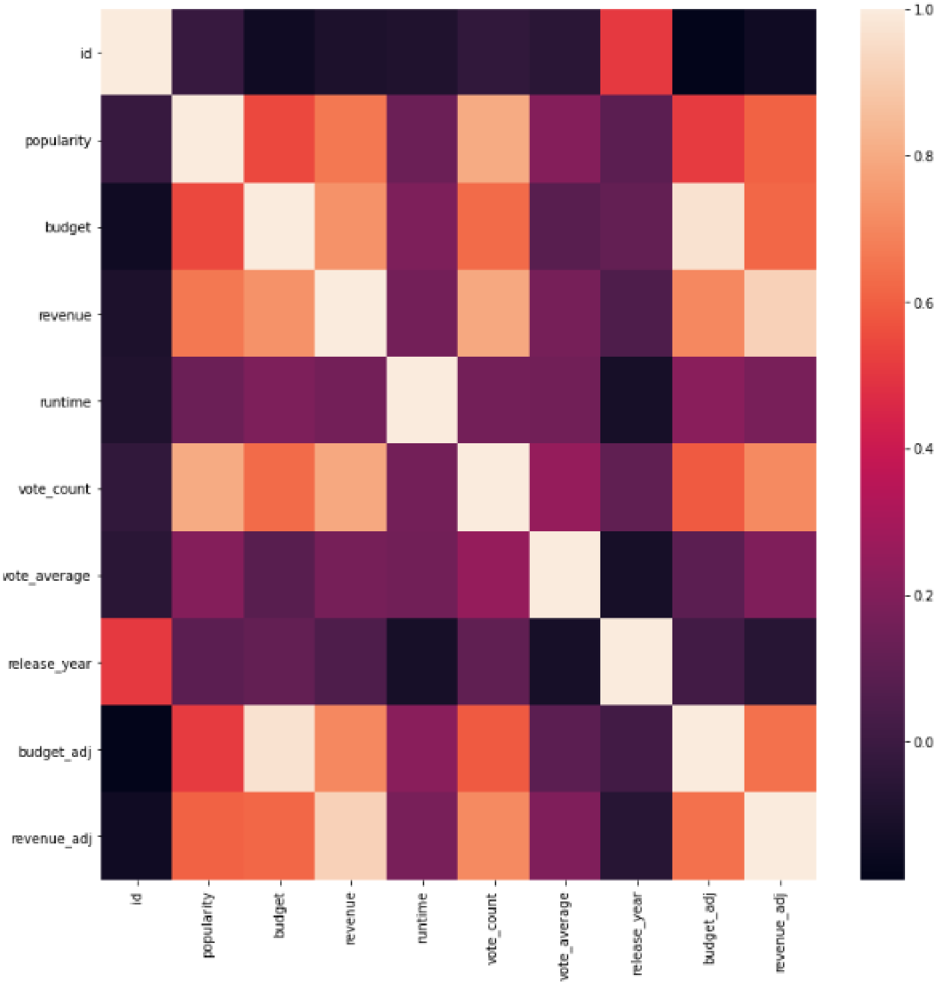


```
                    id  popularity    budget   revenue   runtime  vote_count  \
id            1.000000   -0.014350 -0.141351 -0.099227 -0.088360   -0.035551
popularity   -0.014350    1.000000  0.545472  0.663358  0.139033    0.800828
budget       -0.141351    0.545472  1.000000  0.734901  0.191283    0.632702
revenue      -0.099227    0.663358  0.734901  1.000000  0.162838    0.791175
runtime      -0.088360    0.139033  0.191283  0.162838  1.000000    0.163278
vote_count   -0.035551    0.800828  0.632702  0.791175  0.163278    1.000000
vote_average -0.058363    0.209511  0.081014  0.172564  0.156835    0.253823
release_year  0.511364    0.089801  0.115931  0.057048 -0.117204    0.107948
budget_adj   -0.189015    0.513550  0.968963  0.706427  0.221114    0.587051
revenue_adj  -0.138477    0.609083  0.622505  0.919110  0.175676    0.707942

              vote_average  release_year  budget_adj  revenue_adj
id               -0.058363      0.511364   -0.189015    -0.138477
popularity        0.209511      0.089801    0.513550     0.609083
budget            0.081014      0.115931    0.968963     0.622505
revenue           0.172564      0.057048    0.706427     0.919110
runtime           0.156835     -0.117204    0.221114     0.175676
vote_count        0.253823      0.107948    0.587051     0.707942
vote_average      1.000000     -0.117632    0.093039     0.193085
release_year     -0.117632      1.000000    0.016793    -0.066256
budget_adj        0.093039      0.016793    1.000000     0.646607
revenue_adj       0.193085     -0.066256    0.646607     1.000000

Out[81]: id              1
         popularity      5
         budget          5
         revenue         5
         runtime         0
         vote_count      5
         vote_average    0
         release_year    1
         budget_adj      5
         revenue_adj     5
         dtype: int64
```

Code for graphical representation:

```
import matplotlib.pyplot as plt

import seaborn as sns

corr=df.corr()

f,ax=plt.subplots(figsize=(12,12))

sns.heatmap(corr,vmax=1)

plt.show()
```

Output using graph:



Q. Are any of the pre-processing techniques needed: dimensionality reduction, range transformation, standardization, etc.?

Yes, we need to pre-process data since we have a lot many inconsistencies in dataset. By preprocessing data, we **make it easier to interpret and use**.

This process eliminates inconsistencies or duplicates in data, which can otherwise negatively affect a model's accuracy. Dimensionality reduction can be used in order to process the data

---

Q. Does PCA help visualize the data?

Principal component analysis (PCA) is an unsupervised machine learning technique. Perhaps the most popular use of principal component analysis is dimensionality reduction. Besides using PCA as a data preparation technique, we can also use it to help visualize data.

Q. Do we get any insights from histograms/ bar charts/ line plots, etc.?

Every new visualization is likely to give us some insights into our data. Some of those insights might be already known (but perhaps not yet proven) while other insights might be completely new or even surprising to us. Some new insights might mean the beginning of a story, while others could just be the result of errors in the data, which are most likely to be found by visualizing the data.