```
from google.colab import files
files.upload()
```

```
Choose Files No file chosen  Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving credits.csv to credits.csv
Saving movies.csv to movies.csv
```

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import json
```

```
movies = pd.read_csv('movies.csv',encoding='latin-1')
credits = pd.read_csv('credits.csv',encoding='latin-1')
```

```
movies.head()
```


	budget	genres	homepage	id	keywords	origi
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	

```
movies['genres'][0]
```

```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

```
json.loads(movies['genres'][0])
```

```
[{'id': 28, 'name': 'Action'},
 {'id': 12, 'name': 'Adventure'},
 {'id': 14, 'name': 'Fantasy'},
 {'id': 878, 'name': 'Science Fiction'}]
```

```
credits.head()
```

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James	[{"credit_id": "54805967c3a36829b5002c41"

```
movies_credits = pd.merge(movies,credits,left_on="id",right_on="movie_id")
movies_credits.head()
```


	budget	genres	homepage	id	keywords	origi
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	

5 rows × 24 columns

```
movies_credits.dropna(inplace=True)
```

```
movies_credits.head()
```

	budget	genres	homepage	id	keywords	origi
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	

5 rows × 24 columns

```
movies_credits['genres'][0]
```

```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fa
ntasy"}, {"id": 878, "name": "Science Fiction"}]'
```

```
def extract_func(lst):
  new_list = []
  values = json.loads(lst)
  for value in values:
    new_list.append(value['name'])
  return new_list
```



```
movies_credits['genres'] = movies_credits['genres'].apply(extract_func)
movies_credits['genres'].head()
```

```
0    [Action, Adventure, Fantasy, Science Fiction]
1                     [Adventure, Fantasy, Action]
2                       [Action, Adventure, Crime]
3                  [Action, Crime, Drama, Thriller]
4              [Action, Adventure, Science Fiction]
Name: genres, dtype: object
```

```
movies_credits['keywords'] = movies_credits['keywords'].apply(extract_func)
movies_credits['keywords'].head()
```

```
0    [culture clash, future, space war, space colon...
1    [ocean, drug abuse, exotic island, east india ...
2    [spy, based on novel, secret agent, sequel, mi...
3    [dc comics, crime fighter, terrorist, secret i...
4    [based on novel, mars, medallion, space travel...
Name: keywords, dtype: object
```

```
movies_credits.columns
```

```
Index(['budget', 'genres', 'homepage', 'id', 'keywords', 'original_language',
       'original_title', 'overview', 'popularity', 'production_companies',
       'production_countries', 'release_date', 'revenue', 'runtime',
       'spoken_languages', 'status', 'tagline', 'title_x', 'vote_average',
       'vote_count', 'movie_id', 'title_y', 'cast', 'crew'],
      dtype='object')
```

```
movies_credits['cast'] = movies_credits['cast'].apply(extract_func).apply(lambda x:x[:3])
movies_credits['cast'].head()
```

```
0    [Sam Worthington, Zoe Saldana, Sigourney Weaver]
1       [Johnny Depp, Orlando Bloom, Keira Knightley]
2        [Daniel Craig, Christoph Waltz, Léa Seydoux]
3        [Christian Bale, Michael Caine, Gary Oldman]
4     [Taylor Kitsch, Lynn Collins, Samantha Morton]
Name: cast, dtype: object
```

```
def extract_dir_func(lst):
  new_list = []
  values = json.loads(lst)
  for value in values:
    if value['job'] == 'Director':
      new_list.append(value['name'])
  return new_list
```

```
movies_credits['crew'][0]
```

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id":
1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e30
01f87", "department": "Art", "gender": 2, "id": 496, "job": "Production Design", "nam
e": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "g
ender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boyes"}, {"credit_i
d": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job":
"Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a368
10c9002101", "department": "Production", "gender": 1, "id": 1262, "job": "Casting", "na
```

```
movies_credits['crew'] = movies_credits['crew'].apply(extract_dir_func)
movies_credits['crew'].head()
```

```
0       [James Cameron]
1      [Gore Verbinski]
2          [Sam Mendes]
3    [Christopher Nolan]
4       [Andrew Stanton]
Name: crew, dtype: object
```

```
def collapse(lst):
  new_list = []
  for i in lst:
    new_list.append(i.replace(" ",""))
  return new_list
```

```
movies_credits['genres'] = movies_credits['genres'].apply(collapse)
movies_credits['genres'].head()
```

```
0    [Action, Adventure, Fantasy, ScienceFiction]
1                   [Adventure, Fantasy, Action]
2                     [Action, Adventure, Crime]
3                [Action, Crime, Drama, Thriller]
4             [Action, Adventure, ScienceFiction]
Name: genres, dtype: object
```

```
movies_credits['keywords'] = movies_credits['keywords'].apply(collapse)
movies_credits['keywords'].head()
```

```
0    [cultureclash, future, spacewar, spacecolony, ...
1    [ocean, drugabuse, exoticisland, eastindiatrad...
2    [spy, basedonnovel, secretagent, sequel, mi6, ...
3    [dccomics, crimefighter, terrorist, secretiden...
4    [basedonnovel, mars, medallion, spacetravel, p...
Name: keywords, dtype: object
```

```
movies_credits['cast'] = movies_credits['cast'].apply(collapse)
movies_credits['cast'].head()
```

```
0    [SamWorthington, ZoeSaldana, SigourneyWeaver]
1       [JohnnyDepp, OrlandoBloom, KeiraKnightley]
2        [DanielCraig, ChristophWaltz, LéaSeydoux]
```



```
3        [ChristianBale, MichaelCaine, GaryOldman]
4     [TaylorKitsch, LynnCollins, SamanthaMorton]
Name: cast, dtype: object
```

```
movies_credits['crew'] = movies_credits['crew'].apply(collapse)
movies_credits['crew'].head()
```

```
0       [JamesCameron]
1      [GoreVerbinski]
2          [SamMendes]
3    [ChristopherNolan]
4       [AndrewStanton]
Name: crew, dtype: object
```

```
movies_credits['overview'] = movies_credits['overview'].str.split()
movies_credits['overview'].head()
```

```
0    [In, the, 22nd, century,, a, paraplegic, Marin...
1    [Captain, Barbossa,, long, believed, to, be, d...
2    [A, cryptic, message, from, Bondâ, s, past, se...
3    [Following, the, death, of, District, Attorney...
4    [John, Carter, is, a, war-weary,, former, mili...
Name: overview, dtype: object
```

```
movies_credits['tags'] = movies_credits['overview'] + movies_credits['genres'] + movies_cred
movies_credits['tags'][0]
```

```
['In',
 'the',
 '22nd',
 'century,',
 'a',
 'paraplegic',
 'Marine',
 'is',
 'dispatched',
 'to',
 'the',
 'moon',
 'Pandora',
 'on',
 'a',
 'unique',
 'mission,',
 'but',
 'becomes',
 'torn',
 'between',
 'following',
 'orders',
 'and',
 'protecting',
 'an',
```

```
        'alien',
        'civilization.',
        'Action',
        'Adventure',
        'Fantasy',
        'ScienceFiction',
        'cultureclash',
        'future',
        'spacewar',
        'spacecolony',
        'society',
        'spacetravel',
        'futuristic',
        'romance',
        'space',
        'alien',
        'tribe',
```