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

```
movies=pd.read_csv('/tmdb_5000_movies.csv')
credits=pd.read_csv('/tmdb_5000_credits.csv')
```

```
movies.head()
```

	budget	genres	homepage	id	keywords	o
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":...	

Next steps: Generate code with `movies` | View recommended plots | New interactive sheet

```
1 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 Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight	[{"cast_id": 2, "character": "Bruce	[{"credit_id": "52fe4781c3a36847f81398c3",

Next steps: Generate code with credits | View recommended plots | New interactive sheet

```
1 credits.head(1)['cast'].values
```

```
"Jason Whyte", "order": 13}, {"cast_id": 34, "character": "Venture Star Crew
Chief", "credit_id": "52fe48009251416c750aca63", "gender": 2, "id": 42317,
"name": "Scott Lawrence", "order": 14}, {"cast_id": 35, "character": "Lock Up
Trooper", "credit_id": "52fe48009251416c750aca67", "gender": 2, "id": 986734,
"name": "Kelly Kilgour", "order": 15}, {"cast_id": 36, "character": "Shuttle
```

```
"Basketball Avatar", "credit_id": "52fe48009251416c750acaa3", "gender": 0, "id":
89714, "name": "Ilram Choi", "order": 30}, {"cast_id": 52, "character": "Na\'vi
Child", "credit_id": "52fe48009251416c750acaa7", "gender": 0, "id": 1207249,
"name": "Kyla Warren", "order": 31}, {"cast_id": 53, "character": "Troupe",
"credit_id": "52fe48009251416c750acaab", "gender": 0, "id": 1207250, "name":
"Lisa Roumain", "order": 32}, {"cast_id": 54, "character": "Troupe",
"credit_id": "52fe48009251416c750acaaf", "gender": 1, "id": 83105, "name":
"Debra Wilson", "order": 33}, {"cast_id": 57, "character": "Troupe",
"credit_id": "52fe48009251416c750acabb", "gender": 0, "id": 1207253, "name":
"Chris Mala", "order": 34}, {"cast_id": 55, "character": "Troupe", "credit_id":
"52fe48009251416c750acab3", "gender": 0, "id": 1207251, "name": "Taylor Kibby",
"order": 35}, {"cast_id": 56, "character": "Troupe", "credit_id":
"52fe48009251416c750acab7", "gender": 0, "id": 1207252, "name": "Jodie Landau",
"order": 36}, {"cast_id": 58, "character": "Troupe", "credit_id":
"52fe48009251416c750acabf", "gender": 0, "id": 1207254, "name": "Julie Lamm",
"order": 37}, {"cast_id": 59, "character": "Troupe", "credit_id":
"52fe48009251416c750acac3", "gender": 0, "id": 1207257, "name": "Cullen B.
Madden", "order": 38}, {"cast_id": 60, "character": "Troupe", "credit_id":
"52fe48009251416c750acac7", "gender": 0, "id": 1207259, "name": "Joseph Brady
Madden", "order": 39}, {"cast_id": 61, "character": "Troupe", "credit_id":
"52fe48009251416c750acacb", "gender": 0, "id": 1207262, "name": "Frankie
Torres", "order": 40}, {"cast_id": 62, "character": "Troupe", "credit_id":
"52fe48009251416c750acacf", "gender": 1, "id": 1158600, "name": "Austin Wilson",
"order": 41}, {"cast_id": 63, "character": "Troupe", "credit_id":
"52fe48019251416c750acad3", "gender": 1, "id": 983705, "name": "Sara Wilson",
```

```
movies.merge(credits,on='title')
```

```
         budget              genres                                       homepage      i

   0  237000000     [{"id": 28,                           http://www.avatarmovie.com/   1999
                     "name":
                     "Action"}, {"id":
                     12, "nam...

   1  300000000     [{"id": 12,          http://disney.go.com/disneypictures/pirates/     28
                     "name":
                     "Adventure"},
                     {"id": 14, "...

   2  245000000     [{"id": 28,       http://www.sonypictures.com/movies/spectre/  20664
                     "name":
                     "Action"}, {"id":
                     12, "nam...

   3  250000000     [{"id": 28,                   http://www.thedarkknightrises.com/   4902
                     "name":
                     "Action"}, {"id":
                     80, "nam...

   4  260000000     [{"id": 28,                   http://movies.disney.com/john-carter   4952
                     "name":
                     "Action"}, {"id":
                     12, "nam...

 ...        ...             ...                                                ...      .

4804     220000     [{"id": 28,                                                NaN    936
                     "name":
                     "Action"}, {"id":
                     80, "nam...

4805       9000     [{"id": 35,                                                NaN   7276
                     "name":
                     "Comedy"},
                     {"id": 10749, "...

4806          0     [{"id": 35,      http://www.hallmarkchannel.com/signedsealeddel...  23161
                     "name":
                     "Comedy"},
                     {"id": 18, "nam...

4807          0              []                        http://shanghaicalling.com/  12618
```

4808	0	[{"id": 99, "name": "Documentary"}]	NaN	2597

4809 rows × 23 columns

```
Movies=movies.merge(credits,on='title')
```

```
 Movies['original_language'].value_counts()
```

	count
original_language	
en	4510
fr	70
es	32
zh	27
de	27
hi	19
ja	16
it	14
ko	12
cn	12
ru	11
pt	9
da	7
sv	5
nl	4
fa	4
th	3
he	3
ta	2
cs	2
ro	2
id	2
ar	2
vi	1
sl	1
ps	1
no	1
ky	1
hu	1
pl	1
af	1

```
nb    1
tr    1
is    1
xx    1
te    1
el    1

dtype: int64
```

```
movies=Movies[['movie_id','title','overview','genres','keyw
```

```
movies.head()
```

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	"52fe48009251416
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	"52fe4232c3a3684
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	"54805967c3a3682

Next steps: Generate code with movies | View recommended plots | New interactive sheet

```
movies.isnull().sum()
```

```
                0
  movie_id      0
     title      0
  overview      3
    genres      0
  keywords      0
      cast      0
      crew      0

dtype: int64
```

```python
movies.dropna(inplace=True)
```

```
<ipython-input-19-4d2d825739df>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies.dropna(inplace=True)
```

```python
movies.isnull().sum()
```

```
                0
  movie_id      0
     title      0
  overview      0
    genres      0
  keywords      0
      cast      0
      crew      0

dtype: int64
```

```python
movies.duplicated().sum()
```

```
0
```

```python
movies.iloc[0].genres
```

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

```
def convert(obj):
  l=[]
  for i in obj:
    l.append(i['name'])
  return l
```

```
import ast
ast.literal_eval('[{"id": 28, "name": "Action"}, {"id": 12,
```

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

```
def convert(obj):
  l=[]
  for i in ast.literal_eval(obj):
    l.append(i['name'])
  return l
```

```
movies['genres'].apply(convert)
```

	genres
0	[Action, Adventure, Fantasy, Science Fiction]
1	[Adventure, Fantasy, Action]
2	[Action, Adventure, Crime]
3	[Action, Crime, Drama, Thriller]
4	[Action, Adventure, Science Fiction]
...	...
4804	[Action, Crime, Thriller]
4805	[Comedy, Romance]
4806	[Comedy, Drama, Romance, TV Movie]
4807	[]
4808	[Documentary]

4806 rows × 1 columns

dtype: object

```
movies['genres']=movies['genres'].apply(convert)
```

```
<ipython-input-33-f70b3d855b31>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['genres']=movies['genres'].apply(convert)
```

```
movies.head()
```

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{" "52fe48009251416c7
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{" "52fe4232c3a36847f
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[Action, Adventure, Crime]	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{" "54805967c3a36829b

Next steps: Generate code with movies | View recommended plots | New interactive sheet

```
movies['keywords']=movies['keywords'].apply(convert)
```

```
<ipython-input-35-bfa1f23e8c85>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['keywords']=movies['keywords'].apply(convert)
```

```
#we need three values in cast
def convert3(text):
    L = []
    counter = 0
    for i in ast.literal_eval(text):
        if counter < 3:
            L.append(i['name'])
        counter+=1
    return L
```

```
movies['cast']=movies['cast'].apply(convert3)
```

```
<ipython-input-37-f7f209af84d8>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['cast']=movies['cast'].apply(convert3)
```

```
movies['crew'][0]
```

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "i
d": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a3
6810e3001f87", "department": "Art", "gender": 2, "id": 496, "job": "Production Desig
n", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department":
"Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boye
s"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0,
"id": 900, "job": "Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit
_id": "539c4a4cc3a36810c9002101", "department": "Production", "gender": 1, "id": 126
2, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc",
```

```
def fetch_director(text):
    L = []
    for i in ast.literal_eval(text):
        if i['job'] == 'Director':
            L.append(i['name'])
    return L
```

```
movies['crew'].apply(fetch_director)
```

```
                                            crew
   0                               [James Cameron]
   1                               [Gore Verbinski]
   2                                  [Sam Mendes]
   3                           [Christopher Nolan]
   4                              [Andrew Stanton]
  ...                                          ...
4804                            [Robert Rodriguez]
4805                                [Edward Burns]
4806                                 [Scott Smith]
4807                                 [Daniel Hsia]
4808  [Brian Herzlinger, Jon Gunn, Brett Winn]

4806 rows × 1 columns

dtype: object
```

```python
movies['crew']=movies['crew'].apply(fetch_director)
```

```
<ipython-input-41-d998f4cfc70b>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['crew']=movies['crew'].apply(fetch_director)
```

```python
movies.head()
```

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]
			A cryptic message		[spy, based on	[Daniel	

Next steps: Generate code with movies | View recommended plots | New interactive sheet

```python
def collapse(L):
    L1 = []
    for i in L:
        L1.append(i.replace(" ",""))
    return L1
```

```python
movies['cast'] = movies['cast'].apply(collapse)
movies['crew'] = movies['crew'].apply(collapse)
movies['genres'] = movies['genres'].apply(collapse)
movies['keywords'] = movies['keywords'].apply(collapse)
```

```
<ipython-input-44-aea3176ce85e>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['cast'] = movies['cast'].apply(collapse)
<ipython-input-44-aea3176ce85e>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['crew'] = movies['crew'].apply(collapse)
<ipython-input-44-aea3176ce85e>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['genres'] = movies['genres'].apply(collapse)
```

```
<ipython-input-44-aea3176ce85e>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['keywords'] = movies['keywords'].apply(collapse)
```

```
movies['overview'] = movies['overview'].apply(lambda x:x.sp
```

```
<ipython-input-45-3bf62826a184>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['overview'] = movies['overview'].apply(lambda x:x.split())
```

```
movies['tags'] = movies['overview'] + movies['genres'] + mc
```

```
<ipython-input-46-d64bc4c70271>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
  movies['tags'] = movies['overview'] + movies['genres'] + movies['keywords'] + movi
```

```
new = movies.drop(columns=['overview','genres','keywords','
```

```
new['tags'] = new['tags'].apply(lambda x: " ".join(x))
```

```
new.head()
```

	movie_id	title	tags
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...